Privacy dictionary

Privacy Dictionary: A New Resource for the Automated Content Analysis of Privacy

Asimina Vasalou (corresponding author)

Department of Computer Science, University of Birmingham

Birmingham B15 2TT

Tel. 121 4143744, Fax. 121 4144281

minav@luminainteractive.com

Alastair J. Gill

Department of Sociology, University of Surrey

Guildford, Surrey, GU2 7XH, UK

Tel. 01483 689365, Fax. 01483 689551

A.Gill@surrey.ac.uk

Fadhila Mazanderani

Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, UK

Tel. 01865 287210, Fax. 0 1865 287211

mazanderani@gmail.com

Chrysanthi Papoutsi

Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, UK

Tel. 01865 287210, Fax. 01865 287211

chrysanthi.papoutsi@oii.ox.ac.uk

Adam Joinson

School of Management, University of Bath, Claverton Down, Bath BA2 7AY UK

Tel. 01225 386742, Fax. 01225 386473

a.Joinson@bath.ac.uk

Accepted for publication at the Journal of the American Society for Information Science and Technology

Abstract

This paper presents the privacy dictionary, a new linguistic resource for automated content analysis on privacy-related texts. To overcome the definitional challenges inherent in privacy research, the dictionary was informed by an inclusive set of relevant theoretical perspectives. Using methods from corpus linguistics, we constructed and validated eight dictionary categories on empirical material from a wide range of privacy sensitive contexts. It was shown that the dictionary categories are able to measure unique linguistic patterns within privacy discussions. At a time when privacy considerations are increasing, and online resources provide ever growing quantities of textual data, the privacy dictionary can play a significant role, not only for research in the social sciences, but also in technology design and policy making.

Introduction

Questions surrounding privacy have gained increasing traction across academic disciplines, policy discourses, the media and everyday life. Nevertheless, privacy remains a concept that is notoriously hard to define and study. Relevant interdisciplinary work has illustrated many of its different parameters. Some studies have focused on how privacy is achieved behaviorally through actions of control (Tavani, 2007; Petronio, 2002; Adams & Sasse, 1999; DeCew, 1997; Altman, 1975). Others have discussed its positive psychological effects (Pedersen, 1999; Altman, 1975; Westin, 1967), and noted its governance though social norms (Petronio, 2002; Adams & Sasse, 1999). Privacy not only entails the selective control over the physical realm (e.g. one's sensory presence), but, depending on the interactional context, it can also involve informational (e.g. personal information) or expressive (e.g. one's opinions and values) control (DeCew, 1997). Yet further work has described privacy as underpinned by an optimal desired state, which is dialectic in nature (Altman, 1975). Such conceptual diversity makes privacy research extremely challenging. Although syntheses of the literature have been attempted they have been unable to produce a consistent, uniform theory (Schoeman 1984; Parent 1983). Thus, privacy has been described as a concept in disarray, a chameleon word (Solove, 2008).

At a time when theorists are still grappling with how to define it, privacy has become one of the most contested social issues of the information age (Strickland & Hunt, 2005). In the UK, a DNA profile is held on criminals, including those suspected, but not charged with a crime (Casciani, 2009). Workplace surveillance is an established practice (BBC news, 2003); and social network sites are thriving on people's willingness to disclose and consume personal information (Vasalou et al., 2010). Understanding individuals' privacy perceptions, particularly in relation to technology, has thus become a central question that cuts across a number of disciplines (e.g. human computer interaction, information science, communication studies, computer science). The present research builds on the recognition of a continuing need to advance theory-inclusive and sophisticated methods for studying privacy (Patil et al., 2006). Taking a holistic theoretical perspective, we developed a 'privacy dictionary' that can be implemented for automated content analysis to allow researchers to systematically measure different aspects and uses of privacy language.

The following section describes the methodological landscape against which automated content analysis becomes a useful tool for the study of privacy and outlines the benefits of our approach. We then explain how existing automated content analysis tools work in practice and how a new dictionary, such as our own, may operate within these tools. Next, the theoretical framework underpinning the design of the privacy dictionary is presented. We go on to describe two studies in which 355 dictionary words and eight categories were designed and evaluated. Our main finding is that categories included in the privacy dictionary are able to capture unique linguistic features in privacy language. The paper ends with a discussion of our findings and potential applications of the privacy dictionary in research, policy making and technology development.

The Methodological Landscape of Privacy

A number of methodologies have been employed to shed light in the privacy domain, creating a varied methodological landscape where survey-based methods have traditionally been particularly prevalent. However, while surveys may be able to gather people's self-reported perceptions and attitudes, the assumption that these translate directly into related behaviours remains problematic (Acquisti & Grossklags, 2004). One of the major criticisms raised in relation to attitudinal questionnaires is that they frequently include leading items that bias participants' responses (Harper & Singleton, 2001). This often results in inflated self-reports of privacy concerns that rarely explain privacy protective behaviour (Acquisti & Grossklags, 2004). When such questionnaires are used in experimental settings it has been shown that they can prime particular privacy-related behaviours. For example, one study found that participants avoided answering sensitive questions after completing a privacy concern measure (Joinson et al., 2008). The reliability of these methods has, therefore, been called into question (Patil et al., 2006).

Other methodological approaches have been developed to counter these problems, underpinned by the belief that natural language reveals attention patterns, thoughts, feelings, and provides a way of understanding our social worlds (Chung & Pennebaker, 2007; Tausczik & Pennebaker, 2010; Pennebaker et al., 2003). Some researchers use interviewing and focus groups to probe people as a means of deconstructing and analysing prior violations (e.g. Adams & Sasse, 1999; Raento & Oulasvirta, 2008). While this approach lends itself particularly well to

contexts with persistent privacy problems, helping to identify the source of the breach and participants' judgments, it is limited insomuch as it does not capture naturally occurring privacy practices. These limitations have motivated privacy researchers to develop methods aimed at capturing more nuanced, inclusive and unbiased portrayals of people's concerns, needs and practices. One way this has been done is by looking at privacy concerns and practices as embedded within various domains, in which different manifestations of privacy are gauged through neutral questions framed within these wider contexts, e.g., social network sites, mobile computing and healthcare (e.g. Christidi & Rosenbaum-Elliot, 2010; Mazanderani & Brown, 2010). Another has been to use diary-based approaches such as experience sampling methods (ESM) as a means of prompting privacy responses in real time (e.g. Anthony et al., 2007; Mancini et al., 2009).

In analysing participants' privacy experiences through language, researchers have traditionally turned to qualitative methods such as thematic analysis to interpret their data. Against this context, automated content analysis offers the potential to advance existing analytic tools, either as a method in its own right or in conjunction with other analysis methods. First, automated content analysis can systematically measure specific psychological components, as such serving a parallel function to psychometric measures whose use is well established in the social sciences. Whilst in these latter cases individual questions are the observed items whose submission to statistical procedures, such as factor analysis, informs the researcher about unobserved latent variables, in automated content analysis words and phrases become the observed variables (Lowe, 2004). Indeed, with automated content analysis, it has been possible to reliably identify specific emotional states (Hancock et al., 2007; Gill et al., 2008), predict deception (Hancock et al., 2008) and detect differences in personalities (Oberlander & Gill, 2006). Second, automated content analysis offers a common platform that yields comparable results within and across a large number of different datasets, such as interviews, focus groups and open-ended questions. As the coding is done consistently according to a common frame, the discrepancies that typically emerge due to different interpretations of coding schemes are prevented (Mehl & Gill, 2010). This is particularly useful when researchers want to minimise the subjectivity of individualised qualitative analysis in order to engage in collaborative work across a large body of texts. Third, when analysing more open-ended texts in which specific questions

on privacy have not explicitly been raised, either to prevent priming responses or else if the analysis being conducted is a secondary or post hoc one, automated content analysis can be used in conjunction with other analysis methods. For example, it can be used prior to qualitative thematic coding to pre-identify language of potential interest and hence save time and effort in the coding process (Mehl & Gill, 2010).

Automated Content Analysis: Dictionaries and Software

Automated content analysis software, with particular reference to category frequency software, at its core, uses a dictionary comprised of individual words or phrases that are assigned to one or more linguistic categories. The software will process any given number of texts by counting occurrences of each dictionary word within the text and incrementing the relevant categories to which the words belong. The output of the analysis consists of values for each linguistic category, represented as a percentage of the total words in the text. For example, in parsing the input text "I am", a dictionary that includes the linguistic category "personal pronouns" would increment "personal pronouns" by one, assigning it a value of 50%. This analysis would be repeated for each input text individually yielding a matrix with category values stored (columns) for each case (row).

The categories and words forming part of any dictionary vary depending on the aspects of language that researchers aim to measure and the social psychological phenomena they strive to understand (Tausczik & Pennebaker, 2010). Previous research has developed a number of linguistic categories ranging from functional aspects of language, such as *first person singular pronouns* (e.g., I, my, we), *negations* (e.g., no, never, not), to language that captures the content of communication, e.g. *positive emotions* (e.g., happy, pretty, good), *achievement* (e.g., try, goal, win). Examples of existing dictionaries include the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007), the Affective Norms for English Text (ANEW) (Bradley and Lang, 2007), Diction (Hart & Carroll, 2011) and the General Inquirer (Stone et al., 1966). With the exception of ANEW, these dictionaries are offered together with content analysis software that can be either downloaded on a personal computer or used over the Internet. Researchers who want to design a new dictionary that will operate within existing content analysis software can do so by consulting the manual accompanying each of these tools.

From Theory to Language

To choose words that are meaningful to the semantic analysis of privacy, we must begin from a sound and comprehensive theoretical foundation. Despite theorists' agreement over several shared features, context determines much of the way that privacy has been defined. Therefore, in constructing a dictionary that would encompass all of these manifestations, we need to cast our net wider than existing definitions. Theories of categorization provide a useful frame for meeting this challenge. The classic approach to concept definition identifies sufficient and inclusive criteria (e.g. *control* over *information*). Any instance described by these criteria is considered to be a member of the concept (Rosch, 1978). Many natural language categories, however, do not share a common set of defining features. Taking games as an example, card-games, board games and playing tennis bear a 'family resemblance' structure. Members characterized by more features of the family are better exemplars, thus making membership a matter of degree (Wittgenstein, 2001). Prototype theory evolved from this perspective to propose that concepts, such as privacy, are organized through prototypes that represent the average member of a concept. When new situations are perceived, we evaluate their similarity against the prototype to determine whether they belong to the concept and whether they are good or poor exemplars (Rosch, 1978).

Solove (2006; 2008) proposed that the multifaceted perspectives adopted vis-à-vis privacy exist due to its family resemblance structure. Recent research validated this claim in a unilinear series of experiments (Vasalou et al., 2010). First, 146 participants were asked to report features of privacy. This resulted in an average of 6.6 features with a total of 82 privacy features. If a concept is organized around a prototype, a wide range of features will be reported, none of which are shared across all reports (Fehr, 1988). Crucially, this first study found no agreement on a single unifying feature of privacy. Second, it was determined whether participants could reliably rate the features' importance or centrality with regards to the concept. Once it is shown that features of the concept vary in their degree of centrality, exemplars of the concept can be directly derived from the features (Fehr, 1988). Using a 9-point scale (9-extremely good feature, 1-extremely poor feature), 118 participants were able to reliably rate the privacy features' centrality. In a final step, 62 participants evaluated vignettes that contained either more central or peripheral features of the privacy concept. The vignettes containing more central privacy features

were recognized as better exemplars. Therefore, when participants faced events pertinent to privacy, the internal structure of privacy was found to have an effect on their cognition, which indeed demonstrates that membership is a matter of degree.

The wide-ranging list of privacy features revealed in this research was found to cover the broad scope of psychological and behavioural components discussed in the privacy literature (Vasalou et al., 2010). This ensures that a dictionary built from this foundation will not be representative of a single theoretical view. Moreover, the features reflected the contextual nature of privacy we sought to capture, such that context was explicitly woven into language, expressing environmental (e.g. *personal space*), informational (e.g. *personal information*) and expressive concerns (e.g. *concealing embarrassing details*). The privacy dictionary was therefore developed on the basis of these 82 privacy prototype features.

The remainder of the paper describes how the dictionary was designed and evaluated in a series of two studies. The purpose of Study 1 was two-fold: (1) to choose individual dictionary words that can be grouped into theoretically motivated categories and (2) to evaluate whether the resultant dictionary categories measure differences between privacy language and general language use. Study 2 aimed at replicating the findings of Study 1 by evaluating the dictionary categories within a second linguistic corpus that was independent of that used for the dictionary development.

Study 1

Dataset

To carry out the dictionary development and analysis, we needed a dataset of language both rich in privacy content and general language use. To our knowledge, a repository of data did not exist to satisfy these requirements. For this reason, we built our own dataset by combining two sources of data: (1) one-to-one interviews and focus groups that captured natural patterns of privacy practices and (2) self-reported privacy concerns and violations. In each instance, a within-subject design was followed; for each participant, we collected both a sample of privacy language forming a ‘privacy condition’ and a sample of general language use forming a ‘control condition’.

Privacy practices

Eight contexts that previous research suggests are sensitive to privacy issues were chosen. These were: (1) criminal offences and imprisonment (Pattenden & Skinns, 2010) (2) children and the Internet (Livingstone, 2006); (3) financial exclusion (Leyshon et al., 2006); (4) sexuality (Meerabeau, 2001); (5) sharing in social network sites (Christidi & Rosenbaum-Elliot, 2010); (6) experiences of elderly people with medical care (Costello, 2001); (7) health experiences within medical practices (DeCew, 2000); (8) the role of cultural identity in community participation (Petronio, 2002). We visited the UK Data Archive (hosted on the ESDS site: www.esds.ac.uk), which is a central data bank of previously UK-funded research, or contacted researchers who had worked on these topics to identify pre-existing datasets. The aim was to find qualitative data rich in privacy content, which had been generated by asking questions unrelated to privacy, in order to avoid methodological problems of priming in the responses. The data included fully abided with participants' informed consent and the institutions' ethics approval procedures.

A team of five judges who were knowledgeable in privacy theory selected appropriate transcripts using the following procedure. Two researchers worked on each context. The first judge surveyed the entire panel of transcripts made available in order to identify a maximum of five transcripts per context that involved a diversity of privacy-related issues. This yielded a total of 38 transcripts[1]. Focusing on one transcript at a time, the same judge identified areas in the text where participants expressed privacy-related issues. These segments were examined by the second judge who raised any disagreements concerning the inclusion of a given privacy text. Disagreements between judges were resolved through discussion and only privacy texts that yielded bilateral agreement were included. Language identified as expressing privacy experiences was categorised in the privacy condition while the remaining interview formed part of the control condition.

Privacy concerns and violations

The second source of data consisted of open-ended descriptions of privacy concerns or violations. To bypass the issue of experimenter priming, we collected blog posts in which authors had provided spontaneous descriptions of such events. Data collection was limited to blogspot.com, which is the most popular blogging service[2]. Using Google's search engine, we searched for the keyword 'privacy', limiting the pages to those hosted in the United Kingdom

only. Software written in the Python language was then used to collect and store blog post entries. To obtain data for the control condition, the software automatically collected the blog post immediately preceding the privacy post, from the same blog.

This process resulted in the collection of 859 blog posts. Two judges, knowledgeable in privacy theory, worked together to evaluate their integrity. The first judge read over each privacy related post to verify that it described a privacy concern or violation in its entirety. It was found that many entries included brief and isolated references to privacy whereby the chief topic of the post was irrelevant. These were excluded from the dataset. Moreover, a number of blog posts featured adverts for privacy protection software that were also discarded. Next, we turned to the blog posts forming the control condition and verified each one individually to ensure that they did not include any references to privacy. In cases where privacy was the topic of the post, it was replaced with the previous entry. A second judge went on to evaluate this reduced dataset, raising any objections over the inclusion of a given post. The blog dataset was the product of unilateral agreement between the two judges and it included 129 posts of privacy violations and, 129 non-privacy posts respectively. As with the privacy practices dataset, this data source captured a rich range of contexts: some blog authors described concerns or violations in social network sites and the Internet more generally. Other blog posts focused on particular victims such as children, or people whose sexuality was exposed. The events described were seen as threatening people's quality of life, financial wellbeing and health. Finally, some bloggers described the legal and political dimensions of privacy. Table 1 summarizes the dataset consolidating these two data sources.

Table 1: Total number of words across the two datasets.

Methodology	Dataset type	Privacy condition	Control condition	Total
Interviews and focus groups (N=38)	Privacy practices	65,324	168,472	*233,796*
Self-reports (N=129)	Privacy concerns/ violations	117,551	79,312	*196,863*
		182,875	*247,784*	*430,459*

Dictionary Words and Categories

In designing the dictionary, iterative techniques, similar to those applied in the development of similar dictionaries were used (cf. Pennebaker et al., 2007). After collecting a

panel of relevant words, groups of judges who were knowledgeable in privacy theory decided whether they should be included in or excluded from the dictionary, and how they should be grouped into categories.

The 82 privacy prototype features were first surveyed to identify and retain single word features (e.g. *isolation*). Phrases (e.g. *keeping to oneself*) were reduced to single words, where possible, so as to ensure maximal compatibility with automated content analysis software and dictionaries (e.g. LIWC software: Pennebaker et al., 2007). To give one example, "having control over one's information" was broken down into two linguistic units, *control* and *information*. These revised prototype words were then used as "seed words" over several iterations to generate additional synonyms and antonyms using traditional and computational semantic dictionaries and thesauri. Two judges evaluated the consistency of the additional synonyms and antonyms with the original words, with consensus between judges determining a word's inclusion or exclusion. This resulted in the selection of 730 dictionary words.

In a first step, frequency counts for each dictionary word were calculated on the language contained in the privacy condition of our dataset. This was done to ensure that words achieve an acceptable frequency of usage when people are talking about privacy related issues. One particular problem of using low-frequency words is that of sparse data, which is more likely to lead to skewed distributions. Words used less than two times were excluded resulting in 487 dictionary words. In a second step, three judges conducted 'key word in context' analysis (KWIC) (Rayson, 2009) on the dictionary words as they appeared in the privacy condition and in the control condition. This allowed us to obtain contextual information of the occurrence of the dictionary words. Table 2 provides an example KWIC output for the word 'public'. The KWIC analysis helped identify words for possible exclusion. Despite the frequent use of certain words in discussions around privacy, their high frequency in the control condition indicated that they were ubiquitous in language more generally (e.g. talking). Moreover, the reduction of multi-word prototype features to single words led to some words capturing unintended meanings from those originally envisaged by the judges. For example, the word 'company' was intended to capture the state of 'having or not having company', but instead the analysis of the context in which this word was used revealed that it was more frequently used to refer to a business organization.

Table 2: Examples of KWIC analysis for the word 'public'.

Members had said they were unsure about how	public	their information had become.
We consider it crucial that there is no	public	disclosure of this information.
Blogging is essentially a	public	rather than a private activity.
Are you saying sir, that in a	public	place, I have to ask permission of every person in my picture?
I don't understand why these things should be	public	It 's just bizarre.

In both cases, we sought to determine whether the 'problematic' word was used consistently (i.e., appearing a minimum of two times) in the form of a phrase, either when talking about privacy-related matters or in the more general language use captured through the control condition. Aiming to 'contextualise' single words by replacing them with phrases, we used n-gram software (Banerjee & Pedersen, 2003) to identify phrase clusters, i.e., two-word sequences preceding and following the word under investigation. Consistent phrases within the privacy condition were *included* in the dictionary in place of the single word. Per contra, phrases that were ubiquitous in the general language captured in the control condition were *excluded* from the dictionary. Wherever it was not possible to contextualise a word, the disputed word was removed from the dictionary. This iterative process yielded the final 355 dictionary words and phrases. Table 3 presents several examples of phrase clusters.

Table 3: Cluster examples (the original single dictionary words appear in bold).

Phrases excluded from the dictionary	**Phrases included in the dictionary**
public **confidence**	emotional **support**
security staff	sexual **behaviour**
let **alone**	closely **watched**
deputy **judge**	lack of **control**
I'm **afraid** that	**reasonable** suspicion

The third and final stage in the dictionary development was to construct theoretically sound categories of semantically similar words, which would form the basis of the output of the analyses carried out using the privacy dictionary. This is necessary to enable the measurement of consistent and reliable categories that can provide theoretically meaningful results. For this task an additional judge was recruited who was familiar with linguistics and automated content analysis. While consulting the semantics of each word as it appeared in context (KWIC), four researchers worked together to construct eight categories. These categories were further verified using Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA is a computational semantic technique that represents a word's meaning as a high-dimensional vector space derived

from the word's contextual information within a corpus of text. Using the LSA website (http://lsa.colorado.edu), we performed a matrix comparison based on the semantic space of the TASA corpus of General Reading. The resulting matrix provides a cosine value after comparing the vectors associated with any two of the input words. A cosine close to +1 indicates a very high degree of semantic relatedness between words, whereas -1 indicates semantic dissimilarity. Although these values are theoretical extremes on a scale, they rarely reach +1 or -1. We consulted this matrix to verify that words grouped together had high semantic relatedness. If a word had a very low or negative cosine with two or more words within its respective category, it was reassessed through KWIC analysis in order to assign it to a more appropriate category. The following eight high-level categories are the result of unilateral consensus between the participating researchers.

- **NegativePrivacy** (120 words or phrases; e.g., judgmental, troubled, interfere). This category captures the antecedents and consequences of privacy violations. NegativePrivacy includes words that relate back to privacy concerns, risks, as well as judgments about the source and type of violation (e.g. Adams & Sasse, 1999; Buchanan et al., 2007; Yao et al., 2007).
- **NormsRequisites** (33 words or phrases; e.g., consent, respect, discrete). NormsRequisites encapsulates the norms, beliefs and expectations in relation to achieving privacy. This category can be used to appraise the presence and type of norms that govern each context (e.g. Petronio, 2002; Nissenbaum, 2004)
- **OutcomeState** (39 words or phrases; e.g., freedom, separation, alone). OutcomeState includes words that describe the static behavioural states and the outcomes that are served through privacy. This category is in alignment to Westin's definition of privacy states and functions (e.g. Westin, 1967; Pedersen, 1999).
- **PrivateSecret** (22 words or phrases; e.g., secret, intimate, data). PrivateSecret includes descriptors or words that express the 'content' of privacy. This category can be used to understand precisely what aspects people regard as being private (e.g. DeCew, 1997; Tavani, 2007).
- **Intimacy** (22 words or phrases, e.g., trust, friendship, confide). Intimacy comprises of words that portray and measure different facets of small group privacy. It includes words that refer

to the psychological requisites in opening up to another person, as well as the emotional closeness that develops between people (e.g. Westin, 1967; Schoeman, 1984; Petronio, 2002).

- **Law** (27 words or phrases; e.g., confidentiality, policy, offence). This linguistic category includes words employed to describe legal definitions of privacy (e.g. Regan 1995; Rule 2007).
- **Restriction** (63 words or phrases, e.g. conceal, lock, exclude). Words in this category express the closed, restrictive and regulatory behaviours employed in maintaining privacy. Thus, the Restriction category can be used to measure the behaviours that people take in order to protect their privacy (e.g. Petronio, 2002; Tavani, 2007).
- **OpenVisible** (46 words or phrases, e.g. post, display, accessible). This category includes words that represent the dialectic openness of privacy (e.g. Altman, 1975; Petronio, 2002).

The final dictionary was formatted to be compatible with LIWC 2007, but analysis was conducted using the TAWC open source version of the LIWC word count software (Kramer et al., 2004).

Results and Discussion

The 334 texts of our dataset (167 belonging to the privacy and 167 belonging to the control condition) were processed so that word counts for each dictionary category were saved into eight separate variables. Table 4 presents the mean word occurrence and percentages for each dictionary category.

Table 4: Mean word counts (percentages) for privacy categories.

	Control Condition		**Privacy Condition**	
OpenVisible	3.01	(0.29%)	7.05	(0.63%)
PrivateSecret	1	(0.07%)	5.75	(0.69%)
Intimacy	4.51	(0.29%)	5.66	(0.47%)
NegativePrivacy	2.07	(0.15%)	5.5	(0.53%)
OutcomeState	0.88	(0.08%)	3.35	(0.33%)
Law	0.86	(0.10%)	3.05	(0.20%)
NormsRequisites	0.8	(0.05%)	2.5	(0.21%)
Restriction	1	(0.11%)	2.43	(0.24%)

The output of this analysis was subjected to a series of GLM regressions with the word count of the privacy categories as the dependent variable. The privacy and control condition was the independent variable. The control condition served as the reference category. Given that participants spoke, or wrote, in variable rates (M=1,305; SD=2,035), the log of the total words used in each condition was defined as an offset variable to control for this confound. The Pearson Chi-square/df goodness of fit measure for each regression model ranged between 2 and 7 indicating the presence of overdispersion (i.e. high number of zeros in the dependent variable). Therefore, a negative binomial regression model was fitted to the data. For all analyses, we checked that the conditions of application were respected using residuals analyses (residuals vs. fit and Cook's distance). In all eight regressions, the likelihood ratio chi-square test was significant at the .001 level indicating that the fitted model was significantly better than the intercept-only model.

Table 5: Likelihood ratio chi-square test.

	Likelihood ratio chi-square
Law	33.018***
OpenVisible	16.595***
OutcomeState	81.066***
NormsRequisites	61.531***
Restriction	44.664***
NegativePrivacy	73.081***
Intimacy	12.557***
PrivateSecret	188.853***

In each model, the effect of condition was statistically significant. To interpret differences in category word rates between conditions, the coefficients were converted to rates through the exponential function. The fitted model was subtracted from the intercept model to obtain rate differences between the privacy and control condition. To gain an estimate of word occurrence for a given text segment of 1,000 words, rate difference was multiplied by 1,000. Compared to the control condition, for every 1,000 words, participants speaking about privacy used 5.7 words more for *PrivateSecret*, 3.7 for *NegativePrivacy*, 2.65 for *OpenVisible*, 2.55 for *OutcomeState*, 1.8 for *Intimacy*, 1.6 for *NormsRequisites*, 1.54 for *Restriction*, 1.4 for *Law*.

Table 6: Coefficients and chi-square results.

		B	S.E.	95% Wald Confidence Interval		Wald Chi Square
				Lower	Upper	
Law	Intercept	-7.101	.134	-7.363	-6.839	62882.299***
	Condition	.995	.167	.627	1.282	32.640***
OpenVisible	Intercept	-5.634	.10	-5.836	-5.432	6221.235***
	Condition	.554	.14	.287	.82	16.606***
OutcomeState	Intercept	-7.196	.13	-7.5	-6.9	6117.219***
	Condition	1.480	.17	1.156	1.80	80.341***
NormsRequisites	Intercept	-7.5	.141	-7.77	-7.2	6146.304***
	Condition	1.355	.173	1.01	1.7	60.695***
Restriction	Intercept	-7.197	.13	-7.5	-6.9	6295.453***
	Condition	1.116	.168	.778	1.4	44.507***
NegativePrivacy	Intercept	-6.514	.114	-6.737	-6.29	6527.593***
	Condition	1.254	.145	.968	1.54	74.017***
Intimacy	Intercept	-5.81	.098	-6.001	-5.618	7065.020***
	Condition	.471	.133	.212	.731	12.634***
PrivateSecret	Intercept	-7.26	.13	-7.52	-7.00	5899.306***
	Condition	2.210	.16	1.90	2.52	190.075***

These initial findings are encouraging. The eight dictionary categories were able to measure linguistic patterns in the language contained within the privacy condition when compared to the control condition. Nonetheless, given that the dictionary was designed and then evaluated on the same dataset, there is a possibility of over-fitting our dictionary to one dataset. For this reason, in Study 2 we collected a new dataset, on which the dictionary had not been trained, with the aim to replicate the findings of Study 1.

Study 2

Procedure and Participants

A message was posted on a University intranet website inviting staff and students to take part in an online survey that aimed to capture a taxonomy of everyday life events. In total, 210 people took part of which 143 were female. The mean age was 26.8 (SD=9). After responding to some demographic questions, participants were assigned to one of two conditions. In the control condition, participants were requested to report events that had happened during the previous week as if they were addressing a good friend. These instructions were aimed at producing more

naturalistic language. By contrast, in the privacy condition, participants were asked to describe a past event during which they felt their privacy had been violated either by another person, a group or an organisation. The final dataset comprised of 38,966 words, of which 29,757 belonged to the control condition and 9,209 belonged to the privacy condition.

Results and Discussion

Word counts for the eight dictionary categories were calculated on the 210 texts (105 belonging to the privacy and 105 belonging to the control condition) using the TAWC software. Table 7 presents the mean category word occurrence and percentages for the dictionary categories.

Table 7: Mean word counts (percentages) for privacy categories.

	Control Condition		**Privacy Condition**	
OpenVisible	.11	0.05%	.38	0.48%
PrivateSecret	.12	0.04%	1.13	1.50%
Intimacy	1.12	0.41%	.44	0.49%
NegativePrivacy	.16	0.05%	.84	0.87%
OutcomeState	.30	0.11%	.25	0.29%
Law	.0	0.00%	.08	0.09%
NormsRequisites	.12	0.04%	.23	0.32%
Restriction	.27	0.10%	.33	0.38%

As Table 7 indicates, participants talked very infrequently about the legal dimensions of privacy in both conditions, which resulted in the *Law* category being dropped from subsequent analysis. Seven GLM regressions were calculated with the word count of the privacy categories as the dependent variable. The privacy and control condition was the independent variable with the control condition serving as the reference category. Participants, on average, wrote 185 words (SD=143). The log of the total words used in each condition was defined as an offset variable to control for variable word rates across participants. An examination of the Pearson Chi-square/df goodness of fit measure for each dependent variable indicated that a Poisson regression model was appropriate. Conditions of application were respected using residuals analyses. With the exception of Intimacy, which was non-significant, in the remaining six regressions, the likelihood ratio chi-square test was significant at the .001 level indicating that the fitted model was significantly better than the intercept-only model.

Table 8: Likelihood ratio chi-square test.

	Likelihood ratio chi-square
OpenVisible	65.70***
OutcomeState	14.48***
NormsRequisites	28.28***
Restriction	29.517***
NegativePrivacy	172.60***
Intimacy	2.06, *ns*
PrivateSecret	268.10***

An examination of the five models that yielded significance shows that the effect of condition was statistically significant in the predicted direction. To interpret differences between conditions, the same procedure used in Study 1 was applied in order to convert the log of words to word rates per 1,000 words. Compared to the control condition, for every 1,000 words, when talking about privacy participants used 12.6 words more for *PrivateSecret*, 9.1 for *NegativePrivacy*, 4 for *OpenVisible*, 2.6 for *Restriction*, 2.17 for *NormsRequisites* and 1.9 for *OutcomeState*.

Table 9: Coefficients and chi-square results.

		B	**S.E.**	**95% Wald Confidence Interval**		**Wald Chi Square**
				Lower	**Upper**	
OpenVisible	Intercept	-7.816	.2887	-8.382	-7.250	733.062***
	Condition	2.377	.3291	1.732	3.022	52.149***
OutcomeState	Intercept	-6.867	.1796	-7.219	-6.515	1461.755***
	Condition	1.035	.2632	.519	1.551	15.451***
NormsRequisites	Intercept	-7.736	.2774	-8.279	-7.192	777.968***
	Condition	1.786	.3444	1.111	2.461	26.897***
Restriction	Intercept	-6.969	.1890	-7.339	-6.598	1359.725***
	Condition	1.396	.2535	.899	1.893	30.316***
NegativePrivacy	Intercept	-7.468	.2425	-7.943	-6.992	948.007***
	Condition	2.828	.2647	2.310	3.347	114.179***
PrivateSecret	Intercept	-7.736	.2774	-8.279	-7.192	777.968***
	Condition	3.395	.2920	2.823	3.968	135.226***

These findings lend further support to the function of six out of eight dictionary categories. Two categories that did not function as predicted, *Intimacy* and *Law*, are likely due to the nature of the data analysed. The category *Law* includes technical privacy terms relating to the legal protection of privacy. It is likely that laypeople engage less in legal privacy related discussions, which may be limited to expert legal commentators like some of the blog authors of the first study. The category *Intimacy* measures the positive function of privacy achieved within

groups of intimate others. Since participants were asked to describe events in which their privacy was violated, the importance of intimacy may have been dampened. However, in looking at the word percentages across the control and privacy condition (see Table 7), words from this category, on average, appeared equally frequently. A closer examination of the texts revealed that participants in the privacy condition described situations in which intimacy had been violated, whereas participants in the control condition described intimate events of the previous week. Thus, the instructions employed in the control condition primed intimacy, as a consequence masking differences between conditions.

Discussion

This paper discussed the development and evaluation of a privacy dictionary whose objective is to assist researchers in conducting automated content analysis of texts and transcripts. Most importantly, the prototype perspective that guided the dictionary development integrated numerous theoretical definitions of privacy, thus removing bias that may result from theory-based methods. The privacy dictionary provides a valuable addition to the arsenal of tools available for the study of privacy fulfilling the need for a shared and common methodological platform for privacy research. Through the dictionary, it is now possible to compare a number of different studies. This cumulative information can help to draw meaningful conclusions about particular privacy sensitive domains. In addition, the dictionary complements recent methodological approaches to privacy (whereby participants are not prodded with directive questions) by providing a tool that measures the expression of naturally unfolding experiences. From a purely practical perspective, the dictionary can process large quantities of textual data, and as such supplement laborious and time-consuming human coding by pre-identifying language of interest. In doing so, it also supports new developments in the social sciences and e-research where researchers have increasingly begun to mine naturalistic data from online communities and social media (Tausczik & Pennebaker, 2010). Thus, the privacy dictionary provides researchers with a new resource to measure privacy perceptions as they are expressed in online settings.

Using a contextually rich linguistic dataset, representative of privacy language, the first study found that the eight dictionary categories were used significantly more in the privacy condition relative to the control condition. By order of importance, participants talked about the

realms that privacy protects, e.g., data, secrets (*PrivateSecret* category). They included descriptions anchored on negative privacy experiences, e.g., feeling intruded upon, embarrassed, threatened, (*NegativePrivacy* category), as well as talked about the open behaviours and states that characterise privacy, e.g., post, display, accessible (*OpenVisible* category). Participants described the static behavioural states through which people achieve privacy and the outcomes it serves, e.g. safety, alone (*OutcomeState* category). They used words that referred to intimacy shared with other people or within groups, e.g. trust, closeness (*Intimacy* category), talked about the norms and expectations that govern privacy, e.g., discretion, respect (*NormsRequisites* category) and discussed the various behaviours used to manage and protect privacy, e.g., control, hide (*Restriction* category). Finally, participants used words pertaining to the legal boundaries of privacy, e.g. lawful, offence (*Law* category).

Study 2 used a new dataset that replicated the majority of these findings. In particular, compared to the control condition, participants of the privacy condition used more words from six dictionary categories: *PrivateSecret, NegativePrivacy, OpenVisible, Restriction, NormsRequisites* and *OutcomeState.* The privacy condition of the second dataset contained notably fewer words (9,209 words) than the privacy condition of the first linguistic corpus (182,875 words). Against the more stringent conditions of Study 2, the finding that six categories capture differences between privacy language and non-privacy language indicates that the dictionary categories are robust. As the dataset of Study 2 contained laypeople's descriptions of personal privacy violations, it is not surprising that participants failed to describe the legal dimension of privacy measured through the dictionary category *Law*. The finding that participants of both conditions used equal word rates from the category *Intimacy*, however, is demonstrative of the cautious approach researchers must take to ensure that the dictionary remains an accurate measurement tool. In general, a careful methodology was used to construct the datasets for both studies so that they capture genuine episodes of privacy that will be in turn measurable through language. Yet, Study 2 inadvertently primed descriptions of intimacy in the control condition, a methodological limitation that was only revealed upon closer examination of the nature of the texts. Thus, after the use of any automated content analysis, a second, interpretive stage of analysis (e.g. key word in context analysis) must be undertaken to understand how dictionary words or phrases are used.

In comparing the privacy dictionary categories to the coverage of other popular dictionaries, such as LIWC (Pennebaker et al., 2007), General Inquirer (Stone et al., 1966), or the semantic categories of Wmatrix (Rayson, 2009), it must be acknowledged that the former represent narrower constructs in the psychological landscape. More generally, content analysis categories will often deal with words in the 'long tail' (i.e. those words which form the vast majority of language, but which are used infrequently) and while the words included in the privacy dictionary may not be ubiquitous in everyday language use, our results demonstrate that they are used in a consistent manner when expressing issues surrounding privacy. We note that a similar conclusion has been drawn for some LIWC dictionary categories, e.g., *Sexuality* and *Religion*, whose mean word frequency across 43 studies was 0.2. With the view of understanding how the privacy dictionary may function across datasets of different breadth and depth, differences revealed between the two studies are informative. In both studies, the most frequently used words were from the categories *PrivateSecret, NegativePrivacy* and *OpenVisible*. In Study 2, word rates for these categories were higher which is likely due to the shorter texts collected. Use of words in the remaining five dictionary categories was variable. It is our hope that the widespread use of the privacy dictionary in future research will lead to a cumulative understanding on whether certain dictionary categories appear consistently during the expression of all privacy episodes and if some categories are specific to particular types of contexts.

Potential Applications of the Privacy Dictionary

In light of the particular strengths of automated content analysis (e.g. longitudinal, comparative analysis), we believe that the categories encapsulated in the dictionary can contribute towards investigating a number of longstanding questions about privacy in different contexts. In conducting comparative studies with the dictionary, social scientists may want to analyse privacy preferences across gender, age, socio-economic status or class dimensions, to understand how these differ between groups. This can be further used to support a citizen-centric political agenda by highlighting how citizen privacy priorities differ in the context of technological projects such as the introduction of the NHS Summary Care Record or the national identity cards in the UK for instance. In addition, designers may use the dictionary to develop more appropriate, personalised privacy settings to ensure that privacy is designed-in and safeguarded appropriately according to a diverse range of requirements from different types of

users. There has already been considerable interest in understanding privacy within the highly sensitive contexts of socio-technical environments. In these contexts, researchers have determined that certain components of privacy, such as control and openness, are particularly salient as a result of the constraints and possibilities built into the architecture of technological systems (Palen & Dourish, 2003; Joinson & Paine, 2007; Strickland & Hunt, 2005). This has stimulated an on-going interest in examining whether users are cognizant of the dangers raised in relation to their own technology use (Adams & Sasse, 1999). The privacy dictionary provides the means for comparing technology users' language and the language employed by academics and policy makers to locate where incongruence lies. Explorations into the differences between these two groups can then direct designers' attention to the most problematic aspects of technological design and prove new insights on how to overcome them.

As a tool for longitudinal research, the privacy dictionary can allow researchers to measure the temporal evolution of privacy. For example, using the dictionary, social scientists may want to track people's use of language over time in order to understand how people perceptions, understandings and expectations vis-à-vis privacy have shifted. This can help highlight emerging priorities for policy makers as well as provide historical context for policy debates and decision-making. In exploiting its potential to analyse online materials, technologists may want to measure the levels of privacy language within an online community that has recently undergone design changes (e.g. changes in privacy settings) raising awareness about escalating privacy violations. Technologists can also take preventive steps in launching discussion threads around impeding design changes and through levels of privacy language gauge users' reactions as regards their privacy, further elucidated through a comparison to earlier discussions on successful technological implementations.

Conclusion

In conclusion, privacy has become a critical social issue of the information age (Strickland & Hunt, 2005; Yao et al., 2007). Difficulties in defining privacy have rendered tools for its measurement a key challenge (Patil et al., 2006). In this paper, we proposed a novel technique for measuring privacy through language. The privacy dictionary is a resource that can be used with existing automated content analysis software, such as LIWC 2007. A unifying theoretical framework informed the dictionary words and categories which were first tested and

refined on a dataset of 334 texts balanced between privacy and a control condition, and sampled from a rich variety of contexts e.g. health, social network sites, children and the internet. The dictionary categories were subsequently evaluated with a second dataset containing self-reported privacy violations. It was shown that the dictionary categories could distinguish differences between privacy discussions and general language use. In carving up the space for future research, we provided several examples of possible applications for the dictionary for research, policy and technology development.

Footnotes

(1) To control for dependencies across participants taking part in focus group discussions, the contributions of all participants in a focus group (excluding the facilitator) are included as a single case in our analysis.

(2) In using Google to search for the (approximate) number of results for each of the main four blogging platforms/domains for the previous year the results are as follows: Blogspot.com (850M), Wordpress.com (235M), Livejournal.com (77.6M), and Typepad.com (2.33M).

Supplementary Material

The dictionary is available under the Creative Commons license and can be obtained through correspondence with Asimina Vasalou (minav@luminainteractive.com). More information about the dictionary can be found at www.privacydictionary.info.

Acknowledgements

We thank Anne-Marie Oostveen and Sacha Brostoff whose assistance was invaluable with the transcript analysis, and Francisco Iacobelli for his help with writing the blog collection software. We gratefully acknowledge Will Lowe, Cristina Soriano and the anonymous reviewers of our manuscript whose comments improved our work. This research was funded by the EPSRC through the Privacy Value Networks project (EP/G002606/1) and partially supported by the Future and Emerging Technologies programme FP7-COSI-ICT of the European Commission through the QLectives project (grant no. 231200).

References

Acquisti, A., & Grossklags, J. (2004). Privacy attitudes and Privacy behavior: Losses, Gains, and Hyperbolic Discounting. In Camp & Lewis (Eds.), The Economics of Information Security (pp. 179-186): Kluwer.

Adams, A., & Sasse, A. M. (1999). Taming the wolf in sheep's clothing: privacy in multimedia communications. Proceedings of the seventh ACM international conference on Multimedia (pp. 101-107). ACM.

Altman, I. (1975). The environment and social behavior. Monterey: CA: Brooks/Cole.

Anthony, D., Henderson, T., & Kotz, D. (2007). Privacy in Location-Aware Computing Environments.. IEEE Pervasive Computing, 6 (4), 64-72.

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. Paper presented at the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico.

BBC(2011). Facebook sorry over face tagging launch. Retrieved from June 16, 2011, from http://www.bbc.co.uk/news/technology-12214628

Bradley, M. M. and P. J. Lang (2007). Affective Norms for English Text (ANET): Affective ratings of text and instruction manual. Gainesville, FL., University of Florida.

Buchanan, T., Paine, C., Joinson, A. N., & Reips, U. D. (2007). Development of measures of online privacy concern and protection for use on the Internet. Journal of the American Society for Information Science and Technology, 58(2), 157-165.

Casciani, D. (2009). Q&A: The national DNA database. Retrieved June 16, 2011, from http://news.bbc.co.uk/1/hi/uk/7532856.stm.

Christidi, S., & Rosenbaum-Elliott, R. (2010). Shared spaces and personal corners in social networking websites: the contracted and the cryptically revealed self. European Advances in Consumer Research, 9.

Chung, C., & Pennebaker, J. W. (2007). The Psychological Function of Function Words. In Fiedler (Ed.), Social Communication (pp. 343-359). Psychology Press: New York.

Costello, J. (2001). Nursing older dying patients: findings from an ethnographic study of death and dying in elderly care wards. Journal of Advanced Nursing, 35(1), 59-68.

DeCew, J. W. (1997). In Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Ithaca, Cornell University Press:NY.

DeCew, J. W. (2000). The Priority of Privacy for Medical Information. Social Philosophy and Policy, 17(2), 213-234.

Fehr, B. (1988). Prototype Analysis of the Concepts of Love and Commitment. Journal of Personality and Social Psychology, 55(4), 557-579.

Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The Language of Emotion in Short Blog Texts. In Proceedings of the Conference on Computer Supported Cooperative Work (pp. 1121-1124). ACM.

Hancock, J. T., Curry, L., Goorha, S., & Woodworth, M. T. (2008). On lying and being lied to: A linguistic analysis of deception. Discourse Processes, 45, 1-23.

Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text. In Proceedings of the Conference on Human Factors in Computing Systems (pp. 929-932). ACM.

Harper, J., & Singleton, S. (2001). With a Grain of Salt: What Consumer Privacy Surveys Don't Tell Us. Retrieved June 16, 2011, from http://cei.org/PDFs/with_a_grain_of_salt.pdf

Hart, R. P. & C. Carroll (2011). DICTION: The Text-Analysis Program Thousand Oaks, CA, Sage.

Joinson, A. N., & Paine, C. (2007). Self-disclosure, Privacy and the Internet. In Joinson, McKenna, Postmes & Reips (Eds.), Oxford Handbook of Internet Psychology (pp. 237-252). Oxford University Press: Oxford.

Joinson, A. N., Paine, C., Buchanan, T., & Reips, U. D. (2008). Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. Computers in Human Behavior, 24(5), 2158-2171.

Kramer, A. D. I., Fussell, S. R., & Setlock, L. D. (2004). Text analysis as a tool for analyzing conversation in online support groups. In Proceedings of the Conference on Human factors in computing systems (pp. 1485-1488). ACM.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211–240.

Leyshon, A., Signoretta, P., Knights, D., Alferoff, C., & Burton, D. (2006). Walking with Moneylenders: The Ecology of the UK Home-collected Credit Industry. Urban Studies, 43(1), 161-186.

Livingstone, S. (2006). Children's privacy online: experimenting with boundaries within and beyond the family. In Kraut, Brynin & Kiesler (Eds.), Computers, phones, and the internet: domesticating information technology (pp. 145-167). Oxford University Press: New York, USA.

Lowe, W. (2004). Content analysis and its place in the (methodological) scheme of things Qualitative Methods, 2(1), 25-27.

Mancini, C., Thomas, K., Rogers, Y., Price, B. A., Jedrzejczyk, L., Bandara, A. K., et al. (2009). From spaces to places: emerging contexts in mobile privacy. In Proceedings of the 11th international conference on Ubiquitous computing (pp. 1-10). ACM.

Mazanderani, F., & Brown, I. (2010). Making things private: exploring the relational dynamics of privacy. Paper presented at the Computers, Privacy and Data Protection, Brussels, Belgium.

Meerabeau, L. (2001). The management of embarrassment and sexuality in health care. Journal of Advanced Nursing, 29(6), 1507-1513.

Mehl, M. R., & Gill, A. J. (2010). Computerized Content Analysis. In Gosling & Johnson (Eds.), Advanced Methods for Behavioral Research on the Internet. Washington American Psychological Association Publications.

Nissenbaum, H. F. (2004). Privacy as Contextual Integrity. Washington Law Review, 79(1).

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. Discourse Processes, 43(2), 239-270.

Palen, L., & Dourish, P. (2003). Unpacking "privacy" for a networked world. In Proceedings of the conference on Human factors in computing systems (pp. 129-136). ACM.

Parent, W. (1983). Privacy, morality and the law. Philosophy and Public Affairs, 12, 269–288.

Patil, S., Romero, N., & Karat, J. (2006). Privacy and HCI: methodologies for studying privacy issues. In Proceedings of the conference on Human Factors in Computing Systems (pp. 1719-1722). ACM.

Pattenden, R., & Skinns, L. (2010). Choice, Privacy and Publicly Funded Legal Advice at Police Stations. The Modern Law Review, 73(3), 349-370.

Pedersen, D. M. (1999). Model for types of privacy by privacy functions. Journal of Environmental Psychology, 19(4), 397-405.

Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of LIWC2007. Software manual. Austin, TX.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology, 54, 547-577.

Petronio, S. (2002). Boundaries of privacy: Dialectics of disclosure. State University of New York Press: Albany, NY.

Raento, M., & Oulasvirta, A. (2008). Designing for privacy and self-presentation in social awareness. Personal and Ubiquitous Computing, 12(7), 527-542.

Rayson, P. (2009). Wmatrix: a web-based corpus processing environment. Computing Department, Lancaster University.

Regan, P. (1995). Legislating privacy: Technology, social values, and public policy. University of North Carolina Press: Chapel Hill, NC.

Rosch, E. (1978). Principles of categorization. In Rosch & Lloyd (Eds.), Cognition and categorization (pp. 27-48). Hillsdale, NJ: Erlbaum.

Rule, J. (2007). Privacy in Peril: How We are Sacrificing a Fundamental Right in Exchange for Security and Convenience. Oxford University Press: USA.

Schoeman, F. D. (1984). Philosophical Dimensions of Privacy: An Anthology. Cambridge University Press: Cambridge.

Solove, D. J. (2006). A taxonomy of privacy. University of Pennsylvania Law Review, 154 (3), 477-564.

Solove, D. J. (2008). Understanding Privacy. Harvard University Press.

Strickland, L. S., & Hunt, L. E. (2005). Technology, security, and individual privacy: New tools, new threats, and new public perceptions. Journal of the American Society for Information Science and Technology, 56(3), 221–234.

Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., & Associates. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29(1), 24-54.

Tavani, H. T. (2007). Philosophical theories of privacy: Implications for an adequate online privacy policy. Metaphilosophy, 38(1), 1-22.

BBC News. (2003). New rules for workplace snoopers. Retrieved on June 16, 2011, from http://news.bbc.co.uk/1/hi/technology/2981120.stm

Vasalou, A., Joinson, A. N., & Courvoisier, D. (2010). Cultural differences, experience with social networks and the nature of "true commitment" in Facebook. International Journal of Human-Computer Studies, 68(10), 719-728.

Vasalou, A., Joinson, A. N. and Houghton, D. A Prototype Analysis of Privacy (2010). Retrieved on June 16, 2011 from SSRN: http://ssrn.com/abstract=1865752

Westin, A. (1967). Privacy and freedom. Athenaeum: New York.

Wittgenstein, L. (2001). Philosophical Investigations. Blackwell: Oxford.

Yao, M. Z., Rice, R. E., & Wallis, K. (2007). Predicting user concerns about Online privacy. Journal of the American Society for Information Science and Technology, 58(5), 710-722.