Ambiguity in Privacy Policies and Perceived Privacy Risk

Jaspreet Bhatia

CMU-ISR-18-108
May 2019

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Travis D. Breaux (Chair)
James D. Herbsleb
Eduard Hovy
Joel R. Reidenberg (Fordham University)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Software Engineering.

The research reported in this thesis has been supported by Institute for Software Research at Carnegie Mellon University under award numbers National Science Foundation Award CNS-1330596, National Science Foundation CAREER Award No. 1453139, National Security Agency Award No. 141333 and Office of Naval Research Award No. N00244-16-1-0006. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of Carnegie Mellon University and the U.S.Government.

Keywords: privacy requirements, privacy, privacy policies, natural language, ambiguity, incompleteness, semantic frames, semantic roles, crowdsourcing, perceived privacy risk, multilevel modeling, factorial vignettes

Abstract

Software designers and engineers make use of software specifications to design and develop a software system. Software specifications are generally expressed in natural language and are thus subject to its underlying ambiguity. Ambiguity in these specifications could lead to different stakeholders, including the software designers, regulators and users having different interpretations of the behavior and functionality of the system. One example where policy and specification overlap is when the data practices in the privacy polices describe the website's functionality such as collection of particular types of user data to provide a service. Website companies describe their data practices in their privacy policies and these data practices should not be inconsistent with the website's specification. Software designers can use these data practices to inform the design of the website, regulators align these data practices with government regulations to check for compliance, and users can use these data practices to better understand what the website does with their information and make informed decisions about using the services provided by the website. In order to summarize their data practices comprehensively and accurately over multiple types of products and under different situations, and to afford flexibility for future practices these website companies resort to using ambiguity in describing their data practices. This ambiguity in data practices thus undermines its utility as an effective way to inform software design choices, or act as a regulatory mechanism, and does not give the users an accurate description of corporate data practices, thus increasing the perceived privacy risk for the user.

In this thesis, we propose a theory of ambiguity to understand, identify, and measure ambiguity in data practices described in the privacy policies of website companies. In addition, we also propose an empirically validated framework to measure the associated privacy risk perceived by users due to ambiguity in natural language. This theory and framework could benefit the software designers by helping them better align the functionality of the website with the company data practices described in privacy policies, and the policy writers by providing them linguistic guidelines to help them write unambiguous policies.

Dedicated to

My parents,
for giving me wings.

My husband, Amit,
for being the wind beneath my wings.

Acknowledgments

The time I have spent at CMU as a PhD student has been a life-changing experience for me, and I could not have done it without the support of many people whom I would like to thank.

First, I would like to express my sincere gratitude to my incredible advisor, Prof. Travis Breaux. Travis has been a great advisor, awesome teacher, and most importantly, one of the best mentors I have ever had. I would like to thank Travis for being so involved in all aspects of my PhD: from guiding me towards the right set of literature and helping me set up my experiments, to providing detailed feedback on the papers I write. For all these five years of my PhD, I have always looked forward to my research meetings with Travis. His enthusiasm, optimism and confidence about my research always give me so much hope. Thank you for helping me shape my career, teaching me to be a good researcher and for helping me grow as a person. I am very fortunate to have you as my PhD advisor, and as a mentor for life.

Besides my advisor, I would like to thank my thesis committee members, Prof. Joel Reidenberg, Prof. Jim Herbsleb and Prof. Eduard Hovy, for their insightful feedback on my work and their hard questions which motivated me to think more deeply about my research. The numerous discussions I had with Joel, helped me widen my own vision about privacy from both technology and law perspectives. I would also like to thank ISR staff members Connie and Margaret for all their help in making sure my paperwork was in place, requirements were being met, and my bills were being reimbursed.

I am also grateful to my awesome collaborators and my colleagues: Hanan, Morgan, Shurui, Hemank, Roykrong, Sudarshan, Thomas, Daniel and Mitra, for all the great discussions about research and about life. My PhD would not have been half as fun without you guys. I have also been very lucky to share all these years with my friends Khushboo, Pranav, Pragya (and little Ved), Neeharika, Dheeru, Maria and Mani. I take away from Pittsburgh some great memories and friends for life!

I am eternally indebted to my parents for their innumerable sacrifices and selfless love. Thank you for your support and blessings; for instilling in me the values that have made me the person I am today; and for teaching me perseverance. Whatever I am and whatever I hope to be, I owe it you both.

I would like to thank my sister and brother, Priya and Gurpreet who have been there for me since I was 5 years old and have always kept me grounded. They both have been a constant source of encouragement, support and joy. Thank you for always having my back. I would also like to thank my parents-in-law for their encouragement, positivity and love.

Finally, I would like to thank my husband Amit, who deserves equal credit for this PhD as me. Thank you for pushing me to apply to CMU for my PhD; for believing in my dreams when no one else did, and for your unwavering support and unconditional love every step of the way. And most importantly, thank you for putting up with me through all my insanities in this roller coaster of a PhD. I could not have done this without you!

Contents

List of Tables

List of Figures

Chapter 1

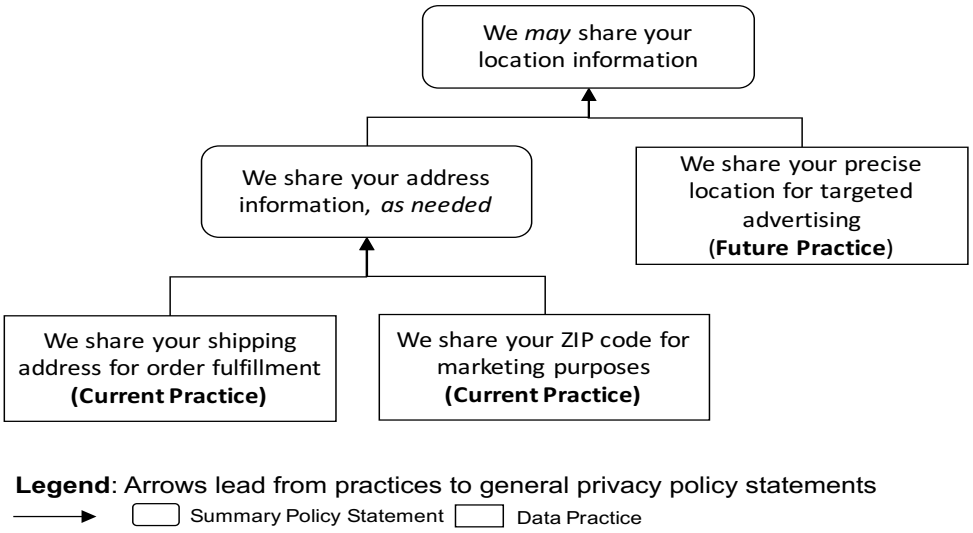
Introduction

Companies and government agencies use personal information to improve service quality by tailoring services to individual needs. To support privacy, regulators rely on the privacy notice requirement, in which organizations summarize their data practices to increase user awareness about privacy. These notices, also called privacy policies, further serve to align company privacy goals with government regulations. In addition, software designers and developers use the data practice descriptions in these privacy policies to inform the design of the website, and to make decisions related to user data such as "what information should be collected from the user?" "what should that information be used for?" "who should be given access to the data?" among other decisions. Users use these privacy policies to better understand the data practices of the company, and in turn make informed decisions about using the website. The underlying ambiguity in privacy policies, however, undermines the utility of such notices to serve as design guidelines for the software designers, and as effective regulatory mechanisms that could be used to check for compliance with the government regulations. Consequently, privacy policies also fail to offer a clear description of the organization's privacy practices to users and in turn effect their ability to make an informed decision about the website. The ambiguity could lead to multiple interpretations of the same data practice by different stakeholders, including the regulators, software designers and engineers, and the users.

Privacy policies pose a challenging requirements problem for organizations, because policies must: (a) be *comprehensive*, which includes describing data practices across physical places where business is conducted (e.g., stores, offices, etc.), as well as web and mobile platforms; and (b) be *accurate*, which means all policy statements must be true for all data practices and systems. Ensuring privacy policies are *comprehensive* and *accurate* means that policy authors can resort to ambiguity when summarizing their data practices, which includes using vague terms to describe their data practices and using incomplete description of the data practices. Variations in data practices may exist because two or more current practices that are semantically different must be generalized into a broader category of statement.

In Figure 1, the data types "shipping address" and "ZIP code" are generalized into "address information," and the purposes "order fulfillment" and "marketing purposes" are combined into a vague condition "as needed," to encompass both practices. To account for future practices, a vague modal verb "may" is added to the general policy statement, while "address" is subsumed by "location information", and the purpose is removed.

Figure 1. Example data practices that are generalized into privacy policy statements



Ambiguity (as shown in Figure 1) can cause different stakeholders to be confused about the actual data practices of the website. For instance, in the example in Figure 1, the ambiguity makes difficult for the stakeholders to accurately predict different aspects of the data practice and does not answer questions such as: (1) "what constitutes their *location information*?" (2) "what are the conditions under which the user's information will be shared?" (due to the presence of the keyword "may"), (3) "with whom is the location information being shared?" (due to the absence of the value for the semantic role "target" i.e. who/what is the receipt of the user's information), and (4) "what will the shared data be used for?" (due to the absence of the value for the semantic role "purpose" i.e. for what purposes will the user's information will be used). This lack of clarity in the information provided in the privacy policy about their data practices could have the following consequences: the software designers would not be able to align the functionality of the system with the company's data practices; the regulators may not be able to accurately align the data practices with government regulations to check for inconsistencies and violations; and finally, it could affect the users' decision making about their use of the website services. Ambiguity can in turn cause users to perceive higher privacy risk, because the flexibility entailed by ambiguous policy statements may conceal privacy-threatening practices. Moreover, ambiguous statements can limit an individual's ability to make informed decisions about their willingness to share their personal information, which may also increase their perceived privacy risk.

Ambiguity can also lead to users or regulators coming to incomplete or inconsistent conclusions, due to the missing or unclear information in the privacy policies which they assume or comprehended in an incorrect way. Consequently, it can lead to misestimation of the privacy risk. For example, in the summary privacy statement "we may share your location information" in Figure 1, the purpose for which the user's location information is shared is missing, which gives the user a chance to make an assumption about the missing purpose. The user may assume that the sharing is being undertaken for a primary purpose, which leads to underestimating the risk. On the other hand, the user may assume that the shared data is used for a secondary purpose, which leads to overestimating the risk, while it remains unknown what the actual data practice is. The overestimation of privacy risk is not a favorable situation for the company,

because it could lead to either the users not using their services due to fear of misuse of their data, or the regulators concluding that the data practice is not in compliance with a regulation. In 2015, social networking website and application Snapchat changed its data practice descriptions in their privacy policy concerning collection, use and retention of users data, stating that "…we may access, review, screen, delete your content at any time and for any reason" and "…publicly display that content in any form and in any and all media or distribution methods," among other such statements which made the users worried about the ways in which their information could be collected, retained and used[1], since the policy was not very clear about it. This led to some users reporting that they had deleted their accounts[2]. In another incident, Google was warned by European regulators for being vague about its data retention practices and not showing commitment towards the European Data Protection Directive[3]. To ensure accuracy, we believe business analysts and system developers, in addition to legal advisors, must participate in deciding which practices to summarize in a privacy policy, and when to use ambiguity to minimize the privacy risk.

Some researchers believe that one can measure the "actual" privacy risk, which is a hypothetical, data subject-independent measure of the above-chance probability that any data subject would experience a privacy harm. The concept of an "actual" privacy risk would require continuous surveillance data on data subjects, which details how a system affects those subject's emotional, psychological and physical well-being. This data would include whether data subjects accept a risk by participating in an activity. Fischhoff et al. argue that human behavior does not reliably reflect an actual risk estimate, if they cannot iterate over the system's design space, including both the possibility of hazards and reliability of safety features [Fischhoff et al. 1978]. In addition, accumulating this surveillance data would introduce a privacy risk paradox, in which the measurement of actual risk would introduce a new, more serious risk by amassing this surveillance data. Finally, the measure of whether a data subject actually experiences a privacy harm, such as whether a data subject's personal information were distorted or mischaracterized, is necessarily a subjective assessment. Fischhoff et al. argue that such assessments are subject to estimator biases and their methods of assessment, if not well documented, can be difficult to reproduce [Fischhoff et al., 1978]. Therefore, while actual privacy risk presents an objective ideal, the concept's general validity and reliability has been criticized in prior work. In this thesis we measure perceived privacy risk, which is based on expressed preferences [Slovic 2000] and which we define as an individual's willingness to share their personal data with others given the likelihood of a potential privacy harm.

In the next section, we discuss in detail our approach to identify and measure ambiguity and the associated perceived privacy risk.

1.1 Proposed Approach

Ambiguity undermines the ability of organizations to align their privacy policies with their actual data practices, which can confuse or mislead users, thus leading to an increase in perceived privacy risk. This thesis examines the presence of ambiguity, which consists of vagueness and

[1] Alex Heath, "Why you don't need to freak out about Snapchat's new privacy policy," Business Insider, 30 October 2015. http://www.businessinsider.com/snapchat-privacy-policy-update-explained-2015-10

[2] Sally French, "Snapchat's new 'scary' privacy policy has left users outraged," Market Watch, 2 November 2015. http://www.marketwatch.com/story/snapchats-new-scary-privacy-policy-has-left-users-outraged-2015-10-29

[3] Zack Whittaker, "Google must review privacy policy, EU data regulators rule," ZDNet, 16 October 2012. http://www.zdnet.com/article/google-must-review-privacy-policy-eu-data-regulators-rule/

incompleteness in data practices, and its effect on perceived privacy risk. The outcome of this thesis is a theory of ambiguity in privacy policies, which includes an approach to: (1) understand, identify and measure vagueness; (2) understand and detect incompleteness using semantic roles; (3) and understand and measure perceived privacy risk due to ambiguity.

We study the concept of *vagueness* in privacy policies, which is caused by the use of vague terms, that reduce the clarity of the data practices. We consider a privacy policy statement as *vague* when words such as "may," "generally," "some," etc. are used to describe the data practices. We studied vagueness present in privacy policies by conducting grounded analysis [Saldaña 2012] on privacy policies, and we measured the relative differences in vagueness of vague terms by performing user studies. Based on the findings from these studies we propose a theory of vagueness which consists of three main parts: a taxonomy of vague terms and their categorization which is based on grounded analysis, a technique to measure the relative inter-and intra-category vagueness using paired comparisons, and an explanation for differences in vagueness based on different semantic functions. We used techniques from natural language processing (NLP), to develop a vagueness scoring tool that is based on the results from the different vagueness studies we conducted (see Chapter 4 for details).

We also analyze incompleteness due to missing contextual information about data actions in data practices using grounded analysis. *Incompleteness* occurs in privacy policies when it does not answer all the questions the users or regulators may have regarding the company's data practices. For example, in context of the data action "share," the questions that one could have include: what type of data is being shared? what is the source of the data being shared? with whom is it shared? for what purpose is it shared? and under what conditions will it be shared? If the data practice does not answer one or more of these questions, the data practice can be considered incomplete with respect to the data action. From our analysis we concluded that context for a given data action can be represented using *semantic frames*. We construct these semantic frames by answering different questions about that data action, which are called *semantic roles* associated with the action. Our approach is a grounded analysis to discover which semantic roles corresponding to a data action are needed to construct complete data practice descriptions. Failure to provide the values for different semantic roles for a given data action can lead to incompleteness in describing the context for that action. We have investigated what semantic roles are expected for different data actions for a complete semantic frame representation, how do these roles help build the context for the action, and how are these roles expressed in privacy policies. We have also developed a hybridized framework to semi-automatically identify privacy goals in privacy policies (see Chapter 5 for details). Both vagueness and incompleteness cause ambiguity and prevent us from making accurate predictions about how the user's data is collected, retained, shared or used by the company. The constructs vagueness and incompleteness, in addition to other factors such as risk likelihood and demographic factors, etc. inform the design of our empirically validated framework to measure perceived privacy risk (see Chapter 6 for details).

An approach to identify and measure ambiguities and the associated privacy risk can benefit software designers, policy writers, regulators and users. Users and regulators use the privacy policies to understand the data practices of the website, that is to understand what the company says it does with their data, whereas what the company actually does with the user data is reflected in their software design. Software designers can therefore use our approach to identify ambiguity in the data practices, and ask for clarifications when required, so that there are no gaps between what the company says it does with user data, and what it actually does. This would

help the website company make sure that the website's functionality is in sync with the data practices described in the privacy policy. Software designers can consequently also use the data practices from the privacy policy to inform their design decisions during the development of the website.

Using the theory and framework proposed herein, policy writers can identify the ambiguity in the data practices and take measures to reduce this ambiguity such that it provides an accurate description of the website's data practices and reduces the assumptions the stakeholders have to make. The theory and the corresponding linguistic guidelines that emerge from this thesis can help policy writers understand when and how to summarize their data practices in order to reduce the ambiguity and the associated privacy risk. For example, if the company is sharing the user's data with a third-party company, the privacy policy should provide details about the purposes for which the data would be shared, if that has been shown to reduce the associated privacy risk. Regulators need a means to identify if the data practices of a website align with the laws and government regulations. Regulators can use the proposed approach to identify ambiguous data practices and score privacy policies for ambiguity. This could help them identify ambiguous data practices which can lead to inconsistencies and non-compliance, and suggest corrective measures to website companies which have a privacy policy with high ambiguity score, or with high associated privacy risk.

In the future, we envision extension points to our approach that can be used with other privacy related research ideas such as those of nutrition labels for privacy [Kelley et al. 2009]. Our results can be used to adjust how they help users make privacy related decisions. The results from our thesis can also be used to augment the findings of NLP and ML tools being developed to automatically process privacy policies [Bhatia et al. 2016b, Sathyendra et al. 2017] by helping these tools process the instances of ambiguity as special cases.

In addition, we envision that the empirically validated framework to measure privacy risk can be used by itself with different contexts, by developers, public policy, regulators and users. System developers, including designers, aim to build systems that users feel comfortable and safe using. In privacy, this includes accounting for Privacy by Design (PbD) [Hustinx 2010], wherein the user's privacy is considered throughout the development of the system. To perform PbD, however, developers need a systematic and scalable framework that can help them understand and measure the privacy risk that users experience while using a software system. Developers can use this privacy risk framework to frame their design choices in a given context and then measure how users perceive the risks that arise due to the context, so that designs can be improved to reduce risk. For instance, if a particular information type or data practice is high risk, designers may introduce risk mitigations to affect the storage and use of that information. This may include limiting collection from the user or encrypting the information before it is stored; and also, the policy writers could pay more attention to describing more clearly the data practices associated with the sensitive information types. This framework can also help regulators identify systems that could put users' privacy at greater risk and suggest corrective measures. Furthermore, known high-risk data practices and information can be used to introduce privacy nudges [Acquisti et al. 2017 and Wang et al. 2014] to users in real-time based on user demographics associated with high perceptions of risk. On the other hand, if data subjects misunderstand a technology and consequently perceive it as high risk, public policy could be used to explain the technology and provide additional guidance to reduce the risk in data handling.

In summary, this thesis aims at building a theory of ambiguity for privacy policies that provides an early, novel foundation upon which to improve the summarization of data practices and readability of these privacy policies, which are known to be hard to read [McDonald and Cranor 2008], in a way that they minimize the associated privacy risk. In addition, it aims to enhance emerging techniques for automating the processing of privacy policies [Bhatia et al. 2016b, Sathyendra 2017].

Chapter 2

Thesis Statement

***Thesis Statement**: Ambiguity undermines the ability of organizations to align their privacy policies with their actual data practices, which can confuse or mislead users, thus leading to an increase in perceived privacy risk. This thesis examines the presence of ambiguity, which consists of vagueness and incompleteness in data practices, and its effect on perceived privacy risk. The outcome of this thesis is a theory of ambiguity in privacy policies, which includes an approach to: (1) understand, identify and measure vagueness; (2) understand and detect incompleteness using semantic frames; (3) and understand and measure perceived privacy risk due to ambiguity and vagueness.*

We present the background and related work in Chapter 3. In Chapter 4, we explain in detail the grounded analysis for identifying vague terms in data practices and the user studies for measuring the relative vagueness of these vague terms in privacy policies that lead to the formation of the theory of vagueness. In Chapter 5 we describe our approach to identify semantic roles for data actions in privacy policies, and the hybridized framework to identify privacy goals semi-automatically. In Chapter 6 we present the empirically validated framework for understanding and measuring perceived privacy risk. In Chapter 7 we present future work. And finally, in Chapter 8 we summarize the research work.

Chapter 3

Background and Related Work

This Chapter reports the background and related work on (1) ambiguity in natural language and in requirements (2) semantic roles and (3) privacy and privacy risk.

3.1 Ambiguity in Natural Language and Requirements

Lakoff notes that natural language (NL) concepts have vague boundaries and fuzzy edges. Consequently, he introduced the term hedging to describe the fuzziness in the truth value of NL sentences, meaning, that they are true to a certain extent, and false to a certain extent, true in certain respects and false in certain other respects [Lakoff 1972]. In natural language processing (NLP), machine learning (ML) systems have been developed as part of the CoNLL-2010 shared task to identify hedge cues and their scopes in Wikipedia and Biomedical texts [Farkas et al. 2010].

Requirements are often written in NL and thus suffer from inherent NL ambiguity [Berry et al. 2003]. For example, Yang et al. report that, out of the 26,829 requirements statements that they analyzed, 12.7% had ambiguity due to a coordinating conjunction (and/or), which is a type of syntactic ambiguity [Yang et al. 2010]. Ambiguity is often considered a potentially dangerous attribute of requirements [Boyd et al. 2005]. Gause and Weinberg note that ambiguity in requirements can lead to subconscious disambiguation, wherein readers disambiguate using their first interpretation, unaware of other possible interpretations [Gause 1989]. This leads different stakeholders with different interpretations of the same requirements. Ambiguity detection is difficult, even if the reader is aware of all the facets of ambiguity [Kamsties 2006].

Table 1 presents Massey et al.'s ambiguity taxonomy that was applied to natural language legal texts [Massey et al. 2014]. In this thesis proposal, we focus on vagueness from the use of vague terms, and incompleteness due to missing semantic roles in context of a data action.

Table 1. Ambiguity Taxonomy for Legal Text

Type	Definition
Lexical	a word or phrase with multiple, valid meanings, also called polysemy
Syntactic	a sequence of words with multiple valid grammatical interpretations regardless of context
Semantic	a sentence with more than one interpretation in its provided context
Vagueness	a statement that admits borderline cases or relative interpretation
Incompleteness	a grammatically correct sentence that produces too little detail to convey a specific or needed meaning
Referential	a grammatically correct sentence with a reference that confuses the reader based on the conduct

Many attempts have been previously made to address the problem of ambiguity in requirements. Fuchs and Schwitter propose Attempto Controlled English, a restricted NL, to align NL specifications with first order logic to reduce the ambiguity in requirements [Fuchs and Schwitter 1995]. However, restricted or formal languages are not as expressive as NL, and incorrectly interpreted NL specifications lead to incorrect formal specifications [Tjong 2013].

Alternatively, Berry et al. introduced the Ambiguity Handbook, which describes ambiguity in requirements and legal contracts, including strategies for avoiding and detecting ambiguity [Berry et al. 2003].

Pattern based techniques have also been used to identify ambiguity in requirements [Kamsties 2001, Denger 2002]. Kiyavitskaya et al. propose a tool that combines lexical and syntactic measures applied to a semantic network to identify ambiguous sentences and determine potential ambiguities [Kiyavitskaya et al. 2008]. Alternatively, object oriented analysis models of the specified system can be used to identify ambiguities [Popescu et al. 2008]. Tjong describes ambiguities found in NL requirements, such as lexical ambiguity, ambiguity due to uncertainty, etc., and guidelines to avoid these ambiguities [Tjong 2008]. The tool called SREE identifies instances of a set of vague words using simple keyword matching and marks it as potentially ambiguous [Tjong and Berry 2013]. In our approach, we do not employ keyword matching, because we do not consider all instances of a vague term to be potentially vague. Instead, we rely on manual annotations to identify vague terms.

Requirements quality evaluation tools, such as IBM Doors and QuARS [Fabbrini et al. 2001] and ARM [Wilson et al. 1997], also identify ambiguous terms. Yang et al. identify speculative requirements and uncertainty cues, using a technique that combines ML and a rule-based approach. They utilize lexical and syntactic features of requirements to identify uncertainty [Yang et al. 2012]. More recently, researchers have used ML based on heuristics drawn from human judgments to identify nocuous coordination and anaphoric ambiguities in requirements [Yang et al. 2010, Yang et al. 2011]. This approach still requires human interpretation to resolve ambiguity. To our knowledge, this prior work to identify vague requirements terms [Berry et al. 2003, Kamsties et al. 2001, Tjong 2008, Tjong and Berry 2013, Fabbrini et al. 2001, Wilson et al. 1997, Yang et al. 2012] does not differentiate the relative vagueness of these terms. We address this limitation with a new vagueness taxonomy and predictions of how vague terms increase and decrease vagueness.

3.2 Semantic Roles

We identify incompleteness in data practices by determining which of the expected roles for a data action are missing values in data practice statements. In order to determine the expected roles that will help us better understand a data action, we need to answer questions associated with that action, such as *who performs the action* and *on what data the action was performed*, among other questions [Jurafsky and Martin 2000]. The answers to these questions can be expressed in many different ways in a statement. For example, consider the following data practice statements:

- `We collect user information.`
- `The user information is logged by us.`
- `We gather information about our users.`
- `The user provides us with their information.`

While the above statements use different action words, such as *collect, log, gather,* and *provide* and have different syntax, they also have similar meaning, which is that the user information is collected by the website. One representation that permits comparison among these statements is called semantic roles [Jurafsky and Martin 2000]. Roles are considered shallow representations, because they rely only on the relationship between a given word or role value and other clauses in the statement, and not among all the words in the statement. Using semantic roles, we represent the fact that there is a collection action taking place, the action is being

performed by the *subject* the website company, and the *object* of the action is the user information. Semantic roles represent the relationship of the different clauses in the statement to the main action, like the subject and object [Jurafsky and Martin 2000]. The context of a data action can be expressed using different semantic roles, such as agent (who initiates and performs an action), patient (what undergoes the action and changes its state), instrument (used to carry out the action), source (where the action originated), among other roles [Gruber 1965].

Semantic roles that are used to describe a data action can be represented together in a knowledge representation technique known as *frames*. Minsky describes a frame as a data structure that is used to represent a stereotyped situation, such as being in a certain kind of living room [Minsky 1981]. Each frame is associated with *slots* or semantic roles, which are filled by *fillers* or semantic role values in specific contexts, and which help readers understand a situation in question. The values for these semantic roles can be atomic values, procedures, or pointers to other frames [Minsky 1981]. Frames can be used to represent knowledge in a succinct manner and to reason in an efficient way [Fikes and Kehler 1985].

According to Fillmore's frame semantics, the meaning of a word cannot be understood in isolation but in conjunction with the related information [Fillmore 1976]. For example, the word "share" can be understood when we have knowledge about who is sharing, what is being shared, and with whom it is being shared. Fillmore's frame semantics are implemented in the FrameNet project [Baker et al. 1998]. The FrameNet corpus contains manually annotated, general purpose semantic frames for the English language, with semantic roles specific to a frame. The frames are evoked by lexical units which are lemmas and their part of speech. The semantic roles associated with each frame are also known as frame elements, which provide information about the frame. Consider the following example from the FrameNet database:

```
Abby bought a car from Robin.
```

In this statement, the frame "commerce_buy" is evoked by the lexical unit "bought (buy.verb)". The frame elements of this frame instantiated in this statement are: buyer (Abby), goods (a car), seller (robin). Similar to FrameNet, our frames are evoked by different categories of data actions, which represent a situation where the user's information is being acted upon by a company. We employ semantic roles that are specific to each such frame and are instantiated when that frame is evoked. The FrameNet resource has been used for automatic semantic role labelling [Das et al. 2014, Roth and Lapata 2015]. Das et al. report an F1 score of 61.4 and 68.49 for frame identification and semantic role value identification respectively for SemEval 2007 data, and F1 score of 80.3 and 79.9 for frame identification and role value identification respectively on the FrameNet 1.5 release [Das et al. 2014]. Semantic role labelling has been used for improving applications such as question-answering [Kaisser and Webber 2007], recognizing textual entailment [Braz et al. 2005], information extraction [Surdeanu et al. 2003] and in requirements engineering, to extract information from software requirements specifications [Wang 2015].

Semantic role labelling (SRL) is a type of shallow semantic parsing with the objective of determining the predicate-argument structure for each predicate in a statement [Jurafsky and Martin 2000, Zhou and Xu 2015]. The techniques used for developing SRL systems can be categorized into two main groups: (1) traditional methods using syntactic features with machine learning classifier, and (2) end to end systems with word embeddings and neural networks. The first and the most widely used method (till recently) is the tradition method which involves extracting syntactic and lexical features from text which are then used with different classifiers to develop a SRL system [Gildea and Jurafsky 2002, Carreras and Màrquez 2005, Cohn and Blunsom 2005, Mitsumori et al. 2005]. The emphasis is on extracting features that can best

describe the properties of the text from the training corpus [Zhou and Xu 2015]. The most important features come from the combination of different syntactic parsers. Pradhan et al. treat SRL as a multi-class classification problem and use features generated from the syntactic parses from Charniak parser [Charniak 2000] and Collins parser [Collins 2003], and then assign constituents of each parse a semantic role label using support vector machine classifier (SVM) [Pradhan et al. 2005]. They then convert the semantic role labels into BIO tags (*beginning-inside-outside* of the semantic role span) [Ramshaw and Marcus 1995], which are used as input features as well with another SVM layer which produces the final SRL tags. The combination of the features from these three different syntactic views leads to significant improvement in performance over features from individual views. In the 2005 CoNLL shared task, 19 teams participated and developed different SRL systems using varied syntactic information such as part of speech tagging, chunking, syntactic parses, and named entities, and various learning algorithms including SVMs, CRFs, maximum entropy frameworks and other such variations [Carreras and Màrquez 2005].

The traditional methods rely heavily on the output of the different syntactic parsers, and Pradhan et al. showed that errors in the syntactic parsing are major sources of errors in the SRL systems [Pradhan et al. 2005]. And therefore, more recently the focus has been on techniques based on word embeddings and neural networks, which try to solve the SRL problem without using feature engineering. Collobert et al. introduced an architecture for SRL, which consists of a word embedding layer, convolution neural network (CNN) layers, and a CRF layer [Collobert et al. 2011]. They used word embedding which are trained on a large corpus of text, to address the problem of data sparsity [Zhou and Xu 2015]. However, they had to use features from parse tree of the Charniak parser [Charniak 2000] in order to perform as well as the traditional methods. They also used CNN layer which does not model long term dependencies as well as other types of neural networks since it only includes words in a limited context [Zhou and Xu 2015]. To overcome this limitation, we plan to use long short-term memory architecture, which can model long term dependencies [Hochreiter and Schmidhuber 1997]. In the past few years the focus has been on developing end to end systems which do not have any intermediate tag, and the only input they use is the statement, the verb of interest, and the word embeddings for the words in the statement. Zhou and Xu have developed such a system which takes as input the word embeddings, and use deep bi-directional LSTMs to perform the task of SRL [Zhou and Xu 2015]. He et al. use deep highway bi-directional LSTMs to develop their SRL system [He et al. 2017]. They also observe that syntactic parser can be used with their system to further improve their results. In this thesis, we evaluate an end to end system to identify information type semantic role which uses LSTMs as the machine learning algorithm and word embeddings as the input to the system.

In the next section, we describe in detail the studies we conducted to identify and measure vagueness in privacy policies and their results.

3.3 Privacy and Privacy Risk

In this section, we review background and related work on privacy, risk perception and privacy risk.

3.3.1 Background on Privacy

Over the course of the last century, multiple definitions of privacy have emerged. Westin describes privacy as when a person, group or company can decide for themselves when, how and to what extent information about them is shared with others. Westin defines four states of privacy: (1) *solitude*, which refers to how one person distances his or herself from others, (2) *intimacy*, where a person chooses to have a close relationship with a small group of people, (3) *anonymity*, where a person can move through public spaces while protecting his or her identity, and (4) *reserve*, where a person can regulate the amount of information about himself or herself that one wants to communicate to others in order to protect against unwanted intrusion [Westin 1967]. Murphy describes the "right to privacy" as being safe from intrusion, the right to make confidential decisions without government interference, the right to prohibit public use of a person's name or image, and to regulate the use of personal information [Murphy 1996]. Nissenbaum argues that privacy and data sharing are contextual, meaning that the factors, data type, data recipient, and data purpose among others affect a person's willingness to share [Nissenbaum 2004, 2009]. Consistent with this argument made by Nissenbaum we observed that contextual factors including data type, type of harm, purposes which provide societal benefits and the person who is experiencing the risk effect users' perception of privacy risk [Bhatia et al. 2018b]. In this thesis, we also propose to study how the presence or absence of different contextual factors, which are also called *semantic roles* associated with the data action effect a user's perception of privacy risk (See Chapters 5 and 6.3).

There are different and conflicting views about the importance of privacy. Solove argues that privacy is "a fundamental right, essential for freedom, democracy, psychological well-being, individuality, and creativity" [Solove 2008]. On the other hand, other scholars, such as Moor, argue that privacy is not a "core value" in comparison to the values of life, happiness, and freedom; rather privacy is an expression of the core value of security and asserts that privacy is instrumental for protecting personal security [Moor 1997].

Studies have shown differences between a user's privacy preferences and their actual behavior in similar situations, called the privacy paradox [Acquisti and Grossklags 2005, Berendt et al. 2005]. This paradox could be explained by the argument made by Slovic et al. that people who see social or technological benefits of an activity tend to perceive a reduction in risks associated with that activity [Slovic 2000]. The studies reported in our paper further support this argument, that perceived benefits from services will reduce the users' perception of privacy risk [Bhatia et al. 2018b].

3.3.2 Risk Perception and Privacy Risk

Risk is a multidisciplinary topic that spans marketing, psychology, and economics. In marketing, risk is defined as a choice among multiple options, which are valued based on the likelihood and desirability of the consequences of the choice [Bauer 1960]. Starr first proposed that risk preferences could be revealed from economic data, in which both effect likelihood and magnitude were previously measured (e.g., the acceptable risk of death in motor vehicle accidents based on the number of cars sold) [Starr 1969]. In psychology, Fischhoff et al. note that so-called revealed preferences assume that past behavior is a predictor of present-day preferences, which cannot be applied to situations where technological risk or personal attitudes are changing [Fischhoff et al. 1978]. To address these limitations, the psychometric paradigm of perceived risk emerged in which surveys are designed to measure personal attitudes about risks and benefits [Slovic 2000]. Two insights that emerged from this paradigm and inform our

approach are: (a) people better accept technological risks when presented with enumerable benefits, and: (b) perceived risk can account for benefits that are not measurable in dollars, such as lifestyle improvements, which includes solitude, anonymity and other definitions of privacy [Slovic 2000]. In other words, people who see technological benefits are more inclined to see lower risks than those who do not see benefits. Notably, privacy is difficult to quantify, as evidenced by ordering effects and bimodal value distributions in privacy pricing experiments [Acquisti et al. 2013]. Rather, privacy is more closely associated with lifestyle improvements, e.g., private communications with friends and family, or the ability to avoid stigmatization. Acquisti et al. observed that estimated valuations of privacy were larger when the participants of the study were asked to consider giving up their personal data for money and smaller when they had to pay money for privacy [Acquisti et al. 2013]. Their studies also showed that the participants' decisions about privacy were inconsistent. Finally, the economist Knight argues that subjective estimates based on partial knowledge represent uncertainty and not risk, also known as ambiguity aversion, wherein respondents are unwilling to accept a risk due to uncertainty in the question or question context [Knight 1921].

Chapter 4

A Theory of Vagueness

Creswell defines a theory as an interrelated set of constructs formed into propositions and hypothesis that specify the relationship among variables, typically in terms of magnitude and direction [Creswell 2008]. To that end, our three-part vagueness theory is: (1) the construct vagueness is described by multiple, exclusive semantic categories; (2) the categories, independently and through composition, predict how vagueness increases and decreases; and (3) semantic functions, called likelihood, authority and certitude, suggest why semantic categories predict vagueness [Bhatia et al. 2016a]. In addition, we used this theory to develop a vagueness scoring mechanism to compare the relative vagueness of privacy policies. The vagueness scores for a set of privacy policies are then compared to those for two benchmarks to determine whether government-mandated privacy disclosures result in notices less vague than those emerging from the market [Reidenberg et al. 2016].

The use of vague terms, such as *may*, *as necessary*, and *generally*, to describe goals in privacy policies introduces uncertainty into the goal's action or the associated information type. Consider the following statements:

- *We will share your personal information, such as your name, email address and phone number, with our marketing affiliates for advertising purposes.*
- *We might share some of your personal information with our third-party affiliates as necessary.*

In the first statement, the modal phrase *will* is certain, whereas the modal phrase *might* in the second statement leaves open the possibility of sharing, and is thus vague. In addition, the first statement elaborates upon what *personal information* is included, *name, email address* and *phone number*, which adds additional clarity missing from the second statement, which mentions sharing *some* of your *personal information*. Similarly, the description of the purpose *advertising purpose*s is more clear than the phrase *as necessary*, which leaves open a range of possible purposes, such as *legal*, *marketing*, etc.

In this section, we report the two studies we conducted and the results which led to the development of a theory of vagueness for privacy policies, and a third study where we used the results from this theory to score privacy policies for vagueness [Bhatia et al. 2016a, Reidenberg 2016]. The first study which we call Study V1 was based on content analysis [Saldaña 2012] to identify vague terms in privacy policy statements and to categorize them into different vagueness categories, and the second study which we call Study V2 used paired comparison technique [David 1988] and Bradley Terry Model [David 1988, Hunter 2004] to measure the relative differences in the vagueness of these vagueness categories and terms by ranking them in the order of vagueness. We then used the results from these two studies to conduct a third study (Study V3), which was aimed at scoring policies for vagueness and comparing it to benchmark policies.

We describe the content analysis study (Study V1) in Section 4.1, the paired comparison study (Study V2) in Section 4.2, the vagueness scoring study (Study V3) in section 4.3, and we

then report the results from these three studies in Section 4.4. We summarize the conclusions from the all the vagueness studies in Section 4.5.

4.1 Content Analysis of Vague Terms

We manually annotated 15 privacy policies (see Table 2) using content analysis [Saldaña 2012] to identify words or phrases that introduce vagueness into policy statements for **Study V1**. We limited our analysis to statements about *collection, use, disclosure* and *retention* of personal information, which have also been discussed by Antón and Earp [Antón and Earp 2004]. These policies are part of a convenience sample, although, we include a mix shopping companies who maintain both online and "brick-and-mortar" stores, and we chose the top employment websites and Internet service providers in the U.S. Table 2 presents the 15 policies by category and date last updated.

Table 2. Privacy Policy Dataset for Vagueness Study

Company's Privacy Policy	Industry Category	Last Updated
Barnes and Noble	Shopping	05/07/2013
Costco	Shopping	12/31/2013
JC Penny	Shopping	05/22/2015
Lowes	Shopping	04/25/2015
Over Stock	Shopping	01/09/2013
AT&T	Telecom	09/16/2013
Charter Communication	Telecom	05/04/2009
Comcast	Telecom	03/01/2011
Time Warner	Telecom	09/2012
Verizon	Telecom	10/2014
Career Builder	Employment	05/18/2014
Glassdoor	Employment	09/09/2014
Indeed	Employment	2015
Monster	Employment	03/31/2014
Simply Hired	Employment	4/21/2010

The policies are first prepared by removing section headers and boilerplate language that does not describe relevant data practices, before saving the prepared data to an input file for an Amazon Mechanical Turk (AMT) task. The task employs an annotation tool developed by Breaux and Schaub [Breaux and Schaub 2014], which allows annotators to select relevant phrases matching a category, in this case, the vague terms belonging to a certain category. I, and two graduate law students, performed the annotation task.

The annotation process employs two-cycle coding [Saldaña 2012]. In the first cycle, I analyzed five policies to identify an initial set of vague terms, and then applied second-cycle coding to group these terms into emergent categories based on the kind of vagueness introduced by related terms. In addition, I developed guidelines to predict into which category a vague term should be placed. The terms, categories and guidelines were shared with the other two annotators, who independently annotated the same five policies. Next, I and the other two annotators met to discuss results, to add new terms to the categories and to refine the guidelines. After agreeing on the categories and guidelines, we annotated the remaining ten policies, before meeting again to reconcile disagreements. Saturation was reached after no new vague terms or new categories were discovered, which occurred after analyzing the first five policies (Barnes and Noble, Lowes, Costco, AT&T, and Comcast).

The resulting vagueness categories and their definitions are:

- *Conditionality* – the action to be performed is dependent upon a variable or unclear trigger
- *Generalization* – the action or information types are vaguely abstracted with unclear conditions
- *Modality* – the likelihood or possibility of the action is vague or ambiguous
- *Numeric Quantifier* – the action or information type has a vague quantifier

This approach is also known as grounded theory in literature [Saldaña 2012]. The guidelines help disambiguate the policy statement in a given context, for example, the phrase "as necessary" when followed by a specific purpose: "We will use your personal information as necessary for law enforcement purposes…" states that the information is used for legal purposes, thus disambiguating the condition "as necessary" in this context.

We use the semi-automated privacy goal-mining framework developed by Bhatia et al. to identify statements with privacy goals [Bhatia et al. 2016b]. This technique was extended to use the Stanford Dependency Parser [Marneffe et al. 2006] to automatically identify which annotated vague terms are attached to either an action or information type in the privacy goal. The resulting vagueness dataset consists only of privacy goals with a vague term attached to either the action or information type.

We applied Fleiss' Kappa, an inter-rater agreement statistic [Fleiss 1971], to the annotations-vagueness category mappings. Because Fleiss' Kappa assumes that categories are exclusive, we compute the Kappa statistic for the complete composition of all vagueness categories assigned to each policy statement. A statement that contains one or more *Modality* category terms is assigned to the singleton category *M*, whereas a statement with terms from a combination of the *Conditionality*, *Generality* and *Modality* categories is assigned to the composite category *CGM*. The Fleiss Kappa for all mappings from annotations to vagueness categories and the three annotators was 0.94, which is a very high probability of agreement above chance alone.

4.2 Ranking Vagueness Categories and Terms using Paired Comparisons:

In **Study V2**, we measured the differences in vagueness within and across vagueness categories and their combinations. Paired comparison is a statistical technique used to compare *N* different items by comparing just two items at once [David 1988]. The overall results are computed by combining data from all paired comparisons. This technique is especially useful when items are comprised of multiple factors, when the comparison context is difficult to control, or when the comparison order influences the outcome. This technique is beneficial when differences between items are small, and when comparison between two items should be as free as possible from any extraneous influence caused by the presence of other entities [David 1988]. To compare *N* entities, a total $N * (N-1)/2$ paired comparisons are performed.

We designed multiple surveys to compare combinations of one or more vague terms, within and across the four vagueness categories. The first survey is an exploratory survey designed to compare statements containing combinations of vague terms from across the four vagueness categories (see Section 4.1). We chose one exemplary vague term from each category. The vague terms were then inserted into a baseline privacy policy statement: "We share your personal information." For example, variants 1 and 2 below show two statements that result from inserting the underlined vague terms selected from the corresponding vagueness categories (in parenthesis):

Variant 1 (Modality, Condition): *We <u>may</u> share your personal information <u>as necessary</u>.*

Variant 2 (Numeric Quantifier): *We share some of your personal information.*

For the four vagueness categories, we have 2^4-1 or 15 category combinations and thus one statement variant per combination. The 15 statement variants yield 105 paired comparisons.

The survey consists of a scenario, and five of 105 paired comparisons (see Figure 2). The scenario frames the survey rationale for the participants.

Figure 2. Paired Comparison Survey Questions

Instructions: A company wants to improve the clarity of their website privacy policies. Therefore, they are considering alternative language to help users better understand what their data practices are. For each numbered question, please read each pair of statements, and identify which of the two statements best represents a more clear description of the company's treatment of personal information.

For example, a clear description of the company's treatment of personal information could be "*We share your personal information such as your name and contact details, as needed for legal purposes*."

In the following statement, any pronouns "We" or "Us" refer to the company, and "you" refers to the user.

1. Which one of the following statements is a more clear description of the company's treatment of personal information than the other?

- We may share your personal information.
- We share some of your personal information, as needed.

The number of participants needed to judge each paired comparison was based on Pearson and Hartley's data for calculating power for paired comparisons [Pearson and Hartley 1962, 1966]. To attain 95% power, at least four participants are needed to judge each paired comparison. We solicited 60 participants to judge each paired comparison. The additional 56 participants only reduce standard error to further delineate between vagueness levels; four participants are sufficient to discover rank order.

We designed four additional surveys based on the design shown in Figure 2 to measure intra-category vagueness. For the intra-category vagueness surveys, each survey has a total $N * (N - 1)/2$ paired comparisons for N vague terms in the corresponding vagueness category. We use the Bradley Terry model, which estimates the probability that one item is chosen over another item using past judgments about the items [David 1988, Hunter 2004], to determine the rank order of the vague terms. Model fitting is either by maximum likelihood, by penalized quasi-likelihood (for models which involve a random effect), or by bias-reduced maximum likelihood in which the first-order asymptotic bias of parameter estimates is eliminated [Turner and Firth 2012]. The Bradley Terry model has been implemented using statistical R package [R 2013, Turner and Firth 2012].

4.3 Scoring Privacy Policies for Vagueness

The objective of **Study V3** was to develop a vagueness scoring model for privacy policies and to determine if the benchmark privacy policies were more or less vague as compared to market privacy policies. We observed that simply counting the number of vague terms in a privacy policy will not provide an adequate measure of vagueness. For example, the AT&T policy

contains 70 vague phrases, which places it at the median of 70 vague phrases and just below Time Warner, which has 85 vague phrases. But this frequency count does not indicate the relative context. Context matters, and a granular scoring model needs to take into account three key variables: (1) the existence of vague terms and their relation to specific categories of data practice (e.g., collection, retention, sharing, and usage); (2) the relative impact that a combination of vague terms may have on overall ambiguity; and, (3) the completeness of the policy. To accomplish this goal, we propose a scoring model based on a relative comparison of vagueness in phrases for each policy. This score is based on a statistical measure that scales the overall vagueness of individual statements in each policy based on the Bradley-Terry model for paired comparisons. To calculate the score for each of the data practice statement with a vague attachment we use the Bradley-Terry coefficients from the study described in Section 4.2 above. The vagueness scores appropriately ignore phrases that do not specifically describe a data processing activity or that do not contain any vague terms. This means that non-relevant language, such as a corporation's philosophy relating to privacy, or unambiguously described data practices will not factor into the vagueness score. For each policy, we can then calculate an aggregate vagueness score by taking the sum of the coefficients for each action-information pair containing vague terms. This policy-specific aggregate score is not, however, sufficient to compare two policies. For example, if a policy is long, it may contain more action-information pairs containing vague terms than a shorter policy, but proportionately be much clearer. To account for this situation, we normalize the aggregate vagueness score by dividing the aggregate score by the total number of action-information pairs in the policy; we call this normalized score the vagueness score. The vagueness score reflects positively on the policy and improves if a policy has more action-information pairs that clearly describe data practices and reflects negatively on the policy and worsens if the policy has more pairs that include vague terms. Moreover, it reflects the total unit vagueness independent of policy length, but relative to the level of contribution to vagueness by the vagueness categories. This can be represented by the following equation:

$$V = \frac{\Sigma\ (BTC_{A\text{-}I})}{\Sigma\ (A\text{-}I)} \qquad (1)$$

where *V* is vagueness score, *BTC* is the Bradley-Terry coefficient, and *A-I* is the action-information pair.

Lastly, in the event that a policy has a high level of vagueness in paragraphs pertaining to key elements that may be masked by clear language elsewhere in the policy, we calculate the vagueness scores for the collection of policy statements addressing each of the four key data practices: *collection*, *retention*, *sharing* and *usage*. These scores are calculated in the same manner as those for the overall policy. Separately, we report on the completeness of the privacy policies using a scale of 0 to 4. For each element missing from the four data practices (collection, retention, sharing and use), the policy is assigned one point. Thus, a policy containing any description for all four elements will score a 0 and a policy missing all four elements will score a 4.

4.4 Summary Results from the Vagueness Studies

In this section, we summarize our results from the three vagueness studies - Study V1, V2, and V3 described above in Sections 4.1, 4.2 and 4.3.

4.4.1 Vagueness Taxonomy from Content Analysis

In this section we describe results from **Study V1**. In Table 3 we present the content analysis results applied to the 15 policies in Table 2. The categorization was done by me and checked by the other two annotators. The frequency represents the number of times the term appeared across all selected statements in the 15 policies. Table 4 presents a breakdown of number of terms per category that appear across all 15 policies and the privacy goal types present in the policy (C: Collection, R: Retention, T: Transfer, U: Use).

Table 3. Taxonomy of Vague Terms

Category	Vague terms	% Freq.
Conditionality (C)	depending, necessary, appropriate, inappropriate, as needed	7.9%
Generalization (G)	generally, mostly, widely, general, commonly, usually, normally, typically, largely, often	4.0%
Modality (M)	may, might, can, could, would, likely, possible, possibly	77.9%
Numeric Quantifier (N)	certain, some, most	10.1%

Table 4. Frequency of Vague Terms Across Policies

	Policy	Vagueness				Goal Types			
		C	G	M	N	C	R	T	U
Shopping	Barnes & Noble	12	4	98	17	55	7	47	48
	Costco	6	7	50	1	47	12	70	43
	JC Penny	6	0	29	5	31	2	31	30
	Lowes	2	0	62	6	61	16	16	54
	OverStock	1	1	19	3	9	2	10	14
Telecom	AT&T	3	0	52	0	41	4	47	77
	Charter Comm.	8	4	81	12	46	16	70	48
	Comcast	20	9	91	9	30	18	68	56
	Time Warner	1	6	47	18	24	12	29	27
	Verizon	14	1	101	12	57	13	83	87
Employment	Career Builder	1	3	28	4	24	14	13	52
	GlassDoor	5	3	42	6	30	13	19	34
	Indeed	0	1	33	4	19	13	25	57
	Monster	3	0	28	1	31	20	23	38
	Simply Hired	1	3	55	8	37	9	12	44

4.4.2 Vagueness Ranking using Paired Comparison

We describe results from **Study V2** in this section. In Section 4.2 we describe a method for rank ordering exemplar terms selected from each vagueness category to measure how vagueness varies within and across categories, and how do vague terms interact in combination to affect overall vagueness. The selected terms are *as needed (C), generally (G), may (M),* and *some (N)*. The survey was conducted on Amazon Mechanical Turk (AMT), and each paired comparison was judged by 60 participants, who were paid $0.12 to judge five paired comparisons at once. We analyze the paired comparisons using the Bradley-Terry (BT) model; the BT model coefficients and standard error appear in Table 5.

Table 5. Bradley Terry Coefficients

Vagueness Category	Coefficient	Standard Error
CN	1.619	0.146
C	1.783	0.146
CM	1.864	0.146
CMN	2.125	0.146
CG	2.345	0.146
CGN	2.443	0.146
MN	2.569	0.146
N	2.710	0.146
M	2.865	0.147
CGMN	2.899	0.147
CGM	2.968	0.147
GN	3.281	0.149
GMN	3.506	0.150
G	3.550	0.150
GM	4.045	0.156

C: Conditionality, G: Generality, M: Modality, N: Numeric Quantifier

Figure 3 presents the BT coefficients and standard error in an annotated scatter plot to show the linear relationship of vagueness categories and their combination. The coefficients show the quantity that each vague term contributes to the overall concept of vagueness. The data practices described with combinations to the left of Figure 3 (CN, C, CM, …) have greater clarity than practices described with combinations to the right of Figure 3 (GMN, G, GM, …). For example, while phrases with both a *conditional* term and *numeric* quantifier (CN) are statistically indistinguishable compared to phrases with only a *conditional* term (C), we observe how the vagueness taxonomy influences overall vagueness. In Figure 3, the red arrow from MN to CMN shows a *condition* term increases clarity and reduces vagueness: statements with both a *modal* term and *numerical quantifier* (MN) are significantly less clear than similar statements with an added *conditional* term (CMN). The blue arrow from MN to GMN shows how *generalization* increase vagueness: the MN statements with the added *generalization* (GMN) are significantly more vague. By comparison, statements with a *generalization* and *modal* term (GM=4.045) are twice as vague as statements with a *condition* and a *modal* term (CM=1.864).

Figure 3. Bradley Terry Coefficients

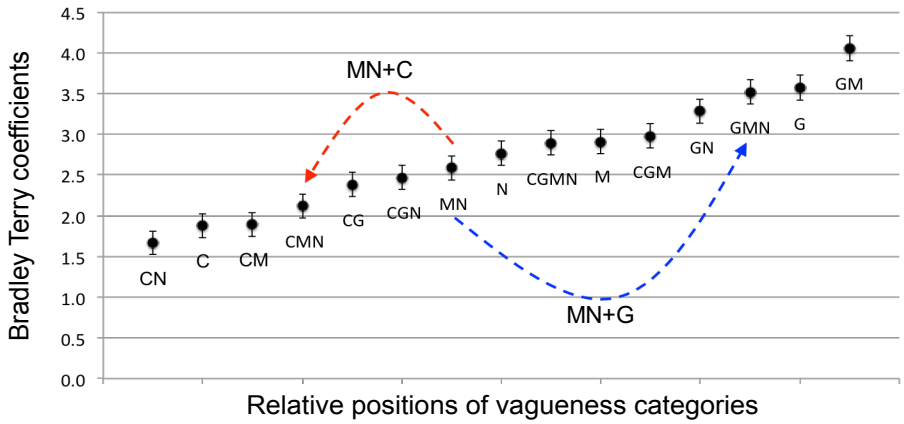


Table 6 presents the BT coefficients for intra-category vagueness: the shaded rows present the model intercepts, which consist of the vague terms in the inter-category survey. In the *Conditionality* category, "as appropriate" was several times more vague than "as necessary". Under *Generalization*, the vagueness appears to increase as the adverbs transition from the

routine (e.g., typical, normal or usual) to the unrestricted (e.g., widely, largely, mostly). Under *Modality*, the past tense verbs "might" and "could" are perceived to be more vague than the present tense variants, "may" and "can", respectively.

Table 6. Bradley Terry Coefficients for Intra-Category Vagueness

	Vague term	Coefficient	Standard Error
Conditionality	as needed	0.00	0.00
	as necessary	0.01	0.15
	as appropriate	0.70	0.14
	depending	0.77	0.14
	sometimes	1.20	0.15
	as applicable	1.37	0.15
	otherwise reasonably determined	1.52	0.15
	from time to time	1.81	0.15
Generalization	typically	-0.38	0.11
	normally	-0.34	0.11
	often	-0.15	0.11
	general	-0.11	0.11
	usually	-0.04	0.11
	generally	0.00	0.00
	commonly	0.03	0.11
	among other things	0.64	0.11
	widely	0.67	0.11
	primarily	0.70	0.11
	largely	1.25	0.13
	mostly	1.71	0.14
Num. Q.	certain	-0.53	0.22
	most	-1.21	0.24
	some	0.00	0.00
Modality	likely	-0.32	0.13
	may	0.00	0.00
	can	0.42	0.13
	would	0.60	0.13
	might	0.76	0.13
	could	0.96	0.14
	possibly	1.78	0.15

4.4.3 Computing Vagueness Scores for Privacy Policies

We apply our scoring model described in Section 4.3 to our privacy policy dataset (Table 2), and two benchmarks, with five policies for each benchmark for **Study V3**. Because the score ratios are designed to compare the clarity of policies against each other and do not provide a minimum level of acceptability for vagueness, the Model Privacy Form adopted under the Gramm Leach Bliley Act can serve as an informative target benchmark for a regulated notice. This model form was adopted by regulatory agencies after careful analysis and testing of language options [Levy and Hastak 2008]. The language used in this standardized privacy disclosure statement has been approved by eight federal financial service regulatory agencies. Financial service providers may use the model form to satisfy their obligations under the Gramm-Leach-Bliley Act, though they are not required to adopt its language. The second benchmark are the companies which are part of the US-EU Safe Harbor Agreement. Out of a total of 15 policies in our dataset, five policies are part of the EU Safe Harbor. The EU Safe Harbor identifies data practices that must be contained and described in a privacy policy to satisfy European data export requirements, but stops short of providing model language like the Model Privacy Form in the United States. The

framework was negotiated between the US and Europe and then approved by the US Department of Commerce. Companies may benefit from the EU Safe Harbor if they include specified provisions in their privacy notices and register with the US Commerce Department.

We report the results of applying the scoring model described in Section 4.3 to the privacy policies of companies that do not have specific notice obligations, and our two benchmarks - national financial institutions that adopted privacy policies based on the Model Privacy Form and Safe Harbor companies in Table 7. When the ratios are in proximity to each other, they indicate that those policies have similar levels of vagueness. Where a ratio is double another, the ratios indicate that the policy with the higher ratio is twice as vague as the policy with the lower ratio.

Table 7. Vagueness Scores for Unregulated Companies Privacy Policy

	Privacy Policy	Total Score	Collect	Retain	Share	Use	Completeness
Unregulated Policies	Costco	1.02	0.68	0.95	1.51	0.63	0
	JC Penny	1.19	1.32	1.44	1.16	1.07	0
	Lowes	1.28	0.87	2.15	2.06	1.25	0
	OverStock	1.71	1.56	1.44	2.03	1.62	0
	AT&T	1.04	0.92	0.45	1.25	0.99	0
	Charter Comm.	1.64	1.54	1.02	1.72	1.84	0
	Comcast	1.80	1.71	1.75	1.96	1.66	0
	Time Warner	2.09	2.1	2.79	1.72	2.17	0
	Verizon	1.38	1.41	0.80	1.48	1.34	0
	Simply Hired	1.56	1.44	0.64	1.12	1.97	0
	Mean	1.36	1.34	1.60	1.45	1.47	0
Financial Institutions using Model Privacy Form	Bank of America	0.96	0.48	2.87	1.03	0	0
	Capital One	0.52	0.58	2.87	0.38	0	0
	Citi Group	0.45	0.58	-	0.43	0	1
	JP Morgan	0.36	0.48	0	0.56	0	0
	PNC	0.35	0.58	-	0.31	0	1
	Mean	0.52	0.54	1.91	0.54	0	
Safe Harbor Companies	Barnes & Noble	2.07	2.19	1.49	2.3	1.78	0
	Career Builder	0.84	0.83	0.81	0.89	0.85	0
	GlassDoor	1.36	1.41	1.23	1.54	1.26	0
	Indeed	0.96	0.8	1.08	1.04	0.94	0
	Monster	0.79	0.86	0.72	1.12	0.58	0
	Mean	1.20	1.22	1.07	1.38	1.08	

Table 7 shows that the most ambiguous policies among the unregulated entities belong to Time Warner, with Comcast, Overstock, and Charter Communications clustered close behind. These policies use large numbers of vague modal verbs and quantifiers. For example, the Comcast policy describes sharing with third-parties using both a modal verb and numeric quantifier: "*In certain situations, third party service providers may transmit, collect, and store this information on our behalf to provide features of our services.*" By contrast, Costco's language describing sharing with third parties is more direct: "*We do not otherwise sell, share, rent or disclose personal information collected from our pharmacy pages or maintained in pharmacist records unless you have authorized such disclosure, or such disclosure is permitted or required by law.*"

By comparison to these most vague policies, the policies belonging to Costco and AT&T are almost twice as clear. Table 7 also shows the vagueness scores for actions to collect, retain, share and use information. The overall mean vagueness across these four data actions varies little from 1.34-1.60; however, the mean variance is not homogenous across practices (collect variance =0.21, retain variance=0.52, share variance=0.10, and use variance=0.30). This variance across practices shows divergent uses of vague terms across companies, with the least consistency across policy descriptions of retention practices, and the most consistency around descriptions of sharing practices. Notably, companies such as Comcast, and Time Warner score higher than average vagueness in all four data practice categories. For the website user, however, Overstock's high vagueness score for sharing (2.03) presents a more significant, or fundamentally different, privacy risk than Comcast's vagueness regarding collection (1.71) and retention (1.75). Vagueness with respect to sharing is significant because third parties are rarely identified in privacy policies and most privacy policies disclaim responsibility for the data practices of the unnamed third parties. Vagueness with respect to collection and retention affords companies greater flexibility in broadening what kinds of information they are potentially collecting. This may or may not present heightened privacy risks. However, when combined with vague sharing terms, website users will not be able to ascertain exactly what information may be at risk of sharing with third parties. All the policies not subject to regulation were complete.

The mean vagueness score for the financial services policies is considerably lower than the Safe Harbor policies: 0.52 vs. 1.20. This striking two-plus fold difference means that financial services policies are more than twice as clear as the Safe Harbor policies. Similarly, the vagueness scores show that the descriptions of three of the four data practices found in the financial services policies have greater clarity than those found in the Safe Harbor policies. As a benchmark, the Model Privacy Form for the financial services industry holds privacy policies to a higher standard of clarity and allows less vagueness than the US-EU Safe Harbor.

All the benchmark policies were complete with the exception of the Citi Group and PNC policies that were silent on data retention.

4.5 Summary Conclusions for the Theory of Vagueness

In this section, we summarize our results for the vagueness studies Study V1, V2, V3 [Bhatia et al. 2016a, Reidenberg et al. 2016].

We categorized the vague terms we identified in privacy policies into four broad categories: conditionality, generalization, modality and numeric quantifier. From the inter- and intra-category vagueness results, we theorize that differences in clarity may be due to one of three semantic functions: *likelihood*, which is the possibility that something is true; *authority*, which is whether an action is discretionary or mandatory; and *certitude*, which is the absoluteness with

which something is true. For example, “likely” is more clear than “possibly,” both of which concern the degree or likelihood that a data practice occurs. *Authority* refers to whether the practice is permitted, required or prohibited, and it may be true that required practices are perceived as more clear than permitted practices: “as needed” is perceived as more clear than “as appropriate.” Similarly, the vague term “may” denotes both permissibility and possibility, and is perceived to be more clear than “can,” which denotes capability and not necessarily authority. Concerning *certitude*, “as needed” and “normally” describe minimal versus routine behavior, respectively. These two vague terms may have a higher degree of absoluteness than “generally,” which assumes the existence of unstated exceptions, and which is perceived to be more vague and less clear than “as needed” and “normally.”

Goals are formulated at different levels of abstraction and refined using sub-goals, which provides a natural mechanism for structuring complex specifications at different levels of concern [Lamsweerde 2009]. A theory of vagueness that accounts for variants of summarization, i.e., *likelihood, authority,* and *certitude*, can be used to augment goal refinement patterns by introducing formalized notions of vague terms. For example, the coarse-grained privacy goal “May share personal information” can be refined into finer-grained sub-goals using OR-refinement to surface the specific situations that a user’s personal information will and will not be shared. Regarding certitude, “mostly” implies larger coverage of cases where a goal will be achieved, whereas “typically” could emphasize common cases at the exclusion of boundary cases, and thus yield a lower frequency of achievement. The vague terms “likely” and “possibly” can indicate planned features for a future system version.

Comparing the vagueness scores for the regulated financial benchmark policies (mean vagueness score=0.52) against the unregulated policies (mean vagueness score=1.36) shows that the unregulated policies have notably higher scores and use significantly more vague language (see Table 7). The findings indicate that more specific regulation of policy language has a positive impact on the clarity with which privacy policies describe data practices.

Chapter 5

Semantic Incompleteness in Privacy Policy Statements

In this Chapter, I describe the work we have done to represent privacy policy statements as semantic frames to identify incompleteness [Bhatia et al. 2016b, Bhatia and Breaux 2018a, Bhatia et al. 2018]. I first describe our study to identify semantic roles that are associated with four categories of data actions: collection, retention, usage, and transfer, and the corresponding incompleteness in the policy statements [Bhatia and Breaux 2018a, Bhatia et al. 2018]. Then, I describe our hybridized framework to identify privacy policy goals which are action, information type pairs [Bhatia and Breaux 2015, Bhatia et al. 2016b]. Finally, I present my work on evaluating a deep learning architecture for automatically identifying information types.

5.1 Semantic Roles in Privacy Policies

Companies describe their data practices in privacy policies to inform users about how their data must be collected, used and transferred for the purposes embodied by the website or software. U.S. regulators may check these policies for compliance with actual data practices, when a data breach or data misuse arises. Consequently, the statements in policies represent legal requirements for software systems. Ideally, users can also use these policies to better understand what the website does with their personal information and to make informed decisions about using the services provided by the website.

A company's data practice description in a privacy policy can govern multiple types of products, across both physical and virtual stores. In addition, policies are drafted to account for current practices, as well as to afford flexibility for future practices that the company envisions. In doing so, companies resort to using ambiguity in the data practice descriptions of their policies. In the worst case, this ambiguity can lead to inaccurate interpretations by users and regulators.

Privacy policy statements correspond to privacy goals and requirements. Incompleteness in requirements can lead to misunderstanding among stakeholders, wherein stakeholders have different interpretations regarding the incomplete information [Dalpiaz 2018]. Incomplete privacy goals convey to developers a potentially inaccurate description of requirements that should be met by the system. Incomplete requirements are one of the most critical challenges faced by software companies and are also a frequent cause of project failures [Fernández and Wagner 2015].

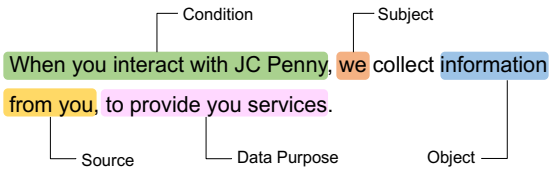
Incompleteness, which is a form of ambiguity, occurs in data practice descriptions when one or more policy statements do not answer all the questions that users or regulators may have regarding the company's data practices. For example, with respect to the data action "share," one could ask: what type of data is shared? With whom will the data be shared? From whom was the data collected? For what purpose is the data shared? Finally, under what conditions will the data be shared? If the data practice description does not answer one or more of these questions, the description can be considered incomplete with respect to the missing information.

Incompleteness in privacy goals and requirements can prevent users from making accurate predictions about how their data is collected, retained, shared or used by the company, consequently causing users to misestimate their personal privacy risk. For example, in the summary privacy statement: "We may share your location information," the purpose for which the user's location information is shared is missing, which requires the user to make assumptions about the missing purpose. The user may assume that the sharing is undertaken for a primary purpose for which the data was collected. For example, the purpose to provide services requested by the user, which leads to underestimating the risk. Alternatively, the user may assume that the shared data is used for an unstated, secondary purpose, either by a first party or third party [Bellman et al. 2004]. Secondary use can lead to overestimation of the privacy risk by users, despite that the third party's data practice remains unknown.

The overestimation of privacy risk is not a favorable situation for a company, because it can lead to either the user not using a service due to fear of data misuse, or it can lead to the regulator concluding that the data practice is not in compliance with a regulation. In 2015, the social networking website and application Snapchat changed its data practice descriptions in their privacy policy concerning collection, use and retention of their user data, stating that "…we may access, review, screen, delete your content at any time and for any reason" and "…publicly display that content in any form and in any and all media or distribution methods." Such statements led users to worry about the ways in which their information could be collected, retained and used, since the policy was extremely permissive. This led some users to report that they had deleted their accounts[4]. In another incident, Google was warned by European regulators about vagueness in their policy concerning data retention practices and about not showing a commitment towards the European Data Protection Directive[5]. Therefore, companies should identify when a data practice is incomplete and take corrective measures to improve the description.

In our research, we identify incompleteness by representing a data practice description namely a data action as a semantic frame. We construct these frames by identifying relevant questions for each data action, which we call semantic roles associated with the action. We propose to develop a network of semantic frames to determine the roles that are expected to complete a data practice description. In so doing, we aim to understand how roles contribute context for an action, and how policy authors choose roles when expressing privacy policies. For example, the following JCPenny privacy policy statement is annotated for semantic roles that describe the data action collect in Figure 4. The condition on the action collect is "when you interact with JC Penny", the object is "information," the source of the information is "you," and the purpose of collection is "to provide you services."

Figure 4. Example statement with annotated semantic roles



[4] Sally French, "Snapchat's new 'scary' privacy policy has left users outraged," Market Watch, 2 November 2015. http://www.marketwatch.com/story/ snapchats-new-scary-privacy-policy-has-left-users-outraged-2015-10-29

[5] Zack Whittaker, "Google must review privacy policy, EU data regulators rule," ZDNet, 16 October 2012. http://www.zdnet.com/article/google-must-review-privacy-policy-eu-data-regulators-rule/

5.1.1 Content Analysis for Identifying Semantic Roles

We manually annotated semantic roles in 15 privacy policies from three domains: health, news and shopping for **Studies SR1, SR2** and **SR3**. We conducted a survey to identify the types of websites that users most frequently use and found that news and shopping websites were most frequently used by our survey participants. Most of our participants reported that they read news online several times a day and shopped for products online a few times a week or more [Bhatia and Breaux 2018b]. In addition, we chose to study the health domain, since it is a highly-regulated domain and deals with sensitive user data. We chose a convenience sample of five policies per domain (see Table 8). For health, we chose companies that provide a diversity of services (DNA testing, online medical records, health clinics, wearable devices, and an online symptom dictionary.) For news, we chose websites with a diversity of U.S. viewpoints. Finally, for shopping we chose companies that maintain both online and "brick-and-mortar" stores. These choices were intended to diversify the observed practices.

Table 8. Privacy Policy Dataset For Semantic Frame Study

Domain	Company Name	Last Updated
Health	23andMe	10/14/2015
	HealthVault	09/2016
	Mayo Clinic	10/06/2014
	My Fitness Pal	06/11/2013
	WebMD	03/20/2015
News	ABC News	10/18/2016
	Bloomberg	07/15/2014
	CNN	07/31/2015
	Fox News	10/26/2016
	Washington Post	01/01/2015
Shopping	Barnes and Noble	08/05/2016
	Costco	12/31/2013
	JC Penny	09/01/2016
	Lowes	08/20/2015
	Overstock	06/20/2017

We annotated the policies in Table 8 using content analysis, in which an analyst assigns codes to text from a coding frame [Saldaña 2012]. Each coded text fragment represents an instance of the code, after which the analyst can review the coded items for insight into the phenomena of interest. Our analysis is limited to statements about collection, retention, usage, and transfer of personal information, which were first studied by Antón and Earp in their seminal paper on privacy goal mining [Antón and Earp 2004].

We prepare the policies for annotation by removing section headers and boilerplate language and itemizing the policy into individual statements. In each statement, we identify the main data action and categorize the statement into one of five categories: *collection, retention, usage, transfer* and *other*. We only analyze the statements which belong to the first four categories, excluding others. Statements that belong to the *others* category are of the following kind, shown with examples from the policy named in parentheses:

- *Definitions* (Costco): "Personal information is information that identifies an individual or that can be reasonably associated with a specific person or entity, such as a name, contact information, Internet (IP) address and information about an individual's purchases and online shopping."

- *User actions* (Barnes and Noble): "You may also access, correct or change the personal information in your community profile(s) on SparkNotes.com at any time, except to change your username."
- *Scope of the privacy policy* (Lowes): "This Privacy Statement applies to the US practices of Lowe's Companies, Inc. and its US operating subsidiaries and affiliates except as outlined below."
- *Customer relations* (Overstock): "If you have questions about your order, you should direct them to us and not to the Vendor."

For **Study SR1** we use the frame-based markup developed by Breaux and Antón to identify semantic roles associated with different data actions [Breaux and Antón 2007]. The tool allows us to use first cycle coding [Saldaña 2012] and to segment the statement by identifying the phrases that correspond to roles, while accounting for variability in the statement due to logical conjunctions and disjunctions. The markup is then parsed to generate lists of roles based on each action and syntactic cue, which we discuss later. Consider the following example, which annotated statement using the tool and which is from the Lowes privacy policy:

```
[[This information] may be used {to [provide a better-tailored shopping
experience]}, |and {for [<market research, | data analytics, | and system
administration> purposes]}.]
```

The guidelines we use to annotate the statements are as follows:

- *Square brackets* are used to denote role fillers that are required to make the statement grammatically correct. For example, in the statement above, the object [this information] is required.
- *Curly brackets* are used to denote clauses that can be removed, which typically correspond to optional roles. For example, `{to [value]}` and `for [value]}` curly-bracketed clauses in the statement above can be removed and the sentence would still be grammatically correct; however, if the words "to" and "for" are present, then the nested role values within the square brackets would be required for the statement to make grammatical sense. For instance, in the statement above, if we remove the roles in the "to" and "for" patterns, the statement would become: "This information may be used." Each statement is enclosed in a square bracket to demarcate sentence boundaries.
- *Angular brackets* are used when a phrase or clause contains alternative sub-clauses among which at most one sub-clause is needed to produce a grammatically correct sentence. For example, the phrase "and for" above applies to all phrases inside the angular brackets.

After the annotation process, we code the extracted phrases in curly brackets using open coding [Saldaña 2012] to assign semantic role names to these phrases. Example annotation-coded pairs are as follows:

- `[this information]`: object
- `{to [provide a better-tailored shopping experience]}`: data purpose
- `{for [<market research, | data analytics, | and system administration> purposes]}`: data purposes

In this statement, the lexical and syntactic patterns `{to [value]}` where value is "provide a better-tailored shopping experience," and `for [value]}` where value is "market research, data analytics, and system administration purposes" are used to specify the data purpose role.

In **Study SR2** we identify the variations in semantic role values, we begin with the coded roles values produced by applying the above method, and then we use open coding [Saldaña 2012] to categorize the role values for the *subject*, *condition*, *source* and *target* roles into

different categories. Bhatia and Breaux categorized the purpose role values for the same policies in a prior study [Bhatia and Breaux 2017].

We answer the research question "what are the different lexical and syntactic triggers that indicate semantic role values within and across website domains?" by extracting all lexical and syntactic patterns from the 15 annotated policies using the frame-based markup tool for **Study SR3**. Next, we analyze the results to determine how the same pattern, when used with different data actions, indicates different semantic roles and how different patterns lead to the same semantic role. Finally, we analyze the syntax of the observed patterns by identifying categories of prepositions [Aarts 2011] present in the patterns and their associated semantic roles.

5.1.2 Content Analysis Results for Identifying Semantic Roles

In this section we describe results from **Study SR1**. We manually annotated semantic roles in 15 privacy policies from three domains: health, news and shopping. We conducted a survey to identify the types of websites that users most frequently use and found We identified a total of 17 unique semantic roles across the 15 policies and across the four categories of data actions. The most frequent semantic roles are defined as follows, with the question answered in parentheses (see Appendix A for the complete list of semantic roles):

- *Subject*: The entity which acts on the information. (Who is performing the data action?)
- *Object*: The data on which the action is being performed. The values of this role were information types in our study. (What is being acted upon?)
- *Purpose*: The goal or justification for which the action is performed. (Why is the information being acted upon?)
- *Condition*: The states or events under which the data action will be performed on the information. (When will the data action be performed?)
- *Source*: The provider of the information in a collection action. (From whom is the information collected?)
- *Target*: The recipient of the information in the transfer action. (Who is the data being transferred to?)

Table 9 presents the frequency of semantic role values for each data action category, across all the policies and domains shown in Table 8 (see Appendix B for policy wise frequency). Note that some actions have multiple instances of the same semantic role attached to them.

From our analysis, we found that transfer actions had the highest number of semantic roles attached, followed by use actions. Policies across all three domains were least descriptive about retention actions. We also observed that the health policies were the most descriptive with a total of 293 actions 1,024 semantic roles across all categories, followed by shopping policies with a total of 281 actions and 878 semantic roles (see Appendix B). The news policies were the least descriptive with only 124 actions and 414 semantic roles across the five policies. This could be because the health domain is highly regulated and mostly deals with sensitive user data, as compared to shopping and news domain. The shopping policies had the highest number of collection actions, whereas the health policies had the highest number of retention, use and transfer actions.

Table 9. Frequency of Semantic Role Values Across Data Action Categories

Semantic Role	Collection	Retention	Use	Transfer
Total Actions	167	63	241	227
action location	2	3	12	8
comparison	0	0	1	4
condition	66	25	60	106
constraint	4	3	13	11
duration	0	4	0	0
exception	0	1	3	14
hypernymy	28	3	14	8
instrument	22	1	6	10
negation	13	4	17	29
object	167	63	241	226
purpose	34	16	190	49
retention location	0	13	0	1
retention property	0	2	0	0
source	50	2	7	4
subject	154	49	196	196
target	6	0	2	141
time of action	4	4	1	3
Total no. of semantic role values	550	193	763	810

In our analysis all of the collection, retention and usage actions across all the three domains have the object role attached, whereas one of the transfer actions is missing the object role in the Costco privacy policy. In our privacy surveys (see Section 6.5), we observe that the participants were the least willing to share their information for transfer actions, and not clearly specifying what information is transferred may further increase the perceived risk.

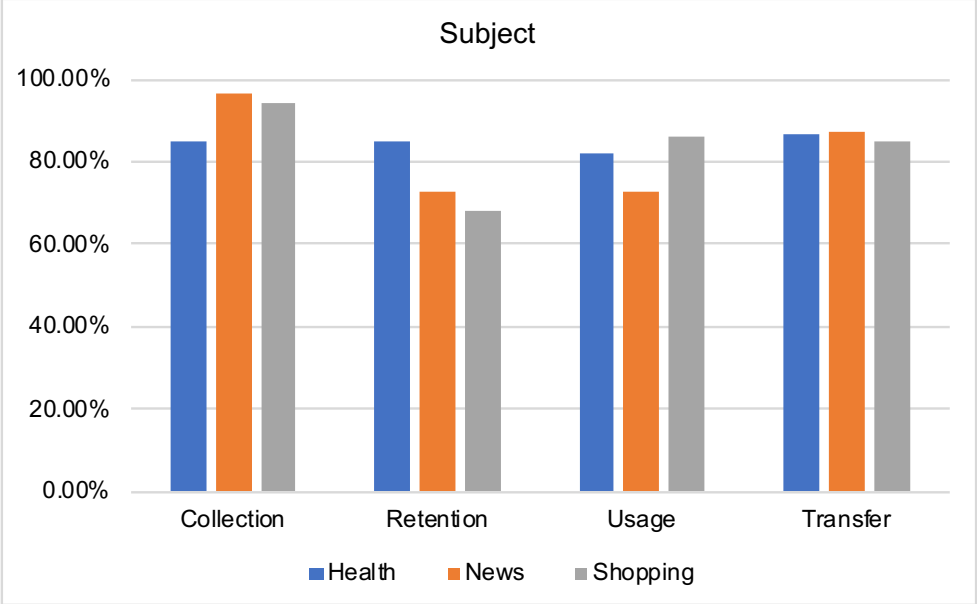
The most frequent action across all three domains was object, followed by subject. The other three most frequent semantic roles are purpose, condition and target across all three domains. The purpose role occurred most frequently with use actions and the condition role with transfer actions. The health and news policies did not contain instances of the semantic role retention property, which was present in the shopping policies. In addition, the role "comparison" was not found in the news policies.

In Figures 5, 6 and 7 we show the frequency of semantic roles subject, condition and purpose for each category of data action across the three domains. Most of the actions across the three domains had the subject role: 84.7% of the actions in health domain, 82.4% in news domain, and 83.5% in shopping domain had the subject role. The condition role was not as frequent with only 40.6% action in health, 31.4% actions in news and 36.0% actions in shopping domain had a condition role attached. Similarly, the purpose role was also not frequently found, 38.0% of health actions, 41.8% of news actions, and only 33.9% of shopping actions had a purpose role.

We observed from our analysis (see Figure 5) that most of the collection actions, specifically 92.1% on average across the three domains, had the subject role attached. In the shopping domain 94.4% and in news 96.7% of collection actions have the subject role attached, as compared to the health domain where only 85.1% of the collection actions have the subject role. This was closely followed by the transfer actions where 86.5% of all the transfer actions across

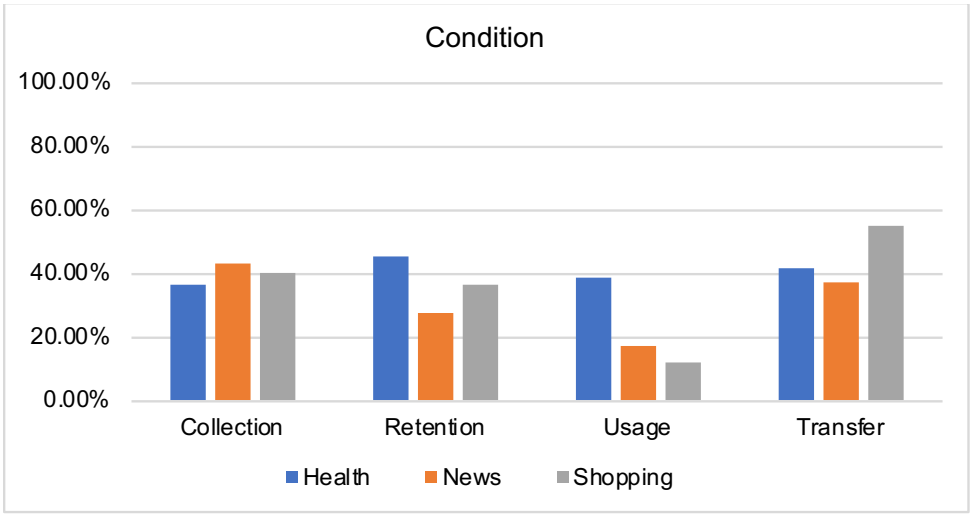
all domains had the subject role, and 87% (health), 87.5% (news) and 85.1% (shopping) transfer actions had the subject role. On the other hand, only 80.1% of usage and 75.3% of retention actions had the subject role on average across all three domains.

Figure 5. Frequency of subject role across action categories and website domains



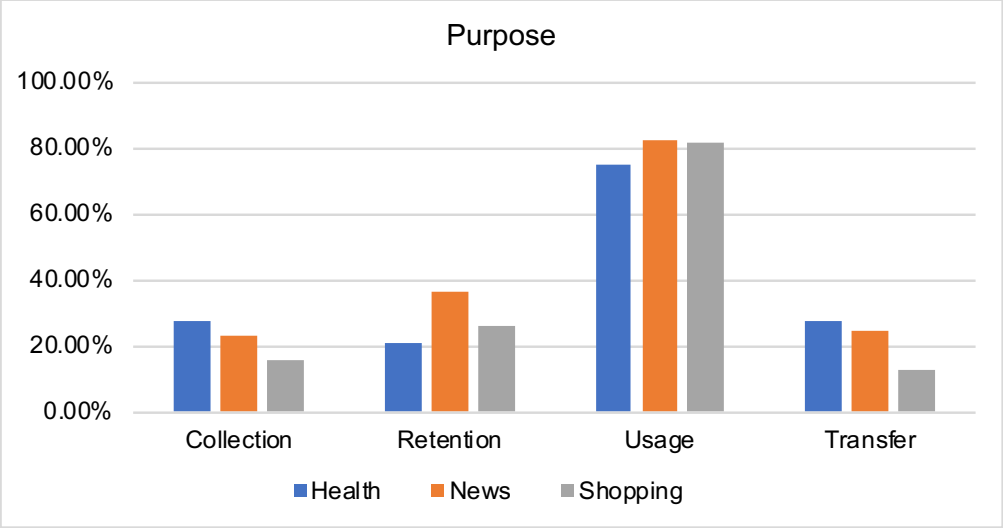
Around 55.2% of the transfer actions in the shopping domain have the condition role attached, whereas only 41.7% of health and 37.5% of news transfer actions have the condition role. On the other hand, 45.5% of the health retention actions have the condition role, as compared to 36.8% of shopping and 27.3% of news retention actions (see Figure 6).

Figure 6. Frequency of condition role across action categories and website domains



A large number of usage actions (79.6%) have the purpose role, whereas only a small number of retention (28.0%), collection (22.2%) and transfer (21.8%) actions have the purpose role attached, across all the three domains (see Figure 7).

Figure 7. Frequency of purpose role across action categories and website domains



From our analysis, we observed that on average the actions in health policies had the maximum number of subjects (84.7%) and conditions (40.6%) attached as compared to actions in news policies (subject: 82.4%, condition: 31.4%) and shopping policies (subject: 83.5%, condition: 36.0%). On average, 41.8% of the news actions had the purpose role, as compared to 38.0% health and 33.9% shopping actions.

From our risk surveys (see Section 6.5), we observed that the privacy risk perceived by the users decreases if the condition and purpose roles are specified. Incomplete description of data practices with missing role values for condition and purpose, could in turn decrease user's willingness to share their information with the website and consequently their use of the services provided by the website.

We further observe that different action words are used to describe data practices belonging to the same data action category. For example, the action words log, submit, gather, and collect are all used to describe collection practices. The action word log is often used when the data collection is implicit, or automated, and occurs when the user is browsing or using the website. For example, in the statement, "Like most web sites, our servers log your IP address, the URL from which you accessed our site, your browser type, and the date and time of your purchases and other activities." The action word submit, however, is often used when the user submits their information to the website, for example, "When you place an international order, you will submit personal information (e.g. your name, email address, billing address, and shipping address) and other order-related information to JCPenny through and to servers located in the United States." This can include the user's name, address, and payment details, in contrast to logged information that includes IP address and browser type. Thus, different action words depict subtle differences in which objects are associated and expected, despite being within the same broader category.

5.1.3 Categories of Values for Semantic Role Values

In this section we describe results for **Study SR2**, in which we categorized role values for *condition, source, target* and *subject* roles.

5.1.3.1 Categories of Values for Condition Role

We identified 280 instances of the condition role across the 15 policies. The condition categories are as follows:

- *First party action*: The data action is conditioned on an action performed by the website company itself.
- *Legal*: The data action is performed, if it is required by law.
- *Merger*: The data action is performed, if the company is part of a merger or acquisition.
- *Scope*: The data action performed is limited by practices described in the privacy policy.
- *Third party action*: The data action is performed in response to an action performed by a third party.
- *User*: The data action is conditioned on an action performed by the user, or a property that the user possesses.

Table 10 presents the condition role categories with examples and frequency across all 15 policies. The most frequent condition category across all the three domains is user, followed by first party action for news and shopping, and legal and vague for health. We also noted that health policies have a higher number of third-party actions as conditions as compared to news and shopping.

Table 10. Condition Categories

Category	Examples	% Frequency		
		Health	News	Shopping
first party action	only if we identify a biometric match to our database of known shoplifters, in the receipt of automatically collected information	7.9%	19.2%	12.9%
legal	if we believe we are required to do so by law, or legal process, as we deem appropriate in response to requests by government agencies	13.5%	7.7%	5.9%
merger	as part of any merger or sale of company assets or acquisition, if some or all of our business assets are sold or transferred	1.6%	5.8%	8.9%
scope	as permitted by this privacy policy	3.2%	1.9%	1.0%
third party	if any of these service providers need access to your personal information, when they no longer need it	9.5%	1.9%	2.0%
user	if you choose to connect your mobile device to the free in-store Wi-Fi available at Lowe's stores, if you are under 18	50.8%	55.8%	61.4%
vague	as necessary	13.5%	7.7%	7.9%
Total number of condition instances		126	52	102

5.1.3.2 Categories of Values for Source Role

The source role describes the information provider. We identified 63 source role instances across all 15 policies, which were categorized using open coding as follows:

- *Technology*: The source of collected information is a device or technology.
- *Third party*: The information about the user is collected from a third-party.
- *First party*: The information about the user is collected from a first party.
- *User*: The information is collected from the user.
- *Vague*: The source of information is present, but unclear.

Table 11 presents the source categories with examples and their frequencies across all policies and domains in our dataset. Users were the main source of information for health and news policies, whereas for shopping websites the source of information were equally likely to be third party sources or the users themselves.

Table 11. Source Categories

Category	Example Role Values	% Frequency		
		Health	News	Shopping
technology	your computer and mobile device, third party cookies	22.2%	0.0%	22.6%
third party	third party sources, public sources	16.7%	14.3%	38.7%
first party	WebMD website	5.6%	7.1%	0.0%
user	you, children under the age of 13	55.6%	78.6%	35.5%
vague	various sources	0.0%	0.0%	3.2%
Total number of source instances		18	14	31

The collection of information from technology, or from third parties is generally automated and the user may be unaware that the collection is taking place. In contrast, information collected from the user can be explicit collection, when the user provides their information to the company directly through a website.

5.1.3.3 Categories of Values for Target Role

We identified a total of 150 instances of the target role, which describes the information recipient in a transfer action, and categorized these instances as follows:

- *First party*: The information is transferred to the first party website company.
- *Third party:* The recipient of the information is a third party.
- *Location*: The target is the location where the information is being transferred.
- *Technology*: The information is being transferred to a technology.
- *User*: The recipient of the information is the user.
- *Vague*: The target of the information is present, but unclear.

Table 12 presents the target categories, examples, and frequencies across the 15 policies in our dataset (see Table 8). Most of the information was transferred to third parties for all the three domains. Health and news websites were not vague about the target of shared information, when specified.

Table 12. Target Categories

Category	Example Role Values	% Frequency		
		Health	News	Shopping
first party	JC Penny, us	4.2%	4.5%	7.0%
third party	third parties, issuer of the Mastercard	90.1%	90.9%	80.7%
location	countries, globally	1.4%	0.0%	3.5%
technology	servers, mobile devices	2.8%	0.0%	5.3%
user	you	1.4%	4.5%	0.0%
vague	others, anyone	0.0%	0.0%	3.5%
Total number of target instances		71	22	57

5.1.3.4 Categories of Values for Subject Role

We identified 595 instances of the subject role across the 15 policies. The subject categories are as follows:

- *First party*: The data action is performed by the website company itself.
- *Third party*: The data action is performed by a third party.
- *User*: The data action is performed by the user.
- *Vague*: It is not clear who performs the action.

Table 13 presents the subject role categories with examples and frequency across all 15 policies. Most of the actions across all domains are performed by the first party companies. It is interesting to note that none of the subjects in the shopping domain were vague.

Table 13. Subject Categories

Category	Examples	% Frequency		
		Health	News	Shopping
first party	we, some of our tools	79.4%	83.3%	76.3%
third party	research contractors, third parties,	11.7%	8.8%	18.8%
user	you, user	7.3%	6.9%	4.9%
vague	whoever has the access code, programs	1.6%	1.0%	0.0%
Total number of subject instances		248	102	245

5.1.4 Lexical and Syntactic Patterns

We describe results for **Study SR3** in this section. Lexical and syntactic patterns are used to coordinate role values in a role phrase or clause. Lexical and syntactic patterns describe how keywords attach to different data actions, and as part of syntactically different statements, they specify similar or different semantic role values. In Study SR3 we identified 74 patterns, with 504 instances across health policies, 235 instances across news policies, and 380 instances across

shopping policies in our dataset. Table 14 presents the five of the most frequent patterns, with example consisting of the semantic role name, followed by a colon and an example role phrase from the policy. For each pattern, we also present the pattern frequency across the 15 policies.

Table 14. Lexical and Syntactic Patterns

Pattern	Semantic Roles	%Frequency		
		Health	News	Shopping
to [value]	purpose: to provide location-based services target: to servers object: to personally identifiable information	31.5%	29.4%	28.4%
if [value]	condition: if Barnes and Noble becomes involved in a merger	6.0%	6.4%	8.2%
with [value]	condition: with your consent object: with other information target: with other companies	5.2%	8.1%	7.9%
when [value]	condition: when you interact with JC Penney	5.2%	8.5%	7.6%
from [value]	source: from you, action location: from our files	4.8%	5.5%	7.6%
Total number of instances		504	235	380

Another frequent pattern in the health domain is for [value] (8%), in the news domain is such as [value] (5.5%) and in the shopping domain is as [value] (3%). We observe that the same lexical and syntactic pattern is used to specify different semantic roles, when attached to different data actions and across different statements. The semantics conveyed by these patterns changes when attached to different data actions and in different contexts. For example, the syntactic pattern with the keyword to [value] can be used to introduce different semantic roles in the context of different data actions:

- `to [data purpose]`

"We will store and use this information to administer the programs and services in which you choose to participate, and as permitted by this Privacy Policy."

- `to [target]`

"In addition, we disclose certain personal information to the issuer of the MasterCard in connection with the administration of the Barnes and Noble MasterCard program."

In addition, different syntactic patterns can be used to introduce the same semantic role. For example, the syntactic pattern `if [value]` and `depending on [value]` can be used to specify the condition role.

- `if [condition]`

"If Barnes and Noble becomes involved in a merger, acquisition, restructuring, reorganization, or any form of … some or all of its assets personal information and your transaction history may be provided to the entities …"

- `depending on [condition]`

"Depending on how you choose to interact with the Barnes and Noble enterprise we may collect personal information …"

In our dataset, we observed that although the patterns `if [value]` and `depending on [value]` both represent the role condition, they cannot be used interchangeably. This is because in our dataset the semantic role values that occur with `if` are specific and the values occurring with `depending on` are comparatively generic set of conditions, which can take one of many possible values.

Table 15 presents the keywords for each of the most frequent roles across the 15 policies.

Table 15. Keywords Used to Specify different Semantic Role Values

Semantic role	Keywords Used
Object	along with, in conjunction with, to, with
Condition	according to, as, as part of, as long as, as well as, along with, at, based on, before, by, depending on, each time, even if, from, if, if and only if, if and when, in connection with, in the good faith belief that, in the event that, in, provided that, once, only if, when, with, without, unless, upon, until
Purpose	as, allowing, in, in an effort to, in order to, for, only as, to, that, so, so that, that, where
Target	among, between, in, only with, outside, to, with
Source	across, from, that, through

Across the three domains, we observed that similar patterns were used to specify the conditions, source and target semantic roles. The most frequent patterns used to specify the condition role for health is `if [value]`, `when [value]`, and `in [value]`; for news is `when [value]`, `if [value]`, and `unless [value]` and for shopping is `if [value]`, `when [value]`, and `in [value]`. The pattern that was used to specify most of the source roles across all the three domains is `from [value]`. The patterns used to specify most of the target roles across the three domains is `to [value]` and `with [value]`.

We noticed from our analysis that the semantic role specified by a pattern is also dependent on the action category with which it occurs. For example, in the shopping policies, the pattern `to [value]` occurs 58 times with usage actions, and in 57/58 times, this pattern coincides with the purpose role. When the pattern is attached to transfer actions, it occurs 36 times and 31/36 times it coincides with a target role. Some of the patterns such as `if [value]`, `depending on [value]`, and `when [value]` are only used to specify the condition role.

We further evaluate our 74 unique patterns under the assumption that the majority of patterns share the syntactic quality of beginning with a preposition. Leveraging this observed pattern quality, we employ preposition categorization [Aarts 2011]. Our analysis across the patterns can be characterized by the following properties of prepositions.

- *Transitive:* A single preposition that takes a noun phrase, an adjective phrase, an adverb phrase, a prepositional phrase, or a clause as a complement, e.g., `to [value]`
- *Intransitive:* A single preposition that does not require a complementing phrase or clause e.g., `when [value]`
- *Deverbal:* A preposition that takes the form of a participle e.g., `during [value]`
- *Complex:* A preposition that consists of two or more words, e.g., `as part of [value]`
- *None:* Pattern does not contain a preposition

The 74 patterns contain 28 transitive preposition patterns, four intransitive preposition patterns, five deverbal preposition patterns, 35 complex preposition patterns and 2 patterns without prepositions. We refer to the transitive, intransitive, and deverbal categories as single-preposition patterns. Our analysis shows that the majority of our patterns can be characterized by

having complex or simple syntactic structure: 47% of patterns fall in the complex category, single-preposition patterns comprise 50%, leaving 3% of patterns without a preposition. We then examined patterns across categories with shared prepositions, specifically complex preposition and single-preposition patterns that end with the same preposition. For example, `as [value]` and `except as [value]`. We found that the last preposition in a complex preposition pattern can in fact diversify the semantic role value from that of the parallel, single-preposition pattern. For example, if we consider the complex pattern `in a manner similar to [value]` and the single-preposition pattern `to [value]`, we know they will both contain prepositional phrases beginning with the preposition "to." Examining the aforementioned "to" patterns in the policies from the health domain, we find that out of the four complex patterns that end in "to," the semantic role value occurs once as a constraint using the pattern `in a manner similar to [value]` and once as a constraint using the pattern `in addition to [value]` and as twice as a purpose using the pattern `in order to [value]`. We find that out of the 159 `to [value]`patterns from the health domain, the semantic role value is 68% purpose, 30% target, and infrequently as object and source. This example suggests that the complex patterns semantic role value is dependent on a noun phrase within the complex preposition. In this example, we note that "a manner similar" and "addition," which both invoke a constraint role value, contrast with "order," which implies a purpose similar to that of the majority of the single-preposition semantic roles for "to" patterns in the health domain.

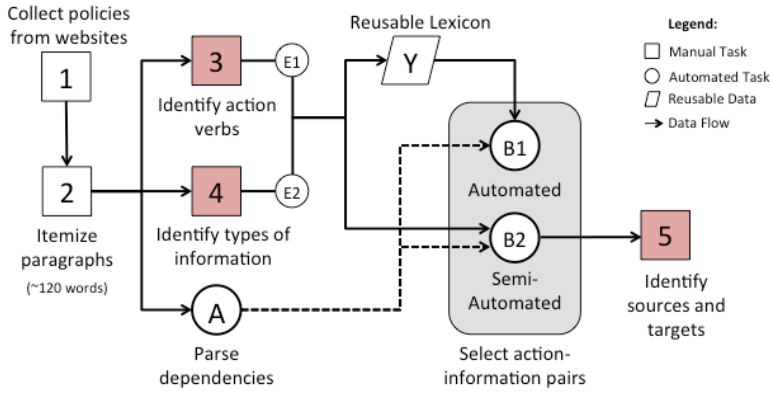
Next, I describe our hybridized framework to identify privacy policy goals.

5.2 Hybridized Framework for Identifying Privacy Policy Goals

We have developed a semi-automated framework to identify privacy goals from policy statements that combines crowd worker annotations, natural language typed dependency parses and a reusable lexicon to improve goal extraction coverage, precision and recall [Bhatia and Breaux 2015, Bhatia et al. 2016b].

Figure 8 provides an overview of our hybrid framework that consists of two kinds of manual tasks (square boxes): tasks performed by an analyst, once (white boxes), or tasks performed by the crowd workers (red boxes); automated steps performed by tools (circles) and a reusable lexicon (parallelogram). The arrows point in the direction of data flows, e.g., illustrating where crowd worker annotations are sent to automated tasks; the solid vs. dotted lines signify separate but overlapping flows.

Figure 8. Task re-composition workflow; red boxes represent crowd worker tasks.

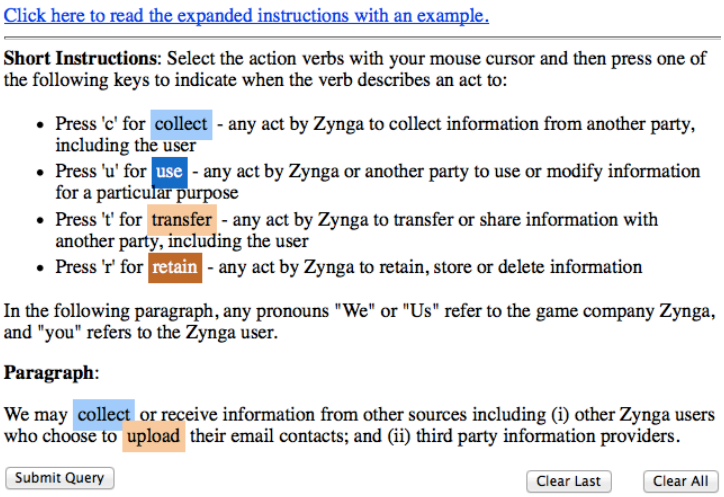


During steps 1 and 2, the analyst prepares the input text to the NLP tools used in steps Y, B1 and B2 and the crowd worker platform, in this case Amazon Mechanical Turk (AMT), which is used in steps 3 and 4. These steps are performed manually by an analyst, once for each policy at present, because they require relatively little time (a few minutes per policy). For step 1, the input text begins as a text file, which may be extracted from a HTML or PDF document. For step 2, the analyst itemizes the text into paragraphs, averaging 90-120 words, that can be annotated in less than one minute by crowd workers, while ensuring that each paragraph's context remains undivided. For example, the analyst ensures that lists are not separated across tasks, and that anaphoric references, such as "it" or "this," are contained in the same paragraph as the noun phrases to which they refer. This invariant can lead to paragraphs that exceed 120 words, which is balanced by smaller 50-60 word paragraphs. The 120-word average limit determines the average time required by one worker to annotate a paragraph, which we set to 60 seconds. This average time provides workers small, but frequent micro breaks between tasks and it allows workers frequent opportunities to stop annotating text whenever they feel fatigue or boredom. Because the tasks are small and independent, workers can stop at any time and workers need not complete all of the tasks for a single policy: subsequent workers can be given tasks that continue where previous workers stopped working. The small tasks also allow us to better distribute the risk of low-performing crowd workers and the associated costs.

5.2.1 Crowd worker Micro Annotations

Steps 3 and 4 are crowd worker micro tasks that ask workers to annotate phrases in one of two ways: for step 3, workers are asked to label action verbs that describe information collection, use, transfer or retention, as shown in Figure 9. We call this **Study SR4**. Following simple instructions, workers see the ~120-word paragraph and are tasked to select and annotate relevant phrases using their mouse and keyboard. The annotated phrases are color coded to correspond to the label selected by the worker. The micro task for step 4 is similar, except that instead of distinguishing among four kinds of actions, workers are asked to identify noun phrases that correspond to any kind of information. In both steps 3 and 4, the results are captured and recorded as part of an AMT batch result, wherein we asked five workers to annotate each paragraph. This number of workers was determined in prior work, which showed worker agreement for 2/5 workers correlates with high precision and recall for these tasks [Breaux and Schaub 2014].

Figure 9. Crowd worker annotations to annotate information types

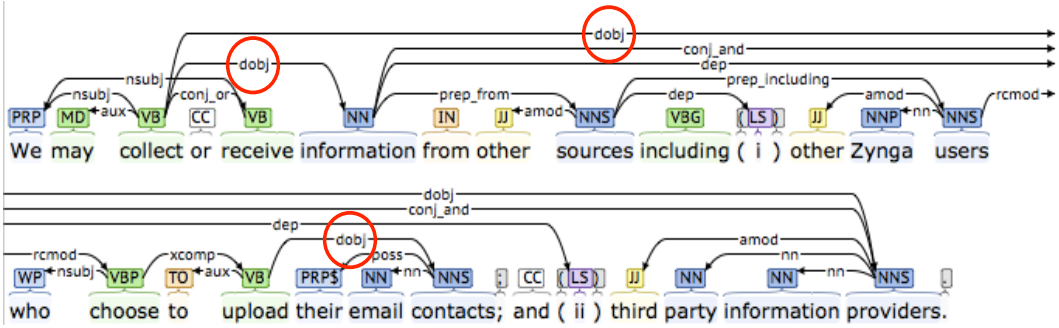


As shown in Figure 8, the results from steps 3 and 4 are combined with a dependency parse of the paragraphs to select action-information pairs in steps A, B1, B2, which we now discuss.

5.2.2 Dependency Parsing and Pair Selection

In step A, we apply typed dependency parsing to the individual sentences from the micro task text input using the Stanford dependency parser [Marneffe et al. 2006], which we call **Study SR5**. Typed dependencies are binary relations between a first term, called the governor, and a second term, called the dependent. We present an example sentence with the corresponding collapsed, CC-processed dependencies (collapsed dependencies with propagation of conjunct dependencies) for each word in the sentence in Figure 10. Commonly found dependency types include *nsubj*, which is the nominal subject of the sentence, and *dobj* or direct object of a verb phrase. One advantage of dependency parsing is that the parser splits phrases along conjunctions and it links modifiers to nouns. However, natural language ambiguity can lead to errors in parsing. For example, Figure 10 presents three dependences *dobj (collect, providers), dobj (collect, information),* and *dobj (upload, contacts).*

Figure 10. Stanford dependency parse of micro task input text



The first dependency *dobj (collect, providers)* is incorrect: the sentence author likely did not mean that the website "collects third party information providers;" rather, the providers are a second example of "from whom" that information is collected. Thus, we assume some degree of inaccuracy produced by the typed dependency parser. However, the second two dependencies are

correct and they indicate prospective goals about which we can ask additional questions: "from whom is information collected" (a collection goal), and "by whom are contacts uploaded" (a transfer goal). Our approach to select action-information type pairs is limited by the accuracy of the Stanford Parser.

We propose two different approaches, denoted by steps B1 and B2, to select the action-information type pairs using the typed dependencies. The typed dependencies are combined with crowd worker annotations in step B2, wherein we perform action–information pair selection to identify actions (typically verbs) that should be paired with information types (typically noun phrases). Step B2, is a semi-automated approach that requires manual annotations for the actions and information types, which are obtained from the crowd workers.

In order to automate this process of obtaining the action and information types, we propose an alternate approach in step B1, which is a fully automated approach. In this step, we use the action and information type lexicon, to identify actions and information types in the policy statements using a simple keyword match between the lexicon and policy terms. We combine these identified actions and information types in each statement with the typed dependencies of the statement to determine if the action and information type are linked by a typed dependency or not. If linked, the corresponding action-information type pair is selected as a candidate partial goal specification. We use two general strategies for both steps B1 and B2 for linking action-information pairs: (1) we first identify direct dependencies, in which both the governor and dependent were separately annotated by either the lexicon for B1 or by crowd workers for B2 in the action and information type tasks; and (2) we identify indirect dependencies that consist of two typed dependencies, each one containing one lexicon- or worker- annotated term and sharing a third term, which may not have been annotated. We only consider terms that have been annotated by the lexicon in step B1 or by two or more crowd workers in step B2 based on prior work that shows 2/5 workers produce high precision and recall for these tasks [Breaux and Schaub 2014]. In Figure 10, for example, dobj (upload, contacts) is a direct dependency, if "upload" was annotated by two or more workers in the action task, and "contacts" was annotated by two or more workers in the information type task. In addition, in Figure 10, dobj (collect, information) and cc_or (collect, receive) comprise an indirect dependency that links receive to information via the cc_or typed dependency for the English conjunction "or". In our evaluation, we are interested in identifying which dependency types are high confidence, meaning, they maximize true positives and minimize false positives.

Next, we introduce the lexicon as a means to collect and reuse knowledge about annotated actions and information types to improve recall (missing true positives in step B2) and to develop the fully automated approach for step B1.

5.2.3 Re-usable Lexicon and Entity Extraction

Lexicons are used to bootstrap requirements analysis by re-using terms frequently seen in particular domains. In our work, we build the lexicon using crowd worker annotations from steps 3 and 4 in Figure 8 for 30 privacy policies to attempt fully automated goal finding, which we call **Study SR6**. The lexicon is constructed from action and information type entities, which are unique textual descriptions needed to identify recurring instances of the same concept. For instance, the entities in the lexicon should enable us to resolve synonyms, plurals and singular forms of information types (e.g., "email address" is basically the same concept as "email addresses"). In steps E1 and E2, we apply an entity extraction technique on the annotated verb and noun phrases provided by the crowd workers to extract the individual entities (information

types) from the annotated phrases. These phrases may consist of ambiguous lists and clauses that obfuscate the unique entities. The entity extractor was first evaluated on 3,850 crowd worker information type annotations [Bhatia and Breaux 2015]. In Section 5.3, we present an extended evaluation on 7,682 annotations from 30 policies and results of applying the acquired lexicon to the re-composition framework.

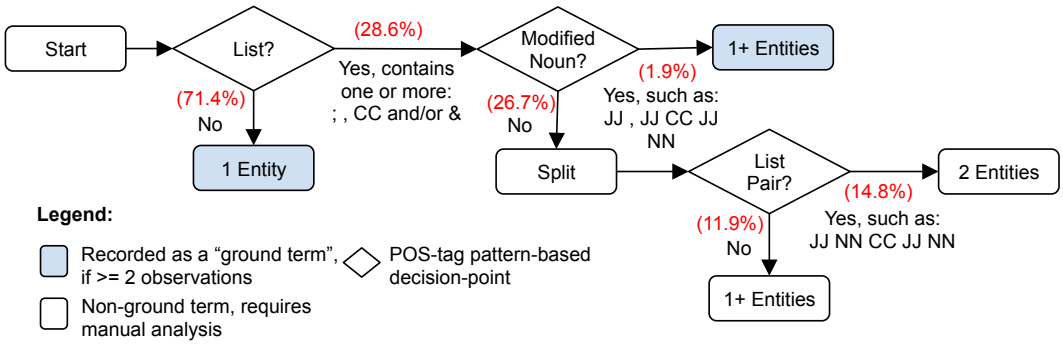
The entity extractor workflow is presented in Figure 11. The extractor first tests whether a worker annotation is a list (i.e., it contains a common list delimiter, such as a comma, semi-colon or POS-tagged English conjunction CC). If an annotation does not contain a list delimiter, then we test whether the annotation describes a single entity by checking the annotation's POS tag sequence against a well-known regular expression NP + CL that matches a noun phrase (NP) followed by a clause (CL) expressed as standard POS tags as follows [Justeson and Katz 1995]:

```
NP=((JJ|RB|VBG|VBD|NN\S?|NN\S?\sPOS)\s)*(NN\S?)

CL= (\s(IN|PRP|TO|VBG|VBN|WDT|WP)\s.*)?
```

Based on our analysis of 30 policies, 71.4% of the worker supplied information type annotations describe single entities, and the remaining 28.6% describe lists. For lists, the extractor checks whether the annotation describes a modified noun, which comprises 1.9% of annotations. This case includes lists of conjoint adjectives followed by a noun (e.g., "aggregate, statistical information"), as well as disjoint lists (e.g., "geographic and demographic information"). Disjoint lists are split to distribute the modifiers separately across the nouns (e.g., to yield "geographic information" and "demographic information"). The remaining 26.7% of annotations are lists of noun phrases, which are split by delimiter. Each delimiter-separated noun phrase is checked against previously seen simple, non-obfuscated entities, called ground terms. In Figure 11, ground terms are automatically identified where the output boxes are colored blue. While the workflow is seemingly complex, it has been shown to be highly effective at extracting entities.

Figure 11. Stanford dependency parse of micro task input text

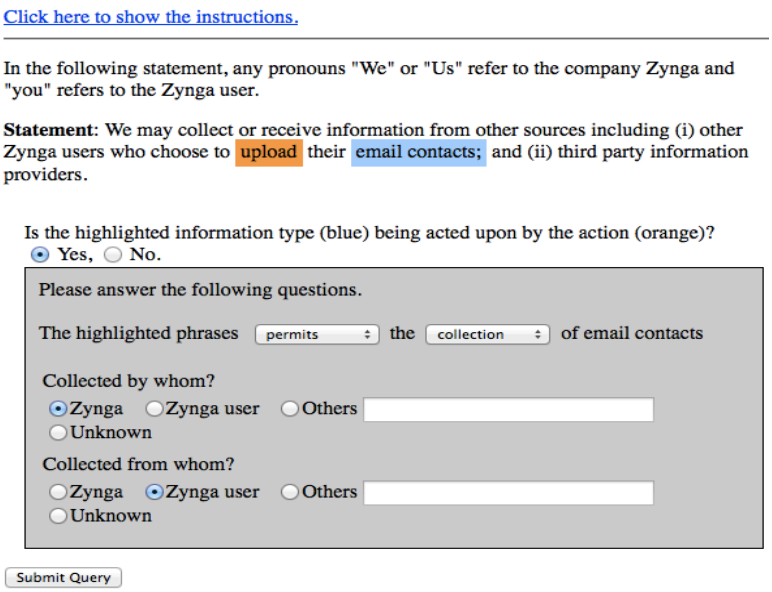


We next discuss the crowd worker pair validation task in step 5.

5.2.4 Validate Pairs and Identify Source and Target

In step 5 (Figure 8), we take the selected action and information type pairs from step B2 and send these pairs to the crowd workers to ask whether the information action and information type are valid pairs, which we call **Study SR7**. If true, we also ask crowd workers to identify the actors who send, receive and use the information based on the coded action type. This validation task helps us to remove the false positive action-information type pairs produced by step B2, because unlike the crowd workers who understand context, the lexicon indiscriminately identifies all candidate pairs based on keyword matches. Figure 12 presents the task interface for step 5, in which workers select the action modality ("permits" or "prohibits"), the action category, and then they complete the source and target questions using radio buttons or free-response text boxes. If the worker selects collection from the drop-down list, the questions ask, "from whom," whereas selecting transfer from the drop-down list asks, "to whom." For use and retention, we ask only "by whom" is the information used or retained.

Figure 12. Crowd worker micro task to validate action-information type pairs



Our framework makes use of crowd worker annotations, to identify the actions and information types in the privacy policy statements, which are linked using the typed dependencies to select the action-information type pairs for each statement. The action and information type annotations are used to build the action and information type lexicon respectively, which are re-used to identify missing crowd worker annotations, and to attempt to fully automate the crowd worker annotation process. We next present the results of our framework evaluation.

5.2.5 Evaluation and Results

We evaluated the framework by answering the following research questions:

RQ1. How do crowd workers compare with expert annotators in performing micro tasks? (Study SR4)

RQ2. How well do typed dependencies combined with crowd worker annotations predict the pairs needed to express partial goal specifications? (Study SR5)

RQ3. How well does the lexicon improve identification of missing annotations or pairs? (Study SR6)

RQ4. How well does lexical reuse increase with each new policy analyzed? (Study SR6)

RQ5. How well do crowd workers identify false positives in a validation task? (Study SR7)

Research question RQ1 evaluates steps 3 and 4 in the framework (see Figure 8) with respect to precision and recall using the expert annotations. This evaluation extends a prior evaluation of these two tasks that examined only a single policy [Breaux and Schaub 2014]. Research question RQ2 evaluates the typed dependency step A and pair selection step B2 against the expert pairs, while RQ3 separately evaluates pair selection steps B1 using the crowd worker annotations and lexicon against the expert pairs. Research question RQ4 evaluates the lexicon independently to assess how it scales over time. Finally, research question RQ5 evaluates step 5 and the ability of the crowd workers to identify false pairs against the expert pairs.

To evaluate our hybrid framework and to answer the research questions, we selected five privacy policies that the first two authors (the experts) analyzed as part of this case study, which we refer to as expert annotations and expert pairs when combined into a partial goal specification:

- AOL Advertising, last updated 4 May 2011
- Facebook API Developer Guidelines, revised 28 June 2013
- Flurry Privacy Policy, updated 9 July 2013
- Waze Privacy Policy, modified 30 May 2013
- Zynga Privacy Policy, last updated 30 Sep 2011

These policies were selected because they were used in two prior case studies to express privacy goals formally in the Eddy language based on Description Logic [Breaux et al. 2014; Breaux et al. 2015]. The policies correspond to different stakeholders in a software composition: the AOL and Flurry policies govern advertising services used by a game provider (Zygna) and a navigation application (Waze). The Facebook policy governs a platform that both Zynga and Waze use for user identification services, when users log in to their applications using their Facebook accounts. Thus, these policies cover two popular data flows and the policy language in each of these policies varies by the role of the covered services (ad services, identity provider, and first-party app developers).

The expert data set was created by two analysts (the first and second authors) by extending the annotations from a prior case study [Breaux et al. 2014]. In this prior study, on average, the first analyst expended 1.09 minutes per statement extracting requirements, whereas the second analyst expended 2.21 minutes per statement [Breaux et al. 2014]. For the new data set, the analysts spent on average 1.9 minutes each per statement to review the previous annotations and extend the dataset. The expert data set serves as the "ground truth" by which we compute precision and recall as measures of the automated steps B1 and B2 shown in Figures 8, above. For all precision and recall calculations, the expert data set contains the sum of true positives and false negatives.

5.2.5.1 Crowd Worker Micro Task Results

We solicited five workers per micro task to identify the actions and information types for **Study SR4**. We recruited US residents as workers on AMT, who had at least a 95% approval rating for over 5,000 tasks. We paid workers $0.15 per task for actions and $0.12 per task for information types to keep the hourly wage close to $8-10 per hour. We allowed up to five minutes to complete each task. Results were accepted or rejected within 24 hours. For the action identification task, the workers required 72 seconds on average to complete a single task, which resulted in an average hourly rate of $8.40. On average, workers required 61 seconds per information type task, with an average hourly rate of $6.30.

Table 16 presents the total cost incurred for the information action and information type identification micro tasks for all policies, including: the total number of tasks (Tasks) in each policy; Amazon charges of 10% (AMT fees), and Total Cost, consisting of worker payments and AMT fees.

Table 16. Cost to Crowdsource Micro Tasks

Policy	Actions Micro Task			Info. Types Micro Task		
	Tasks	AMT fees	Total Cost	Tasks	AMT fees	Total Cost
Waze	34	$2.55	$28.05	34	$2.04	$22.44
Zynga	32	$2.40	$26.40	32	$1.92	$21.12
Flurry	33	$2.48	$27.23	33	$1.98	$21.78
FB	32	$2.40	$26.40	32	$1.92	$21.12
AOL	18	$1.35	$14.85	18	$1.08	$11.88

Table 17 presents the number of annotations acquired from steps 3 and 4: for each policy, we present the total number of sentences in the policy, the total number of sentences with annotated actions only, with annotated information types only, with both an annotated action and information type, and finally the overall total number of annotated actions and information types. For sentences with only the annotated actions or information types and not both, these sentences would not yield an action-information type pair based on an expert analysis of the text.

Table 17. Summary of Micro Task Annotations

Policy	Total Sentences	Sentences with:			Annotations	
		Only Actions	Only Info Types	Both	Actions	Info Types
Waze	117	5	36	56	117	146
Zynga	97	4	28	52	103	125
Flurry	135	22	32	49	106	111
FB	136	15	25	57	129	166
AOL	76	6	6	50	96	87

Table 18 presents the precision and recall for both actions and information types as compared to the expert annotations. On average, workers were able to identify the actions and information types with high recall of 0.84 and 0.92, respectively and average precision of 0.87 and 0.83, respectively. Notable in Table 18, the Flurry policy includes nomenclature specific to the advertising industry that crowd workers are likely unfamiliar with, which may explain the lower precision and recall for that policy as compared to the other policies.

Table 18. Crowd-Sourced Annotations Compared to Expert

Policy	Actions		Information Types	
	Precision	Recall	Precision	Recall
Waze	0.88	0.83	0.62	0.91
Zynga	0.91	0.79	0.95	0.98
Flurry	0.73	0.64	0.97	0.84
FB	0.98	0.96	0.88	0.90
AOL	0.86	0.98	0.71	0.95
Average	0.87	0.84	0.83	0.92

5.2.5.2 Dependency Parse and Pair Selection Results

We now present the dependencies parser results and results of our techniques for selecting action-information type pairs from **Study SR5**.

Table 19 presents results from a naïve approach to produce typed dependencies from the five policies to illustrate the scope of the pair selection challenge. This includes the number of unfiltered dependencies per policy (Total Dependencies); the subset of the total in which the governor or dependent are a verb and noun pair (Dependencies w/ Verbs & Nouns); the three most common dependency types found in the direct selection method described in 5.2.2, which are dobj (direct object of a verb phrase), nsubjpass (syntactic subject of a passive clause) and vmod (verb heading a phrase); and the number of pairs identified in the expert analysis, which represents our evaluation target. As can be seen from Table 19, the space of dependencies is quite large and, assuming perfect recall, the precision of a naïve approach to pair selection would be very low.

Table 19. Naïve Approach to Identify Relevant Pairs – Parser

Policy	Total Dependencies	Dependencies w/ Verbs & Nouns	dobj, nsubjpass, vmod	Expert Pairs
Waze	3286	794	365	101
Zynga	2758	655	352	93
Flurry	3268	845	398	81
FB	3389	765	339	91
AOL	1720	452	216	81

In Table 20, we present a slightly more informed approach to identify action and information type pairs using typed dependencies and lexicon (B1 in Figure 8). The column Expert Pairs lists the total number of action and information type pairs identified by the experts, manually. The column Lexicon and Parser Pairs lists the total number of pairs automatically obtained by pairing actions and information types from the lexicons that share a direct or indirect dependency based on the parser output. The columns Precision and Recall are computed by comparing the Lexical and Parser Pairs to the Expert Pairs, which serves as the ground truth. The lexicon-based approach was able to identify the action-information type pairs with an average recall of 0.80, however, the average precision was very low at 0.20. The large number of false positives obtained using the lexicon can be attributed to the fact that at present the lexicon does not have the ability to disambiguate the meaning of a term in the given context, and thus identifies all instances of a term in a statement.

Table 20. Naïve Approach to Identify Relevant Pairs - Parser and Lexicon

Policy	Expert Pairs	Lexicon and Parser Pairs	Precision	Recall
Waze	101	360	0.22	0.77
Zynga	93	424	0.19	0.86
Flurry	81	432	0.15	0.79
FB	91	306	0.22	0.75
AOL	81	229	0.22	0.79

From Table 19 and 20, we see that semantic dependencies alone, even direct dependencies without human guidance, produce a large number of false positives compared to the evaluation target. In addition, while the lexicon contains terminology from prior worker annotations, it lacks the workers' direction in reducing the dependencies to within a reasonable reach of the evaluation target. To inform our approach, we analyzed the direct and indirect dependencies to determine the most frequent dependency patterns found in the re-combinations and how often they lead to true or false positives. We found three direct dependency patterns and five indirect dependency patterns that constitute 71.81% of the total true positive re-combinations. We describe these patterns in Table 21 as follows: the pattern name, the typed dependency sequence, the frequency of the pattern across all five policies, and the number of true and false positive action-information type pairs for each pattern measured against the expert pairs.

Table 21. Typed Dependency Patterns

	Pattern Name	Typed Dependency Sequence	Frequency	True Positive	False Positive
Direct	Direct Object	dobj	195	188	7
	Passive nominal subject	nsubjpass	34	32	2
	Verbal modifier	vmod	24	22	2
Indirect	Conjunction and with direct object	conj_and , dobj	25	15	10
	Conjunction or with direct object	conj_or, dobj	17	12	5
	Passive nominal subject with list.	nsubjpass, prep_such as	1	1	0
	Direct object with verbal modifier	dobj, vmod	13	2	11
	Direct object with preposition	dobj, prep_*	20	10	10

The three direct dependency patterns (direct object, passive nominal subject and verbal modifier) in Table 21 on average constitute 59.1%, 10.3% and 7.3%, respectively, of the direct dependency re-compositions across all five policies for the hybrid approach. These three patterns led to true positives in 99.6% of the instances studied. The only instance where the direct object pattern yields an incorrect result was "You must immediately revoke an end-advertiser's access to your app upon our request." (from Facebook privacy policy) In this sentence, revoke is annotated as an information action and access is annotated as an information type by the

workers. The pair (revoke-access) is linked by a direct object dependency, which is a true dependency yet a false positive because "access" is not an information type.

The five indirect dependency patterns describe 41.1% of the total indirect dependency re-compositions in the hybrid approach action and information type pairs. On average, the direct dependency patterns led to true positives in 87.9% and indirect dependency patterns led to true positives in only 44.3% of the instances.

As observed from Table 19 and 20, there is no simple approach to using the parser to identify the action-information type pairs. By adding our crowd worker annotations for both the actions and information types, however, we identified a set of high confidence pairs that consist of the direct and indirect pairs defined in Section 5.2.2. Ideally, these pairs will contain all true positives and minimal false positives and omit minimal false negatives. In Table 22, we present our baseline measure (Total Annotated Pairs), which is the number of all possible pairs, which assumes naively that every annotated information action is crossed with every information type that occurs in the same sentence, followed by the total number of high confidence pairs based on dependency parsing and worker annotations, and the total number of expert pairs. The hybrid approach greatly reduces the number of pairs as compared to the naïve approaches presented in Tables 19 and 20.

Table 22. Action-Information Type Pairs from Hybrid Approach

Policy	Total Annotated Pairs, Possible	High Confidence Pairs	Expert Pairs	Precision	Recall
Waze	379	107	101	0.73	0.77
Zynga	467	120	93	0.64	0.83
Flurry	237	71	81	0.79	0.69
FB	301	111	91	0.75	0.91
AOL	239	106	81	0.74	0.96

In Table 22, the Flurry policy has the lowest precision and recall among all analyzed policies. This is because workers annotated both the information action and information type in only 36.3% of the sentences in the Flurry policy (see Table 17), whereas in other policies workers annotated 52.3% on average. The actions in the Flurry policy that were not identified by the workers were context-sensitive – e.g., "get back" (a colloquialism), and "export" and "request" (both software functions), to name a few – which were also different from the action words frequently found in other policies. Thus, the workers biased by terminology commonly found in other policies may have not expected and thus missed these terms.

Our analysis of the remaining 143 false positive pairs after the expert analysis shows that 14/143 pairs contain an action that was part of a data purpose. For example, in the sentence "We use personal information to create new services", the action "create" marks the beginning of the purpose for which the information type personal information is being used. We observed that 12/143 pairs were pairs where a technology was being used to perform an information action. For example "This information is collected by the use of log-files." In this case, the log-files are a container for information and a technology. Manually excluding such pairs from our analysis would improve average precision from 0.73 to 0.78, which offers promise for future work.

In addition, we manually analyzed the 75 false negative sentences from all five policies in which the action and information type pairs were identified by the experts but were missed by our crowd workers. Our analysis shows that out of the 75 sentences, the workers did not annotate

an information action in 37/75 of these sentences; in 20/75 of these sentences the workers did not annotate an information type; and in 5/75, the workers did not annotate both the action and the information type. In Section 4.3, we discuss how we make use of the reusable lexicon to identify these missing annotations and reduce the number of false negatives.

In the remaining 13/75 sentences, the workers had identified the information action and the information type, but the parser could not determine a direct or indirect dependency between the information action and information type in the pairs, thus, they were not included in our high confidence pairs. On further inspection, we found that this was because of incomplete worker annotations. For example, in the sentence, "You can retrieve recommendations created for a particular End User by passing the device identifier of the End User" the workers annotated the information action retrieved but missed its corresponding information type (recommendations). Instead, they annotated the information type device identifier, but missed its corresponding information action (passing). The annotated information action and type pair (retrieved-device identifier) is not linked by a direct or indirect dependency relationship and was therefore excluded from the high confidence pairs.

In the next section, we discuss the reusable lexicon's impact on identifying missing actions and information types.

5.2.5.3 Impact of Reusable Lexicon on Pair Selection

We now present the results of the reusable lexicon in **Study SR6**. We built the lexicon from 30 policies spanning five domains: employment, news, social networking, shopping, and telecommunications. The five policies listed above were not part of the selected policies. The entity extractor successfully extracted entities from 97.8% of crowd worker annotations. In Table 23, we present the number of actions and information types that were missing from the crowd worker annotations and identified using the lexicon, and the corresponding number of missing high confidence direct and indirect pairs that result from applying the lexicon to each of the five policies that are used for evaluation.

Table 23. Results for Reusable Lexicon

Policy	New Actions, Missed	New Information Types, Missed	New High Confidence Pairs, Missed
Waze	116	17	58
Zynga	165	19	74
Flurry	88	99	78
FB	81	79	51
AOL	20	36	26

Table 24 presents the number of false negative pairs produced from worker annotations reported in Section 5.2.5.1, the number of true positive pairs identified using the high confidence pairs from the lexicon reuse reported in Table 23, and the precision and recall without the lexicon reported in Table 22, and the precision and recall with the lexicon. The results in Tables 23 and 24 were computed using all the terms in the action lexicon and information type lexicon.

The lexicon-produced high confidence pairs identified 37.34% of the pairs that were FNs from the worker annotations and improves the average recall by 8.8% to 0.90. However, the lexicon also significantly reduces the average precision by 31% to 0.50 based on the expert pairs.

But as it has been previously noted by Berry et al., it is difficult to achieve both high precision and high recall with NLP techniques for requirements engineering, and a NLP tool for requirements engineering should be tuned to favor recall over precision because errors of commission are generally easier to correct than errors of omission [Berry et al. 2012]. We therefore aim at minimizing the number of false negatives, even if that means accepting a few false positives.

Table 24. Impact of Lexical Reuse on Precision and Recall

Policy	False Negative Pairs	True Positive Pairs	Precision w/o Lexicon	Precision w/ Lexicon	Recall w/o Lexicon	Recall w/ Lexicon
Waze	23	6	0.73	0.51	0.77	0.83
Zynga	16	6	0.64	0.43	0.83	0.92
Flurry	25	12	0.79	0.46	0.69	0.84
FB	8	3	0.75	0.52	0.91	0.95
AOL	3	1	0.74	0.60	0.96	0.98

When using all the terms from the action and information type lexicons, precision drops by 31% for an 8.8% increase in the recall over the hybrid pair results in Table 22. This decrease in precision is due to the effect of context in terminological reuse. Phrases, such as "send" or "receive" may indicate information collection and transfer in one context but be used to describe non-information transactions in another context. To find the optimal subset of the lexicons that leads to an increase in recall without a steep decrease in the precision, we conducted an experiment based on lexicon partitions portioned from increasing increments of 10%. In Table 25, we present the Precision and Recall for different lexicon partitions. The column, Action Lexicon shows the partition of the action lexicon that was used for the respective experiment. In Table 25, x% Action Lexicon means that, top x% of the terms in the action lexicon were used for the analysis. Similarly, x% Info. Type Lexicon means that the top x percent of the terms in the information type lexicon were used for the analysis.

Table 25. Impact of Lexical Reuse on Precision and Recall

Action Lexicon	10% Info. Type Lexicon		50% Info. Type Lexicon		100% Info. Type Lexicon	
	Precision	Recall	Precision	Recall	Precision	Recall
10% Action Lexicon	0.63	0.88	0.61	0.89	0.60	0.89
50% Action Lexicon	0.56	0.89	0.54	0.90	0.53	0.90
100% Action Lexicon	0.53	0.90	0.52	0.90	0.51	0.91

From our experiments with the lexicon partitions, we conclude that the precision decreases and recall increases as the partition size of the lexicon for the experiment increases. Further, the decrease in precision is greater than the increase in recall. The precision drops by 14.3% and the recall increases by 5.4% over the hybrid pair results in Table 22 when we use the top 10% terms in the action and information type lexicons. The precision further decreases as we increase the partition sizes and the precision drops by 31% when the entire action and information type lexicons are used, for an increase of 8.8% increase in recall.

Even after lexical reuse, some information actions from the expert annotations could not be identified using the hybrid approach and lexical reuse. These actions include "based on," "get back," "complete" (a user profile), and "be visible," which are context-sensitive or require rich interpretations, such as multiple inferences or tacit knowledge (e.g., "be visible" suggests that others can see the information that has been visible, and this inference constitutes a form of information transfer). The lexicon is missing some information types, which include domain-specific information not present in the lexicon, for example, "customized audience" and "identifiable-route." Missed information types also include anaphora, such as "it," that refer to an information type in the prior sentence, which was identified by the experts, but not by the workers.

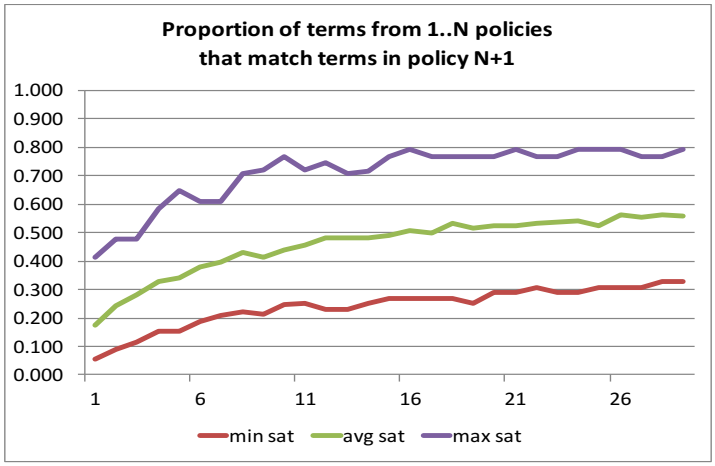
The first and third authors evaluated the lexicon to determine the scale of false positives that the lexicon can introduce when used without worker annotations. These two authors analyzed the Waze and Zynga policies to identify those instances of actions and information types that appear in the lexicon, but were not annotated by the workers (i.e., to find possible false negatives). They identified 909 actions and 450 information types in the two policies, among which only 15% of the actions and 12.2% of the information types were false negatives. From this analysis, we conclude that worker ability to distinguish between true positives and false positives is an improvement over the lexicon alone, and the lexicon alone could greatly inflate the number of false positives, if used without worker annotations.

In summary, the low precision due to the lexicon can be attributed, in part, to the ambiguity of terms and the role of context in determining when data processing events take place, and to the noise in worker responses. Information actions that are ambiguous include "assist," "solicit," "permit," and "allow," among others. Terms that workers annotated as information types that should be excluded include "third parties," "campaign" and "network." In the case of "campaign" and "network," these are activities and technologies that imply some type of information but are not themselves the implied information type. We also observed that false negatives in the worker data include words that occur less frequently across policies, including actions, such as "permit" and "export," and information types, such as "payment," "ads." Thus, limiting the lexicon to the most frequent words and phrases, will in turn hinder the ability of the lexicon to identify false negatives.

5.2.5.4 Results of Scaling Reusable Lexicon

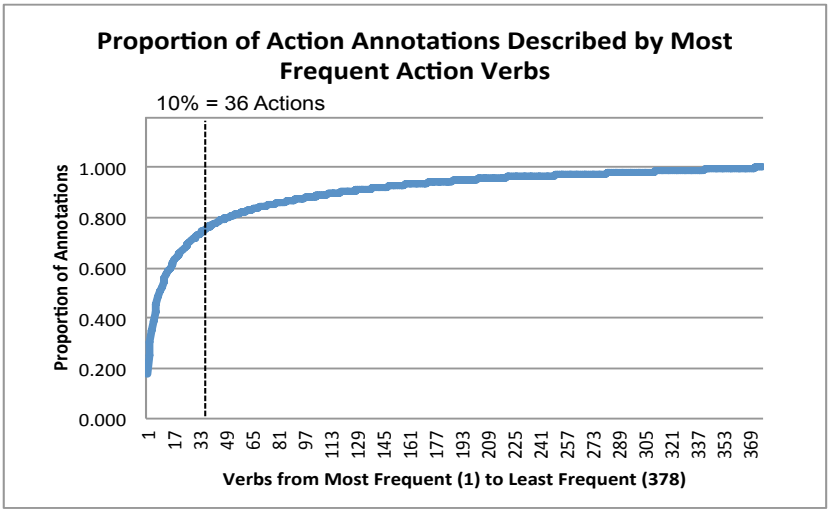
We also examined the extent to which the lexicon can predict actions and information types in additional privacy policies in **Study SR6**. This analysis shows that privacy policies have unique entities that are not shared across policies. Figure 13 presents the saturation (sat.) of information type entities for the same 30 policies: at any point along the x-axis, we observe the percent reuse of information types in a policy *N* based on the last *N-1* policies previously seen. This result is based on 100 pseudo-random permutations of the orders of the 30 annotated policies. We observe that near 14-15 policies, the average maximum threshold for saturation of 77% is achieved, meaning, every new policy contributes a sufficient number of unique terms to the lexicon that 23% of the new policy would not appear in any previously seen policy in the best case, and 71% of the policy terms would be new in the worst case. At present, this observation suggests that the lexicon cannot entirely replace crowd workers, because there will always be new terms that the lexicon has never encountered. In figure 13, we show the % reuse of information types described in N-1 policies for each N'th policy along x-axis.

Figure 13. Saturation of information types in lexicon



We further analyzed the action verbs from the same 30 policies and found 377 unique verbs identified by crowd workers. Only a small subset of these verbs dominate the results, with 10% of action verbs describing 75% of the annotations (see Fig. 14, which shows the number of annotations per verb on the y-axis in logarithmic scale, and each indexed verb along the x-axis). There is ambiguity, however, with 28% of verbs coded by two or more actions (collect, use, transfer and retain) and 5% of verbs coded as sharing-ambiguous, meaning they were coded as both collect and transfer by two or more workers. For these verbs, it may be difficult for crowd workers to determine from the text who is providing and who is receiving the relevant information. Finally, some of the verbs were also used to describe use-related purposes, which is one source of reduced precision.

Figure 14. Number of annotations per verb along the y-axis (log scale), and each unique verb of 380 verbs along x-axis



5.2.5.5 Validation Task Results

The results from Section 5.2.5.2 show that the high confidence pairs from our hybrid approach contain some number of false positives (see step A and B2 in Figure 8, and Tables 22 and 23). One objective of step 5 in our framework (see Figure 8) is to identify these false positives and remove them from the results to achieve higher average precision. In **Study SR7** which is the validation task, we ask workers whether the action and information type pair from the high confidence pairs is a valid pair (true positive), or whether it is an invalid pair (false positive). If a valid pair, then we ask crowd workers to identify the modality and the actors who send, receive, and use the information based on action labels provided by the previous crowd workers in step 4.

We solicited five workers per task to identify the valid information actions and information type pairs. We recruited US residents as workers on AMT, who had at least a 95% approval rating for over 5,000 tasks. We paid workers $0.12 per classification task. We allowed up to five minutes to complete the task. Results were accepted or rejected within 24 hours. The workers completed each task in 39.6 seconds on average, resulting in an average hourly rate of $10.95.

In the analysis of the validation task, we mark a pair as false positive, if more workers annotated it as false positive than the number of workers who annotated it as true positive. Table 26 presents the validation task results as follows: the number of high confidence pairs obtained using our hybrid approach (see results in Section 5.2.5.2 from step B2 in Figure 8); the number of false positive pairs identified by three or more workers, the number of false positive pairs identified by the experts; the number of ambiguous pairs, in which 2/5 and 3/5 workers yielded conflicting annotations; the precision without validation reported in Table 22; and the precision with validation from crowd workers. As shown in Table 26, the crowd workers greatly reduced the number of false positives produced by the direct and indirect dependency patterns.

Table 26. Pairs Validation Result

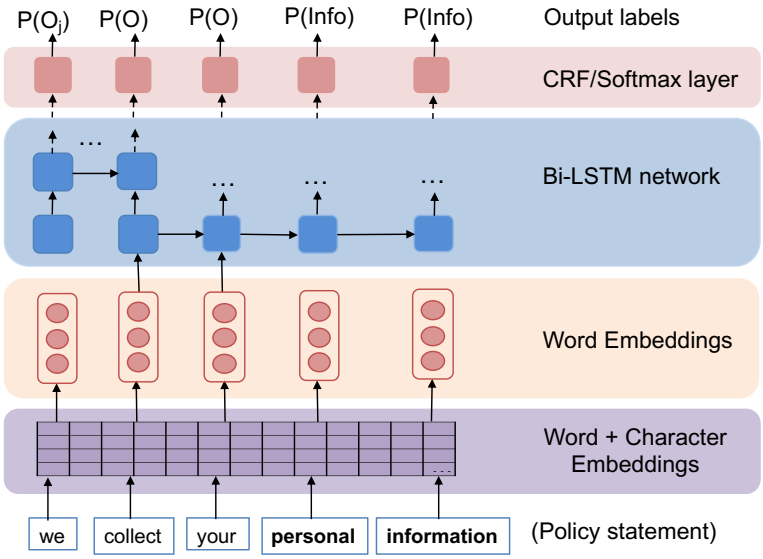
Policy	High Confidence Pairs	False Pairs by Worker	False Pairs by Experts	Ambiguous Pairs	Precision w/o Validation	Precision w/ Validation
Waze	107	20	30	12	0.73	0.88
Zynga	120	44	43	13	0.64	0.94
Flurry	71	11	16	4	0.79	0.92
FB	111	27	28	20	0.75	0.91
AOL	106	32	28	17	0.74	0.99

Next, we describe results from evaluating the identification of information types from privacy policy statements using a deep learning architecture.

5.3 Evaluating Deep Learning Approach for Information Type Identification

Identifying privacy goals using the approach defined in Section 5.2 can be time and effort intensive. We therefore did an exploratory study which we call Study SR8 to determine if a deep learning architecture could be used to automatically identify information types in an end to end system with no manual effort. We evaluated the architecture showed in Figure 15 for information type identification. The code for this architecture was modified from the codebase implemented by Genthial [Genthial 2017, Ma and Hovy 2016]

Figure 15. Deep Learning Architecture for Information Type Identification



In this model, the character and word embeddings for a given word are concatenated to build the embedding for each word. These word embeddings are then used as input to the deep neural network that uses Bi-directional LSTMs to compute the scores for each of the two tags: O (Outside) and Info (Information Type). The last layer uses either a Conditional Random Field or Softmax to predict the final tag for each word.

The datasets for this evaluation were creating using the method described in Study SR4 in Section 5.2.1. To prepare a policy for the dataset, we download the policy and remove the boilerplate language. Next, we itemize each statement and then construct 120-word paragraphs with the itemized statements to create the input file for crowdsourcing. The input file is then used to conduct a crowdsourcing task, where crowd workers are asked to identify and annotate the information types in each paragraph. Each paragraph is annotated by five crowd workers. We then collect this data and use the information type annotations that have been annotated by all the five workers to build our dataset. In Table 27 below, we present the number of policies in the training, development and testing datasets.

Table 27. Datasets for Evaluating Information Type Identification

Dataset	Training	Development	Test
No. of Policies	46	17	11
No. of "Info" Tags	3727	1303	907
No. of "O" Tags	36380	10438	8195

The model presented in Figure 15 predicted a total of 911 words as information types on the test dataset. Out of the 911 predicted information type words, 619 words were information types in the ground truth test dataset. However, the total number of information types in the ground truth test set were 907. Thus, the average precision is 619/911=0.68 and recall is 619/907=0.68.

While analyzing the results we observed that the model was in most cases correctly able to identify the frequently occurring information types such as personal information, name, credit card number among other such information types. The model was not able to identify the infrequent adjectives associated with root information types. The adjectives that constitute the information type which could not be identified include: non-identifiable, diagnostic, other, etc..

Analyzing the false positive we found that some of the words tagged as information types were indeed not information types, including words such as transfer, data integrity and security. However, while analyzing the false positives we also identified errors in the ground truth dataset which was built from the crowd workers annotations. For example, in the statement "We delete that information...," the model tagged the word *information* as an information type, whereas the test dataset had word *information* tagged with the O tag.

5.4 Summary Conclusion for Semantic Roles

We manually annotated and analyzed fifteen privacy policies across health, news and shopping domains to identify the different semantic roles and their values attached to the four different categories of data actions: collection, retention, use and transfer. From a total 698 instances of data actions, we identified 17 unique semantic roles which occur 2,316 times. The health policies were the most descriptive of the three domains, with 293 actions and 1,024 semantic roles, followed by shopping with 281 actions and 878 semantic roles. And the news policies were least descriptive with 124 data actions and 414 semantic roles instances across all actions.

The expected roles for the four categories of data action *were subject, information, condition,* and *purpose*. In addition, collection actions frequently have the source role to indicate from where the information was collected, and transfer actions have the target role to indicate to where the information was transferred. Missing values for these roles in a data practice statement leads to incompleteness in the data practice description and thus become a source of ambiguity. From our analysis, we observe that all the three domains had similar distribution of semantic roles. The health policy actions had the most subjects (85%) and conditions (41%) attached, whereas the news policies had the most purposes (42%) attached. In our dataset, on average 25% of the retention statements across all three domains were incomplete with respect to the subject role. In addition, 55% of transfer statements were incomplete with respect to the condition role, and 20% of usage statements were incomplete with respect to the purpose role. Most of the actions across all domains were performed by first party companies, followed by third party companies. We also observed that the most frequent source of user information were the users themselves in health (56%) and news (77%) policies, whereas for shopping policies the source was equally likely to be user (36%) and third parties (39%). When the information is being sourced from third parties, the user might not be aware of the actions being performed on the user's information and thus feel at greater risk. This was also evident from our risk study (See Chapter 6 Section 5), wherein participants perceived greater privacy risk when the information was being collected from them, as compared to the information being collected from third parties.

In our analysis, we identified a total of 74 unique lexical and syntactic patterns that occurred a total of 1,119 times in our dataset and can be used to specify semantic roles. We also observed that multiple lexical and syntactic patterns can be used to specify the same semantic role, for example the `if [value]` and depending `on [value]` pattern, among other such patterns, can be used to specify the condition semantic role. In other instances, we found that the same pattern can be used to specify different semantic roles, for example, the pattern to [value] can be used to specify the purpose, target, object and constraint roles. We also observed that in some cases, the semantic role specified by a pattern can be predicted from the action category it occurs with. For instance, in the shopping policies, the pattern `to [value]` specifies a data purpose in 98.3% of instances when attached to a usage action and specifies a target in 86.1% of instances when attached to a transfer action. Other patterns, such as `if [value]` and `when [value],` are

used to specify a condition, irrespective of the action category to which they are attached. It was also interesting to note that same patterns were used frequently across all three domains to specify the semantic roles. For example, the most frequent patterns used to specify condition semantic role across all three domains were `if [value]` and `when [value]`, and the patterns `to [value]` and `with [value]` were used to specify the target roles. Finally, we used preposition categorization to analyze the 74 unique patterns and observe the relationship between the syntax-based category and a pattern's observed semantic role value.

We have developed and evaluated a method that combines crowdsourcing and natural language processing (NLP) to extract goals from privacy policies. We find that untrained crowd workers can be used to elicit most of the actions and information types that were identified by the experts, which leads to high recall for the crowdsourcing action and information type micro tasks, when compared to the expert annotations. Moreover, we discovered that many false positives are due to natural ambiguities in the text and task description that are difficult to remove. A complementary finding about the performance of dependency parsing alone suggests that context and tacit knowledge are required to identify relevant actions and information type pairs. The crowd worker annotations, which are reasonably low cost to acquire, can be used as guidance for selecting parser dependencies to identify a set of high confidence pairs. Our results also show that these high confidence pairs contain most of the true positives as compared to the expert annotations, a minimal number of false positives that the hybrid approach identifies but were not identified by the experts and omit a minimal number of false negatives, that were identified by the expert annotators but missed by the hybrid approach.

The lexicon produced mixed results with respect the lexicon's utility in finding missing annotations. The lexicon increased recall, but at a high cost of precision, because the lexicon lacks contextual cues to distinguish when particular action and information type phrases are true positives. We observed that the lexicon reaches a saturation limit of between 42-84% in the domain of privacy policies, which suggests the lexicon will likely never become complete. Alternatively, the lexicon may be used to find annotations for common words and phrases that can be used to further reduce the number of tasks sent to crowd workers and thus the overall framework cost or can be used to solicit a higher number of workers to complete the reduced number of tasks for the same cost, thus reducing the probability of false negatives.

We also observed an improvement in precision as we sent the selected high confidence pairs back to the crowd for acceptance or rejection. Improvements in the front end of the framework (steps 3, 4 and B1 or B2 in Figure 8) could further reduce the number of pairs that need to be sent to the workers in step 5, further improving the overall performance.

We observed that the techniques used in the hybridized framework are complementary and can be used to address each other's weaknesses for improved performance. While we believe this technique could be applied to other domains with similar results, future work is needed to evaluate our approach in such domains.

In our evaluation of a deep learning model to automatically identify information types, we observed that information types can be identified with some amount of accuracy using deep learning approaches. The complexity of privacy policy statements makes it difficult to sometimes reliably extract semantic roles. This complexity is manifested in the form of statements that are very lengthy, have long lists of semantic roles, and have syntactic ambiguity. In addition crowdsourced data can be noisy and requires improvements. Our training datasets is also limited. We envision annotating more policies to build a larger dataset and also use data

augmentation techniques to apply transformation to our existing dataset to synthetically generate more data for training the deep learning model.

Chapter 6

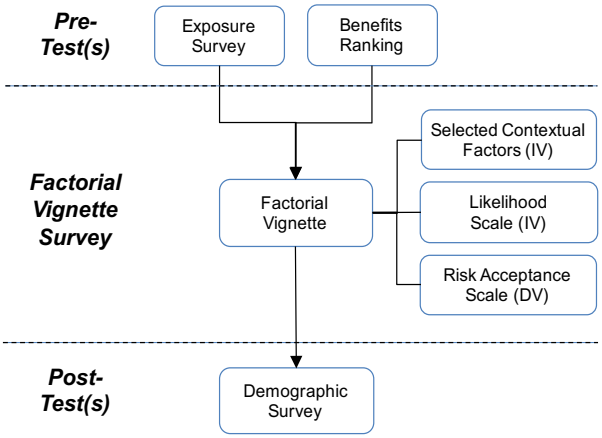
Privacy Risk Measurement Framework

In this Chapter, we describe the privacy risk measurement framework we developed to understand and measure perceived privacy risk, along with the study to measure the effect of vagueness and risk likelihood on perceived privacy risk [Bhatia et al. 2016a, Bhatia et al. 2016c, Bhatia and Breaux 2018b].

6.1 Framework for Measuring Perceived Privacy Risk

The privacy risk measurement framework for measuring privacy risk consists of a collection of surveys that are tailored to fit an information technology scenario. The surveys can be administered to actual or potential users of a system, to data subjects, or the general public. As shown in Figure 16, the framework consists of pre-tests, one or more vignette surveys, and post-tests. The pre-tests could measure participants' online behavior, their exposure to privacy risks and how they rank the technological benefits or privacy harms. The exposure surveys ask participants to report the frequency of their participation in online activities, such as online shopping or banking or searching for employment. In addition, the exposure survey asks participants about their experiences of privacy harms. The exposure survey is conducted as a pre-test prior to asking participants about their risk tolerances, or as a separate study to inform vignette design. Each vignette consists of a scenario with multiple contextual factors, a risk likelihood scale, and a risk acceptance scale. The scenarios situate participants in the context of a specific cost-benefit tradeoff. Finally, the vignette survey is followed by a post-test demographic survey to compare the sample population against standard demographics, such as age, gender, education level, and income. The post-survey helps determine the extent to which the collected risk measures will generalize to the population of interest.

Figure 16. Empirically validated framework to measure perceived privacy risk



We now discuss factorial vignette survey design, followed by the statistical method used to analyze the data, called multilevel modeling and lastly, the privacy risk study for measuring the effect of vagueness on perceived privacy risk.

6.2 Factorial Vignette Survey Design

Factorial vignettes provide a method to measure the extent to which discrete factors contribute to human judgment [Auspurg and Hinz 2014]. The factorial vignette method employs a detailed scenario with multiple factors and their corresponding levels, designed to obtain deeper insights, into a person's judgment and decision principles, than is possible using direct questions (i.e., with a prompt "Please rate your level of perceived risk" and a scale). Our factorial vignette survey design measures the interactions between the different independent variables, and their effect on a dependent variable, the person's *willingness to share* their personal information. This includes whether the different independent variables alone, in combination, or none of these factors affect willingness to share.

The factorial vignettes are presented using a template in which factors correspond to independent variables and each factor takes on a level of interest. For each factorial vignette survey (see Section 6.4), the factor levels replace an independent variable in the survey. The factors are often presented in the context of a scenario, which serves to situate the survey participant in a specific context. For example, a vignette may ask a participant to think about an online shopping experience with a website they routinely use, or to think about applying for a job online at an employment website. While the primary scenario does not change across vignettes, the embedded factors do change. For example, if we are interested in whether privacy risk changes when the vagueness changes, the survey designer can introduce a new factor `$VS` with four levels: conditionality, generalization, modality and numeric quantifier. For a between-subjects variable, a participant only sees and judges one level of the factor, whereas for a within-subjects variable, the participant sees all factor levels. In Figure 17, we present a vignette for an example study with two independent variables, which are vagueness (`$VS`), and data type (`$DT`), and a dependent variable, which is willingness to share (`$WtS`). The variable `$DT` is a within-subjects variable, which means that all the participants see and rate all the levels of this variable, whereas the variable `$VS` is between-subject variable, and each participant sees and rates only one level of this variable. In this vignette, the place holders for the variables are replaced by the values of the levels of these variables for each participant. For instance, for the variable vagueness, the variable placeholder `$VS` will be replaced by a statement with one category of vagueness. The semantic scale for `$WtS` consists of eight options starting from Extremely Unwilling (0) to Extremely Willing (8), part of the scale has been omitted for brevity (…).

Figure 17. Example Factorial Vignette

Please rate your willingness to share your information below with the Federal government, given the following statement about sharing of your information:

$VS

When choosing your rating for the information types below, consider the **$VS** above.

	Extremely Willing	Very Willing	Willing	Somewhat Willing	Somewhat Unwilling	...
Age Range	○	○	○	○	○	
Home Address	○	○	○	○	○	

Kaplan and Garrick define risk as a function of the probability and consequence, where consequence is the measure of damage [Kaplan and Garrick 1981]. More recently, NIST defines risk as the likelihood times the impact of an adverse consequence or harm [Stoneburner 2002]. One approach to measure probability or likelihood is to describe the number of people affected by the adverse consequence: the greater the number of people affected, the greater the probability is that the consequence may affect a randomly selected person. When considering how many people are affected by a consequence, prior research shows that lay people can map ratios (e.g., 1/10,000) to physical people much better than they can map probabilities (e.g., 0.0001%) [Fischhoff et al. 1978]. To evaluate this conclusion, we pilot tested a between-subjects risk likelihood factor with ratio-based likelihood levels. The risk likelihood had four levels, which were the ratios of people who experienced the privacy harm: 1/4, 1/10, 1/100 and 1/1,000. In the pilot study, we found no significant effects among the ratios, which suggests that participants perceive no greater privacy harm when the harm affects 1/4 people versus 1/1,000 people.

As an alternative to ratios, we designed a new risk likelihood scale based on construal-level theory from psychology. Construal-level theory shows that people correlate increased unlikelihood along four dimensions of increased spatial, temporal, social and hypothetical distances, than they do with shorter psychological distances along these four dimensions [Wakslak and Trope 2009]. We chose spatial and social distance as correlate measures of likelihood as follows: a privacy harm affecting only one person in your family is deemed a psychologically closer and more likely factor level than one person in your city or one person in your country, which are more distal and perceived less likely. The risk likelihood levels used in the framework are as follows, ordered from most likely and least hypothetical to least likely and most hypothetical:

- Only one person in your family
- Only one person in your workplace
- Only one person in your city
- Only one person in your state
- Only one person in your country

The evaluation of the risk likelihood scale is reported later in Section 5.4.

Risk has been described in terms of an individual's willingness to participate in an activity [Fischhoff et al. 1978], for example, one accepts the risk of a motor vehicle accident each time they assume control of a motor vehicle as the driver. To measure privacy risk, we propose to estimate a computer user's *willingness to share* data, including but not limited to personal data. The independent variable willingness to share (`$WtS`) is estimated from survey participant ratings on an eight-point, bipolar semantic scale, labeled at each anchor point: 1=*Extremely Unwilling*, 2=*Very Unwilling*, 3=*Unwilling*, 4=*Somewhat Unwilling*, 5=*Somewhat Willing*, 6=*Willing*, 7=*Very Willing* and 8=*Extremely Willing*. This scale omits the midpoint, such as "Indifferent" or "Unsure," which can produce scale attenuation when responses are prone to cluster, and which can indicate vague or ambiguous contexts rather than a respondent's attitude [Kulas and Stachowski 2013].

6.3 Multilevel Modeling Analysis Method

Multilevel modeling is a statistical regression model with parameters that account for multiple levels in datasets, and limits the biased covariance estimates by assigning a random intercept for each subject [Gelman and Hill 2007]. Multilevel modeling has been used to study interactions among security and privacy requirements [Bhatia et al. 2016a, Hibshi et al. 2015].

In our studies, the main dependent variable of interest is *willingness to share*, labeled `$WtS`. We conducted multiple studies, that have different independent variables of interest that affect our dependent variable `$WtS`. For the within-subject design, subject-to-subject variability is accounted for by using a random effect variable `$PID`, which is a unique identifier for each participant. Equation 2 below is our main additive regression model with a random intercept grouped by participant's unique identifier. The additive model is a formula that defines the dependent variable `$WtS`, *willingness to share*, in terms of the intercept α and a series of components, which are the different independent variables (`$IV1, $IV2` and so on). Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Equation 2 is simplified as it excludes the dummy (0/1) variable coding for reader convenience.

$$\$\mathrm{WtS} = \alpha + \beta_1 \$\mathrm{IV}_1 + \beta_2 \$\mathrm{IV}_2 + \cdots + \epsilon \qquad (2)$$

We analyze the data from our studies in R [R Core Team 2015] using the package lme4 [Bates et al. 2015]. We test the multi-level models' significance using the standard likelihood ratio test: we fit the regression model of interest; we fit a null model that excludes the independent variables used in the first model; we compute the likelihood ratio; and then, we report the chi-square, p-value, and degrees of freedom [Gelman and Hill 2007]. We performed a priori power analysis for each study using G*Power [Faul et al. 2007] to test for the required sample size for repeated measures ANOVA.

6.4 Risk Likelihood, Vagueness and Perceived Privacy risk

In this section, we describe the study design and results for the study we conducted to understand and measure how changes in vagueness and risk likelihood effect users' perception of privacy risk, which we call Study PR1.

6.4.1 Privacy Risk Perception Survey Design

In **Study PR1**, we designed our factorial vignette survey (described in Section 6.2) to measure the interactions between two independent variables, *vagueness* and *likelihood of privacy violation*, and their effect on a dependent variable, the Internet user's *willingness to share* their personal information. This includes whether vagueness or likelihood of violation alone, or neither of these two factors affect willingness to share. For this study, we chose to control several factors that affect willingness to share. For example, Nissenbaum argues that privacy and information sharing are contextual, meaning that the factors, data type, data recipient, and data purpose, affect willingness to share [Nissenbaum 2009]. We chose to control these factors by examining a single context that many Internet users engage in: shopping for products online [Horrigan 2008]. As suggested by Fischhoff et al., we presented the survey participants with numerous benefits while they were judging the specific privacy event [Fischhoff et al. 1978]. We conducted a brief one-hour, four-person focus group to elicit benefits of online shopping (as opposed to visiting a physical store), without considering potential harms of online shopping. The elicited benefits include: convenience, discounts and price comparisons, anonymous and discreet shopping, certainty that the product is available, wider product variety, and informative customer reviews.

As described in Section 6.2, we designed our risk likelihood scale to combine spatial and social distance as a correlate measure of likelihood (see Table 28): a privacy harm affecting *only one person in your family* is deemed a psychologically closer and more likely factor level than *one person in your city* or *one person in your country*, which are more distal and perceived less likely.

Table 28. Vignette Factors and Their Levels

Factors	Levels
Risk Likelihood ($RL)	only one person in your family
	only one person in your workplace
	only one person in your city
	only one person in your state
	only one person in your country
Vague Statement ($VS)	(C) We share your personal information as necessary.
	(G) We generally share your personal information.
	(M) We may share your personal information.
	(N) We share some of your personal information.

Factorial vignettes are presented using a template in which factors correspond to independent and dependent variables and each factor takes on a level of interest. The two independent factors are *Risk Likelihood* and *Vague Statement* with the levels described in Table 28. Figure 18 shows the vignette template: for each participant, each factor is replaced by one level. Because the independent variables are within-subjects factors, each participant sees and responds to all combinations of levels (4x5=20). Within-subject designs reduce subject-to-subject variability thereby increasing power.

For each vignette, participants rate their willingness to share their personal information on an eight-point, bipolar semantic scale, labeled: Extremely Willing, Very Willing, Willing, Somewhat Willing, Somewhat Unwilling, Unwilling, Very Unwilling and Extremely Unwilling.

Figure 18. Template used for vignette generation

(fields with $ sign are replaced with values selected from Table 28)

Please rate your willingness to share your personal information with a shopping website you regularly use, given the following benefits and risks of using that website.

Benefits: convenience, discounts and price comparisons, anonymous and discreet shopping, certainty that the product is available, wider product variety, and informative customer reviews

Risks: In the last 6 months, **$RiskLikelihood** experienced a privacy violation while using this website.

When choosing your rating, given the above benefits and risks, also consider the following website's privacy policy statements. Website privacy policies are intended to protect your personal information.

	Extremely Willing	Very Willing	Willing	Somewhat Willing	Somewhat Unwilling	...
$VagueStatement	○	○	○	○	○	

Before the vignettes, participants are presented a pre-survey to elicit their demographic characteristics (gender, age, race, education, income) and frequency of online behavior in six activities: using social networking sites; shopping for products or services; paying bills, checking account balances, or transferring money; searching for health information; using dating websites; and searching for jobs. The semantic scale response options for frequency of online behavior are: *a few times a day, once a day, few times a week, few times a month, few times a year, and never*.

In our study, the main dependent variable of interest is *willingness to share*, labeled `$WtS` in our model. The two fixed independent variables, which are within-subject factors, are risk likelihood labeled `$RL` (with five levels) and vague statement labeled `$VS` (with four levels). The independent exploratory variable `$Shopping` is based on the pre-test online behavior question about online shopping frequency and has two levels: S1 for participants who shop online a few times a week or more, and S0 for participants who shop less than a few times a week. For the within-subject design, subject-to-subject variability is accounted for by using a random effect variable `$PID,` which is unique to each participant.

The data is analyzed in R [R 2013] using the package lme4 [Bates et al. 2015]. Each participant sees all 20 combinations of our two within subject factors. Thus, our analysis accounts for dependencies in the repeated measures, calculates the coefficients (weights) for each explanatory independent variable, and tests for interactions. As described in Section 5.3 we test the multi-level models' significance using the standard likelihood ratio test: we fit the regression model of interest; we fit a null model that excludes the independent variables used in the first model; we compute the likelihood ratio; and then, we report the chi-square, p-value, and degrees of freedom [Gelman and Hill 2006]. We performed a priori power analysis using G*Power [Faul et al. 2007] to test for the required sample size for repeated measures ANOVA. The power analysis estimate is at least two participants per combination of the within-subject factors to achieve 95% power, and a medium effect size [Cohen 1988].

6.4.2 Perceived Privacy Risk Survey Results

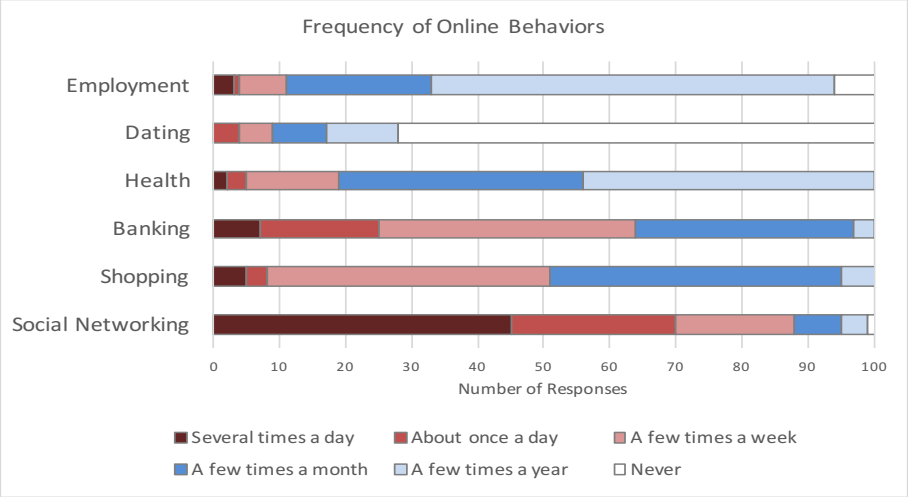
In Study PR1, we were interested in understanding and measuring how vagueness and risk likelihood affect user willingness to share personal information. We recruited 102 participants

using Amazon Mechanical Turk (AMT), where we paid $3 for completing the survey. We now discuss our results from the privacy risk perception survey (see Section 6.4.1).

6.4.2.1 Descriptive Statistics

A total 102 participants responded to our risk perception survey: 45.1% are female and 54.9% are male; 84.3% reported "white" as their ethnicity; 87.3% reported having at least some college level education; and 84.3% reported having annual household income less than $75,000. Figure 19 shows frequency of online behavior by participants. While 70% of respondents report viewing social networking sites daily, while 33% in a separate survey reported sharing personal information on these sites a *few times a week* or more.

Figure 19. Frquencies of Online Behaviors



6.4.2.2 Willingness to Share

Equation 3 below is our main additive regression model with a random intercept grouped by participant's unique ID, the independent within-subjects measure `$RL`, which is the likelihood of a privacy violation, and `$VS`, which is the vague privacy statement with a single vague term from one of the four categories (see Table 3 in Section 4.4.1). The additive model is a formula that defines the dependent variable `$WtS`, willingness to share, in terms of the intercept α and a series of components, which are the independent variables. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Eq. 3 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience.

$$\$WtS = \alpha + \beta_{RL}\$RL + \beta_{VS}\$VS + \epsilon \qquad (3)$$

To compare dependent variable `$WtS` across vignettes, we establish the baseline level for the factor `$RL` to be "only one person in your family" who experiences the privacy violation and, for the factor `$VS`, we set the vagueness category to *Condition*, "We share your personal information as needed". The intercept (α) is the value of the dependent variable, `$WtS`, when the independent variables, `$RL` and `$VS` take their baseline values.

We found a significant contribution of the two independent factors, for predicting the `$WtS` ($\chi^2(7)=875.15$, $p<0.000$), over the null model, which did not have any of the independent variables. In our model, we did not observe any effect of the interaction term `$RL*$VS`, ($\chi^2(12)=4.7$, $p=0.97$), which means vagueness and risk likelihood did not interact to affect the

willingness to share. In Table 29, we present the *Model Term*, the corresponding model-estimated *Coefficient* (along with the p-value, which tells us the statistical significance of the term over the corresponding baseline level), and the coefficient's *Standard Error*. In our survey, the semantic scale option *Extremely Unwilling* has a value of 1, and *Extremely Willing* has a value of 8. A positive coefficient in the model signifies an increase in willingness to share and a negative coefficient signifies a decrease in willingness to share.

Table 29. Study PR1 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Intercept (Family+Condition)	3.133***	0.164
Risk - only 1 person in your workplace	0.162*	0.080
Risk - only 1 person in your city	0.968***	0.080
Risk - only 1 person in your state	1.517***	0.080
Risk - only 1 person in your country	2.118***	0.080
Vagueness - generalization	-0.729***	0.072
Vagueness - modal	-0.155*	0.072
Vagueness - numeric	-0.218**	0.072

*p≤.05 **p≤.01 ***p≤.001

The results in Table 29 show that `$WtS` is significantly different and increasing for decreasing levels of `$RL`, as compared to the baseline level "only 1 person in your family". For the `$RL` level "only 1 person in your workplace", the `$WtS` increases by 0.16 over the baseline level, which is "only 1 person in your family", which denotes an increasing willingness to share. For the baseline `$VS` level "Condition," however, the `$WtS` is at the maximum. The `$VS` level "Generalization" shows a 0.73 decrease in the value of the dependent variable `$WtS`, as compared to the baseline level, which means generalization reduces the willingness to share.

6.4.2.3 Effect of the Online Behavior Shopping

We computed a new, two-level independent exploratory variable `$Shopping` based on the participant responses to the online behavior questions. The two levels correspond to the frequency that respondents shop online: *S1*, which is a few times a week or more, and *S0*, which is less than a few times a week. The new additive model in Eq. 4, below, has a component for the `$Shopping` variable. The new model in Equation 4 improves the prediction of the `$WtS` over the model in Eq. 3 ($\chi^2(1)=4.3$, $p<0.05$), which means respondents who shop more often express increased certainty about their willingness to share their personal information.

$$\$WtS = \alpha + \beta_{RL}\$Rl + \beta_{VS}\$VS + \beta_{S}\$Shopping + \epsilon \quad (4)$$

We found that participants who shop online a few times a week or more, are also more willing to share their personal information (`$WtS` is 0.62 higher than other participants), which means they may be more likely to comprehend the presented benefits of shopping while evaluating the risk.

6.5 Semantic Roles and Perceived Privacy Risk

We conducted three studies **Study PR2, PR3** and **PR4** to measure the effect of the presence or absence of different semantic roles across all four data action categories on the perceived privacy risk. To that end, we fixed the values of the subject role and object role to be "we," and "personal information," respectively. Table 30 presents the factors and corresponding factor level values for **Study PR2**. Figure 20 presents the factorial vignette survey text.

Table 30. Study PR2 Vignette Factors and Their Levels

Factors	Factor Level
Risk Likelihood ($RL) Between subject	only one person in your family
	only one person in your workplace
	only one person in your city
	only one person in your state
	only one person in your country
Data actions ($DA) Within subject	(C) Collection: collect
	(R) Retention: retain
	(U) Usage: use
	(T) Transfer: share
Semantic Role ($SR) Within subject	(DP) Data Purpose: to provide you services
	(Cond.) Condition: when you create an account with us
	(Source) Source: from you

Figure. 1. Template used for vignette generation

(fields with $ sign are replaced with values selected from Table 30 and Table 31)

Please rate your willingness to share your personal information with a shopping website you regularly use, given the following benefits and risks of using that website.

Benefits: convenience, discounts and price comparisons, anonymous and discreet shopping, certainty that the product is available, wider product variety, and informative customer reviews

Risks: In the last 6 months, **$RiskLikelihood** experienced a privacy violation while using this website.

When choosing your rating, given the above benefits and risks, also consider the following website's privacy policy statements. Website privacy policies are intended to protect your personal information.

	Extremely Willing	Very Willing	Willing	Somewhat Willing	Somewhat Unwilling	...
$Policy Statement	○	○	○	○	○	

The baseline policy statement for our survey was "We $DataAction your personal information," which includes the semantic roles subject and object associated with the data action. The policy statements $Policy Statement for each of the four actions are generated by adding one or more of the semantic roles from Table 30 to the baseline statement. For this survey, we have three different semantic roles, and therefore a total of eight policy statements for each action including the baseline statement, with all combinations of one or more of the semantic roles. For example, the *collection* statement with the roles *data purpose* and *condition* would be: "When you create an account with us, we collect your personal information to provide you services."

The next study, **Study PR3** has the same three dependent variables: risk likelihood, data action and semantic roles. The levels for the *risk likelihood* and *data action* variables are the same as the first study. Table 31 presents the additional factors and factor levels for the semantic roles used in Study PR3.

Table 31. Study PR3 Vignette Factors and Their Levels

Factors	Factor Level
Semantic Role ($SR) Within subject	(Cond.) Condition: with your consent
	(Source) Source: from you
	(Target) Target for the data action Transfer: third parties

In the grounded study, we categorized the role values for the condition, source and target roles (see Sections 5.1 and 5.1.3) The semantic role value categories can affect a user's perception of privacy risk. A user may be more willing to share their information, if the data action is *required by law*, as compared to if the action is performed *as necessary*, which is a vague condition. The most frequent roles in our policy statements after the *subject* and *object* roles were condition, source and target. The next semantic role risk study (**Study PR4**) has three pages with all the role value categories for a particular semantic role on each page. Table 32 presents the factor (a semantic role), the breakout for each semantic role category, followed by the factor levels, which is the semantic role value per category.

Table 32. Study PR4 Vignette Factors and Their Levels

Factors	Category	Factor Level
Condition (`$Cond`) Within subject	first party action	as part of your member profile
	legal action	if we are required to do so by law
	merger action	as part of a merger
	scope	as permitted by this privacy policy
	third party action	if third party service providers need access to your information
	user	with your consent
	vague	as necessary
Source (`$Source`) Within subject	technology	from your computer and mobile device
	third party	from third party sources
	user	from you
	vague	from various sources
Target (`$Target`) Within subject	first party	to us
	third party	to third parties
	location	globally
	technology	to servers
	vague	to others

Re-using the survey design from Figure 20, the `$Policy Statement` is generated by adding the semantic role value category to the baseline statement, "we transfer your personal information" for the *condition* and *target* roles, and "we collect your personal information" for the *source* role.

6.5.1 Results for Semantic Roles Privacy Risk Studies

Studies PR2 and PR3 for semantic roles privacy risk studies measure the effect of the presence and absence of the condition, source, purpose and target roles on the participant's willingness to share their information.

Equation 5 is our main additive regression model for semantic roles privacy risk Studies PR2, and PR3 with a random intercept grouped by participant's unique ID (ϵ), the independent within-subjects measure `$RL`, which is the likelihood of a privacy violation, `$DA`, which is the data action, and `$SR`, which is the semantic role (see Tables 30 and 31). The additive model formula defines the dependent variable `$WtS` (willingness to share) in terms of the intercept α and a series of components, which are the independent variables. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Eq. 5 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience.

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$SR + \epsilon \quad (5)$$

Tables 33 and 34 present the results for the baseline statement "We `$DataAction` your personal information." In Tables 33 and 34, the row baseline + semantic role(s) presents the value of the coefficient for the statement which is constructed by adding the semantic role(s) to the baseline statement. A positive coefficient signifies an increase in `$WtS` and a negative coefficient represents a decrease in `$WtS` over the baseline.

Table 33. Study PR2 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Intercept (DataAction-collect)	4.588***	0.378
Risk: only 1 person in your workplace	-0.242	0.524
Risk: only 1 person in your city	-0.697	0.524
Risk: only 1 person in your state	0.197	0.524
Risk: only 1 person in your country	0.021	0.524
Data Action: retain	0.097	0.068
Data Action: transfer	-0.413***	0.068
Data Action: use	0.039	0.068
Baseline+condition	0.006	0.096
Baseline+condition+purpose	0.397***	0.096
Baseline+condition+purpose+source	-0.444***	0.096
Baseline+condition+source	0.016	0.096
Baseline+purpose	0.478***	0.096
Baseline+purpose+source	0.313***	0.096
Baseline+source	-0.794***	0.096

*p≤.05 **p≤.01 ***p≤.001, 4=Somewhat Unwilling

We observe that adding the source role to the baseline statement (e.g., from you) decreases the participant's willingness to share. In addition, specifying the purpose role in any situation increases the willingness to share. Participants were less willing to provide their information when their data can be transferred as compared to when their data is collected by the website. Table 34 presents the modeling results for Study PR3.

Table 34. Study PR3 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Intercept (DataAction-collect)	3.795***	0.354
Risk: only 1 person in your workplace	0.078	0.496
Risk: only 1 person in your city	1.340	0.481
Risk: only 1 person in your state	0.791	0.488
Risk: only 1 person in your country	0.088	0.488
Data Action: retain	-0.222	0.088
Data Action: transfer	-1.341	0.088
Data Action: use	-0.328	0.088
Baseline+condition	0.744***	0.088
Baseline+source	0.081	0.088
Baseline+target	-0.141	0.149
Baseline+condition+source	0.784***	0.088
Baseline+condition+target	0.684***	0.149
Baseline+source+target	-0.104	0.149
Baseline+condition+source+target	0.659***	0.149

*p≤.05 **p≤.01 ***p≤.001, 4=Somewhat Unwilling

In Study PR3, we observe that adding the condition role, which concerns seeking consent from the user before their data is acted upon, considerably increases the participant's willingness to share their information. In both surveys, we did not observe any statistically significant difference among the levels of the factor *risk likelihood*.

We now report results from Study PR4 to measure the effect of role values on perceived privacy risk. The policy statements for this survey were generated by adding the role value category to the baseline statement, "we transfer your personal information" for the condition and target roles, and "we collect your personal information" for the source role.

In equations 6.1, 6.2, and 6.3 below we present our main additive regression models for Study PR4, with a random intercept grouped by participant's unique ID (ϵ), the independent within-subjects measure `$RL`, which is the likelihood of a privacy violation, and `$DA`, which is the data action, and `$Cond` which is the condition role, `$Source` which is the source role, `$Target` which is the target role, (see Table 32).

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$Cond + \epsilon \quad (6.1)$$

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$Source + \epsilon \quad (6.2)$$

$$\$WtS = \alpha + \beta_{R}\$RL + \beta_{DA}\$DA + \beta_{DA}\$Target + \epsilon \quad (6.3)$$

The baseline for the condition category is "first party," the baseline source is "technology," and the baseline target is "first party." The results appear in Table 35.

We observe from Table 35 that when information will be transferred on condition of a user consent action, as required by law, or as permitted by the policy, elsewhere, the user's willingness to share increases above the baseline. On the other hand, third-party condition ("if third party service providers need access to your information") decreases the willingness to share below the baseline, whereas the differences between merger and vague condition as compared to the baseline condition are not statistically significant. We observed that the user's willingness to share increases when the information is collected from the user, directly, as compared to when it is collected from their computer or mobile device. With respect to the target role, the user's willingness to share decreases when the information is transferred to third parties, or the target role value is vague.

Table 35. Study PR4 Multilevel Modeling Results

Term	Coeff.	Stand. Error
Semantic Role: Condition, baseline: "first party action"		
Intercept (first party)	3.113***	0.355
Condition: legal	1.788***	0.196
Condition: merger	-0.188	0.196
Condition: scope	0.775***	0.196
Condition: third party	-0.875***	0.196
Condition: user	2.213***	0.196
Condition: vague	-0.150	0.196
Semantic Role: Source, baseline: "technology"		
Intercept (technology)	2.325***	0.399
Source: third party	0.100	0.173
Source: user	2.000***	0.173
Source: vague	0.163	0.173
Semantic Role: Target, baseline: "first party"		
Intercept (first party)	3.245***	0.330
Target: location	-1.775***	0.159
Target: technology	-0.050	0.159
Target: third party	-1.438***	0.159
Target: vague	-1.525***	0.159

*p≤.05 **p≤.01 ***p≤.001, 4=Somewhat Unwilling

6.6 Summary Conclusions from the Perceived Privacy Risk Study

The terms in the vagueness taxonomy are associated with two semantic roles: the action performed on the information and the information type. In Study PR1 while we did not observe an interaction between risk likelihood and vagueness on willingness to share personal information, there may be an interaction with respect to specific roles, e.g., vague disclosure recipients may be perceived as higher risk ambiguities, than the type of information disclosed.

We conclude from the results that *willingness to share* increases as a participant's social and physical distance from the person experiencing the privacy violation (`$RL`) increases. This means that the users' perception of privacy risk increases, when they think about a person from their family or workplace experiencing the violation, as compared to the experience of a person somewhere in their state or country. We also found that the *willingness to share* is highest for the least vague category *Condition*, as compared to other vague categories, and *willingness to share* was the lowest for *Generalization*, which is the most vague category in Figure 3, Section 4.4.2 and Table 29 in Section 6.4.2.2. Furthermore, there was no statistically significant difference between willingness to share for *Modality* and *Numeric Quantifier* (p=0.38), which have similar vagueness measures. The inverse decrease in *willingness to share* due in the presence of increased vagueness is in contrast to Acquisti and Grossklags, who found that a user is less likely to protect their personal information in presence of benefits with missing information about data use [Acquisti and Grossklags 2005]. The explanation offered is that the missing information leads the user to not think about the risk [Acquisti and Grossklags 2005]. In our study, the vague terms are signals that information is missing, which may explain why users reduce their willingness to share.

We also conducted three studies (Study PR2, Study PR3, Study PR4) to measure the effect of presence or absence of semantic roles and their categories of values on privacy risk. We observe

that that describing the purpose for which the user's data will be acted upon considerably increases the user's willingness to share their information. Similarly, specifying that the user's data will be acted upon only under the condition that the user has consented, increases the willingness to share information. In Study PR2, adding the source role with the value "from you" decreased the user's willingness to share their information. In this survey, there was no other value of the *source* role. One explanation may be that participants assume that the source suggests the collected information is more sensitive or personal, or that it is collected automatically without user consent. In Study PR3, we observed that adding the condition role, which concerns seeking consent from the user before their data is acted upon, considerably increases the participant's willingness to share their information. In Study PR3 we also saw an increase in participant's willingness to share their information when the source was added to the baseline statement, as compared to Study PR2 where the condition was "when you create an account with us." The participants see multiple statements on the same page in the survey which includes the statements with conditions. The condition value in Study PR3 "with your consent" could have primed participants to think about the other statements more positively.

In Study PR4, we observe that participant's willingness to share increases when the information is collected from the user directly, as compared to when the information is collected from third parties, or when the source of the information is vague. Participants were also shown multiple sources from which their information could be collected, including from their devices, third parties, and instances where the source role value is vague. These additional sources may have implied that "from you" excludes automated sources in which participants would not be directly involved in the collection process, in other words, there was an anchoring effect. By comparing the sources from which their information is collected, the users may have felt that they have more control over their information, when they directly provide it to the website, as compared to information about them that can be collected by the website from other sources outside their control. Participants were most willing to share their information when they consented to the transfer, or when the transfer was required by law. In addition, participants perceived the least risk when the information was being transferred to the first party company, compared to other targets.

Chapter 7

Future Work

In this section, we list two future directions. The first is the development of a text-based dialogue system using semantic frames as an intermediate representation. The second is supporting the standardization of privacy by design principles by using the findings from my privacy risk framework.

7.1 Dialogue Systems using Semantic Frames Representation

Internet users cannot reasonably review the hundreds of privacy policies governing the services they use and not all users have the same questions about how their data is used. Furthermore, tracing data practices to program functions could help developers write code that is consistent with the data practices. A better way to communicate with users and developers would be to use a closed domain dialogue system that can converse with and answer policy-relevant questions. The semantic frame-based representation that I developed (described in Section 5.1) can be used as an intermediate representation for this system.

Presently, the frames are limited to a single statement and data action. To answer policy questions, we need to combine semantics from multiple statements. Consider the example question, "is my home address being used for anything other than shipping?" To answer this question, we would need to combine all purpose role values from all frames that contain home address as the information type, in addition to the statements that describe purposes for parts of a home address, such as street address, state, and zip code. This requires an ontology of concepts that appear in privacy policies [Bhatia et al. 2016d, Evans et al. 2017, Hosseini et al. 2018]. Another challenge is how to use the context of the conversation to ensure the user and the system are in agreement. This raises the following technical challenges in dialogue systems:

7.1.1 Grounding:

The process of updating the common ground in a conversation based on contributions in the conversation is called grounding [Stent and Bangalore 2014]. This includes performing contribution tracking, wherein the system builds on a partial understanding of previous contributions, and revises and reframes its contributions. For example, if the user asks, "for what purposes is information being used?", the system can ask a follow up question, "did you mean contact information?" to clarify what category of information the user is asking about given the conversation history. Grounding strategies include: asking clarification questions, using confirmation strategies, such as repeating the information and receiving an affirmation, providing expansion of meaning by providing additional information, if the user cues that they have not understood the presented information.

7.1.2 Compound Contributions

Compound contributions are semantic or syntactic units of conversation that expand or complete a given contribution. For instance, a user might ask, "with whom is the contact information being shared?" and then after the system provides the answer, the user might ask, "who else?" expanding their earlier contribution to the conversation. This would require the system to have capabilities of (a) incremental interpretation: producing and accessing semantic representations for partial constituents and (b) incremental representation: accessing lexical, syntactic and semantic information presented in the constituent units thus far.

7.2 Supporting Privacy by Design using Privacy Risk Measurements

I envision using the privacy risk framework to support the standardization of the privacy by design paradigm. Below, I describe two aspects of privacy by design and how the risk framework can support these aspects:

7.2.1 Privacy as a Default Setting

One of the principles of privacy by design is to set default setting that automatically protect the privacy of the user. However, designers cannot always anticipate a user's perception of privacy risk, nor should designers necessarily treat all personal data as high-risk. In traditional practice, designers may use their personal judgment or user personas to make design decisions related to user privacy, which may also underestimate the risk, e.g., when the designer estimates the risk of a person of a different ethnicity or age range. The proposed privacy risk measurement framework can be used to better inform design decisions regarding setting default made during software development by surveying prospective users. For example, in context of a social networking website, more risky information types such as personal identifiers, home address, and pictures posted by a user should by default be visible to only a limited set of users. In addition, such data should not be used by the service for any secondary purposes other than registration. Whereas, less risky information such as age range and country of residence could be visible to a broader audience by default. Similarly, for risky data practices such as unrestricted sharing of user information with third parties to provide personalized services and targeted advertisements the default should be set to "no sharing." The sharing in this case should be permitted with explicit user consent as it decreased the perceived privacy risk. However, the first party could use user's profile to provide relevant content to the user, if the perceived risk is measured to be low.

The privacy risk framework I have developed can thus be used to survey expected users with the relevant contextual factors (such as information types, target, purpose, conditions) to inform the design of default settings for various applications. These default settings should be such that the user should not have to perform any additional actions to achieve the maximum degree of privacy in the product.

7.2.2 Proactive not Reactive Design

Privacy by design aims to measure privacy risks before they happen and take measures to mitigate the risk, rather than wait for the risks to materialize and then take remedial actions. The privacy risk framework I have developed can be used to understand how privacy risk varies under different contextual factors. This in turn can help designers allocate resources more carefully when collecting, sharing, and securing high and moderate risk data. For instance, data

redaction techniques could be applied to high and moderate risk data before sharing it with third parties for secondary purposes perceived to be low-benefit to data subjects. Similarly, data belonging to high risk users such as children should be collected, retained, transferred and used more restrictively.

Some of the challenges in supporting the standardizing of the two aspects describe above are:

(1) **Identifying relevant factors and confounds:** For each new product or service type, we would have to first determine the relevant contextual factors that affect the risk in the given scenario. Given the complexity of online services, in addition to the factors identified affecting privacy risk, there could be additional factors that act as confounding variables for privacy risk measurements.

(2) **Survey results to design decisions:** Mapping the results from the surveys to actionable design decisions is a complex activity. In addition to the findings from the surveys, we have to take into account (a) company policies (b) applicable laws (c) the quality of the features the product aims to provide and (d) the highly interconnected ecosystem of products and services that are provided by a single company or collection of companies. For instance, an important feature of a shopping website is providing personalized recommendations, which are based on a user's past browsing and buying history. If the website does not use a user's past history and profile to provide relevant recommendations the quality of experience for the users of the website goes down. In addition, software requirements which effect the design of most of the systems are cross-cutting. These cross-cutting requirements are a consequence of highly complex and interconnect ecosystem of products and services either provided by the same company or different companies such as advertising platforms that are used by multiple corporations. Thus, the privacy risk framework is the first step towards going from empirical privacy risk measurements to actionable items for design decisions. However, more work is required to study how these different aspects (privacy risk measurements, company policies, laws, product features and complex ecosystem of products and services) can be leveraged together to make design decisions that systematically decrease privacy risk for the users.

Chapter 8

Conclusions

Ambiguous privacy policies fail to provide their users with adequate or appropriate notice of treatment of their personal information, undermine their ability as regulatory mechanisms, and can in turn lead to an increase in privacy risk as perceived by the users. These concerns motivate our proposed thesis which is to identify and measure ambiguity in privacy policies which includes vagueness and incompleteness, and to develop an empirically validated framework to measure the associated perceived privacy risk.

In this thesis, we propose a theory of vagueness which consists of three main parts: a taxonomy of vague terms and their categorization which is based on grounded analysis, a technique to measure the relative inter-and intra-category vagueness using paired comparisons, and an explanation for differences in vagueness based on different semantic functions. We measure incompleteness in privacy policies by identifying semantic roles that describe the context for a given data action. We have also developed a semi-automated hybridized framework to identify privacy goals from privacy policy statements. In addition, in this thesis we also present an empirically validated framework to measure the effect of different contextual factors on users' perception of privacy risk. Using this framework, we show that increase in vagueness leads to an increase in perceived privacy risk and the presence of semantic roles condition and purpose decrease privacy risk.

In summary, we introduce an approach to identify and measure ambiguity and the associated privacy risk in this thesis. We envision that the results and observations from our studies can be used to provide companies with mechanisms to improve drafting, enable regulators to easily identify ambiguous privacy policies especially ambiguity associated with high risk components such as sensitive data types, empower regulators to more effectively target enforcement actions, and help software designers make better and more informed decisions about software design during the software development phase taking into account the perceived privacy risk.

Appendix A: Extracted Semantic Roles

We identified 17 total semantic roles in our analysis, six of which are described in Section 5.1.

The remaining roles are as follows:

- *Action location*: The location where the action is performed.
- *Comparison*: Comparison of the action with other action(s).
- *Constraint*: The restrictions on the action.
- *Duration*: The duration for which the action will be performed.
- *Exception*: Describes an exception to the action.
- *Retention property*: This role describes how the information is retained. Example role value from Costco policy: separately from other member databases.
- *Hypernymy*: A more generic semantic role value with specific values.
- *Instrument*: The medium with which the action is performed.
- *Negation*: The presence of this role signals that the action will not be performed.
- *Retention location*: The location at which the object of the retention action is retained.
- *Time of action*: The time at which the action is performed.

Appendix B: Semantic Roles Frequency

The following table presents statistics, including: the total number of data actions identified in each data action category (Total Actions); the number of role value instances for the most frequent roles and the total number of roles attached to each data actions category (Total Roles), for each policy.

Table B.I. Frequency of Semantic Roles Across Health Policies

Policy	Category	Total Actions	Subject	Object	Cond-ition	Pur-pose	Total Roles
Healthvault	C	7	5	7	2	3	23
	R	9	8	9	4	0	28
	U	14	13	14	4	11	48
	T	9	8	9	4	3	35
Mayo Clinic	C	1	0	1	0	1	4
	R	1	0	1	0	1	2
	U	17	16	17	11	11	64
	T	40	34	40	20	14	137
MyFitness	C	6	6	6	2	3	21
	R	0	0	0	0	0	0
	U	13	10	13	2	12	13
	T	19	18	19	7	8	19
WebMD	C	14	14	14	5	3	52
	R	8	8	8	3	2	26
	U	21	19	21	4	16	80
	T	15	13	15	6	2	57
23andMe	C	19	15	19	8	3	65
	R	15	12	15	8	4	47
	U	40	28	40	20	29	126
	T	25	21	25	8	4	89
Total		293	248	293	118	130	1024

C: Collection, R: Retention, U: Usage, T: Transfer

Table B.II. Frequency of Semantic Roles Across News Policies

Policy	Cat-egory	Total Actions	Subject	Object	Cond-ition	Pur-pose	Total Roles
ABC News	C	5	5	5	3	1	16
	R	1	1	1	0	0	2
	U	2	2	2	1	2	7
	T	6	6	6	1	3	22
Bloomberg	C	2	2	2	0	0	7
	R	2	1	2	0	0	5
	U	9	6	9	0	9	24
	T	4	4	4	0	0	14
CNN	C	5	5	5	1	2	17
	R	0	0	0	0	0	0
	U	18	13	18	4	16	57
	T	6	5	6	3	2	20
Fox News	C	7	7	7	5	0	22
	R	7	5	7	2	4	24
	U	12	10	12	4	9	43
	T	9	8	9	3	2	33
Washpost	C	11	10	11	4	4	43
	R	1	1	1	1	0	4
	U	10	6	10	0	6	27
	T	7	5	7	5	1	27
Total		124	102	124	37	61	414

C: Collection, R: Retention, U: Usage, T: Transfer

Table B.III. Frequency of Semantic Roles Across Shopping Policies

Policy	Cat-egory	Total Actions	Subject	Object	Cond-ition	Pur-pose	Total Roles
Barnes and Noble	C	30	29	30	16	6	89
	R	7	6	7	4	3	24
	U	22	20	22	4	17	69
	T	24	18	24	12	1	76
Costco	C	16	13	16	4	2	38
	R	4	1	4	0	0	10
	U	16	14	16	5	12	49
	T	28	24	27	20	4	97
JC Penny	C	20	19	20	9	2	69
	R	1	1	1	0	0	2
	U	19	13	19	0	17	51
	T	12	10	12	4	3	40
Lowes	C	14	14	14	3	2	52
	R	5	3	5	2	2	13
	U	12	10	12	0	10	34
	T	15	14	15	10	2	52
Overstock	C	10	10	10	4	2	32
	R	2	2	2	1	0	6
	U	16	16	16	1	13	46
	T	8	8	8	3	0	29
Total		281	245	280	102	98	878

C: Collection, R: Retention, U: Usage, T: Transfer

Appendix C: Semantic Roles Frequency

The following table presents the study IDs, brief description of the study and the section the study is presented in.

Table C.I Studies

Study ID	Study Description	Sections
Study V1	Building a vagueness taxonomy using content analysis	Section 4.1, 4.4.1
Study V2	Ranking vagueness using Bradley Terry model.	Section 4.2, 4.4.2
Study V3	Scoring policies for vagueness	Section 4.3, 4.4.3
Study SR1	Identifying semantic roles and incompleteness for privacy policies	Section 5.1, 5.1.1, 5.1.2
Study SR2	Categorizing condition, source, target and subject role values	Section 5.1, 5.1.1, 5.1.3
Study SR3	Analyzing lexical and semantic patterns for semantic role specification	Section 5.1, 5.1.1, 5.1.4
Study SR4	Crowd worker micro annotations to identify actions and information types	Section 5.2, 5.2.1, 5.2.5.1
Study SR5	Selecting Action-information type pairs using dependency parsing	Sections 5.2, 5.2.2, 5.2.5.2
Study SR6	Building and re-using action and information type lexicon	Sections 5.2, 5.2.3, 5.2.5.3, 5.2.5.4
Study SR7	Validating action-information type pairs using crowdsourcing	Sections 5.2, 5.2.4, 5.2.5.5
Study SR8	Evaluating deep learning model for identification of information types	Sections 5.3
Study PR1	Measuring the effect of vagueness and risk likelihood on perceived privacy risk	Section 6.4
Study PR2	Measuring the effect of presence and absence of semantic roles on perceived privacy risk	Section 6.5
Study PR3	Measuring the effect of presence and absence of semantic roles and their values on perceived privacy risk	Section 6.5
Study PR4	Measuring the effect of semantic role value categories on perceived privacy risk	Section 6.5

Bibliography

[Aarts 2011] Bas Aarts. Oxford modern English grammar. Oxford University Press, 2011 [

[Antón and Earp 2004] A.I. Antón, J.B. Earp, "A requirements taxonomy for reducing web site privacy vulnerabilities," *Req'ts Engr. J.,* 9(3):169-185, 2004.

[Acquisti and Grossklags 2005] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Security and Privacy,* vol. 3, no. 1, pp. 26–33, 2005.

[Acquisti et al. 2013] A. Acquisti, L.K. John, G. Lowestein. "What is the price of privacy," *Journal of Legal Studies*, 42(2): Article 1, 2013.

[Acquisti et al. 2017] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. Cranor, S. Komanduri, P. Leon, N. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson, "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online," *ACM Comput. Surv.* 50, 3, Article 44 (August 2017). Available at SSRN: https://ssrn.com/abstract=2859227

[Auspurg and Hinz 2014] K. Auspurg and T. Hinz. *Factorial Survey Experiments*. Sage Publications, 2014.

[Baker et al. 1998] Collin F. Baker, Charles J. Fillmore, and John B. Lowe, "The Berkeley FrameNet Project," In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1 (ACL '98), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 86-90.

[Bates et al. 2015] D. Bates, M. Maechler, B. Bolker, S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Soft*., 67(1): 1-48, 2015.

[Bauer 1960] R.A. Bauer, "Consumer behavior as risk-taking, dynamic marketing for changing world," *American Marketing Association*, Chicago, 389, 1960.

[Bellman et al. 2004] S. Bellman, E. J. Johnson, S. J. Kobrin, and G. L. Lohse, "International Differences in Information Privacy Concerns: A Global Survey of Consumers," Information Society 20, no. 5 (2004): 313-24.

[Berendt et al. 2005] B. Berendt, O. Günther, and S. Spiekermann, "Privacy in e-commerce: Stated preferences vs. actual behavior," *Communications of the ACM*, vol. 48, no. 4, pp. 101–106, 2005.

[Berry et al. 2003] D.M. Berry, E. Kamsties, M.M. Krieger. "From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity," Univ. of Waterloo, Tech. Rep., Nov. 2003.

[Bhatia and Breaux 2015] Jaspreet Bhatia, Travis D. Breaux, "Towards an Information Type Lexicon for Privacy Policies," *IEEE 8th International Workshop on Requirements Engineering and Law (RELAW),* Ottawa, Canada, pp. 19-24, Aug. 2015.

[Bhatia et al. 2016a] J. Bhatia, T.D. Breaux, J.R. Reidenberg, T.B. Norton, "A Theory of Vagueness and Privacy Risk Perception," *24th IEEE International Requirements Engineering Conference (RE'16)*, Beijing, China, 2016.

[Bhatia et al. 2016b] J. Bhatia, T.D. Breaux, F. Schaub. "Privacy goal mining through hybridized task re-composition," *ACM Trans. Soft. Engr. Method*., 25(3): Article 22, 2016.

[Bhatia et al. 2016c] Jaspreet Bhatia, Travis D. Breaux, Liora Friedberg, Hanan Hibshi, Daniel Smullen, "Privacy Risk in Cybersecurity Information Sharing." ACM 3rd Workshop on Information Sharing and Collaborative Security (WISCS), Vienna, Austria, October 2016.

[Bhatia et al. 2016d] Jaspreet Bhatia, Morgan C. Evans, Sudarshan Wadkar, Travis D. Breaux, "Automated Extraction of Regulated Information Types using Hyponymy Relations." IEEE 3rd International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), Beijing, China, Aug. 2016.

[Bhatia and Breaux 2017] Jaspreet Bhatia, Travis D. Breaux, "A Data Purpose Case Study of Privacy Policies," *25th IEEE International Requirements Engineering Conference*, RE:Next! Track, Lisbon, Portugal, 2017.

[Bhatia and Breaux 2018a] Jaspreet Bhatia, Travis D. Breaux, "Semantic Incompleteness in Privacy Policy Goals," Distinguished Paper Award, 2*6th IEEE International Requirements Engineering Conference,* Banff, Canada, 2018.

[Bhatia and Breaux 2018b] Jaspreet Bhatia, Travis D. Breaux, "Empirical Measurement of Perceived Privacy Risk." Accepted to: ACM Transactions on Computer-Human Interaction (ACM TOCHI)

[Bhatia et al. 2018] Jaspreet Bhatia, Morgan C. Evans, Travis D. Breaux, "Identifying Incompleteness in Privacy Policies using Semantic Frames," Invited Paper, Requirements Engineering Journal.

[Bengio et al. 2001] Yoshua Bengio, R´ejean Ducharme, and Pascal Vincent, "A Neural Probabilistic Language Model," In *Advances in Neural Information Processing Systems* 13 (NIPS'00), pages 932938, MIT Press, 2001.

[Boyd et al. 2005] S. Boyd, D. Zowghi, and A. Farroukh, "Measuring the expressiveness of a constrained natural language: an empirical study," *13th IEEE Int'l Req'ts Engr. Conf.,* pp. 339-352, 2005.

[Braz et al. 2005] R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons, "An inference model for semantic entailment in natural language," In National Conference on Artificial Intelligence (AAAI), pages 1678–1679, 2005.

[Breaux and Antón 2007] T.D. Breaux and A.I. Antón, "Impalpable constraints: Framing requirements for formal methods," Technical Report Technical Report TR-2006-06, Department of Computer Science, North Carolina State University, Raleigh, North Carolina, February 2007.

[Breaux and Schaub 2014] T.D. Breaux, F. Schaub. "Scaling requirements extraction to the crowd: experiments on privacy policies," *22nd IEEE Int'l Req'ts Engr. Conf*., pp. 163-172, 2014.

[Breaux et al. 2014] Travis D. Breaux, Hanan Hibshi, and Ashwini Rao. "Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements," Requirements Engineering Journal 19, 3 (September 2014), 281-307. DOI:http://dx.doi.org/10.1007/s00766-013-0190-7.

[Breaux et al. 2015] Travis D. Breaux, Daniel Smullen, and Hanan Hibshi. "Detecting Repurposing and Over-collection in Multi-Party Privacy Requirements Specifications," 23rd IEEE International Requirements Engineering Conference. IEEE Computer Society, Washington, D.C., 166-175. DOI:10.1109/RE.2015.7320419.

[Carreras and Màrquez 2005] Xavier Carreras and Lluís Màrquez, "Introduction to the CoNLL-2005 shared task: semantic role labeling," *Conference on Computational Natural Language Learning (CONLL '05)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 152-164.

[Charniak 2000] Eugene Charniak, " A maximum-entropyinspired parser," *1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Cohn and Blunsom 2005] Trevor Cohn and Philip Blunsom, "Semantic role labelling with tree conditional random fields," *Ninth Conference on Computational Natural Language Learning (CONLL '05)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 169-172.

[Cohen 1988] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 1988.

[Collins 2003] Michael Collins, "Head-driven statistical models for natural language parsing," *Comput. Linguist.,* 29(4):589–637, December 2003.

[Collobert et al. 2011] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, "Natural language processing (almost) from scratch," *Journal of Marchine Learning Research*, 12:2493–2537, November.

[Cranor 2006] L.F. Cranor, P. Guduru, and M. Arjula, "User interfaces for privacy agents," *ACM Trans. Comput.-Hum. Interact.* 13, 2 (June 2006), pp. 135-178.

[Creswell 2008] R. Creswell. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 3rd ed. SAGE Publications, 2008.

[Dalpiaz 2018] Fabiano Dalpiaz, Ivor van der Schalk, Garm Lucassen, "Pinpointing Ambiguity and Incompleteness in Requirements Engineering via Information Visualization and NLP," Requirements Engineering: Foundation for Software Quality 2018, pp. 119-135.

[Das et al. 2014] Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith, "Frame-semantic parsing," Comput. Linguist. 40, 1, March 2014

[Das et al. 2017] Arjun Das, Debasis Ganguly, and Utpal Garain, "Named Entity Recognition with Word Embeddings and Wikipedia Categories for a Low-Resource Language," *ACM Trans. Asian Low-Resour. Lang. Inf. Process,* 16, 3, Article 18 (January 2017), 19 pages. DOI: https://doi.org/10.1145/3015467

[David 1988] H. A. David. *The Method of Paired Comparisons*. 2nd ed. Oxford University Press, 1988.

[Denger 2002] C. Denger. *High Quality Requirements Specifications for Embedded Systems through Authoring Rules and Language Patterns*. M.Sc. Thesis, Fachbereich Informatik, Universität Kaiserslautern, Germany 2002.

[Evans et al. 2017] M. C. Evans, J. Bhatia, S. Wadkar, T. D. Breaux, "An Evaluation of Constituency-based Hyponymy Extraction from Privacy Policies," Accepted To: *25th IEEE International Requirements Engineering Conference (RE'17),* Lisbon, Portugal, 2017.

[Fabbrini et al. 2001] F. Fabbrini, M. Fusani, S. Gnesi, and G. Lami, "The linguistic approach to the natural language requirements, quality: benefits of the use of an automatic tool," *26th IEEE Comp. Soc.-NASA GSFC Soft. Engr. W'shp,* pp. 97-105, 2001.

[Farkas et al. 2010] R. Farkas, V. Vincze, G. Móra, J. Csirik, G. Szarvas, "The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text," *14th Conf. Comp. NL Learning-Shared Task*, pp. 1-12, 2010.

[Faul et al. 2007] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. Methods*, 39(2): 175-191, 2007.

[Fernández and Wagner 2015] Daniel Méndez Fernández, Stefan Wagner, Naming the pain in requirements engineering: A design for a global family of surveys and first results from Germany, Information and Software Technology, Volume 57, 2015, Pages 616-643.

[Fikes and Kehler 1985] R. E. Fikes and T. Kehler, "The role of frame-based representation in knowledge representation and reasoning," Communications of the ACM 28(9), pp.904-920, 1985.

[Fillmore 1976] C. J. Fillmore, "Frame Semantics and the Nature of Language," Annals of the New York Academy of Sciences, 280: 20–32, 1976.

[Fischhoff et al. 1978] B. Fischhoff, P. Slovic, S. Lichtenstein, S. Read, B. Combs, "How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits," *Policy Sci.* 9: 127-152, 1978.

[Fleiss 1971] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psych. Bulletin,* 76(5): 378-382, 1971.

[Fuchs and Schwitter 1995] N. E. Fuchs, R. Schwitter, "Specifying logic programs in controlled natural language," *Workshop on Comp. Logic for NLP*, pp. 3-5, 1995.

[Gause 1989] D.C. Gause, G.M. Weinberg. *Exploring Requirements: Quality Before Design.* Dorset House, 1989.

[Gelman and Hill 2007] A. Gelman and J. Hill, "Data analysis using regression and multilevel/hierarchical models," *Policy Anal.,* pp. 1-651, 2007.

[Gildea and Jurafsky 2002] Daniel Gildea and Daniel Jurafsky, " Automatic labeling of semantic roles," *Comput. Linguist.* 28, 3 (September 2002), 245-288. DOI=http://dx.doi.org/10.1162/089120102760275983

[Goldberg and Hirst 2017] Yoav Goldberg and Graeme Hirst. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers. 2017

[Goodfellow et al. 2016] Ian Goodfellow and Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT Press 2016.

[Graves et al. 2009] Alex Graves, Marcus Liwicki, Santiago Fernandez, Roman Bertolami, Horst Bunke, and J¨urgen Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.

[Gruber 1965] J.S. Gruber. *Studies in Lexical Relations*. Ph.D. thesis, MIT, 1965.

[Genthial 2017] Guillaume Genthial, "Sequence Tagging with Tensorflow." Downloaded on May 2018. https://guillaumegenthial.github.io/sequence-tagging-with-tensorflow.html

[He et al. 2017] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer, "Deep Semantic Role Labeling: What Works and What's Next," *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[Hibshi et al. 2015] H. Hibshi, T. D. Breaux, and S. B. Broomell, "Assessment of risk perception in security requirements composition," *2015 IEEE 23rd Int. Requir. Eng. Conf. (RE)*, pp. 146-155, 2015.

[Hochreiter et al. 2001] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, S.C. Kremer and J.F. Kolen, 1, Wiley-IEEE Press, 2001, pp.237-243.

[Hochreiter and Schmidhuber 1997] Sepp Hochreiter; Jürgen Schmidhuber, “Long short-term memory,” *Neural Computation*. 9 (8): 1735–1780, 1997.

[Horrigan 2008] J. Horrigan, “Online shopping,” PEW Internet and American Life Project, Feb. 13, 2008.

[Hosseini et al. 2018] Bokaei Hosseini M., Breaux T.D., Niu J. (2018) Inferring Ontology Fragments from Semantic Role Typing of Lexical Variants. In: Kamsties E., Horkoff J., Dalpiaz F. (eds) Requirements Engineering: Foundation for Software Quality. REFSQ 2018. Lecture Notes in Computer Science, vol 10753. Springer.

[Hunter 2004] D. R. Hunter, “MM algorithms for generalized Bradley–Terry models,” *The Annals of Statistics*, 32(1): 384–406, 2004.

[Hustinx 2010] Peter Hustinx, “Privacy by design: delivering the promises,” *Identity in the Information Society,* Volume 3, Issue 2, pp 253–255, August 2010.

[Jurafsky and Martin 2000] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

[Kaisser and Webber 2007] Michael Kaisser and Bonnie Webber, “Question answering based on semantic roles,” In Proceedings of the Workshop on Deep Linguistic Processing (DeepLP '07), Association for Computational Linguistics, Stroudsburg, PA, USA, 41-48.

[Kamsties 2006] E. Kamsties, “Understanding ambiguity in requirements engineering,” *Engr. & Managing Soft. Req’ts,* pp.245-266, 2006.

[Kamsties et al. 2001] E. Kamsties, D. Berry, B. Paech, “Detecting ambiguities in requirements documents using inspections,” *1st Workshop on Inspection in Soft. Engr. (WISE'01),* pp. 68-80, 2001.

[Kaplan and Garrick 1981] S. Kaplan and B. J. Garrick, “On the quantitative definition of risk,” *Risk Analysis*, 1 (1): 11-27, 1981.

[Kelley et al. 2009] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder, “A "nutrition label" for privacy,” *5th Symposium on Usable Privacy and Security (SOUPS '09)*, ACM, New York, NY, USA, Article 4, 12 pages. DOI=http://dx.doi.org/10.1145/1572532.1572538

[Kiyavitskaya 2008] N. Kiyavitskaya, N. Zeni, L. Mich, D. M. Berry, “Requirements for tools for ambiguity identification and measurement in natural language requirements specifications,” *Req’ts Engr. J.,* 13(3): 207–240, 2008.

[Knight 1921] F.H. Knight. *Risk, Uncertainty, and Profit.* Houghton Mifflin Company, 1921.

[Kulas and Stachowski 2013] J. T. Kulas and A. A. Stachowski, “Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators,” *J. Res. Pers.,* vol. 47, no. 4, pp. 254-262, Aug. 2013.

[Lakoff 1972] G. Lakoff, “Linguistics and natural logic,” *The Semantics of Natural Language*, pp. 545– 665, 1972.

[Lamsweerde 2009] A. van Lamsweerde. *Requirements Engineering - From System Goals to UML Models to Software Specifications*. Wiley 2009.

[Levy and Hastak 2008] Alan Levy and Manoj Hastak, “Consumer Comprehension of Financial Privacy Notices: A Report on the Results of the Quantitative Testing. Interagency notice

research project," December 15. http://www.sec.gov/comments/s7-09 -07/s70907-21-levy.pdf.

[Limsopatham and Collier 2016] N. Limsopatham, N. H. Collier, "Bidirectional LSTM for Named Entity Recognition in Twitter Messages," Proceedings of the *2nd Workshop on Noisy User-generated Text*, 145-152. https://doi.org/10.17863/CAM.7201

[Ma and Hovy 2016] Xuezhe Ma and Eduard Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pages 1064-1074, Berlin, Germany. August 2016.

[Marneffe et al. 2006] M. C. de Marneffe, B. MacCartney, C. D. Manning. "Generating typed dependency parses from phrase structure parses," *Intl. Conf. Lang. Res. & Eval.,* pp. 449-454, 2006.

[Massey et al. 2014] A. Massey, R.L. Rutledge, A.I. Antón, P.P. Swire, "Identifying and classifying ambiguity for regulatory requirements," *22nd IEEE Int'l Req'ts Engr. Conf,* pp. 83-92, 2014.

[McDonald and Cranor 2008] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *I/S – A Journal of Law and Policy for the Information Society*, 4(3): 540-565, 2008.

[Mikolov et al. 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality," In Proceedings of *NIPS*, 2013.

[Mitsumori et al. 2005] Tomohiro Mitsumori, Masaki Murata, Yasushi Fukuda, Kouichi Doi, and Hirohumi Doi, "Semantic role labeling using support vector machines," In Proceedings of the *Ninth Conference on Computational Natural Language Learning (CONLL '05)*, *Association for Computational Linguistics*, Stroudsburg, PA, USA, 197-200.

[Miwa and Bansal 2016] Makoto Miwa and Mohit Bansal, "End-to-end Relation Extraction using LSTMs on Sequences and Tree Structures," Proceedings of *ACL 2016*, Berlin, Germany.

[Moor 1997] J. H. Moor, "Towards a theory of privacy in the information age*," Computers and Society,* vol. 27, no. 3, pp. 27–32, 1997.

[Murphy 1996] R. S. Murphy, "Property rights in personal information: An economic defense of privacy," *Georgetown Law Journal*, vol. 84, p. 2381, 1996.

[Nissenbaum 2004] H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, 79, 2004

[Nissenbaum 2009] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books, 2009.

[Nguyen and Grishman 2015] Thien Huu Nguyen and Ralph Grishman, "Relation Extraction: Perspective from Convolutional Neural Networks," in Proceedings of *NAACL Workshop on Vector Space Modeling for NLP*, Denver, Colorado, June, 2015.

[Pearson and Hartley 1962] E.S. Pearson, H. O. Hartley (eds). *Biometrika Tables for Statisticians*. v. I, 2. Aufl. Cambridge University Press, 1962.

[Pearson and Hartley 1966] E.S. Pearson, H. O. Hartley (eds). *Biometrika Tables for Statisticians*. v. I, 3. Auflage. Cambridge University Press, 1966.

[Popescu 2008] D. Popescu, S. Rugaber, N. Medvidovic, D. M. Berry, "Reducing ambiguities in requirements specifications via automatically created object-oriented models," *Lecture Notes Comp. Sci.,* 5320: 103-124, 2008.

[Pradhan et al. 2005] Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H. Martin, and Daniel Jurafsky, "Semantic role chunking combining complementary syntactic views," In Proceedings of the *9th Conference on Computational Natural Language Learning*, CONLL '05, pages 217–220, Stroudsburg, PA, USA. Association for Computational Linguistics.

[R 2013] R Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, 2013.

[R Core Team 2015] R Core Team, "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing,* Vienna, Austria. 2015. URL http://www.R-project.org/.

[Ramshaw and Marcus 1995] L. A. Ramshaw and M. P. Marcus, "1995. Text chunking using transformationbased learning," In Proceedings of the *Third Annual Workshop on Very Large Corpora,* pages 82–94. ACL.

[Reidenberg et al. 2016] Joel R. Reidenberg, Jaspreet Bhatia, Travis D. Breaux, Thomas B. Norton, "Ambiguity in Privacy Policies and the Impact of Regulation," *The Journal of Legal Studies,* 45(S2): S163-S190, June 2016.

[Roth and Lapata 2015] Michael Roth and Mirella Lapata, "Context-aware Frame-Semantic Role Labeling," Transactions of the Association for Computational Linguistics, v. 3, p. 449-460, August 2015

[OECD 2013] OECD, "The OECD Privacy Framework", 2013.

[Saldaña 2012] J. Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2012.

[Sathyendra 2017] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh, "Identifying the Provision of Choices in Privacy Policy Text", *Conference on Empirical Methods in Natural Language Processing (EMNLP),* Copenhagen, Denmark, Sep 2017

[Slovic 2000] P. Slovic. *The Perception of Risk.* Earthscan Publication, 2000.

[Solove 2008] Daniel J. Solove. *Understanding Privacy*. Harvard University Press, 2008.

[Starr 1969] C. Starr, "Social benefit versus technological risk," *Science*, 165, pp. 1232-1238, 1969.

[Stent and Bangalore 2014] A. Stent, S. Bangalore. Natural Language Generation in Interactive Systems. Cambridge: Cambridge University Press, 2014.

[Stoneburner 2002] Gary Stoneburner, Alice Y. Goguen, and Alexis Feringa, "Risk Management Guide for Information Technology Systems," *SP 800-30, Technical Report, NIST*, Gaithersburg, MD, United States, 2002.

[Surdeanu et al. 2003] Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth, "Using predicate-argument structures for information extraction," In Proceedings of 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL '03), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 8-15.

[Sutskever et al. 2014] I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks," In *NIPS'2014*.

[Tackstrom 2015] Oscar Tackstrom, Kuzman Ganchev, and Dipanjan Das, "Efficient inference and structured learning for semantic role labeling," *Transactions of the Association for Computational Linguistics* 3:29–41, 2015.

[Tjong 2008] S.F. Tjong. *Avoiding Ambiguities in Requirements Specifications*. PhD Thesis, Univ. of Nottingham, 2008.

[Tjong and Berry 2013] S.F. Tjong, D.M. Berry, "The design of SREE - a prototype potential ambiguity finder for requirements specifications and lessons learned," *REFSQ,* pp. 80-95, 2013.

[Turian et al. 2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio, "Word representations: a simple and general method for semi-supervised learning," In Proceedings of the *48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, Association for Computational Linguistics, Stroudsburg, PA, USA, 384-394.

[Turner and Firth 2012] H. Turner, D. Firth, "Bradley-Terry models in R: the BradleyTerry2 package," *J. Stat. Soft.,* 48(9): 1-21. 2012.

[Wang et al. 2014] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh, "A field trial of privacy nudges for Facebook," In Proceedings of the *SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, ACM, New York, NY, USA, 2367-2376. DOI=http://dx.doi.org/10.1145/2556288.2557413

[Wilson et al. 1997] W. M. Wilson, L. H. Rosenberg, L. E. Hyatt, "Automated analysis of requirement specifications," *19th ACM/IEEE Int'l Conf. Soft. Engr.,* pp. 161-171, 1997.

[Yang et al. 2010] H. Yang, A. Willis, A. de Roeck, B. Nuseibeh. "Automatic detection of nocuous coordination ambiguities in natural language requirements," *25th IEEE/ACM Int'l Conf. Auto. Soft. Engr.,* pp. 53-62, 2010.

[Yang et al. 2011] H. Yang, A. de Roeck, V. Gervasi, A. Willis, and B. Nuseibeh. "Analysing anaphoric ambiguity in natural language requirements," *Req'ts Engr. J.,* 16: 163-189, 2011.

[Yang et al. 2012] H. Yang, A. De Roeck, V. Gervasi, A. Willis and B. Nuseibeh, "Speculative requirements: Automatic detection of uncertainty in natural language requirements," *20th IEEE Int'l Req'ts Engr. Conf.,* pp. 11-20, 2012.

[Wakslak and Trope 2009] C. Wakslak and Y. Trope, "The effect of construal level on subjective probability estimates," *Psychol. Sci*., vol. 20, no. 1, pp. 52-58, Jan. 2009.

[Westin 1967] A. F. Westin. *Privacy and Freedom*. New York, NY: Atheneum, 1967.

[Wang 2015] Yinglin Wang, "Semantic information extraction for software requirements using semantic role labeling," 2015 IEEE International Conference on Progress in Informatics and Computing (PIC), Nanjing, 2015, pp. 332-337.

[Zhou and Xu 2015] Jie Zhou and Wei Xu, "End-to-end learning of semantic role labeling using recurrent neural networks," *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1127–1137 2015.