An Extended Investigation of the Similarity Between Privacy Policies of Social Networking Sites as a Precursor for Standardization

Emma Cradock[1], David Millard[1] and Sophie Stalla-Bourdillon[1]

[1] *University of Southampton, UK*

ABSTRACT

Privacy policies are unsatisfactory in communicating information to users. Social networking sites (SNS) exemplify this, attracting growing concerns regarding their use of personal data, whilst lacking incentives to improve their policies. Standardization addresses many of these issues, but is only possible if policies share attributes that can be standardized. This investigation assessed the similarity of two attributes (the clauses and the coverage of forty recommendations made by the UK Information Commissioner) between the privacy policies of the six most frequently visited SNS globally. Similarity was also investigated by looking at whether any recommendations were not addressed by all SNS and if there were any themes of information discussed in the policies, but not included in the ICO Code. We found that similarity in the clauses was low, yet similarity in the recommendations covered was high. This indicates that SNS use different clauses, but to convey similar information. There were a number of recommendations which none of the SNS addressed. There were also four themes of information which all six SNS addressed, which were not recommended in the ICO Code. This paper proposes the policies of SNS already share attributes, indicating the feasibility of standardization at a thematic level currently. Five recommendations are made to begin facilitating this.

Keywords: Privacy Policies, Standardization 2, Web Science, Data Protection, Social Networking Sites

ISSN 2332-4031; DOI 10.1561/106.00000006

1 Introduction

Technological and computational advancements have facilitated the processing of personal data on a scale like never before. This dramatic increase in the potential of personal data (Rowland *et al.*, 2012) has led to concerns over its effect on the privacy of individuals. 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 (Rowland *et al.*, 2012). One right given to individuals under the European Union (EU) Data Protection Directive (DPD) (Parliament and Council., 1995) is the right to information about personal data processing. This has led to the adoption of privacy policies as the de facto means of compliance. However, despite the many benefits of a well-executed privacy policy, their current role in making data processing transparent has been heavily criticized (McDonald and Cranor, 2008). The growing concern SNS attract regarding their use of personal data, and its effect on user privacy (Anderson, 2009), typifies the need for more informative policies. Policies need to create transparency and at a minimum, effectively communicate the information that individuals are legally entitled to. However, with individuals continuing to use their services, SNS lack the incentives to make improvements. This is exacerbated by the difficulties which creating a concise and legally compliant policy entail. As a suggestion for improvement, standardizing the privacy policies of SNS addresses many of these issues, and begins the groundwork for further improvements (Cranor, 2012). However, standardization is only possible if policies share attributes on which standards can be built. Using thematic analysis [8] and cross-document structure theory (Aleixo and Pardo, 2008) our research investigates the similarity of the policies of SNS, to answer the following five research questions:

1. What is the similarity between the privacy policies of the top six SNS globally, in the clauses they use?
2. 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)?
3. Are there any recommendations of the ICO Code, which all privacy policies do not address?
4. Are there any themes of information addressed in all the privacy policies not included in these forty recommendations from the ICO Code?
5. To what extent 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 (Office., 2010). Whilst this paper focuses on EU and UK law (and policies of US-based companies), privacy policies are not a concept exclusive to these states. Indeed, the Organisation for Economic Co-operation and Development (OECD) (an international forum of 34 democracies) has identified 'openness' as a principle of good practice when processing personal data (Economic Cooperation and Development., 2013). Privacy policies are often the manifestation of this principle by its members, some of which have incorporated it into their national law. However, despite their common usage, the role of privacy policies in informing users is unsatisfactory. They have been heavily criticized for being long (McDonald and Cranor, 2008), legalistic, complex (Office., 2010) and ineffective in helping users understand their rights. The result is that they are not read, defeating their purpose. Arguably, organizations are complicit in this, as with individuals still using their services, they lack strong incentives to improve their policies. Even if incentivized, creating a concise and compliant policy is not easy, given the supranational nature of the web, where data is processed in numerous jurisdictions, each with differing requirements (Rowland *et al.*, 2012). However, if executed well, privacy policies can promote transparency and reduce information asymmetry [49] by communicating information that enables users to make effective privacy choices. Indeed, evidence suggests users are privacy aware and active (Hargittai *et al.*, 2010), just that they do not view privacy policies as a means of expressing consent [39].

2.2 Social Networking Sites

Various types of websites, from search engines to e-commerce sites, would benefit from improved privacy policies. However, for this investigation, we chose to select one 'type' of website to investigate. Social Networking Sites (SNS) were chosen as they are the second most frequently visited 'type' of website globally (behind search engines) (Alexa., 2014). 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 [44]. However, as businesses, the trade-off for free use is the data harvested from users, which can be monetized to support the provision of the service. Despite this trade-off, a 2011 survey [45] found that 72% of SNS users worry that they are giving away too much data online. Indeed, it is not just the data that users knowingly share, or the data observed without their awareness that SNS can access, but also the information that SNS can derive and infer about users from seemingly innocuous data. A 2013 study found that Facebook 'likes' (which were publicly available by default at the time, through Facebook's API) could be used to accurately predict a variety of attributes about an individual, including ethnic origin, religious beliefs and sexual orientation [31]. This issue has also been recognized by an OECD roundtable of 65 privacy experts from around the world. Experts were from governments, privacy enforcement authorities, academia, businesses and the Internet technical community. They acknowledged that 'increasing amounts of data are not collected from the individuals concerned, but are instead observed, derived and inferred' [36]. Therefore, although SNS rely on personal data, it is questionable how much data this entitles them to. Especially given the criticisms they have received regarding their collection and use of personal data [3]. Thus, SNS are a prime example of the need for more informative policies. Improving their ability to communicate information about the collection and use of personal data would support users in making better judgments on how much information they are willing to give away for use of these services.

2.3 Suggested Improvements

The question then, is how best to improve privacy policies, and various suggestions have been made. Becker et al (Becker *et al.*, 2014) looked at using visualizations for certain aspects of the policy, to aid communication and improve trust. Another suggestion is taking an approach similar to the creative commons model for intellectual property rights licences (Robinson *et al.*, 2009). Alternatively, a technical approach could be taken, such as making policies machine-readable, as was the aim of the Platform for Privacy Preferences Project (P3P) [10]. Developed and officially recommended by W3C in 2002, P3P 1.0 provides a markup language for websites to use to encode their natural language privacy statements into a machine-readable XML format (Olurin *et al.*, 2012). The P3P User agent was then able to translate a website's P3P policy into a human-readable format and check the policy against user preferences. If there was a conflict, warnings were provided (Olurin *et al.*, 2012). However, P3P only achieved limited adoption (Olurin *et al.*, 2012) due to its complexity (Schwartz, 2009), and the lack of industry participation (Cranor, 2012). This led to closure of the P3P working group in 2006 (Cranor, 2012). Yet, some believe P3P-based techniques have considerable potential, with the challenge being to design formalized privacy policy languages (Olurin *et al.*, 2012).

2.4 Standardization

Thus, as a suggestion for improvement, standardization has the most potential here, offering benefits to various stakeholders as well as beginning the groundwork for other improvements. Indeed, a United States of America (US) Federal Trade Commission (FTC) report (Commission., 2010) called for privacy policies to be clearer, shorter and more standardized. Benefits of standardization to users include the facilitation of comparisons between different policies, increasing familiarity with terminology, and the location of particular information (Cranor, 2012). Studies have also shown that standardized presentations can have significant positive effects on a reader's enjoyment of privacy policies compared to non-standardized presentations [30]. Benefits for organizations include allowing them to verify their compliance with the law (Cranor, 2012) and a reduction in the hassle of creating policies completely by themselves. Although these benefits may not incentivize organizations to improve their policies in the same way consumer demand would, they do reduce the deterrents or roadblocks to doing so. Standardization also allows for large-scale analysis of policies (Cranor *et al.*, 2013). This allows regulators s compliance, gain a better understanding of policies, and to move away from human annotation, which is currently required to understand and compare policies. Standardizing elements of policies also begins the groundwork for other suggested improvements. It begins the process of information reduction and refinement, which is required to develop formalized privacy policy languages (Olurin *et al.*, 2012), or standardized descriptions for a creative commons model approach [39]. However, to succeed, standardization requires policies to share attributes, upon which standards can be built. Given the fragmented evolution of the privacy policies of SNS, in their creation by different organizations, governed by differing jurisdictional legal requirements, the shared attributes required for standardization may not be present. Thus, prior to attempting standardization, it is important to assess similarity of the data in question, to ascertain whether standardization is possible, at what level, and where the similarities and differences between the privacy policies lie.

2.5 Comparative Analysis of Privacy Policies

There is not a vast amount of literature on comparing privacy policies (of SNS or otherwise), but one study that did was by Wu et al (Wu *et al.*, 2010). Here, the researchers adapted a privacy taxonomy previously used for data storage policies, and extended it to SNS. They applied the taxonomy to the policies of six SNS (Facebook, LinkedIn, MySpace, Orkut, Twitter, and YouTube) to compare how the published policies protected user privacy in reality. Based on the taxonomy, Wu et al (Wu *et al.*, 2010) asserted that privacy policies are formed by four elements: purpose, visibility, granularity and retention (of the data). However, despite adaptation, this taxonomy was still primarily aimed at providing a means for thinking about data privacy technologically (and specifically for data repositories) (Barker *et al.*, 2009), opposed to legally. The taxonomy elements were created from principles of handling data from various sources, rather than from concrete legal requirements. Given that one of the main benefits of a standardized policy is to help organizations comply with the law, this is also likely to also be a huge incentive for the adoption of such a policy. Therefore, finding similarity with a complete set of elements indicating legal compliance is more appropriate for supporting conclusions about the potential for a legally compliant standardized privacy policy.

2.6 Information Commissioner's Office Code

The UK Information Commissioner's Office (ICO) is an independent authority, set up to uphold information rights in the public interest in the UK. In an effort to help make policies more informative, ICO has issued a 'Privacy notices code of practice' (Office., 2010) aimed at helping organizations 'collect and use information appropriately by drafting clear and genuinely informative privacy notices'. The ICO Code has been released as part of the Information Commissioner's role, as head of ICO, under section 51 of the UK Data Protection Act (DPA) (Britain., 1998). The DPA is the implementation of the EU Data Protection Directive (DPD) (Parliament and Council., 1995) in the UK. This section requires the Information Commissioner to promote good practice and empowers him/her (after consultation) to prepare such codes. The ICO Code provides recommendations, aimed at aiding those processing information in complying with the DPA (Britain., 1998). The ICO Code itself states that it can be used as a 'checklist to evaluate an existing privacy notice', which is why its recommendations were chosen as one of the attributes for this investigation. Comparing the policies for the presence of these recommendations is more appropriate for supporting conclusions about the potential for a legally compliant standardized policy, than comparing them for the elements used in the Wu et al. (Wu *et al.*, 2010) study alone. Whilst the four elements Wu et al. (Wu *et al.*, 2010) used are reflected in the ICO Code, the Code provides further, more specific recommendations, to aid compliance with UK (based on EU) law.

3 Methodology

3.1 Selection of the SNS

The most frequently visited SNS were chosen, as ranked by Alexa.com (Alexa., 2014), a web analytics website that publishes a global traffic rank for major websites. As these SNS are used the most, any improvement to their policies has the potential to benefit the most individuals. Alexa allows visitors to browse websites by 'category' and their category 'Social Networking' was used for the purpose of this investigation. Due to the time available in which to manually analyze the data, a limited number of SNS could be chosen. In deciding how many to investigate, we found that the top five SNS were also ranked in the top thirty of all websites globally. Whereas, the sixth ranked SNS (Flickr), was ranked 164th. Because there was such a steep drop in the popularity of the ranked SNS, from the 5th to the 6th (and onwards), it seemed rational (given the time available) to investigate the top five ranked SNS in addition to the 6th. This would account for any confounding variables that might be linked to popularity. The 7th ranked SNS onwards were therefore excluded from the investigation. As a result, the six SNS policies selected were: Facebook (FB) (Facebook., 2013), Twitter (T) [50], LinkedIn (L) (LinkedIn., 2014), Pinterest (P) (Pinterest., 2014), Google+ (G+) (Google., 2014), and Flickr (F) (Flickr., 2015). Their policies as available in August 2014 were analyzed.

Alexa (Alexa., 2014) does not provide exact metrics on the number of users worldwide, instead ranking sites by traffic estimates based on data from their global traffic panel. Their global traffic rank is a measure of how a website is doing relative to all other sites on the web over the past three months. It is calculated using a proprietary methodology that combines a site's estimated average of daily unique visitors and its estimated number of pageviews over the past three months. Indeed, finding comparable and exact numbers of SNS users worldwide is not easy. This is because the definition of 'users' can be broken down into subcategories, such as 'registered users' (i.e. people who simply 'have' an account) and 'active users' (i.e. people who have an account and use it). Furthermore, 'active users' can be defined using different parameters. Statista

SNS	Alexa Rank	Year of Release	Headquarters Location
FB	1st	2004	Menlo Park, California, US
T	2nd	2006	San Francisco, California, US
L	3rd	2003	Mountainview, California, US
P	4th	2010	San Francisco, California, US
G+	5th	2011	Mountainview, California, US
F	6th	2004	San Francisco, California, US

Table 1: Details of SNS

(statista., 2016) detail the number of active users worldwide as of January 2016 for four of the SNS investigated: Facebook (1550 million); Twitter (320 million); LinkedIn (100 million); and Pinterest (100 million). However, they do not provide numbers for Google+ and Flickr. Finding user numbers for these for the same period and using the same 'user' definition is difficult. Flickr stated in June 2015 [20] that they have a 'community of more than 112 million photographers'. However, this figure is not for the same period, and also does not say how it was measured. If it is the number of registered accounts, it is not comparable to statista's figures. Even if it is for 'active users', it cannot be deemed comparable without details of the parameters used to calculate this, which are not listed. The effect of how 'user' is defined is exemplified in the context of Google+. This is because for every Google account created, a Google+ profile is created automatically. Thus, counting 'registered users' could reflect even less about the user base of Google+ than other SNS. Indeed, a digital marketing firm (Consulting., 2015) analyzed 516,246 randomly selected Google+ profiles and found that 90.1% of these had never posted anything on the service.

3.2 Extent of the Data

Online privacy policies often take a layered approach and use hyperlinks to link to further explanatory information. For example, within their policy, LinkedIn stated that 'You may choose the parts of your profile that search engines index or completely opt out of this feature in your LinkedIn account settings'. When clicked, 'settings' redirects the user to their account settings. For this investigation, we had to decide whether the content we assessed for similarity would extend to the information contained on the pages following these hyperlinks (second layer). At the familiarization stage of the investigation (Stage 1), we examined the hyperlinks. However, we found that they were generally links to: more advice on the topics discussed (from the SNS itself and outside sources); links to account settings and other pages on the website; other SNS policies; other SNS services; and online contact forms. Furthermore, ICO have stated that when using a layered approach, the first layer of a privacy policy should contain the 'key privacy information' with 'more detailed information available elsewhere' (Office., 2016), opposed to new information. Therefore, we took a pragmatic approach and examined only the 'first layer' of the policies i.e. the content on the page when their 'privacy policy' or 'privacy' links were clicked.

3.3 Attributes for Comparison

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, the clauses used by the SNS proved appropriate, as they convey enough information to make comparison meaningful. A 'clause' is defined as 'a part of a treaty, law or contract' (Press., 2011). 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 upon which the potential for standardization could be assessed. For comparison, a second attribute was also measured for similarity, the coverage of forty recommendations from the ICO 'Privacy notices code of practice' (Office., 2010).

Because the world is divided into legal jurisdictions, we wanted to select one jurisdiction from which we could extract an appropriate subset of legal recommendations. The EU proved interesting, because of its single law (DPD) aimed at harmonizing data protection laws throughout EU member states. This law means that findings in relation to one EU country should be more generalizable to other countries within the EU, than to those outside. However, implementations of the DPD sometimes differ between member states, who decide the means to achieve the DPD's aims. In fact, the implementations of the DPD's information obligation has varied significantly in terms of what information should be provided, in what form, and at what time (Office., 2010). Because of this, we decided to pick one member state's legislation to examine. Whilst any could have been chosen, the UK seemed appropriate. This was because the researchers were familiar with its legislation and aware that ICO had produced specific guidance aimed at creating legally compliant privacy policies. Furthermore, comparing the US privacy policies with legal requirements from a jurisdiction where they are not headquartered provides for an interesting juxtaposition. Conclusions about the possibility of a global standardized policy will be strengthened where similarity between policies originating from the US and recommendations for compliance based on UK and EU law can be found. Because the UK and US are both members of the OECD, who have identified 'openness' as a principle of good practice (Economic Cooperation and Development., 2013), they could be predisposed to similarities. However, both have very different policy contexts. The US currently has no single comprehensive federal (national) law regulating data protection. Whilst, the US FTC's Fair Information Practice Principles (Commission., 2000) have significantly shaped how privacy policies are written by US web companies, specific guidance on policies equivalent to ICO's (in stating exactly what they should contain) is often sector-based. For example, the model privacy form (Securities and Commission., 2009) for compliance with the US Gramm-Leach-Bliley Act is aimed at financial institutions, as this is whom the Act regulates. These differences are another reason why investigating the similarities is interesting, but also why ICO's sector neutral guidance is appropriate. As the UK was the chosen, the ICO Code provided the most appropriate set of recommendations for this investigation. As the independent body charged with upholding information rights in the UK, ICO's guidance should lead to compliance with UK law.

3.4 Measuring Similarity

A combination of thematic analysis and cross-document structure theory (CST) were used. Thematic analysis (Braun and Clarke, 2006) is used to pinpoint, examine and record themes within data, and occurs in the six standard stages of thematic analysis outlined below. More detail is included in each stage, to discuss how these stages were adapted for this investigation. Cross-document structure theory (CST) (Aleixo and Pardo, 2008) is a formal discourse theory for multi-document analysis, which establishes relationships among segments of different documents about the same topic. CST has two classification scenarios, binary and full. Binary classification is simply interested in the existence of cross-document relations, regardless of type. Whereas, full classification cares about the type of cross-document relationship. For this investigation, only binary classification was completed. The coding and analysis in Stages 1-5 was completed by the primary researcher only, based on a framework agreed by all researchers. The coding and analysis was then discussed with the other researchers to produce the Stage 6 Final Report. 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 (Jaccard, 1912).

3.5 The Six Stages of Thematic Analysis

Stage 1: Familiarization with data: Here researchers immerse themselves in the data, gaining familiarity with its depth and breadth (Braun and Clarke, 2006). Thus, the policies were read multiple times, first passively then actively, recognizing meanings and patterns to support the subsequent phases of analysis.

Stage 2: Generating Initial Codes: This phase involved the production of initial 'codes' from the data. 'Codes' are defined as *'the most basic segment of the raw data that can be assessed in a meaningful way regarding the phenomenon'* (Boyatzis, 1998). Unlike some legal documents, which are broken down into numbered clauses, privacy policies are only broken into sections, which meant that the clauses had to be identified for the purpose of this investigation. 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 initially 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 (Ryan and Bernard, 2003), similar to the thematic analysis 'compare and contrast' approach (Glazer, 1978). 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 the policies had been worked through and all sentence pairs compared, the table containing the allocated clauses was repeatedly checked, until no more clauses were moved, known as achieving theoretical saturation (Strauss, Corbin, *et al.*, 1990). 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. Also 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's policy 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. The logic of removing these was to normalize the data. Including them could inflate the number of clauses some SNS had and skew the results. Jaccard's similarity coefficient was then used to measure similarity between the clauses for the answer to research question 1.

Stage 3: Searching for Themes Among Codes: This phase re-focuses the analysis at broader themes, and here involved sorting the clauses into potential themes (Braun and Clarke, 2006). Rather than coding inductively (initially) to produce themes from the clauses, forty ICO Code (Office., 2010) recommendations were used as themes, into which the data was placed for the purposes of research questions 1-3. The ICO 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 (using the processes in Stages 1 and 2). Although forty-six were identified, six were too broad or vague to assess in the context of this investigation e.g. *'Any further information necessary, in the specific circumstances, to enable the processing in respect of the individual to be fair'*. These six 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 code recommendations, or into a category of 'miscellaneous'. To create an objective description of what it meant to 'address' a Code recommendation, we created a table containing the recommendation and a definition for thematic coding. Table 2 shows three examples of Code recommendations and their definitions, which formed the thematic codes.

As the focus of this investigation was to see whether the policies contained clauses **addressing** the recommendation, we did not investigate whether the SNS was legally complying with them. 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 from LinkedIn’s 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 had included information in its policy ‘addressing’ the recommendation, it would take further investigation (outside the scope of this paper) to assess whether the assurances obtained are legally compliant. There are two reasons why this type of in-depth legal analysis is outside the scope of this study. Firstly, because of the time it would take to complete such an analysis on compliance with each of the forty recommendations. Secondly, because that was not the aim of this study, which is to assess similarity as a potential for standardization. Future work will explore areas identified in this study upon which further in-depth legal analysis would be beneficial.

The reason we chose not to code inductively (and produce themes from the clauses) for the first part of this stage is that the purpose here was to assess similarity with the long-term aim of creating a standardized policy. Because one of the main benefits of a standardized policy for organizations is to help them comply with the law, this is likely to be a huge incentive in their adoption of such a policy. Thus, a standardized policy will be impotent unless it aids organizations in compliance with the law. So whilst generating themes from the policies may be useful at later stages (as discussed in Section 6), here we wanted to see how well the recommendations could be used as a framework for the policies in their current form. It could be argued that coding the clauses into the recommendations risks creating a circular argument in the research design. Ergo, because the aim of the recommendations is to address the law, you would expect to find them in the policies. However, as studies (Van Alsenoy *et al.*, 2015), experience and indeed our results in Section 4.3 show, this is not always the case. Furthermore, the juxtaposition of US policies being compared with UK (based on EU) law may also mean this is not the case. Thus, by coding the clauses into ICO Code recommendations, we can identify any recommendations warranting further investigation, due to their lack of presence in the policies. Once this coding was completed, the ‘miscellaneous’ category contained a number of clauses, which could not be allocated to an ICO Code recommendation. We then coded inductively on these clauses. This identified themes that the policies included, above and beyond the ICO Code recommendations. This provided the answer to research question 4. These results identified areas for future work, to understand why these were not present in the ICO Code, and what the implications of these exceptions are. Jaccard’s similarity coefficient was then used to measure similarity between the policies for the answer to research question 2.

ICO Code Recommendations	Definition For Thematic Coding	Example of Clause
Tell people how long you or other organisations intend to keep the data.	The privacy policy refers to how long it (or organisations 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**

Table 2: Examples of ICO Code Recommendations

Stage 4: Reviewing Themes: This stage involves two levels. Level one involves reading the collated clauses for each theme and considering whether they form a coherent pattern (Braun and Clarke, 2006). 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 (Braun and Clarke, 2006). Because the themes used here were pre-determined from the ICO Code, at this stage each clause that had been allocated to a recommendation was checked for coherence. As were the themes generated from the ‘miscellaneous’ clauses.

Stage 5: Defining and Naming Themes: This stage names themes and paraphrases their content, clearly defining what themes are, and what they are not (Braun and Clarke, 2006). As the themes were pre-determined recommendations, they were simply named as their recommendation in full. The thematic codes we had already produced (examples provided in Table 2) were used as their definitions. For the themes generated from the 'miscellaneous' clauses', these were named and defined, and are discussed below in Section 4.4.

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 3 displays the results from Stage 2, and shows 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 removed, indicating that increased policy length did not necessitate repetition. Table 3 also shows that the descending order of SNS in terms of number of clauses identified and number of clauses remaining, stays the same (Facebook, LinkedIn, Google+, Twitter, Pinterest, Flickr). However, the order varies in terms of the number and percentage of clauses removed. Flickr in particular had just over a quarter of clauses removed. Given only 89 were identified initially (the lowest), this is a significant amount.

	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

Table 3: Number of Clauses identified, removed and remaining

4.1 Similarity in Clause Coverage

Figure 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 (Clauset *et al.*, 2009). In answering research question 1 (Section 1), Table 4 shows that the similarity between the SNS in the clauses they use was low, with the range between 8-27%. 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%. Interestingly, Table 3 shows that Flickr and Facebook were at separate ends of the continuum in terms of number of clauses identified, with Facebook having the most and Flickr the least. This may explain their dissimilarity. Whereas, Table 3 shows that Pinterest and Twitter would sit next to each other on this continuum, with a similar number of clauses. This may indicate why they have a higher similarity.

The investigation highlighted three prominent reasons for differences between SNS in the clauses they used:

1. **Functionality:** Differences in the functionality offered between SNS, such as LinkedIn's use of Polls and Facebook's 'Instant Personalization', resulted in a number of clauses communicating information about these. Other SNS would not include these clauses within their policy, because they do not offer the functionality. Therefore, they would not need to communicate information about these.

2. **Semantics:** Different words were often used between policies to discuss the same topics, but without being defined. For example, when discussing the termination of an account, the words 'close', 'delete' and 'deactivate' were all used across different policies. Whilst Facebook confirmed 'delete' meant permanent deletion, Pinterest only stated users had the ability to 'close your account at any time'. Without defining what 'close' meant, it was difficult to ascertain whether the clauses were comparable, meaning that they had to be treated as different.

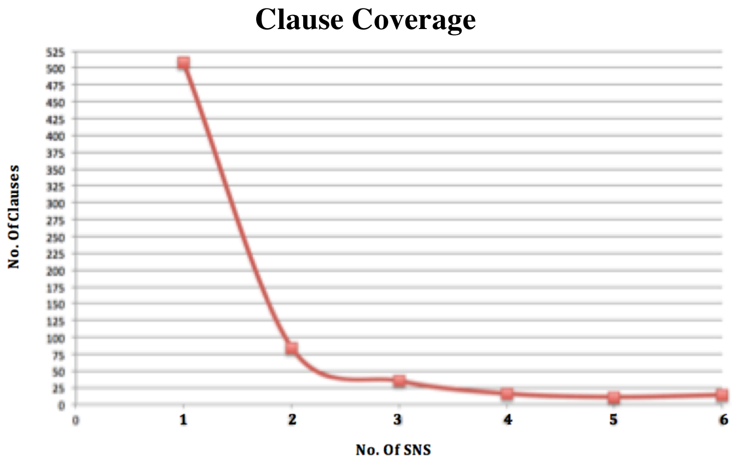
3. **Elaboration:** Some SNS elaborated on certain topics more than others. For example, although all SNS included a link to follow if users had any questions, comments or complaints, some also included their physical address and information regarding the complaints/question procedure. Equally, SNS provided 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.

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+)	■	■	■	■	■	■

Table 4: Jaccard Similarity of Clause Coverage

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.

SNS in the specific recommendations they addressed, than their overall similarity with the ICO Code. These percentages corroborate Graph 2, that there were certain recommendations that SNS collectively did, or did not, address.



Graph 1: Shows number of clauses covered by SNS

Figure 1: Shows number of clauses covered by SNS

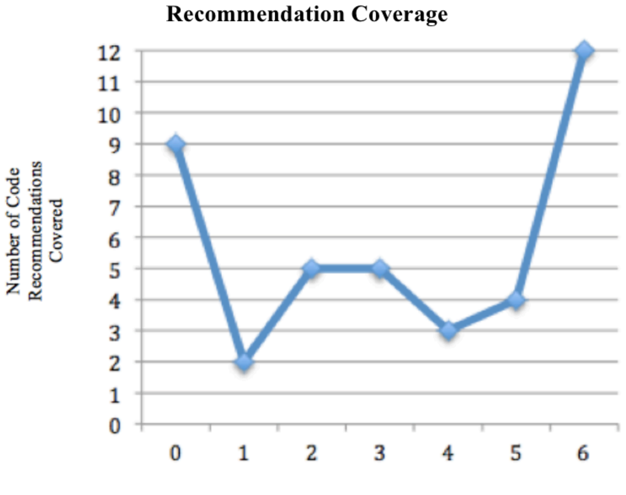


Figure 2: Shows number of recommendations covered by SNS

Graph 2: Shows number of recommendations covered by SNS

Table 4: 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+)						

Table 5: 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+)							

4.2 Similarity in Recommendation Coverage

Interestingly, Graph 2 shows that when looking at recommendation coverage, the largest percentages of recommendations covered, were for those covered by **none** (22.5%), or **all six** of the SNS (30%). These percentages show similarity between SNS in terms of the ICO Code recommendations that they do (and do not) address. Unlike Graph 1, there is no evidence of a power-law relationship in Graph 2 between the number of SNS and how many recommendations they address. Instead, the majority of recommendations were either addressed by all SNS, or none.

Table 5 shows the similarity of SNS with the Code, ranges from **45-65%**, which may be because the ICO Code

However, failing to address a recommendation and a lower similarity with the ICO 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 and thus the policy would not need to address it, or because the recommendation was applicable, but the SNS failed to address it. For example, none of the SNS 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 policies 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

4.2 Similarity in Recommendation Coverage

Interestingly, Figure 2 s hat when looking at recommendation coverage, the lar ercentages of recommendations covered, were for those ne (22.5%), or all six of the SNS (30%). These show similarity between SNS in terms of the ICO tions that they do (and do not) address. U 1 is no evidence of a power-law relationship 2 number of SNS and how many rec s. Instead, the majority of recomme by all SNS, or none.

Table 5 shows the similar 45-65%, which may be because the ICO Code recommendations are based on UK and EU law, whereas the SNS are based in the US. However, in answering research question 2 (Section 1), Table 5 shows that similarity, between the SNS themselves in addressing the code recommendations, ranges from 52%-81%. This evidences a higher percentage of similarity amongst the SNS in the specific recommendations they addressed, than their overall similarity with the ICO Code. These percentages corroborate Figure 2, 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 ICO 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 and thus the policy would not need to address it, or because the recommendation was applicable, but the SNS failed to address it. For example, none of the SNS 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 policies 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 address this in their policy. Which scenario is correct cannot be ascertained without further investigation and legal analysis, outside the scope of this paper, for the reasons explained above. Such analysis would also require access to information, which SNS often do not divulge in full, such as what data they collect. Table 5 also shows that the least-similar (with each other) in addressing ICO Code recommendations were jointly Facebook and Pinterest (52%) and LinkedIn and Pinterest (52%). The most similar were Facebook and LinkedIn (81%). Interestingly, Facebook and LinkedIn had the highest numbers of clauses (Table 3) and although Pinterest did not have the lowest, it did have the second lowest with only 21 more than Flickr, who had 60 (Table 3). 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 or lower similarity with the ICO Code should not be assumed to mean a less legally compliant policy.

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

Table 5: Jaccard

4.3 ICO Code Recommendations Not Present in Privacy Policies

Analysis identified nine ICO Code recommendations not addressed by any of the SNS. These were as follows:

Try and predict whether you will be likely to do things with it (the data) in future without drawing up a long

list of future possible uses if you are unlikely to use it for those purposes. LinkedIn and Facebook were the only SNS to touch on the future of the services, although not enough to address this recommendation. Facebook stated that: *'Granting us permission to use your information not only allows us to provide Facebook as it exists today, but it also allows us to provide you with innovative features and services we develop in the future that use the information we receive about you in new ways'*. Therefore, although this indicates that Facebook envisages there will be future uses of data, it does not specify what these may be, just that an individual has given permission for them. This is a particularly wide licence, lacking specificity on what these features may be. Facebook has a track record of such licences, with the Electronic Privacy Information Center (EPIC) filing a complaint with the US FTC about this (Center., 2015).

LinkedIn was more focused on the possibility that they would collect new types of information in the future, rather than put data they have collected to new uses, stating that: *'LinkedIn is a dynamic, innovative environment, which means we are always seeking to improve the services we offer you. We often introduce new features, some of which may result in the collection of new information (for example, when the Endorsements feature launched, we began collecting information about skills for which Members were endorsed and the individuals who endorsed them). Furthermore, new partnerships or corporate acquisitions may result in new features, and we may potentially collect new types of information'*.

About the right to complain to the Information Commissioner if there is a problem. None of the SNS mentioned the right to complain to the UK Information Commissioner. This is likely to be because this is the data protection authority for the UK, and it could be impractical for the SNS to list the equivalent for every county using its service. Furthermore, because of the aforementioned issues relating to legal jurisdiction on the web, SNS users may not have the right to complain to the Information Commissioner, or have an equivalent. Interestingly, both Facebook and LinkedIn do refer to California's 'Shine the Light law', which is only applicable to California residents. This may be because both are based in California, even though it would certainly not be the only law applicable.

Have separate notices aimed at different groups of individuals you deal with. Each SNS only provided one policy (although versions in different languages were often available). This may be because it only deals with one 'group' of individuals, as exactly what 'group' means is unclear. The example given in the ICO Code is that a local authority may use information about old age pensioners to administer free access to local leisure facilities. Or use information about shopkeepers to collect business taxes (Office., 2010). In this sense, this recommendation may not apply to SNS, because they may not have different uses for different user data, but use all user data in the same ways.

Where individuals are required by law to provide their personal details be open with people and explain clearly why their information is being collected and what it will be used for. None of the SNS addressed this. As discussed earlier, this may be because there is no information the SNS are required by law to obtain, or that there is, but that the SNS has not included the details of this within their policy.

In marketing contexts, when organisations ask for permission to share customer information with third parties e.g. companies in the same group, this should be backed up with more detail information such as the names of the companies involved for those who want it. This is certainly applicable to the SNS and all discussed sharing information with third parties, however the specific third parties were not detailed. It may be that this information was provided (and this recommendation addressed) on another layer of the policy accessible by following a hyperlink. Or that this is communicated at the time when permission is obtained (although most SNS simply used their privacy policy as the information provided prior to consent). Therefore, although it can be concluded that it was not addressed in the first layer of the policy, this information may have been available elsewhere. Although, the practice of listing this information elsewhere is questionable, given ICO's guidance on key information being in the first layer.

If an organisation intends to collect personal information with the intention of selling or renting it you should make it clear to individuals that the information they provide could be supplied to anyone and used for any purpose and tell them this when they provide their details; and

That if their information is rented, individuals are told that if the business is insolvent, bankrupt, being closed down or sold that their information will be returned to its owner.
None of the SNS addressed either of these. However, they all discussed the transfer of information if the business is insolvent etc. This may mean these recommendations do not apply because none of the SNS sell or rent information. However, with companies like GNIP (GNIP., 2016) selling access to social media data from various SNS, specifically listing that access to Twitter, Facebook, Flickr and Google+ data is available, this is unlikely. Thus, for these four SNS at least, these recommendations would appear to apply.

Avoid using confusing terminology e.g. technical language. Pinterest, LinkedIn and Google all stated at the beginning of their policy that they had tried to keep their policies as simple as possible. Pinterest acknowledged that some of their terms were a little technical, and Google advised that if readers were not familiar with terms like cookies etc. they should read about them first. However, all policies used technical language, such as 'cookies', 'API' and 'plugins' etc., and therefore arguably failed in addressing this recommendation. The recommendation is to 'avoid' rather than 'try to avoid', which is more akin to what the policies did. Some technical language was used and then followed by an explanation of what it meant. For example, Twitter, Facebook and Pintersest all provided a definition of 'cookies'. However, the term API was often used without definition. It must be noted that complying with this recommendation would prove almost impossible for SNS, and websites in general. As Robinson, et al., (Robinson *et al.*, 2009) acknowledges, national laws require full descriptions of data processing activities, which prove difficult to describe in a form the consumer can understand.

If you collect information from vulnerable individuals (such as children) have appropriate privacy notice to their level of understanding – would they understand the consequences? None of the SNS offered policies aimed at vulnerable individuals, which could be because they do not collect information from them. Pinterest, Twitter and LinkedIn all stated that they had a minimum age to use their site. Google and Flickr did not mention a minimum age in their policy, which may be because they have a generic policy, applicable to many services. Some services may be available to children and some may not be. However, upon brief further investigation (which will not change the results or scope of this study but was merely meant to ascertain whether SNS should have been providing notices specific to children), elsewhere on its site, Google+ mentions a minimum age of 13 to use its services and in Yahoo's terms of service it states a minimum age of 13 to use the SNS provided by Yahoo. Despite having a specific policy for its service, Facebook also stated that it had a minimum age of 13 in its terms of service rather than its privacy policy. The common factor of the minimum age of 13 may be because of the US FTC's Children's Online Privacy Protection Act 1998 (States., 1998 15 U.S.C. 6501–6505), which applies to the online collection of personal information from children under 13, and places additional requirements upon websites which do. Stating a minimum age of 13 is a bold statement that the website is not aimed at children, which the FTC will consider during investigations. This indicates, that regarding children, SNS may not be required to provide notices to their level of understanding. However, there may be other vulnerable individuals that require an appropriate privacy policy, especially given the amount of information, which an SNS can obtain. Other vulnerable individuals could include those with mental health or learning difficulties, over the age of 13. As all of the SNS only had one privacy policy, despite being used by vulnerable individuals (Holmes and O'Loughlin, 2014), they are certainly not in compliance with this recommendation.

4.4 Themes of Information Addressed by All Privacy Policies and Not Included in the Code

As mentioned in Section 3.5, if a clause was not allocated to an ICO Code recommendation, then it was placed into a category of 'miscellaneous'. Inductive thematic analysis of these clauses identified four themes, which appeared in **all** six privacy policies, but not in the ICO Code.

The Process of Updating the Privacy Policy

Whilst the code mentions that policies should be regularly reviewed, it does not advise that websites should include the process of doing so, or involve users in this process. However, all SNS mentioned something about their process of revising the policy, with some mentioning more than others. In particular Facebook was very detailed:

'If we make changes to this Data Use Policy we will notify you (for example, by publication here and on the Facebook Site Governance Page). If the changes are material, we will provide you additional, prominent notice as appropriate under the circumstances. You can make sure that you receive notice directly by liking the Facebook Site Governance Page... Unless we make a change for legal or administrative reasons, or to correct an inaccurate statement, we will give you seven (7) days to provide us with comments on the change. After the comment period, if we adopt any changes, we will provide notice (for example, on the Facebook Site Governance Page or in this policy) of the effective date'.

The detail Facebook provided may be linked with the fact that they had the largest number of clauses, as length did not appear to be a concern. Or it could be due to the criticism they have received regarding previous revisions of their policy (Anderson, 2009). Either way, this theme of detailing the process of updating the policy was common to all six SNS.

Functionality

As mentioned in relation to research question one, a large number of clauses were used to explain the functionality of SNS. This was often required to explain the purpose for using or sharing information. For example, both Facebook and LinkedIn had to explain Platform technology (an underlying system on which application programs can run) in order to explain who they share information with through it, and the purpose of this. Technically, the SNS could fulfill the Code's recommendations without explaining their functionality in detail. However, explaining these functionalities and their role in the collection, use and sharing of personal data can provide the contextual information, which makes policies more informative and enables users make decisions regarding their personal data.

List of Personal Data Collected

Interestingly, the Code does not mention that organizations should list exactly **what** personal data they collect, despite requiring that they state the purpose of obtaining, using and disclosing it. However, this was clearly present in some form in all of the SNS, who provide descriptions of what information they collect. For example, photos, associated metadata, messages, responses to ads etc. This may not have been included in the Code because it is not solely aimed at SNS, or even websites, but governs data collection both online and offline. In the offline context of filling in a questionnaire, an individual may be aware of what information they are providing, as they write it down. However, online users may not be so aware of what information is collected, especially as *'increasing amounts of data are not collected from the individuals concerned, but are instead observed, derived and inferred'* (Economic Cooperation *et al.*, 2014). Thus, it is impossible for SNS to be transparent about data processing without talking about this in some form.

How They Receive Information

Another theme present only in the policies is how the data is received e.g. through friends, through a users computer etc. Again, this may be for the same reason as above, that the Code is not aimed specifically at online contexts, where the sources of information are not as transparent and users may not be aware of *how* information is collected (Economic Cooperation *et al.*, 2014). However, providing this information clearly makes data processing more transparent.

5 Discussion

It is only by combining the answers to the first four research questions, that an understanding of their implications regarding the 'bigger picture' of privacy policies can be understood, and the fifth research question (about the extent to which standardization is possible) can be answered. Research question one indicated that there is not a lot of similarity between the privacy policies in the clauses that they use, with similarity ranging between **8-27%**. Although there were common clauses that could be drawn out, a power-law relationship existed between the number of common clauses and the number of SNS. This meant that there were only a small percentage of clauses shared by all six SNS, with the majority of clauses bespoke to one SNS. From looking at this alone, one would conclude that the similarity between SNS is so low that standardization seems, if not impossible, definitely a long way off. Much work would be required to create the level of similarity required for standardization by clause.

However, research question two showed that if you look at similarity in terms of themes of information addressed, rather than specific clauses used to do so, the similarity between SNS is far higher, ranging from **52-81%**. This indicates that SNS express similar themes of information, but in different ways. This indicates that standardization is not as impossible as research question one suggests. Indeed, it was noted that in relation to research question one, the differences in clause coverage were largely due to differences in functionality, semantics and amounts of elaboration between SNS. By looking at similarity in terms of theme, these differences are not as influential, particularly in relation to the semantics of the language used and the amount of elaboration. For example, if one SNS used ten clauses to address a recommendation and another SNS used two, with differing language, they have both still addressed the same theme.

Research questions two and three also showed that there were recommendations that every SNS addressed and recommendations that were addressed by none. However, it cannot be assumed that failure in addressing a recommendation is due to a SNS failing or choosing not to, as the particular recommendation may not be relevant to the SNS. Furthermore, some recommendations were almost impossible to comply with in the context of SNS, such as the recommendation to 'avoid using confusing terminology'. Because of the technical functionality of SNS, to properly convey what information will be used or disclosed for, a level of technical language is required.

Research question four showed a number of themes present in all the policies, which were not recommendations in the ICO Code. However, these clearly provided additional information to users regarding their privacy and personal data. One reason for this may be that the Code was not aimed exclusively at online environments and therefore, makes assumptions about people's awareness of the information they provide. Furthermore, its age may also attribute to this. Dated December 2010, the Code is currently under review and in the process of being updated. It will be interesting to see whether the final updated version includes any of these themes.

So, whilst research question three shows that the ICO Code provides a number of themes, which **are** present in the policies, it cannot be treated as an exhaustive list of what **should** be included in a policy, when all stakeholders are considered. Therefore, when looking to standardize policies of SNS or other types of website, a thorough examination should include the relevant policies, to ascertain a full list of themes, in addition to examining other sources.

6 Recommendations

Therefore, the outcomes of this investigation indicate that standardization is possible, albeit at a thematic level currently. It is also only possible if the issues raised are addressed. Moving forward, five initial recommendations are made to facilitate this:

1. **Begin with an asexhaustiveaspossible list of themes, which a SNS should address, rather than focusing on clauses initially.** Because the investigation showed high similarity between the policies in the ICO Code recommendations they covered, SNS policies are already in a better 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 in a standardized order on the left. The SNS clauses can then be 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 from which themes can be gleaned.

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

3. **Definitions, explanations and examples of technical terms should be standardized so that each policy uses the same ones.** For example, when referring to 'cookies' there should be a set, approved definition, explanation and example of a cookie. Given that it is almost impossible to avoid using technical terms in relation to describing the activities of 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.

This will lessen confusion and possibly support 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) should be standardized i.e. close account always means one thing, as does delete account. This would also lessen confusion and increase transparency.

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 an ICO Code recommendation was not addressed 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 this issue, making SNS justifications clear to users, regulators and researchers and evidencing that they have considered all the applicable requirements when constructing a privacy policy.

7 Limitations

One limitation is that this investigation only looks at the policies at a single point in time (August 2014), and whilst other comparative analyses have been discussed (Section 2.5), these did not compare the policies for the same elements as this investigation. This makes it difficult to provide information on the evolution of privacy policies at different points in time. However, this investigation has provided the starting point for future work, which could repeat this investigation on the policies as they change, to allow for a discussion on the evolution of privacy policies. Another limitation of the investigation is that the coding and analysis in Stages 1-5 was completed by one person and then discussed with the other researchers. Whilst this provided methodological consistency, coding in parallel could have enriched the work by providing multiple perspectives. Therefore, future work could conduct inter-rater reliability or intra-rater reliability on the findings. Furthermore, as discussed in Section 3.2, we chose to focus on the first layer of the policies for various pragmatic reasons. This could have impacted the findings, because the nine ICO recommendations that we did not find addressed in the 'first layer' of the policies could have been found in these further layers. Future work could extend this analysis to all layers of the policies on the same domain, to conclude whether certain elements have been addressed at all. This analysis could include a full legal analysis of what the implications of these exceptions are, whether the SNS are in contravention with the code, and whether it is appropriate that this information is not in the first layer of the policies.

8 Conclusion and Future Works

In conclusion, this paper proposes that the privacy policies of SNS demonstrate homogeneity and promising potential for standardization, at a thematic level initially. In answering research question one; analysis initially showed that similarity between the policies in the clauses they used was low. However, analysis of a second attribute showed that similarity between SNS was far higher in the themes of information addressed. Research question three showed that there were some ICO Code recommendations which none of the SNS addressed, but research question four showed that the policies included extra material, beyond the Code's recommendations. This analysis enabled us to conclude that the policies share attributes required for standardization to be possible, albeit on a thematic level initially.

To support this standardization in practice, the recommendations made in Section 6 could be followed up with further work on how to put these into action and standardize policies on a thematic level. For example, computer supported algorithms could be used to test whether this standardization can take place in practice, and by computers opposed to manually. Whilst this is outside the scope of this study, which sought to provide the starting point for this type of work, tests such as these would be valuable future work.

Indeed, by following the recommendations in Section 6, SNS could begin standardizing their policies by theme. Following this level of standardization, further levels could be explored, such as standardizing some of the clauses, or the specific information to be provided within themes. For example, the Code recommendation of detailing 'Who to contact if they want to complain or know about how their information will be used', could be standardized so that organizations have to provide the same specific information, like: telephone number; written address; email address etc. Furthermore, although standardization by clause is not currently feasible, overcoming the differences in functionality, semantics and elaboration with the Section 6 recommendations would allow for another assessment of similarity. This could investigate the potential for standardization by clause.

To make sure this standardized policy legally compliant (opposed to just re-arranging the policies in a standardized order), further investigations could seek to understand the reasons why some ICO Code recommendations were not addressed. It could also seek to understand whether the extra themes the policies discussed should be (or are) legal requirements. This could include investigating why they are not mentioned in the ICO Code. Once this is completed, a standardized policy for compliance by SNS with UK law could be created. As mentioned this will give indications for compliance with EU law, but will not be completely generalizable to all EU member states. Although, the DPD is being updated and replaced by the EU General Data Protection Regulation (GDPR) (Commission., 2016), which will be directly applicable in all EU member states without implementation. Therefore, this study could be replicated, but based on the GDPR's requirements, as the direct applicability of the GDPR means that the results would be generalizable to all EU member states. Furthermore, because this was an initial study, testing whether there was similarity between US policies with UK (based on EU) law, the ICO Code provided the most appropriate set of recommendations. However, extensions of this work could also measure similarity against thematic codes from other legal jurisdictions, or a superset of legal requirements from multiple jurisdictions. This would begin work towards a global standardized privacy policy. Standardizing the privacy policies of SNS would be a major success for individual management of personal data online. Our work demonstrates the potential for this at a thematic level currently, and we hope it will result in a step closer to the standardization of privacy policies of SNS in the future.

Acknowledgments

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

References

Aleixo, P. and T. Pardo. 2008. “Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts.” Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana – TIL(Oct.): 298–303.

Alexa. 2014. “Actionable Analytics for the Web.” http://www.alexa.com. (Accessed on 08/01/2014).

Anderson, H. 2009. “A privacy wake-up call for social networking sites.” *Ent. L. R.* 20(7): 245–248.

Barker, K., M. Askari, M. Banerjee, K. Ghazinour, B. Mackas, M. Majedi, S. Pun, and A. Williams. 2009. “A data privacy taxonomy.” In: *British National Conference on Databases*. Springer. 42–54.

Becker, J., M. Heddier, A. Oksuz, and R. Knackstedt. 2014. “The effect of providing visualizations in privacy policies on trust in data privacy and security.” In: *2014 47th Hawaii International Conference on System Sciences*. IEEE. 3224–3233.

Boyatzis, R. E. 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage.

Braun, V. and V. Clarke. 2006. “Using thematic analysis in psychology”. *Qualitative research in psychology*. 3(2): 77–101.

Britain., G. 1998. “Data Protection Act 1998.”

Center., E. P. I. 2015. “Complaint, Request for Investigation, Injunction and Other Relief”. http://epic.org/privacy/inrefacebook/EPIC-FacebookComplaint.pdf. (Accessed on 07/21/2015).

Clauset, A., C. R. Shalizi, and M. E. Newman. 2009. “Power-law distributions in empirical data”. *SIAM review*. 51(4): 661–703.

Commission., E. 2016. “Reform of EU data protection rules”. http://ec.europa.eu/justice/data-protection/reform/index_en.htm. (Accessed on 01/29/2016).

Commission., F. T. 2000. “Privacy Online: Fair Information Practices in the Electronic Marketplace.” https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000.pdf. (Accessed on 01/29/2016).

Commission., F. T. 2010. “Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework for Businesses and Policymakers.” https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-bureau-consumer-protection-preliminary-ftc-staff-report-protecting-consumer/101201privacyreport.pdf. (Accessed on 01/29/2016).

Consulting., S. T. 2015. “Hard Numbers for Public Posting Activity on Google Plus.” https://www.stonetemple.com/real-numbers-for-the-activity-on-google-plus/. (Accessed on 02/02/2016).

Cranor, L. 2012. “Necessary but not sufficient: Standardized mechanisms for privacy notice and choice”. *Journal on Telecommunications and High Technology Law*. 10: 273.

Cranor, L., K. Idouchi, P. Leon, M. Sleeper, and B. Ur. 2013. “Are they actually any different? Comparing thousands of financial institutions’ privacy practices”. In: *The Twelfth Workshop on the Economics of Information Se-Curity (WEIS 2013)*.

Economic Cooperation, O. for and Development. 2013. “The OECD Privacy Framework.” http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf. (Accessed on 01/27/2016).

Economic Cooperation, O. for, D. W. P. O. Security, and P. I. T. D. Economy. 2014. “Protecting Privacy in a Data-driven Economy: Taking Stock of Current Thinking.” http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/iccp/reg(2014)3&doclanguage=en. (Accessed on 06/23/2015).

Facebook. 2013. “Data Use Policy.” https://www.facebook.com/privacy/explanation. (Accessed on 08/01/2014).

Flickr. 2015. “Thank You, Flickr Community.” http://blog.flickr.net/en/2015/06/10/thank-you-flickr-community/. (Accessed on 01/02/2016).

Glazer, B. 1978. “Theoretical sensitivity: Advances in the methodology of grounded theory. Mill Valley”.

GNIP. 2016. “GNIP: Unleash the Power of Social Data.” https://gnip.com. (Accessed on 01/02/2016).

Google. 2014. “Privacy Policy. Community.” https://www.google.co.uk/intl/en/policies/privacy/archive/20140331/. (Accessed on 01/02/2016).

Hargittai, E. *et al.* 2010. “Facebook privacy settings: Who cares?” *First Monday*. 15(8).

Holmes, K. M. and N. O’Loughlin. 2014. “The experiences of people with learning disabilities on social networking sites.” *British Journal of Learning Disabilities*. 42(1): 1–5.

Jaccard, P. 1912. “The distribution of the flora in the alpine zone.” *New phytologist*. 11(2): 37–50.

LinkedIn. 2014. “Your Privacy Matters.” https://www.linkedin.com/legal/privacy-policy?trk=uno-reg-guest-home-privacy-policy. (Accessed on 08/01/2014).

McDonald, A. M. and L. F. Cranor. 2008. “The Cost of Reading Privacy Policies.” *ISJLP*. 4: 543.

Office., I. C. 2010. “Privacy notices code of practice.” http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Detailed_specialist_guides/PRIVACY_NOTICES_COP_FINAL.ashx. (Accessed on 01/02/2016).

Office., I. C. 2016. "Use a layered approach." https://ico.org.uk/about-the-ico/privacy-notices-transparency-and-control/use-a-layered-approach/. (Accessed on 01/02/2016).

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

Parliament, E. 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". http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML. (Accessed on 12/21/2014).

Pinterest. 2014. "Privacy Policy." https://about.pinterest.com/en/privacy-policy. (Accessed on 08/01/2014).

Press., O. U. 2011. "Oxford English Mini Dictionary."

Robinson, N., H. Graux, M. Botterman, and L. Valeri. 2009. "Review of the European Data Protection Directive." *Information Commissioner's Office*.

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

Ryan, G. W. and H. R. Bernard. 2003. "Techniques to identify themes." *Field methods*. 15(1): 85–109.

Schwartz, A. 2009. "Looking back at P3P: lessons for the future." *Center for Democracy & Technology, https://www. cdt. org/files/pdfs/P3P_ Retro_ Final_ 0. pdf*.

Securities and E. Commission. 2009. "Final Model Privacy Form Under the Gramm-Leach-Bliley Act." https://www.sec.gov/divisions/marketreg/tmcompliance/modelprivacyform-secg.htm. (Accessed on 01/29/2016).

States., U. 1998 15 U.S.C. 6501–6505. "Children's Online Privacy Protection Act 1998."

statista. 2016. "Leading social networks worldwide as of January 2016, ranked by number of active users (in millions)." http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. (Accessed on 02/02/2016).

Strauss, A., J. Corbin, *et al.* 1990. *Basics of qualitative research*. Vol. 15. Newbury Park, CA: Sage.

Van Alsenoy, B., V. Verdoodt, R. Heyman, E. Wauters, J. Ausloos, and G. Acar. 2015. "From social media service to advertising network: a critical analysis of Facebook's Revised Policies and Terms." (Accessed on 02/02/2016).

Wu, L., M. Majedi, K. Ghazinour, and K. Barker. 2010. "Analysis of social networking privacy policies." In: *Proceedings of the 2010 EDBT/ICDT Workshops*. ACM. 32.