

University of Vermont

UVM ScholarWorks

Graduate College Dissertations and Theses

Dissertations and Theses

2023

Group-Level Frameworks for Data Ethics, Privacy, Safety and Security in Digital Environments

Juniper Lovato
University of Vermont

Follow this and additional works at: <https://scholarworks.uvm.edu/graddis>



Part of the [Computer Sciences Commons](#), [Philosophy Commons](#), and the [Social and Behavioral Sciences Commons](#)

Recommended Citation

Lovato, Juniper, "Group-Level Frameworks for Data Ethics, Privacy, Safety and Security in Digital Environments" (2023). *Graduate College Dissertations and Theses*. 1780.
<https://scholarworks.uvm.edu/graddis/1780>

This Dissertation is brought to you for free and open access by the Dissertations and Theses at UVM ScholarWorks. It has been accepted for inclusion in Graduate College Dissertations and Theses by an authorized administrator of UVM ScholarWorks. For more information, please contact schwrrks@uvm.edu.

GROUP-LEVEL FRAMEWORKS FOR DATA ETHICS, PRIVACY,
SAFETY AND SECURITY IN DIGITAL ENVIRONMENTS

A Dissertation Presented

by

Juniper L. Lovato

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Complex Systems and Data Science

October, 2023

Defense Date: September 7, 2023
Dissertation Examination Committee:

Randall Harp, Ph.D., Advisor
Peter S. Dodds, Ph.D., co-Advisor
Sarah Nowak, Ph.D., Chairperson
Chris Danforth, Ph.D.
Jeremiah Onaolapo, Ph.D.
Holger Hooch, DPhil, Dean of the Graduate College

© Copyright by Juniper L. Lovato
October 2023

ABSTRACT

In today’s digital age, the widespread collection, utilization, and sharing of personal data are challenging our conventional beliefs about privacy and information security. This thesis will explore the boundaries of conventional privacy and security frameworks and investigate new methods to handle online privacy by integrating groups. Additionally, we will examine approaches to monitoring the types of information gathered on individuals to tackle transparency concerns in the data broker and data processor sector. We aim to challenge traditional notions of privacy and security to encourage innovative strategies for safeguarding them in our interconnected, dispersed digital environment.

This thesis uses a multi-disciplinary approach to complex systems, drawing from various fields such as data ethics, legal theory, and philosophy. Our methods include complex systems modeling, network analysis, data science, and statistics.

As a first step, we investigate the limits of individual consent frameworks in online social media platforms. We develop new security settings, called *distributed consent*, that can be used in an online social network or coordinated across online platforms. We then model the levels of observability of individuals on the platform(s) to measure the effectiveness of the new security settings against surveillance from third parties. Distributed consent can help to protect individuals online from surveillance, but it requires a high coordination cost on the part of the individual. Users must also decide whether to protect their privacy from third parties and network neighbors by disclosing security settings or taking on the burden of coordinating security on single and multiple platforms. However, the coordination burden may be more appropriate for systems-level regulation.

We then explore how groups of individuals can work together to protect themselves from the harms of misinformation on online social networks. Social media users are not equally susceptible to all types of misinformation. Further, diverse groups of social media communities can help protect one another from misinformation by correcting each other’s blind spots. We highlight the importance of group diversity in network dynamics and explore how natural diversity within groups can provide protection rather than relying on new technologies such as distributed consent settings.

Finally, we investigate methods to interrogate what types of personal data are collected by third parties and measure the risks and harms associated with aggregating personal data. We introduce methods that provide transparency into how modern data collection practices pose risks to data subjects online.

We hope that the collection of these results provides a humble step toward revealing gaps in privacy and security frameworks and promoting new solutions for the digital age.

CITATIONS

Material from this dissertation has been published in the following form:

Lovato, J., Allard, A., Harp, R., Onaolapo, J., Hébert-Dufresne, L.. (2022). Limits of individual consent and models of distributed consent in online social networks. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2251-2262. DOI: 10.1145/3531146.3534640.

Lovato, J., Mueller, P., Suchdev, P., S. Dodds, P.. (2023). More Data Types More Problems: A Temporal Analysis of Complexity, Stability, and Sensitivity in Privacy Policies. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1088-1100, DOI: 10.1145/3593013.3594065

Material from this dissertation has been submitted for publication to npj Complexity on June 21, 2023, in the following form:

Lovato, J., Hébert-Dufresne, L., St-Onge, J., Harp, R., Salazar Lopez, G., Rogers, S., Ul Haq, I., and Onaolapo, J.. Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks. npj Complexity.

A Triptych of Dedications:

1. To my father Andrew Leo Lovato
for paving the way one adobe a day
2. To my muse and mentor Ginger R. Richardson
for being the best friend I will ever have
3. To my husband, my greatest supporter and partner in crime

“Consider the subtleness of the sea; how its most dreaded creatures glide under water, unapparent for the most part, and treacherously hidden beneath the loveliest tints of azure. Consider also the devilish brilliance and beauty of many of its most remorseless tribes, as the dainty embellished shape of many species of sharks. Consider, once more, the universal cannibalism of the sea; all whose creatures prey upon each other, carrying on eternal war since the world began. Consider all this; and then turn to the green, gentle, and most docile earth; consider them both, the sea and the land; and do you not find a strange analogy to something in yourself? For as this appalling ocean surrounds the verdant land, so in the soul of man there lies one insular Tahiti, full of peace and joy, but encompassed by all the horrors of the half-known life.” - Herman Melville, Moby Dick

ACKNOWLEDGEMENTS

There is a cosmos to thank. Throughout writing this dissertation, I have been incredibly fortunate to have the strong support of colleagues, family, friends, and pets. I thank you all with my whole heart. First, one of these would be possible without the support and encouragement of my husband. You are mon amour, mon chum. Thank you to my pets, Dante, Virgil, and Fox Mulder, for being the best family we could ever wish for.

I want to thank my advisors, Professor Randall Harp and Professor Peter Dodds, whose very different expertise was invaluable in formulating this multidisciplinary body of work. And my committee members, especially Chris Danforth, for all your support, mentorship, and feedback. Your thoughtful collaboration and kindness pushed me to think critically about my work from multiple perspectives. I would also like to thank my Master's thesis advisor Professor Russell Winslow and committee member Sherry Martin for their great contributions to my scholarship and academic development over my academic career.

I would not have been able to do this work without my co-authors. Thank you for your collaboration, forbearance, and cooperation. Special thanks to Antoine Allard, Amanda Casari, Julia Ferraioli, Philip Mueller, Parisa Suchdev, Jonathan St-Onge, Gabriela Lopez, Sean Rogers, Ijaz Ul Haq, Carter Ward, Avi Chawla. A special thanks to Julia Zimmerman for creating the visual abstracts for this thesis.

To my colleagues at the University of Vermont, Jeremiah Onaolapo, James Bagrow, Nick Cheney, Josh Bongard, Chris Skalka, Linda Schadler, Kirk Dombrowski, Bryn Geffert, Alice Patania, Jean-Gabriel Young, Regina Toolin, Kendall Fortney, and Melissa Parr.

To my former colleagues at the Santa Fe Institute and St. John's College, special thanks to Ginger Richardson, John Miller, Cris Moore, John German, Melanie Mitchell, Paige Prescott, Josh Garland, Carla Shedivy, Gabby Beans, David Krakauer, Eric Rupley, Mirta Galesic, Laura and Tim Taylor, Liz Bradley, Liz Hobson, Patrisia Brunello, JP Gonzales, Somdatta Sinha, Caroline Buckee, Jennifer Dunne, Sander Bais, and Miguel Fuentes. To my former colleagues at the New Mexico Lieutenant Governors' Office, a special thanks to Samantha Johnson (for encouraging me to go to grad school way back when), Eric Vasquez, Josh Rosen, Carmella Casados, and of course, Diane Denish. To my wonderful teachers and mentors Richard Bank, Robert Jessen, Tony Gerlicz, Seth Biderman, and Joe Ray Sandoval.

A hearty love for mi familia and my hometown. Que ¡Viva Santa Fe! Special thanks to my father Andy, my mother Anhara, my big brother Todd, my lovely sister-in-law Mari, my brother Niko, my niece and nephew Alma and Calvin, and all my aunties and uncles, cousins, and of course thank you to grandma and grandpa, Mansi, and my elders. Thank you to the best in-laws in the world, Danielle, Lysiane, Gab, Guy, Violette, and Romane.

A big thanks to my dearest childhood friends who have stuck by my side through the best and the worst, especially thanks to Kelly McReynolds, Colleen Martin, Mike Stupin, Brooke and John Scripps, Anastasia Woldridge, Zac Hogan, Tara Khozein and all of those who we have loved and lost over the years.

Thanks to those who have supported my professional and academic research, and thanks to the Alfred P. Sloan Foundation, MassMutual, Google Open Source, and the National Science Foundation. All of the opinions in this thesis are my own and do not necessarily reflect those of my funders or supporters.

TABLE OF CONTENTS

Dedication	iii
Acknowledgements	iv
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Overview	1
1.2 Road-map of Thesis	7
1.3 Introduction to Privacy	7
1.4 Introduction to Security	21
1.4.1 Privacy and Security in the Digital Age	33
Bibliography	42
2 Group Privacy: Distributed Consent In Online Social Networks	48
2.1 Introduction	50
2.1.1 A Critique on the Adequacy of Individual Consent for Data Transactions	53
2.1.2 A Theory of Consent	57
2.1.3 Consent in Data Transactions	59
2.2 Results	62
2.2.1 A General Overview of the Problems with Individual Consent	62
2.2.2 A Threat Model for Leaky Individual Data in Social Networks	65
2.2.3 A Model of Distributed Consent and Network Observability .	66
2.2.4 A Model of Coordinated Consent Across Platforms	74
2.3 Discussion	77
2.3.1 Implications	77
2.3.2 Conclusion	78
2.4 Research Methods	81
2.4.1 Data.	81
2.4.2 Observability Model.	82
2.4.3 Distributed Consent Model.	82
2.4.4 Consent Passport Model.	83
2.4.5 Data and Code Availability.	83
Bibliography	84

3	Group Online Safety: Diverse Misinformation in Online Social Networks	89
3.1	Introduction	91
3.2	Background	95
3.2.1	Bias Types	95
3.2.2	Misinformation	96
3.2.3	Deepfakes	98
3.2.4	Ethical Considerations	99
3.2.5	Research Questions	102
3.3	Results	104
3.4	Mathematical Model	110
3.5	Discussion	117
3.6	Data Availability	119
3.6.1	Survey Methodology	119
3.6.2	Data	121
3.6.3	Analytical Methods	124
	Bibliography	125
4	Grouped Data: Issues of Consumer Data Aggregation Online	131
4.1	Introduction	133
4.2	Background	139
4.2.1	Research on Data Brokers	139
4.2.2	Research on Risk and Harms of Combining Data Types	140
4.2.3	Analysis of Privacy Policies	141
4.3	Results	142
4.3.1	Descriptive Statistics of Privacy Policy Data	143
4.3.2	PII Data Types Lexicon	144
4.3.3	Word level results: Measures of turbulence	145
4.3.4	Topic level results: Measures of complexity	150
4.3.5	Topic level results: Topic prevalence over time	152
4.3.6	Network level results: Measures of sensitivity	154
4.4	Discussion	156
	Bibliography	157
4.5	Supplementary Materials	160
4.5.1	Summary Statistics: Full corpus, PII data types corpus, network analysis	160
4.5.2	Research Methods: Word level	160
4.5.3	Research Methods: Topic level	162
4.5.4	Research Methods: Network level	164
4.6	Lexicon of Personally Identifiable Information (PII)	165

4.7	Lexicon of Negation Words	167
4.8	Topics	167
4.9	Code availability statement	169
5	Conclusion	171
5.1	Précis of the Thesis	171
5.1.1	Summary of Group Privacy: Distributed Consent	171
5.1.2	Summary of Group Correction: Diverse Misinformation	172
5.1.3	Summary of Grouped Data: Issues in data aggregation on users	173
5.2	Limitations and future work	174
5.3	Discussion	176
	Bibliography	176
	Complete Bibliography	178

LIST OF FIGURES

1.1	Visual Abstract for Chapter 1 by Julia Zimmerman	1
1.2	The Social Optimization Problem by Julia Zimmerman	31
2.1	Visual Abstract of Chapter 2 by Julia Zimmerman	48
2.2	Cartoon networks that demonstrate the information network, the observed network, and the protected network	51
2.3	A series of figures which show the fraction of the observed individuals in the network as a function of their levels of adoption of distributed consent.	71
2.4	A series of figures which show the fraction of the observed individuals in the network as a function of their levels of adoption of distributed consent.	73
2.5	We show a multilayer network by doubling the original data, mimicking a two-platform ecosystem.	74
3.1	Human vs. the Machine by Julia Zimmerman	89
3.2	Illustration of the problem considered in this work.	92
3.3	Question where survey participants are asked after the debrief of the survey if they think the videos they watched are real or fake.	94
3.4	A confusion matrix showing our participant guesses about the state of the videos vs. the real state of the video.	105
3.5	Bootstrap MCC samples from observed confusion matrices to compare MCC scores of user and video feature pairs.	109
3.6	Spread of diverse deepfake on configurations of a degree-heterogeneous mixed membership stochastic block model with equal group size and densities (in-group density is set to Q and across the group to $1 - Q$). Other parameters are given in the plots, with panels (b) and (c) using the correction rate highlighted in (a) around a value of 1.7.	115
3.7	Preview of questions from our survey.	121
3.8	Example video clip from the Facebook Deepfake Detection Challenge (DFDC) dataset. The person depicted is fake.	122

4.1	Data Types Monster by Julia Zimmerman	131
4.2	1997 co-occurrence network of PII relevant terms.	134
4.3	PII Data Types by Julia Zimmerman	137
4.4	Timeline of selected privacy legislation in the U.S. and E.U. from 1998-2022	139
4.5	a.) Frequency Distribution of words that rise more than ten times in a 7-year period, b.) Frequency Distribution of words that fall more than 15% in a year	146
4.6	a.) Frequency distribution of stable words (change less than 2% in frequency over a 20-year period), b.) Frequency distribution of words that emerge (occur 20 times after not occurring the previous year) . .	147
4.7	HIPAA Timeline and History	148
4.8	a.) Frequency distribution of words related to health, b.) Frequency distribution of words related to insights	149
4.9	This figure represents privacy policies from 1997-2019 and their complexity measured by year via a compression ratio.	151
4.10	Topic prevalence 1997-2019. Topics extracted from the full corpus after negation filtering through hSBM topic modeling.	153
4.11	a.) 2000-2019 co-occurrence network of PII terms network density by year. b.) 2000-2019 co-occurrence network of PII terms network modularity by year.	155
4.12	a.) The linear plot is a Zipf-ranked degree distribution of the 2019 co-occurrence network of PII terms. b.) Zipf-ranked strength distribution of the 2019 co-occurrence network of PII terms, logarithmic y axis . .	156

LIST OF TABLES

3.1	Accuracy scores of machine deepfake detectors versus primed human deepfake detectors versus non-primed human deepfake detectors. . . .	106
3.2	Matthew’s Correlation Coefficient (MCC) is a correlation measure between a participant’s guess about the video being real or fake (0,1) versus the actual state of the video (real 0, fake 1).	110
3.3	Descriptive statistics of video data (N=5,000).	122
3.4	Descriptive demographics of survey participants (N=2,016).	123
4.1	Overview of privacy policies corpus: the full corpus that is cleaned. Minimum description length (MDL) from the hSBM topic model, text description length (TDL) = total number of words x (\log_2 (unique words)), compression factor = MDL/TDL. UW = unique words, CF = Compression Factor, PP = number of privacy policies in the corpus.	161
4.2	Overview of privacy policies PII data types corpus: includes a corpus filtered for just PII data type terms.	162
4.3	Network analysis metrics on privacy policy text from 1997-2019. . .	170

CHAPTER 1

INTRODUCTION

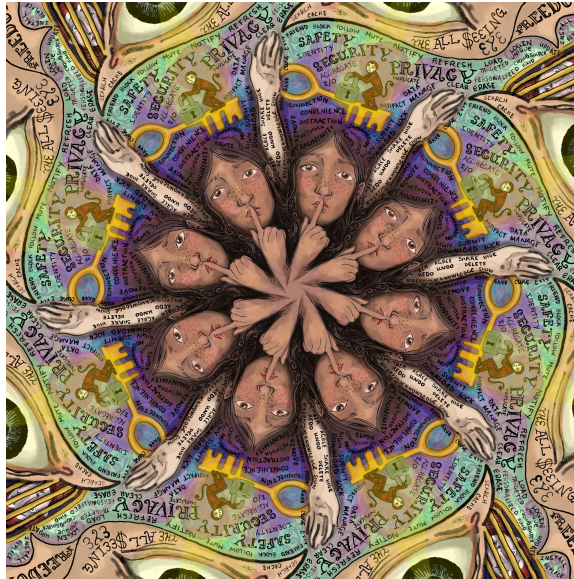


Figure 1.1: Visual Abstract for Chapter 1 by Julia Zimmerman

1.1 OVERVIEW

Data has the potential for great benefit and harm. On the one hand, data harnesses immense potential for improving systems' performance and increasing knowledge and transparency, especially as it concerns the structure and dynamics of information sys-

tems. On the other hand, data has also posed significant potential and actual harm to data subjects and many new unforeseen ethical, privacy, and security challenges for human and automated information systems. Over time, the precipitous rise in data use in our information systems urgently calls for an overarching and fair ethical framework for addressing the moral, privacy, safety, and security problems and associated risks arising from the increased data use in the digital age. Ethical frameworks for data use will need to be developed for those using data and deploying its insights. Specific ethical concerns may differ for each context. Still, the core concerns of protecting data subjects (data subjects here include individuals, groups, and systems together) from harm must remain the central objective. In addition to ethical frameworks, there needs to be more transparency in data collection, use, and dissemination by data processors to hold them accountable and check if they are upholding the ethical standards set forth by data privacy laws and ethical frameworks.

Privacy and security are concepts that have been given much scholarship. However, the conceptualization of privacy and security has primarily originated from a framework grounded in notions of physical safety in the home, liberty of the individual, and bodily autonomy of the individual [93, 88, 87]. There are several issues with this framework being used in the digital age. Namely, traditional frameworks must effectively address the practical issues concerning data’s distributed, networked, and non-physical nature. There are also issues due to a lack of data processor transparency, which makes it difficult for data subjects to understand what information is being collected, used, or shared and adequately assess risk. Part of the vulnerabilities and privacy gaps are due to information systems’ distributed, networked, and non-physical nature. However, digital technological innovation progresses much

faster than previous physical technological innovations, so we must adjust our expectations for the rate of technological change. Moreover, technological innovation changes much faster than legal institutions' ability to regulate and create new laws. New frameworks for privacy and security must better suit privacy and security in the digital age and build mechanisms for more adaptive and contextual regulatory frameworks.

Data is non-physical: It is commonly assumed that the exact boundaries that apply to physical spaces also extend to information systems regarding traditional privacy and security frameworks. However, the distributed, networked, and non-physical nature of information systems in the digital age questions this assumption. Traditional privacy frameworks are grounded in protecting the home and the body. The home and the body are seen as physical boundaries where it is acceptable to expect privacy. However, in a digital age the body can be monitored or tracked without physical touch, and the activities of private life at home are now easily accessed via ambient surveillance technology and online data generated in the home but quickly disperses to far corners of the digital world [68, 104, 105]. Data subjects may go about their everyday activities assuming privacy in these traditional spaces where they no longer retain privacy. Part of the issue is a byproduct of the rapid progress of technology, which innovates faster than a data subject's ability to build heuristics to assess risk associated with new technology [88]. There is a conceptual mismatch between privacy and security of information systems qua physical privacy and security. This mismatch leads to significant vulnerabilities and potential harm to data subjects due to the gap between physical and informational boundaries and dynamics.

Data is distributed and networked: Data privacy often assumes that individual data subjects have complete control over their data and the ability to trade privacy for online services. However, this assumption overlooks that personal data can contain sensitive information about groups or network neighbors and can be accessed by a wider audience than the immediate data processor. It also overlooks the ability of data processors to infer and predict private information about data subjects using generalizations through their network neighbors and network attributes like homophily.

Lack of Transparency: The data broker industry is a multi-billion-dollar market that involves data brokers and processors who collect, buy, and sell consumer information. Unfortunately, this industry lacks transparency, making understanding the data types being collected, utilized, and sold challenging. Consequently, it is challenging to determine the risk posed to data subjects. **In Chapter 4, we will look at methods for investigating privacy policies as a means to find what personally identifiable data types data brokers say they collect, use, and sell. We will also introduce some methods to measure how these collection practices change over time, how complex the privacy policies are over time, and the sensitivity of the data types collected concurrently.**

The theoretical framing of this thesis comes from data ethics, which studies the moral problems related to data. *Data ethics* is an emerging multi-disciplinary research field that explores the positive and negative impacts of collecting, using, processing, storing, possessing, disseminating, disposing of, aggregating, and creating models and byproducts from data [63]. Data can be considered a collection of values (datum) that, when aggregated, transmit information related to a data subject. Data ethics frame-

works help develop best practices to maximize the utility of data while minimizing harm to data subjects (e.g., individuals, groups, and society featured in the data). These frameworks are often utilized by lawmakers interested in data protection and regulation, industries (e.g., for-profit companies, non-profit organizations, data scientists, and researchers who use data) setting best practices and industry standards, and data subjects interested in their potential or realized benefit or harm.

Much of the research on privacy, safety, and security focuses on risk or harm to the individual or the system. However, many interactions in data ethics, privacy, and security occur in a socially networked digital environment where information is distributed and not delimited by individuals. We will take a multi-scale approach but focus primarily on the group-level structure and dynamics of privacy and security in the digital age. We will also provide methods for greater transparency and some strategies for mitigating group-level harms grounded in data ethics.

Data ethics, privacy, and security are complex adaptive systems with multi-level interdependencies. Moreover, data is distributed and networked across many levels of society. The interactions across and between individuals, groups, and systems must be considered to understand the large-scale dynamics and impact of data flow, fully understand the potential harms and benefits to society, and build meaningful and effective mechanisms for protecting data subjects.

This thesis is written from the perspective of complex systems and uses methods and conceptual framing from many fields in a transdisciplinary tradition when possible.

Conceptual Framing: Data ethics and Fairness, Accountability, and Transparency (FAccT) is an emerging research field that has begun formulating frameworks

around important technical and social issues related to adequately handling data and new technological innovations such as artificial intelligence. The next decade will be crucial as these frameworks begin to take hold and formal definitions of critical concepts like fairness, accountability, and transparency are solidified into laws and norms. Necessary groundwork, however, has already begun with seminal works in data ethics and fairness by scholars such as D’Ignazio et al.’s conceptualization of data feminism. They wrote an essential work using a theoretical foundation from feminist theory to ground data ethics [28]. This thesis will primarily utilize foundational theory from ethical philosophy [54, 101, 31, 35] and data ethics [68, 88, 28, 93, 38, 96], to ground the practical concerns and case studies outlined in this thesis. Many other fields, such as legal theory, sociology, political theory, economics, and computational social science, will also inform this thesis.

Methodological Framing: Our analytical methods come primarily from the theoretical foundation of complex systems, computer science, data science, and multi-scale thinking. It will use methods from complex systems [92, 66, 62], data science and statistics, [51] computational social science [61], networks [67], and modeling complex systems [80]. Empirical methodologies primarily come from practices in survey design from social research methods [85].

This thesis explores group-level privacy and security in digital information. Groups must protect themselves and make informed decisions about their collective data. We need to examine the dynamics of these systems at a group level to understand how data ethics, privacy, and security interact. This work overviews group data ethics, privacy, and security through three critical case studies.

1.2 ROAD-MAP OF THESIS

This thesis can be outlined as follows: the remainder of **Chapter 1** introduces privacy and security, explores ethical frameworks for data privacy and security in the digital age, and outlines gaps in the field. **Chapter 2** explores group privacy and notions of distributed consent in online social networks and how individuals can coordinate consent. **Chapter 3** focuses on group correction and diverse misinformation, how biases impact our ability to detect diverse misinformation, and how diverse groups can help one another detect misinformation through group correction. **Chapter 4** will look at transparency issues in the data broker and data processor industry and take a multi-scale approach to assess the turbulence, sensitivity, and complexity of data collected by brokers and processors by textually analyzing their privacy policies. **Chapter 5** will outline future work and propose possible solutions for re-conceptualizing data ethics, privacy, and security.

1.3 INTRODUCTION TO PRIVACY

Privacy is an elusive concept long debated in legal and ethical scholarly works. Nevertheless, privacy is a necessary right, and functionally defining privacy is essential to hold violations against privacy accountable. Indeed, privacy is a concept that is so socially integral that it is thought to be necessary for protecting personhood, autonomy, freedom of speech, flourishing, safety, and security, and the development of democracy [87, 29]. Privacy is an essential right in that it is essential to enjoying other rights. Privacy ensures that data subjects are not harmed by a physical vi-

olation or undue information flow, protects autonomy and freedom, and also helps to protect subjects from inequity (e.g., informational power imbalance). However, we have a hard time talking about privacy without the boundaries of the concept becoming functionally restricted, partly due to our definitions being too narrow or broad. The idea of privacy can be difficult to define, and it is also to borrow a term by Rittel and Webber, a wicked problem. Wicked problems are often defined [77] as ones that are complex, have many interdependencies, and are dynamic, making them especially difficult to solve. In this thesis, we will primarily be concerned with information privacy.

In the context of US law, issues relating to privacy date back to some of the earliest philosophical and religious writings with differentiation between public and private boundaries [64, 44]. In the United States (US), privacy norms formed almost immediately when the first colonies were established. During this time, the home constituted the primary location for privacy [64, 73].

As mentioned above, the concept of privacy is greatly debated and has changed over time in the context of new technology and norms. However, some clear themes emerge from privacy literature and traditional privacy literature. Privacy is a human need to protect and control the house, body, family life, and personal information to maintain freedom, personhood, control, and self-determination. However, this need is not absolute. If the individual's need conflicts with societal interests, then the individual's need may be forgone. This tradeoff may be the case in an example of national security. When we enter civil society, there is naturally a tradeoff between individual privacy or autonomy in exchange for the security provided by the societal structure [58].

There are several key definitions of privacy in the traditional privacy literature. It was in 1891 when Warren and Brandeis defined privacy as the right to be let alone. In Alan Westin’s 1968 work *Privacy as Freedom*, he defines privacy as follows:

the claim of individuals, groups, or institutions to determine for themselves when, how, and what extent information about them is communicated to others. Viewed in terms of the relation to the individual to social participation, privacy is the voluntary and temporary withdrawal of a person from the general society through physical or psychological means, either in a state of solitude or small-group intimacy or, when among larger groups, in a condition of anonymity of reserve [100].

Two dimensions of privacy can be observed: physical privacy, such as controlling who can enter your home or touch your body, and informational privacy. In Holvast’s 2007 work, *History of Privacy* [46], they define privacy as “the individual’s right to self-determination, within certain borders, to his home, body, and information” [46]. Holvast states privacy has four main functions: personal autonomy as a form of emotional release, space for self-evaluation and decision-making, and the need to limit and protect communication. These needs and functions are vital parts of maintaining stable personhood.

Daniel Solove states that the traditional judicial and scholarly work on privacy summarizes into six main ideas: “(1) the right to be left alone; (2) limited access to the self; (3) secrecy; (4) control of personal information; (5) personhood; and (6) intimacy” [87].

The right to be left alone was first described by Warren and Brandeis (as the right to be let alone) in 1891 in their seminal work *The Right to Privacy* [99].

Here, they state that every individual should be able to consent to whether their information is made public. Warren and Brandies argue that personal information is not a public product but a domestic one. One of the main arguments in their treatise is that privacy should not be solely thought of as the protection of body and property but that it should fall under a more general right, which is the right of the individual to be let alone (which includes more emotional dimensions of individual privacy). They claim that the right to privacy protection in common law was outdated and no longer met the demands of society. In the common law, privacy is only protected from “physical interference with life and property” [99]. In common law at the time, the right to life talks about protection from assault, and the right to protection of property relates to an individual’s property (tangible or intellectual) and livestock.

In this work, they advocate expanding the conceptualization of privacy also to protect “thoughts, sentiments, and emotions, expressed through the medium of writing or of the arts, so far as it consists in preventing publication” [99] as an extension of the right of the individual to be let alone (the right to an inviolate personality). For example, private letters sent in the mail are considered in this framework to be domestic products, even if they are in transit outside the home. The interception of such a product could harm the individual through pain and suffering and the possible hindrance of future flourishing.

They equate this right to not being assaulted, imprisoned, maliciously prosecuted, or defamed. The Warren and Brandeis framework also extended the privacy boundary to include protection from invasion via photography, invasive press, invasive business practices, recordings of sound or scenes, and other modern recording devices. This 1890s treatise was foundational in conceptualizing privacy in legal frameworks for the

right to privacy. It also was one of the first conceptualizations of privacy that took the use of technology into account.

Limited access to the self is the right to protect oneself from unwanted access by others and the right to being apart from others. David O'Brien states that limited access to self is mostly formulated as a choice or the ability to control who has access to your personal life [70]. One criticism of this perspective is that access to oneself is not always an active choice. Some privacy is accidental, involuntary, or compulsory [87].

Limited access should not be confounded with solitude; it includes solitude (being apart from individuals) but also extends the right to being apart from governmental agencies and press interference. The concept of limited access was popularized by E.L. Godkin [40, 41] in the 1880s. Godkin stated that “nothing is better worthy of legal protection than private life, or, in other words, the right of every man to keep his affairs to himself, and to decide for himself to what extent they shall be the subject of public observation and discussion” [40]. Godkin also argued that individuals could decide how much of their private life (thoughts, feelings, affairs, doings) would be made public [41]. This conceptualization had been expanded a bit through time by scholars like Sissela Bok to broaden limited access as protection against unwanted physical access, access to personal information, or attention (such as an unwanted gaze) [15].

Secrecy is a conceptualization of privacy where privacy is violated if there is public disclosure of something intended to be concealed.

In a legal context, The conceptualization of privacy as concealing had a significant impact on the US constitutional right to privacy, which emerged out of US Supreme Court cases related to contraceptive rights like *Griswold v. Connecticut* and *Roe v.*

Wade (where the courts decision states it was to protect a zone of privacy and protect an individual’s intent to conceal the disclosure of personal matters).

Control of personal information is, as Alan Westin describes, a “claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others” [100]. In other conceptualizations of controlling personal information, this also extends to how the information is collected, used, and shared. In this framework, acceptable disclosure usually occurs through an informed consent process [31]. The control of personal information is similar to the limited access framework but is specific to information. This thesis will primarily concern information privacy but will question the limit to which we have control over the flow of this information. **In Chapter 2, we will discuss the limits of individual consent when information implicates groups.**

In a legal context, this conceptualization is particularly important for protecting bodily autonomy. It has been used in several important cases like *Union Pacific Railway Co. v. Botsford*, where the United States Supreme Court decided that a plaintiff could not be required to undergo a surgical examination. This conceptualization also relies on consent as a tool to protect personhood. [31] The personhood conceptualization of privacy has been pivotal in cases related to contraception like in *Planned Parenthood v. Casey* where the US Supreme Court stated, “These matters, involving the most intimate and personal choices a person may make in a lifetime, choices central to personal dignity and autonomy, are central to the liberty protected by the Fourteenth Amendment. At the heart of liberty is the right to define one’s own concept of existence, of meaning, of the universe, and of the mystery of human

life. Beliefs about these matters could not define the attributes of personhood were they formed under compulsion of the state” [10].

Finally, privacy can also be framed as a type of **intimacy**. This conceptualization posits that privacy is not just an individual right but is essential for forming intimate relationships and group formation. Jeffrey Rosen states, “In order to flourish, the intimate relationships on which true knowledge of another person depends need space as well as time: sanctuaries from the gaze of the crowd in which slow mutual self-disclosure is possible” [79]. In this quote by Rosen, he acknowledges that privacy extends beyond the individual.

According to Solove’s taxonomy of privacy [89], there are also four basic types of harmful activities: information collection, information processing, information dissemination, and invasion.

Information collection harms include surveillance of a data subject’s activities, interrogating a data subject for information, or persistent online behavior tracking. Information processing includes activities like aggregating data, identifying data subjects inside a dataset, using the information for a purpose other than the one stated or agreed to, and lack of proper security and disclosure. Dissemination of information harm involves activities like sharing information with a third party. Dissemination harms can happen through a breach of confidentiality where the information is intentionally shared when there was a promise not to share the information, sharing information intentionally through disclosure. Exposure harm can happen through activities like accidental disclosure and drastically increasing accessibility to information that may be sensitive or could add to the sensitivity of another dataset if combined. Appropriation harm can happen through activities like assuming some-

one’s identity. Distortion harms involve activities like spreading misinformation or disinformation about someone. Intrusion harms involve forcefully invading someone’s private affairs or personal space. **In Chapter 3, we will discuss the issues of privacy concerns related to appropriation and distortion at length using deepfake videos as a case study for this type of privacy violation.** Note that we will primarily be concerned with information privacy in this thesis.

According to Solove, privacy is difficult to define because it does not have a universal value in all contexts and depends on societal norms. However, legal precedent has constrained the concept of privacy in inflexible terms in cases related to emerging privacy issues. This presents a problematic tension because there is a need for a functional definition to use privacy in law, but privacy is dynamic and contextually related to innovation. These dynamics make it challenging to define privacy statically and set a legal precedent representing modern or future realities. In addition, the pace at which the contextual environment evolves about privacy (mainly due to rapid technological advancements) is faster than the pace at which new laws are enacted to define privacy within those parameters. This, in turn, leaves a considerable vulnerability in the gap between these events.

The traditional method of defining concepts is historically categorical and works to define significant slippery concepts by what separates them from other concepts. It uses an Aristotelian method of defining terms through naming genera and differentia [3] to extract which elements of that term are essential and unique. Solove proposes a new conceptualization method to conceptualize privacy, especially in legal contexts. This method, he suggests, should take a pragmatic approach, be bottom-up, and conceptualize privacy using Wittenstein’s notion of “family resemblance” which

is grounded in context [101]. Pragmatism is the notion that understanding and meaning are inseparable from our transactions with the world. Moreover, our ability to learn about the world and make meaning of concepts relies on our experience of the world and the practical consequences that concepts or knowledge lead [74, 52].

For Wittgenstein, linguistic meaning is not an objectively inherent link between the meaning of a word and the cluster of nodes the word references. Instead, a word is pragmatic, and its meaning is grounded in how people use it in practice and the resulting consequences of its use. A word is not tied to an intrinsic singular meaning, but the uses of the word form “a complicated network of similarities overlapping and crisscrossing” [101]. The conceptualization of the term can be understood by the relationships between the similarities in the context of their usage occurrences. This new way of conceptualizing is noteworthy for our concerns because it can help address conceptual gaps in privacy in the context of technological innovation (digital privacy). This pragmatics framework provides terms, such as privacy, with a flexible outer boundary capable of reflecting emerging conceptualizations while remaining dynamic. This model is much more friendly to expanding definitions of privacy, allowing them to adapt to new circumstances and technologies in the digital age. We will discuss methods of integrating this pragmatic framework into privacy practices in Section 1.4.1.

History of Information Privacy Laws and Regulations in the US

The timeline below will review major privacy law events that helped move us to where we are now. This timeline aims to demonstrate how information privacy laws have progressed over time but have moved relatively slowly in their reaction to emerging

technology. One of the first cases that concerned wiretapping as a violation of privacy in the home was *Olmstead v. United States* in 1928. During this case, the US Supreme Court ruled that because nothing physical was taken from the home, there was not a violation of privacy. It took nearly forty years after *Olmstead v. United States* for the Supreme Court to recognize electronic surveillance like wiretapping as a privacy violation with *Katz v. United States* in 1967.

In this timeline, laws protecting sensitive information such as educational, financial, information on minors, and health-related information emerge over time. There have also been significant laws related to consumer data protection in recent years in Europe and the US. However, these regulations all came into effect decades after exploitative data collection practices had occurred. Other emerging technological developments related to technology such as facial recognition [9], deepfakes [57], use of big data [72], artificial intelligence (AI) and generative AI [24], surveillance technology [105], behavioral tracking [12], biometric sensing [104], and automated technology [30] pose immediate harm to data subjects but have yet to prompt large-scale or federal legal protections in the US.

Below is a timeline of some important events in the history of privacy in the US (this, of course, is not an exhaustive list):

- **1789** The US Constitution came into effect in 1789, which includes implicit rights to privacy in the First, Third, Fourth, and Fifth Amendments [46].
- **1790** US citizens are concerned that the US Census is a form of government intrusion which violates privacy. Citizens worry to what extent the information would remain confidential [46].
- **1792** The Post Office Act worked to safeguard the privacy of the mail and was

signed into law by President George Washington [73].

- **1873** US politicians complained that the press was becoming invasive and violating their privacy [46].
- **1890** Samuel Warren and Louis Brandeis wrote an article *The Right to Privacy* in the *Harvard Law Review*, which was a seminal article in the development of privacy rights in the US [46].
- **1905** *Pavesich v New England Life Insurance Company*: is a supreme court case where a man's photo was featured in a life insurance advertisement without his consent, and the life insurance firm was fined for unauthorized use of the man's photograph [65].
- **1914** The Federal Trade Commission Act (FTCA) was established to protect people against deceptive commercial practices. This agency has been a leading force for consumer privacy regulation and enforcement. FTC is an independent federal agency that has enacted federal privacy laws like the Fair Credit Reporting Act. The FTC can hold companies accountable through legal action if they violate consumer privacy rights or mislead or harm consumers through faulty or deceptive acts or practices related to consumer privacy [25].
- **1928** *Olmstead v. United States*: a supreme court case where wiretapping was used to procure evidence. The courts ruled against Olmstead, who was wiretapped, because the search and seizure amendment of the constitution did not apply in this case because nothing physical was taken from Olmstead's home. This was an important case where new technology disrupted our notions of privacy and its boundaries [39].

- **1948** The UN Declaration of Human Rights declares, “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks” [5].
- **1965** *Griswold v. Connecticut*: The Supreme Court struck down Connecticut’s “Comstock law” that banned contraception, establishing the right to privacy in intimate practices [78].
- **1967** *Katz v. United States*: electronic surveillance is used to collect evidence against Katz. The courts ruled that the electronic surveillance used to collect the evidence went against the Fourth Amendment of the US Constitution. The court stated “the individual’s actual, subjective expectation of privacy and the extent to which that expectation was one that society is prepared to recognize as ‘reasonable’ ” [46].
- **1968** Alan Westin published *Privacy and Freedom* and defines privacy as “the claim of individuals ...to determine for themselves when, how and to what extent information about them is communicated” [100].
- **1972** *Eisenstadt v. Baird*: the Supreme Court ruled that unmarried individuals have the right to possess contraception based on the right to privacy and the Fourteenth Amendment [4].
- **1973** The HEW Advisory Committee Report on Automated Personal Data Systems The Department of Health, Education, and Welfare (HEW) Secretary’s Advisory Committee on Automated Personal Data Systems (SACAPDS) introduced the Fair Information Practices, a set of principles still used in privacy

legislation today [71].

- **1974** Family Educational Rights and Privacy ACT is a federal law that safeguards the privacy of student education records [36].
- **1974** The Privacy Act of 1974 is a federal law that was passed on December 31, 1974. It sets forth a Code of Fair Information Practice for federal agencies regarding collecting, maintaining, using, and sharing personally identifiable information [13].
- **1977** The Privacy Protection Study Commission was created in 1974 to assess the effectiveness of the Privacy Act. It presented a final report with recommendations for enhancement and concluded in 1977 [22].
- **1981** The Organization for Economic Co-operation and Development (OECD) has established a crucial framework for safeguarding privacy, which lays out principles for ensuring proper privacy protection. Eight principles were established to limit the flow of personal data by the following means: (1) Collection Limitation Principle: Collection of data should be limited and collected by lawful and fair means with consent; (2) Data Quality Principle: data collected should be relevant to the explicit use of the data and should be kept up to date and be accurate; (3) Purpose Specification Principle: the purpose of the data collection should be made clear at the time of collection and should not be used for another purpose; (4) Use Limitation Principle: data should not be shared unless given consent to share the information or by authority of the law; (5) Security Safeguards Principle: data should be adequately secured from destruction or dissemination; (6) Openness Principle: data controllers should be open to changes in regulation about personal data and prepared to make changes; (7)

Individual Participation Principle: individuals should have the right to access data about them and challenge the data in order to make changes or revoke the access; (8) Accountability Principle: a data controller should be accountable for following the principles above [11].

- **1986** The Telephone Consumer Protection Act (TCPA) and the National Do Not Call Registry regulate telemarketing and automated telephone dialing. The National Do Not Call Registry prohibits certain types of telemarketing while allowing marketing to non-automated telephone numbers [97].
- **1991** The Common Rule, which governs biomedical and behavioral research involving human subjects, has been updated. It is now the standard of ethics for both government and most non-government-funded research in the US, overseen by Institutional Review Boards [59].
- **1995** The European Union established the Data Protection Directive to oversee personal data handling within its jurisdiction properly [14].
- **1996** The Health Insurance Portability and Accountability Act (HIPAA) of 1996 aims to simplify the sharing of healthcare information, safeguard Personally Identifiable Information kept by the healthcare and health insurance sectors against theft and fraud, and address restrictions on health insurance coverage [45].
- **1998** COPPA (Children’s Online Privacy Protection Act) is a federal law in the US regulating the collection of personal information online about children under 13, regardless of their location. The law requires websites to have specific information in their privacy policies, obtain parental consent, and ensure the

safety and privacy of children online [37].

- **1999** Chief Counselor for Privacy and First Chief Privacy Officer created in US Federal Government [20].
- **1999** The Financial Modernization Act of 1999, also known as the Gramm-Leach-Bliley Act (GLBA), is a federal law that mandates financial institutions to disclose their methods of sharing and safeguarding private customer information [17].
- **2002** The E-Government Act of 2002 required federal agencies to conduct a Privacy Impact Assessment for any new technology or data collection involving personally identifiable information [83].
- **2003** California implemented data breach notification laws in 2003, which require businesses and state agencies to disclose any security breaches compromising Californian's personal information. Other US states and jurisdictions have adopted similar legislation [86].
- **2010** The Red Flags Rule was created by the Federal Trade Commission (FTC) and National Credit Union Administration (NCUA) to prevent identity theft [81].
- **2020** The California Consumer Privacy Act CCPA regulates how CA businesses handle residents' personal data. It was passed in 2018 and went into effect on January 1, 2020 [26].

1.4 INTRODUCTION TO SECURITY

It is important here to distinguish the concepts of privacy and security. Privacy and security are often conflated with one another but have meaningful differences from

one another and often hold separate ethical concerns. Derek E. Bambauer [8] argues that separating privacy from security has practical consequences in legal settings. Privacy, for Bambauer, deals with the competing normative decision-making processes and frameworks of access to information. Privacy tasks the normative decision maker to choose between different normative philosophies and decide how rights and entitlements to information ought to be weighed.

Conversely, security involves implementing privacy decisions (the set of mechanisms protecting the information). Security in this framework mediates between the state of the information and the normative privacy decision. This is important because there are different strategies for reacting to issues of privacy and security and different types of penalties and culpability claims. For example, an information privacy violation is disclosing information via a decision or choice when direct disclosure of the information is not permissible (e.g., Cambridge Analytica obtained Facebook data from an academic source [47]). In contrast, a security violation is the disclosure of information inadvertently through incompetent protection of information (e.g., a hacker accessing private records through a non-robust server via a data breach). Bambauer argues that security violations should be punished more harshly than privacy violations because incompetence is worse than malice. Bambauer also argues for higher culpability for security violations over privacy violations because security flaws make all parties worse off (there are two victims: the breached organization via the inadvertent disclosure and being hacked and the data subject through their privacy being violated).

According to Wagner et al. [96], there is another important distinction between privacy and security. In the context of risk assessment, the harm to data subjects is

the primary privacy concern. Security threats, however, only consider harm to the data subject as a secondary concern. The primary concern is to mitigate risk.

The conceptualization of security in practice in the context of cyber security and information security is often guided by the CIA Triad—Confidentiality, Integrity, and Availability concepts and threat models [33, 103]. The CIA Triad builds a framework around security controls focused on confidentiality, integrity, and availability. Confidentiality in this context deals with protecting information from unauthorized access. Integrity concerns the accuracy of the information in that it reflects the truth, is complete, and has not been modified in an unauthorized manner. Finally, accessibility concerns the risks associated with inaccessibility or loss of information and the need to secure the possible loss by taking actions such as backing up information.

Threat models, typically used in cyber security, require identifying the systems security requirements and potential vulnerabilities and, measuring their potential threat level, then making a plan to mitigate these threats [103].

It should be noted that the concept of security also suffers from an ambiguity problem [7, 102, 94, 55] just like privacy and may also benefit from a more pragmatic approach to conceptualization. David Baldwin states that security is an important “concept, which has been used to justify suspending civil liberties, making war, and massively reallocating resources” but is concerned about the lack of attention the conceptualization of security has received in academic literature [7]. To Fischer and Green, “security implies a stable, relatively predictable environment in which an individual or group may pursue its ends without disruption or harm and without fear of such disturbance or injury” [34].

Kleinig et al. have highlighted a tension concerning security that raises the ques-

tion of whose security is at stake [55]. There may even be nested security subjects at play. They state that even though individual security is essential:

The focus [of security] is generally on some form of collective security – such as public safety or, more commonly, national security – and the latter in particular is said to justify a significant number of constraints on our individual liberty interests (including privacy)... In fact, there appear to be at least three related distinctions in play here. Firstly, there is the security of the individual person in relation to other persons and collective entities, including the state. Secondly, there is the (so to speak, internal and external) collective security of a population of such persons... Thirdly, there is the (internal) collective security of a population of collective entities in relation to some of their own number, for example, the community of nation-states. Note that qua community, the community of nation-states, does not have a need for external collective security; there is no external collective entity (for example, invading Martians) [55].

Kleinig et al. reflect on the conceptualization of security as simply “security against X so that we can be secure to Y” [55]. For this thesis, we will adopt this simple conceptualization of security with the understanding that the concept of security is difficult to pin down. Much more work needs to be done to give this broad concept more theoretical backing.

Tension between security and privacy

Richard Ullman points out that the “tradeoff between liberty and security is one of the crucial issues of our era. In virtually every society, individuals and groups seek

security against the state, just as they ask the state to protect them against harm from other states.” Ullman continues to say, “In addition to examining security tradeoffs. It is necessary to recognize that security may be defined not merely as a goal but as a consequence—this means that we may not realize what it is or how important it is until we are threatened with losing it. In some sense, therefore, security is defined and valorized by the threats which challenge it” [94]. Moreover, if the threats are dynamic and change, this may also make security a difficult concept to pin down.

The tension between privacy and security is important for the functioning civil society. John Locke, in his second treatise, states:

If Man in the State of Nature be so free, as has been said; If he be absolute Lord of his own Person and Possessions, equal to the greatest, and subject to no Body, why will he part with his freedom? Why will he give up this Empire, and subject himself to the Dominion and Controul of any other Power? To which 'tis obvious to Answer, that though in the state of nature he hath such a right, yet the enjoyment of it is very uncertain, and constantly exposed to the invasion of others. For all being Kings as much as he, every Man his Equal, and the greater part no strict Observers of Equity and Justice, the enjoyment of the property he has in this state is very unsafe, very insecure. This makes him willing to quit a Condition, which however free, is full of fears and continual dangers: And 'tis not without reason, that he seeks out, and is willing to joyn in society with others who are already united, or have a mind to unite for the mutual Preservation of their Lives, Liberties and Estates, which I call by the general Name, Property [58].

In a liberal society, there is a presumption that people will trade some level of liberty and privacy to enter civil society and be protected by those who have a duty to protect the citizens of that society to ensure human flourishing. This is one of the bases on which modern democracy is founded. There is always a tension between how to weigh rights and duties in a liberal society, the extent to which privacy and security are really at odds, and how to assign the weights is highly debated and an issue we do not have time to discuss in this thesis in detail. Kleinig et al. also reflect on this tension between privacy and security [55]:

In considering the relations between liberty and security – how they are to be “played off” against each other – it is very common, almost standard, to use the metaphor of a scale in which liberty/privacy and (possibly national) security are placed in opposing pans, one to be “balanced” against the other in zero-sum fashion. . . . Securing the right balance is not something that can be determined in the abstract, or once and for all, but something that will change depending on the gravity of a threat and the level of risk (to security) [55].

They conclude that this balancing metaphor may be dangerous and tradeoffs between privacy and security should be considered by careful judgement [55]. The point here is that there is no explicit means to weigh rights and duties against one another. Again, the careful judgment of the appropriate tradeoff will need to consider the context and be made through careful deliberation.

The right to security is also considered to be an essential right. Here we are talking about the concept of security generally as a means to understand how it is conceptualized. In later chapters, we will be primarily focused on information

security. According to Henry Shue, security is a basic right that “is essential to the enjoyment of all other rights” [84]. Examples of these rights include the right to free association, the right to assemble, privacy, and the right to what Shue calls basic physical subsistence (safety, food, basic shelter, and education) [84]. He goes on to say that “a moral right provides the rational basis for a justified demand that the actual enjoyment of a substance be socially guaranteed against standard threats” [84]. To Shue, a right justifies a citizen in demanding that society make practical arrangements to establish and maintain that right. An arrangement usually takes the form of law in addition to norms. He states that the right must only be a standard and reasonable guarantee. However, if, for example “people who walk alone after dark are likely to be assaulted, or if infant mortality is 60 per 1000 live births, we would hardly say that enjoyment of, respectively, security or subsistence had yet been socially guaranteed . . . a right involves a rationally justified demand for social guarantees against standard threats” [84]. This, in turn, creates a duty for those in power to create and protect adequate arrangements to secure against these standard threats.

Identity thefts, data breaches, and unethical data processing practices are becoming increasingly ubiquitous. There is cause for concern that our civil liberties, rights to privacy and security, and digital personhood may be in harm’s way. Knowledge and power inequality between the data subject and the data processor functionally removes individuals’ ability to control their personal information and digital safety and privacy realistically. This inequality often leads to cases where individuals are fighting for justice against large entities with little to no effective arrangements for their safety and security by the state. The power imbalance between individuals and large entities makes it difficult for an individual to have a fair chance at seeking justice

for misuse of their information.

According to Avivah Litan, an analyst from Gartner Inc., only 1 in 700 identity theft cases results in a conviction [60]. These disheartening statistics represent individuals in helpless situations. Cybercrime harms our economy as well. In 2018, the Federal Trade Commission processed 1.4 million fraud reports, which totaled over \$1.4 billion in losses [21], and among the most common types of fraud is identity theft. According to the Identity Theft Resource Center 2018 report on the Non-economic Impacts of Identity Theft, individuals facing digital crises find little help from the companies who leaked their data, credit reporting agencies, or governmental legal systems. These legal failures are due to a severe lack of infrastructure and regulations surrounding the digital protections of online users.

In the legal context, current privacy laws (such as the Fourth Amendment) were created for a physical space where the individual is responsible for protecting their physical, private, personal boundaries like a home, family, or body. This analogy, however, does not translate well to a digital realm where personal boundaries are fuzzy and interwoven. As a result, seeking legal justice for digital privacy breaches is very difficult. Moreover, those who have the duty to protect the digital civil society need to be more effectively making arrangements to do so.

In 2019, only about 10% of US adults reported not using the Internet [2]. Online interaction is integral to most Americans' daily occupational and personal routines. For example, in 2019, 72% of Americans reported actively using social media platforms, filling many roles in a digital user's daily life [18]. Users depend on these platforms to build community, interact with others, view the news, learn, explore new trends, shop, and consume entertainment. In many ways, social media plat-

forms are the new medium our society uses to form personal identity and community. Therein, we create spatially unbounded identities, unlike the ones developed in our physical and social space. This is a momentous shift for our society. In 2005, only 5% of Americans were using such platforms.

Given such a large adoption of social media to form digital personhood, build networked counter publics [48], and our dependence on digital media outlets to develop our digital personhood, it is critical to develop appropriate social norms, laws, and technologically relevant regulations specific to the digital arena. Invoking a central question by Durkheim here is important: How can we maintain integrity and coherence in this newly formed, quickly evolving digital society [29]? Online, every small detail about your life, from the items you purchase, the people you connect with, the media you consume, to the content you create, all leave behind unique traces of information that, when combined can provide a digital dossier of a user [88]. These digital profiles, in turn, can influence a user's behavior, target them to purchase items, steal their identity, profile them for crimes, or deny a user of a future opportunity. How do we innovate in a digital world by leveraging incredible amounts of data while allowing for the free development of personality, unencumbered by surveillance and malevolent use of these same data? In 2019, 81% of American adults reported that they are online daily, and 28% said they are online almost constantly [75]. This daily activity produces a cornucopia of data and meta-data. This immeasurably valuable new resource is currently being inexpensively collected, hosted, combined, and analyzed for its maximum future return on investment. So how do we protect ourselves in this digital state of nature and maintain digital autonomy? The onus of this protection falls on individuals to protect both security and privacy online.

One may argue that someone who cares about safety and privacy can leave the digital age. Many social media platforms, such as Facebook and WhatsApp, present participation in their digital state of nature as a take-it-or-leave-it proposition in their terms of service. For example, in many instances, when an online user is faced with a consent request, there is no room for negotiation; thus, there is a lack of any meaningful choice. They have a binary option of consenting away their privacy and security or being unable to use the online service. The financial model for many online services solely depends on collecting and selling data or ads to consumers. In many instances, these companies are gaming the individual's desire for short-term gratification in exchange for valuable consumer information [32, 91, 82]. These services are difficult to opt out of since social media platforms are an important social ecology where people form personhood and maintain personal relationships. However, there is little to no power on the part of the individual to negotiate the terms with these companies. Online privacy then turns into an unfortunate social optimization problem, where the user must choose between the pressures of disclosing too much personal information (being digitally crowded) and being socially isolated [1]. If indeed the digital world is supposed to be a part of our civil society, an individual should not even have to make that decision.

Examples of this tension play out with specific technology, like the recent use of end-to-end encrypted messaging services like Signal or WhatsApp. Where the online user has chosen privacy at the cost of security, these platforms are presented as technology free from governmental and corporate control. End-to-end encryption means that no one outside of the message can read the content of the message (this does not include inferences on metadata in some cases or screenshots from one of



Figure 1.2: The Social Optimization Problem by Julia Zimmerman

the participants in the message). This also means that all users on these messaging services can delete all messages for both parties. In the case of end-to-end encryption, the users are opting for total privacy. This presents a tension between privacy and security because complete privacy, in this case, means there is a lack of security from law enforcement when there is a threat or illegal act in the messages. In the history of privacy, we saw that protecting sealed mail was a significant value. Still, in this instance, once the mail was sent to the recipient, the sender no longer had control over the message's contents, which means it would be challenging to hold the other parties in the chat accountable and provide safety in this space.

Privacy in a civil society is a fundamental right, and combating public surveillance is essential. However, individuals are now deciding what technology to use in response to a lack of regulation and protection of privacy in the digital world. Consequently, some of these individuals may be taking on a duty that is not theirs and maybe over-correcting in a way that threatens their security.

Privacy is not a virtue in and of itself. Privacy works in an ecosystem of other rights that work together to help ensure human flourishing. Security is difficult to

uphold in a space with little regulatory boundaries where illegal activity (like human trafficking coordination) can occur without any consequences behind end-to-end encryption.

One right alone is not good in and of itself, and it is morally suspect to present privacy as a stand-alone right. The world is a complex system where a tapestry of rights works together to form a civil society. This is not to say that those who use these services generally participate in an immoral activity. The criticism here is that the legal frameworks that should have been effectively arranging safe and private digital spaces online have failed, leaving a duty to individuals that should not be in their hands. How do we balance the potential tradeoffs between these rights? Kleinig et al suggest:

At the end of the day, there are reasons for thinking that both terms (balance and tradeoff) fail to do justice to the complexity involved in clashes between security and liberty – even though the term ‘tradeoff’ comes rather closer to the mark. That is also Waldron’s position. What is required is judgment, and judgment is not a matter of algorithmically drawing conclusions from premises but of incrementally bringing reasons to bear on one another – point and counterpoint – until we can reach a conclusion that is defensible [55].

Moreover, the standard threats in the digital age are becoming ubiquitous. To hark back to Shue again, if “people who walk along after dark are likely to be assaulted, to if infant mortality is 60 per 1000 live births, we would hardly say that enjoyment of, respectively, security or subsistence had yet been socially guaranteed ... a right involves a rationally justified demand for social guarantees against standard

threats” [84]. From the above, there are standard threats that are not reasonably or effectively made arrangements for, which is a failure on those with the duty to protect civil society, digital or otherwise.

1.4.1 PRIVACY AND SECURITY IN THE DIGITAL AGE

The conceptualization of privacy needs to pragmatically adapt to new circumstances and technologies in the digital age through new laws and frameworks.

Updating privacy and security frameworks is critical because new ambient surveillance technology (e.g., geolocation tracking, RFID tracking, internet of things tracking, smart cars, video surveillance, facial recognition, inferences, thermal prints) challenges the historical conceptualizations of privacy as something that should only be reasonably expected inside the home or pertaining to one’s body. Historically there was a notion that privacy could only be reasonably expected in personal spaces. Still, as the practical experience of being in public changes, these notions of privacy may need to develop more pragmatic approaches to protect people’s right to a reasonable expectation of privacy in a given context and their integrity.

The digital age challenges classical notions of physical personal boundaries for several reasons. Namely, the diffuse and networked nature of data, the development of ambient tracking technology (especially biometric tracking), the power of predictive inference and shadow networks, the potential harm of aggregated behavioral data, and a lack of transparency.

Data is distributed and networked: Data privacy often assumes that individual data subjects have complete control over their data and the ability to trade privacy for online services. However, this assumption overlooks that personal data can

contain sensitive information about groups or network neighbors and can be accessed by a wider audience than the immediate data processor.

The online social ecology’s densely interconnected nature creates a significant challenge to the traditional framework that takes an individualistic boundary. When personal information is shared online by a user, they also leak personal information about others in their social network (digital or otherwise) because network neighbors share information about their contacts (e.g., sharing contact lists of people when joining a new platform, network properties of homophily, sharing group photos) [38, 6]. As per Bagrow et al.’s estimation, “due to the social flow of information, we estimate that approximately 95% of the potential predictive accuracy attainable for an individual is available within the social ties of that individual only, without requiring the individual’s data” [6]. On the other hand, these network properties are also tools that groups can utilize to provide security from threats such as misinformation and disinformation. **In Chapter 3 we will explore how diverse network structures can work to create *herd privacy* where the biases of your network neighbors can help to protect you from your blind spot and detect types of misinformation that you may overlook.**

As a result, the individual cannot have complete control over their data flow. Moreover, groups need more control over their data in the traditional framework. If the network effect strongly influences user privacy, consider group-level mechanisms to control data in addition to individual controls such as consent. **In Chapter 2, we present a framework of distributed consent that considers the dispersed nature of personal data and the privacy of groups online. Distributed consent can help limit a group’s observability in an online social network**

and provide a level of *herd correction* for those in the network surrounded by neighbors who practice distributed consent.

The potential harm of aggregated behavioral data: The harms associated with collecting data increase superlinearly as different data types are aggregated together. The harms become more than the sum of their constituent parts. Datasets become more sensitive and pose more privacy risks to a subject as more data types are combined. For example, a dataset of names may not be considered sensitive. Once combined with associated social security numbers, the data becomes much more sensitive than each data point on its own. Further, if combined with a personal address, this information collection becomes more sensitive because it can be used to assume a person’s digital identity. With the collection of even more information, such as browsing history, this aggregate of information can be used to build a digital dossier [88] on the data subject, which can be used to make inferences about them, predictions about their future behavior, and manipulate their future behavior [16, 90] with targeted advertising and other techniques.

It is well studied that aggregation of PII and associated data types increase the data privacy risk to the data subject [96]. Privacy risk pertains particularly to the potential or actual harm to individual data subjects. According to Wagner et al. [96], the impact of privacy risk can be broken down into composite categories: scale, sensitivity, expectation, and harm. The number of individuals implicated can be quantified scale. Sensitivity can be quantified by the number of data types involved, entropy, and the average privacy setting. The expectation of risk can be quantified as deviation from the expectation of how the data will be handled. Finally, harm can be quantified as damages awarded or perceived harm. **We will discuss measures**

for sensitivity and aggregating data in Chapter 4.

The development of ambient tracking technology: With classical notions of privacy, a data subject essentially yields their expectation of privacy when in public because they cannot hold a reasonable expectation for privacy in public spaces. New surveillance technology, such as thermal tracking [55], is challenging the traditional concept of privacy in public spaces. By monitoring people’s behaviors and bodies, the idea of public spaces being free from surveillance is being tested. For example, biometric tracking can continuously monitor your body without physical contact. Facial recognition, as another example, is becoming readily used in public. In most US states, facial recognition is allowable and is used by law enforcement. One might argue that this rapidly progressing biometric and surveillance technology may be developing so quickly that a common person would not reasonably expect to be monitored in such a way in a public space and may indeed violate privacy. In addition, there is little regulation on the accuracy of this technology, which is a concern due to the likelihood of false positives for technologies like facial recognition [43, 42].

Moreover, as Kleinig et al. state, this surveillance technology can be further exacerbated by dissemination. If a video recording of someone “occurs in a public space, though one might presume that what can be seen using such a camera would be considered an intrusion into a private domain. Posting it on a video-sharing site like YouTube would aggravate the invasion” [55].

The potential harm of public surveillance: Helen Nissenbaum argues that a new framework of “contextual integrity” benchmarks must be adopted and implemented into privacy frameworks to protect our integrity in public spaces. Contextual integrity, like the pragmatic notion of privacy mentioned earlier, creates a more flex-

ible outer boundary for privacy benchmarks. It can reflect the appropriate context for practical privacy issues and align with norms within those contextual spheres. This framework gets rid of the old-fashioned idea that there are spheres of public spaces that are not governed by norms of information flow. Taking this approach, the outer boundaries are now defined by several different spheres of life (e.g., education, politics, markets, etc.) that all have their contextual norms, and a reasonable expectation for privacy should follow the norms for that given sphere. In practice, people will move in and out of different spheres as they experience and live their daily lives. These distinct spheres have norms, governance structures, expectations, and practices generally grounded in ordinary experience [95]. Consequently, these spheres will also hold their informational norms and contextualized notions of appropriate expectations of privacy while within that sphere.

Nissenbaum partially bases her argument on contextual integrity on the notion of “spheres of justice” by Michael Walzer, outlined in his works on distributive justice, complex equality, and the pluralist theory of justice [98]. In distributive justice, the good (e.g., wealth, education, safety, privacy, employment, etc.) is distributed across relatively autonomous spheres through complex equality, which is how goods are distributed according to unique sets of norms for that particular sphere. For example, in the sphere of mental healthcare, it may be appropriate for me to share private details with my therapist under the agreement of confidentiality. However, it would not be a fair or reasonable expectation for that relationship to be reciprocal (this pair-wise relationship is not bidirectional).

According to Nissenbaum, in this framework, a privacy violation is when “either contextual norms of appropriateness or norms of flow have been breached” [68]. Infor-

mation is marked with the original sphere and the norms therein. There is a violation, for example, if the information flow from one sphere is distributed according to the criteria of a different sphere (e.g., educational information produced in an educational context being sold in the marketplace and treated as consumer information). When examining the contextual integrity of a privacy violation, it's crucial to identify the parties responsible for collecting, processing, using, and sharing the information, as well as those who have received the information. Additionally, it's important to analyze the content of the information, the roles of the individuals involved, their relationships, the institutions involved, the social environment, and the norms and expectations of the information's original sphere. Some protections in US law do regulate these types of information already to particular kinds of sensitive data such as those relating to children, financial data, educational information, and health information. Some laws restrict the flow of information from one sphere to another, such as insider trading.

One issue with the contextual integrity framework is that contextual integrity has low moral authority because norms often do not have firm boundaries until they are made into law. There is also a limitation if the norms in these contexts must be more dynamic to capture technological innovations. There may still be a vulnerability gap between common experience and expectations for privacy and emerging threats from new technology (e.g., if the technology is too complex or technical for the common audience to understand). However, this is an important step in integrating context into notions of privacy.

Part of the solution to this privacy gap issue will come from building assessment tools and benchmarks that capture contextual and dynamic privacy concerns. These

benchmarks help build consistent tools to measure risk and harm quantitatively, which capture changing environments. It is understandable how a dynamic benchmark may seem like a contradiction, yet integrating it into contextual framing such as that suggested by Nissenbaum [68] would help to allow these benchmarks to remain dynamic and reflective of modern privacy issues.

Wagner et al. [96] propose one promising assessment framework that presents a general framework for measuring privacy risk. This framework particularly concerns measuring the potential harm and risk caused by a particular dataset. They propose four components of privacy that should be measured quantitatively (in their framework, methods for quantification for all four components are just proposed, not fully developed). The four components are scale, sensitivity, user expectations, and harm.

Size in this framework is the number of data subjects implicated in the data. The argument is that risk (and potential harm) increases as the number of individuals are implicated in the privacy violation. **This is a simple measure of how many individuals are in the dataset on the face of it; however, as we will see in Chapter 2, once we consider the networked and distributed properties of groups and the power of inference on social data, size becomes much more difficult to measure.**

Sensitivity is the type of data involved and the extent of the potential harm imposed on data subjects. The data type refers to information about an individual that spans different spheres of life. For example, identifiers (e.g., name, IP address) are information types. They argue that mixing data types (e.g., identifies mixed with financial information) has a property where the sensitivity has a “higher than a linear combination of individual sensitivities” [96]. Meaning that the sensitivity when data

types are combined is more than the sum of its constituent parts. **Later in this thesis, we will address methods to quantitatively measure sensitivity over time in Chapter 4.**

Expectation is a measure that looks at the data subject’s reasonable expectation for privacy. The authors note that this measure tries to quantify how creepy a privacy violation would feel to a data subject. This measure depends on social norms and the contextual sphere in which the data was taken. This measure could potentially be quantified using the contextual spheres privacy framework presented by Nissenbaum [68] where a deviation from expectation would be the use of data collected in one sphere being used, processed, or shared in another contextual sphere. A measure of magnitude could also be the number of contextual spheres the data has traversed (data flow) and the number of contextual spheres combined into a single dataset (similar to sensitivity measure). This measure may also need to include qualitative measures to gauge the norms and expectations of the data subject. **This measure is not explicitly addressed in this thesis, but future work will explore mixed method tools to measure the expectations of data subjects. A person’s expectation may also depend upon the heuristics and biases that the data subject holds. In Chapter 3, we will explore how these biases may impact our decision-making and how diverse groups can work together to mitigate their impact.**

Harm can be measured in many ways, according to Wagner et al., such as financial damages, discrimination, distress, anxiety, and reputation. Again the harms need to be considered dynamically because they may be cumulative. For example, if a snapshot of a data subject’s online behavior is taken, that may implicate harm in

the form of exposed static information about the subject. Still, suppose persistent tracking of their online behavior is recorded. In that case, their data can be used to build a digital dossier which can be used to make inferences and predictions about the data subject and consequently be used to influence their future behavior. Here harm is cumulative. Due to its highly contextual nature, harm like expectation may need to include qualitative measures to gauge a data subject’s perceived harm, distress, and anxiety. Measures like financial damages may be more appropriate for quantitative measurement. Reputation and discrimination may be quantified depending on the circumstances of the privacy violation. Reputation harm, for example, may be measured through sentiment analysis of customer reviews. The disparate impact also may be used to measure discrimination [50, 56, 53, 69, 49, 19].

Finally, Wagner et al. add a measure of the *likelihood* of a privacy violation occurring as the likelihood of an attack, the likelihood of adverse effects, and the potential exploitability of the dataset.

Currently, there are many promising and important assessment tools [76, 23, 69, 27, 49, 56] that aim to audit or measure the accuracy and fairness of technological systems, provide benchmarks, and measure risk and potential harm, such as the framework we explored by Wagner et al. As this is an emerging field, much more work must be done to holistically provide the quantitative tools to dynamically assess privacy in context. Future work must explore how to build dynamic assessment tools that better capture the contextual nature of the harms and risks associated with modern privacy concerns. Moreover, these tools must understand how privacy harms and risks impact data subjects from the individual, group, and societal levels to avoid creating additional back doors for privacy violations and privacy gaps in society.

This thesis aims to provide three case studies for analyzing privacy with particular attention to the group level to demonstrate and address important security gaps.

BIBLIOGRAPHY

- [1] Altman, I. (1977). Privacy regulation: Culturally universal or culturally specific? *Journal of Social Issues*, 33(3):66–84.
- [2] Anderson, M., Perrin, A., Jiang, J., and Kumar, M. (2019). 10% of Americans don’t use the internet. Who are they. *Pew Research Center*.
- [3] Apostle, H. G. (1980). Aristotle’s categories and propositions.
- [4] Appleton, S. F. (2016). The Forgotten Family Law of Eisenstadt v. Baird. *Yale JL & Feminism*, 28:1.
- [5] Assembly, U. G. et al. (1948). Universal declaration of human rights. *UN General Assembly*, 302(2):14–25.
- [6] Bagrow, J. P., Liu, X., and Mitchell, L. (2019). Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3(2):122.
- [7] Baldwin, D. A. (2018). The concept of security. In *National and International Security*, pages 41–62. Routledge.
- [8] Bambauer, D. E. (2013). Privacy versus security. *J. Crim. L. & Criminology*, 103:667.
- [9] Barrett, L. (2020). Ban facial recognition technologies for children-and for everyone else. *BUJ Sci. & Tech. L.*, 26:223.
- [10] Bell, C. (1993). Planned Parenthood of Southeastern Pennsylvania, et al. v. Robert P. Casey, et al. *Feminist L. Stud.*, 1:91.
- [11] Bennett, C. J. (2012). The accountability approach to privacy and data protection: Assumptions and caveats. In *Managing privacy through accountability*, pages 33–48. Springer.
- [12] Berger, D. D. (2010). Balancing consumer privacy with behavioral targeting. *Santa Clara Computer & High Tech. LJ*, 27:3.
- [13] Beverage, J. (1976). The Privacy Act of 1974: an overview. *Duke law journal*, 1976(2):301–329.
- [14] Birnhack, M. D. (2008). The EU Data Protection Directive: An engine of a global regime. *Computer Law & Security Review*, 24(6):508–520.
- [15] Bok, S. (2011). *Secrets: On the ethics of concealment and revelation*. Vintage.
- [16] Calo, R. (2013). Digital market manipulation. *Geo. Wash. L. Rev.*, 82:995.
- [17] Carow, K. A. and Heron, R. A. (2002). Capital market reactions to the passage of the Financial Services Modernization Act of 1999. *The Quarterly Review of Economics and Finance*, 42(3):465–485.
- [18] Center, P. R. (2019). Social Media Fact Sheet. *Pew Research Center*.

- [19] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.
- [20] Clearwater, A. and Hughes, J. T. (2013). In the Beginning-An Early History of the Privacy Profession. *Ohio St. LJ*, 74:897.
- [21] Commission, F. T. (2018). Consumer sentinel network data book 2018.
- [22] Commission, U. S. P. P. S. (1977). *Personal Privacy in an Information Society: The Report of the Privacy Protection Study Commission*, volume 2. The Commission.
- [23] Costanza-Chock, S., Raji, I. D., and Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583.
- [24] Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- [25] Davis, G. C. (1962). The Transformation of the Federal Trade Commission, 1914-1929. *The Mississippi Valley Historical Review*, 49(3):437–455.
- [26] Definitions Under CCPA (2020). California Consumer Privacy Act (CCPA). *Policy*.
- [27] Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.
- [28] D’ignazio, C. and Klein, L. F. (2020). *Data feminism*. MIT press.
- [29] Durkheim, E. (2014). *The division of labor in society*. Simon and Schuster.
- [30] Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- [31] Faden, R. R. and Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.
- [32] Farmer, J. D. and Geanakoplos, J. (2009). Hyperbolic discounting is rational: Valuing the far future with uncertain discount rates.
- [33] Fenrich, K. (2008). Securing your control system: the CIA triad is a widely used benchmark for evaluating information system security effectiveness. *Power Engineering*, 112(2):44–49.
- [34] Fischer, R., Halibozek, E., Halibozek, E. P., and Walters, D. (2012). *Introduction to security*. Butterworth-Heinemann.
- [35] Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A*, 374(2083):20160112.
- [36] Fry, B. G., Therese, M., Weckmueller, B., et al. (1997). The family educational rights and privacy act of 1974. *Student records management: A handbook*, 43.

- [37] Gadbow, T. (2016). Legislative update: Children’s Online Privacy Protection Act of 1998. *Child. Legal Rts. J.*, 36:228.
- [38] Garcia, D. (2017). Leaking privacy and shadow profiles in online social networks. *Science Advances*, 3(8):e1701172.
- [39] Gerety, T. (1977). Redefining privacy. *Harv. CR-CLL Rev.*, 12:233.
- [40] Godkin, E. L. (1880). Libel and its legal remedy. *J. Soc. Sci.*, 12:69–80.
- [41] Godkin, E. L. (1890). The Rights of the Citizen, IV—To His Own Reputation. *Scribner’s Magazine*, 8(1):58–67.
- [42] Goldberg, R. D. (2020). You Can See My Face, Why Can’t I? Facial Recognition and Brady. *HRLR Online*, 5:261.
- [43] Goodwin, G. L. (2019). Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains, Statement of Gretta L. Goodwin, Director, Homeland Security and Justice, Testimony Before the Committee on Oversight and Reform, House of Representatives. In *United States. Government Accountability Office*, number GAO-19-579T. United States. Government Accountability Office.
- [44] Habermas, J. and Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press.
- [45] HIPPA (1996). Health insurance portability and accountability act of 1996. *Public law*, 104:191.
- [46] Holvast, J. (2007). History of privacy. In *The history of information security*, pages 737–769. Elsevier.
- [47] Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8):56–59.
- [48] Jackson, S. J. and Foucault Welles, B. (2015). Hijacking# myNYPD: Social media dissent and networked counterpublics. *Journal of Communication*, 65(6):932–952.
- [49] Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.
- [50] Jain, R. K., Chiu, D.-M. W., Hawe, W. R., et al. (1984). A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 21.
- [51] James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.
- [52] James, W. (1975). *Pragmatism*, volume 1. Harvard University Press.
- [53] Klein, B., Ogbunugafor, C. B., Schafer, B. J., Bhadracha, Z., Kori, P., Sheldon, J., Kaza, N., Sharma, A., Wang, E. A., Eliassi-Rad, T., et al. (2023). COVID-19 amplified racial disparities in the US criminal legal system. *Nature*, pages 1–7.
- [54] Kleinig, J. (1982). The ethics of consent. *Canadian Journal of Philosophy*,

- 12(sup1):91–118.
- [55] Kleinig, J., Mameli, P., Miller, S., Salane, D., and Schwartz, A. (2011). *Security and privacy: global standards for ethical identity management in contemporary liberal democratic states*. ANU Press.
 - [56] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, 30.
 - [57] Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101:761.
 - [58] Locke, J. (2015). *The second treatise of civil government*. Broadview Press.
 - [59] Menikoff, J., Kaneshiro, J., and Pritchard, I. (2017). The common rule, updated. *N Engl J Med*, 376(7):613–615.
 - [60] Mihm, S. (2003). Dumpster-Diving for Your Identity. *The New York Times Magazine*, pages 42–42.
 - [61] Miller, J. H. and Page, S. (2009). Complex adaptive systems. In *Complex Adaptive Systems*. Princeton university press.
 - [62] Mitchell, M. (2009). *Complexity: A guided tour*. Oxford university press.
 - [63] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.
 - [64] Moore Jr, B. (2023). *Privacy: Studies in social and cultural history*. Taylor & Francis.
 - [65] Moosavian, R. (2021). Pavesich v New England Life Insurance Co (1905).
 - [66] Morin, E. (1992). From the concept of system to the paradigm of complexity. *Journal of social and evolutionary systems*, 15(4):371–385.
 - [67] Newman, M. (2018). *Networks*. Oxford university press.
 - [68] Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, 140(4):32–48.
 - [69] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.
 - [70] O’Brien, D. M. (1979). Privacy, law, and public policy.
 - [71] of Health, U. S. D. and Services, H. (1973). *Secretary’s Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens: Report*. MIT Press.
 - [72] O’neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
 - [73] Pedersen, K. (2019). Manipulating the New Hampshire Mail: Political Power and the American Postal Service, 1792-1829.
 - [74] Peirce, C. S. (1905). What pragmatism is. *The monist*, 15(2):161–181.
 - [75] Perrin, A. and Kumar, M. (2019). About three-in-ten US adults say they are

- ‘almost constantly’online. *Pew Research Center*.
- [76] Radiya-Dixit, E. and Neff, G. (2023). A Sociotechnical Audit: Assessing Police Use of Facial Recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1334–1346.
 - [77] Rittel, H. (1967). Wicked problems. *Management Science*, (December 1967), 4(14).
 - [78] Roraback, C. G. (1989). Griswold v. Connecticut: A Brief Case History. *Ohio NUL Rev.*, 16:395.
 - [79] Rosen, J. (2011). *The unwanted gaze: The destruction of privacy in America*. Vintage.
 - [80] Sayama, H. (2015). *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks.
 - [81] Schein, M., Avery, R. J., and Eisenberg, M. D. (2022). Missing the mark: The long-term impacts of the Federal Trade Commission’s Red Flag Initiative to reduce deceptive weight loss product advertising. *Journal of Public Policy & Marketing*, 41(1):89–105.
 - [82] Schermer, B. W., Custers, B., and van der Hof, S. (2014). The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics and Information Technology*, 16(2):171–182.
 - [83] Seifert, J. W. and Relyea, H. C. (2008). E-government act of 2002 in the United States. In *Electronic Government: Concepts, Methodologies, Tools, and Applications*, pages 154–161. IGI Global.
 - [84] Shue, H. (2020). *Basic rights: Subsistence, affluence, and US foreign policy*. princeton University press.
 - [85] Singleton Jr, R., Straits, B. C., Straits, M. M., and McAllister, R. J. (1988). *Approaches to social research*. Oxford University Press.
 - [86] Skinner, T. H. (2003). California’s Database Breach Notification Security Act: The First State Breach Notification Law Is Not Yet A Suitable Template For National Identity Theft Legislation. *Rich. JL & Tech.*, 10:1.
 - [87] Solove, D. J. (2002). Conceptualizing privacy. *California law review*, 90:1087–1155.
 - [88] Solove, D. J. (2004). *The digital person: Technology and privacy in the information age*, volume 1. NyU Press.
 - [89] Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania law review*, pages 477–564.
 - [90] Susser, D., Roessler, B., and Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, 4:1.
 - [91] Thaler, R. (2005). Advances in behavioral economics. *Russel Sage Foundation*.
 - [92] Thurner, S., Hanel, R., and Klimek, P. (2018). *Introduction to the theory of complex systems*. Oxford University Press.

- [93] Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1):20–36.
- [94] Ullman, R. H. (1983). Redefining security. *International security*, 8(1):129–153.
- [95] Wacquant, L. J. and Bourdieu, P. (1992). *An invitation to reflexive sociology*. Polity Cambridge.
- [96] Wagner, I. and Boiten, E. (2018). Privacy risk assessment: from art to science, by metrics. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2018 International Workshops, DPM 2018 and CBT 2018, Barcelona, Spain, September 6-7, 2018, Proceedings 13*, pages 225–241. Springer.
- [97] Waller, S. W., Heidtke, D. B., and Stewart, J. (2013). The Telephone Consumer Protection Act of 1991: Adapting Consumer Protection to Changing Technology. *Loy. Consumer L. Rev.*, 26:343.
- [98] Walzer, M. (2008). *Spheres of justice: A defense of pluralism and equality*. Basic books.
- [99] Warren, S. and Brandeis, L. (1989). The right to privacy. In *Killing the Messenger: 100 Years of Media Criticism*, pages 1–21. Columbia University Press.
- [100] Westin, A. F. (1968). Privacy and freedom. *Washington and Lee Law Review*, 25(1):166.
- [101] Wittgenstein, L. (1968). *Philosophical investigations*. Macmillan.
- [102] Wolfers, A. (1952). “National security” as an ambiguous symbol. *Political science quarterly*, 67(4):481–502.
- [103] Xiong, W. and Lagerström, R. (2019). Threat modeling—A systematic literature review. *Computers & security*, 84:53–69.
- [104] Zimmerman, H. (2017). The data of you: Regulating private industry’s collection of biometric information. *U. Kan. L. Rev.*, 66:637.
- [105] Zuboff, S. (2019). Surveillance capitalism and the challenge of collective action. In *New labor forum*, volume 28, pages 10–29. SAGE Publications Sage CA: Los Angeles, CA.

CHAPTER 2

GROUP PRIVACY: DISTRIBUTED CONSENT IN ONLINE SOCIAL NETWORKS

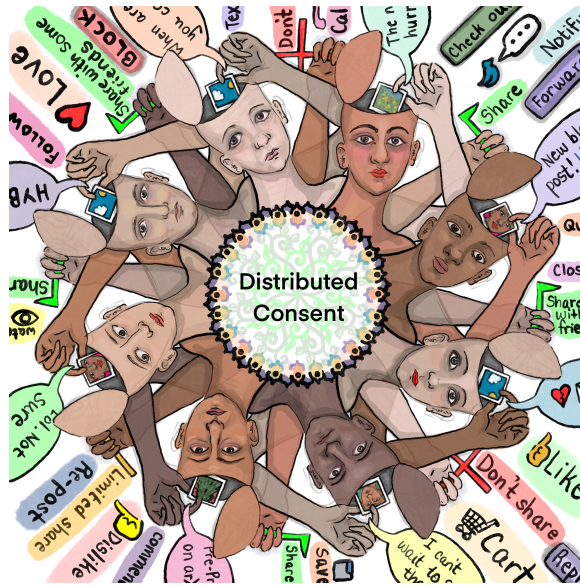


Figure 2.1: Visual Abstract of Chapter 2 by Julia Zimmerman

PREFACE

This chapter challenges the traditional concept of individual consent for controlling informational privacy in online settings. The distribution and networking of information implicate groups, and privacy and security trade-offs are explored. Users must decide whether to protect their privacy from third parties and network neighbors by disclosing security settings publicly or taking on the burden of coordinating security on single and multiple platforms. Demographics are not considered as the dynamics of distributed consent are examined. However, Chapter 3 will explore how diversity in the dynamics of network spreading.

The chapter also examines privacy settings on social media platforms and how they can be altered to protect groups and limit third-party surveillance. New security settings can help individuals safeguard their data on a single platform, but coordination is necessary across platforms to remain unobserved by network neighbors. These settings can enable privacy-conscious individuals to create subgroups that protect themselves and others in the community.

Legal regulations should be implemented on platforms and third parties to protect online group privacy; more options for collectively making consent decisions would help. Our research demonstrates significant coordination costs for individuals in safeguarding their privacy across platforms, emphasizing the need for laws to protect user privacy. Material from this chapter has been published or made publicly available in the following form:

Lovato, J., Allard, A., Harp, R., Onaolapo, J., Hébert-Dufresne, L.. (2022). Limits of individual consent and models of distributed consent in online social networks.

In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2251-2262. DOI: 10.1145/3531146.3534640.

ABSTRACT

Personal data are not discrete in socially-networked digital environments. A user who consents to allow access to their profile can expose the personal data of their network connections to non-consented access. Therefore, the traditional consent model (informed and individual) is not appropriate in social networks where informed consent may not be possible for all users affected by data processing and where information is distributed across users. Here, we outline the adequacy of consent for data transactions. Informed by the shortcomings of individual consent, we introduce both a platform-specific model of “distributed consent” and a cross-platform model of a “consent passport.” In both models, individuals and groups can coordinate by giving consent conditional on that of their network connections. We simulate the impact of these distributed consent models on the observability of social networks and find that low adoption would allow macroscopic subsets of networks to preserve their connectivity and privacy.

2.1 INTRODUCTION

One key focus of the burgeoning field of data ethics concerns how big data and networked systems challenge classic notions of privacy, bias, transparency, and consent [34]. In particular, the traditional privacy model, which relies on individual

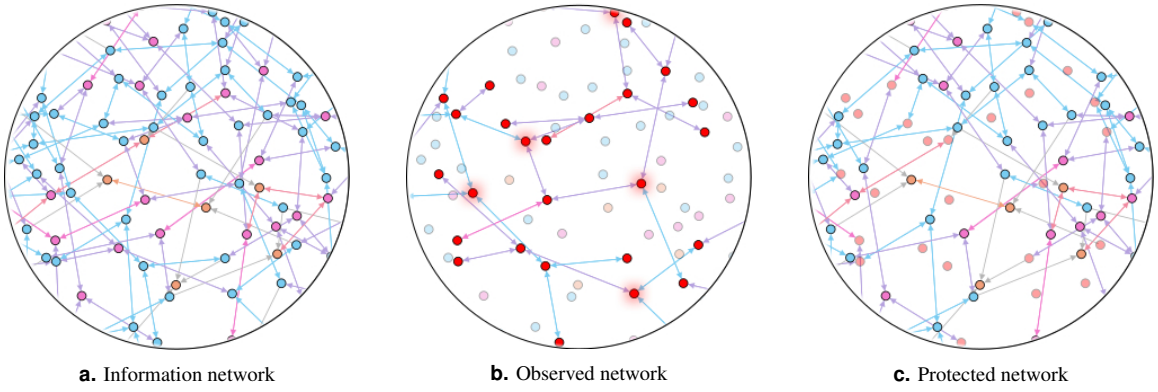


Figure 2.2: (a.) Information flow across a network with our basic implementation of distributed consent. Blue nodes have the lowest security settings and are susceptible to surveillance from third-party applications or websites. Purple nodes have stricter security settings but share their posts and data with all their neighbors. Orange nodes follow a distributed consent model and only share their data with purple nodes or other orange nodes. (b.) The same network where a third party directly observes a handful of low-security accounts is highlighted in red with shading. All nodes sharing their data with directly observed accounts are also de facto observed and shown in red. Nodes at a distance $L > 1$ can also be observed if the third party leverages some statistical procedure, inferring data up to a distance of two from directly observed nodes. (c.) We show the remaining unobserved or protected subnetwork.

self-determination and individual consent, we argue, is no longer appropriate for the digital age. First, the traditional privacy model requires that consent be *informed*, which may not be possible in the context of large data sets and complicated technologies. Second, the traditional privacy model presumes *individual* control over personal information, even though the flow of information in networked systems precludes anyone from having such control over any piece of data. While the modern information environment shows both conditions as problematic, and while we briefly discuss the information condition, we focus most of our attention on the individuality condition.

Individual consent (by which we mean requiring that individual end-users consent in order for some action or outcome to be permissible) has many limitations—notably, we live in a highly networked [38] and advanced technological society, where digital decisions and actions are interconnected and affect not just ourselves but our digital community as a whole. In a digital age, individual consent is flawed [11] and ineffectual when protected class data and social profiles can be easily inferred via our social networks [24, 6, 10, 39, 53]. The individual consent model works most effectively in a physical space with accepted boundary norms [63, 49], linear contacts between two discrete parties, and no externalities. This, however, does not translate well to a digital realm where personal data boundaries are fuzzy and interwoven. The current over-reliance on individual consent online has also led to a negative externality of less legitimate consent due to consent desensitization, in part, because users are now faced with a deluge of consent requests [54]. Thus, a new approach to data privacy and consent in this context is needed.

A new data privacy model will need to consider several factors: the networked virtual space that we occupy; integration of group consent; and a mechanism for

distributed moral responsibility when data privacy is breached or data are processed, combined, or manipulated in unethical manners [22]. In this Chapter, we will focus on distributed consent in particular and evaluate, in a mathematical model, its potential to increase online social networks’ general privacy. We aim to cover the latter data privacy concerns in future work. In addition, future work could explore the potential for early adopters of distributed consent to influence their network neighbors (e.g. via cascade effects towards a contagious taste for privacy).

2.1.1 A CRITIQUE ON THE ADEQUACY OF INDIVIDUAL CONSENT FOR DATA TRANSACTIONS

We will call the means by which information is shared *personal data transactions*. Broadly speaking, a personal data transaction is any transaction in which one party gives or reveals personal information to another, so the category is vast; it includes every behavior and every speech act that imparts information of some kind to another.

We can narrow the broad category of personal data transactions in three ways for our purposes here. First, we are interested in those transactions for which the *primary purpose of the actions or transaction is the transfer of information*. We will not attempt to give a complete conceptual framework here, but we can provide some indication of what we mean. If *A* gives *B* money, and *B* gives *A* a shirt, information has been exchanged, but the primary purpose of *A*’s giving money, and *B*’s giving the shirt was not the information exchange—it was the exchange of goods. If *A* asks *B* how their day was and *B* replies “pretty good,” information has been exchanged, but it is possible that the primary purpose was not the information exchange but rather

the demonstration of caring or the strengthening of solidarity. So we will focus on data transactions where the primary purpose behind the transaction is the transfer of information itself—though it is important to know that the information might be valued because of further downstream uses of the information. If *A* asks *B* about where *B* grew up, what *B*’s mother’s maiden name is, and the name of *B*’s childhood pet, then *B*’s responses would be data transactions—even if *A*’s ultimate reason in asking the questions were to get access to *B*’s online banking accounts.

Second, we are interested in those transactions which are about *personal information*. As with the previous case, we cannot precisely conceptualize personal information here. The general idea is that personal information is information about an identifiable living person [44]. If *A* downloads all of the 1911 Eleventh Edition of the *Encyclopaedia Britannica* and gives it to *B*, that is a data transaction in the broad sense but is likely not a transfer of personal information in the sense that we are interested in here (assuming that *B* is not an entry in the encyclopedia). If *A* gives *B* information about *A*’s whereabouts over the past 24 hours or information about *A*’s food preferences, that is personal information. It is important to note that personal information need not be information about the person giving it to someone else; *A* can give *B* information about *C*’s food preferences or whereabouts, which would constitute a data transaction. Here again, the boundaries are difficult to layout precisely.

Where do the limits of personal data lie? Suppose *A* were to give *B* a copy of a biography of Barack Obama: *A*’s giving *B* the biography is outside of the scope of data transactions that we are interested in because the personal information about Obama contained in the biography is presumed to be *public*. Data transactions are

valuable insofar as some party is gaining something of value, and discrete information that is previously known is not valuable. Moreover, that suggests a third way to restrict the domain of data transactions that we are interested in: we are concerned only with data transactions that deal with *non-public information*.

So when we talk about personal data transactions, we are talking about exchanges of personal information between two parties where the exchange of information is the (or a) primary purpose of the exchange, where the information is personal, and where the information is non-public. Personal data transactions of this form make up a significant and increasing portion of our modern lives. The following examples are just two of many but highlight personal data exchanges that have the potential to be highly impactful on our moral lives.

As one example, consider our use of social media. When we make an account on social media, we potentially engage in three different kinds of personal data transactions. First, there are the data transactions between the person with the account (the “end-user”) and the social media company or platform. Second, there are the data transactions between one end-user and other end users on the platform between whom there are some network connections. Third, there are data transactions between end-users and third-party auxiliaries (e.g., individuals, apps, bots) that receive or process data to enhance the end user’s social network experience. As examples of these, consider Facebook: anyone creating a Facebook account shares with the site all of the information they intentionally put on the site (their profile information, their network connections/friends, their posts). There is also a host of information that they might not be aware they are putting on the site (their location, the amount of time they spend reading posts or watching ads). The Facebook user also builds

networks of social connections (nodes) on the platform.

Selected information is shared among various nodes on the friendship network in accordance with the end user's preferences. End-user A might share a lot of information with friends B and C . User A is close friends with B and C and may want to share information such as profile data, location data, and the content of all of their posts. User A is an acquaintance with users D and E and may want data sharing to be limited to only some posts or only some photos, or no location information or information about A 's friend network. The third kind of information flow is that which is shared with third parties. For example, third parties might be designers of games or quizzes that can be run on the Facebook platform, such that those third parties can then collect information Facebook possesses about the end-user playing the game.

Cambridge Analytica [29] is an example of a third party who acquired information from end-users through games or apps on the Facebook platform; Facebook has since changed some of its policies on third-party data acquisition. Cambridge Analytica was a political data processing firm whose business model centered around gathering user data and using this information to influence the user's voting behavior. In 2018, this firm gathered significant attention because of the personal information it collected from over 50 million Facebook users. The firm used the bait of a personality survey to sway 270,000 users to download an app (in the Terms of Service, these users consented to have their data shared with academics), which was used as an entryway to scrape the user's profile and also the user's friends profiles. The firm collected information on user identities, tracked their online behavior, and gained information on their social ties; they created digital dossiers on users and, in turn, used that information to

purchase paid advertising to influence their voting behavior.

Third-party websites can also allow users to sign in with their Facebook accounts, and those third parties can then get some user information when people use their Facebook accounts to log in. While not every social media site offers the same options for how information is shared across those three modalities, having an account on any social media site entails sharing information in each of those three different ways.

2.1.2 A THEORY OF CONSENT

The fact that people have putatively consented to all of the personal data transactions that we identified above is supposed to be doing a lot of normative work: it takes information gathering actions that would have been impermissible and supposedly makes them permissible. As we have seen, our modern lives are filled with personal data transactions. And while it has been argued that the ubiquity of these personal data transactions is such as to entail that we are living within a *de facto* surveillance state [68], there is at least one important *prima facie* difference: surveillance states are typically imposed on subjects without their consent, so that personal information in a surveillance state is gathered without any consideration to the preferences of the surveilled, whereas (it is claimed) we freely consent to the personal data transactions that we are subject to. (Note that we are not here endorsing this argument; we present it merely in order to motivate our analysis of the role of consent in personal data transactions.) In this way, consent for personal data transactions functions analogously to how consent functions in medical ethics and how consent functions in sexual ethics.

To borrow a phrase used by both Heidi Hurd and Larry Alexander, consent is a

kind of “moral magic:” “it transforms acts from impermissible to permissible” [28, 2]. This is true in generic ways; *A*’s entering *B*’s house can be either a trespass or a permissible visit based on whether *B* has consented to *A*’s entering. It is particularly true in matters of sexual ethics, where the moral status of a sexual act crucially depends on whether the act is consented to at the time that it is performed [19]. Furthermore, it is equally valid for medical ethics, where invasive medical procedures and treatments are impermissible unless consented to by the patient.

Regardless of the legality, it is reasonable to think that if *A* were to persistently and systematically surveil *B* (going through *B*’s trash, compiling every public record about *B*, recording all of *B*’s public movements), and *B* were not a public figure who might be an appropriate target of community scrutiny, then that would be an impermissible form of information gathering. Of course, if *B* were to consent to their information being gathered in this way by *A* (say they are the willing subject of a documentary), then the information-gathering would thereby become permissible. So consent would have the same kind of moral magic as it does in other contexts [43].

Likewise, personal information transactions open up the risk of one’s autonomy or integrity being violated. After all, personal information is personal—and as such, it potentially enables others to identify them and predict or control one’s behavior in a way that compromises one’s autonomy and capacity to act on one’s intentions and one’s own conception of the good. It is true that we share personal information with others around us all the time; the mere sharing of personal information does not compromise one’s autonomy. However, we typically share personal information with those we trust to use the information correctly. We share information that is anodyne enough to not threaten our ability to pursue our own goals; we do not typically share

personal information with those we know intend to use that information to circumvent our autonomy or act contrary to our interests. This is why privacy is important even for those who do not think themselves to have a strong taste for privacy: autonomy matters because it is the means by which we pursue the good, and privacy is connected to autonomy. Getting legitimate consent to personal data transactions helps to ensure that one's autonomy is safeguarded.

Finally, personal data transactions often involve commercial parties, either as one of the parties to the data transaction (as happens when you upload your information to Facebook) or as the platform in which data transaction occurs (as when you share your data with your friends through Facebook). Commercial parties have an interest in limiting their liability. To this end, obtaining putative consent to acquire and process data helps limit the legal liability they face. Companies and commercial agents have a vested interest in preventing the end users of social media sites and data technologies from complaining. Obtaining consent helps support the argument that the end-users have, indeed, forfeited their right to complain about any consequences arising from the personal data transactions.

2.1.3 CONSENT IN DATA TRANSACTIONS

Before discussing the limitations of the individual consent model for personal data transactions, it is worth addressing one question: Why would we ever have thought that consent was relevant for data transactions at all? After all, the argument goes, data transactions are just one species of transaction. Moreover, transactions are necessarily mutually consensual; if they were not, they would not be transactions. An exchange in which A gets a beer and B gets a dollar is a transaction if they both

agree, but it is robbery or coercion if either A or B does not consent to the exchange. (The fact that both A and B receive something is irrelevant; if A breaks into B 's house and steals B 's property, it is no less a robbery just because A left something behind in exchange.) Likewise, the objection goes, that if A and B are engaged in a personal data transaction, then it is irrelevant to ask whether the transaction is consensual; if it were not, it would not be a personal data transaction but would instead be a data theft or something similar. If this is right, it is as unnecessary to ask whether a data transaction is consensual as it would be for the cashier at a clothing store to explicitly ask every patron whether they consent to trade their money for the clothes they wish to buy.

One reply to this objection is to say that we actually *do* care about obtaining consent for some exchanges precisely because we want to ensure that the exchange is a “transaction” rather than a “theft.” When the stakes are high, there is a possibility of future risk. In other words, we do seek to clarify the consensual nature of the exchange, but when the stakes are low and there is a low perception of risk, we are less concerned [56, 1, 21]. This reply is correct, but we think an important point is in danger of being obscured. In the case of ordinary transactions, there is less danger of the transaction being non-consensual precisely because there is little danger of the parties to the exchange not knowing what they are exchanging; when A gives B money, and B gives A a shirt, both parties are aware that the transaction happened because both parties have clearly gained something and have lost something.

Moreover, there are positive actions that A and B both perform, without which the transaction cannot take place; A must hand over the money, and B must hand over the shirt. However, we can easily imagine transactions in which these conditions

are not met—perhaps, for instance, it is not clear among the parties precisely what has been gained and what has been lost after a transaction. (One might think of the “sale” of the island of Manhattan by Native people that Peter Minuit orchestrated on behalf of the Dutch; what exactly is one selling if the land is still there after the transaction is done?) Alternatively, imagine transactions for which no positive actions need to be taken, like arrangements that automatically deduct money from a bank account or other store of value without anyone needing to do anything. It is certainly not challenging to sway our intuitions towards thinking of these “transactions” as theft or, at a minimum, theft-adjacent. The same issues arise with data transactions (whether legitimate or not) we might wonder whether the exchange of data between A and B is actually a transaction if one or both parties are not clear on precisely what has been gained or lost—but, as we will see, this is a common feature of most modern personal data exchanges. It is certainly not like a transaction for a shirt, where one minute you have a dollar and the next minute you do not; after a personal data exchange is complete, you have just as much of your own personal data as you started with. Personal data exchanges often do not require any positive action on behalf of the parties; we lose data in personal data exchanges all the time, through no action of our own. So while it is acceptable to say that valid transactions are consensual, it is also true that not every data exchange is a data transaction in that strict sense. It is reasonable to ask for explicit consent to data exchanges so that all parties are confident that it is a data transaction and not a mere exchange.

It is a reasonable strategy, but it is nevertheless a failed one. As we will see, there are systematic reasons why personal data exchanges cannot be justified by individual consent.

2.2 RESULTS

2.2.1 A GENERAL OVERVIEW OF THE PROBLEMS WITH INDIVIDUAL CONSENT

We can now provide a very general overview of the problems with individual informed consent when applied to data transactions. Note that we are assuming for the sake of this discussion that the relevant data are, in fact, adequately subject to control by individuals. This is a problematic assumption; data often implicate multiple individuals or are otherwise ‘co-owned’ and thus are not properly the things that individual consent can govern. This important point requires a fuller discussion, but we make this simplifying assumption here because social networks cannot function in anything like their current form without it.

Consent, in this context, should not be mistaken for a state of mind or an attitudinal event [32, 65] and it must meet certain criteria in order to be considered a valid. The legitimacy of consent hinges on a number of criteria [9]:

1. the subject has sufficient accurate information and understands the nature of the agreement,
2. the agreement is entered into without coercion,
3. the agreement is entered into knowingly and intentionally,
4. the agreement authorizes a specific course of action.

In the context of personal data transactions, digital consent also rests on the four

criteria mentioned above. The user agrees to the specified service terms and privacy policy outlined by the data processor. Notably, the four criteria listed above fail in the context of personal data transactions and classic Privacy Policies and Terms of Service (ToS) agreements.

First, most users entering into consent agreements know very little about data processing or the risks of handing over their data. The dense legal and technical nature of ToS agreements task non-experts to consent to something they do not understand [17]. This dynamic takes advantage of asymmetry in technical and legal knowledge.

Second, it is difficult to opt-out of these services since online platforms are an important social ecology where people form personhood, maintain personal relationships, and build valuable networked counter-publics [23, 31, 30, 51]. Nevertheless, there is little to no power on the part of the individual to negotiate the ToS with these companies, as consent in these ToS is typically presented on a take-it-or-leave-it basis and offers no conditions of choice [55, 45]. Online privacy then turns into an unfortunate social optimization problem [63], where the user must choose between the pressures of disclosing too much personal information (being digitally crowded) and being socially isolated [4].

Third, the volume of consent requests a user faces has led to a troublesome externality where the user is fatigued and habitually agrees to everything due to consent desensitization [17]. This delegitimizes the premise that each act of putative consent reflects the individual user’s autonomous judgment.

Fourth, the language in ToS agreements is typically so broad and open-ended that data processors have the flexibility to manipulate the data in many ways. The

consent scope cannot be so broad as to allow actions that the user could not have considered or would otherwise not have consented to. An adequately limited scope of consent also implies that there should be some mechanism for a user to check if their data are indeed following the agreed-upon course of action. However, data processors often make it very difficult [33], if not impossible, to track personal data, know what they have collected or how it is being processed, and hold them accountable for misuse [58, 13].

A fundamental assumption for individual consent is that the user has power over their personal data and can trade their personal privacy in exchange for using an online service [14]. Perhaps more importantly, a significant concern with the individual consent model is that personal data, in this context, are distributed information that contains information about more than a single individual and spans a broader communication boundary than the user is aware [49]. In reality, these data may not belong wholly to the individual. Therefore, it is not appropriate for the individual to act alone in controlling the course of action or the flow of these data. Perhaps the first step to understanding the impact of this issue in an online social media context is to understand how different levels of consent impact the flow of networked data and observability in the first place. It will be important for us to understand if the network effect has a strong influence on privacy to justify group-level consent settings.

2.2.2 A THREAT MODEL FOR LEAKY INDIVIDUAL DATA IN SOCIAL NETWORKS

The densely interconnected nature of online social ecology creates a significant problem with the model of individual consent. When users share personal information online, they are also leaking personal information about others in their social network (digital or otherwise) [24, 6]. According to Bagrow et al., “due to the social flow of information, we estimate that approximately 95% of the potential predictive accuracy attainable for an individual is available within the social ties of that individual only, without requiring the individual’s data” [6].

One example of leaky data is when a user attempts to sign on to a new online service. They may be prompted to skip the hassle of entering their personal information manually and instead opt to use an existing account to act as a secured access delegation [64] in order to gain quicker access to the new third-party online service. The online service can ask to gain access to the user’s online social network, phone or email contacts, location, and other personal data. Through these leaky data, third parties can be granted access to a wealth of knowledge about people who never consented to share their information with that particular service.

Similarly, attackers can breach social accounts via various methods, including phishing attacks, malware, and data breaches [18, 60, 27]. They can also create fake social accounts and then use those fake accounts to befriend real accounts. To extend their reach, those attackers can then monitor the activity of other accounts that are directly connected to compromised or fake accounts. By leveraging network effects, attackers can further indirectly observe the social activity of groups of accounts that

are several hops away from the captive accounts, starting with accounts that are one hop away [16, 61].

Leveraging those captive accounts, the attackers traverse the social graph or segments of it and record profile information and social activity that would later be used in future phishing and spam attacks, influence manipulation attempts [66], and disinformation campaigns [5], among others. Given the massive size and connected nature of online social networks, the potential reach of attackers and the resulting harm both have the capacity to rise to catastrophic levels. We describe examples of real-world incidents that demonstrate the severity of this problem in our discussions.

2.2.3 A MODEL OF DISTRIBUTED CONSENT AND NETWORK OBSERVABILITY

To account for the distributed nature of personal data (i.e., the distributed online self), we consider a simple model of distributed consent. Imagine a social network platform where individuals have the following privacy options:

0. Individuals share their data with all their connections and are vulnerable to third-party surveillance (similar to Facebook accounts with access for “Apps, Websites and Games” turned on).
1. Individuals share their data with all their connections but are not directly vulnerable to third-party surveillance.
2. Individuals only share their data with their connections whose privacy levels are set at least 1.

N. Individuals only share their data with connections whose privacy levels are set at least to $N - 1$.

Options 2 and greater are currently unavailable on popular social media platforms but are a first-order implementation of what we call distributed consent. Simple implementations of this concept could be an attractive setting to adopt if the platforms wish to address privacy concerns and keep users. Individuals who pick this option are stating that they want to be part of a local group that agrees on minimal privacy settings. It is a consent conditional on the consent of their neighbors in the network structure.

Imagine now that a third party wishes to observe this population, either by releasing a surveillance application on the social network or by explicitly gaining control of their accounts through similar malware. Say they can directly observe a fraction φ of individuals with privacy level set to 0 through this attack. They can then leverage these accounts to access neighboring individuals' data with the privacy level set to 1 or 0, therefore using the network structure to observe more nodes indirectly. They can further leverage all of these data to infer information about other individuals further away in the network, for example, through statistical methods, facial recognition, and other data sets.

Deep surveillance processes are relevant to other network systems where it is possible to indirectly observe nodes within a certain distance from directly observed nodes. Deep surveillance allows us, for example, to monitor an entire power grid without monitoring the voltage and line currents everywhere in the system [67]. It was recently shown that the surveillance process itself could be generally modeled through the concept of depth-L percolation [3]: Monitoring an individual allows one to monitor

their neighbors up to L hops away. Depth-0 percolation is a well-studied process known as site percolation. The third-party would then be observing the network without the help of any inference method and by ignoring its network structure. With depth-1 percolation, they would observe nodes either directly or indirectly by observing neighbors of directly observed nodes, e.g., by simply observing their data feed or timeline. Depth-2 percolation would allow one to observe not only directly monitored nodes but also their neighbors and neighbors’ neighbors, e.g., through statistical inference [6]—and so on, with deeper observation requiring increasingly advanced methods. The model is illustrated in Fig. 4.2 and detailed in our Methods section.

To study the interplay of consent and observability on social networks, we combine our distributed consent and depth- L percolation models on subsets of Facebook friendship data informed by empirical work on the demographic population’s taste for privacy. We assume that one-third of the population has a taste for privacy [35]. These data are anonymized with all metadata removed and are simply used to capture the density and heterogeneity of real online network platforms. We use distributed consent with a security level up to $N = 2$ and an observation process with $L = 2$ (observing nodes two hops away from the compromised account). As we will see, those values mean that the third party is more sophisticated than our distributed consent mechanism, and no one is *guaranteed* to be unobservable. We then set 1% of accounts with the lowest security setting to be compromised and directly observed such that that between 90% and 100% of the population will be observed given the default security settings. We then ask to what extent distributed consent can preserve individual privacy even when a large fraction of nodes can be directly observed and

third parties can infer data of unobserved neighbors. How widely should distributed consent be adopted to guarantee connectivity *and* privacy of secure accounts?

The results of our simulations are shown in Fig. 2.3 and Fig. 2.4. We focus on the number of observed nodes of different security levels and on the size of the giant component of unobserved nodes. This last quantity refers to the size of the largest subpopulation of accounts that are not observed and maintain connected pathways of any lengths between one another, thus preserving both their individual privacy and the global connectivity objective of the social network. In classic percolation theory, only one giant component can span the entire system [59], meaning only one subset of nodes can scale with the total size of the social network. Yet, from recent results on network observability [3], we also know that giant *observed* components can co-exist with giant *unobserved* components. This is where distributed consent can play a large role: Even if a third-party surveillance system scales with the size of the social network, it is theoretically possible for accounts to maintain their individual privacy and global connectivity at scale.

Figure 2.3 shows the results of our model in populations facing either a very strong attack (1% of compromised accounts, chosen to observe almost the full population) or a more modest attack (0.05% of compromised accounts, chosen for about 50% observation). Against a strong attack, we find that while extremely low adoption levels of distributed consent have little impact on the observability of the system, moderate adoption (roughly 1 in 5 users) can lead to a transition where observability now drops sharply with the adoption of distributed consent; see Fig. 2.3(a, e). There are few unobserved nodes at a low adoption rate of distributed consent. All are mostly disconnected from each other and therefore observable through compromised

neighbors. At higher levels of adoption rate, the system transitions to an unobservable and connected phase where privacy can co-exist with connectedness and information flow; see Fig. 2.3(b, f). With large-scale adoption of distributed consent (say one-third of users), we find that close to half of all accounts are now protected even against very strong attacks, even if their privacy settings only prevent about 22% of data flow around them.

Against the more modest attack, increases in the adoption of distributed consent lead to an increase in smaller but smoother and more reliable protection. With 33% adoption, populations can halve the size of their observed population and conversely double the size of their unobserved component.

To understand these results, notice that any user with privacy settings set to a greater value than the percolation depth will be unobservable. If we had set simulated naive attackers who observe at a depth $L = 1$ only, adopters of distributed consent would have been fully protected. Indeed, users using the security setting N will only share their data with users using settings of $N - 1$ or more, and this statement holds for all N . We thus know that users using setting N will be at least N steps away from users using the lowest setting, which are the only directly observable nodes. Users with security level set to $1 < N < L$ can however be observed indirectly through their relationships. At low levels of adoption of distributed consent, a large amount of luck is required to remain unobservable (e.g., having zero connections with low-security users). At higher levels of adoption, users of distributed consent connect to and therefore protect, one another. However, these connections are localized and do not spread throughout the entire system. We find that when roughly 25% of nodes with a taste for privacy adopt distributed consent, a large macroscopic component

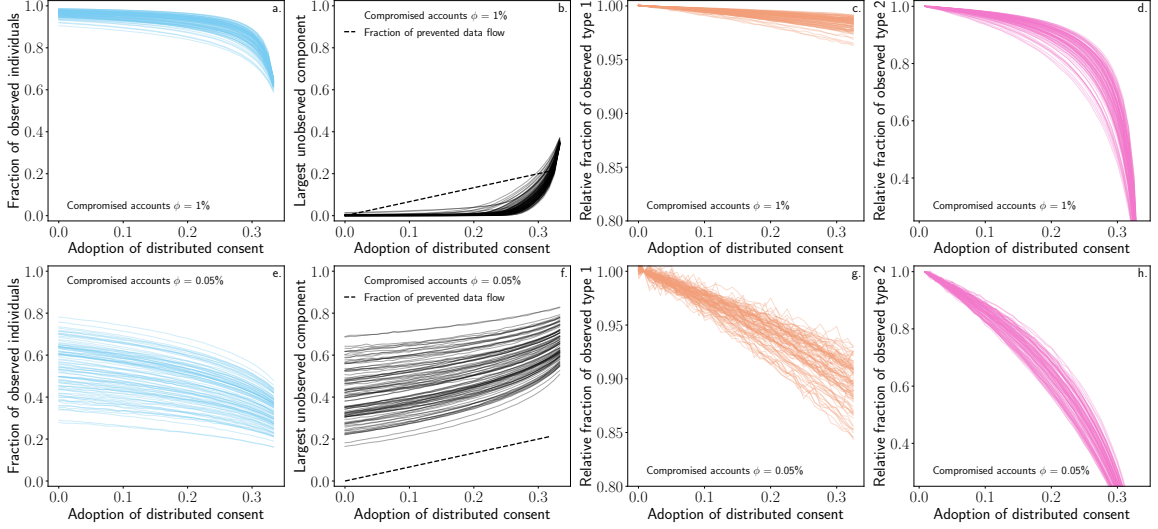


Figure 2.3: We use the anonymized Facebook100 dataset [62]. We assume that one-third of the population has a taste for privacy [35], split between security options 1 and 2 (i.e., classic or distributed consent) according to the adoption rate of distributed consent shown on the horizontal axis. At the same time, the remaining two-thirds will use the default setting with the lowest security, option 0. We set 1% (top row) or 0.05% (bottom row) of accounts with security option 0 to be directly observable by a third-party app, which can also observe neighbors up to two hops away in the network. These values are chosen to model attacks that observe nearly the entire population (top row) or about a half of it (bottom row). We then vary the adoption rate and measure the total fraction of observed accounts (blue curves), the relative size of the largest unobserved connected component (black curves), and the fraction of observed individuals with security option 1 (orange curves) or two (pink curves).

of connected unobservable nodes emerge even against the strongest attack. This component reflects a parallel, protected community that is unobservable but still connected to the rest of the social network.

Even though a phase transition in connected unobservable nodes occurs at a fairly low level of distributed consent adoption, these nodes provide secondary protection to other users. The pervasive adoption of group consent is required to fully protect a network. Again, a single observable neighbor is all it takes for one vulnerable node to be indirectly observed. Because of this and because of the density of most

online networks platforms, it is extremely hard to completely protect vulnerable nodes even if distributed consent provides some secondary protection to all nodes. We thus see the co-existence of both observed and unobserved connected components at the medium adoption level of distributed consent. Interestingly, these components are interconnected, with data flowing both ways across observable and unobservable components, yet the users in the latter remain fully protected from statistical inference of their data.

Importantly, the macroscopic but unobservable component that we see emerge with increased adoption of distributed consent does not only contain adopters of distributed consent. Early adopters of distributed consent provide some low amount of *herd privacy* to the population, protecting otherwise vulnerable users; see Fig. 2.3(c, g). Users with lower privacy settings can thus also benefit since the adoption of distributed consent in one’s neighborhood reduces the probability that one of their neighbors is directly or indirectly observed, thereby reducing the probability that they are themselves observed. However, as long as a majority of users rely on the default lax security settings, this effect will be limited as a single compromised neighbor is sufficient to observe a node. However, we do find much stronger herd privacy effects against more moderate attacks. Moreover, similar effects provide non-linear returns on relative protection of distributed consent users, Fig. 2.3(d, h), consistent with our previous observation of the emergence of an unobserved component.

Finally, in Fig. 2.4, we reproduce the large attack against a population with a stronger taste for privacy and compare our previous results with those obtained by halving the fraction of unprotected nodes (two thirds to one third). Unsurprisingly, the stronger the taste for privacy, the stronger the effects of distributed consent; both

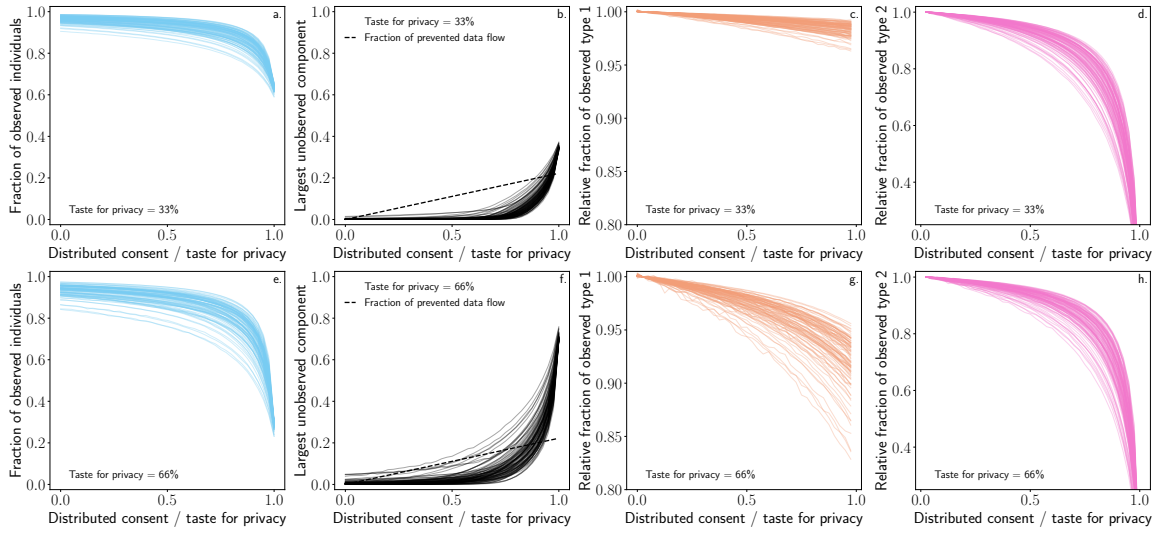


Figure 2.4: We reproduce results from Fig. 2.3 but using populations with different taste for privacy. The results in the top row of this figure match the top row of Fig. 2.3 but are now shown as a function of the nodes with a taste for privacy (security level greater than 0) that opt for distributed consent (security level greater than 1). We then change the fraction of the population with a taste for privacy from 33% (top row) to 66% (bottom row). Qualitatively, the results are very similar. Therefore, the key quantity that drives the macroscopic effects of distributed consent is not its total adoption but its relative adoption within individuals that cannot be directly observed.

at the macroscopic level (the fraction of observed nodes and the size of the unobserved component in panels a, b, e, and f) and at the microscopic level (herd privacy effects shown in panels c, d, g, and h). To make the comparison easier, we plot all quantities against the fraction of users with a taste for privacy (any security level other than the lowest) that opt for distributed consent. This ratio collapses all results on similar curves and therefore appears to be the critical quantity in determining the privacy level of a population.

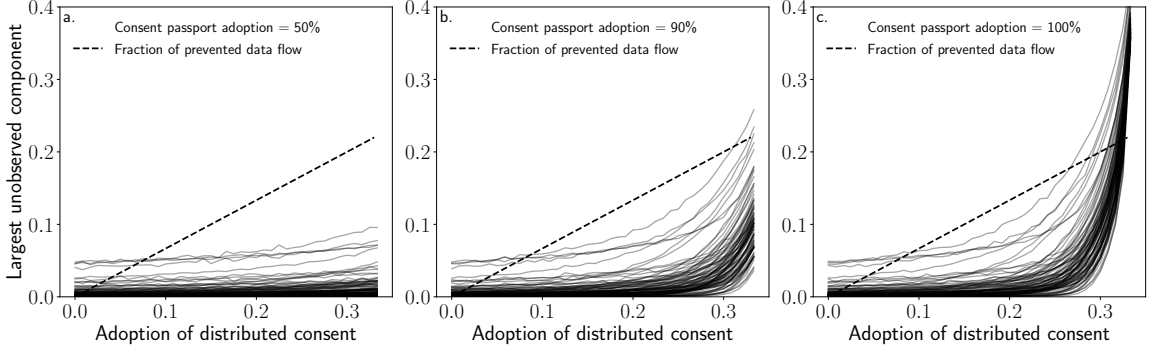


Figure 2.5: We again use the anonymized Facebook100 dataset [62]. We now create a multilayer network by doubling the original data, mimicking a two-platform ecosystem. We use the same parameters as in Fig., 2.3 assuming that for each platform, only one-third of accounts have changed their default security setting to options 1 or 2 (i.e., classic or distributed consent) according to the adoption rate of distributed consent shown on the horizontal axis. The remaining two-thirds will use the default setting with the lowest security, option 0. To account for the doubled network density and the fact that users can be observed on either platform, we now set 0.25% of accounts with security option 0 to be directly observable by a third-party app that can also observe neighbors up to two hops away through any layer of the network. We then vary the adoption rate of distributed consent (horizontal axis) and, within that subset of the population, vary the adoption of consent passport (different panels). In the resulting systems, we measure the relative size of the largest unobserved connected component. Randomly protecting users on a single platform does not protect anyone if spyware can jump network layers, but distributed consent coordinated across platforms through a consent passport can restore our ability to create an unobservable component within the multilayer ecosystem.

2.2.4 A MODEL OF COORDINATED CONSENT ACROSS PLATFORMS

Currently, there is a David and Goliath problem [37]; the inequality of knowledge and power between the user and the data collector functionally takes away the individual’s ability to control their personal information realistically. Part of the problem we believe is that the terms of the agreement are predicated solely on the platform’s norms. Users do not have the option to opt-out and pay for the service in exchange

for limiting information flow for most online social networks. There is a high cost on the user’s side to read and understand all ToS agreements presented to them [41].

There is little to no power on the part of the individual to negotiate the ToS with tech companies. Online consent in its current state creates an unfortunate social optimization problem, where the user must choose between the pressures of disclosing too much personal information (being digitally crowded) and being socially isolated [4, 8]. Moreover, this also implies that most platforms have little power to change the digital ecosystem on their own if users can be exposed to other platforms with laxer security and privacy policies. Multiple platforms create a multilayer network where information flows through social connections and across platforms through different layers of the network. This complex network structure exposes users to whichever platform offers them the least privacy.

Our second model focuses on coordinated consent across platforms. Inspired by previous work on automated ToS tools [36, 48, 52, 26, 7, 46, 42], we envision a *consent passport* model where instead of relying on every platform to offer advanced privacy settings and therefore asking users to adjust their settings on every platform individually, users could use a consent passport stating their desired privacy settings before they join a platform. This could take the form of a key enabling a user to set their privacy baseline criteria based on their taste for privacy. The key will act to shift the burden [15, 25, 57] away from the user to read and understand the legalese of the platform’s privacy policy and ToS. The consent key would restrict login and present a warning [12] when the user attempts to enter sites that do not meet a minimum privacy criteria in their ToS and privacy settings. This key is intended for users who have a taste for privacy but do not want to fall prey to consent

desensitization. The consent passport will need to be dynamic in nature so that users can remain autonomous in their decision-making and easily opt to enter a platform with discordant privacy settings if they trust the site or decide that the privacy cost is worth the risk. Importantly, this ensures that a given user’s security settings will be coordinated across platforms, which might be the only way to confidently protect them in a complex multilayer ecosystem.

To include a consent passport within our simulation model, we turn our network data into a multilayer infrastructure by duplicating the Facebook networks’ structure to represent two platforms where users are randomly assigned security setting independently on each platform. However, a subset of users adopts a consent passport which guarantees that they will follow our previous model of distributed consent on *both* platforms. We again use depth- L percolation with $L = 2$ in the resulting systems to simulate a third party’s ability to observe network users. As a worst-case scenario, we allow the depth- L percolation process to be able to jump between layers of the networks freely, e.g., observing a Facebook neighbor of an individual directly observed through Instagram. As detailed in our Methods section, we classify any user as observed if they are observed on either platform.

Figure 2.5 shows that distributed consent alone cannot protect you if your security settings are not coordinated. Indeed, we parametrize the system such that the networks are roughly equally observable. However, there is no emergence of a giant unobservable component even with medium adoption of consent passports (50% adoption among distributed consent users, left panel). It is only at very high adoption of consent passports that we start seeing a non-linear benefit in unobservability (90% adoption, middle panel) and only at near-perfect adoption that we protect more

users than we prevent data flow (95% and above, right panel). These results demonstrate that multi-platform ecosystems are a much more complex beast to protect; users participating in multiple platforms are only as secure as their weakest security settings. Therefore, coordinating privacy settings across platforms is a critical part of the solution. We discuss this problem further in the next section.

While our current results illustrate how mathematical and computational models can contribute to the study of consent and privacy policies, it is critical to keep all of their assumptions and approximations in mind before drawing conclusions about real-world applications. In the case of our model, we have assumed that all users generate and share similar amounts of data, that all users with a given security setting areas are susceptible to privacy breaches, and that security settings are uncorrelated with the connectivity in the network. In practice, one might imagine that high-profile accounts tend to have higher security settings but that they might also face more frequent attacks. Accounting for correlations between network structures, taste for privacy, and susceptibility to breaches could be included in the model, but these mechanisms first require further empirical studies.

2.3 DISCUSSION

2.3.1 IMPLICATIONS

In recent real-world incidents, attackers reportedly leveraged online social networks to target groups of people. In 2020, BBC News reported that a foreign intelligence agent allegedly used LinkedIn, a prominent online social network, to locate and befriend “former US government and military employees” [50]. Additionally, the same BBC

article reported that Germany’s intelligence agency stated that foreign agents “used LinkedIn to target at least 10,000 Germans” in 2017. And another example comes to mind: the Cambridge Analytica case [66], in which political entities made attempts to sway the political stance of groups of people via Facebook, another large online social network.

In view of these real-world examples, the need for a model of distributed consent, as presented in this Chapter, becomes more apparent. Although the proposed model would not completely stop the attackers, it offers a better level of protection to users of online social accounts than the status quo. In other words, the broad reach of attackers within the threat model presented previously could be restricted with the deployment of the model of distributed consent. We hope that online social network platforms will consider and adopt models where users can have more power to coordinate their privacy with their network neighbors and across platforms. We also hope that policymakers would actively push for adopting similar models to help make online social networks safer for all users.

2.3.2 CONCLUSION

Altogether, in this work, we provided a philosophical critique of individual consent in the context of data transactions and used a modeling framework to suggest potential solutions.

As part of our philosophical critique, we listed four criteria for the legitimacy of informed consent. We argued that none of the four criteria are met by individual consent within online media’s complex ecosystem. Further, a fundamental problem is that if personal data are distributed across individuals, so should be their consent.

Our results based on computational models and simulations suggest that even the simplest implementation of distributed consent could allow users to protect themselves and the flow of their data in the network. They do so by consenting to share their data conditionally on the consent or security settings of their contacts, thereby not sharing their data with users who might, in turn, make them available to third parties. This simple condition allows users to authorize a specific course of action for their own personal data (criterion 4).

While this protection disconnects them from some other users, only a relatively low level of adoption of distributed consent is required to create a connected macroscopic sub-system within existing online network platforms. This sub-system consists of different individuals, including some that are granted secondary protection despite their low-security settings and remain connected to the rest of the system such that information still flows throughout the entire population of users. Via this protected sub-system, distributed consent removes the *de facto* coercion (criterion 2) involved in forcing individuals to choose between relinquishing control of their data or simply not participating in a platform.

Beyond the actual protection mechanism, this new consent model may also have interesting behavioral impacts on the users. Exposing users to this type of coordinated privacy setting might prompt them to reflect on their personal data's distributed nature and its flow through online media. This realization may encourage users to openly voice their social boundaries to their social network or restrict sending sensitive information to social neighbors who do not share their taste for privacy [35]. Imagine a user publishing a post to their social network before enacting the new privacy settings, urging those who want to remain connected to change their settings as well.

Beyond the utility of limiting the social network’s observability, this measure could also serve as an important educational tool on the interconnectedness of personal data (criterion 1).

In a modest form, distributed consent could allow concerned users to protect themselves without entirely leaving a platform. It would also let platforms maintain a large critical mass of observable users that chose to remain vulnerable and who are not granted sufficient protection through their contacts.

That being said, legitimate consent criterion 1 (understanding the consent agreement) and criterion 3 (or consent fatigue) remains an issue that will need additional consideration in future work. An important caveat is that a useful implementation of distributed consent might require platforms to provide additional education regarding data privacy. Regarding criteria 1 and 3, the consent passport attempts to shift the agreement’s burden to platforms rather than users. In doing so, it may provide additional protection in complex multi-platform ecosystems. There are many types of privacy violations that are not solved by distributed consent. These data are still leaky; individual users can still aggregate information about their neighbors without their explicit consent. Finally, while the distributed consent model goes beyond the strict individuality of the traditional privacy model, it does so modestly; it still models the agents, choices, and values as fundamentally individual. Obviously, there is no silver bullet to solve this multi-scale complex problem; data privacy is a significant societal issue with multi-level interdependencies that must be considered thoughtfully and ethically. Much work remains to be done in this area.

Future work should extend to look at possible collective behavior around the adoption of new consent models. Indeed, the greatest hurdle to herd-like immunity

against network observability is our assumption that only one-third of the population has a taste for privacy such that two-thirds of users will never deviate from the default lax security settings. Users signaling their adoption of distributed consent and potentially influencing their network neighbors to do the same could then spark a contagious taste for privacy whose co-evolution with observability could be modeled using tools from network epidemiology [47].

Beyond new notions of consent, effective data privacy measures will need to take a systems-level approach and integrate a mechanism for distributed moral responsibility [22] that will simultaneously involve both top-down and bottom-up interventions. Doing so will involve a synergy between increased governmental and professional regulation, technological intervention, distributed consent, and citizens’ empowerment. Increasing data privacy and protection is not only an essential public service but a democratic imperative [23, 51, 20]. Access to data privacy and protection is a growing global issue [40], and it must be investigated through further multidisciplinary collaboration.

2.4 RESEARCH METHODS

2.4.1 DATA.

We use network data from the anonymous Facebook100 [62] data set without any associated metadata and for the sole purpose of having realistic network structures from a social media platform. The original data set presents 100 complete and independent networks of Facebook “friendships” from 100 American colleges and universities collected as a single-day snapshot in September 2005. Figures 2.3 and 2.5 show simulations of our models indepen-

dently on the 95 Facebook networks with more than 2000 nodes, showing the individual averages obtained from each set of parameters on each network.

2.4.2 OBSERVABILITY MODEL.

Our observability model runs on a directed (or undirected [3]) network of potential data flow where a link from i to j means that user j receives data from user i . We simulate an observability process by selecting a fraction φ of users whom a third party directly observes. For an observability process of depth $L = 0$, the simulation is now over, and a fraction φ of users have been observed. For an observability process of depth $L = 1$, all currently unobserved users whose data are received by directly observed users are now also observed (call those users the first generation of indirectly observed users). For an observability process of depth $L = 2$, all currently unobserved users whose data are received by the first generation of indirectly observed users are now also observed. While the model can be extended to any depth l , Figs. 2.3 and 2.5 both use $L = 2$. In these figures, we measure the fraction of observed individuals (directly and indirectly observed), the largest component of unobserved individuals connected through uninterrupted data flow, and the fraction of observed individuals with a given security setting.

2.4.3 DISTRIBUTED CONSENT MODEL.

Our distributed consent passport constrains data flow in social media networks and the possibility of users being directly or indirectly observed. We define a general distributed consent model but implement it using only three distinct security settings: Users at level 0 share their data with all network neighbors and can be directly observed by a third party; users at level 1 share their data with all network neighbors but can *not* be directly observed by a third party; users at level 2 only share their data with neighbors at level 1 or 2 and can

not be directly observed by a third party. Security levels are randomly assigned to nodes in a network (uniformly at random) at the start of every run of the model with the following probabilities given an adoption frequency x of distributed consent: $2/3$ of users are assigned to level 0, $1/3 - x$ are assigned to level 1, and x are assigned to level 2. In Fig. 2.3, a fraction $\varphi = 1\%$ of users at security level 0 are then directly observed, and the observability model then runs as defined above but with the directionality of data flow limited by security level 2.

2.4.4 CONSENT PASSPORT MODEL.

We extend the distributed consent passport to a general multilayer network where users are part of multiple platforms. In Fig. 2.5 we do this by doubling the original network to obtain a two-platform system where social connections exist on two platforms at once. Given an adoption frequency y of consent passport, we force a fraction y of the fraction x of adopters of security level 2 to have level 2 on both platforms. On both platforms independently, we then distribute uniformly at random the remaining $(1 - y)x$ fraction of adopters of level 2 without the passport, then the $1/3 - x$ fraction of level 1 users, and the $2/3$ fraction of level 0 users. In Fig. 2.5, we then select $\varphi = 0.25\%$ of users at security level 0 on at least one platform to be directly observed. The observability model then runs as usual but indirectly observing the network neighbors j of observed node i if data flows from j to i on *at least* one platform.

2.4.5 DATA AND CODE AVAILABILITY.

All data and all codes for model implementation and figure replication are available from <https://github.com/antoineallard/distributed-consent>.

BIBLIOGRAPHY

- [1] Acquisti, A., John, L. K., and Loewenstein, G. (2013). What Is Privacy Worth? *The Journal of Legal Studies*, 42(2):249–274.
- [2] Alexander, L. (1996). The Moral Magic of Consent (II). *Legal Theory*, 2(3):165–174.
- [3] Allard, A., Hébert-Dufresne, L., Young, J.-G., and Dubé, L. J. (2014). Coexistence of Phases and the Observability of Random Graphs. *Phys. Rev. E*, 89:022801.
- [4] Altman, I. (1977). Privacy regulation: Culturally universal or culturally specific? *Journal of social issues*, 33(3):66–84.
- [5] Asmolov, G. (2018). The disconnective power of disinformation campaigns. *Journal of International Affairs*, 71(1.5):69–76.
- [6] Bagrow, J. P., Liu, X., and Mitchell, L. (2019). Information Flow Reveals Prediction Limits in Online Social Activity. *Nat. Hum. Behav.*, 3(2):122–128.
- [7] Bannihatti Kumar, V., Iyengar, R., Nisal, N., Feng, Y., Habib, H., Story, P., Cherivirala, S., Hagan, M., Cranor, L., Wilson, S., Schaub, F., and Sadeh, N. (2020). Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020*, page 1943–1954, New York, NY, USA. Association for Computing Machinery.
- [8] Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. *First Monday*, 11i9.
- [9] Beauchamp, T. L. and Faden, R. R. (1986). *A History and Theory of Informed Consent*. Oxford University Press, Oxford, United Kingdom.
- [10] Borgesius, F. J. Z. (2016). Singling out people without knowing their names—Behavioural targeting, pseudonymous data, and the new Data Protection Regulation. *Computer Law & Security Review*, 32(2):256–271.
- [11] Borgesius, F. Z. (2015). Informed consent: We can do better to defend privacy. *IEEE Security & Privacy*, 13(2):103–107.
- [12] Calo, R. (2011). Against notice skepticism in privacy (and elsewhere). *Notre Dame L. Rev.*, 87:1027.
- [13] Cate, F. H. (1999). Principles of internet privacy. *Conn. L. Rev.*, 32:877.
- [14] Cohen, J. E. (2000). Examined Lives: Informational Privacy and the Subject as Object. *Stan. L. Rev.*, 52(5):1373–1438.
- [15] Colnago, J., Feng, Y., Palanivel, T., Pearman, S., Ung, M., Acquisti, A., Cranor, L. F., and Sadeh, N. (2020). Informing the Design of a Personalized Privacy Assistant for the Internet of Things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13, New York, NY, USA. Association for Computing Machinery.
- [16] Crețu, A.-M., Monti, F., Marrone, S., Dong, X., Bronstein, M., and de Montjoye, Y.-A. (2022). Interaction data are identifiable even across long periods of time. *Nature Communications*, 13(1):1–11.
- [17] Custers, B., van Der Hof, S., Schermer, B., Appleby-Arnold, S., and Brockdorff, N. (2013). Informed consent in social media use — the gap between user expectations and

- EU personal data protection law. *SCRIPTed*, 10(3):435–457.
- [18] Dhamija, R., Tygar, J. D., and Hearst, M. (2006). Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 581–590, New York, NY, USA. Association for Computing Machinery.
 - [19] Dougherty, T. (2014). Fickle consent. *Philosophical Studies*, 167(1):25–40.
 - [20] Dutt, R., Deb, A., and Ferrara, E. (2018). “Senator, We Sell Ads”: Analysis of the 2016 Russian Facebook Ads Campaign. In *International conference on intelligent information technologies*, Advances in Data Science. ICIIT 2018, pages 151–168, New York, NY, USA. Springer.
 - [21] Farmer, J. D. and Geanakoplos, J. (2009). Hyperbolic discounting is rational: Valuing the far future with uncertain discount rates. *Cowles Foundation Discussion Paper*, No. 1719.
 - [22] Floridi, L. (2016). Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions. *Philos. Trans. Royal Soc. A*, 374(2083):20160112.
 - [23] Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Soc. Text*, 25/26:56–80.
 - [24] Garcia, D. (2017). Leaking Privacy and Shadow Profiles in Online Social Networks. *Sci. Adv.*, 3:e1701172.
 - [25] Grünwald, E. and Pallas, F. (2021). TILT: A GDPR-Aligned Transparency Information Language and Toolkit for Practical Privacy Engineering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 636–646, New York, NY, USA. Association for Computing Machinery.
 - [26] Guarino, A., Lettieri, N., Malandrino, D., and Zaccagnino, R. (2021). A Machine Learning-Based Approach to Identify Unlawful Practices in Online Terms of Service: Analysis, Implementation and Evaluation. *Neural Comput. Appl.*, 33(24):17569–17587.
 - [27] Halfond, W. G., Viegas, J., Orso, A., et al. (2006). A classification of SQL-injection attacks and countermeasures. In *Proceedings of the IEEE international symposium on secure software engineering*, volume 1, pages 13–15.
 - [28] Hurd, H. M. (1996). The Moral Magic of Consent. *Legal Theory*, 2(2):121–146.
 - [29] Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8):56–59.
 - [30] Jackson, S. J. and Banaszczyk, S. (2016). Digital Standpoints: Debating Gendered Violence and Racial Exclusions in the Feminist Counterpublic. *J. Commun. Inq.*, 40(4):391–407.
 - [31] Jackson, S. J. and Foucault Welles, B. (2015). Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics. *J. Commun.*, 65(6):932–952.
 - [32] Kleinig, J. (1982). The Ethics of Consent. *Can. J. Philos.*, 12(sup1):91–118.
 - [33] Lapowsky, I. (2019). One Man’s Obsessive Fight to Reclaim His Cambridge Analytica Data.
 - [34] Leonard, P. G. (2018). Emerging Concerns for Responsible Data Analytics: Trust, Fairness, Transparency and Discrimination. *SSRN Electronic Journal*, 941:151–168.

- [35] Lewis, K., Kaufman, J., and Christakis, N. (2008). The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network. *J. Comput.-Mediat. Commun.*, 14(1):79–100.
- [36] Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., and Torroni, P. (2019). CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.
- [37] Loos, M. and Luzak, J. (2016). Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers. *Journal of consumer policy*, 39(1):63–90.
- [38] Marwick, A. E. and Boyd, D. (2014). Networked privacy: How teenagers negotiate context in social media. *New media & society*, 16(7):1051–1067.
- [39] Matzner, T. (2014). Why Privacy is Not Enough Privacy in the Context of “Ubiquitous Computing” and “Big Data”. *Journal of Information, Communication and Ethics in Society*, 12(2):93–106.
- [40] Mba, G., Onalapo, J., Stringhini, G., and Cavallaro, L. (2017). Flipping 419 Cybercrime Scams: Targeting the Weak and the Vulnerable. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW ’17 Companion, page 1301–1310, Perth, Australia. International World Wide Web Conferences Steering Committee.
- [41] McDonald, A. M. and Cranor, L. F. (2008). The cost of reading privacy policies. *Isjlp*, 4:543.
- [42] Micklitz, H.-W., Palka, P., and Panagis, Y. (2017). The empire strikes back: digital control of unfair terms of online services. *Journal of consumer policy*, 40(3):367–388.
- [43] Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 79:119.
- [44] Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, Stanford, CA, USA.
- [45] Obar, J. A. and Oeldorf-Hirsch, A. (2017). Clickwrap Impact: Quick-Join Options and Ignoring Privacy and Terms of Service Policies of Social Networking Services. In *Proceedings of the 8th International Conference on Social Media & Society*, SMSociety17, New York, NY, USA. Association for Computing Machinery.
- [46] Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T. B., Russell, N. C., Story, P., Reidenberg, J., and Sadeh, N. (2018). PrivOnto: A semantic framework for the analysis of privacy policies. *Semantic Web*, 9(2):185–203.
- [47] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87(3):925–979.
- [48] Pearson, S. and Tsiavos, P. (2014). Taking the Creative Commons beyond copyright: developing Smart Notices as user centric consent management systems for the cloud. *International Journal of Cloud Computing* 2, 3(1):94–124.
- [49] Petronio, S. (1991). Communication boundary management: A theoretical model of managing disclosure of private information between marital couples. *Communication theory*, 1(4):311–335.
- [50] Ponniah, K. (2020). How a Chinese agent used LinkedIn to hunt for targets.
- [51] Rouvroy, A. and Pouillet, Y. (2009). The Right to Informational Self-Determination and

- the Value of Self-Development: Reassessing the Importance of Privacy for Democracy. In Gutwirth, S., Poulet, Y., De Hert, P., de Terwangne, C., and Nouwt, S., editors, *Reinventing Data Protection?*, pages 45–76. Springer Netherlands, Dordrecht.
- [52] Santos, C., Nouwens, M., Toth, M., Bielova, N., and Roca, V. (2021). Consent Management Platforms Under the GDPR: Processors and/or Controllers? In Gruschka, N., Antunes, L. F. C., Rannenber, K., and Drogkaris, P., editors, *Privacy Technologies and Policy*, pages 47–69, Cham. Springer International Publishing.
- [53] Sarigol, E., Garcia, D., and Schweitzer, F. (2014). Online Privacy as a Collective Phenomenon. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, page 95–106, New York, NY, USA. Association for Computing Machinery.
- [54] Schermer, B. W., Custers, B., and van der Hof, S. (2014). The Crisis of Consent: How Stronger Legal Protection May Lead to Weaker Consent in Data Protection. *Ethics Inf. Technol.*, 16:171–182.
- [55] Schwartz, P. M. (2000). Internet Privacy and the State. *Conn. L. Rev.*, 32(3):815–859.
- [56] Skirpan, M. W., Yeh, T., and Fiesler, C. (2018). What’s at Stake: Characterizing Risk Perceptions of Emerging Technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.
- [57] Sloan, R. H. and Warner, R. (2014). Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L.*, 14:370.
- [58] Solove, D. J. (2004). *The Digital Person: Technology and Privacy in the Information Age*. New York University Press, New York, NY, USA.
- [59] Stauffer, D. and Aharony, A. (2018). *Introduction to percolation theory*. Taylor & Francis, London.
- [60] Stone-Gross, B., Cova, M., Cavallaro, L., Gilbert, B., Szydlowski, M., Kemmerer, R., Kruegel, C., and Vigna, G. (2009). Your Botnet is My Botnet: Analysis of a Botnet Takeover. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, page 635–647, New York, NY, USA. Association for Computing Machinery.
- [61] Thielman, S. (2015). Surveillance reform explainer: can the FBI still listen to my phone calls.
- [62] Traud, A. L., Mucha, P. J., and Porter, M. A. (2012). Social Structure of Facebook Networks. *Physica A*, 391(16):4165–4180.
- [63] Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1):20–36.
- [64] Wang, N., Xu, H., and Grossklags, J. (2011). Third-Party Apps on Facebook: Privacy and the Illusion of Control. In *Proceedings of the 5th ACM Symposium on Computer Human Interaction for Management of Information Technology*, CHIMIT '11, New York, NY, USA. Association for Computing Machinery.
- [65] Westen, P. (2017). *The logic of consent: The diversity and deceptiveness of consent as a defense to criminal conduct*. Routledge, Oxfordshire, England, UK.
- [66] Wylie, C. (2019). *Mindf*ck: Cambridge Analytica and the Plot to Break America*.

- Random House, New York, NY.
- [67] Yang, Y., Wang, J., and Motter, A. E. (2012). Network observability transitions. *Physical Review Letters*, 109(25):258701.
- [68] Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. Public Affairs, Boston, MA, USA.

CHAPTER 3

GROUP ONLINE SAFETY: DIVERSE MISINFORMATION IN ONLINE SOCIAL NETWORKS

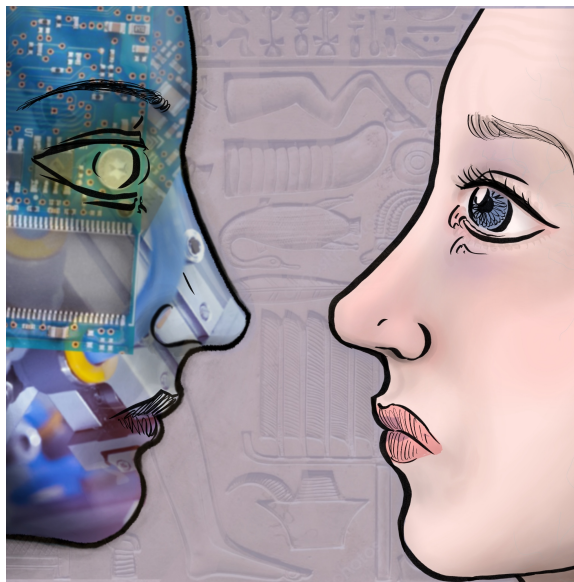


Figure 3.1: Human vs. the Machine by Julia Zimmerman

PREFACE

This chapter will discuss how groups can work together to protect themselves from the harms of online misinformation by using group correction mechanisms. We focus on the relationship between a group’s diversity and their vulnerability to misinformation on online social networks. Additionally, we will examine the ethical concerns surrounding the manipulation of an individual’s likeness through deepfakes. Unlike Chapter 2, this chapter delves into the importance of group diversity in network dynamics. It explores how natural diversity within groups can provide protection rather than relying on new technologies such as distributed consent settings. Material from this chapter has been made publicly available in the following form:

Lovato, J., Hébert-Dufresne, L., St-Onge, J., Harp, R., Salazar Lopez, G., Rogers, S., Ul Haq, I., and Onaolapo, J.. (2023). Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks. arXiv preprint arXiv:2210.10026. (submitted)

ABSTRACT

Social media users are not equally susceptible to all types of misinformation. In this paper, We call “diverse misinformation” the complex relationships between human biases and demographics represented in misinformation. To investigate how users’ biases impact their susceptibility to misinformation and their ability to correct each other, we analyze human classification of computer-generated videos (deepfakes) as a type of diverse misinformation. We chose deepfakes as a case study for three reasons: 1) their classification as misinformation is more objective; 2) we can control the demographics of the personas presented; 3) deepfakes are a real-world concern with associated harms that need to be better understood. Our paper presents an observational survey (N=2,016) where U.S.-based participants (using

a Qualtrics survey panel) are exposed to videos and asked questions about their attributes, not knowing some might be deepfakes. Our analysis investigates the extent to which different users are duped and which perceived demographics of deepfake personas tend to mislead. Importantly, we find that accuracy varies significantly by demographics, and participants are generally better at classifying videos that match them (especially for white participants). We extrapolate from these results to understand the potential population-level impacts of these biases using an idealized mathematical model of the interplay between diverse misinformation and crowd correction. Our model suggests that a diverse set of contacts might provide “herd correction” where friends can protect each other’s blind spots. Altogether, human biases and the attributes of misinformation matter greatly, but having a diverse social group may help reduce susceptibility to misinformation.

3.1 INTRODUCTION

There is a growing body of scholarly work focused on distributed harm in online social networks. These scholarly works focus on a wide range of topics from leaky data [7], and group security and privacy [48] to hate speech [31], misinformation [19] and detection of computer-generated content [33]. Social media users are not all equally susceptible to these harmful forms of content. Our level of vulnerability depends on our own biases. We define “diverse misinformation” as the complex relationships between human biases and demographics represented in misinformation. This paper explores deepfakes as a case study of misinformation to investigate how U.S. social media users’ biases influence their susceptibility to misinformation and their ability to correct each other. We choose deepfakes as a critical example of the possible impacts of diverse misinformation for three reasons: 1) their status of being misinformation is binary; they either are a deepfake or not; 2) the perceived demographic attributes of the persona presented in the videos can be characterized

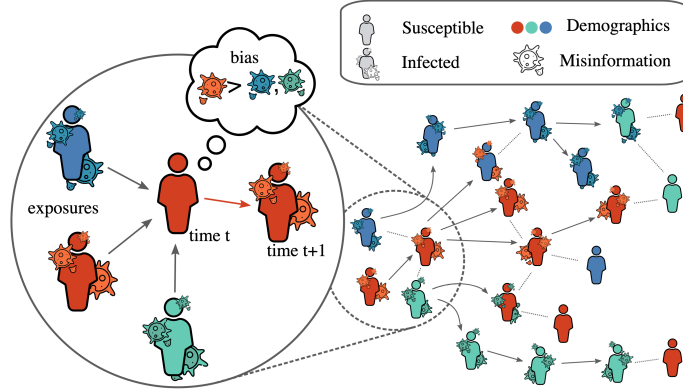


Figure 3.2: Illustration of the problem considered in this work. Populations are made of individuals with diverse demographic features (e.g., age, gender, race; here represented by colors), and misinformation is likewise made of different elements based on the topics they represent (here shown as pathogens). Through their biases, certain individuals are more susceptible to certain kinds of misinformation. The cartoon represents a situation where misinformation is more successful when it matches an individual’s demographic. Red pathogens spread more readily around red users with red neighbors. In reality, the nature of these biases is still unclear, and so are their impacts on online social networks.

by participants; 3) deepfakes are a current real-world concern with associated negative impacts that need to be better understood. Together, this allows us to use deepfakes as a critical case study of diverse misinformation to understand the role individual biases play in disseminating misinformation at scale on social networks and in shaping a population’s ability to self-correct.

We present an empirical survey (N=2,016 using a Qualtrics survey panel [10]) observing what attributes correspond to U.S.-based participants’ ability to detect deepfake videos. Survey participants entered the study under the pretense that they would judge the communication styles of video clips. Our observational study is careful not to prime participants at the time of their viewing video clips so we could gauge their ability to view and judge deepfakes when they were not expecting them (not explicitly knowing if a video is fake or not is meant to emulate what they would experience in an online social media platform). Our survey also investigates the relationship between human participants’ demographics

and their perception of the video person(a)’s features and, ultimately, how this relationship may impact the participant’s ability to detect deepfake content.

Our objective is to evaluate the relationship between classification accuracy and the demographic features of deepfake videos and of survey participants. Further analysis of other surveyed attributes will be explored in future work. We also recognize that data used to train models that create deepfakes may introduce algorithmic biases in the quality of the videos themselves, which could introduce additional biases in the participant’s ability to guess if the video is a deepfake or not. The Facebook Deepfake Detection Challenge dataset that was used to create the videos we use in our survey was created to be balanced in diversity in several axes (gender, skin-tone, age). We suspect that if there are algorithmic-level biases in the model used resulting in better deepfakes for personas of specific demographics, we would expect to see poorer accuracy across the board for all viewer types when classifying these videos. We do see that viewer groups’ accuracy differs based on different deepfake video groups. However, our focus is on the perception of survey participants towards deepfakes’ identity and demographics to capture viewer bias based on their perception rather than the model’s bias and classification of the video persona’s racial, age, and gender identity. Our goal is to focus on viewers and capture what a viewer would experience in the wild (on a social media platform), where a user would be guessing the identity features of the deepfake and then interrogating if the video was real or not with little to no priming.

This paper adopts a multidisciplinary approach to answer these questions and understand their possible impacts. First, we use a survey analysis to explore individual biases related to deepfake detection. There is abundant research suggesting the demographics of observers and observed parties influence the observer’s judgment and sometimes actions toward the observed party [27, 46, 12, 41, 49]. In an effort to avoid assumptions about any demographic group, we chose four specific biases to analyze vis-à-vis deepfakes: **(Question 1) Priming bias:** How much does classification accuracy depend on participants being

Use the following definition of deepfakes to answer the following questions:
 Deepfakes, sometimes referred to as deep learning fakes, are synthetic images or videos in which the original person is replaced with features of another person.

These are more advanced, and thus often more believable, than traditional photoshop methods as they use techniques from deep learning to generate the new visual, copying everything from facial expressions and mannerisms to the audio of a person's voice.

Do you think the primary person in Video #1 was real? (the first one you watched)
 Note: If you are unsure, make your best guess.

☐ Yes, they are real

☐ No, they were fictionally created for this video

Do you think the primary person in Video #2 was real? (the second one you watched)
 Note: If you are unsure, make your best guess.

☐ Yes, they are real

☐ No, they were fictionally created for this video

Figure 3.3: Question where survey participants are asked after the debrief of the survey if they think the videos they watched are real or fake. The performance metric we use to measure participant accuracy is the ratio of the correct guesses to the entire pool of guesses where $\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$.

primed about the potential of a video being fake? Our participants are not primed on the meaning of deepfakes and are not explicitly looking for them. Importantly, we do not explicitly vary the priming of our participants but we compare their accuracy to a previous study with a similar design but primed participants [33]. **(Question 2) Prior knowledge:** Does accuracy depend on how often the viewer uses social media and whether they have previously heard of deepfakes? Here, we ask participants to evaluate their own knowledge and use their personal assessment to answer this research question. **(Question 3) Homophily bias:** Are humans better classifiers of video content if the perceived demographic of the video persona matches their own identity? **(Question 4) Heterophily bias:** Inversely, are humans more accurate if the perceived demographic of the video persona does not match their own? We then use results from the survey to develop an idealized mathematical model to theoretically explore population-level dynamics of diverse misinformation on online social networks. Altogether, this allows us to hypothesize the *mechanisms* and *possible impacts* of diverse misinformation, as illustrated in Fig. 3.2.

Our paper is structured as follows. We outline the harms and ethical concerns of diverse misinformation and deepfakes in Section 4.2. We explore the possible effects through which demographics impact susceptibility to diverse misinformation through our observational study in Section 4.3. We then investigate the network-level dynamics of diverse misinformation using a mathematical model in Section 3.4. We discuss our findings and their implications in Section 4.4. Our full survey methodology can be seen in Section 3.6.1.

3.2 BACKGROUND

3.2.1 BIAS TYPES

It is important to understand human biases as they impact the transmission and correction of misinformation and its potential impacts on polarization and degradation of the epistemic environment [23]. In social networks, it has been shown that there are human tendencies toward homophily bias [22, 42]. Indeed, there are differences in user demographic groups’ abilities to detect deepfakes and misinformation (e.g., age) [54]. Previous work has also shown that biases impact people’s accuracy as an eyewitness through the own-race bias (ORB) phenomenon [13, 15, 51]. It is an open question whether deepfake detection also demonstrated the own-race bias (ORB) phenomenon.

Subsequently, these biases impact how social ties are formed and, ultimately, the shape of the social network. For example, in online social networks, homophily often manifests through triadic closures [45] where friends in social networks tend to form new connections that close triangles or triads. Understanding individuals’ and groups’ biases will help understand the network’s structure and dynamics and how information and misinformation spread on the network depending on its level of diversity. For example, depending on the biases and the node-specific diversity of the connections it forms, one may have a

system that may be more or less susceptible to widespread dissemination as it would in a Mixed Membership Stochastic Block Model (MMSBM) [1]. A Mixed Membership Stochastic Block Model is a Bayesian community detection method that segments communities into blocks but allows community members to mix with other communities. Assumptions in an MMSBM include a list of probabilities that determine the likelihood of communities interacting. We explore these topics in more detail in Section 3.4.

Previous work has demonstrated that homophily bias towards content aligned with one’s political affiliation can impact one’s ability to detect misinformation [73, 16]. Traberg et al. show that political affiliation can impact a person’s ability to detect misinformation about political content [73]. They found that viewers misclassified misinformation as being true more often when the source of information aligned with their political affiliation. Political homophily bias, in this case, made them feel as though the source was more credible than it was.

In this paper, we investigate the accuracy of deepfake detection based on multiple homophily biases in age, gender, and race. We also explore other bias types, such as heterophily bias, priming, and prior knowledge bias impacting deepfake detection.

3.2.2 MISINFORMATION

Misinformation is information that imitates real information but does not reflect the genuine truth [43]. Misinformation has become a widespread societal issue that has drawn considerable recent attention. It circulates physically and virtually on social media sites [78] and interacts with socio-semantic assortativity. In contrast, assortative social clusters will also tend to be semantically homogeneous [60]. For instance, misinformation promoting political ideology might spread more easily in social clusters based on shared demographics. Motivations vary broadly to explain why people disseminate misinformation, which we refer to as

disinformation when specifically intended to deceive. Motivations include 1) purposefully trying to deceive people by seeding distrust in information, 2) believing the information to be accurate and spreading it mistakenly, and 3) spreading misinformation for monetary gain. In this paper, we will primarily focus on deepfakes as misinformation meaning the potential of a deepfake viewer getting duped and sharing a deepfake video. Disinformation is spreading misinformation with the intent to deceive. In this paper, we do not assume that all deepfakes are disinformation since we do not consider the intent of the creator. A deepfake could be made to entertain or showcase technology. We instead focus on deepfakes as misinformation meaning the potential of a deepfake viewer getting duped and sharing a deepfake video, regardless of intent.

There are many contexts where online misinformation is of concern. Examples include: 1) misinformation around political elections and announcements (*political harms*); 2) misinformation on vaccinations during global pandemics (*health-related harms*) [20, 68]; 3) false speculation to disrupt economies or speculative markets [40]; 4) distrust in news media and journalism (*harms to news media*) [19, 58]; 5) false information in critical informational periods such as humanitarian or environmental crises [77], and 6) propagation of hate speech online [31] which spreads harmful false content and stereotypes about groups (*harms related to hate speech*).

Correction of misinformation: There are currently many ways to try to detect and mitigate the harms of misinformation online [79]. On one end of the spectrum are automated detection techniques that focus on the classification of content or on observing anomaly detection in the network structure context of the information or propagation patterns [66, 63]. Conversely, crowd-sourced correction of misinformation leverages other users to reach a consensus or simply estimate the veracity of the content [4, 52, 2]. We will look at the latter form of correction in an online social network to investigate the role group correction plays in slowing the dissemination of diverse misinformation at scale.

Connection with deepfakes: The potential harms of misinformation can be amplified by computer-generated videos used to give fake authority to the information. Imagine, for instance, harmful messages about an epidemic conveyed through the computer-generated persona of a public health official. Unfortunately, deepfake detection remains a challenging problem, and the state-of-the-art techniques currently involve human judgment [33].

3.2.3 DEEPFAKES

Deepfakes are artificial images or videos in which the persona in the video is generated synthetically. Deepfakes can be seen as false depictions of a person(a) that mimics a person(a) but does not reflect the truth. Deepfakes should not be confused with augmented or distorted video content, such as using color filters or digitally-added stickers in a video. Creating a deepfake can involve complex methods such as training artificial neural networks known as generative adversarial networks (GANs) on existing media [71] or simpler techniques such as face mapping. Deepfakes are deceptive tools that have gained attention in recent media for their use of celebrity images and their ability to spread misinformation across online social media platforms [59].

Early deepfakes were easily detectable with the naked eye due to their uncanny visual attributes and movement [53]. However, research and technological developments have improved deepfakes, making them more challenging to detect [19]. There are currently several automated deepfake detection methods [75, 39, 34, 80, 9]. However, they are computationally expensive to deploy at scale. As deepfakes become ubiquitous, it will be necessary for the general audience to identify deepfakes independently during gaps between the development of automated techniques or in environments that are not always monitored by automated detection (or are offline). It will also be important to allow human-aided and human-informed deepfake detection in concert with automated detection techniques.

Several issues currently hinder automated methods: 1) they are computationally expensive; 2) there may be bias in deepfake detection software and training data—credibility assessments, particularly in video content, have been shown to be biased [36]; 3) As we have seen with many cybersecurity issues, there is a “cat-and-mouse” evolution that will leave gaps in detection methodology [64].

Humans may be able to help fill these detection gaps. However, we wonder to what extent human biases impact the efficacy of detecting diverse misinformation. If human-aided deepfake detection becomes a reliable strategy, we need to understand the biases that come with it and what they look like on a large scale and on a network structure. We also posit that insights into human credibility assessments of deepfakes could help develop more lightweight and less computationally expensive automated techniques.

3.2.4 ETHICAL CONSIDERATIONS

As deepfakes improve in quality, the harms of deepfake videos are coming to light [32]. Deepfakes raise several ethical considerations: 1) the evidentiary power of video content in legal frameworks [19, 62, 28]; 2) consent and attribution of the individual(s) depicted in deepfake videos [35]; 3) bias in deepfake detection software and training data [36]; 4) degradation of our epistemic environment, i.e., there is a large-scale disagreement between what community members believe to be real or fake, including an increase in misinformation and distrust [19, 58]; and 5) possible intrinsic wrongs of deepfakes [24].

It is important to understand who gets duped by these videos and how this impacts people’s interaction with any video content. The gap between convincing deepfakes and reliable detection methods could pose harm to democracy, national security, privacy, and legal frameworks [19]. Consequently, additional regulatory and legal frameworks [61] will need to be adopted to protect citizens from harms associated with deepfakes and uphold

the evidentiary power of visual content. False light is a recognized invasion of privacy tort that acknowledges the harms that come when a person has untrue or misleading claims made about them. We suspect that future legal protections against deepfakes might well be grounded in such torts, though establishing these legal protections is not trivial [62, 21].

In the same way, any laws and regulations against deepfake manipulation are only as powerful as our ability to detect the video manipulation and a plaintiff’s ability to show that the defendant created the video to deceive the courts. Other state-level laws [44] have emerged, making specific deepfake manipulations a crime. Still, these efforts are particular in scope, have been criticized as difficult to enforce, possibly an infringement on the constitutional rights to free speech, and are limited to specific regions in the US [5]. To further define the importance of deepfake detection for legal contexts, it is important to determine which attributes drive a jury or judge to detect a deepfake video themselves.

Deepfake harms in the U.S. legal framework have been primarily concerned with the invasion of personal privacy. This calls into question an ethical consideration about the extent to which a person has autonomy and control over their own likeness. The boundaries are fuzzy in this regard. It seems clear that Generative Adversarial Networks (GAN) revenge porn, where one person distributes fake photographic or video representations of another person engaged in sexual activity without the other person’s consent, would violate privacy [70]. However, where does the boundary fall? Does a GAN art video fall into this scope? What about a parody exhibiting a well-known politician or celebrity? There are also questions about the implications for copyright and intellectual property infringement especially when using deepfakes to mimic the likeness as a substitute for their human labor without fair compensation or adequate informed consent.

The ethical implications of deepfake videos can be separated into two main categories: the impacts on our epistemic environment and the moral relationships and obligations people have with others and themselves. Consider first the epistemic environment, which

includes our capacity to take certain representations of the world as true and our taking beliefs and inferences to be appropriately justified. Audio and video are particularly robust and evocative representations of the world. They have long been viewed as possessing more testimonial authority (in the broader, philosophical sense of the phrase) than other representations of the world. This is true in criminal and civil contexts in the United States, where the admissibility of video recordings as evidence in federal trials is specifically singled out in Article X of the Federal Rules of Evidence [69] (State courts have their own rules of evidence, but most states similarly have explicit rules that govern the admissibility of video recordings as evidence). The wide adoption of deepfake technology would strain these rules of evidence; for example, the federal rules of evidence reference examples of handwriting authentication, telephone conversation authentication, and voice authentication but do not explicitly mention video authentication. Furthermore, laws are notorious for lagging behind technological advances [65], which can further complicate and limit how judges and juries can approach the existence of a deepfake video as part of a criminal or civil case.

Moreover, suppose juries are aware of the possibility that a video is a deepfake. That might affect the degree to which they treat the video representation as authoritative, regardless of the video's being admitted as evidence. To that end, it is important to know how reliable people are concerning deepfake identification and what factors might be relevant.

Belief and knowledge transmission often happens in non-regulated environments and is particularly important for democratic societies. A wide prevalence of deepfake videos would harm such belief and knowledge transmission and represent a degradation of our epistemic environment. In addition, legal contexts do not exhaust the epistemic environment; people often take video representations of the world as more truthful or authoritative.

Deepfake videos do not merely degrade the epistemic environment, however, because deepfakes do not merely represent the world; they also represent other agents in the world. As such, deepfakes entail implications for our moral obligations towards one another. Be-

cause one person can create a deepfake video representing another person, the question is raised about whether consent is necessary for such a video to be morally permissible. This is relevant to issues of autonomy, attribution, and intellectual property (e.g., using an actor’s likeness to create an advertisement without their permission or using an artist’s style in an advertisement without attribution or compensation). To what extent, that is, is control over our likeness within the proper scope of our autonomy? For example, some state statutes that outlaw revenge porn do not specify whether the photographic or video representation has to be authentic. Whether people believe the image to be real might be morally significant. However, compare a crude animation of a person engaged in sexual activity with a highly realistic deepfake video of the same activity. The fact that most people disbelieve the crude animation seems morally relevant. Potential harms and moral wrongs of the deepfake video depend on whether the video is viewed as veridical. Moreover, there is every reason to suspect that these social harms are not distributed equally but that more socially vulnerable people are subject to greater social harm.

These considerations make it critical to identify the factors that lead people to classify things as real or fake.

3.2.5 RESEARCH QUESTIONS

Our paper asks four primary research questions regarding how human biases impact deepfake detection. **(Q1) Priming:** How important is it for an observer to know that a video might be fake? **(Q2) Prior knowledge:** How important is it for an observer to know about deepfakes, and how does social media usage affect accuracy? **(Q3-Q4) Homophily and heterophily biases:** Are participants more accurate at classifying videos whose persona they perceive to match (homophily) or mismatch (heterophily) their own demographic attributes in age, gender, and race?

To address our four research questions, we designed an IRB-approved survey (N=2,016) using video clips from the Deepfake Detection Challenge (DFDC) Preview Dataset [26, 25]. Our survey participants entered the study under the pretense that they would judge the communication styles of video clips (they were not explicitly looking for deepfake videos in order to emulate the uncertainty they would experience in an online social network). After the consent process, survey participants were asked to watch two 10-second video clips. After each video, our questionnaire asked participants to rate the pleasantness of particular features (e.g., tone, gaze, likability, content) of the video on a 5-point Likert scale. They are also asked to state their perception of the person in the video by guessing the video persona’s gender identity, age, and whether they are white or a person of color.

After viewing both videos and completing the related questionnaire, the participants were then debriefed on the deception of the survey, given an overview of what deepfakes are, and then asked if they thought the videos they just watched were real or fake. After the debrief questions, we collected information on the participants’ backgrounds, demographics, and expressions of identity.

Our project investigates features or pairings of features (of the viewer or the person(a) in the video) that are the most important ones needed to determine an observer’s ability to detect deepfake videos and avoid being duped. Conversely, we also ask what pairings of features (of the viewer or the person(a) in the video) are important to determine an observer’s likelihood of being duped by a deepfake video.

Our null hypothesis asserts that none of the features or pairing of features we measure in our survey produce biases that are significantly important in a user being duped by a deepfake video or being able to detect a deepfake video. We then measure our confidence in rejecting this null hypothesis by measuring a bootstrap credibility interval for a difference in means test between the accuracy of two populations (comparing Matthew’s Correlation Coefficient scores). In all tests, we use 10,000 bootstrap samples and consider a comparison

significant if the difference is observed in 95% of samples (i.e., in 9,500 pairs). With this method, our paper aims to better understand how potential social biases affect our ability to detect misinformation.

3.3 RESULTS

Our results can be summarized as follows. (Q1) If not primed, our survey participants are not particularly accurate at detecting deepfakes (accuracy = 51%, essentially a coin toss). (Q3-Q4) Accuracy varies by some participants’ demographics and perceived demographics of video persona. In general, participants were better at classifying videos that they perceived as matching their own demographic.

Our results show that of the 4,032 total videos watched, 49% were deepfakes, and 1,429 of those successfully duped our survey participants. A confusion matrix showing the True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) rates can be seen in Fig. 3.4. We also note that the overall accuracy rate (where accuracy = $(TP+TN)/(TP+FP+FN+TN)$) of our participants was 51%. This translates to an overall Matthew’s Correlation Coefficient (MCC) score of 0.334 for all participant’s guesses vs. actual states of the videos. MCC [50, 14] is a simple binary correlation between the ground truth and the participant’s guess. Regardless of the metric, our participants performed barely better than a simple coin flip (credibility 94%). All summary statistics for our study and all confusion matrices for our primary and secondary demographic groups can be found in *Appendix SI2* and *Appendix SI3* respectively [47]. Next, we explain our findings in detail.

Q1: Priming bias: Our results suggest that priming bias may play a role in a user’s ability to detect deepfakes. Compared with notable prior works [33, 6, 18], our users were not explicitly told to look for deepfake videos while viewing the video content. Our survey

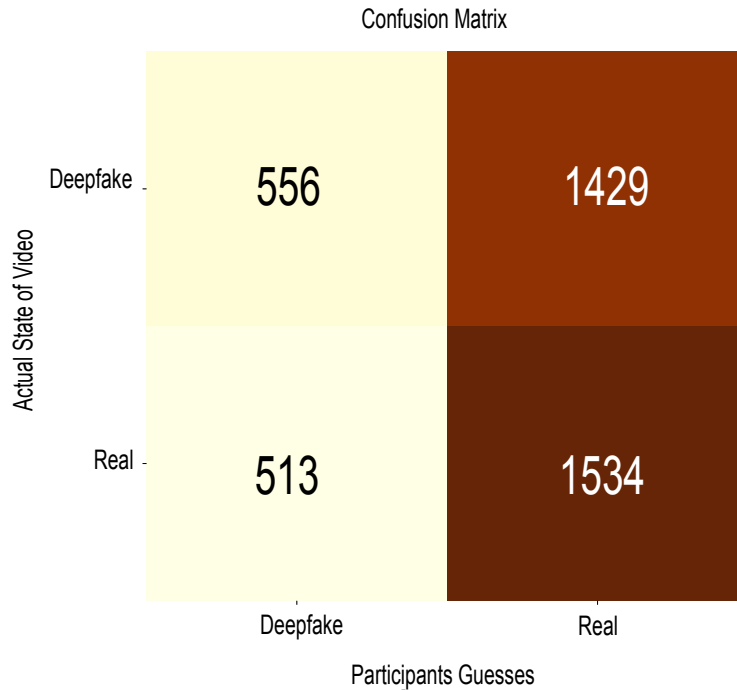


Figure 3.4: A confusion matrix showing our participant guesses about the state of the videos vs. the real state of the video. Participants in our study watched two videos followed by a questionnaire and a debriefing on deepfakes. They were then asked to guess whether the videos were deepfakes or real. Out of 2,016 participants and 4,032 total videos watched, 1,429 videos duped our participants, meaning they saw a fake video they thought was real. The top right panel shows the participants who were duped by deepfakes. The confusion matrix is defined by the number of true positives in the top left, false negatives in the top right, false positives in the bottom left, and true negatives in the bottom right.

takers participated in a deceptive study where they thought they answered questions about effective communication styles. They were debriefed only after the survey was completed and then asked if they thought the video clips were real or fake. Priming, on the contrary, would mean that when the user watched the two video clips, they would be explicitly looking for deepfakes.

Other works measured primed human deepfake detectors to compare them to machines and humans with machine aid. For example, in a study by Azur et al. [6], humans were deployed as deepfake evaluators. The participants were explicitly asked to view images and

look for fake images. Participants in the study were also required to pass a qualification test where they needed to correctly classify 65% real and fake images to participate in the study [6] fully. In a more recent study by Groh et al. [33], participants viewed video clips from the Facebook Deepfake Detection Challenge Dataset (DFCD), as in our study. They were asked to explicitly look for deepfake videos and then tested regarding how this compared to machines alone and machines aided by humans. Groh et al. reported an accuracy score of 66% for primed humans, 73% for a primed human with a machine helper, and 65% for the machine alone. In another study, Chen et al. [18] also showed that hybrid systems that combine crowd-nominated and machine-extracted features outperform humans and machines alone.

Our results show that the non-primed participants were only 51% accurate at detecting if a video was real or fake. One important takeaway from previous studies is that human-machine cooperation provides the best accuracy scores. The previously mentioned prior studies were performed with primed participants. We believe a more realistic reflection of how deepfake encounters would occur “in the wild” would be with observers who were not explicitly seeking out deepfakes. Future work is needed to investigate how non-primed human deepfake detectors perform when aided by machines.

Type	Accuracy
Non-Primed Human	51%
Primed Human [33]	66%
Machine Only [33]	65%
Primed Human with Machine Helper [33]	73%

Table 3.1: Accuracy scores of machine deepfake detectors versus primed human deepfake detectors versus non-primed human deepfake detectors. We compare primed and non-primed survey participants and their abilities to detect deepfakes. Our results show that humans who are not primed to find deepfakes reach an accuracy of 51% (MCC Score 0.334, 35% participants duped). The accuracy scores of our survey participants are 15% points below those of primed human deepfake detectors from previous work [33].

Q2 Prior knowledge effect: We also ask if participants are better at detecting a

deepfake if they have prior knowledge about deepfakes or more exposure to social media.

Our results show no significant results due to prior knowledge or frequent social media usage and therefore we cannot draw any conclusions as to the compatibility with our data for this particular question given that our credibility score for this metric fell below 95% credibility.

We see that participants who are frequent social media users (i.e., use social media once a week or more) had a higher MCC score ($MCC = 0.0396$) than those who used social media less frequently ($MCC = -0.0110$). Participants who knew what a deepfake was before taking the survey ($MCC = 0.0790$) also had a higher score than those unfamiliar with deepfakes ($MCC = 0.0175$). However, in both comparisons, the difference is not deemed significant to our data given that bootstrap samples reject the null only with 83% and 94% credibility, respectively.

Q3-4: Homophily versus heterophily bias: We then focus on the potential impacts of heterophily and homophily biases on a participant’s ability to detect if a video is real or a deepfake. We look at the Matthew’s Correlation Coefficients (MCC) for all user groups and compare their guesses on videos that either match their identity (homophily) or do not match their own identity (heterophily). Results of these MCC scores related to homophily and heterophily bias can be seen in Fig. 3.5.

One of our demographic subgroups, namely white participants, was significantly more accurate when guessing the state of video personas that match their own demographic. We test our null hypothesis by comparing the answers given by a certain demographic of participants when looking at videos that match and do not match their identity. In doing so, we only observed a significant homophily bias for white participants, which can be seen in Table 3.2. In that case, the null hypothesis that they are equally accurate on videos of white personas and personas of color fall outside of a 99% credibility interval, which can be seen in Fig. 3.5.

We further break down this potential bias in two dimensions (overall demographic classes of the participants and video persona) in Table 3.2. We then see more significant results. Here we compare subgroups of our survey participants (e.g., male vs. female viewers, persons of color vs. white viewers, and young vs. old viewers) to see which groups perform better when watching videos of a specific sub-type (e.g., videos of men, videos of women, videos of persons of color, videos of white people, videos of young people, and videos of old people).

By gender, we find that male participants are significantly more accurate than female participants in when watching videos with a male persona. Similarly, by race, we find that participants of color are significantly more accurate than white participants when watching videos that feature a persona who is a person of color. Lastly, young participants have the highest accuracy score overall for any of our demographic subgroups. Of course, these results may be confounded with other factors, such as social media usage, which can be more prominent in one group (e.g., young participants) than another (e.g., older participants). More work needs to be done to understand the mechanisms behind our results.

In summary, our main significant results (rejecting the null hypothesis with 95% credibility) on human biases in deepfake detection are as follows.

- White participants show a significant homophily bias, meaning they are more accurate at classifying videos of white personas than they are at classifying videos of personas of color.
- When viewing videos of male personas, male participants in our survey are significantly more accurate than female participants.
- When viewing videos of personas of color, participants of color are significantly more accurate than white participants.
- When viewing videos of young personas, participants between the ages of 18-29 are

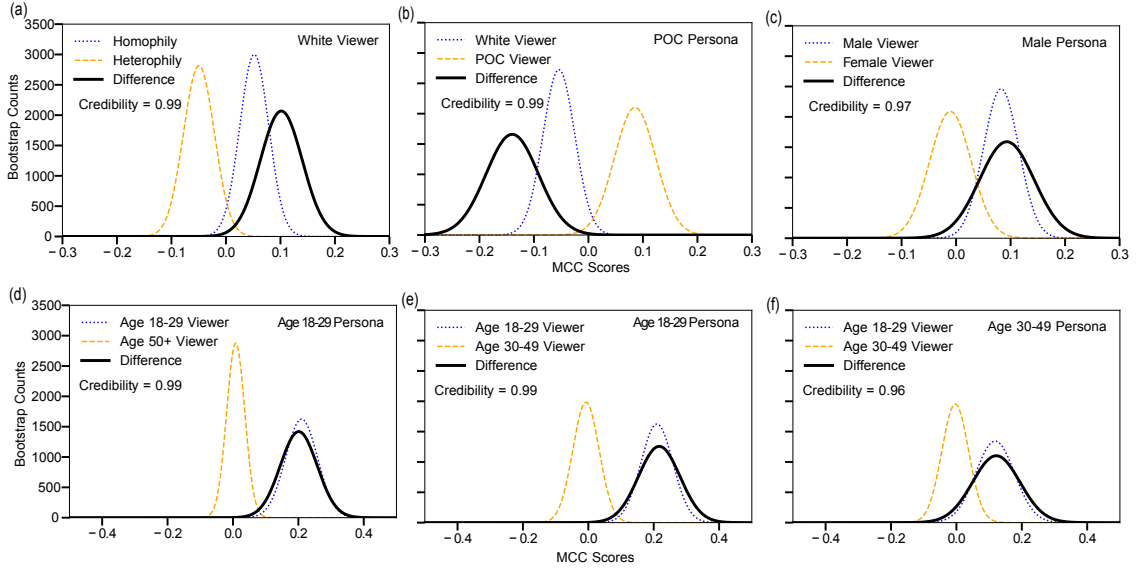


Figure 3.5: Bootstrap MCC samples from observed confusion matrices to compare MCC scores of user and video feature pairs. Our significant results are as follows. **(a)** White users were found to have a homophily bias and are better at classifying videos of a persona they perceive as white. **(b)** Consequently, videos of personas of color are more accurately classified by participants of color. **(c)** Similarly, videos of male personas are better identified by male users. Across multiple age classes, we find that participants aged 18-28 years old are better at identifying videos that match them than older participants (panels **(d)** and **(e)**) or even better at classifying videos of persona perceived as 30-49 years old than participants from that same demographic (panel **(f)**).

Video/User Demographics	MCC of User	N	Credibility
White Viewer/Homophilic Videos	0.0518	1372	0.99
White Viewer/Heterophilic Videos	-0.0498	1224	
Male Persona/Male Viewer	0.0827	918	0.97
Male Persona/Female Viewer	0.0567	1188	
POC Persona/POC Viewer	0.0858	708	0.99
POC Persona/White Viewer	-0.0544	1143	
Age 18-29 Persona/Age 18-29 Viewer	0.1475	303	0.99
Age 18-29 Persona/Age 30-49 Viewer	0.0354	264	
Age 18-29 Persona/Age 50+ Viewer	-0.0198	694	
Age 30-49 Persona/Age 18-29 Viewer	0.1168	282	0.96
Age 30-49 Persona/Age 30-49 Viewer	-0.0037	607	

Table 3.2: Significant (above a 95% credibility) categories of interest. Matthew’s Correlation Coefficient (MCC) is a correlation measure between a participant’s guess about the video being real or fake (0,1) versus the actual state of the video (real 0, fake 1). We use a bootstrap approach to then test the credibility of a superior accuracy (frequency of bootstrap pairs that produce a superior accuracy). Bootstrap distributions can be seen in Fig. 3.5. Note that “heterophilic videos” (row 2) include video personas which the viewers classified as “maybe POC” or “uncertain,” while POC persona (rows 5 and 6) did not.

significantly more accurate than participants above the age of 30; surprisingly, participants aged 18-29 are also more accurate than participants aged 30-49 even when viewing videos of personas aged 30-49.

3.4 MATHEMATICAL MODEL

In essence, the results shown in Table 3.2 illustrate how there is no single demographic class of participants that excels at classifying all demographics of video persona. Different participants can have different weaknesses. For example, a white male participant may be more accurate at classifying white personas than a female participant of color, but the female participant of color may be more accurate on videos of personas of colors. To consider the implications of this simple result, we take inspiration from our findings and formulate an

idealized mathematical model of misinformation to better understand how deepfakes spread on social networks with diverse users and misinformation.

Models of misinformation spread often draw from epidemiological models of infectious diseases. This approach tracks how an item of fake news or a deepfake might spread, like a virus, from one individual to its susceptible network neighbors, duping them such that they can further spread misinformation [74]. However, unlike infectious diseases, an individual’s recovery does not occur on its own through its immune system. Instead, duped individuals require fact-checking or correction from their susceptible neighbors to return to their susceptible state [67]. We integrate these mechanisms with the core finding of our study: Not all classes of individuals are equally susceptible to misinformation.

Specifically, we build a simple model to investigate two mechanisms where we expect an impact of heterogeneous susceptibility to misinformation based on individual demographic characteristics. (1) Individuals with increased susceptibility should be preferentially duped, but this effect exists only if misinformation can spread (above a certain contagion threshold) but not saturate the population (below certain transmissibility). (2) Individuals with a diverse neighborhood are also more likely to have friends who can correct them should they be duped by misinformation.

Our model uses a network with a heterogeneous degree distribution and a structure inspired by the mixed-membership stochastic block model [1]. This stylized structure captures the known heterogeneity of real networks and its modular structure of echo chambers and bridge nodes with diverse neighborhoods [57]. We then track individuals based on their demographics. These abstract classes, such as 1 or 2, could represent a feature such as younger or older social media users. We also track their state, e.g., currently duped by a deepfake video (infectious) or not (susceptible). We also track the demographics of their neighbors to know their role in the network and exposure to other users in different states.

Our model has two critical mechanisms. First, inspired by our survey, individuals get

duped by their duped neighbor at a rate λ_i dependent on their demographic class i . Second, as in crowd-sourced approaches to correction of misinformation based on the “self-correcting crowd” [4, 52, 2], duped individuals can be corrected by their susceptible neighbors at a fixed rate γ . The dynamics of the resulting model are tracked using a heterogeneous mean-field approach [55] detailed in Box 1 and summarized in Fig. 3.6.

This model has a simple interesting behavior in homogeneous populations and becomes much more realistic once we account for heterogeneity in susceptibility. In a fully homogeneous population, $\lambda_i = \lambda \forall i$, if misinformation can, on average, spread from a first to a second node, it will never stop. The more misinformation spreads, the fewer potential fact-checkers remain. Therefore, misinformation invades the entire population for a correction rate γ lower than some critical value γ_c , whereas misinformation disappears for $\gamma > \gamma_c$.

The invasion threshold for misinformation is shown in Fig. 3.6(a). In heterogeneous populations, where different nodes can feature different susceptibility λ_i , the discontinuous transition from a misinformation-free to a misinformation-full state is relaxed. Instead, a steady state of misinformation can now be maintained at any level depending on the parameters of misinformation and the demographics of the population. In this regime, we can then further break down the dynamics of the system by looking at the role of duped nodes in the network, as shown in Fig. 3.6(b). The key result here is that very susceptible individuals with a homogeneous assortative neighborhood (e.g., an echo chamber) are at the highest risk of being duped. Conversely, nodes in the same demographic class but with a mixed or more diverse neighborhood are more likely to have resilient susceptible neighbors able to correct them if necessary.

Box 1: Mathematical model of diverse misinformation and herd correction on social networks

We wish to explore the potential impacts of our results on the spread of diverse misinformation on social networks. We consider that multiple independent streams of misinformation spread simultaneously; i.e., there are multiple sets of deepfakes, each with its own demographical biases. We also consider that social networks are often very heterogeneous with a skewed distribution of contacts per user and modular with denser connections among users of the same demographics.

We account for the above using three stylized patterns for the network structure. First, we divide the network into two demographic classes of equal size, simply labeled 1 and 2. Second, we assume a power-law distribution p_k of contacts k per user with $p_k \propto k^{-\alpha}$ regardless of demographics. Third, we use a mixed-membership stochastic block model to generate the network structure: Half of the nodes of each demographic always interact following their demographics, and half act as bridge nodes connecting randomly. The probability that a contact falls within a single demographic class is proportional to Q , while contacts across classes occur proportionally to $1 - Q$; with $Q > 0.5$ for modular structure.

According to the above, we can write the fraction of nodes $p_{k,\ell}^1$ which are of demographic class 1 with k contacts of class 1 and ℓ contacts of class 2:

$$p_{k,\ell}^1 \propto \frac{1}{2}(k + \ell)^{-\alpha} \left[\frac{1}{2} \binom{k + \ell}{k} Q^k (1 - Q)^\ell + \frac{1}{2} \binom{k + \ell}{k} (1/2)^{k+\ell} \right].$$

Box 1: Mathematical model of diverse misinformation and herd correction on social networks

On this contact structure, we then define a simple dynamical process where individuals are exposed to misinformation through each of their duped network neighbors, and themselves get duped at a rate λ_i based on their demographic class i . We also introduce an important mechanism where non-duped neighbors can correct their duped neighbors at a rate γ [4, 52, 2], e.g., we assume that your network neighbors can fact-check something you diffuse online and potentially correct your opinion. The fraction of individuals of a certain type (i, k, ℓ) that are duped, $D_{k,\ell}^i$, can be followed in time using a set of ordinary differential equations:

$$\frac{d}{dt}D_{k,\ell}^i = \lambda_i \left(p_{k,\ell}^i - D_{k,\ell}^i \right) (k\theta_{i,1} + \ell\theta_{i,2}) - \gamma D_{k,\ell}^i (k\phi_{i,1} + \ell\phi_{i,2}) .$$

where $\theta_{i,j}$ and $\phi_{i,j}$ represent the probabilities that a connection from an individual of demographic i to an individual of demographic j connects to a duped or non-duped individual, respectively. They can be calculated, for example, as

$$\theta_{1,2} = \sum_{k,\ell} k D_{k,\ell}^2 / \sum_{k',\ell'} k' p_{k',\ell'}^2 \quad \text{or} \quad \phi_{2,1} = \sum_{k,\ell} \ell \left(p_{k,\ell}^1 - D_{k,\ell}^1 \right) / \sum_{k',\ell'} \ell' p_{k',\ell'}^1 .$$

Box 1: Mathematical model of diverse misinformation and herd correction on social networks

These quantities close the system of equations and allow us to simulate a relatively simple model that manages to capture the heterogeneity (γ) and community structure (Q) of social networks, as well as demographic-specific susceptibility to misinformation ($\{\lambda_i\}$) and fact-checking among the population (γ). Our results are summarized in Fig. 3.6 and further analyzed in the Appendix [47].

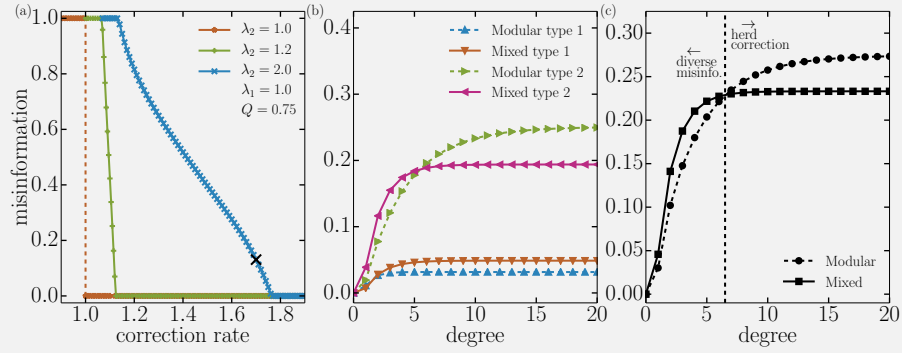


Figure 3.6: Spread of diverse deepfake on configurations of a degree-heterogeneous mixed membership stochastic block model with equal group size and densities (in-group density is set to Q and across the group to $1 - Q$). Other parameters are given in the plots, with panels (b) and (c) using the correction rate highlighted in (a) around a value of 1.7.

Consider now that diverse misinformation spreads. We assume just two types of misinformation (say young or older personas in two deepfake videos) targeting each of our two demographic classes (say younger and older social media users). We show this thought experiment in Fig. 3.6(c) where we use two complementary types of misinformation: One with $\lambda_1 = \lambda_2/2 = 1.0$ and a matching type with $\lambda_2 = \lambda_1/2 = 1.0$. We run the dynamics of these two types of misinformation independently as we assume they do not directly interact, and, therefore simply combine the possible states of nodes after integrating the dynamical

system. For example, the probability that a node of type 1 is duped by both pieces of misinformation would be the product of the probabilities that it is duped by the first and duped by the second. By doing so, we can easily study a model where multiple, diverse pieces of information spread in a diverse network population.

For diverse misinformation in Fig. 3.6(c), we find two connectivity regimes where the role of network structure is critical. For low-degree nodes, a diverse neighborhood means more exposure to *diverse misinformation* than a homogeneous echo chamber, such that the misinformation that best matches the demographics of a low-degree user is more likely to find them if they have a diverse neighborhood. For high-degree nodes, however, we find the behavior of herd correction: A diverse neighborhood means a diverse set of neighbors that is more likely to contain users who are able correct you if you become misinformed [77, 11, 76].

In the appendix, we analyze the robustness of herd correction to the parameters of the model. We show mathematically that the protection it offers is directly proportional to the homophily in the network (our parameter Q). By simulating the dynamics with more parameters, we also find that herd correction is proportional to the degree heterogeneity of the network. As we increase heterogeneity, we increase the strength of the friendship paradox. “Your friends have more friends than you do”[29], which means they get more exposed to misinformation than you do but also that they have more friends capable of correcting them when duped.

Our stylized model is meant to show how one can introduce biases in simple mathematical models of diverse misinformation. Future modeling efforts should also consider the possible interactions between different kinds of misinformation [17]. These can be synergistic, [37] parasitic [38], or antagonistic [30]; which all provide rich dynamical behaviors. Other possible mechanisms to consider are the adaptive feedback loops that facilitate the spread of misinformation in online social networks [72].

3.5 DISCUSSION

Understanding the structure and dynamics of misinformation is important as it can bring a great amount of societal harm. Misinformation has negatively impacted the ability to disseminate important information during critical elections, humanitarian crises, global unrest, and global pandemics. More importantly, misinformation degrades our epistemic environment, particularly regarding distrust of truths. It is necessary to understand who is susceptible to misinformation and how it spreads on social networks to mitigate its harm and propose meaningful interventions. Further, as deepfakes deceive viewers at greater rates, it becomes increasingly critical to understand who gets duped by this form of misinformation and how our biases and social circle impact our interaction with video content at scale. We hope this work will contribute to the critical literature on human biases and help to better understand their interplay with machine-generated content.

The overarching takeaways of our results can be summarized as follows. If not primed, humans are not particularly accurate at detecting deepfakes. Accuracy varies by demographics, but humans are generally better at classifying videos that match them. These results appear consistent with findings of the own-race bias (ORB) phenomenon[51], where overall, we see that participants are better at detecting videos that match their own attributes. Consistent with ORB research [3], our study results also show that white participants display a greater accuracy when presented with videos of white personas. We also see that persons of color are significantly more accurate than white participants when viewing deepfakes of personas of color and more accurate overall than white participants (see supplementary material) [47]. Our study adds several extra dimensions of demographic analysis by using gender and age. We see that male participants are significantly better at detecting videos of male personas than female viewers. With age, we see that when viewing videos of young personas, participants between the ages of 18-29 are significantly more accurate than par-

ticipants above the age of 30; surprisingly, participants aged 18-29 are also more accurate than participants aged 30-49 even when viewing videos of personas aged 30-49. Combining these results, more work needs to be done to understand better how interventions such as education about deepfakes, cross-demographic experiences and exposure, and exposure to the technology impact a user’s ability to detect deepfakes.

In this observational study, we also explored the potential impacts of these results in a simple mathematical model and extrapolated from our survey to hypothesize that a diverse set of contacts might provide “herd correction” where friends can correct each other’s blind spots. Friends with different biases can better correct each other when duped. This modeling result is a generalization of the self-correcting crowd approach used in the correction of misinformation [4].

In future work, we hope to investigate how non-primed human deepfake detectors perform when aided by machines. We want to investigate the mechanisms behind why some human viewers are better at guessing the state of videos that match their own identity. For example, do viewers have a homophily bias because they are more accustomed to images that match their own, or do they simply favor these images? We also would like to empirically investigate our survey via a more robust randomized controlled experiment and model results on real-world social networks with different levels of diversity to measure the spread of diverse misinformation in the wild. Consequently, we would be interested in testing possible educational or other intervention strategies to mitigate adversarial misinformation campaigns. Our simple observational study is a step towards understanding social biases’ role and potential impacts in an emerging societal problem with many multilevel interdependencies.

3.6 DATA AVAILABILITY

Our full survey questionnaire, code, data, and codebook can be found on our GitHub repository.

<https://github.com/juniperlovato/DiverseMisinformationPaper>

Due to the nature of this research, participants of this study did not consent for their personally identifiable data to be shared publicly, so the full survey’s raw supporting data is not available. Aggregated and anonymized data needed for analysis can be found in our repository.

METHODS

3.6.1 SURVEY METHODOLOGY

We first ran a pilot stage of our observational study. We conducted a simple convenience sample of 100 participants (aged 18+) to observe the efficacy of our survey. We then ran *phase 1* (April-May 2022) of the full survey using a Qualtrics survey panel of 1,000 participants who matched the demographic distribution of U.S. social media users. We then ran *phase 2* (September 2022) of the full survey, again using Qualtrics and the same sampling methodology. The resulting *full study* from phases 1 and 2 is a 2,016-participant sample.

Survey participants did not know before the start of the survey that the videos could potentially be deepfakes. The survey was framed for participants as a study about different communication styles and techniques that help make video content credible. Participants were told that we were trying to understand how aspects of public speaking, such as tone of voice, facial expressions, and body language, contribute to the effectiveness and credibility of

a speaker. The survey’s deceptiveness allowed us to ask questions about speaker attributes, likeability, and agreeableness naturally without priming the participants to look specifically for deepfakes [8]. We chose to make our survey deceptive not to prime the participants but also because this more closely replicates the deceptiveness that a social media user would encounter in the real world.

We designed our survey using video clips from the Deepfake Detection Challenge (DFDC) Preview Dataset [26, 25]. In our survey, we ask the participants to view two random video clips, which are approximately 10 seconds in length each. Each video clip may be viewed unlimited times before reading the questions but not again after moving to the questions. The information necessary to answer these questions relies solely on the previously shown video clip. A link to the full survey and survey questions is available in **Appendix 1** [47].

After viewing both videos, the participants are then asked to complete a related questionnaire about the communication styles and techniques of the videos. The questions ask about attributes of the video, such as pose, tone, and style and are asked to rate them on a Likert scale from very pleasant to very unpleasant. We also asked them to rate their agreement with the video content and credibility. We also ask participants to identify the perceived gender expression of the person(a) in the video, to identify what age group they belong to, and to ask if they perceive the person in the video to be a person of color or not.

Following the viewing of both videos and completion of a related questionnaire, the participants are debriefed on the deception of the survey, given a short explanation of deepfake technology, and then asked if they think the videos were real or fake (as seen in Fig. 3.7).

Lastly, we collect demographic information on the survey participants’ backgrounds and expressions of identity. We also ask participants how knowledgeable they already were on deepfakes, how often they use social media, and their political and religious affiliations. We also asked participants if they knew that the survey was about deepfakes before taking the

Do you think the primary person in Video #1 was real? (the first one you watched)

Note: If you are unsure, make your best guess.

☐ Yes, they are real
☐ No, they were fictionally created for this video

Do you think the primary person in Video #2 was real? (the second one you watched)

Note: If you are unsure, make your best guess.

☐ Yes, they are real
☐ No, they were fictionally created for this video

Figure 3.7: Preview of questions from our survey.

survey (survey participants who were primed were subsequently dropped from the analysis).

Sampling: Survey responses from 2,016 participants were collected through Qualtrics, an IRB-approved research panel provider, via traditional, actively managed, double-opt-in research panels [10]. Qualtrics’ participants for this study were randomly selected stratified samples from the Qualtrics panel membership pool that represents the average social media user in the US [56]. Our survey respondents represent the following categories and demographic breakdown in Table 4.3.

3.6.2 DATA

Secondary Data

For this project, we use the publicly available Facebook AI Research Deepfake Detection Challenge (DFDC) Preview Dataset (N = 5,000 video clips) [26, 25]. For our purposes, we filtered out all videos from the dataset that featured more than one person(a). The video clips may be deepfake or real; see Table 3.3. Additionally, some of the videos have been purposefully altered in several ways. Here is the list of augmenters and distractors:

- Augmenters: Frame-rate change, Quality level, Audio removal, add audio noise, Brightness/contrast level, Saturation, Resolution, Blur, Rotation, Horizontal flip.



Figure 3.8: Example video clip from the Facebook Deepfake Detection Challenge (DFDC) dataset. The person depicted is fake.

- Distractors: Dog filter, Flower filter, add overlaid images, shapes, or dots, add additional faces, add text.

A video's deepfake status (deepfake or not) was not revealed to the respondents during or after the survey. Many augmenters and distractors were noticeable to the respondents but were not specifically revealed.

Table 3.3: Descriptive statistics of video data (N=5,000).

Video	Binary	Number	Percent
Real	0	2,500	50%
Deepfake	1	2,500	50%

Original Data

We transformed all survey response variables of interest into numerical form to analyze our survey results. All Likert survey questions were converted from 'Very unpleasant,' 'Unpleasant,' 'Neutral,' 'Pleasant,' and 'Very pleasant' to an ordinal scale of 1,2,3,4,5.

Participants selected education levels from 'Some high school,' 'High school diploma or equivalent,' 'Some college,' Associate's degree (e.g., A.A., A.E., A.F.A., AS, A.S.N.),

Table 3.4: Descriptive demographics of survey participants (N=2,016).

Type	Sub-Group	% of Sample
Gender	Female	45%
Gender	Male	55%
Gender	Non-Binary	0.5%
Ages	18-29	17%
Ages	30-49	27%
Ages	50-64	29%
Ages	65+	27%
Demographic	Non-Hispanic White	67%
Demographic	Non-Hispanic Black	10%
Demographic	Hispanic	14%
Demographic	Other	9%

‘Vocational training,’ ‘Bachelor’s degree (e.g., B.A., BBA BFA, BS),’ ‘Some postgraduate work,’ ‘Master’s degree (e.g., M.A., M.B.A., M.F.A., MS, M.S.W.),’ ‘Specialist degree (e.g., EdS),’ ‘Applied or professional doctorate degree (e.g., M.D., D.D.C., D.D.S., J.D., PharmD),’ ‘Doctorate degree (e.g., EdD, Ph.D.)’ was transformed to an ordinal scale of 1-11 respectively.

Participants selected income levels from ‘Less than \$30,000,’ ‘\$30,000-\$49,999,’ ‘\$50,000-\$74,999,’ ‘\$75,000+’ were transformed to an ordinal scale of 1-4 respectively.

Participants selected their social media usage levels from ‘I do not use social media,’ ‘I use social media but less than once a month,’ ‘Once a month,’ ‘A few times a month,’ ‘Once a week,’ ‘A few times a week,’ ‘Once a day,’ ‘More than once a day’ were transformed to an ordinal scale of 1-8 respectively. Variables were split into the category of frequent social media users 5-8 and infrequent social media users 1-4.

Participants selected their knowledge of deepfake from ‘I did not know what a deepfake was,’ ‘I somewhat knew what a deepfake was,’ ‘I knew what a deepfake was,’ ‘I consider myself knowledgeable about deepfakes’ was transformed to an ordinal scale of 1-4 respectively. Variables were split into users who are knowledgeable about deepfakes 3-4 and users

who are not knowledgeable about deepfakes 1-2.

All nominal and categorical variables were transformed into binary variables. Categorical variables (some survey questions included write-in answers) were combined into coarser-grained categories for analysis, such as participant racial/ethnic identity (transformed to Person of Color or White), U.S. state of residence (transformed to U.S. regions), employment (transformed to occupational sectors), religious affiliation (transformed into religious affiliations), and political affiliation (transformed to major political affiliations).

We allowed survey participants to identify their gender identity, the results of which were largely binary. Unfortunately, our sample was insufficient to perform meaningful analysis on a larger non-binary gender identity spectrum. Variables with an N under 30 were subsequently dropped. Our survey participants were given two video clips to view and critique; in our analysis, we decided to analyze the first or second video in the same way.

3.6.3 ANALYTICAL METHODS

Matthews Correlation Coefficient: To understand the relationship between the participant’s guesses on the status of the video (fake or real) and the actual state of the video (fake or real), we ran a Matthews Correlation Coefficient (MCC) [50, 14] to compare what variables most significantly impact a participant’s ability to guess the actual state of the video correctly. MCC is typically used for classification models to observe the classifier’s performance. Here we treat human participant subgroups as classifiers and measure their performance with MCC. MCC takes the participant subgroup’s guesses and the actual answers and breaks them up into the following categories: number of true positives (TP), number of true negatives (TN), number of false positives (FP), and number of false negatives (FN). The MCC metric ranges from -1 to 1, where 1 indicates total agreement between

participant guess about the video and the actual state of the video, -1 indicates complete disagreement between participant guess about the video and the actual state of the video, and 0 indicates something similar to a random guess. To calculate the MCC metric for our human classifiers, we then use the following formula:

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC is considered a more balanced statistical measure than an F1, precision, or recall score because it is symmetric, meaning no class (e.g., TP, TN, FP, FN) is more important than another.

To compare MCC scores, we bootstrap samples from pairs of confusions matrices and compare their MCC scores. This process generates 10,000 bootstrapped samples of differences in correlation coefficients. We then compare the null hypothesis (difference equal to zero) to the bootstrapped distribution to measure the significance of biases and get a credibility interval on their strength.

Accuracy Rate: The performance metric we use to measure participant accuracy is the ratio of the correct guesses to the entire pool of guesses where

$$accuracy = (TP + TN) / (TP + FP + FN + TN)$$

BIBLIOGRAPHY

- [1] Airoldi, E. M., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21.
- [2] Allen, J., Arechar, A. A., Pennycook, G., and Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393.
- [3] Anthony, T., Copper, C., and Mullen, B. (1992). Cross-racial facial identification: A social cognitive integration. *Personality and Social Psychology Bulletin*, 18(3):296–301.

- [4] Arif, A., Robinson, J. J., Stanek, S. A., Fichet, E. S., Townsend, P., et al. (2017). A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors . In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, Cscw '17*, page 155–168, New York, NY, USA. Association for Computing Machinery.
- [5] Artz, K. (2019). Texas Outlaws ‘Deepfakes’—but the Legal System May Not Be Able to Stop Them .
- [6] Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? . *Int J Methods Psychiatr Res*, 20(1):40–49.
- [7] Bagrow, J. P., Liu, X., and Mitchell, L. (2019). Information flow reveals prediction limits in online social activity. *Nat. Hum. Behav.*, 3(2):122–128.
- [8] Barrera, D. and Simpson, B. (2012). Much ado about deception: Consequences of deceiving research participants in the social sciences . *Sociological Methods & Research*, 41(3):383–413.
- [9] Blue, L., Warren, K., Abdullah, H., Gibson, C., Vargas, L., et al. (2022). Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction . In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2691–2708, Boston, MA.
- [10] Boas, T. C., Christenson, D. P., and Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics . *Political Science Research and Methods*, 8(2):232–250.
- [11] Bode, L. and Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media . *Journal of Communication*, 65(4):619–638.
- [12] Bond, J., Julion, W. A., and Reed, M. (2022). Racial Discrimination and Race-Based Biases on Orthopedic-Related Outcomes: An Integrative Review . *Orthopaedic Nursing*, 41(2):103–115.
- [13] Bothwell, R. K., Brigham, J. C., and Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, 15(1):19–25.
- [14] Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric . *PLOS One*, 12(6):e0177678.
- [15] Brigham, J. C., Maass, A., Snyder, L. D., and Spaulding, K. (1982). Accuracy of eyewitness identification in a field setting. *Journal of Personality and Social Psychology*, 42(4):673.
- [16] Calvillo, D. P., Garcia, R. J., Bertrand, K., and Mayers, T. A. (2021). Personality factors and self-reported political news consumption predict susceptibility to political fake news . *Pers. Individ. Differ.*, 174:110666.
- [17] Chang, H.-C. H. and Fu, F. (2018). Co-diffusion of social contagions. *New Journal of Physics*, 20(9):095001.
- [18] Cheng, J. and Bernstein, M. S. (2015). Flock: Hybrid Crowd-Machine Learning Classifiers. In *Association for Computing Machinery, CSCW '15*, page 600–611, New York,

- NY, USA. Association for Computing Machinery.
- [19] Chesney, B. and Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security . *Calif. L. Rev.*, 107:1753.
 - [20] Chou, W.-Y. S., Oh, A., and Klein, W. M. (2018). Addressing health-related misinformation on social media. *The Journal of the American Medical Association*, 320(23):2417–2418.
 - [21] Citron, D. K. (2022). *The fight for privacy: protecting dignity, identity, and love in the digital age* . W.W. Norton & Company, first edition.
 - [22] Currarini, S. and Mengel, F. (2016). Identity, homophily and in-group bias. *European Economic Review*, 90:40–55.
 - [23] Dandekar, P., Goel, A., and Lee, D. T. (2013). Biased assimilation, homophily, and the dynamics of polarization. *Proc. Natl. Acad. Sci. U.S.A.*, 110(15):5791–5796.
 - [24] de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*, 34(4):1311–1332.
 - [25] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., et al. (2020). The DeepFake Detection Challenge Dataset. *arXiv:2006.07397*.
 - [26] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv:1910.08854*.
 - [27] Ebner, N. C., Ellis, D. M., Lin, T., Rocha, H. A., Yang, H., et al. (2020). Uncovering susceptibility risk to online deception in aging. *The Journals of Gerontology: Series B*, 75(3):522–533.
 - [28] Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, pages 1–21.
 - [29] Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477.
 - [30] Fu, F., Christakis, N. A., and Fowler, J. H. (2017). Dueling biological and social contagions. *Scientific Reports*, 7(1):1–9.
 - [31] Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., and Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):3.
 - [32] Greengard, S. (2019). Will deepfakes do deep damage? *Communications of the ACM*, 63(1):17–19.
 - [33] Groh, M., Epstein, Z., Firestone, C., and Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci. U.S.A.*, 119(1).
 - [34] Güera, D. and Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE.
 - [35] Harris, D. (2018). Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Technology Review*, 17:99.
 - [36] Haut, K., Wohn, C., Antony, V., Goldfarb, A., Welsh, M., et al. (2021). Could you become more credible by being White? Assessing Impact of Race on Credibility with

- Deepfakes . *arXiv:2102.08054*.
- [37] Hébert-Dufresne, L. and Althouse, B. M. (2015). Complex dynamics of synergistic coinfections on realistically clustered networks . *Proc. Natl. Acad. Sci. U.S.A.*, 112(33):10551–10556.
 - [38] Hébert-Dufresne, L., Mistry, D., and Althouse, B. M. (2020). Spread of infectious disease and social awareness as parasitic contagions on clustered networks . *Physical Review Research*, 2(3):033306.
 - [39] Jung, T., Kim, S., and Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access*, 8:83144–83154.
 - [40] Kimmel, A. J. (2004). Rumors and the financial marketplace. *J Behav Financ*, 5(3):134–141.
 - [41] Klaczynski, P. A., Felmban, W. S., and Kole, J. (2020). Gender intensification and gender generalization biases in pre-adolescents, adolescents, and emerging adults . *British Journal of Developmental Psychology*, 38(3):415–433.
 - [42] Kossinets, G. and Watts, D. J. (2009). Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450.
 - [43] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.
 - [44] Legislature of the State of Texas (2019). § Senate Bill No. 751.
 - [45] Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470.
 - [46] Lloyd, E. P., Hugenberg, K., McConnell, A. R., Kunstman, J. W., and Deska, J. C. (2017). Black and White lies: Race-based biases in deception judgments. *Psychological Science*, 28(8):1125–1136.
 - [47] Lovato, J., Hébert-Dufresne, L., St-Onge, J., Harp, R., Salazar Lopez, G., et al. (2022a). Supplementary materials for Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks . *Available upon request*.
 - [48] Lovato, J. L., Allard, A., Harp, R., Onaolapo, J., and Hébert-Dufresne, L. (2022b). Limits of individual consent and models of distributed consent in online social networks . In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2251–2262.
 - [49] Macchi Cassia, V. (2011). Age biases in face processing: The effects of experience across development . *British Journal of Psychology*, 102(4):816–829.
 - [50] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme . *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.
 - [51] Meissner, C. A. and Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. . *Psychology, Public Policy, and Law*, 7(1):3.
 - [52] Micallef, N., He, B., Kumar, S., Ahamad, M., and Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the COVID-19 infodemic . In *2020*

- IEEE International Conference on Big Data (Big Data)*, pages 748–757. Ieee.
- [53] Mori, M. (1970). The uncanny valley: the original essay by Masahiro Mori. *IEEE Spectrum*.
 - [54] Nightingale, S. J., Wade, K. A., and Watson, D. G. (2022). Investigating age-related differences in ability to distinguish between original and manipulated images. . *Psychology and Aging*, 37(3):326–337.
 - [55] Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200.
 - [56] Pew Research Center (2021). Social media fact sheet. *Pew Research Center: Washington, DC, USA*.
 - [57] Red, V., Kelsic, E. D., Mucha, P. J., and Porter, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks . *SIAM Review*, 53(3):526–543.
 - [58] Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers’ Imprint*, 20(24):1–16.
 - [59] Roose, K. (2018). Here come the fake videos, too. *The New York Times*, 4.
 - [60] Roth, C., St-Onge, J., and Herms, K. (2022). Quoting is not Citing: Disentangling Affiliation and Interaction on Twitter . In Benito, R. M., Cherifi, C., Cherifi, H., Moro, E., Rocha, L. M., and Sales-Pardo, M., editors, *Complex Networks & Their Applications X*, Studies in Computational Intelligence, pages 705–717. Springer International Publishing.
 - [61] S.3805 – 115th Congress (2017–2018) (2018). Malicious Deep Fake Prohibition Act of 2018.
 - [62] Schwartz, G. T. (1990). Explaining and Justifying a Limited Tort of False Light Invasion of Privacy . *Case W. Res. L. Rev.*, 41:885.
 - [63] Sedhai, S. and Sun, A. (2015). HSpam14: A collection of 14 million tweets for hashtag-oriented spam research . In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 223–232.
 - [64] Shillair, R. and Dutton, W. H. (2016). Supporting a cybersecurity mindset: getting internet users into the cat and mouse game . *Social Science Research Network*.
 - [65] Solove, D. J. (2002). Conceptualizing privacy. *California Law Review*, 90:1087.
 - [66] Starbird, K., Maddock, J., Orand, M., Achterman, P., and Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing . *IConference 2014 proceedings*.
 - [67] Tambuscio, M., Ruffo, G., Flammini, A., and Menczer, F. (2015). Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks . In *Proceedings of the 24th international conference on World Wide Web*, pages 977–982.
 - [68] Tasnim, S., Hossain, M. M., and Mazumder, H. (2020). Impact of rumors and misinformation on COVID-19 in social media. *J Prev Med Public Health*, 53(3):171–174.
 - [69] The Committee on the Judiciary House of Representatives (2019). Federal Rules of Evidence.
 - [70] The People of the State of California (2019). Assembly Bill No. 602 to the Civil Code, relating to privacy, § 1708.86. 2019 .

- [71] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.
- [72] Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion . *PLOS One*, 13(9):e0203958.
- [73] Traberg, C. S. and van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility . *Pers. Individ. Differ.*, 185:111269.
- [74] van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public . *Nature Medicine*, 28(3):460–467.
- [75] Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932.
- [76] Vraga, E. K. and Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645.
- [77] Walter, N., Brooks, J. J., Saucier, C. J., and Suresh, S. (2021). Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis . *Health Communication*, 36(13):1776–1784.
- [78] Watts, D. J., Rothschild, D. M., and Mobius, M. (2021). Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci. U.S.A.*, 118(15):e1912443118.
- [79] Wu, L., Morstatter, F., Carley, K. M., and Liu, H. (2019). Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explorations Newsletter*, 21(2):80–90.
- [80] Zotov, S., Dremluga, R., Borshevnikov, A., and Krivosheeva, K. (2020). DeepFake Detection Algorithms: A Meta-Analysis. In *2020 2nd Symposium on Signal Processing Systems*, pages 43–48.

CHAPTER 4

GROUPED DATA: ISSUES OF CONSUMER DATA AGGREGATION ONLINE



Figure 4.1: Data Types Monster by Julia Zimmerman

PREFACE

This chapter explores tools and methods to interrogate platforms’ privacy policies to measure the risks and harms associated with grouping data types of personally identifiable information (PII) about data subjects together. This chapter looks at ways in which we can get more transparency on the activities of data processors. We will look at methods for investigating privacy policies to find what personally identifiable data types brokers say they collect, use, and sell. We will also introduce some methods to measure how these collection practices change over time, how complex the privacy policies are over time, and measure the sensitivity of the data types collected concurrently. These Natural Language Processing (NLP) measurements may help to provide transparency and mitigate some of the consent fatigue we explored in Chapter 2. Material from this chapter has been published or made publicly available in the following form:

Lovato, J., Mueller, P., Suchdev, P., S. Dodds, P.. (2023). More Data Types More Problems: A Temporal Analysis of Complexity, Stability, and Sensitivity in Privacy Policies. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1088-1100, DOI: 10.1145/3593013.3594065

ABSTRACT

Collecting personally identifiable information (PII) on data subjects has become big business. Data brokers and data processors are part of a multi-billion-dollar industry that profits from collecting, buying, and selling consumer data. Yet there is little transparency in the data collection industry which makes it difficult to understand what types of data are being collected, used, and sold, and thus the risk to individual data subjects. In this

study, we examine a large textual dataset of privacy policies from 1997-2019 in order to investigate the data collection activities of data brokers and data processors. We also develop an original lexicon of PII-related terms representing PII data types curated from legislative texts. This mesoscale analysis looks at privacy policies over time on the word, topic, and network levels to understand the stability, complexity, and sensitivity of privacy policies over time. We find that (1) privacy legislation may be correlated with changes in stability and turbulence of PII data types in privacy policies; (2) the complexity of privacy policies decreases over time and becomes more regularized; (3) sensitivity rises over time and shows spikes that appear to be correlated with events when new privacy legislation is introduced.

4.1 INTRODUCTION

Data brokers and data processors (DBDPs) form a multi-billion-dollar industry that collects, buys, and sells personally identifiable information (PII) from individuals worldwide. According to the market research company eMarketer, the size of the data broker industry was approximately 300 billion dollars in 2020. They have also recently projected the industry’s size to nearly double within three years [19].

Even with new legislation emerging in the U.S. and the E.U., the data broker industry lacks sufficient transparency and regulation. Moreover, many data processors may not be officially listed by definition as data brokers on any U.S. state registry, even though they play a significant role in collecting, processing, and sharing (often for monetary gain) data that they collect from consumers. It is important to note that many data processors do not sell raw data but rather generate income by serving as a data pass-through selling insights or using the data to generate income from advertising revenue [14]. According to the Vermont Data Broker Act of 2018, a data broker is defined as “a business, or unit or units of a business, separately or together, that knowingly collects and sells or licenses

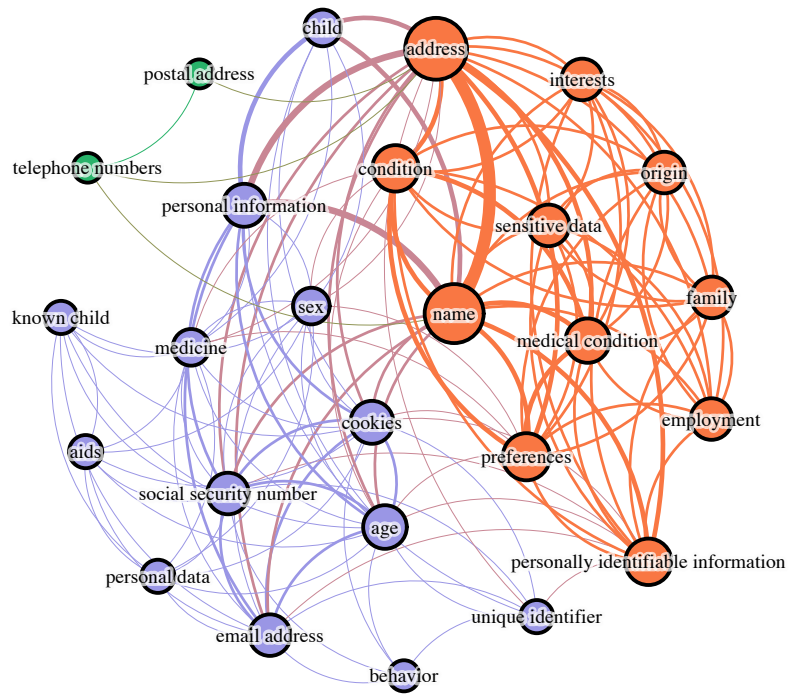


Figure 4.2: 1997 co-occurrence network of PII relevant terms. All nodes with a degree less than one are filtered from the network (words that do not co-occur with other words). Networks are partitioned by modularity, signified by the nodes' color. The node size ranges from 30 to 60 based on the weighted degree.

to third parties the brokered personal information of a consumer with whom the business does not have a direct relationship” [2]. For example, Facebook may not be considered a data broker by definition. However, it is one of the largest data processors in the world and benefits financially from the data assets they collect from social media users. For the purposes of our project, we will consider both data brokers and data processors (we will collectively call them DBDPs) because both significantly impact the risks and potential harms to data subjects (individuals whose data is collected by DBDPs) associated with collecting personal data.

Lack of transparency and information about the DBDP industry [7] is a significant hurdle in assessing the harms associated with collecting PII. For our purposes, we are primarily concerned with harm to U.S. data subjects.

There are several issues related to the lack of transparency in the DBDP industry, namely: (1) It is difficult to properly assess the magnitude and impact of harm to data subjects because it is currently unknown how much data and what data types have been collected by DBDPs over time; (2) There is little to no mandatory financial reporting for the DBDP industry, so the market value of the data they collect and sell is unknown which makes it difficult to assess damages and loss; (3) Finally, it is difficult to track which third parties have been granted access to PII data by DBDPs and track the flow of information that is shared.

Moreover, many DBDPs function under the justification that PII data is collected with sufficient consent from individuals. However, the legitimacy of this consent is a concern, particularly considering the necessary criteria for informed consent and the issue of group consent for data that implicates social groups [22].

The legitimacy of individual informed consent is dependent on (1) the data subject understanding the agreement; (2) entering into it without coercion; (3) entering the agreement intentionally and deliberately; (4) and the agreement authorizing a specific course of action

for the data being collected [30]. If DBDPs were to follow the individual informed consent criteria, particularly for criteria 4, they would need a mechanism for the data subject to track the status of their data to assess if the consent agreement was being upheld. However, there is currently no way in the U.S. to track your PII data and fully understand how it is being processed and shared by DBDPs. Information asymmetry between DBDPs and data subjects makes it challenging for data subjects to follow their data, know what types of data are being collected, used, and shared, and hold firms accountable if data are being processed against the terms of their original consent agreement or the law.

Transparency will remain an issue until there is transparent mandatory reporting by DBDPs on the data collected, used, and shared, a mechanism for consumers to track their data, and disclosure of detailed digital assets by DBDPs (currently, only publicly traded DBDPs need to disclose this information to the Securities and Exchange Commission and data assets are lumped together with all intangible assets like patents and intellectual property). In place of these resolutions, researchers must find other means to infer the activities of DBDPs. Typically, the only reporting required for DBDPs, in the United States includes declaring they are a data broker on a State registry (required by law in states like Vermont [2] and California [1]), and most of these DBDPs also publish a privacy policy on their public website.

In this study, we perform a mesoscale exploratory data analysis on a temporal privacy policy dataset (includes years 1997-2019) and extract time series to explore the following measures: (1) descriptive summaries of the text such as basic summary statistics and topics, (2) frequency distributions of words as a measure of stability, (3) complexity as measured by corpus compression factor, (4) and co-occurrence network structures as a measure of sensitivity.

We investigate text from privacy policies to infer DBDP activities and understand what PII data types DBDPs collect and how this activity has changed over time. Our study looks

at a temporal textual dataset from Amos et al. [3] that includes over one million privacy policy snapshots from the Internet Archive’s WayBack Machine from 1997-2019 spanning over 100,000 websites. We also use legislative text from nine U.S. state laws on data privacy to manually extract an original lexicon of personally identifiable information (PII) terms represented in the corpora to represent PII data types 4.3.



Figure 4.3: PII Data Types by Julia Zimmerman

Finally, we compare signals in our results to privacy regulation events in the U.S. and the E.U. (we include some E.U. regulations events like GDPR because they have a global impact on privacy policy text [4]) and find correlations between changes in privacy policies and new governmental regulations. A timeline of significant governmental regulations in the U.S. and E.U. from 1997-2022 can be seen in Figure 4.4. This multidisciplinary project uses methods from complex systems and data science, natural language processing (NLP), computational social science, and network science.

Future work will examine the flow of these data by inferring the data types collected

by DBDPs in the collection section of the text and then inferring the data types shared by DBDPs in the share section of the text. We will then investigate tort law to find examples of the magnitude of monetary harms associated with each relevant PII data types to estimate the potential monetary harm inflicted on data subjects through collecting and selling their PII data.

This chapter investigates four primary research questions:

Research Questions:

- **Q1:** What PII-related words, topics, and co-occurrence network structures appear in privacy policies over time?
- **Q2:** What PII data types are collected from consumers, and how stable is the representation of those PII data types over time?
- **Q3:** How complex or regular are privacy policies over time?
- **Q4:** What PII data types are collected concurrently from consumers, and how has the level of sensitivity (and potential risk to data subjects) changed over time?

This chapter is structured as follows. We outline related work around privacy policy research, risks associated with data aggregation, DBDPs, and data valuation in Section 4.2. We investigate the PII data types lexicon on the temporal privacy policy dataset measured by its summary statistics, stability, complexity, and sensitivity over time. We also report the results of our mesoscale analysis on the structure and dynamics of the policies over time in Section 4.3. We discuss our findings, their broader implications, and future work in Section 4.4. Finally, We outline a large temporal dataset of privacy policies and a lexicon of PII relevant terms and explore methods used in Appendices 4.5, 4.6, 4.7, and 4.8.

4.2 BACKGROUND

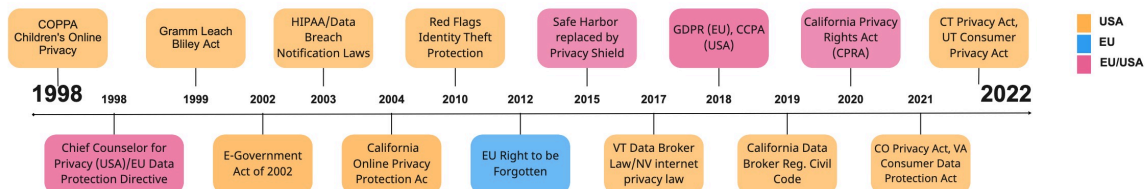


Figure 4.4: Timeline of selected privacy legislation in the U.S. and E.U. from 1998-2022

4.2.1 RESEARCH ON DATA BROKERS

Otto et al. [25] conducted a critical case study of ChoicePoint, a data broker well known for a massive data breach and consumer information mishandling. In 2005, ChoicePoint disclosed that it had sold the personal records of thousands of individuals to identity thieves, resulting in hundreds of fraud and identity theft cases. This scandal shed light on ChoicePoint's practices and gave the researchers access to data about the firm's activity that would not otherwise be available.

In the case study, the authors track the flow of data collected, bought, and sold by ChoicePoint. Their analysis of how the data flows from one place to another highlights the extensive breadth of data sources collected, purchased, and sold by the broker. It also highlights that the data types sold were highly sensitive, including D.N.A. sequences and social security numbers. These data were sold, without restriction, to a wide variety of buyers, which ranged from individuals to large agencies. The only reason the authors have such a detailed account of the data flow for this case study is that ChoicePoint had several large data breach events. Brokers who have not faced large events such as bankruptcy or security breaches are not required to disclose this information. Roderick and Crain [7, 28]

argue that this lack of transparency and an asymmetry in power takes away consumers' ability to protect their data. More transparency would open up additional analysis of this industry.

4.2.2 RESEARCH ON RISK AND HARMS OF COMBINING DATA TYPES

The harms associated with collecting data increase superlinearly as different types of data are aggregated together. The harms become more than the sum of their constituent parts. Datasets become more sensitive and pose more privacy risks to a subject as more data types are combined. For example, a dataset of names may not be considered sensitive. Once combined with associated social security numbers, the data becomes much more sensitive than each data point on its own. Further, if combined with a personal address, this information collection becomes more sensitive because it can be used to assume a person's digital identity. With the collection of even more information, such as browsing history, this aggregate of information can be used to build a digital dossier [32] on the data subject, which can be used to make inferences about them, predictions about their future behavior, and manipulate their future behavior [6, 35] with targeted advertising and other techniques.

It is well studied that aggregation of PII and associated data types increase the data privacy risk to the data subject [39]. Privacy risk pertains particularly to the potential or actual harm to individual data subjects. According to Wagner et al. [39], the impact of privacy risk can be broken down into composite categories: scale, sensitivity, expectation, and harm. Scale can be quantified by the number of individuals implicated. Sensitivity can be quantified by the number of data types involved, entropy, and the average privacy setting. The expectation of risk can be quantified as deviation from the expectation of how the data will be handled. Finally, harm can be quantified as damages awarded or perceived

harm.

In this project, we will primarily focus on the impact of sensitivity measured by the number of different PII data types collected together over time by DBDPs. As stated above, aggregations of different data types display a higher level of risk than the sum of the component data type risks in isolation. In future work, we will measure harm through tort damages as defined by Wagner et al. [39].

4.2.3 ANALYSIS OF PRIVACY POLICIES

The current over-reliance on data subjects to take on the burden of reading privacy policies and consent online has led to a negative externality of less legitimate consent due to consent fatigue [29]. McDonald et al. [23] estimates that if every individual were to fully read every privacy policy they encountered online in the United States during 2008, the national opportunity cost would be 781 billion dollars.

Several researchers have built automated methods to help data subjects parse through the legal language of privacy policies in shorter time periods [33, 8]. We believe these are indeed very important tools. Still, we also recognize that sometimes the devil is in the details regarding consent agreements. A balance between easily understood and readable versions [15, 26] of privacy policies and thorough disclosure of what information is being collected, used, and shared is also needed.

Other studies have investigated the text of privacy policies to understand how they have adapted to legal frameworks over time [21, 40, 34, 3, 38, 18] and how well they align with consumer values [9]. In 2021, Amos et al. [3] crawled over one million privacy policies on the Way Back Machine from the past 20 years to bring transparency to these issues. This dataset is made publicly available by request. They also created an automated trend detection tool to identify terms and concepts that show shifts in the language of privacy

policies. Linden et al. [21] also conducted a text analysis on privacy policies to examine how these policies have changed since the introduction of the EU General Data Protection Regulation (GDPR). In 2020, Srinath et al. created the web crawler *Privateer* [33], which made a large-scale corpus of web privacy policies.

4.3 RESULTS

Our results can be broken up into three mesoscale categories:

Result #1. Word Level: In result #1, we highlight the stability of the privacy policy PII data type terms over time as represented by frequency trends. Several sensitive data types related to location, behavior, and internet activity appear to rise over time. Stable words predominantly describe customer information. Falling words may result from other, more specific words taking over. For example, “coordinates” appears to be falling in frequency over time, but other geolocation words are taking over the word space. For the most part, new words appear to reflect biometric information that may be tied to new technology or methods for inference, such as “faceprint” and “voiceprint.”

Result #2. Topic Level: In result #2, we explore the complexity of the privacy policies as represented by the compression factor of privacy policies compared by year. We see a steady decrease in complexity over time from 2000-2019. In our analysis of topic prevalence over time, the usage of topics is relatively stable over time, which speaks to the rigid and established language used in privacy policies. Moreover, we see trends related to new technologies and legislation as privacy policies are added and amended over time.

Result #3. Network Level: In result #3, we investigate the level of sensitivity and risk (number of PII data types collected together) in the privacy policies as represented by the density of the word co-occurrence network graph. We also consider the network’s density alongside modularity because there are networks of different sizes across time. Modularity

allows for a quantitative comparison of community structures via the number of classes across networks of different sizes. We see that network density rises over time.

We address our high-level research questions as follows. Question 1: What PII words, topics, and co-occurrence network structures appear in privacy policies over time? Will be addressed by results #1-3. Question 2: What PII data types are collected from consumers, and how stable is the representation of those PII data types over time? Will be addressed in result #1. Question 3: How complex or regular are privacy policies over time? Will be addressed in result #3. Finally, question 4: What PII data types are collected concurrently from consumers, and how has the level of sensitivity (and potential risk to data subjects) changed over time? will also be addressed in result #3.

4.3.1 DESCRIPTIVE STATISTICS OF PRIVACY POLICY DATA

The privacy policy corpus is provided by Amos et al. [3] in SQLite format and weighs in at 48.24GB. It includes over 1 million snapshots of privacy policies from the Internet Archive WayBack Machine from 1997-2019. The data includes many fields, but for our purposes, we use the following columns: the full text of the privacy policy snapshot in markdown, the year it was collected, the URL, and the category of the institution (as classified by Webshrinker’s API) for the policy. The necessary information was extracted from the corpus using SQL queries and converted to a comma-separated values file format using the Pandas library. On average, a website has 8.4 privacy policy snapshots ($M = 6$). 79.4% of snapshots are from 2010 or later. An overview of summary statistics about the privacy policy dataset can be found in Table 4.1 in Appendix 4.5, and all associated code for our analysis can be found in 4.8.

To explore the dataset for our research questions, we cleaned and segmented the data in

several ways: (1) We filter all sentences that include negation (negation words lexicon can be found in Appendix 4.7); (2) Used NLTK for Tokenization and removal of punctuation (Regexp \w); (3) Made all text in the privacy policies lowercase; (3) Split the dataset by year; (4) For some analysis, we filter all of the corpora to extract just the terms in our PII Lexicon. PII lexicon words can be found in Appendix 4.6. We decided to keep non-sentence text, like headings or tables, and treat it the same as other parts of the text, as they could potentially use slightly different wording and thus help us match on specific terms.

4.3.2 PII DATA TYPES LEXICON

For our project, we manually curated an original lexicon of 287 terms that are associated with personally identifiable information (PII) as defined by legislative definitions sections from nine U.S. state-level laws concerning consumer data privacy. These laws include:

1. California Consumer Privacy Act (CCPA) of 2018 (Cal. Civ. Code §§ 1798.100 et seq.)
2. California Consumer Privacy Rights Act (CPRA) of 2020 (Proposition 24 A.B. 1490)
3. Colorado Privacy Act of 2021 (Colo. Rev. Stat. § 6-1-1301 et seq)
4. Connecticut Act Concerning Personal Data Privacy and Privacy Online Monitoring 2022 (Senate Bill No. 6 File No. 238 Cal. No. 189)
5. Utah Consumer Privacy Act 2022 (S.B. 227)
6. Virginia Consumer Data Protection Act 2021 (2021 H.B. 2307/2021 S.B. 1392)
7. California Data Broker Registration Civil Code 2019 (Cal. Civ. Code §§ 1798.99.80 et seq)

8. Nevada internet privacy laws 2021 and 2017 (NRS § 603A.300 and 2021 S.B. 260)
9. Vermont Protection of Personal Information Act 2017 (9 V.S.A § 2446-2447).

To create our lexicon, we manually extract defined terms by reading legislative documents and collecting words defined as personally identifiable information (PII) in all legislative documents' definitions sections. The full lexicon of PII data type terms can be found in Appendix 4.6. PII lexicon summary statistics can be found in Appendix 4.5. An overview of descriptive statistics on the filtered PII data types corpus can be seen in Appendix 4.5 Table 4.2. Note that the number of unique words present in each year shows the number of unique PII data types used in the corpus for that year which increases steadily over time. This lexicon should be treated as a dynamic list that will adjust as new legislative text is introduced over time.

4.3.3 WORD LEVEL RESULTS: MEASURES OF TURBULENCE

In this section, we look at the frequency distribution of our PII data types lexicon as a time series to investigate what PII data types demonstrate rapid increases in frequency, decreases in frequency, stability in frequency, and introduction of new words over time. We will also look at groups of related words to demonstrate how the PII data types can be used to query related groups of words to investigate their frequency distributions in the context of more specific legislative events.

Rise: Rising looks at PII data types whose frequency increases quickly over time. We define rising words as words that rise in frequency by more than ten times in a 7-year period, as seen in Figure 4.5a. Our results show that PII terms such as “beacons,” “behavioral,” “geolocation,” “inferences,” “religion,” “latitude,” “longitude” rise quite dramatically in the

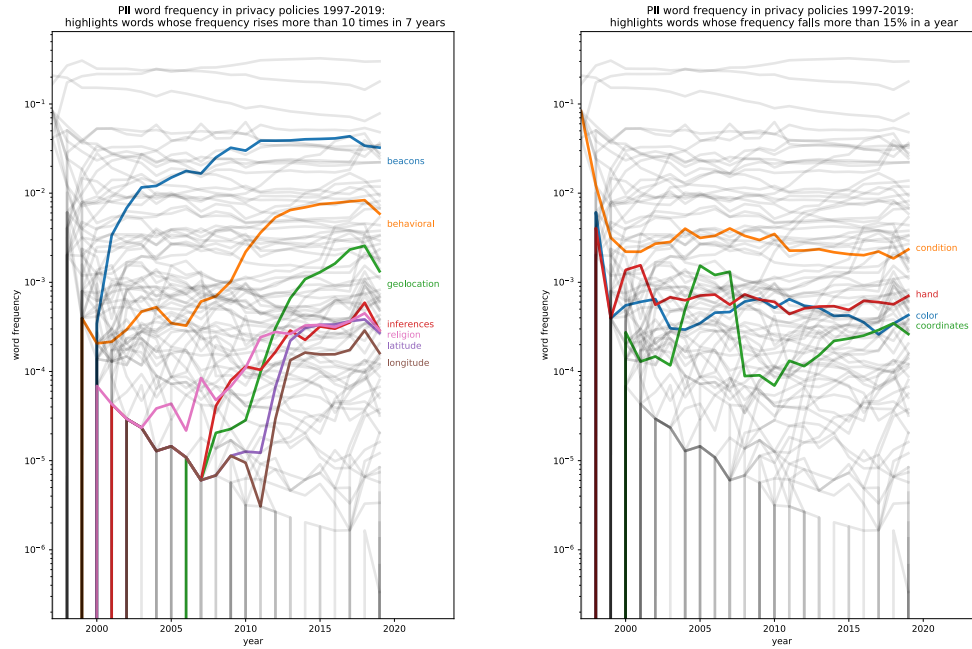


Figure 4.5: a.) Frequency Distribution of words that rise more than ten times in a 7-year period, b.) Frequency Distribution of words that fall more than 15% in a year

dataset over time. It is worth noting that these terms represent rather sensitive types of information related to geolocation data, browsing behavior, and inferences. Notably, these data types represent indirect information collection methods, meaning it is not information directly given to the DBDPs but gathered from monitoring the data subject’s behavior and digital traces over time.

Fall: Falling looks at PII data types whose frequency drops quickly over time. Falling words are defined as words that fall in frequency more than 15% in a year, as seen in Figure 4.5b. Our results show that terms such as “condition,” “hand,” “coordinates,” and “color” fall quite dramatically in the dataset over time. The decreased use of words like “coordinates” may be explained by the rising use of other geolocation words taking their

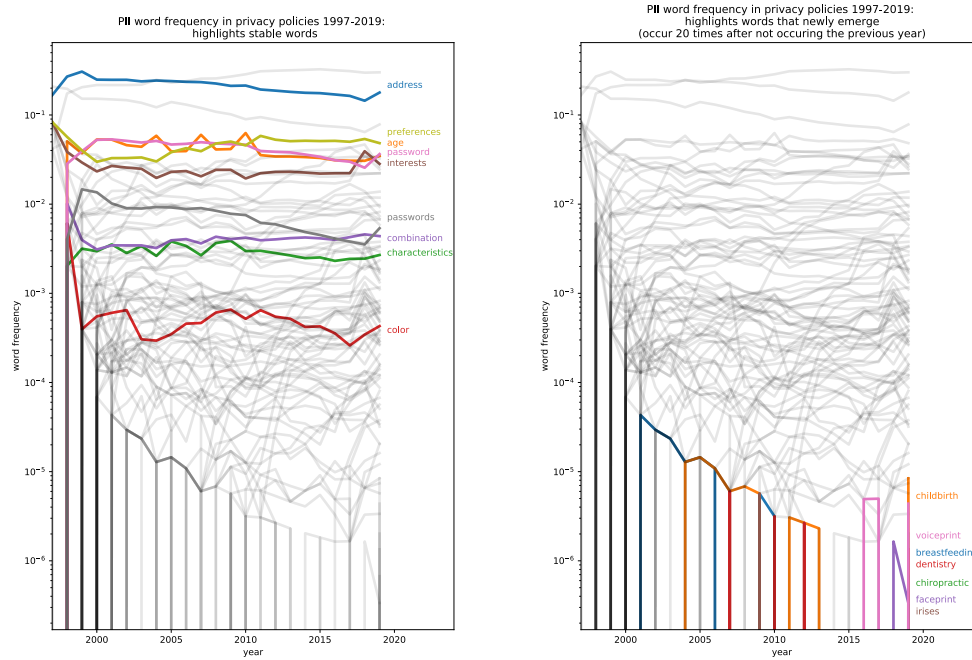


Figure 4.6: a.) Frequency distribution of stable words (change less than 2% in frequency over a 20-year period), b.) Frequency distribution of words that emerge (occur 20 times after not occurring the previous year)

place in the word space as seen in the rising words results.

Stability: Stability explores PII data types whose frequency distributions remain at similar frequencies over time with little change. Here stability is defined by the frequency distribution of words that change less than 2% in frequency over a 20-year period, as seen in Figure 4.6. Our results show that PII terms such as “address,” “preferences,” “age,” “password,” “interests,” “passwords,” “combination,” “characteristics,” and “color” remain steadily present in the privacy policies over time. These words generally represent consumer identifiers and characteristics which would be expected for DBDPs to collect while conducting regular business with data subjects.

Emergence: Emergence looks at new PII data types that appear in privacy policies over

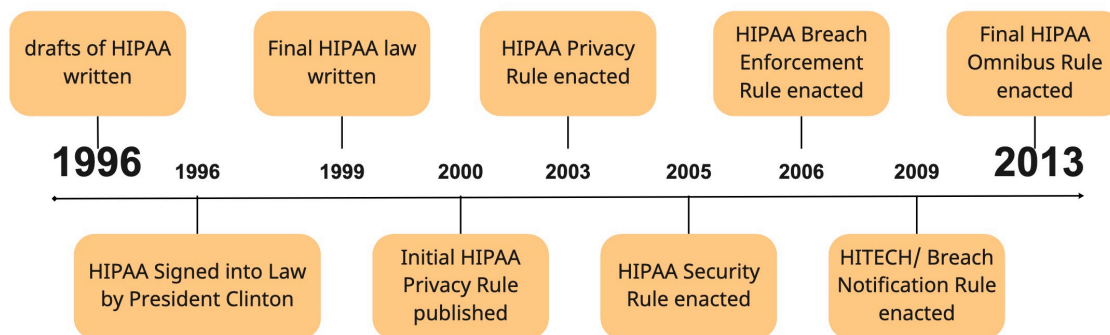


Figure 4.7: HIPAA Timeline and History

time. We define emergent words as words that occur at least 20 times in one year after not occurring the previous year, as seen in Figure 4.6. Our results show that PII-relevant terms such as “childbirth,” “voiceprint,” “breastfeeding,” “dentistry,” “chiropractic,” “faceprint,” “irises” newly emerge in the dataset primarily in the years 2016-2019. The emerging words appear to represent new biometric technologies and health-related words.

Frequency distribution of health-related words: We can use the frequency distribution of PII terms to explore groups of related words. We explore groups of words related to health and inference as an example to showcase how the time series of these frequency distributions may be used to understand sub-topics of PII data types better. Our first example explored PII data types related to health, and we compared the frequency distribution to potentially correlated legislative events. Our results show that a higher frequency distribution of words related to health and medicine may correlate with the Health Insurance Portability and Accountability Act (HIPAA) events. HIPAA restricts the healthcare industry from disclosing protected health PII without patient consent.

HIPAA has a long and complicated history, beginning with its introduction in 1996 and final enactment in 2013. Figure 4.7 shows a timeline of notable HIPAA events. Comparing our HIPAA timeline with frequency distributions of health-related words, as seen in Figure 4.8a, we do see spikes in the use of health-related words during periods when HIPAA

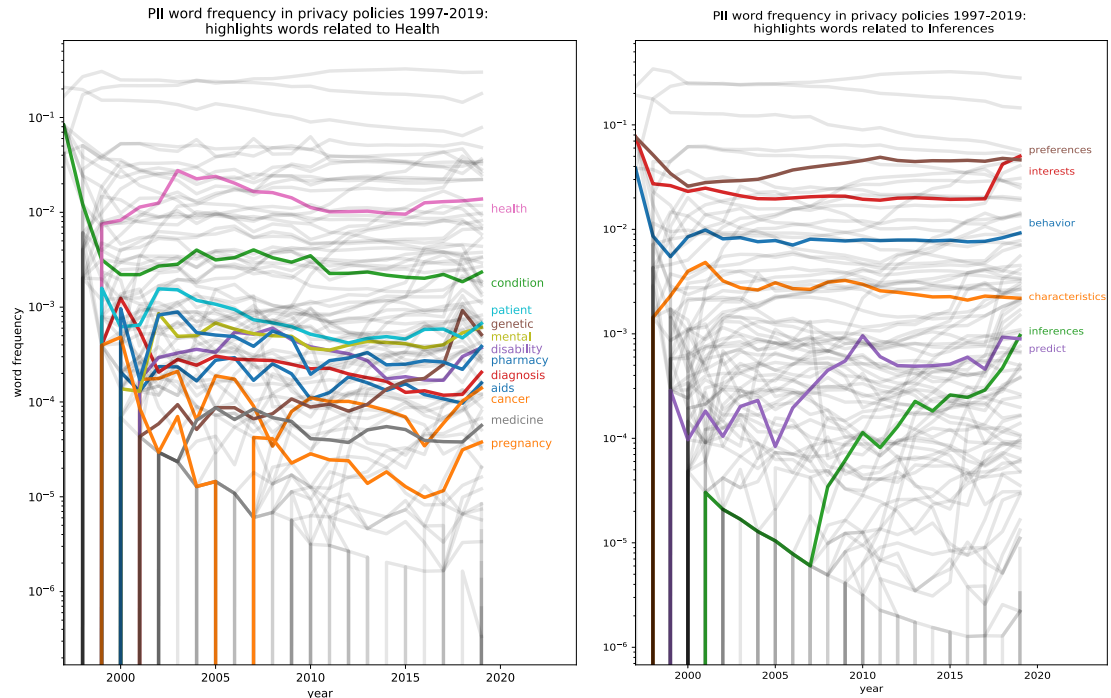


Figure 4.8: a.) Frequency distribution of words related to health, b.) Frequency distribution of words related to insights

regulations were being rolled out (note that there will be a natural delay between legislation being enacted or effective and privacy policies reacting to them). We can see words like “health,” “patient,” and “pharmacy” start to spike around 2002-2004. A notable exception in the frequency distribution behavior of health words is the word “genetic,” which rises steadily from 2001-2019.

Frequency distribution of insights-related words: As mentioned earlier, Insights are ways in which DBDPs can collect PII data and sell the insights gleaned from the PII data to gain from the data collection financially but avoid being classified as a data broker. In Figure 4.8b, we look at words related to inferences and insights. Most insight words are

used quite frequently compared to other words and remain fairly stable with two notable exceptions, “predict” is quite peaky and spikes to its highest point in 2010 but has been on the rise since its emergence in 1999. The term “inferences” emerged in 2001 and began to rise rapidly starting in 2007.

4.3.4 TOPIC LEVEL RESULTS: MEASURES OF COMPLEXITY

Topic Complexity: In result #2, we explore the complexity of the privacy policies as represented by the compression factor as seen in Figure 4.9 (full details can be seen in Appendix 4.5 Table 4.1). We see a trend of the compression factor [16] going down over time which can be interpreted as the privacy policies containing more text regularity [36, 11]. The policies being more compressible means that they look more textually similar to one another and, therefore, more compressible over time. The compression factor in our model signals the complexity of the privacy policies for that year (compression factor = bits of the Minimum Description Length (MDL) divided by bits of a non-compressed description). We see a fairly steady decrease in the complexity of the policies over time, which may be related to new regulatory pressures which require policies to disclose certain information in more precise ways (e.g., CCPA, GDPR) to comply. Note that our measure of complexity is from the perspective of machine readability and compressibility (more work would need to be done to conclude if this differs from human readability measures). In future work, we would like to test if we can use the compression factor and other machine-readable anomaly measures as a means to detect individual privacy policies or clusters of privacy policies which exhibit irregularities as a possible means of detecting non-compliance.

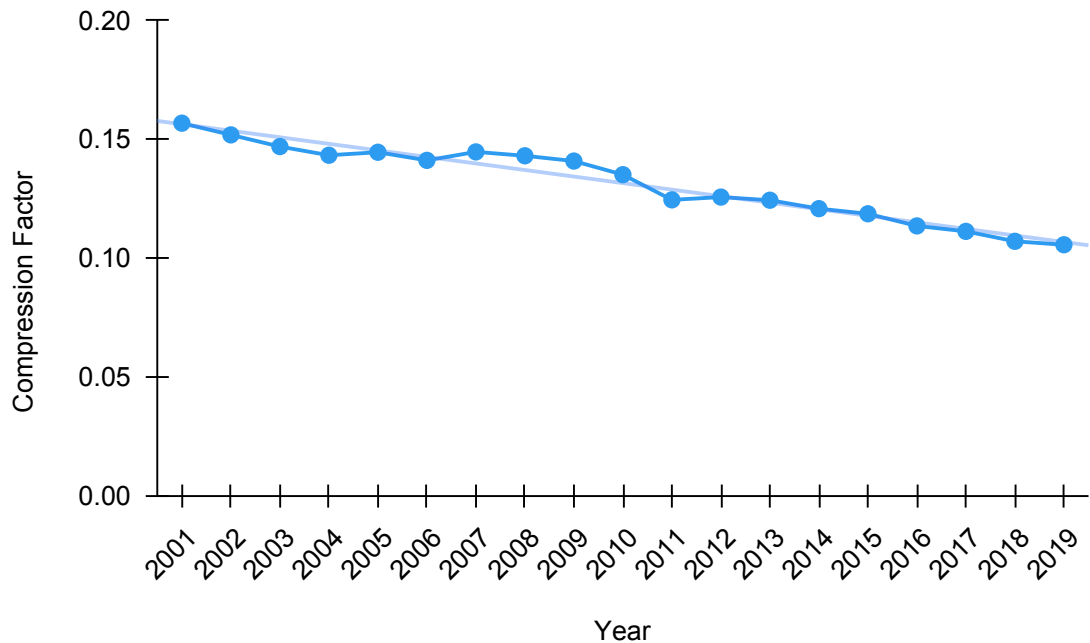


Figure 4.9: This figure represents privacy policies from 1997-2019 and their complexity measured by year via a compression ratio. The blue line indicates the compression factor from the full text of each year. To compute the compression ratio, we get the Minimum Description length (MDL) and convert the output from nats into bits ($\text{MDL bits} = \text{MDL nats} / \log_e 2$). We then get the size of the original file in bits which we calculate as the total number of words in the corpus multiplied by the log of the number of unique words in the corpus [16, 13]. To compute the compression ratio, we divide the MDL in bits by the original file size in bits. We then compare the compression ratio in bits over time for the privacy policies as extracted from year-by-year corpora using the hSBM topic model [13].

4.3.5 TOPIC LEVEL RESULTS: TOPIC PREVALENCE OVER TIME

In our time-series analysis of topic prevalence, we find several trends that confirm our previous findings (seen in Figure 4.10). We find “cookies” to dominate the topic landscape throughout the entire time frame. Manually entered PII attributes like address, name, and age show high prevalence but dwindle over time as newer tracking mechanisms emerge and enable the collection of more valuable data. Behavioral data like preferences and movement are the third most prevalent topic overall and rise slightly over time.

Interestingly we see the topic “beacons” rise rapidly from 2007 through 2009, which appears to correlate with the introduction of Facebook Beacon in 2007 [17] and the shutdown of Facebook Beacon after a class action lawsuit in 2009 that claimed the use of Facebook beacons violated user privacy. Notably, the topic prevalence (and word frequency distribution) of beacons remained fairly steady from 2011-2017, with dips occurring from 2018-2019. This points to updates to privacy policies being more often additive than subtractive - overly broad policies do not seem to cause serious problems to organizations.

In addition, we find “location-geolocation-latitude” rising significantly in prevalence from 2011 on, with slope increasing over time. This coincides with a steep rise in the availability of smartphones after the iPhone and the first Android models were introduced in 2008. The Android Open Source Project opened access to the smartphone market to smaller companies in 2010. Smartphones drove mobile usage of many digital services and contained GPS locators and thus making much richer location information available than stationary, or laptop computers did previously.

We find usage of the most prevalent topics, except “address-name-age”, stable over time, which is explained by the narrow purpose of an established, formal language used

Topic prevalence over time

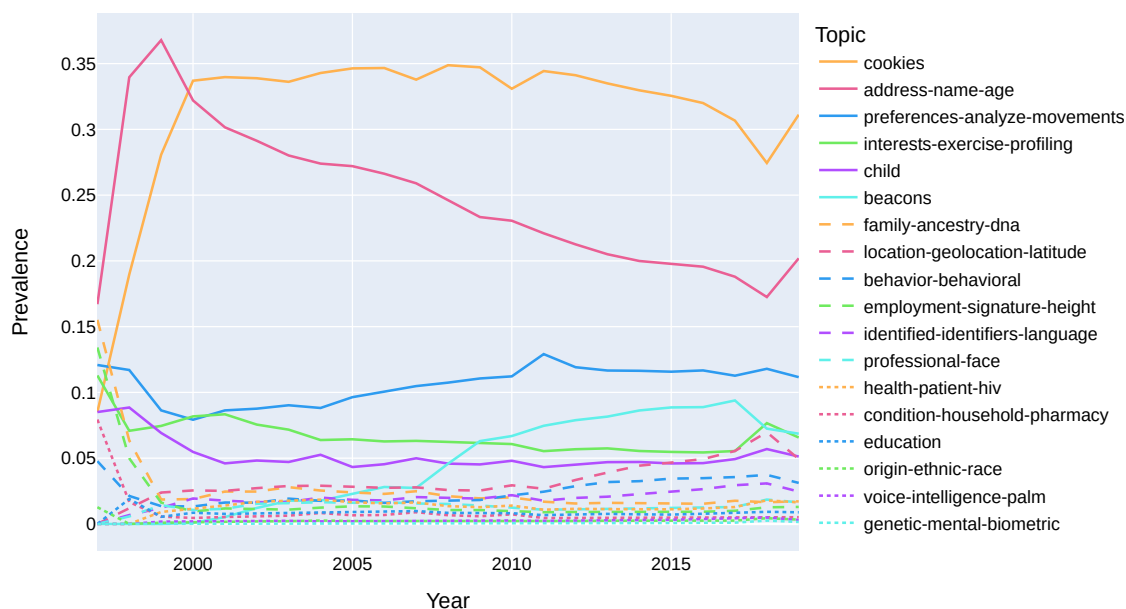


Figure 4.10: Topic prevalence 1997-2019. Topics extracted from the full corpus after negation filtering through hSBM topic modeling.

in privacy policies. The top five topics prevalent over time are as seen in Figure 4.10: cookies, address-name-age, preferences-analyze-movement, interests-exercise-profiling, and child. Throughout the entire time frame, the top two topics were prevalent. However, their prevalence decreased over time. This suggests that in 2019, privacy policies were used for a greater variety of purposes compared to 20 years ago. A list of vector representations of all analyzed topics can be found in Appendix 4.8.

4.3.6 NETWORK LEVEL RESULTS: MEASURES OF SENSITIVITY

In result #3, we investigate the level of sensitivity and risk (# of PII data types collected together) in the privacy policies as represented by the density of the word co-occurrence network graph in relation to modularity and the number of classes. We choose these two network measures to represent the sensitivity of the co-occurrence in the network because they represent the proportion that words are presented in privacy policies together and can be seen as a representation of the number of PII data types collected together.

Network density is defined as the proportion of the number of edges in a network compared to the number of potential edges between all pairs of nodes in the graph. Both the measure of network density and modularity will allow us to examine a fuller picture of the graph density qualified by the modularity density, which accounts for the network's density in relation to the size of the network.

We see that density in the network rises over time. That density and modularity show spikes during the critical time when new privacy legislation is introduced. As we use these network density measures as a representation of sensitivity, we can say that the co-occurrence of PII data type words does appear to be increasing. However, it is unclear whether this is a measure of increased concurrent PII data type collection activity or perhaps just a signal of compliance with new legislative mandates for more precise PII data type disclosure. It will be interesting, looking forward, to see whether this trend continues.

In Figure 4.11a, we can see between 2000-2019, there is a fairly steady increase in network density (we exclude the years 1997-1999 because their networks are too small). We can also see that there is an increase in density from 2013-2019. This makes sense in light of our timeline of legislative policy changes in the US and EU (as seen in fig. 4.4). During

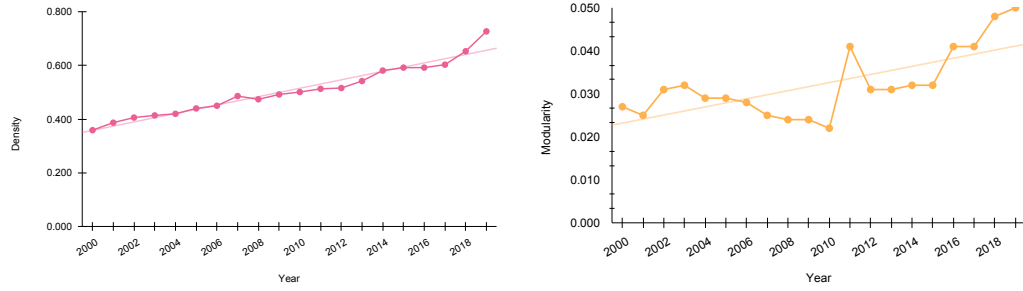


Figure 4.11: a.) 2000-2019 co-occurrence network of PII terms network density by year. b.) 2000-2019 co-occurrence network of PII terms network modularity by year.

2012-2019, we saw wide-sweeping data privacy laws come onto the scene, which requires that DBDPs begin to disclose more information (CCPA in 2018, for example, requires they list the data types they collect) in their privacy policies.

Network modularity is a scalar value that ranges from -1 to 1 and measures the density of links within communities compared to that of links between the communities [5]. Modularity measures the strength of the division of a network into classes. The modularity coefficient is usually computed by finding an optimal partition of the network into classes.

A high modularity score means the network has dense connections between nodes in the same class but sparse connections between nodes in different classes. In Figure 4.11b, we can see a rise in modularity between 2002-2005 (which coincides with the E-government act of 2002, the State Data Breach Notification Laws of 2003, and the California Online Privacy Protection Act of 2004). Modularity then went down again between 2006-2010 (which is a period with no new broad sweeping privacy laws), and then in 2011, there was a substantial spike. From 2012-2019, we see a steady yearly increase in modularity (which coincides with several large sweeping privacy laws such as GDPR and CCPA). Additional network measures can be seen in Appendix 4.5 Table 4.3.

We also analyze the degree and strength distributions of the co-occurrence graph for 2019 (our largest year), based on negation-filtered data, and compare them to the distri-

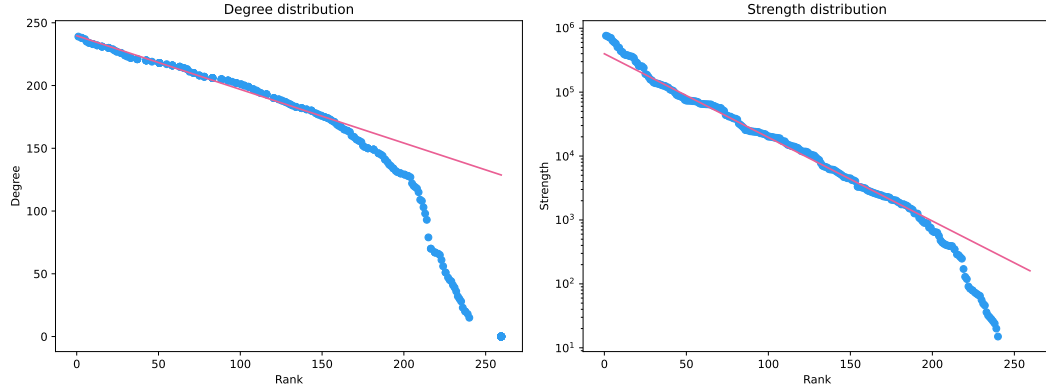


Figure 4.12: a.) The linear plot is a Zipf-ranked degree distribution of the 2019 co-occurrence network of PII terms. b.) Zipf-ranked strength distribution of the 2019 co-occurrence network of PII terms, logarithmic y axis

butions found by Fudolig et al. in [12]. They find power laws in degree and strength distributions in an analysis of word co-occurrence in tweets.

As shown in Figure 4.12, we do not find anything close to a power law in either. We find an approximately linear degree distribution and an approximately exponential strength distribution. Notable differences in the corpora that could lead to this difference are that (1) privacy policy documents are multiple orders of magnitude longer than tweets; (2) legal specialists write privacy policy documents to protect companies from being sued and thus contain much more specific and less varied language than tweets.

4.4 DISCUSSION

Our results can be summarized as follows: (1) privacy legislation appears to be associated with turbulence and rates of change of privacy policies use of PII data type terms; (2) complexity of privacy policies decreases in years, which means that policies are becoming more regular over time; (3) sensitivity rises over time and shows spikes during the critical time when new privacy legislation is introduced. We find an increase in mentions of health

and location data that is only partially explained by new legislation, which shows that new technologies with major privacy implications will be reflected in privacy policies even without changes in legislation. We also find evidence for insights as a mechanism for selling data without being classified as a data broker, which is becoming more common.

Looking at the basic properties of the dataset, we find that there is stability in the language and terms used in privacy policies over time and that the language is relatively consistent across companies, which makes this a promising dataset to extract more meaning from in future work.

The collection of PII data has become ubiquitous and poses a significant risk to data subjects. However, the lack of transparency about DBDPs who collect data on data subjects remains a large issue. Lack of transparency makes it difficult to assess the level of harm associated with harvesting PII data and hold DBDPs accountable for data misuse. We hope this study humbly contributes critical insight into an important body of work related to privacy policies and data broker and data processor transparency. Much work remains to be done in this area, and we hope this analysis offers a small step toward bringing light to this important issue.

In future work, we would like to investigate the flow of PII data types across sections of the privacy policies. With this future work, we hope to see what PII data types are present in the collect section of a privacy policy and compare that to the use and share section in privacy policies. We want to see what data is collected and then shared with third parties.

BIBLIOGRAPHY

- [1] (2018a). California Consumer Privacy Act 2018.
- [2] (2018b). Vermont Data Broker Act of 2018 9 V.S.A. 2430 et seq.
- [3] Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., and Mayer, J. (2021). Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of the Web Conference 2021*, WWW '21, pages 2165–2176, New York, NY, USA. ACM.

- [4] Birnhack, M. D. (2008). The EU Data Protection Directive: An engine of a global regime. *Computer Law & Security Review*, 24(6):508–520.
- [5] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.*, 2008(10):P10008.
- [6] Calo, R. (2013). Digital market manipulation. *Geo. Wash. L. Rev.*, 82:995.
- [7] Crain, M. (2016). The limits of transparency: Data brokers and commodification. *New Media & Society*, 20(1):88–104.
- [8] Cranor, L. (2003). P3P: Making privacy policies more useful. *IEEE Security & Privacy*, 1(6):50–55.
- [9] Earp, J., Anton, A., Aiman-Smith, L., and Stufflebeam, W. (2005). Examining Internet Privacy Policies Within the Context of User Privacy Values. *IEEE Trans. Eng. Manage.*, 52(2):227–237.
- [10] Feicheng, M. and Yating, L. (2014). Utilising social network analysis to study the characteristics and functions of the co-occurrence network of online tags. *Online Inform. Rev.*, 38(2):232–247.
- [11] Friedrich, R. (2021). Complexity and Entropy in Legal Language. *Aip. Conf. Proc.*, 9:671882.
- [12] Fudolig, M. I., Alshaabi, T., Arnold, M. V., Danforth, C. M., and Dodds, P. S. (2022). Sentiment and structure in word co-occurrence networks on Twitter. *Applied Network Science*, 7(1):1–27.
- [13] Gerlach, M., Peixoto, T. P., and Altmann, E. G. (2018). A network approach to topic models. *Sci. Adv.*, 4(7):eaq1360.
- [14] González Cabañas, J., Cuevas, A., and Cuevas, R. (2017). FDVT: Data valuation tool for Facebook users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3799–3809, Denver, CO, USA. ACM.
- [15] Graber, M. A., D Alessandro, D. M., and Johnson-West, J. (2002). Reading level of privacy policies on internet health web sites. *J. Fam. Practice*, 51(7):642–642.
- [16] Hébert-Dufresne, L., Young, J.-G., Daniels, A., and Allard, A. (2022). Network Onion Divergence: Network representation and comparison using nested configuration models with fixed connectivity, correlation and centrality patterns. *arXiv preprint arXiv:2204.08444*, pages 1–15.
- [17] Jamal, A., Coughlan, J., and Kamal, M. (2013). Mining social network data for personalisation and privacy concerns: a case study of Facebook’s Beacon. *International Journal of Business Information Systems*, 13(2):173–198.
- [18] Jensen, C. and Potts, C. (2004). Privacy policies as decision-making tools. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 471–478, Vienna, Austria. ACM.
- [19] Kirkpatrick, K. (2021). Monetizing your personal data. *Commun. ACM*, 65(1):17–19.
- [20] Latapy, M. (2008). Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theor. Comput. Sci.*, 407(1-3):458–473.
- [21] Linden, T., Khandelwal, R., Harkous, H., and Fawaz, K. (2020). The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47–

- [22] Lovato, J. L., Allard, A., Harp, R., Onaolapo, J., and Hébert-Dufresne, L. (2022). Limits of Individual Consent and Models of Distributed Consent in Online Social Networks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2251–2262, Seoul, South Korea. ACM.
- [23] McDonald, A. M. and Cranor, L. F. (2008). The cost of reading privacy policies. *Isjlp*, 4:543.
- [24] Newman, M. (2018). *Networks*. Oxford University Press, Oxford, UK.
- [25] Otto, P. N., Anton, A. I., and Baumer, D. L. (2007). The ChoicePoint Dilemma: How Data Brokers Should Handle the Privacy of Personal Information. *IEEE Security & Privacy Magazine*, 5(5):15–23.
- [26] Pollach, I. (2007). What’s wrong with online privacy policies? *Commun. ACM*, 50(9):103–108.
- [27] Price, M., Legrand, A. C., Brier, Z. M., and Hébert-Dufresne, L. (2019). The symptoms at the center: Examining the comorbidity of posttraumatic stress disorder, generalized anxiety disorder, and depression with network analysis. *J. Psychiatr. Res.*, 109:52–58.
- [28] Roderick, L. (2014). Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry. *Critical Sociology*, 40(5):729–746.
- [29] Schermer, B. W., Custers, B., and van der Hof, S. (2014). The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics Inf. Technol.*, 16(2):171–182.
- [30] Schouten, R. and Kumer, K. D. (2017). *Informed Consent*. Oxford University Press, Oxford, United Kingdom.
- [31] Serrano, M. A., Boguñá, M., and Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci.*, 106(16):6483–6488.
- [32] Solove, D. J. (2004). *The Digital Person*, volume 1. New York University Press, New York, NY.
- [33] Srinath, M., Wilson, S., and Giles, C. L. (2020). Privacy at scale: Introducing the privaseer corpus of web privacy policies. *arXiv preprint arXiv:2004.11131*, pages 1–10.
- [34] Stallings, W. (2020). Handling of Personal Information and Deidentified, Aggregated, and Pseudonymized Information Under the California Consumer Privacy Act. *IEEE Security & Privacy*, 18(1):61–64.
- [35] Susser, D., Roessler, B., and Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, 4:1.
- [36] Van den Bosch, A., Content, A., Daelemans, W., and De Gelder, B. (1994). Measuring the complexity of writing systems*. *J. Quant. Linguist.*, 1(3):178–188.
- [37] Veling, A. and Van Der Weerd, P. (1999). Conceptual grouping in word co-occurrence networks. In *IJCAI*, volume 99, pages 694–701, Dorpsweg 78, 1697 KD Schellinkhout, The Netherlands.
- [38] Wagner, I. (2023). Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021. *ACM Transactions on Privacy and Security*, pages 1–33.
- [39] Wagner, I. and Boiten, E. (2018). Privacy Risk Assessment: From Art to Science, by

- Metrics. In *Lecture Notes in Computer Science*, pages 225–241. Springer International Publishing, Cham, Switzerland.
- [40] Zaeem, R. N. and Barber, K. S. (2020). The Effect of the GDPR on Privacy Policies. *ACM Trans. Manage. Inf. Syst.*, 12(1):1–20.
- [41] Zhang, L. and Peixoto, T. P. (2020). Statistical inference of assortative community structures. *Phys. Rev. Research*, 2(4):043271.

4.5 SUPPLEMENTARY MATERIALS

4.5.1 SUMMARY STATISTICS: FULL CORPUS, PII DATA TYPES CORPUS, NETWORK ANALYSIS

In this section, we outline three tables with summary statistics: (1) summary statistics of the entire privacy policy corpus can be seen in Table 4.1; (2) summary statistics of the PII privacy policy filtered corpus can be found in Table 4.2; and summary statistics of the network properties of the co-occurrence networks can be found in Table 4.3.

4.5.2 RESEARCH METHODS: WORD LEVEL

We first filter the corpus for both analyses to extract the terms in our PII data types lexicon. We then separate the privacy policy text by year, with 1997 as the first in the series and 2019 as the final corpus in the comparison. Frequency distribution is defined as the total count of a particular PII data type term and how many times it is used in a given corpus/total count of all words used in the given corpus.

Rising looks at PII data types whose frequency increases quickly over time. We define rising words as words that rise in frequency by more than ten times in a 7-year period. Falling looks at PII data types whose frequency drops quickly over time. Falling words are

Table 4.1: Overview of privacy policies corpus: the full corpus that is cleaned. Minimum description length (MDL) from the hSBM topic model, text description length (TDL) = total number of words x ($\log_2(\text{unique words})$), compression factor = MDL/TDL. UW = unique words, CF = Compression Factor, PP = number of privacy policies in the corpus.

Year	Words	UW	PP	MDL	TDL	CF
1997	4743	1148	9	8516.52666	48212.15355	0.1766468832
1998	80529	3901	144	191616.9197	960681.036	0.199459459
1999	413998	8059	602	991443.7243	5372197.486	0.1845508708
2000	2452053	19883	2886	5528846.65	35013472.48	0.1579062646
2001	4089391	26230	4442	9399703.064	60027885.0	0.156588943
2002	6009825	32451	6161	13659820.04	90063089.1	0.1516694594
2003	7558026	36958	7608	16829015.37	114682456.3	0.1467444622
2004	10051958	43265	9771	22154128.03	154809328.4	0.1431058984
2005	12485455	49056	11740	28084926.16	194550132.5	0.1443582988
2006	16721665	58317	15307	37314163.2	264731194.6	0.1409511382
2007	21688278	67080	19093	50251529.0	347741067.2	0.1445084683
2008	26488167	76818	23379	61419768.27	429880615.1	0.1428763385
2009	31196291	83596	27277	71748737.18	510095117.9	0.1406575649
2010	40154734	96377	34760	89700229.98	664817889.1	0.1349245131
2011	56365399	111383	46260	117489747.6	944975469.9	0.1243310026
2012	63254136	125455	50572	134453798.0	1071323311	0.1255025412
2013	72057214	138920	55591	152872571.3	1231017863	0.1241838773
2014	80135508	143880	59937	165640922.4	1373082386	0.1206343655
2015	88188734	148063	63820	179455129.4	1514716611	0.1184743919
2016	96232665	150585	66029	187656253.0	1655222869	0.1133721969
2017	94350462	146704	60343	179866323.9	1619294397	0.1110769754
2018	98640479	146546	49514	180949278.4	1692768675	0.1068954555
2019	114104999	163745	51791	208407776.4	1976423109	0.1054469438

defined as words that fall in frequency by more than 15% in a year. Stability explores PII data types whose frequency distributions remain at similar frequencies over time with little change. Here stability is defined by the frequency distribution of words that change less than 2% in frequency over a 20-year period. Emergence looks at new PII data types that appear in privacy policies over time. We define emergent words as words that occur at least 20 times in one year after not occurring the previous year.

Table 4.2: Overview of privacy policies PII data types corpus: includes a corpus filtered for just PII data type terms.

Year	No. words	Unique words	No. policies
1997	25	13	5
1998	494	27	128
1999	2526	45	546
2000	14522	64	2671
2001	23182	71	4156
2002	33963	76	5723
2003	42670	76	7059
2004	77980	78	9079
2005	68965	78	10912
2006	91992	78	14321
2007	165744	80	17940
2008	146335	80	21983
2009	176552	84	25719
2010	316706	85	32996
2011	326307	85	44209
2012	374757	86	48449
2013	433677	84	57515
2014	491032	83	57822
2015	547066	84	61707
2016	609593	84	64064
2017	602959	86	58672
2018	607413	86	48337
2019	2972358	92	50576

4.5.3 RESEARCH METHODS: TOPIC LEVEL

Stochastic block model topic model:

We conduct our topic modeling by first filtering all of the corpora to extract just the terms in our PII lexicon. We do this for the entire corpus, which includes privacy policies from 1997-2019, and then split it by year to compare individual years to the whole text. The filtering identifies words that are present in our PII lexicon and deletes all words that are not in the lexicon. The result leaves the word frequency of the PII terms and their relative

placement in the text. We run the filtered text through a stochastic block model topic model to extract topic groups in the corpora.

To understand what topic appears in the corpus, we will use the hSBM topic model implemented in Graph-tool [13]. This Stochastic Block Model is a generative model that generalizes the Erdős-Rényi model to have groups. It is a random Poisson graph model in that node degrees within any group are distributed according to a Poisson distribution [24]. The SBM is widely used in complex systems because it generates different network structures (e.g., core-periphery, community structure, and hub-and-spoke). Not only is the SBM flexible, but it is also extensible, as we can see below with the SBM Topic Model. SBM topic model is a method that combines topic modeling and community detection through a word-document matrix (which is a bipartite network) that assumes communities represent topics. It uses maximum likelihood estimation to find a hierarchical clustering that fits the data.

A topic model is a method of extracting high-level information from textual data. In the hSBM topic model, we infer topics through the community structure of the document to word groupings. The hSBM model for topic modeling is a non-parametric symmetrical formulation that hierarchically clusters words and documents in a corpus. The tool divides words and documents into hierarchical groups with a list of constituent words per topic and the weight of the contribution of each of those words.

Complexity measure:

From the graph-tool hSBM [13], we get the Minimum Description Length (MDL) and convert the output of nats into bits (MDL bits = MDL nats x 1.4426950408889). We then get the size of the original file, which we calculate as the total number of words in the corpus multiplied by the \log_2 of the number of unique words in the corpus [16]. To compute the compression ratio, we divide the MDL in bits by the original file size in bits. We then

compare the compression ratio in bits over time for the privacy policies.

4.5.4 RESEARCH METHODS: NETWORK LEVEL

Co-occurrence networks:

To investigate what words are used in the corpus and later to see what data types are collected concurrently by DBDPs, we use co-occurrence network analysis to represent the use of relevant terms within the same text. We first filter all the corpora to extract the terms in our PII lexicon. co-occurrence networks utilize network analysis to represent the relationship between objects that co-occur in the same environment. [27, 10, 37, 12] To create a word co-occurrence network, we add an edge between words that appear in the same privacy policy. Edges start with a weight of one. If any two terms from our PII lexicon appear together in additional documents, we increase the weight based on the number of documents they co-occur.

To process our network visualizations, we remove all nodes with a degree less than one (words that do not co-occur with other words). Then networks are partitioned by modularity, which is signified by the color of the nodes. [5] Finally, the node size ranges from 1 to 40 based on the weighted degree. Most of our co-occurrence networks are partitioned into 2-4 classes. The co-occurrence networks seem to partition words into groups that can be generalized into two themes: PII data about a data subject’s internet activity and PII data about the data subject’s characteristics. An example of this can be seen in Figure 4.2. In future work, we would like to take a closer look at the co-occurrence networks and investigate the dynamics of individual node attributes and how they change over time.

For visualization, we use a network backbone algorithm developed by Serrano et al. [31] to reduce the number of edges while preserving the relevant network structure. The algorithm works with a random network null model to compute the statistical significance of

each link and drops links if their significance is below a certain threshold.

4.6 LEXICON OF PERSONALLY IDENTIFIABLE INFORMATION (PII)

‘identifiers’, ‘alias’, ‘online identifier’, ‘internet protocol (ip) address’, ‘account name’, ‘social security number’, ‘passport number’, ‘customer records information’, ‘identification number’, ‘signature’, ‘electronic mail address’, ‘address’, ‘telephone number’, ‘protected health information’, ‘state identification card number’, ‘education’, ‘employment history’, ‘bank account number’, ‘face’, ‘financial information’, ‘records of personal property’, ‘products purchased’, ‘health condition’, ‘consuming histories’, ‘services purchased’, ‘eye color’, ‘retina scans’, ‘network activity’, ‘internet activity’, ‘search history’, ‘geolocation’, ‘visual’, ‘thermal’, ‘olfactory’, ‘professional’, ‘medical condition’, ‘characteristics’, ‘aggregated data’, ‘predispositions’, ‘behavior’, ‘specific location’, ‘aptitudes’, ‘facial recognition’, ‘physiological’, ‘behavioral’, ‘audio’, ‘dna’, ‘iris’, ‘retina’, ‘hand’, ‘palm’, ‘vein patterns’, ‘voice recordings’, ‘minutiae template’, ‘health status’, ‘keystroke patterns’, ‘keystroke rhythms’, ‘gait rhythms’, ‘sleep’, ‘health’, ‘exercise’, ‘cross-context behavioral advertising’, ‘targeted advertising’, ‘dark pattern’, ‘personal information’, ‘racial’, ‘real name’, ‘account number’, ‘postal address’, ‘financial account number’, ‘internet protocol address’, ‘gender identity’, ‘email address’, ‘security question’, ‘color’, ‘religion’, ‘sex’, ‘sexual orientation’, ‘marital status’, ‘national origin’, ‘ancestry’, ‘genetic information’, ‘retaliation for reporting patient abuse in tax-supported institutions’, ‘age’, ‘religious dress’, ‘pregnancy’, ‘gender’, ‘childbirth’, ‘breastfeeding’, ‘mental characteristics’, ‘physical characteristics’, ‘hiv/aids’, ‘cancer’, ‘genetic characteristics’, ‘geolocation data’, ‘record of cancer’, ‘history of cancer’, ‘gender expression’, ‘abilities’, ‘mental condition’, ‘predict’, ‘biological’, ‘purchasing tenden-

cies', 'aggregate consumer information', 'first name', 'voice', 'electronic network activity',
 'biological characteristic', 'interaction with an advertisement', 'browsing history', 'employ-
 ment', 'race', 'health records', 'citizenship', 'military or veteran status', 'medical identifica-
 tion number', 'access code', 'preferences', 'protected classifications', 'psychological trends',
 'commercial information', 'medical information', 'attitudes', 'intelligence', 'name', 'driver's
 license', 'sensitive personal information', 'precise geolocation', 'locate', 'geographic area',
 'radius', 'sensitive data', 'profiling', 'consumer's social security', 'driver's license', 'state
 identification card', 'account log-in', 'credit card', 'health insurance information', 'pass-
 word', 'credentials allowing access to an account', 'combination', 'racial origin', 'ethnic ori-
 gin', 'philosophical beliefs', 'union membership', 'text messages', 'genetic data', 'biometric
 data', 'personally identifiable information', 'security code', 'fingerprint', 'device identifier',
 'ip address', 'cookies', 'beacons', 'pixel tags', 'customer number', 'unique pseudonym', 'tele-
 phone numbers', 'persistent identifier', 'probabilistic identifier', 'family', 'child', 'identifier
 template', 'de-identified data', 'health-care information', 'health-care provider', 'medicine',
 'pharmacy', 'chiropractic', 'nursing', 'physical therapy', 'podiatry', 'dentistry', 'optometry',
 'occupational therapy', 'healing arts', 'identified', 'identifiable individual', 'online identi-
 fier', 'personal data', 'products obtained', 'automated processing', 'request for pregnancy
 disability leave', 'analyze', 'economic situation', 'personal preferences', 'religious beliefs', 're-
 liability', 'location', 'movements', 'religious beliefs', 'physical health condition', 'diagnosis',
 'credit card number', 'postal address', 'citizenship status', 'genetic', 'last name', 'mobile
 ad identifiers', 'health-care', 'patient', 'fingerprints', 'products considered', 'physical de-
 scription', 'voiceprint', 'eye retinas', 'unique identifier', 'consuming tendencies', 'faceprint',
 'driver's license number', 'services considered', 'global positioning system', 'latitude', 'longi-
 tude', 'coordinates', 'interests', 'financial account', 'user alias', 'irises', 'request for leave for
 an employee's own serious health condition', 'condition', 'physical', 'diagnosis', 'insurance
 policy number', 'immigration status', 'known child', 'sex life', 'height', 'aids', 'medical diag-

nosis', 'religious grooming practices', 'identifiable individual', 'debit card', 'ethnic', 'origin', 'medical treatment', 'inferences', 'medical history', 'mental health', 'physical health', 'mental', 'biometric information', 'email content', 'physical representation', 'biological pattern', 'mother's maiden name', 'interaction with an internet website application', 'deoxyribonucleic acid', 'purchasing histories', 'disability', 'targeting of advertising', 'movements', 'hair color', 'digital representation', 'initials', 'specific geolocation data', 'driver authorization card number', 'identification card number', 'debit card number', 'health insurance identification number', 'user name', 'request for family care leave', 'date of birth', 'place of birth', 'unique biometric', 'human body', 'hiv', 'biometric', 'language', 'household', 'driver's license number', 'government-issued identification number', 'driver license', 'nondriver state identification card number', 'individual taxpayer identification number', 'military identification card number', 'gait patterns', 'unique personal identifier', 'passwords', 'personal identification number', 'services obtained', 'wellness program', 'health promotion', 'disease prevention', 'health insurance policy number'

4.7 LEXICON OF NEGATION WORDS

"don't", "never", "nothing", "nowhere", "noone", "none", "not", "hasn't", "hadn't", "can't", "couldn't", "shouldn't", "won't", "wouldn't", "don't", "doesn't", "didn't", "isn't", "aren't", "ain't", "in*", "un*", "dis*", "mal*"

4.8 TOPICS

All topics we trace over time, as extracted from the full corpus, omitting words with weight < 0.001 :

address-name-age: address (0.5), name (0.241), age (0.097), password (0.085), physical (0.038), gender (0.019), passwords (0.013), locate (0.004), reliability (0.001)

cookies: cookies (1.0)

preferences-analyze-movements: preferences (0.71), analyze (0.252), movements (0.038)

interests-exercise-profiling: interests (0.664), exercise (0.292), profiling (0.045)

employment-signature-height: employment (0.789), signature (0.131), height (0.045), citizenship (0.036)

professional-face: professional (0.901), face (0.099)

beacons: beacons (1.0)

location-geolocation-latitude: location (0.926), geolocation (0.039), latitude (0.008), longitude (0.008), coordinates (0.007), inferences (0.006), fingerprint (0.004), radius (0.002)

identified-identifiers-language: identified (0.323), identifiers (0.223), language (0.207), combination (0.079), characteristics (0.056), audio (0.044), hand (0.029), color (0.013), visual (0.011), predict (0.011), attitudes (0.002)

behavior-behavioral: behavior (0.561), behavioral (0.439)

family-ancestry-dna: family (0.912), ancestry (0.056), dna (0.032)

condition-household-pharmacy: condition (0.439), household (0.271), pharmacy (0.083), cancer (0.061), medicine (0.054), diagnosis (0.047), aids (0.025), initials (0.016), dentistry (0.002), chiropractic (0.001)

origin-ethnic-race: origin (0.278), ethnic (0.155), race (0.155), sex (0.149), racial (0.107), religion (0.099), disability (0.057)

health-patient-hiv: health (0.951), patient (0.046), hiv (0.003)

education: education (1.0)

child: child (1.0)

genetic-mental-biometric: genetic (0.265), mental (0.261), biometric (0.238), physiological (0.075), sleep (0.06), pregnancy (0.054), fingerprints (0.021), biological (0.01), breast-

feeding (0.005), predispositions (0.003), retina (0.002), olfactory (0.002), childbirth (0.002),
voiceprint (0.002)

voice-intelligence-palm: voice (0.645), intelligence (0.257), palm (0.052), nursing (0.036),
iris (0.01)

4.9 CODE AVAILABILITY STATEMENT

Code associated with this project can be found on our [Github repository](#).

Table 4.3: Network analysis on privacy policy text from 1997-2019. Metrics on the graph of co-occurrences of PII lexicon words for each year individually and also all years. All nodes with a degree less than one are filtered from the network (words that do not co-occur with other words). Here we explore the network measures of modularity [5], blocks [41], average clustering coefficients [20], average degree, and density. ACC = average clustering coefficient, Mod = modularity.

Year	Nodes	Edges	Mod	Blocks	ACC	Degree	Density
1997	26	192	0.199	3	0.891	14.7690	0.591
1998	65	743	0.096	4	0.798	22.8620	0.357
1999	104	1748	0.029	3	0.828	33.6150	0.326
2000	143	3636	0.027	2	0.837	50.9930	0.359
2001	157	4742	0.025	2	0.833	60.4080	0.387
2002	168	5697	0.031	2	0.830	67.8210	0.406
2003	117	6454	0.032	3	0.824	72.9270	0.414
2004	185	7156	0.029	3	0.830	77.3620	0.420
2005	187	7660	0.029	3	0.829	81.9250	0.440
2006	192	8255	0.028	3	0.835	85.9900	0.450
2007	190	8725	0.025	3	0.833	91.8320	0.486
2008	198	9249	0.024	2	0.837	93.4240	0.474
2009	201	9885	0.024	2	0.835	98.3580	0.492
2010	205	10469	0.022	3	0.839	102.1370	0.501
2011	216	11913	0.041	3	0.843	110.3060	0.513
2012	221	12533	0.031	3	0.848	113.4210	0.516
2013	220	13045	0.031	2	0.850	118.5910	0.542
2014	216	13498	0.032	2	0.853	124.9810	0.581
2015	219	14127	0.032	2	0.855	129.0140	0.592
2016	220	14253	0.041	2	0.860	129.5730	0.592
2017	221	14670	0.041	3	0.860	132.7600	0.603
2018	221	15868	0.048	2	0.874	143.6020	0.653
2019	240	20847	0.050	3	0.894	173.7250	0.727
All Years	251	22334	0.039	2	0.894	177.9600	0.712

CHAPTER 5

CONCLUSION

5.1 PRÉCIS OF THE THESIS

This thesis gives a practical and theoretical overview of data ethics, privacy, and security using three critical case studies that showcase the need for a mesoscale understanding of data ethics, privacy, and security. As previous works in this field primarily focus on the individual and societal dimensions of privacy and security, this thesis focused on the group level to address significant privacy and security gaps. Future work will combine these multiscale analyses in a more unified framework that addresses all levels. The chapters of this thesis can be summarized as follows:

5.1.1 SUMMARY OF GROUP PRIVACY: DISTRIBUTED CONSENT

In socially-networked digital environments, personal data can be easily exposed to non-consented access when a user grants access to their profile. This means that the traditional informed and individual consent model may not be suitable for social networks where obtaining informed consent from all users affected by data processing may not be possible.

Additionally, information is distributed across users, making it difficult to obtain individual consent. We propose two new models of consent for data transactions: a platform-specific model called “distributed consent” and a cross-platform model known as a “consent passport.” Both models allow individuals and groups to coordinate their consent based on the consent of their network connections. We have tested the impact of these models on social networks and found that low adoption rates would allow macroscopic subsets of networks to preserve their connectivity and privacy.

In chapter 2, we question the traditional idea of individual consent to control informational privacy in online environments where information is distributed, networked, and affects network neighbors. We also investigate the privacy settings of social media platforms and how they could be modified to safeguard groups and limit visibility from third-party surveillance. Although these new security settings may help individuals protect their data from being leaked to third parties within a single platform, when working across platforms, security settings must be coordinated to ensure that the data subject remains unnoticed by their network neighbors. Consent may not be a practical mechanism for safeguarding individual or group-level data online or across multiple platforms. Instead, more appropriate mechanisms for protecting group privacy involve legal regulation of these platforms and third parties when coordination costs are high and more tools for collectively sharing consent decisions when coordination costs are low.

5.1.2 SUMMARY OF GROUP CORRECTION: DIVERSE MISINFORMATION

In chapter 3, we delve into the topic of group coordination in order to protect against the harms of online misinformation through group correction. Our investigation centers around the connection between a group’s diversity and its vulnerability to misinformation

on social networking platforms. Additionally, we discuss the ethical concerns associated with manipulating a data subject’s information through deepfakes.

Chapter 3 explored how different biases and demographics can make social media users more or less prone to falling for certain types of misinformation. Specifically, the study focuses on “diverse misinformation,” which refers to the complex interplay between human biases and demographics represented in false information. To investigate this phenomenon, we analyzed how people classified computer-generated videos (deepfakes) as a form of diverse misinformation. Deepfakes were chosen as a case study due to their objective classification as misinformation, the ability to control the demographics of the personas presented, and potential real-world harms.

The study surveyed 2,016 US-based participants exposed to videos and asked questions about their attributes without knowing that some might be deepfakes. The analysis found that accuracy in identifying deepfakes varied significantly by demographics and that participants were generally better at classifying videos that matched their demographics (especially for white participants). The study then used an idealized mathematical model to explore how diverse social groups might help reduce susceptibility to misinformation. The model suggests that diverse contacts may provide “herd correction,” where friends can protect each other’s blind spots. Overall, this study highlights the importance of considering human biases and the attributes of misinformation when trying to reduce susceptibility to false information.

5.1.3 SUMMARY OF GROUPED DATA: ISSUES IN DATA AGGREGATION ON USERS

Collecting personally identifiable information (PII) on individuals has become a lucrative industry, with data brokers and processors profiting from buying and selling consumer data.

Unfortunately, the lack of transparency in this industry makes it challenging to understand what types of data are being collected, used, and sold, which puts individuals at risk. To better understand the data collection activities of data brokers and processors, we conducted a study using a large dataset of privacy policies from 1997 to 2019. Additionally, we created a lexicon of PII-related terms based on legislative texts to aid in our analysis. Our findings show that privacy legislation may affect the stability and turbulence of PII data types in privacy policies. Over time, privacy policies have become less complex and more regularized and sensitive, with spikes in sensitivity occurring when new privacy legislation is introduced.

In chapter 4, we explored various tools and techniques to examine the privacy policies of platforms. This helps determine the risks and potential harms of aggregating different types of personally identifiable information (PII) about individuals. Our focus was on ways to increase transparency regarding the activities of data processors. We discussed methods for investigating privacy policies to determine what types of PII data brokers claim to collect, use, and sell. Additionally, we introduced techniques that can measure changes in collection practices over time and the complexity and sensitivity of the data types collected. Using Natural Language Processing (NLP) measurements, we can address some of the consent fatigue issues highlighted in Chapter 2.

5.2 LIMITATIONS AND FUTURE WORK

One considerable limitation of this work is that notions of groups essentially still seen as a collection of individuals may miss the emergence of complex behavior exhibited by groups over time. Future work will look at complex notions of group social ontology and group epistemology [5] to better understand how groups make collective decisions and to what extent they can make decisions as a collective. It is a question however to what extent a group is a metaphysical agent that can make decisions.

Future work will focus on the following areas:

Group Technological Design: What would technology look like if designed by communities with more group-minded or collectivist values [6, 7, 8, 2, 4, 3]? This work will utilize methods such as designed fiction [1] to work with communities to imagine and co-design systems from another perspective.

Group decision making and ownership: This project will explore technological means by which groups can effectively control data and make decisions [9, 10] as a group about shared data. Examples of critical shared data or information might be group data ownership of culturally shared digital assets (e.g., ancestral knowledge, genetic sequences). This will require understanding the trade-off between coordination costs and the technical challenges of collectively controlling the flow of information as a group.

The technological mechanism may have different forms depending on the nature of the data and the aims. The aim, for example, may be to use collective ownership to secure the longevity and accessibility of the open data for as long as possible. On the other hand, if sensitive, the aim of group ownership may be to limit the flow of data for cases when the information could harm group members.

In the open data preservation case (like in the case of a public library), this is a group storage mechanism, but for sensitive material like genetic data or sensitive religious, cultural information, in order to get the entire digital asset elected members of the collective need a mechanism to consent to release the material.

Both of these technological aims have potential issues that must also be considered in the design. For example, how to handle ransoming the asset by one individual when shared by a group? Pieces must be small enough and redundant so that one person cannot ransom the asset and hold majority control.

5.3 DISCUSSION

The conceptualization of privacy has come mainly from the perspective of the individual. More narrative work on what these values look like to groups needs to be done. In general, most of the powerful technology dominating the digital world was created and controlled by individuals from similar demographic backgrounds. It is crucial to create counterfactual narratives about what this technology could look like as an avenue for future development and recognition that these forms of technology do not align with the values of all people. More work needs to be done to include other cultures' voices in the technology design process.

Moreover, until we as a society start to take the digital world seriously as part of the civil society that we should have a duty to protect, we will be vulnerable in that space. Overcoming conceptual barriers such as physicality is a difficult leap for many. However, the standard threats to our society coming from digital spaces have become large enough that we must drastically change how we regulate technology and protect the digital world. It is not acceptable to force individuals to take on the duty of protecting themselves in the digital realm against tech Goliaths. This is the duty of those regulators of the State, and it is a monumental failure with significant consequences that need to be addressed immediately for our civil society to flourish and persist. Those with this duty: allowing the technological firms and data processors that are generating harm for profit to monitor and regulate themselves, will indeed be seen as one of the most significant failings in history if left to continue.

BIBLIOGRAPHY

- [1] Blythe, M. (2014). Research through design fiction: narrative in real and imaginary abstracts. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 703–712.

- [2] Dillon, G. L. (2016). Indigenous futurisms, Bimaashi Biidaas Mose, flying and walking towards You. *Extrapolation.*, 57(1/2):1.
- [3] Duarte, M. E. (2017). *Network sovereignty: Building the internet across Indian country*. University of Washington Press.
- [4] Kukutai, T. and Taylor, J. (2016). *Indigenous data sovereignty: Toward an agenda*. ANU press.
- [5] Lackey, J. (2021). *The epistemology of groups*. Oxford University Press, USA.
- [6] Lewis, J. E. (2013). The future imaginary. *Presentation at TEDxMontreal, Montreal, Que., September, 14*.
- [7] Lewis, J. E. (2016). A brief (media) history of the indigenous future. *Public*, 27(54):36–50.
- [8] Lewis, J. E., Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., et al. (2020). Indigenous protocol and artificial intelligence position paper.
- [9] Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.
- [10] Schneider, N., De Filippi, P., Frey, S., Tan, J. Z., and Zhang, A. X. (2021). Modular politics: Toward a governance layer for online communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26.

FULL REFERENCE LIST IN ALPHABETICAL ORDER

COMPLETE BIBLIOGRAPHY

- [1] Social Media Fact Sheet. *Pew Research Center*, 2019.
- [2] Vermont Data Broker Act of 2018 9 V.S.A. § 2430 et seq.
- [3] Legislature of the State of Texas. § Senate Bill No. 751, 2019.
- [4] Pew Research Center. Social media fact sheet. *Pew Research Center: Washington, DC, USA*, 2021.
- [5] The Committee on the Judiciary House of Representatives. Federal Rules of Evidence, 2019.
- [6] The People of the State of California. Assembly Bill No. 602 to the Civil Code, relating to privacy, § 1708.86. 2019, 2019.
- [7] Alessandro Acquisti, Leslie K. John, and George Loewenstein. What Is Privacy Worth? *The Journal of Legal Studies*, 42(2):249–274, June 2013.
- [8] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.
- [9] Edo M Airolidi, David Blei, Stephen Fienberg, and Eric Xing. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21, 2008.
- [10] Antoine Allard, Laurent Hébert-Dufresne, Jean-Gabriel Young, and Louis J. Dubé. Coexistence of Phases and the Observability of Random Graphs. *Phys. Rev. E*, 89:022801, February 2014.
- [11] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36):eabf4393, 2021.
- [12] Irwin Altman. Privacy regulation: Culturally universal or culturally specific? *Journal of social issues*, 33(3):66–84, 1977.
- [13] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. Privacy Policies over Time: Curation and Analysis of a Million-Dataset. In *Proceedings of the Web Conference 2021*, WWW '21, pages 2165–2176, New York, NY, USA, April 2021. ACM.
- [14] Monica Anderson, Andrew Perrin, Jingjing Jiang, and Madhumitha Kumar. 10% of Americans don’t use the internet. Who are they. *Pew Research Center*, 2019.
- [15] Tara Anthony, Carolyn Copper, and Brian Mullen. Cross-racial facial identification:

- A social cognitive integration. *Personality and Social Psychology Bulletin*, 18(3):296–301, 1992.
- [16] Hippocrates G Apostle. Aristotle’s categories and propositions. 1980.
 - [17] Susan Frelich Appleton. The Forgotten Family Law of Eisenstadt v. Baird. *Yale JL & Feminism*, 28:1, 2016.
 - [18] Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie S. Fichet, Paul Townsend, et al. A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Cscw ’17, page 155–168, New York, NY, USA, 2017. Association for Computing Machinery.
 - [19] Kenneth Artz. Texas Outlaws ‘Deepfakes’—but the Legal System May Not Be Able to Stop Them.
 - [20] Gregory Asmolov. The disconnective power of disinformation campaigns. *Journal of International Affairs*, 71(1.5):69–76, 2018.
 - [21] UN General Assembly et al. Universal declaration of human rights. *UN General Assembly*, 302(2):14–25, 1948.
 - [22] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*, 20(1):40–49, 2011.
 - [23] James P Bagrow, Xipei Liu, and Lewis Mitchell. Information flow reveals prediction limits in online social activity. *Nat. Hum. Behav.*, 3(2):122–128, January 2019.
 - [24] David A Baldwin. The concept of security. In *National and International Security*, pages 41–62. Routledge, 2018.
 - [25] Derek E Bambauer. Privacy versus security. *J. Crim. L. & Criminology*, 103:667, 2013.
 - [26] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020*, page 1943–1954, New York, NY, USA, 2020. Association for Computing Machinery.
 - [27] Susan B. Barnes. A privacy paradox: Social networking in the United States. *First Monday*, 11i9, 2006.
 - [28] Davide Barrera and Brent Simpson. Much ado about deception: Consequences of deceiving research participants in the social sciences. *Sociological Methods & Research*, 41(3):383–413, 2012.
 - [29] Lindsey Barrett. Ban facial recognition technologies for children-and for everyone else. *BUJ Sci. & Tech. L.*, 26:223, 2020.
 - [30] Christine Bell. Planned Parenthood of Southeastern Pennsylvania, et al. v. Robert P. Casey, et al. *Feminist L. Stud.*, 1:91, 1993.
 - [31] Colin J Bennett. The accountability approach to privacy and data protection: Assumptions and caveats. In *Managing privacy through accountability*, pages 33–48.

- Springer, 2012.
- [32] Dustin D Berger. Balancing consumer privacy with behavioral targeting. *Santa Clara Computer & High Tech. LJ*, 27:3, 2010.
 - [33] James Beverage. The Privacy Act of 1974: an overview. *Duke law journal*, 1976(2):301–329, 1976.
 - [34] Michael D Birnhack. The EU Data Protection Directive: An engine of a global regime. *Computer Law & Security Review*, 24(6):508–520, 2008.
 - [35] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.*, 2008(10):P10008, October 2008.
 - [36] Mark Blythe. Research through design fiction: narrative in real and imaginary abstracts. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 703–712, 2014.
 - [37] Taylor C Boas, Dino P Christenson, and David M Glick. Recruiting large online samples in the United States and India: Facebook, mechanical turk, and qualtrics. *Political Science Research and Methods*, 8(2):232–250, 2020.
 - [38] Leticia Bode and Emily K Vraga. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638, 2015.
 - [39] Sissela Bok. *Secrets: On the ethics of concealment and revelation*. Vintage, 2011.
 - [40] Jerenda Bond, Wrenetha A Julion, and Monique Reed. Racial Discrimination and Race-Based Biases on Orthopedic-Related Outcomes: An Integrative Review. *Orthopaedic Nursing*, 41(2):103–115, 2022.
 - [41] Frederik J Zuiderveen Borgesius. Singling out people without knowing their names—Behavioural targeting, pseudonymous data, and the new Data Protection Regulation. *Computer Law & Security Review*, 32(2):256–271, 2016.
 - [42] Frederik Zuiderveen Borgesius. Informed consent: We can do better to defend privacy. *IEEE Security & Privacy*, 13(2):103–107, 2015.
 - [43] Robert K Bothwell, John C Brigham, and Roy S Malpass. Cross-racial identification. *Personality and Social Psychology Bulletin*, 15(1):19–25, 1989.
 - [44] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS One*, 12(6):e0177678, 2017.
 - [45] John C Brigham, Anne Maass, Larry D Snyder, and Kenneth Spaulding. Accuracy of eyewitness identification in a field setting. *Journal of Personality and Social Psychology*, 42(4):673, 1982.
 - [46] Ryan Calo. Digital market manipulation. *Geo. Wash. L. Rev.*, 82:995, 2013.
 - [47] Ryan Calo. Against notice skepticism in privacy (and elsewhere). *Notre Dame L. Rev.*, 87:1027, 2011.
 - [48] Dustin P Calvillo, Ryan JB Garcia, Kiana Bertrand, and Tommi A Mayers. Personality factors and self-reported political news consumption predict susceptibility to political fake news. *Pers. Individ. Differ.*, 174:110666, 2021.

- [49] Kenneth A Carow and Randall A Heron. Capital market reactions to the passage of the Financial Services Modernization Act of 1999. *The Quarterly Review of Economics and Finance*, 42(3):465–485, 2002.
- [50] Fred H Cate. Principles of internet privacy. *Conn. L. Rev.*, 32:877, 1999.
- [51] Definitions Under CCPA. California Consumer Privacy Act (CCPA). *Policy*, 2020.
- [52] Ho-Chun Herbert Chang and Feng Fu. Co-diffusion of social contagions. *New Journal of Physics*, 20(9):095001, 2018.
- [53] Justin Cheng and Michael S. Bernstein. Flock: Hybrid Crowd-Machine Learning Classifiers. In *Association for Computing Machinery, CSCW '15*, page 600–611, New York, NY, USA, 2015. Association for Computing Machinery.
- [54] Bobby Chesney and Danielle Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *Calif. L. Rev.*, 107:1753, 2019.
- [55] Wen-Ying Sylvia Chou, April Oh, and William MP Klein. Addressing health-related misinformation on social media. *The Journal of the American Medical Association*, 320(23):2417–2418, 2018.
- [56] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- [57] Danielle Keats Citron. *The fight for privacy: protecting dignity, identity, and love in the digital age*. W.W. Norton & Company, first edition, 2022.
- [58] Andrew Clearwater and J Trevor Hughes. In the Beginning-An Early History of the Privacy Profession. *Ohio St. LJ*, 74:897, 2013.
- [59] Julie E. Cohen. Examined Lives: Informational Privacy and the Subject as Object. *Stan. L. Rev.*, 52(5):1373–1438, May 2000.
- [60] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the Design of a Personalized Privacy Assistant for the Internet of Things. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.
- [61] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583, 2022.
- [62] Matthew Crain. The limits of transparency: Data brokers and commodification. *New Media & Society*, 20(1):88–104, July 2016.
- [63] L.F. Cranor. P3P: Making privacy policies more useful. *IEEE Security & Privacy*, 1(6):50–55, November 2003.
- [64] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.
- [65] Ana-Maria Crețu, Federico Monti, Stefano Marrone, Xiaowen Dong, Michael Bernstein, and Yves-Alexandre de Montjoye. Interaction data are identifiable even across long periods of time. *Nature Communications*, 13(1):1–11, 2022.
- [66] Sergio Currarini and Friederike Mengel. Identity, homophily and in-group bias. *Eu-*

- ropean Economic Review*, 90:40–55, 2016.
- [67] Bart Custers, Simone van Der Hof, Bart Schermer, Sandra Appleby-Arnold, and Noellie Brockdorff. Informed consent in social media use — the gap between user expectations and EU personal data protection law. *SCRIPTed*, 10(3):435–457, December 2013.
 - [68] Catherine D’ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020.
 - [69] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proc. Natl. Acad. Sci. U.S.A.*, 110(15):5791–5796, 2013.
 - [70] G Cullom Davis. The Transformation of the Federal Trade Commission, 1914-1929. *The Mississippi Valley Historical Review*, 49(3):437–455, 1962.
 - [71] Adrienne de Ruiter. The Distinct Wrong of Deepfakes. *Philosophy & Technology*, 34(4):1311–1332, jun 2021.
 - [72] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872, 2021.
 - [73] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’06, page 581–590, New York, NY, USA, 2006. Association for Computing Machinery.
 - [74] Grace L Dillon. Indigenous futurisms, Bimaashi Biidaas Mose, flying and walking towards You. *Extrapolation.*, 57(1/2):1, 2016.
 - [75] B Dolhansky, J Bitton, B Pflaum, J Lu, R Howes, et al. The DeepFake Detection Challenge Dataset. *arXiv:2006.07397*, 2020.
 - [76] B Dolhansky, R Howes, B Pflaum, N Baram, and CC Ferrer. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv:1910.08854*, 2019.
 - [77] Tom Dougherty. Fickle consent. *Philosophical Studies*, 167(1):25–40, 2014.
 - [78] Marisa Elena Duarte. *Network sovereignty: Building the internet across Indian country*. University of Washington Press, 2017.
 - [79] Emile Durkheim. *The division of labor in society*. Simon and Schuster, 2014.
 - [80] Ritam Dutt, Ashok Deb, and Emilio Ferrara. “Senator, We Sell Ads”: Analysis of the 2016 Russian Facebook Ads Campaign. In *International conference on intelligent information technologies*, Advances in Data Science. ICIIT 2018, pages 151–168, New York, NY, USA, 2018. Springer.
 - [81] J.B. Earp, A.I. Anton, L. Aiman-Smith, and W.H. Stufflebeam. Examining Internet Privacy Policies Within the Context of User Privacy Values. *IEEE Trans. Eng. Manage.*, 52(2):227–237, May 2005.
 - [82] Natalie C Ebner, Donovan M Ellis, Tian Lin, Harold A Rocha, Huizi Yang, et al. Uncovering susceptibility risk to online deception in aging. *The Journals of Gerontology: Series B*, 75(3):522–533, 2020.
 - [83] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press, 2018.

- [84] Don Fallis. The Epistemic Threat of Deepfakes. *Philosophy & Technology*, pages 1–21, 2020.
- [85] J Doyne Farmer and John Geanakoplos. Hyperbolic discounting is rational: Valuing the far future with uncertain discount rates. *Cowles Foundation Discussion Paper*, No. 1719, 2009.
- [86] Federal Trade Commission. Consumer sentinel network data book 2018. 2018.
- [87] Ma Feicheng and Li Yating. Utilising social network analysis to study the characteristics and functions of the co-occurrence network of online tags. *Online Inform. Rev.*, 38(2):232–247, February 2014.
- [88] Scott L Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477, 1991.
- [89] Kim Fenrich. Securing your control system: the CIA triad is a widely used benchmark for evaluating information system security effectiveness. *Power Engineering*, 112(2):44–49, 2008.
- [90] Robert Fischer, Edward Halibozeck, Edward P Halibozeck, and David Walters. *Introduction to security*. Butterworth-Heinemann, 2012.
- [91] Luciano Floridi. Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions. *Philos. Trans. Royal Soc. A*, 374(2083):20160112, December 2016.
- [92] Nancy Fraser. Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Soc. Text*, 25/26:56–80, 1990.
- [93] Roland Friedrich. Complexity and Entropy in Legal Language. *Aip. Conf. Proc.*, 9:671882, June 2021.
- [94] Bobbye G Fry, M Therese, BL Weckmueller, et al. The family educational rights and privacy act of 1974. *Student records management: A handbook*, 43, 1997.
- [95] Feng Fu, Nicholas A Christakis, and James H Fowler. Dueling biological and social contagions. *Scientific Reports*, 7(1):1–9, 2017.
- [96] Mikaela Irene Fudolig, Thayer Alshaabi, Michael V. Arnold, Christopher M. Danforth, and Peter Sheridan Dodds. Sentiment and structure in word co-occurrence networks on Twitter. *Applied Network Science*, 7(1):1–27, February 2022.
- [97] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.
- [98] Tianna Gadbaw. Legislative update: Children’s Online Privacy Protection Act of 1998. *Child. Legal Rts. J.*, 36:228, 2016.
- [99] David Garcia. Leaking Privacy and Shadow Profiles in Online Social Networks. *Sci. Adv.*, 3(8):e1701172, 2017.
- [100] Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):3, 2022.
- [101] Tom Gerety. Redefining privacy. *Harv. CR-CLL Rev.*, 12:233, 1977.
- [102] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. A network approach to

- topic models. *Sci. Adv.*, 4(7):eaaq1360, July 2018.
- [103] Edward L Godkin. Libel and its legal remedy. *J. Soc. Sci.*, 12:69–80, 1880.
 - [104] Edward L Godkin. The Rights of the Citizen, IV—To His Own Reputation. *Scribner’s Magazine*, 8(1):58–67, 1890.
 - [105] Rebecca Darin Goldberg. You Can See My Face, Why Can’t I? Facial Recognition and Brady. *HRLR Online*, 5:261, 2020.
 - [106] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. FDVT: Data valuation tool for Facebook users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3799–3809, Denver, CO, USA, May 2017. ACM.
 - [107] Gretta L Goodwin. Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains, Statement of Gretta L. Goodwin, Director, Homeland Security and Justice, Testimony Before the Committee on Oversight and Reform, House of Representatives. In *United States. Government Accountability Office*, number GAO-19-579T. United States. Government Accountability Office, 2019.
 - [108] Elias Grünewald and Frank Pallas. TILT: A GDPR-Aligned Transparency Information Language and Toolkit for Practical Privacy Engineering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 636–646, New York, NY, USA, 2021. Association for Computing Machinery.
 - [109] Mark A Graber, Donna M D Alessandro, and Jill Johnson-West. Reading level of privacy policies on internet health web sites. *J. Fam. Practice*, 51(7):642–642, 2002.
 - [110] Samuel Greengard. Will deepfakes do deep damage? *Communications of the ACM*, 63(1):17–19, 2019.
 - [111] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci. U.S.A.*, 119(1), 2022.
 - [112] Alfonso Guarino, Nicola Lettieri, Delfina Malandrino, and Rocco Zaccagnino. A Machine Learning-Based Approach to Identify Unlawful Practices in Online Terms of Service: Analysis, Implementation and Evaluation. *Neural Comput. Appl.*, 33(24):17569–17587, dec 2021.
 - [113] Laurent Hébert-Dufresne and Benjamin M Althouse. Complex dynamics of synergistic coinfections on realistically clustered networks. *Proc. Natl. Acad. Sci. U.S.A.*, 112(33):10551–10556, 2015.
 - [114] Laurent Hébert-Dufresne, Dina Mistry, and Benjamin M Althouse. Spread of infectious disease and social awareness as parasitic contagions on clustered networks. *Physical Review Research*, 2(3):033306, 2020.
 - [115] Laurent Hébert-Dufresne, Jean-Gabriel Young, Alexander Daniels, and Antoine Al-lard. Network Onion Divergence: Network representation and comparison using nested configuration models with fixed connectivity, correlation and centrality patterns. *arXiv preprint arXiv:2204.08444*, pages 1–15, 2022.
 - [116] Jurgen Habermas and Jürgen Habermas. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT press, 1991.

- [117] William G Halfond, Jeremy Viegas, Alessandro Orso, et al. A classification of SQL-injection attacks and countermeasures. In *Proceedings of the IEEE international symposium on secure software engineering*, volume 1, pages 13–15, 2006.
- [118] Douglas Harris. Deepfakes: False pornography is here and the law cannot protect you. *Duke Law & Technology Review*, 17:99, 2018.
- [119] Kurtis Haut, Caleb Wohn, Victor Antony, Aidan Goldfarb, Melissa Welsh, et al. Could you become more credible by being White? Assessing Impact of Race on Credibility with Deepfakes. *arXiv:2102.08054*, 2021.
- [120] Heidi M. Hurd. The Moral Magic of Consent. *Legal Theory*, 2(2):121–146, 1996.
- [121] Jan Holvast. History of privacy. In *The history of information security*, pages 737–769. Elsevier, 2007.
- [122] Jim Isaak and Mina J Hanna. User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.
- [123] Sarah J Jackson and Brooke Foucault Welles. Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics. *J. Commun.*, 65(6):932–952, November 2015.
- [124] Sarah J. Jackson and Sonia Banaszczyk. Digital Standpoints: Debating Gendered Violence and Racial Exclusions in the Feminist Counterpublic. *J. Commun. Inq.*, 40(4):391–407, September 2016.
- [125] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385, 2021.
- [126] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 21, 1984.
- [127] Arshad Jamal, Jane Coughlan, and Muhammad Kamal. Mining social network data for personalisation and privacy concerns: a case study of Facebook’s Beacon. *International Journal of Business Information Systems*, 13(2):173–198, 2013.
- [128] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.
- [129] William James. *Pragmatism*, volume 1. Harvard University Press, 1975.
- [130] Carlos Jensen and Colin Potts. Privacy policies as decision-making tools. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 471–478, Vienna, Austria, April 2004. ACM.
- [131] T. Jung, S. Kim, and K. Kim. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access*, 8:83144–83154, 2020.
- [132] Kevin Ponniah. How a Chinese agent used LinkedIn to hunt for targets, August 2020.
- [133] Allan J Kimmel. Rumors and the financial marketplace. *J Behav Financ*, 5(3):134–141, 2004.
- [134] Keith Kirkpatrick. Monetizing your personal data. *Commun. ACM*, 65(1):17–19, December 2021.
- [135] Paul A Klaczynski, Wejdan S Felmban, and James Kole. Gender intensification and gender generalization biases in pre-adolescents, adolescents, and emerging adults.

- British Journal of Developmental Psychology*, 38(3):415–433, 2020.
- [136] Brennan Klein, C Brandon Ogbunugafor, Benjamin J Schafer, Zarana Bhadracha, Preeti Kori, Jim Sheldon, Nitish Kaza, Arush Sharma, Emily A Wang, Tina Eliassi-Rad, et al. COVID-19 amplified racial disparities in the US criminal legal system. *Nature*, pages 1–7, 2023.
 - [137] John Kleinig, Peter Mameli, Seumas Miller, Douglas Salane, and Adina Schwartz. *Security and privacy: global standards for ethical identity management in contemporary liberal democratic states*. ANU Press, 2011.
 - [138] John Kleinig. The Ethics of Consent. *Can. J. Philos.*, 12(sup1):91–118, January 1982.
 - [139] Gueorgi Kossinets and Duncan J Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009.
 - [140] Tahu Kukutai and John Taylor. *Indigenous data sovereignty: Toward an agenda*. ANU press, 2016.
 - [141] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.
 - [142] Jennifer Lackey. *The epistemology of groups*. Oxford University Press, USA, 2021.
 - [143] Jack Langa. Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, 101:761, 2021.
 - [144] Issie Lapowsky. One Man’s Obsessive Fight to Reclaim His Cambridge Analytica Data, January 2019.
 - [145] Larry Alexander. The Moral Magic of Consent (II). *Legal Theory*, 2(3):165–174, September 1996.
 - [146] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theor. Comput. Sci.*, 407(1-3):458–473, November 2008.
 - [147] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.
 - [148] Peter G. Leonard. Emerging Concerns for Responsible Data Analytics: Trust, Fairness, Transparency and Discrimination. *SSRN Electronic Journal*, 941:151–168, April 2018.
 - [149] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.
 - [150] Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, et al. Indigenous protocol and artificial intelligence position paper. 2020.
 - [151] Jason Edward Lewis. The future imaginary. *Presentation at TEDxMontreal, Montreal, Que., September, 14, 2013*.
 - [152] Jason Edward Lewis. A brief (media) history of the indigenous future. *Public*, 27(54):36–50, 2016.
 - [153] Kevin Lewis, Jason Kaufman, and Nicholas Christakis. The Taste for Privacy: An Analysis of College Student Privacy Settings in an Online Social Network. *J. Comput.-*

- Mediat. Commun.*, 14(1):79–100, October 2008.
- [154] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47–64, January 2020.
 - [155] E Paige Lloyd, Kurt Hugenberg, Allen R McConnell, Jonathan W Kunstman, and Jason C Deska. Black and White lies: Race-based biases in deception judgments. *Psychological Science*, 28(8):1125–1136, 2017.
 - [156] John Locke. *The second treatise of civil government*. Broadview Press, 2015.
 - [157] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, et al. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2691–2708, Boston, MA, August 2022.
 - [158] Marco Loos and Joasia Luzak. Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers. *Journal of consumer policy*, 39(1):63–90, 2016.
 - [159] Juniper Lovato, Laurent Hébert-Dufresne, Jonathan St-Onge, Randall Harp, Gabriela Salazar Lopez, et al. Supplementary materials for Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks. *Available upon request*, 2022.
 - [160] Juniper L. Lovato, Antoine Allard, Randall Harp, Jeremiah Onaolapo, and Laurent Hébert-Dufresne. Limits of Individual Consent and Models of Distributed Consent in Online Social Networks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2251–2262, Seoul, South Korea, June 2022. ACM.
 - [161] Viola Macchi Cassia. Age biases in face processing: The effects of experience across development. *British Journal of Psychology*, 102(4):816–829, 2011.
 - [162] Alice E Marwick and Danah Boyd. Networked privacy: How teenagers negotiate context in social media. *New media & society*, 16(7):1051–1067, 2014.
 - [163] Brian W Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.
 - [164] Gibson Mba, Jeremiah Onaolapo, Gianluca Stringhini, and Lorenzo Cavallaro. Flipping 419 Cybercrime Scams: Targeting the Weak and the Vulnerable. In *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion*, page 1301–1310, Perth, Australia, 2017. International World Wide Web Conferences Steering Committee.
 - [165] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.
 - [166] Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3, 2001.
 - [167] Jerry Menikoff, Julie Kaneshiro, and Ivor Pritchard. The common rule, updated. *N Engl J Med*, 376(7):613–615, 2017.
 - [168] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. The role of the crowd in countering misinformation: A case study of the COVID-19

- infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 748–757. Ieee, 2020.
- [169] Hans-W Micklitz, Przemysław Pałka, and Yannis Panagis. The empire strikes back: digital control of unfair terms of online services. *Journal of consumer policy*, 40(3):367–388, 2017.
 - [170] Stephen Mihm. Dumpster-Diving for Your Identity. *The New York Times Magazine*, pages 42–42, 2003.
 - [171] John H Miller and Scott Page. Complex adaptive systems. In *Complex Adaptive Systems*. Princeton university press, 2009.
 - [172] Melanie Mitchell. *Complexity: A guided tour*. Oxford university press, 2009.
 - [173] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679, 2016.
 - [174] Barrington Moore Jr. *Privacy: Studies in social and cultural history*. Taylor & Francis, 2023.
 - [175] R Moosavian. Pavesich v New England Life Insurance Co (1905). 2021.
 - [176] Masahiro Mori. The uncanny valley: the original essay by Masahiro Mori. *IEEE Spectrum*, 1970.
 - [177] Edgar Morin. From the concept of system to the paradigm of complexity. *Journal of social and evolutionary systems*, 15(4):371–385, 1992.
 - [178] Mark Newman. *Networks*. Oxford University Press, Oxford, UK, October 2018.
 - [179] Mark Newman. *Networks*. Oxford university press, 2018.
 - [180] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, Stanford, CA, USA, 2009.
 - [181] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.
 - [182] Helen Nissenbaum. A contextual approach to privacy online. *Daedalus*, 140(4):32–48, 2011.
 - [183] David M O’Brien. Privacy, law, and public policy. 1979.
 - [184] Cathy O’neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.
 - [185] Jonathan A. Obar and Anne Oeldorf-Hirsch. Clickwrap Impact: Quick-Join Options and Ignoring Privacy and Terms of Service Policies of Social Networking Services. In *Proceedings of the 8th International Conference on Social Media & Society, SMSociety17*, New York, NY, USA, 2017. Association for Computing Machinery.
 - [186] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
 - [187] State of California. California Consumer Privacy Act 2018, 2018.
 - [188] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. PrivOnto: A semantic framework for the analysis of privacy policies. *Semantic Web*, 9(2):185–203, 2018.

- [189] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press, 1990.
- [190] Paul N. Otto, Annie I. Anton, and David L. Baumer. The ChoicePoint Dilemma: How Data Brokers Should Handle the Privacy of Personal Information. *IEEE Security & Privacy Magazine*, 5(5):15–23, September 2007.
- [191] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87(3):925–979, Aug 2015.
- [192] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200, 2001.
- [193] Siani Pearson and Prodromos Tsiavos. Taking the Creative Commons beyond copyright: developing Smart Notices as user centric consent management systems for the cloud. *International Journal of Cloud Computing* 2, 3(1):94–124, 2014.
- [194] Kelly Pedersen. Manipulating the New Hampshire Mail: Political Power and the American Postal Service, 1792-1829. 2019.
- [195] Charles S Peirce. What pragmatism is. *The monist*, 15(2):161–181, 1905.
- [196] Andrew Perrin and Madhu Kumar. About three-in-ten US adults say they are ‘almost constantly’online. *Pew Research Center*, 2019.
- [197] Sandra Petronio. Communication boundary management: A theoretical model of managing disclosure of private information between marital couples. *Communication theory*, 1(4):311–335, 1991.
- [198] Irene Pollach. What’s wrong with online privacy policies? *Commun. ACM*, 50(9):103–108, September 2007.
- [199] Matthew Price, Alison C. Legrand, Zoe M.F. Brier, and Laurent Hébert-Dufresne. The symptoms at the center: Examining the comorbidity of posttraumatic stress disorder, generalized anxiety disorder, and depression with network analysis. *J. Psychiatr. Res.*, 109:52–58, February 2019.
- [200] Evani Radiya-Dixit and Gina Neff. A Sociotechnical Audit: Assessing Police Use of Facial Recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1334–1346, 2023.
- [201] Veronica Red, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011.
- [202] Regina Rini. Deepfakes and the Epistemic Backstop. *Philosophers’ Imprint*, 20(24):1–16, 2020.
- [203] Protection Regulation. General data protection regulation. *Intouch*, 25:1–5, 2018.
- [204] Horst Rittel. Wicked problems. *Management Science*, (December 1967), 4(14), 1967.
- [205] Leanne Roderick. Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry. *Critical Sociology*, 40(5):729–746, January 2014.
- [206] Kevin Roose. Here come the fake videos, too. *The New York Times*, 4, 2018.
- [207] Catherine G Roraback. Griswold v. Connecticut: A Brief Case History. *Ohio NUL Rev.*, 16:395, 1989.

- [208] Jeffrey Rosen. *The unwanted gaze: The destruction of privacy in America*. Vintage, 2011.
- [209] Jeffrey Rosen. The right to be forgotten. *Stan. L. Rev. Online*, 64:88, 2011.
- [210] Camille Roth, Jonathan St-Onge, and Katrin Herms. Quoting is not Citing: Disentangling Affiliation and Interaction on Twitter. In Rosa Maria Benito, Chantal Cherifi, Hocine Cherifi, Esteban Moro, Luis M. Rocha, and Marta Sales-Pardo, editors, *Complex Networks & Their Applications X*, Studies in Computational Intelligence, pages 705–717. Springer International Publishing, 2022.
- [211] Antoinette Rouvroy and Yves Poullet. The Right to Informational Self-Determination and the Value of Self-Development: Reassessing the Importance of Privacy for Democracy. In Serge Gutwirth, Yves Poullet, Paul De Hert, Cécile de Terwangne, and Sjaak Nouwt, editors, *Reinventing Data Protection?*, pages 45–76. Springer Netherlands, Dordrecht, 2009.
- [212] S.3805 – 115th Congress (2017–2018). Malicious Deep Fake Prohibition Act of 2018, December 21 2018.
- [213] Cristiana Santos, Midas Nouwens, Michael Toth, Nataliia Bielova, and Vincent Roca. Consent Management Platforms Under the GDPR: Processors and/or Controllers? In Nils Gruschka, Luís Filipe Coelho Antunes, Kai Rannenberg, and Prokopios Drogkaris, editors, *Privacy Technologies and Policy*, pages 47–69, Cham, 2021. Springer International Publishing.
- [214] Emre Sarigol, David Garcia, and Frank Schweitzer. Online Privacy as a Collective Phenomenon. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, page 95–106, New York, NY, USA, 2014. Association for Computing Machinery.
- [215] Hiroki Sayama. *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks, 2015.
- [216] Mara Schein, Rosemary J Avery, and Matthew D Eisenberg. Missing the mark: The long-term impacts of the Federal Trade Commission’s Red Flag Initiative to reduce deceptive weight loss product advertising. *Journal of Public Policy & Marketing*, 41(1):89–105, 2022.
- [217] Bart W. Schermer, Bart Custers, and Simone van der Hof. The crisis of consent: How stronger legal protection may lead to weaker consent in data protection. *Ethics Inf. Technol.*, 16(2):171–182, March 2014.
- [218] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z Tan, and Amy X Zhang. Modular politics: Toward a governance layer for online communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26, 2021.
- [219] Ronald Schouten and Kimberly D. Kumer. *Informed Consent*. Oxford University Press, Oxford, United Kingdom, May 2017.
- [220] Gary T Schwartz. Explaining and Justifying a Limited Tort of False Light Invasion of Privacy. *Case W. Res. L. Rev.*, 41:885, 1990.
- [221] Paul M. Schwartz. Internet Privacy and the State. *Conn. L. Rev.*, 32(3):815–859, May 2000.

- [222] Surendra Sedhai and Aixin Sun. HSpam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 223–232, 2015.
- [223] Jeffrey W Seifert and Harold C Relyea. E-government act of 2002 in the United States. In *Electronic Government: Concepts, Methodologies, Tools, and Applications*, pages 154–161. IGI Global, 2008.
- [224] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proc. Natl. Acad. Sci.*, 106(16):6483–6488, April 2009.
- [225] Ruth Shillair and William H Dutton. Supporting a cybersecurity mindset: getting internet users into the cat and mouse game. *Social Science Research Network*, 2016.
- [226] Henry Shue. *Basic rights: Subsistence, affluence, and US foreign policy*. princeton University press, 2020.
- [227] Royce Singleton Jr, Bruce C Straits, Margaret M Straits, and Ronald J McAllister. *Approaches to social research*. Oxford University Press, 1988.
- [228] Timothy H Skinner. California’s Database Breach Notification Security Act: The First State Breach Notification Law Is Not Yet A Suitable Template For National Identity Theft Legislation. *Rich. JL & Tech.*, 10:1, 2003.
- [229] Michael Warren Skirpan, Tom Yeh, and Casey Fiesler. What’s at Stake: Characterizing Risk Perceptions of Emerging Technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.
- [230] Robert H Sloan and Richard Warner. Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L.*, 14:370, 2014.
- [231] Daniel J Solove. *The Digital Person*, volume 1. New York University Press, New York, NY, October 2004.
- [232] Daniel J Solove. Conceptualizing privacy. *California Law Review*, 90:1087, 2002.
- [233] Daniel J Solove. Conceptualizing privacy. *California law review*, 90:1087–1155, 2002.
- [234] Daniel J Solove. A taxonomy of privacy. *University of Pennsylvania law review*, pages 477–564, 2006.
- [235] Sophie J. Nightingale, Kimberley A. Wade, and Derrick G. Watson. Investigating age-related differences in ability to distinguish between original and manipulated images. *Psychology and Aging*, 37(3):326–337, May 2022.
- [236] Mukund Srinath, Shomir Wilson, and C Lee Giles. Privacy at scale: Introducing the privaseer corpus of web privacy policies. *arXiv preprint arXiv:2004.11131*, pages 1–10, 2020.
- [237] William Stallings. Handling of Personal Information and Deidentified, Aggregated, and Pseudonymized Information Under the California Consumer Privacy Act. *IEEE Security & Privacy*, 18(1):61–64, January 2020.
- [238] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *IConference 2014 proceedings*, 2014.

- [239] Dietrich Stauffer and Ammon Aharony. *Introduction to percolation theory*. Taylor & Francis, London, 2018.
- [240] Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. Your Botnet is My Botnet: Analysis of a Botnet Takeover. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09, page 635–647, New York, NY, USA, 2009. Association for Computing Machinery.
- [241] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, 4:1, 2019.
- [242] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS One*, 13(9):e0203958, 2018.
- [243] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th international conference on World Wide Web*, pages 977–982, 2015.
- [244] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. Impact of rumors and misinformation on COVID-19 in social media. *J Prev Med Public Health*, 53(3):171–174, 2020.
- [245] Richard Thaler. *Advances in behavioral economics*. Russel Sage Foundation, 2005.
- [246] Sam Thielman. Surveillance reform explainer: can the FBI still listen to my phone calls, 2015.
- [247] Stefan Thurner, Rudolf Hanel, and Peter Klimek. *Introduction to the theory of complex systems*. Oxford University Press, 2018.
- [248] Tobias Matzner. Why Privacy is Not Enough Privacy in the Context of “Ubiquitous Computing” and “Big Data”. *Journal of Information, Communication and Ethics in Society*, 12(2):93–106, 2014.
- [249] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.
- [250] Cecilie Steenbuch Traberg and Sander van der Linden. Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Pers. Individ. Differ.*, 185:111269, 2022.
- [251] Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social Structure of Facebook Networks. *Physica A*, 391(16):4165–4180, August 2012.
- [252] Zeynep Tufekci. Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1):20–36, 2008.
- [253] Richard H Ullman. Redefining security. *International security*, 8(1):129–153, 1983.
- [254] United States. Department of Health and Human Services. *Secretary’s Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens: Report*. MIT Press, 1973.
- [255] United States. Privacy Protection Study Commission. *Personal Privacy in an Information Society: The Report of the Privacy Protection Study Commission*, volume 2.

- The Commission, 1977.
- [256] Antal Van den Bosch, Alain Content, Walter Daelemans, and Beatrice De Gelder. Measuring the complexity of writing systems*. *J. Quant. Linguist.*, 1(3):178–188, January 1994.
 - [257] Sander van der Linden. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3):460–467, 2022.
 - [258] Anne Veling and Peter Van Der Weerd. Conceptual grouping in word co-occurrence networks. In *IJCAI*, volume 99, pages 694–701, Dorpsweg 78, 1697 KD Schellinkhout, The Netherlands, 1999.
 - [259] L. Verdoliva. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.
 - [260] Emily K Vraga and Leticia Bode. Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5):621–645, 2017.
 - [261] Loïc JD Wacquant and Pierre Bourdieu. *An invitation to reflexive sociology*. Polity Cambridge, 1992.
 - [262] Isabel Wagner and Eerke Boiten. Privacy Risk Assessment: From Art to Science, by Metrics. In *Lecture Notes in Computer Science*, pages 225–241. Springer International Publishing, Cham, Switzerland, 2018.
 - [263] Isabel Wagner. Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021. *ACM Transactions on Privacy and Security*, pages 1–33, 2023.
 - [264] Spencer Weber Waller, Daniel B Heidtke, and Jessica Stewart. The Telephone Consumer Protection Act of 1991: Adapting Consumer Protection to Changing Technology. *Loy. Consumer L. Rev.*, 26:343, 2013.
 - [265] Nathan Walter, John J Brooks, Camille J Saucier, and Sapna Suresh. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication*, 36(13):1776–1784, 2021.
 - [266] Michael Walzer. *Spheres of justice: A defense of pluralism and equality*. Basic books, 2008.
 - [267] Na Wang, Heng Xu, and Jens Grossklags. Third-Party Apps on Facebook: Privacy and the Illusion of Control. In *Proceedings of the 5th ACM Symposium on Computer Human Interaction for Management of Information Technology*, CHIMIT ’11, New York, NY, USA, 2011. Association for Computing Machinery.
 - [268] Samuel Warren and Louis Brandeis. The right to privacy. In *Killing the Messenger: 100 Years of Media Criticism*, pages 1–21. Columbia University Press, 1989.
 - [269] Duncan J Watts, David M Rothschild, and Markus Mobius. Measuring the news and its impact on democracy. *Proc. Natl. Acad. Sci. U.S.A.*, 118(15):e1912443118, 2021.
 - [270] Peter Westen. *The logic of consent: The diversity and deceptiveness of consent as a defense to criminal conduct*. Routledge, Oxfordshire, England, UK, 2017.
 - [271] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.
 - [272] Ludwig Wittgenstein. *Philosophical investigations*. Macmillan, 1968.
 - [273] Arnold Wolfers. “National security” as an ambiguous symbol. *Political science quar-*

- terly, 67(4):481–502, 1952.
- [274] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explorations Newsletter*, 21(2):80–90, nov 2019.
 - [275] Christopher Wylie. *Mindf*ck: Cambridge Analytica and the Plot to Break America*. Random House, New York, NY, 2019.
 - [276] Wenjun Xiong and Robert Lagerström. Threat modeling—A systematic literature review. *Computers & security*, 84:53–69, 2019.
 - [277] Yang Yang, Jianhui Wang, and Adilson E Motter. Network observability transitions. *Physical Review Letters*, 109(25):258701, 2012.
 - [278] Razieh Nokhbeh Zaeem and K. Suzanne Barber. The Effect of the GDPR on Privacy Policies. *ACM Trans. Manage. Inf. Syst.*, 12(1):1–20, December 2020.
 - [279] Lizhi Zhang and Tiago P. Peixoto. Statistical inference of assortative community structures. *Phys. Rev. Research*, 2(4):043271, November 2020.
 - [280] Hannah Zimmerman. The data of you: Regulating private industry’s collection of biometric information. *U. Kan. L. Rev.*, 66:637, 2017.
 - [281] Sergey Zotov, Roman Dremluga, Alexei Borshevnikov, and Ksenia Krivosheeva. DeepFake Detection Algorithms: A Meta-Analysis. In *2020 2nd Symposium on Signal Processing Systems*, pages 43–48, 2020.
 - [282] Shoshana Zuboff. *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. Public Affairs, Boston, MA, USA, 2019.
 - [283] Shoshana Zuboff. Surveillance capitalism and the challenge of collective action. In *New labor forum*, volume 28, pages 10–29. SAGE Publications Sage CA: Los Angeles, CA, 2019.