

Data Driven, Personalized Usable Privacy

THÈSE N° 7841 (2017)

PRÉSENTÉE LE 21 AOÛT 2017

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE DE SYSTÈMES D'INFORMATION RÉPARTIS

PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Hamza HARKOUS

acceptée sur proposition du jury:

Prof. R. Guerraoui, président du jury

Prof. K. Aberer, directeur de thèse

Prof. F. Schaub, rapporteur

Prof. J. Grossklags, rapporteur

Prof. J.-P. Hubaux, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2017



Acknowledgements

First, I would like to thank Karl Aberer, my Ph.D. supervisor, for giving me the opportunity in the first place. If there is something I owe to Karl, it is pushing me to master the “*so what?*” question to find meaning behind the research. Through our meetings, Karl would almost always come up with something that would inspire my research, even when that was not immediately apparent. He supplied me as a researcher with two essential conditions to thrive: freedom to pursue my ideas and resources to accomplish them.

I would like to also thank Prof. Rachid Guerraoui, my research supervisor during my master studies and the president of my jury. Rachid was one of the reasons I came to EPFL, and I learnt a lot from him about being sharp, accurate, and concise.

Next, I would like to thank the jury members for their efforts and for all their feedback and discussions: Prof. Jean Pierre Hubaux (for his stimulating questions), Prof. Jens Grossklags (for his in-depth scrutiny and for making the trip to Lausanne), and Prof. Florian Schaub (for his detailed feedback and for being an insightful collaborator too).

I am also grateful to the two postdocs I worked closely with at EPFL: Rameez Rahman and Rémi Lebret. Rameez was an inspiration in so many ways. He is a very-well read person with exceptional writing skills, which I tried to learn from during his presence. Rameez also has great insights into all kinds of techno-social systems, and he was the ideal person to throw ideas at and get feedback from. In all, he is a great friend with a bright mind. Rémi is also someone whom I was lucky to have as an officemate. In my pursuit of integrating Deep Learning in my research, I would have struggled a lot without his expert feedback about neural networks and how to train and tune them properly. I also learnt from his work-life balance, hoping to apply it in the coming years.

Next, I would like to thank the LSIR members, starting with Chantal, who helped me a lot throughout the Ph.D. and made my stay a smooth one. I also thank my friend Hao Zhuang for the thoughtful discussions in our joint projects and for enlightening me about the Chinese culture. I was also glad to have Amit as a collaborator in the last years of the Ph.D., and I continue to be amazed with his perseverance on all fronts. Many thanks to Berker, Alexandra, Jean-Eudes, Hung, Julia, Alevtina, Alex, Martin, Matteo, Mehdi, Michele, Jean-Paul, Julien, Thanasis, Rammohan, Panayiotis, Tri, Tam, Thang, Tian, and Jérémie for all the fruitful and funny moments we had and for all your feedback throughout the years.

During the Ph.D., I also had the chance to work with amazing people from the industry and academia, with whom I gained great experiences. This includes the team at Privately, especially Deepak and Francois and the Graspeo team, especially Andrii and Evgeny. I am grateful for my bright friend and collaborator, Kassem Fawaz, who has been always ready to hear from me on all matters of life since my Bachelor studies and until now. Kassem is a talented, down-to-earth researcher, whom I wish a wonderful career. I am also indebted for all the advice received from Professors Fadi Zaraket, Zaher Dawy, and Hassan Artail during my graduate applications.

These years would not have been as enjoyable had it not been for all my friends in Switzerland and in Lebanon, including Mohamad, Rajai, Abbass, Rida, Ali, Gharib, Hadi, Hussein, Mahdi, Serj, Mahmoud, Elie, Taha, Pedram, Fazel, Mojtaba, Mohsen, Mohammad, and several others. You all know who you are to me, and I am very grateful for all the cool times we had at the university and beyond.

Next, I can never thank my wife, Walaa, enough. For all the unconditional love you provided, for all the nice times we had, for all the weekends you spent home because I was working, for all the unproportional effort you spent even when we split tasks, for the uncertainty you bore throughout, and for all those sacrifices that I am reluctant to even mention here, I am eternally grateful. Without you as the chief strategist of our life, the last years would have been exponentially more difficult and the coming years would be much tougher to think about.

Finally, I am incredibly thankful to my parents, Noha and Hussein, for bringing me to this life and for raising me the way I am. They always strived to provide me with the best, especially when it comes to education. I am grateful for their patience as I have been regularly away from them since finishing school. My mother is a symbol for sacrifice as she raised me and my brother and sisters, giving up her own interests for ours. My father is a symbol for all kinds of generosity; we never feel insecure when he is beside us. My gratitude goes to my brother and sisters for all the love they provided me with and for all the nice times we continue to enjoy together. My thanks to all those who kept me in their prayers and thoughts, especially my grandparents, Walaa's family, and my extended family.

Lausanne, 12 July 2017

H. H.

Abstract

We live in the “inverse-privacy” world, where service providers derive insights from users’ data that the users do not even know about. This has been fueled by the advancements in machine learning technologies, which allowed providers to go beyond the superficial analysis of users’ transactions to the deep inspection of users’ content. Users themselves have been facing several problems in coping with this widening information discrepancy. Although the interfaces of apps and websites are generally equipped with privacy indicators (e.g., permissions, policies, ...), this has not been enough to create the counter-effect. We particularly identify three of the gaps that hindered the effectiveness and usability of privacy indicators:

- **Scale Adaptation:** The scale at which service providers are collecting data has been growing on multiple fronts. Storage technologies are increasingly capable and less costly. The profitable data economy has contributed to the birth of new data collectors. Users, on the other hand, have limited time, effort, and technological resources to cope with this scale.
- **Risk Communication:** Although providers utilize privacy indicators to announce *what* and (less often) *why* they need particular pieces of information, they rarely relay what can be potentially *inferred* from this data. Users have become habituated to repetitive dialogs that do not communicate the potential risks. Without this knowledge, users are less equipped to make informed decisions when they sign in to a site or install an application.
- **Language Complexity:** The information practices of service providers are buried in complex, long privacy policies, which are aimed to cover the company from a legal perspective. Generally, users do not have the time and sometimes the skills to decipher such policies, even when they are interested in knowing particular pieces of it.

In this thesis, we approach usable privacy from a data perspective. Instead of static privacy interfaces that are obscure, recurring, or unreadable, we develop techniques that bridge the understanding gap between users and service providers. Towards that, we make the following contributions:

- **Crowdsourced, data-driven privacy decision-making:** In an effort to combat the growing scale of data exposure, we consider the context of files uploaded to cloud services.

We propose C3P, a framework for automatically assessing the sensitivity of files, thus enabling realtime, fine-grained policy enforcement. C3P works on top of unstructured data and allows privacy preserving crowdsourcing of users' sharing decisions.

- **Data-driven app privacy indicators:** We introduce PrivySeal, which involves a new paradigm of dynamic, personalized app privacy indicators that bridge the risk understanding gap between users and providers. Through a variety of data analysis and visualization techniques, PrivySeal communicates risks by showing users the far-reaching insights that can be inferred from their data. Through PrivySeal's online platform, we also study the emerging problem of interdependent privacy in the context of cloud apps and provide a usable privacy indicator to mitigate it.
- **Automated question answering about privacy practices:** We introduce PriBot, the first automated question-answering system for privacy policies, which allows users to pose their questions about the privacy practices of any company with their own language. PriBot is based on a novel deep learning architecture of classifiers that we developed. Through a user study, we show its effectiveness at achieving high accuracy and relevance for users, thus narrowing the complexity gap in navigating privacy policies.

A core aim of this thesis is paving the road for a future where privacy indicators are not bound by a specific medium or pre-scripted wording. We design and develop techniques that enable privacy to be communicated effectively in an interface that is approachable to the user. For that, we go beyond textual interfaces to enable dynamic, visual, and personalized privacy interfaces that are fit for the variety of emerging technologies.

Key words: privacy, machine learning, human-computer interaction, anonymity, privacy indicators, interdependent privacy, deep learning, chatbots, privacy policies, internet of things, decision-making

Résumé

Nous vivons dans le monde de la “confidentialité inversée”, où les fournisseurs de services acquièrent des connaissances sur leurs utilisateurs à partir de leurs données, sans même que ces derniers n’en soient conscients. Ce nouveau monde est possible grâce aux progrès réalisés dans les techniques d’apprentissage automatique, qui permettent aux fournisseurs d’aller au-delà de l’analyse superficielle des actions de leurs utilisateurs pour aboutir à une inspection approfondie du contenu des utilisateurs. Bien que les interfaces des applications et des sites Web soient généralement équipées d’indicateurs de confidentialité (p. Ex., Autorisations, politiques, ...), cela n’est généralement pas suffisant pour contrer ces problèmes.

Nous identifions en particulier trois lacunes qui entravent l’efficacité et l’utilité des indicateurs de confidentialité :

- **Adaptation à grande échelle :** L’échelle à laquelle les fournisseurs de services collectent des données a augmenté de plusieurs façons. Les technologies de stockage ont accru en capacité et sont devenues moins coûteuses. L’économie des données, devenue très rentable, a contribué à la naissance de nouveaux collecteurs de données. Les utilisateurs ont, en revanche, un temps, des efforts et des ressources technologiques limités pour faire face à une telle échelle.
- **Communication de risque :** Bien que les fournisseurs utilisent des indicateurs de confidentialité pour annoncer le *quoi* et (moins souvent) le *pourquoi* ils ont besoin d’informations particulières, ils relèvent rarement ce qui peut potentiellement être *déduit* de ces données. Les utilisateurs sont devenus habitués à des dialogues répétitifs qui ne communiquent pas les risques potentiels. Sans cette connaissance, les utilisateurs sont moins équipés pour prendre des décisions éclairées lorsqu’ils se connectent à un site ou installent une application.
- **Complexité du langage :** Les pratiques en matière d’informations des fournisseurs de services sont camouflées dans des politiques de confidentialité complexes et longues qui visent à couvrir l’entreprise d’un point de vue juridique. Généralement, les utilisateurs n’ont ni le temps et parfois ni les compétences nécessaires pour déchiffrer de telles politiques, même s’ils s’intéressent à des parties en particulier.

Dans cette thèse, nous abordons la confidentialité utilisable du point de vue des données. A la place d’interfaces de confidentialité qui sont obscures, récurrentes ou illisibles, nous développons des techniques qui permettent de combler l’écart de compréhension entre les

utilisateurs et les fournisseurs de services. Pour ce faire, nous apportons les contributions suivantes :

- **Prise de décision en matière de confidentialité dépendante des données** : Dans le but de lutter contre l’augmentation de l’exposition aux données, nous considérons le contexte des fichiers téléchargés sur les services en nuage. Nous proposons C3P, un cadre pour évaluer automatiquement la sensibilité des fichiers, permettant ainsi une mise en application très détaillée et en temps réel de la politique de confidentialité. C3P fonctionne sur des données non structurées et permet de collecter les décisions de partage des utilisateurs tout en préservant la confidentialité.
- **Indicateurs de confidentialité de l’application axés sur les données** : Nous introduisons PrivySeal, qui propose un nouveau paradigme d’indicateurs dynamiques et personnalisés de la confidentialité des applications, permettant de combler le fossé entre les utilisateurs et les fournisseurs. Grâce à une variété d’analyses de données et de techniques de visualisation, PrivySeal communique les risques en montrant aux utilisateurs les connaissances approfondies qui peuvent être déduites de leurs données. Grâce à la plate-forme en ligne de PrivySeal, nous étudions également le problème émergent de la confidentialité interdépendante dans le contexte des applications en nuage et fournissons un indicateur utilisable de confidentialité pour atténuer ce problème.
- **Système de questions-réponses automatiques sur les pratiques de confidentialité** : Nous présentons PriBot, le premier système automatisé de réponses aux questions sur les politiques de confidentialité, qui permet aux utilisateurs de poser leurs questions avec leurs propres mots sur n’importe quelle entreprise. PriBot utilise une nouvelle architecture de classifieurs basés sur l’apprentissage profond. Grâce à une étude menée sur des utilisateurs, nous montrons son efficacité à atteindre une grande précision et une pertinence dans les réponses, réduisant ainsi l’écart de complexité dans la lecture des politiques de confidentialité.

L’objectif principal de cette thèse est d’ouvrir la voie à un avenir où les indicateurs de confidentialité ne sont pas limités à un support spécifique ou un message préétabli. Nous concevons et développons des techniques permettant à la confidentialité d’être communiquée efficacement avec une interface accessible à l’utilisateur. Pour cela, nous allons au-delà des interfaces textes pour proposer des interfaces de confidentialité dynamiques, visuelles et mains libres qui conviennent à la variété des technologies émergentes.

Mots clefs : vie privée, apprentissage automatique, interactions homme-machine, anonymat, indicateurs de confidentialité, interdépendance dans la protection des données, apprentissage profond, agents conversationnels, règles de confidentialité, internet des objets, prise de décision

Contents

Acknowledgements	i
Abstract (English/Français)	iii
1 Introduction	1
1.1 A Tale of Two Viewpoints	1
1.2 An Intellectual Luxury Good?	2
1.3 Orwell vs. Kafka	3
1.4 Problems	3
1.5 Contributions	7
I Adapting to Scale	13
2 Context-aware, Crowdsourced Cloud Privacy	15
2.1 Overview	15
2.2 System Model	17
2.3 Context Vocabulary and Sharing Policies	19
2.4 Crowd-Sourcing and Risk Evaluation	21
2.5 Evaluation and Experiments	28
2.6 Implementation	36
2.7 Related Work	43
2.8 Summary	45
II Communicating the Risk	47
3 A Primer on Cloud Apps Privacy	49
3.1 Overview	49
3.2 Privacy Issues in Third Party Cloud Apps	50
3.3 Third-party Cloud Apps Ecosystem	51
3.4 Summary	56
4 PrivySeal: Breaking the Knowledge Imbalance	57
4.1 Overview	57

4.2	Privacy Risk of 3rd Party Google Drive Apps	58
4.3	New Permission Models	63
4.4	Evaluating the Models	70
4.5	PrivySeal: A Privacy-Focused App Store	79
4.6	Anatomizing Developers' Behavior	80
4.7	Recommended Best Practices	84
4.8	Related Work	85
4.9	Summary	87
5	A Usability Approach to Interdependent Privacy	89
5.1	Overview	89
5.2	Models and Preliminaries	91
5.3	Collaborators' Impact	93
5.4	User Study	96
5.5	Large Networks' Simulations	108
5.6	Related Work	115
5.7	Summary	116
III	Handling Language Complexity	119
6	PriBot: Automated QA for Privacy Policies	121
6.1	Overview	121
6.2	System and Data Overview	123
6.3	Policy Pre-processing	126
6.4	Question-Answering Approaches	127
6.5	Evaluation Methodology & Dataset	137
6.6	Accuracy Evaluation	138
6.7	User Study	141
6.8	Friendly Summary Generation	144
6.9	PriBot Implementation	145
6.10	Discussion	147
6.11	Related Work	149
6.12	Summary	151
7	Conclusion	153
A	Study Material for Chapter 2	157
B	Study Material for Chapter 5	181
B.1	Introductory Material	181
B.2	Material for Modules	186
B.3	Final Survey	193

C Study Material for Chapter 6	195
C.1 Introductory Material	195
C.2 Answer Evaluation	199
C.3 Final Survey	205
D Example Cases for PriBot	207
Curriculum Vitae	229

1 Introduction

1.1 A Tale of Two Viewpoints

Historically, there has never been a consensus on whether privacy is an *innate* need for humans. On the one hand, prominent figures, including Vint Cerf (one of the internet pioneers), claim that privacy may be an anomaly of the 20th-century [Sha15, Fer13, Fer15]. Proponents of this opinion argue that, for thousands of years, people have been prioritizing money, prestige, or convenience over solitude or privacy. For instance, until 1500 A.D., most homes in the western world did not have internal walls separating rooms. The desire for warmth led to the development of the brick chimney, along with the needed support beams. Only then did walls start to spread inside homes [Fer15]. A single bed for the whole family and its guests was also the norm in Europe until families could afford to buy multiple beds. This transition was mainly driven by hygiene reasons and by the spread of contagious diseases [Fer15].

This standpoint has its opponents. Despite the socioeconomic factors that have contributed to the evolution of privacy, prominent scholars, like Irwin Altman, perceive it as a universal human characteristic [ABL15, Alt77]. It is also regarded as a necessary ingredient in guarding human dignity [Blo64], enabling relationships' intimacy [Ger78], and protecting personal liberty [All11]. Moreover, ancient religious texts contained several references to privacy as a trait to aspire to. In the Old Testament (Numbers 24:5), the biblical Israelites are praised for not positioning their tents' openings facing each other: "*How fair are your tents, O Jacob ...*". The Quran (49:12) has also tackled the right for privacy: "*...do not spy or backbite each other. Would one of you like to eat the flesh of his brother when dead? You would detest it.*"

The activity of *reading* itself has seen an interesting historical evolution that illustrates how privacy preserving options emerge. According to Nicholas Carr, in the world that predated the commoditization of books, reading has mostly been a public, vocal activity [Car11]. Scribes wrote books based on hearing, and they neither separated the word with spaces nor paid attention to word order. Silent reading, as we know it today, did not become the standard until well after the collapse of the Roman Empire. However, once people witnessed the proliferation of books, their ability to read privately became a key factor for personal instruction and im-

provement. Interestingly, this could have been much more difficult without the modifications in how text is inscribed. The inter-word spaces, the punctuation marks, and the word order were some of these new features that enabled private reading [Car11].

Despite the differing viewpoints, this historical prologue has a few takeaways. First, privacy-enhancing technologies of the ancient times, like the internal walls or —loosely speaking—the inter-word spaces, were not solely motivated by privacy per se. Second, the level of privacy has always been a result of compromises that people face in their daily life, whether related to their purchasing power or their desire for recognition. Third, it has rarely been the case that people voluntarily chose the privacy-invading tools or technologies as the norm for their life despite having a choice of something else. When privacy-equipped options became affordable (e.g., private baths or individual rooms), people generally adopted such options. Even in our times, when an app like Whatsapp introduced end-to-end encryption to a billion users, we did not see people fleeing away from the app. It was an additional feature that strengthened the bond with the app for many.

1.2 An Intellectual Luxury Good?

Despite people's tendency to go with the privacy preserving option, this is not always feasible. In fact, privacy has frequently been described as a *luxury good* [McG59, Sch68]. The essayist Phyllis McGinley puts it nicely:

“The poor might have to huddle together in cities for need's sake, and the frontiersman cling to his neighbor for the sake of protection. But in each civilization, as it advanced, those who could afford it chose the luxury of a withdrawing place. Egyptians planned vine-hung gardens, the Greeks had their porticos and seaside villas, the Romans put enclosures around their patios. . . . Privacy was considered as worth striving for as hallmarked silver or linen sheets for one's bed” [McG59]

With time, emerging privacy solutions continued to suffer from the same issue: *limited accessibility*, whether this is financial, intellectual, political, or other types of accessibility. Even today, financial affordability cannot be neglected whenever we analyze the adoption of privacy- or security-enhancing technologies. Access to home internet is still subject to restrictive quota limits in a lot of developing countries. Hence, a security measure, which is as simple as updating a computer with the latest packages, is seen as an economic burden. The same goes for upgrading a smartphone to a newer model to be eligible for continued security updates.

In the recent years, several important factors have exacerbated the problem of privacy accessibility, going beyond the financial aspect. From the scale at which data is being collected to the advanced machine learning models extracting insights from this data, privacy has further emerged as an *intellectual* luxury good. Its affordability is limited to the few who know the

far-reaching implications of their activities and the actual shareholders of their data.

1.3 Orwell vs. Kafka

Daniel Solove differentiates in his book “Nothing to Hide: The False Tradeoff Between Privacy and Security” between two ways in which privacy is perceived: the *Orwellian* view and the *Kafkaesque* view [Sol11]. In the Orwellian view (based on George Orwell’s *Nineteen Eighty-Four* [Orw09]), privacy is the protection against surveillance and its associated harms, such as inhibition and social control. The government (a.k.a. the Big Brother) is the main presumed adversary, and people are less concerned about protecting information which they do not mind being known, such as their demographics or the places they visit.

The Kafkaesque view, named after Franz Kafka, embodies the other aspects of privacy that frequently go unnoticed. This view is inspired by *The Trial*, a novel by Kafka about a man who is investigated and prosecuted—for reasons he does not know—by a mysterious, inaccessible authority. According to Solove, this perspective on privacy highlights the problems due to information *processing* (storage, use, or analysis) rather than information *collection*.

Under the Kafkaesque umbrella falls data *aggregation*, i.e., the ability to combine seemingly benign data from multiple sources and to repurpose it in new contexts. For example, the industry of data brokering has the sole purpose of linking as much data as possible through a large set of heuristics, before selling it to interested parties. Second, user *exclusion* occurs when people are not allowed to know how their information is handled, nor given a choice to correct it. Third, the *secondary* use of data is another aspect where information is processed for reasons that it has not been intended for. Fourth, personal data can suffer from *distortion*, where the data depicts an incomplete and often wrong picture about individuals by reducing them to a limited, subjective set of features [Sol11].

These practices get another name when they are viciously used to advance the benefits of the big corporations and induce unfairness in the society. The writer Cathy O’Neil calls them the **“Weapons of Math Destruction”** [O’N16]. From the models used for screening job applicants to those used to create e-scores for people—thus deciding their loans’ interest rates and insurance premiums—the repercussions are directly impacting lives in one way or another. In an algorithmic world where the past of individuals and groups determines their future, users’ privacy, through minimal exposure to data hunters, becomes more and more essential.

1.4 Problems

On a high level, in this thesis, we ask the following question:

In the age of large-scale data analysis, how can we empower users to make better privacy decisions without loading them with a huge cognitive burden?

We consider privacy in its broad sense, which includes the Orwellian interpretation in addition to the (often marginalized) Kafkaesque one. In fact, the majority of users understand the presence of government as a surveillance power, which makes them aware of the Orwellian aspect. Nevertheless, the very reasons that are molding privacy into an intellectual luxury good make the Kafkaesque aspects of privacy more crucial than ever. Highlighting such aspects is a core part of this thesis, and addressing them contributes towards mitigating the problem of privacy inaccessibility itself.

In our work, we have identified and worked towards bridging the following major gaps that users encounter in their digital environments: (i) **the scale adaptation gap**, (ii) **the risk understanding gap**, and (iii) **the language complexity gap**.

The Scale Adaptation Gap

The scale at which people's data is being collected has seen orders of magnitude increase since the beginning of the digital revolution. One reason for that is the **advancement in storage technologies**, which can now comfortably hold the records of billions of users. Together with the increasing efficiency of database technologies, hosting data and querying it are no more the bottlenecks for data collectors.

The **data collectors themselves have proliferated**. Instead of solely worrying about governmental surveillance or being watched by their neighbors, people have to deal with an increasing number of potential *adversaries*. For someone who buys a smartphone, these adversaries include the entity that sold them the device, the device manufacturer, the third party service providers that their device supports, the advertisement providers incorporated with the apps, etc.

The sources from which data can be analyzed have expanded too. Instead of solely depending on *structured* data records, providers have bolstered their analytics arsenal to target *unstructured data*, which is present in documents, photos, videos, etc. They benefited from the huge advancements in natural language processing and visual computing, which have taken a recent boost with the deep learning era. Images are now an important source for recognizing objects [SVI⁺16], activities [YNS⁺15], landmarks [CLLH16], faces [TYRW14], emotions [CBDC14], etc. Documents are also a huge trove for deciphering opinions [Cam16], interests [CLL⁺15], connections [SKV15], etc.

The increased capabilities of service providers fueled the asymmetry of their relationship with users. On one end are the automated services with vast computing powers. On the other end are the users with limited time and cognitive capabilities, a phenomenon that has been termed as “bounded rationality” [Sim72]. The vast majority of users have not been able to cope with the scale at which the data collection is happening. Although the human brain itself has been shown to manifest *plasticity* properties [Car11], where it neurologically adapts to new media, the scale and diversity of data collection has so far rendered that phenomenon

ineffective.

The Risk Understanding Gap

In addition to the vast scale data collection, **the types of insights that could be inferred from the data** have also become beyond the grasp of most users. With the emergence of machine learning as the mainstream method of user understanding and profiling, the accessibility of privacy has suffered a new major setback. Bathing in public, reading aloud, or publishing an article under one's name are activities for which the repercussions are relatively well-understood by the users. This is not the case with granting access to an over-privileged application or consenting to share personal information with third parties. Such activities might seem benign to the majority of users, who might not notice their potential negative impacts until they are personally affected.

The opacity of data analysis has escalated to the extent that it warranted itself a new term: **“Inverse Privacy”**. Depicted by Gurevich et al., a piece of data is inversely private *if some party has access to it, but you do not* [GHW16]. This typically results from the legitimate analysis of data collected by banks, healthcare providers, governmental agencies, shopping malls, employers, etc. Derived from this data are insights such as the health status or the credit score of an individual. Such information has been so far siloed inside the institutions generating it, due to business and privacy reasons. Gurevich et al., contrast this concept against the more traditional **“Partial Privacy”**, where the user has access to the data, but a limited number of other parties does too.

Faced with this new reality, Hubaux and Juels proposed that researchers should prepare a post-confidentiality agenda. They suggested the development of a new category of privacy-enhancing technologies, namely **“Fair-use PETs” (F-PETS)**, which allow users to verify the fair use of their information by service providers [HJ16]. One example they give is introducing protocols into algorithmic decision making, which allow proving that the decisions taken do not violate social norms expressed as laws, policies, or regulations. This is in comparison to the **“Confidentiality-oriented PETs” (C-PETS)** that solely focus on confidentiality, such as encrypted email, ad blockers, or location obfuscation techniques.

Nevertheless, the need for the verifiability of the decisions made by algorithms does not preclude the need for better communicating the possibilities of such decisions. The mere existence of such practices is obscure for the majority of users, and this obscurity is for a reason: relaying the real possibilities with today's data could result in discouraging users from using certain applications.

To put this in perspective, consider the analogy between the need of data for service providers and the need of cadavers (corpses for dissection) for medical researchers. Both data and corpses are indispensable for the respective parties. In the early 19th century, due to the shortage of cadavers, some resorted to the business of secretly murdering people (e.g., by

suffocation) before selling the bodies for dissection purposes. Grave robbing was the less bad and more common practice for achieving the same goal, to the extent that special techniques were used to deter grave robbers (such as guards, graveyard watchtowers, or huge stone slabs to cover the graves). It was only later that medical schools started rejecting grave robbing as a means for anatomists to get corpses, and body donors had to provide their consent *a priori*. Even today, whether that consent is an informed one is not that clear. In his book, “When Breath Becomes Air”, the neurosurgeon, Paul Kalanithi, puts this as follows [Kal16]:

Yet the best-informed people—doctors—almost never donated their bodies. How informed were the donors, then? As one anatomy professor put it to me, “You wouldn’t tell a patient the gory details of a surgery if that would make them not consent.” Even if donors were informed enough—and they might well have been, notwithstanding one anatomy professor’s hedging—it wasn’t so much the thought of being dissected that galled. It was the thought of your mother, your father, your grandparents being hacked to pieces by wisecracking twenty-two-year-old medical students.

Replacing medical students by the analysts crunching people’s data, one can get to the status of risk communication nowadays. There are two differences though. First, data looting—a practice that is well and alive in the unregulated part of the industry—is not regarded as nefariously as grave robbing. Second, those analysts are mainly optimizing the benefits of their companies, compared to the researchers advancing the medical field and saving lives.

The Language Complexity Gap

Even when the service provider aims to communicate all the potential privacy risks to the users, the means might fall short. Language is at the core of this problem. Consider privacy policies, the *de facto* standard for notice and choice online. They are intended to inform users how companies collect, store, and manage their personal information. **Such policies are typically excessively long and hard to follow** [Cat10, Fed12, GSF⁺16, MC08, Pre14]. In 2008, McDonald *et al.* estimated it would take an average user 201 hours to read all the privacy policies encountered in a year [MC08]. Since then, we have witnessed the smartphone revolution and the rise of the Internet of Things (IoT). Thus, in 2017, users would likely have to spend substantially more time on reading the privacy policy of each website, app, device, or service they interact with.

The issue is not only limited to the length and the presentation. **Policies are also rife with legal jargon** that is meant to protect the provider on the legal front [MEAS13, Mei13]. Such language is usually different from that of ordinary users, who are supposedly the intended audience of such policies. Attempting to be informative to users and compliant with the law at the same time, these policies have so far significantly leaned towards the latter role.

To handle this, significant works have been done towards standardization (via labels [CLM⁺02,

KBCR09], icons [CGA06, HZH11], etc.). Their intuition is that standardized interfaces with less text result in reducing the language complexity. However, these attempts have also seen limited spread/usage. The main reason behind that was the rare adoption from the service providers, especially with the lack of incentives and the absence of regulations. Another reason is the difficulty of shaping a standard interface that appeals to the vast majority of users coming from different countries and educational backgrounds.

This leads us to observe that **current mechanisms for relaying privacy practices are static in nature**. Be it app permissions or privacy policies; such interfaces cannot easily capture the growing complexity of information processing while remaining swiftly comprehensible and skimmable by the average user. The response to having more information to say with more text to show is clearly not the path forward. Moreover, due to this static nature, repeatedly subjecting the user to similar content with each privacy decision leads to undesirable habituation effects [SBDC15]; i.e., the effectiveness of the static notices decays with time.

This becomes even more challenging with the miniaturization trend of electronic devices that started with mobile phones and reached its peak with the Internet of Things (IoTs) [Fed15]. The less the screen estate is, the more difficult it is to communicate via written text. Coupling the long, complex text with a voice interface—which is a natural alternative—is a match “made in hell”; together they result in multiplying the required cognitive load on users’ behalf.

1.5 Contributions

This thesis has culminated in the development of three main systems, each of them contributing to bridging one or more of the gaps above:

C3P: Context-Aware Crowdsourced Cloud Privacy

C3P is our answer to the growing scale of privacy decisions that users should make [HRA14]. It is motivated by Nissenbaum’s approach to privacy through *contextual integrity* [Nis04]: the sensitivity of a piece of data is dependent on the context in which the data is shared.

We developed C3P as **the first automated sensitivity assessment framework for unstructured data**—namely users’ files (documents, photos, etc.). C3P leverages the wisdom of the crowd—in a privacy-preserving way—to compute a file’s sensitivity without accessing the file itself. This is achieved by (1) modeling each unstructured file through a bag of features, extracted from its content, metadata, and sharing environment, (2) querying the service about sensitivity in an anonymized way, and (3) contributing privacy decisions to the service provider through an anonymized protocol. To determine the sensitivity of files—while accounting for variable user attitudes towards privacy—we use *Item Response Theory*, a psychometric technique for modeling latent traits of people and items. We show the efficacy of C3P in the context of privacy-preserving file sharing within cloud storage services.

On a high level, C3P's goal is automating privacy decisions based on the user's context to reduce the number of interventions the user has to take. This framework is a general one, which naturally extends to other domains. For example, it can be easily transplanted to dynamically adjust the level of access that smartphone apps have to users' files.

C3P, which will be the topic of Chapter 2 has appeared in this paper [HRA14]:

C3P: Context-aware CrowdSourced Cloud Privacy.

Hamza Harkous, Rameez Rahman, and Karl Aberer.

In: Privacy Enhancing Technologies. PETS 2014.

Lecture Notes in Computer Science, vol 8555. Springer, Cham.

PrivySeal: A Personalized Privacy Assistant for Cloud Apps

The next part of the thesis is mainly focused on bridging the risk communication gap between the users and the service providers. To further motivate this part, we note that there is a clearly uneven power balance in the existing digital ecosystems, whether in the mobile case (e.g., Android and iOS), in the cloud case (e.g., Google Drive and Dropbox), or in other similar platforms.

The Economist Magazine reported in 2014 on the current state of things [Eco14]:

“Today the IT sector looks like a very flat inverted pyramid: the bottom, where economies of scale rule, is made up of just a few powerful platforms; the top, where creativity and agility are at a premium, is becoming ever more fragmented. There is not much in between.

As software eats more and more industries, they will increasingly take on this shape, predicts Philip Evans of Boston Consulting Group. By lowering transaction costs, IT allows big chunks of the economy to reshape themselves and turn into what he calls “stacks”—industry-wide ecosystems that will have large platforms at one end of their value chains and a wide variety of modes of production at the other, from startups to social enterprises and communities to user-generated content.”

Accordingly, from a business and growth point of view, the platforms itself are interested in scaling their ecosystems with more apps so that they attract a wider user base and reap the benefits of the network effect. The current status of privacy notices is directly impacted by the fact that platforms tend to favor such a steady growth over slow, but privacy-focused steps. This is also one reason why a lot of platforms tolerate the presence of over-privileged apps, i.e., those which access more data than is needed for them to function.

PrivySeal, our next contribution, looks at the possibilities of breaking this imbalance. As

a case study, we take *3rd Party Cloud* apps (shortly *3PC* apps), an ecosystem which we were the first to anatomize. By analyzing the top Google Drive apps on Chrome Store, we discovered that around two-thirds of them are over-privileged and that 79% require full access to users' data [HRKA16].

We divide our contributions into two parts:

- **Exposing the far-reaching implications of data sharing:** Our primary goal is to assess the efficacy of the current models in deterring users from installing over-privileged apps and to test alternative models that could improve that efficacy. We analyze three different permission models. In experiments with 210 real users, we discover that **the most successful permission model is our novel ensemble method that we call Far-reaching Insights**. Far-reaching Insights inform the users about the data-driven insights that apps can make about them (e.g., their topics of interest, collaboration and activity patterns, etc.). Thus, they seek to bridge the gap between what third parties can potentially know about users and users' perception of their privacy leakage. The efficacy of Far-reaching Insights in bridging this gap is demonstrated by our results, as Far-reaching Insights prove to be, on average, twice as effective as the current model in discouraging users from installing over-privileged apps.

In an effort to promote general privacy awareness, we deployed PrivySeal, a publicly available, privacy-focused app store that uses Far-reaching Insights. Based on the knowledge extracted from data of the store's users (over 115 gigabytes of Google Drive data from 1440 users with 662 installed apps), we also delineate the ecosystem for 3PC apps from the standpoint of developers and cloud providers. Finally, we present several general recommendations that can guide other future works in the area of privacy for the cloud.

This study, which we will detail in Chapter 4 has appeared in [HRKA16]:

*The Curious Case of the PDF Converter that Likes Mozart:
Dissecting and Mitigating the Privacy Risk of Personal Cloud Apps.*

Hamza Harkous, Rameez Rahman, Bojan Karlas, and Karl Aberer. In
Proceedings on Privacy Enhancing Technologies (PoPETs), 2016.

- **Exposing the impact of collaborators on the user's privacy:** Another risk that is rarely evident and poorly communicated to 3PC apps' users is that their privacy is not solely determined by their own decisions. Whenever a user grants access to a new vendor, she is inflicting a privacy loss on herself and on her collaborators too. We study this issue, benefiting from PrivySeal's platform. By analyzing a real dataset of 183 Google Drive users and 131 third party apps, we discover that collaborators inflict a privacy loss which is at least 39% higher than what users themselves cause. We take a step toward minimizing this loss by introducing the concept of History-based decisions. Simply put, users are informed at decision time about the vendors which have been previously

granted access to their data. Thus, they can reduce their privacy loss by not installing apps from new vendors whenever possible.

Next, we realize this concept by introducing a new privacy indicator, which can be integrated within the cloud apps' authorization interface. Via a web experiment with 141 participants recruited from CrowdFlower, we show that our privacy indicator can significantly increase the user's likelihood of choosing the app that minimizes her privacy loss.

Finally, we explore the network effect of History-based decisions via a simulation on top of large collaboration networks. We demonstrate that adopting such a decision-making process is capable of reducing the growth of users' privacy loss by 70% in a Google Drive-based network and by 40% in an author collaboration network. This is despite the fact that we neither assume that users cooperate nor that they exhibit altruistic behavior. To our knowledge, our work is the first to provide quantifiable evidence of the privacy risk that collaborators pose in cloud apps. We are also the first to mitigate this problem via a usable privacy approach.

This work —portrayed in Chapter 5 —has been published in this paper [HA17]:

“If You Can't Beat Them, Join Them”:

A Usability Approach to Interdependent Privacy in Cloud Apps.

Hamza Harkous and Karl Aberer. In Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, CODASPY '17, pages 127–138, New York, NY, USA, 2017. ACM.

To pave the way for these two chapters, we give an overview of the ecosystem of third party cloud apps in Chapter 3.

PriBot: A Question Answering Chatbot for Privacy Policies

The last core component of this thesis is focused on the third major gap we identified: *how to deal with the prevalent language complexity problem in privacy notices?* We turn our focus on the flagship case of this issue: privacy policies.

We address this problem by proposing PriBot, the first question-answering (QA) system for privacy policies. In a fully automated approach, PriBot takes a previously unseen privacy policy and uses it to answer, in real time with high accuracy and relevance, user questions that are posed in free form.

We make multiple contributions to overcome challenges related to the discrepancy of language between user's questions and policies as well as the lack of privacy-related QA datasets. In particular, we propose two algorithms based on deep learning for extracting answers from a privacy policy for a given question.

Our user study, with 1,186 participants, shows that PriBot’s top three responses are relevant answers for 91% of 120 real-world privacy questions posted on Twitter. Our best algorithm further outperforms traditional methods by 15% regarding the accuracy of generated results. We present a practical implementation of PriBot, which is ready for public use and discuss real-world applications of the proposed approach.

At the time of writing of this thesis, PriBot **is under submission** [HFL⁺17]:

PriBot: Answering Free-form Questions about Privacy Policies with Deep Learning

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub,
Kang G. Shin, and Karl Aberer. Technical report, 2017

We have outlined the vision for PriBot in [HFSA16]:

Pribots: Conversational Privacy with Chatbots.

Hamza Harkous, Kassem Fawaz, Kang G. Shin, and Karl Aberer.
In Workshop on the Future of Privacy Notices and Indicators, SOUPS 2016,
Denver, CO, USA, June 22, 2016. USENIX Association, 2016.

PriBot **will be the subject of Chapter 6**. We **conclude this thesis in Chapter 7**, by giving a summary of our findings and an overview of potential future directions.

Adapting to Scale Part I

2 Context-aware, Crowdsourced Cloud Privacy

2.1 Overview

In this chapter, we embark on the task of bridging the scale adaptation gap: i.e., *how to help users manage their privacy with the expanding amount of data that is being shared and the limited time available?*. Towards that, we consider the case of cloud storage providers, provided that they have acquired a ubiquitous presence in our digital lives. Given the pervasiveness of useful cloud services such as storage, online document editing, media streaming, etc., data which would normally be on the user's local machine, now invariably lies in the cloud. As we have seen in the previous years, these platforms are not always a safe haven, especially with the large-scale exposure of data to governmental agencies by the big players in this domain [GM13].

2.1.1 Motivation and Challenges

In order to inform potential solutions, we identify three major stumbling blocks towards privacy provision in the cloud:

a) Privacy vs. Services Dilemma: To tackle privacy concerns, some cloud computing companies provide the users with the option of client-side encryption to protect the data before it leaves the users' device, thus preventing any other entity from data decryption, including the cloud provider itself. However, while this approach eliminates most of the data privacy concerns, its main disadvantage is that the user cannot readily utilize existing cloud services. For some services, attempts exist at designing alternatives that operate over encrypted data, benefiting from the recent breakthroughs in homomorphic encryption [Gen09]. In addition to resulting in services orders of magnitude less efficient than their counterparts, homomorphic encryption is provably not sufficient for constructing several essential services involving multiple users [VDJ10]. Furthermore, resorting to homomorphic encryption as the ultimate solution requires rewriting most of the cloud applications' code to operate over the encrypted data. New versions of existing \LaTeX compilers, photo filters, music recommenders, etc., based

on homomorphic encryption, will need to be programmed with the goal of keeping all data private, which is evidently non-realistic.

b) Difficulty of manually assessing data privacy levels: Users cannot be expected to individually assess the sensitivity level for each item before they share it as that can require a lot of investment in terms of time and effort, coupled with technical expertise. A recent survey [Pro13] has shown that, in one out of four organizations, the management has little or no understanding of what constitutes sensitive data. Evidently, this fraction is expected to be significantly higher for individual users.

c) General lack of awareness about privacy: This includes limited notions about privacy being restricted to hiding “sensitive” content, such as personal identification numbers, credit card details, etc. Often, the metadata associated with the data item, the location and device from which the item is shared, the entity with whom the data is shared, etc., can be as important as the content of the data itself.

In our solution for privacy provision in the cloud, we seek to overcome the above hurdles.

2.1.2 Approach and Contributions

How do we address the “stumbling blocks” that we identified in Section 2.1.1? First, we show how we can use a centralized solution to facilitate crowdsourcing for privacy *without requiring the revelation of users’ preferences*. We argue that to achieve this, cryptographic methods are infeasible, and we present a novel design that allows users to reveal their preferences to the central server privately. We show how an existing psychologically grounded method for analyzing users’ preferences and data properties, can be rigorously used to analyze this crowdsourced information. Users can then reap the benefits of this crowdsourced information as the server analyzes it to provide them with sensitivity indicators when they share new data.

By crowdsourcing the solution, users are no longer isolated individuals who lack privacy awareness. They can now be guided by the *Wisdom of the Crowd*. Also, they do not have to exert manual effort to find the sensitivity associated with each item they share, as the server can guide them automatically. Furthermore, they need not worry about getting stuck with “bad” crowdsourced information, i.e., about the majority of users being as clueless about privacy as them. This is because the psychometric method we use for analyzing this information, Item Response Theory, ensures that computed parameters of data items do not only apply to a specific sample of people. The solution would ensure, for example, that sharing compromising photos of oneself with the public is deemed risky even when the majority of the participants in the system are doing so. Only a few conservative users in the system are enough to keep the system risk-averse. Finally, we validate our design with both simulation and empirical data, thus showing the feasibility of our solution.

Specifically, we make the following main contributions in this chapter:

- We propose a privacy framework, called Context-aware Crowdsourced Cloud Privacy (shortly C3P), which is specific to the cloud scenario and incorporates the nuances of data sharing, such as the *Privacy vs. Services Dilemma* and *Lack of Privacy Awareness and Effort* on the part of most users.
- We create a realistic vocabulary for a personal cloud, and use it to create “Human Intelligence Tasks” on the *Amazon Mechanical Turk*. We measure people’s responses, in terms of their privacy attitudes, against the *Item Response Theory* (IRT) and find a good fit. We thereby demonstrate that Item Response Theory, a well-used psychometric model for diverse purposes, can be applied fruitfully in the cloud scenario.
- Our solution depends on crowdsourcing the contexts and policies associated with shared items. The sensitivity associated with different items is determined by grouping together same (or similar) contexts and analyzing different policies set by people with different privacy attitudes. However, we also have to ensure the privacy of this aggregated context information. Towards that aim, we provide a lightweight mechanism based on *K-Anonymity* [Swe02] for privately calculating the similarity between items in a centralized way, without depending on infeasible cryptographic methods.
- We perform a set of experiments using synthetic data, with various graphs for user activities, item distribution, and types of users (honest vs. malicious).
- Finally, we use the *Enron* email dataset for evaluating C3P [Nui]. This dataset gives us a good model of users sharing activities and the diversity of data items (and their contexts). Under both datasets, we show that our scheme bootstraps quickly and provides accurate privacy scores in varying conditions.

2.2 System Model

2.2.1 Interacting Entities

We consider a system involving interactions between two types of entities: *end-users* and *cloud service providers (CSPs)*. The end-user can play one of two roles: *data sharer* or *data observer* while the cloud provider can only be a data observer. A data sharer is an end-user who shares *data items* she possesses. A data observer is any entity that is given access to observe the shared items by the data sharer.

We assume that the user sends her data to a single CSP, called the *intermediary* that acts as the repository for this user’s data. The user can select to give other 3rd party providers access to her data through that CSP (e.g. when the latter has an API that the other providers can use). The interaction between these two types of entities is in the form of data sharing operations. Each such operation is initiated by an end-user s_0 who shares a data item d (e.g. document, picture, etc.) with a CSP or another user s_1 . Additionally, the data sharer intends from the sharing operation to obtain a certain service of interest, such as music streaming, document viewing,

file syncing, etc. The network is dynamic, in the sense that these entities can enter and leave the network, and the user items can be shared over time, not necessarily concurrently.

2.2.2 Threat Model

We assume that the user is interested in hiding her sensitive data from the CSPs. Existing privacy threat models, concerned with structured data, consider an adversary who attempts at discovering quantifiable sensitive information, such as location, browsing history, credit card information, etc. In our model, we do not set an a priori definition of sensitive information due to the heterogeneity of the shared data items we consider. Instead, we develop a protocol that quantifies the sensitivity of a certain sharing operation (determined by its context), based on the protection mechanisms that people use. Furthermore, we assume that the CSP is *honest but curious*, in the sense that it follows the protocol, but it can arbitrarily analyze the protocol transcript offline to infer extra information.

2.2.3 Our Conceptual Framework

We now discuss the key concepts and components that underlie C3P, our conceptual framework for privacy provision in the cloud.

Context Vocabulary In Section 2.3, we use the notion of *Context vocabulary* to define the contexts of items shared in a given domain. A context accounts for the content features of the items, the metadata associated with it, and the environment of the sharing operation (e.g. data observers, the device used, etc.).

Sharing Policy People can share different data items with different policies, where a policy is in the range $[0, 1]$ and **0** signifies full transparency while **1** signifies full obscurity. We discuss this in more detail in Section 2.3.2.

Crowd-Sourcing In C3P, after each sharing operation, the context of the item and the policy applied are eventually aggregated at the cloud via a privacy preserving mechanism. This aggregation is required so that the *Lack of Privacy Awareness* may be overcome, and individual decisions could be guided by the *Wisdom of the Crowd*.

Risk Evaluation Based on the processing and analysis of the crowdsourced information, the system can guide others about the privacy risk that is posed by sharing different items in different contexts. Towards that aim, we use *Item Response Theory* (IRT) which is a well-known psychometric function that has been widely used in psychology, education, public health, and computerized adaptive testing.

Policy Recommendation The final component in C3P is a suite of risk mitigation applications. By this, we mean system recommended policies that can guide the general user in minimizing risk while still availing services. In this work, we do not focus on Policy Recommendation and

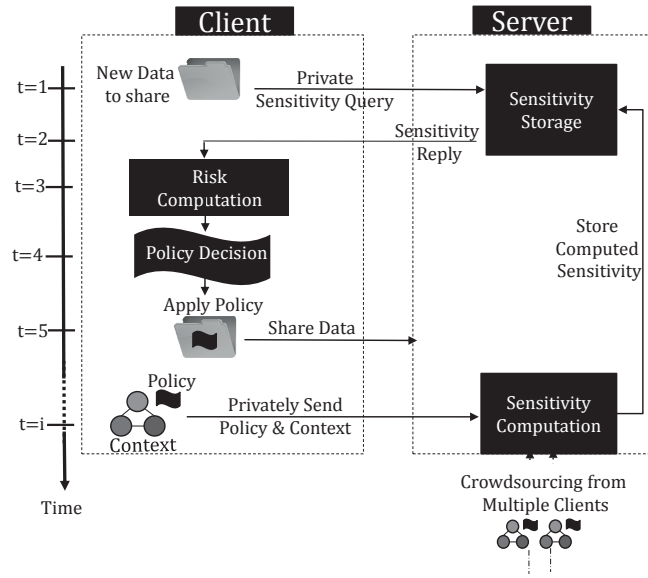


Figure 2.1 – Sequence diagram of the system

leave it for future work.

In Figure 2.1, we show a sequence diagram summarizing the steps taken in one run of the system. The client contacts the server with a private query about the sensitivity of the current context ($t=1$), in a way that the server remains oblivious to the actual context. Upon receiving the response with the sensitivity ($t=2$), the client locally computes the privacy risk of sharing the data ($t=3$) and decides on the relevant protection mechanism ($t=4$). Next, the client sends the data at $t=5$. At a later round ($t=i$), the client sends the context along with the used policy after it makes sure that the server cannot associate the context with the actual sharing operation. The server determines the similarity of this item with other items that users have crowdsourced to it. Using psychometric functions, the server computes the sensitivity associated with the item being shared, which is used to respond to future sensitivity queries.

2.3 Context Vocabulary and Sharing Policies

We begin by describing the fundamental C3P building blocks, which refer to the context in which an item is shared and the policy with which the item is shared.

2.3.1 Context Vocabulary

We introduce the technical notion of “Context”, which includes the metadata associated with a particular data item, user supplied information about the data item (such as tags), and the environment features in which the data is being shared (such as the device information or the relationship with the observer). Furthermore, “Context” also includes information

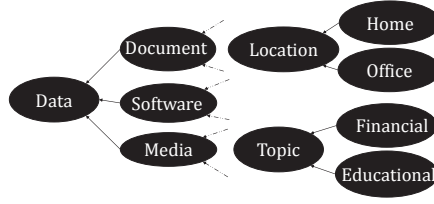


Figure 2.2 – An example vocabulary for data sharing in the personal cloud

extracted through content analysis of the data item, such as topic modeling in the case of a text document and face recognition in the case of images.

For an illustration of “Context”, consider a case where Bob shares a *word document* on *financial risk*, authored by the *sharer* (i.e. Bob himself) on Bob’s *laptop* and shared with a *colleague*. The words in italics capture the context of the data item. For a specific domain, “Context Vocabulary” is the set of all features that can be used to represent any shared item in that domain. Put in another way, the context vocabulary is the vocabulary that can be used to represent all possible contexts in a given domain. We give an example of such a vocabulary in Figure 2.2.

The general template for a context of an item would be a tuple of the general form:

$$(f_1=\text{value}_1, f_2=\text{value}_2, \dots, f_m=\text{value}_m)$$

containing m features. Thus, the context of the data item in the above example would be

(file_type=word document, topic=financial risk, sender_device=laptop,
author=sender, observer=colleague).

It should be noted that there are usually two kinds of features associated with a data item. The first are those which are by default associated with the data item, e.g., *data type*, and other metadata information, e.g., *author*, which are available (or can be extracted by anyone) if the data item is shared completely transparently as in plaintext. We term these as *explicit* features. The second are defined by the sharer when sharing the data item or are based on a private knowledge base (e.g. *observer* or other tags associated with the item). We term these as *implicit* features.

We note here that it is not necessary (or even usual) for all data items to have all context features available. An item’s context is defined by whatever features are available. For example, if we have a *pdf* file which does not have its *author* present, then obviously the *author* feature in the file’s context would be empty.

2.3.2 Sharing Policies

When a user decides to share a data item, this is done with a policy. This policy ranges from 0 to 1, where 0 signifies full transparency while 1 signifies full obscurity. For example, if the

user decides to encrypt a file, then this would be symbolized by a policy value of 1. On the other hand, sharing an unencrypted file while hiding some meta-data features (e.g., author, modified_by, etc.) would result in a policy value between 0 and 1. Between these two extremes lie the other obfuscation methods.

2.4 Crowd-Sourcing and Risk Evaluation

As shown in Figure 2.1, a client can privately query the server about the sensitivity of a specific sharing operation and get a response based on that. In this section, we describe these parts of C3P in more detail. Informally speaking, the privacy guarantee that we achieve throughout is that, at any time, the server has multiple contexts that can be associated with each sharing operation. Accordingly, the context of each operation is never deterministically disclosed to the server.

2.4.1 Privacy Aware Querying

Directly sending the context to the server allows it to associate the sharing operation with that context, which we aim to avoid. Instead, we describe a scheme, where the client queries the server about multiple dummy contexts, in a way that hides the actually requested one¹.

QuerySet Formation

We denote by `targetContext` the context for which the client is querying. This context is sent as part of a `QuerySet`, containing other contexts, which we term as *homonyms*. As shown in Figure 2.3a, suppose that the `targetContext` is $c_1: (f_1=x_1, f_2=v_1, f_3=w_1)$. The client forms a list of alternative values for each feature, e.g. $L = [\{x_1, x_2, x_5\}, \{v_1, v_3, v_6\}, \{w_1, w_2, w_3\}]$ so that, in total, each feature has k possible values. Then the homonyms are formed by producing the cartesian product of all the sets in L . This results in contexts having different combinations of feature values. With m features per context, L has k^m contexts in total.

The choice of the alternative feature values is not totally at random. In order to allow `targetContexts` to appear faster in multiple `QuerySets`, thus approaching the privacy condition formalized in this section, the client keeps a `Pending List (PL)`, containing previously queried `targetContexts`. The feature values of those contexts are used in forming the set L (cf. Figure 2.3a). In particular, we select at random a maximum of $\lceil p \times k \rceil$ values² of those values per feature. The rest of the potential feature values are sampled at random from the domain of each feature.

The client sends this `QuerySet` to the server. The server, on receiving a `QuerySet`, responds

¹This querying step is partially similar to other obfuscation techniques (e.g., [HN09]), but its guarantees differ due to the specifics of our context.

² p is a constant ($0 < p < 1$) (we take $p = 2/3$ in our experiments).

with a subset of all those contexts for which it knows the sensitivity³. If the sensitivity of the `targetContext` is also returned by the server, the client decides to apply a policy on the data item based on the sensitivity value; otherwise, the client can choose to do the same uninformed. From an implementation perspective, this sharing policy can either be applied automatically or used to suggest settings that the user can approve. In this work, we focus on computing the sensitivity, and we give a practical example of how it can be practically integrated in Section 2.6. Next, the actual data item is sent to the server. Once the server receives the actual data item, it can try to infer as much as it can from the `targetContext` and the item. We distinguish between two parts of the `targetContext`:

- *exposed* part: This part consists of those *explicit* features as defined in Section 2.3.1, which the client did not choose to hide.
- *unexposed* part: This part contains all the *implicit* features and the subset of the *explicit* features which have been hidden by the client according to the sharing policy.

It is evident to notice that, by the construction of the `QuerySet`, the server is not able to deterministically infer any feature of the *unexposed* part of the context. In particular, the server has k possible values for each *unexposed* feature. Accordingly, assuming there are u features in the unexposed part, we will have k^u contexts that match the exposed part of the `targetContext` (remember that each feature had k values in the `QuerySet`). We call this set of contexts the *Anonymity Subset* (A_i) of the `targetContext` c_i , and we illustrate its contents with an example in Figure 2.3b. With respect to the server, one of the elements of this subset is the `targetContext`, but no element can be ruled out without further information.

To formulate our findings, we present the following definition:

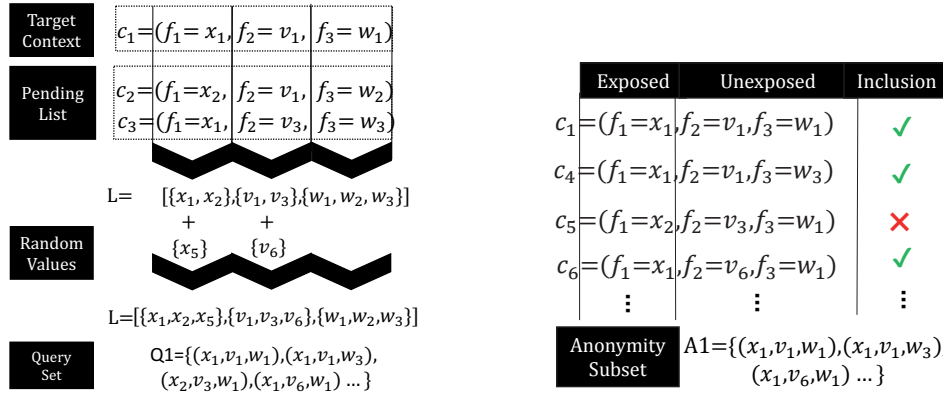
Definition 1. *We say that a context c can be validly associated with the sharing operation of item d_i if c has appeared in A_i and if the server cannot assert with certainty that c exclusively belongs to one or more *Anonymity Subsets* other than A_i .*

For example, based on Figure 2.3, c_1 , c_4 , c_5 , and c_6 are example contexts that can be validly associated with the current sharing operation. Given what the server currently receives (i.e., the queries and the data), it cannot gain any evidence that these contexts are dummy contexts.

Hence, at this stage, we have the following guarantee:

Guarantee 1. *At the querying phase, the server receives k^u contexts that can be validly associated with the current sharing operation.*

³We shall discuss how the server calculates this sensitivity in Section 2.4.3.



(a) QuerySet formation: Consider the given targetContext. We first populate the possible feature values from the Pending List, to get $\{x_1, x_2\}$, $\{v_1, v_3\}$, and $\{w_1, w_2, w_3\}$. Our goal is to have $k = 3$ values per feature. Hence, we amend these values with random feature values from the domain (i.e., x_5 and v_6 in this case). The list L contains k possible values for each feature. The QuerySet is formed from cartesian product of all the sets in L .

(b) Anonymity Subset definition: We start with QuerySet that has been formed in Figure 2.3a (we show 4 of its members). The actual value of feature f_1 has been exposed (e.g., when f_1 corresponds to the *file author*, and the user did not hide this metadata feature). Hence, the contexts which have $f_1 \neq x_1$ (e.g., c_5) are excluded. There are k^u contexts still included as a result ($u = 2$ is the number of unexposed features.)

Figure 2.3 – An example showing the formation of the QuerySet and of the Anonymity Subset

Crowdsourcing

Up till now, we have shown how the client privately queries the server about the sensitivity. In order to compute this sensitivity, the server relies on crowdsourcing, through *privately* collecting targetContexts along with the corresponding sharing policies (together called the Crowdsourcing Information (CI)) from different clients. We alternatively say that a context c is *crowdsourced* when $CI(c)$ is sent to the server.

A (non-malicious) client should not send dummy information in this phase in order not to affect the accuracy of the sensitivity computation⁴. Hence, the server should receive the correct contexts and sharing policies from the client. Thus, we now present the scheme in which client sends the CI in a way that continues to maintain Guarantee 1 for all the sharing operations. As a result, the server will be able to know, for example, that a client *Bob* shared a *financial document* with a *colleague* in a *plaintext form*, but it will not be able to link the document topic or his relationship with the observer to a specific sharing operation.

One way that guarantee might be weakened is if the client sends the CI in a way that allows the server to discover the Anonymity Subset in which the context was the targetContext. For example, sending $CI(c)$ directly after c has appeared in a single Anonymity Subset A_1 will reveal to the server that c corresponds to data d_1 . In this case, all the other homonyms in A_1

⁴We do not discuss the case of malicious clients here, but we do study the effect of such clients in the simulations later.

will no more be validly associated with it. Hence, the first intuitive measure for preventing this association is to wait until a context appears in multiple `Anonymity Subsets` before sending the `CI`.

However, this measure is not sufficient. Consider the case of two contexts c_x and c_y , both only appearing in `Anonymity Subsets` A_4 and A_6 . Suppose that we require that a context appears in at least two `Anonymity Subsets` before it is sent. Then, both $CI(c_x)$ and $CI(c_y)$ will be sent directly after item d_6 (with `Anonymity Subset` A_6) is sent. At this point, the server is sure that one of c_x and c_y is the `targetContext` for A_4 and the other for A_6 . All of the other $k^u - 2$ contexts that have appeared in A_4 and A_6 are no more possible candidates for being the actual `targetContext` from the viewpoint of the server. Hence, Guarantee 1 for these two sharing operations is weakened as the $k^u - 2$ contexts are now deterministically excluded from A_4 and A_6 .

The guarantee would be weakened further if there was a third item d_8 that has been subsequently sent, with its context c_8 appearing in A_4 and A_8 . From the server's viewpoint, A_4 is no more a valid possibility for c_8 due to the mapping deduced when c_x and c_y were sent. Therefore, the server can deterministically associate A_8 with c_8 , and the actual context for d_8 is revealed.

The main weakness in this naive method is that it does not account for the fact the server can link multiple sending instances and reduce the possibility of mapping to a single case. Our strategy to counteract that and keep Guarantee 1 is to verify that crowdsourcing the next context preserves the property that each sent context item is still validly associated with all the `Anonymity Subsets` it has appeared in.

At this point we add another definition:

Definition 2. We say that there is a **valid mapping** from a list of contexts to a list of `Anonymity Subsets` if each context in the former can be **validly associated** with a distinct `Anonymity Subset` from the latter.

Suppose the client has just completed the sharing operation i , and is attempting to crowdsource the contexts that have not been sent yet, which are kept in its `Pending List (PL)`. We also denote by `SL` the `Sent List`, containing all contexts that have been crowdsourced previously⁵, and by \mathcal{G} the group of all client's `Anonymity Subsets` up to (and including) A_i . Towards achieving Guarantee 1, a context $\hat{c} \in PL$ can be crowdsourced only when the following two conditions are true:

1. c appears in at least r `Anonymity Subsets`
2. For each $A \in \mathcal{G}$, there exists a valid mapping from the list $SL' = SL \cup \{c\}$ of contexts to the list $\mathcal{G} \setminus A$ of `Anonymity Subsets`.

⁵We assume throughout that such lists of contexts contain distinct elements; i.e., each user sends one (context,decision) tuple for each context.

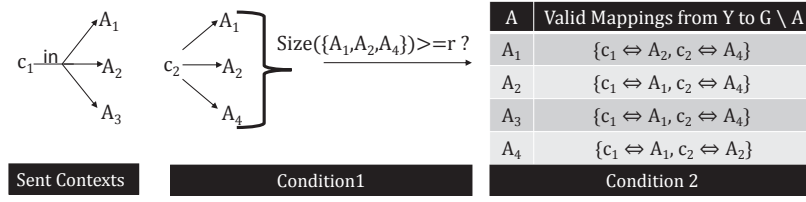


Figure 2.4 – Checking privacy conditions before crowdsourcing context c_2

Going back to the previous example, after each of c_x and c_y has appeared in two Anonymity Subsets, condition 1 is satisfied assuming $r = 2$. However, condition 2 is not satisfied since excluding A_4 will lead to $\mathcal{G} \setminus A_4 = \{A_6\}$, and then we cannot map each context to a distinct anonymity set.

Figure 2.4 illustrates with another example how the two conditions can be verified. For each targetContext, the client maintains a list of the Anonymity Subsets it has appeared in. In addition, it maintains two lists: $U1$, containing the targetContexts that have not satisfied the first condition yet⁶, and $S1U2$, containing the list of items that have satisfied the first condition but not the second. The figure shows a valid mapping that exists for each Anonymity Subset in \mathcal{G} when c_2 is considered for crowdsourcing. It is worth noting that $PL = U1 \cup S1U2$. Also, as discussed in Section 2.4.1, when the contexts of PL appear in more Anonymity Subsets, the above privacy conditions will be satisfied faster; hence, they were used in the construction of the QuerySet.

Theorem 1. *Checking conditions 1 and 2 allows preserving Guarantee 1.*

Proof. Consider any context \hat{c} that is about to be crowdsourced. Let \mathcal{Z} be the list of all contexts that appeared in the elements of \mathcal{G} . In order to preserve Guarantee 1, all the contexts in \mathcal{Z} should remain validly associated with the Anonymity Subsets they have appeared in. We can write \mathcal{Z} as $\mathcal{Z} = (SL') \cup (\mathcal{Z} \setminus SL')$; i.e., it is the union of the contexts that should have been sent after this operation and those that have not been sent yet.

Condition 2 above implies that after \hat{c} is sent, we can still claim that: for each $A \in \mathcal{G}$, there is a possibility that the targetContext of A has not been sent yet. Hence, each context $c \in SL'$, can still be *validly associated* with all the r subsets it appeared in.

It is evident that there is no new information being sent about the contexts in $(\mathcal{Z} \setminus SL')$. Therefore, all the contexts in \mathcal{Z} can still be validly associated with the Anonymity Subsets they appeared in. Accordingly, Guarantee 1 is preserved. \square

⁶regardless of whether the second condition is satisfied

2.4.2 On the Privacy Guarantee

The goal of C3P is to provide a light-weight, non-cryptographic mechanism for sensitivity querying and for delivering crowdsourced information to the central server. So far, we have shown that the provided privacy guarantee is that the server cannot determine—with certainty—that a certain context is associated with a specific sharing operation. The server has k^u contexts that can be associated with each sharing operation, where u is the size of the *unexposed* part of the context.

To give more intuition on why this guarantee makes sense in our setting, we take an example of a user sharing an image with the following context:

```
(file_type=image, scene=indoor, faces=wife, location=home,
taken_with=user_camera, camera_type=smartphone, time_captured=recent,
observer=family ).
```

Assume that the sharing policy consists of stripping the image metadata. Hence, the *unexposed* features are:

```
{location, taken_with, camera_type, time_captured, observer },
```

and the features that are *exposed* when the server receives the image are:

```
{file_type, scene, faces}.
```

If the user wants to only participate in the querying phase and not in the crowdsourcing phase, then the privacy guarantee could be:

Each targetContext, for which the user requests the sensitivity, is anonymized among k^u homonyms ($u = 5$ in this case).

If the user participates in the crowdsourcing phase, then the contexts that reach the server in that phase need to be accurate. Hence, the server must receive the full context along with the user's sharing policy. Our scheme guarantees that the server will not deterministically associate the crowdsourced context with this sharing operation.

Without this guarantee, the server will know that the current receiver of the image is a family member, that this sender is sharing his indoor photos with that observer, that this particular photo has been taken at the senders' home, etc.

With our scheme, the user has the ability to repudiate. The server cannot deterministically associate the current receiver with the observer feature, will not know whether this photo is indeed an indoor photo, and will not be sure of the location where the photo was taken⁷.

Discussion: We note that an alternative scheme for crowdsourcing that includes encrypting the context before sharing it would not work. In C3P, the server is required to use a similarity

⁷We assume that the server does not have additional background information about the user.)

function to match the context with other ones sent by people in order to compute the context sensitivity. Even if we encrypt the context before we send it, the server will be able to know it by computing its similarity with all the possible contexts in the vocabulary (as the latter are not large enough to prevent being iterated over easily). Another place where encryption might be applied is in the querying phase, where *Private Information Retrieval (PIR)* techniques with constant communication complexity might replace the QuerySet technique. However, as the complexity gain is absent, and the privacy guarantee obtained by the querying phase is limited by the crowdsourcing phase, we do not resort to the encryption-based method, which is more complex to implement.

2.4.3 Sensitivity and Risk Evaluation

When the server receives the Crowdsourcing Information, it seeks to determine the sensitivity associated with this item based on same or similar items shared with different policies in the past by different users. The client, upon receiving this sensitivity, locally computes the privacy risk of sharing. In this chapter, for computing the sensitivity, we use *Item Response Theory (IRT)*, a well-known psychometric function, which we describe next.

Sensitivity Computation by the Server

Item Response Theory (IRT) is a modern test theory typically used for analyzing questionnaires to relate the examinees' probability of answering a question correctly (or in general a correct response probability P_{ij}) to two elements: (1) the difficulty of the question (or in general a latent threshold parameter β_i of item i) and (2) the examinees' abilities to answer questions (or in general a latent parameter θ_j for each person j). In contrast to *Classical Test Theory (CTT)*, which measures a person's ability based on averages and summations over the items, IRT has two distinguishing features: (1) the group invariance of calculated item parameters (i.e. a single item's parameters do not only apply to the current user sample, assuming the social norms will not vary significantly) and (2) the item invariance of a person's latent trait (i.e. the trait is invariant with respect to the items used to determine it) [Bak01].

In this work, we apply IRT by mapping the item's difficulty to the sensitivity, the user's trait to the privacy attitude (or willingness to expose the items), and the response probability to the policy level of the item. Although this mapping has been done in the context of Facebook profile items [LT10, QCP⁺12], we are the first to tailor it to multi-featured items based on unstructured data.

We focus on the unidimensional IRT models, which make three main assumptions about the data: (1) unidimensionality (i.e. there is a single underlying trait θ that determines the person's response), (2) local independence (i.e. for each underlying trait θ , there is no association between responses to different items), and (3) model fit (i.e. the estimated item and person parameters can be used to reproduce the observed responses) [RF05]. An IRT model is termed

as *dichotomous* if the responses to the questions are binary ones (correct/incorrect) and *polytomous* if there are multiple levels of the response (e.g. a five-level Likert scale with responses: strongly disagree/disagree/neutral/agree/strongly agree).

The Rasch model, one of the most common IRT models, assumes that the probability of correct response is a function of θ and β only and that the items are equally discriminative for testing the underlying trait. It is particularly advantageous with smaller sample sizes, due to its simplicity and few parameters, and, as we show in Section 2.5.1, it also fits well in the scenario of cloud data sharing. The parameters of the dichotomous Rasch model for an item i and a person with parameter θ are related by the following function, called the *Item Response Function (IRF)*: $P_i = 1 / (1 + e^{-(\theta - \beta_i)})$.

With polytomous models, we will make the assumption that the policies chosen by the users are on the same scale for all the items. It is similar to the case of Likert scale, where the same set of categories are applied for each item in the test. Accordingly, the most suitable model for us, and whose fit to the cloud scenario will be demonstrated in Section 2.5.1, is the Rasch Rating Scale Model. For estimating the parameters of the different models, we used *Marginal Maximum Likelihood estimation*, which is an expectation-maximization algorithm. The estimation technique relies on having enough responses for multiple items by different people. For more details about item response theory models, the reader is referred to the following works [Bak01, RF05, NO11].

Risk Computation by the Client

The sensitivity is an indication of the magnitude of privacy loss incurred when data is lost. The client can combine this measure with another measure of the *likelihood* that this event happens, using information that is kept locally, such as the level of *trust* for the current observer, the level of *protection* (i.e. the policy as we show do in Section 2.6), etc. The privacy risk is then a combination of the sensitivity and the likelihood.

2.5 Evaluation and Experiments

2.5.1 Experiments for Validating IRT

Since we shall be using Item Response Theory (IRT) to calculate the sensitivity of shared items, the first question that needs to be answered is this: *Can IRT be meaningfully applied in the cloud scenario in which people share data items in a variety of contexts?* In order to investigate this and to empirically ground our design and subsequent experiments, we validated IRT for the cloud scenario using real people's feedback on Amazon Mechanical Turk. Next, we explain our methodology for this validation.

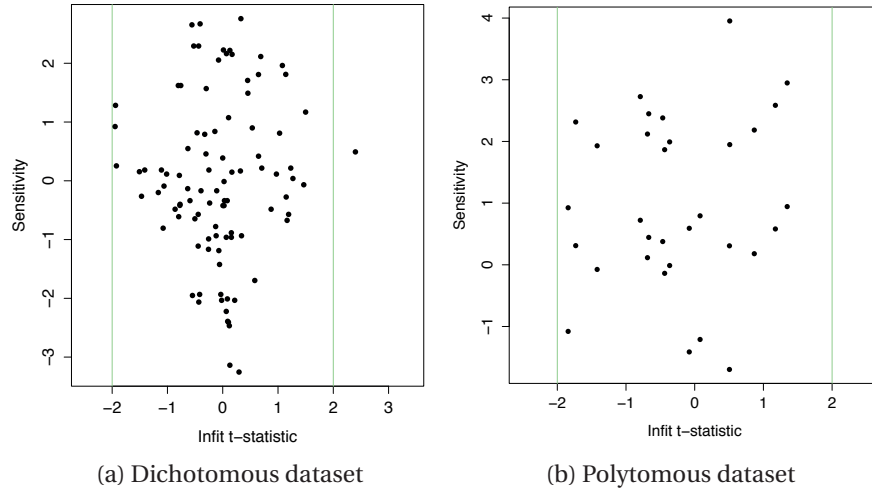


Figure 2.5 – Bond-and-Fox Pathway Map on the mTurk data (a dot represents a context item)

Methodology

We created a realistic vocabulary for the personal cloud, and, based on it, we developed a list of questions that we submitted as *Human Intelligence Tasks* (HITS) on Amazon mTurk⁸. We created two separate HITS for the dichotomous and polytomous cases of IRT. For the dichotomous case, we asked 96 questions to which we received answers from 81 people. For the polytomous case (with 3 categories), we asked 16 questions to which we received answers from 50 people⁹. Here each question represents a *context item* while the users' responses represent their *policies* for sharing in the given context.

We analyzed the results using the eRm Package in R [MH07]. For testing the model fit, we used the standardized (STD) and the mean square (MSQ) infit statistics. An infit statistic is a weighted fit statistic derived from the squared standardized residual between the observed data and the one predicted by the model [DA09]. The STD infit indicates whether the data fits the model *perfectly* and is also an approximate t-statistic. In Figure 2.5, we show the STD infit statistic in the two cases of dichotomous and polytomous items, along with the sensitivity value of items (threshold values in the polytomous case) in each graph, also called the *Bond-and-Fox Pathway Map*. We notice that all the values in the polytomous case and all but one in the dichotomous case lie between -2 and 2, which are the typically acceptable bounds [DA09]. We also derived the MSQ infit which serves as an indication of whether the data fits the model *usefully*, i.e. if it is productive for measurement. We found that the MSQ infit was in the range [0.7, 1.312] for dichotomous items and [0.683, 1.287] for polytomous items, which are both within the typically accepted [0.5, 1.5] range [DA09].

Having shown the applicability of IRT to the cloud sharing scenario, we proceed to the evalua-

⁸The vocabulary and the survey are shown in Appendix A

⁹The numbers of respondents is generally considered a good number for testing IRT [Lin94]

tion of C3P.

2.5.2 Synthetic Datasets

In this section we detail, our methodology for evaluating our framework with synthetic data, followed by the experimental results and discussion.

Methodology

The context items in this dataset were generated by selecting a generic vocabulary with 5 features per context. Each feature of a context had 5 possible values for a total of 3125 possible contexts. From these contexts, we selected 200 ones at random. There are 500 sharers (or people) who share these items. In total, for each experiment, we allocated 30000 sharing instances, each of which represents a data item (corresponding to the context item) shared by a certain person with another person at a specific time. The item to share at each instance is drawn according to a predetermined item distribution (zipf with exponent 2, or uniform, depending on the experiment).

In our implementation, the distance (and hence similarity) between each pair of contexts is based on considering the hamming distance over their features¹⁰. The people connections for sending data were modeled using two types of graphs: (i) small world (using the Watts-Strogatz model with a base degree of 2 and $\beta = 0.5$) and (ii) random (with an average degree of 4). Our simulation is a discrete event based simulation, punctuated by sharing events.

The person who instantiates a sharing event is selected randomly from the graph, weighted by her degree, so that people who have more neighbors share more items than those with less. The data receiver is selected randomly from the list of neighbors of the sender. Each person sends data at a time rate modeled by a Poisson process so that the time between her two sharing instances is exponentially distributed with an average of 3, 6, or 12 hours, depending on the experiment.

At each sharing instance, the context item's QuerySet is sent according to our scheme. The server maintains clusters of contexts it receives, grouped according to a *similarity parameter* (whose value of 1 implies that each cluster's contexts differ by one feature from their cluster center, etc.). When the server receives a new context, it either maps it to an existing cluster or assigns it as the center of a new one. All the contexts of a certain cluster are assumed to have the same sensitivity. The server replies with all the sensitivities it knows for the clusters to which the contexts in the QuerySet were mapped. If the reply contains the requested item, this is considered as a *Hit*.

In the crowdsourcing phase, upon receiving new Crowdsourcing Information (CI) from a

¹⁰System designers can use any similarity measure best suited for their needs, e.g., those dealing specifically with semantic similarity. However, that is beyond the scope of this work.

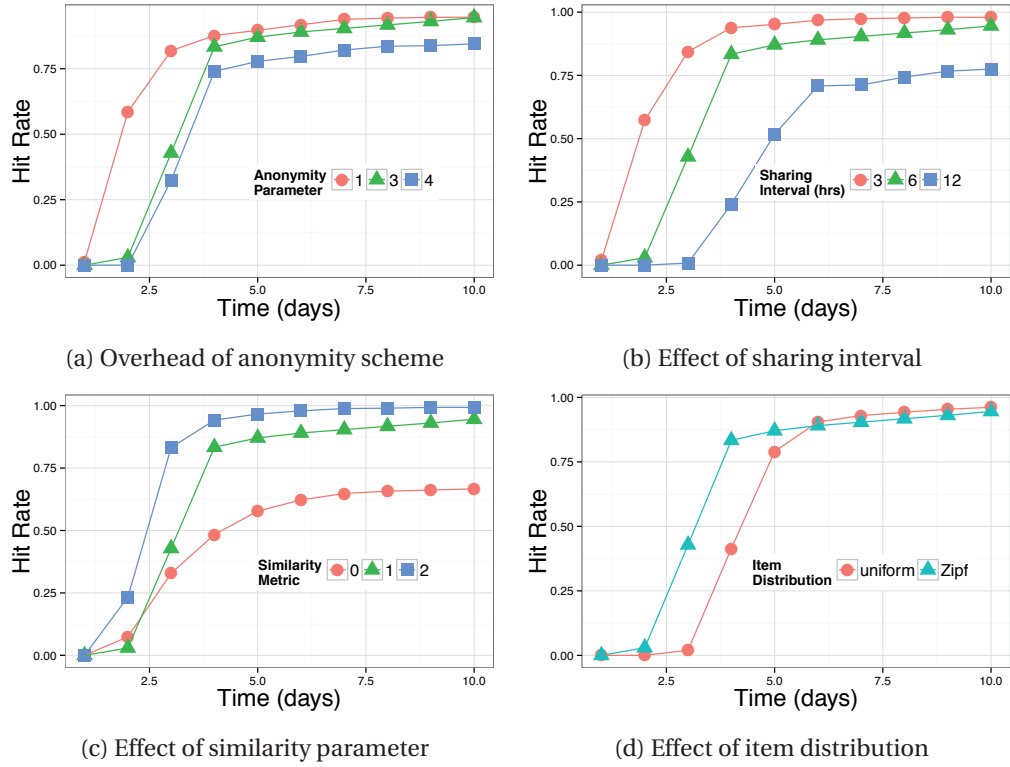


Figure 2.6 – Synthetic dataset graphs

client, the server matches it to a cluster S and tries to compute the sensitivity for S if it is not yet computed. To achieve an acceptable sample for IRT, we require that (1) S has a minimum of 15 contexts with their policies, (2) that there are 4 other clusters satisfying requirement 1, and (3) that each of these 4 clusters has at least 8 CIs by people who have also sent CIs appearing in S . The sensitivities are computed using the *marginal maximum likelihood estimation* technique. In all the experiments, unless otherwise stated, the default setting is a small world social network with zipf item distribution, six hours average sharing interval, and a similarity parameter of 1. In addition, the value for parameter r is equal to k , which is 3 by default. Hence, k is the anonymity parameter we use henceforth.

Results and Discussion

Figure 2.6a shows the Hit Rate of users queries over time, where Hit Rate is defined as:

$$\text{Hit Rate} = \frac{\text{\# of queried items with available sensitivity}}{\text{total \# of queried items}} \quad (2.1)$$

The Hit Rate is calculated per day unless otherwise specified.

Anonymity Overhead In Figure 2.6a we can see that the Hit Rate for anonymity parameter 3 is better than the Hit Rate for 4. As discussed earlier, anonymity parameter k implies that a `targetContext` for sensitivity must have appeared in k different `Anonymity Subsets` and that k different values for each feature in the `targetContext` must be present in the `QuerySet`. The above conditions suggest that the lower the anonymity parameter value, the more `targetContexts` would be sent to the server for crowdsourcing, and thus the more quickly would IRT be able to respond with sensitivity values. The anonymity parameter 1 implies no anonymity at all. We plot this curve to see the “overhead” of our K-anonymity scheme on top of the time required by IRT. Simply put, the curve for the anonymity parameter 1 represents the time it takes IRT to provide Hit Rates when there is no anonymity scheme in place. Thus the difference between the curves for anonymity parameters 1 and 3 represents the overhead of our anonymity scheme in terms of reduced Hit Rate. However, we see that the curve for 3 converges to the same Hit Rate as 1 in ten days time. This suggests that our anonymity scheme bootstraps quickly and does not pose significant overhead¹¹.

Sharing Interval Effect Figure 2.6b shows the Hit Rate with different sharing intervals in hours. An interval of 3 means that all users query for the sensitivity of an item every 3 hours on average. It can be seen from the graphs that initially, the longer the interval, the slower the increase in the Hit Rate. This is most noticeable around the 5th day when the Hit Rate with an interval 12 is still around 0.5 and lags significantly behind. Eventually, as the server collects more and more items, the Hit Rates of all sharing intervals converge to similar values.

Similarity Parameter Impact Figure 2.6c shows the Hit Rate with different similarity parameters. The similarity parameter has been defined in Section 2.5.2. A similarity parameter of 0 signifies that there is no (zero) difference between two context items while calculating sensitivity¹². Precisely, what this means is that: to calculate the sensitivity of an item, IRT would require that other contexts, which are exactly the same as this context, be shared with different policies. A similarity parameter 1 implies that two items that differ by a distance of 1 would be considered the same while 2 implies that items differ by a distance of 2 would be considered the same. This, in turn, implies that IRT would be able to more quickly calculate the sensitivity of an item (as opposed to case 0) since there would be more items which are considered the same. Thus we can see in Figure 2.6c that Hit Rate with similarity parameter 0 is the worst since IRT does not have enough items for calculation.

Item Distribution Effect In Figure 2.6d, we investigate the effects of the “item distribution” on the Hit Rate. By “item distribution” we mean the distribution of the context items, i.e., the different contexts in which users share data. This is an important feature because different

¹¹This overhead can be further reduced through bootstrapping the system with initial data collected from surveys, thus increasing the Hit Rate at the beginning.

¹²This was the case for example in the experiments for validating IRT in Section 2.5.1

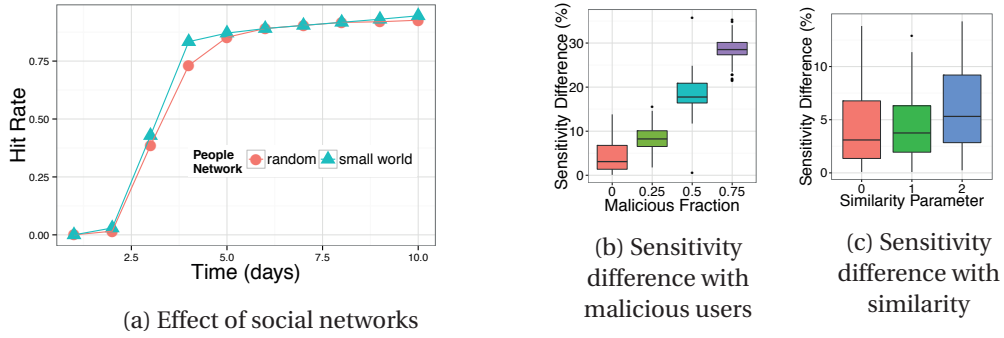


Figure 2.7 – Hit rate and sensitivity difference under various conditions

organizations and different systems would naturally have different item distributions. For our experiments, we use two different item distributions. One is the *zipf* distribution, which has been shown to be most common in social networks [LKG⁺08]. The other is the *random* distribution in which all context items are randomly distributed. A look at the Figure 2.6d reveals that a *zipf* distribution “bootstraps” faster than a random distribution. The Hit Rate with random distribution lags behind *zipf* by a day, i.e., it reaches the same Hit Rate a day later, till the fifth day. We argue this is because, given a *zipf* distribution, users share more similar items, and thus the crowdsourcing is more effective, and IRT is able to calculate the sensitivity of items quickly. Given a random distribution, it takes more time for IRT to accumulate enough similar items for the calculation of sensitivity. However, as the times goes by and more items are accumulated, both random and *zipf* converge to around the same values.

Effect of Social Graph In Figure 2.7a we observe the effect of changing the underlying social network. We use two graphs for the social network structure: small world and random. These affect the sharing patterns of the users. We see that the effect of the underlying graphs on the Hit Rate is not significant and both lead to similar values, with the small world doing slightly better than the random network.

Effect of Malicious Users on Sensitivity Figures 2.7b and 2.7c show the effect of malicious users and changing similarity parameters on the sensitivity values. For this particular set of experiments, we begin by assigning different sensitivity values to the items and also different attitudes to the users. As the experiment runs, the policy of the users on sharing the items is dictated by their attitude and the item sensitivity given at the beginning. The sensitivity of the items is then calculated by our scheme. We then measure the absolute difference between the actual sensitivity of the items and the calculated sensitivity. Ideally, there should be no significant difference between the actual sensitivity and the calculated sensitivity. However, differences could arise under certain conditions. The first condition is the presence of *malicious users*. A malicious user sends random policies for items, i.e., she does not have a fixed

attitude but rather a random and unpredictable one¹³. Figure 2.7b shows the effect of such malicious users on our scheme. The figure shows the box plots for each item's normalized sensitivity difference in terms of percentage. We observe that when there are no malicious users, the difference is pretty low (in the range [2%, 6%]), with most items' calculated sensitivity very near the actual sensitivity (the individual dots represent the outliers). This keeps getting progressively worse as the proportion of malicious users increases. Finally, with a fraction of 0.75 malicious users, most of the items' calculated sensitivity differs by as much as 30% from the actual sensitivity.

Effect of Similarity Parameter on Sensitivity In Figure 2.7c, we see that the effect of different similarity parameters on the calculated sensitivity. We observe that, with similarity parameters 0 and 1, the difference between actual and calculated sensitivity is very low. The reader will recall that similarity parameter 0 means that two items would only be grouped together if they are identical. Therefore, when IRT calculates sensitivity value of an item, it does so on the basis of other identical items for which it has received policies from different users. Thus the calculated sensitivity value would be in high agreement with actual sensitivity. With increasing similarity parameter values, the system would group together items which are not identical, therefore sacrificing accuracy in sensitivity calculation. We observe that the difference between actual and calculated sensitivity with similarity parameter 2 is greater than 0 and 1. However, as we discussed while explaining the results of Figure 2.6c, a higher value for the similarity parameter signifies a better Hit Ratio. Therefore, we discover that there is a *tradeoff* between accuracy of calculated sensitivity and Hit Rate, as far as similarity parameter is concerned.

Overall, the above results signify that our approach has a reasonable overhead, with varying item distributions and social graphs, and is resistant to a considerable fraction of randomly behaving users.

2.5.3 Enron Experiments

We want to evaluate our scheme in a realistic setting. However, as there is no dataset of users sharing activities in the cloud that is publicly available, we use the Enron email dataset [Nui]. Sharing data of various types with certain entities in the cloud is analogous to sharing attachments via emails. Specifically, what we get from this dataset, is a variety of items (hence the variety of contexts in which real people share these items with others) and also the level of trust that they have in each other. We explain these points as well as our data extraction and analysis methodology below.

¹³We note that, strictly speaking, random choices do not always stem from having intentions to attack the system, but we use the term "malicious" to signify inconsistencies in the users' privacy attitudes.

Methodology

The dataset was obtained in the form of 130 personal storage folders (pst)¹⁴. It was processed using the PST File Format SDK¹⁵ and the MIME++ toolkit¹⁶. We only considered emails with attachments, whose metadata was extracted using the GNU Libextractor library¹⁷. Precisely, the main metadata we extracted from files is: (revision history, last saved by, resource type, file type, author name, and creator). We then collected all the email addresses mentioned in the dataset and grouped the ones corresponding to the same person, based on the patterns of occurrence of email aliases. Next, the emails of each person were used to obtain the list of companies she is affiliated with according to the email domain, filtering out public email services (e.g. AOL, Yahoo). We matched all the processed metadata with a specific vocabulary we created for the Enron Dataset.

In total, the obtained dataset contained 2510 people sharing 184 distinct contexts over 19415 sharing instances. Moreover, for each file sender, we calculated a measure of the trust associated with each receiver based on the frequency of emails exchanged with her. The trust value $T(i, j) = F(i, j) / \text{Max}(i)$, where $F(i, j)$ is the number of emails sent from user i to user j , and $\text{Max}(i)$ is the maximum number of emails sent by user i to any receiver. In our experiments, the policies we associate with each sending event are dictated by this degree of trust between the sender and the receiver. We use a similar timing scale as the synthetic experiments, where each person shares all his items in the sequence the emails were originally sent but with a rate modeled as a Poisson process.

Results and Discussion

Anonymity Overhead Figure 2.8a shows the Hit Rate of users queries over time, where Hit Rate is the same as defined in Equation 2.1. The graph is over a period of 10 days. We can see that with anonymity parameter 1, i.e. with no anonymity scheme in place, the Hit Rate jumps very quickly. However, anonymity parameter 3 and 4 eventually catch up, and all the curves show a Hit Rate of 1 by the third day. We argue that this improvement in Hit Rate over the case of synthetic experiments (see Figure 2.6a) is because the sharing contexts in the Enron dataset are not diverse and more similar items are collected faster, thus leading to an increase in the Hit Rate.

Similarity Parameter Impact In Figure 2.8b we can see that with the similarity parameter equal to 2, the Hit Rate remains at 0 consistently. Our investigation into this reveals to us the reason behind this strange result. We discover that the context items shared in the Enron dataset are not very diverse. Hence, having a similarity value of 2 means that most items

¹⁴from <http://info.nuix.com/Enron.html>

¹⁵<http://pstsdk.codeplex.com/>

¹⁶<http://www.hunnysoft.com/mimepp/>

¹⁷<http://www.gnu.org/software/libextractor/>

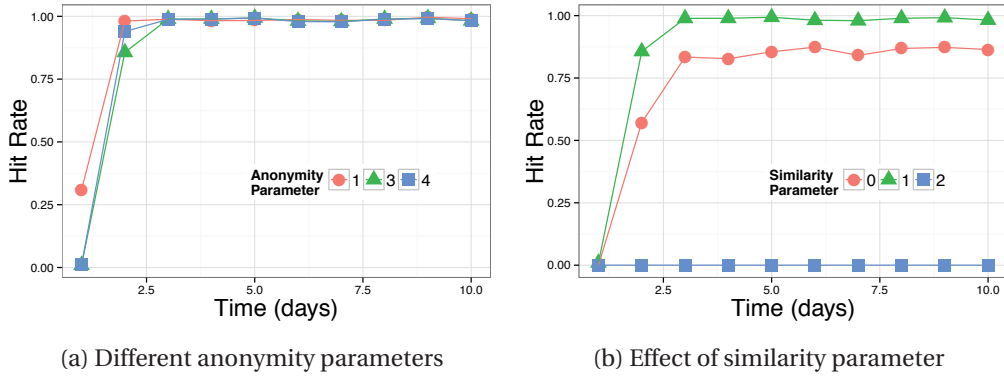


Figure 2.8 – Enron dataset graphs

are clustered together since most items in the Enron dataset differ from each other by at most 2 features. As most items are clustered in very few clusters, this means that IRT is not able to work since it does not find enough different items for sensitivity calculation. The number of different items that IRT requires for working differs on the implementation being used. Our implementation requires that there must be at least 5 different items for IRT to work. In the case of similarity 2, these clusters are not available. However, with similarity 1, enough clusters are found and this, in turn, implies that IRT would be able to more quickly calculate the sensitivity of an item (as opposed to case 0) since there would be more items which are considered the same. Therefore, similarity 1 shows better Hit Rate than similarity 0. These results suggest that using IRT in a scenario where most people share similar items, the similarity parameter should be low. However, we note that the Enron dataset, being an email dataset, does not have the same diversity as would be available in a cloud setting.

2.6 Implementation

We have developed a prototype backend system based on C3P, in addition to two clients, a desktop app (called PrivyShare Desktop) and a web client (called PrivyShare Web), showcasing its potential.

2.6.1 PrivyShare Desktop App

This app, whose main screen is shown in Figure 2.9 allows users to analyze the risk associated with files before they are uploaded to any cloud service. It also gives the option to apply fine-grained policies instead of encryption only. This enables users to balance the privacy-utility tradeoff and to still use certain services with their cloud files (e.g., via 3rd party cloud applications). PrivyShare Desktop supports any cloud storage provider as long as the app for that provider is already installed on the user's operating system. We leverage the fact that each cloud storage provider typically has an assigned folder on the user's filesystem. Hence, when a

Common Features	Text-specific	Image-specific
Recency (day, week, year, more)	Topic Identifier	Small Faces Count
Creator (social group)		Medium Faces Count
Recipient (social group)		Large Faces Count
Device Type (phone, tablet, camera, desktop)		Location available
File Type		Photo Type (drawing, design, picture)
Title available		
Company available		
Contributors available		

Table 2.1 – Features extracted from files in PrivyShare Desktop app; some of the features are extracted from both text files (e.g., word documents, presentations, worksheets, etc.) and images files. Other features are specific to the filetype.

file is to be uploaded to that service, our app transfers the file to the corresponding service folder.

A typical flow of actions for an upload goes as follows:

- A user drops the file into the icon of one of the existing cloud provider (see Figure 2.9). The app extracts the features of the file at the client side and builds the QuerySet according to the C3P protocol. These features are shown in Table 2.1. For feature extraction, we used a combination of Apache Tika¹⁸, OpenCV¹⁹, and Mallet²⁰, which were tailored to our purposes. Then the client sends the QuerySet to the server and receives the sensitivity result. If the sensitivity data is available, it is used to compute the risk, which appears via the risk meter on the right side of the screen. The risk itself is chosen to be a normalized product of the sensitivity and the policy level applied. The policy level ranges from 0 when the file is encrypted to 1 when the file is sent in plain text. In between, there is a range of predefined fine-grained policies (e.g., based on the fraction of metadata removed or the auxiliary information uploaded with the file as we describe below).
- The user has the option to see the most important metadata fields of the file, which are visualized to make them more comprehensible (Figure 2.10). For example, the location metadata shows the actual location of the photo on the map. In addition, the user can hide each of the metadata fields (Figure 2.11), which is one type of the fine-grained policies provided. For metadata hiding we used a combination of Apache POI²¹ and Apache Commons Imaging²².
- Another option provided is to encrypt the file on the client side before uploading (Figure 2.12). In order to still some services on the web apps of storage services (e.g., seeing

¹⁸<https://tika.apache.org/>

¹⁹<http://opencv.org/>

²⁰<http://mallet.cs.umass.edu/>

²¹<https://poi.apache.org/>

²²<https://commons.apache.org/proper/commons-imaging/>

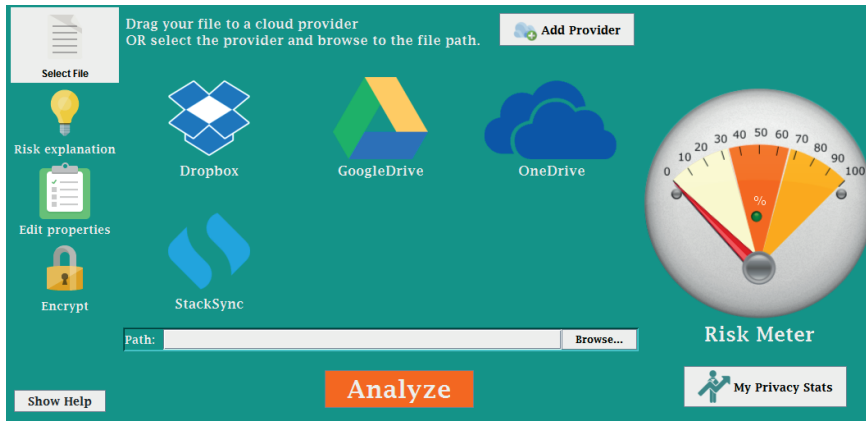


Figure 2.9 – Main screen of PrivyShare Desktop app.



Figure 2.10 – Page for viewing and editing the metadata in PrivyShare Desktop app; more metadata fields are adjustable in a secondary menu.



Figure 2.11 – Ability to toggle metadata visibility by clicking on it.



Figure 2.12 – Encryption options page in PrivyShare Desktop. Clicking on the key are will toggle encryption. Clicking on the thumbnail area will toggle adding a thumbnail.

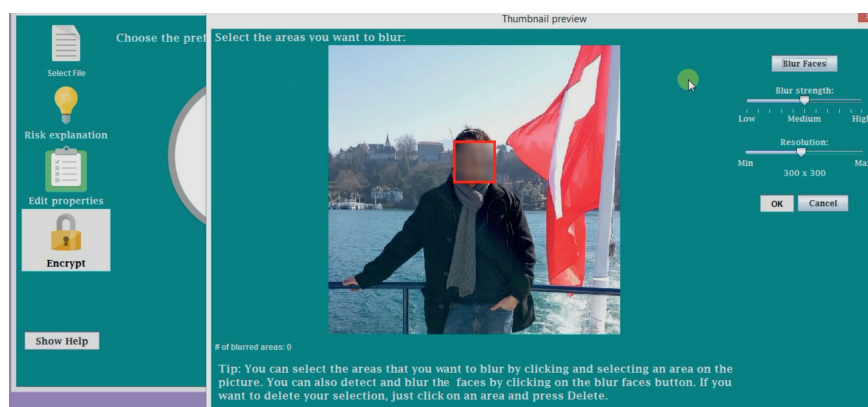


Figure 2.13 – Page for editing thumbnails. Faces can be detected and blurred automatically. The resolution is low by default but can also be adjusted along with the blur strength.

an image thumbnail or searching through text files), we give the option to attach auxiliary info to the encrypted file. In the case of photos, the app can add a low-resolution thumbnail, which allows the photos to be navigable on the cloud storage web apps, without exposing the photo itself (Figure 2.13)²³. Faces in the thumbnail are detected automatically, and the user can also manually blur additional areas. In the case of textual files, the thumbnail of the first page/slide is also attachable. The app gives the option to attach a textual summary to the uploaded file, which can be indexed on the server side.

- When the user confirms the upload, the file is synchronized from its original folder to the folder of the cloud storage service. For example, in Figure 2.14, we see the image and its thumbnail in the original folder. In Figure 2.15, we see the thumbnail with the

²³We note here that, in the current web interfaces of cloud storage services, these thumbnails appear as individual files next to the original files when sorted by name. We have plans to add support for a web interface of PrivyShare that connects to the various cloud providers' APIs and shows these files as one file with auxiliary information.

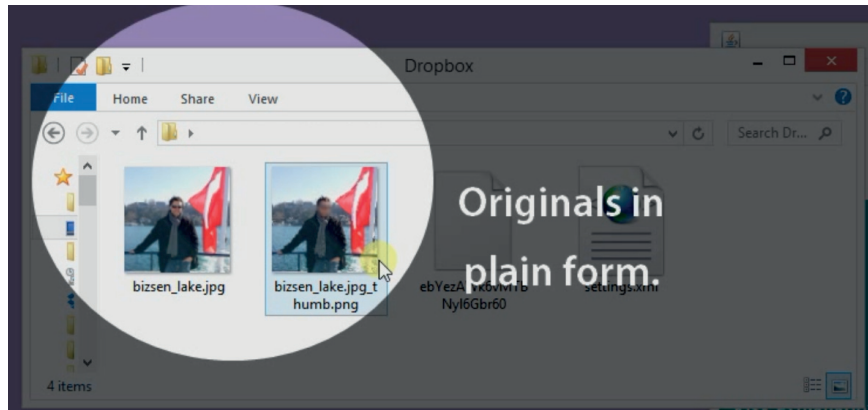


Figure 2.14 – Original image which is intended to be synchronized to the cloud storage service, along with the generated thumbnail.

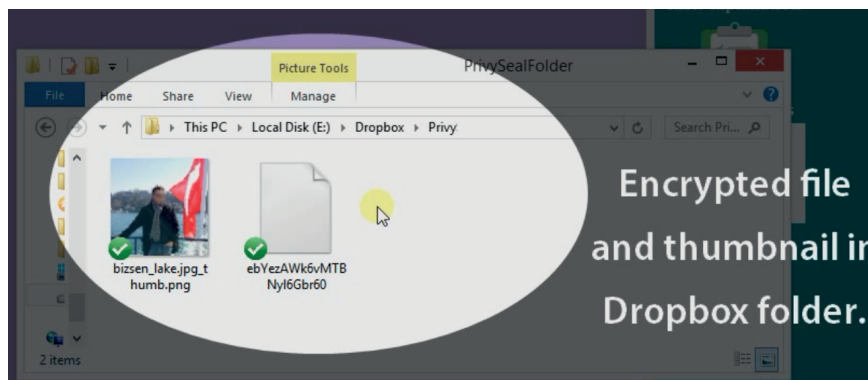


Figure 2.15 – Dropbox folder, to which the image is encrypted and synchronized, along with a low-resolution thumbnail.

blurred face and the full encrypted photo in the Dropbox folder of the user. Whenever the original file changes, the policy is reapplied, and the file is synchronized again. For synchronizing files, including encrypted ones, we utilized JFileSync²⁴.

The goal behind PrivyShare Desktop was to show what is possible to achieve on the client side before uploading files to cloud storage services. Users, however, are not expected to update the fine-grained policy for each file. Instead, they can set a predefined risk threshold they tolerate, and the fine-grained policy is applied based on the sensitivity of the file and the risk threshold. A natural extension of this app is to add the possibility of visualizing the risk across and setting the policy across a larger number of files, which takes the app further towards the goal of helping the users adapt to the scale of data collection by reducing the time spent on configuring their privacy.

²⁴<https://github.com/mgoellnitz/JFileSync3>

2.6.2 PrivyShare Web App

One shortcoming of PrivyShare Desktop is that, despite the users having the option to set fine-grained policies, it is not apparent to them how this reflects on the services they can still use with their files online. Moreover, with the emergence of web apps, users are less willing to install additional apps on their machines. In order to cope with these issues, we developed a web app, called PrivyShare Web, which is utilized by simply going to the app's URL in their browser.

The web app allows users to drop multiple files, whose features are extracted on the client-side of the browser (see Figure 2.17). This includes metadata features (e.g., location) and content (e.g., faces, topic). Then users can apply a policy level, which allows them to encrypt files, depending on their determined sensitivity (via the C3P protocol). The policy level ranges from “zero” (where all files are in plain text) to “high” (where all files are encrypted). For client-side file encryption in the browser, we used SpiderOak Crypton²⁵. The policy level can be automatically set based on users' preconfigured risk threshold, as discussed earlier. It is worth noting that, compared to PrivyShare Desktop, the policy level is fine-grained at the level of all files and not per file (i.e., it adjusts which of the files are encrypted, rather than what policy to apply to a specific file).

In order to educate the user about the privacy-utility tradeoff, our app requests access to the users' list of installed 3rd party apps (e.g., PDF converters or image editors that can import files from cloud storage services). In the example, we show how this works with the user's preinstalled Google Drive apps²⁶. The user can see the list of apps whose functionality might be affected by the current upload (Figure 2.19). By clicking on a specific app, the user can zoom into the actual files which are affected (Figure 2.20).

Both PrivyShare Desktop and PrivyShare Web serve as prototypes showing what is possible with a sensitivity as a service and that fine-grained control over files uploaded to cloud storage is possible. We believe though that there is further need for conducting usability studies for such implementations, to assess how they are perceived by real users, which we leave for future works. Moreover, we note that, despite the various options we illustrated, we envision the system to be highly automated, applying the policy level based on the estimated sensitivity. The users' effort is simply to approve or change the recommended settings before the final upload occurs.

²⁵<https://spideroak.com/solutions/crypton-framework>

²⁶PrivyShare Web gets access to the list of Google Drive through a separate authentication mechanism via OAuth

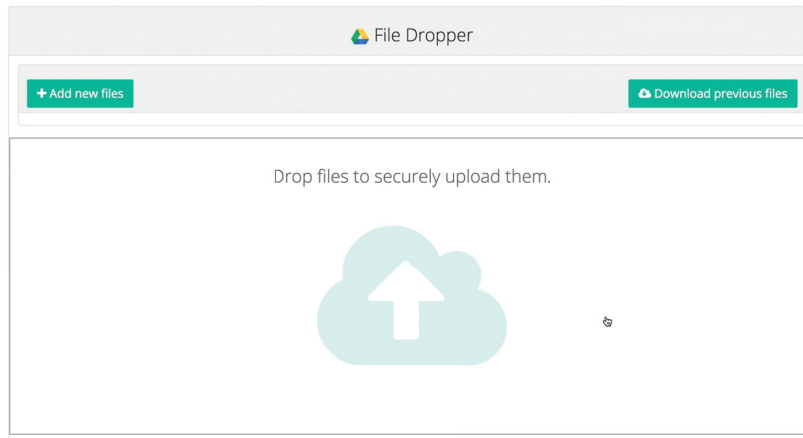


Figure 2.16 – Upload area of PrivyShare Web app.

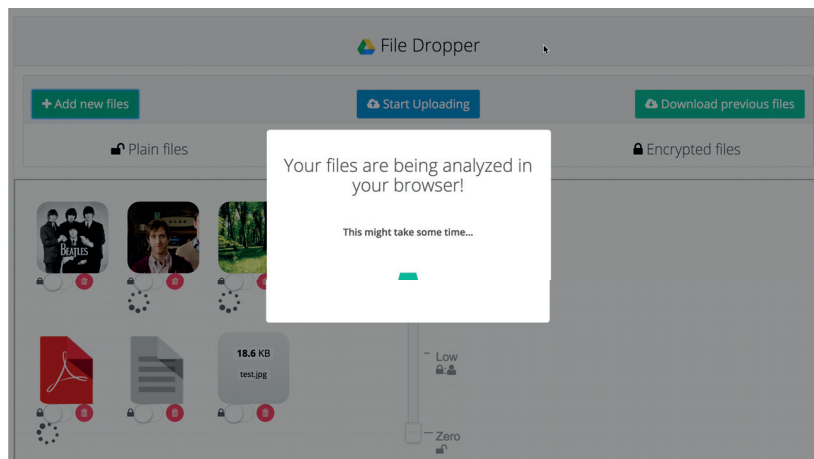


Figure 2.17 – File Analysis interface once the files are dropped into the upload area.

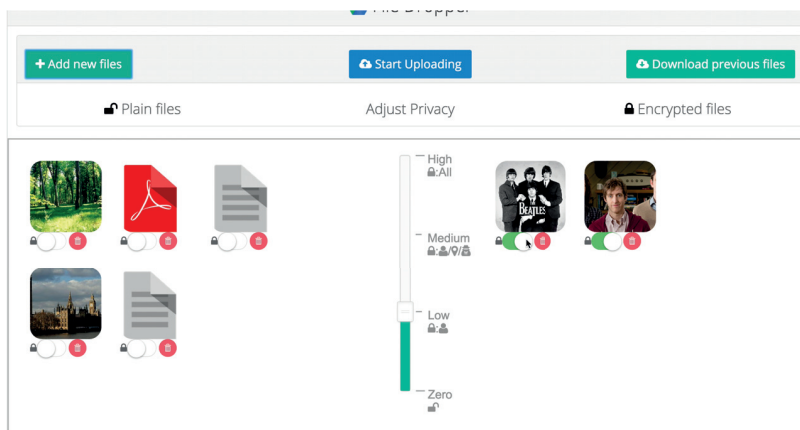


Figure 2.18 – Ability to set the policy level; files which are to be encrypted (based on their sensitivity) automatically move to the right.

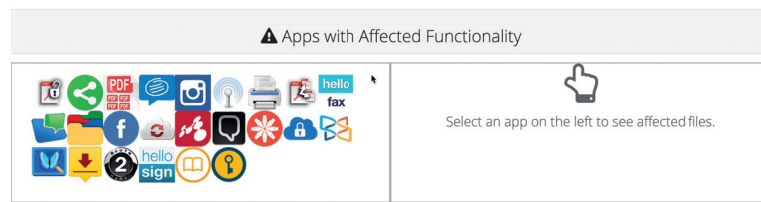


Figure 2.19 – The part of PrivyShare web showing the subset of the user’s apps that can be potentially affected by the current upload (i.e., based on the file types the apps have access to).

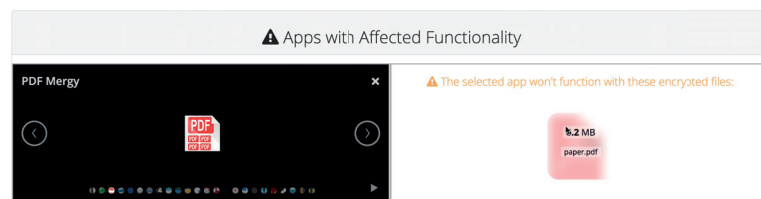


Figure 2.20 – Clicking on a specific app will show the specific files that will no more be usable, due to client-side encryption.

2.7 Related Work

Context-based Privacy Modeling

The concept of contextual privacy has received attention recently. Nissenbaum’s work on “contextual integrity” [Nis11, Nis04] was one of the notable contributions that called for articulating context-based rules and expectations and for embedding some of them in law.

Context-aware systems have been applied in multiple scenarios, ranging from social networks to ubiquitous computing. Schaub et al., showcased a model that abstracts context information about the user, the privacy-sensitive items, and the entities involved in the environment. They took into account information privacy aspects as well as physical aspects, which is of particular importance in the ubiquitous computing applications, such as ambient assisted living. Schaub et al., later developed a system for ambient calendar displays that adapts to the people present in a room and their privacy preferences [SKL⁺14].

Bilogrevic et al., introduced SPISM, a system for (semi-)automatically sharing information on mobile social networks [BHA⁺13]. SPISM allows subscribers (e.g. services or other users) to request contextual information (e.g., location) of other subscribers. Access to this information is granted at multiple granularity levels.

C3P’s differentiating factor compared to these previous works is being the first work to address privacy modeling of unstructured data. The problem setting is also more challenging as the system combining the context information is an honest-but-curious remote server. In the other systems, the device determining data sensitivity is either the same as the device where sensitive data resides or is trusted by the latter.

Interest in context-aware data sharing continued to evolve afterward. The subsequent work by Yuan et al., zoomed into the problem of privacy in photo sharing, with a similar approach [YTE17]. One of their main contributions is that they showed, via a user experiment with actual photos, the efficacy of machine learning in predicting the users' policies based on image semantics. In the field of mobile permissions, there have been also later works on automatically predicting the default app permissions [LAS⁺16] and on continuously managing realtime permissions of Android apps [WBT⁺17, ODPSM⁺17].

Risk Estimation

One of the relevant attempts at privacy risk estimation was in the field of social networks. Liu and Terzi [LT10] used IRT in quantifying the privacy risk of exposing user profile items on Facebook (e.g. birthday, political affiliation, etc.). Quercia et al. [QCP⁺12] also modeled the process of information disclosure in Facebook using IRT and found correlations between disclosure scores and personality traits. In this chapter, we have shown that IRT applies in the case of privacy aware sharing in the cloud. Our work is also distinct in that it utilizes context extraction to work with any data file, rather than being limited to a predefined set of profile items as on social networks.

Privacy in Cloud Storage

The concerns about the privacy of user data in the cloud were confirmed by Ion et al. [ISK11] through surveys highlighting the users' beliefs about the intrinsic insecurity of the cloud. They discovered that users are less interested in issues like data deletion, country of storage, and storage outsourcing. On the other hand, they tend to trust their local storage for sensitive data rather than the cloud. Users were also found to hold false assumptions, such as that the service provider is liable in case of data loss or that it does not have the right to access or modify their data. Users were still interested in better security options and are willing to pay for systems providing more privacy. These findings provide a strong motivation for systems like C3P as it can help reduce users' concerns and expand the user base of cloud storage services.

Several systems have been proposed earlier for controlling the privacy of online information. Könings et al., developed PrivacyJudge, a system for giving the users control on who has access to their information and how long it is kept online [KPSW11]. They combine cryptographic approaches for enforcing access control with privacy labels to convey how the data should be treated. A client-centric system was also developed by Ion et al. [IBC⁺13], enabling users to share text and image files on web-based platforms while keeping the data encrypted and hiding the fact that confidential data has been exchanged from a casual observer.

From one angle, our work is complementary to theirs as we design C3P to help users decide on what data to keep confidential. From another perspective, our work is distinct as we allow multiple levels of protection that can be controlled by the user.

Crowdsourcing Privacy

The work by Garg et al. [GPKC13] highlights the *peer produced privacy* paradigm, which treats privacy as a community good and considers individuals who share the risk of information sharing. The authors argue that such an approach can result in more socially optimal privacy decisions. Agarwal and Hall [AH13] have used crowdsourcing in their work on the *Protect-MyPrivacy* project to recommend adequate permissions on iOS. One of their interesting findings is that as few as 1% of the users, classified as experts, make enough decisions that can support the crowdsourced privacy recommendation system. Our work shares similar motivations to these, among which are the suboptimal privacy decisions taken by average users and the inability of users to keep track of the changing contextual factors affecting privacy.

2.8 Summary

In this chapter, we have depicted C3P, a new framework for preserving the privacy of data shared to the cloud. The core premise of C3P is that, by providing the users fine-grained estimations of their files' sensitivity, we open the door for automatically applying fine-grained privacy protection mechanisms. We strived to do that with a system that continuously learns from its users, with minimal extra data required. A natural extension of this work is investigating the best way to handle these automated policy recommendations mechanisms and to balance the privacy provided with the services intended by the user.

Communicating the Risk Part II

3 A Primer on Cloud Apps Privacy

3.1 Overview

The popularity of consumer cloud storage providers (CSPs) over the previous decade has been steadily rising. Dropbox, Google Drive, and One Drive have each amassed hundreds of millions of users [Oom17]. In order to further appeal to their users, the CSPs have been transitioning from being pure *service providers* to becoming *app ecosystems*. Hence, they now offer APIs for developers to import and process users' files stored in the cloud.

Consider, for example, a web app called PandaDoc¹, which allows creating, editing, and signing documents online. When a user uses PandaDoc from her laptop browser, she can import files stored in her Google Drive instead of her hard drive. Such a pattern is increasingly more prevalent with the growing number of 3rd Party Cloud apps (or 3PC apps) that are tightly integrated with cloud storage services. Dropbox alone claims that hundreds of thousands of apps² have been integrated with its platform.

Moreover, we are witnessing the rise of *cloud-first* devices, such as Chromebooks. With 2 million devices sold in Q1 of 2016, Chromebooks have already outsold Apple's range of Macs for the first time in the US market³. In cloud-first devices, users install 3rd party apps from a web store. Such apps are hence designed to easily integrate with cloud storage services.

Even in the enterprise setting, 3rd party cloud apps are increasingly popular. This is first because companies are officially adopting the likes of *Dropbox Business*, *OneDrive for Business*, and *Google Drive for Work*. Second, it is due to employees utilizing their personal cloud accounts to share company's files (resulting in the rise of *Shadow IT*). Various reports from cloud application security providers state that organizations use from 10 to 20 times more cloud apps than their IT department thinks [Sky16, Ela16].

This chapter serves as an overview of the ecosystem of 3PC apps, along with the associated

¹<https://pandadoc.com>

²<https://www.dropbox.com/business/app-integrations>

³<http://www.theverge.com/2016/5/19/11711714/chromebooks-outsold-macs-us-idc-figures>

privacy issues. The reader will find this common ground useful for the upcoming two chapters, where we will dive more into our efforts at mitigating the risk communication gap.

3.2 Privacy Issues in Third Party Cloud Apps

3.2.1 The Threat

While the CSPs themselves are few in number, and, at least, have clearly defined privacy policies, users are now faced with a new kind of privacy adversary: the 3rd party app vendors. With every app authorization decision that users make, they are trusting a new vendor with their data and increasing the potential attack surface. Elastica, the cloud application security provider, estimates that the average financial impact on a company as a result of a cloud-storage data breach is \$13.85M, including remediation costs [Ela15]. In 2015, the data breach at Anthem, a US insurance company, has reportedly cost more than \$100M, with 80M unencrypted health records leaked. This was a result of an exfiltration exploit leveraging a popular public cloud storage application [Ela16]. Even on the personal level, the risk extends from breaches exposing financial information and health records to unnoticeable, continuous profiling based on stored files.

3.2.2 Over-privileged and Full-access Apps

As observed in other 3rd party apps ecosystems, users often end up exposing more data than is needed to unaccountable apps. For instance, a user's favorite PDF converter is highly likely to get access to her music library and discover her taste in Mozart or obtain her geo-tagged photos and know where she went on the weekend. Throughout this thesis, we refer to such apps as *over-privileged apps*. Giving such over-privileged apps superfluous access can potentially result in users' data being abused. This has recently been the case in the health apps market where the top 20 most visited apps were found to be sharing users' data with 70 analytics and advertising companies [SD13]. It is also the case that a considerable percentage of 3PC apps request full access to users' cloud files. Even if these apps are not over-privileged, they still place the users' data in the hands of more and more parties, thus increasing the likelihood of data leaks. In this thesis, we will be providing usable privacy solutions to curtail the users' adoption of over-privileged apps (Chapter 4) and to hamper the unnecessary spread of users' data to more parties (Chapter 5).

3.2.3 Unique Properties of the Cloud Ecosystem

Given that these problems have been also manifested in other ecosystems, such as mobile or social networking apps, one would definitely ask whether the cloud apps ecosystem has some unique features that warrant a particular study like ours. This is indeed the case.

Unlike studies on mobile app ecosystems, where the permissions concern the user's list of

contacts, current location, or photos, the cloud permissions allow the 3rd party apps to get access to any file the user has stored in the cloud. Thus, instead of profiling the current user context, such apps can get far-reaching insights inferred from her documents, concerning her financial, legal, or health-related outlook for example.

This can also be done without the user noticing. While data processing on smartphones can be detected from excessive battery usage for example, 3PC apps can perform the analysis completely on the server side once they get access to the users' cloud files, without consuming resources from the users' devices. Put simply, the scale and the quality of data that can be collected is both a privacy nightmare for unaware users and a goldmine for advertisers.

Finally, collecting the permissions of cloud apps at scale is challenging. Unlike other ecosystems where app permissions of thousands of apps can be easily mined via traditional web crawling, each 3rd party cloud app has a unique interface that links to the service providers. Hence, this limits the corpus of apps that one can study. Still, while our contributions are shown in this context of 3PC apps, our various techniques, especially on the level of risk communication, can be easily transplanted to other ecosystems.

3.3 Third-party Cloud Apps Ecosystem

3.3.1 System Model

There are four main entities that interact in the third-party cloud app system:

1. a *user* u who uses that app for achieving a certain service
2. a *cloud storage provider* (CSP) hosting the user's *data*
3. a *data subject* to whom the files belong and whose privacy is being considered. We further define two levels of data subject granularity:
 - *individual-level granularity*: i.e., the user herself is interested in guarding her own data privacy,
 - *team-level granularity*: i.e., a group of users are interested in guarding the privacy of team-owned data (e.g., using an enterprise version of cloud storage services)
4. a *vendor* v that is responsible for programming and managing a 3PC app. These vendors register their apps with the CSPs. The apps themselves are hosted on any website the vendors choose (i. e., not hosted by the CSP itself).

Each user has access to a set F_u of files stored at the CSP. A subset of these files is owned exclusively by the data subject while the other subset is composed of files that are each shared with at least one other *collaborator*. We denote the set of all collaborators of user u by $C(u)$. For simplicity reasons, we will assume throughout this thesis that the files of all data subjects, as

Notation	Explanation
u	user
v	vendor
$C(u)$	Collaborators of u
F_u	set of files of u
$F_{u,v}$	set of files of u accessible by vendor v

Table 3.1 – Summary of notations used

well as the collaborators for each file, are all fixed from a reference step $t = 0$. Using the CSP's API, the vendor v can get access, at step $t \in \mathbb{N}$, to the subject's data upon *user authorization*, which consists of u accepting a list of *permissions*. We will alternatively refer to this as *app installation*, and we will assume that exactly one app is installed in each step t . Permissions are named differently across various providers, but, in general, we can categorize them into three categories:

- **per-file access**: where the user has to authorize the vendor for each file access individually. This is typically done via a file picker provided by the CSP itself.
- **full-access**: where the vendor gets access to all users' data. In the interface, this is worded, for instance, as “View the files in your Google Drive” or “access to the files and folders in your Dropbox”.
- **per-type access**: where the vendor gets access to all files of a specific type. For example, Dropbox words it as “access to images in your Dropbox”. Some platforms, like Google Drive, do not provide app developers with such fine-grained options.

The authorization can also give v access to files shared with the collaborators of u . Similarly, collaborators of u can install apps that expose files shared with u to new vendors. We denote the set of files of u accessible by vendor v at step t as $F_{u,v}(t)$. Table 3.1 contains a summary of the used notations.

3.3.2 The Case of Google Drive

In the two upcoming chapters, we will be taking Google Drive as a case study, and we anatomize this ecosystem in detail. Nevertheless, the solutions we develop are applicable to other cloud platforms as well. This is because all these platforms have similar interfaces to authorize 3rd party apps and have comparable permissions schemes.

In Google Drive, any developer can register an app that accesses Google Drive API at Google Developers Console for free. She then receives a *Client ID* and *Client Secret* that need to be included in the app code to access Google APIs. The developer can then specify in her code a

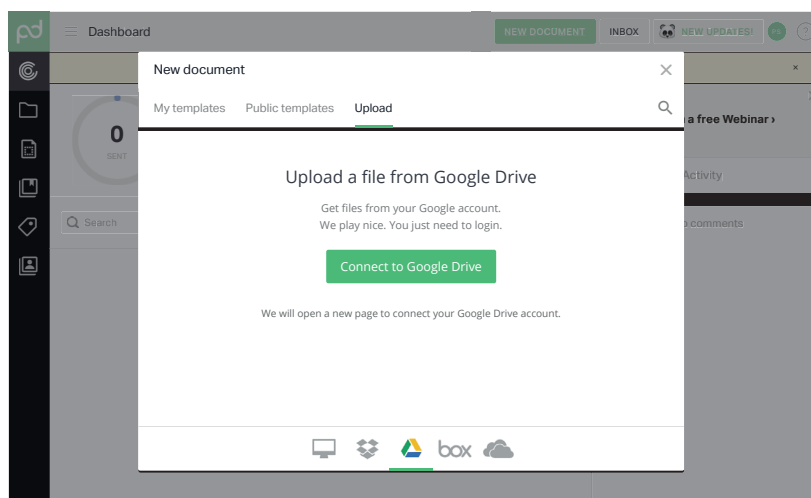


Figure 3.1 – Example of how Pandadoc permissions’ request looks like. Notice that this window only appears after the user decides to upload a new document and chooses the Google Drive icon on the bottom. This interface is app-dependent.

set of Google permissions (a.k.a. scopes) she wants to obtain. The app itself can be hosted on any website the developer chooses; i.e., it is not hosted by Google itself. The developer can also submit a request for featuring the app on Google Chrome Web Store, which has a section for apps that work with Google Drive. In the store, apps are presented along with screenshots and descriptions of their functionality (provided by the developer). The store also allows users to rate and review apps. Apps can be also submitted to other web stores hosted by Google, such as the Add-ons Stores for Google Docs, Google Sheets, or Google Slides and the Google Apps Marketplace for enterprises. However, there are a lot of apps that exist outside these stores too. Unlike the mobile ecosystems where users find it cumbersome and technically difficult to install apps from outside the official stores, the 3PC apps stores’ act as a simple aggregator website that facilitates discoverability.

An app can request permission to access Google Drive data at any time of its operation, and not necessarily at the beginning. For example, the user can be presented with a button in a side menu that reads “Upload a file from Google Drive” (cf. Figure 3.1 for an example). Clicking on that button redirects to a Google-hosted page that presents the set of permissions requested by the app, as shown in Figure 3.2. The user has to accept all these permissions to connect the app to her Google Drive. She cannot select a subset of them at installation time or later. However, she may revoke the app authorization completely from her Google account settings. As we see later, the absence of a standard location and interface for hosting apps and triggering the permissions request is one of the reasons that makes the automated, large-scale privacy analysis of apps infeasible.

The main permissions pertinent to Google Drive are presented in Table 3.2, along with the Google-provided description for each. This short description is also presented to the user, and

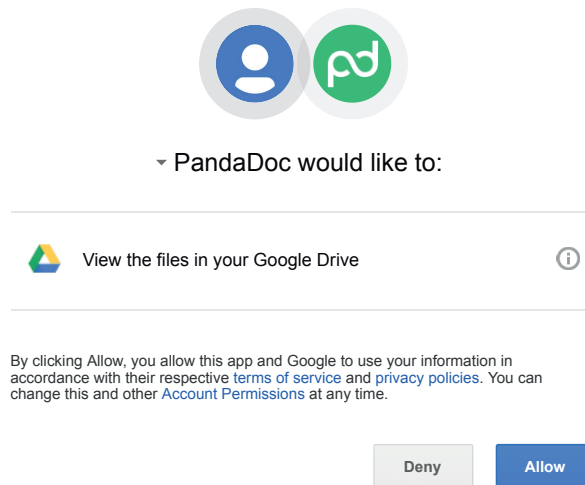


Figure 3.2 – Example of the current permissions interface of Google Drive

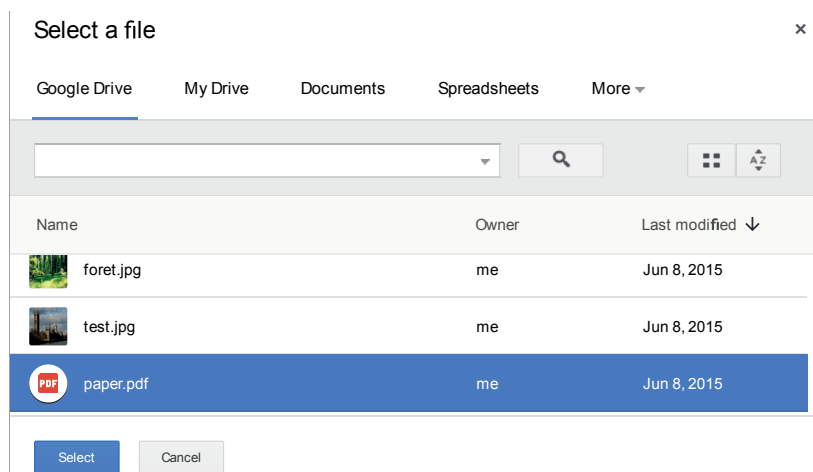


Figure 3.3 – Google Drive file picker interface

a longer explanation is available via clicking the info button (i) next to each permission.

As far as files' data is concerned, an app can request full access (DRIVE and DRIVE_READONLY permissions) or on a per-file basis (DRIVE_FILE). In the case of full access, an app can access any file directly via Google Drive API without the need for user intervention. For example, this type of access enables an app to obtain all the user's files and download them to its server.

When developers request DRIVE_FILE only, the explicit approval for each new file(s) is mediated by an interface provided by Google. For example, the developer presents the user with a Google-hosted file picker popup (Figure 3.3) so that she can select (and thus approve access to) the file. Alternatively, the file can be opened from Google Drive's interface via the "Open with" option in the context menu of the file (cf. Figure 3.4).

Permission	Short Name
View and manage the files in your Google Drive.	DRIVE
View the files in your Google Drive.	DRIVE_READONLY
View metadata for files in your Google Drive.	DRIVE_METADATA_READONLY
View and manage Google Drive files that you have opened or created with this app .	DRIVE_FILE
Add itself to Google Drive.	ADD_DRIVE
View and manage its own configuration data in your Google Drive.	DRIVE_APPDATA
View your Google Drive apps.	DRIVE_APPS_READONLY

Table 3.2 – Requested permissions with the short reference name

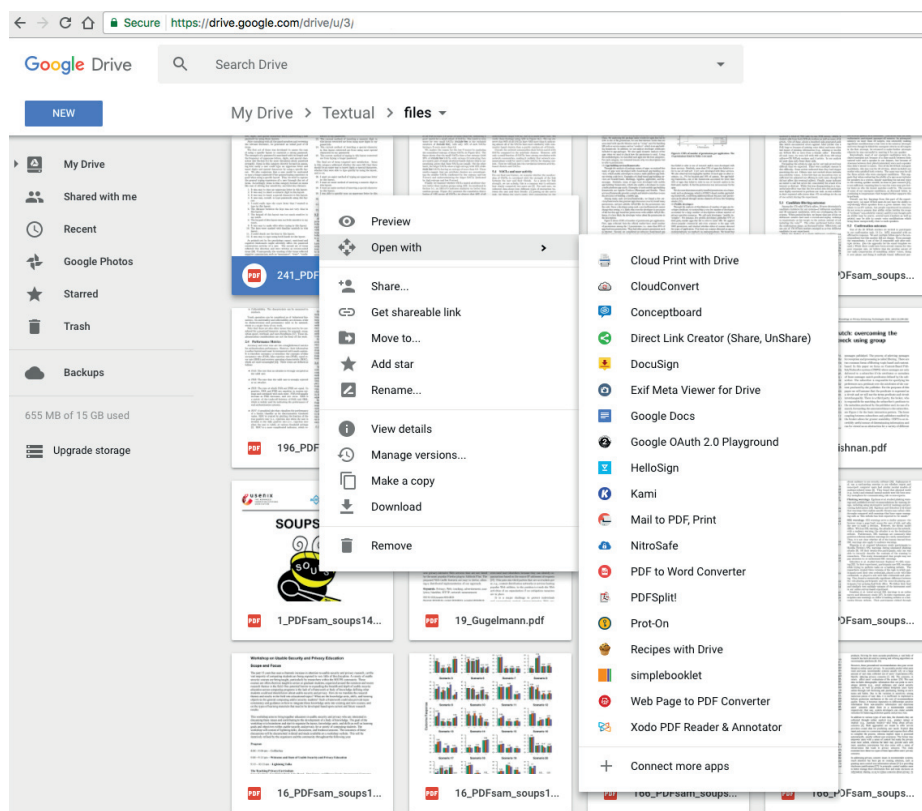


Figure 3.4 – Example of the interface where the app is allowed to access a single file on the Google Drive website itself. The permission requested by apps to appear in this menu is the ADD_DRIVE permission. For accessing the files via this interface, the apps need per-file access (i.e., DRIVE_FILE permission).

The developer can alternatively request access to file metadata only via `DRIVE_METADATA` or `DRIVE_METADATA_READONLY` (allows accessing filenames, editing dates, photos' Exif information, etc.). Additionally, the developer can request access to the list of apps the user has authorized before via `DRIVE_APPS_READONLY`. It is worth noting that the permission list is not limited to Google Drive API and that it typically includes permissions from other Google APIs, such as access to user's profile information, email address, contacts list, etc.

3.4 Summary

Given this overview of the ecosystem of 3rd party cloud apps and the special case of Google Drive, we will delve in the upcoming two chapters into two problems, both revolving around better risk communication: (1) how to deter users from installing over-privileged apps and (2) how to lead them to take better decisions by accounting for previous decisions by them or by their collaborators.

4 PrivySeal: Breaking the Knowledge Imbalance

4.1 Overview

Based on the previous chapter, it has become clear that users sacrifice some of their privacy to get functionalities that third party cloud apps (or 3PC apps) provide. Moreover, it is apparent that such apps might acquire more data than is needed for them to function.

In this chapter, we scrutinize such apps by taking Google Drive's app ecosystem as a case study. Our goal behind this is twofold:

1. First, we aim to explore this platform from a privacy angle, a task that has not been done systematically before. More specifically, we study the ecosystem from the standpoint of all relevant parties, namely, the users, app developers and cloud providers. We seek to characterize (a) the spread of over-privileged apps in the ecosystem, (b) conditions determining developers misbehavior, and (c) the steps cloud providers can take to mitigate users' privacy risks.
2. Second, we leverage app permissions as a medium for experimenting with various models of risk communication. We present three different permission models namely: (a) *Delta Permissions*, (b) *Immediate Insights*, and (c) *Far-reaching Insights*. The first model, i.e., *Delta Permissions*, informs users about the unneeded permissions that over-privileged apps are using. The second model, *Immediate Insights* presents randomly selected examples from the user's data such as portions of text or image files, photo locations, etc., that over-privileged apps can get access to. The third model, *Far-reaching Insights*, has been motivated by the novel concept of *Inverse Privacy* [GHW15]. Inverse privacy refers to the situation when a user is not aware of the information that an external entity has about the user. Based on this definition, Far-reaching Insights communicate to users the inferences which can be made by the apps with superfluous permissions using advanced text and image analysis techniques. These include but are not limited to user collaboration and activity patterns; the top faces, locations, and concepts that appear in users' photos, etc.

Overall, we make the following specific contributions in this chapter:

i. Far-reaching Insights sensitize users with intimate details, and promote privacy-aware behavior: Through extensive user experiments, we discover that our first simple model, *Delta Permissions* fails to deter users from installing over-privileged apps. Put bluntly, *telling users that their privacy is being infringed does not help*. The second model, *Immediate Insights*, does twice as well in discouraging users from installing over-privileged apps. However, the clear winner is our novel model, Far-reaching Insights, which can be twice as effective in deterring users from installing over-privileged apps as Immediate Insights. Overall, our analysis reveals various factors that can deter users from installing over-privileged apps. For instance, we discover that within Far-reaching Insights, *Relational Insights* (that reveal users' relations with other people) reduce by half the installation of over-privileged apps, as compared to *Personal Insights* (that reveal information about the *users themselves*) (Section 4.4).

ii. PrivySeal helps us profile developer behavior and helps users safeguard their data: In an effort to promote privacy-awareness in the general public, and to help users safeguard their privacy, we present PrivySeal, a privacy-focused app store that uses Far-reaching Insights to warn users about over-privileged apps. This store is available for public use and has been used by over 1440 registered by until November 2015. A considerable fraction of these users has prior experience of using Google Drive 3rd party apps. By automatically getting their previously installed apps' metadata, we anatomize current developers' behavior, point towards potential avenues of misbehavior, and present suggestions to deter misbehavior (Sections 4.6 and 4.7).

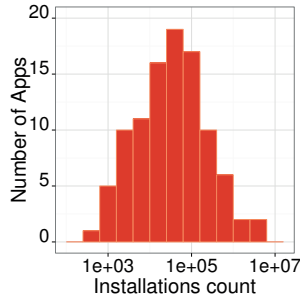
iii. Shared wisdom: Finally, based on our analysis we present several easy to implement practical suggestions that can be adopted by cloud providers and by those others working in the domain of privacy to safeguard users privacy in the cloud (Section 4.7).

The remainder of this chapter is organized as follows. In Section 4.2, we describe in detail our app permissions review process and results. In Section 4.3, we present our three permission models, before evaluating them in Section 4.4. Based on the privacy-focused store we have developed (Section 4.5), we analyze app developer behavior in Section 4.6. Finally, we give our recommendations for the community in Section 4.7.

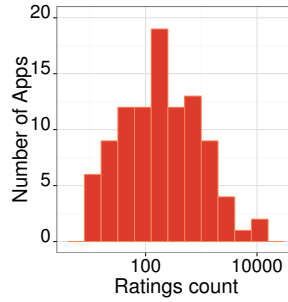
4.2 Privacy Risk of 3rd Party Google Drive Apps

The question that comes next is: “**what is the extent of risk that actual users are exposed to?**” To answer this, we examined a sample of 3rd party Google Drive apps to determine the percentage of apps that request extra permissions¹. We proceeded to Google Chrome Web Store, which has a section for apps that work with Google Drive. The store features apps on its main page, that change with time. At the time of this study, there were around 420 apps on the store that are labeled as “*Works with Google Drive*”. We selected 100 featured apps at

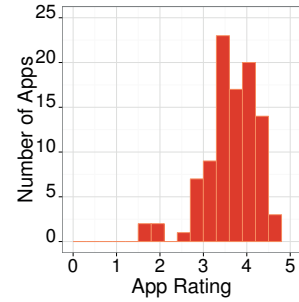
¹We refer the reader to Chapter 3 for a primer on the Google Drive ecosystem and the different permissions.



(a) Distribution of the app installation counts (on a log scale) in the reviewed dataset



(b) Distribution of the app rating counts (on a log scale) in the reviewed dataset



(c) Distribution of the app rating values in the reviewed dataset

random from the main page (during May 2015), and we manually reviewed them one by one. Hence, our sample represents around one-fourth of the whole set of apps in the store, which is one of the main avenues for finding Google Drive apps. As we discuss later in Section 4.6.1, we discovered that, in our real word sample of 1440 users, around one-fourth of the installed apps are from the Google Chrome Store.

Figure 4.1a shows the distribution of the installation counts of apps in our dataset on a log scale, where it is clear that the apps follow closely a normal distribution (this has been individually confirmed using q-q plots). The average number of installations was 194,600 and the median was 29,350. Figure 4.1b shows that the number of ratings follows a similar distribution, with a mean of 736 and a median of 181. The ratings value distribution is shown in Figure 4.1c, with a mean of 3.66 and a median of 3.72. Overall, this shows the diversity of the apps in our APRs dataset and that it represents a wide range of apps.

4.2.1 Permissions Review Process

We now explain the App Permissions Review (APR) methodology we followed, and we refer the reader to Figure 4.2 for the corresponding flowchart. Our methodology is inspired by Google Drive's guide for choosing authentication scopes².

Each APR aims to get: (a) **set P of requested permissions**, (b) **set S of sufficient permissions** for the app functionality. We start each APR by going to the app's website, linked from the store, and testing the app manually. For each app, we first find the step where the app can be connected to Google Drive (if this is not upon the initial sign up). Then, we record the set P of requested permissions and authorize the app to access a test Google Drive account created for this purpose, and we record the permissions requested. If `DRIVE_FILE` (i.e. minimal per-file access) is the only Google Drive permission requested, the app review is complete ($S = \{\text{DRIVE_FILE}\}$). Otherwise, we continue to check the app's interface for all file pickers

²<https://developers.google.com/drive/v3/web/about-auth>

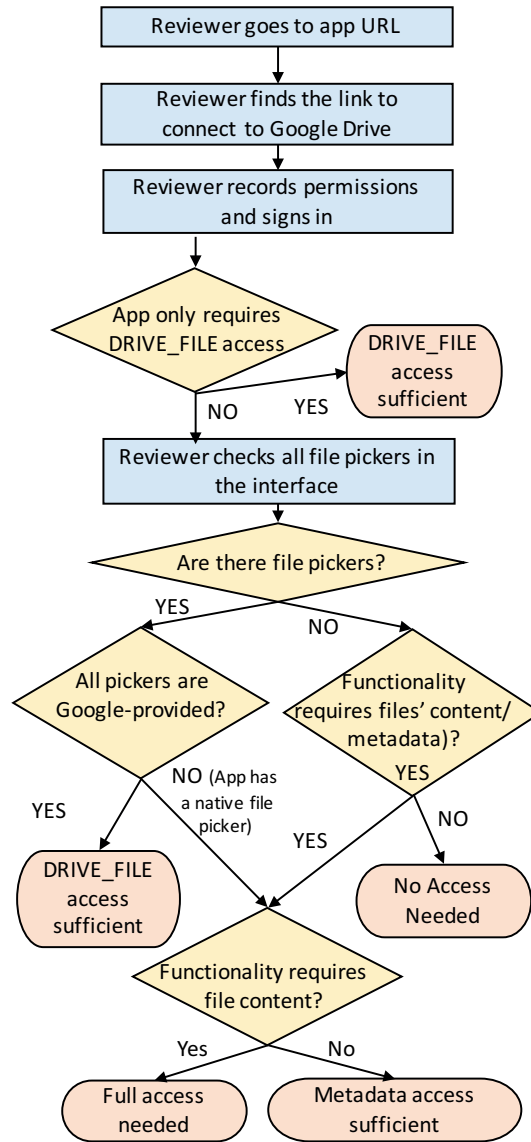


Figure 4.2 – Flowchart of the APR process, inspired by Google Drive guide for choosing authentication scopes

that allow importing files from Google Drive (in almost all the cases, there is at most one file picker).

In the first case where the app solely uses Google’s official file picker of Figure 3.3 (e.g., an app that allows users to convert specific files to PDF format), we set $S = \{\text{DRIVE_FILE}\}$. In the second case where we find that there are no file pickers in the interface and that the app functionality does not require access to any file, the app is labeled as not requiring any file permissions ($S = \{\}$). In the case where the app includes a custom file picker, we decide that (a) $S = \{\text{DRIVE}\}$ if the app’s declared functionality necessitates files’ content (e.g., a photo collage

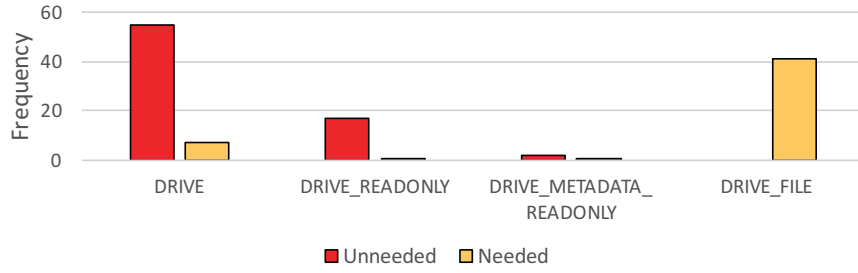


Figure 4.3 – Permissions' usage in APRs

app with custom photos browser) or (b) $S = \{\text{DRIVE_METADATA}\}$ if the functionality does not need the content (e.g., an app that visualizes who has access to a selected folder). Similarly, if the app has no file picker, we decide that (a) $S = \{\text{DRIVE}\}$ if the app's declared functionality necessitates file content (e.g., malware scanning apps for Google Drive that do not need a file picker) or (b) $S = \{\text{DRIVE_METADATA}\}$ if the functionality necessitates file metadata (e.g., an app that visualizes all collaborators with access to user's files). Finally, we label an **app as over-privileged** if either (a) the set S is empty, and P is not or (b) if the set P contains at least one permission that is more demanding than all permissions in S (the permissions in Figure 4.3 are listed from the most demanding on the left to the least demanding on the right). We also determine the set of **unneeded permissions** U , composed of each permission in P that is more demanding than all permissions in S . The set of **needed permissions** is given by $N = P \setminus U$.

4.2.2 Review Results

Analyzing the APRs, we found that **64 out of 100 apps** request unneeded permissions. In other words, the developers could have requested less invasive permissions with the current API provided by Google. In total, 76 out of the 100 apps requested full access to all the files in the user's Google Drive. Moreover, the 64 over-privileged apps have all requested full access. Accordingly, in our sample, around **84% (64/76) of apps requesting full access** are over-privileged.

As shown in Figure 4.3, the top permission that is needlessly requested is the full read and write access to Google Drive (in 55 apps), followed by the full read access permission (in 17 apps). This further increases the magnitude of data that can be exploited with the extra permissions. On the other hand, the per-file access permission is the top permission that is actually needed when requested. This happens in 41 of the apps. However, in **16 of these 41 apps**, we have found that the developer also requested full access to the user's data. Accordingly, developers are sometimes mixing full-access with partial access (which is a subset of the former). We note that such mixing of permissions can either be the result of developer incompetence or aimed at deceiving the user. Regardless, such apps pose a risk which can be potentially exploited.

Another outcome of the APR was that `DRIVE_FILE` was the top alternative permission (in 48 apps) that could replace the unnecessary permissions requested. `DRIVE_METADATA_READONLY` was the alternative for one app only. This indicates that, simply, the correct usage of the current Google Drive API (which does provide per-file access), can eliminate the major part of the privacy risk. Nevertheless, it is evident that developers are generally not following concept of least-privilege (see the later work of Fischer et al., [FBX⁺17] on some of the possible causes for this issue).

4.2.3 Automating the APR Process

Being an external party, we do not have access to the full list of Google Drive apps with their permissions. Hence, the *first* task we had to do was to find the position in each app where Google Drive permissions are requested. This is not always on the main page of the app, and sometimes finding it requires navigating multiple menus and/or pages (cf. Figure 3.1 for an example). Automating this task involves building an advanced web crawler that can retrieve the permissions from a large number of such apps by smartly searching for the sign-in button.

The *second* task was checking the functionality of the app to see if it matches the requested permissions. Automating the process of over-privilege detection or real-time private data leakage detection has been tackled in the mobile apps scenario (e.g. in [FCH⁺11] and [EGC⁺10]). However, in the mobile scenario (or any similar architecture), the user's device hosts the data, the 3rd party apps, and the detection/monitoring solution. Cloud apps present a radically different scenario as the data is hosted by the CSP, the 3rd party app is served at a developer-specified location, and any detection/monitoring app would operate from outside. The only part of the code that the 3rd party app exposes is the client side code. Hence, all techniques that check the app's code (e.g., via static/dynamic analysis) or its inputs/outputs cannot be transplanted to the cloud app case as they would evidently underestimate what APIs/permissions the app might need³.

One automated way we perceive for over-privilege detection is to cluster apps of similar functionality and identify the ones which request more permissions than others in the same cluster. Even then, the data collected manually would be used as the ground truth to evaluate the automated method. Detecting actual data leakage is much more challenging in the cloud apps scenario as the app can send users' data to other parties from the server side (which is impossible to monitor via external solutions).

Faced with these limitations, manual expert reviews are the closest we can get to assessing the apps' needed permissions. Still, we do not claim that this method is perfectly accurate as a developer might be working, for example, on a non-advertised feature that requires new permissions. However, we conjecture that APRs are accurate with the vast majority of the

³In a concurrent work, Fernandes et al., faced similar issues while studying the ecosystem of smart home applications.

reviewed apps⁴.

Finally, as our main purpose in this work is to characterize the ecosystem and suggest alternative permission models, automating both the app permissions collection and the over-privilege detection tasks falls out of the scope of this work. We note though that we are concurrently working on the specific research problem of designing automated APRs.

4.3 New Permission Models

In the light of the risk that over-privileged apps pose, we propose in this section three alternatives to the existing permission model in Google Drive before evaluating their efficacy in mitigating the risk in the next section.

4.3.1 Delta Permissions

Our first model is based on the following hypothesis: *“When users are informed about the unneeded permissions being requested by apps, they are less likely to authorize such apps.”*

Hence, this model replaces the current permissions interface displayed in Figure 4.4 with a new interface, presented in Figure 4.5. We call this permission model *Delta Permissions (DP)*, and it reveals to the user the distinction between permissions that are needed for the app functionality and those others (the delta) that are unnecessarily requested.

4.3.2 Immediate Insights

The second model is based on the following hypothesis: *“When users are shown samples of the data that can be extracted from the unneeded permissions granted to apps, they are less likely to authorize these apps.”* Accordingly, we show users randomly selected data examples, directly extracted from their Google Drive, such as excerpts of text or image files, photo locations, or people she collaborated with. An instance of this model, which we call *Immediate Insights (IM)*, is given in Figure 4.6. On the left, we have the same previous *DP* interface. On the right, we have the *Insights Area*, where we show a question that says: “What do the **unneeded permissions** say about you?”, followed by an answer in the form of a visual with short explanatory text. In this figure, the Insights Area visualizes the location where a randomly chosen user photo was taken. In the following, we describe the design of the *IM* Insights:

Image: We show an image selected at random from the set of user’s image files.

Location: We randomly choose a photo from the set of user’s image files, such that it includes a GPS location in its Exif data. Then we show that photo on a map centered at that location (as

⁴From our experience over one year, rarely did apps introduce new features that required new permissions. Moreover, in Section 4.5, we discuss how to further alleviate the repercussions of inaccuracies in a real-world deployment by allowing developers to submit rebuttals.

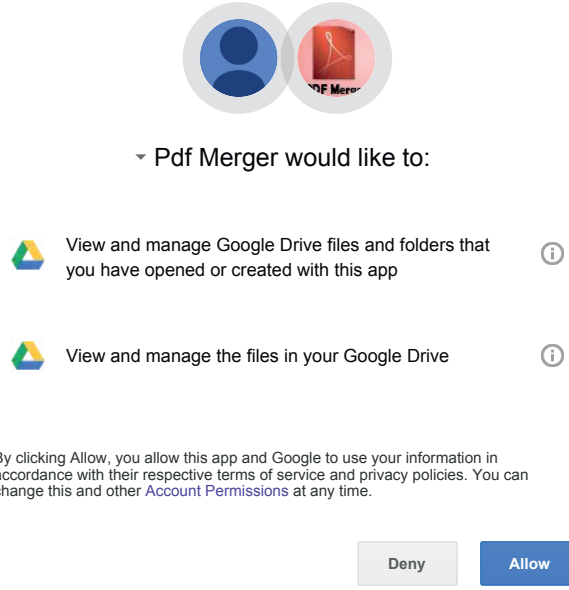


Figure 4.4 – Example of the current permissions interface of Google Drive

in Figure 4.6).

Text: We show the user an excerpt from the beginning of a randomly chosen textual file.

Collaborator: We show the profile picture and the name of a randomly chosen collaborator.

4.3.3 Far-reaching Insights

The third model is based on the following hypothesis: “*When the users are shown the far-reaching information that can be inferred from the unneeded permissions granted to apps, they are less likely to authorize these apps.*” These are insights that go beyond examples and include what can be inferred by running more involved algorithms, such as sentiments towards entities, objects identified in photos, faces detected, etc. Hence, we denote this model by *Far-reaching Insights* (or shortly *FR Insights*). The interface layout is the same as that of Figure 4.6, but with the Insights Area containing an *FR* insight instead of an immediate insight. In this work, we have designed six types of *FR* insights that can be extracted from users’ data. Below, we will provide the details for generating each of these insights. Towards that goal, we highlight two file categories of interest: (1) textual files, such as PDF documents, word-processing documents, spreadsheets, presentations, text files, etc., and (2) image files, such as JPEG, PNG, TIFF, etc. We represent the set of textual files as $TF = TF_1, TF_2, \dots, TF_K$ and the set of image files as IF_1, IF_2, \dots, IF_L .

Entities, Concepts, and Topics (ECT): The first type of insights we form is based on applying various NLP techniques to extract *named Entities (E)*, *Concepts (C)*, and *Topics (T)* from users’ textual files.

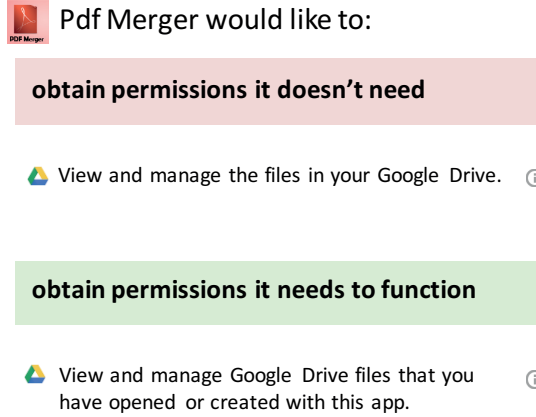


Figure 4.5 – Example of Delta Permissions interface

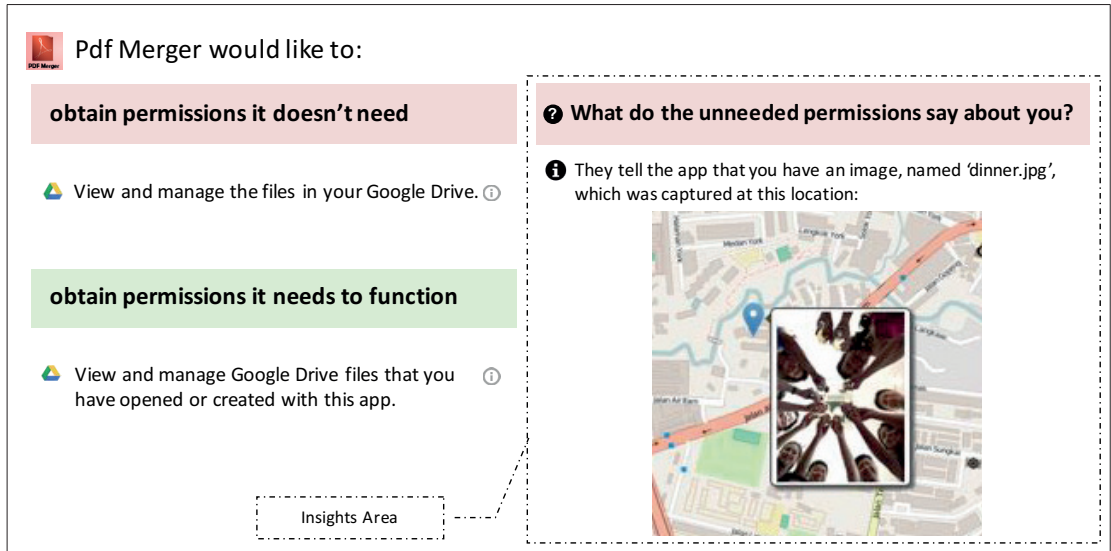


Figure 4.6 – Example of Immediate Insights interface; the same layout is used for Far-reaching Insights, with the insights area content changing accordingly

i. Entities: We get the *top named entities* (e.g., people, places, companies, etc.) present in the user's textual files. Such entities are recognized using Named Entity Recognition (NER), which is a traditional problem in natural language processing that involves locating and classifying elements in text into predefined categories [DBE07]. For this task, we perform text extraction on each file, and we then pass the text to a AlchemyAPI's service. Given the text of file TF_j , this service returns a set of entities, along with the frequency of occurrence $f_{i,j}$ of each entity e_i in TF_j . We normalize this frequency for each entity by dividing it by f_{max_j} , which is the frequency of the most recurrent entity in TF_j :

$$f_{norm_{i,j}} = \frac{f_{i,j}}{f_{max_j}} \quad (4.1)$$

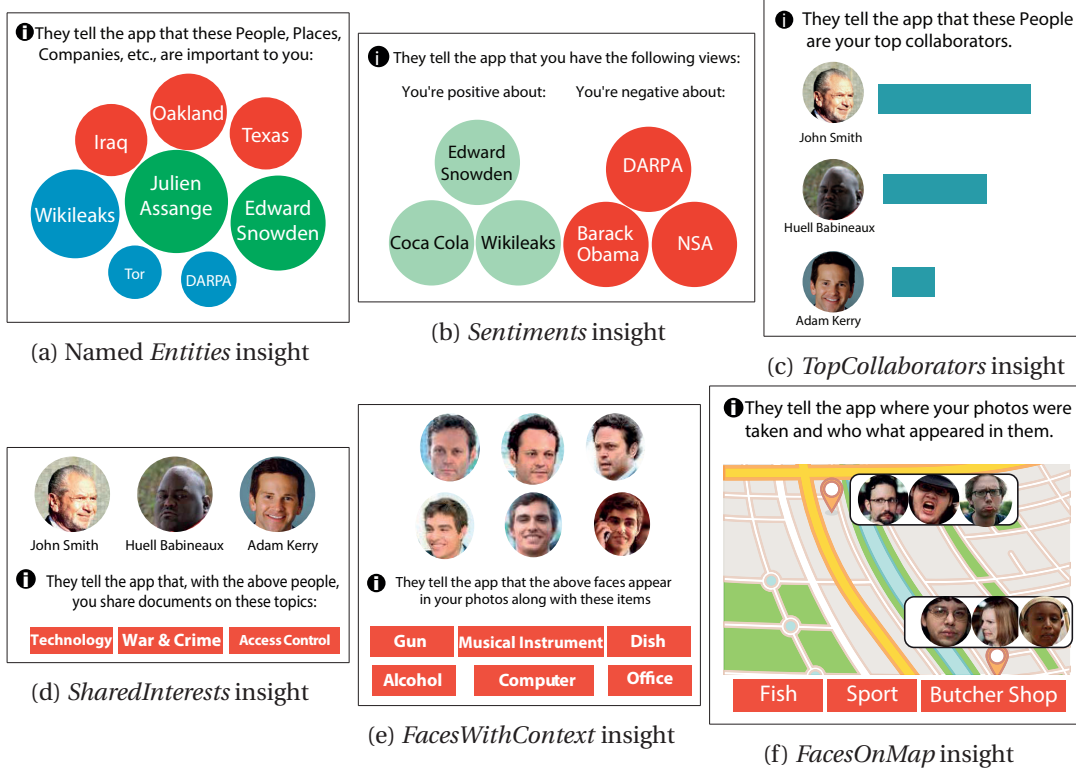


Figure 4.7 – Visualizations for the various Far-reaching Insights

Then, we compute an overall score for entity e_i across all the files in TF , by summing its individual normalized frequencies:

$$score(e_i) = \sum_{j=1}^K f_norm_{i,j} \quad (4.2)$$

As shown in Figure 4.7a, we visualize the entities with the highest scores as a set of circles, each of a diameter proportional to the score of the corresponding entity. Different types of entities (e.g., people, places, etc.) have different circle color.

ii. Concepts: We also extract concept tags from users' documents. These concepts are high-level abstractions, not necessarily mentioned in the text. For example, the sentence "My favorite brands are BMW, Ferrari, and Porsche", would be tagged by the concept "Automotive Industry". AlchemyAPI was again used for this task, returning, for each file TF_j , a set of concepts, each denoted as c_i along with a relevance score $r_{i,j} \in [0, 1]$. We used the following scoring method to rank the concepts across the user's documents:

$$score(c_i) = \sum_{j=1}^K r_{i,j} \quad (4.3)$$

Similar to the case of entities, we represent concepts by circles, each of a diameter proportional to the score of the concept.

iii. Topics: Topics are higher level abstractions, used to classify documents into general categories, such as technology, art, business, etc. We used AlchemyAPI, which returns a maximum of 3 topics per file TF_j (each denoted as t_i), along with a relevance score $r_{i,j} \in [0, 1]$ for each of them. A topic comes in the form of “ $a_1 / a_2 / \dots / a_n$ ”, representing a hierarchy among the labels (e.g., “/hobbies and interests/astrology” or “/finance/investing/venture capital”). To extract the top topics based on a user’s documents, we use the same scoring method as that of concepts:

$$score(t_i) = \sum_{j=1}^K r_{i,j} \quad (4.4)$$

We represent topics by circles, similar to the case of entities, where the diameter of a circle is proportional to the score of the topic. Topics sharing the top level label are colored similarly.

We combine these three together due to their similar goal of profiling users’ interests. When we use the *ECT* insight, one of E , T or C is randomly displayed to the user in the Insights Area.

Sentiments: For each entity that occurs in TF , it is possible also to estimate whether the text relays a positive, neutral, or negative sentiment about that entity. Towards that end, we use the sentiment analysis service of AlchemyAPI. For each TF_i , we select the entities labeled with positive or negative sentiments (each such entity also has a sentiment score $s_{i,j} \in [-1, 1]$ with 1 corresponding to the most positive sentiment and -1 to the most negative one.). We then compute the overall sentiment score s_i of entity e_i across all the user documents TF :

$$s_i = \sum_{j=1}^K s_{i,j} \quad (4.5)$$

The sentiments with the highest positive and negative scores are then shown to the user, as presented in Figure 4.7b.

Top Collaborators: The next insight we added displays the top collaborators a user has, based on the analyzed files (Figure 4.7c). We define collaborators as people who share files with the user, regardless of who initiates the sharing operation. These typically include close work colleagues, intimate friends, or people the user goes out with and shares pictures with afterward. In the interface, this insight is visualized as a horizontal bar chart of the top collaborators with the bar lengths representing the relative frequency of the user’s collaboration with each of them.

Shared Interests: In this insight, we try to represent the user’s mutual topics of interests with a group of people. Towards that end, we perform the following steps:

- We determine the top topics as we have done in the *ECT* insight.
- Then we select from these topics a subset S_t that only includes the ones which appeared in shared files.
- Via Google Drive API, we extract, for each topic t_i , a list $U(t_i)$ of collaborators (based on files it appeared in).
- We select from each $U(t_i)$ the most frequent collaborators (i.e., those appearing in most documents with this topic).

Users then get a visualization similar to Figure 4.7d, where we show the three top topics from S_t along with the top collaborators for these topics.

Faces with Context: We now come to the insights that are based on features inside the user's images. The first insight of this type shows a group of faces, representing the most frequent people appearing in the user's images, alongside the concepts that appear in the same images (see Figure 4.7e). One can imagine that such information might be valuable, for example, to advertisers that aim to extract the user's interests in certain people, products, or services.

To construct this insight, we performed two steps:

i. Face clustering: It is evident that showing the user random faces detected in her photos will not create the same effect as when these faces are actually people she cares about. Our plan to achieve the latter case involves three steps:

- We use a face clustering algorithm in order to group together photos of the same person. As a result, we get a list of groups G , where each group $G_i \in G$ is comprised of the faces that belong to a person identified as p_i . The algorithm used is by Zhu et al., [ZWS11] implemented by the OpenBR framework. [KKK⁺13]
- From each group G_i , we exclude the faces with width (height) less than $\frac{1}{15}$ of the total image width (height).
- We exclude groups with less than 3 faces in total.
- We sort the groups by the number of faces in each of them.

ii. Image concept recognition: In order to identify the concepts inside each photo, we used a classifier from the Caffe library [JSD⁺14]. The classifier uses a pre-built deep learning network, that is based on the architecture used by Krizhevsky et al., [KSH12] that won the Imagenet 2012 contest.

Based on the above, we show the user the top groups (i.e. with most faces) along with the most recurring concepts in these groups (as in Figure 4.7e).

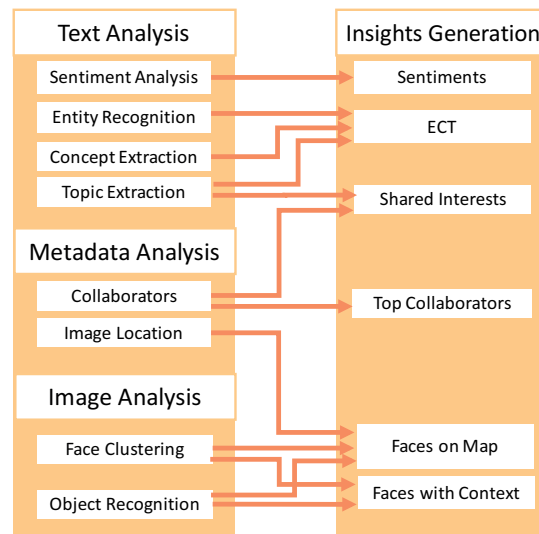


Figure 4.8 – Component diagram mapping the used analysis techniques to the generated insights

Faces on Map: In addition to the image content itself, image metadata can also be sensitive, especially the geographical location where the image is captured. Hence, this insight, shown in Figure 4.7f, consists of showing the faces of people overlaid on a map, centered at the geographical area where these faces appeared. Below the map is a list of the top concepts that appeared in the photos taken in that area. In our actual implementation, the visual is animated, moving between different areas to show the user the places that different photos were taken at. To construct this visualization, we had to cluster the images into different geographical areas. For that, we used the OPTICS algorithm (Ordering Points to Identify the Clustering Structure) by Ankerst et al., [ABKS99]. OPTICS allows finding density-based clusters in spatial data and is tailored for detecting meaningful clusters with data of varying density. After getting the cluster results, the zoom level on the map is animated to show one cluster to the user at a time.

The component diagram of Figure 4.8 summarizes the techniques used for generating each of the *FR* insights.

Further Notes:

We note that the reasoning behind designing lightweight models such as *DP* and *IM* was that we wanted to examine whether designing heavyweight insights such as *FR* Insights is worth the effort for the potential adopters of our approach. If users respond equally favorably (or badly) to both the heavy and the lightweight approaches, the *FR* insights need not be adopted. We also note that, for an app that does not request unneeded permissions (even if it requests full access), the Insights Area will simply show a text saying that the app does not require any extra permissions. We also note that we follow Google Drive’s approach of requesting per-

missions “At Setup” [SBDC15] (i.e., at the first time of app authorization). This is unlike other ecosystems (e.g., iOS or Android M), which require a “Just in Time” approach (i.e., permissions are requested only when the actual functionality is needed). This is because, in Google Drive, many apps are supposed to work with the user’s data even when she is offline. Hence, granting access in an interactive manner for individual permissions is not always feasible.

4.4 Evaluating the Models

4.4.1 Experimental Setup

We designed an experiment with actual users to test the hypothesis of whether the new models can better deter the users from installing over-privileged apps as compared to the existing one and to discover factors that influence users’ decisions.

User Recruitment:

To recruit users, we primarily used our university’s mailing list. The users were briefed about an app that is related to protecting the privacy of their data against 3rd party apps on Google Drive. The news about the app was also reported on the university’s website and was picked up by several technology websites. The website described itself as an app for Google Drive that aims at exposing what 3rd party web apps can needlessly get about users.

Via our website, the users can sign in to their Google account and then grant full Google Drive access to our app. Next to the “sign in” button, we linked to our privacy policy, explaining what data the app gets and what it keeps. Only those users who had at least 10 files containing text or 20 images were allowed to continue. This is to ensure that they possess at least a minimal level of experience with Google Drive. Figure 4.9 shows the density plot of the percentages of users’ analyzed files that are textual. Although there is a significant fraction of users with no image files, there are many users with a balanced fraction of textual and image files. Next, users who agreed to participate in our experiment were randomly assigned to one of the groups described below. As a motivation to complete the experiments, the users were enrolled in a lucky draw, where they could win one of five gift cards to a mobile app store of their choice.

Methodology:

The first goal of the experiment is to investigate the efficacy of the three new permission models by comparing them to the existing Google Drive permission model as well as to each other. The second goal is to perform micro-comparisons among the different types of *IM* and *FR* insights. Accordingly, we went for a mixed between-subject and within-subject design. The reason for not going for a complete between-subject design was the large number of participants needed for statistically significant results with 12 independent groups (for all micro-comparisons). The reason for not going for a complete within-subject design was to

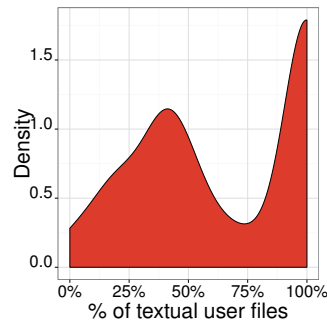


Figure 4.9 – Density plot of the percentage of textual files for our experiment’s users

avoid any participants’ bias that can result from showing the existing interface they are used to and the new interface we developed in the same experiment. Accordingly, we had four groups in our experiment. A user is assigned to one single group, and the only permission interface she sees during the experiment is that of its group. The groups were:

1. **Baseline group (*BL* group):** Users in this group were presented with a clone of the original interface that Google shows upon installing the app (shown in Figure 4.4). This group serves as the control group, and we briefly refer to it as *BL*
2. **Delta Permissions group (*DP* group):** Users in this group were presented with the modified interface, previously shown in Figure 4.5.
3. **Immediate Insights Group (*IM* group):** Users in this group were presented with the modified interface of Figure 4.6, with the Insights area containing one of the *IM* insights of Section 4.3.2.
4. **Far-reaching Insights Group (*FR* group):** Users in this group were presented with the modified interface, of Figure 4.6, with the Insights Area containing one the *FR* insights described in Section 4.3.3.

A user experiment was divided into multiple *tasks*. In each task, the user was requested to select an app with a specified *goal*. For example, the goal would read “Select the app which allows you to extract the ZIP files on your Google Drive”, and the corresponding app would be “ZIP Extractor”. The user would then choose this app among other apps that are listed in the interface. We show this interface in Figure 4.10, and we note that it is similar to the actual Google Chrome Web Store.

Moreover, only one app of those listed satisfies the given goal, and it is highlighted in the interface. This part of the setup only serves a gamification purpose to keep the user interested. Once the user selects the app, she is presented with a permission interface that corresponds to its group (i.e. that of Figure 4.4 for the *BL* group, Figure 4.5 for the *DP* group, and Figure 4.6 with a randomly selected visual for the *IM* or *FR* Insights groups). The user is then presented

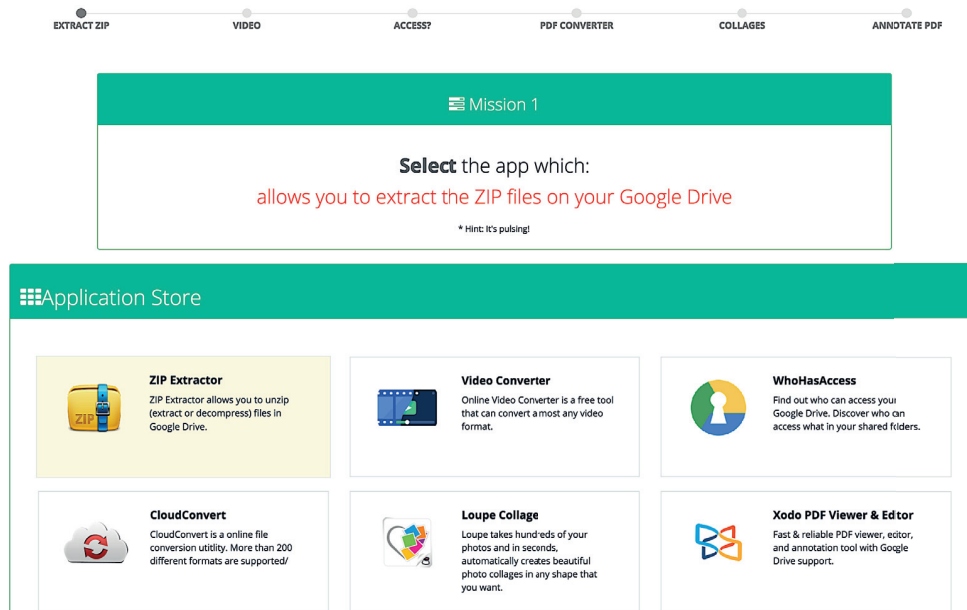


Figure 4.10 – Task interface presented for the users in the experiment, where they had to select the app satisfying the given purpose (already highlighted for them)

with a question that says: “Based on permissions below, would you be likely to install this app?”. She can choose between “Permissions are too invasive” (accept) and “I’m OK with these permissions” (reject). Figures 4.11 and 4.12 show screenshots of the interface for the *FR* and *IM* experimental groups. We worded the question so that we avoid all users rejecting the installations of all apps. We rather aimed that users would reject apps whose permissions they consider as too invasive, thus allowing us to make within-subject comparisons. After answering the question, the user is directed to the next task with another app, until she completes the whole set of tasks.

The apps used in the experiment were obtained from the Google Drive section of the Chrome Web Store. For experimental purposes, we modified the permissions requested by these apps to be able to test various conditions. Unlike in the store, we removed elements such as ratings, user reviews, and screenshots and kept a minimal interface, allowing the users to focus solely on the app permissions. We also avoided using apps from popular vendors to avoid the bias resulting from users being influenced by famous brands. These steps were taken to study the effect that the permission model has on the user’s decisions, without the influence of extraneous factors⁵.

Moreover, the apps were presented to the users in randomized order to compensate for the effects of learning and fatigue. For reference, the permissions that each app requested are presented in Table 4.1. A user assigned to the *BL* or *DP* groups had to install 5 apps in 5 tasks.

⁵Incidentally, the user might confront a scenario exactly as in the experiments if she does not find the app from the store, but lands on a certain site that has the option of authenticating with Google Drive.

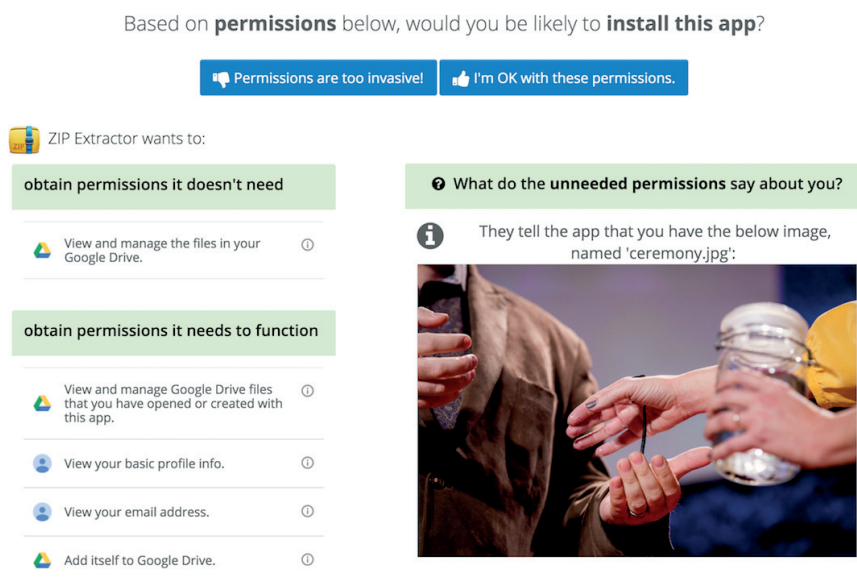


Figure 4.11 – Example of the interface shown to users of the *IM* group, with the decision dialog on top

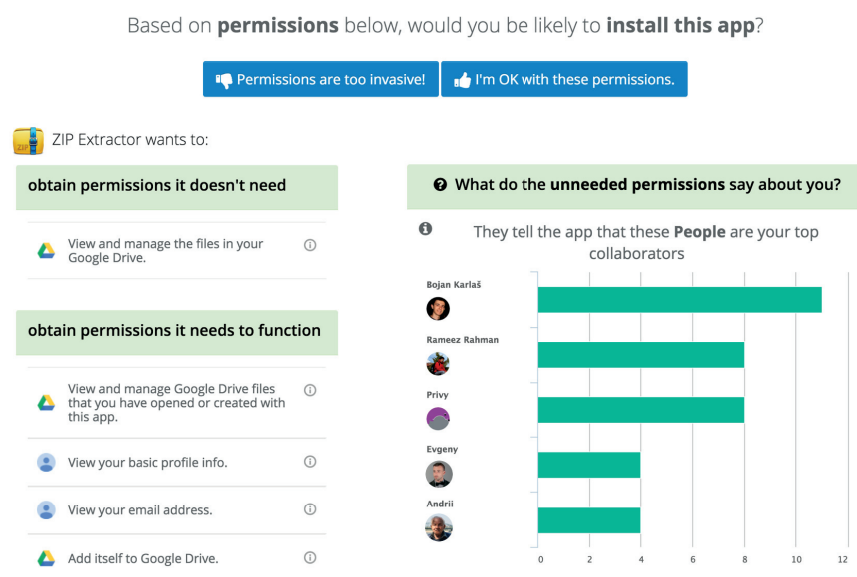


Figure 4.12 – Example of the interface shown to users of the *FR* group, with the decision dialog on top

For the *IM* Insights and *FR* groups, we added additional apps. This is because we wanted to compare the effects of the different kinds of insights. The permissions of the additional apps were fixed to those of (ZIP Extractor), but the insights displayed were changing. For each user, the insights were assigned at random to each of those added apps. In total, users assigned to the *IM* Insights and *FR* Insights groups had to complete 8 and 10 tasks respectively.

App	DRIVE	DRIVE_ METADATA	DRIVE_ FILE	Experimental Group
ZIP Extractor	R,U		R,N	1,2,3,4
Xodo PDF Viewer & Editor		R,U	R,N	1,2,3,4
WhoHasAccess	R,U	R,N		1,2,3,4
Video Converter	R,U			1,2,3,4
Cloud Convert	R,U		NR,N	1,2,3,4
HelloFax	R,U		R,N	3,4
Heap Note	R,U		R,N	3,4
Photo to Cartoon	R,U		R,N	3,4
PDFUnlock	R,U		R,N	4
HelloSign	R,U		R,N	4

Table 4.1 – Permissions of apps in the experiment. *R*: Requested, *N*: Needed, *U*: Unneeded, *NR*: Not Requested

At the end of the experiments, users were presented with a survey, which consisted of a set of multiple choice survey questions, in addition to a free form to provide feedback.

Data Protection and Ethics:

Respecting the user privacy when working with cloud data is of fundamental importance. Our experiments were done according to a code of ethics protecting this privacy. In particular, after generating the insights from a user's files, these files are deleted immediately from our apps' servers. As per our displayed privacy policy, only the insights' data presented to the user is kept in the app database. Moreover, the user is given the option to delete her insights data at any time with a single click in the app's menu. The database dump we ran our analysis on was isolated from the one to which the deployed web server connects. Also, we use the *https* protocol so that users can securely connect to our system. Before data analysis, we anonymized any occurrence of names and emails in the database by applying a one-way MD5 hash on them. At all times, we refrained from manually checking the database for users' insights. All the images used in this chapter are in the public domain, and the insights shown do not belong to real users. For further transparency, all the libraries and frameworks used for building the tool and data analysis were listed and linked to from the main page of the website. Although an IRB review was not performed beforehand, this work was subsequently reviewed by our university's IRB, which did not object to publishing the results.

4.4.2 Results

We got 210 users in total who successfully completed this part of the experiment. Out of them, 55 were in the *BL* Group, 50 in the *DP* Group, 54 in the *IM* Insights group, and 51 in the *FR* group. We start by interpreting the results of our user experiment and comparing the efficacy of the various permission models. The metric we used in our evaluation is the *Acceptance*

Likelihood AL, defined as:

$$AL = \frac{\#(Accepts)}{\#(Accepts) + \#(Rejects)}, \quad (4.6)$$

where *Accepts* denotes the cases where users were fine with the permissions, and *Rejects* denotes the cases where they found them too invasive. *Accepts* and *Rejects* are aggregated across users and tasks for the permission model under consideration. The lower the *AL*, the better the performance in deterring users from installing over-privileged apps.

In order to compare the effect of different interfaces, we plotted in Figure 4.13 the Acceptance Likelihood for the *BL* and *DP* groups and also for each particular insight of the *IM* and *FR* Insights groups. To evaluate the significance of the *AL* differences among the interface types, we fit a generalized linear mixed model (GLMM) with the user's decision (Accepting/Rejecting the app installation) as the binary response variable and the interface type as the fixed effect. Participants' IDs and apps' names were fitted as random effects to control for the potential between-participants and between-apps variabilities. The model was fit assuming a binomial distribution and a logit link function, using the *glmer* function in the *lme4* package in *R* [BMBW15]. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. The significance of differences among *AL* values was determined using Tukey Honest Significant Difference test with the *glht* function of the *multcomp* package [HBW08]. The difference between the *AL* of any two interfaces in Figure 4.13 is significant if the corresponding row and column intersect at a $p\text{-value} \leq 0.05$ in Figure 4.14.

Inefficacy of Baseline and Delta Permissions: We first found that the Delta Permissions and Baseline approaches performed closely (*AL* of 0.42 and 0.39 respectively) without a statistically significant difference ($p\text{-values} = 0.77$). Hence, we found no evidence of any advantage that the *DP* can introduce, which means that telling our experiment's participants explicitly about unneeded permissions did not help deter them from installing over-privileged apps. We also observe that both these interfaces had a significantly higher *AL* (i.e. $p\text{-values} \leq 0.05$) than all the insights, except for the Collaborator insight. This highlights the fact that showing well-selected insights will result in deterring more users compared to the case of not showing any insights.

The Power of Relational Insights: The next interesting outcome is that there is a category of insights (Category 1) composed of *{Image, Text, ECT, and Sentiments}* that are all associated with a significantly higher acceptance likelihood than the category composed of *{FacesWith-Context, TopCollaborators, and SharedInterests}* (Category 2)⁶. Since this is a very interesting result, we investigate further to analyze the defining characteristics of these two naturally clustered categories. The main feature of Category 1, which includes both *IM* and *FR* insights, is that insights in this category are restricted to characterizing the user *herself*, such as showing text excerpts from her documents, topics appearing in them, or images she has in her files.

⁶The number of users who had location-tagged photos was low; hence, we could not obtain highly significant results in the case of *Location* and *FacesOnMap* insights.

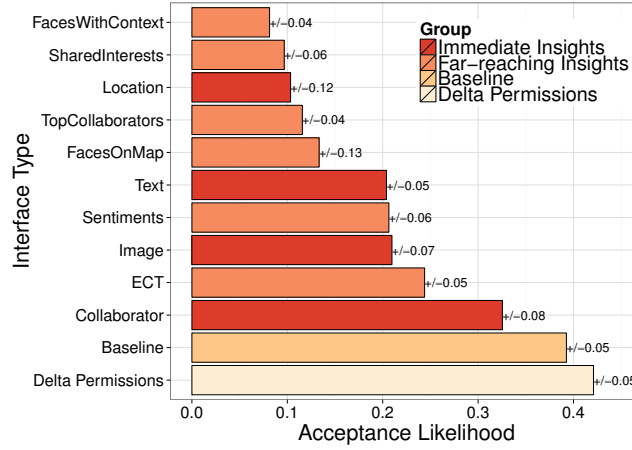


Figure 4.13 – AL for the different types of interfaces; numbers next to each bar are the error values at 95% confidence interval

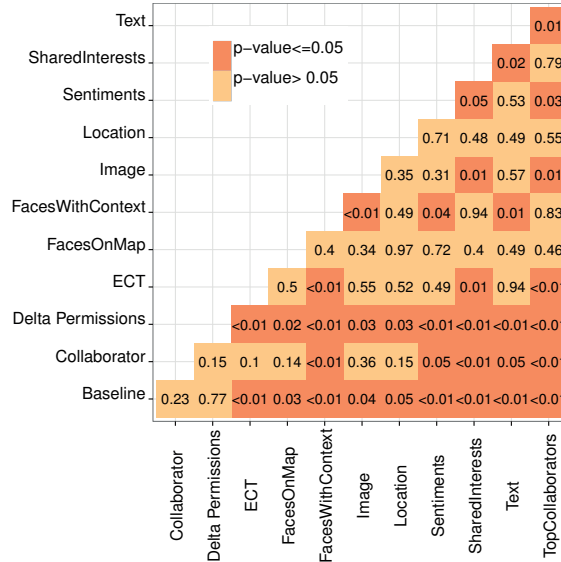


Figure 4.14 – p -values of pairwise tests; if p -value ≤ 0.05 , we consider the visuals on the corresponding row and column as different; the difference direction is obtained from Figure 4.13

Hence, we denote this category as *Personal Insights*. On the other hand, the defining feature of Category 2 insights, which are all Far-reaching, is that they extend to characterizing the relationships of the user *with other people*. For instance, *FacesWithContext* shows the most prominent faces in user's photos along with the items appearing with them. *SharedInterests* shows the people who collaborate with the user and the type of topics they share. Also, *TopCollaborators* identifies the most frequent people the user interacts with. We denote these

as *Relational Insights*. From our results, we can conclude that Relational Insights promote greater privacy awareness in users, as such insights are more likely to dissuade them from installing over-privileged apps.

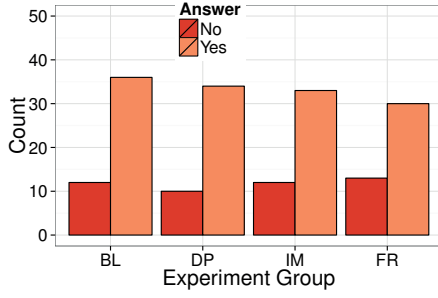
Impact of Face Recognition: Delving deeper into more results brought forth by the comparison of different insights, one can notice that showing examples of user's images ($AL = 0.21$) is significantly less deterring than showing the important faces and listing the concepts in the image ($AL = 0.08$) with pairwise comparison $p\text{-value} < 0.01$. This highlights the fact that users are sensitive towards the output of face detection and object recognition in photos. Given that services such as Google Photos, OneDrive, and Flickr already apply such techniques to facilitate search, the above result highlights that they can also be used by these companies to easily implement solutions such as ours for raising users' privacy awareness when sharing data.

Influence of High-Level Textual Insights: Contrary to the case of images, in the case of textual documents, showing the high-level entities or concepts extracted from the text does not seem to have a significant difference as compared to simply showing direct excerpts from the text ($p\text{-value} = 0.94$). Only when the relationship factor is introduced does the AL significantly decrease (as in the case of *SharedInterests*).

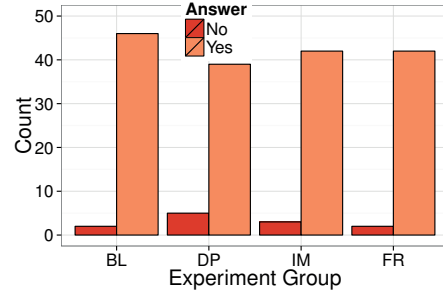
Superiority of Far-reaching Insights: By aggregating the results over all the experiments with *FR* Insights, we obtained a lower AL value compared to *IM* Insights ($AL = 0.161$ and 0.226 respectively). To check the statistical significance of this difference, we followed the previous methodology and fit a GLMM model, but with the fixed effect being the experimental group instead of the specific interface. We confirmed that the AL difference is significant with a pairwise comparison $p\text{-value} = 0.004$. We also noticed from Figure 4.14 that the best Far-reaching insight, *FacesWithContext* ($AL = 0.081$), performed more than twice better than the best Immediate Insight, *Text* insight ($AL = 0.206$) (ignoring the insights where the difference is not statistically significant). Overall, these results demonstrate the superiority of our novel approach of *FR* Insights. Nevertheless, *IM* Insights are still significantly better than the *BL* and *DP* models. This goes in line with the findings of [HHWS14], which showed the goodness of an approach similar to Immediate Insights in the case of Android permissions, even though they did not have Delta Permissions as a building block.

4.4.3 Survey

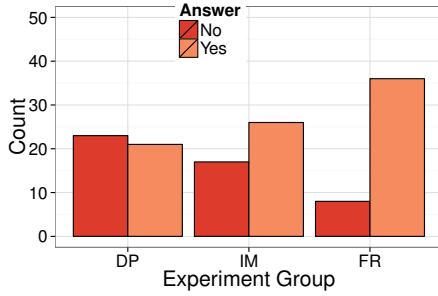
In the following, we discuss the most important results based on users' answers. Figure 4.15a shows that although the majority of users understand what the text of the different Google permissions means, at least one-fourth of users expressed that they do not fully understand these permissions. Figure 4.15b allowed us to verify whether the experimental permission interfaces were intuitive to the users. More than 90% of users answered affirmatively, indicating that our experiments' interface was user-friendly. Figure 4.15c showed that the users in the *FR* group were the ones that expressed the most surprise at what the apps can know about



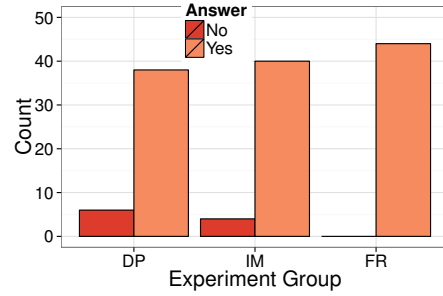
(a) Q: I understand what the different Google permissions mean.



(b) Q: I found the interface in these missions intuitive.



(c) Q: I was surprised that apps know more about me than I expected.



(d) Q: I would be interested in such a Store before installing real apps.

them, which is justified given the low Acceptance Likelihood in this group. Finally, more than 90% of users (and 100% of the *FR* group) expressed interest in using a similar interface to the one they saw in the experiment (Figure 4.15d). Overall, the survey results were in line with the experimental findings. Furthermore, surveyed users expressed the interest in “*adding recommendations for whether one should install 3rd party apps*”, in “*implementing similar functionalities in the Google Play Store and iOS App Store*”, and in “*highlighting apps that actually misbehave rather than only the over-privileged ones*”. These ideas can potentially be realized in future works.

4.4.4 Limitations

First, our design of the experiment abstracted several other factors that users take into account when installing apps. The interplay between ratings, app’s brand, and permissions has been studied before ([KCS13] and [HHWS14]), and it might be worth revisiting in a future work in the context of our new permission models. Second, our experiment’s advertisement included a mention of privacy as we wanted participants to focus on the app permissions. Evidently, this might have made participants more alert towards this issue. Both these points can imply that the real values of the *AL* might be different in reality, where privacy might not be the main factor. Nevertheless, even if the absolute values of *AL* have been impacted, the relative advantages of new permission models still hold. Moreover, we also note that the users in the *FR* and *IM* groups had to do more tasks than the *BL* and *DP* groups, which

might have resulted in more user fatigue and habituation in the *FR* and *IM* groups. This was counteracted first via task randomization at design time and second by the very nature of insights that change at every step. For further validation, we computed the *AL* values of Figure 4.13, considering only the first 5 tasks each user performed. We did not see any major deviation from the results with all tasks included. Finally, our user recruitment strategy was primarily targeted towards our university’s network, and our study was only for English speakers. It would be interesting to see how the results compare in a more general sample (linguistically, demographically and geographically).

4.5 PrivySeal: A Privacy-Focused App Store

Driven by the magnitude of the risk posed by over-privileged apps in Google Drive, we were motivated to bring the advantages of the Far-reaching Insights interface to the user community of this platform. One approach towards achieving that would be for Google itself to implement a scheme similar to ours and to integrate it within the app authorization process. However, we decided not to wait and chose an alternative approach, which is independent of the company’s plans and is ready for user utilization immediately. We built PrivySeal, a privacy-focused store for Google Drive apps, which is readily available at <https://privyseal.epfl.ch>.

PrivySeal allows users to navigate a list of apps, click on those of interest, and check whether they are over-privileged via our *FR* Insights interface. Users can also search by keyword for apps, specifying criteria such as the app being least-privileged. The component diagram for PrivySeal is shown in Figure 4.16. Similar to the APRs we conducted, we have included a “Review Wizard” inside PrivySeal for indicating the requested, needed and unneeded permissions along with the alternative permissions the developer could have used. This responsibility is currently given to a small set of expert developers and is moderated by the store administrators. Developers who would like to object to existing APRs of their apps can submit rebuttals.

Currently, PrivySeal has more than 100 apps and 1440 registered users, with a geometric mean of around 50 new users per month (whose vast majority is signing up out of interest in the app after reading article(s) about it). We finally note that PrivySeal gets access, as is the case with other apps, to users’ data to generate insights. Hence, users are assumed to trust the provider of such a “Privacy-as-a-Service” solution. However, this assumption of trust will hold if a solution such as PrivySeal is hosted by the CSP itself (which already possesses the data), or an enterprise protecting its documents from 3rd party apps. The assumption of trust is also valid if the users choose to trust a *single* entity (such as PrivySeal) to protect themselves from *multiple* other unaccountable over-privileged entities that they would otherwise be forced to trust.

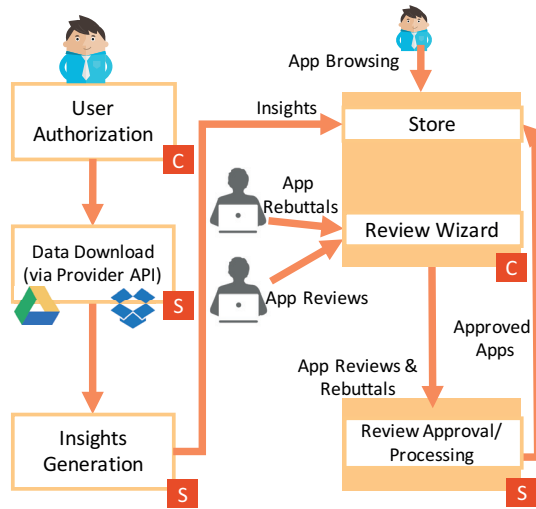


Figure 4.16 – Component diagram of PrivySeal (components labeled by S are server-side and by C are client-side)

4.6 Anatomizing Developers' Behavior

After studying the users' privacy decisions, we now move to investigate the developers' landscape, building on the apps data that our store's users have installed. In total, we obtained data from 1440 registered users of our privacy store.

4.6.1 Current Developer Behavior

The “DRIVE_APPS_READONLY” permission requested by our app allowed us to get the list of apps previously installed by users, along with the information that Google Drive API gives about the apps⁷. We found 662 unique apps installed by users in our dataset. For each app, we obtained the following:

- i. *Access Level*: which indicates whether the app had *Partial Access* or *Full Access* to the user's drive *on authorization time*. Since an app can change the permissions it requests from future users, our dataset had instances of the same app installed with different access levels by different users. We denote such apps as having an access level of *Both*.
- ii. *App Location*: which indicates whether the app is (1) on *Google Chrome Web Store*, (2) on Google's *Other Web Stores* (namely the Add-ons Stores and the Google Apps Marketplace for enterprises), or (3) *Outside Web Stores* of Google. This categorization is inferred by following the *productUrl* field present in the app information, which either leads to one of the stores or is absent.

Figure 4.17 shows how the apps in our dataset were distributed over the different locations and

⁷For details, we refer the reader to: <https://developers.google.com/drive/v2/reference/apps>

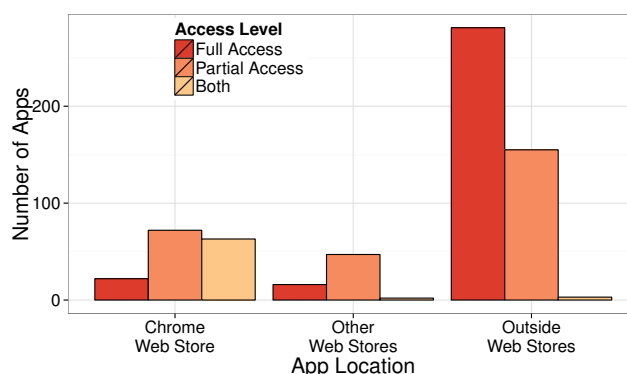


Figure 4.17 – Change of access level with app location

the number of apps requesting the different access levels. From this figure, one can observe the following:

Developers Changing Behavior: The first surprising outcome from this dataset is that around 40% of apps on Chrome Web Store (63 apps) had *Both* as access level, signifying that many developers have changed the requested permissions at least once. To check the current access levels of these apps, we reviewed them one-by-one. We discovered that 59 of these apps (i.e., 94%) have changed from requesting *Partial Access* in the past, to requesting *Full Access* currently. Hence, we can deduce that when developers change the access level, *there is a high probability that it is associated with getting more data instead of the other way round*. Highlighting this change of access level on installation time can further serve for more informed user decisions.

Developer Deterrence through Official Stores: Apps outside the Web Stores requested *Full Access* almost twice as much as they requested *Partial Access* (281 full vs. 155 partial). This was not the case in the Chrome Web Store, where we observe only a slightly higher number of apps with *Full Access* (81 full vs. 76 partial - counting apps that fall under the *Both* access level but which currently request *Full Access*). So we can see that developers with apps outside the Web Stores are more prone to asking for *Full Access*. This can be explained by the conjecture that the store acts as a medium where the apps receive more exposure. Hence, *developers there are likely to be under the pressure of being evaluated through reviews and ratings, and thus tend to avoid abusing the permissions*, while developers outside Web Stores are under no such pressure. Although *Full Access* does not necessarily mean that apps are over-privileged, our APRs have shown that 84% of apps that request *Full Access* are over-privileged apps. In the case of the Other Web Stores, the number of apps that requested *Partial Access* is around thrice the number that requested *Full Access* (16 full vs. 47 partial). This is mainly because that these Add-ons apps are generally expected to provide functionality for Google Docs (or other native Google file types), so deviating from this and requesting permissions for all Google Drive files will be easily detected by the community. Similarly, the community of enterprises, which is highly sensitive towards privacy will deter developers from requesting *Full Access* in

the Google Apps Marketplace.

Deterring Developers in the Wild: The majority of apps in our sample do not come from any Google Web Store (24% from Google Chrome Store, 10% from the Other Web Stores, and 66% from outside the Web Stores). This is also the case for 75% of the apps requesting full data access. These can be apps on other platforms, such as mobile platforms, for example, where there are other types of application stores. These apps can also be ones that are not present in any store but still have Google Drive integration. Hence, we can infer that improving the Chrome Web Store privacy indicators might not be a sufficient solution for deterring the majority of developers. *There is a need for alternative solutions, focused on Google Drive permissions in specific, and independent of the various stores.*

4.6.2 Potential Developer Misbehavior

Although it is clear that full access to users' data can expose various far-reaching insights about the user, it is not completely apparent what seemingly benign permissions, such as metadata-only access can reveal about the user. In the previous section, we have shown that the *TopCollaborators* insight, which can be extracted just from the metadata, has resulted in an Acceptance Likelihood of 0.13, which is around three times lower than what the current Google permission scheme (Baseline) attains. Hence, users are remarkably deterred by seeing what they expose when they give access to their file metadata. Therefore, it is worthwhile to explore the potential of information leakage through metadata-only access. In this section, we show that metadata-only access on its own can allow developers to gather deeper insights about the user's topics and concepts of interest. This calls for extending the *FR* Insights with such information to better inform the user about the potential risk of giving unneeded access to metadata. Towards that, we analyze and compare insights inferred from users' file metadata to what can be inferred from file contents (the data).

Upon user u signing up to our app, the following operations are executed as part of the data analysis:

- i For each analyzed file, the filename is processed by removing its extension and replacing punctuation marks by spaces.
- ii The names of all the analyzed files are grouped into a comma-separated list $L_{FN}(u)$
- iii Topic and concept analysis are applied to $L_{FN}(u)$. The service used for text analysis allowed us to extract topics in the form " $a_1/a_2/\dots/a_n$ ", representing a hierarchy among the labels (e.g., "/law, government and politics/espionage and intelligence/surveillance" or "/finance/investing/venture capital"). In this section, we differentiate between *General Topics* where we only consider a_1 , and *Specific Topics*, where we consider a_n . Accordingly, *General Topics* would indicate user's interest in *law, government and politics* or *finance* for example while *Specific Topics* could indicate the user's interest in

surveillance or *venture capital*. At the end of this step, we filter the results to restrict our analysis of metadata to a maximum of 3 *General Topics*, 3 *Specific Topics*, and 5 *Concepts* for each user's list $L_{FN}(u)$.

iv. From the user's files' contents, we extract the top 5 *General Topics*, top 10 *Specific Topics*, and top 20 *Concepts*. These choices are motivated by the general observation that one's *Concepts* of interest are usually more in number than the *Specific* abstract topics one cares about, which are in turn more than the *General Topics* of interest.

For each user u , we compared the list $D(u)$ of labels (i.e., concepts/topics) extracted from the files' contents with the list $M(u)$ of labels extracted from the list $L_{FN}(u)$ of filenames. We selected precision as the evaluation metric as we are mainly interested in determining whether labels extracted from metadata serve as a good approximation of labels extracted from the data. Inspired by the multi-label classification literature [TKV10], we computed precision using the micro-averaging method, i.e., directly across all labels. A label occurrence is considered as true positive if it belongs to $M(u) \cap D(u)$ and a false positive if it belongs to $M(u) \setminus D(u)$. $tp(l)$ is the number of true positives for a label l , and $fp(l)$ is that of false positives, both across all users. Let LT also be the set of all labels found across user's data and metadata. The overall precision is thus given by the following equation:

$$P_{micro} = \frac{\sum_{l \in LT}^N tp(l)}{\sum_{l \in LT}^N (tp(l) + fp(l))} \quad (4.7)$$

We used this method instead of macro-averaging (i.e., computing the precision per label and then taking the average) because we are interested in estimating the users' interests more than the ability to predict each and every label. For this experiment, we only considered who signed in to our app and had at least 10 textual files with associated concepts/topics. Hence, our sample contained 200 users. Interestingly, the results for *General Topics* indicate that 0.78 of the metadata labels across users match with their top 5 topics of interest. In the case of *Specific Topics*, on average, nearly two of the three extracted metadata labels also appear in the top 10 *Specific Topics* extracted from data ($P_{micro} = 0.61$). Finally, the fraction of metadata *Concepts* that also appear in the data is around one-third ($P_{micro} = 0.31$). However, this does not necessarily imply that the other two-thirds of concepts appearing in the metadata are not relevant to the user. In fact, we have noticed that a lot of these metadata labels are semantically similar to those in the data.

In sum, we have observed that metadata on its own can be considerably accurate in revealing part of users' interests. It can be easily abused by sophisticated adversaries who conceal their misbehavior through only requesting seemingly benign permissions (for metadata access). Therefore, this calls for extending the *FR* insights in the case of metadata-only access to match the developer's potential. For instance, *SharedInterests*, which was shown earlier to convey inferences from content, can also be used as an insight based on the collaborators and the potential mutual topics inferred exclusively from the files' metadata.

4.7 Recommended Best Practices

In addition to PrivySeal, there are several steps that can serve to mitigate the potential of misbehavior in Google Drive and similar services. These solutions serve to help the user both before and after installing the apps.

Fine-Grained APIs: The availability of finer grained permissions (such as access to a specific file type) evidently reduces the amount of data in the hands of the developer and is in line with the principle of least privilege. One disadvantage of such detailed permissions is that they become more difficult for users to comprehend in a short amount of time. However, providing developers with the means to request such fine-grained controls should not necessarily result in a more complicated interface. This can be achieved via multi-layered interfaces [SBDC15]. For example, instead of the app indicating that it needs to “View the files in your Google Drive”, it can indicate that it needs to “View files of specific types in your Google Drive”. Users that are interested in knowing these file types can then click on an additional button (such as the info button ⓘ in the current interface of Figure 4.4).

Transparency Dashboard: A post-installation technique which can potentially deter developers from actually abusing the users’ data is for the cloud platform to provide what we call “Transparency Dashboards”. These dashboards allow the user to see which files have been downloaded by each 3rd party app and when such operations took place. Such a monitoring solution for all apps can only be achieved by the platform itself.

Insights Based on Used Data: Unlike external solutions (e.g., ours) that can only determine what data can be *potentially* accessed, the cloud platform can provide users with insights based on the data that developers have *previously* downloaded. Such an interface will help users better pinpoint adversarial apps that needlessly retrieve files outside the scope of their functionality.

A Privacy Preserving API Layer: It is not uncommon nowadays to find APIs that work as an additional layer on top of one or more existing cloud APIs (e.g., Cloud Elements Documents Hub). Hence, one solution to build a privacy-preserving API is to create it as a layer on top of one or multiple existing platforms’ APIs. This new API can provide finer-grained access control, allow permissions reviews from the community, and implement transparency dashboards. By building this layer on top of existing cloud APIs that already offer various services, one can circumvent the problem of attracting developers who might otherwise loathe using a solution that only serves to protect privacy.

4.8 Related Work

4.8.1 Privacy in Other App Ecosystems

To our knowledge, this work was the first that studied the problem of user privacy in the context of 3rd party apps on top of cloud storage providers. In the case of other ecosystems, there are related works that have studied the current state of privacy notices (e.g. [CYA12, HMSW13, FGW11, PXY⁺13]).

Felt et al., [FGW11] performed a study on permissions in the Chrome Extensions and Android ecosystems. They found that only 14 out of 1000 extensions request the most dangerous permissions and that the average Android app asks for fewer than 4 of the 56 dangerous permissions. Interestingly, they discovered that users are shown at least one dangerous permission during installation of almost each extension and Android app. Moreover, 5 out of 50 extensions they inspected manually were over-privileged. In another work, Felt et al., built *Stowaway*, an automated tool for over-privilege detection in compiled Android apps [FCH⁺11]. *Stowaway* works by comparing the API calls used to the permissions requested. One-third of the 940 analyzed apps were shown to be over-privileged. One of the justifications proposed was the insufficient documentation available for developers, which leads them to request a few extra permissions. In this dataset we studied in this work, we have seen that 3rd party Google Drive apps are, on average, almost twice as likely to be over-privileged. Moreover, the implications of the extra permissions are mostly full-access to the users' files, which makes the problem highly significant.

Chia et al., conducted a large-scale analysis of Facebook apps, Chrome extensions, and Android apps to study the effectiveness of user-consent permissions systems [CYA12]. They observed that the community ratings are not reliable indicators of app privacy in these ecosystems and showed evidence of attempts at misleading users into granting permissions via free apps or apps with mature content. Huber et al., developed *AppInspect*, a framework for automating the detection of malpractices in 3rd party apps within Facebook's ecosystem and used network traffic analysis to spot web trackers and identify leaks of sensitive information to other third parties [HMSW13].

On a higher level, the case of 3rd party apps in Google Drive differs from these platforms in that it is not possible to perform large-scale analysis, firstly due to the absence of a standard application format and secondly due to the difficulty of automatically finding the triggering button for permission requests in different apps. Aside from the above, client-side traffic analysis is not sufficient to detect all cloud data leaks as the apps can send data to third parties after it arrives at the server side, to which outsiders do not have access.

4.8.2 Nudging Privacy

Our work on improving the privacy notices can be classified under the general umbrella of *privacy nudges*. In their book [TS08], Thaler and Sunstein use the term *nudge* to signify “any aspect of the choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentive”. In that sense, they defined nudges as interventions that are cheap to avoid, rather than being mandates that force the user to take a specific action. Acquisti later proposed to transplant this approach of *soft paternalism* from behavioral economics to the context of privacy, and termed it as privacy nudges [Acq09].

Although there has been a significant division around the approach of nudging in general ([SW11, Sun12]), debating its relation to freedom of choice and human anxiety, Acquisti et al., argue in their recent survey on privacy and security nudges that every design choice made is itself a nudge [AAB⁺17]. Success stories in A/B testing tell us that minor changes in a certain button or a photo on a website can result in a significant modification in user engagement. Hence, they state that the question is not about the ethics of nudging in the first place, but whether the inevitable nudges are themselves ethical. Another argument they make is that notices should strike a balance between greater transparency and overwhelming information.

In our approach of Far-reaching Insights, we tried to achieve that balance in two ways. First, the insights themselves are aggregated to reveal high-level abstractions from users’ data. Second, we do not pack all the relevant information for a specific insight in the interface at the same time. For example, each time, we display a subset of the most important people and photos in the corresponding insights. We believe that this also serves for reduced user habituation.

There have also been several previous works that also fall under the category of privacy nudges. Particularly relevant is the investigation of improvements to the existing permissions schemes. In the case of the Android ecosystem, Kelley et al., argued that the privacy information should be a part of the app decision-making process and should not be left till after the user makes her decision [KCS13]. Hence, they appended a list of “Privacy Facts” to the app description screen, textually indicating that the app, for example, collects contacts, location, photos, credit card details, etc., and found that it assisted users in choosing apps that request fewer permissions. Harbach et al., proposed to integrate personal examples from the user’s data in the permissions request screen to expose the data apps can get access to [HHWS14]. This involved showing random pictures, call logs, location, and contacts from user’s data that correspond to each permission. They showed that this created more awareness among their study’s participants and instilled a negative effect, making them more alert when installing apps.

Another work of interest was that by Tan et al., who investigated the effect of explanations given by developers when requesting permissions on iOS [TNT⁺14]. They found that the mere presence of the explanation lead to a higher approval rate of the permission and that the *content* of the explanation did not make a noticeable difference. In the context of Facebook,

Wang et al. [WLS⁺13] introduced privacy nudges to aid users while posting statuses to Facebook through showing random profile pictures of friends who can see the post, introducing a time interval before the actual post is sent, or showing the post sentiment. Personalizing warning notices has been studied before in the context of LED signs [WRKS94] and was shown to significantly increase compliance when compared to impersonal signs.

In this work, we go further, and we show that well-crafted visuals showing far-reaching insights extracted from users' data can be more effective than randomly selected data. We also show through pairwise comparisons among the insights themselves that the choice of the displayed insight highly affects the interface's effectiveness, which was not proven in the case of textual explanations for example ([TNT⁺14]). It is also worth mentioning that, in our experiments, the number of users who were involved with their personal accounts in the experiment was more than five times the number of users in [HHWS14] and [KCS13]. Furthermore, we also provide a readily available solution for the public in the form of a privacy-focused app store.

Moreover, our work is in line with the best practices recommended by the recent work of Schaub et al., who developed a design space for privacy notices to assist researchers in increasing the impact of their schemes [SBDC15]. For instance, we implemented the multi-layered notice concept by showing data of textual and visual modalities. We also developed various visuals to ensure that the permissions dialog is *polymorphic*, which was also shown recently to have an effect on reducing the habituation effect in the user's brain [AKJ⁺15]. We further hypothesize that our interfaces go beyond reducing habituation to making the user curious about the content of the Far-reaching insights as these are not readily imagined by the users as is the case of random examples (e.g., [HHWS14]).

4.9 Summary

In this chapter, we characterized the various factors that have an impact on user privacy in the ecosystem of 3rd party apps for the cloud. We considered Google Drive as an example case study and comprehensively anatomized the ecosystem from the viewpoint of users, developers, and the cloud provider. For users, we carefully devised a set of experiments and tested existing and novel risk communication models to analyze the factors that influence users' decisions in app installation. Our results provide interesting insights into how user privacy can be improved and how CSPs can develop better risk indicators. We also presented a privacy-aware store for cloud apps, which already has over 1440 registered users. From our store users and people who took part in our experiments, we had the unique and unprecedented opportunity to first-hand study real users cloud data. Based on this data, we were able to characterize the current behavior of 3rd party app developers and also point out avenues for developer misbehavior. Finally, based on our analysis, we provided several suggestions for CSPs that can help in safeguarding users' privacy and protecting their data from needless leakage and exploitation.

5 A Usability Approach to Interdependent Privacy

5.1 Overview

In the previous chapter, we have shown our approach for better risk communication in the context of *over-privileged* third party apps. However, users' privacy is not solely determined by these apps. For instance, the mere fact that users are granting more apps full access to their cloud accounts is itself increasing the shareholding parties of their files, and subsequently the likelihood of data exposure to unintended parties.

An additional intricacy is that when users grant access to a 3rd party cloud app, they are not only sharing their personal data but also others' data. This is because cloud storage providers are inherently collaborative platforms where users share and cooperate on shared files. Hence, protecting these files is not solely in the hands of the user.

Skyhigh Networks, a Cloud Access Security Broker (CASB), reports that 37.2% of documents (across 23 million users) are shared with at least one other user. In organizations, documents are shared, on average, with accounts from 849 external domains [Sky15]. Moreover, around 23% of cloud documents were found by Elastica (another CASB) to be “broadly shared”, which means that they are shared (a) among all employees, (b) with external partners and clients, or (c) with the public [Ela16]. Interestingly, 12% of those documents contained compliance-related or confidential data.

This further highlights what has been termed as the *interdependent privacy problem* [BC13], where the decisions of friends can affect the user's privacy and vice-versa. This concept was initially proposed in the context of third-party social networking apps, such as Facebook. However, while 1.92% of Facebook apps request friends' personal information, this is much more pronounced in 3rd party cloud apps, where all apps accessing one's files get access to the part which is shared too. Moreover, unlike Facebook apps, due to the collaborative nature of cloud apps, the CSPs do not provide an option for users to control whether their collaborators' apps can get access to data they own.

Research Questions

In this chapter, we aim to lead the users to reduce the total exposure of their data to third party shareholders (i.e., apps' vendors) by better communicating to them the existence of those shareholders. We go beyond over-privileged apps, and we tackle this issue in all apps that are granted blanket access to users' files (i.e., excluding the per-file access apps where the user gives consent for each sharing operation). We are driven by the rationale that users will inevitably continue to install apps to achieve various services. Hence, instead of stopping them, we aim to lead them to select apps from vendors in a way that minimizes their privacy risk.

We achieve this by leading users to take what we term as *History-based decisions*. Such decisions account for the vendors who previously obtained access to the user's data, whether directly (with her consent) or via her collaborators. Our strategy consists of introducing privacy indicators to the current permissions interfaces that help users minimize the number of vendors with access to their data. Our "usable privacy" approach is guided via a data-driven study and is evaluated via a data-driven simulation.

In essence, we tackle the following research questions:

- From a practical perspective, are the collaborators' decisions significant enough to be accounted for in users' app adoption decisions?
- Do users already account for entities with access to their data? If not, to what extent can the usage of privacy indicators lead to users taking History-based decisions?
- How significant is the effect of adopting these privacy indicators in the case of large networks of users and teams?

Contributions

Towards addressing these questions, we continue to take Google Drive as our case study. We make the following contributions:

- In Section 5.3, we analyze a real-world dataset of Google Drive users, and we show that the median privacy loss that collaborators cause by installing apps can be much higher than that inflicted by the user's own app adoption decisions (39% higher with 5% of shared files and 523% higher with 60% of shared files). To our knowledge, this is the first usage of a real-world dataset to give a concrete evaluation of interdependent privacy in any ecosystem.
- Driven by the significant impact of collaborators, we design new privacy indicators for helping users mitigate the privacy risk via History-based decisions (cf. Section 5.4). We assess these indicators via a web experiment with 141 users. We show that they significantly increase the likelihood that users choose the option with minimal privacy loss, even if not

all of these users are motivated by privacy per se. To the best of our knowledge, this is also the first work to investigate a usable-privacy approach to mitigating the problem of interdependent privacy. The few studies on this issue have mainly approached it from a theoretical perspective, such as developing game-theoretic or economic models [BC13, PG14] or from a behavioral perspective, such as studying the factors affecting real users' monetary valuation of others' privacy [PG15, PG16].

- We explore the potential of History-based decisions by performing a simulation on two large user networks. We show that the network-effects of our approach result in curtailing the growth privacy loss by 70% in a synthetic Google Drive-based collaboration network and by 40% in a real author collaboration network. We also simulate the effect of such decisions in a teams' network. We demonstrate that teams can reduce the privacy loss by up to 45% by solely accounting for team members' decisions (cf. Section 5.5).

5.2 Models and Preliminaries

In Chapter 3, we gave an overview of the various interacting parties in the ecosystem of third party cloud apps: the user, the cloud storage provider (CSP), the data subject whose privacy is being considered (i.e., the individual or the team), and the third party cloud app vendor. In this section, we further add the specific details related to the problem we tackle in this chapter.

5.2.1 User Model

A user is further assumed to be *self-interested*, i. e., only caring about optimizing the privacy of the data subject (a.k.a., privacy egoist), and *non-cooperative*, i. e., does not coordinate her decisions with others. We do not assume that the risks of installing each app are known to the users or calculated a priori. In fact, unlike other 3rd party app ecosystems, the risk of each cloud app cannot be automatically estimated based on techniques such as taint tracking [EGC⁺10] or code analysis [FCH⁺11] because the main app's functionality is typically implemented on the server side (which cannot be accessed by external entities). Such assumptions constitute the *worst case* in the scenarios we consider, and further privacy optimizations can be obtained by relaxing them.

We also assume that the mental model for privacy-concerned users matches the possible permission granularities they are given. Accordingly, privacy-concerned users can have one of the following privacy-goal granularities¹:

- **per-type privacy goal**: where users aim to optimize the data subject's privacy independently for different file types. For example, in an ecosystem like Dropbox, where per-type access

¹We list here the mental models that match the status of app permissions nowadays. It is possible to add other variants, such as caring for the privacy of a certain set of files for example, but this is not achievable with the existing models. We also note that per-file access already achieves the least privilege possible; hence we do not consider a corresponding privacy mental model.

is an option, users might follow the separation-of-concerns principle. Hence, they might install photo-related apps from a set of vendors that is different from the set authorized for document processing.

- **all-files privacy goal:** where users aim to reduce the privacy risk for their entire set of files. This can be in the case of ecosystems which do not have the option of per-type access, like Google Drive. It can also be the case that a user of Dropbox has this goal in mind despite being presented with finer-grained app permissions.

5.2.2 Threat Model

We consider the 3rd party app vendors as the adversary (and not the CSP). The privacy indicator we introduce is best implemented by the CSP, which already has access to the users' and collaborators' data. Alternatively, this can be a feature within Cloud Access Security Brokers (e.g., SkyHigh Networks, Netskope, etc.), which are already trusted by thousands of enterprises to protect their cloud data against other 3rd parties. Moreover, we consider the protection against over-privileged apps as an orthogonal problem, which we have considered in Chapter 4. We rather focus on the interdependent privacy problem, which covers all vendors with full access and is an issue in least-privileged apps too.

Furthermore, in this chapter, we will focus on content-related permissions. In the case of Google Drive, there are two such two access levels: (1) full access, which can be achieved with the `DRIVE_READONLY` or `DRIVE` permissions and (2) per-file access via the `DRIVE_FILE` permission. Google Drive does not offer the per-type permissions option.

5.2.3 Privacy Loss Metrics

In order to quantify the privacy loss that a user incurs with time, we introduce now the *Vendors File Coverage (VFC)* metric². Consider a user u and a set V of vendors at a certain time step. For notation simplicity, we will omit the time step henceforth. $VFC_u(V)$ is computed as the summation of the files' fractions shared with each of these vendors:

$$VFC_u(V) = \sum_{v \in V} \frac{|F_{u,v}|}{|F_u|} \quad (5.1)$$

Intuitively, $VFC_u(V)$ increases as vendors in V get access to more files of u . It has the range $[0, |V|]$.³

If we consider the set V_u of vendors explicitly authorized by user u , we can define the *Self-*

²Table 5.1 summarizes the notation employed in this chapter

³We do not normalize $VFC_u(V)$ by $|V|$. Our rationale is that multiple vendors with access to all the user's files induce a higher privacy loss than one vendor with such access.

Notation	Explanation
u	user
v	vendor
$C(u)$	Collaborators of u
V_u	set of vendors authorized by u
$V_{C(u)}$	set of vendors authorized by collaborators of u
$VFC_u(V)$	file coverage due to the vendors in set V
F_u	set of files of u
$F_{u,v}$	set of files of u accessible by vendor v

Table 5.1 – Summary of the notations used

Vendors File Coverage as:

$$\text{Self-}VFC_u = VFC_u(V_u) \quad (5.2)$$

Similarly, if we consider the set $V_{C(u)}$ of vendors authorized by the collaborators $C(u)$ of u , we can define the *Collaborators-Vendors File Coverage* as:

$$\text{Collaborators-}VFC_u = VFC_u(V_{C(u)}) \quad (5.3)$$

Finally, the *Aggregate VFC_u* for a user u is that due to all vendors authorized by u or its collaborators:

$$\text{Aggregate-}VFC_u = VFC_u(V_u \cup V_{C(u)}) \quad (5.4)$$

Throughout this work, we will use the terms *privacy loss* and *VFC* interchangeably. As will become evident in Section 5.4, this metric choice allows relaying a message that is simple enough for users to grasp, yet powerful enough to capture a significant part of the privacy loss. Obviously, one can resort to a deeper inspection of content or metadata sensitivity (as we did in Chapter 2) had the purpose been finding the best privacy model in general. However, our rationale to not do so is the following: for instigating a behavioral change, telling users that a company has 30% of their files is more concrete than a black-box description informing them that the calculated loss is 30% and constitutes less information-overload than presenting them with detailed loss metrics.

5.3 Collaborators' Impact

At this point, we are in a position to handle the first research question on the extent of collaborators' contribution to a user's privacy loss. Hence, we want to test the following hypothesis:

H1: The collaborators' app adoption decisions have a significant impact on the user's privacy loss.

If this hypothesis is valid in practice, it provides a strong motivation for designing privacy notices that aid users in accounting for their collaborators' decisions, which is what we will study in Section 5.4. Towards that, we will be dissecting the privacy loss, quantified by *VFC*, that users incur in a realistic 3rd party cloud apps dataset.

5.3.1 Dataset

One of the main challenges when studying the privacy loss in 3rd party cloud apps is the absence of public datasets with realistic file distributions, collaborator distributions, sharing patterns, 3rd party app installations, etc. We benefit in this section from an anonymized dataset that we have constructed via the PrivySeal⁴ service in Chapter 4. We build our analysis on it to evaluate the *VFC* of users in a realistic context.

Out of the database of registered users in PrivySeal, we selected those which had a minimum of $N_{files_min} = 10$ files in total, at least $P_{min_shared} = 5\%$ of files that are shared, and a minimum of $N_{apps_min} = 1$ third party app installed. The dataset, henceforth referred to as the *PrivySeal Dataset*, was anonymized and contained metadata-only information. It consisted of a subset of the files' metadata of 183 PrivySeal users in addition to the Google Drive apps installed by those users prior to authorizing PrivySeal's app (the `DRIVE_APPS_READONLY` permission was requested by PrivySeal). The dataset specifically contained:

- list of user IDs (anonymized via a one-way hash function);
- IDs of files in each user's Google Drive,
- list of anonymized collaborators' IDs for each file ID;
- list of apps with full access installed by each user;
- the vendor of each app.

In total, the number of users in addition to collaborators was 3422. Overall, these users had installed 131 distinct Google Drive apps from 99 distinct vendors. Figure 5.1 characterizes the PrivySeal Dataset. Particularly, it displays 4 distributions in this dataset, which realistically model the system under study:

- number of files per user, which follows a skewed distribution with a median of 67 files
- sharing pattern: percentage of shared files out of all user files, which also follows a skewed distribution with a median around 18%

⁴<https://privyseal.epfl.ch>

- number of collaborators across all user files (a.k.a., the degree of the user node in the collaboration network): where 75% of the users had less than 23 collaborators
- number of vendors authorized per user: also follows a skewed distribution with a median of 1 vendor per user

5.3.2 Results

We computed the *Self-VFC*, the *Collaborators-VFC*, and the *Aggregate-VFC* (as defined in Section 5.2.3) for users in the PrivySeal Dataset⁵. As we did not have the actual number of apps for each collaborator of users in the dataset, we assigned to these collaborators a set of apps from a random user of the dataset. We show in Figure 5.2 how these metrics evolve as we gradually consider populations that collaborate more frequently. With $P_{min_shared} = 5\%$, we had a median of 1.39 for *Collaborators-VFC*, which was 39% higher than a median of 1.00 for *Self-VFC*. The significance of the median difference is evidenced by the non-overlapping box-plot notches. This difference became much larger when we considered users that share more files. We had a 100% median difference at $P_{min_shared} = 10\%$ and 523% median difference at $P_{min_shared} = 60\%$. Such results indicate that:

- The collaborators' app adoption decisions contribute a core component to the user's privacy loss, thus confirming our hypothesis *H1*.
- The higher the number of collaborators is, the higher the magnitude of loss these collaborators can potentially inflict.

Both conclusions motivate the need for taking collaborators' decisions into account when designing privacy indicators for cloud apps, which is what we will embark on next.

⁵To avoid double counting, we considered the vendors authorized by both the user and her collaborators in computing *Self-VFC* but not in computing *Collaborators-VFC*.

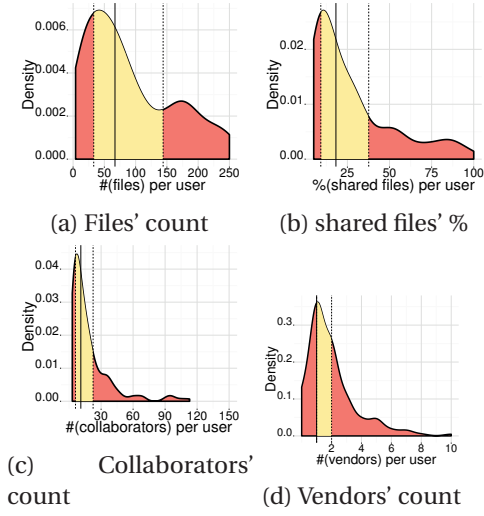


Figure 5.1 – Density plots for various parameters, computed per user ($P_{shared_min} = 5$). Median line is shown, and the light orange area represents the range between the 25% to 75% quantiles.

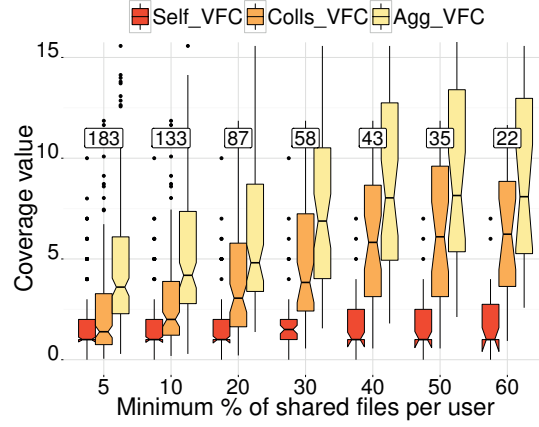


Figure 5.2 – Evolution of metrics with populations that share more files ($N_{files_min} = 10$, $N_{apps_min} = 1$). The numeric labels denote the corresponding number of users in the dataset.

5.4 User Study

Up till now, we have confirmed that, if users want to minimize their privacy loss, they are better off not ignoring the app installation decisions of collaborators. In this section, we tackle the next research question, where we investigate the potential of privacy indicators in leading users to minimize their exposure to 3PC app vendors. We show first our design methodology for the privacy indicators, and we follow that by a web experiment that investigates the efficacy of these indicators in realistic scenarios.

5.4.1 History-based Privacy Indicators

We call our proposed privacy indicators “*History-based Insights*” (*HB Insights*) as they allow users to account for the previous decisions taken by them or by their collaborators. We continue to consider Google Drive as a case study, and we show this indicator in the context of Google Drive apps’ permissions in Figure 5.3b. Compared to the current interface provided by Google (Figure 5.3a), we added a new part to highlight the percentage of user files readily accessible by the vendor (computed based on $VFC_u(\{v\})$ for each vendor v).

As we prove in the next section, selecting the vendor that already has the largest percentage of user files is the optimal strategy to minimize the privacy loss in our context. We denote this strategy as “*History-based decisions*”. Following the best practices in privacy indicators’ design [SBD15], our indicator was multilayered, with both textual and visual components. The wording of the main textual part was brief and general enough to hold for both the data

percentage exposed by friends and that exposed by the user. We used a percentage value rather than a qualitative measure to facilitate making comparisons among apps based on this value. The visual part showed the percentage as a progress bar with a neutral violet color. The bottom textual part was added in a smaller font to provide further explanation for those interested. We used the term “company” in our interface instead of “vendor” as it is more commonly understood by the general audience.

5.4.2 Proof of Optimal User Strategy

Before proceeding to our user study, we provide in this section a proof of the optimal user strategy for minimizing the privacy risk, given our assumptions. We follow the notation summarized in Table 5.1. Let us consider that each 3PC app vendor has probability p of exposing users’ data. As we do not assume that users are provided with a per-vendor risk estimation utility, we set this probability to be the same for all vendors. In general, at a time t , a user u would have exposed her data to a set V of vendors, such that each vendor v has access to a fraction $f_{u,v}(t) = \frac{|F_{u,v}(t)|}{|F_u|}$ of the files. Without loss of generality, we will consider henceforth that the user has an all-files privacy goal (cf. Section 5.2.1). However, the same reasoning applies in the case of a per-type privacy goal. In that case, we simply replace “files” by “files of a specific type” (e.g. photos, documents). We will also be assuming that the users themselves are the data subjects (i. e., we consider individual-level subjects).

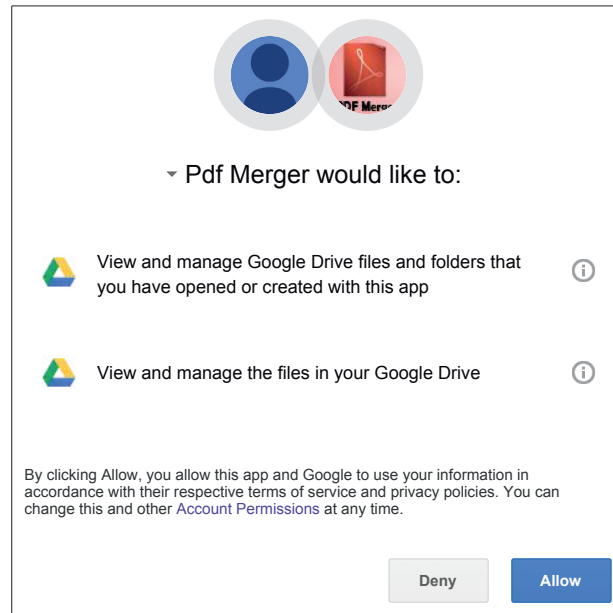
For a vendor v , we quantify the user’s privacy risk magnitude as $p * f_{u,v}(t)$, i. e., the fraction of user files possessed by the vendor multiplied by the probability that the vendor exposes the user’s files. This vendor could have obtained access due to app installations by the user herself or by her collaborators. A user’s privacy risk magnitude at time t can thus be defined as the sum of the risk magnitude across vendors in V : $\text{Risk}(t) = \sum_{v \in V} p * f_{u,v}(t)$.

When a user installs an app from a vendor \hat{v} at time $t + 1$, the vendor gets access to the whole set of user’s files. Hence, the risk magnitude is increased by $p * (1 - f_{u,\hat{v}}(t))$. Given that p is constant, the risk magnitude can be minimized by choosing \hat{v} , such that $\hat{v} = \arg\max_v f_{u,v}(t)$ (which can also be written as $\hat{v} = \arg\max_v VFC_u(\{v\}, t)$). Hence, the optimal, greedy strategy to minimize the risk is to select the vendor that already has the largest fraction of user files, thus minimizing $p * (1 - f_{u,\hat{v}}(t))$. We call this strategy: “History-based decisions”.

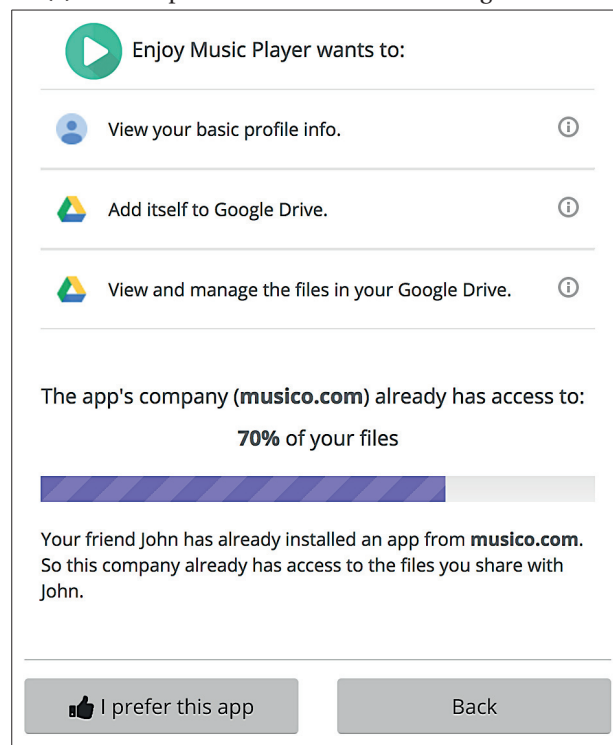
5.4.3 Methodology

To evaluate the new permissions interface, we performed an online web experiment (rather than a lab study) as we were mainly motivated by obtaining a large sample of users that is also geographically and culturally diverse. The hypothesis we wanted to test is:

H2: Introducing the new privacy indicator significantly increases the probability that users take History-based decisions.



(a) Current permissions interface of Google Drive



(b) Proposed “History-based insights” interface, with the buttons from the user study in the bottom

Figure 5.3 – Comparison figures showing the original permissions interface of Google Drive and the “History-based insights” interface

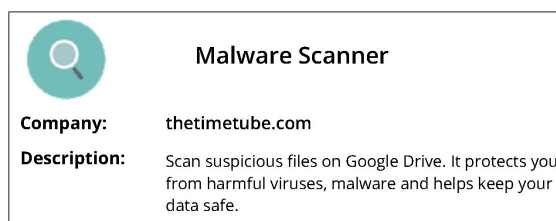


Figure 5.4 – Example app displayed in the list of apps

In addition, the study allowed us to build a realistic user decision model based on the choices taken by participants in different conditions. We will utilize this model in Section 5.5 to simulate the app choices in a large user network and to study the effect on the overall *VFC* in the network. We structured our study to have (1) an Introductory Survey, (2) a series of App Installation Tasks, and (3) a Concluding Survey.

User Recruitment: We recruited users via CrowdFlower’s crowdsourcing platform. In our study, we restricted participation, via the platform’s filtering system, to the highest quality contributors (Performance Level 3). We also geographically targeted countries where English is a main language as our interface was only in English. To further guarantee quality responses, each user was rewarded a small amount of \$0.5 for merely completing the study and an additional amount of \$1.25 that was manually bonused for those who did not enter irrelevant text in the free-text fields.

Instructions: Participants were first presented with introductory instructions that explained the context of the study (i. e., cloud storage services and 3rd party apps that can be connected to them). They were asked to only continue if they had good familiarity with cloud storage services (e. g., Google Drive, Dropbox, etc.). We did not explicitly require that participants have experience with 3rd party cloud apps. However, we educated them about such apps throughout the instructions, particularly showing them two examples of 3rd party apps in action (PandaDoc for signing documents and iLoveIMG for cropping photos). These apps were displayed via animated GIFs that play automatically and do not rely on the user clicking. We used limited deception by neither mentioning the focus of the study on participants’ privacy nor giving hints about selecting apps based on the installation history. The advertised purpose was to “check how people make decisions when they install 3rd party apps.”

Introductory Survey: After checking the instructions, users were presented with an introductory survey, where they first entered general demographic information. This survey was also front-loaded with questions about cloud storage services (several of which required free-text input) to discourage users who had not used these services from continuing to the actual study.

5.4.4 Study Overview

Next, users could proceed to the study page. We used a split-plot design in the study. Participants were randomly assigned to one of two groups:

1. **Baseline Group (BL):** where the permissions interface used is that currently provided by Google Drive (Figure 5.3a).
2. **History-based Group (HB):** where the *History-based Insights* permissions interface (Figure 5.3b) is used.

In each group, the study consisted of 3 modules, which cover the main conditions that can occur when users desire to install a cloud app. On a high level, the modules investigate the following questions:

1. **Module 1:** are users likely to select apps from the same vendor they installed from before?
2. **Module 2:** are users likely to select apps from vendors that her collaborators have used before?
3. **Module 3:** do users consider the differences in access levels obtained by vendors that collaborators installed?

In all modules, whenever the user was asked to *choose* an app, she was presented with a list of 12 apps (Figure 5.4 shows an example app). Only two of these apps were relevant to the task purpose, and they were placed on top of the list (randomly positioned as first or second). With this setup, we wanted to mimic the realistic setup of app browsing while not squandering the user's effort on finding apps. All apps had the same full access permissions too (namely DRIVE permission). Unlike in Chrome Store, we removed elements such as ratings, user reviews, and screenshots and kept a minimal interface. This is all in order to reduce the distractions from factors outside the study. We refer the reader to the work of Kelley et al., [KCS13] who investigated the effects of those elements on users' decisions for Android apps.

To account for fatigue and learning effects, modules 1, 2, and 3 were presented in a random order for users. We piloted our experimental setup in two stages: with colleagues and with online users from the CrowdFlower community itself. For reviewing the online pilot testers work, we embedded a Javascript code for session recording in our study's web page, which allowed us to view the user's mouse and keyboard actions on our side.

Demographics

We had 157 users who completed the study. Based on manually reviewing the users' inputs, we removed 16 users who were inputting irrelevant free-text in the survey in the study. We

Age	18-62	(median 31 years)
Gender	35.5%	Female
	64.5%	Male
Occupation	59.6%	full-time employees
	14.2%	student
	6.4%	part-time worker
	8.5%	self-employed
	5.0%	homemaker
	6.4%	Unemployed/retired
IT Experience	41.8%	Have worked or studied in IT
Degree	19.1%	High school
	7.1%	Trade/tech./vocational training
	51.1%	Associate or Bachelor's degree
	22.7%	Post Graduate Degree
Countries	35.0%	USA
	37.5%	IND
	7.5%	GBR
	6.9%	DEU
	6.9%	CAN
	7.4%	AUS+IRL+ NLD + PAK

Table 5.2 – Demographics in our user study; $N = 141$

thus report the results of 141 users, 72 of which were in the *BL* group and 69 in the *HB* group. In Table 5.2, we describe the participants' demographics based on the introductory survey. Of these participants, 66.4% were males and 33.6% were females. They were between 18 and 62 years old, with a median of 31. Moreover, 42.3% of the participants had worked or studied in IT before. Participants were mostly from India (37%), USA (35%), Britain (7%), Germany (7%), and Canada (7%).

CrowdFlower presents the users with an optional satisfaction survey after completing the study, and 49 users took this survey. On average, the study received 4.2/5 for instructions clarity, 3.8/5 for questions' fairness, 3.8/5 for ease of job, 3.6/5 for pay sufficiency (before the bonus was rewarded). This ensures that participants' behavior has not been affected by either a lack of time to complete the task or the task design in general.

5.4.5 Study Details and Results

We now move to the detailed description of the modules and the results obtained. These modules are summarized in Figure 5.5, to which we refer henceforth. We also show sample screenshots from the online study in Figure 5.6. For a more elaborate documentation of the study steps, we refer the reader to Appendix B. The results are also presented in Table 5.3.

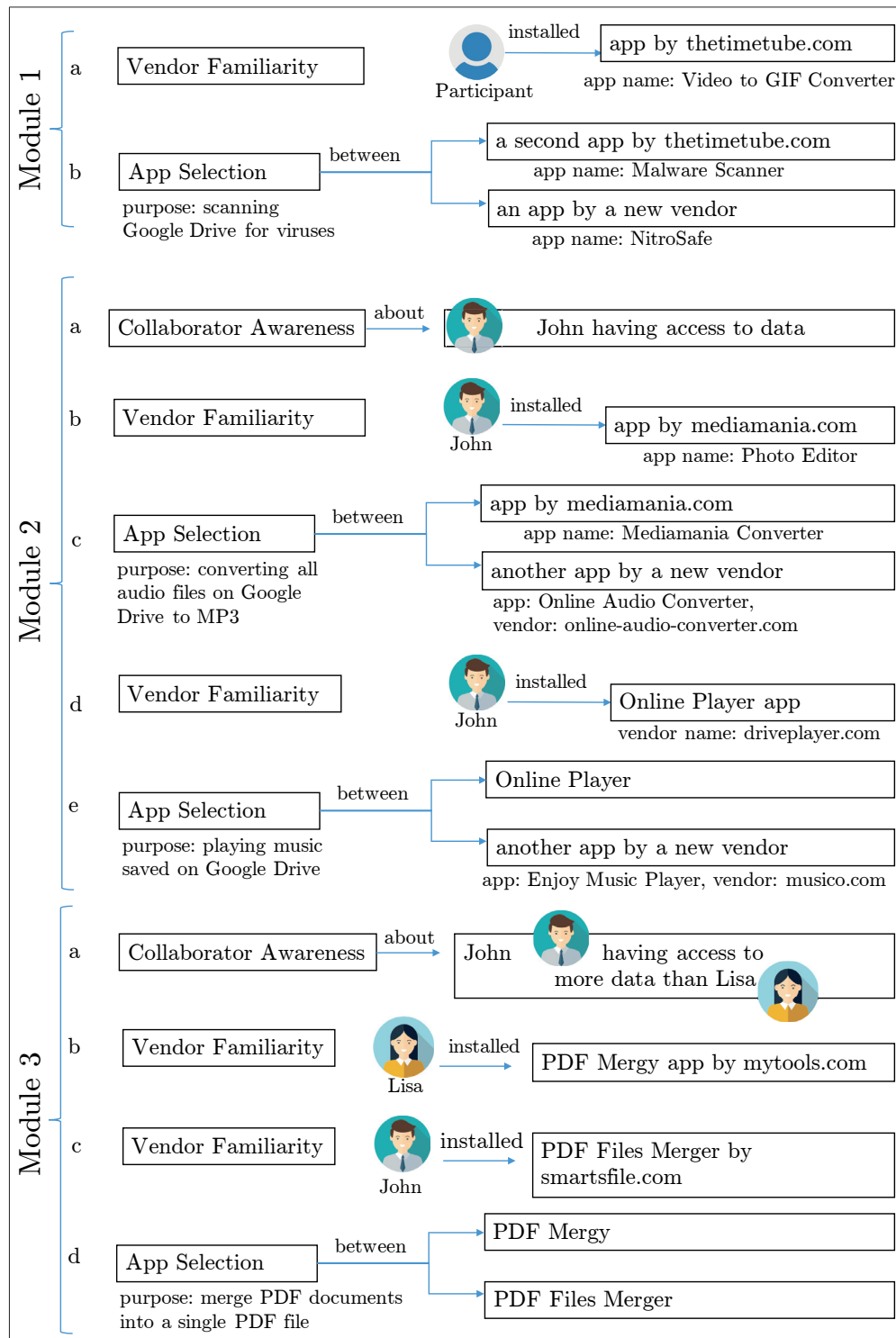


Figure 5.5 – Summary of the experiment modules; a sample of the questions corresponding to each step are available in Figure 5.6.

- Take a breath.
- You will now play a role of someone who has a Google Drive account and has **already stored files** on it (like your images from your trips with the family, some official documents, music files, etc.).
- At some point, you'll be asked to choose some apps to connect to your Google Drive account. Although no apps will be actually installed, we ask you to think as if they were real apps and that this is a real Google Drive account.
- Once you install an app, assume **it is still installed throughout this experiment.**
 - For example, if the first task says: "you have installed an application from the company *pandadoc.com*", this application is still there when you go to the next task. So if the second task says: "who has access to your data?", the answer would be "*pandadoc.com company*".
- Similarly, whenever you are informed that your friends have installed applications, consider that **these applications are still connected to their Google Drive throughout this experiment.**
- When this experiment ends and **you move to the next experiment, you will start from scratch** (i.e. with no apps installed).


(a) Instructions presented to users at the beginning of a module

As explained, we now start from scratch. Consider that this is the first app you will install. Please install any application from the company: **thetimetube.com**. (Only one such app exists, and you can click on the app to view its info.)

(b) Module 1; Task *a* : Vendor Familiarity


You now need an app that **allows scanning your Google Drive files for viruses.** Two such apps exist below. Check them both by clicking on them. Then choose the one that you prefer to install.
Be prepared to give a reason for that choice.

(c) Module 1; Task *b*: App Selection

Google Drive allows you to share files with friends. You decided to share **all your photos** on Google Drive with your friend **John** . Up till now, who is the friend who has access to your data?

☐ John
☐ Lisa

(d) Module 2; Task *a*: Collaborator Awareness

Task:
Your friend **John**  has installed an application called **Photo Editor** and has given its company access to all his files (including shared files). Write below the name of the company that owns this application. (You can click on the app to view its info.)

(e) Module 2; Task *b*: Vendor Familiarity

Assume that you have shared all your photos with **John** . Additionally, you have shared with **Lisa**  some of your photos. Who has more files from you in their Google Drive?

☒ John
☐ Lisa

(f) Module 3; Task *a*: Collaborator Awareness

Figure 5.6 – Screenshots from the user study

Module 1 (Self-History Scenario)

This module tests whether the user is more likely to select an app from the same vendor she has just installed from before. In step (a), the user is made aware she installed an app from a specific vendor v (Figure 5.6b). In step (b), she is asked to install⁶ an app that satisfies the given purpose (Figure 5.6c) among a list of apps. Two of the listed apps were relevant, and one of them was from vendor v itself.

Despite the participants being informed one step earlier that they installed an app from “thetimetube.com”, that did not make a difference in the *BL* case: half of the users still chose the app from the new vendor “nitrosafe.org” (cf. Table 5.3. In the absence of traditional signals that users follow for deciding on apps (reviews, ratings, permissions), participants apparently made decisions that canceled out, making the two apps equally favored across participants. **The vast majority of users were not approaching the installation from the angle of keeping their data with fewer shareholders.** Based on their provided justifications, they rather looked for other cues, such as selecting the app that, in their opinion, has a more comprehensive description, a more professional logo, a better sounding name, or a more trustable URL.

Still, 12 users have explicitly mentioned in their text input that they chose an app *because* it is from the same vendor they have dealt with earlier. Even then, neither of them has alluded to a privacy motivation behind the choice. These 12 participants mainly provided cross-app compatibility, interface familiarity, and satisfaction with the previous vendor as justifications. For example, one participant wrote: *“I favored Malware Scanner due to the fact that the name ‘thetimetube.com’ was in the last app installed, and I tend to install apps from the same company due to cross-app compatibility usually found in apps by the same company.”*

Interestingly, two users justified their installation of the app from the new vendor (nitrosafe.org) by writing that they had just installed an app from the same company before. This indicates that **even when users try to account for previous decisions, they might find it difficult to remember the previous app vendors.** Given that our study had a short time span separating the current from the previous installation, we expect that such mistakes would be even more common in real scenarios when app installation instances are separated by longer time spans.

The *HB* group witnessed a much larger proportion of users who favored the option with less privacy loss. 72.2% of the participants selected the app from the “thetimetube.com” (the vendor which already has access). The difference of 22.8% compared to the *BL* group is statistically significant (Fisher’s exact test, p -value = 0.005). Many of the participants who chose the app from “thetimetube.com” reported that they were motivated by the 100% access that the app already has. We counted around 40 such users (i. e., 57% of the *HB* group). Some of them went further and explicitly mentioned that their selection was motivated by giving data to fewer data owners (i. e., more privacy). For example, one user wrote: *“This company has access to all my files, so I would choose them as I don’t want to have 2 companies with full*

⁶Users were informed that this is a role-playing study, and no apps were actually installed.

Scenario	BL		HB group		Δ	p -value
	<i>VwA</i>	<i>NV</i>	<i>VwA</i>	<i>NV</i>		
Self History	50.0%	50.0%	75.4%	24.6%	25.4%	0.003
Collaborator's app	52.8%	47.2%	88.4%	11.6%	35.6%	< 0.001
Collaborator's vendor	58.3%	41.7%	82.6%	17.4%	24.3%	0.002
Multiple collaborators	44.4%	55.6%	82.6%	17.4%	38.2%	<0.001

Table 5.3 – App selection statistics in the study; *VwA*: vendor with access; *NV*: new vendor. The comparisons in each experimental group were planned contrasts, and the p -values of difference between the percentages of users who selected each app type were computed using Fisher's exact test

access to my files".

In a nutshell, we were able to verify our hypothesis in this scenario: **the new privacy indicator leads users to more frequently choose the app from a vendor they already authorized**. Furthermore, we have discovered that the *HB* Insights interface has indirectly made users think about various positive effects brought by using apps from the same vendor. This eventually lead them to make more privacy-preserving decisions.

Module 2.1 (Collaborator's App Scenario)

This module tests the likelihood that the participant selects the same app that her collaborator had used. In step (a), the participant is made aware that she had shared all her photos with a friend f (Figure 5.6d). For more familiarity, we also added a picture for each of the two fictitious friends throughout the study. In step (b), the user is made aware that her friend f has installed an app a_0 (Figure 5.6e) from vendor v . She is asked to type the name of the app's vendor ("paste" option was disabled in the input field to further ensure the participant is aware of the vendor). In step (c), the user is asked to install an app with a certain purpose (similar to Figure 5.6c). One of the two matching apps is app a_0 .

Similar to the previous module, the *BL* group witnessed an almost even split between "Online Player", installed previously by the friend, and "Enjoy Music Player", from a new vendor (cf. Table 5.3. We also noticed that 20 participants in this group justified their decision by mentioning that their friend has used the app. Still, neither of them alluded at privacy reasons in their justifications. Instead, the two most prevalent motivations were (1) considering the friend's use of the app as a *recommendation* or (2) achieving *compatibility* with their friends' app, which facilitates data sharing within the app itself. Quoting one user: "*This is the same app my friend is using so it should be quite compatible for us to both share.*"

In addition to having a significant 35.6% difference in the case of the *HB* group, we noticed that

32 users mentioned the existing data access as a reason for choosing the app “Online Player”. Also, 26 users referred to the fact that the friend has installed this app before (including those who mentioned both of the previous reasons). Unlike the *BL* group’s justifications though, where the friend’s recommendation and the app’s compatibility prevailed, the privacy issue was explicitly brought up by at least 10 users. One participant put it as follows: “*Thanks to John, they have already access to 70% of my data. Sharing the last 30% isn’t as bad as sharing 100% of my data with driveplayer.com.*”

Module 2.2 (Collaborator’s Vendor Scenario)

We proceed in steps (d) and (e) as in the previous scenario’s steps (b) and (c), with the difference that a *new* app from v is included among the options in step (e) instead of the exact same app a_0 . One interesting insight from this scenario is that **the line between the company and the app is blurred in the minds of several users** who used the two entities interchangeably. In fact, 3 users in the *BL* group and 7 participants in the *HB* group justified their choices by mentioning that their friend installed the *same app* before, which was not the case. For example, one user wrote: “*this app already has access to my files, and I don’t want to install any new app.*”

Module 3 (Multiple Collaborators Scenario)

Given collaborators f_{more} and f_{less} , where the user shares much more data with f_{more} , this scenario checks the likelihood of the participant authorizing an app that f_{more} has installed. In step (a), the participant is made aware that f_{more} has access to more data than f_{less} (Figure 5.6f). In steps (b) and (c), the participant is made familiar with the apps each of the friends installed (similar to Figure 5.6e). In step (d), the user is asked to select an app with a specific purpose. The two friends’ apps are the only ones matching, and the choice is to be made between them (similar to Figure 5.6c).

In the *BL* group, we had 44.4% of the participants choosing the app installed by f_{more} . Still, this percentage is relatively close to an equal split between the two apps. Out of this percentage, 13 users justified their choice by mentioning that they were encouraged to follow the choice of friend f_{more} . Even though they did not mention privacy, the larger number of files shared with f_{more} was often used as a justification. For example, one participant wrote: “*This is the app that John already uses, and he has access to all of my files. The PDF Mergy app is used by Lisa, but she only has access to part of my files.*”

In the *HB* group, around 82.6% chose the app previously installed by the friend f_{more} , which is significantly more than those in the *BL* case (Fisher’s exact test, p -value < 0.001). Looking at the justifications, around 37 users explicitly mentioned the higher access level that this app already possesses as a reason for their choice. Privacy was additionally mentioned by 8 of these users. Quoting one of them: “*PDF Mergy already has access to 70% of my files. Using*

PDF Files Merger would unnecessarily increase third party app access to my files.” However, we still had 2 users who went for the app with less existing access, with one of them saying he favors the app that only “*had accessed 30% of files before installation*”. What was interesting though is that **almost all users who mentioned friends were making a comparison between the two friends’ existing access level**, regardless of their final choice.

5.4.6 Concluding Survey

At the end of the user study, users were presented with a final set of questions. We asked them whether they would like to be notified when a friend installs an app that gets access to their shared files. Around 92% of users in the *BL* group and 90% of users in the *HB* group agreed. We further asked the participants whether they are fine with a collaborator being notified when they install applications that access files shared with that collaborator. The percentage of people who agreed dropped to 75% in the *BL* and 78% in the *HB* group. The relatively small difference between the answers to these two questions highlights that **only a minority of users is not willing to make the trade-off of contributing to the overall system**. Such users can be given the option to not use privacy indicators based on their friends’ decisions.

Next, users were asked the following question “*Assume you have installed an application called YouMusic from a company called Musicana and gave it access to all your files on Google Drive. Now you are considering installing an application called YouVideo from the same company. How do you think that this application will affect your privacy:*”. Only 11% of each group replied by “*negatively*”. The vast majority in both groups either perceived the avoidance of a new vendor as a positive outcome or considered that the privacy loss will remain the same. Interestingly, the users in the *BL* showed a similar reasoning in justifying their choices as the *HB* group although the latter were primed about these aspects via the privacy indicators. This indicates that **the privacy indicators match the first intuition for a large fraction of users**.

5.4.7 Discussion and Limitations

Overall, we found out that, in the three modules, participants in the *HB* group were significantly more likely to install the app with less privacy loss (i. e., the app from the vendor with the largest share of the user’s files) than those in the *BL* group. Despite showing the efficacy of History-based Insights, our study still has its limitations. To get a large, diverse sample size, we resorted to a web experiment based on role-playing with hypothetical data. It would be interesting to see how such results extrapolate to the case where users’ own data is in question and the users are in immediate need of installing an app. On the one hand, the users might be more alerted towards their own data privacy. On the other hand, users are typically seeking an app to satisfy an immediate need. A longitudinal study with actual users’ data is also well-suited to study the effect of our new privacy indicators over time.

Moreover, in our design, we have abstracted several factors (e.g., ratings and reviews), which

have been previously studied in similar ecosystems [KCS13], in order to focus on one factor. These factors might have diluted the effect of the privacy indicator. Still, we conjecture that, although the absolute values of our findings might not strictly apply, the differences between the two groups will still be practically significant.

Additionally, in this chapter, we have investigated only one type of history-based privacy indicators. Evidently, such indicators can be integrated at different stages of the app installation process. For example, they can be part of the recommendation strategy for suggesting alternative apps. They can also be included in the apps' search interface. Apps can also be labeled as "privacy preserving" in the web store based on this metric. It is also possible that the privacy indicator is only shown when the vendor has existing access to the user's data. This might serve to reduce the habituation effect and the information overload. The best choice among these deployment scenarios needs further investigation.

Furthermore, it is important to note that, although our experimental interface mentions the collaborators' name in the explanation under the progress bar, this does not have to be the case in actual deployments. We hypothesize that removing the name will not have a significant impact on the results as it was not highlighted in the interface. This allows the CSP to relay such information to the users without exposing sensitive data about particular collaborators. The CSP can resort to more sophisticated anonymization methodologies, such as showing a non-exact percentage that can be mapped to multiple collaborators. Exploring the impact of these techniques is left for future work. Moreover, we note that this anonymization might not be needed at all in the enterprise settings, where apps installed by team members are supposed to be visible for the administrators. As we show in Section 5.5.3, a significant reduction in privacy loss can be achieved without even accounting for decisions by users external to the team.

Finally, the privacy indicator in our study has addressed two granularity levels: full and per-file access. However, the same indicator can be extrapolated to the case of per-type access. For example, the interface can say: "The app's company already has access to 70% of your *photos*" (instead of *files*).

5.5 Large Networks' Simulations

In the previous section, we showed the significant change that our privacy indicator can effect through encouraging users to make History-based decisions. We will tackle the next research question, where we investigate the impact of adopting such privacy indicators on the privacy risk in realistic scenarios with large user networks. As we are not in the position of the CSP to study an actual implementation of the *HB* Insights interface over time, we will perform a simulation of potential users' installation behavior. We will base this on both the crowdsourced decision model inferred from the user study and on new collaboration networks that we construct.

5.5.1 Simulation Data

Collaboration Networks

For the purposes of this simulation, we constructed the following three networks. The first network is an inflated version of the *PrivySeal Dataset*. The second is a large collaboration network with a more realistic degree distribution. The third network allows us to study the case of collaboration within teams.

- **Inflated Google Drive Network:** We used the standard degree-driven approach for network topology generation to construct a larger Google Drive network based on the one in the *PrivySeal Dataset* of Section 5.3.1 [MV02]. Based on an input user degrees' distribution from that dataset, we particularly used the *Configuration Model* as described by Newman [New03] and implemented by the library NetworkX [SS08] for inflating the graph. This model generates a random pseudograph (a graph with parallel edges and self-loops) by randomly assigning edges to match an input degree sequence. We removed the self-loops and parallel edges a posteriori from the generated graph. In the end, we had a collaboration graph with 18,000 users and 138,440 edges. This graph is, by construction, a connected graph, with an average node degree of 15.
- **Paper Collaboration Network:** In an effort to have a realistic, large collaboration network without resorting to graph inflation, we relied on the Microsoft Academic Graph, which consists of records of scientific papers along with the authors and their affiliations [SSS⁺15]. We used a snapshot of 50,000 papers, and we constructed the collaboration graph based on it. We ended up with 41,000 collaborators and 199,980 edges. The graph itself is not a connected graph but is rather constructed of around 1700 connected components. The average node degree is 4. Our rationale is that this graph captures a realistic scenario of users collaborating on authoring documents, which is, in fact, an activity achieved via cloud services nowadays. Hence, it is fit for showing the efficacy of our privacy indicators.
- **Team Collaboration Network:** We used the same academic graph to construct a network of teams. A team is defined as a frequently collaborating group of people. Motivated by research around community detection [MV13], we use Strongly Connected Components (SCCs) to label teams in our graph. We ended up with 16,400 users split over 1700 teams. Unlike the previous two networks where users themselves are the data subjects (whose privacy is to be optimized), members of each team in this network consider their team as the data subject.

Sharing and Installation Patterns

In order to closely model the user characteristics in Google Drive, we assigned to each user in the collaboration networks a file sharing distribution and a number of apps corresponding to a user with a matching degree in the *PrivySeal Dataset*.

Apps

As we wanted to perform the simulation with a much larger number of users than we had in the dataset described in Section 5.3.1, we also needed a larger collection of apps. Given that Google Chrome Store has only around 500 apps that are tagged by the “Works with Google Drive” tag, we decided to also include all Google Chrome Apps in the dataset (i. e., even those that do not have this tag). As far as the simulation is concerned, this step is justified since the only realistic information that we will rely on is the distribution of vendors per app. It is fair then to assume that this distribution does not differ significantly between the general category and the Google Drive category. Hence, we augmented the PrivySeal Dataset via apps from the Google Chrome Store to arrive at 1000 apps. In addition to the app’s installation count and vendor name, we also collected the set of “*Related Apps*” that the store displays for each app. This is because, in our simulation, we will assume that users have the choice to choose the app itself or one of its related apps. Again, this is a fair assumption as these related apps are mostly the apps which deliver a close functionality to the app itself, and we will only rely on them to model the alternatives at each simulation step.

User Decision Models

For the purpose of this simulation, we define 3 user decision models:

- **Fully Aware Model (FA):** the user always makes the decision that minimizes the privacy loss of the data subject, taking into account all previous installation decisions by her and by her collaborators.
- **Experimental History-based Model (EHB):** the user takes decisions similar to what a random user of the *HB* experimental group does. In specific, we model those users as taking a history-based decision with probability q and making a random app choice with probability $1 - q$. We set q based on the number of users who mentioned the app’s existing access in *writing* as a reason for their choice in each module of Section 5.4. Based on Module 1’s users’ responses, we set $q = 0.57$ when the user encounters a vendor she previously authorized. Based on Module 2, we set $q = 0.70$ whenever the user is presented with one vendor previously authorized by a single collaborator. Based on Module 3, we set $q = 0.67$ for the cases where the user is presented with multiple vendors previously authorized by her collaborators. In all of these cases, the user will select the vendor with the minimal resulting *Aggregate VFC_u* with probability q .
- **Experimental Baseline Model (EBL):** the user takes decisions similar to what a random user of the *BL* experimental group does. As users in practice are rarely informed of what their friends have installed before, we do not integrate this knowledge into the model. Hence, we only account for the case of Module 1, where the user’s previous decisions are concerned. Based on the fraction of users who mentioned the app’s existing access as a motivation for their choice, we set the probability of taking history-based decision in this model as

Algorithm 1 Simulation Steps

```

1: Initialize  $VFC_u$  value to 0 for each user
2: for  $t \leftarrow 0$  to  $N$  do  $\triangleright N$  is total number of steps
3:   select a random user  $u_0$  based on user's app installation frequency
4:   select a random new app  $a_0$  based on app's installation count
5:    $A_{rel} := \{a_0\} \cup$  (set of related apps of  $a_0$ )
6:    $V_{rel} :=$  set of vendors of apps in  $A_{rel}$ 
7:    $r :=$  a random rational number in the range  $[0,1]$ 
8:   if user had installed apps by vendors in  $V' \subset V_{rel}$  then
9:     if ( $r < q(\text{group}, \text{'same vendor'})$ ) then  $\triangleright q$  is a function of the user decision model;  $\text{group}$  is the experimental group
10:      select a random vendor  $\hat{v} \in V'$ 
11:      install the app  $\hat{a}$  in  $A_{rel}$  from vendor  $\hat{v}$ 
12:     else
13:       install app  $a_0$ 
14:     end if
15:   else if  $\exists (c \in C(u_0))$  who installed apps by vendors in  $V' \subset V_{rel}$  then
16:     compute  $VFC_{u_0}(\{v\})$  for each vendor  $v$  in  $V'$  at this time step
17:     select the vendor  $\hat{v} \in V'$  with highest  $VFC_{u_0}(\{v\})$  at this time step
18:     if ( $r < q(\text{group}, \text{'collaborator vendor'})$ ) then
19:       install the app  $\hat{a}$  in  $A_{rel}$  from vendor  $\hat{v}$ 
20:     else
21:       install app  $a_0$ 
22:     end if
23:   else
24:     install app  $a_0$ 
25:   end if
26:   for all  $u \in \{u_0\} \cup C(u_0)$  do
27:     update  $\text{Aggregate } VFC_u$  for  $u$   $\triangleright$  recompute it via Equation 5.2.3
28:   end for
29:   update the average  $\text{Aggregate } VFC$  over all users
30: end for

```

$$q = 0.18.$$

In the special case of the team collaboration network, users who take history-based decisions account for their own decisions and the decisions of their team members only. We do not consider that users account for decisions taken by members of other teams. This is to demonstrate the potential of the privacy indicators under strict conditions.

5.5.2 Simulation Details

We now move to the description of the simulation itself, which is detailed in Algorithm 1. We had three simulation groups, named after the three decision models: *FA* group, *EHB* group, and the *EBL* group. The simulation was run until the average number of apps installed across by users reached 30 apps⁷. On a high level, at each simulation step, the following actions are performed:

⁷Comparatively, mobile users have accessed 26.7 smartphone apps on average per month in the fourth quarter of 2014 [The15].

- A user is selected from the collaboration network via a weighted random sampling based on the assigned app installation frequencies (line 3). This accounts for the diversity of users' installation frequencies. An app a_0 is selected from the simulation apps' dataset via a weighted random sampling based on the actual app installations count in Google Chrome Store (line 4). That way, popular apps are installed more frequently (as is the case in practice).
- A user decision is simulated. The user is assumed to be choosing the app a_0 or one of its related apps. This choice is made depending on the user's decision model, as explained previously.
- Finally, the average *Aggregate VFC* is computed based on all users' *Aggregate VFC_u*.

5.5.3 Simulation Results

To demonstrate the simulation results, we show three types of figures per collaboration network. On a high level, in Figures 5.7a, 5.8a, and 5.9a, we show how the privacy loss (quantified using the average *Aggregate-VFC*) in each group evolves as users install more apps. In Figures 5.7b, 5.8b, and 5.9b, we show ratios of the privacy loss in the two experimental groups *EHB* and *FA* with respect to the baseline *EBL* group. Finally, Figures 5.7c and 5.8c, and 5.9c show the actual events contributing to the privacy loss growth, where we can specifically check the fraction of apps coming from new vendors, those coming from vendors previously authorized by the user, and those from vendors previously authorized by collaborators.

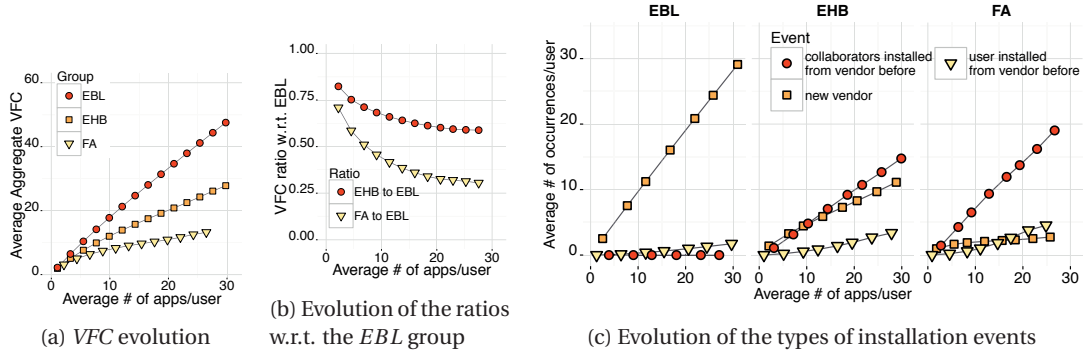
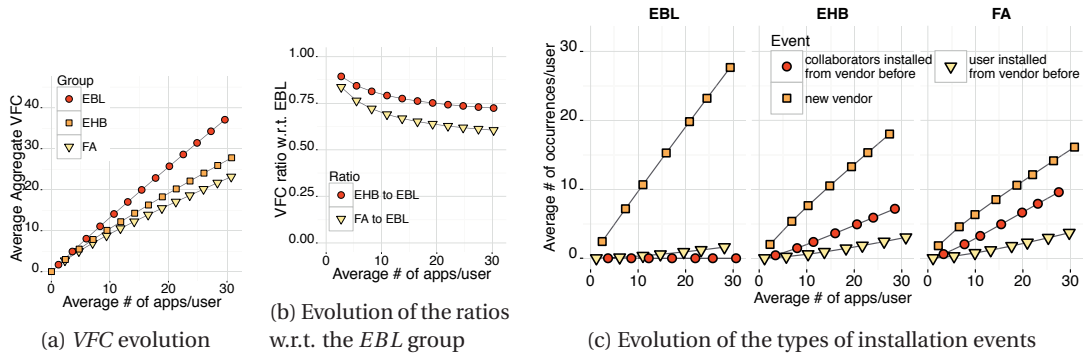
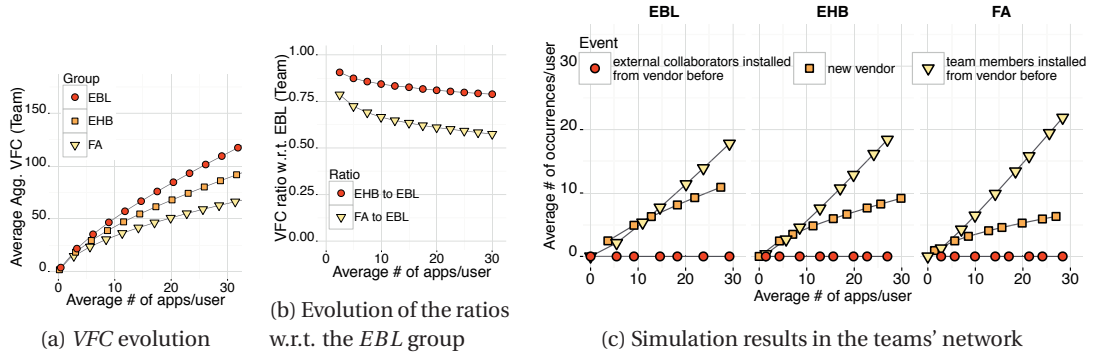
Results for Individuals' Networks

Based on these metrics we start by analyzing the results for the individuals' networks, where we observe the following:

Curtailed growth of privacy loss: From Figures 5.7a and 5.8a, we notice that the growth of the privacy loss is visibly curtailed in the cases of *EHB* and *FA* groups compared to the baseline *EBL* group. This significant divergence demonstrates the efficacy of our *HB* privacy indicators.

Impact of the network effect: Looking into the ratios in Figures 5.7b and 5.8b, we see that the privacy loss in the *EHB* group has dropped by 41% in the inflated network and by 28% in the authors-based network (both with respect to the baseline). In the *FA* group, where users always optimize their privacy, the privacy loss has dropped by 70% in the inflated network and by 40% in the authors-based network. This higher impact in the case of the inflated network is because it is a connected graph, unlike the authors-based network, which is composed of smaller connected components. Nevertheless, we can state that, although our privacy indicators have a larger effect on highly connected networks, they are still significantly effective in less connected networks, like the authors-based dataset.

Importance of accounting for collaborators' decisions: To dive further into events that lead

Figure 5.7 – Simulation results in the **inflated network**Figure 5.8 – Simulation results in the **author-based network**Figure 5.9 – Simulation results in the **teams' network**

to the observed privacy loss patterns, we look into Figures 5.7c and 5.8c. First, we observe that users in the *EBL* group are mainly installing new apps from vendors that had no previous access to their data. This is reflected in the almost linear increase of privacy loss in Figures 5.7a and 5.8a. Second, we observe that, in the case of the inflated network, users have frequently been installing apps from vendors with existing access through their collaborators. In fact, as apparent in Figure 5.7c, this event outnumbers the event of installing from a new vendor. Third, the number of installations from collaborators' vendors is also significant in the case of the authors-based dataset. While it does not outnumber the installations from new vendors

(due to the low-graph connectivity), this is still enough to lead to 28% and 40% decrease in the privacy loss in the *EHB* and the *FA* groups respectively. Finally, we note that, although the users are more frequently encountering vendors authorized by their collaborators than by themselves, the latter event is still significantly impacting the results. This is because users still incur an incremental privacy loss with vendors authorized by their collaborators while this loss is zero with vendors they have previously authorized. Accordingly, the obtained optimizations are a result of users' accounting for their own and for others' decisions.

Results for the Teams' Network

We now discuss the results for the case of the collaboration network where users work in teams and aim to protect the privacy of the team's data. We observe the following, based on Figure 5.9:

Inherent usage of similar apps: From Figure 5.9c, it is clear that the dominant event is that of users installing apps which have been authorized by other team members before. This is even in the case of the baseline group (*EBL*), which was not the case in the individuals' networks. We justify that by the fact that we selected apps at each simulation step to match their realistic installation frequencies. In practice, apps' installation counts follow a long-tail distribution, and users tend to mostly install a limited set of apps. That is why team members will naturally tend to install a set of similar apps.

Curtailed growth of privacy loss: Still, we observe that the trend of slower growth of privacy loss also applies in the case of teams (Figure 5.9a). As we also observe in Figure 5.9b, the privacy loss has decreased by 23% for the *EHB* group and by 45% for the *FA* group, both with respect to the baseline group. This implies that there is ample room for privacy optimization in teams too.

Effect due to internal collaborators: We finally observe that the privacy loss decrease was achieved via decisions taken by each team's members independently, without relying on other teams' decisions. This highlights the fact that *HB* privacy indicators can still be effective even when users do not account for others' decisions. Obviously, taking the external members' decisions into account can lead to further optimizations.

In sum, our simulations provide further evidence of the efficacy of using History-based privacy indicators in a large network of collaborators. It is worth noting too that, although users in our study were following the *EHB* decision model, we believe that, in an actual deployment of such indicators, the model will move closer to the *FA* model. This is because users are more protective when their personal data is at risk than when they are put in a role playing scenario about fictitious data. Moreover, users in our study were exposed to this indicator for the first time. When users are educated more about this feature, they might be more likely to take advantage of it.

5.6 Related Work

5.6.1 Interdependent Privacy

The problem of interdependent privacy has been tackled before in the context of social apps. The main approaches were high-level game-theoretic or economic modeling.

Biczók and Chia first introduced the concept of interdependent privacy and studied its presence in Facebook 3rd party apps' ecosystem [BC13]. They also modeled its impact via a game theoretic, (2-player, 1-app) model and showed how positive externalities (improved user experience due to users installing the same app) and the negative externalities (privacy loss) could affect the users' app usage behavior. One of the problems they hint at is the absence of risk signaling, which is what this chapter has tackled.

Pu and Grossklags [PG14] later presented a more elaborate economic model that additionally accounts for larger groups of users and the interplay among various social network parameters. They also consider the users' "other-regarding preferences", i.e., how much they care about privacy harms they inflict on their peers. They showed that app rankings do not accurately reflect the level of interdependent privacy harm the app can cause (a phenomenon that does not extrapolate to our case of cloud apps, where *all* apps have the potential to inflict interdependent privacy harm). They additionally conclude that even rational users who consider their friends' well-being might adopt apps with invasive privacy practices at a certain stage. One commonality with our work is that they also perform a simulation to study the effect of app adoption decision at scale. However, our simulations are based on graphs that mimic the real-world collaboration networks (rather than scale-free networks). We also consider the worst case of selfish users, and we model users' decisions based on the user study that we performed. In that sense, the insights derived here can be seen as complementary to the insights derived in that paper.

In a subsequent work, Pu and Grossklags [PG15] used a conjoint study approach to quantify the monetary value which individuals associate with their friends' personal data. They found that individuals place a significantly higher value on their own personal information than their friends' personal information. This further supports our assumption of self-interested users in this work. The same authors also built on a user survey in [PG16] to assess the factors affecting users' own privacy concerns as well as friends' privacy concerns in the context of social app adoption. In particular, they found evidence of negative association between past privacy invasion experiences and the trust in 3rd party apps handling of their own data. They also found partial support for a positive effect of privacy knowledge on concerns for users' own privacy and their friends' privacy.

Other works have also investigated the issue of interdependent privacy in the context of location privacy [OHS⁺16] and genomic privacy [HAHT17]. In this work, we are focused on quantifying the interdependence of privacy in the context of cloud apps before addressing it from a usable-privacy perspective, thus bridging the gap between the theoretical studies and

the end-user needs.

5.6.2 Privacy Nudges

This chapter can be categorized under the line of work on privacy nudges, which we have discussed in the previous chapter (Section 4.8). In that chapter, we presented the first study the privacy of 3rd party cloud apps and we exposed that almost two-thirds of those apps are over-privileged. We also introduced a novel privacy indicator for deterring users from installing *over-privileged* apps by showing them Far-reaching insights that apps can needlessly infer from their data (e.g., top topics, faces, or locations of interest). This work, however, helps users improve their privacy by reducing the vendors with access to their data, even if the functionality delivered by the vendor abides by the least-privilege principle. Hence, it complements these approaches and can be deployed alongside any of them.

In addition to the related works on privacy nudges that we discussed in Section 4.8, we also note the relevant work by Almuhiemedi et al., who showed the effectiveness of regularly alerting users about sensitive data collected by their apps, in encouraging users to review and adjust the permissions [ASS⁺15]. Applying such an approach in the context of 3rd party cloud apps can also be effective in increasing transparency towards the apps' practices (see our discussion on transparency dashboards in Section 4.7).

Highlighting the app's vendor name in the interface, a step which fits naturally into the *HB* insights interface has been previously shown by Bravo-Lillo et al., to be effective in nudging users to pay attention to potential malicious vendors appearing in security warnings [BLKC⁺13]. They also demonstrated that interacting with the vendor's field was even more effective. In our user study, the vendor's name appeared multiple times as a justification for installing/not installing the app. However, the problem we tackle here is orthogonal to the problem of inferring (mis)trust from the vendor's name; it is rather about the ability of users to remember vendors across multiple, typically far-apart, installations.

5.7 Summary

The findings in this work are the first to concretely delineate the various aspects of interdependent privacy in 3PC apps. One of the major outcomes is that a user's collaborators can be much more detrimental to her privacy than her own decisions. Consequently, accounting for collaborators' decisions should be a key component of future privacy indicators in 3rd party cloud apps. We have shown the impact of History-based Insights as a privacy-enhancing technology in this context, especially that, based on our user study, users are less likely to account for previous decisions on their own. Our privacy indicators would optimally be implemented by the CSPs themselves as they control the authorization interface and the application stores. The indicators can also be realized by third party privacy providers with access to users' data. Our approach can also be easily mapped to other ecosystems. In the mobile apps' scenario, it

can enable users to reduce the number of vendors with access to her contacts. It can also be extended to the case where the goal is protection against 4th parties (e.g., ad providers and data brokers). There, the user can account for data previously held by a 4th party with which the app vendor cooperates.

Handling Language Complexity **Part III**

6 PriBot: Automated QA for Privacy Policies

6.1 Overview

The previous two parts of this thesis were contributions towards two main problems: how to protect the data before it leaves the user’s device in the first place and how to protect the data before it is transferred from the cloud storage provider to a third party app. In this final part, we consider the complementary problem: how to better understand what privacy and security guarantees the apps promise. We tackle this from the angle of privacy policies, which are at the core of the online *notice* and *choice* paradigm for virtually all apps and websites (i.e., not only restricted to third party cloud apps).

Users, however, rarely read these policies when they sign up to new services. Multiple recent events have shown that this is not due to indifference but rather to the difficulty of sifting through all the information spread across multiple pages of text [Isa17, Bra17]. For example, *Unroll.me*, a free service for removing unwanted subscriptions from users’ email, was recently reported to be selling information mined from those emails to third parties (e.g., *Uber*) [Isa17]. After the media reports, a lot of user backlash occurred. However, the reported practices were already covered by the company’s privacy policies that users had agreed too. Still, it was never clear for the majority of users that the company’s business model hinged on selling their emails’ contents to other parties while still advertising the company as a mailbox “clean-up” service.

Given this inherent user interest in knowing the data practices of service providers on hand and the difficulty in attaining that within the different user time, effort, and knowledge constraints, on the other hand, we see an immediate need to address the shortcomings of privacy policies with a new perspective. Our motivation is not only the user frustration with the current status quo but also the emerging technologies, which are adding new kinds of user needs. In particular, this is manifested through UI-limited devices and automated customer support.

UI-limited Devices. As part of our daily routines, we interact with a large number of UI-limited devices and services, such as voice-activated digital assistants (e.g., Amazon Alexa and Google Assistant) and IoT devices (e.g., smart thermostats or door locks). Many of these devices rely on *voice commands* as their primary input method. With such conversation-first devices, the existing techniques of linking to a privacy policy or reading it aloud are not usable; they might require the user to access privacy information and controls on a different device, which is not desirable in the long run [SBDC15]. The miniaturization trend has further led to new classes of small screen devices, such as smartwatches and other wearables. Previous efforts on standardizing privacy policy content into tables similar to nutrition-labels [KBCR09] are difficult to fit on such devices. Even the machine learning techniques of analyzing privacy policies, such as [ZB14], have so far followed a one-size-fits-all approach by providing all users with the same interface, regardless of their privacy interests.

Customer Support. Similar issues, related to the inadequacy of current methods for privacy notice delivery, are emerging in another domain: customer support. As a new trend in the industry, automated customer support allows companies to respond to customer inquiries in real time and around the clock. Through chatbots and other automated interfaces (e.g., Twitter bots), companies interact with their customers on a variety of topics, including privacy. We have sampled Twitter for companies' tweets that mention the term "privacy policy" in response to customer inquiries. We found that the number of such tweets has increased 20x between 2008 and 2016. Existing approaches for presenting privacy policies to users cannot apply in this context; they are incapable of providing concise answers to privacy-related questions posed in natural language.

Our vision to address such users' concerns is by treating the privacy policy as an unstructured data source rather than a user interface. Our approach breaks the concept of privacy policy into a three-layered framework. On the bottom is the *Data Layer*, where we have the textual content of these policies. In the middle, is the *Machine Learning layer*, which involves the data analysis algorithms that we built to classify the content of privacy policies and to answer questions about them. On the top is the *User Interface (UI) Layer*, which leverage the results of the automated analysis to relay the information to the user in novel ways.

In this chapter, we develop an instantiation of this framework via PriBot, the **first automated Question-Answering (QA) system for privacy policies**. The PriBot system takes a previously unseen privacy policy and uses it to answer, in real time with high accuracy and relevance, user questions that are posed in free form.

At the core of PriBot, we propose a novel, deep-learning-based algorithm for matching users' questions with answers from the privacy policy. This algorithm accounts for the intricacies of privacy policies that render traditional retrieval or QA systems suboptimal in this case. Specifically, our approach can handle questions about high-level issues, such as services sharing user information with third parties, and fine-grained issues, such as whether companies release

data to legal authorities. PriBot can provide the answers in two alternative forms: via excerpts extracted from the privacy policy itself, and via high-level interpretations of these excerpts.

PriBot provides substantial benefits for both users and service providers. It allows the user to pose a question, in natural language, about the privacy practices of a certain provider. PriBot, whether integrated as a chatbot, a voice assistant, or a social media bot, responds back with a relevant and concise answer from the company's privacy policy. Moreover, PriBot enables providers to deliver privacy notices through their UI-limited devices and services. Furthermore, companies can use PriBot to add automated support for their privacy-related questions from consumers without requiring a large knowledge-base of hard-coded rules. PriBot assists companies by streamlining processes and training for customer service representatives to handle privacy inquiries.

Contributions

In this chapter, we make the following key contributions:

- We propose **a novel deep learning approach for ranking answers with respect to questions in the privacy domain** (Sections 6.3 and 6.4). Our approach consists of a hierarchy of neural-network classifiers that accounts for both the high-level aspects and the specific pieces of information present in privacy policies. We evaluate its performance against multiple traditional QA techniques that we transplanted into the privacy policies context. We show that our best algorithm consistently surpasses the predictive accuracy of traditional QA retrieval techniques on standard metrics (Section 6.6). We also demonstrate its high user-perceived utility via a user study, where it returned a satisfying answer among the top-3 for 91% of the questions (Section 6.7).
- **We create a new test dataset of user-posed, privacy-related questions from Twitter.** These questions represent real-world and unaltered user-company interactions about privacy practices. They further have the advantage of avoiding subject bias, which is likely to happen when eliciting privacy-related questions from individuals (Section 6.5).
- **We implement PriBot as an online, real-time, text and voice-activated chatbot** (Section 6.9). The chatbot delivers answers segmented from the policy itself or answers generated from high-level interpretations of the policy. Our chatbot is the first practical system for answering questions about privacy policies. It will be available to the public via a web interface with the release of this thesis.

6.2 System and Data Overview

In this section, we describe the PriBot approach at a high level, along with the datasets we utilize to build it.

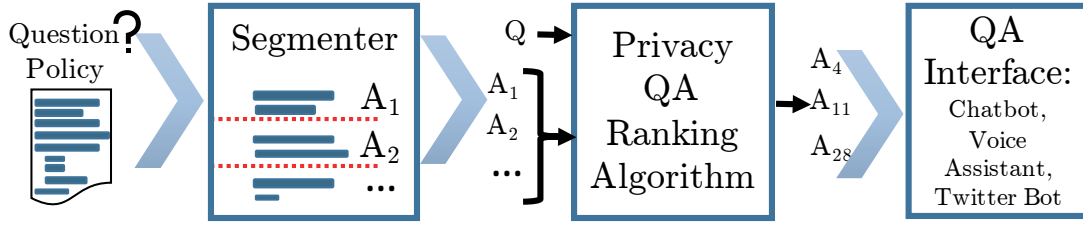


Figure 6.1 – High-level system overview of PriBot.

6.2.1 System Overview

From a high-level perspective, PriBot takes as input the user’s questions along with a privacy policy URL to retrieve suitable answers. Fig. 6.1 provides a high-level description of PriBot. The operation of PriBot involves two main components, preprocessing the policy and the QA ranking algorithm.

Policy Pre-processing. PriBot pre-processes the privacy policy through its *segmenter* component. The segmenter first extracts the policy content from the policy’s webpage and then partitions it into a set of adequately-sized and semantically coherent fragments. These segments constitute a pool of candidate answers from which PriBot chooses the answer to the user’s question. We elaborate further on the policy pre-processing in Section 6.3.

QA Ranking Component. At runtime, the main task of PriBot is to match the user’s question to one or more segments of the privacy policy. PriBot involves a QA ranking component, specialized for the privacy setting, to rank the policy segments according to their “closeness” to the posed question. In this chapter, we contribute a new architecture of neural networks, termed the Hierarchical model, that accounts for the multi-level complexity of privacy policies (Section 6.4.3). We compare it against two other models that we build: the Retrieval model, inspired by traditional information retrieval (Section 6.4.2) and the SemVec model that takes a neural-network approach to retrieval (Section 6.4.3).

Friendly Summary Approach. We also propose an answer-generation algorithm that summarizes the retrieved policy fragments into a short-form, more readable response. Employing this approach, PriBot automatically labels the segments matching the user’s question according to the type and purpose of privacy practice. It uses these labels to generate an answer via a custom grammar; these answers are considerably shorter and more readable. We elaborate on this answering approach in Section 6.8.

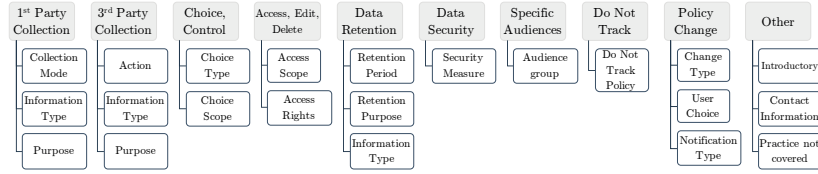


Figure 6.2 – The OPP-115 dataset hierarchy [WSD⁺16].

6.2.2 Data Overview

To answer the wide range of free-form questions from users, PriBot is data-driven by design. Prior to this work, there were no publicly available datasets for privacy-related question answering that we can readily use to train PriBot’s neural networks. To address this challenge, we leverage the Online Privacy Policies (OPP-115) dataset, introduced by Wilson *et al.* [WSD⁺16]. We re-purpose this classification-oriented dataset as a building block for our QA ranking approaches (SemVec and Hierarchical). We do not apply this dataset in the traditional classification sense as it is inapplicable for the QA context. Instead, we use this dataset to train a set of custom neural networks to extract feature vectors from the user questions and policy segments, which we use later for answer ranking.

OPP-115 Dataset for QA Purposes

The OPP-115 dataset contains 115 privacy policies manually annotated by skilled annotators (law students). In total, the dataset has 23K annotated data practices. The annotations were at two levels. First, paragraph-sized segments were annotated according to one of the 10 high-level categories in Fig. 6.2 (e.g., 1st Party Collection). Then, annotators select parts of the segment and annotate them using attribute-value pairs, e.g., Information Type: Location Information Type, Purpose: Advertising, etc. In Fig. 6.2, we only show the mandatory attributes that should be present in all segments. There were also optional attributes that can sometimes occur in each category. In total, there were 16 distinct mandatory attributes and 94 distinct values across all attributes. For space constraints, we do not show all the attribute values in Fig. 6.2. We use this dataset to train a set of neural networks to predict the higher-level categories and the attribute-value pairs given a policy fragment as input.

Combined Policies Dataset

We have collected a corpus of 16,500 privacy policies of Android apps from Google Play and augmented it with a diverse corpus of 1,000 privacy policies collected by Ramanath *et al.* [RLSS14]. We combine these two policies datasets with the raw policies from the OPP-115 dataset, to form an unlabeled corpus of around 16,615 policies, to which we refer henceforth as the Combined Policies Dataset. We utilize our Combined Policies Dataset to seed the Retrieval model and to segment the policies into coherent answers.

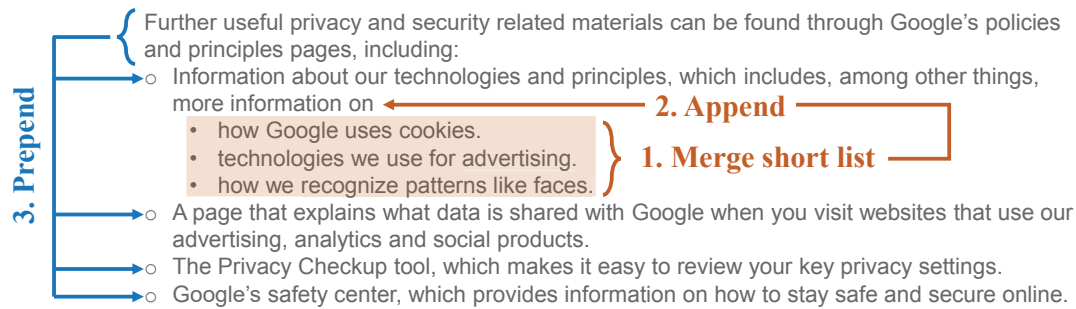


Figure 6.3 – List merging during the policy segmentation.

6.3 Policy Pre-processing

The policy preprocessing includes two steps: extraction and segmentation. Other than the link to the privacy policy, the preprocessing requires no other information or prior knowledge.

Policy Extraction. Given the URL of the privacy policy (provided by the administrator of PriBot for instance), the segmenter scrapes the web page containing the privacy policy using a headless browser (we use PhantomJS in our implementation). Then, it removes all irrelevant HTML elements including the scripts; header, footer, side and navigation menus; comments; CCS and header texts ($<h1>$ to $<h6>$).

Second, the segmenter handles the lists inside the policy. Recalling that segments constitute the answers pool of PriBot, lists require a special treatment as counting an entire list as a segment could result in overly long segments. On the other hand, treating each list item as an independent segment is problematic as list elements are typically not self-contained. Fig. 6.3 shows a list from Google's privacy policy¹. Obviously, each of the inner list items (shaded) cannot function as a standalone segment.

Our handling of the lists involves two techniques, one for small list items (the inner list of Fig. 6.3) and another for longer list items (the outer list of Fig. 6.3). The segmenter traverses every list in the policy ($$ tags). If the largest list element is less than 20 words, then the segmenter combines the elements with the introductory statement of the list, thereby converting the list ($$ element) into a paragraph ($<p>$ element). This is highlighted by the first two steps of Fig. 6.3. For those lists with long items, the segmenter transforms the list into a set of paragraphs. Each paragraph is a distinct list element prepended by the list's introductory statement as highlighted by the third step in Fig. 6.3. The resulting paragraphs constitute better standalone answers compared to simply taking the list items out of their context.

¹<https://www.google.com/policies/privacy>, last modified on Aug. 29, 2016 and retrieved on Feb. 14, 2017

Policy Segmentation. The segmenter performs initial and coarse-level segmentation by breaking down the policy according to the HTML `<div>` and `<p>` tags. The `<p>` elements include paragraphs in the original policy and the newly constructed ones from the lists. The output of this step is the initial set of segments. Some of resulting segments might still be long; the segmenter further breaks them down by passing each of them to an out-of-the-box text segmenter.

Our choice for the text segmenter is GraphSeg [GNP16], a recently proposed unsupervised algorithm that generates semantically coherent segments. GraphSeg is an attractive choice since it makes no assumptions about the structure of the input text. It relies on word embeddings to generate segments as cliques of related (semantically similar) sentences. To employ GraphSeg, our custom word embeddings are generated by training the fastText language model [BGJM16] on the Combined Policies Dataset. Finally, the segmenter outputs a series of fine-grained segments.

6.4 Question-Answering Approaches

The primary challenge of PriBot is to develop QA ranking algorithms that match the user's question with relevant and accurate answers. The complexity of this task can be motivated by the question: *"To whom do you expose my content?"* While simple, this sentence embodies four issues to be considered while developing the QA ranking algorithms. First, the terms in the question might not occur at all in the privacy policy, despite being common in everyday usage. Second, even if a term like "content" occurs in the policy, it does not indicate the general topic (e.g., whether the context is about third-party sharing or first party collection). Third, words tend to be used differently in the question's context than a policy's context. Policies typically use a pronoun like "you" to indicate choices the user has and "we" to indicate what it does; users mention "you," in contrast, to refer to the company. Finally, users typically pose questions in general terms, without specifying their intent in exact words. The answers should not exhibit the same level of generality as to not miss the information of interest to the user.

With these challenges in mind, we developed three QA ranking approaches. We start with our unsupervised approach (called Retrieval), inspired by the state-of-the-art in term-matching retrieval algorithms. Retrieval constitutes the baseline with which we compare the rest of our QA approaches. We then describe SemVec, our approach that uses word embeddings and neural networks, to bridge the semantic similarity gap between questions and the policy segments. Finally, we present our further optimized approach, Hierarchical, that utilizes a hierarchy of classifiers, accounting for the discrepancy between questions' and answers' structures.

6.4.1 Problem Formulation

The input consists of a user question Q about a privacy policy P . PriBot first generates a pool of candidate answers $\{A_1, A_2, \dots, A_M\}$ by segmenting the policy (*cf.* Section 6.3). Each question comprises the terms $\{q_1, q_2, \dots, q_n\}$. Similarly, an answer is composed of the terms $\{a_1, a_2, \dots, a_n\}$. A subset \mathcal{G} of the answer pool constitutes the ground-truth. We consider an answer A_k as *correct* if $A_k \in \mathcal{G}$. We consider A_k as *incorrect* if \mathcal{G} is not empty and $A_k \notin \mathcal{G}$. If \mathcal{G} is empty, we denote the case as *unanswerable*. This can happen when the answer to the question does not exist in the privacy policy. We formulate the problem as a one-shot question-answering problem, in which PriBot returns the best answer without requesting further question refinements from the user. This formulation does not preclude conducting a dialog with our system, provided that users pose standalone questions. The ranking algorithms introduced below, rank each potential answer (policy segment), A , by computing a proximity score $s(Q, A)$ between A and each question, Q . Before passing the question and candidate answers to the ranking algorithms, PriBot processes them with traditional tokenization techniques (Treebank tokenizer), converts their text into lower case and replaces the numbers with words.

6.4.2 Baseline Retrieval Approach

To obtain a reasonable baseline to evaluate our QA algorithm against, we developed a ranking mode that builds on the BM25 algorithm [Rob04], which represents the state-of-the-art in ranking models based on term-matching. It has been employed successfully across a range of collections and search tasks, such as the TREC evaluations [BGH⁺97]. Given both a question Q (composed of terms q_i) and an answer A , we compute the score $s(Q, A)$ as:

$$\sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, A) \cdot (k_1 + 1)}{f(q_i, A) + k_1 \cdot \left(1 - b + b \cdot \frac{|A|}{\text{avgal}}\right)}, \quad (6.1)$$

where $f(q_i, A)$ is the frequency of term q_i in A , $\text{IDF}(q_i)$ is the inverse document frequency computed as $\log(N/\text{df}(q_i))$, N is the total number of answers, $\text{df}(q_i)$ is the number of answers containing q_i , and avgal is the average answer length (in words). The default parameters were taken as $k_1 = 1.6$ and $b = 0.75$.

This algorithm is highly dependent on the presence of distinctive words in the question that link it to the answer but has the advantage of being unsupervised. In other words, it does not require training over a dataset of annotated privacy policies. We take a step to leverage this by pre-computing the IDF value for each word using a large corpus of unlabeled data. In particular, we segment the Combined Policies Corpus of Section 6.2.2 to automatically generate answers. Then, we use those answers to compute the IDF, instead of solely relying on one policy's answers. With this, Retrieval provides an even stronger baseline against which we can evaluate more elaborate approaches.

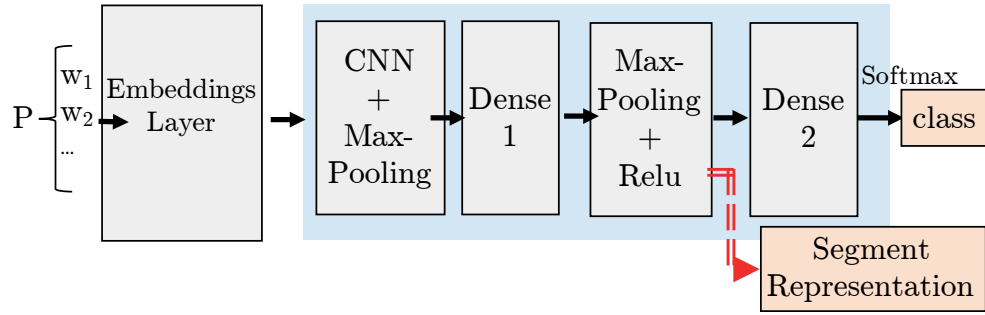


Figure 6.4 – CNN classifier building block.

6.4.3 Neural Network-based QA

Our second class of approaches relies on supervised learning, which leverages the OPP-115 dataset, to address the challenges we described earlier (related to term matches and usage). The idea consists of training a supervised model on a *proxy task* and then using this trained model as the basis for a QA algorithm. We employ neural network models, which have been shown to be effective in matching *non-factoid* questions with answers [FXG⁺15, TdSXZ16]. Our rationale behind using neural networks is bridging the semantic gap between users and privacy policies. This can enable users to pose free form questions, without requiring the questions to correspond to terminology used in the privacy policies. In what follows, we present our two QA approaches that are based on neural networks: SemVec and Hierarchical.

Handling the Vocabulary Discrepancy

Traditional QA approaches assume that a representative set of questions and answers are both available at training time. In our setting, the distribution of the vocabulary terms that people use in their questions differs from that of the privacy policies' vocabulary. To handle this issue, we take advantage of the generalization power of *word embeddings*, which are vectors representing the “embeddings” of English words in the d_i -dimensional and continuous vector space. With these vectors, words of similar semantics, such as “delete” and “erase” are close in the embeddings space. We use these vectors as inputs to the neural network. Note the departure from the Term-Matching approach where the features are the terms themselves.

Specifically, we utilize the GloVe vectors [PSM14], which include 400,000 terms, pre-trained on a large and general corpus, consisting of the whole English Wikipedia and the Gigaword corpus. We made the decision to fix the embeddings of the input words during the training process. This is to account for words that do not occur in the specific vocabulary of the training set (i.e., the privacy policies in our case). If the embeddings were set to be trainable, words like “delete” and “erase” might no longer be close in the embeddings space if “erase” does not occur enough in the privacy policies. Another advantage of the word embeddings is that they indirectly mitigate the issue of discrepancy in pronouns usage in the questions vs. the policies. Pronouns are typically close in the embeddings space. Hence, replacing the pronoun “you” in

a sentence by another pronoun “I” will not shift the combined sentence representation.

Segment Classification with CNNs

A building block of our two approaches is a multi-class classifier that uses convolutional neural networks (CNNs). This block is driven by the OPP-115 dataset, which is rich in class information for each segment of the policy. We utilize the classification network as a “proxy task” that will help us obtain a rich representation of the segments in the policy. An alternative building block for our QA problem would have been a proxy task directly trained to evaluate “phrase similarity.” In that case, the training data can be constructed by labeling two phrases as “similar” if the skilled annotators gave them the same attribute-value label. For example, the two phrases “*better able to adapt our services and provide you with a better experience*” and “*internal market research to help us better serve you*” were both labeled as {Purpose: Analytics-Research}. The phrase similarity task assumes, however, that phrases are single-labeled. In practice, the phrases often contain a mixture of attribute-values, such as mentioning both the type and purpose of information collection in one sentence.

The architecture we use for this building block is presented in Fig. 6.4. An input segment is composed of terms, which are represented using their embeddings (this happens at the embeddings layer). Next, the embeddings pass through a Convolutional layer. This layer applies a nonlinear function (a rectified linear unit (ReLU) in our case) to each window of k words in a phrase. Thus, it transforms each k -words into a d_c -dimensional vector, which accounts for the co-occurrences of words in this window. The max-pooling layer combines these vectors by taking the maximum value observed in each of the d_c channels over all the windows (to detect the most important features). This vector passes through the first dense (i.e., fully connected) layer, which is again followed by a max-pooling operation and a ReLU activation function. Finally, the vector arrives at the second dense layer. A *softmax* operation is applied to the output of this layer to obtain probability distribution across the possible output classes. This architecture is a simplified variant of the sentiment classification network introduced by Kim [Kim14]. We particularly reduced the network parameters to account for the smaller dataset we have, thus improving the classifier’s performance.

Semantic Vector QA Approach (SemVec)

Our deep-learning model builds on the previously introduced CNN classifier introduced in Section 6.4.3. We train that classifier with the segments from the OPP-115 dataset, labeled with attribute values. An example segment is “*geographic location information or other location-based information about you and your device*” that was labeled as {Information Type: Location}. We take all the attribute-values as a single pool and train the classifier to distinguish among them.

After training this model, we extract a “*semantic vector*”, which is a representation vector that

Table 6.1 – SemVec parameters and aggregate metrics

Parameter	Value	Metric	Av. Value
Embeddings size	200	Precision	0.56
Num. of filters	750	Recall	0.54
Filter size	3	F1	0.54
Dense Layer size	200	Support	200
Batch size	40		

accounts for the distribution of value labels in the input text. We extract this representation as the vector at the output of the ReLU activation layer preceding the second dense layer (as shown in Fig. 6.4). As we use a classification task for training, phrases with similar classification distribution exhibit similar segment representations.

This representation further accounts for the co-occurrence of the same words in multiple contexts, despite the fact that the dataset is not designed for multi-label classification. The semantic vector allows us to rank the similarity between a question (denoted by vector r_Q) and an answer (denoted by vector r_A) in our QA system using the Euclidean similarity, defined as $1/(1 + \|r_A - r_Q\|_2)$. This approach has been previously shown to be successful in the image retrieval domain, where image representations learned from a large-scale image classification task were effective in visual search applications [RSCM14]. However, we are the first to apply it for analyzing legal text, such as privacy policies.

Model Tuning. In order to obtain the semantic vectors, we proceeded in training our model. The training data for the CNN classifier involves around 15,000 (*segment, label*) tuples, extracted from the annotated OPP-115 dataset. We reserve 9,000 tuples as a validation set. The value-labels fall into 81 classes as we select classes with a minimum of 10 labels in each. We tuned the parameters of the classification network by running a grid-search and selecting the best-performing model. The selected parameters along with the aggregate statistics are presented in Table 6.1. In Table 6.2, we present the results of value-level classification performed across 81 classes. For space constraints only the first 50 classes are shown. We also show in Figure 6.5 the confusion matrix between all the predicted and the true labels. We ran stratified cross validation with five folds on the whole dataset to obtain the final results. We obtain a micro-average precision of 0.56 (i.e., the classifier is on average predicting the right label across the 81 classes in 56% of the cases – compared to 3.6% precision for a random classifier). Although we do not rely on the classification outputs per se, this performance is crucial for the adequacy of the segment representations we obtain.

Hierarchical Classification Approach

One limitation of the SemVec approach discussed in Section 6.4.3 is that it is trained to detect similarities between policy excerpts representing specific data practices, without accounting

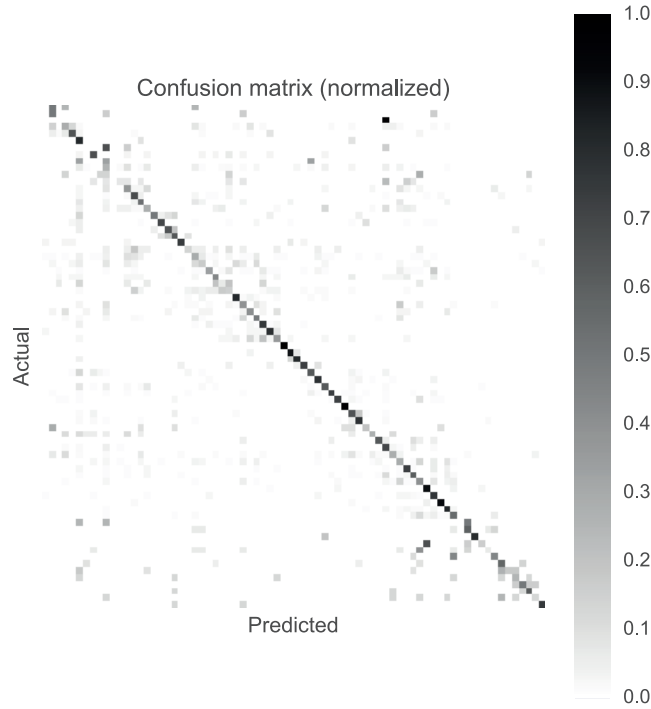


Figure 6.5 – Confusion matrix for the value-level classifier among the 81 classes.

for high-level categories. For example, consider the following two privacy policy excerpts:

1. “third-party application providers may automatically collect real-time *geographic location information or other location-based information about you and your device* . . .”
2. “we automatically collect and store certain information (...) including: (...) *your mobile device’s geographic location (specific geographic location if you’ve enabled collection of that information . . .)*”

The annotators labeled the italicized parts of those excerpts as {Information Type: Location}. However, the whole segment that includes those excerpts was labeled as 3rd Party Collection in the first case and as 1st Party Collection in the second case. To tailor for these differences, we propose Hierarchical, a new model that accounts for both the high-level categories and the finer-grained attribute values.

Hierarchical consists of building a hierarchy of classifiers that are individually trained on handling specific parts of the problem, as shown in Fig. 6.6. Given a segment x and its labels from the OPP-115 dataset, we first train a classifier that provides us with a probability distribution across the high-level categories C (i.e., $P(C|x)$). This classifier uses the same model we presented in Fig. 6.4. Next, we have a set of classifiers that are trained on the value-level; each classifier is trained to classify among the values of a single attribute. For exam-

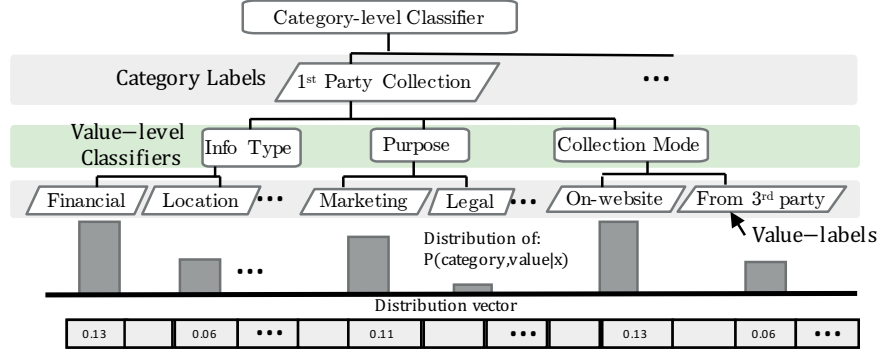


Figure 6.6 – Overview of Hierarchical QA Approach.

ple, we have an Information Type classifier that provides a probability distribution across the values: financial, location, survey data, user profile, health, personal identifiers, demographics, cookies, contact information, device IDs, online activities, and computer information. More formally, we have the set $V = \{V^1, \dots, V^K\}$ of all possible subsets of values, and the classifiers will produce $P(V^j|x) \forall V^j \in V$, where V^j is the set of values under a given attribute.

As we train the two classifiers (category- and value-level) independently, we obtain the joint probability distribution as $P(C, V|x) = P(C|x) \cdot P(V|x)$. This distribution accounts for both the distribution across the value labels as well as for the high-level categories.² The similarity score between two segments is computed as the proximity between these distributions. For this, we rely on the Hellinger distance, which quantifies the similarity between two probability distributions $J = (j_1, \dots, j_k)$ and $L = (l_1, \dots, l_k)$,

$$H(J, L) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{j_i} - \sqrt{l_i})^2}. \quad (6.2)$$

We compute the similarity between a question Q and an answer A as $1/(1 + H(P_A, P_Q))$, where with $P_A = P(C, V|A)$ and $P_Q = P(C, V|Q)$.

Model Tuning For training the value-level classifiers, we used similar parameters to those of Table 6.1. For the high-level classifier, we tune the parameters using grid-search. We used a training set of 4,700 categories labels of segments from the OPP-115 dataset and a tuning set of 1,170 labels. Similar to Wilson et al., OPP-115 dataset [WSD⁺16], we consider the attributes of the *Other* category in Fig. 6.2 as separate classes. Compared to Table 6.1, the best performing model had 2,000 filters and a dense layer of size 100. We report the classification results in Table 6.3. As we do not have access to the actual split between the training and the testing sets in the OPP-115 paper [WSD⁺16], we report the results based on five folds stratified cross-validation, and we obtain a micro-average precision of 0.75. By comparison,

²The product is only computed for attribute-value labels descending from each category.

the best performing SVM model trained on word embeddings in the OPP-115 paper [WSD⁺16] had a micro-average precision and recall of 0.66.³ Therefore, our CNN-based classification approach significantly improves performance over prior work.

Addressing the generality issue. As we described earlier, users' questions can often be generic. Matching such questions with general answers is not desirable as these answers would be from introductory or generic sentences, lacking the specific details. We address this issue on the levels of category and value labels. First, we exclude clearly generic answers from consideration. This is done by removing answers that are classified into the `Introductory/Generic` category with a probability exceeding a threshold t (we choose $t = 0.3$ in our evaluation). These answers typically do not contain relevant details to users' questions and come at the top or the end of the policy. Second, we benefit from the generic labels at the value level in the OPP-115 dataset. For example, (e.g., `Generic Personal Information` is the annotated label when the particular information type is unspecified). Our strategy is to exclude such classes at training time rather than prediction time. We found that this forces the question and the answer to be classified under more precise labels, even when the question is generic. That way, general questions are more likely to be matched with specific answers. This does not affect the case when the question is already labeled with specific labels as these are already matched together.

³Note that although the testing set was not available, we evaluated our model with different random sets of similar sizes and the performance was consistently superior.

Table 6.2 – Classification results for the value-level classifier across 81 classes (first 50 are shown)

Category	Prec.	Recall	F1	Support
action-third-party-receive-shared-with	0.83	0.66	0.73	589
action-first-party-collect-on-website	0.43	0.30	0.36	509
purpose-basic-service-feature	0.48	0.48	0.48	466
personal-information-type-contact	0.84	0.77	0.80	441
personal-information-type-generic-personal-information	0.70	0.59	0.64	414
purpose-advertising	0.80	0.62	0.70	407
purpose-additional-service-feature	0.41	0.24	0.30	395
purpose-marketing	0.63	0.68	0.65	367
personal-information-type-cookies-and-tracking-elements	0.56	0.73	0.64	365
purpose-analytics-research	0.66	0.57	0.61	358
personal-information-type-user-online-activities	0.61	0.57	0.59	301
purpose-service-operation-and-security	0.62	0.57	0.60	251
choice-type-opt-in	0.42	0.52	0.46	221
purpose-personalization-customization	0.67	0.72	0.69	203
choice-scope-collection	0.20	0.29	0.24	188
personal-information-type-ip-address-and-device-ids	0.46	0.61	0.52	174
choice-scope-first-party-use	0.30	0.34	0.32	163
choice-type-dont-use-service-feature	0.27	0.29	0.28	157
personal-information-type-location	0.60	0.60	0.60	154
purpose-legal-requirement	0.75	0.81	0.78	142
personal-information-type-computer-information	0.67	0.71	0.69	142
personal-information-type-financial	0.72	0.70	0.71	139
choice-type-opt-out-link	0.62	0.50	0.56	130
access-type-edit-information	0.75	0.66	0.70	120
choice-type-browser-device-privacy-controls	0.78	0.70	0.74	113
audience-type-children	0.89	0.79	0.84	112
security-measure-generic	0.64	0.82	0.72	110
personal-information-type-demographic	0.74	0.58	0.65	110
access-scope-user-account-data	0.29	0.17	0.21	101
choice-scope-use	0.19	0.25	0.22	97
action-third-party-collect-on-first-party-website-app	0.15	0.24	0.19	79
personal-information-type-user-profile	0.22	0.39	0.28	77
choice-type-opt-out-via-contacting-company	0.51	0.64	0.57	77
action-third-party-track-on-first-party-website-app	0.16	0.34	0.22	77
choice-scope-first-party-collection	0.15	0.13	0.14	69
choice-scope-third-party-sharing-collection	0.44	0.40	0.42	68
purpose-merger-acquisition	0.66	0.94	0.78	63
choice-type-third-party-privacy-controls	0.41	0.24	0.30	59
change-type-privacy-relevant-change	0.74	0.73	0.73	55
notification-type-general-notice-in-privacy-policy	0.62	0.78	0.69	51
choice-scope-both	0.14	0.10	0.11	51
choice-type-first-party-privacy-controls	0.19	0.42	0.26	48
action-first-party-collect-in-mobile-app	0.34	0.58	0.43	48
action-third-party-see	0.25	0.40	0.30	43
personal-information-type-personal-identifier	0.23	0.60	0.34	42
notification-type-personal-notice	0.79	0.50	0.61	38
audience-type-californians	0.77	0.69	0.73	35
notification-type-general-notice-on-website	0.69	0.71	0.70	34
security-measure-secure-data-transfer	0.77	0.61	0.68	33
security-measure-data-access-limitation	0.48	0.39	0.43	33
Average	0.56	0.54	0.54	8956

Table 6.3 – Classification results at the category level.

Category	Prec.	Recall	F1	Support
1 st Party Collection	0.77	0.82	0.79	328
3 rd Party Sharing	0.76	0.82	0.79	268
User Choice/Control	0.68	0.73	0.70	101
Introductory/Generic	0.79	0.65	0.71	136
Data Security	0.77	0.82	0.80	62
Specific Audiences	0.73	0.72	0.72	85
Privacy Contact Info	0.71	0.71	0.71	56
Access, Edit, Delete	0.72	0.53	0.61	40
Practice Not Covered	0.60	0.32	0.42	37
Policy Change	0.81	0.88	0.84	33
Data Retention	0.75	0.21	0.33	14
Do Not Track	0.78	1.00	0.88	7
Average	0.75	0.75	0.74	1167

6.5 Evaluation Methodology & Dataset

We assess the performance of PriBot’s QA ranking approaches from two angles: the *predictive accuracy* (Section 6.6) of the QA models and the *user-perceived utility* (Section 6.7) of the provided answers. This is motivated by research around recommender systems evaluation, where the model with the best accuracy is not always rated to be the most helpful by the users (see the work by Knijnenburg *et al.*, [KWH10]).

In particular, we evaluate PriBot using a set of real-world, user-posed, and privacy-related questions, collected from Twitter. These questions represent user-company interactions about privacy practices. This approach has the advantage of avoiding subject bias, which is likely to happen when eliciting privacy-related questions from individuals, who will not pose them out of genuine need. We crawled Twitter to obtain questions that represent realistic privacy-related concerns of users, in their words. This collection methodology allows us to achieve high ecological validity as we collect real-world privacy questions from real user interactions with companies in a medium in which PriBot could be employed and without explicitly soliciting users’ inputs via a survey.

Instead of directly searching for questions on Twitter, we searched for reply tweets that direct the users to a company’s privacy policy (e.g., using queries such as "filter:replies our privacy policy" and "filter:replies we privacy policy"). We then backtracked these reply tweets to the (parent) question tweets asked by customers to obtain a set of 4,743 pairs of tweets. As customary in computational social science research [OLT16], we distill this initial dataset, filtering the noise via heuristics, to reduce the human labeling effort. In our case, we filtered the question tweets to keep those containing question marks and at least four words (excluding links, hashtags, mentions, numbers and stop words). We also selected the pairs where the reply tweet included a valid link. This link is almost always directing users to the privacy policy, which automates our answer generation during the evaluation of PriBot. This stage resulted in 260 pairs of valid question-reply tweets.

Next, the author and another member of the research team manually annotated each of the tweets in order to remove question tweets (a) that are not related to the privacy policies, (b) to which the replies are not from the official company account, and (c) with inaccessible privacy policy links in their replies. The level of agreement (Cohen’s Kappa) among both annotators for the labels *suitable* vs. *unsuitable* was almost perfect ($\kappa = 0.84$) [LK77]. The two annotators agreed on 231 of the question tweets, tagging 182 as *suitable*. As we will evaluate the answers to these questions with a user study, our estimates of an adequately sized study lead us to randomly sample 120 tweets out of the tweets labeled as *suitable*. We provide these tweets in Appendix C, and we henceforth refer to them as the *Twitter QA Dataset*. The size of our question test set ($\# = 120$) is consistent with other QA evaluation contexts. For example, the Text REtrieval Conference (TREC) included a question answering track each year (till 2007) to compare the performance of different QA systems on a given question set. The number of test questions of a particular type (e.g., those of list or definition types) is usually

less than one hundred [VB03, DKL07].

6.6 Accuracy Evaluation

In this section, we evaluate the *predictive accuracy* of the QA models by comparing their predicted answers against expert-generated ground-truth answers for the questions of the Twitter QA Dataset from Section 6.5. We consider four evaluation conditions. The first three correspond to the three QA approaches (Retrieval, SemVec and Hierarchical) and the fourth to a control approach, Random. In the latter, questions are answered with randomly chosen segments from the policy.

6.6.1 Ground-truth Generation

To generate the ground-truth, skilled annotators, represented the author and another member of the research team, were given the user's question (tweet), and the segments comprising a policy (generated from the privacy policy URL in the answer tweet based on Section 6.3). Each policy consists of 45 segments on average ($min=12$, $max=344$, $std=37$). Each annotator selected, *independently*, the subset of these segments which they consider as best responding to the user's question. This annotation took place *prior* to PriBot generating its answers to avoid any bias. While deciding on the answers, the annotators accounted for the fact that multiple segments of the policy might answer a question. For generic or ambiguous questions, they included, in the answer set, both the general and specific segments that address the question. Moreover, given that the annotation is a highly time-demanding task (takes around 6 minutes per question per annotator, excluding consolidation time), we limit the labeling to 60 questions (i.e., to around 2700 segments in total).

After finishing the individual annotations, the two experts met and consolidated the differences in their labels to reach an agreed-on set of segments, each assumed to be answering the question. We call this the *ground-truth* set for each question. The annotators agree on at least one answer in 88% of the questions for which they found matching segments, thus signifying a substantial overlap. For questions with the ground-truth containing two or more answers, the annotators agreed on at least two of these answers in 76% of the questions.

6.6.2 Accuracy Results

We generated, for each question, the predicted ranked list of answers according to each QA model. In what follows, we describe our approach for evaluating the predictive accuracy of these models.

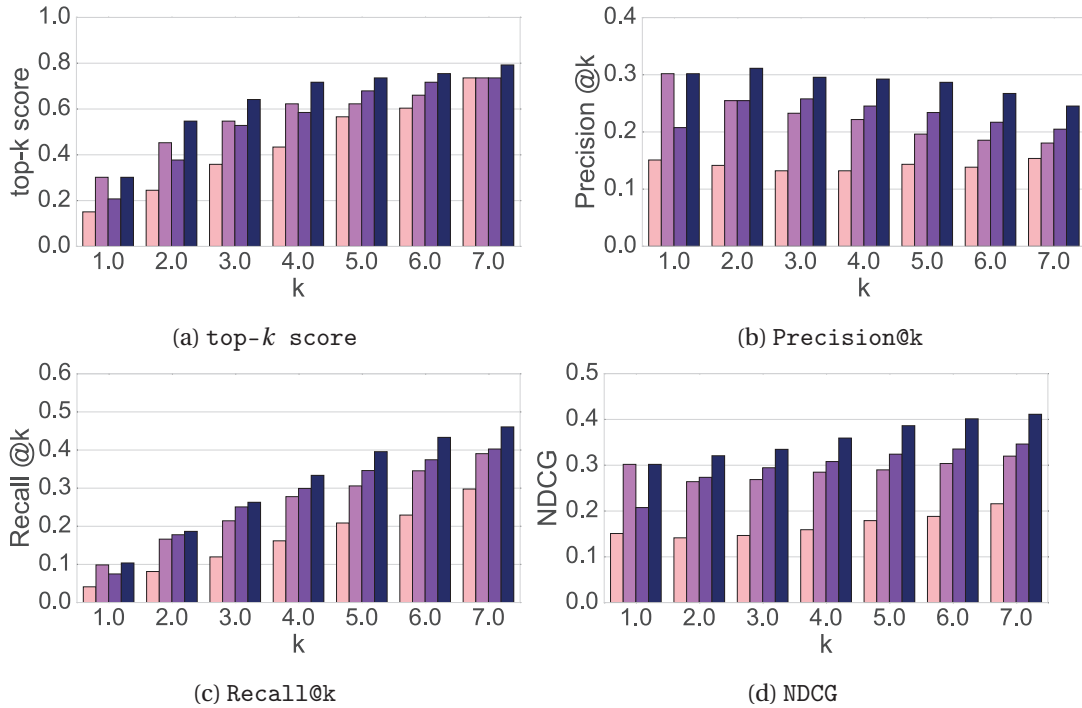


Figure 6.7 – Metrics for various values of k , with the color code being:

Random Retrieval SemVec Hierarchical

QA Metrics

We start by reporting a widely used and easily interpretable metric, the top- k score, which denotes the portion of questions having at least one correct answer in the top k returned answers. The top- k score has a direct usability implication: the higher it is for low values of k , the less information the user has to process before reaching a correct answer. We show in Fig. 6.7a how the top- k score varies as a function of k . Hierarchical model has the best performance over the other three models. It achieves 0.30, 0.55, 0.64, and 0.72 for k varying from 1 to 4. As k increases, the differences between the various models become less significant. At $k = 7$, the Random model is already as good as the Retrieval and the SemVec models, which highlights the importance of achieving higher top- k score at lower k . We also observe that the Retrieval and SemVec models are trading places for the second best performing model, with the Retrieval model being better than SemVec at $k = 1$. This is not entirely surprising. We seeded Retrieval, which is based on term-matching, with a large corpus of unsupervised policies, thus improving its performance on answers with matching terms. However, it falls short when retrieving the other answers with semantic similarity (those with non-matching terms) as evident from the higher values of k .

To confirm this observation, we evaluated Precision@ k (i.e., average proportion of retrieved

Table 6.4 – MAP and MRR metrics with the expert evaluation.

	Random	Retrieval	SemVec	Hierarchical	Emb.Variant
MAP	0.23	0.33	0.34	0.40	0.28
MRR	0.32	0.47	0.40	0.50	0.35

top- k answers that are correct ⁴) in Fig. 6.7b and the Recall@ k (average proportion of correct answers that are retrieved in the top- k) in Fig. 6.7c. As evident from both plots, the SemVec model exhibits a clear advantage over Retrieval with respect to both metrics for $k > 1$. Hence, our decision to consider the semantic similarity between questions and answers pays off as it exhibits a superior performance over term-matching. It is also clear from both plots that Hierarchical performs the best among the QA approaches.

Policy Length

The previous metrics (namely top- k score, Precision@ k , and Recall@ k) suffer from two shortcomings. They do not capture how presenting the users with more choices affects their user experience as they need to process more text. Also, because of their nature, these metrics attain higher values for short policies that have few potential answers. A better metric that accounts for both shortcomings is the *Normalized Discounted Cumulative Gain (NDCG)* [JK02]. Intuitively, it indicates that a relevant document’s usefulness decreases logarithmically with the rank. The DCG_k part is computed as $DCG_k = \sum_{i=1}^k \frac{rel_i}{\log_2(i+1)}$, where rel_i is 1 if answer $A[i]$ is correct and 0 otherwise. NDCG at k is obtained by normalizing the DCG_k with the maximum possible DCG_k across all values of k . We show in Fig. 6.7d the average NDCG across questions for each value of k . It is clear that the Hierarchical model consistently exhibits superior NDCG and that the SemVec approach is clearly advantageous over the Retrieval model for $k > 1$. This indicates that the neural networks models are poised to perform better in a system where low values of k matter the most. To further focus on the effect of policy length, we categorize the policy lengths (#segments) into *short*, *medium*, and *high*, based on the 33rd and the 66th percentiles (we split at #segments of 28 and 46). We then compute a metric independent of k , namely the Mean Average Precision (MAP), which is the mean of the area under the precision-recall curve across all questions. Informally, MAP is an indicator of whether all the correct answers tend to get ranked highly. We see from Fig. 6.8 that, for short policies, our 3 main models perform very closely, which makes sense given the smaller number of potential answers. With medium-sized policies, the Hierarchical model is better by a large margin. This margin is still considerable with long policies, where Retrieval model becomes worse than Random model and SemVec model as good as Random model.

⁴We note that the absolute value of Precision@ k is inherently low as it depends on the total number of correct answers per policy.



Figure 6.8 – Variation of MAP across policy lengths.

Pre-trained Embeddings Choice

As discussed in Section 6.4.3, we utilize the GloVe word embeddings, which are trained on 6 billion tokens. Our motivation is that they can capture the relations between layman terms used in everyday language and legal terms present in privacy policies. We favored that over word embeddings built using privacy policies alone, due to the limited size of available policies' corpora (which are at least 3 orders of magnitude smaller). To validate this decision, we evaluated the performance of a variant of the Hierarchical model that uses the word embeddings of Section 6.3, trained on the 16,615 privacy policies we obtained. We show in Table 6.4 how this model, named *Emb.Variant* fares in comparison to the other models. In addition to the MAP metric, we also computed the Mean Reciprocal Rank (MRR), as $MRR = \frac{1}{M} \sum_{j=1}^M \frac{1}{rank_j}$, where $rank_j$ is the rank of the first correct answer for question j and M is the total number of questions. Table 6.4 shows that this *Emb.Variant* model performed worse than all but the Random model. We also evaluated the performance of the classification model of Section 6.4.3 with those policies' embeddings. We noticed that their performance approaches the case of GloVe embeddings, scoring 0.73, 0.74, and 0.74 on the precision, recall, and F1 metrics respectively (1%-2% lower compared to Table 6.3). This shows that the use of the more generic, but also more expansive, GloVe embeddings rather than privacy policy specific embeddings results in a more robust overall performance on the QA task, likely due to the larger size of data it is trained on.

6.7 User Study

We conducted a user study to assess the *user-perceived utility* of PriBot's answers. Our methodology is simple; for each QA pair, we collect the evaluations from 10 different individuals (via Amazon MTurk). We repeat this assessment for each of the four different conditions (Retrieval, SemVec, Hierarchical and Random). We evaluated the top 3 responses of each QA approach to each question. Thus, we assess the utility of 360 answers to 120 questions (randomly chosen from the Twitter dataset). With four modes, we have 1,440 QA pairs, for which we obtained 17,790 user evaluations.

Study Design

We used a between-subject design by constructing four surveys, each corresponding to a different evaluation condition. We provide the study materials in Appendix C. The survey starts with a series of demographics questions. Second, we include an open-ended Cloze reading comprehension test (based on [Cra13] and employed by Wilson *et al.* [WSR⁺16] and Reidenberg *et al.* [RBC⁺15]). The test consists of an English paragraph with five missing words. The respondents fill (by typing in) each blank with a single word best fitting the context. We utilize this test to assess the reading level of participants and weed out responses failing in more than 2 blanks.

Next, we display a series of 17 QA pairs (each on a different page). Of these, 15 are a random subset of the pool of 360 QA pairs (of the evaluated condition); a participant does not receive two QA pairs with the same question. The other two are randomly positioned anchor questions serving as attention checkers. The first corresponds to a QA pair with a clearly relevant answer, and the second has a clearly irrelevant answer. We shuffle the order of the QA pairs per user, to account for ordering effects (participant fatigue and practice). Additionally, we enforce a minimum duration of 15 seconds for the respondent to evaluate each QA pair.

Participant Recruitment

After obtaining IRB approval, we recruited participants using Amazon Mechanical Turk (MTurk). More than 90% of the respondents were from North America (U.S. and Canada). We limited the respondent pool in MTurk to those with 95% success rate in their previous tasks. Across all the conditions, the average completion time of the survey was 14 minutes. In total, 1,186 individuals participated in our study. We compensated each respondent with \$1. We show the breakdown per group in Table 6.5. We limit the Random group to 60 questions (180 QA pairs) and collected five responses per QA pair. While not fully representative of the general population, our set of participants exhibited high intra-group diversity, but little difference across the respondent groups. For all respondents, the average age is 37 years ($std=12$), 59% are males, 41% are females, more than 90% have some level of college education and more than 90% reported being employed (less than 15% are teachers or students).

QA Pair Evaluation

To evaluate the relevance for a QA pair, we display the question and the candidate answer as shown in Fig. 6.9. In our pilot studies, the respondents were confused about the context of the question as it might depend on the company. Hence, we decided to display the company's name and Twitter Bio, which resulted in a better understanding of the context. Although there is a potential bias due to the perceived quality of answers to questions pertaining to well-known brands, this should not favor one QA approach over another as the same set of questions is presented in the four conditions. Moreover, many of the tweets (questions) and

Question: .@AskTarget ok thanks but I assume that means yes you all do sell patient names and addresses?

Answer: We may share your personal information with other companies, or organizations which are not part of Target. These companies and organizations may use the information we share to provide special offers and opportunities to you. To opt out of our sharing of your personal information with such companies and organizations, go to the Choices section of this privacy policy.

How relevant is the candidate answer to the given question?

- ☐ **Definitely Relevant:** It perfectly answers the question.
- ☐ **Partially Relevant:** It answers the bulk of the question, though there might be more to say.
- ☐ **Undecided:** I find it too difficult to give a judgment on this pair.
- ☐ **Partially Irrelevant:** It doesn't answer the question; only has a slight clue.
- ☐ **Definitely Irrelevant:** It totally misses the topic of the question.

Figure 6.9 – An example of a QA pair.

Table 6.5 – A breakdown of the top- k relevance score by evaluation group (N is the number of participants per group).

Group	N	top- k Relevance Score		
		$k = 1$	$k = 2$	$k = 3$
Random	66	0.36	0.59	0.75
Retrieval	268	0.52	0.77	0.87
SemVec	271	0.50	0.75	0.84
Hierarchical	291	0.55	0.77	0.91

the policy segments (answers) already contained references to the company's name.

We asked the respondent to rate the relevance of the candidate answer to the question on a 5-point Likert scale (1=Completely Relevant to 5=Definitely Irrelevant). Fig. 6.9 depicts the explanation of each item on the Likert scale. We denote a respondent's evaluation of a **single** answer candidate corresponding to a QA pair as relevant if s/he chooses either Definitely Relevant or Partially Relevant. We consider the response as irrelevant if the respondent chooses Partially Irrelevant or Definitely irrelevant. Finally, we label the candidate answer of a QA pair as relevant when more than half of the individual evaluations for the pair are relevant. At this point, each QA pair from the four approaches is tagged as either "relevant" or "irrelevant".

User Study Results

As in the previous section, we compute the top- k score for relevance (i.e., the portion of questions having at least one relevant answer in the top k returned answers). Table 6.5 shows this score for the four QA approaches with $k = [1, 2, 3]$. We show in Table 6.6 the MAP and the

Table 6.6 – MAP and MRR metrics for user-perceived utility.

	Random	Retrieval	SemVec	Hierarchical
MAP (3 answers)	0.54	0.74	0.70	0.75
MRR (3 answers)	0.55	0.74	0.70	0.75

MRR metrics, which are independent of k . We compute these metrics by considering the top 3 answers returned by each model (the ones evaluated in the study).

For all the metrics, the Hierarchical approach clearly outperforms the rest. The respondents regarded at least one of the top 3 answers as relevant for 91% of the questions, with the first answer being relevant in 55% of the cases. The Retrieval model comes at a close second, edging the SemVec approach. Our explanation for this result is twofold. First, as we have seen in Section 6.6.2, the Retrieval model has a better predictive accuracy at $k = 1$, likely due to the notable portion of the answers which contain terms present in the question. For instance, this is the case when users ask about particular personal information, such as “Does anybody know if @EE sell on emails?”. Second, users are likely to perceive answers with matched terms more positively than answers with non-matched terms but with close meaning.

Still, the main takeaway from these results is that PriBot answers the users’ questions in a satisfactory manner to them, which highlights its practical significance.

6.8 Friendly Summary Generation

The evaluation shows that Hierarchical model outperforms other ranking models in terms of retrieval accuracy and perceived utility of responses. In addition, we note that the joint probability distribution across categories and values (Section 6.4.3) has a high-level interpretation. It informs us about the prominent classes present in each segment. We exploit this property in order to generate *abstractive answers* (i.e., user-friendly summaries) from the *extractive answers* we already have (from the Hierarchical approach). This model, termed as Friendly-summary, serve the users better by returning simpler and shorter answers.

Fig. 6.10 illustrates our approach. Given an existing answer and the computed joint probability distribution, we first remove the low-quality labels, the ones with probability lower than a threshold (equal to 0.05 in our implementation). Then, we group the labels under the high-level category they belong to, as shown in the figure. Next, for each high-level category, we generate a summary based on all the labels descending from it. This grouping serves to preserve the coherence of the generated content. The summary consists of one or more sentences and is generated based on our custom grammar. Our grammar consists of an optional introductory sentence about the high-level category, followed by a statement about each label. Statements describing the same attributes (e.g., information type) follow a similar structure; they start with an optional introductory phrase, followed by a specific phrase about

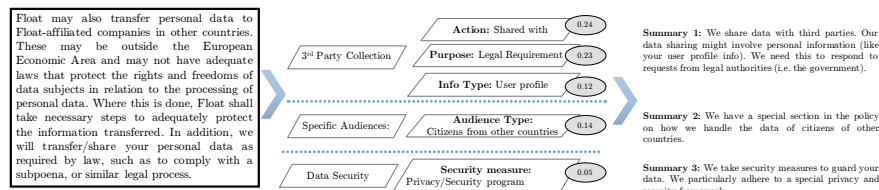


Figure 6.10 – The flow diagram of the Friendly-summary approach.

the value (e.g., user profile information). Example summaries are shown in Fig. 6.10. We rank these answers according to the label of highest joint probability present in them.

Friendly-summary has the benefit of responding back to the user in a simplified language while keeping the essence of the extracted answer. These properties are highly desirable in conversational settings, such as voice-activated devices. We evaluated the perceived utility of this approach with the same user study protocol and setup as Section 6.7 to compare against the extractive approaches. Based on the evaluation by 290 users, we obtained a top- k score of 0.52, 0.69, and 0.83 for a k of 1, 2, and 3 respectively. Comparing with the results of Table 6.5, we see that the top-1 score of Friendly-summary is the second after the Hierarchical approach. This is despite the fact that Friendly-summary produces generic answers instead of returning policy extracts. The answers are also 2.5 times shorter on average. Finally, the Friendly-summary achieves better readability with an 8.7 score on the Flesch-Kincaid Grade Level compared to 12.6 for the answers from the policy.

6.9 PriBot Implementation

To demonstrate the practical feasibility of our approach, we implemented PriBot as an online, text and voice-activated chatbot. As shown in Fig. 6.11, the user poses a question about the company of his/her choice to which the chatbot responds with the top-ranked answer according to the Hierarchical approach. In addition, the interface allows checking two more answers via an accordion interface, expandable via a user click. The chatbot also provides the option of narrating the answer to illustrate its functionality with voice-activated devices. We provide several examples of the answers returned by PriBot in Appendix D.

Chatbot System

Fig. 6.13 shows the flow of the chatbot. It follows a client-server architecture; the client consists of the user-facing interface and the back-end server consists of a chatbot server as well as a machine learning (ML) server. We implemented the chatbot server using *rivescript-js* to script the chatting logic and *Node.js* to interact with the ML server. The ML server employs the *Keras* neural network library (with a *TensorFlow* backend) for training and running the classifiers of the Hierarchical approach. We run the back-end server on an HP Z280 workstation with two bridged Nvidia Titan X GPUs. The chatbot server receives the question and the policy's

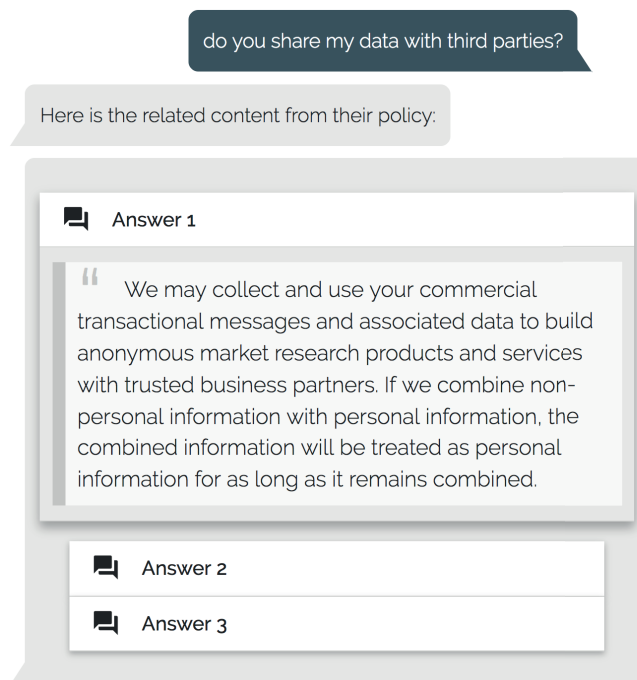


Figure 6.11 – A screenshot of PriBot

identifier from the client and passes them to the ML server. If the question belongs to a previously unseen privacy policy, the ML server runs the pre-processing logic of Section 6.3. Then, the ML server passes each segment of the policy through the `Hierarchical` classifiers to obtain their vector representations. The server caches each segment and its associated vector in *Redis*, an in-memory data store. This is a one-time effort per policy. If the question belongs to a pre-processed policy, the ML server passes the question through the same workflow of the segments to obtain its vector representation. Next, an answer ranking component determines the top three segments matching the user's question. The chatbot server returns these segments (the answers) to the client for display to the user.

Timing Measurement

To assess how practical our chatbot is, we measure its delay to generate a response and the time it takes to read out the response through a speech synthesizer. We measure the chatbot's answer generation delay for each of the 120 Twitter questions and the three QA approaches, `Hierarchical`, `SemVec` and `Retrieval`; the distribution is shown in Fig. 6.12a. Practically, all approaches return answers in almost real-time, with the `Hierarchical` taking around 120ms longer than the fastest model, which is likely an acceptable tradeoff given the higher accuracy and utility of the responses it provides. Compare that to the 6 minutes taken by the skilled annotator to decide on an answer.

Fig. 6.12b shows the distribution of the time for IBM Watson's Text to Speech Engine to narrate

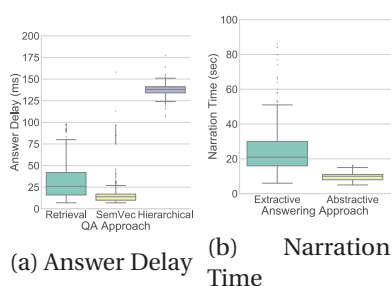


Figure 6.12 – Chatbot timing measurements.

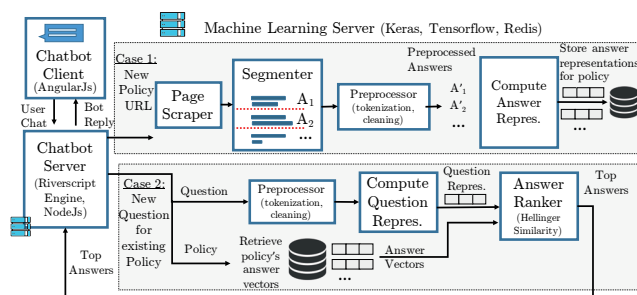


Figure 6.13 – The system flow diagram of our chatbot implementation.

a chatbot's answer. We compare between extractive approaches (Retrieval, SemVec and Hierarchical), which are based on the same segmentation process that returns excerpts of the policy, and the abstractive approach of Friendly-summary that constructs answer in a different way. We consider the top 3 answers for the 120 Twitter questions of the extractive approaches and compare them to those of the abstractive approach. The 75th percentile of the narration times is less than 30 seconds for the extractive methods and less than 10 seconds for the abstractive approach. This shows that the chatbot returns reasonably short responses for the user questions, making it also suitable for voice-based conversational systems.

6.10 Discussion

In this section, we discuss the implications and the challenges for the demonstrated ability to answer free-form questions from privacy policies.

Ecological Validity

Similar to studies involving users, we employed several decisions that could have impacted the evaluation of PriBot. First, we tested it in the wild with a social media dataset that is rife with informal language, complaints, cynicism and jokes. Some questions (e.g., questions 33 and 34 in Appendix C) contained spelling mistakes and Internet acronyms which resulted in returning wrong segments (we intentionally kept these cases). Note that research on adaptive language behavior in HCI [PHB⁺06] indicates that users tend to align their language to that of computer systems' with limited perceived capability. We conjecture that in a practical scenario, where users are aware they are communicating with an automated agent, their wording is expected to be different, and PriBot's performance even better. Thus, our evaluation is both realistic as well as conservative in assessing accuracy.

An avid reader might notice the differences between the predictive models' accuracy (Section 6.6) and the users' perceived quality (Section 6.7). This is, actually, consistent with the observations from research in recommender systems where the prediction accuracy does not

always match user's satisfaction [TMA⁺04, MAC⁺02, KWH10]. In addition, users are known to be biased by the very fact that they are evaluating a specific answer since it can be considered as a recommendation [KJ13]. We have tested this aspect by asking the experts to take the user study with the same 120 questions. We found that the expert-perceived utility of the models, based on the survey, was consistently higher than the predictive accuracy measured from the ground truth of Section 6.6 (a 5–10% difference). We can partially attribute this difference to the potential persuasive effect of the recommendations on the experts themselves.

UI Considerations

Like any QA automated system, PriBot may make mistakes in practice. However, a well-implemented UI can mitigate the negative effects of potential mistakes. One optimization is to show automatically classified labels (e.g., 1st party collection or data security) next to the answers' IDs. The user can then glance quickly at multiple answers and click on the one matching her intention. This allows providing more potentially correct answers without overloading the user. Another approach consists of PriBot asking the user for a clarification about the intended high-level category when there are multiple close matches. This also suits the voice-based UI. To improve the future predictions, PriBot actually includes an additional button for users to give feedback on a returned answer's quality. Furthermore, we hypothesize that “incorrect” answers from an evaluation perspective can be of educational value in practice, potentially covering other user concerns.

Another UI-related challenge is displaying potentially conflicting answers to the users. One answer could describe a general sharing clause while another specifies an exception (e.g., one answer specifies “share” and another specifies “do not share”). To mitigate this issue, we used the same CNN classifier of Section 6.4.3 and exploited the fact that the OPP-115 dataset had optional labels of the form: “does” vs. “does not” to indicate the presence or absence of sharing/collection. This classifier had a cross-validation F1 score of 95%. Hence, we can use this classifier to detect potential discrepancies between the top-ranked answers. The UI of PriBot can thus highlight the potentially conflicting answers to the user.

Deployment and Legal Aspects

A standalone chatbot is only one of many applications for PriBot. PriBot can be integrated within a platform like Amazon Alexa or Google Assistant as another way of delivering information about privacy policies of third party apps. Also, it can function as a means to facilitate the comparison of the privacy practices of different companies. PriBot can answer a standard set of questions from a range of policies to provide researchers, users and regulators with a scalable and quantifiable way to compare privacy policies.

However, PriBot is not intended to replace the legally-binding privacy policy or terms of services. It offers a complementary interface for users to easily inquire the contents of a

privacy policy. Following the trend of automation in legal advice [Goo17], insurance claim resolution [Lev17], and privacy policy presentation [ZB14, LFL16], 3rd parties, e.g., automated legal services firms, can deploy PriBot as a solution for their users. As is the standard in similar cases, these companies should amend PriBot with a disclaimer specifying that PriBot is an assistant and does not represent the actual service provider [Hwa13]. The same can be done in the case of a company deploying PriBot to directly interact with users.

Companies and service providers can also deploy PriBot internally as an assistant tool for their customer support agents to handle privacy-related inquiries. Putting the human in the loop allows for a favorable trade-off between the utility of PriBot and its legal implications. For a wider discussion on the issues surrounding automated legal analysis, we refer the interested reader to the works of McGinnis and Pearce [MP14] and Pasquale [PC15].

Finally, we note that the quick responses given by PriBot can play a role in privacy policy debates. In fact, Fig. 6.11 showed the case of *Unroll.me*, a free service for removing unwanted subscriptions from users' email. Recently, there has been a lot of backlash against this company after it was reported to be selling information mined from those emails to third parties (e.g., selling data to *Uber* about the billing emails for customers of its competitor *Lyft*) [Isa17]. As Figure 6.11 shows, PriBot answers the question about third-party sharing with the particular policy segment that describes the practice reported on in the news. This is despite the fact that the question's only intersection with the answer is the word "with".

6.11 Related Work

Several studies have evaluated the effectiveness of privacy policies. Good et al., showed that users have a limited understanding of these policies and that they have a little desire to read them [GDG⁺05]. In a later work, Good et al., found that giving a short summary notice, in addition to the End User License Agreement, before installing the software significantly reduced the number of installations [GGMK07]. Moreover, they showed that presenting users with a short summary notice after installation led to a significant number of uninstalls.

Privacy Policy Analysis

Accordingly, there have been numerous attempts to create easy-to-navigate and alternative presentations of privacy policies. We differentiate between approaches that pursued manual designs of notices and approaches that investigated automated approaches of analyzing the policies.

Manual Designs

The Platform for Privacy Preferences (P3P) was one of the early works at making privacy policies more accessible [CLM⁺02, RC99]. In P3P, websites encode their policies in a machine-

readable format, and software agents (typically in the browser) parse this information and display it to the user. Examples of such agents were Privacy Bird [CGA06] and Privacy Bird Search [BCKM04]. Still, P3P did not have wide adoption, and its working group was closed in 2006, due to the weak adoption from the industry, the limited spread of such agents, and to the issues of user comprehension and enjoyability [Cra12, KBCR09].

In a later work, Kelley et al., proposed using nutrition labels as a paradigm for displaying privacy notices [KBCR09]. In their study, users were able to find out about data practices more quickly and had a more enjoyable information seeking experience compared to natural language policies. Icons representing the privacy policies have also been proposed [HZH11, CGA06].

In the financial sector, the standardization approach of privacy notices has found applicability [GHH⁺12] in the United States. However, in the general industry, no approach, till now, has reached the standardization point, and service providers have not voluntarily opted into one of the proposed approaches. Zimmeck and Bellovin suggest three reasons behind that: (i) the absence of industry incentives to move from the de facto standard of privacy policies, (ii) the support for natural language policies from the U.S. governmental agencies, and (iii) the stronger expressivity of natural language and its ability to relay industry-specific nuances [ZB14].

(Semi-)Automated Analysis

Early works on this problem studied the problem of analyzing privacy policies for limited purposes. Costante et al., targeted the problem of listing the set of data collected by the website [CdHP13]. In another work, Costante et al., [CSPdH12] developed a solution for assessing the completeness of privacy policies according to a predefined set of categories. Stamey and Rossi showed a system for topic modeling on privacy policies, which includes an ambiguous term extractor [SR09]. Still, these works have been focused on studying a single aspect of the policies with a limited scope. In this work, we tackle the problem on a much wider scale and with specialized natural language processing techniques, rather than out-of-the-box mechanisms (which we also show as sub-optimal).

Recently, several efforts have explored the potential of automated analysis of privacy policies. For example, Liu et al., have used deep learning to model the vagueness of words in privacy policies [LFL16]. Zimmeck et al., have been able to show significant inconsistencies between app practices and their privacy policies via automated analysis [ZWZ⁺17]. These studies, among others [SSWS16, LWSS16], have been largely enabled by the release of the OPP dataset by Wilson et al., [WSD⁺16], containing 115 privacy policies extensively annotated by law students.

Our work is the first to use the OPP dataset for the task of answering free-text questions by users. This work also falls in the line of automatically generating a data-driven interface to

privacy policies. In that regard, the closest work to ours is that of the *Privee* system. Privee used crowdsourced data from the “Terms of Service; Didn’t Read” service⁵ to train a machine learning classifier on grading a privacy policy according to 6 classes [ZB14]. This system was deployed in the form of a Google Chrome extension. In contrast, we automate the analysis at a much more fine-grained level, with more than 90 categories. We also exploit the emergence from a highly rich annotated dataset in building the core QA algorithms of PriBot. In addition, we believe that a standalone web application has much more potential to satisfy wider users need than the limited interface of Chrome extensions.

Automated Question Answering

QA techniques typically tackle two types of questions: *factoid* questions asking about short facts (e.g., *what is the highest peak in Europe?*), and *non-factoid* questions, which are typically complex and open-ended (e.g., *what losses can auto insurance protect you from?*). Evidently, our work falls into the second category. In the past few years, deep learning methods have shown superior results to traditional retrieval techniques in this domain. Researchers have explored several combinations of Convolutional Neural Networks (CNNs) [FXG⁺15], Long short-term memory networks (LSTMs), Attention-based Networks [TdSXZ16], and pointwise neural networks [RHL16] for this problem in general.

PriBot follows the spirit of many of these previous efforts in that we use word embeddings as features of a neural network to compute the proximity between sentences. However, the contributions of PriBot stem from addressing the *unique* challenges associated with the privacy policies domain (e.g., the absence of a QA dataset), in developing a custom QA architecture, and in demonstrating the quality of PriBot’s generated answers with a large-scale user study.

6.12 Summary

In this chapter, we propose PriBot, the first privacy-specific QA system that is readily available for public use. It answers user’s free-form questions from previously unseen privacy policies. We developed two deep-learning algorithms that allow PriBot to return answers from the privacy policy. We evaluated PriBot using a dataset of 120 real-world questions that we collected from Twitter. Our user study of 1186 participants revealed the high accuracy and relevance of PriBot’s answers. Further, we provide a proof-of-concept implementation of PriBot as a user-facing chatbot and evaluate its usability.

We envision that the techniques developed in PriBot will open the door for the future of automated privacy specialists in different fields. It has the potential to replace repeated work that takes several minutes with a considerably accurate replacement that takes milliseconds. This can result in a significant efficiency boost in the customer service domain for example.

⁵tosdr.org

Such improvements cannot be achieved if each company relies on its small dataset of privacy related conversations.

7 Conclusion

In this thesis, our goal was to improve the accessibility and usability of data privacy. We tackled this in an age where users' personal information is being collected at an enormous scale, data practices are more and more obscure, and risk communication is significantly lagging behind. We started with three obstacles in mind: *scale adaptation*, *risk communication*, and *language complexity*. Throughout the previous chapters, we presented our efforts for mitigating these obstacles. The takeaway messages from this thesis can be summarized as follows:

Feasibility of Sensitivity Assessment over Unstructured Data. We have shown with C3P that one can tame the complexity of sensitivity analysis in the case of unstructured files by accounting for the various context and content features that are associated with such files. We have found that this goal can be attained in a privacy-preserving manner too. Another interesting finding is that Item Response Theory, despite not being as complex as the recent machine learning models, is well suited to model the users' privacy attitudes and the files' sensitivity levels.

Improving Risk Communication with Data Analysis and Visualization. We have shown via PrivySeal that risk communication for unstructured data is not only achievable but can also be highly effective. Our Far-reaching Insights approach serves towards reducing the opacity of data processing by exposing, to the user, the possible repercussions of granting access to over-privileged 3rd party cloud apps. We have also found that not all insights are created equal. For instance, *relational* insights, displaying information about users' relations with others, were more effective than *personal* insights about the user only.

Enlightening the Users on Interdependent Privacy Risks via New Privacy Indicators On top of exposing the extent of interdependent privacy in third party cloud apps, we sought a way to curtail the privacy loss via privacy indicators. We were able to show that small changes in the permissions interface (i.e., our History-based Insights) can help the users better account

for previous decisions.

Building Better Privacy Interfaces with Machine Learning. We have demonstrated that privacy policies do not have to stay to be the de facto method for communicating data practices on the long run. Policies can be rather used as the fuel to power the machine learning models, which, in turn, can be leveraged to give the users the specific information they are looking for. We have shown the efficacy of this approach with users in the case of question answering, where our new model scored high relevance and accuracy metrics.

On a high level, one of the key differentiators of our work was striving for demonstrability by putting the systems, whenever possible, in the hands of end-users with minimal effort needed on their behalf. This was the case for example with *PrivySeal*, which grew to around 1750 registered users at the time of writing this thesis. It is also the case with *PriBot*, our newly released web application.

Looking forward, we highlight both the direct extensions of our work and the long term ramifications. To begin with, although the privacy issue in cloud file sharing has been less pronounced recently, there are several emerging problems that share a similar DNA. For example, conducting privacy preserving analytics over users' interactions with smartphones has been on the agenda of the top manufacturers. We envision that a lot of interesting analytics can also be conducted on unstructured data, and there is much work to be done there on offline feature specification and extraction with machine learning.

When it comes to deterring users from installing over-privileged applications, we have admittedly attempted at breaking the knowledge imbalance between users and providers with Far-reaching insights. Hence, using this exact technique is likely to not be in the interest of platform providers as it can curtail their growth. However, we envision that these techniques can be alternatively provided by third parties, which offer privacy as a service. These third parties can be used by the interested individuals or teams in an organization. On the other hand, highlighting the effects of interdependent privacy and informing users about the existing parties with access to their data is less of a burden for the platform providers. In the future, we hope that providers go beyond *listing* the apps that the user has authorized to *visualizing* the parties with access to their data. This includes parties enabled by other users and parties enabled by the apps themselves (e.g., ad providers, data brokers, etc.). This gives users a more transparent overview of their data.

Moreover, in the case of privacy policies, our work on *PriBot* scratches the service of what is possible. We see a significant opportunity of complementing these policies with existing datasets, such as customer services' logs, to attain much higher accuracy. Another interesting extension is to go beyond rule-based summarization to abstractive summarization, where the policies are encoded in a simpler language. That is another avenue where deep learning research on summarization, neural machine translation, and text generation can be leveraged. Additionally, putting *PriBot* and other similar systems in the hands of user would provide

a valuable data source for adjusting the models based on users' feedback. Towards that end, there is work to be done on the adequate interfaces for collecting feedback and proper methods for promptly integrating such feedback within the models themselves. Furthermore, we plan to tackle the UI design challenges of PriBot by investigating the trade-off between accuracy and usability. We also plan to improve the accuracy of PriBot through a multi-stage conversation. For example, we want to narrow down the answers by allowing PriBot to ask whether the user is concerned about the third parties or the first party.

To look further, one of the major outcomes of this thesis is that *unstructured privacy interfaces* are feasible and effective. We distinguish these interfaces from the traditional *structured privacy interfaces*, such as permission dialogs with pre-scripted text, predesigned privacy labels, or privacy information within application stores. It has often been the case in the user experience domain that "the best interface is no interface" [Kri15]. This has been further reaffirmed with the emergence of new devices where "the only interface is no interface". Accordingly, privacy indicators should not be restricted by a form or a medium. In fact, the emerging trend of virtual assistants calls for virtual privacy specialists that can effectively communicate to the user the data practices of all the new parties receiving users' information. In such specialists, we see a lot of potential for applying our data-driven techniques in risk communication. Talking to the user using their data as a language is one of the most effective methods as we have seen, not only because it swiftly relays the message but also because it reduces habituation and serves as an educational tool.

A Study Material for Chapter 2

In this appendix, we present the material used in the user study reported in Chapter 2. The full vocabulary on which we build is presented in Figure A.1. The questions were split over 9 surveys with a similar structure. The surveys are all presented next (we kept them as they were originally, with the replicated instructions).

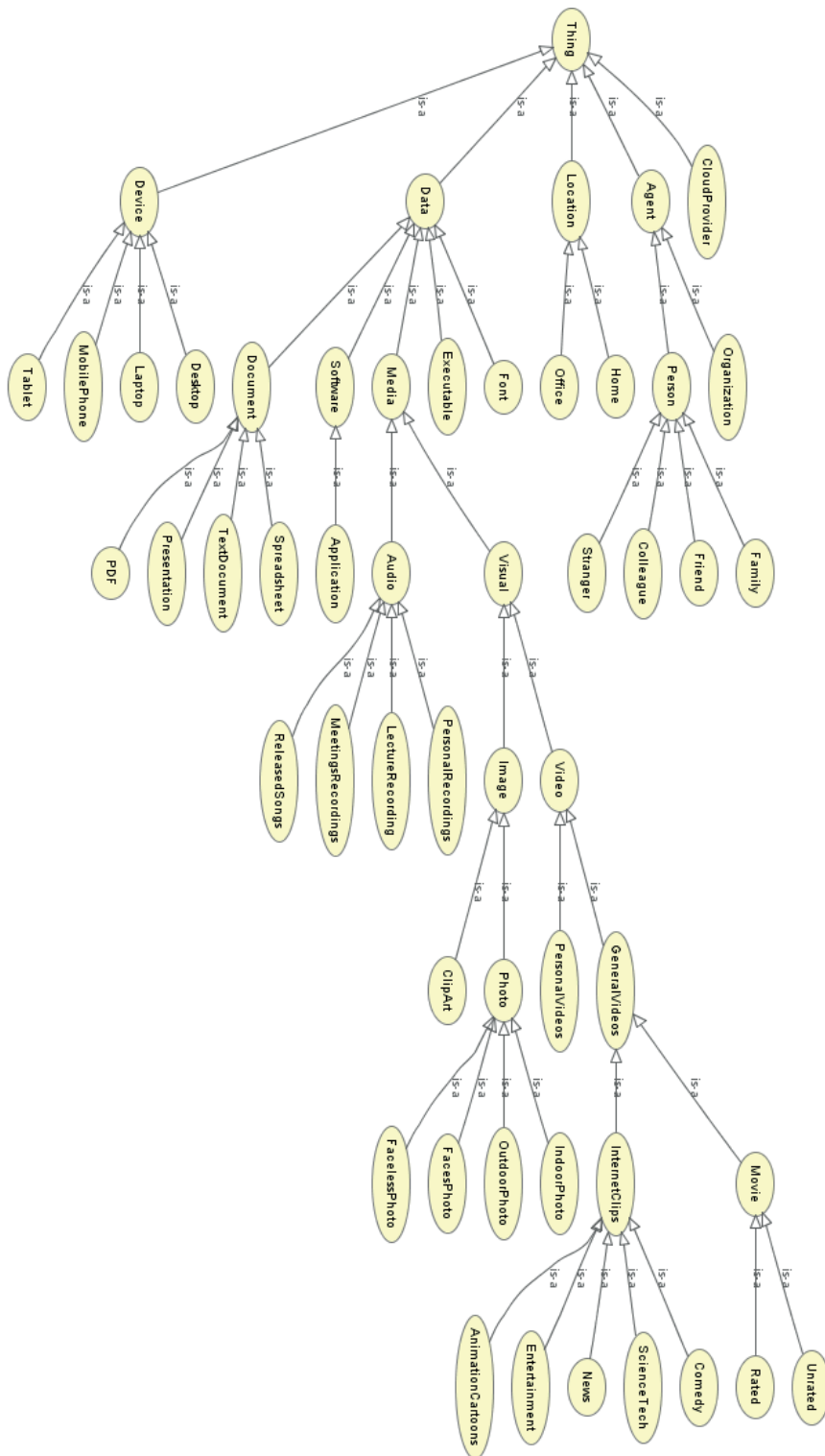


Figure A.1 – Vocabulary used in the user study

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1. You have a recent photo of you with your family in your living room.

With whom of the following would you **AVOID** sharing this photo for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2. You have a word document for a tutorial you downloaded from the Internet.

With whom of the following would you **AVOID** sharing this document for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have presentation slides for a project you are working in at your organization/school/university. If you would share these slides with your friends, which of the following features in the slides would you choose to **HIDE**?

- ☐ Author name
- ☐ File Revision History (who edited it and when)
- ☐ Organisation/Company
- ☐ Name of the computer
- ☐ None of the above

4. You have a PDF that you created. If you would share this file with your friends, which of the following features in this PDF would you choose to **HIDE**?

- ☐ Author name
- ☐ Revision History (who edited it and when)
- ☐ Your organisation/company
- ☐ Name of your computer
- ☐ None of the above

5. Spideroak is a company with which it is possible to send your files for backing them up. Before you send these files, Spideroak allows you to lock them so that nobody other than you can access these files. In addition, once it receives the data, it locks it to prevent external hackers from accessing it. Even if the US authorities request this data from Spideroak, it cannot give it to them since only you can unlock the data. Put another way, it can only provide them with locked data, that only you can unlock

Select all the items for which you would **NOT TRUST** Spideroak enough to send them to their servers for storage:

- ☐ a photo of you with your family in your living room
- ☐ a word document for a tutorial you downloaded from the internet
- ☐ a photo of a landscape scene you have captured
- ☐ a photo of some family you don't know, captured in their house
- ☐ a song of a famous artist

- ☐ a photo of you with friends in a park
- ☐ I would trust it with all of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1. You have a photo of a landscape scene you have captured.

With whom of the following would you **AVOID** sharing this photo for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2. You have a photo of a family you don't know, captured in their house.

With whom of the following would you **AVOID** sharing this photo for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have a photo of your friend you have taken. If you would share this photo with your friends, which of the following features in the photo would you choose to **HIDE**?

- ☐ Author name
- ☐ Photo location
- ☐ Device used to capture
- ☐ Date taken/modified
- ☐ None of the above

4. You have captured a video clip of your birthday by your camera. If you would put it to the public on some site, which of the following features in the video would you choose to **HIDE**?

- ☐ Author name
- ☐ Video Location
- ☐ Device used to capture
- ☐ Date created/modified
- ☐ None of the above

5. Spideroak is a company with which it is possible to send your files for backing them up. Before you send these files, Spideroak allows you to lock them so that nobody other than you can access these files. In addition, once it receives the data, it locks it to prevent external hackers from accessing it. Even if the US authorities request this data from Spideroak, it cannot give it to them since only you can unlock the data. Put another way, it can only provide them with locked data, that only you can unlock

Select all the items for which you would **NOT TRUST** Spideroak enough to send them to their servers for storage:

- ☐ a software that belongs to your organization/company
- ☐ a birthday party video of your brother or sister
- ☐ a movie which contains some unrated contents
- ☐ an audio recording of your organization's/company's meetings
- ☐ a video of you and your family captured at home
- ☐ a video of you and your colleagues have captured in your organization/company

☐ I would trust it with all of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1. You have a song of a famous artist.

With whom of the following would you **AVOID** sharing this song for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2. You have a photo of you with your friends in a park.

With whom of the following would you **AVOID** sharing this photo for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have an excel sheet prepared by your colleague and you. If you would share it with your colleagues, which of the following features in the sheet would you choose to **HIDE**?

- ☐ Author name
- ☐ Revision history (who edited it and when)
- ☐ Your organisation/company
- ☐ Name of your computer
- ☐ None of the above

4. You have a cartoon movie downloaded from the Internet. If you would share it your family, which of the following features in the movie would you choose to **HIDE**?

- ☐ Name of your computer
- ☐ Data accessed
- ☐ None of the above

5. Dropbox is a company with which it is possible to send your files for backing them up. Once it receives the data, it locks it to prevent external hackers from accessing it. Unlike Google, computers at Dropbox do not analyze your data to provide you with personal ads. Dropbox is also obliged to hand in your data to the US authorities if the latter demand.

Select all the items for which you would **NOT TRUST** Dropbox enough to send them to their servers for storage:

- ☐ a photo of you with your family in your living room
- ☐ a word document for a tutorial you downloaded from the internet
- ☐ a photo of a landscape scene you have captured
- ☐ a photo of some family you don't know, captured in their house
- ☐ a song of a famous artist
- ☐ a photo of you with friends in a park
- ☐ I would trust it with all of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1. You have an installation file for a new software.

With whom of the following would you **AVOID** sharing this file for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2 You have a PDF file written by your superior (e.g. teacher/boss).

With whom of the following would you **AVOID** sharing this file for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have a scientific video you got from the net . If you would share it with your friends, which of the following features in the video would you choose to **HIDE**?

- ☐ Name of your computer
- ☐ Data accessed
- ☐ None of the above

4. You have a photo of you with your friends taken at home. If you would share it with your colleagues at work/university, which of the following features in this photo would you choose to **HIDE**?

- ☐ Location
- ☐ Date taken
- ☐ Device used to capture
- ☐ None of the above

5. Dropbox is a company with which it is possible to send your files for backing them up. Once it receives the data, it locks it to prevent external hackers from accessing it. Unlike Google, computers at Dropbox do not analyze your data to provide you with personal ads. Dropbox is also obliged to hand in your data to the US authorities if the latter demand.

Select all the items for which you would **NOT TRUST** Dropbox enough to send them to their servers for storage:

- ☐ a software that belongs to your organization/company
- ☐ a birthday party video of your brother or sister
- ☐ a movie which contains some unrated contents
- ☐ an audio recording of your organization's/company's meetings
- ☐ a video of you and your family captured at home
- ☐ a video of you and your colleagues have captured in your organization/company
- ☐ I would trust it with all of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1. You have recorded an audio file using your phone at home.

With whom of the following would you **AVOID** sharing this file for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2. You have recorded a video file using your phone at home.

With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have downloaded a word file from the Internet. If you would share it with your colleagues at office, which of the following features in this file would you choose to **HIDE**?

- ☐ Author name
- ☐ File Revision History (who edited it and when)
- ☐ Organisation/Company/School
- ☐ Name of the computer
- ☐ None of the above

4. You have a technical video for your organization/company. If you would share it in public, which of the following features in this video would you choose to **HIDE**?

- ☐ Name of the computer
- ☐ Data accessed
- ☐ None of the above

5. Google is a company with which it is possible to send your files for backing them up. Once it receives the data, it locks it to prevent external hackers from accessing it. However, computers at Google analyze your data to provide you with personal ads. Google is also obliged to hand in your data to the US authorities if the latter demand for it.

Select all the items for which you would **NOT TRUST** Google enough to send them to their servers for storage:

- ☐ a photo of you with your family in your living room
- ☐ a word document for a tutorial you downloaded from the internet
- ☐ a photo of a landscape scene you have captured
- ☐ a photo of some family you don't know, captured in their house
- ☐ a song of a famous artist
- ☐ a photo of you with friends in a park
- ☐ I would trust it with all of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1.You have received an excel spreadsheet your colleague at work/university/college has prepared.

With whom of the following would you **AVOID** sharing this file for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2.You have just downloaded a kids' movie from the Internet.

With whom of the following would you **AVOID** sharing this file for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have audio recordings of meetings at your organization in your office computer. If you would share you it in public, which of the following features in this audio would you choose to **HIDE**?

- ☐ Name of your computer
- ☐ Organisation/Company
- ☐ File creator
- ☐ None of the above

4. You have downloaded a song from the Internet to your mobile phone. If you would shared it with your friends, which of the following features in this song would you choose to **HIDE**?

- ☐ Device used
- ☐ Your rating of the song
- ☐ None of the above

5. Google is a company with which it is possible to send your files for backing them up. Once it receives the data, it locks it to prevent external hackers from accessing it. However, computers at Google analyze your data to provide you with personal ads. Google is also obliged to hand in your data to the US authorities if the latter demand for it.

Select all the items for which you would **NOT TRUST** Google enough to send them to their servers for storage:

- ☐ a software that belongs to your organization/company
- ☐ a birthday party video of your brother or sister
- ☐ a movie which contains some unrated contents
- ☐ an audio recording of your organization's/company's meetings
- ☐ a video of you and your family captured at home
- ☐ a video of you and your colleagues have captured in your organization/company
- ☐ I would trust it with all of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1. You have audio recordings of some lectures on your tablet.

With whom of the following would you **AVOID** sharing this audio for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2. You have a video about recent news on your work/university/school computer.

With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have a video clip of your birthday party on your laptop.

With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

4. You have a photo of you during a wedding party you have recently attended.

With whom of the following would you **AVOID** sharing this album for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

5. You have presentation slides of a technical course on your home computer.

With whom of the following would you **AVOID** sharing this slides for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1.You have a comedy video clip you got from the Internet .

With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2.You have a technical video related to your work/university/school.

With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have audio recordings of meetings at your organization on your office/university computer.

With whom of the following would you **AVOID** sharing this audio for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

4. You have an application for converting photos to cartoons on your tablet.

With whom of the following would you **AVOID** sharing this application for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

5. You have an application on your laptop for converting photos to cartoon. If you would share it with your friends, which of the following features in this application would you choose to **HIDE**?

- ☐ Name of the computer
- ☐ Data accessed
- ☐ None of the above

6. You have a video in your laptop that contains some unrated contents. If you would share it with your friends, which of the following features in this video would you choose to **HIDE**?

- ☐ Name of your computer
- ☐ Data downloaded
- ☐ Date accessed

☐ None of the above

Thank you for agreeing to take part in this survey by the *Distributed Information Systems lab* at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. This survey intends to assess people's attitudes towards privacy of online data sharing.

The questions will ask you about your viewpoints towards sharing different types of information from your devices with internet companies that store such information (a.k.a cloud providers) and with other people. They **keypoint** to answering these questions is understanding that **the privacy attitude is typically related to the contexts of sharing your data**. Borrowing an example from the phone call scenario: knowing that you have called a specific person at a certain time might be as sensitive to you as the content of the call itself. .



1.You have recorded a video file using your phone at work/university/college.
With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

2.You have a movie which contains some unrated contents on your laptop/desktop.
With whom of the following would you **AVOID** sharing this movie for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

3. You have captured a video of your friends on your tablet. This video contains funny scenes while you were on a trip.

With whom of the following would you **AVOID** sharing this video for privacy reasons:

- ☐ Your close friends
- ☐ Your colleagues at work/university/school (non-friends)
- ☐ Members of your Family
- ☐ The public (as in posting it online)
- ☐ None of the above

4. You have captured a video of you and your family using your tablet. If you would share it with your friends, which of the following features in this video would you choose to **HIDE**?

- ☐ Author name
- ☐ Location
- ☐ Device used to capture
- ☐ Date captured
- ☐ None of the above

5. You have an audio recording of your lectures in your laptop. If you would share it with your colleague, which of the following features in this audio would you choose to **HIDE**?

- ☐ Device used to record
- ☐ Data recorded
- ☐ None of the above

B Study Material for Chapter 5

In this appendix, we provide the material for the study reported in Chapter 5.

B.1 Introductory Material

First the participants were presented with the instructions presented in Figure B.1 and B.2. Next, they answer the introductory survey in Figures B.3 and B.4.

Instructions

Overview

What is this about?

Cloud storage services (e.g. Dropbox, Google Drive, OneDrive, etc.) are now being used by a lot of people for various purposes. People store their documents, photos, or music on these services, so that they can access these files from any device at any time.

Also, many companies have developed applications that you can connect to your Dropbox or Google Drive accounts. For example, you can allow an image editing app to access your Dropbox in order to edit an image you have stored there. You can also use an app for signing documents you already have on Google Drive.

These apps are made by companies other than Google or Dropbox, but they provide users with a lot of services by connecting to their Google Drive or Dropbox accounts. That's why they are called **3rd Party Apps**.

We are conducting a user experiment to check how people make decisions when they install such 3rd party apps.

Are you eligible?

To qualify for completing the tasks, you will have to be:

- a user with a good familiarity with one of the cloud storage services.
- have actually utilized these services to store files

How is the study paid? (Read this please)

- All participants will automatically receive the **first payment** after completing the study.
- All the participants who complete the study without randomly filling answers will receive an amount equal to **2.5 times the first payment as a bonus (for example first payment will be \$0.5 and bonus will be \$1.25)**
 - This is especially important in the part where the answer is in the form of **text input**. You will be provided with **guidelines** on how to not answer these questions. After **we manually check** these answers for quality, we will issue the bonus to all participants who have answered according to the guidelines.

Why is this important?

- Your contribution will lead to better process of cloud apps' installation.
- You will affect how apps are presented to the users in the future.

Example Apps

To make sure you understand how these apps work, here are two examples:

- This is an application called *ILoveIMG* that allows importing photos from Google Drive and then cropping them.

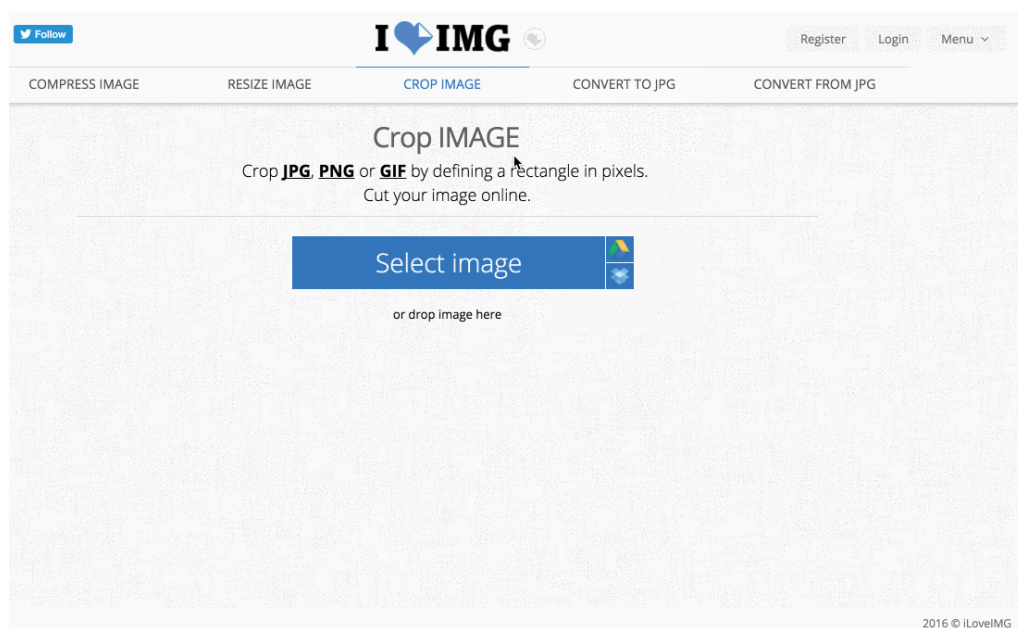
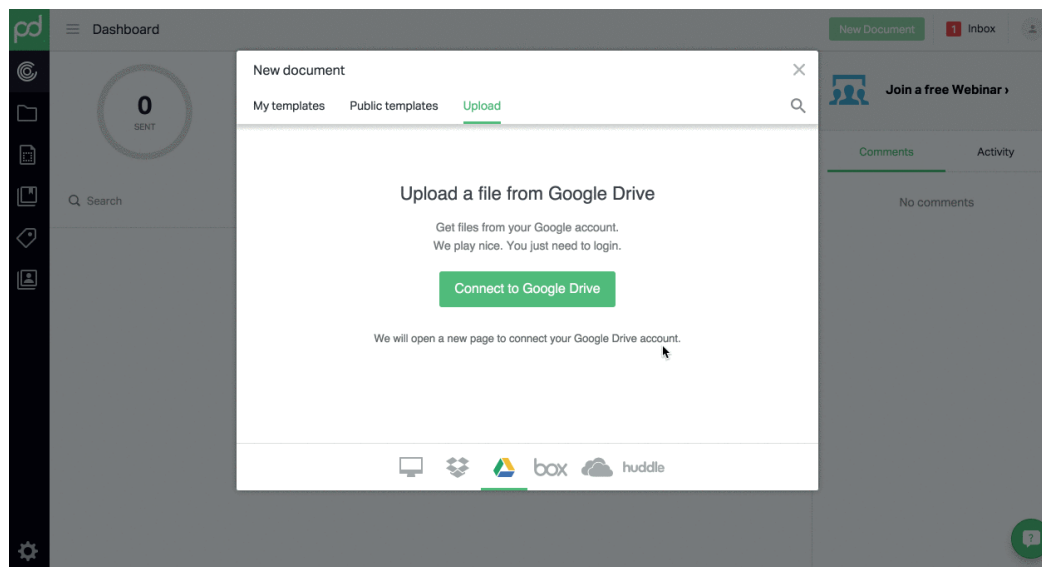


Figure B.1 – First part of the instructions given to the participants at the beginning of the study

- Here is another application called *PandaDoc* that allows importing a letter from Google Drive and then signing that letter.



Process

- You will first go to our study's webpage on the given link.
- You will answer a survey.
- You will be given a set of tasks related to cloud services.
- Then you will be asked to fill a final short survey.
- At the end, you will be given a code that you have to enter in CrowdFlower interface in order to get the payment.

Do Not

- **Do Not Refresh** the study's page or close it until the end of the study where you get your completion code. Refreshing it will make you lose the progress and start your task from the beginning.
- **Do Not press the Back or Forward buttons** on your keyboard to go to the previous or next page during the study. Only click on the given buttons in order to go to the next step.
- **Do Not use a tablet or smartphone to do this study:** only use a computer/laptop/desktop

Do's

- Please try to do the experiment without interruption as some parts are related.
- **Think well about your decisions** during the experiment.
- **Check all the possible choices** before making your mind on a decision.

Figure B.2 – Second part of the instructions given to the participants at the beginning of the study

Survey

Please answer in **English** the following questions (and the subsequent study).
 Make sure you have read well the instructions on CrowdFlower. You can also see them on this page:
<https://privyseal.epfl.ch/#/hisExp/instructions>

1.

What is your CrowdFlower Contributor ID?

2.

What is your age (in years)?

3.

What is your gender?

- ☐ Male
☐ Female

4.

What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

- ☐ High school
☐ Trade/technical/vocational training
☐ Associate or Bachelor's degree
☐ Post Graduate Degree

5.

What is your occupation?

- ☐ full-time employee
☐ student
☐ part-time worker
☐ self-employed
☐ homemaker
☐ Unemployed/retired

6.

Have you worked before or studied in a field related to IT (Information Technology)?

- ☐ Yes
☐ No

7.

Please list the names of cloud storage service(s) that you use to store your files?

8.

Select all the purposes for which you use these providers?

- ☐ Storing photos
☐ Storing documents
☐ Sharing photos with others
☐ Collaborating on documents

9.

Have you previously used real-time collaboration services (e.g. Google Docs)?

- ☐ Yes
☐ No

10.

Suppose you have friends who don't use cloud storage services. What is the one feature/advantage that you can mention to convince them to use such services.

11.

Have you previously given 3rd party applications access to some/all of your cloud files?

- ☐ Yes
☐ No

Figure B.3 – First part of the initial survey that the participants fill

12.
If you answered Question 11 by 'Yes', to how many 3rd apps approximately have you given access?

☐ 1-5
☐ 5-10
☐ more than 10

13.
If you answered Question 11 by 'Yes', what was the purpose of those 3rd party application(s)?

14.
If a 3rd party application requests the permission to: "View and manage the files in your Google Drive (or Dropbox)", what do you think that means?

☐ The app can immediately view all my files.
☐ I can give the app access to specific files when I want

15.
If a 3rd party application requests the permission to: "View and manage Google Drive (or Dropbox) files and folders that you have opened or created with this app.", what do you think that means?

☐ The app can immediately view all my files.
☐ I can give the app access to specific files when I want

16.
With how many people overall have you shared files previously?

☐ 0
☐ 1-5
☐ 5-10
☐ more than 10

Done. Go to experiment

Figure B.4 – Second part of the initial survey that the participants fill

B.2 Material for Modules

The participants are then presented with the modules summarized in Figure B.5. Figure B.7 to B.17 show screenshots of different steps in the modules. The captions in those figures refer to the numbering in Figure B.5. Before each module, the instructions in Figure B.6 are shown.

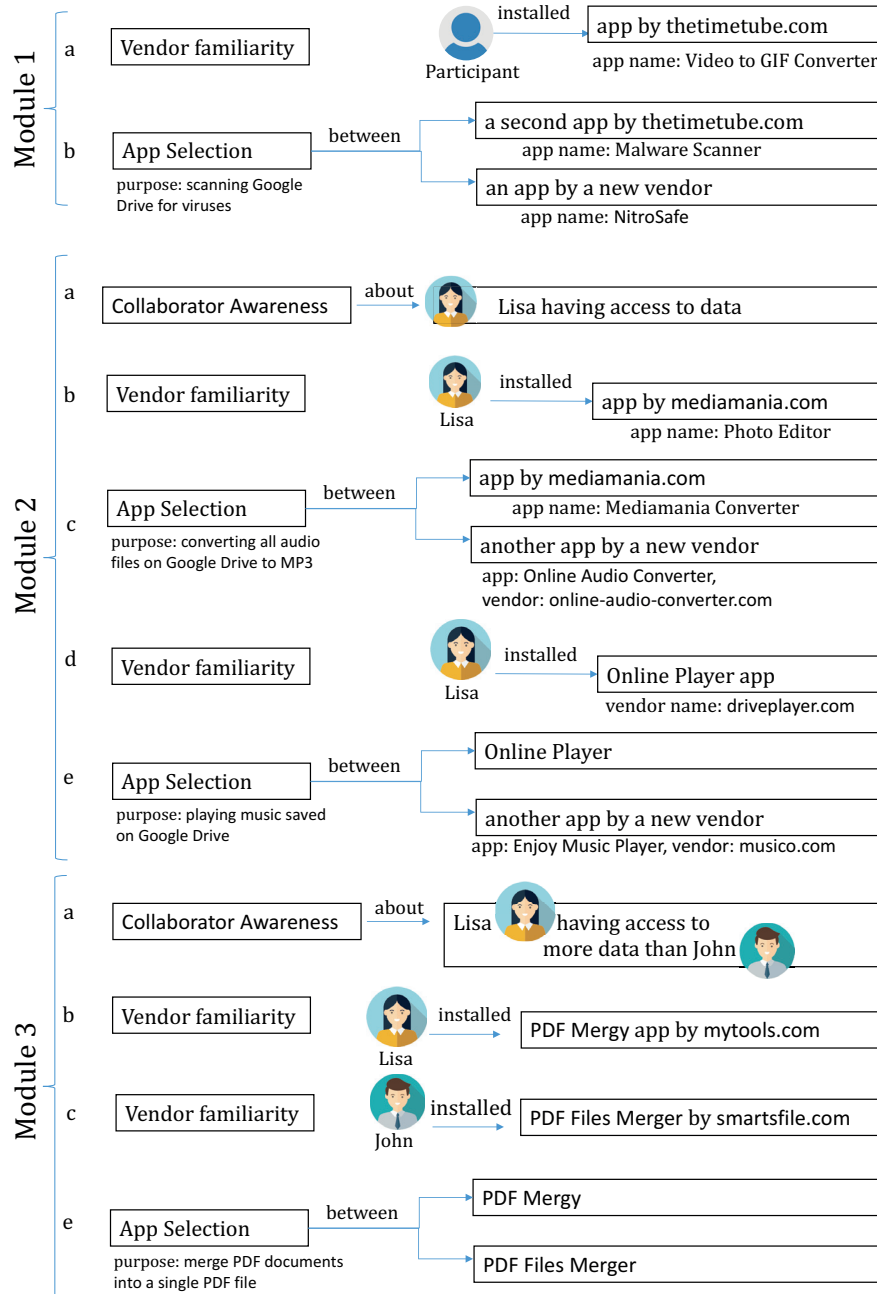


Figure B.5 – High level overview of the modules of the experiment; we refer to the parts in this figure in the coming figures.

- Take a breath.
- You will now play a role of someone who has a Google Drive account and has **already stored files** on it (like your images from your trips with the family, some official documents, music files, etc.).
- At some point, you'll be asked to choose some apps to connect to your Google Drive account. Although no apps will be actually installed, we ask you to think as if they were real apps and that this is a real Google Drive account.
- Once you install an app, assume **it is still installed throughout this experiment**.
 - For example, if the first task says: "you have installed an application from the company *pandadoc.com*", this application is still there when you go to the next task. So if the second task says: "who has access to your data?", the answer would be "*pandadoc.com company*".
- Similarly, whenever you are informed that your friends have installed applications, consider that **these applications are still connected to their Google Drive throughout this experiment**.
- When this experiment ends and **you move to the next experiment, you will start from scratch** (i.e. with no apps installed).

Figure B.6 – Instructions that participants see before each module

Module 1

Task:
As explained, we now start from scratch. Consider that this is the first app you will install. Please install any application from the company: **thetimetube.com**. (Only one such app exists, and you can click on the app to view its info.)





 <p>Video to GIF Converter</p> <p>Company: thetimetube.com</p> <p>Description: This app allows you to create animated GIFs from videos directly. You can open a video file from your Google Drive and computer.</p>	 <p>Online Audio Converter</p> <p>Company: online-audio-converter.com</p> <p>Description: Convert audio files on your Google Drive from any format to another.</p>
 <p>NitroSafe</p> <p>Company: nitrosafe.org</p> <p>Description: Malware scanning for Google Drive: Searches for malware, viruses, trojans and other nasty files in your Google Drive.</p>	 <p>PDF Mergy</p> <p>Company: mytools.com</p> <p>Description: Allows to merge PDF files from your Google Drive with a simple interface.</p>

Figure B.7 – Module 1-a: the participant is asked to install an application from a specific company.

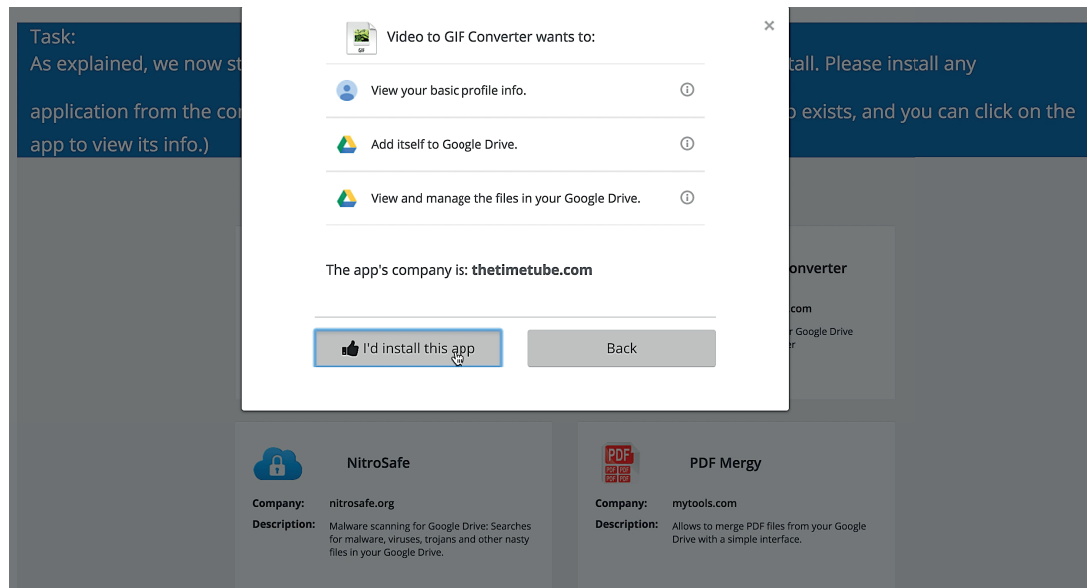


Figure B.8 – Module 1-a: the participant sees the traditional permissions interface to install the app.

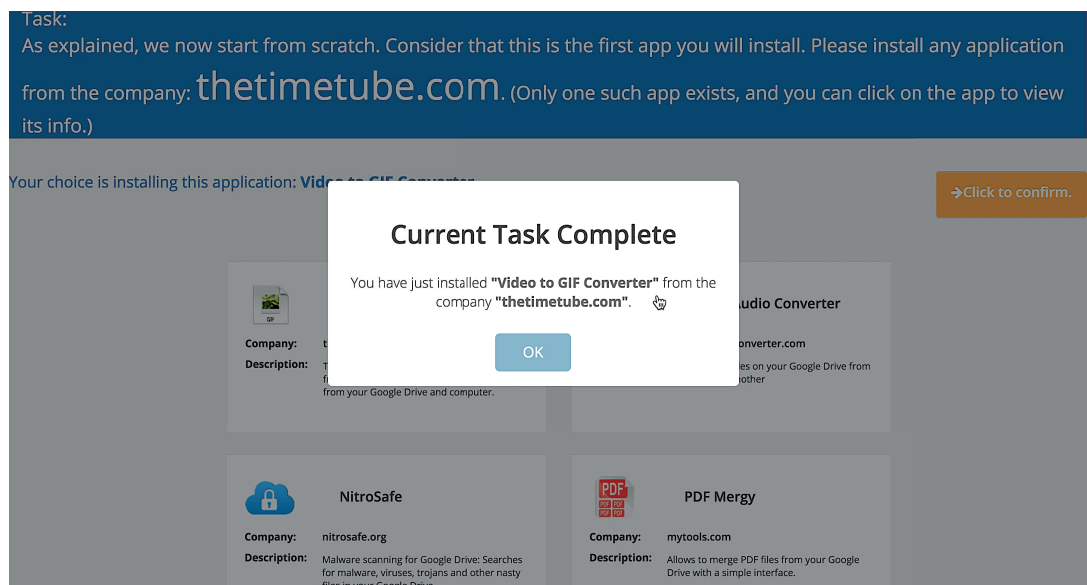


Figure B.9 – Module 1-a: the participant is notified again of the company name in order to get familiarized with it.

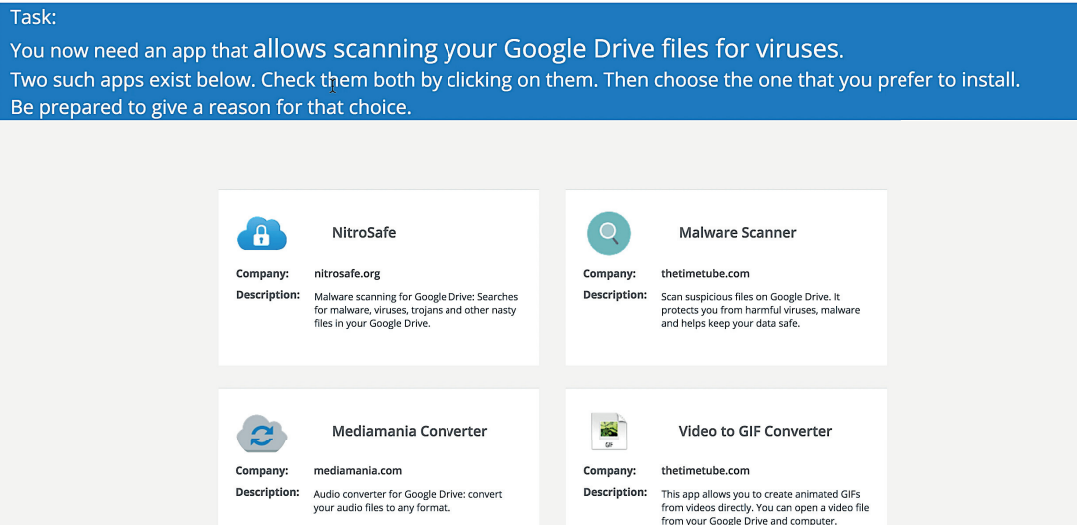


Figure B.10 – Module 1-b: the participant is asked to install an app of a certain purpose; two apps satisfy this purpose and are shown (in a random order) on the top row.

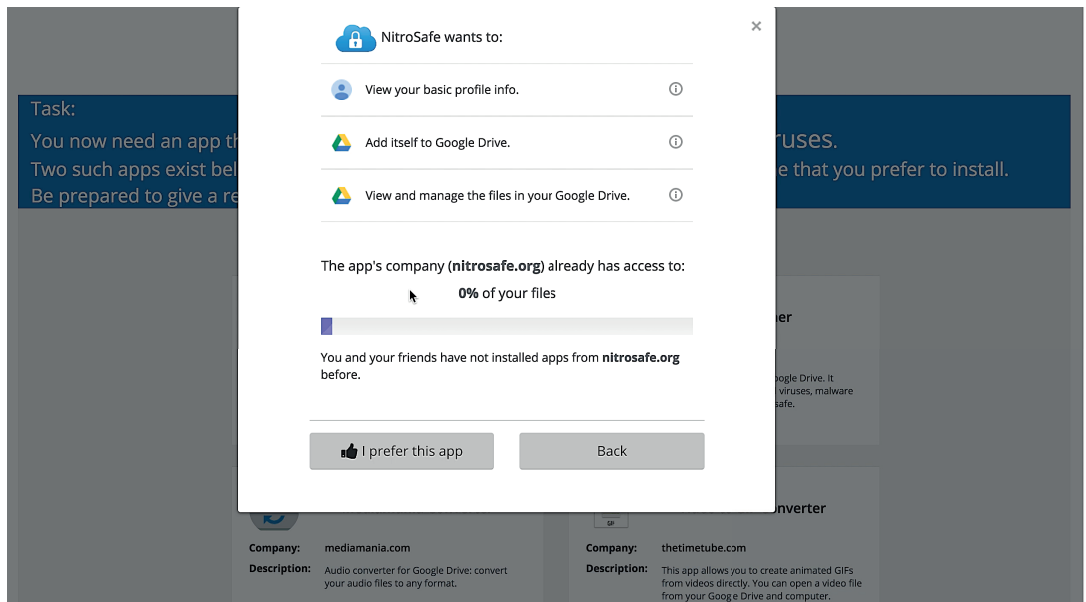


Figure B.11 – Module 1-b: the participant is shown the permissions interface, depending on the experimental group; here we show the case of *HB* group.

Task:
You now need an app that allows scanning your Google Drive files for viruses.
Two such apps exist below. Check them both by clicking on them. Then choose the one that you prefer to install.
Be prepared to give a reason for that choice.

Please provide a **brief reason** for favoring the app **Malware Scanner** over the app **NitroSafe**

Examples of bad reasons (that don't qualify for bonus):

- This app is better.
- Malware Scanner allows me to do the indicated functionality (like scan my files, listen to music, convert documents, etc.)
- Malware Scanner is very good and does what I need.
- I like Malware Scanner more.
- Malware Scanner is easy to use.

Style of a good reason (to qualify for bonus):

- Your reason should indicate something about/related to Malware Scanner that made you favor it over NitroSafe




	NitroSafe		Malware Scanner
Company:	nitrosafe.org	Company:	thetimetube.com
Description:	Malware scanning for Google Drive: Searches	Description:	Scan suspicious files on Google Drive. It

Figure B.12 – Module 1-b: the participant is asked to provide a justification for their choice.

Module 2

For this module, we show the case of Module 2-d and 2-e in Figure B.5. The case of Module 2-b and 2-c follow a similar structure.

Task:
Your friend **Lisa**  has installed an application called **Online Player** and has given its company access to all her files (including shared files). Write below the name of the company that owns this application. (You can click on the app to view its info.)

Company: → → Click to confirm.





 <p>Online Player</p> <p>Company: driveplayer.com</p> <p>Description: Music and Audio files player for Google Drive.</p>	 <p>PDF Files Merger</p> <p>Company: smartfile.com</p> <p>Description: PDF merge - Allows you to merge your Google Drive pdf files easy and fast.</p>
 <p>NitroSafe</p>	 <p>Online Audio Converter</p>

Figure B.13 – Module 2-d: the participant is made aware that a collaborator has installed an app from a specific company by requesting her to type company URL in the box. We randomly selected the name between Lisa or John for each user in our study. Here, we show the case of Lisa. Afterwards, the participant gets a confirmation as in Figure B.9.

Task:
You now need an app that allows playing your music files that are saved on Google Drive. Two such apps exist below. Check them both by clicking on them. Then choose the one that you prefer to install. Be prepared to give a reason for that choice.







 <p>Online Player</p> <p>Company: driveplayer.com</p> <p>Description: Music and Audio files player for Google Drive.</p>	 <p>Enjoy Music Player</p> <p>Company: musico.com</p> <p>Description: With Enjoy Music Player, you can play music and audio easily from your Google Drive.</p>
 <p>Video to GIF Converter</p> <p>Company: thetimetube.com</p> <p>Description: This app allows you to create animated GIFs from videos directly. You can open a video file from your Google Drive and computer.</p>	 <p>NitroSafe</p> <p>Company: nitrosafe.org</p> <p>Description: Malware scanning for Google Drive: Searches for malware, viruses, trojans and other nasty files in your Google Drive.</p>

Figure B.14 – Module 2-e: the participant is asked to install an app of a certain purpose; two apps satisfy this purpose and are shown (in a random order) on the top row. One of these apps has been installed earlier by Lisa. The participant sees an installation interface similar to Figure B.11 and has to justify as in Figure B.12.

Module 3

Task:


Assume that you have shared all your photos with **Lisa** . Additionally, you have shared with **John**  some of your photos. Who has more files from you in their Google Drive?

☐ Lisa
☐ John

[→Click to confirm.](#)

Figure B.15 – Module 3: at the beginning of the module, the participant is made aware that Lisa has more of the shared files than John.

Task:

Your friend **Lisa**  has installed an application called PDF Mergy and has given its company access to all her files (including shared files). Write below the name of the company that owns this application. (You can click on the app to view its info.)

Company:

[→Click to confirm.](#)





 <p>PDF Mergy</p> <p>Company: mytools.com</p> <p>Description: Allows to merge PDF files from your Google Drive with a simple interface.</p>	 <p>Video to GIF Converter</p> <p>Company: thetimetube.com</p> <p>Description: This app allows you to create animated GIFs from videos directly. You can open a video file from your Google Drive and computer.</p>
---	---

Figure B.16 – Module 3-a: the participant is made aware that Lisa has installed an app called PDF Mergy; a confirmation like in Figure B.9 is then shown.

Task:
Your friend **John**  has installed an application called **PDF Files Merger** and has given its company access to all his files (including shared files). Write below the name of the company that owns this application. (You can click on the app to view its info.)


Company: → → Click to confirm.



PDF Files Merger

Company: smartsfile.com

Description: PDF merge - Allows you to merge your Google Drive pdf files easy and fast.




Malware Scanner

Company: thetimetube.com

Description: Scan suspicious files on Google Drive. It protects you from harmful viruses, malware and helps keep your data safe.

Figure B.17 – Module 3-b: the participant is made aware that John has installed an app called PDF File Merger; a confirmation like in Figure B.9 is then shown.


Task:
You now need an app that allows you to merge multiple PDF documents on your Google Drive into a single PDF file.
Two such apps exist below. Check them both by clicking on them. Then choose the one that you prefer to install. Be prepared to give a reason for that choice.



PDF Files Merger

Company: smartsfile.com


Description: PDF merge - Allows you to merge your Google Drive pdf files easy and fast.



PDF Mergy


Company: mytools.com

Description: Allows to merge PDF files from your Google Drive with a simple interface.



Video to GIF Converter

Company: thetimetube.com



Enjoy Music Player

Company: musico.com

Figure B.18 – Module 3-c: the participant is asked to install an app of a certain purpose; two apps satisfy this purpose and are shown (in a random order) on the top row. One of these apps has been installed earlier by Lisa and the other by John. The participant sees an installation interface similar to Figure B.11 and has to justify as in Figure B.12.

B.3 Final Survey

After finishing the modules, the participants are requested to fill a final survey, shown in Figure B.19.

Thanks for completing the experiment. We have some final questions for you. After you answer them, click on the button below to access the **Completion code** that you can enter in CrowdFlower for getting your reward.

- 1.**
If you knew that your friend (let's call him John) has connected an application to his Google Drive and has given it access to all files (including those you share with him), would you prefer that you are notified about that?

☐ Yes
☐ No
☐ Indifferent
- 2.**
You were asked several times to choose between two apps to install. Select all the reasons from below that affected your decisions.

☐ Permissions the apps requested
☐ Whether the app name looked professional/cool
☐ How much the app's company knew previously about me
☐ Whether the company has a professional-looking name
☐ I was actually confused and selecting randomly
☐ Others (please specify):
- 3.**
Given two apps with the same permissions. App A is from a company that you have previously given access to your Google Drive. App B is from a company that you haven't given access in the past. Which app, in your opinion, would give you better privacy?

☐ App A
☐ App B
☐ Both apps give me the same privacy.
- 4.**
Assume you have installed an application called *YouMusic* from a company called *Musicana* and gave it access to all your files on Google Drive. Now you are considering installing an application called *YouVideo* from the same company. How do you think that this application will affect your privacy:

☐ It will affect my privacy **negatively**.
☐ It will affect my privacy **positively**.
☐ It will not have an effect on my privacy.

Please justify your answer:
- 5.**
If you connect an application to your Google Drive and you give it access to all files (including those you share with your friends), would you prefer that your friends are notified about that?

☐ Yes
☐ No
☐ Indifferent
- 6.**
(Optional) Please feel free to provide any general feedback below.

Done. Show me the finishing code.

Figure B.19 – Final survey presented to the participants

C Study Material for Chapter 6

In this appendix, we provide the material for the study reported in Chapter 6.

C.1 Introductory Material

The participants are presented with the instructions in Figure C.1, followed by a demographics survey (Figure C.2 and C.3) and a reading test (Figure C.4).

We are researchers from the University of Michigan in the U.S. and EPFL in Switzerland. We are trying to understand the mindset of individuals regarding the privacy practices of the different services, apps and websites they use.

To help us in our research, we need you to evaluate a set of **question-answer pairs** about the privacy policies of different companies. In particular, we need your feedback regarding **how relevant/satisfying** you find the displayed answer to each question.

Please Note: We have a minimum duration of 15 seconds before you can move to the next question.

Before we jump in to the main study task, we will ask you to answer few background questions. Please answer those questions honestly. The answers to those questions will NOT impact your participation in this study or the received compensation.

>>

Figure C.1 – General instructions at the beginning of the study

Demographics

Please answer the following questions honestly. The answers to those questions will NOT impact your participation in this study or the received compensation.

Please provide your age (in years):

Please provide your gender:

☐ male

☐ female

☐ I prefer not to tell

Which of the following best describes your highest achieved education level?

☐ No high school

☐ Some high school

☐ High school graduate

☐ Some college - no degree

☐ Associates/2 year degree

☐ Bachelors/4 year degree

☐ Graduate degree - Masters, PhD, professional, medicine, etc.

Figure C.2 – First part of the demographics survey

Which of the following best describes your primary occupation?

- ☐ Administrative support (e.g., secretary, assistant)
- ☐ Art, writing, or journalism (e.g., author, reporter, sculptor)
- ☐ Business, management, or financial (e.g., manager, accountant, banker)
- ☐ Computer engineer or IT professional (e.g., systems administrator, programmer, IT consultant)
- ☐ Education (e.g., teacher)
- ☐ Engineer in other fields (e.g., civil engineer, bio-engineer)
- ☐ Homemaker
- ☐ Legal (e.g., lawyer, law clerk)
- ☐ Medical (e.g., doctor, nurse, dentist)
- ☐ Retired
- ☐ Scientist (e.g., researcher, professor)
- ☐ Service (e.g., retail clerks, server)
- ☐ Skilled labor (e.g., electrician, plumber, carpenter)
- ☐ Student
- ☐ Unemployed
- ☐ Decline to answer
- ☐ Other

Please provide your place of residence:

- ☐ North America (United States, Canada)
- ☐ South America
- ☐ Europe
- ☐ Asia & Middle East
- ☐ Africa
- ☐ Australia

Figure C.3 – Second part of the demographics survey

Please read the text below and think of the word which best fits each space. Use only one word in each space.

In fiction, robots have a personality, (1) _____ reality is disappointingly different. Although sophisticated (2) _____ to assemble cars and assist during complex surgery, modern robots are dumb automations, (3) _____ of striking up relationships with their human operators. However, change is (4) _____ the horizon. Engineers argue that, as robots begin to make (5) _____ a bigger part of society, they will need a way to interact with humans. To this end they will need artificial personalities.

(1)

(2)

(3)

(4)

(5)

Figure C.4 – Cloze reading test, with 5 blank values, used to judge the reading comprehension abilities of the participants and to remove the low-performing ones.

C.2 Answer Evaluation

The participants are then shown the instructions in Figure C.5, followed by 17 question-answer pairs. Figure C.6 shows an example of each pair.

- For the rest of this study, we will present you with **17 question-answer pairs**.
- At each step, we will give you a **question** somebody asked about the **privacy practices** of a company.
- Below the question, you will find a **candidate answer, which might or might not be relevant**.
- Please read both the question and the candidate answer carefully, and indicate **how relevant do you find the candidate answer to the posed question**.
- We have given you the **company name and Twitter bio** to supply you with additional context, in case you need it.
- You have to choose one of 5 choices, ranging from **Definitely Relevant** to **Completely Irrelevant**.
- **Beware:** There are some steps, where the evaluation is very obvious, so if you are filling it without reading carefully, this will lead to **rejecting your HIT**.

>>

Figure C.5 – Instructions for the answer evaluation part

Company: nswpolice

Twitter Description: The Official Twitter account of the NSW Police Force. Please do not report crime here. For emergencies, contact 000 or 131 444 for non urgent matters.

Question: @nswpolice what sort of data do you track outside the questions for visitors to the link?

Answer: Please note that mobile phone users accessing this website should be aware that most phone applications enable their user's mobile phone details (and in some phone applications, also their user's geographic location) to be recorded by the NSWPF's website host. That information will be kept secure, confidential and private, in accordance with the applicable laws.

How relevant is the candidate answer to the given question?

☐ **Definitely Relevant:** It perfectly answers the question.

☐ **Undecided:** I find it too difficult to give a judgment on this pair.

☐ **Partially Relevant:** It answers the bulk of the question, though there might be more to say.

☐ **Definitely Irrelevant:** It totally misses the topic of the question.

☐ **Partially Irrelevant:** It doesn't answer the question; only has a slight clue.

Figure C.6 – Example question-answer pair, given to the participants for evaluation

We list below the 120 questions around privacy policies that we use in our evaluation (collected from Twitter and agreed on during the annotations by the author and another member of the research team.).

1. @Monsterjobs_uk Are you legally responsible for any loss of claimants' data that may occur on Universal Job Match?
2. @Kenshoo I know I can just check your website, but are you taking any personal data while you are looking for our search queries?
3. Just tried to change my details at @theregister and it now wants my address and phone number. Why?
4. @NorthumbrianH2O thanks. Do you pass on customer addresses to 3rd parties? Got interiors catalogue addressed to me here. How did they know?
5. @EE May I please request what companies my details have been passed on to? I am getting calls from various elec suppliers. Many thanks.
6. @creditkarma Does cancelling an account also delete all associated data (especially SSN) from your system? Want to know before I sign up. :)
7. @yewknee @Simplify does it simplify sharing your banking data with advertisers?
8. @TechSmith Do you collect information and provide it to third parties?
9. @moneysupermarktUK Hi - if I use your service will there be ANY telephone calls from you or 3rd party company or is it 100% web based?
10. @Viber So, can everyone in your contacts see the photos you have shared on Viber even if not originally shared with them? @fit_gurl
11. @TradeMe Isn't releasing information under the Privacy Act voluntary? I.e to protect users you could make the Police follow formal process?
12. @FitbitSupport is data stored on the cloud? I heard of leaks from the cloud.
13. @SparkMailApp Is there more information about sync settings via cloud. Security? Possible to delete what's been synced?
14. @nswpolice what sort of data do you track outside the questions for visitors to the link?
15. @weebly shocked at the unsolicited emails and calls I'm getting since I signed up with your web service. Did you REALLY sell my information?
16. @Prezi Guys, small question: the email addresses you collect from Prezi accounts, are they being used for third parties? If so, why?
17. @myen Are Evernote notes encrypted at rest?
18. @quip quick question if I connect my accounts can you access all my info? Or it still remains just for me to see?
19. .@AngiesList so do you sell your mailing list to everyone??? My junk email has increased exponentially since joining. #sheesh
20. @fullcontact do you have a warrant canary statement that you've never provided users' address books to authorities? If not, can you?
21. @getspeedify, does Speedify encrypt my traffic or is it an unencrypted VPN? Also, do you keep logs of users? If so, for how long?
22. @duckduckgo You don't log user info, but what bout cookies ? Do you use them every time we log on ?

23. @AskSubaruCanada So that makes it ok for you to give them personal info to spam customers with every day? Where was the opt out option?
24. @skulpt_me Very interested in your device! Can you tell us how our personal info and data are used and/or resold with your app?
25. @loseit how about you guys? Do you #share the data we log in your app? #Privacy <https://t.co/qTH6Ir905A>
26. @EuclidAnalytics are you able to isolate a MAC address' data and provide it to law enforcement?
27. @threatspikelabs what data do u store, if any, how long u store for? 3rd party compliance? If served with warrant what's ur steps to protect
28. @Adobe do you sell Mail Addresses??!!
29. @FreePPICheck do you keep or pass over any personal information after completing your ppi check??
30. Hey @Optus why am I getting calls from people wanting to sell me funeral insurance? Have you sold my phone number to a call centre?
31. @nest Do you keep your customer's emails private after you have obtained an email during installation setup?
32. @smallpdf are your applications HIPAA compliant?
33. @SpotifyCares where is the opt out option for shareing personal info. If I opt out of the terms I am told that I cannot use Spodify?
34. @msg @ProductHunt @service Curious how personal info is protected, assuming u have to give it out for most cust serv resolutions?
35. Also, @HotDocOnline you make no mention on your site of how patient data is secured. Would you like to elaborate in public?
36. Does anybody know if @EE sell on emails? I've been inundated with junk mail since I got my new contract the other week...
37. @carmillaseries how secure is the merch store? I want to buy myself something for my birthday but I'm afraid to use my credit card online.
38. @carshare hey folks, what's the best way to reach you about security disclosure of your service and potential access to customer data?
39. @submittable why do I suddely need to enable cookies? Is there an opt out, or am I switching back to stamps?
40. Latest @bankofireland iPhone app update wants constant background access to my location. For security, marketing, or something else?
41. MT @Remind101 @cellyme ...What do you do with the phone numbers that are archived? Your current policy?
42. @AirsidesInsider Could you provide some technical docs about the security used for the #MobilePassport apps? Saved locally, encryption, etc.
43. Ok @HRBlock, why would @Ghostery report over 16 advertising trackers from your supposedly secure online tax application?

44. @angrybirds is this true? <http://t.co/gCBoFicZ> you send people's contacts to 3rd parties without permission?
45. @LinkedInHelp so someone is selling my email address from you and signing me up? Are you saying that you've nothing to do with this email?
46. @floatapp how secure are your servers? Can you direct me to the security info on your website please?
47. @Gopit_Search @PrivacyMatters what's your privacy statement and what data do you store from your users?
48. .@AskTarget ok thanks but I assume that means yes you all do sell patient names and addresses?
49. @opendns I'd like to know if there's any security concern, like whether the DNS provider can track my browsing. What are the privacy issues?
50. @swiftkey can your app NOT collect my passwords and credit card numbers? How is that legal?
51. @HootSuite_Help If I sign up: Where's the option on your website to opt out of your sharing my personal info?
52. Also, can anyone at @TrustifyPI guarantee that the emails people put in to check against this "list" won't be sold off? No TOS on app.
53. @SagiGidali Hi Sagi, what's your stance on keeping customer logs, and where is your company/customer data based for legal reasons?
54. @Telstra just wondering what's the extent of your monitoring on customers (me). Link me to pds if possible?
55. @TTChelps Yes, how will you manage my travel records and contact info, under what circumstances will you release to 3rd parties?
56. @AirbnbHelp you already have my phone number, my linkedin and my profile pic + feedback from previous hosts. So why you want my ID? @Airbnb
57. @hushmail Q: does hushmail collaborate with NSA to spy on its users' emails? <http://t.co/aZqiuL6sja> looking for alternative to Google mail
58. @EtsyHelp is it safe putting my banking info to etsy?? I have unpaid student loans, and I don't want them paying off @Etsy for my info :P
59. @troyhunt @FreedomeVPN do they log traffic ? Port open for upload ? Rhanks :)
60. @onavo Can I ask - why is it free? and how do u guarantee that data is anonymised? Ta :)
61. So @stripe has entered the UK market (nice) shouldn't they declare if they share user data with any one?
62. @asiaelle @graceishuman I like Evernote for some things but I worry about data security. Who can see my pages ?
63. The @nest needs to collect your in and out patterns for all your family. Who owns that data? Can it be subpoenaed? federated? @mdrasch #IoT
64. @tigerVPN Can't wait for the IOS app! It seem that I can't find info if u log or no log to guarantee our privacy?
65. @duckduckgo -is it really private. Does duckduckgo-follow me and watch me and spy on me? Can I truly search in total privacy?

66. @Telus @Shawhelp @Shawinfo Can you verify if you provide customer information without a warrant to law enforcement but upon request? #bcpoli
67. @nest Is Nest sharing data with Google still optional? Will it remain that way for the foreseeable future?
68. @duggan My money is on marketing plus hubris. @bankofireland, how confident are you that data won't leak?
69. .@fitbit What are you doing to protect customers' privacy?
70. @TMobileHelp where's the part of my contract where I gave you permission to log my urls and location?
71. @Netflixhelps The Perfect World Peanut Labs offer to earn zen for signing up. Will my information be confidential or is it shared with PWE?
72. Hey @indeed, do you sell info to 3rd party sites?
73. @Jawbone how do you protect all the information tracked by your wearables? #CIS210
74. I understand that @airbnb want to see my ID's before booking, but can I know what they are doing with that data ? #privacy #matters
75. @privatwifi Will do more research, but how do I know VPN software isn't gathering my data/personal info?
76. @mysms Can you elaborate on prvcy policy? Can employees access sms? If so, when/when would they? What safeguards exist to prevent abuse? Thx!
77. @mrgunn @colwizSupportWhat kind of data do you collect from users? What do you do with it? Who stands to gain from it financially? How?
78. @Viber is that truth you are spying on users's calls ??
79. @22seven - do you, would you, could you ever share personal data with SARS? Is there anything in your T&C's prohibiting you to do so?
80. Hey @GoDaddy why do you guys sell my information every time I buy a domain from you? I'm assuming I gave you permission at some point?
81. @sprintcare @sprint - What's up with the new privacy contract you guys just did? Giving away our info to random businesses ? #Sketchy #Smh
82. @opera Do you track and store people's data like Google, Microsoft, Facebook and all their other cronies? If you don't I'll switch to you.
83. Does @Official_GDC have a privacy statement anywhere re: the personal data of attendees & how it's shared w/ exhibitors, speakers, etc.?
84. Is this correct? @VodafoneIreland cc rep said they don't retain customer information predating your existing or last contract. cc: @ComReg
85. Hey @HostGator can you not sell my phone number to telemarketers? Paying for privacy protection should protect me from YOU TOO
86. With the rapid rise in so called encrypted messaging apps, how do you feel @viber competes on security? #cgc14
87. @UnrollmeHelp is there a way i can be sure that you're not just the nsa reading my mails? #faq

88. @evernotehelps are my notes being saved encrypted on your servers per default? Or is only manually encrypted text encrypted?
89. @Telstra So nothing more than absolutely necessary to meet legal interception and not used unless required under law?
90. @Viber if I sign up with viber, do you load my iphone contacts into your servers? Thanks #viber
91. @AutomaticHelp hi! how is the sensitive data my Automatic collects encrypted and stored?
92. @truecallerhelp good, thx. The question was different: will they be removed from your servers or not? Which is the procedure to remove them?
93. Sean: What Mobile Apps Know & Transmit About You: @AngryBirds sends my contacts to third parties? #WTF #FAIL <http://t.co/IKVYc6l7>
94. @simpletaxca What personal information do you retain after I do my taxes on your site?
95. @MailChimp Does Mail Chimp retain our list and use or sell them elsewhere ?
96. @FreePPICheck if I were to give you my phone #, how many people will you sell it to?
97. @TripCase how safe is the personal info?
98. @fitbit @FitbitSupport Where can I go to see who you sold my private health data too? <http://t.co/Rd64dKWGFb>
99. @VentraChicago How do you use our personal data once we've registered with Ventra? #AskVentra
100. @bitly Thanks for the response. Does bitly have access to the links? What if I wanted to send a file to a friend and it's personal?
101. @jobsdotie Also, no data privacy guarantee, or info on who gets access to my CV (is it just the advertising company or also jobs.ie staff?).
102. @nest Privacy question, does Nest share usage patterns or anything with third parties? cc @joshmend
103. The @tapjoy ad framework uploads my UDID *and* my MAC address? Is that *really* necessary? :|
104. Dear @eBay I don't appreciate my buyers having access to my phone number! Why is this person calling me at 9pm? #totallyinappropriate
105. Soooo, I joined @mint about a week ago and today I've received 7 credit card offers in the mail... selling my info much?
106. Hi @auspost how can a person opt out on Australia Post storing phone numbers ?
107. .@MGMResortsIntl do you sell your mlife member contact information? Receiving calls from sports betting tip line since staying @ NYNY.
108. Im conflicted 2 clicking "accept" to your policy changes. Why do you need my birth date, friends info, etc.? @SpotifyUSA #TaylorMightBeRight
109. @AdblockPlus <https://t.co/3Awum5BwRF> How much control will users have over their personal data? #tracking #adblock
110. @clue just curious. do you share users information with third parties? i'm getting targeted ads for tampons etc. after using your app.

111. @theTunnelBear nice! Given the recent U.K. change in law, do you have any details about your logging privacy etc?
112. Is customer credit card info stored encrypted? Even if it is storing password in plain text kind of kills that. @RSComponents
113. @davidsbridal when people sign up for your site, do you sell their info? I have received numerous unsolicited calls on my cell phone.
114. @Cabelas Do you or your partners sell, share, or otherwise disseminate your customer's mailing addresses directly or indirectly to the @NRA?
115. @Spotify Due to the complete lack of respect for privacy in the new T&C I wonder where to delete my account? stopped subscription already.
116. @hubspot Why must I enable third party cookies in your browser settings? Is there a workaround? DM pls
117. @Viber Hi, is your service secured against spying by the nsa and gchq?
118. .@automatic Who owns my driving data? Will my driving behavior be aggregated and sold?
119. @ProtonMail @ProtonMailHelp E.g.: If I use my IOS to access emails, what data is stored by you guys? Is there a link to explain?
120. @Nosgoth - is it mandated to link our steam account to Square in order to play the game ? What securities are in place for protection of acc

C.3 Final Survey

Finally, the participants are presented with a short survey to check their legal expertise (Figure C.7).

Legal Background

Please answer the following questions honestly. The answers to those questions will NOT impact your participation in this study or the received compensation.

How easy or difficult is it for you to understand legal texts, e.g., a privacy policy of a website or a legal contract.

☐ Very Difficult

☐ Difficult

☐ Neutral

☐ Easy

☐ Very Easy

What is your level of legal training?

☐ No legal training

☐ No legal training, but my background in another field provides me with some legal experience

☐ Knowledgeable in legal matters, but no formal legal training

☐ Studied law

☐ Received other legal or paralegal training

Do you work in a position that requires legal expertise (e.g., working as a paralegal, lawyer or attorney)?

☐ Yes

☐ No

Figure C.7 – Final survey for checking the legal expertise of the participants

D Example Cases for PriBot

In this appendix, we show PriBot in action with questions about multiple companies.

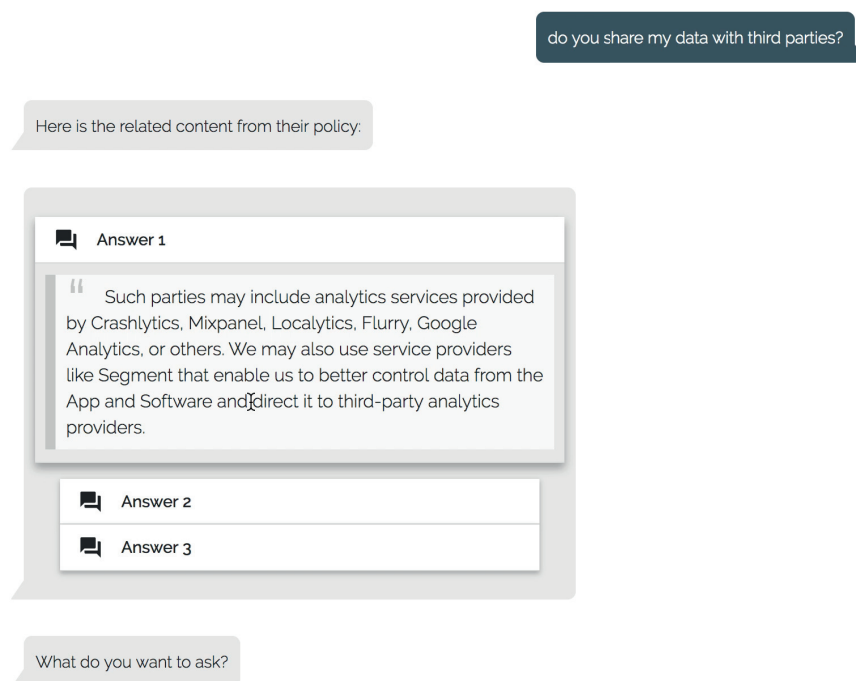


Figure D.1 – The first answer about third-party sharing in the case of the headphones company “Bose” [Bra17]

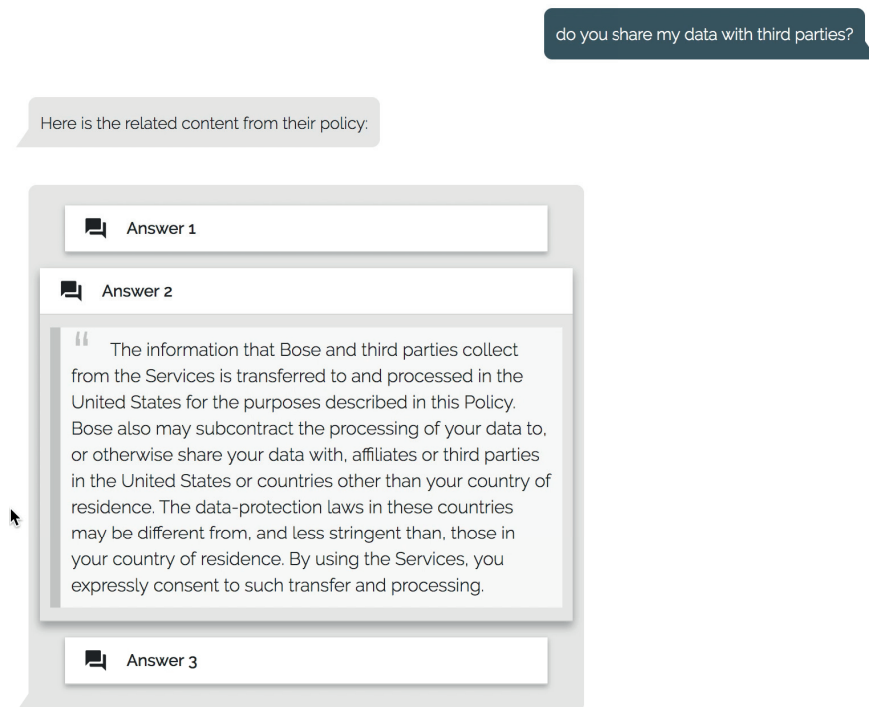


Figure D.2 – The second answer about third-party sharing in the case of the headphones company “Bose” [Bra17]

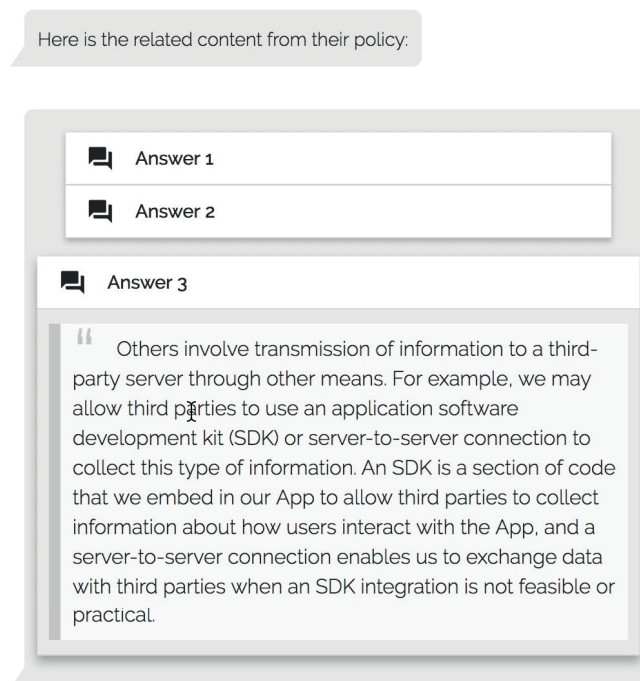


Figure D.3 – The third answer about third-party sharing in the case of the headphones company “Bose” [Bra17]

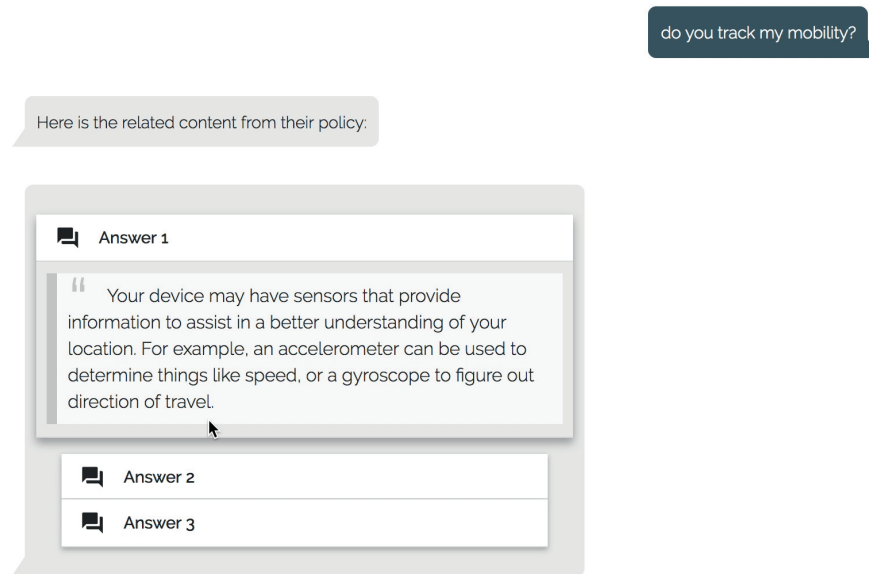


Figure D.4 – The first answer about mobility tracking in the case of Google. Notice the semantic matching between mobility and location data.

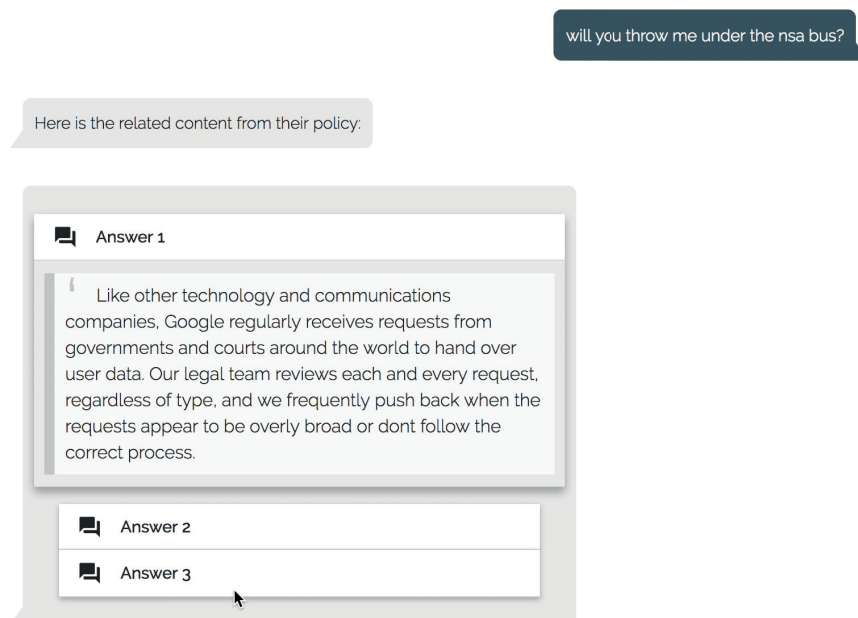


Figure D.5 – The first answer to an informal question about providing data to the NSA in the case of Google. “NSA” is semantically matched to “government”.

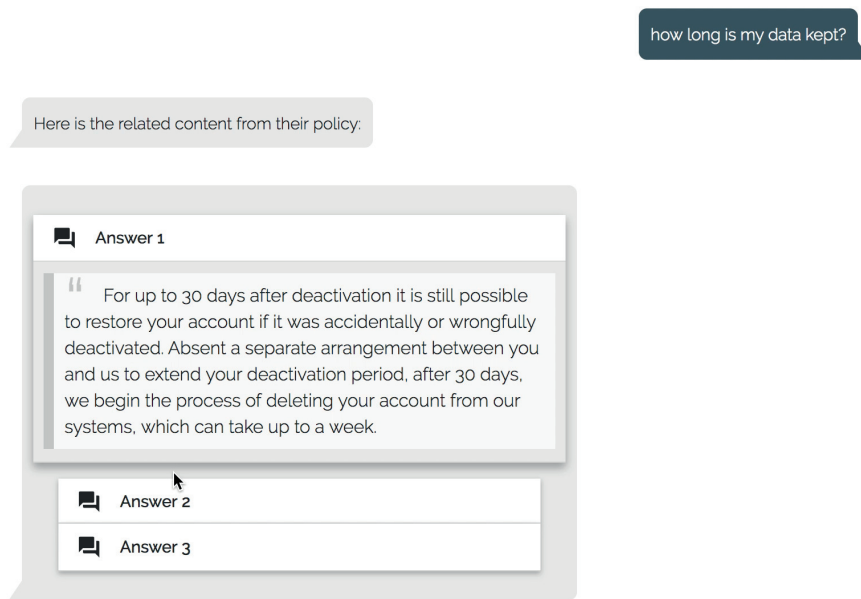


Figure D.6 – The first answer to a question about data retention in the case of Twitter. Notice the absence of matched terms between the question and the answer.

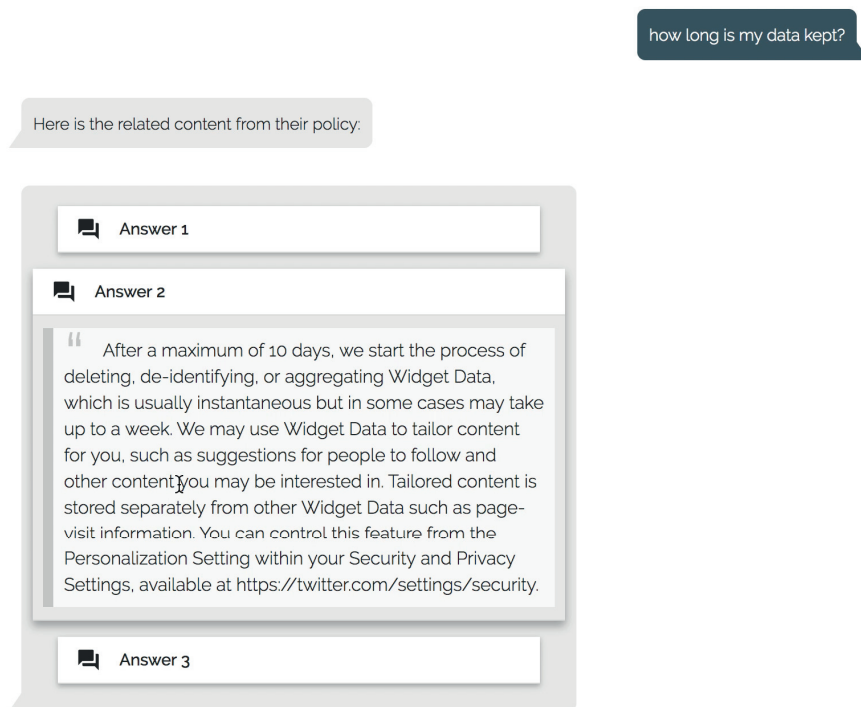


Figure D.7 – The second answer to a question about data retention in the case of Twitter. Notice the different duration given for a different data type, showing the interesting patterns that can be observed by comparing the top answers.

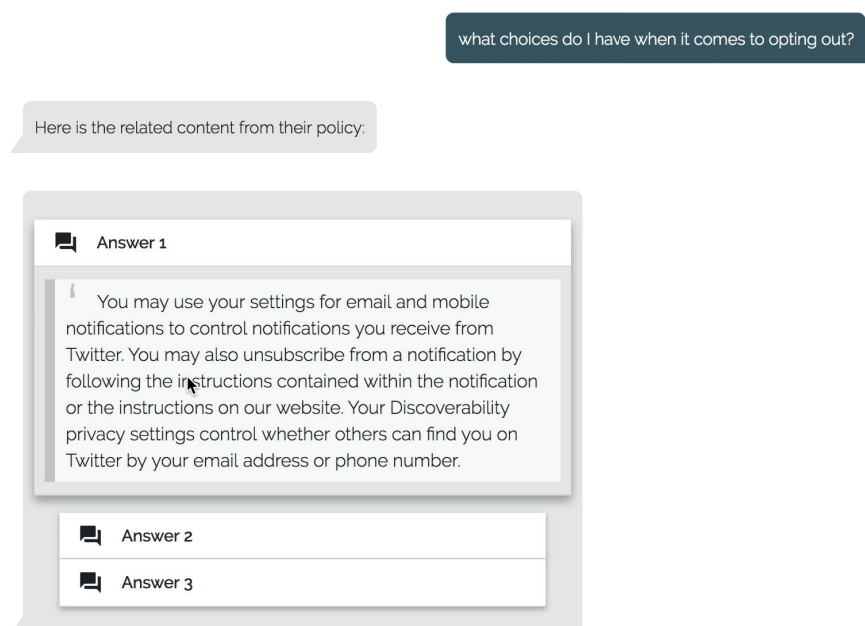


Figure D.8 – The first answer to a question about data control in the case of Twitter. Again, opting out is a form of data control, and the answers are in this spirit.

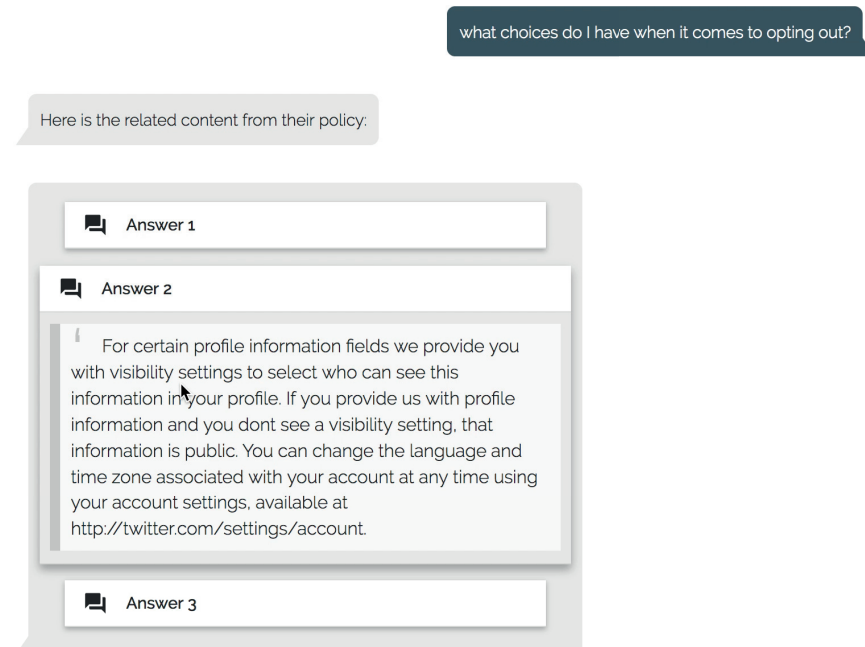


Figure D.9 – The second answer to a question about data control in the case of Twitter. Other forms of data control are provided in this answer.

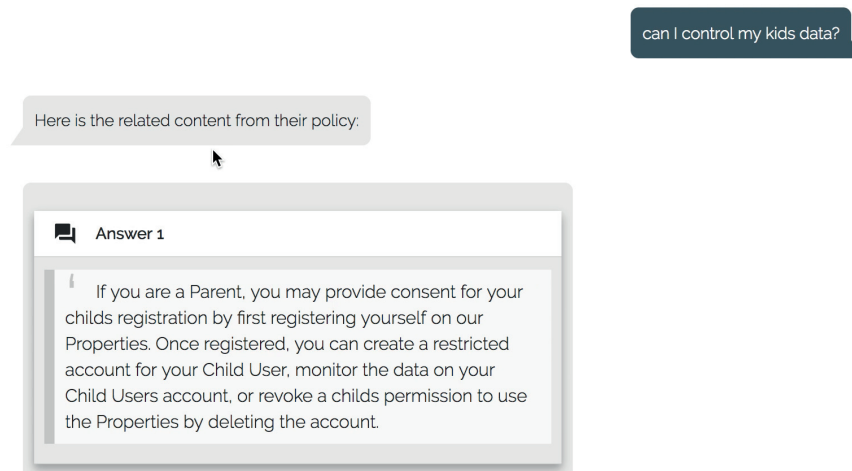


Figure D.10 – The first answer to a question about data control in the case of Khan Academy. Notice the semantic matching between “kids” and “child” and the high level understanding of the user’s interest in data control options.

Bibliography

- [AAB⁺17] Alessandro Acquisti, Idris Adjerid, Rebecca Hunt Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. Nudges for privacy and security: Understanding and assisting users' choices online. 2017.
- [ABKS99] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pages 49–60, New York, NY, USA, 1999. ACM.
- [ABL15] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.
- [Acq09] Alessandro Acquisti. Nudging privacy: The behavioral economics of personal information. *IEEE Security & Privacy*, 7(6), 2009.
- [AH13] Yuvraj Agarwal and Malcolm Hall. ProtectMyPrivacy: detecting and mitigating privacy leaks on iOS devices using crowdsourcing. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pages 97–110. ACM, 2013.
- [AKJ⁺15] Bonnie Brinton Anderson, C. Brock Kirwan, Jeffrey L. Jenkins, David Eargle, Seth Howard, and Anthony Vance. How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2883–2892, New York, NY, USA, 2015. ACM.
- [All11] Anita Allen. *Unpopular privacy: what must we hide?* Oxford University Press, 2011.
- [Alt77] Irwin Altman. Privacy regulation: culturally universal or culturally specific? *Journal of Social Issues*, 33(3):66–84, 1977.
- [ASS⁺15] Hazim Almuhiemedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your location

- has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 787–796, New York, NY, USA, 2015. ACM.
- [Bak01] Frank B Baker. *The basics of item response theory*. ERIC, 2001.
- [BC13] Gergely Biczók and Pern Hui Chia. Interdependent privacy: Let me share your data. In *Financial Cryptography and Data Security*, pages 338–353. Springer, 2013.
- [BCKM04] Simon Byers, Lorrie Faith Cranor, Dave Kormann, and Patrick McDaniel. Searching for privacy: Design and implementation of a P3P-enabled search engine. In *International Workshop on Privacy Enhancing Technologies*, pages 314–328. Springer, 2004.
- [BGH⁺97] M Beaulieu, M Gatford, Xiangji Huang, S Robertson, S Walker, and P Williams. Okapi at TREC-5. *NIST Special Publication*, pages 143–166, 1997.
- [BGJM16] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
- [BHA⁺13] Igor Bilogrevic, Kévin Huguenin, Berker Agir, Murtuza Jadliwala, and Jean-Pierre Hubaux. Adaptive information-sharing for privacy-aware mobile social networks. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 657–666, New York, NY, USA, 2013. ACM.
- [BLKC⁺13] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. Your attention please: designing security-decision UIs to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 6. ACM, 2013.
- [Blo64] Edward J Bloustein. Privacy as an aspect of human dignity: An answer to dean prosser. *NYUL Rev.*, 39:962, 1964.
- [BMBW15] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.
- [Bra17] Russell Brandom. Bose Connect app is sharing private listening data, claims lawsuit. <http://www.theverge.com/2017/4/19/15356108/bose-connect-private-listening-song-list-sharing-data-lawsuit>, 2017. Accessed: 2017-04-27.
- [Cam16] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

- [Car11] Nicholas Carr. *The shallows: What the Internet is doing to our brains*. WW Norton & Company, 2011.
- [Cat10] F. H. Cate. The limits of notice and choice. *IEEE Security Privacy*, 8(2):59–62, March 2010.
- [CBDC14] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.
- [CdHP13] Elisa Costante, Jerry den Hartog, and Milan Petković. What websites know about you. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 146–159. Springer, 2013.
- [CGA06] Lorrie Faith Cranor, Praveen Guduru, and Manjula Arjula. User interfaces for privacy agents. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(2):135–178, 2006.
- [CLL⁺15] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. In *AAAI*, pages 2210–2216, 2015.
- [CLLH16] David J Crandall, Yunpeng Li, Stefan Lee, and Daniel P Huttenlocher. Recognizing landmarks in large-scale social image collections. In *Large-Scale Visual Geo-Localization*, pages 121–144. Springer, 2016.
- [CLM⁺02] Lorrie Cranor, Marc Langheinrich, Massimo Marchiori, Martin Presler-Marshall, and Joseph Reagle. The platform for privacy preferences 1.0 (P3P1. 0) specification. *W3C recommendation*, 16, 2002.
- [Cra12] Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273, 2012.
- [Cra13] Lorrie Cranor. *Cambridge English Proficiency Certificate of Proficiency in English CEFR level C2, Handbook for Teachers*. University of Cambridge, 2013.
- [CSPdH12] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. A machine learning solution to assess privacy policy completeness:(short paper). In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, pages 91–96. ACM, 2012.
- [CYA12] Pern Hui Chia, Yusuke Yamamoto, and N. Asokan. Is this app safe?: A large scale study on application permissions and risk signals. In *Proceedings of the 21st International Conference on World Wide Web, WWW '12*, pages 311–320, New York, NY, USA, 2012. ACM.
- [DA09] Rafael Jaime De Ayala. *Theory and practice of item response theory*. Guilford Publications, 2009.

- [DBE07] Doug Downey, Matthew Broadhead, and Oren Etzioni. Locating complex named entities in web text. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07*, pages 2733–2739, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.
- [DKL07] Hoa Trang Dang, Diane Kelly, and Jimmy J Lin. Overview of the TREC 2007 question answering track. In *Trec*, volume 7, page 63, 2007.
- [Eco14] Economist. Something to stand on. <http://www.economist.com/news/special-report/21593583-proliferating-digital-platforms-will-be-heart-tomorrows-economy-and-even>, 2014. Accessed: 2017-04-06.
- [EGC⁺10] William Enck, Peter Gilbert, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N. Sheth. Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation, OSDI'10*, pages 1–6, Berkeley, CA, USA, 2010. USENIX Association.
- [Ela15] Elastica Cloud Threat Labs. Q2 2015 shadow data report. pages 1–14, 2015. <https://www.elastica.net/q2-2015-shadow-data-report/>.
- [Ela16] Elastica Cloud Threat Labs. 1H 2016 shadow data report. 2016. <https://www.elastica.net/1h-2016-shadow-data-report/>.
- [FBX⁺17] Felix Fischer, Konstantin Böttinger, Huang Xiao, Christian Stransky, Yasemin Acar, Michael Backes, and Sascha Fahl. Stack Overflow considered harmful? the impact of copy&paste on Android application security. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 121–136. IEEE, 2017.
- [FCH⁺11] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. In *Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS '11*, pages 627–638, New York, NY, USA, 2011. ACM.
- [Fed12] Federal Trade Commission. Protecting consumer privacy in an era of rapid change, March 2012.
- [Fed15] Federal Trade Commission. Internet of things, privacy & security in a connected world, Jan. 2015.
- [Fer13] Gregory Ferenstein. Google's Cerf says “privacy may be an anomaly”. historically, he's right. <https://techcrunch.com/2013/11/20/googles-cerf-says-privacy-may-be-an-anomaly-historically-hes-right/>, 2013. Accessed: 03-11-2017.
- [Fer15] Gregory Ferenstein. The birth and death of privacy: 3,000 years of history told through 46 images. <https://medium.com/the-ferenstein-wire/the-birth-and-death-of-privacy-3-000-years-of-history-in-50-images-614c26059e>, 2015. Accessed: 03-11-2017.

- [FGW11] Adrienne Porter Felt, Kate Greenwood, and David Wagner. The effectiveness of application permissions. In *Proceedings of the 2Nd USENIX Conference on Web Application Development*, WebApps'11, pages 7–7, Berkeley, CA, USA, 2011. USENIX Association.
- [FXG⁺15] Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, pages 813–820, 2015.
- [GDG⁺05] Nathaniel Good, Rachna Dhamija, Jens Grossklags, David Thaw, Steven Aronowitz, Deirdre Mulligan, and Joseph Konstan. Stopping spyware at the gate: a user study of privacy, notice and spyware. In *Proceedings of the 2005 symposium on Usable privacy and security*, pages 43–52. ACM, 2005.
- [Gen09] Craig Gentry. *A fully homomorphic encryption scheme*. PhD thesis, Stanford University, 2009.
- [Ger78] Robert S Gerstein. Intimacy and privacy. *Ethics*, 89(1):76–81, 1978.
- [GGMK07] Nathaniel S Good, Jens Grossklags, Deirdre K Mulligan, and Joseph A Konstan. Noticing notice: a large-scale experiment on the timing of software license agreements. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 607–616. ACM, 2007.
- [GHH⁺12] Loretta Garrison, Manoj Hastak, Jeanne M Hogarth, Susan Kleimann, and Alan S Levy. Designing evidence-based disclosures: A case study of financial privacy notices. *Journal of Consumer Affairs*, 46(2):204–234, 2012.
- [GHW15] Yuri Gurevich, Efim Hudis, and Jeannette M. Wing. Inverse privacy (revised). Technical Report MSR-TR-2015-37, May 2015.
- [GHW16] Yuri Gurevich, Efim Hudis, and Jeannette M. Wing. Inverse privacy. *Commun. ACM*, 59(7):38–42, June 2016.
- [GM13] Glenn Greenwald and Ewen MacAskill. NSA Prism program taps in to user data of Apple, Google and others. *The Guardian*, 7(6):1–43, 2013.
- [GNP16] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised text segmentation using semantic relatedness graphs. In *The Fifth Joint Conference on Lexical and Computational Semantics*, pages 125–130, Stroudsburg, Pa., 2016. Association for Computational Linguistics.
- [Goo17] Joanna Goodman. Legal technology: the rise of the chatbots. <https://www.lawgazette.co.uk/features/legal-technology-the-rise-of-the-chatbots/5060310.article>, 2017. Accessed: 2017-04-27.

- [GPKC13] Vaibhav Garg, Sameer Patil, Apu Kapadia, and L Jean Camp. Peer-produced privacy protection. In *IEEE International Symposium on Technology and Society*, pages 147–154, 2013.
- [GSF⁺16] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 321–340, Denver, CO, 2016. USENIX Association.
- [HA17] Hamza Harkous and Karl Aberer. "If you can't beat them, join them": A usability approach to interdependent privacy in cloud apps. In *Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, CODASPY'17*, pages 127–138, New York, NY, USA, 2017. ACM.
- [HAHT17] Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. Quantifying interdependent risks in genomic privacy. *ACM Transactions on Privacy and Security (TOPS)*, 20(1):3, 2017.
- [HBW08] Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.
- [HFL⁺17] Hamza Harkous, Kassem Fawaz, Remi Lebrete, Florian Schaub, Kang G. Shin, and Karl Aberer. PriBot: Answering free-form questions about privacy policies with deep learning. Technical report, 2017.
- [HFSA16] Hamza Harkous, Kassem Fawaz, Kang G. Shin, and Karl Aberer. PriBots: Conversational privacy with chatbots. In *Workshop on the Future of Privacy Notices and Indicators, SOUPS 2016, Denver, CO, USA, June 22, 2016*. USENIX Association, 2016.
- [HHWS14] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14*, pages 2647–2656, New York, NY, USA, 2014. ACM.
- [HJ16] Jean-Pierre Hubaux and Ari Juels. Privacy is dead, long live privacy. *Commun. ACM*, 59(6):39–41, May 2016.
- [HMSW13] Markus Huber, Martin Mulazzani, Sebastian Schrittwieser, and Edgar Weippl. Appinspect: Large-scale evaluation of social networking apps. In *Proceedings of the First ACM Conference on Online Social Networks, COSN '13*, pages 143–154, New York, NY, USA, 2013. ACM.
- [HN09] Daniel C Howe and Helen Nissenbaum. TrackMeNot: Resisting surveillance in web search. *Lessons from the Identity trail: Anonymity, privacy, and identity in a networked society*, 23:417–436, 2009.

- [HRA14] Hamza Harkous, Rameez Rahman, and Karl Aberer. C3P: Context-aware crowd-sourced cloud privacy. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 102–122. Springer, 2014.
- [HRKA16] Hamza Harkous, Rameez Rahman, Bojan Karlas, and Karl Aberer. The curious case of the PDF converter that likes Mozart: Dissecting and mitigating the privacy risk of personal cloud apps. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, volume 2016, 2016.
- [Hwa13] Tim Hwang. The laws of (legal) robotics. Technical report, Robot, Robot & Hwang LLP, 2013.
- [HZH11] Leif-Erik Holtz, Harald Zwingelberg, and Marit Hansen. Privacy policy icons. In *Privacy and Identity Management for Life*, pages 279–285. Springer, 2011.
- [IBC⁺13] Iulia Ion, Filipe Beato, Srdjan Capkun, Bart Preneel, and Marc Langheinrich. For some eyes only: Protecting online information sharing. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy, CODASPY'13*, pages 1–12, New York, NY, USA, 2013.
- [Isa17] Mike Isaac. Uber's C.E.O. plays with fire. <https://www.nytimes.com/2017/04/23/technology/travis-kalanick-pushes-uber-and-himself-to-the-precipice.html>, 2017. Accessed: 2017-04-27.
- [ISKC11] Iulia Ion, Niharika Sachdeva, Ponnurangam Kumaraguru, and Srdjan Capkun. Home is safer than the cloud!: Privacy concerns for consumer cloud storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pages 13:1–13:20, 2011.
- [JK02] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.
- [JSD⁺14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [Kal16] Paul Kalanithi. *When breath becomes air*. Random House, 2016.
- [KBCR09] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 4:1–4:12, New York, NY, USA, 2009. ACM.
- [KCS13] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13*, pages 3393–3402, New York, NY, USA, 2013. ACM.

- [Kim14] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIG-DAT, a Special Interest Group of the ACL*, pages 1746–1751, 2014.
- [KJ13] Bart P Knijnenburg and Hongxia Jin. The persuasive effect of privacy recommendations for location sharing services. In *Twelfth Annual Workshop on HCI Research in MIS*, 2013.
- [KKK⁺13] J.C. Klontz, B.F. Klare, S. Klum, A.K. Jain, and M.J. Burge. Open source biometric recognition. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*, pages 1–8, Sept 2013.
- [KPSW11] Bastian Konings, David Piendl, Florian Schaub, and Michael Weber. Privacy-Judge: Effective privacy controls for online published information. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 935–941. IEEE, 2011.
- [Kri15] Golden Krishna. *The Best Interface is No Interface: The Simple Path to Brilliant Technology*. Pearson Education, 2015.
- [KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [KWH10] Bart P Knijnenburg, Martijn C Willemsen, and Stefan Hirtbach. Receiving recommendations and providing feedback: The user-experience of a recommender system. In *International Conference on Electronic Commerce and Web Technologies*, pages 207–216. Springer, 2010.
- [LAS⁺16] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhiemedi, SA Zhang, Norman Sadeh, Alessandro Acquisti, and Yuvraj Agarwal. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Symposium on Usable Privacy and Security*, 2016.
- [Lev17] Ari Levy. Microsoft CEO Satya Nadella: for the future of chat bots, look at the insurance industry. <http://www.cnn.com/2017/01/09/microsoft-ceo-satya-nadella-bots-in-insurance-industry.html>, 2017. Accessed: 2017-04-27.
- [LFL16] Fei Liu, Nicole Lee Fella, and Kexin Liao. Modeling language vagueness in privacy policies using deep neural networks. In *2016 AAAI Fall Symposium Series*, 2016.
- [Lin94] John Michael Linacre. Sample size and item calibration stability. *Rasch measurement transactions*, 7(4):328, 1994.

- [LK77] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [LKG⁺08] Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social networks*, 30(4):330–342, 2008.
- [LT10] Kun Liu and Evimaria Terzi. A framework for computing the privacy scores of users in online social networks. *ACM Transactions on Knowledge Discovery from Data*, 5(1):6, 2010.
- [LWSS16] Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *2016 AAAI Fall Symposium Series*, 2016.
- [MAC⁺02] Sean M McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K Lam, Al Mamunur Rashid, Joseph A Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125. ACM, 2002.
- [MC08] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *ISJLP*, 4:543, 2008.
- [McG59] Phyllis McGinley. A lost privilege. *Province of the Heart*, page 56, 1959.
- [MEAS13] Aaron K Massey, Jacob Eisenstein, Annie I Antón, and Peter P Swire. Automated text mining for requirements analysis of policy documents. In *Requirements Engineering Conference (RE), 2013 21st IEEE International*, pages 4–13. IEEE, 2013.
- [Mei13] Gabriele Meiselwitz. Readability assessment of policies and procedures of social networking sites. In *International Conference on Online Communities and Social Computing*, pages 67–75. Springer, 2013.
- [MH07] P Mair and R Hatzinger. Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9):1–20, 2007.
- [MP14] John O McGinnis and Russell G Pearce. The great disruption: How machine intelligence. will transform the role of lawyers in the delivery of legal services. *Fordham L. Rev.*, 82:3041–3481, 2014.
- [MV02] Milena Mihail and Nisheeth K Vishnoi. On generating graphs with prescribed vertex degrees for complex network modeling. *Position Paper, Approx. and Randomized Algorithms for Communication Networks (ARACNE)*, 142, 2002.
- [MV13] Fragkiskos D. Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95 – 142, 2013. Clustering and Community Detection in Directed Networks: A Survey.

- [New03] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
- [Nis04] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.
- [Nis11] Helen Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.
- [NO11] Michael L Nering and Remo Ostini. *Handbook of polytomous item response theory models*. Taylor & Francis, 2011.
- [Nui] Nuix. The Enron PST Data Set cleansed of PII by Nuix and EDRM. <http://info.nuix.com/Enron.html>. Accessed: 2017-05-03.
- [ODPSM⁺17] Katarzyna Olejnik, Italo Ivan Dacosta Petrocelli, Joana Catarina Soares Machado, Kévin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. SmarPer: Context-aware and automatic runtime-permissions for mobile devices. In *Proceedings of the 38th IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.
- [OHS⁺16] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, Mathias Humbert, and Jean-Pierre Hubaux. Quantifying interdependent privacy risks with location data. *IEEE Transactions on Mobile Computing*, 2016.
- [OLT16] Alexandra OLTEANU. *Probing the Limits of Social Data: Biases, Methods, and Domain Knowledge*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2016.
- [O’N16] Cathy O’Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group (NY), 2016.
- [Oom17] Ooma. Business in the cloud. <http://www.ooma.com/blog/business-in-the-cloud/>, 2017. Accessed: 2017-05-03.
- [Orw09] George Orwell. *Nineteen eighty-four*. Everyman’s Library, 2009.
- [PC15] Frank Pasquale and Glyn Cashwell. Four futures of legal automation. *UCLA L. Rev. Discourse*, 63:26, 2015.
- [PG14] Yu Pu and Jens Grossklags. An economic model and simulation results of app adoption decisions on networks with interdependent privacy consequences. In *Decision and Game Theory for Security*, pages 246–265. Springer, 2014.
- [PG15] Yu Pu and Jens Grossklags. Using conjoint analysis to investigate the value of interdependent privacy in social app adoption scenarios. In *Proceedings of the International Conference on Information Systems, ICIS 2015*, 2015.

- [PG16] Yu Pu and Jens Grossklags. Towards a model on the factors influencing social app users' valuation of interdependent privacy. *PoPETs*, 2016(2):61–81, 2016.
- [PHB⁺06] Jamie Pearson, Jiang Hu, Holly P Branigan, Martin J Pickering, and Clifford I Nass. Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1177–1180. ACM, 2006.
- [Pre14] President's Concil of Advisors on Science and Technology. Big data and privacy: A technological perspective. report to the President, Executive Office of the President, May 2014.
- [Pro13] Protiviti. Knowing how – and where – your confidential data is classified and managed. Technical report, Protiviti Inc., 2013.
- [PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [PXY⁺13] Rahul Pandita, Xusheng Xiao, Wei Yang, William Enck, and Tao Xie. Whyper: Towards automating risk assessment of mobile applications. In *Proceedings of the 22Nd USENIX Conference on Security, SEC'13*, pages 527–542, Berkeley, CA, USA, 2013. USENIX Association.
- [QCP⁺12] Daniele Quercia, Diego Las Casas, Joao Paulo Pesce, David Stillwell, Michal Kosinski, Virgilio Almeida, and Jon Crowcroft. Facebook and privacy: The balancing act of personality, gender, and relationship currency. In *International AAAI Conference on Weblogs and Social Media*, 2012.
- [RBC⁺15] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, 30:39, 2015.
- [RC99] Joseph Reagle and Lorrie Faith Cranor. The platform for privacy preferences. *Communications of the ACM*, 42(2):48–55, 1999.
- [RF05] Bryce B Reeve and Peter Fayers. Applying item response theory modeling for evaluating questionnaire item and scale properties. *Assessing quality of life in clinical trials: methods of practice*, 2:55–73, 2005.
- [RHL16] Jinfeng Rao, Hua He, and Jimmy Lin. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, pages 1913–1916, New York, NY, USA, 2016. ACM.

- [RLSS14] Rohan Ramanath, Fei Liu, Norman M. Sadeh, and Noah A. Smith. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 605–610, 2014.
- [Rob04] S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60:503–520, 2004.
- [RSCM14] Ali Sharif Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *arXiv preprint arXiv:1412.6574*, 2014.
- [SBDC15] Florian Schaub, Rebecca Balebako, Adam L. Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa, July 2015. USENIX Association.
- [Sch68] Barry Schwartz. The social psychology of privacy. *American Journal of Sociology*, 73(6):741–752, 1968.
- [SD13] Emily Steel and April Dembosky. Health apps run into privacy snags. <http://www.ft.com/cms/s/0/b709cf4a-12dd-11e3-a05e-00144feabdc0.html>, 2013. Accessed: 2015-08-16.
- [Sha15] David Shariatmadari. Privacy is starting to seem like a very 20th-century anomaly. <https://www.theguardian.com/commentisfree/2015/nov/07/privacy-seems-20th-century-aberration-but-worth-mourning>, 2015. Accessed: 03-11-2017.
- [Sim72] Herbert A Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.
- [SKL⁺14] Florian Schaub, Bastian Könings, Peter Lang, Björn Wiedersheim, Christian Winkler, and Michael Weber. PriCal: context-adaptive privacy in ambient calendar displays. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 499–510. ACM, 2014.
- [SKV15] Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. Portrait of a privacy invasion. *Proceedings on Privacy Enhancing Technologies*, 2015(1):41–60, 2015.
- [Sky15] Skyhigh Networks. Cloud adoption and risk report. 2015. http://info.skyhighnetworks.com/rs/274-AUP-214/images/WP_Skyhigh_Cloud_Adoption_Risk_Report_Q4_2015.pdf.

- [Sky16] Skyhigh Networks. Cloud report | Skyhigh Networks, 2016. <https://www.skyhighnetworks.com/cloud-report/>.
- [Sol11] Daniel J Solove. *Nothing to hide: The false tradeoff between privacy and security*. Yale University Press, 2011.
- [SR09] John W Stamey and Ryan A Rossi. Automatically identifying relations in privacy policies. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 233–238. ACM, 2009.
- [SS08] Daniel A Schult and P Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008.
- [SSS⁺15] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion*, pages 243–246, New York, NY, USA, 2015. ACM.
- [SSWS16] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. Automatic extraction of opt-out choices from privacy policies. In *2016 AAAI Fall Symposium Series*, 2016.
- [Sun12] Cass R Sunstein. The Storrs lectures: Behavioral economics and paternalism. *Yale LJ*, 122:1826, 2012.
- [SVI⁺16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [SW11] Evan Selinger and Kyle Whyte. Is there a right way to nudge? the practice and ethics of choice architecture. *Sociology Compass*, 5(10):923–935, 2011.
- [Swe02] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
- [TdSXZ16] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016.
- [The15] The Nielsen Company. So many apps, so much more time for entertainment. <http://www.nielsen.com/us/en/insights/news/2015/so-many-apps-so-much-more-time-for-entertainment.html>, 2015.

- [TKV10] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.
- [TMA⁺04] Roberto Torres, Sean M McNee, Mara Abel, Joseph A Konstan, and John Riedl. Enhancing digital libraries with TechLens. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, pages 228–236. IEEE, 2004.
- [TNT⁺14] Joshua Tan, Khanh Nguyen, Michael Theodorides, Heidi Negrón-Arroyo, Christopher Thompson, Serge Egelman, and David Wagner. The effect of developer-specified explanations for permission requests on smartphone user behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 91–100. ACM, 2014.
- [TS08] Richard Thaler and Cass Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press, 2008.
- [TYRW14] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.
- [VB03] Ellen M Voorhees and L Buckland. Overview of the TREC 2003 question answering track. In *TREC*, volume 2003, pages 54–68, 2003.
- [VDJ10] Marten Van Dijk and Ari Juels. On the impossibility of cryptography alone for privacy-preserving cloud computing. In *Proceedings of the 5th USENIX conference on Hot topics in security*, pages 1—8, 2010.
- [WBT⁺17] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *Proceedings of the 38th IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.
- [WLS⁺13] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. Privacy nudges for social media: An exploratory Facebook study. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW ’13 Companion*, pages 763–770, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [WRKS94] Michael S. Wogalter, Bernadette M. Racicot, Michael J. Kalsher, and S. Noel Simpson. Personalization of warning signs: The role of perceived relevance on behavioral compliance. *International Journal of Industrial Ergonomics*, 14(3):233 – 242, 1994.

- [WSD⁺16] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard H. Hovy, Joel R. Reidenberg, and Norman M. Sadeh. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.
- [WSR⁺16] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, pages 133–143, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
- [YNS⁺15] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*, pages 3995–4001, 2015.
- [YTE17] Lin Yuan, Joël Régis Theytaz, and Touradj Ebrahimi. Context-dependent privacy-aware photo sharing based on machine learning. In *Proc. of 32nd International Conference on ICT Systems Security and Privacy Protection (IFIP SEC 2017)*, 2017.
- [ZB14] Sebastian Zimmeck and Steven M Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *USENIX Security*, volume 14, 2014.
- [ZWS11] Chunhui Zhu, Fang Wen, and Jian Sun. A rank-order distance based clustering algorithm for face tagging. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '11*, pages 481–488, Washington, DC, USA, 2011. IEEE Computer Society.
- [ZWZ⁺17] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M Bellovin, and Joel Reidenberg. Automated analysis of privacy requirements for mobile apps. In *24th Annual Network and Distributed System Security Symposium, NDSS 2017*, 2017.

Hamza Harkous

hamzaharkous.com

✉ hamza.harkous@gmail.com
📄 medium.com/@hamzaharkous
linkedin.com/in/hamzaharkous
github.com/harkous

Residence: Switzerland
Nationality: Lebanese



Profile

I am a Postdoc at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, working at the intersection of **Privacy, Machine Learning, and Human Computer Interaction**.

I have been researching and developing AI-driven systems, focused on the Privacy and Security domains. My portfolio includes [PrivySeal](#), an automated privacy assistant for installing cloud apps, [PriBot](#), a chatbot for answering privacy policy questions with deep learning, and [Modemos](#), a machine-learning stack for assisting online children safety.

My research has been published in top venues and has been honored by several awards from the community. I am also a full-stack developer and an avid user/system experimenter who loves conceiving beautiful, usable products while leaving the noise out.

Education

2012 – 2017 **Ph.D.**, *Computer, Communication and Information sciences*.

Thesis: “Data-Driven, Personalized, Usable Privacy”, Distributed Information Systems Laboratory, EPFL

2010 – 2012 **MSc**, *Communication Systems*.

Specialization in Networking and Mobility, EPFL

2006 – 2010 **BE**, *Computer and Communications Engineering*, (Minor in Mathematics).

American University of Beirut (AUB) - Lebanon

Grants/Awards

Jun. 2017 Best Presentation Award at [SwissText 2017](#)

Jun. 2017 1st Runner-up Award for Best Demo Award at EPFL’s [IC Research Day](#)

Mar. 2017 Outstanding Paper Award at [ACM CODASPY 2017](#)

Jun. 2016 Distinguished Poster Award at [USENIX SOUPS 2016](#)

Jan. 2016 Lead a \$370,000 [CTI](#) research grant for an online children safety platform

Apr. 2014 Venture Kick II Award with Graspeo startup team (CHF 20,000)

Jun. 2010 Dean’s Award for Creative Achievement, AUB

Sept. 2006 4-year Merit Scholarship Award from AUB (top 10 among 8000 applicants).

Experience

June 2017 – **Postdoctoral Researcher**.

present Distributed Information Systems Laboratory, EPFL

July 2015- **Scientific Collaborator**, [Privately Sarl](#), Lausanne, Switzerland.

- Present ◦ building an AI-driven platform for data collection and labeling
- building the frontend for incident detection on social media, within the [Oyoty](#) project

July 2015- **Consultant**, [Koemei SA](#), Martigny, Switzerland.

- Jan. 2016 ◦ assisted in the product and analytics strategy for Koemei’s video analytics technology

Mar. 2012 – **Intern**, [Nokia Research Center](#), Espoo, Finland.

- Sept. 2012 ◦ designed and developed Nokia’s [patented](#) authentication mechanism for anonymous, local peer-to-peer communities

2012 – 2017 **Teaching Assistant**, EPFL.

- supervised 14 students' semester projects, 4 internships, and 3 Master's theses; assisted in 5 teaching courses

Sep. 2010 – **Research Scholar**, *Distributed Programming Lab*, EPFL.

- Feb. 2012 ◦ developed a new distributed system for private polling using Secure Multi-party Computations

Projects

PriBot (pribot.org): I created PriBot, the first question-answering chatbot for privacy policies. It takes a previously unseen privacy policy and uses it to answer user questions that are posed in free form. It further simplifies the policy with high-level summaries generated from the legalese text. In our experiments with 1200 users, 91% of the questions had an answer among the top-3 given by PriBot.

PrivySeal (privyseal.epfl.ch): I created PrivySeal, a web app that tells users what apps can needlessly know about them from their cloud data. Through machine learning and visualization techniques, our “**Far-reaching Insights**” scheme was twice as effective in **detering users from installing misbehaving apps** as the current model. Over 1750 users have signed up to PrivySeal so far.

Modemos (modemos.epfl.ch): I built the Modemos web platform, which uses deep learning in order to detect hateful/obscene posts in addition to inappropriate images on social media.

AI-Driven Data Analysis Platform: We are developing a new platform that aids researchers in performing **large-scale image and text analysis**. Towards that, we are building novel visualization techniques, powered with machine learning, in order to speed up data annotation and to improve classifier's accuracy.

PrivyShare-Web: A **browser-based application** that analyzes the content and metadata of files that users desire to upload to Google Drive. It then enables them to set rules for **encrypting uploaded files** and shows them what apps still work with their protected data.

PrivyShare-Desktop: A cross-platform **desktop application** that analyzes files' contents and metadata and **assesses their sensitivity** before letting you securely upload them to any cloud provider.

Selected Publications

- Selected Papers
- **Harkous, H.**, Fawaz, K., Lebre, R., Schaub, F., Shin, K.G., Aberer, K.: “**PriBot: Answering Free-form Questions about Privacy Policies with Deep Learning**”. *Under Submission*, 2017
 - **Harkous, H.**, Aberer, K.: “**If You Can't Beat Them, Join Them: A Usability Approach to Interdependent Privacy in Cloud Apps**”. In *7th ACM Conference on Data and Applications Security and Privacy (CODASPY'17)*. (**Outstanding Paper Award**)
 - **Harkous, H.**, Rahman, R., Karlas, B., Aberer, K.: “**The Curious Case of the PDF Converter that Likes Mozart: Dissecting and Mitigating the Privacy Risk of Personal Cloud Apps**”. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2016.
 - **Harkous, H.**, Fawaz, K., Shin, K.G., Aberer, K.: “**PriBots: Conversational Privacy with Chatbots**” In *Workshop on the Future of Privacy Notices and Indicators*, at *SOUPS 2016*
 - **Harkous, H.**, Rahman, B., Aberer, K.: “**Data-Driven Privacy Indicators**”. In *Workshop On Privacy Indicators*, at *SOUPS 2016*
 - **Harkous, H.**, Rahman, R., and Aberer, K.: “**C3P: Context-Aware Crowdsourced Cloud Privacy**.” In *14th Privacy Enhancing Technologies Symposium (PETS)*, 2014.
 - Gambs, S., Guerraoui, R., **Harkous, H.**, Huc, F., and Kermarrec, A.-M., “**Scalable and Secure Polling in Dynamic Distributed Networks**.” In *31st International Symposium on Reliable Distributed Systems (SRDS)*, 2012.

- Patents
- **Harkous, H**, Leppänen, K. J., Turunen, M. T., Ginzboorg, P., and Niemi, P. V., “[Methods and Apparatus for Data Security in Mobile Ad Hoc Networks](#).” (2014). U.S. Patent No. 20,140,122,882. Washington, DC: U.S. Patent and Trademark Office.
 - Helg, F., **Harkous, H**, “Method to Detect Incidents from Social Network Use” (2016) (*provisional*)

Technical Experience

Proficient With

Languages Python, JavaScript, HTML
Technologies Node.js, AngularJS, MongoDB, Keras Deep Learning Framework, Docker, Microservices, Redis

Have Experience With

Languages Java, C, C++, Matlab, R, Bash Scripting, SQL, CSS
Technologies Nginx, Adobe Illustrator, Spring Framework

