Context, Causality, and Information Flow: Implications for Privacy Engineering, Security, and Data Economics

by

Sebastian P Benthall

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Management and Systems

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John Chuang, Co-chair
Professor Deirdre Mulligan, Co-chair
Professor Helen Nissenbaum
Professor David Wagner

Spring 2018

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Context, Causality, and Information Flow: Implications for Privacy Engineering, Security,
and Data Economics

Abstract

Context, Causality, and Information Flow: Implications for Privacy Engineering, Security, and Data Economics

by

Sebastian P Benthall
Doctor of Philosophy in Information Management and Systems

University of California, Berkeley

Professor John Chuang, Co-chair
Professor Deirdre Mulligan, Co-chair

The creators of technical infrastructure are under social and legal pressure to comply with expectations that can be difficult to translate into computational and business logics. This dissertation bridges this gap through three projects that focus on privacy engineering, information security, and data economics, respectively. These projects culminate in a new formal method for evaluating the strategic and tactical value of data: data games. This method relies on a core theoretical contribution building on the work of Shannon, Dretske, Pearl, Koller, and Nissenbaum: a definition of situated information flow as causal flow in the context of other causal relations and strategic choices.

The first project studies privacy engineering's use of Contextual Integrity theory (CI), which defines privacy as appropriate information flow according to norms specific to social contexts or spheres. Computer scientists using CI have innovated as they have implemented the theory and blended it with other traditions, such as context-aware computing. This survey examines computer science literature using Contextual Integrity and discovers, among other results, that technical and social platforms that span social contexts challenge CI's current commitment to normative social spheres. Sociotechnical situations can and do defy social expectations with cross-context clashes, and privacy engineering needs its normative theories to acknowledge and address this fact.

This concern inspires the second project, which addresses the problem of building computational systems that comply with data flow and security restrictions such as those required by law. Many privacy and data protection policies stipulate restrictions on the flow of information based on that information's original source. We formalize this concept of privacy as Origin Privacy. This formalization shows how information flow security can be represented using causal modeling. Causal modeling of information security leads to general theorems about the limits of privacy by design as well as a shared language for representing specific privacy concepts such as noninterference, differential privacy, and authorized disclosure.

The third project uses the causal modeling of information flow to address gaps in current theory of data economics. Like CI, privacy economics has focused on individual economic contexts and so has been unable to comprehend an information economy that relies on the flow of information across contexts. Data games, an adaptation of Multi-Agent Influence Diagrams for mechanism design, are used to model the well known economic contexts of principal-agent contracts and price differentiation as well as new contexts such as personalized expert services and data reuse. This work reveals that information flows are not goods but rather strategic resources, and that trade in information therefore involves market externalities.

For my grandmother, Theodora, who taught me to trust in God.
For my father, Timothy, who taught me that Logic is God.
For my mother, Susan, who taught me to see it through the bullshit.

Contents

List of Figures

List of Tables

Acknowledgments

The *sina non qua* figure in the production of this dissertation is the inimitable Michael Carl Tschantz. I will be grateful for life to him for getting me hired by Helen Nissenbaum as a researcher on DARPA agreement number FA8750-16-2-0287. This act snatched me from the jaws of failure and delivered me to New York City, where I could work towards completion with renewed vigor.

Michael is a man who has, to use the Platonic metaphor, left the cave and seen the world in sunlight. Through his uncompromising feedback, he has been a great teacher to me in mathematical discipline.

Helen is a firm taskmaster in addition to a visionary thinker. Her intellectual courage and ambition gave me the permission I needed to express my own. I thank her.

Chapter 2 of this dissertation began as the brainchild of Helen Nissenbaum and the brilliant Seda Gürses. It is with their permission that I include our collaborative work here. Seda has been another great teacher to me, and is proof that intellectually one *can* have it all: instrumental effectiveness, moral value, social relevance, and aesthetic beauty. Through this work I've learned that what matters in scholarship is not ones individual powers but rather ones complementarity with collaborators of high caliber.

Chapter 4 was originally written as a technical report in collaboration with Michael Tschantz with guidance from Anupam Datta. It took several years of graduate study to discover my fascination with the intersection of computer science research and the law. I thought it was a far-fetched dream. Then I met Anupam, who had been doing it successfully for over a decade. I aspire to one day have a perspective as correct as his.

Chapter 5 contains what I feel is this dissertation's most original contribution: a framework for modeling the consequences of information flow in strategic games. The seed of this work was nurtured by John Chuang in my first year at the I School, and my collaboration with him on "Strategic Bayesian Networks" is one of my happiest memories of Berkeley. Years later John became my advisor. I thank him for years of patience, honesty, sensitivity, kindness, and clarity of purpose.

I am also deeply grateful to David Wolpert, who discovered my work with John on "Strategic Bayesian Networks" and emailed me to inform me that the work had already been done by Daphne Koller ten years prior as "Multi-Agent Influence Diagrams" (MAIDs). MAIDs are now a cornerstone of the theoretical contribution of this dissertation. Besides this tip, Dr. Wolpert's wide-ranging and foundational work has been a source of inspiration for me.

What words can do justice to Deirdre Mulligan? None! I thank her.

I thank David Wagner for his timely approval.

So many colleagues at Berkeley helped me along my way. In particular, I thank Daniel Aranki, Nick Doty, and Jen King for teaching me about privacy through their research before I ever knew I would join them in their field of study.

Completing a dissertation at the remarkable institution of UC Berkeley is as much a bureaucratic challenge as it is an intellectual one. It has been all the more challenging for

me as I have conducted most of this business from New York. This would have been impossible without the vigor, availability, and thoroughness of Catherine Cronquist Browning.

Attendees at the 10th Annual Privacy Law Scholars Conference (2017) gave their helpful comments on Chapters 2 and 4. Special thanks goes to Blase Ur and Michael Hintze, our discussants. Relatedly, I am grateful to Chris Hoofnagle for his warm welcome into that community and early advice on how to contribute to privacy scholarship.

I thank Ignacio Cofone, Cathy Dwyer, Yafit Lev-Aretz, Amanda Levendowski, John Nay, Julia Powles, Ira Rubinstein, Madelyn Sanfilippo, Andrew Selbst, Yan Shvartzshnaider, Katherine Strandberg, Ari Waldman, Elana Zeide, Bendert Zevenbergen, and other members of the Privacy Research Group at the NYU School of Law for their helpful comments on Chapter 5.

I gratefully acknowledge funding support from the U.S. Defense Advanced Research Projects Agency (DARPA) under award FA8750-16-2-0287. The opinions in this paper are those of the author and do not necessarily reflect the opinions of any funding sponsor or the United States Government. I am also grateful to Tracy Figueroa and Finn Brunton for continuing to administer this grant from NYU after Helen transitioned to Cornell Tech.

I also gratefully acknowledge John Scott and JC Herz at Ion Channel for taking me into their company as a data scientist and not flinching when completing my doctorate took much, much longer than expected. Their support as employers have meant much more than the paycheck, which I depended on for much of my dissertation work. Their understanding of business, technology, and government rivals that of any professors who could advise me and contributed significantly to the framing and direction of this work.

There is an old Budokon saying, “How you do anything is how you do everything.” My academic life improved immeasurably when I began to train physically in movement arts. I thank Professor Felippe Moraes, Sensai Petra Lee Ghin, and Coach Mike Martin from Elements Athletics for their support of my well-being and their guidance in new modes of embodied thought.

I thank Dr. Yacine Kouyate for consolations regarding the mathematics of the Cosmos, and Floyd Toliver and Allyson Erick for all the great parties at Birdland Jazzista Social Club.

Lastly, I thank Daria, my beloved wife, who made completion both possible and worth it.

Chapter 1

Introduction

> For the rational study of the law the blackletter man may be the man of the present, but the man of the future is the man of statistics and the master of economics. It is revolting to have no better reason for a rule of law than that so it was laid down in the time of Henry IV. It is still more revolting if the grounds upon which it was laid down have vanished long since, and the rule simply persists from blind imitation of the past.
>
> - Oliver Wendell Holmes, Jr., “The Path of the Law”, 1897 [66])

This dissertation addresses one of the the great scientific, economic, and political challenge of our time: the design and regulation of the social and technical platforms such as the major commercially provided web services made available through the Internet. Its core contention is that no scholarly domain has yet comprehended the data economics driving the development and impact of this infrastructure well enough to predict and regulate its systemic impact. This scholarship labors to combine the necessary expertise in the correct mixture for progress.

1.1 The problem

Lessig [79] has argued that cyberspace is regulated in four modalities: technical infrastructure, social norms, the law, and the market. Each modality has its corresponding fields of academic inquiry. The construction of technical infrastructure is guided by principles from electrical engineering, computer science, and statistics. Social norms are studied in philosophy and sociology. The law is both a practice and the study of statutes and judgements. Market institutions are design according to principles of economics.

It would be convenient for scholars if these domains were as distinct from each other in practice as they are in theory. Of course, they are not, and so each branch of scholarship is partial and inadequate to the task of managing the Internet. Thoughtful leadership of technical infrastructure now comes primarily from the corporate leadership of private entities that are not tied to the academic traditions that tie down the institutions of public expertise. Since what is beyond public understanding is beyond democratic legal rule, the status quo is that these corporations are only weakly governed by law and social norms. This has created a public crisis and an opportunity for scientific advance [15].

1.2 The form of solution

The challenge is to construct a solution to this problem. What would such a scientific theory of information infrastructure need to be, to suffice? It must develop a new formal theory of strategic information flow and demonstrate its applicability across and at the intersection of all four regulatory modalities. This is what this dissertation does (See Figure 1.1).

1.2.1 Social norms and technical architecture

A priori, we know that at the very least a solution must be consistent with the mathematical foundations of electrical engineering, computer science, and statistics. These are robust and objective truths that are proven by scientific practice and everyday experience. Formal, mathematical specification is a prerequisite to technical design, and we won't shy away from this desideratum.

We also know that our theory must be sensitive to social norms. Herein lies the first major obstacle: the sociologists, anthropologists, and philosophers who best understand social norms are often alienated by mathematical formalism. Most (though not all) would reject the possibility that norms may be understood well enough by technologists that the latter could be adequately responsive to them.

But not all. Contextual integrity (CI) is a social theory of privacy norms that has been actively offered to computer scientists as a guide to technical design. This work at the intersection of social theory and privacy engineering is the subject of Chapter 2 of this dissertation, "Contextual Integrity through the Lens of Computer Science", co-authored by myself, Seda Gürses and Helen Nissenbaum [16]. It is a survey of computer science papers that have been inspired by CI. We compare the original social theory with its technical projection, and identify both opportunities for contextual integrity to be made more technically actionable and calls to action for computer scientists to better accomodate social processes and meaning.

Contextual integrity, which is described in detail in Section 2.2.1, defines privacy as contextually appropriate information flow. It envisions a social world differentiated into multiple social contexts or spheres, each with well-defined roles, purposes, and meanings of information. In Chapter 2, we discover that computer scientists see the world differently. They design infrastructure for a messy reality in which information crosses contextual boundaries.

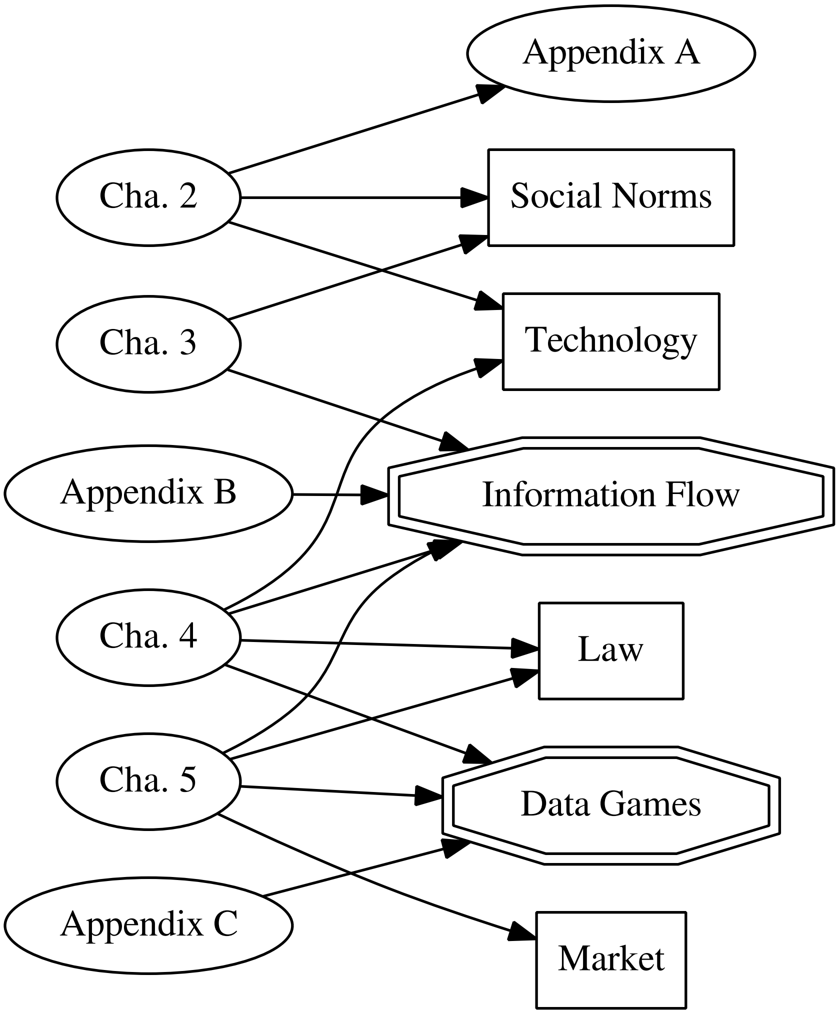


Figure 1.1: A graphical outline of the dissertation. Boxed components are the four modalities of the regulation of cyberspace. Round components refer to chapters and appendices of this document. Octagonal components refer to formal constructs. Taken as a whole, the dissertation argues that these formal constructs constitute scientific progress towards an understanding of strategic data flow that is applicable across all four modalities.

In short, technical and social platforms are not well regulated by social norms because they exist outside of our shared understanding of social contexts. Chapter 3 briefly assesses these conclusions and identifies the heart of the problem: the ambiguity inherent in our socially shared understanding of *information flow*. That understanding is based on a flawed ontology in which data's meaning is contained within it. This is not the case [68]; the meaning of information depends on the context, or contexts, in which it flows. Effective design and regulation of information infrastructure requires a scientific definition of information flow that addresses this fact precisely.

1.2.2 Law, technical architecture and situated information law

Such a definition of information flow is provided in Chapter 4, "Origin Privacy: Causality and Data Protection", originally written as a technical report with Michael Tschantz and Anupam Datta. It is targetted at another dyad in the regulation of information infrastructure: the interaction between technical architecture and information law.

This chapter builds on previous work in the automation of regulatory compliance. Businesses and other organizations are motivated to comply with privacy policies even as their information processing systems get more complex [11] [40] [123]. This business interest has driven scholarhsip in the mathematical formulation and implementation of law in computer science.

The concept of information implicit in privacy policies has ambiguities that are similar to those identified in Chapter 3: information is sometimes restricted based on its topic, and other times restricted based on its origin. What theory of information can capture how information flows from source to destination with meanings that depend on context?

Building on Dretske [43] and Pearl [108], we posit a definition of information flow as a causal flow within a greater context of causally structure probabilistic events. The context provides each event with *nomic associations*, or law-like, regular correspondences with other events in the causal system. The nomic associations are what make information useful for inference and thereby meaningful.

Chapter 4 develops this theoretical understanding of situated information flow as a tool for understanding the security of information processing systems. It develops the Embedded Causal System (ECS) model, a generic way of modeling an information processing system embedded in an environment. We use this model to define variations of well-known security properties such as noninterference and semantic security in terms of Pearlian causation, and prove sufficient conditions for systems to have these properties. We consider a new class of security properties based on Origin Privacy, the principle that a system designer must control information flow based on that information's origin. We further extend this model to the GDPR's regulation of biometric data and differential privacy. We also introduce a game theoretic variation on the model which relates the causal security properties to the tactical and strategic presence of an attacker. Developing this style of strategic modeling is the project of Chapter 5.

1.2.3 Information law and data economics

It is just a fact that most information infrastructure today is developed by industry, military, or both. No sufficient theory of infrastructure design can deny that there is a strategic aspect to its creation and use. Though there is a tradition of economics literature on the business of information goods [125] [141] [5], this scholarship has not captured the economics of data flow and reuse in a way that is commensurable with technical design and available to legal scholars for regulatory design. Chapter 5 fills this gap.

Chapter 5 works with the definition of situated information flow introduced in Chapter 4 and builds it into a framework for economic mechanism design (detailed in Appendix C) using the Multi-Agent Influence Diagrams of Koller and Milch [76]. This framework is the basis for a formal method for determining the value of information flow within a **data game**.

This framework can capture well-understood economic contexts such as principal-agent contracts and price differentiation. It can also support the modeling and analysis of newly introduced economic situations, such as the provision of personalized expert services and the value of the cross-context use of personal data. The model of cross-context use of personal data reveals that when firms trade in the personal data of their users, they can benefit at their user's expense. Because the user is not a party to this transaction, effects on their welfare may be considered a market externality, which invites the discussion of what form of regulation is appropriate to fix the market deficiency.

What this chapter demonstrates is that information is not, contrary to many traditional economic models, a kind of good that is consumed. Rather, information is a strategic resource, something that changes the very nature of the game played by economic actors. A thorough study of data economics through the modeling, simulation, and empirical analysis of data games may be the scholarly lens needed to understand how technical architecture, social norms, law, and the market interact and resolve in equilibrium.

Chaper 6 concludes this dissertation with a discussion of future work.

Chapter 2

Contextual Integrity through the Lens of Computer Science

Abstract: The theory of Privacy as Contextual Integrity (CI) defines privacy as appropriate information flow according to norms specific to social contexts or spheres. CI has had uptake in different subfields of computer science research. Computer scientists using CI have innovated as they have implemented the theory and blended it with other traditions, such as context-aware computing. This survey examines computer science literature using contextual integrity and discovers: (1) the way CI is used depends on the technical architecture of the system being designed, (2) 'context' is interpreted variously in this literature, only sometimes consistently with CI, (3) computer scientists do not engage in the normative aspects of CI, instead drawing from their own disciplines to motivate their work, and (4) this work reveals many areas where CI can sharpen or expand to be more actionable to computer scientists. We identify many theoretical gaps in CI exposed by this research and invite computer scientists to do more work exploring the horizons of CI.[1]

2.1 Introduction

Privacy is both an elusive moral concept and an essential requirement for the design of information systems. The theory of contextual integrity (CI) is a philosophical framework that unifies multiple concepts of privacy – as confidentiality, control, and social practice [57] – and has potential as a systematic approach to privacy by design [99]. Indeed, over the

[1] This chapter was originally published as:
Sebastian Benthall, Seda Gürses and Helen Nissenbaum (2017), "Contextual Integrity through the Lens of Computer Science", Foundations and Trends® in Privacy and Security: Vol. 2: No. 1, pp 1-69.
I am grateful to my collaborators, Dr. Seda Gürses and Dr. Helen Nissenbaum, for permission to include our joint work here.

last decade, computer scientists in a variety of subfields such as security, HCI, and artificial intelligence have approached the challenge of technical privacy design by applying CI.

This is a structured survey and review of this body of work. This survey has threefold aims: 1) to characterize the different ways various efforts have interpreted and applied CI; 2) to identify gaps in both contextual integrity and its technical projection that this body of work reveals; 3) perhaps most significant, it aims to distill insights from these applications in order to facilitate future applications of contextual integrity in privacy research and design. We call this, "making CI more actionable for computer science and computer scientists."

Over the last 20 years, privacy by design [24, 27, 111] and privacy engineering [56] have become research topics that span multiple sub-disciplines in computer science [115, 33]. Prior work has shown that computer scientists often stick to a single definition of privacy, for example, confidentiality, secrecy, or control over personal information. Although reducing privacy to a narrow definition has generated interesting work, it has limitations in addressing the complexities of privacy as an ethical value. In the wild narrow definitions offer analytic clarity, yet they may stray too far from a meaningful conception of privacy, that is, a conception that people actually care about.

The theory of contextual integrity (CI) was offered as a rigorous philosophical account of privacy that reflected its natural meaning while also explaining its moral force. Generally CI characterizes privacy as appropriate information flow, and appropriate flow characterized in terms of three parameters: actors (subject, sender, recipient), information type, and transmission principles. This definition immediately sets it apart from definitions in terms of subject control or stoppage of flow. Besides allowing for a more expressive framing of privacy threats and solutions than other approaches, the additional factors allow for greater specificity – hence less ambiguity – in prescribing and prohibiting certain flows. Because CI allows formal representation of flow constraints, it may serve to bridge privacy needs experienced by humans, *in situ*, with privacy mechanisms in digital systems. Although CI's account of privacy's ethical importance plays a lesser role in the work we have surveyed it remains important as a normative justification for Privacy by Design (PbD) initiatives grounded in privacy as contextual integrity.

With this survey we aim for more than a description of leading scientific applications of CI; in addition, we seek an exchange of ideas. In assessing how these applications have engaged with CI we ascertain, in one direction of exchange, how true to the letter they have been and how the framework might have been better or more fully reflected in the work. Equally, in the other direction, we assess how these frontrunners may materially inform future developments of CI itself. Such insights are crucial to enhancing the capacity of CI both to challenge and inspire scientific work and technical design, thus making CI more actionable for computer scientists. We conclude the survey by providing prescriptive guidance going forward.

On the one hand, our findings reveal that, for the most part, computer scientists engaged in technical design do not take up contextual integrity in its full theoretical scope. They often give specificity and depth to some of its concepts while bracketing others, rarely addressing normative dimensions of CI explicitly. Another common departure from CI is how researchers interpret context, which often maps onto their respective disciplinary assumptions, strategies, and literatures. In reviewing the literature, it was not our aim to declare any of these approaches "wrong" or "misguided". Instead, our aim was to record our findings, identify different perspectives, and note discrepancies as opportunities to learn from, revise, expand, and improve CI, to guide future research practice, and, in our terms, to make CI more "actionable."

On the other hand, the forays into implementing systems using contextual integrity led to significant innovations and improvisation that we believe can inform contextual integrity theory. For example, the papers we have surveyed have elaborated on different types of contexts, most prominently those that arise as a consequence of interactions (with people and machines) or those that come to bear as changes occur in the environmental conditions surrounding users. In doing so, computer scientists tease out different socio-technical situations that may impact how informational norms play out in a social sphere.

Relatedly, the desire of computer scientists to design systems that observe and adapt to changes throughout time is common. This is reflected in the use of technical mechanisms that capture changes to contexts, social norms and environments and that respond to the evolving conditions. Finally, papers we surveyed often position users as central actors, highlighting their role and agency in engaging and transforming informational norms in a context and throughout time. These concrete instances of socio-technical contexts, adaptivity and user agency shed light on issues that, with some elaboration, could enhance the analytical power of CI for privacy by design.

The authors have also taken up known challenges to CI, as in the case of papers that propose solutions to applying the framework when information flows from one social situation to the other or when multiple contexts are co-located. The question of multiple contexts is acute especially in technologies that act as infrastructures, e.g., platforms that host multiple applications or that host actors and actions from multiple social contexts. The solution space proposed by the authors include mechanisms to negotiate information flows across contexts or agents introduced to reason about multiple contexts in a single application.

Overall, when authors delegate responsibility of governing CI to technical elements that act semi-autonomously in an adaptive environment, they also raise novel questions for CI. It is not uncommon that researchers design agents that reason about contexts, auditing mechanisms that ensure informational norms are not violated, or apps that take active part in negotiating permissions. Can technical mechanisms be seen as actors in CI? If so, are they acting in the same context as they are "serving" or are they in a different context? These are some of the hard questions revealed by the works we studied, and that require input when considering future applications of CI in technical systems.

Just as important to our analysis is what the authors have not attended to in applying the CI framework. We reference here CI's account of privacy's ethical legitimacy, which identifies stakeholder interests, societal values, and contextual ends, purposes, and values as the basis for such legitimacy. Although omission of this aspect of CI might not be problematic for the technical accounts of privacy given by computer scientists, it nevertheless warrants explanation and justification; we discuss these issues at the end of the paper. Finally, we found no reference to information theoretic advances in privacy technologies, e.g., differential privacy or privacy preserving machine learning. Given the growing role that machine learning and artificial intelligence is playing in information systems, we believe there is a great potential in exploring how CI may be applied in systems that have the ability to infer and reason using data.

Finding that these papers both narrow and expand the CI framework, our review concludes with issues that will be important to address if CI is going to be useful to a wider spectrum of computer scientists. These include attending to questions such as: how to be more technically precise about the nature of contexts in order to relate the definition given in CI with more concrete notions used by computer scientists; how to advance normative concepts in CI – i.e. ends, purposes, and values – by taking advantage of well developed methods in the scientific study of privacy, including user studies, models, threat models and threat agents; how to use CI in systems where data not only flows, but also persists in a single place; and, how to apply contextual integrity to systems that function as infrastructure across multiple contexts.

2.2 Privacy and context in computing

In this section we provide theoretical background in privacy and context in computer science that situates our findings. Section 2.2.1 details contextual integrity as a philosophical theory that aspires to be robust to changes wrought by technology, rooted in legal, ethical, and social theory. Our study is primarily of how this framework has been used by computer scientists. As an inductive result we discovered that this work often drew from different conceptions of context as relevant to privacy that came from different subdisciplines in CS. In particular, we found computer scientists working at the intersection of CI, introduced in computer science literature with Barth et al. [9], and in the tradition of ubiquitous computing that has given a central place to context and its implications for privacy design [39, 42]. We detail the latter in Section 2.2.2. We note in Section 2.2.2.2 how the connection between context and privacy has become recognized by policy-makers [69, 52]. We speculate that this policy recognition was responsible for an uptick in the interest of computer scientists in context and privacy. The resulting creative synthesis of multiple traditions offers an opportunity for realizing new theoretical insights and opening new research problems.

2.2.1 Contextual integrity

The practice of privacy may be as old as social life itself but the contemporary need for a concept of privacy rich enough to drive policy and precise enough to shape architecture follows in the wake of advances in technologies that have disrupted how we create, collect, communicate, disseminate, interpret, process, and utilize data. (It is worth noting, however, that rarely, if ever, is it raw technology that stirs agitation; instead agitation is a response to technology embedded within particular social practices and particular political ecosystems.) In the US the contemporary need to sharpen the concept and strengthening protection is often dated back to 1895 with Warren and Brandeis's historic call to define a legal right to privacy in the wake of new photographic and printing technologies. In 1973, the landmark Report to the Secretary of Housing, Education, and Welfare issued Principles of Fair Information Practices (FIPs) following the rise of massive computerized databases.

In the 1990s and 2000s, such systems extended to video, audio, and online surveillance, RFID, and biometrics systems. Subsequently, public attention has turned to hyperbole over "big data" – database technologies, computational power, and scientific advances in information and data processing. Dramatically amplifying the privacy impacts of these technologies are transformations in the software engineering industry – with the shift from shrink-wrap software to services– spawning an agile and ever more powerful information industry. The resulting technologies like social media, user generated content sites, and the rise of data brokers who bridged this new-fangled sector with traditional industries, all contribute to a data landscape filled with privacy perils.

Approaches to privacy that depended on neat divisions of spaces or data into public and private have been severely challenged by these developments. Long entrenched definitions of privacy rights as rights to control information or rights to secrecy, that is, to block access, were overly simplistic, either too easily challenged by those with competing interests or over-claiming on the part of data subjects. An account that captured the complex contingencies of legitimate privacy claims was needed – one that benefitted from conceptual building blocks of existing theories but offered a greater expressive agility, to resist incursions while allowing the positive potential of novel socio-technical systems to be realized. Contextual integrity intends to provide such an account. For one, it addresses gaps in prior entrenched conceptions allowing it to identify privacy threats to which other accounts were blind (e.g. "privacy in public"). It also offers a view on the nature and sources of disruptive information flows in order to distinguish that that constitute threats from those that do not.

The theory of privacy as contextual integrity (CI) introduces three key concepts into the privacy vocabulary:

1. **Contexts.** These refer to social contexts, not formally constructed but discoverable as natural constituents of social life. As theorized in sociology, social theory, and and social philosophy, they have been assigned various labels, including, social domains, social spheres, fields, or institutions. (Throughout this survey, we will use the term *sphere* to denote this sense of context.) Societal recognition of distinct contexts, such as healthcare, family and home life, politics, religion, commercial marketplace, and education is clearly evidenced in distinctive structures of law and regulation.

   For the framework of contextual integrity, contexts are formally characterized in terms of key elements, which include, paradigmatic activities, roles (or capacities), practices, and norms. Distinguishing contexts from one another, are contextual goals, ends, purposes, and values, around which activities and norms are oriented, and to which respective contexts are committed.

2. **Contextual informational (privacy) norms.** Among contextual norms, these govern information flows and, according to contextual integrity, are likely to map onto people's privacy expectations. Informational norms are well-formed only if they refer to five parameters: sender, recipient, and information subject, information types (topics, attributes), and transmission principle. The parameters of actors and attributes range over contextual ontologies, distinctive to respective social contexts, if not unique. Thus, in healthcare context, senders, recipients, and subjects range over agents acting in the capacities, such as, doctor, nurse, patient, surgeon, psychotherapist, etc. and topics may range over symptoms, diagnoses, and drug prescriptions. Transmission principles condition the flow of information from party to party, including those commonly associated with privacy, such as, *with permission of data subject*, *with notice,* or *in confidence,* in addition to those less salient, such as, *required by law*, *with a warrant,* and *entitled by recipient*.

   Privacy as contextual integrity is respected when entrenched informational norms are followed. When these norms are violated (e.g. by disruptive information flows due to newly functioning technical systems) there is a prima facie case for asserting that privacy has been violated. The framework of contextual integrity allows, however, for the legitimacy of disruptive flows to be defended, as described below.

3. **Contextual ends, purposes, and values.** These may be considered the "essence" of a context, without which respective contexts would not be comprehensible. How would one properly describe a school, say, without indicating its purpose? These – let us call them – teleological factors are also important in defending the legitimacy of informational norms, particularly useful when comparing novel information flows against past expectations, or when no competing alternative is obvious, they are useful in evaluating the ethical legitimacy of given flows taken alone.

According to the CI framework, privacy norms can be assessed in terms of how they affect the interests of relevant parties ("stakeholders") and how they impinge on societal values, such as equality, justice, fairness and political liberties. In addition to these considerations the norms governing flow can be evaluated in terms of their impacts on the attainment of contextual ends, purposes, and values – either promoting or confounding them. For example, informational norms enabling (and enforcing) a secret ballot protects autonomous voting in elections and, as such, promotes ends and values of democracy.

This structure ensures that though CI is conservative, in the sense that it presumes in favor of entrenched norms, it nevertheless has built into it a set way for systematically evaluating and updating norms. This is done by examining balance of interests, general ethical and political values, and contextual values and purposes. It follows that norms adapt to their environment, crucial for an account of privacy to remain relevant in the face of advancing technologies of information and computational technologies. Societal and environmental shifts can destabilize entrenched privacy norms in many ways, either revealing that they are no longer optimal in achieving contextual ends and values or have nothing to say about disturbing information practices. Although such circumstances constitute challenges for ethicists and social policy makers, they do not necessarily constitute challenges to CI itself, which copes with novel or disruptive flows by presenting new norms for consideration.

Adapting norms to novel or disruptive flows may involve adjusting any of the parameters. For example, the increasing digital mediation of transactions, communications, and interactions (including social media) creates new data recipients, which forces the reconsideration of norms. The same goes for increasing specialization and fracture of skills and functions within traditional contexts, healthcare being a prime example. The one-on-one physician-patient relationship paradigmatic of the distant past has been replaced by an immensely complex care and treatment ecosystem, involving specialists, insurance companies, pathologists, public health officials, wireless pacemaker service providers, and, with that, the emergence of new informational norms – in the ideal, to serve contextual ends and values.

A note on terminology: We refer to aspects of CI that deal with evaluating the *legitimacy* of norms as its normative, prescriptive, or ethical aspects. These aspects are contrasted with what we might call its *descriptive* or conceptual aspects, referring to the the structure of informational norms.

2.2.2 Context in computing

CI bridges two worlds. In one, it is an account of privacy as a term that has accrued meaning to describe alarm over wide-ranging, technology induced practices of surveillance, data accrual, distribution, and analysis. The alarm is due to disruptive practices that violate privacy expectations and create or amplify imbalances in power. CI posits contextual informational norms to model privacy expectations and explains when such expectations are morally legitimate and warrant societal protection. In the other, CI offers a formal structure for expressing rules of information flow (informational norms) and for building computational models of people's privacy expectations in well-defined settings (contexts.) The first is a world inhabited by humanists, social scientists, lawyers, and regulators; the second is inhabited by mathematicians, computer scientists, and engineers. Perhaps because it seeks to map a meaningful conception of privacy onto a conception that strives for formal rigor contextual integrity has been taken up by computer scientists interested in privacy design and engineering.

Although philosophical versions of contextual integrity appeared in articles, dating back to 1998 [96] and, later, in the book, *Privacy in Context* [99], it was not represented in computer science literature until Barth et al. [9]. This paper, which we introduce as one of our survey exemplars in Section 2.3.3.1, formalized the fragment of CI known as context-specific (or, contextual) information norms. The authors, which include Nissenbaum, developed a logical framework for expressing and reasoning about norms and showed that this framework is adequate for expressing regulations drawn from U.S. sectoral privacy laws, such as, HIPAA, GLBA, and COPPA.

In fact, contextual integrity is but one source of influence that has drawn computer scientists to engage with the idea of context as it relates to privacy. Two others are worth discussing because they have roused the interest of computer scientists in context and shaped how they conceive of context with respect to privacy. We therefore find in computer science loose interpretations of contextual integrity that consider these other forms of context. They are: the field ubiquitous computing and the Obama White House Bill of Consumer Privacy Bill of Rights [69] (also the World Economic Forum and FTC Reports around a similar time [52, 27]. Grasping these influences has been important in advancing our own ability to analyze the articles we have chosen for this survey.

2.2.2.1 Context in ubiquitous computing

Contextual integrity is not the only research tradition linking context and privacy in computer science. Many contemporary issues in human-computer interaction around mobile devices and IoT were anticipated in earlier waves of research into "ubiquitous computing". This research program envisioned a world in which computation was not restricted to specialized workstations but, instead, was embedded in everyday objects and practices, enabling user interaction through sensors and actuators. Within ubiquitous computing research interest emerged in developing technologies that were responsive to social and environmental context, that is to say, 'context-aware' computing.

In their "anchor article" on context-aware computing, Dey et al. [39] extensively analyze definitions of 'context' in the literature of their field and settle on the following for their own work:

> **Context:** any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects. [39]

This definition of context, specifically referring to the concrete *situation* of persons and objects, starkly contrasts with the notion historically evolved, abstract, and normative social *spheres* of CI. In our survey of the computer science literature invoking contextual integrity, we found that several papers conceive of 'context' in ways that have more in common with context-aware computing than with context as defined in CI. This has led to interesting synthetic work, in addition to incipient disjunctures.

That computer scientists have taken up the tradition of context-aware computing in their work on privacy as contextual integrity is not surprising. Early in this field, contributors Ackerman et al. [3] anticipated that context-aware computing would lead to privacy by design, arguing that technical systems and legal frameworks would be co-designed. But hints of the connection between context-aware computing and contextual integrity, which was not formulated as a framework until later, were present at least as early as Dey et al.'s anchor article [39]:

> As computational capabilities seep into our environments, there is an ever-growing and legitimate concern that technologists are not spending enough of their intellectual cycles on the social implications of their work. There is hope, however. Context-aware computing holds a promise of providing mechanisms to support some social concerns. For example, using context to tag captured information may at first seem like an intrusion to individual privacy ("I do not want to meet in a room that records what I say."). However, that same context can be used to protect the concerns of individuals by providing computational support to establish default access control lists to any information that has been captured [78], limiting distribution to those who were in attendance.

Most relevant to this article are the implicit connections these authors draw between the design of technology responding to a particular situation (certain people meeting in an office room) and a general expectation of privacy. The norm that literal, unfiltered information about what happens in meetings is available only to those who attended could be attributed to the abstract social sphere of office meetings. This prefigures a result of our study, which finds computer scientists taking up 'context' in ways that reflect both senses of the word, and in so doing implicitly drawing connections between them.

This early work on context-aware computing reflected the state of the art in sensors and the kind of sensing they made possible: explicitly representing context as a kind of fixed container in which people could act. This was famously critiqued in a paper by Dourish [42], connecting how 'context' is approached in ubiquitous computing to broader questions in philosophy of science. Dourish argued that representing context (e.g. location or time in which an application is used), as a kind of container for activity whose boundaries are delineable draws from the *positivist* tradition in social science that sees context as stable and separable from the activity taking place within it. Dourish contrasts this understanding of context with a different one deriving from the *phenomenological* tradition of social science. According to it, context is *occasioned*, "relevant to particular settings, particular instances of action, and particular parties to that action", not an abstract category that generalizes over settings, actions, and parties. This kind of context arises from and is maintained by its activity, sometimes dynamically adjusting along with the activities themselves. For example, a private conversation between friendly colleagues at work can shift from a formal, professional discussion into an informal, personal discussion and back again. These shifts will occur as and through changes in the conversational activity, such as changes in tones of voice or comments such as, "Well, we should really get back to work; I have to go in twenty minutes."

Dourish investigates how this conception of context ties into the sociological mystery of how social order comes into being. There is a tension between explanations of social order that attribute it to rules, expectations, and conventions that have a broader reality beyond particular occasions of interaction (what we might call a 'top-down' ordering), and explanations that see all social order as arising from interaction itself as an achievement of the social actors ('bottom-up' ordering).

While it may be argued that top-down and bottom-up ordering are always co-occurring, often one or the other process is emphasized in scholarly work. Contextual integrity, in its original articulations [97, 99], tends to emphasize the top-down pressure of contextual ends, purposes, and values shaping norms that in turn guide information flows. In contrast, while Dourish acknowledges the role of top-down orderings, he highlights the bottom-up processes that make each context occasioned and dynamic, in the spirit of his interactional, phenomenological objection to static representations of context. We find that both ways of thinking about context are prominent in the literature that we review, even though we have limited this review only to computer science literature that refers to contextual integrity.

2.2.2.2 Context in privacy policy

While we were looking specifically for computer science papers that referenced contextual integrity, it was interesting to find many papers that took "privacy in context" as an idea (which also happens to be the title of Nissenbaum's book about contextual integrity [99]), but that do not draw from the framework of contextual integrity. If the direct origin of "privacy in context" was not contextual integrity, what was it?

Our contention is that interest was prompted by the general uptake of context and contextual integrity in the formulation of several policy documents from 2010 and later. For example, the White House Report, "Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy" [69], lists "Respect for Context" as one of its seven principles: "Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data." Around the same time, a report issued by the U.S. Federal Trade Commision also invoked context when it stipulated that data collection by companies should be restricted to what was appropriate for the "context of interaction" or else they should make "appropriate disclosures". In a similar vein the World Economic Forum's 2012 report, Rethinking Personal Data invokes the importance of context for policy governing data in numerous places (e.g. pages 5, 10, 15, 16, 17, 19, and more.) [52]

We have found significant variation in how computer scientists have interpreted the term "context", often reflecting their disciplinary background and research agendas. Some follow contextual integrity quite closely. Others cite 'contextual integrity' and 'privacy in context', possibly to situate the privacy-context connection within a scholarly lineage without following CI substantively. Most of these papers were written after the policy arena acknowledged CI theory, while in parallel ubicomp researchers had already established a concept of context. For some computer scientists, context as *situation*, reminiscent of ubiquitous computing research, informs their work, as it does some of the work on privacy regulation (albeit outside the focus of this paper.) Overall, though computer scientists have, characteristically, explored the relationship between various forms of context and privacy in a rigorous and pragmatic way, they have not made the definition of context the subject of explicit theoretical commitment.

Nevertheless this work in computer science at the boundaries of contextual integrity makes important contributions to the theory itself. Inspired broadly by contextual integrity, computer scientists have explored aspects of the relationship between privacy and context in detail. Our systematic study of these works has found in the variations and commonalities within this literature insights that can inform and inspire further developments in contextual integrity.

2.3 The study

The main objective of this study is to characterize the different ways CI has been interpreted and applied in computer science, reveal its technical projection, and thereby, capture gaps in CI itself. The long term objective of this study is to identify ways that CI can be made more actionable for computer scientists and systems developers. In order to do so we systematically reviewed literature coming out of different subfields of computer science explicitly stating the use of contextual integrity in their problem or solution definition. We made use of techniques proposed by Kitchenham and Charters [75] to make our study as comprehensive and transparent as possible.

In their projects invoking CI, computer scientists have taken on the hard task of translating an elaborate philosophical framework into computer science research practice – in which different theoretical and methodological traditions apply. This renders the translation of CI into technical contexts a non-trivial task. For these reasons alone an assessment of current uses of the theory in CS is valuable for understanding how well the theory translates, what new questions arise when applied in a technical context, and what obstacles become evident. Through this survey we evaluate its uptake in computer science and begin to sharpen the theory to make it more actionable for researchers who want to use it in the future.

2.3.1 Research Questions

Driven by the motivations listed above, we decided to focus on four research questions as we take stock of the use of CI in computer science research and assess it:

2.3.1.1 RQ1. For what kind of problems and solutions do computer scientists use CI?

As an initial question for our inquiry, we wanted to know if there were any particularly notable categories of problems being addressed by computer scientists using contextual integrity. Computer science is a broad field; researchers may have found contextual integrity useful for solving particular kinds of problems, focus on certain domains, or be more likely to invoke CI in certain subfields of computer science.

2.3.1.2 RQ2. How have the authors dealt with the conceptual aspects of CI?

Contextual integrity is partly a conceptual theory that is predictive of social concerns about privacy that originate and manifest themselves especially with technological change. The theory posits *social contexts* as evolved abstract spheres of activity characterized by *ends, purposes, and values*. Social contexts have *information norms*, parameterized by *actors (senders, recipients, and subjects), information types,* and *transmission principles*. Contextual integrity identifies privacy as appropriate information flow; such flow would be characterized by contextual informational norms.

We wanted to know to what extent the computer science researchers using contextual integrity used this conceptualization of privacy. Do the researchers define context in the way contextual integrity does, or in other ways? And do they define privacy in terms of appropriate information flow according to norms?

2.3.1.3 RQ3. How have the authors dealt with the normative aspects of CI?

Contextual integrity is a normative framework of privacy. It argues that privacy is an important value because appropriate information flow promotes the data subject's interests in balance with those of others as well as societal and ethical values, and maintains the ability of social contexts to fulfill their purposes.

We wanted to know if computer scientists using contextual integrity take up this normative aspect of the theory. If not, from where do they perceive the normative clout of privacy coming? How do they evaluate whether privacy is addressed effectively through their proposed mechanisms or solutions?

2.3.1.4 RQ4. Do the researchers expand on CI?

In developing technical systems computer scientists have to make a number of substantive and specific design decisions. This is also the point at which the rubber meets the road: how does a researcher translate a philosophical theory into a formulation useful for technical design? In executing this translation computer scientists are likely to attend to concrete questions that CI may not provide explicit guidance for. In these moments, researchers are likely to identify gaps in CI and propose techniques to make up for these gaps. What are the gaps that researchers identify, how do they expand on these, and how do they stretch the theory explicitly or implicitly?

2.3.2 Study Methodology

In compiling and revising the relevant papers, we followed empirical research methodologies recommended for use in software engineering studies [75]. In order to answer our research questions, we conducted the following four steps:

- Based on our research questions (Section 2.3.1), we iteratively developed an initial template of analytic questions using a selection of CI articles.

- We searched in online repositories for papers using CI as its reference theory. To ensure we have a reasonable collection, we searched digital libraries (Google Scholar, IEEE Xplore, ACM DL) for papers that appeared in CS venues that had CI in their title or main body. To cast a wider net, we included the key terms "contextual integrity" and "context AND privacy". For those papers that explicitly invoked CI, we combed through later publications that cited them to see whether the use of CI propagated. We carefully evaluated the inclusion of papers that only reference CI without making further use of the framework. In the process, we found a number of papers that refer to context and contextual norms that do not refer to Nissenbaum's work and removed these from the study. Evaluating whether and how CI may have proven useful in these papers is out of the scope of the current work. Some papers claimed they used CI and integrated other conceptions of "context" in CS, we kept these papers in our study. We initially categorized papers with respect to the subfields of computer science from which they originated. The represented fields of research included security engineering (including privacy engineering and access control); artificial intelligence (including papers on multi-agent systems, machine learning, semantic web, social network analysis and community detection); systems (distributed systems, pervasive and mobile computing); HCI (usable security and privacy, ubiquitous computing); and software engineering (requirements engineering and business process design).

- Once we had completed our search, we tested the completeness and consistency of the template based on close reading of additional articles. Once the template was stable (see Appendix A), we (the authors) independently read each paper and answered each question of the analytic template for it. We did a comparative analysis of the answers in order to distill those aspects of the papers that answered our research questions. At this stage, we also concluded a quality assessment of each paper with respect to its contributions to computer science and removed those that failed our assessment. We documented all of our analysis and discussions in an online repository.

- We used the output of the templates to complete a thematic analysis of each paper. We consolidated what we had discovered into major categories of themes, one for each research question. Our work indicated the most productive way to interpret these questions. For RQ1, we found the most significant way we could characterize the variety of problems addressed in the literature was by looking at the kind of technological architecture researchers were designing. For RQ2, we focused on how researchers characterized "context" in their work. We split this concept of 'context' down into many finer-grained variables in order to show the variability between papers. For RQ3, we looked specifically for sources of normativity used by each paper and coded them accordingly. For RQ4, we analyzed the ways in which the papers expanded on contextual integrity. Our analysis did not reveal that the initial categorization of papers according to subfields in CS revealed further insights for our study.

In the remainder of this Section, we provide detailed accounts of select papers as illustrations of how we thematically analyzed each paper in accordance with the steps described above.

2.3.3 Three exemplars of analysis

In order to provide the reader with a demonstration of how we got to the different themes in our results, we pulled out three of the papers to serve as exemplars. We selected these three papers as they deeply engage CI; they stem from different subfields in CS with varying methods and techniques; and, they allow us to demonstrate the rather different ways in which the authors have elaborated on CI. The curious reader is encouraged to read these full papers which are rich in ideas and thoughtful in their use of CI. All other papers are are analyzed according to respective categories and themes, extracted through the template that had guided our reading of them.

2.3.3.1 Privacy and Contextual Integrity: Framework and Applications (Barth, Datta, Mitchell, and Nissenbaum)

The first published computer science paper to reference contextual integrity was coauthored by Helen Nissenbaum and therefore can be said to be an authoritative expression of the theory. It is not, strictly speaking, a paper about the design of a technological artifact. Rather, it is an articulation of a subset of the principles and parameters of contextual integrity in a formal logic (something further discussed in Section 2.4.1 under RQ1). Formalization is a prerequisite to computational implementation, and so this paper demonstrated the potential of contextual integrity as a guide to the design of computational systems. For the purposes of our study it is just as notable what it did not formalize into logic, as this has left open many challenges to computer scientists seeking to use contextual integrity.

After grounding the work in an exposition on contextual integrity theory, the first major contribution of the paper is a careful translation of principles of contextual integrity into formal logic. The particular flavor is Linear Temporal Logic (LTL), a type of logic which is capable of expressing formulae of relationships of variables arranged in time. This translation refines the ontology of contextual integrity by making explicit that information flows have a temporal order. This allows the authors to define specific transmission principles that condition appropriate flow on previous flows (further discussion under RQ4 in Section 2.4.4). The logical specification allows a particular history or trace of information flows to be audited for appropriateness according to formal rules.

One of the benefits of having to make the logic of contextual integrity explicit is that it brings to light aspects of the theory that are easy to take for granted but which have far-reaching implications. The paper explicitly models both the knowledge available to each actor at different points in time as well as the ways that different attributes are related to each other via inference. This paper therefore provides an epistemic model that is only implicit in other accounts of CI. Having provided a formal language for expressing policies in the style of CI's context-specific information norms, the authors go on to prove a number of theorems about the computational complexity of auditing traces based on these policies, testing for the possibility of complying with the policy, and comparing policies.

The authors do not tie their formalization back to the origin of norms through the evolution of social sphere and its ends, purposes, and values. Rather, after formalizing the aspects of contextual integrity that they are able to, they validate their work by showing that it is expressive of United States sectoral privacy laws: HIPAA, GLBA, and COPPA (see Datta et al. [35] for further work along these lines). They also argue that the expressivity of their formalization compares favorably with other proposed access control policy languages such as XACML, ECAP, and P3P.

This paper is particularly notable as the first published computer science paper concerning contextual integrity. Explicitly only a formalization of *part* of CI, Barth et al. [9] provide a way of expressing norms as policies that can be used in computational tests for compliance. This sets a precedent for computer science papers using contextual integrity to consider 'context' in a structured, abstracted, and normative way (see RQ2 in Section 2.4.2). It sets aside parts of contextual integrity that account for how norms form through adaptive social processes. By focusing on regulatory compliance, it brackets the social source of privacy norms (RQ3 in Section 2.4.3). If there is something lost in this usage of contextual integrity in computer science, it may be recovered through other uses and understandings of social context that have influenced technical research.

2.3.3.2 Android Permissions Remystified: A field Study on Contextual Integrity (Wijesekera et al.)

The potential role that permissions in mobile platforms can play in providing users with control and transparency over how their information flows to apps and mobile platforms has recently attracted much research. For a long time, Android and iOS platforms asked users for permissions at install time. Recently they have extended the framework to also make it possible to prompt users for permissions during runtime. Prior research has shown that few people read the Android install-time permissions and even fewer comprehend the complexity that the permissions are loaded with – for example, the permission to access Internet may be bundled with the permission to load ads. Prompting users too frequently causes “habituation”. Limiting prompts, however, raises questions about which ones to select in order to effectively protect user privacy. Wijesekera et al. [143] leverage contextual integrity to approach the usable security and privacy problems that arise when interfacing a permission request model to the users (a user interface and technical platform problem as discussed in RQ1 in Section 2.4.1).

In their study, the authors examine how applications are currently accessing user data and assess whether or not it corresponds to users’ expectations. To do so, they instrumented Android phones to log whenever an application accesses a permission-protected resource and distribute these to 36 participants who use the phones for a week. For each permission request, they keep a log of the app name and permission, and further “contextual factors” which include whether the requesting application was visible to the user (running with or without user interaction, notifications, in the foreground or background); screen status (whether the screen was on or off); connectivity (the wifi connection state); location (the user’s last known coordinates); the UI elements that were exposed to the user during the request; history of interactions with applications; and, the path to the specific content that was requested[2]. After the week, study subjects participated in an exit survey where they were shown a sample of screenshots to inquire about their expectations relating to requested permissions. The authors use the outcomes of the study to start specifying a classifier to automatically determine whether or not to prompt the user based on contextual factors.

During the one week study, the 36 phones logged 27 million application requests to protected resources, translating to 100,000 requests per user/day. The authors found that 75.10% of the permissions were requested by apps that are invisible to the user (most of these were requested when the screen was turned off, which is 95% of most phones lifetime)[3]. Using the data they collected, they analyze which permissions are requested and the different ways in which certain information can be accessed (e.g., there are multiple ways to access location). They argue that due to invisibility, frequency and exposure, what the authors’ have dubbed as users’ contextual integrity – meaning what they expect from apps and their permission behavior – is violated.[4]If we were we to describe the study using

[2] For example, if Spotify requests a wi-fi scan while the user is playing Solitaire, then visibility is set to false, the history shows that prior to the Spotify prompt, the user had viewed Solitaire, the screen status was on etc.

[3] The applications making the most permission requests are Facebook, Google Location Reporting and Facebook Messenger.

[4] The user study provides greater insights as to when users feel that their expectations are not met which is worth reading but too detailed for the study at hand.

The result of the exit survey shows that users’ decision to block a permission request was based on a variety of contextual factors. When asked why they would want to block certain permissions, 53% of survey subjects stated that they didn’t think the permission was necessary for the functionality of the application. However, users do not categorically deny permission requests based solely on the type of resources being accessed by an app. They also take into consideration how much they trust the application and whether they are actively using it. Moreover, the status of the screen and whether the app is in the foreground has an impact on whether users are more likely to appreciate the permission type in their decision.

The authors use these insights to develop a classifier that can infer when the user is likely to deny a permission request and prompt for explicit run-time permissions. Their classifier makes use of originating

Dourish's vocabulary, we would say that the authors study which of these factors users consider to be "contextual" to their interaction with their apps and mobile devices. This sets this study apart from typical context-aware computing papers that have a more static view of what counts as context. A more detailed discussion on how contexts are handled in the different papers can be found in Section 2.4.2.

While the authors lean on CI, they do not make explicit use of the parameters part of the conceptual framework nor invoke its normative aspects. Implicitly, we can interpret the model that underlies the study to treat users as senders, apps as recipients, type of data as a kind of contextual factor. Moreover, we can regard the permission prompts as implementing transmission principles that make select information flows conditional on user's approval. However, by evaluating appropriateness of information flows with respect to an app, rather than the social context that the app serves, the study also falls short of understanding user expectations with respect to information flows that may be initiated by the organization, be it sharing user data with other companies or users (see RQ3). In general, relying on users expectation as a normative source leaves out other potential sources of information norms which may have been very useful in further pruning those prompts that request permissions for inappropriate information flows. As a result, the authors clearly deviate from the normative ambitions of the framework and hold its conceptual premises only in our interpretation.

Foregrounding apps does reveal interesting results that go beyond what is typically in the scope of a CI analysis (see RQ4). First, the authors find that users wanted to accept some permissions because they were convenient and others they wanted to reject because they requested access to privacy sensitive information (e.g., SMS messages) regardless of the social context. Second, users were more likely to expect and accept requests that occurred in the foreground of an application than in the background, and they were more likely to want to block a permission if it was from an app or process in the background, too frequent or when the phone screen was locked. In other words, users consider additional factors when it comes to evaluating the appropriateness of an information flow. This result stands to inform CI by pointing out the need to acknowledge technical and operational contexts, which we discuss in Section 4.2.

application, permission and visibility for prompting users as well as personalization factors to meet users' contextual expectations. They complete the study of this classifier with a short evaluation of its accuracy.

In their reading of contextual integrity, the authors abstract away the social contexts of apps (see RQ2). They are not concerned with the information norms an app may be subject to due its social context, e.g., is it appropriate for a health app to collect user location? Rather, they equate privacy violations with occurrences of the collection of personal information in ways that defy user expectations in the context of an interaction based on their list of contextual factors. Starting from this definition, they go on to study those permissions and contextual factors that are most likely to defy users' expectations and that may be good candidates and situations for prompting users at run-time.

2.3.3.3 Implicit Contextual Integrity in Online Social Networks (Criado and Such)

In this fascinating paper coming from the field of Multi-Agent Systems in Artificial Intelligence applied in the context of Online Social Networks (OSNs) the authors theorize and develop an agent that responds to the problem of "Implicit Contextual Integrity".[5] The main motivation for the authors is to introduce mechanisms that address issues related to "inappropriate exchanges and undesired disseminations" that happen due to lack of effective privacy controls in OSNs (see RQ1). Pointing to numerous studies in computer science, the authors argue that contextual integrity, a model upon which effective computational mechanisms can be built, is the right framework for developing effective controls for OSN users. However, prior computational models have assumed the existence of well-defined contexts with predefined roles and explicit norms. These are not always available in OSNs, as context, roles and associated informational norms are "implicit, ever changing and not a-priori known".

In order to support users with these implicit norms, roles and contexts, the authors propose an Information Assistant Agent (IA-agent) that can infer the context that the user is in and the information norms belonging to that context. In describing their solution, they first present an information model for Implicit Contextual Integrity and then characterize the IA-agent. The agent uses the information model and further modules to learn implicit contexts, relationships and the information sharing norms. It uses this information to warn users before they make potentially inappropriate information exchanges (within a context) or engage in undesirable dissemination of information previously exchanged (across contexts).

Criado and Such leverage a plethora of techniques available to them to compose a formalization of appropriateness that can be used by the IA-Agent. First, they assume that each information exchange can be mapped to a range of finite topics, e.g., that a post about a tennis match is about sports. The frequency with which certain topics gets mentioned by members of a context is crucial – messages pertaining to topics that are rarely mentioned are considered inappropriate and vice versa. Some exception is made, however, to infrequently communicated topics: if reciprocity underlies a given communication between members of the context, then the information flow is reconsidered as appropriate. Furthermore, if information on a topic has been previously disclosed in a given context, then a repeat disclosure in that context is not seen as inappropriate, and hence is not regarded as entailing new privacy risks. Appropriateness of a topic may increase if members of a context start exchanging messages on the subject. It may also decay, as information flows pertaining to the topic decrease or disappear.

A message may flow to people in multiple contexts, in which case it is assumed to be flowing to the context with most recipients. For example: if Mary is Alice's friend and workmate, and Alice sends a message to Mary and three other people from her work context, then it is assumed to be a message flowing in the work context. The agent also

[5] The authors have written two papers with the same title. Here we refer to the longer and more detailed version published in the Information Sciences Journal 325 (2015) 48-69.

takes into consideration whether the information in a message is known in the different contexts shared by the recipients of that message. Since the IA-agent needs to keep track of frequency, past mentions and reciprocity, the corresponding design requires keeping track of all past communications.

In summary, the total appropriateness of a given information flow is based on three different metrics: appropriateness of topic to individuals, appropriateness of an information flow in a context, and appropriateness of a message across contexts. Four modules allow the IA-Agent to complete its tasks:

1. Community finding algorithm: identifies new contexts made up of densely connected members.

2. Passing time function: updates appropriateness of information flows over time also based on the knowledge about different topics in a context.

3. Message sending function: uses received messages to update the different appropriateness and knowledge functions.

4. Message reception function: processes messages before they are sent to either flag them to the user as inappropriate, avoid undesirable dissemination of previously exchanged information, and update appropriateness and knowledge functions.

The authors conclude the paper with experiments based on simulations of exchanges among multiple IA-agents. The results show that the agents are able to infer information sharing norms even if a small proportion of the users follow the norms; agents help reduce the exchange of inappropriate information and the dissemination of sensitive information; and, they minimize the burden on the users by avoiding raising unnecessary alerts.

This paper mostly remains faithful both to the definition of CI as well as its parameters (see RQ2). The model includes sender, receiver, messages, topics and context. The authors make no explicit comments about the transmission principle, however, one could argue that the agent implements transmission principles: information may flow as long as it passess the contextual norms of a context, or norms of dissemination across contexts. Otherwise, the user is presented with an alert which gives her an opportunity to double-check on the appropriateness of an information flow.

The authors assume that contexts emerge in interaction, an approach very much aligned with Dourish [42]. Contexts are not predefined, but as communities of users establish connections and communications they are detected by the “community finding module”. Hence, users’ communication patterns, networking patterns as well as the IA-agent become sources of normativity (see RQ3). This implies a division of labor between the user and the agent: the agent plays an active role in maintaining informational norms and the user is still able to practice discretion when it comes to determining what is considered an appropriate information flow in a context.

In addition, the authors introduce a number of distinctions and parameters that go beyond those of CI (see RQ4). The distinction between inappropriate exchange and undesirable dissemination allows the authors to express norms with respect to information flows within and *across multiple contexts*.

The different functions for frequency, reciprocity and prior knowledge give the authors the tools to explore adaptivity of informational norms throughout time and in multiple contexts. This allows the authors to capture norm development and also make explicit the role that users play in maintaining norms. In many ways, the “implicit CI model” the authors introduce is complementary to CI, in that it provides means to extend social norms in a context with changes to those norms through interactions over time. Adaptivity, multiple-contexts, temporality and user engagement in contextual norms are further discussed in Section 2.4.4.

The model underlying the IA-Agent also exhibits some differences in interpretation of aspects of CI. The agent relies on frequency of exchanges on topics as a means to infer norms. Norms are not the same as the most frequent information flows, nor would such a definition do justice to topics that are pertinent but infrequently exchanged.

Finally, the proposed IA-agent helps maintain contextual integrity but is outside of the scope of CI analysis. The appropriateness of information flows to the OSN provider, the provider of the IA-Agents as well as other third parties is not discussed. It is as if CI only applies to social relations but the service providers are outside of the scope of CI. This leaves out questions like whether an IA-Agent should compile and keep all past communications of all members of a social network, and if so, who can have access to the Agent’s memory? This aligns with industrial practices where OSN companies claim that they are only facilitating information flows deemed appropriate by their users. It is possible to argue that what norms should apply to an IA-Agent is too much to ask of a single CS paper. However, this type of scoping is not exceptional among the papers we found and worthy of a lengthier discussion which we come back to in Section 2.5.

2.4 Results

Through our study of 20 computer science papers invoking contextual integrity we discovered a variety of themes and innovations in privacy engineering that also reflect on improvements to the privacy framework. After parsing each paper into our review template (see Section 3), we coded our results and surfaced a number of recurring themes. We then consolidated these themes into answers to our research questions. We discuss those answers in this section.

The papers included in the survey were: Barth et al., 2006 [9]; Barth et al., 2007 [10]; Criado and Such, 2015 [30]; Datta et al., 2011 [35]; Jia et al., 2017 [71]; Kayes and Iamnitchi, 2013a [73]; Kayes and Iamnitchi, 2013b [74]; Krupa and Vercouter, 2012 [77]; Netter et al., 2011 [94]; Omoronyia et al., 2012 [102]; Omoronyia et al., 2013 [103]; Salehie et al., 2012 [117]; Samavi and Consens, 2012 [118]; Sayaf et al., 2014 [120]; Shih and Zhang, 2010 [127]; Shih et al., 2015 [128]; Shvartzshnaider et al., 2016 [129]; Tierney and Subramanian, 2014 [137]; Wijesekera et al., 2015 [143]; and Zhang et al., 2013 [146].

The three categories we derived to answer RQ1, RQ3, and RQ4, and the themes within each category as they apply to each paper, are in Table 2.2 at the end of Section 2.4.4. A separate table (Table 2.1) is provided for our results for RQ2 in Section 2.4.2. In Sections 2.4.1-2.4.4, we detail each set of results and its relation to our research questions. Blank fields in the tables stand for cases where none of the themes in our taxonomy were applicable to respective papers in question.

2.4.1 (RQ1) Architecture

Our first research question was "What kind of problems and solutions do computer scientists use CI for?" CI is a philosophical theory of privacy with social and legal implications that are designed to apply across a wide range of technologies. Computer scientists do not have the luxury of this insensitivity to technical detail. Their work reveals how specific classes of technical architecture have different socially meaningful implications for privacy.

There was variation in the kind of system described in each paper. Far from being neutral with respect to the way CI was used by the papers, focus on different technical architectures resulted in different manifestations of the privacy theory. Some themes within this category were **user interfaces and experience** (2.4.1.1), **infrastructure** (2.4.1.2), and **decentralized** architectures (2.4.1.3).

2.4.1.1 User Interfaces and Experiences

Four papers surveyed (Shih and Zhang [127]; Shih et al. [128]; Zhang et al. [146]; Wijesekera et al. [143]) studied the experience of users with applications or with designing user facing interfaces or applications. Since contextual integrity theory operates at the level of social norms and says little about user interfaces, and user experience in different situations, these papers raise the question of how user-facing apps and their interfaces are related to broader normative questions about what is appropriate information flow in differentiated socio-technical situations. One paper (Zhang et al. [146]) explicitly drew its motivation from the FTC's view of the importance of "context of interaction" rather than a broader social or normative view of privacy. Nevertheless, these cited contextual integrity as part of its motivation and study set up. This prompts contextual integrity theorists to address the theoretical connection between 'context of interaction' and social spheres.

In general, these papers were not concerned with modeling social norms of a large population of users. Rather, they were more concerned with individual user's activity and their interaction with a device in different situations. Situations could be environmental conditions (e.g., where the user is located, night or day); social situations (e.g., work, home, among friends); or, technical situations (e.g., whether an app is in use when it asks for permissions), comparable to conceptions of context a la Dey et al. [39].

Two themes were common to these papers: they implicitly highlighted that in addition to what may be appropriate flow of information in different *social spheres*, users may have further criteria for what information flows they expect or prefer in different *situations* (See Section 2.4.2 for further discussion on this topic). Second, these papers aspired to generalize their results in order to provide recommendations concerning infrastructural design that can be implemented to respect contextual integrity. For example, Wijesekera et al. [143] propose introducing techniques to improve Android permission models to better cater to user preferences and expectations in different interactional situations.

2.4.1.2 Infrastructure

Many of the papers in our sample were about formal models for or techniques specific to systems that serve as *infrastructure*. By infrastructure, we mean technology that is designed to cater to a large set of users and diversity of applications. We distinguish between social and technical platforms since they raise different kinds of challenges to applying CI in practice.

Social platform: A social platform is a technology that mediates social interaction as an affordance or service. In the papers we surveyed, it was used synonymously with online social networks (OSN) and social ecosystems (Kayes and Iamnitchi [73]), examples varying from Facebook, Snapchat, SMS, to Amazon and Google Play store reviews, and email.

From the perspective of contextual integrity and privacy, what is most pressing about social platforms is how they can potentially mediate activity pertaining to multiple social spheres. Friends, family, classmates, work associates, and so on may all interact using the same social platform.

This poses a challenge to contextual integrity because while the framework is well tailored to evaluating the ethics and impact of a particular technology by identifying the (singular) context it is in, social platforms are designed to mediate more than one social context and perhaps to create entirely new social spheres through their use.

In order to accommodate the uses of the technology in multiple spheres simultaneously, computer scientists are challenged with modeling not just the norms within a single social sphere, but contexts in general and how they interact. Contexts may be very fluid in social platforms. Papers we reviewed looked at scenarios where contexts may collapse; multiple contexts may produce conflicting norms [137]; contexts and social norms may change over time [94] [120]; and, as in the case of Criado and Such [30], how contexts may emerge in interaction. Contextual integrity scholars have not yet provided much guidance on how to deal with the fluidity of social contexts and its impact on how to interpret informational norms, leaving computer scientists to come up with creating solutions themselves.

Note that the definition of a social platform is agnostic about the particular implementation or location of the technology that undergirds a social service. The technology may be distributed, federated or centralized; include apps on a smartphone; web pages in a browser; servers hosted in a "cloud", and telecommunications infrastructure supporting information flow via Internet protocols. This technical complexity is addressed in what we call *technical platforms*.

Technical platform: A technical platform is a technology that mediates the interactions between other heterogenous technologies connecting multiple users. Examples include Android smartphones [143], Platforms for Smart Things [71], the Web (and web browsers) and Smart Grids [117].

A difficulty in defining "technical platform" is that the technology in question is often designed as a "stack" with multiple layers, each layer being a 'platform' on which the next one operates. Hence there are many technologies, such as Facebook, that are both in a sense "applications" that stand on technical platforms and are also technical platforms in their own right as they mediate other applications through a developer API. This is in part due to a design principle that has influence on the Internet [25] that recommends having as few controls as possible introduced on each layer to allow for a wider range of possibilities at the higher layers.

From the perspective of contextual integrity, the challenge with analyzing technical platforms is that they necessarily involve the participation of many social actors who may access and process data (and especially personal information) flowing through them. In contemporary applications the actors involved with operating the technical platform are subject to a number of technical, legal and social norms, some of which are substantial to the social contexts their users see themselves as operating in. We tentatively propose the concept of "operator context" that defines the roles and norms of the operator of communications infrastructure that acts between users.

Formal models: By formal models we refer to papers that conceptualize frameworks that can be part of an infrastructure that serves many different social contexts or technologies, but the implementation details of which are either irrelevant or considered only at an abstract level. Such papers come with verification of the consistency and completeness of the formal model as well as a prototype to show the feasibility of actually implementing the system. These papers provide useful insight into how CI can be operationalized, raising issues at the logical level that are difficult to surface in more empirical work.

Examples include papers on access control models that preserve contextual integrity in an enterprise, like Barth et al. [9] and Barth et al. [10]; frameworks that describe and evaluate ontologies to audit privacy relevant processes in a linked data environment [118]; or adaptive systems that monitor when new threats arise, reconfiguring information flows to continue matching user privacy requirements [102], or that identify when information norms themselves change [129].

2.4.1.3 Decentralization

The rare paper in our sample dealt with a specifically challenging technical feature: decentralized architectures. We highlight this theme, however rare, because of the way it positions technology relative to social spheres and interactions. While user interfaces and experiences are connected to individual users (and their expectations), social platforms are central and common to a large number of diverse social contexts. Decentralized architectures have an interactivity and topological complexity that mirrors that of society itself, and trust and reputation mechanisms come to play a greater role in the absence of a centralized entity that can arbiter information norms. We look forward to more papers on this theme and CI.

2.4.2 (RQ2) What did they talk about when they talked about context?

Our second research question was "How have the authors dealt with the conceptual aspects of CI?" Contextual integrity theory has a specific understanding of context as social sphere, parameterized by roles, norms, purposes, and values. The norms are parameterized by their actors (senders, receivers, and subjects in contextually defined roles), information topics, and transmission principles. We wanted to know whether and how computer science papers used this conceptualization of privacy. We found that while several papers drew closely from the concepts in CI, others represented context very differently. As we have discussed, many computer scientists interpreted 'context' in a way that draws from the research field of ubiquitous computing (See Section 2.2.2.1). Because of these discrepancies, we have chosen to focus on the nuanced differences in how context is represented rather than on which of the parameters are used.

We have coded the way each paper has defined and used context across five binary dimensions, which we have named: substantiality, domain, stability, valence, and epistemology. Within each dimension there are two opposed poles.

- **Substantiality**. Some papers discuss contexts as an **abstract** type or ideal of a situation. Others discussed contexts as **concrete** happenings and situations. *Example: hospitals in general are an abstract context. Mount Sinai Beth Israel hospital in Manhattan is a concrete context.*
- **Domain**. Some papers discuss **social** contexts, defined by configurations of people to each other. Others discuss **technical** contexts, defined by objective properties of mechanical devices and the environment they were in. Some papers understood contexts as combining **both** social and technical factors. *Example: a classroom with a teacher and students is a social context. A language education mobile app that prompts the user with questions and sends results back to a server for analysis is a technical context.*
- **Stability**. We draw on Dourish [42] for this distinction. Some papers treat context as a **representational** problem, as if they were stable, delineable, and distinct from the activity that contained them. Others treat them more as an **interactional** problem,

as arising from interactions between people and things, defined by specific circumstances. *Example: The Oval Office in the White House is a stable context. A flash mob is an interactional context.*

- **Valence**. Some papers see the **normative** aspects of privacy as being inherent in context. Others treat contexts merely **descriptively**, without normative force. *Example: A conference Code of Conduct is an account of norms inherent in a context. A list of attendees, keynote speakers, and program committee members is a description of the context.*

- **Epistemology.** Some papers adopt a **model-building** approach to defining contexts. They posit a schema or model of context and derived conclusions from it. Other papers take a more **empirical** approach, deriving context definitions from data. A parameterized definition of a context, e.g., context is location, time, and activity, is an example of a model based approach, whereas applying traffic and topic analysis to communications in order to surface contexts is an example of an empirical approach that can be used to characterize different contexts.

We note that as far as CI is concerned, it is essential that contexts be understood as **normative**, as one important trait of contexts is that they have ends, purposes, and values. They are **social** contexts, pertaining to relationships between people in defined roles, but they are oriented around functions, purposes, aims, goals, activities, values, etc. As these social norms evolve in society in general and then are applied to particular cases of information flow, contextual integrity conceptualizes contexts **abstractly**. "Context" interpreted to mean *sphere*, as discussed above, has these three properties (i.e. they are normative, social, and abstract). To the extent the papers draw on different meanings of context, they diverge from CI. For example, when the literature interprets context as *situations*, as discussed in Section 2.2.2.1, it conceptualizes contexts as **concrete** and at least partly **technical**. Our study has surfaced that computer scientists, in trying to make CI actionable, have encountered the problem of applying abstract social norms to concrete socio-technical situations.

Table 2.1: Results from RQ2. Stability codes: REPRESENT = Representational; INTERACT = Interactional.

Paper	Substantiality	Domain	Stability	Valence	Epistemology
Barth et al. [9]	ABSTRACT	SOCIAL	REPRESENT	NORMATIVE	MODEL
Barth et al. [10]	ABSTRACT	BOTH	REPRESENT	NORMATIVE	MODEL
Criado and Such [30]	CONCRETE	SOCIAL	INTERACT	NORMATIVE	EMPIRICAL
Datta et al. [35]	ABSTRACT	SOCIAL	REPRESENT	NORMATIVE	MODEL
Jia et al. [71]	CONCRETE	TECHNICAL	REPRESENT	DESCRIPTIVE	EMPIRICAL
Kayes and Iamnitchi [73]	CONCRETE	SOCIAL	REPRESENT	DESCRIPTIVE	EMPIRICAL
Kayes and Iamnitchi [74]	CONCRETE	SOCIAL	REPRESENT	NORMATIVE	MODEL
Krupa and Vercouter [77]	ABSTRACT	SOCIAL	REPRESENT	DESCRIPTIVE	MODEL
Netter et al. [94]	CONCRETE	SOCIAL	REPRESENT	NORMATIVE	EMPIRICAL
Omoronyia et al. [102]	ABSTRACT	BOTH	REPRESENT	DESCRIPTIVE	MODEL
Omoronyia et al. [103]	ABSTRACT	BOTH	REPRESENT	DESCRIPTIVE	MODEL
Salehie et al. [117]	CONCRETE	BOTH	INTERACT	DESCRIPTIVE	EMPIRICAL
Samavi and Consens [118]	ABSTRACT	BOTH	REPRESENT	NORMATIVE	MODEL
Sayaf et al. [120]	CONCRETE	SOCIAL	INTERACT	DESCRIPTIVE	EMPIRICAL
Shih and Zhang [127]	BOTH	BOTH	REPRESENT	DESCRIPTIVE	EMPIRICAL
Shih et al. [128]	CONCRETE	TECHNICAL	INTERACT	DESCRIPTIVE	EMPIRICAL
Shvartzshnaider et al. [129]	ABSTRACT	SOCIAL	REPRESENT	NORMATIVE	BOTH
Tierney and Subramanian [137]	BOTH	SOCIAL	REPRESENT	NORMATIVE	MODEL
Wijesekera et al. [143]	CONCRETE	TECHNICAL	INTERACT	DESCRIPTIVE	EMPIRICAL
Zhang et al. [146]	CONCRETE	TECHNICAL	INTERACT	DESCRIPTIVE	EMPIRICAL

The first paper of our study in publication order is Barth et al. [9], which we have detailed in Section 2.3.3.1 of this paper. Helen Nissenbaum is a coauthor and the paper includes a summary of contextual integrity theory. The technical contribution of the paper focuses on a “fragment” of contextual integrity. It is this technical contribution that we have assessed according to the criteria above. The Barth et al. [9] technical presentation of context is as one that is **abstract, social, representational, normative,** and **modeled**. Their work models the specific normative logic of contextual integrity.. It shows how norms and laws can be represented as abstract policies amenable to automated enforcement.

This paper is one end of a spectrum. Other papers in our sample drew their understandings of context from other traditions, including ubiquitous computing (discussed in Section 2.2.2.1 of this paper).

Following Dourish [42], some papers eschewed explicit abstract representational modeling of context for what resembles interactional views of context derived from empirical data about user behavior or human-computer interaction. Several papers considered the narrow context of a user and their device, as opposed to social relations more generally. Most papers did not see norms as inherent to the contexts they studied, but rather saw contexts descriptively. (Some of these papers sourced their normativity from other factors, see Section 2.4.3). Our paper exemplars (2.3.3.1-2.3.3.3) provide deeper explanations of the dimensions used to classify contexts here.

What we have discovered in answer to RQ2 is the distribution of papers across these dimensions. This tells us how well contextual integrity as a conceptual theory of privacy has made it into computer science. CI conceptualizes contexts as **normative** and **social**. Papers that have modeled context as either purely technical or purely descriptive have missed some of the core intent of CI.

To the extent that it sees the formation and maintenance of a social context as an adaptive social process, we argue that contextual integrity is consistent with the **interactional** view of contexts from Dourish [42], though in its concrete application it has a tendency to work from a **representation** of context. We believe this leads to deep sociological questions about how social norms and purposes, which can seem **abstract** and theoretical, can form from **concrete** human interactions.

We note with special interest Criado and Such [30], detailed in Section 2.4.1.3, which stands out as a paper that addresses a particularly difficult challenge. It is the only paper in our sample that manages to be both concrete, interactional, and empirical as well as socially normative. We see this as an important innovation in the use of contextual integrity in computer science.

2.4.3 (RQ3) Source of Normativity

Our third research question was, “How have the authors dealt with the normative aspects of CI?” In contextual integrity, the normative (in the sense of prescriptive or ethical) force of information norms comes from the purposes, ends, and values associated with each social sphere. This complex metaethical theory rarely finds its full expression in the computer science literature. Instead, the papers in our sample take a variety of narrower positions, implicitly or explicitly, on the source of normative values that motivate the importance of privacy.

The subsections here explain the themes we found in this category. **Compliance and policy** refers to when normativity was taken from legal authority or some other unquestioned source of policy. **Threats** refers to the computer security research practice of positing a threat model to motivate research. **User preferences and expectations** locates the source of normativity in the subjective perspective of individual users. **Engagement** refers to designs that allow users to dynamically engage with each other to determine norms.

2.4.3.1 Compliance and Policy

Some of the papers in our sample took their motivation from the practical problem of compliance with legal regulation, such as HIPAA. These papers effectively outsource their normative questions to the legal system. They at times argue as if compliance is relevant because it is internalized as a business interest [10]. One line of this compliance-based research is contiguous with other work on formalizing privacy regulations in ways that are less closely tied to contextual integrity [41]. Datta et al. [35] synthesize the contributions of this research trajectory.

Other papers are less specific about source of the specific form of their restrictions, but nevertheless have an explicit mechanism for stating *policy*. Some computer research in this field culminates in the articulation of a policy language, which is valid for its expressivity, not for the specific character of the content of any particular expression it allows.

In both the cases of compliance and policy, normativity is exogenous to the technical design.

2.4.3.2 Threats

Some of the papers motivated their research goals in terms of privacy *threats*. These presumably adopted this stance as a continuation of practices from security research, which typically posits a threat model of potential attacks and adversarial capabilities before detailing how a technical improvement can mitigate these threats.

Taking this position alleviates the researcher from having an overarching theory of privacy; they can instead work from specific cases that are plausible or undisputed cases of privacy violation.

2.4.3.3 User Preferences and Expectations

Some papers motivated their research either explicitly or implicitly in terms of whether a technical design was able to meet user preferences or expectations of privacy. Preferences and expectations are not the same thing, but they are related in that they depend primarily on the individual subjectivity of the user. A user's expectation is the outcome they desire or is in their acknowledged interest, and a number of papers explore the expectations users have in different social or interactional context. User preferences on the other hand were often used to study what kind of controls users may prefer to have or exercise when using systems. User perceptions also played a role in papers where researchers explored what information flows users noticed or how they perceived them [143] [146].

Measuring user expectations and preferences as a way of assessing the appropriateness of information flow is consistent with contextual integrity. This can be done explicitly through survey methods, as is done by Shvartzshnaider et al. [129]. In CI, appropriateness is a function of social norms, and these norms do codify social expectations and values. Certainly in some cases user expectations will track social expectations. But though they are related, we caution researchers against conflating social norms with user expectations and preferences. This is because individual users are more prone to becoming unreflectively habituated to a new technology than society as a whole. Also, individual user preferences may at times be opposed to the interests of society. We have identified elaborating on the relationship between individual preferences and social norms as a way to improve CI.

2.4.3.4 Engagement

Some papers explicitly articulated mechanisms through which users could engage with a system to define what's normative for the system. Rather than accept a policy or threat model exogenously or see an individual's opinions and satisfaction as the ends of design, these papers allowed for the establishment of norms to be a dynamic social process accomplished through use of the technology itself. For a more in depth discussion of how this can work, see the more detailed discussion of Criado and Such [30] in Section 2.3.3.3. Another example is Tierney and Subramanian [137] who describe a marketplace or library of abstract context definitions, complete with roles and access controls corresponding to transmission principles, that are developed by a community of context designers. Users can then instantiate the context template that best fits their social needs.

2.4.4 (RQ4) Expanding Contextual Integrity

Our fourth research question was "Do the researchers expand on contextual integrity?" The rigors of computer science led many paper authors to innovate and improvise as they used contextual integrity in their designs. We grouped these innovations into the category **Expanding Contextual Integrity**. We found many papers were engaged in developing mechanisms for technological **adaptation** to changing social conditions (2.4.4.1). Some addressed the challenges associated with technologies that operated within **multiple contexts** at once (2.4.4.2). Some developed ideas concerning the **temporality and duration** of information and how this affects privacy (2.4.4.3). Others were particularly concerned with **user decision making** (2.4.4.4) with respect to privacy and information controls. While all these innovations are compatible with contextual integrity as outlined in Nissenbaum [99], we found the detail with which the paper authors engaged these topics showed ways to expand contextual integrity.

We note that many of these themes echo discoveries made with respect to our other three research questions. For example, those papers that addressed the design of social infrastructure (see Section 2.4.1) had to address the problem of how to handle multiple contexts in the same technology, and as they did so they had to make decisions about how to represent context that did not necessarily accord with CI's concept of context as social sphere (see section 2.4.2). Of the four research questions, this one reflects the technical accomplishments discussed in Sections 2.4.1-2.4.3 back on CI in order to identify the limits of the framework itself. Table 2.2 shows how themes from different research questions were distributed across the papers in the survey.

2.4.4.1 Adaptation

The most common way in which computer science papers expanded on contextual integrity was to address questions of social adaptation.

As noted in Section 2.2.1 above, CI theorizes that norms are the result of a process of social adaptation. Social spheres have ends, purposes, and values robustly as a function of their evolution. Norms within these spheres are legitimate to the extent that they serve their contextual purposes, but environment changes (such as the prevalence of new digital technologies) are the stimulus for further adaptation of norms. To the extent that CI has a conservative impulse, it is to warn against the agitation caused by disruptive technologies that change the environment too quickly for social evolution to adapt.

This grand theory of privacy is not actionable for computer scientists. In the papers we found that dealt with *adaptation*, the researchers were interested in designing technology that is responsive to social change at a much smaller scale in both space and time. Criado and Such [30] discuss the adaptation of an informal sports discussion group emerging out of a collegial working forum. If large-scale evolution of social spheres and privacy norms depends on variation on the level of social interaction, it is challenging to design technology that keeps up with this variation. If large scale agitation about threats to privacy happens when technology disrupts a shared social expectation, then small scale agitation can occur when technology fails to address emerging norms. For computer scientists to deal with these challenges, they have to be more specific about these processes of adaptation than CI currently is.

Many of the papers we reviewed concerned themselves with the problem of maintaining contextual privacy under conditions of social change. Few adopted the theory proposed by Nissenbaum [98]; instead these papers proposed their own mechanisms to account for and capture changes in context and norms. Most often these did not take into account the stability of contextual ends, purposes, and values. Rather, they generally took on the problem of having technology react appropriately to exogenous social change. Criado and Such [30] design agents that guess rules for appropriate information flow from regularities in user behavior. Shvartzshnaider et al. [129] experiment with a method for empirically surveying for opinions on social norms and translating results into a logical specification. Such mechanisms could be used to build a system that is robust to changes in social opinion.

Papers that addressed social adaptation were likely to also use concrete, interactional and empirical concepts of context (see Section 2.4.2). Some designed methods to have users engage in the process of determining norms (see 2.4.4). In general, technical systems that are adaptive to changes in social behavior can be prone to the failure of maladaptation. To be actionable for these designs, CI would benefit from more specificity regarding the process of social evolution that legitimates the norms of social spheres.

2.4.4.2 Multiple Contexts, and Context Clash

Another common way in which computer science papers expanded on contextual integrity is that many discussed technologies that recognized the existence of multiple contexts at once. This was common for those papers that addressed the design of social infrastructure (see Section 2.4.1), for example. Contextual integrity as a privacy framework posits many different social spheres with different norms of information flow. But as it is currently resourced, CI provides little conceptual clarity as to how different contexts relate to each other, and no normative clarity as to how this multiplicity of contexts affects the appropriateness of information flow.

As a result, many of the paper in our study improvised solutions to the problems associated with representing multiple social contexts. In some, system users were registered or detected as being in one or another context, with shifting access control policies in a context-appropriate way, something the agent in the Criado and Such [30] paper is tasked with reasoning about. Some papers accommodated the relationship between contexts through a mechanism of context adaptation (see above). Others addressed the specific problem of what happens when information *flows between contexts*. For example, Sayaf et al. [120] raised the privacy concern that a photograph might move from a context where it was interpreted as a swimsuit advertisement into one where it was sexually objectified.

All the papers that dealt explicitly with the problem of using CI when multiple contexts affected a situation used a concrete and empirical concept of context (see Section 2.4.2). This points to an insight about CI that we see as a research finding: a more actionable CI would address how situations (concrete context) can be empirically analyzed to determine which sphere or spheres (abstract, normative, social contexts) apply. For example, could a system that monitors communication within a university in general classify a particular message as belonging to a classroom, employment, or social sphere? It may be possible to formulate this as a machine learning problem.

2.4.4.3 Temporality and Duration (Read/Write)

Several of the papers in our sample extended contextual integrity by explicitly addressing restrictions or allowances on information flow based on the timing of flows. For example, a flow might be allowed after the sender has received permission, but not before, or until certain actions are completed in the future. These extensions are not a challenge to contextual integrity as a theory; they are fully within the scope of what is possible as a *transmission principle*. However, the specific elaborations of the relationship between timing and information flow policies were notable.

A related theme which does more conceptual work within contextual integrity is that of data's *duration*. In technical terms, this was expressed in our sample as restrictions of reading, writing, and deleting data, as found in Kayes and Iamnitchi [73]. These operations stretch the idea of information "flow" so much that they perhaps require an entirely different notion, that of information "stock".

Another line of research discusses the relationship between temporality and the possibility of privacy policy enforcement. Datta et al. [35] note that some aspects of privacy policies cannot be completely enforced at the time when information flows because the policy mandates what happens *after* the flow. For example, some policies impose restrictions on how information is used.

2.4.4.4 User decision making

Contextual integrity as a theory of privacy abstracts away from individuals in order to provide a normative framework that is independent of specific actors and their interests. It is this stability that gives it much of its normative power. Nevertheless, many computer science papers that used contextual integrity were concerned with user's individual decision making.

While voluntarity is one factor that can affect the transmission principles of information norms, contextual integrity has little to say about the role of the individual in shaping norms and social contexts more generally. These computer science papers put emphasis back on the individual and her decisions in context.

Table 2.2: Results from RQ1, RQ3, and RQ4. RQ1 Codes: FM = Formal Model; I:S = Infrastructure:Social; I:T = Infrastructure:Technical; UI = User Interface. DC = Decentralized. RQ3 Codes: COMP = Compliance. ENGAGE = Engagement. UP = User Preferences. UE = User Expectations. RQ4 Codes: TEMP = Temporality. ADAPT = Adaptation. MC = Multiple Contexts. UDM = User Decision Making.

Paper	**RQ1. Architecture**	**RQ3. Source of Normativity**	**RQ4. Expanding CI**
Barth et al. [9]	FM	COMP	TEMP
Barth et al. [10]	FM	COMP	TEMP
Criado and Such [30]	I:S	ENGAGE	ADAPT, MC
Datta et al. [35]	FM	COMP	TEMP
Jia et al. [71]	I:T	THREATS	
Kayes and Iamnitchi [73]	I:S	ENGAGE, COMP, UP	MC, ADAPT, TEMP
Kayes and Iamnitchi [74]	I:S		
Krupa and Vercouter [77]	DC	ENGAGE, UP	ADAPT
Netter et al. [94]	I:S	THREATS, UP	
Omoronyia et al. [102]	I:S	THREATS, UP	ADAPT
Omoronyia et al. [103]	FM	THREATS, UP	ADAPT
Salehie et al. [117]	I:T	THREATS	ADAPT
Samavi and Consens [118]	I:T	COMP, UP	TEMP
Sayaf et al. [120]	I:S	UP	MC, TEMP
Shih and Zhang [127]	UI	UP	MC
Shih et al. [128]	UI	UP	UDM
Shvartzshnaider et al. [129]	FM	UE	ADAPT
Tierney and Subramanian [137]	I:S	ENGAGE	
Wijesekera et al. [143]	UI	UE, UP	APPS AS ACTORS
Zhang et al. [146]	UI	THREATS	MC

2.5 Findings and discussion

We have summarized the achievements of computer scientists in developing contextual integrity. The research we have reviewed has variously used parts of contextual integrity and innovated on the relationship between privacy and context. Through our analysis, we have identified new research questions and opportunities at the intersection of CI and computer science.

In the time since contextual integrity first emerged, it has attracted useful insights from legal and ethical theorists as well as social scientists. Some of the toughest challenges have come from those seeking to apply CI to problems in their home fields, whether law and public policy or computer science, design, and engineering – the focus of this paper. Like most efforts to apply theory and other abstractions to concrete or real-world challenges, these, too, require that a distance be traveled to leverage the theoretical constructs of contextual integrity, to concrete privacy challenges of computer science, design, and engineering. In traveling this distance, the efforts we have surveyed reveal unanswered questions, conceptual gaps, and realities that do not align fully with the CI model. These findings call attention to several specific ways to expand, explain, and adjust CI in order to make it more responsive to the needs of computer science and engineering researchers seeking to inform their work with a meaningful account of privacy.

In this final section, we present theoretical gaps in CI that our literature survey has exposed, systematically organizing them into four subsections, each associated with our four research questions: RQ1 - Architecture; RQ2 - Character of Contexts; RQ3 - Privacy as a Moral Imperative; and RQ4 - Expanding CI. In each subsection, we describe the nature of the theoretical gaps, i.e between theory and practical application, followed by a discussion of lessons learned that could translate into guidance for those embarking on new technical privacy research and design projects. The task is challenging because although the parameterized informational norms of contextual integrity offer greater specificity than other normative privacy theories, there remains significant room for interpretation. This room for interpretation, on the one hand, is what distinguishes contextual integrity from accounts of privacy that are not adaptive in the face of historical, cultural, social, and even personal variations, but it can be frustrating for those looking for precise, literal rules that are both correct and directly implementable.

For each research question we also have sections that we have titled, “call to action,” in which we discuss the lessons learned from past applications that can positively inform further forays into using CI in privacy research. We encourage the creative spirit we observed in our survey and recommend lessons learned and open questions to inspire future researchers in the field of context integrity.

2.5.1 Architecture: Towards a Modular Contextual Integrity

Corresponding to our **RQ1,** we have discovered that the way contextual integrity is used in technical design depends on the architectural properties of the technology being designed. This presents an opportunity for faceting CI into more specialized programs that are targeted at specific classes of technical problems. At the same time, our study revealed that the technical designs of computer science researchers often bracketed the social roles of those operating technical and social platforms, despite these being central to public discussion and scholarship on privacy and technology.

2.5.1.1 Theoretical gaps

We see the demand for "modular contextual integrity", faceting CI and giving guidelines for design and research at specific levels of the technical stack, be it in designing user interfaces or experiences,technical or social platforms, or devising formal conceptualizations. Providing these guidelines may require that we derive frameworks of heuristics and principles from CI's conceptual and normative facets. We expect to do this in tandem with further elaboration of the fundamental concept of "context" (see Section 2.5.2) and the concept of Transmission Principle, both distinctive of CI and often not well understood, despite their importance for CI's power as a normative, or ethical theory.

An example of a promising strategy to address this problem is to identify and describe social spheres specific to the design, provision, operation, and use of technology. This is especially relevant in those papers where the designers explicitly delegated responsibilities for enabling contextual integrity to technical elements. In the case of Criado and Such [30], the agent co-regulates norms. In Wijesekera et al. [143] apps actively take part in asking for information flows and the authors consider a classifier that would reason as to when information flows may breach contextual integrity. In Samavi and Consens [118], the authors produce an auditing mechanism that checks logs for potential breaches to contextual integrity. These mechanisms are very different with respect to the degree of autonomy they provide to technical agents and those who are operating them. However they all invoke the question: to what extent these mechanisms are subject to the norms of the context they are co-regulating, acting in or auditing? Should these mechanisms be subject to other contextual norms (pertaining to intelligent agents and their administrators)? In the practical world, this is comparable to the question of whether operators of systems and the technical infrastructures they deploy can simply posit themselves as (providers of) communication channels that are not bound by the social context of their users. The papers we surveyed consistently treat them as a product of but not as subject to the application of the CI. We have raised the possibility of an "operational" context, with an 'operator' role empowered with certain privileges and responsibilities with respect to information flowing on the platform. Further guidance on this matter will be pertinent to enabling a holistic application of CI to technical designs.

On a related matter, further guidance is also necessary with respect to systems that provide infrastructure to multiple contexts: Such systems are expected to reflect on the normative aspects of CI while promoting a logic that can provide the flexibility for multiple social or technical contexts with potentially diverging informational norms to co-exist. What role the normative and conceptual aspects of CI can and should play in infrastructure underlying multiple contexts is an open research question.

2.5.1.2 Calls to action

We call computer science researchers to be as explicit as possible about how the technologies they design are situated in the broader complex of platforms, operators, users, and moderators. If there is an implicit hierarchy (such as users whose activities are logged, agents that track all conversations, and auditors who use these records, or an operating system with many dependent applications), computer scientists can be explicit about this and address the privacy and information flow issues resulting from this differentiation of roles. If there are critical roles in the operation of the system (such as an auditor or operator), can privacy tools inspired by contextual integrity be built for them?

Most papers we selected did not focus on social spheres but on situations, proposing techniques that surface or implement informational norms that arise in a representational or interactional context. This focus often meant that in their models the authors did not consider normative rules applying to a specific context, abstracting the social sphere away. Some of this is justified: the intention is to develop designs that are flexible enough to function in different social spheres. It is also possible that the authors are more comfortable making normative judgements about what constitutes a relevant situation, e.g., some combination of location and activity, these also being things that the researchers can measure using sensor data. However, the numerous research papers showing user concerns due to context clash in online social networks, as well as ever increasing public calls for curation of user generated content suggests that lack of attention to informational norms in social spheres may have negative consequences and should not be an afterthought. One way to guarantee this in abstract formal models as well as in infrastructure, is to at least provide a placeholder in associated conceptual frameworks that can be used to express and enforce normative rules when these systems are implemented. Better would be to also consider how well a proposed technique can sustain divergent informational norms pertaining to different social spheres that an infrastructure comes to play a role in. Developing and evaluating systems that enable a fluid interaction between informational norms in a social sphere and user preferences presents itself here as an interesting research question.

2.5.2 Diverse concepts of context

Our investigation into how computer scientists conceptualize contexts when they employ CI revealed diverse and divergent theoretical assumptions. Some researchers were well aligned with CI's concept of contexts as abstract, normative, social spheres; others drew on other traditions such as ubiquitous computing's concept of context as situations including users and technology. Still others supported users co-creating contexts through their engagement with each other and the technology. Some drew inspiration from multiple sources in order to provide a new technical solution to privacy.

This variety of work demonstrates that privacy and context are closely linked. It also demonstrates that *context* is a polysemous (many-meaning) term. The different senses of context have different implications for privacy by design. Our survey suggests that no one sense of context supports either a complete normative theory or technical design, and that there is a rich design space at the interplay between diverse specific meanings.

2.5.2.1 Theoretical gaps

Our investigation revealed inductively that computer scientists use diverse meanings of context that vary across many dimensions. "Context" can refer to something abstract or concrete, social or technical, representational or interactional, normative or descriptive, and a priori modeled or empirically discovered. Only a subset of this space of meanings are addressed by CI in its current form, specifically, in the way it conceptualizes contexts as abstract, normative, social spheres continuously evolving within differentiated society.

It is not surprising that technical designs are concerned with the concrete circumstances of both users and technical applications. In order for CI to be actionable in this sense, what is needed is a theoretical account of how social spheres relate to sociotechnical situations. This account may well address other tensions between the many senses of "context". For example, an advanced CI would be able to guide how to infer from the observed, descriptive details of a situation a model of the norms appropriate to guide behavior within it. This is a philosophical problem, but one that is made urgent by the demands of existing research on privacy by design.

Another theoretical challenge to CI is raised by the Dourish [42] critique of ubiquitous computing. CI's model of contexts as social spheres parameterized by roles, information types, and transmission principles does suggest what Dourish describes as a positivist model of social contexts: contexts as containers of social action with specific expectations and prescriptions associated with them. To the theorist, we raise the question: what is the relationship between the situated, interactional account of context in Dourish and the social spheres in CI? The theory of "Implicit Contextual Integrity" invented by Criado and Such [30] has suggested that the spirit of CI can be extended to social situations that evolve on a much smaller and more specific scale than is currently suggested by CI. Philosophical theorists can work to make this claim more precise.

2.5.2.2 Calls to action

Computer scientists need not wait for theoretical prescriptions to continue to do good work at the intersection of CI and privacy by design. There is much to be done designing systems that address the reality that supporting users' privacy requires skillful creation and moderation of context. We anticipate that the best work will be explicitly aware of the challenges of matching concrete situations with the abstract spheres from which CI posits users get their normative expectations. Applications of CI are especially likely to be relevant in the smart environment applications, where sensors and actuators will interact with many users at once, making it hard to rely on individual preferences and expectations.

Indeed, any one of the dimensions of variation in the meaning of context (abstract or concrete, social or technical, representational or interactional, normative or descriptive, and a priori modeled or empirically discovered) presents a technical problem to computer scientists wishing to implement privacy by design based on CI. One concrete issue that persisted throughout many papers is the scoping of context. Papers, for example, that focused only on the social context, be it either due to their focus on user interfaces and experiences or social platforms, neither considered what we call the operational context, nor did they pay attention to how social informational norms may be impacted by flows of personal information to third parties. It would be very valuable to have studies that not only consider norms of information flow among users or towards an app provider, but also flows to other third parties, like other services, companies or governments. If a study intends to focus only on a subset of the information flows, than the limitations of the results due to this decision should be made explicit. We call upon computer scientists to work on pragmatic solutions to the problems these conceptual discrepancies pose to designers and users.

2.5.3 Privacy as a moral imperative: between bits and norms

One major finding from our investigation of **RQ3** is that of the papers in our review used the normative aspect of contextual integrity as a basis for their technical contributions. In contextual integrity, the normativity of privacy comes from the ends, purposes, and values of social contexts (spheres) as they have evolved over time. These ends, purposes, and values legitimize the norms that determine the appropriateness of information flow, even as technology changes what those norms should be. Computer scientists sometimes acknowledge this aspect of contextual integrity, but they do not ground their technical contributions in it. Rather they draw on other sources of normativity, such as threats, user expectations, or legal compliance, to motivate their work.

For a number of reasons, these moves are understandable. Computer science has not traditionally equipped itself to deal with the hard problems surrounding the origins of ethics and morals. Threat modeling is narrowly pragmatic and has proven to be suitable for engineering purposes. User expectations are measurable and so attractive to those concerned with empirical validity. Considerations of legal compliance are part of the real business logic of functioning organizations. By focusing on these sources of normativity, computer scientists make their research more actionable. But these methods also carry the risk of falling short of socially maintained norms of privacy. Threat modeling may miss the mark; user expectations can be habituated by technology that works at odds with social principle; laws may be unjust. The burden is on contextual integrity theorists to show how its social and philosophical theory of social norms relates to these more concrete factors. In turn, we call computer scientists to stretch towards the social and philosophical sources of normativity. Our survey has shown that such ambition can lead to new technical innovation.

2.5.3.1 Theoretical gaps

Contextual integrity theorists need to address how their metaethical theory, whereby norms arise from the evolution of social spheres, ties in with the concrete sources of normativity used by computer scientists. We have identified three areas that need elaboration.

Contextual integrity must provide a way of translating from the information norms of social spheres into a characterization of enumerated and discrete privacy *threats*. This is connected to the task of deriving mid-level theories of CI for modules of the technical stack (see Section 2.5.1.1).

Contextual integrity must also articulate the special place for user expectations, preferences, and control within the general framing of appropriate flow. This would require greater fleshing out of situations where user control is legitimate, given its importance in the sphere of technical device usage. It would also address questions of how to resolve conflicts between user preferences and social norms, and between users with different preferences and expectations.

Finally, CI theorists must clarify the relationship between social spheres and the law. While there is in the United States an attractive synergy between the structure of sectoral privacy laws (like HIPAA, GLBA, and the like) with the view of society differentiated into social spheres, the relationship between social spheres and the law is less clear in jurisdictions of omnibus data protection laws such as the EU's GDPR [111]. The CI theorist must address what circumstances social norms provide important guidelines to appropriate information flow that are not covered by law, and what advantages they provide to technology designers who heed them.

2.5.3.2 Calls to action

Computer scientists need not wait for passive instruction on the normative goals of their work. Rather, the problem of measuring *social norms*, in contrast to user's expectations, is one that requires technical sophistication. Shvartzshnaider et al. [129] is one example of a paper that takes this task on explicitly in service of contextual integrity.

Computer scientists are particularly well situated to study users' perception and expectation of informational norms in different social spheres (and not independent of them). Developing and evaluating techniques to do so remains an open question. Coming back to Dourish, exploring how far users and different actors can be brought into engage in the evolution of informational norms is another avenue of exploration that has not been exhausted by researchers in our survey. Computer scientists may also consider designing systems that support communities of users' to determine their own norms.

Many of the studies did not consider conflicts among actors: these could be conflicts in informational norms across contexts, in different situations, even for individual users due to how their expectations evolve in relation to norms throughout time. Discrepancies between ideal vs. actual norms may also lead to conflicts. Looking at these conflicts not as something to be designed away but as productive points of departure for engagement presents itself as another interesting research question.

2.5.4 Expanding and sharpening contextual integrity

This leads to our findings from **RQ4**, where we look for aspects of CI that computer science researchers expanded on through their work. Computer scientists sometimes worked through mechanisms of technical adaptation to social change as they tried to respect privacy norms that were grounded in descriptions of concrete social and technical interaction.

2.5.4.1 Theoretical gaps

CI theorists must develop the framework's account of normative change and adaptation. The work we surveyed suggests a need for technical systems that automatically recognize contexts and that are sensitive to the norms of their users, even though social contexts and norms change. What principles can CI offer to adaptive system designers to ensure that these coevolving sociotechnical systems maintain their legitimacy with respect to the purposes of some more abstract social sphere? Do such systems challenge the theory that social contexts are robust in their ability to maintain their purpose? On what grounds would such a system be considered maladaptive? Is there any danger that technology will derail the social processes that reproduce contexts, or can society always be trusted to correct its own use of technology over time? What if powerful actors leverage existing systems with appropriate flows for ends, purposes and values that lack legitimacy? These thorny theoretical questions are both profound and of practical import for system design.

CI must also address the critical "sore" point in the present-day when many systems and devices span multiple contexts. Our inquiry here into the many relevant senses of 'context' sheds light on this phenomenon. Contexts can clash when the norms of multiple social spheres applicable to the same situation conflict with each other. Information can also flow inappropriately between different situations. A more actionable version of CI will address these complex privacy challenges specifically.

CI also needs to better account for the relationship between privacy and time. Some papers in our survey tried to adapt CI to systems in which data did not only flow, but also was stored, processed, and deleted. Current versions of CI do not recognize that sometimes merely holding data (sometimes for great durations) can pose privacy threats. We are considering developing a concept of exposure to characterize this threat. Relatedly, there is a nuance discovered by Datta et al. [35] that is not observed within CI: that it may not be apparent whether a case of information flow is inappropriate at the time that it flows because prescriptions refer to actions in the future. A more mature version of CI would account for the conditions under which parties can be aware of their privacy violations, and how ambiguities can be resolved.

Related to the questions resulting from the ambiguity or incompleteness of privacy norms are questions concerning the relationship between user choice and privacy. CI can in principle accommodate a wide range of preferences and a pluralistic society despite presupposing robust social agreement on information norms and the nature of social spheres. Technology is often designed to maximize adoption to diverse users and consequently can give (or restrict) users' control over how their data flows. A refinement of CI would address how user control and user diversity relate to social norms.

2.5.4.2 Calls to action

Computer scientists have already made significant contributions to CI by providing valuable exemplars of research on adaptation, multiple contexts, temporality and duration, and user decision making. This work is invaluable for the evolution of CI.

We see further potential at the intersection of information theoretic approaches to privacy and contextual integrity. Many of the papers made use of techniques coming from machine learning, access control, formal methods, and user surveys, however, while inferences from information flows were a concern in some papers [102] [103] [35], we were missing works that looked at evaluating or enforcing desired norms using information theoretic models. It is one thing to have a policy that expresses a norm that limits the flow of information about race, gender, class, religion and other sensitive attributes; it is another to guarantee that this information cannot be inferred otherwise. One could also imagine novel protections like differential privacy being used to develop elaborate transmissions principles. The numerous papers we surveyed demonstrate that computer scientists have actively applied and contributed to the evolution of contextual integrity using novel techniques. We hope these results serve to provide inspiration and guidance to all those researchers who are committed to leveraging or further developing CI in theory and practice.

Chapter 3

Disciplinary Bridge: Regulating Information

Chapter 2 is an exposition of social theory and a technical literature survey. The authorial perspective starts from the standpoint of Contextual Integrity, a theory of social norms, and from it looks at computer science research. The following chapters of the dissertation have a different style. In contrast, each of Chapters 4 and 5 is intended to stand alone as a technical article, sharing Appendices B and C between them. Though they both draw on Contextual Integrity, the authorial perspective has different assumptions.

This chapter serves as a disciplinary bridge between these perspectives. As a warning to the reader, this bridge is not meant to sustain heavy intellectual loads, so to speak. It is a rope bridge that may be traversed carefully by those interested in precarious adventure and panoramic view. Its purpose is to convince the reader that the transition into a technical discussion of information flow as defined in Section 4.4 is not an abandonment of the theory and mechanisms of social norms, but rather is motivated by the understanding of them developed in Chapter 2. In the terms of Chapter 1, it is a bridge between the regulatory modality of social norms and the formal construction of situated information flow (see Figure 1.1)

3.1 Expanding contextual integrity

Chapter 2 concludes in Section 2.5 with both calls to action for computer scientists interested in open problems in privacy by design and pointers to theoretical gaps in contextual integrity. It calls for "Contextual Integrity theorists" to fill these gaps, inventing a category of researcher that had not existed before the article's publication. Our intention as authors was to frame Contextual Integrity's blossoming from a bounded theory into an expanding field of inquiry.

We discovered that contextual integrity should be expanded:

1. (Section 2.5.1.1) ...to account for social and technologal platforms that span multiple social spheres, perhaps by introducing an “operator” context.

2. (Section 2.5.2.1) ...to account for more of the meanings of ‘context’, that range from abstract social spheres to concrete sociotechnical situations.

3. (Section 2.5.3.1) ...for clarity on how social norms form to reflect ends, purposes, and values in society, and the relationship between these norms and the law.

4. (Section 2.5.4.1) ...to address the challenging cases where multiple social contexts collide or clash.

There is a reflexive, meta-theoretical irony to these gaps. Contextual integrity is an account of social norms that depends on the idea of a society divided into social spheres. These spheres are characterized by social roles, information attributes, and contextual purposes that robustly coevolve (see Section 2.2.1). Norms of information flow are defined *within* a context in terms of its internal ontology. The is meant to reflect why and how privacy is socially meaningful: it is necessary for the maintenance of these meaningful social structures.

The gaps in the theory of contextual integrity reflect changes in society’s actual situation: technical and social platforms now literally fill the gaps between previously distinct social spaces. This makes the ways these platforms conform to or violate privacy difficult to conceptualize. If it is true that social norms are indexed to particular social spheres, the fact that technical infrastructure traverses social spheres does not make the infrastructure *against* social norms so much as *beyond* them, and consequently beyond society’s capacity to regulate it through norms alone. Contextualized normative expectations are specifically vulnerable to infrastructure, and that is why we look to other modalities (technical design, the law, and the market) to protect society’s interests.

3.2 Information in and out of context

The remainder of this dissertation grapples with the question of how technical and social platforms can be designed and regulated given that many of our socially comfortable expectations of information flow are ontologically mismatched to them. At the heart of this question is a deep theoretical question: what is information flow, and why is it valuable? What does information *mean*?

It is surprising that such a foundational question has not yet met with a scientific, transdisciplinary answer. The meaning of data is of great pragmatic concern to science and industry. Through data mining techniques and other innovations [82] [92], “data’s meaning has become a moving target” [68] because the inferences data enable depend on its sources and what other data it is combined with. These techniques are driving business and technical innovation and legal regulation. Scholarship has perhaps not yet caught up.

Disciplinary fracturing is part of the problem. A consensus definition of semantic information has eluded philosophers [51]. Linguistic analysis of the modern use of the word "information" has concluded that it is a confused creole of distinct and incompatible meanings [101]. Library and Information Sciences (LIS) has fruitfully analyzed the term and discovered that information can be both a process and a thing [22]. In LIS, Brier [21] provides a comprehensive account of "cybersemiotics" that traces the relationship between hierarchical layers of semiotics ranging from the basic information theoretic sense developed by Shannon [124] to social and linguistic meaning based on the social theory of Luhmann [80]. Brier [21] argues that the latter, being a property of open systems, will always be to some extent indefinite. However true, this theory of social meaning is unsatisfying for those engaged in the practices of privacy engineering, public policy, or business. Surely something more actionable can be discovered about the contours and logic of social meaning. Meanwhile, in the natural sciences scholars in physics [144] and biology [38] have expressed the aspiration to ground an understanding of inference and signification in nature in mathematical information and computational theory.

The approach taken in this dissertation is to take inspiration from contextual integrity and philosophy of information in defining information flow, but to posit an analytically tractable, mathematically specific definition that is grounded in constructs that are well known in computer science, statistics, and social scientific methodology. I will refer to this definition as "situated information flow". Specifically, this account of information flow, which is introduced in Section 4.4, builds on Dretske [43], who argued that an essential characteristic of information that it enables inference due to regularities in the environment that produces it, which Dretske calls "nomic associations".

Contextual integrity is enriched by a Dretskian view of information flow because it helps explain how the practices that maintain the integrity of a social sphere and the meaning of information flows dynamically reinforce each other. If the meaning of information depends on regularities in the system through which it flows, then the fact that actors are exchanging information to fill well-understood roles for well-understood purpose *is* the reason why the information they exchange has the meaning that it does. As a corollary, *when* social practices change, for example due to a new kind of sociotechnical intermediary, the possible inferences from information change, resulting in new and sometimes unexpected meanings. The meaning of information flows and the sociotechnical practices around them are mutually constitutive.

To make this more precise, it is necessary to define both the regularity of these practices and how they vary. As has been known since Shannon [124], information only flows when a signal has many different potential values. It is, to paraphrase Bateson [13] the difference that makes the difference. The mathematics of probability and statistics, which provide formal tools for understanding the relationships between variables whose values are uncertain, are intimately connected to the mathematics of information for precisely this reason.

Pearl [108] provides a robust and widely used formal account of structural flows of probabilistic influence through causal relationships. The position of this dissertation is that

Pearlian causation and Bayesian networks can provide a useful and tractable formalism for understanding the meaning and value of information flows. The advantage of this formalism is that it can model the relationships between both technical components and social practices in an apples-to-apples way. This is illustrated first in Chapter 4 by applying the framework to computer security in embedded systems and then in Chapter 5 by modeling information economics. In order to model how social practices, and in particular strategic practices, change the meaning of information flows, I draw on Koller and Milch [76] who expand Bayesian networks into a game theoretic formalism: Multi-Agent Influence Diagrams (MAIDs). I extend MAIDs into **data games**, a formal method for mechanism design that elucidates the value of data.

The aspiration of this work is to develop a scientific definition of information flow that is useful for understanding the interaction of the different modalities of cyberspace regulation (social norms, the market, the law, and technology). In order to achieve this, it's necessary to develop the definition in a way that is both precise and general. Mathematical formulations accomplish this precision and generality, and mathematical analysis of information flows and MAIDs are provided in Appendices B and C. These mathematics, which are objectively and proveably true, are intended to transcend any particular narrowly defined scholarly discipline. This aim of this work is a contribution to computational social science [15] under conditions of disciplinary collapse [14].

Shortcomings of this model and room for development are discussed in the Conclusion of this dissertation, Chapter 6.

Chapter 4

Origin Privacy: Causality and Data Protection

Abstract Many privacy and data protection policies stipulate restrictions on the flow of information based on that information's original source. We formalize this concept of privacy as Origin Privacy. This formalization shows how information flow security can be represented using causal modeling. Causal modeling of information security leads to general theorems about the limits of privacy by design as well as a shared language for representing specific privacy concepts such as noninterference, differential privacy, and authorized disclosure. These considerations raise questions for future work about whether policies should be design with respect to the feasibility of automating their enforcement.

4.1 Introduction

Many policies (e.g. HIPAA, GLBA, FERPA, and Executive Order 13526 in the United States and the GDPR in the European Union) place restrictions on the collection, flow, and processing of personal information. When engineers build technical systems that collect and use personal data, they are under business and social pressure to translate prescriptive privacy policies, fitted to their case by lawyers, ethicists, and other policy-makers, into engineering requirements [10, 50, 136, 122]. The goal of engineering a privacy policy is to enable automated enforcement of some or all of the policy. This reduces the cost of protecting privacy. To automate enforcement of a privacy policy, that policy must first be translated into a machine-readable language with a precise syntax and semantics.

Prior research has explored the feasibility of translating classes of privacy clauses into formal logic for enforcement. Some attempt to formalize a wide range of policies, such as those expressible within the framework of Contextual Integrity [10, 130]. Others focus on particular kinds of clauses, such as those restricting information based on its purpose

[139] or its use [36]. This article is concerned with clauses in privacy policies that restrict information based on its *origin*. We consider the origin of data to be the processes that created or transmitted that data to the system governed by the privacy policy, that is, the data's provenance.

Section 4.2 will show how existing policies motivate this work by defining restricted classes of information in terms of the processes that originated them. This policy analysis reveals that origin clauses are most often used to identify a broad class or type of information that is then subject to restrictions or exemptions. In addition, information topic (what the information refers to or is about) is also frequently used alongside information origin. We show that the distinction between information origin and information topic is subtle but important for the purposes of determining the conditions under which a formalized policy can be enforced.

In Section 4.3, we derive, from the policies analyzed, a general ontology of systems, processes, and messages. This ontology is intended to bridge between policies as intuitively articulated in natural language and the mathematical formalisms we will use in the rest of the paper. Using this ontology, we propose Origin Privacy as a framework for understanding privacy requirements and the knowledge necessary to enforce them by design.

Section 4.4 shows how this informal specification can be mapped to a well established formal representation of causal models [106]. In addition to presenting the formal theory, we show that causal modeling makes clear distinctions between two elements of information flow that are sometimes conflated: causal flow and nomic association, where "nomic" means law-like, or regular. These two aspects of information flow correspond to information origin and information topic.

In Section 4.5, we combine the ontology with causal modeling to develop the Embedded Causal System (ECS) model. This model represents a computational system embedded in a larger environment. This is motivated by the need to consider technical systems in their environments (possibly interacting with third parties) when assessing their privacy, fairness, and security properties [142]. We show demonstrate conditions under which an ECS model is secure according to the well-established formal security model of noninterference [54]. We also formalize semantic security for an ECS model and reproduce the result that it is, in general, impossible for a system designer to guarantee semantic security on a statistical database given auxiliary knowledge. It is well known that information can be revealing of other sensitive information given auxiliary knowledge. Our theorem reflects the conditions under which auxiliary knowledge is possible. The contibutions from the model are due to explicit causal modeling of the generative process of data input into the system as well as the operations of the system itself.

In Section 4.6, we build on these results to develop security models for cases where information is restricted based on its origin. We find these models analogous to noninterference and semantic security, and demonstrate sufficient conditions under which an ECS has these properties.

Section 4.7 shows a case study of using Origin privacy. We show that in a case of biometric sensing with Internet of Things devices, origin privacy specification can be used to enforce GDPR compliance. Section 4.8 shows how differential privacy is a special case of origin privacy. Section 4.9 demonstrates how a game theoretic layer can be added to the ECS model to show how system security properties relate to the impact on system users. Section 4.10 addresses directions for future work.

Contributions. The contributions of this Chapter include:

- An analysis of privacy policies, specifically with respect to how they determine protected classes of information through information topic and information origin.
- An informal ontology and articulation of Origin Privacy appropriate for use by policy designers.
- Disambiguation of the concept of "information flow" into causal flow and nomic association components through causal modeling.
- The Embedded Causal System (ECS) model, a model of a causal system embedded in its environment suitable for proving properties of information flow security under these conditions.
- Proofs of conditions for noninterference and semantic security in causal and embedded causal systems.
- Formal security models for origin noninterference and origin semantic security, with proofs of sufficient conditions.
- Demonstration of the use of Origin Privacy in a biometric Internet of Things use case.
- Relaxed security models based on mutual information and proofs of their formal relationship to differential privacy.
- A demonstration of the use of ECS models in game theoretic modeling using Multi-Agent Influence Diagrams.

This chapter refers to two technical appendices. Appendix B proves several supplemental theorems in information theory that are used in a proof of the relationship between causal modeling and differential privacy. Appendix C outlines the major findings of Koller and Milch [76] on Multi-Agent Influence Diagrams, and introduces a new concept: tactical independence.

4.2 Policy motivations

To motivate a consideration of Origin Privacy as a flavor of privacy specification, we look to existing policies that include rules that restrict information flow based on its origin. We build on prior work in logical specification of privacy laws as inspiration for this approach [9, 41].

4.2.1 Policy example: HIPAA Psychotherapy Notes

Some laws define a class of information in terms of the process that creates it. A straightforward example from law are psychotherapy notes as defined under HIPAA[1]:

> Psychotherapy notes means notes recorded (in any medium) by a health care provider who is a mental health professional documenting or analyzing the contents of conversation during a private counseling session or a group, joint, or family counseling session and that are separated from the rest of the individual's medical record. Psychotherapy notes excludes medication prescription and monitoring, counseling session start and stop times, the modalities and frequencies of treatment furnished, results of clinical tests, and any summary of the following items: Diagnosis, functional status, the treatment plan, symptoms, prognosis, and progress to date.

In this definition, there is a reference to a process involving "documenting or analyzing [...] a [...] counseling session". Any information with provenance beginning with its creation by a health care provider who is a mental health professional documenting a conversation during a counseling session separately from an individual's medical record, barring some exceptions, are psychotherapy notes.

4.2.2 Policy example: GLBA

Some laws include references to information origin in the definition of what information is protected. We find this in particular in the Privacy Rule of the Gramm–Leach–Bliley Act, which applies to "nonpublic personal information" (NPI)[2]. This class of information is defined as personally identifiable financial information that is

1. provided by a consumer to a financial institution;
2. resulting from any transaction with the consumer or any service performed for the consumer; or
3. otherwise obtained by the financial institution.

Reading progressively through each of these criteria, the concept of *origin* helps us understand the differences between them and how that affect their enforceability. Criterion 1 explicitly refers to the channel of transmission and does not refer to any specific meaning or category of the information transmitted (though examples are provided in the law, these are constrained only by what is normally transmitted in the process of procuring a financial service). It sets clear guidelines as to the process of creation and transmission. Criterion 2 refers to broad classes of ways in which the information is transmitted to the

[1] 45 CFR §164.501.

[2] GLBA, 15 U.S. Code §6809

governed entity. Criterion 3 is written as if to cover all other cases where a financial institution could collect individualized information, regardless of the process of transmission. It is agnostic to the process of transmission. It raises the question of whether information can be personal financial information without having the specific provenance of a transaction or service with a financial institution. For example, information about a person's income may be personal financial information no matter how it is discovered.

4.2.3 Policy example: PCI DSS

Though not a law, we consider the Payment Card Industry Data Security Standard (PCI DSS) to be a security regulation [20]. Established as a proprietary information security standard by the Payment Card Industry Security Standards Council, it applies to organizations that use major branded credit cards such as Visa, Mastercard, and American Express. Though referenced in the laws of some U.S. states, it is mandated mainly by Visa and Mastercard themselves through their dealings with merchants[3].

We note two aspects of the PCI DSS that make it an effective regulatory standard. First, the PCI DSS governs only types of data the enforcing industry is responsible for generating: cardholder data and sensitive authentication data [105]. This data is generally not meaningful outside of the operations that the payment card industry enables, because the potential uses of the data are a function of the system that created the data in the first place. This allows the payment card industry to straightforwardly enforce contractual obligations on those that use the data. This is in contrast with legal regimes that regulate the flow of more generally meaningful information between persons and organizations.

A second feature of PCI DSS is that it is explicit about the application of the standard to networks of technology and business processes, which it calls the *cardholder data environment* (CDE) [105]:

> The PCI DSS security requirements apply to all system components included in or connected to the cardholder data environment. The cardholder data environment (CDE) is comprised of people, processes and technologies that store, process, or transmit cardholder data or sensitive authentication data.

which PCI DSS further defines as [105]:

> The first step of a PCI DSS assessment is to accurately determine the scope of the review. At least annually and prior to the annual assessment, the assessed entity should confirm the accuracy of their PCI DSS scope by identifying all locations and flows of cardholder data, and identify all systems that are connected to or, if compromised, could impact the CDE (for example, authentication servers) to ensure they are included in the PCI DSS scope. All types

[3] Minnesota Session Laws - CHAPTER 108–H.F.No. 1758. Nevada Revised Statutes, Chap. 603A §215. Wash. Rev. Code §19.255.020 (2011). See [104]

> of systems and locations should be considered as part of the scoping process, including backup/recovery sites and fail-over systems.

This is a clear example of how a privacy policy can be specific about the technical system to which it applies.

4.2.4 Other policies

The examples above have been chosen for their representativeness of the concepts developed in this paper. Other information protection policies do not define protected information solely in terms of its origin but rather depend in whole or in part on a definition of information topic. For example, the Family Educational Rights and Privacy Act (FERPA) defines "education records" as

> those records, files, documents, and other materials which–
>
> (i) contain information directly related to a student; and
>
> (ii) are maintained by an educational agency or institution or by a person acting for such agency or institution.

The Children's Online Privacy Protection Rule (COPPA) includes a broad definition of "personal informal" as any individually identifiable information collected on-line, but includes special restrictions on personal information collected from a child, which could be read as a restriction based on origin.

The United States Executive Order 13526 of 2009 gives certain government officials authority to classify documents they determine pose a risk to national security. In some cases, the information may be classified as soon as it is produced. In all cases, the information's classification prevents unauthorized access. Derivative information carries the classification of the original information. Information may be declassified when the conditions for classification no longer hold. Information restrictions on information therefore depend partly on its procedural history, but also on predictions made about the effects of disclosure.

Our analysis of Origin Privacy explores the limits of privacy by design and what policies can be automatically enforced on a system bound by laws of statistical causation. We will demonstrate the ambiguity of the term "information" and how this renders the application of many policies that restrict information based on information topic indeterminate.

4.3 Origin Privacy

We will now provide an informal definition of Origin Privacy. First, we will introduce an ontology appropriate to the design of technical systems and inspired by the policy examples above. Three concepts are introduced in this section. *Systems* are assemblages of people and technologies through which information flows and is transformed. *Processes* are regular events, implemented as a person or technology's behavior, which act on information. Processes pass information to other processes as data. The *origin* of data is the history of processes that led to its creation. Origin privacy includes any privacy restrictions made on information flow conditioned on that information's origin.

4.3.1 Systems

Regulations mark out spaces or domains in which information may flow more freely than others, or where restrictions specifically apply. For example, HIPAA applies to "covered entities", including health care providers[4]. These spaces may be defined by legal jurisdiction, or they may be defined through networks of contractual relations. More often than not, the regulated space is not bounded geographically but rather by relationships between personnel, institutions, and technical systems, all of which are participating in information flows and processing.

The Payment Card Industry Data Security Standard (PCI DSS) is explicit about this in its definition of the covered system in terms of the "people, processes, and technologies that store, process, or transmit" data (Section 4.2.3 provides further details) [105]. We generalize this concept and refer to assemblages of people, processes, and technologies such as these *information processing systems*, or just *systems*.

There will inevitably be people, processes, and technologies that interact with a governed system without being included within it. We refer to these external factors as the *environment* of the system. Systems have inputs and outputs, which are events that pass data from and to the environment. We will find it useful to model the world of a system within its environment as a larger, superordinate system. This causally embedded system framework is formalized in Section 4.5.1.

The use of the term "system" here is abstract and intended to provide a precise analytic frame. How it gets applied to an empirical case can vary. For example, in healthcare, we might consider a single hospital with its procedures as a system, or a network of covered entities collectively as a system. A complete discussion of systems and how they may be nested or interconnected is beyond the scope of this paper.

4.3.2 Processes

For the purpose of these definitions, we will assume that technical systems and their environments consist of many different components implementing information processes. People in their professional roles are, for the purposes of this framework, included as another way of implementing processes.

[4] "Covered entity means: (1) A health plan. (2) A health care clearinghouse. (3) A health care provider who transmits any health information in electronic form in connection with a transaction covered by this subchapter." 45 CFR §160.103

The outputs of a process may be the input to another. These messages between processes are *data*. When a process A sends data to another process B under one or more conditions, we say that there is a channel from A and B, or, equivalently, that (the state of) A directly causes (the state of) B.

The structure of processes and their dependence on each other implies a causal graph [106] with directed edges representing the channels between processes. We will discuss the implications of causal modeling of systems more thoroughly in Section 4.4.

Systems are composed of contiguous processes, meaning that for any two processes in the system there will be an (undirected) path between the two through channels that includes only other processes in the system.

While causal modeling has been used in a security context (e.g., Feng et al. [49]), formal security research more often relies on static analysis of programs (e.g., McLean [85]). We choose a causal models in this chapter because these models can represent both programs and their subprograms, as well as non-technical environmental processes that generate data.[5] An example of such a process is a medical examination conducted via an interpersonal interaction between patient and doctor.

4.3.3 Origin and provenance

The data resulting from a process depends causally on a history of prior processes. Sometimes data that is an input to a system is generated by a process that is not immediately part of the system. Data can flow through a series of relays before it reaches the system on which a privacy policy is being enforced. The entire history of the information as it flows from its creation to the system input is the information's *provenance*. The governed system may or may not have access to assured metadata (such as cryptographic certificates) about the provenance of its data.

We consider the origin of data to be the processes that have caused it, either directly or indirectly. For the purposes of enforcement of a policy, these processes may be either in the governed system or outside of it, in an originating system or the system's environment.

4.3.4 Origin privacy

Given the above ontology, we can now provide a definition of origin privacy. Origin privacy includes any and only those information flow restrictions implemented in a system that are conditioned on the provenance of system inputs.

[5] While it may be the case that programs are as expressive as causal models, this discussion is beyond the scope of this paper.

4.4 Information flow and causal models

We have motivated Origin Privacy as a concept of privacy that is useful when considering how to design information processing systems to be compliant with laws and other rules regarding the flow of personal information. The ontology in Section 4.3 is motivated by policies that specifically mention systems and restrict information flow based on its origin.

In this section we will introduce philosophical and formal concepts with which we will make our definitions and claims about origin privacy precise. Philosophically, privacy depends on appropriate information flow [98], where information is defined as that which allows somebody to learn about something else based on its regular associations with it. We find a formalization of this idea in Bayesian networks, a common formalism in statistics which represents the relationships between random variables with a directed acyclic graph. Bayesian networks have two attractive properties which we will explain. First, it is easy to derive some independence relations between variables from the graph structure of a Bayesian network. Second, this formalism supports an intervention operation that gives it robust causal semantics. All of these conceptual tools will be used in proofs later in this paper. We will close this section by showing how this formalism rigorously clarifies an ambiguity in the term ‘information flow’, which refers to both causal flow and nomic associations between variables. We adopt the term *situated information flow* for this sense of information flow in causal context.

4.4.1 Philosophy: contextual integrity and information flow

Origin Privacy is intended to be broadly consistent with the contextual integrity [98] philosophy of privacy in so far as it defines privacy as *appropriate information flow*. Specifically, contextual integrity maintains that privacy expectations can be characterized by norms about how information about persons flows in particular social contexts, or spheres.

In this article, we restrict our analysis of Origin Privacy to cases where expectations of privacy have been articulated as *policies* in natural language and endowed with a social system of enforcement. We consider laws and contracts as kinds of policies. Policies may or may not express social norms as per contextual integrity; addressing the conditions under which policies reflect social norms is beyond the scope of this article. However, we maintain that some policies are specifically *privacy* policies because they, like privacy norms, prescribe personal information flows.

As a way of bridging from contextual integrity through privacy policies to the specification of privacy-preserving mechanisms, we address *information flows* in general. Despite its wide use, the phrase “information flow” is rarely given a precise definition. However, philosophical and formal work on information flows have provided general insights that can bridge between social, legal, and technical theories of privacy. We will build on these insights to make arguments about origin privacy.

There is a long history of literature on information flow in computer security and privacy research [85, 55, 12, 140, 133]. These take their inspiration from Shannon's classic formulation of information theory [124]. Dretske's philosophical formulation of information flow [44] also draws on Shannon's information theory. In this work, we are explicitly bridging between philosophy and engineering principles, finding common ground between.

According to Dretske's theory, a message carries information about some phenomenon if a suitably equipped observer could learn about the phenomenon from the message. In other words, a message carries information about anything that can be learned from it. For an observer to learn from it, the message must have a *nomic* connection with its subject, where here "nomic" means “law-like” or “regular” [43]. Messages, in this understanding, get their meaning from the processes that generate them, because these processes connect the content of messages reliably to other events. There is a logical connection between the definition of information flow and the structure of the regular dependence between events. The formal theory of the causal dependence between events has been worked out in the literature on causal graphical models [106].

4.4.2 Causal probabilistic graphical models

There is the a well known formalism for representing the causal relationship between uncertain events: the *Bayesian network*, or probabilistic graphical model, framework [106]. As we will see, this form of modeling can be used to represent processes within a system, as well as in its environment. Before showing the relationship between Bayesian networks and Origin Privacy, we will present them and a few of their formal properties, drawing heavily on Koller and Milch [76] for our choice of formal notation and wording.

4.4.2.1 Bayesian networks

A Bayesian network represents the joint probability distribution of a set of random variables with a graph. Consider variables $X_1, ..., X_n$ where each X_i takes on values in some set $dom(X_i)$. We use $\mathcal{X}$ to refer to the set $X_1, ..., X_n$ and $dom(\mathcal{X})$ to refer to their joint domain.

A Bayesian network (BN) represents the distribution using a graph whose nodes represent the random variables and whose edges represent direct influence of one variable on another.

Definition 1 (Bayesian network)**.** A Bayesian network over variables $\mathcal{X} = X_1, ..., X_n$ is a pair (G,Pr). G is a directed acyclic graph with n nodes, each labeled for one of the variables in $\mathcal{X}$. We use $Pa(X)$ to denote the parents of X in the graph. Pr is a mapping of each node X to a conditional probability distribution (CPD), $Pr(X|Pa(X))$.

Example 2. (Figure 4.1) Alice will be on time for work W if she sets her alarm A early enough and traffic T allows. Bad traffic can be caused by construction C or an accident on the road D.

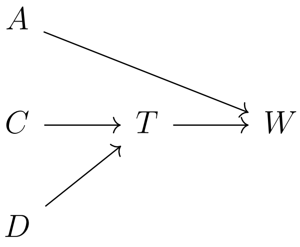


Figure 4.1: Alice's commute to work.

Given the discussion of information flows above, we can see why this is relevant to privacy. The conditional dependence functions between random variables are the *nomic relations* between events and messages. If two variables are conditionally dependent on each other, and this conditional dependence is known to the observer of one of the variables, then the observer can infer something (have knowledge of) the other variables. Hence, by our definitions, the variables carry information about each other. If privacy is appropriate information flow, then the privacy of a system will depend on the causal relationships between its components and the environment.

A directed edge between one variable and another indicates a possible conditional dependence between them. Strictly speaking, it does not necessitate that there is a conditional dependence between them, it only necessitates that there is a conditional probability distribution function defined between them. But it does guarantee that at least one such conditional probability distribution does exist, and under reasonable conditions *most* possible functions (in a measure-theoretic sense) will exhibit the conditional independence [88]. As functions ensuring independence are quite rare in the space of all possible conditional probability functions, this quirk in the notation has not prevented this formalism from being useful in identifying independence in practice.

4.4.2.2 D-separatedness

A useful property of probabilistic graphical models is that some aspects of the joint probability distribution of all variables represented in the graph can be read easily from the graph's structure. Of particular interest in the analysis of the joint probability distribution is when and under what conditions two random variables are independent.

Definition 3 (Path)**.** A *path* between two nodes X_1 and X_2 in a graph to be a sequence of nodes starting with X_1 and ending with X_2 such that successive nodes are connected by an edge (traversing in either direction).

Definition 4 (Head-to-tail, tail-to-tail, head-to-head)**.** For any three nodes (A, B, C) in succession on a path, they may be *head-to-tail* ($A \to B \to C$ or $A \leftarrow B \leftarrow C$), *tail-to-tail* ($A \leftarrow B \to C$), or *head-to-head* ($A \to B \leftarrow C$).

We will find it useful to refer to a special kind of paths, *direct paths*.

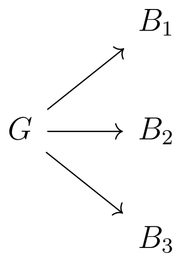
Definition 5 (Direct path)**.** A *direct path* from X_1 to X_2 is a path starting with X_1 and ending with X_2 such that all triples are head-to-tail.

Definition 6 (Ancestors and descendants)**.** If there is a direct path from X_1 to X_2, then X_1 is an ancestor of X_2 and X_2 is a descendant of X_1.

Let $descendants(X)$ be the set of descendants of X.

There are two ways in which a variable A can be conditionally dependent on another variable B without one of them being a descendant of the other. The variables may share an unobserved common cause or they may share an observed common effect.

Example 7. One building in a neighborhood loses power, B_1. One can guess that other buildings B_i around nearby lost power, because power in each building is dependent on the electric grid G. All the buildings may be affected by the common cause of a grid failure.



Example 8. (Figure 4.1) Suppose we observe that Alice is late for work W, as per our earlier example. This could be due to many reasons, including traffic T and missing her alarm A. Traffic may be due to construction C or an accident D. The probability of any particular cause is conditionally dependent on the others, because if any one cause is ruled out, the others are more likely.

The existence of a path between two nodes is necessary for their probabilistic dependence on each other. It is not sufficient, particularly when considering their dependence *conditional on other variables*. For this reason, paths in a Bayesian network can be blocked or unblocked based on a set of variables that is otherwise known or observed, the *conditioning set*.

Definition 9 (Blocked path)**.** A path is considered to be *blocked* if either:

- it includes a node that is in the conditioning set C where the arrows point to it do not meet head-to-head, or

- it includes a node where arrows do meet head to head, but neither this node nor any of its descendants is in the conditioning set

Definition 10 (D-separation)**.** If every path from X_1 to X_2 given conditioning set C is blocked, then X_1 and X_2 are d-separated.

Theorem 11. If X_1 and X_2 are d-separated conditioned on set C, then $X_1 \perp\!\!\!\perp X_2 | C$.

Proof. Proof is discussed in Koller and Milch [76]. □

The converse (that independence implies d-separatedness) is not true in general because specific conditional distribution functions can imply independence. Similarly, it is not generally true that the absence of d-separatedness implies conditional dependence. However, is has been shown that conditional distribution functions implying conditional independence are rare in a measure-theoretic sense [53, 88, 76].

4.4.2.3 Intervention

We have used the terms *Bayesian network* and *causal model* interchangeably. This is because Bayesian networks support a causal interpretation through one additional construct, *intervention* [107]. An intervention on a Bayesian network sets the values of one or more of its values. Unlike an observation of a variable, an intervention effectively creates a new graphical model that cuts off the influence of a set variable on its parents and vice versa. Descendants of the set variable are affected by the intervention according to the probability distribution of the original model.

Definition 12 (Intervention)**.** An *atomic intervention* setting variable X_i to x_i' on a Bayesian network $\mathcal{W}$ creates a new network $\mathcal{W}'$ with post-intervention probability distribution $Pr_{x_i'}$

$$Pr_{x_i'}(X_1, X_2, ..., X_n) = \begin{cases} \frac{Pr(X_1,X_2,...,X_n)}{Pr(X_i=x_i'|Pa(X_i))} & \text{if} X_i = x_i' \\ 0, & \text{otherwise} \end{cases}$$

Theories of causation based on manipulation and intervention have been influential in philosophy [145] and have been shown to be effective theories in psychology of causation [131] including the role of causation in moral reasoning [132], suggesting interventionist causation as a potential bridge between computer science and ethical domains such as privacy and fairness. Cowgill and Tucker [29] discuss the evaluation of algorithmic impact using counterfactuals, which draws on a different but compatible theory of causality [114].

4.4.3 Ambiguity of information flow

We have drawn a connection between information flow in the philosophical sense relevant to Contextual Integrity and Bayesian networks. A Bayes network is a way of representing the nomic dependencies between phenomena. They are “nomic” because they describe probability distributions that generalize over particular instances of a system’s

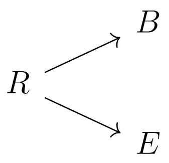


Figure 4.2: Test score ranks (R) distributed to Bob (B) and Eve (E).

functioning. These nomic relations are factored out as an explicit structure of causal relationships.

This reveals an ambiguity in the very concept of *information flow*, illustrated in the following example.

Example 13. Alice, a teacher tells every student privately their test score's rank R (first in class, second in class, etc.) after every test, with class participation used as a tie-breaker. Alice sends a message B to Bob with the information that he has the second highest rank in the class. Alice also sends a message E to Eve that she has the highest rank in the class. From her message and knowledge of the test environment, Eve learns from her message that Bob was told that he was, at best, second in class. Did information about Bob flow to Eve?

A formal representation of this example makes the root of the ambiguity clear. Consider a three node Bayesian network where R is the test results, B is the message sent to Bob, and E is the message sent to Eve (Fig 4.2).

There is causal flow along the edges from R to B and from R to E. But an observer of a single variable aware of the system's laws (nomic connections, graphical structure) can learn nomic associations of a message that inform about variables that are not in the message's causal history. Despite E neither causing nor being caused by B, E reveals information about B.

The phrase "information flow" is ambiguous because the word "information" is ambiguous [101]: it can refer to both a message and the contents of a message. We do not favor either sense. Rather, we propose that to resolve this ambiguity, one has to recognize how the systematic creation and transfer of messages–represented in general by a graph of causal flows–gives each message its meaningful contents. In our formalism, a situated information flow is a causal flow that, by virtue of its place in a larger causal structure, has nomic associations.

Our analysis of privacy policies shows how they variously restrict the flow of information based on its contents as well as its causal history or origin. This is consistent with our analysis of "information flow" as refering to one causal flow within a larger system of causes that give it contents. This scientific formulation of information flows is not yet native to the the language of the law. That the law refers variously to aspects of information flow based on contents and causal history reflects how both are essential to the meaning of the term.

In the following sections we will precisely model a system in its environment in order to disambiguate the different aspects of information flow and understand the conditions under which a system can be free of information leaks. We will measure the strength of nomic associations using a well-understood measure, *mutual information*. Mutual information captures how much about one random variable can be learned from observing another.

Definition 14 (Mutual information)**.** The mutual information of two discrete random variables X and Y is

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

In particular, $I(X,Y) = 0$ if and only if $X \perp\!\!\!\perp Y$.

Mutual information is a technical term with a specific mathematical meaning. It is no etymological accident that it forms part of the analytic definition of “information flow” that we have developed in this section. In the Appendix B, we derive several theorems relating to the mutual information of variables organized in a Bayesian Network. We will use these theorems in proofs about system security in the following sections.

4.5 Bayesian networks and information flow security

In this section, we formalize the ontology from Section 4.3 that we derived from privacy policies. Our formalization uses the causal graphical modeling tools outlined in Section 4.4. We show that several known results in information flow security have dual results within our formal causal modeling of systems in their environments. We demonstrate this for concepts of noninterference [54] and the impossibility of guaranteeing secrecy given the possibility of arbitrary auxiliary knowledge [45, 46].

Where our analysis goes beyond these known results in information flow security are that our models explicitly take into account the relationship between a technical system and its environment. In each case, our theorems prove conditions of the security properties that could not be discovered by static program analysis in isolation. For example, it is well known that information can be revealing of other sensitive information given auxiliary knowledge. Our Theorem 36 reflects the conditions under which auxiliary knowledge is possible.

Because the CPD defined by Pr between each random variable and its parents can be an arbitrary function, including deterministic logical operations, it is possible to encode a system of computational components, including sensors, data processors, and actuators, as a BN. Earlier we defined Origin Privacy in terms of systems, processes, and messages. These concepts map easily into this formalism: systems are Bayesian networks; processes are random variables or events; the inputs and outputs of processes are determined by links connecting them to other processes; messages are the instantiation of particular random variables, which are available as inputs to later variables.

4.5.1 Embedded Causal System (ECS) Model

In the subsequent sections we will use the following standard notation and model, which we will refer to as the Embedded Causal System (ECS) Model.

Definition 15 (World)**.** A *world* $\mathcal{W}$ is a set of random variables with a conditional probability distribution that may be modeled as a Bayesian network $(G_{\mathcal{W}}, Pr_{\mathcal{W}})$.

$$\mathcal{W}$$

Definition 16 (System)**.** A subset of the world $\mathcal{Y} \subset \mathcal{W}$ is the *system*.

Definition 17 (Environment)**.** The *environment* of $\mathcal{Y}$ is the set of nodes in the world that are not in the system, $\mathcal{E} = \mathcal{W} - \mathcal{Y}$

$$\begin{tikzcd}
\mathcal{Y} \arrow[d, bend left] \\
\mathcal{E} \arrow[u, bend left]
\end{tikzcd}$$

(In this and a few other diagrams, we will include cycles because these are diagrams of blockmodels over other networks. In a blockmodel of G, a partition $\{P_1, P_2, \ldots\}$ of the set of original nodes $\mathcal{X}$ is treated as a new set of nodes, with an edge between P_1 and P_2 iff there exists X_1 and X_2 such that X_1 is in P_1, X_2 is in P_2, and (X_1, X_2) is in G_{edges})

It is common in security research to consider *systems* as units of analysis; these systems contain *programs*; system input is data and system output is the result of the programs operating on the data [85]. These programs are represented using a formal approximation of a programming language in order to prove security properties of systems. Systems in our formalism also have inputs and outputs, which are defined by their position relative to the environment.

Definition 18 (Sensors and inputs)**.** A *sensor* is an edge $(A, B) \in G_{\mathcal{W}}$ such that $A \in \mathcal{E}$ and $B \in \mathcal{Y}$, An *input* is the head node of a sensor, B. Denote the set of all inputs with $\mathcal{S}$.

Definition 19 (Actuators and outputs)**.** An *actuator* is an edge $(A, B) \in G_{\mathcal{W}}$ such that $A \in \mathcal{Y}$ and $B \in \mathcal{E}$. An *output* is the tail node of an actuator, A. Denote the set of all outputs with $\mathcal{A}$.

$$\begin{tikzcd}
\mathcal{S} \arrow[r] & \mathcal{Y} \setminus (\mathcal{S} \cup \mathcal{A}) \arrow[r] & \mathcal{A} \arrow[dl, bend left] \\
 & \mathcal{E} \arrow[ul, bend left] &
\end{tikzcd}$$

See the definition of *orderly* (Definition 25) for a set of further constraints on inputs and outputs that are necessary for proving security properties of ECS models.

For some security related applications of this model, it is necessary to distinguish between “high-side” and “low-side” inputs and outputs. High-side variables denote sensitive variables that should not be leaked.

Definition 20 (High and low sides)**.** Inputs $\mathcal{S}$ are partitioned into high $\mathcal{S}_H$ and low $\mathcal{S}_L$ side variables. Similarly, outputs $\mathcal{A}$ are each partitioned high-side $\mathcal{A}_H$ and low-side $\mathcal{A}_L$ variables.

$$\begin{tikzcd}
\mathcal{S}_{\mathcal{H}} \arrow[r] & \mathcal{Y}\setminus(\mathcal{S}\cup\mathcal{A}) \arrow[r] \arrow[dr] & \mathcal{A}_H \arrow[ddl, bend left=60] \\
\mathcal{S}_{\mathcal{L}} \arrow[ur] & & \mathcal{A}_L \arrow[dl, bend left] \\
& \mathcal{E} \arrow[ul, bend left] \arrow[uul, bend left=60] &
\end{tikzcd}$$

Note that in the above diagram and throughout this paper, we will sometimes refer to a set of random variables such as the set of all high-side inputs $\mathcal{S}_H$ as if it is a single random variable. This is well-motivated, because for any set of random variables $\mathcal{X} = \{X_0, X_1, X_2, ...\}$ one can define a new random variable whose domain $Dom(\mathcal{X})$ is the cross product $DomX_0 \times DomX_1 \times ...$ and whose probability distribution is the joint probability distribution $Pr(X_0, X_1, ...)$.

Example 21. A hospital uses a system to manage its medical records. It takes input from many health care professionals through many different forms of treatment. Most medical records are considered a low-side input because they can be accessed by other professionals treating the patient. Psychotherapy notes are a high-side input because they have special restrictions on their use.

Example 22. An intelligence agency has many classified sensors, such as satellites and drone imagery, which contain information that is critical for national security. These are high-side inputs. They also use many data sets that are available publicly and commercially. These are low-side inputs.

4.5.2 System design

In many cases what we are interested in is the possibility of a *system designer* inventing a system subject to certain constraints.

We are interested in the ways that an ECS enables inferences, and how these inferences depend on what is known or observable about the system and its environment. We define some terms here to denote properties of the conditioning set that we will use in later proofs. Intuitively, conditioning sets can be interpreted as observed states of the world that are available to an adversary trying to learn high-side information.

The condition of being *present* captures the intuition that system designers cannot account for the ways that downstream uses of system outputs may be used to reveal sensitive information.

Definition 23 (Present)**.** A system $\mathcal{Y}$ with conditioning set $\mathcal{C}$ is *present* iff

- No descendants of $\mathcal{A}$ are in the conditioning set $\mathcal{C}$, and
- No descendant of $\mathcal{A}$ is in $\mathcal{S}$.

The term "present" indicates that the attacker is able to condition on variables prior to and during the operation of the system, but not variables in the "future" of the system. Requiring that the system outputs are not ancestors of the the system inputs guarantees that the system is in fact positioned in a particular place in time, so to speak.

The condition of being *covered* captures the intuition that in general we do not expect attackers to have the ability to observe systems *as they are functioning*, even if we allow them to know exactly how a system works because they know the causal relationships between the system components.

Definition 24 (Covered)**.** A system is *covered* if no $Y \in \mathcal{Y}$ is in the conditioning set $\mathcal{C}$.

We also specify a condition on the relationship between sensors, actuators, the system, and the environment. In some cases we will not allow an input to be caused by a system variable. We will also not allow an output to cause a system variable. When both these conditions hold, we call a system *orderly*.

Definition 25 (Orderly)**.** A system is *orderly* iff:

- $\forall X \in \mathcal{W}, S \in \mathcal{S}, X \in Pa(S) \implies X \in \mathcal{E}$
- $\forall X \in \mathcal{W}, A \in \mathcal{A}, A \in Pa(X) \implies X \in \mathcal{E}$

Given only the subgraph represented by $\mathcal{Y}$, under some condition it will be the case that the high-side inputs and the low-side outputs are conditionally independent. We will name this property *safety*.

Definition 26 (Safe)**.** A system $\mathcal{Y}$ is *safe* given conditioning set $\mathcal{C}$ iff when considering it as a subgraph, there are no unblocked paths between $\mathcal{A}_L$ and $\mathcal{S}_H$.

If there are no unblocked paths between $\mathcal{A}_L$ and S_H in the system subgraph, then these variables are d-separated and so the system is safe. We assume for our purposes that a system designer can guarantee its safety through sound engineering alone.

For example, consider the system defined by the graph in Figure 4.3.

$$\begin{tikzcd}
\mathcal{S}_H \arrow[r] & \mathcal{A}_H \\
\mathcal{S}_L \arrow[r] \arrow[ur] & \mathcal{A}_L
\end{tikzcd}$$

Figure 4.3: A system that is safe so long as its high-side actuators are not observed.

When $\mathcal{A}_H$ is not in the conditioning set, the path between $\mathcal{S}_H$ and $\mathcal{A}_L$ is blocked, so the high-side input and low-side outputs are independent.

In general, system designers can guarantee safety by removing any direct paths from $\mathcal{S}_H$ to $\mathcal{A}_L$ and then ensuring that the system is covered.

Theorem 27. If a system subgraph $\mathcal{Y}$ has no direct paths from $\mathcal{S}_H$ to $\mathcal{A}_L$ and is covered and orderly, then it is safe.

Proof. Assume there are no direct paths from $\mathcal{S}_H$ to $\mathcal{A}_L$ in the system subgraph. So any unblocked path between $\mathcal{S}_H$ and $\mathcal{A}_L$ must be indirect.

Suppose there is a path p that begins with an incoming edge to $\mathcal{S}_H$, as in $\mathcal{S}_H \leftarrow \cdots \mathcal{A}_L$. Because the system is orderly, incoming edges to into input $\mathcal{S}_H$ must come from nodes that are not in the system $\mathcal{Y}$, as in $\mathcal{S}_H \leftarrow \mathcal{E} \cdots \mathcal{A}_L$. Because these nodes are not in the system subgraph, they cannot be in p.

Therefore, the path p must begin with an outgoing edge to $\mathcal{S}_H$. By a parallel argument, the path must end with an incoming edge to $\mathcal{A}_L$, as in $\mathcal{S}_H \rightarrow \cdots \rightarrow \mathcal{A}_L$.

Because the path p is indirect and it begins with an outgoing edge and ends with an incoming edge, there must be some node X such that X is on the path and X is a common effect node, as in $\mathcal{S}_H \rightarrow \cdots \rightarrow X \leftarrow \cdots \rightarrow \mathcal{A}_L$.

Because the system is covered, X must be unobserved. This implies that the path p is blocked. Therefore, there are no unblocked paths between $\mathcal{S}_H$ and $\mathcal{A}_L$. Thus, by Theorem 11, these variables are independent and the system is safe. □

Information flow security literature often considers systems or programs in isolation from their environment. In practice, systems are always connected with an environment, which is why we have developed ECS. So Theorem 27 is not enough to show the conditions of security in an ECS model because its inputs and outputs are not connected with variables in an environment. For this, we turn to a well established formal security model, noninterference.

4.5.3 Noninterference

Noninterference, introduced by [54], is a security policy model widely used in computer science. Sabelfeld and Myers [116] define noninterference informally as “a variation of confidential (high) input does not cause a variation of public (low) output.”

More formally, model a program C as taking an input state $s = (s_h, s_l)$, as producing an output in a set $S \cup \{\bot\}$, where $\bot$ stands for nontermination and $\bot \notin S$. We use $[\![C]\!] : S \rightarrow S \cup \{\bot\}$ to denote such a semantics. An equivalence relation $=_L$ holds when two inputs they agree on the low values ($s =_L s'$ iff $s_l = s'_l$). The attacker's power is characterized by a relation $\approx_L$ such that if two behaviors are related by $\approx_L$ they are indistinguishable to an attacker.

Following Tschantz et al. [140] and Datta et al. [36], we expand the definition of the operator $[\![\cdot]\!]$ to be a function to and from probability distributions over states, which affords a probabilistic definition of noninterference.

Definition 28 ((Probabilistic) noninterference)**.** For a given semantic model, C is exhibits *noninterference* or *is secure* iff for all s_1 and s_2 in S, $s_1 =_L s_2$ implies $[\![C]\!](s_1) \approx_L [\![C]\!](s_2)$.

This definition admits a wide range of possible semantics for the attacker's equivalence relation $\approx_L$.

We will choose a particular semantics relevant to the ECS model. We impose a probability distribution over inputs $\mathcal{S}$. With it we can construct the variable $\mathcal{A} = [\![C]\!](\mathcal{S})$. We use $\mathcal{Y}$ to denote the minimal Bayesian network relating $\mathcal{S}$ to $\mathcal{A}$ and treat it as the system model. As per the ECS model, we partition the inputs and outputs into high and low sides, $(\mathcal{S}_H, \mathcal{S}_L)$ and $(\mathcal{A}_H, \mathcal{A}_L)$, respectively. Define attacker indistinguishability $\approx_L$ as probabilistic indistinguishability of the low-side outputs when conditioned on inputs:

Definition 29 (ECS attacker indistinguishability)**.**

$$\mathcal{A} \approx_L \mathcal{A}' \text{ iff } Pr(\mathcal{A}_L) = Pr(\mathcal{A}'_L)$$

Conceptually, we are modeling the execution of the program C as the realization of the random variables in $\mathcal{Y}$. C implies a probability distribution $Pr(\mathcal{A}|\mathcal{S})$. It also implies a probability distribution of $\mathcal{A}$ conditional on $\mathcal{S} = s$ realized. $[\![C]\!](s) = Pr(\mathcal{A}|\mathcal{S} = s)$, where s is an instantiation of $\mathcal{S}$. For $s \in \mathcal{S}, a \in \mathcal{A}$, let $(s_h, s_l, a_h, a_l) =_L (s'_h, s'_l, a'_h, a'_l)$ iff $s_l = s'_l$.

Definition 30 (ECS Noninterference)**.** For a given ECS model, $\mathcal{Y}$ exhibits *noninterference* or *is secure* iff

$$\forall s_1, s_2 \in \mathcal{S}, s_1 =_L s_2 \implies P(\mathcal{A}|\mathcal{S} = s_1) \approx_L P(\mathcal{A}|\mathcal{S} = s_2)$$

Corollary 31. $\mathcal{Y}$ is secure by noninterference iff

$$\mathcal{A}_L \perp\!\!\!\perp \mathcal{S}_H \mid \mathcal{S}_L$$

Proof. $\mathcal{Y}$ is secure by noninterference iff

$$\forall s_1, s_2 \in \mathcal{S}, s_1 =_L s_2 \implies P(\mathcal{A}|\mathcal{S} = s_1) \approx_L P(\mathcal{A}|\mathcal{S} = s_2)$$

iff

$$\begin{aligned}&\forall s_H, s'_H, s_L, s'_L \in \mathcal{S}, s_L = s'_L\\ &\implies P(\mathcal{A}|\mathcal{S}_H = s_H, \mathcal{S}_L = s_L) \approx_L P(\mathcal{A}|\mathcal{S}_H = s'_H, \mathcal{S}_L = s'_L)\end{aligned} \tag{4.1}$$

iff

$$\begin{aligned}&\forall s_H, s'_H, s_L \in \mathcal{S}\\ &P(\mathcal{A}_L|\mathcal{S}_H = s_H, S_L = s_L) = P(\mathcal{A}_L|\mathcal{S}_H = s'_H, S_L = s_L)\end{aligned} \tag{4.2}$$

iff

$$A_L \perp\!\!\!\perp S_H|S_L$$

□

Under what conditions is a system secure by noninterference?

We can prove that if the system designer can guarantee that the system is present and safe, then it is secure by noninterference.

Lemma 32. If an ECS model is present, then there can be no unblocked path between $\mathcal{S}_H$ and $\mathcal{A}_L$ that includes an outgoing edge from $\mathcal{A}_L$.

Proof. Proof by contradiction.

Suppose an unblocked path exists between $\mathcal{S}_H$ and $\mathcal{A}_L$ such that the edge connecting to $\mathcal{A}_L$ was outgoing. That is, suppose the path was of the form:

$$S_H \cdots \leftarrow \mathcal{A}_L$$

for some S_H in $\mathcal{S}_H$.

Consider the sequence of nodes on the path $S_H, X_1, X_2, \ldots, X_n, \mathcal{A}_L$ and the direction of the arrows between them, with $X_n \rightarrow \mathcal{A}_L$, and labeling S_H with X_0.

Because the system is present, no descendant of $\mathcal{A}_L$, is in $\mathcal{S}$. Therefor the must be at least one edge on the path such that $X_{i-1} \rightarrow X_i$.

Count down from n to 1 and identify the first X_i such that $X_{i-1} \rightarrow X_i$,

X_i will be a descendant of $\mathcal{A}_L$ because there is a direct path between $\mathcal{A}_L$ and it. X_i cannot be $X_0 = S_H$, which is in $\mathcal{S}$.

Therefore node X_i will be the common cause of a head-to-head connection. That is,

$$S_H \cdots X_{i-1} \rightarrow X_i \leftarrow X_{i+1} \leftarrow \cdots \leftarrow \mathcal{A}_L$$

Because the system is present, this node X_i must not be in the conditioning set. The path must therefore have a head-to-head connecting node that is not in the conditioning set. So it is a blocked path, resulting in a contradiction. □

Theorem 33. Given an ECS model, if the system is present, safe, and orderly then the system is secure by noninterference.

Proof. Consider possible paths between $\mathcal{A}_L$ and $\mathcal{S}_H$ while $\mathcal{S}_L$ is in the conditioning set.

By Lemma 32, a present system can have no unblocked paths from $\mathcal{S}_H$ to $\mathcal{A}_L$ that end with an outgoing edge from $\mathcal{A}_L$. So any unblocked path from $\mathcal{S}_H$ to $\mathcal{A}_L$ must end with an incoming edge into $\mathcal{A}_L$.

Because the system is orderly, if $X \rightarrow A$ for any A in $\mathcal{A}$, then $X \in \mathcal{Y}$. Therefore, unblocked paths must include nodes in the $\mathcal{Y}$ subgraph.

Because the system is safe, there are no unblocked paths between $\mathcal{S}_H$ and $\mathcal{A}_L$ consisting of only nodes in the system $\mathcal{Y}$ subgraph.

So any unblocked path between $\mathcal{S}_H$ and $\mathcal{A}_L$ must include both nodes that are in $\mathcal{Y}$ and nodes the are in $\mathcal{E}$. We have already ruled out paths that include descendants of $\mathcal{A}$. So such a path must include ancestors of $\mathcal{S}$. The path must begin with $\mathcal{S}_H$, go into the environment, then re-enter the system via $\mathcal{S}_L$, then go to $\mathcal{A}_L$. Because the system is orderly, incoming edges into $\mathcal{S}_L$ must be $E \in \mathcal{E}$ and outgoing edges must be $Y in \mathcal{Y}$. That is,

$$S_H \leftarrow \cdots E \rightarrow S_L \rightarrow Y \cdots \leftarrow A_L$$

for some S_H in $\mathcal{S}_H$, S_L in $\mathcal{S}_L$, and A_L in $\mathcal{A}_L$.

S_L is in the conditioning set and part of a head-to-tail structure $E \rightarrow S_L \rightarrow Y$ on the path. Therefore the path is blocked.

So with $\mathcal{S}_L$ in the conditioning set, all paths between $\mathcal{A}_L$ and $\mathcal{S}_{\mathcal{H}}$ must be blocked.

So $\mathcal{A}_L \perp\!\!\!\perp \mathcal{S}_H | \mathcal{S}_L$. By Corollary 31, the system is secure by noninterference. □

It is therefore possible to implement an ECS that is secure in the sense of noninterference as long as a few conditions of the system (covered, safe, and orderly) are guaranteed. Privacy policies that restrict information flow (e.g. by guaranteeing confidentiality) of data based on how it was inputted into the system can be modeled in this framework. In the next section, we will show that policies that impose restrictions on information flow based on the content of information cannot be as easily restricted; to be effective there must be independence relations in the environment of the system. Thus the viability of privacy policies depends on the distinction noted in Section 4.4.3 between causal flow and association.

4.5.4 Preventing associations

We have proven that under certain conditions an ECS is secure by noninterference (Theorem 33). Noninterference is a widely respected formal security model in among computer security researchers. One reason for this is that it is a criterion that depends only one the internals of the system. Computer scientists can guarantee that a system, such as a program, is secure by noninterference without considering how that program will be used in practice. What if we wanted to hold systems to a higher standard that takes into consideration the processes that generate a system's data? For this we need a stronger security property.

We can be more specific and introduce a security policy model that is strictly stronger than noninterference.

$$\begin{tikzcd}
E \arrow[r] \arrow[dr] & S_H \arrow[r] & A_H \\
 & S_L \arrow[ur] \arrow[r] & A_L
\end{tikzcd}$$

Figure 4.4: A system that is not propitious when E is unobserved.

Definition 34 (ECS semantic security)**.** For a given ECS model, $\mathcal{Y}$ is exhibits *semantic security* iff

$$\mathcal{S}_H \perp\!\!\!\perp \mathcal{A}_L$$

Semantic security is a well known property of cryptographic systems that means, intuitively, that an attacker intending to determine the contents of a message might as well not look at an encrypted signal. The term has taken on a wider use in the differential privacy literature as it has been introduced as a desideratum for statistical databases along the lines of that proposed by Dalenius in 1977 [32, 45]. Dwork and Naor [45, 46] show that it is impossible to guarantee the semantic security of such a database given arbitrary auxiliary knowledge.

We draw on the spirit of this literature in our definition of ECS semantic security. The principle difference between ECS semantic security and noninterference is that the latter is concerned with the Independence of system outputs from sensitive inputs *conditioned* on the inputs, whereas the former takes into consideration how environmental correlations may allow system outputs to reveal system inputs.

Noninterference does not imply semantic security. Put another way, the same system can be secure by noninterference but semantically insecure. Consider the system in Figure 4.4.

This system is safe when A_H is not in the conditioning set. It is secure by noninterference because when S_L is in the conditioning set, the path from S_H to A_L that goes through E is blocked. But when S_L is not in the conditioning set, this path is open and therefore A_L can be conditionally dependent on S_H.

Our conjecture is that semantic security cannot be guaranteed by the system designer alone. We are able to prove sufficient conditions for semantic security by including a general property of the world, including the environment outside the system.

Definition 35 (Propitious)**.** The world $\mathcal{W}$ is *propitious* iff there is no unblocked path between S_H and S_L.

Theorem 36. If a system is present, safe, and orderly, and the world is propitious, then the system is semantically secure.

Proof. By Theorem 33, the system is secure by noninterference, implying that

$$S_H \perp\!\!\!\perp A_L | S_L$$

So no unblocked paths run from S_H to A_L when S_L is in the conditioning set.

Recall from the proof of Theorem 33 that this was because we ruled out all possible paths from S_H to A_L.

Paths running from S_H through S_L to A_L were blocked because S_L was in the conditioning set.

Consider any such path, now unblocked as S_L is not in the conditioning set.

If it is unblocked, then there is an unblocked path between S_H and S_L, which contradicts the assumption that the world is propitious.

Therefore there are no unblocked paths between S_H and A_L and so $A_L \perp\!\!\!\perp S_H$.

□

4.6 Formalizing origin privacy

We have defined origin privacy as privacy policies that place restrictions on information based on its provenance. This is in contrast to policies that restrict information based on its content. Another way to put this difference is that origin-based privacy policies restrict information based on the structure of causal flows, while information content based policies restrict information based on its nomic associations.

The problem with restricting information flows based on information content is well illustrated by the problem of guaranteeing semantic security in an ECS system. Any sensitive information content can potentially have nomic associations with otherwise inocuous inputs due to causal paths in the environment. Guaranteeing the absence of associations depends on properties of the environment that may be outside the system designer's control. Noninterference, on the other hand, is an achievable property for a system designer. However, it is defined in such a way that it can be guaranteed even when some kinds of harmful information leaks are probable in practice.

In our legal analysis in Section 4.2 we identified some policies that restrict information based on its provenance, or origin, rather than its information content. In Section 4.3, we have identified the origin of information as the chain of processes causing its input into the system. Taking the concept of "high-side" input as those inputs to a system that are treated with special sensitivity, we can model an example world that meets the most basic requirement of an origin based policy roughly like so:

$$\begin{tikzcd}
O \arrow[r] & R_0 \arrow[r] & \cdots R_n \arrow[r] & S_H \arrow[r] & A_H \\
& & & S_L \arrow[ur] \arrow[r] & A_L
\end{tikzcd}$$

In this model, the original value O is connected only to the high-side input S_H by a direct path of relays $R_0, \ldots, R_n$. We can define the origin property as:

Definition 37 (Origin restricted)**.** A system $\mathcal{Y}$ with inputs $\mathcal{S}$ is *origin-restricted* for a protected variable O iff all direct paths from O to $\mathcal{S}$ end in $\mathcal{S}_H$, and there is at least one such path.

In what sense is an origin restricted system secure? We would like the low-side output of an origin restricted system to be independent of the sensitive variable. As we have seen in Sections 4.5.3 and 4.5.4, there are multiple security models that make different assumptions about the conditions of security. We can use analogous security models for origin privacy.

Definition 38 (Origin noninterference)**.** $\mathcal{Y}$ is secure by origin noninterference with respect to a sensitive variable $O \in \mathcal{E}$ iff

$$A_L \perp\!\!\!\perp O | S_L$$

Theorem 39. Given an ECS model, if the system is present, safe, orderly, and origin restricted then the system is secure by origin noninterference.

Proof. Because the system is origin restricted, there is at least one direct path from O to $S_H \in \mathcal{S}_H$.

If there were an unblocked path from A_L to O that included an outgoing edge from A_L, this would extend into an unblocked path to S_H, violating the condition imposed by Lemma 32. Therefore there is no unblocked path from A_L to O that includes an outgoing edge from A_L.

Because the system is orderly, incoming edges to A_L must go to nodes in the system. Therefore, any unblocked path from O to A_L must go through $\mathcal{S}$

Because the system is safe, there is no unblocked path from $\mathcal{S}_H$ to A_L.

Because the system is orderly, any path from $E \in \mathcal{E}$ to $Y \in \mathcal{Y}$ going through $\mathcal{S}_L$ will include a head-to-tail triplet centered on S_L. Conditioning on this node S_L blocks the path.

Therefore there is no unblocked path between O and A_L, and the system is secure by origin noninterference. □

Definition 40 (Origin semantic security)**.** $\mathcal{Y}$ is secure by origin noninterference with respect to a sensitive variable $O \in \mathcal{E}$ iff

$$A_L \perp\!\!\!\perp O$$

Theorem 41. Given an ECS model, if the system is present, safe, orderly, and origin-restricted and the world is propitious, then the system is secure by origin semantic security.

Proof. By Theorem 39, the system is secure by origin noninterference.

The system is origin restricted, implying that there is at least one direct path from O to S_H. S_L cannoth be on this path because the system is orderly. As no node on this path is in the conditioning set, it is not blocked.

Mirroring the proof to 36, we consider any path ϕ between O and A_L that was blocked by conditioning on S_L. Such path must have a node S_L either within a head-to-tail triplet or as a common cause.

Suppose ϕ includes S_L in a head-to-tail triplet. Then there is a subpath of ϕ there is an unblocked path between S_L and O. But there is also an unblocked path from O to S_H, implying that there is an unblocked path from S_L to S_H. This contradicts the condition that the world is propitious.

Suppose ϕ includes a S_L as a common cause node. Because the system is orderly, both outgoing edges must go to nodes in $\mathcal{Y}$. The path ϕ must therefore enter the system through a node in S_H. That implies a subpath of ϕ within the system runs unblocked from S_H to S_L. That contradicts the condition that the world is propitious.

Because no unblocked path between O and A_L is possible, $O \perp\!\!\!\perp A_L$ and the system has origin semantic security.

□

This demonstrates that origin restrictions do prevent associations between low-side outputs and the sensitive environmental variable under the condition that the systems are otherwise secure.

4.7 Use case: IoT and biometric data

In this section we introduce a use case of Origin Privacy that we have identified through legal analysis and conversations with stakeholders.

Example 42 (Smart building biometric sensing)**.** In an "Internet of Things" instrumented building, many sensors collect information about the contents of rooms, including photograph and other imagery such as infrared scanning to identify the number and size of people present. This information is useful to control the environment in the room (heating, ventilliation). However, this data can also be potentially identified using auxiliary information, such as a facial recognition database. This processed data reveals the identities of persons in the room. In some cases this may be intentional, as when it is used for building security. In other cases these revelations may be unexpected and constitute an invasion of privacy.

We chose this example because it highlights the way smart building technology interacts with privacy policies around photography and biometric data.

4.7.1 GDPR biometric data

Here we focus particularly on the EU's General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679). In general, the GDPR places a number of restrictions on the processing of *personal data*, which it defines thus:

> 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier

> such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person; (Article 4 §1, GDPR)

We interpret this definition as referring to the topic or content of information; personal data is any information *relating to* a natural person. As we have argued, a system designer cannot guarantee that a system does not process information relating to a natural person since these relations may be caused by nomic associations that are external to the system itself.

Noting this difficulty with ensuring compliance, we can nevertheless continue to work with the more specific requirements relating to biometric data. In particular, the GDPR makes a distinction between photographs and biometric data:

> The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person. (Recital 51, GDPR)

By definition under the GDPR, biometric data is a form of personal data that results from particular kinds of processes:

> ’biometric data’ means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data; (Article 4 §14, GDPR)

Unlike the definition of personal data, the definition of biometric data is an origin requirement because it refers to the causal flow of data from a class of processes. Using these legal requirements, we can now use Origin Privacy to formalize their semantics with respect to a system.

4.7.2 Formalizing GDPR requirements

Consider the smart building example as an ECS. Let S_P be the photographic input to the system. Let S_D be a database of identified photographs, originating from an external process E_F. Let Y_B be a component of the system $\mathcal{Y}$ caused by S_P and S_D; it includes imagery from S_P that has been identified using the database S_D.

$$\begin{tikzcd}
E_F \arrow[r] & S_D \arrow[r] & Y_B \arrow[r] & A_H \\
 & S_P \arrow[ur] \arrow[rr] & & A_L
\end{tikzcd}$$

We note that the photographic input S_P may indeed "relate to" natural persons in a systematic way if, for example, certain persons frequent the smart building on a regular schedule. Since these regularities (nomic associations) are unknown to the system designer, there is little she can do to guarantee the treatment of this information as personal data. What the system designer does have access to is the identified faces database, S_D. The *process of identification* that results in the biometric building surveillance data Y_B requires data from an identified source such as S_D.

The system designer knows about the origin of S_D. Specifically, she knows that this data is sourced from E_F, a process that personally identifies the facial images within it. Knowing the environmental source is sensitive, they can impose the conditions for noninterference between E_F and A_L: that no unblocked path exist within Y between inputs originating in E_F and A_L. This implies that both S_D and Y_B be d-separated from A_L within $\mathcal{Y}$. Note that in the diagram above, E_F is indeed d-separated from A_L when the system is *covered*, i.e. when none of its components are in the conditioning set. Intuitively, Y_B is subject to restricted flow because it originates from the sensitive process E_F; it inherits this origin from one of its parent components, S_D.

We build on this result to model more complicated aspects of GDPR compliance. For example, processing of personal information, including biometric information, is generally legal given the consent of the identified person. We can introduce identified sources into the ECS model by denoting the set of natural persons $\mathcal{I}$, and denoting a process that generates data from an identified person X_i for $i \in I$. We can then place conditions on any data that is a result of causal flow from this source, as in this example specification:

Example 43 (Disclosure specification)**.** In the system, all outputs $A \in \mathcal{A}_L$ such that A is a descendant of X_i must also be a descendant of Z_i, where $i \in I$ is the identifier of a natural person, X_i is personally identifiable information, and Z_i is a disclosure agreement identifiable with that person.

To the extent the GDPR controls on biometric information are use restriction or topic restrictions as opposed to an origin restriction, they cannot be automatically enforced based on origin alone. However, considering GDPR through the rigor of Origin Privacy clarifies some of its requirements as formal specification shows what knowledge is needed by the system designer for origin based policy enforcement.

4.8 Relationship to Differential Privacy

In this section we will show the connection between origin privacy and differential privacy. Thus far we have defined origin privacy strictly in terms of conditional independence. This has been inspired by the formal security model of *noninterference*.

When assessing computational systems that process personal information, it is possible for privacy requirements to be looser than this strict security requirement. The particular case of privacy preserving statistical analysis of databases of personal information has motivated differential privacy as a formal privacy model [45, 46].

Formally, an algorithm $\mathcal{A}$ is ϵ-differentially private if for all subsets $S \subset image(\mathcal{A})$ and datasets D_1 and D_2 that differ on a single element, the probability of the output of the algorithm run on the datasets being in S differs by at most a multiplicative factor of e^ϵ and an additive factor δ [47].

Definition 44 (Differential Privacy ((ϵ, δ)-DP))**.**

$$Pr[\mathcal{A}(D_1) \in S] \leq e^\epsilon Pr[\mathcal{A}(D_2) \in S] + \delta$$

When $\delta = 0$, then $\mathcal{A}$ is ϵ-differentially private (ϵ-DP).

Consistent with our origin privacy approach, we will investigate how differential privacy can be assessed given a model of system and its environment as a causal Bayes network. We will draw on prior results relating differential privacy and causality and variations on differential privacy expressed in terms of mutual information.

4.8.1 Mutual information differential privacy

Cuff and Yu [31] demonstrate that there is an equivalence between differential privacy and what they define as mutual information differential privacy:

Definition 45 (Mutual information differential privacy (ϵ-MI-DP))**.** A randomized mechanism $P_{Y|X^n}$ satisfies ϵ-mutual information differential privacy if,

$$sup_{i,P_{X^n}} I(X_i, Y|X^{-i}) \leq \epsilon$$

where X^{-i} is the dataset X excluding variable X_i.

They prove that ϵ-MI-DP is equivalent to differential privacy in the sense that ϵ-MI-DP implies (ϵ, δ)-DP) for some δ, though ϵ-MI-DP is weaker than ϵ-DP. McSherry [86] argues that that ϵ-MI-DP falls short of the desiderata of ϵ-DP. Nevertheless, we will proceed with the ϵ-MI-DP because it is suggestive of probabilistic structure may be used to infer a privacy relevant bound.

That the mutual information limit is conditioned on every other member of the database is an indication of a disappointing fact about differential privacy, which is that its beneficial properties with respect to preserving privacy are not robust to cases when entries in the database are correlated with each other.

The value of using MI-DP for our purposes is that the properties of mutual information are well-understood, and we can derive a number of useful theorems about mutal information between variables in a Bayesian network.

4.8.2 Randomizing database inputs

We can now combine the previous results to show how a system designer can develop a Bayesian network model that guarantees differential privacy. A common way of achieving differential privacy is by randomizing inputs to the database prior to aggregation; this is done in practice in Erlingsson et al. [48]. We can model randomization explicitly as in the following example.

Let X_i be a set of variables representing the personal information to be aggregated. Let Y_i be a random variable over the same domain as X_i that is almost but not quite independent of X_i; we'll say that the mutual information between X_i and Y_i is bounded by ϵ_i. Then aggregate all the Y_i variables into a database, Z, that is available for querying.

$$\begin{tikzcd}
X_0 \arrow[r, "\epsilon"] & Y_0 \arrow[ddr] & \\
X_1 \arrow[r, "\epsilon"] & Y_1 \arrow[dr] & \\
X_2 \arrow[r, "\epsilon"] & Y_2 \arrow[r] & Z \\
X_3 \arrow[r, "\epsilon"] & Y_3 \arrow[ur] & \\
X_4 \arrow[r, "\epsilon"] & Y_4 \arrow[uur] &
\end{tikzcd}$$

In the above diagram, we annotate an arrow between variables A and B with the upper bound on the mutual information $I(A, B)$ afforded by the conditional probability distribution. In this case, we have set all the ϵ_i equal to each other, ϵ.

We can now use this graphical structure to prove that this system is 2ϵ-MI-DP). We can prove this using the Data Processing Inequality,

Proposition 46 (Data Processing Inequality)**.** If three variables are in a Markov chain

$$X \rightarrow Y \rightarrow Z$$

where $X \perp\!\!\!\perp Z|Y$, then $I(X, Y) \geq I(X, Z)$

A standard proof of this is in the appendix. We will also use the Path Mutual Information Theorem (PMIT), that we prove as Theorem 70 in Appendix B.

Example 47. For the structure described above, Z is 2ϵ-MI-DP.

Proof. Note that because X_i and X^{-i} are joined only by a path with a common effect node, $I(X_i, X^{-i}) = 0$.

It follows that:

$$\begin{aligned} &I(X_i, Z|X^{-i}) \\ &= I(X_i; Z, X^{-i}) - I(X_i, X^{-i}) \\ &= I(X_i; Z, X^{-i}) \\ &= I(X_i; Z) + I(X_i; X^{-1}|Z) \end{aligned} \tag{4.3}$$

By DPI and the graphical structure, we know that for all i

$$I(X_i, Z) \leq I(X_i, Y_i) = \epsilon$$

By PMIT, we know the mutual information of two variables connected by a path with all of its common effect nodes observed is bounded by the mutual information of steps along the path. In this case, it entails that:

$$I(X_i; X^{-1}|Z) \leq I(X_i, Y_i) = \epsilon$$

By substitution, we know that:

$$I(X_i, Z|X^{-i}) = I(X_i; Z) + I(X_i; X^{-1}|Z) \leq 2\epsilon$$

As this holds for all i, it follows that Z is 2ϵ-MI-DP. □

4.8.3 Generalizing ϵ-security

We can generalize the two formal security models that we introduced in Section 4.5.3 to models that allow for ϵ mutual information between sensitive inputs and low-side outputs.

Definition 48 (ϵ-noninterference)**.** A system is secure by ϵ-noninterference iff

$$I(A_L, S_H|S_L) \leq \epsilon$$

.

Definition 49 (ϵ-semantic security)**.** A system is ϵ-semantically secure iff

$$I(A_L, S_H] \leq \epsilon$$

.

Recall that two variables are perfectly independent if and only if their mutual information is zero, implying that 0-noninterference is equivalent to noninterference, and 0-semantic security is equivalent to ECS semantic security.

We can now show that differential privacy follows from an application of ϵ-noninterference with one important caveat. We have defined noninterference in terms of a system's high-side and low-side inputs. Schematically, we have considered the "high side" to be a part of the system in need of special information flow restrictions, while the "low side" is what's available to less restricted access.

This model is intended to be generalized to cases where there are multiple categories of restricted information. In particular, to prove that ϵ-noninterference implies differential privacy, we must consider *each entry individually* to be noninterferent with respect to the other entries in the data set.

Theorem 50. For an ECS model, if for all D_i, ECS is secure by ϵ-noninterference with respect to D_i as a high-side input and D^{-i} as low-side input, then A is mutual information differentially private with respect to data set D

$$\begin{tikzcd}
D_i \arrow[dr, "\epsilon"] & \\
D^{-i} \arrow[r] & A
\end{tikzcd}$$

Proof. If for all D_i, ECS is secure by ϵ-noninterference with respect to D_i as a high-side input and D^{-i} as low-side input, then

$$\forall i, I(D_i, A_L|D^{-i}) \leq \epsilon$$

which implies that

$$sup_{i,P_{X^n}} I(D_i, A_L|D^{-i}) \leq \epsilon$$

which is the condition for ϵ-MI-DP. □

It is not generally the case that if each of a database's entries D_i is ϵ-semantically secure from the output A that the output will be ϵ-differentially private. A bound on $I(D_i; A)$ does not imply a bound on $I(D_i; A|D^{-i})$, as is apparent from Equation 4.4.

$$I(D_i; A|D^{-i}) = I(D_i, D^{-1}; A) - I(D_i; A) \tag{4.4}$$

4.9 Incentives in ECS

We have motivated ECS modeling by showing how it captures the implicit ontology of privacy policies and enables reasoning about the security properties of systems embedded in an environment. We have used Bayesian networks as a modeling tool because they clarifying the relationship between two aspects of information flow. Bayesian networks also provide a robust, qualitative means of determining dependency relations between their variables, which we have used in our proofs about the relationship between various system and privacy properties. In this section, we will show how the same ECS framework can be extended to include strategic actors and their incentives.

To accomplish this, we will use the Multi-Agent Influence Diagram (MAID) framework developed by Koller and Milch [76]. In brief, a MAID is a Bayesian network with two important extensions.

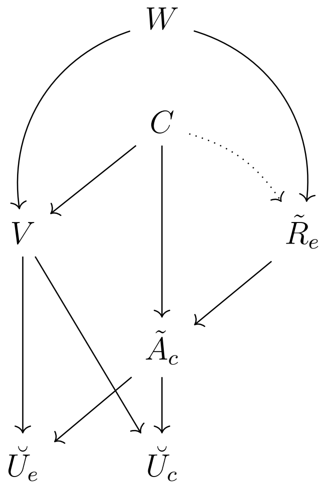
- Some of the variables are reserved as *decision variables* and assigned to one of several agents. An agent's assignment of CPDs to its decision variables is that agent's *strategy*; replacing each decision variable with the CPD from a *strategy profile* transforms the MAID into a Bayesian network.

- Some of the variables are reserved as *utility variables*. These are assigned to one agent each, and are summed when realized into the agent's total utility. Utility variables must not have children.

A formal definition of MAIDS, strategies, and other useful properties is given in Appendix C, which includes an account of the graphical notation we will use in this Section.

4.9.1 Expert services model

As an illustration of an information game that can be represented as a MAID, consider the following diagram, which is a generalized model of an expert service. (This model will be analyzed in more detail in the following chapter, in Section 5.4.3.) The services, which include health services, legal services, as well as some software based services like search engines, involve a client, c, who presents knowledge of their situation to an expert, e. The expert has access to general knowledge relevant to the client's condition, and recommends some action based on this knowledge. The client can choose to take this recommendation. In the idealized case considered here, the client and the expert have perfectly aligned incentives and so the expert will use their knowledge to the best of their ability.



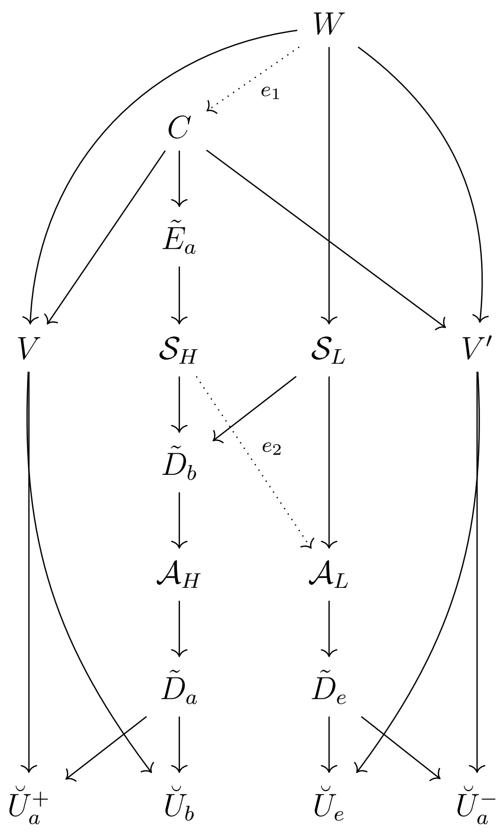
In this model, W is the generalized knowledge of the world that is available to e at their decision variable representing their recommendation, $\tilde{R}_e$. The action taken by the client $\tilde{A}_c$ and the action value function V determine the utility of both the expert and the client, $\breve{U}_e$ and $\breve{U}_c$. The value function is influence both by general facts about the world W and the particular situation of the client, C. The client knows their situation but not the general expert knowledge.

The client's action is informed by the general knowledge only through the recommendation of the expert. The expert may or may not know the client's specific situation; this is represented by the dotted arrow between C and $\tilde{R}_e$.

The value of such a model is that a qualitative analysis can readily provide insights into the influence of an information flow on outcomes. Since we know the expert's utility depends on influencing the client to make the best possible action at $\tilde{A}_c$, and that the value of this action depends on V, the expert's effectiveness will be limited by how well they can predict V given the knowledge availabel to them at $\tilde{R}_e$. Without information about their client's specific situation C, their advice can at best be perfectly general. But with access to information C, the expert can improve their recommendation and outcomes for both players.

4.9.2 Expert ECS Model

We now combine the expert service model with the ECS model. We will embed the expert insider an ECS and give them access to personal information of the client via the high-side input. They will also have access to general knowledge through the low-side input. An adversary will have access to the low-side output, but not the high-side output. Using this model, we will be able to test how the security properties we have analyzed in Section 4.5 and Section 4.6 can be motivated in terms of the way they affect the incentives of interacting with the system.



This game has three players, Alice (a), Bob (b), and Eve (e).

Alice and Bob have perfectly aligned incentives, and Eve is an attacker who is adversarial to Alice. We specify that the following relations hold:

$$\breve{U}_a^+ = \breve{U}_b$$
$$\breve{U}_a^- = -\breve{U}_e$$
$$\breve{U}_a = \breve{U}_b - \breve{U}_e$$

At the center of this model is an ECS, with high- and low- side sensors ($\mathcal{S}_H, \mathcal{S}_L$) and actuators ($\mathcal{A}_H, \mathcal{A}_L$).

In this model, Alice is aware of her personal information C and decides at $\tilde{E}_a$ what if any of it to divulge secretly (via $\mathcal{S}_H$) into the system because she wants an expert recommendation from Bob. Bob has access to general expertise W through a low-side input ($\mathcal{S}_L$). These inputs are both available to Bob at his decision node $\tilde{D}_b$, at which he chooses a recommended action for Alice, which he passes through the high-side output $\mathcal{A}_H$.

Alice will use the information about the recommendation taken from the high-side output $\mathcal{A}_H$ to choose an action. The utility of this action will depend on the action values

V, which are a function of two variables: the personal characteristics C of Alice and other general information about the world, W.

Eve will make a decision $\tilde{D}_e$ based on the low-side output of the system, A_L. Eve's utility U_e depends on this decision and an action value function V' that is analogous to Alice's action value function V, in that it depends on C and W.

The system in this diagram is $\mathcal{Y} = \{\mathcal{S}_H, \mathcal{S}_L, \tilde{D}_b, \mathcal{A}_H, \mathcal{A}_L\}$.

The diagram has two dotted edges, e_1 and e_2 Each dotted edge may be either included in the graph (open) or excluded from the graph (closed). The diagram therefore describes four distinct models: none open, e_1 open, e_2 open, and both open. We will analyze each case in Section 4.9.2.1.

4.9.2.1 Analysis

First, we can analyze the expert ECS model presented in Section 4.9.2 in terms of the system design properties introduced in Section 4.5.2.

Note that no descendent of $\mathcal{A}$ is in $\mathcal{S}$. Therefore, the expert ECS model is present if none of $\tilde{D}_a, \tilde{D}_e, \breve{U}_a^+, \breve{U}_b, \breve{U}_e, \breve{U}_a^-$ are in the conditioning set $\mathcal{C}$.

The system is covered if none of $\mathcal{S}_H, \mathcal{S}_L, \tilde{D}_b, \mathcal{A}_H, \mathcal{A}_L$ are in the conditioning set $\mathcal{C}$.

It is plain from observation that the system is orderly. It is also clear that the system is origin-restricted with respect to the personal characteristics C: there is one direct path from C to $\mathcal{S}$ and it goes to $\mathcal{S}_H$.

We are left with several candidates for the conditioning set: $W, C, V, V', \tilde{E}_a$. Recall that the world is propitious if there are no unblocked paths from $\mathcal{S}_H$ to $\mathcal{S}_L$. There are two ways an unblocked path can happen under the conditions discussed so far. One is that either V or V' is in the conditioning set. Another is that the edge e_1 is open.

It is clear that the system is safe if edge e_2 is closed and not safe if edge e_2 is open.

Suppose that there are no variables in the conditioning set. Then by the reasoning above, the following properties of the expert ECS system hold:

- If e_1 and e_2 are closed, then the system is origin secure with respect to C both semantically and by noninterference.
- If e_1 is open, then the system will be origin secure with respect to C by noninterference, but may not be semantically secure.
- If e_2 is open, then the system may not be origin secure with respect to C either by noninterference nor semantically.

Though we have been able to show that these security properties hold on the expert ECS model, this model also reveals how these security properties do not provide all desireable guarantees a system might provide in terms of the incentives of Alice and Bob.

What can be shown is that given that the expert ECS system is semantically secure, it is also the case that $\tilde{E}_a$ and $\tilde{D}_e$ are tactically independent (see Definition 79), meaning that for any strategy specifying decision rules to each decision variable, in the induced probability distribution $\tilde{E}_a$ and $\tilde{D}_e$ are independent. In other words, at the level of tactics, Alice's choice to reveal her information to the ECS will not depend on Eve's choice of how to use the system's low-side outputs adversarially.

However, despite these security properties, we can show with this model that $\tilde{E}_a$ may *strategically rely* on $\tilde{D}_e$ (see Definition 76). This means that Alice's choice of decision rule at $\tilde{E}_a$ can depend on Eve's choice of decision rule at $\tilde{D}_e$. This can be a problem for system designers if their goal is to guarantee that the presence of Eve has no deterring effect on Alice's choice to reveal her data to the ECS.

Further exploration of the relationship between system security properties and incentives of players in this causal formalism is left to future work.

4.10 Discussion and future work

We have analyzed privacy policies and discovered that they variously restrict information based on origin and topic. We developed an informational ontology that reflects the assumptions underlying many of these policies. We have shown that this informal ontology can be formalized in terms of causal graphical models. These models show that the two aspects of information flow correspond to precisely defined concepts of causal flow and nomic association. Both senses of information flow are accomodated by an understanding of situated information flow as a causal flow in causal context, as represented by a Bayesian network.

We developed a model of system security, the ECS model, which represents a system embedded in its environment. This model can demonstrate and extend known results in computer security, such as those concerning noninterference, semantic security, and differential privacy. It also allows us to formally define a new privacy property, origin privacy, which assumes that system designers have some control over the paths through which information enters their systems. We demonstrate how the ECS model can be used to elucidate a case of implementing GDPR compliance on biometric data. We demonstrated preliminery results on how the ECS model can be extended into game theoretic form to account for how strategically acting agents interact with systems with or without relevant security properties.

These contributions are suggestive of several lines of future work.

4.10.1 Specifiability Criterion

One direction for future work is to question what these results mean for the design and interpretation of legal policies. Are privacy policies based on information topic harder for consumers to understand than policies based on information origin? How do end users interpret ambiguous language in privacy policies?

One benefit of using causal models as opposed to program analysis for considering information flow security of technical systems is that it shows that information flow security is not only a technical problem. Because information leaks based on nomic associations may be properties of any mathematically defined causal system, these results extend to human institutions as well. In general, mathematical results about the limits of computational enforceability will generalize to the possibility of enforceability through non-computational means.[6] Privacy policies and social expectations that restrict information based on its contents may be unenforceable or ambiguous in general.

Concretely, the brittleness of these kinds of restrictions is exposed by advances in large-scale data analysis, or "big data", which routinely confound our expectations about what information is about. Data about a person's purchases of scent-free hand lotion, which we might have assumed to be innocuous, has been shown to be correlated with sensitive information about early-stage pregnancy [63]. Ad targeting systems that use all available correlational information in the data they collect risk violating people's privacy due to unexpected discoveries from automated learning processes.

By demonstrating the limits of what security properties can be enforced, and by whom, this framework can shed light on how privacy policies should be written and which parties should be held liable in the event of a violation. This may build on other work in identifying legal liability in cases of perceived inappropriate behavior of a sociotechnical system [34].

4.10.2 Observer capabilities

We have shown that nomic associations between system outputs and sensitive environmental variables can lead to violations of privacy policies. In order for these threats to be material, nomic associations must be known to attackers as auxiliary information. A natural next step in this line of inquiry is a more systematic study of how observer's capabilities for learning nomic associations factor into information flow security considerations.

In our models in this article, we have used Bayesian networks as models of the objective frequencies of variable outcomes. Possible inferences from observed variables have been interpreted as those inferences possible *in fact* from the causal structure of the world. Information flow security guarantees were possible when the system designer was assumed to have some true knowledge about the system's environment, such as the origin of its inputs.

In practice, most systems will be embedded in environments that are only partially known. In the cases of fraud and spam detection, and other adaptive machine learning systems, a model of the origin of inputs is trained continuously from collected data. Probabilistic graphical models are indeed one of the many machine learning paradigms used in these applications [18].

A direction for future work is developing a theory of information flow security and

[6] This claim assumes the Church-Turing thesis.

privacy under conditions where observer knowledge of nomic associations is itself a function of system inputs. In simple cases this may reduce to single program analyses already established in work on use privacy [36]. In cases where multiple systems interact, there may be novel problems.

4.10.3 Incentives based policy design

Prior work has been done on game theoretic models of information flow policies [10], mechanism design with differential privacy [87], and semantic interpretations of differential privacy framed in terms of the incentives of data subjects [72]. We see potential to continue this line of work using statistical models of systems, their internal processes, and their environment, including the process that generate their input data.

We have shown how Multi-Agent Influence Diagrams (MAIDS) [76] can be used in game theoretic modeling of security problems. Feng et al. [49] have used Bayesian networks to model security risks and perform a security vulnerability propagation analysis. They build their causal model from observed cases and domain experts. We anticipate new frameworks for aligning the incentives of system designers and data subjects which are sensitive to risk of data misuse (see [1]). An application of this work is automated policy design for smart buildings and cities, where interacting data subjects and control systems must share information while minimizing data misuse.

Our work in Section 4.9 is a first step in developing a new way of assessing the value of security properties based on their impact on game outcomes. This method of modeling information value through data games is the subject of the next chapter.

We have unpacked the assumptions of privacy policies to develop a general ontology of systems, processes, and information. We have then formalized this ontology using the assumption that these systems are subject to the laws of probability and an interventionist account of causation [145]. Using these simple assumptions, we have confirmed well-known results about information flow security about programs and systems in isolation. We have also gone beyond these results by showing explicitly what their consequences are when systems are embedded in an environment, developing a general security modeling framework for this purpose.

We prove the relative effectiveness of origin based information flow restrictions over association based information flow restrictions using the causally embedded system framework. We show how origin privacy would be applied in a GDPR and Internet of Things use case. We anticipate new lines of inquiry extending from the intersection of causal modeling and information flow security, including evaluation of policy enforceability, modeling the role of observer knowledge in privacy and security, and automated policy generation through incentives-based design.

Chapter 5

Data Games and the Value of Information

In this chapter, I develop a general, formal model of economic information flow. This builds on prior work identifying the gaps in social theoretical understanding of privacy (Chapter 2) and advancing a formal definition of situated information flow compatible with concepts of security and privacy in computer science (Chapter 4). I argue that this model is well suited to capturing the economic impact of information flows through mechanism design, which can inform both regulation and privacy by design.

Section 5.1 considers the social theory of privacy and notes, based on prior work in Chapters 2 and 3 of this dissertation, that cross-context information flows remain an unresolved theoretical problem in privacy. When societal expectations are organized according the boundaries of social contexts, they cannot easily anticipate flows that violate those contexts. In particular, the kinds of information flows in technical infrastructure and their impact on society are difficult to conceptualize and therefore difficult to regulate socially.

Section 5.2 outlines legal frameworks for data protection. These do offer rationales for preventing cross-context information flow in particular contexts, such as health and legal advice, through confidentiality. These sectoral privacy laws have not prevented cross-context flows that fall through the gaps of the law, such as those facilitated by data brokers. These flows are driven by actors who, unregulated by the law of society or the law of the state, are beholden instead by the law of the market.

Section 5.3 addresses the existing economics of privacy and information. This literature is also organized into analysis of single economic contexts. I argue that this is due to a lack of formal modeling tools for addressing the complex reality of economic information flow. Drawing on the formal model of situated information flow in Chapter 4 based on Dretske, Pearl, and Nissenbaum, and Multi-Agent Influence Models [76], I develop a framework for mechanism design of games involving information flows: **data games**.

Section 5.4 uses the data game framework developed in Section 5.3 to model simple economic contexts involving personal information flow. This section includes models of:

a principal hiring and agent of uncertain quality; price differentiation based on personal information; and the provision of expert advice to a client. These models demonstrate the expressivity of data games.

Section 5.5 builds on the prior sections to model a case of cross-context information flow and its social effects. The model shows how one firm purchasing an information flow from another firm can have a negative impact on consumers who are otherwise not involved in that transaction. This shows that the cross-context information flows can have market externalities, suggesting that the economy of information flow is prone to market failure.

Section 5.6 concludes the chapter with a discussion of broader implications and future work in a broadly conceived data economics.

5.1 The limitiations of contextualized privacy

Privacy is a many-meaninged term difficult [134] to concisely define. Mulligan et al. [89] argue that privacy is an essentially pluralistic and contestable concept that should be defined at the 'retail' rather than 'wholesale' level. We will call a theory of privacy that maintains that the term has many distinct and perhaps irreconciliable meanings a *particularist* account.

The field of contextual integrity [98] accounts for variety of meanings of the term by (1) defining privacy as *appropriate information flow* and (2) noting that what is "appropriate" depends on socially situated expectations or *norms*, that are indexed to social contexts. According to this theory, privacy refers to these social expectations that vary from context to context. This theory of privacy both stands up to empirical tests [83] and has been useful in privacy engineering (e.g. [129], [16]). We will refer to this kind of theory, in which privacy has a single meaning that is parameterized by social context a *contextualized* account.

Both particularist and contextualist accounts have trouble addressing legal, ethical, and technical privacy challenges arising from social platforms, technologies that mediate multiple different contexts [16]. The commonly felt awkwardness of social media due to the unexpected participation of different audiences known as *context collapse* [84] [37] is a symptom of the more general problem that the digital infrastructure mediating so many of our social and commercial interactions is often indifferent to our contextualized social expectations because it is not "in" any one social context. In many cases technology makes our *situation* much more complex and interconnected in ways that go beyond any social expectations of our social *spheres*.

Contextual Integrity is socially meaningful and psychologically compelling. For most people, our ubiquitous and complex technical infrastructure, in actuality, is neither. It is perhaps for precisely this reason that social norms are not enough to regulate privacy in our technical infrastructure.

Beyond society's expecations of privacy, there are also legal limits to the collection and use of personal data.

5.2 Information in the law

This section will briefly survey relevant legal positions on information and data protection. This analysis will show how the law has yet to discover a regulatory framework that deals squarely with the economic effects of personal data flow.

5.2.1 Information as property

There is a sense of the word "information" that corresponds to physical records–papers on file, patterns recorded electrically in databases. This sense of information as a *thing* [22] perhaps encourages privacy solutions that frame personal information as a good that could be protected by property rights and thereby allocated more efficiently [90]. Private property rights create a legal relationship between a person and a thing that transcends social spheres; robbery of private property is illegal in almost all social contexts.

The closest existing legal framework for property rights in information are intellectual property rights laws. However, intellectual property rights such as those for copyright, patents, and trade secrets are motivated by the economic incentivization of innovation, not by privacy. They are not designed to protect ownership of data in general. For example, copyright specifically does not pertain to mere data or the organization of facts.[1] So a data subject does not by default own facts about themselves. Databases may be protected as a compilation if the selection of the data constitute individual, creative expression.

Samuelson [119] argues that intellectual property law is a poor fit for protecting privacy because property rights are alienable whereas privacy interests are not. In other words, when party A sells property to party B, it is generally with no restrictions on whether and how party B sells that property to party C. With personal data, individuals commonly have an interest in sharing their information with one party with the specific expectation that it is not resold or reused for an unknown purpose.

Despite the appeal of metaphors that consider data to be a kind of commodifiable good, like oil [65], data's real properties and the interest people have in their personal data defies these metaphors. Data in general presents a conceptually more difficult case than the kinds of intellectual goods considered in intellectual property law. I will make the case that this is due to data's ontological slipperiness, a slipperiness that demands a new method of economic reasoning.

5.2.2 Confidentiality and sectoral privacy law

United States law has many provisions for the confidentialiality of personal information gathered in specific professional contexts. For example, HIPAA has special provisions for psychotherapy notes that do not apply to personal health information more generally. Attorney-client privilege, which protects personal information disclosed to ones lawyer,

[1] Feist v. Rural, 499 U.S. 340 (1991)

is another example of strongly protected confidentiality [59] [7] [113]. Confidentiality in these domains is meant to ensure that the protected client can freely divulge personal information to the service provider without concern that their information may be used in a secondary way that harms them. This is necessary for the effective execution of these services. It is notable that in all these cases of expert services, data protection is mandated by law, not left for market selection or self-regulation.

These confidentiality cases are perhaps the clearest cut examples of contextual privacy. These are examples of *sectoral* privacy laws, meaning laws that apply only to a single business sector. This indexes them into a particular social context, the one implicated by that sector's activity. In the language of Contextual Integrity, it is clear to which abstract social *sphere* each law applies.[2] Notably, these laws generally do not apply to data collection performed by online services.

Furthermore, confidentiality is a restriction on information flow. Restrictions on information flow, when observed, prevent the collapse of otherwise separate social *situations* in a more complex and perhaps conflicted one. Contextual integrity is specific about how information norms need not be restrictive, and sectoral privacy laws indeed do recognize cases where some kind of information flow is mandatory (for example, when a hospital disclose medical records to comply with law enforcement procedure). Throttling information flow between situations is what keeps sphere or sector based information flow rules enforceable in part because it simplifies the information semantics.

5.2.3 Notice and consent

On-line services that do not fall under the rubric of any sectoral privacy laws are regulated in the United States by the Federal Trade Commission (FTC). The FTC has encouraged a self-regulatory regime of "notice and consent" whereby on-line services must transparently describe how they will use personal data and get consent before collecting it. The company must abide by the terms of the notice, even in the case of a corporate merger [64], or risk being in violation of the FTC Act Section 5 prohibitions against unfair or deceptive practices.

The effectiveness of the notice and consent framework has been widely panned as ineffective. Notices change over time and obscure rather than reveal the ways third parties use data [8]. Notices under-inform, are impractically burdensome on users, are hard to understand, and do not account for how one person's choice to reveal their personal data may be revealing about others who have not consented [112] [135]. The complexity of notices may indeed reflect the complexity with which collected personal data may be used in practice [121]. Nevertheless, the scholarly consensus is that the notice and consent framework does little to protect privacy.

[2] Though business sectors can be assigned to social contexts in CI, it is important to distinguish business sectors from social contexts because only the latter form legitimate societal norms [100].

5.2.4 GDPR, purpose-binding, and data minimization

The European Union's General Data Protection Regulation, which at the time of this writing has not yet gone into effect, promises to set a significant new standard for data protection in on-line services. While it protects only EU citizens, its extraterritorial enforcement means that many companies that are not based in the EU must still take significant steps to be compliant or risk facing heavy fines.

A notable feature of the GDPR is its use of "purpose binding" [62] [61]: data subjects must consent to particular purposes of use by data processors before the data may be collected. Exceptions to this rule are also framed in terms of purposes (such as the purpose to protect the "vital interests" of the data subject). While purpose binding is not a new legal standard (it is a feature of the EU's Data Protection Directive, which the GDPR supersedes, as well as the U.S. Fair Information Practice Principles), its purpose binding clauses are empowered by the addition of obligations of *data minimization*, the requirement that data may not be held or processed in excess of what is needed for the original purposes of collection [58], and the obligation of privacy by design of information processing systems [33].

The efficacy of this regulation is still untested. However, it compares favorably with existing U.S. law. Narrowing the complexity of notices to be about particular purposes may be an improvement over the more complex legal and technical conditions in notices typical under the FTC's notice and comment framework. While some U.S. sectoral privacy policies include purpose restrictions on information use [138], the fact that the GDPR is an omnibus law means that its purpose restrictions apply even to those businesses that fall through the gaps of sectoral regulation. Truly, the GDPR formalizes new privacy *rights*, which are akin to but unlike other rights like property rights, and protects them by placing obligations on data controllers and data processors.

5.3 Economics and mechanism design

Privacy is a complex social phenomenon and the importance of nuanced social theories like contextual integrity cannot be overstated. However, it is also a fact that technical infrastructure that spans social contexts is most often developed by private companies that are more responsive to economic principles than social norms. Having motivated the inquiry by reflecting on philosophical and legal theories of privacy, I will now turn to the economics of privacy, as economics are at the core of the social and legal questions that have concerned other scholars. Though narrower in scope, the field of economics has provided a rich literature on privacy that lends precision to claims about how interests or incentives shape outcomes.

Modern economics of privacy concerns itself mainly with the economics of personal information as it is used by businesses employing information technology. Specially, it most often addresses what Acquisti, Taylor, and Wagman [5] call *tangible* impacts of privacy, those impacts that have objectively measurable and modelable costs and benefits and effects on market structure. While others acknowledge the possible importance of *intangible* impacts, such as a psychological concern about how ones personal information may be used (which may be modeled as a subjective preference for privacy [23] [26]) and other more global social effects, we will limit the discussion in this paper to tangible impact.

Even so narrowly scoped, there are many different economic contexts in which the presence or absence of personal information is critically relevant. There are so many different contexts, each represented in their own sophisticated scholarly literatures, some [5] argue that a comprehensive economics of privacy cannot be achieved. Essentially, this is an argument that the economics of privacy should be contextualized, echoing the contextualized account of privacy outlined in Section 5.1. But what if we want to understand the economic impact of information flowing *between economic contexts*? In order to accomplish this, we need an economic framework that can model many different kinds of economic contexts, as well as a the ways in which they may interact.

More concretely, economics has so far failed to come up with a theory explaining why people *buy and sell data*, and how they price it. Such a question is critical for explaining the way personal information transfers from business to business in the case of, say, online behavioral advertising. I posit that this is in part because of the slippery ontological properties of data: it is not a thing that one can hold as property and its main economic value is informing the actions of other agents (such as pricing decisions). In other words, the value of data is often the value of the strategic advantage provided by the data. This may help explain why often companies are often more interested in buying and selling flows of data, such as those provided by a web-based Application Programming Interface (API), than any particular piece of data.

One tool in the economics toolkit for understanding policy decisions is mechanism design [70] [95]. Mechanism design is an “inverted game theory”, wherein the designer defines a range of possible economic games and chooses the structure of the game that maximizes some predetermined goal or objective function. The objective is a function of the outcome of the game assuming the players are operating according to strategies that are rationally optimized for their self interest, such as the strategies of a Nash Equilibrium. This in turn provides insight about what kind of rules can be imposed on an economic transaction such that socially prefered outcomes result, even when the economic actors are self-interested.

In this section, I develop a framework, data games, for mechanism design of economic situations involving information flows. This framework will extend the Multi-Agent Influence Diagram (MAID) framework [76], which is a game-theoretic extension of Bayesian Networks. This framework, which was briefly introduced in Chapter 4, models information flow in a social context as information is understood by engineers and economists. For regulatory regimes to be most effective, they must be reducible, in the scientific sense, to something like this model.

5.3.1 Formalizing information flow mechanisms

We have motivated the need for a general framework for mechanism design for economics contexts involving (personal) information flow. In this section, I will specify that framework. Summarizing prior work (see Section 4.4), I synthesize a formal representation of situated information flow from Nissenbaum, Dretske [43], Shannon, and Pearl [106]. In this representation, information flow is a causal flow that carries nomic associations due to other causal flows in its context, which can be represented precisely using Bayesian networks. I then propose the use of Multi-Agent Influence Diagrams, a game theoretic extension to Bayesian networks, as a framework for mechanism design in privacy economics [76].

5.3.2 Formal theory of information flow

An upshot of CI is that it identifies privacy as a property of information flows, which when unpacked proves to be a more substantive claim than it may first appear. When we speak about "consumer information" or "personal information", we are faced with the ambiguity of the meaning of the word "information", which can mean alternatively either a medium of representation (such as paper or electronic records, "data") or a mathematical relationship between events or objects such that one is sufficient for inferences about the other [101].

Section 4.4 provides a mathematical analysis of the concept of *information flow* on robust foundations: Dretske's philosophical theory of information flow and Pearl's account of statistical causation.

Dretske's [43] definition that a message carries information about something it represents if and only if it messages of its kind carry a regular or “nomic” relationship with what is represented. Dretske develops this philosophical account of information flow to be consistent with classical information theory [124], in which an information channel establishes a correspondence between the probability distributions of two random events. The emphasis on the regularity of the probabilistic relationship suggests the need for an account of how messages can flow in a structured way.

Just such a theory of structured probabilistic relationships can be found in Pearl's theory of statistical probability and causation [106], and more generally theory around Bayesian networks. Bayesian networks provide a formulation of precisely how causally linked events can be correlated without being directly caused by each other. For example, two events that share a common cause can be correlated. This means that the nomic associations of a message depend not just on who sent the message but how the message is situated in a larger context of messages.

Information flow therefore decomposes into two related parts, *causal flow* of events and their relationship to each other, and *nomic associations* between events. Both of these properties of information flow can be deduced from a model of information's context as a Bayesian network.

A fully specified Bayesian network, complete with conditional probability distributions at every node, will determine not just the existence of a nomic assocation (or, equivalently, a conditional independence), but also the strength fo the association. Many measures of associative strength are possible, but one useful measure that is very well understood is Shannon's *mutual information*:

Definition 51 (Mutual information)**.** The mutual information of two discrete random variables X and Y is

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log \frac{p(x,y)}{p(x)p*y)}$$

In particular, $I(X,Y) = 0 \iff X \perp\!\!\!\perp Y$.

See Appendix B for theorems concerning the ways bounds on mutual information between variables can be read off of Bayesian networks.

5.3.3 Data Games

Introduced briefly in Chapter 4, the Multi-Agent Influence Diagram (MAID) framework developed by Koller and Milch [76] provides a game-theoretic extension to Bayesian networks. As a formalism, it is well suited for modeling how information flows, which we have detailed as causal flows with nomic associations, play a role in strategic games. A full account of the formalism is in the Appendix C.

MAIDs have three types of nodes: chance variables, decision variables, and utility variables. Chance variables are much like the nodes in a Bayesian network: a CPD is defined for each chance variable that conditions on its parent nodes.

Utility variables are much like chance variables, but they are each assigned to an agent $a \in \mathcal{A}$ and they may not have children. The utility for each player in the game defined by the MAID is the sum of the values of the utility nodes assigned to them.

Decision variables are assigned to an agent $a \in \mathcal{A}$. Their CPD functions are not defined as part of the MAID. Rather, the choice of CPD function for each decision variable is a strategic choice of the agent. The strategy profile σ_a for each agent is their assignment of a CPD to each decision variable. Taken to together, the strategy σ of all the players induces a MAID into a Bayesian network, from which the expected utilitites of all players may be computed.

I define a data game as a MAID adapted into a framework for mechanism design. This is done with an extension to the formalism. We introduce the mechanic of an *optional edge*, represented in our diagrams as a dotted edge.

$$\begin{tikzcd} A \arrow[r, dotted] & B \end{tikzcd}$$

A dotted edge represents a potential information flow whose value is the focus of the study. An optional edge means a diagram represents two distinct MAIDs, one with the edge "open" or present in the graph, and one with the edge "closed" or absent. We will look at the outcomes of the open and closed cases and evaluate them according to values like efficiency and equity.

Intuitively, there's a difference between information that becomes suddenly available, as in a data breach, and well-establish information flows to which everyone is accustomed, such as security cameras in malls. In both cases the information flow will have an effect on outcomes, but the cases are subtly different. This difference is reflected in data games by the distinction between the tactical and strategic value of information (this is formalized in Appendix C.2). The tactical value of information is its value to an agent assuming all other agent's strategies remain fixed. The strategic value of information is the difference in an agent's utilities in the open and closed cases, considering a strategic equilibrium of all players in each case.

In the cases discussed in this chapter, I will consider the strategic value of information flow except when specifically stated otherwise.[3]

5.4 Single context economic models

In this section, I will present data games corresponding to single economic contexts. Two of these correspond to well-known phenomena in privacy economics: pricipal-agent contracts and price differentiation. One of these is a new economic model of personalized expert services. These data games make explicit the relationship between information flow and contextual outcomes. This reveals the strategic value of information flow in each economic context.

5.4.1 Agent quality uncertainty

One of the first contexts studied under the term "privacy economics" was labor markets [109]. In labor, insurance, and credit markets, a firm must evaluate natural persons for their individual capacities (to perform a certain kind of work, to avoid risk, or to repay a loan) and decide whether to invest resources in them. The firm generally benefits from having more information about the persons under consideration. The effect of privacy, or lack of it, is uneven across the population being considered by the firm. Paradigmatically, more suitable employees are benefited if their suitability is known to potential employers, while conversely less suitable employees are harmed by the same. Analogous results hold for credit and insurance.

We can model this interaction with the graph in Figure 5.1.

In this model, V represents the value to a principal of a service or contract with an agent. For simplicity, in the model V is normalized with a predetermined price, so the

[3] Section 4.9.2.1 discusses a case where tactical and strategic properties of a system are different: modeling the reaction to a pontential security threat.

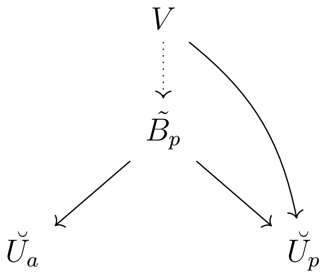


Figure 5.1: Graphical model of principal-agent contract data game.

value of V may be negative. At $\tilde{B}_p$, the principal decides whether or not to buy the contract; $dom(\tilde{B}_p) = \{0, 1\}$.

The utility awarded to the principal is the normalized value of the contract if the principal buys and zero otherwise.

$$U_p = \begin{cases} V & \text{if } \tilde{B}_p = 1 \\ 0, & \text{otherwise} \end{cases} = \tilde{B}_p V$$

The utility for the agent is, for simplicity, a fixed amount (for example, 1) if the principal buys the contract, and zero otherwise; so $U_a = \tilde{B}_p$.

This model affords some simplifications through backwards induction. The optimal strategy for the principal is to buy the contract if the expected value of it is positive. If the dotted edge is open, then the principal is able to use the knowledge of V to make this tactical decision.

If the dotted edge is closed, then the optimal decision $\hat{B}_p$ depends only on the distribution of V.

$$\begin{aligned} \hat{B}_p &= \arg\max_{b_p \in 0,1} \mathbb{E}(U_p) \\ &= \arg\max_{b_p \in 0,1} b_p \mathbb{E}(V) = \begin{cases} 1 & \text{if } \mathbb{E}(V) \geq 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5.1}$$

If the dotted edge is open, then the decision to buy the contract will be better informed.

$$\begin{aligned} B_p | (\hat{V} = v) &= \begin{cases} 1 & \text{if } v \geq 0 \\ 0, & \text{otherwise} \end{cases} \\ &= [v \geq 0] \end{aligned} \tag{5.2}$$

$\mathbb{E}(\cdot)$	Open	Closed
U_p	$\mathbb{E}(V\|V \geq 0)P(V \geq 0)$	$[\mathbb{E}(V) \geq 0]V_E$
U_a	$P(V \geq 0)$	$[\mathbb{E}(V) \geq 0]$
$U_a\|V \geq 0$	1	$[\mathbb{E}(V) \geq 0]$
$U_a\|V < 0$	0	$[\mathbb{E}(V) \geq 0]$

Table 5.1: Outcomes for open and closed cases of the principal-agent contract data game with $V_E = \mathbb{E}(V)$.

Example 52. Let V range over $\{-1, 1\}$ with even odds. $V_E = 0$, $[\mathbb{E}(V) \geq 0] = 1$. So the utility to the agent in the closed case, whatever their quality, is 1, and the expected utility to the principal is 0. In the open case, the principal's utility is $\mathbb{E}(V|V \geq 0)P(V \geq 0) = 1 * (.5) = .5$. The high-quality agents get utility 1, and the low quality agent gets utility 0.

From this example, we can see that principals and agents who can offer more valuable contracts benefit from more openness, while agents with low-value contracts suffer.

5.4.1.1 Values

Early work on privacy economics reasoned that flow of personal information in labor markets leads to greater economic efficiency [109]. The MAID model in Section 5.4.1 corroborates this result. More flow of personal information (the open condition) brings greater utility to the principal on average, and this is a form of market surplus.

It must also be noted that personal information flow has an unequal effect on the contract agents. Less valuable contract agents are negatively impacted by the flow of their personal information. In this narrowly considered economic context there is a global tradeoff between economic productivity, lubricated by flows of personal information, and equality.

This model is general enough to extend to cases where agents are not natural persons but rather firms. Indeed, the situation may be flipped: a single natural person may have to choose among many firms in order to, for example, contract an improvement to their home. The model therefore generalizes from cases of privacy economics to other cases where there is quality uncertainty and the buyer has market power. A question for policy designers is whether individual privacy is any more worthy of protection than information about firms to those who would hire their services, and why.

One reason to be wary of a hiring or other contract choice depending on personal information is indirect discrimination. If contract value is negatively correlated with membership in a protected class of persons, choosing contracts solely on the basis of value might compound an injustice, which Hellman argues there is a duty to avoid doing [60]. Modeling historical injustice with MAIDs is a problem left for future work.

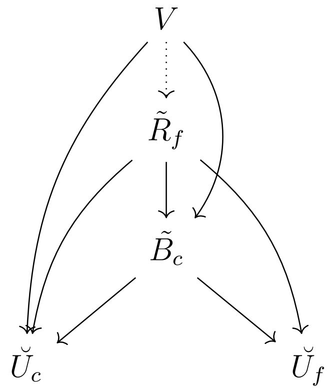


Figure 5.2: Graphical model of price differentiaion data game.

5.4.2 Price differentiation

It is well known that personal information is used by on-line retailers for price differentiation [125, 141]. According the classic economic theory, when a firm charges all consumers at the same price it leaves some unserved by the market (because the price exceeds their demand) and some accruing a consumer surplus (because their demand exceeds the price). With differentiated prices, a firm can charge individual or groups of consumers closer to their reservation prices. This reduces deadweight loss by charging consumers with very low willingness to pay a price they can afford, while transforming consumer surplus formerly accrued by those with high reservation price to the firm as producer surplus.

We can model this context graphically with Figure 5.2.

In this model, V represents a consumer's demand for a product. $\tilde{R}_F$ is the price offered by the firm for the product (costs normalized out) based on the available information S. At $\tilde{B}_c$, the consumer decides whether or not to buy the product; $dom(\tilde{B}_c) = \{0, 1\}$.

The firm's utility is the offered price of the product if it is bought and zero otherwise; $U_f = B_c R_p$.

The consumer's utility is their demand minus the price if they buy the product, and zero otherwise; $U_c = B_c(V - R_p)$.

Once again, we can consider two cases. In the "closed" case, the firm does not know the demand of the individual consumer. They only know the general distribution. The consumer will buy the product if and only if the price is lower than their demand or reservation price; $\hat{B} = [V > R]$. The firm must choose $\hat{R}$ that maximizes their expected revenue:

$$\hat{R} = \arg\max_{r \in \mathbb{R}} \mathbb{E}(r[V > r])$$

If $V \geq \hat{R}$, then the consumer will find the price agreeable and purchase the good,

$\mathbb{E}(\cdot)$	Open	Closed
U_f	$V_E - \epsilon$	$\hat{R}P(V \geq \hat{R})$
U_c	ϵ	$(V - \hat{R})[V \geq \hat{R}]$
$U_c \vert V \geq V_E$	ϵ	$V - \hat{R}$
$U_c \vert V < V_E$	ϵ	0

Table 5.2: Outcomes of open and closed cases in the price differentiation data game with $V_E = \mathbb{E}(V)$.

accruing $V - \hat{R}$ utility. Otherwise, they will not purchase the good.

In the "open" case, the producer knows the reservation price $V = v$ when deciding their price $\hat{R}$.

$$\hat{R} = \arg\max_{r \in \mathbb{R}} r[v > r] = v - \epsilon$$

This value approaches v from below, and for simplicity of presentation we will use ϵ to represent a vanishingly small value.

Example 53. Let V be a uniform distribution ranging $[0, 1]$.

In the closed condition, $U_f = r(1 - r) = r - r^2$, implying the $\hat{R} = .5$ and $U_f = .25$. $\mathbb{E}(U_c) = .125$.

In the open condition, $\hat{R} = V - \epsilon$ and $\mathbb{E}(U_f) = .5 - \epsilon$. Consumer utility is ϵ.

In this price differentiation case, the strategic value of information to the producer is positive. The strategic value of the information to the consumer depends on the consumers' reservation price: it may be negative, or it may be very slightly positive. In general, allowing information flow for price differentiation is better for producers than for consumers.

5.4.2.1 Values

This model shows the tradeoffs of allowing information flow in the economic context of price differentiation. The outcomes for different agents can be inferred from the model. It's clear that information flow for the purpose of price differentiation primarily serves the firm selling a good or service. Arguably, this is valuable because it allows firms to recoup fixed costs for product development with greater sales.

However, this model shows that price differentiation is on the whole bad for consumers. While it's true that consumers with low willingness to pay have access to the good with price differentiation, they are charged a price that makes them almost indifferent to the transaction. Meanwhile, consumer surplus has drained from those who valued to the good highly.

This is a case where the purpose of a market context may be hotly contested by different actors within it. If the contextual purpose of the market transaction is to satisfy as much

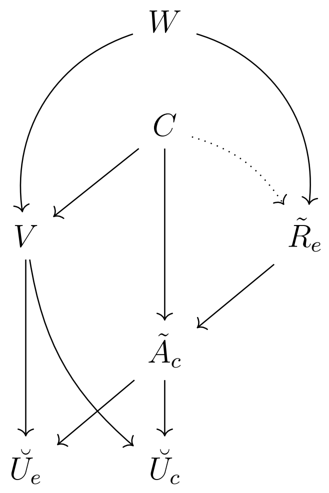


Figure 5.3: Graphical model of the expert service data game.

consumer demand as possible while rewarding productive suppliers, then allowing information flow for price differention is wise policy. But this may be contested by consumer advocates who would argue that consumer satisfaction is more important than economic growth. This context is so raw with economic intent it may be that not societal consensus is possible.

5.4.3 Expertise

Doctors, lawyer, and financial services professionals all have something in common. Their clients consult them for their expertise. In the schematic interaction we'll consider in this section, we'll consider the case where these clients must divulge personal information to an expert in order to get a personalized recommendation or response.

In many of these domains, there are already strong data protection laws in place in the United States. HIPAA in health care, GLBA in personal finance, and FERPA in education all place restrictions on institution's ability to disclose personal information that's collected as a part of that instition's normal professional service. (See Section 5.2.2.) Notably, there is no similar data protection law for search engine queries, which may also be considered a kind of expert recommendation service.

The MAID modeling tool we have been using can capture the difference in knowledge between the client and the expert and the consequences that has for the service market. This model was introduced briefly in Section 4.9.1 and is shown here again as Figure 5.3.

In this model, W are facts about the world that determines the relationship between personal qualities of clients and the best course of action taken by them. For example, this may be thought of as parameters in a function from symptoms of illness to appropriate prescribed remedies. Its domain is some flavor of n-by-m matrices, where n is the number of personal characteristics or types in the model, m is the number of actions available to the client, and $W_{i,j}$ is the reward to a client with of type i of action j.

The variable C encodes those personal qualities known to and communicable by the client. The domain of this variable is an integer from 1 to n.

The variable V encodes the value to the particular client of a variety of courses of action that they might take. It depends on W and C, and in the simple version of the model considered here is a deterministic function: V is the row of W indexed by C.

$\tilde{R}_e$ is the strategically determined decision of the expert to recommend a course of action based on their knowledge of W and optionally C. Its domain is an integer from 1 to m. $\tilde{A}_c$ is the decision of which action the client takes. It also has domain from 1 to m. U_c and U_e are the utilities awarded to the client and expert, respectively, which take their value from the vector V indexed by the value of $\hat{A}_c$.

Perhaps idealistically, we have modeled the utility of the expert as depending only on the utility of the client. We imagine that the client pays for the expertise up front, that this is normalized into the value of the action taken V, and that the expert benefits from the positive recommendations of satisfied clients. Future work and other models may explore other possible configurations of incentives, including conflicts of interest. For an action taken $a \in A$, we will specify that $U_c = V(a)$.

We will once again consider two cases. In the closed case, there is no edge from C to $\tilde{R}_e$. In this case, the expert still has specialized knowledge (the value of W), but no personal information about the client with which to tailor their recommendation. Their best recommendation is the action that would benefit a random client the most in expectation.

$$\hat{R}_{closed} = \arg\max_{a \in A} \mathbb{E}(V(a)|W)$$

The client, on the other hand, has access to information about their symptoms C but not the expert knowledge W. By the assumption of the model, the client does have access to the expert's recommendation, $\hat{R}_{closed}$. So their choice of action is:

$$\hat{A}_{closed} = \arg\max_{a \in A} \mathbb{E}(V(a)|C, \hat{R}_{closed})$$

In the alternative "open" condition, there is an edge between S and $\tilde{R}$.

$$\hat{R}_{open} = \arg\max_{a \in A} \mathbb{E}(V(a)|W, C)$$

$$\hat{A}_{open} = \arg\max_{a \in A} \mathbb{E}(V(a)|C, \hat{R}_{open})$$

The specific utility outcomes depend heavily on the parameters of the model. We can make a few general observations about bounds. If the model is such that individual symptoms carry no information about the value of actions taken even with expert knowledge

$\mathbb{E}(\cdot)$	Open	Closed
U_e	$\mathbb{E}(V(\hat{A}_{open}))$	$\mathbb{E}(V(\hat{A}_{closed}))$
U_c	$\mathbb{E}(V(\hat{A}_{open}))$	$\mathbb{E}(V(\hat{A}_{closed}))$

Table 5.3: Outcomes for open and closed cases of the expert service data game.

taken into account ($V \perp\!\!\!\perp C|W, V \perp\!\!\!\perp C$), then the welfare outcomes in the closed case and the open case will be the same.

If the expert knowledge W has information about the action values given the symptoms ($I(W;V|C) > 0$), then the expert recommendation $\hat{R}_{open}$ will generally be better in expectation than $\hat{A}_{closed}$ and indeed $\hat{A}_{open} = \hat{R}_{open}$.

Note that there is an interaction between the strategies of the expert and the client. The optimality of the expert's strategy at $\tilde{R}_e$ depends on how its signal will be "interpreted" at $\tilde{A}_c$. Interpreting $\tilde{R}_e$ as a recommendation implies some correspondence between the value taken at that variable and the values of actions according to V. But in some cases an alternative encoding of the information in W (and C) may be more efficient.

An example of the instantiated model will illustrate these points.

Example 54. Let n and m both equal 2. Let the domain of W be binary 2-by-2 matrices with the restriction that each row contains one 0 and one 1. Let the distribution of W be uniform over the four possible matrices in its domain.

$$dom(W) = \{\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\}$$

In the closed case, $\tilde{R}_e$ does not depend on C. $\tilde{R}_e$ is therefore a strategically chosen encoding of only the information in W. Notably, whereas the random variable has 2 bits of information, $\tilde{R}_e$ ranges over 0 and 1 and can carry at most 1 bit of information.

One such encoding, communicating one bit from W, is:

$$\hat{R}_{closed}|(W = w) = \begin{cases} 1 & \text{if } w_{0,1} = 1 \\ 0, & \text{otherwise} \end{cases}$$

At $\tilde{A}_c$, the client knows both C and $\tilde{R}_e$ and must choose an action that optimizes their expected utility. At this node the client knows if they are of type 0 or 1. If they are of type 0, they know the recommendation applies to them, with certainty of a reward of 1 if they take the suggested action.

$$\hat{A}_c|(C = 0) = \hat{R}_c$$

But if the client knows they are of type 1 (probability $.5$), then the recommendation does not encode information about the value of their action. Whatever action they take has an even chance of having utilty of 0 or 1. Expected utility in the closed case is $.5 * 1 + .5 * .5 = .75$.

In the open case, the expert's recommendation $\tilde{R}_e$ depends on both W and C. From this information, the expert can deduce the one bit of information relevant to the client's decision, which is V.

$$\hat{R}_{open}|(W = w, C = c) = \begin{cases} 1 & \text{if } w_{c,1} = 1 \\ 0, & \text{otherwise} \end{cases}$$

In this case, when $\hat{A} = \hat{R}_{open}$, the value of the action is guaranteed to be 1, which implies that the expected utility in the open case is 1.

The strategic value of the information flow from C to R is the difference in expected utilities in the two cases, which in this example is $1 - .75 = .25$.

In this example, the expert chooses a strategy at $\tilde{R}$ that maximizes the flow of information, in the Shannon sense of the term, to $\tilde{A}$ about another variable of interest, V. Or, formally:

$$\hat{R} = \arg\max_{R} I(R; V)$$

In the closed case, the limited domain of $\tilde{R}$, which permits the flow of at most one bit of information, restricts the expert's ability to provide an adequate recommendation to the client. If the number of bits in R was greater than or equal to the number of bits in W, the expert would be able to communicate the entirety of their expertise to the client, who could then make a perfect judgment of action taking C into account.

This information theoretic lens provides a new view into personalized expert advice. Personalization is useful to the client only because the client lacks expertise, but this lack of expertise is due in part because the expert cannot communicate all the information they know to the client. Personalization allows the expert to provide the highest value information through the narrow bandwidth of communication. The constraints on information flow are due mathematically to the Data Processing Inequality and its consequences for Bayesian networks, which analyzed in depth in Appendix B.

5.4.3.1 Values

This model of expert services has been simplified to exclude cases of expert conflicts of interest that might engage societal values in mechanism design. We accomplished this simplification by directly aligning client and expert incentives. Despite this simplification, the model shows some of the difficulty in modeling the welfare outcomes of expert services. The principle difficulty is that the outcomes depend on general facts and the quality of expertise in a particular domain. Because it is hard to encode an actual field of expertise into a simple model we can prove only very general properties of such a field.

Despite these difficulties, this simple model shows that when expert and client incentives are aligned, greater flow of information from client to expert enables better outcomes for both parties. In a later section, we will elaborate on this model by introducing the possibility of a breach of confidentiality.

5.5 Cross-context information flow and secondary use of personal data

In the above models, we have shown how in a variety of economic contexts the flow of information can have tangible effects on welfare outcomes. What these models have in common is that they show that the relevance of information on outcomes depends on the process by which it is generated and how elements of that process affect outcomes. While we have provided narrative stories for each example where we have described the information flows in terms of particular types of documents or events (job applications, symptoms, etc.), what really gives information its semantics are its associational relationships with other variables. These are given by the conditional probability distribution governing the model.

A reason for modeling games with information flow in this way is to model the economic impact of the *secondary use* of data. What economic reason makes secondary data use so popular for businesses, even though it risks affronting social norms and risks legal violation?

5.5.1 Cooperating Firms: Price differentiation and agent quality

Consider the following model, constructed as a combination of the agent quality uncertainty and price differentiation models. Here c is a natural person who is *both* potentially a consumer of firm f's products and potentially involved in a contract with principal p. The value of this person's contract V and their willingness to pay for the product D both depend on a prior variable W that encapsulates many factors about the background of the person.

$D \rightarrow \tilde{R}_f$ represents the ability of the firm to known the customer's demand before choosing their price. There is also a principal that decides at $\tilde{B}_p$ whether or not to buy a contract with the customer, who in this case is also an agent for hire. In this model, the principal cannot know the value of the contract V directly. Rather, there is a new edge $\tilde{R}_f \rightarrow \tilde{B}_p$ that represents the option of the product-selling firm f to share its pricing information with the contract principal p.

Why would two companies ever interact in this way? If the principal does *not* know the value of a potential contract V directly, then the pricing information $\hat{R}_f$ potentially contains information about V in a way that the principal p can use. Here, "contains information about" can be read to mean "has mutual information with", i.e. $I(\hat{R}_f, V) \geq 0$. This information may be valuable for the principal by allowing them to avoid bad contracts.

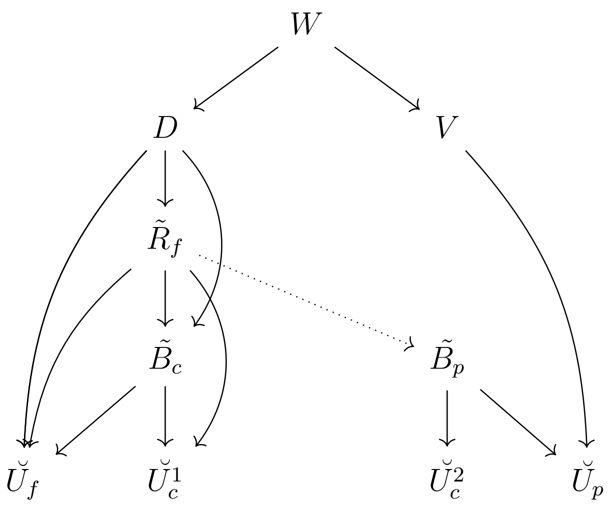


Figure 5.4: Graphical model of cross-context flow data game.

Since the principal and the producing firm's utility's do not interact directly in any other way, we can imagine that the principal would be willing to purchase the pricing data from the producing firm for the *value of the data to the principal*. Though this data relates directly to a natural person, it is not data collected from that person; it is data derived from the producing firm's pricing algorithm. Nevertheless, sharing this data has a function analogous to sharing personal data that could be used in a hiring decision or in offering a loan.

In this model, the firm's incentives are the same as in the simple price differention case in Section 5.4.2. By assumption, the firm knows the customer's demand D, and therefore prices at $\hat{R} = D - \epsilon$.

$\mathbb{E}(\cdot)$	Open	Closed
U_f	$D_E - \epsilon$	$D_E - \epsilon$
U_c^1	ϵ	ϵ
U_p	$\mathbb{E}(V \| V \geq 0) P(V \geq 0 \| \hat{R})$	$[\mathbb{E}(V) \geq 0] V_E$
U_c^2	$P(V \geq 0)$	$[\mathbb{E}(V) \geq 0]$
$U_c^2 \| V \geq 0$	1	$[\mathbb{E}(V) \geq 0]$
$U_c^2 \| V < 0$	0	$[\mathbb{E}(V) \geq 0]$

Table 5.4: Outcomes of opened and closed condition for cross-context flow data game with $D_E = \mathbb{E}(D)$ and $V_E = \mathbb{E}(V)$.

Example 55. Let W vary over $\{0, 1\}$ with even probability, with the value corresponding to one of two socioeconomic classes, *low* and *high*.

In this example, we will assume that higher class people have access to better education and wealth, and therefore both have higher reservation prices and offer higher value contracts to creditors, employers, and insurance providers.

$$D(w) = \begin{cases} 10 & \text{if } w = 1 \\ 1, & \text{otherwise} \end{cases}$$

$$V = \begin{cases} 1 & \text{if } w = 1 \\ -1, & \text{otherwise} \end{cases}$$

The firm has access to the reservation price D at $\tilde{R}_f$ and so will maximize their utility by pricing at slightly below the customer's willingness to pay.

$$\hat{R}(v) = v - \epsilon$$

The optional edge being considered in this case runs from $\tilde{R}$ to $\tilde{B}$. In the closed case, the principal has no information about V on which to decide except the base rate provided by the game structure. The expected value to the principal of providing the contract is 0.

In the open case, $\tilde{B}$ is conditional on $\tilde{R}$. Crucially, V is conditionally dependent on $\hat{R}$. In particular:

$$P(V = 1|\hat{R} > 1) = 1$$
$$P(V = 0|\hat{R} \leq 1) = 1$$

The optimal strategy for the principal is to hire the agent when the firm reveals to them that they offered a high price for the good, and to reject the agent otherwise. The high quality contract is purchase half the time with reward 1 to the principal. So the strategic value of the information flow from $\tilde{R}$ to $\tilde{B}$ to the principal is $.5$. The strategic value to the average customer/agent of this information flow is negative, as it results in many customer/agents not getting hired, whereas in the closed case all potential agents get hired.

In this simple example, in strategic equilibrium the firm's offered price and the customer/agent's contract quality are highly correlated. This means that a *causal flow* between the firm's price and the principal's hiring decision carries a *nomic association* with the contract value. That association has strategic value for the principal similar to the value of having the contract value's causally flowing directly to the hiring decision.

The secondary use of data to determine the social class of natural persons is not an academic hypothetical. Facebook has filed a patent to use information about user's hardware specifications, presumabely collected originally to optimize service performance, to predict social class, presumably useful for targeting advertisements [2]. The cross-context use of personal data for targetted advertising is arguably the fundamental business proposition of on-line advertising companies like Facebook and Google [91].

The strategic value of the information to the principal can be interpreted as the price at which the principal would be willing to purchase this information flow from the firm. This price depends on the causal structure of the environment, including the distribution of qualities of natural persons. Even though natural persons are on average worse off as a result of this information flow, this is not a factor in its value or price to the principal. This can be considered a market externality, though because it involves a new flow of information, it may also correct a market inefficiency.

5.5.1.1 Values

The outcomes of this game are similar to the outcomes in the simple principal agent case. The difference is that the information flow runs between two cooperating firms. This models the flow of information from one economic context (price differentiation on a consumer good) to another (a principal-agent hiring decision). Because the two contexts are part of a shared causal environment, the data from from one context can carry meaningful information relevant to another.

It's notable that the natural person in this example, who is both a customer and an agent to be hired, is (on average) disadvantaged by the hypothetical transaction taking place between two firms. This is a market externality, though the flow of information corrects a market inefficiency in the second economic context. There is a tradeoff between market efficiency, which is good for firms, and the privacy of natural persons, and especially the most vulnerable natural persons.

5.5.2 Secondary use of queries to experts

Section 5.5.1 detailed the potential negative impact on vulnerable natural persons from an information flow that crosses economic contexts. It is possibly because of the potential negative impact of secondary uses of information that so many market segments are protected by sectoral privacy and confidentiality laws. HIPAA in health care, GLBA in personal finance, and FERPA in education all place restrictions on firm's ability to disclose personal information. (See Section 5.2.2.) What these sectors all have in common is their firms require significant disclosures of personal information to provide personalized service. The expertise model in Section 5.4.3 captures why personal information is necessary for the functioning of these services in fundamental mathematical and economic terms: personalization allows expert service providers to deliver more value to clients given a tight information bottleneck relative to the body of knowledge of the expert.

Section 5.5.1 provides a template for understanding why disclosure of sensitive information from the context of expert services into other domains could have negative externalities for natural persons. Indeed, information we provide our doctors, lawyers, financial advisors, and search engines is sensitive precisely because this information is potentially impactful in other contexts in ways that are surprising and/or unwelcome.

Whereas the firms in 5.5.1 both benefit from the sale of personal information, the sale of personal information from expert services may have secondary effects that negatively impact experts. If clients are aware of a harmful information flow, they may be reluctant to engage the expert. In the the terminology introduced earlier, an expert may get tactical value from selling personal information of its clients, but if clients can adjust their behavior according to new expectations, the strategic value of this information flow to experts will be negative.

5.6 Discussion

This chapter has demonstrated "data games", a framework for modeling economic games with information flow. This framework expands MAIDs with optional edges, which results in a system for modeling mechanism design with Bayesian Networks. This framework can model well understood cases of privacy economics (principal/agent, price differentiation) as well as the less understood case of expert services. The framework makes it clear how the fundamental limits of information theory, as well as the nature of information flow as causal flow with nomic associations, relates to the economics of information services. The models show that sometimes personal information flow improves market efficiency at the expense of consumers and riskier agents. These models allow for a direct comparision between social values and the outcomes of policies allowing or disallowing personal information flows. (Section 5.4.)

This framework can also model cases where information flows between economic contexts (Section 5.5.) In particular, secondary use of personal information can play an economic role similar to primary use of personal information if and when the processes that generate the data result in reliable and useful statistical correlations. These correlations can occur when society is stratified into socioeconomic classes, as they are in reality.

Central to this modeling system is a conceptual shift in how to understand the role of information flow in economics. These models make clear that the strategic choices of agents in the economy is one of the elements that determines the causal structure that gives information its meaning. This indicates a major source of confusion in economics of information. Information is not a good that is bought and sold for consumption. Information is a strategic resource, part of the social and economic fabric. When information flows are bought and sold, it changes the strategic landscape of the economy. Market externalities abound as information flows effect many parties who are not party to transactions.

Beyond these general conclusions, there are a number of more specific implications of these models which indicate directions of future work.

5.6.1 Privacy concerns and privacy competence

A robust empirical result is that there are different segments of the general population that have different privacy concerns. These are often presented as the marginally concerned, pragmatic majority, and privacy fundamentalists [4] [17] [126]. This matches results from the economic models: in this chapter, some populations are more vulnerable than others to the negative effects of personal information flow. Further work is needed to test to what extent different preferences or concerns about information flow are determined by the economic situation of data subjects. Class differences may have significant effects, with implications for value-driven policy design [81].

There is also evidence that consumers are generally not making privacy decisions in rational and informed self-interest but, rather, become much more concerned with their privacy when told facts about how personal information is used [67]. There is a disconnect between consumer expectations and fact. This may be because the most prominent privacy threats are beyond user comprehension.

Many serious privacy threats, whether they be Big Data analytics drawing conclusions from aggregating for an unforeseen business end, a network of companies engaged in secondary uses of data shared between them, or an illicit dark web of hackers and fraudsters, are due to cross-context information flows in which the data subject plays little active role [93]. Section 5.5 shows the mechanics of how companies can gain strategic advantage by reusing personal data to the detriment of consumers. However, if consumer privacy expectations are tied to the normative expectations in specific social spheres, perhaps because these expectations are encoded as mental models or causally structured frames, then consumers can not be expected to be competent stewards of their own personal information. Consumers cannot act strategically in their interest, individually let alone collectively, unless they are aware of how their information is being used.

The true mechanics of information flow, represented here by Bayesian networks, are opaque and largely unknown. The framework provided here can be extended to take into account different degrees of knowledge about the causal structure that gives information flow its meaning. This sheds light on the well known problem of the ineffectiveness of privacy self-management [135]. Further work is needed to understand the implications of knowledge and information assymetry in data economy market equilibria.

5.6.2 Market failure

By the preceding argument, consumers are not competent to make decisions about how to control their personal information because their privacy expectations are tied to contexts that are routinely violated in practice. Potential secondary uses of personal data depend on associational properties of the data that are beyond users comprehension. In the case of large, data-rich firms, these associational properties are discovered through aggregation and data mining by the very firms that attract consumer interaction through expert services that they offer. This data is then used in two-sided advertising markets, which act as intermediaries in many other economic contexts, further complicating any prediction of the benefits and harms of disclosure. Quantitative, let alone qualitative, prediction of these harms and benefits is beyond what an individal can accomplish.

In the absence of a more concrete culprit for privacy threats, security considerations raise a general case for needing to limit secondary use of personal information. On the one hand, we can consider security to be another context where personal information is used, perhaps in a secondary way. Uses of personal information which are harmful to all affected consumers include those that facilitate security threats like spearphishing (when attackers use personal information to manipulate a person to reveal security-related information or otherwise be a vector for a further attack) and identity theft. On the other hand, it is the possibility of harmful secondary use *across all potential contexts* that makes security of personal information so important in the first place. Security in this sense is necessary for an implementation of confidentiality.

The conditions appear to be ripe for classic market failure As has been mentioned, property rights for personal data are weak if not nonexistant (see Section 5.2). Metaphors aside, personal data is not a good being *produced* by anybody in the privacy economics ecosystem. It is rather information in the *strategic* sense of allowing some market actors to perform more effectively. There is no sense in which the market of personal information has the properties that would lead us to believe the market would allocate resources efficiently, because if there was perfect information in the data market, it would cease to exist.

As an alternative to regulating personal data as a kind of property, some have proposed regulating personal data through tort [109, 26]. Certainly some meanings of "privacy", such as those that refer to protection from libel, are enforceable in this way. To the extent that considering personal data to be a *thing* is misleading, it may more more effective to craft data protection regulation through the framework of dignitary privacy [110]. However, as we have discussed it seems unlikely that the scope of consumer harm or benefit can be adequately assessed given the scale of the empirical problems involved.

Another alternative is strengthened data protection laws for two-sided markets involving targeting, such as advertising in social media platforms. As we have noted, in most expert service sectors, including health care, finance, education, and so on, there are existing sectoral data protection laws ensuring confidentiality. The existence of these laws is an indication that without them, these expert service markets would degrade through market failure. If protecting confidential information from secondary use (through austere prohibitions on disclosure and security investments) is a form of service *quality*, and this quality is difficult for consumers to assess independently, then this information assymetry about service quality would result in a market failure along the lines of Akerlof's market for "lemons" [6]. Since unregulated two-sided markets are in the senses described above equivalent to providing unrestricted secondary use to other firms, perhaps present economic conditions are just such a market failure.

5.6.3 Purposes

As discussed in Section 5.2.4, the EU's GDPR attempts to limit the kinds of privacy violations due to secondary use of personal information through purpose restrictions, which place restrictions on the goals for which collected data may be used. Personal data may be processed only for purposes to which the data subjects consent (with some exceptions). Data minimization requirements reduce the amount to which data is unintendedly exposed to other purposes. As a way of creating agreement between the expectations of data subjects and the activities of data processors, this can be seen as a refinement of the notice-and-consent framework. It may be argued that purposes are easier to understand than the complexity of legal and technical reality.

The rising importance of purpose binding as a privacy requirement raises the question of how the purpose of data processing can be formalized to facilitate privacy engineering. Tschantz et al. [138] [139] do so by positing that "an action is for a purpose if the action is part of a plan for achieving that purpose." They then go on to formalize this in terms of a Markov Decision Process (MDP), a way of modeling the relationship between actions, environment, policies, and outcomes that allows for a formal definition of optimal policy. A promising direction for future work is to formalize purpose binding in terms of Bayesian causality and incentives, extending the mechanism design framework introduced in this chapter. Can purpose be modeled in a data game?

Chapter 6

Conclusion and Future Work

This dissertation introduces data games, a formal method for determining the value of information in strategic games involving situated information flow. This method depends on a core theoretical contribution: the definition of information flow as a channel in the context of a causal system. This definition is built on the theoretical contributions of Shannon, Dretske, Pearl, Nissenbaum, and Koller, and motivated by legal analysis. It's applicability has been demonstrated through several studies in computer security and data economics. Among other uses, this definition and method is able to answer a question opened by the systematic study of Contextual Integrity's projection in computer science: the problem of cross-context information flows. This theoretical contribution therefore stands to inform future efforts in Privacy-by-Design [58].

One implied by this definition and method is that in the data economy information is not a good; it is a strategic resource. A change in information flow is a change in the strategic landscape and the structure of the social field. Because a new information flow can change the nomic associations of many connected events and actions in the economic field, any transaction may have a wide variety of externalities. A consequence of this is that traditional market economics developed for tangible goods are a poor fit for the information economy. More work is needed to develop information economics in a way that takes the causal structure of information flows into account.

Though the definition of information flow developed in this dissertation depends on mathematical results from information theory, statistics, and computer science, it has implications for the structure of the information economy and the design and regulation of sociotechnical systems. Among other purposes, it is intended to inform the pragmatic matter of information law. Hence, I conclude this dissertation by explicitly positioning it as a scientific contribution, in two senses. First, in the Bourdieusian sense [19]: as a claim to the arbitration of the real, as opposed to as a way of participating in an academic discipline which do not make such claims. Second, simply by virtue of its dependence on mathematics that are common in computer science and statistics while also addressing social phenomena, I position it as a contribution to the emerging field of computational social science, which gets its legitimacy from its acceptance of the objectivity of mathematical proof and the shared use of computational instrumentation [15].

I believe the formal constructs developed in this dissertation are a good start to a scientific understanding of social information flow. But it is only a start. I end with a coda in the same spirit of Chapter 3: an expression of where this work resonates on a different intellectual plane.

The definition of information flow used in this dissertation depends on models of the world in terms of Bayesian networks and Pearlian causation. It may be objected that the world can only ever approximately be described in terms of probabilistic events, especially considering the unresolved tensions over the interpetation of probability as either subjectivist or frequentist. To address these questions, consider: Pearlian causation has been successful not only as a statistical theory of causation useful in machine learning and the social sciences, but also in philosophy [145] and cognitive psychology [131], including cognitive psychology of moral judgement [132]. A reason to use Pearlian causation in modeling the world is that it formalizes how we experience and reason about the world implicitly. We cannot escape the constraints that are implicit in the structure of our experience. And our implicit models of how the world works are structured in ways that can be characterized as Pearlian models.

Critically, our models of the world are often wrong. This detail marks the inadequacy of the models presented in this dissertation: they assume that all agents have a shared understanding of the true causal structure of the world. This is hardly ever the case, and so there are a number of open problems resulting from the need to model differences in observer capabilities. (This point has been made already in Section 4.10.2).

A key consideration for future work is that whereas the true nomic associations of an event depend on the true causal structure of the world, the intepreted meaning of an event to an observer will depend on their model of that structure. In terms of Shannon information theory, the world encodes itself into an event that the observer decodes upon observation. When there is a mismatch between the encoding and the decoding, information is lost.

Proper analytic treatment of this subject will require more formal work. Speculatively, we can hypothesize that the social need for shared and accurate models of how information is generated is one of the reasons why society differentiates into social spheres: contextual roles, purposes, and information norms maintain the matching between social processes generating information and expectations of meaning society depends on to function. Developing and testing this hypothesis may have the happy result of putting contextual integrity onto the robust theoretical and empirical footing of computational sociology. This too must rest for now and await future work.

Bibliography

[1] Privacy risk management for federal information systems. NISTIR 8062 (Draft), 2015. NIST.

[2] Facebook patents tech to bucket users into different social classes, Feb 2018. URL https://www.cbinsights.com/research/facebook-patent-socioeconomic-detection/9375681.

[3] Mark Ackerman, Trevor Darrell, and Daniel J Weitzner. Privacy in context. *Human–Computer Interaction*, 16(2-4):167–176, 2001.

[4] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 1–8. ACM, 1999.

[5] Alessandro Acquisti, Curtis R Taylor, and Liad Wagman. The economics of privacy. *Journal of Economics Literature*, 52, 2016.

[6] George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, pages 488–500, 1970.

[7] Ronald J Allen, Mark F Grady, Daniel D Polsby, and Michael S Yashko. A positive theory of the attorney-client privilege and the work product doctrine. *The Journal of Legal Studies*, 19(2):359–397, 1990.

[8] Solon Barocas and Helen Nissenbaum. On notice: The trouble with notice and choice. In *First International Forum on the Application of and Management of Personal Electronic Information*, 2009.

[9] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and Contextual Integrity: Framework and applications. In *IEEE Symposium on Security and Privacy*, 2006. doi: 10.1109/SP.2006.32.

[10] Adam Barth, John Mitchell, Anupam Datta, and Sharada Sundaram. Privacy and utility in business processes. In *Proceedings of the 20th IEEE Computer Security Foundations Symposium*, CSF '07, pages 279–294. IEEE Computer Society, 2007. ISBN 0-7695-2819-8. doi: 10.1109/CSF.2007.26.

[11] Adam Barth, John Mitchell, Anupam Datta, and Sharada Sundaram. Privacy and utility in business processes. In *Computer Security Foundations Symposium, 2007. CSF'07. 20th IEEE*, pages 279–294. IEEE, 2007.

[12] Gilles Barthe, Pedro R. D'Argenio, and Tamara Rezk. Secure information flow by self-composition. In *CSFW '04: Proceedings of the 17th IEEE Computer Security Foundations Workshop*, page 100, 2004. ISBN 0-7695-2169-X. doi: http://dx.doi.org/10.1109/CSFW.2004.17.

[13] Gregory Bateson. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology*. University of Chicago Press, 1972.

[14] Sebastian Benthall. Designing networked publics for communicative action. *Interface*, 1(1):3, 2015.

[15] Sebastian Benthall. Philosophy of computational social science. *Cosmos and History: The Journal of Natural and Social Philosophy*, 12(2):13–30, 2016.

[16] Sebastian Benthall, Seda Gürses, Helen Nissenbaum, et al. Contextual integrity through the lens of computer science. *Foundations and Trends® in Privacy and Security*, 2(1):1–69, 2017.

[17] Bettina Berendt, Oliver Günther, and Sarah Spiekermann. Privacy in e-commerce: stated preferences vs. actual behavior. *Communications of the ACM*, 48(4):101–106, 2005.

[18] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[19] Pierre Bourdieu. *Science of science and reflexivity*. Polity, 2004.

[20] Martin Bradley and Alexander Dent. Payment card industry data security standard. 2010.

[21] Søren Brier. *Cybersemiotics: Why information is not enough!* University of Toronto Press, 2008.

[22] Michael K Buckland. Information as thing. *Journal of the American Society for Information Science (1986-1998)*, 42(5):351, 1991.

[23] Ryan Calo. The boundaries of privacy harm. *Ind. LJ*, 86:1131, 2011.

[24] Ann Cavoukian et al. Whole body imaging in airport scanners: Activate privacy filters to achieve security and privacy. 2009.

[25] David D Clark, John Wroclawski, Karen R Sollins, and Robert Braden. Tussle in cyberspace: defining tomorrow's internet. In *ACM SIGCOMM Computer Communication Review*, volume 32, pages 347–356. ACM, 2002.

[26] Ignacio N Cofone. The dynamic effect of information privacy law. *Minn. JL Sci. & Tech.*, 18:517, 2017.

[27] Federal Trade Commission et al. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers, Mar 2012.

[28] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[29] Bo Cowgill and Catherine Tucker. Algorithmic bias: A counterfactual perspective. *Working Paper: NSF Trustworthy Algorithms*, 2017.

[30] Natalia Criado and Jose M Such. Implicit contextual integrity in online social networks. *Information Sciences*, 325:48–69, 2015.

[31] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54. ACM, 2016.

[32] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

[33] George Danezis, Josep Domingo-Ferrer, Marit Hansen, Jaap-Henk Hoepman, Daniel Le Metayer, Rodica Tirtea, and Stefan Schiffner. Privacy and data protection by design-from policy to engineering. *arXiv preprint arXiv:1501.03726*, 2015.

[34] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. Discrimination in online advertising: A multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency*, pages 20–34, 2018.

[35] Anupam Datta, Jeremiah Blocki, Nicolas Christin, Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kaynar, and Arunesh Sinha. Understanding and protecting privacy: Formal semantics and principled audit mechanisms. In *International Conference on Information Systems Security*, pages 1–27. Springer, 2011.

[36] Anupam Datta, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. *arXiv preprint arXiv:1705.07807*, 2017.

[37] Jenny L Davis and Nathan Jurgenson. Context collapse: theorizing context collusions and collisions. *Information, Communication & Society*, 17(4):476–485, 2014.

[38] Terrence W Deacon. Steps to a science of biosemiotics. *Green Letters*, 19(3):293–311, 2015.

[39] Anind K Dey, Gregory D Abowd, and Daniel Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, 16(2):97–166, 2001.

[40] Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kaynar, and Anupam Datta. Experiences in the logical specification of the hipaa and glba privacy laws. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*, pages 73–82. ACM, 2010.

[41] Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kirli Kaynar, and Anupam Datta. Experiences in the logical specification of the HIPAA and GLBA privacy laws. In *WPES*, pages 73–82, 2010.

[42] Paul Dourish. What we talk about when we talk about context. *Personal and ubiquitous computing*, 8(1):19–30, 2004.

[43] Fred Dretske. *Knowledge and the Flow of Information*. MIT Press, 1981.

[44] Fred I Dretske. The epistemology of belief. *Synthese*, 55(1):3–19, 1983.

[45] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006. ISBN 3-540-35907-9. doi: 10.1007/11787006_1.

[46] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107, 2008.

[47] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[48] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[49] Nan Feng, Harry Jiannan Wang, and Minqiang Li. A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis. *Information sciences*, 256:57–73, 2014.

[50] Kathi Fisler, Shriram Krishnamurthi, and Daniel J Dougherty. Embracing policy engineering. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pages 109–110. ACM, 2010.

[51] Luciano Floridi. Semantic conceptions of information. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[52] World Economic Forum. Rethinking personal data: A new lens for strengthening trust, 2012.

[53] Dan Geiger and Judea Pearl. Logical and algorithmic properties of independence and their application to bayesian networks. *Annals of Mathematics and Artificial Intelligence*, 2(1-4):165–178, 1990.

[54] Joseph A. Goguen and Jose Meseguer. Security policies and security models. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 11–20, 1982.

[55] James W. Gray, III. Toward a mathematical foundation for information flow security. In *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, pages 21–34, 1991.

[56] Seda Gürses and Jose M del Alamo. Privacy engineering: Shaping an emerging field of research and practice. *IEEE Security & Privacy*, 14(2):40–46, 2016.

[57] Seda Gürses and Claudia Diaz. Two tales of privacy in online social networks. *IEEE Security & Privacy*, 11(3):29–37, 2013.

[58] Seda Gürses, Carmela Troncoso, and Claudia Diaz. Engineering privacy by design. 2011.

[59] Geoffrey C Hazard Jr. An historical perspective on the attorney-client privilege. *California Law Review*, pages 1061–1091, 1978.

[60] Deborah Hellman. Indirect discrimination and the duty to avoid compounding injustice. *Virginia Public Law and Legal Theory Research Paper*, 2018.

[61] Michael Herrmann, Mireille Hildebrandt, Laura Tielemans, and Claudia Diaz. Privacy in location-based services: An interdisciplinary approach. *SCRIPTed*, 13:144, 2016.

[62] Mireille Hildebrandt. Slaves to big data. or are we? 2013.

[63] K. Hill. How target figured out a teen girl was pregnant before her father did. *Forbes*, 2012.

[64] Jamie Hine. Mergers and privacy promises, Mar 2015. URL https://www.ftc.gov/news-events/blogs/business-blog/2015/03/mergers-privacy-promises.

[65] Dennis D Hirsch. The glass house effect: Big data, the new oil, and the power of analogy. *Me. L. Rev.*, 66:373, 2013.

[66] O. W. Holmes. The path of the law. *Harvard Law Review*, 1897.

[67] Chris Jay Hoofnagle and Jennifer M Urban. Alan westin's privacy homo economicus. *Wake Forest Law Review*, 2014.

[68] Eric Horvitz and Deirdre Mulligan. Data, privacy, and the greater good. *Science*, 349(6245):253–255, 2015.

[69] White House. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. *White House, Washington, DC*, pages 1–62, 2012.

[70] Leonid Hurwicz and Stanley Reiter. *Designing economic mechanisms*. Cambridge University Press, 2006.

[71] Yunhan Jack Jia, Qi Alfred Chen, Shiqi Wang, Amir Rahmati, Earlence Fernandes, Z Morley Mao, Atul Prakash, and Shanghai JiaoTong Unviersity. Contexiot: Towards providing contextual integrity to appified iot platforms. In *Proceedings of The Network and Distributed System Security Symposium*, volume 2017, 2017.

[72] Shiva Prasad Kasivisiwanathan and Adam Smith. On the 'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1):1–16, 2014.

[73] Imrul Kayes and Adriana Iamnitchi. Aegis: A semantic implementation of privacy as contextual integrity in social ecosystems. In *Privacy, security and trust (pst), 2013 eleventh annual international conference on*, pages 88–97. IEEE, 2013.

[74] Imrul Kayes and Adriana Iamnitchi. Out of the wild: On generating default policies in social ecosystems. In *Communications Workshops (ICC), 2013 IEEE International Conference on*, pages 204–208. IEEE, 2013.

[75] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. Technical report, Keele University and University of Durham, 2007.

[76] Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

[77] Yann Krupa and Laurent Vercouter. Handling privacy as contextual integrity in decentralized virtual communities: The privacias framework. *Web Intelligence and Agent Systems: An International Journal*, 10(1):105–116, 2012.

[78] Tessa Lau, Oren Etzioni, and Daniel S Weld. Privacy interfaces for information management. *Communications of the ACM*, 42(10):88–94, 1999.

[79] Lawrence Lessig. *Code: And other laws of cyberspace*. ReadHowYouWant. com, 2009.

[80] Niklas Luhmann. *Social systems*. Stanford University Press, 1995.

[81] Mary Madden, Michele Gilman, Karen Levy, and Alice Marwick. Privacy, poverty, and big data: A matrix of vulnerabilities for poor americans. *Wash. UL Rev.*, 95:53, 2017.

[82] Bradley Malin and Latanya Sweeney. Re-identification of dna through an automated linkage process. In *Proceedings of the AMIA Symposium*, page 423. American Medical Informatics Association, 2001.

[83] Kirsten Martin and Helen Nissenbaum. Measuring privacy: an empirical test using context to expose confounding variables. *Colum. Sci. & Tech. L. Rev.*, 18:176, 2016.

[84] Alice E Marwick and Danah Boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1): 114–133, 2011.

[85] John McLean. Security models and information flow. In *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, pages 180–187, 1990.

[86] Frank McSherry. On "differential privacy as a mutual information constraint", Jan 2017. URL https://github.com/frankmcsherry/blog/blob/master/posts/2017-01-26.md.

[87] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.

[88] Christopher Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 411–418. Morgan Kaufmann Publishers Inc., 1995.

[89] Deirdre K Mulligan, Colin Koopman, and Nick Doty. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Phil. Trans. R. Soc. A*, 374(2083):20160118, 2016.

[90] Richard S Murphy. Property rights in personal information: An economic defense of privacy. *Geo. LJ*, 84:2381, 1995.

[91] Arvind Narayanan. When the business model *is* the privacy violation, Apr 2018. URL https://freedom-to-tinker.com/2018/04/12/when-the-business-model-is-the-privacy-violation/.

[92] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

[93] Arvind Narayanan, Vincent Toubiana, Solon Barocas, Helen Nissenbaum, and Dan Boneh. A critical look at decentralized personal data architectures. *arXiv preprint arXiv:1202.4503*, 2012.

[94] Michael Netter, Moritz Riesner, and Günther Pernul. Assisted social identity management-enhancing privacy in the social web. 2011.

[95] Noam Nisan et al. Introduction to mechanism design (for computer scientists). *Algorithmic game theory*, 9:209–242, 2007.

[96] Helen Nissenbaum. Protecting privacy in an information age: The problem of privacy in public. *Law and philosophy*, 17(5-6):559–596, 1998.

[97] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

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

[99] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[100] Helen Nissenbaum. Respecting context to protect privacy: Why meaning matters. *Science and engineering ethics*, pages 1–22, 2015.

[101] Geoffrey Nunberg. Farewell to the information age. *The future of the book*, pages 103–138, 1996.

[102] Inah Omoronyia, Liliana Pasquale, Mazeiar Salehie, Luca Cavallaro, Gavin Doherty, and Bashar Nuseibeh. Caprice: a tool for engineering adaptive privacy. In *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pages 354–357. ACM, 2012.

[103] Inah Omoronyia, Luca Cavallaro, Mazeiar Salehie, Liliana Pasquale, and Bashar Nuseibeh. Engineering adaptive privacy: on the role of privacy awareness requirements. In *Proceedings of the 2013 International Conference on Software Engineering*, pages 632–641. IEEE Press, 2013.

[104] PCI Security Standards Council. Who has to comply with the pci standards? URL https://pcissc.secure.force.com/faq/articles/Frequently_Asked_Question/Who-has-to-comply-with-the-PCI-standards/. n.d. Web. 15 May 2017.

[105] PCI Security Standards Council. Payment card industry (pci) data security standard, 2016. v3.2.

[106] Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. 1988.

[107] Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[108] Judea Pearl. *Causality*. Cambridge university press, 2009.

[109] Richard A Posner. The economics of privacy. *The American economic review*, 71 (2):405–409, 1981.

[110] Robert Post. Data privacy and dignitary privacy: Google spain, the right to be forgotten, and the construction of the public sphere. *Duke Law Journal*, 2017.

[111] General Data Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Official Journal of the European Union (OJ)*, 59:1–88, 2016.

[112] Joel R Reidenberg, N Cameron Russell, Alexander J Callen, Sophia Qasir, and Thomas B Norton. Privacy harms and the effectiveness of the notice and choice framework. *ISJLP*, 11:485, 2015.

[113] Neil M Richards and Daniel J Solove. Privacy's other path: recovering the law of confidentiality. *Geo. LJ*, 96:123, 2007.

[114] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[115] Ira S Rubinstein. Privacy and regulatory innovation: Moving beyond voluntary codes. *ISJLP*, 6:355, 2010.

[116] Andrei Sabelfeld and Andrew C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.

[117] Mazeiar Salehie, Liliana Pasquale, Inah Omoronyia, and Bashar Nuseibeh. Adaptive security and privacy in smart grids: A software engineering vision. In *Proceedings of the First International Workshop on Software Engineering Challenges for the Smart Grid*, pages 46–49. IEEE Press, 2012.

[118] Reza Samavi and Mariano P Consens. L2tap+ scip: An audit-based privacy framework leveraging linked data. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on*, pages 719–726. IEEE, 2012.

[119] Pamela Samuelson. Privacy as intellectual property? *Stanford Law Review*, pages 1125–1173, 2000.

[120] Rula Sayaf, Dave Clarke, and Richard Harper. \mathrm {CPS}ˆ 2: A contextual privacy framework for social software. In *International Conference on Security and Privacy in Communication Systems*, pages 25–32. Springer, 2014.

[121] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, 2015.

[122] Shayak Sen, Saikat Guha, Anupam Datta, Sriram K. Rajamani, Janice Tsai, and Jeannette M. Wing. Bootstrapping privacy compliance in big data systems. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*, SP '14, pages 327–342. IEEE Computer Society, 2014. ISBN 978-1-4799-4686-0. doi: 10.1109/SP.2014.28.

[123] Shayak Sen, Saikat Guha, Anupam Datta, Sriram K Rajamani, Janice Tsai, and Jeannette M Wing. Bootstrapping privacy compliance in big data systems. In *Security and Privacy (SP), 2014 IEEE Symposium on*, pages 327–342. IEEE, 2014.

[124] CE Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[125] Carl Shapiro and Hal R Varian. *Information rules: a strategic guide to the network economy*. Harvard Business Press, 1998.

[126] Kim Bartel Sheehan. Toward a typology of internet users and online privacy concerns. *The Information Society*, 18(1):21–32, 2002.

[127] Fuming Shih and Mi Zhang. Towards supporting contextual privacy in body sensor networks for health monitoring service. In *W3C Workshop on Privacy and data usage control*, volume 4, 2010.

[128] Fuming Shih, Ilaria Liccardi, and Daniel Weitzner. Privacy tipping points in smartphones privacy preferences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 807–816. ACM, 2015.

[129] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. Learning privacy expectations by crowdsourcing contextual informational norms. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP'16)*, 2016.

[130] Yan Shvartzshnaider, Zvonimir Pavlinovic, Thomas Wies, Lakshminarayanan Subramanian, Prateek Mittal, and Helen Nissenbaum. The vaccine framework for building dlp systems. *arXiv preprint arXiv:1711.02742*, 2017.

[131] Steven Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.

[132] Steven A Sloman, Philip M Fernbach, and Scott Ewing. Causal models: The representational infrastructure for moral judgment. *Psychology of Learning and Motivation*, 50:1–26, 2009.

[133] Geoffrey Smith. Recent developments in quantitative information flow (invited tutorial). In *Proceedings of the 2015 30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, LICS '15, pages 23–31. IEEE Computer Society, 2015. ISBN 978-1-4799-8875-4. doi: 10.1109/LICS.2015.13.

[134] Daniel J Solove. A taxonomy of privacy. *U. Pa. L. Rev.*, 154:477, 2005.

[135] Daniel J Solove. Introduction: Privacy self-management and the consent dilemma. *Harv. L. Rev.*, 126:1880, 2012.

[136] P. Swire and A. Anton. Engineers and lawyers in privacy protection: Can we all just get along? IAPP Privacy Perspectives, Jan 2014.

[137] Matt Tierney and Lakshminarayanan Subramanian. Realizing privacy by definition in social networks. In *Proceedings of 5th Asia-Pacific Workshop on Systems*, page 6. ACM, 2014.

[138] Michael Carl Tschantz, Anupam Datta, and Jeannette M Wing. Formalizing and enforcing purpose restrictions in privacy policies. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 176–190. IEEE, 2012.

[139] Michael Carl Tschantz, Anupam Datta, and Jeannette M. Wing. Purpose restrictions on information use. In *Proceedings of the 18th European Symposium on Research in Computer Security (ESORICS)*, volume 8134 of *Lecture Notes in Computer Science*, pages 610–627. Springer Berlin Heidelberg, 2013.

[140] Michael Carl Tschantz, Amit Datta, Anupam Datta, and Jeannette M. Wing. A methodology for information flow experiments. In *Computer Security Foundations Symposium*. IEEE, 2015.

[141] Hal R Varian. Economics of information technology. *University of California, Berkeley*, 2001.

[142] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. 2017.

[143] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android permissions remystified: A field study on contextual integrity. In *USENIX Security Symposium*, pages 499–514, 2015.

[144] David H Wolpert. Physical limits of inference. *Physica D: Nonlinear Phenomena*, 237(9):1257–1281, 2008.

[145] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[146] Frances Zhang, Fuming Shih, and Daniel Weitzner. No surprises: measuring intrusiveness of smartphone applications by detecting objective context deviations. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 291–296. ACM, 2013.

Appendix A

Research Template

The literature survey used for the study documented in Chapter 2 used the following list of prompts as a research instrument. Readers independently recorded answers to the prompts for each of the papers in the study before comparing results.

1. Provide a short summary of the objectives of the paper.
2. What subfield are the authors situated in?
3. What are the technical elements of the framework the authors are proposing? (Technique, system, model, mechanism, tool, or platform.)
4. What problem are they solving?
5. Do they explicitly address context?
6. What parameters are recognized?
7. Are further parameters introduced?
8. How CI is challenged or extended?

Appendix B

Information theory theorems

This appendix contains proofs for several theorems extending the well-known Data Processing Inequality in information theory to configurations of random variables beyond a triplet Markov Chain. The motivation for these theorems is the desire to model information flow through a world modeled as a Bayesian network, where information flow is determine by causal flows and nomic associations. Nomic association here is measured as mutual information between two variables. The Chain rule for mutual information and the Markov properties of a Bayesian network make it possible to prove several theorems that are as far as we know new.

B.1 Triplet Structures

The Data Processing Inequality is a standard theorem in information theory. It concerns the mutual information of three variables arranged in a Markov Chain.

Definition 56. (Cover and Thomas [28]) Random variables X, Y, Z are said to *form a Markov chain in that order* (denoted $X \rightarrow Y \rightarrow Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, Z form a Markov chain $X \rightarrow Y \rightarrow Z$ if the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y) \tag{B.1}$$

Theorem 57 (Data Processing Inequality)**.** Given a probability model defined by the following (Markov Chain):

$$\begin{tikzcd} X \arrow[r] & Y \arrow[r] & Z \end{tikzcd}$$

where $X \perp\!\!\!\perp Z|Y$, then it must be that $I(X;Y) \geq I(X;Z)$.

Proof. From [28]. By the Chain Rule, mutual information can be expanded in two different ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned} \tag{B.2}$$

Since X and Z are conditionally independent given Y, we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$, we have

$$I(X;Y) \geq I(X;Z) \tag{B.3}$$

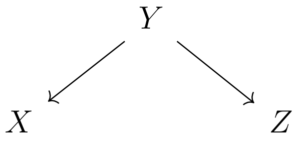
We have equality if and only if $I(X;Y|Z) = 0$ (i.e. $X \to Z \to Y$ forms a Markov Chain). Similarly, one can prove that $I(Y;Z) \geq I(X;Z)$. □

Bayesian networks are a generalization of Markov chains. A Bayesian network models the probability distribution between many random variables as a directed acyclic graph. The conditional probability distribution of each variable is defined in terms of the graphical parents $pa()$ of each variable, i.e. $P(X_i) = P(X_i|pa(X_u))$. The joint distribution is

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i)) \tag{B.4}$$

We can now prove several theorems that are similar to the Data Processing Inequality but for other probabilistic structures besides Markov chains.

Theorem 58 (Data Sourcing Inequality)**.** Given a probability model defined by the following (Common Cause):



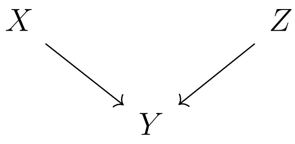
then it must be that $I(X;Y) \geq I(X;Z)$.

Proof. The implication of the common cause structure is that

$$p(x,y,z) = p(x|y)p(y)p(z|y). \tag{B.5}$$

It follows that $X \perp\!\!\!\perp Z|Y$. The rest of the proof is identifical to the previous proof. □

Theorem 59 (Unobserved common effect inequality)**.** Given variables X, Y, Z with the common cause structure



then it must be that $I(X,Y) \geq I(X,Z) = 0$.

Proof. The implication of the structure is that

$$p(x,y,z) = p(x)p(y|x,z)p(z). \tag{B.6}$$

It follows that $X \perp\!\!\!\perp Z$, therefore $I(X;Z) = 0$. Because the mutual information of two variables is always nonnegative,

$$I(X,Y) \geq I(X;Z)$$

□

We note that while a similar property holds for a "collider" or common effect structure, it's proof is different from the chain and common cause cases because, in general, it is not the case that $X \perp\!\!\!\perp Z|Y$ for a common effect structure. For example, when X and Z are both fair coin tosses and $Y = X \oplus Z$, X and Z are independent from each other but not when conditioned on Y.

When a common effect is in the conditioning set, the two causes depend probabilistically on each other. The extent to which these dependencies are limited can be characterized by a few equations.

Lemma 60. Given variables X_1, X_2, Y with the common effect structure $X_1 \to Y \leftarrow X_2$, then $I(X_1;X_2,Y) = I(X_1,Y|X_2)$.

Proof. By the Chain Rule for mutual information,

$$I(X_1;X_2,Y) = I(X_1;X_2) + I(X_1;Y|X_2)$$

Because of the common effect structure, $I(X_1;X_2) = 0$. Therefore, $I(X_1;X_2,Y) = I(X_1;Y|X_2)$. □

Lemma 61. Given variables X_1, X_2, Y with the common effect structure $X_1 \to Y \leftarrow X_2$, then

$$\begin{aligned} I(Y;X_1,X_2) &= I(X_1;X_2,Y) + I(X_2;Y) \\ &= I(X_2;X_1,Y) + I(X_1;Y) \end{aligned} \tag{B.7}$$

Proof.

$$\begin{aligned} &I(X_1;X_2,Y) \\ &= I(X_1;Y|X_2) \\ &= H(Y|X_2) - H(Y|X_1,X_2) \\ &= H(Y|X_2) - H(Y) + I(Y;X_1,X_2) \\ &= I(Y;X_1,X_2) - I(X_2;Y) \end{aligned} \tag{B.8}$$

which implies that

$$I(X_1;X_2,Y) + I(X_2;Y) = I(Y;X_1,X_2)$$

The proof works symmetrically for $I(X_2;X_1,Y) + I(X_1;Y) = I(Y;X_1,X_2)$ □

Lemma 62. Given variables X_1, X_2, Y with the common effect structure $X_1 \rightarrow Y \leftarrow X_2$, then $I(X_1, X_2|Y) \leq I(X_1; X_2, Y)$.

Proof.

$$
\begin{aligned}
& I(X_1, X_2|Y) \\
& = H(X_1|Y) - H(X_1|X_2, Y) \\
& = H(X_1) - I(X_1; Y) - H(X_1) + I(X_1; X_2, Y) \\
& = I(X_1; X_2, Y) - I(X_1; Y) \\
& \leq I(X_1; X_2, Y)
\end{aligned}
\tag{B.9}
$$

□

Theorem 63. Given variables X_1, X_2, Y with the common effect structure $X_1 \rightarrow Y \leftarrow X_2$, then

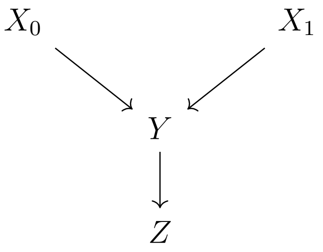
$$I(X_1, X_2; Y) \geq I(X_1; X_2, Y) = I(X_1; Y|X_2) \geq I(X_1; X_2|Y)$$

Proof. Follows from Lemmas 60, 61, and 62. □

B.2 Quartet Structures

While triplet structures (chain, common effect, and common cause) are the building blocks of larger paths in Bayesian networks, an analysis of larger, quarter structures will help us develop general theorems about the mutual information along paths.

Recall that a both with a common effect is not blocked if *either* the common effect *or* a descendant of the effect is in the conditioning set. Let's look at the following structure, which we will call a *wishbone* structure.



Here, Y is a common effect of X_0 and X_1, and Z is a descendant of Y. How much information flows from X_0 to X_1 when Z is known?

Theorem 64. For variables X_0, X_1, Y, Z in a wishbone structure,

$$I(X_0; X_1|Z) \leq I(Y; Z)$$

Proof. Consider the quantity $I(X_0, Y; X_1, Z)$, expanded by the Chain Rule. One expansion is:

$$\begin{aligned}
&I(X_0, Y; X_1, Z)\\
&= I(Y; Z) + I(X_0; Z|Y) + I(Y; X_1|Z) + I(X_0, X_1|Y, Z)\\
&= I(Y; Z) + I(X_0, Y; X_1)
\end{aligned} \tag{B.10}$$

Another expansion is:

$$\begin{aligned}
&I(X_0, Y; X_1, Z)\\
&= I(X_0; Z) + I(Y; X_1|X_0) + I(X_0; X_1|Z) + I(Y; X_1|X_0, Z)\\
&\geq I(Y; X_1|X_0) + I(X_0; X_1|Z)
\end{aligned} \tag{B.11}$$

By Theorem 63, we know that $I(Y; X_1|X_0) = I(X_0, Y; X_1)$ for three variables in a common effect structure, as they are for these variables in the wishbone structure.

So we can set the two expansions equal to each other and reduce:

$$\begin{aligned}
I(Y; X_1|X_0) + I(X_0; X_1|Z) &\leq I(Y; Z) + I(X_0, Y; X_1)\\
I(X_0, Y; X_1) + I(X_0; X_1|Z) &\leq I(Y; Z) + I(X_0, Y; X_1)\\
I(X_0; X_1|Z) &\leq I(Y; Z)
\end{aligned} \tag{B.12}$$

□

B.3 Paths

We can now look at mutual information of nodes connected by longer paths. We start with an arbitrariliy long Markov chain.

$$X_0 \longrightarrow X_1 \longrightarrow \ldots \longrightarrow X_{n-1} \longrightarrow X_n$$

Theorem 65 (Chain Data Processing Inequality)**.** Given a Markov chain of variables

$$X_1, ..., X_n$$

such that $X_1 \rightarrow \ldots \rightarrow X_n$. It must be the case that

$$I(X_1, X_n) \leq \min_i I(X_i, X_{i+1}).$$

Proof. For all i, by the Chain rule for mutual information and the independence properties of the Markov chain,

$$\begin{aligned}
I(X_0, ..., X_i; X_{i+1}, ..., X_n) &= \\
\sum_{j=i+1}^{n} I(X_0, ..., X_i; X_j | X_{i+1}, ..., X_j) &= \\
I(X_0, ..., X_i; X_{i+1}) &= \\
\sum_{j=i}^{1} I(X_{i+1}; X_j | X_i, ...X_j) &= \\
I(X_i; X_{i+1}) + \sum_{j=i-1}^{1} I(X_{i+1}; X_j | X_i, ...X_j) &= \\
I(X_i; X_{i+1}) &
\end{aligned} \tag{B.13}$$

The Chain rule can expand the variables in arbitary order. So we can also derive (using the fact that mutual information is always nonnegative):

$$\begin{aligned}
&I(X_0, ..., X_i; X_{i+1}, ..., X_n) \\
&= I(X_0, .., X_i; X_n) + \sum_{j=n-1}^{i+1} I(X_0.., X_i; X_j | X_{j+1}.., X_j) \\
&\geq I(X_0, ..., X_i; X_n) \\
&= \sum_{j=0}^{n-1} I(X_n; X_j | X_{j-1}, ..., X_0) \\
&= I(X_n; X_0) + \sum_{j=1}^{n-1} I(X_n; X_j | X_{j-1}, ..., X_0) \\
&\geq I(X_n; X_0)
\end{aligned} \tag{B.14}$$

Combining these two results and generalizing across all i,

$$\forall i, I(X_0; X_n) \leq I(X_i, X_{i+1}) \tag{B.15}$$

which entails that which is to be proven,

$$I(X_0; X_n) \leq \min_i I(X_i, X_{i+1}) \tag{B.16}$$

□

Our goal is to generalize this theorem to Bayesian paths with other structures, just found as in the previous section we found equivalents to the Data Processing Inequality in other triplet structures.

Definition 66 (Path)**.** A *path* between two nodes X_1 and X_2 in a graph to be a sequence of nodes starting with X_1 and ending with X_2 such that successive nodes are connected by an edge (traversing in either direction).

In this section, we will only consider paths isolated from any other variables. We are interested in how to derive useful bounds on the mutual information of a path based on the mutual information of links within the path.

Definition 67 (Mutual information of a path)**.** The *mutual information of a path* between two nodes X and Y is $I(X, Y)$.

Theorem 68 (Unobserved Path Data Processing Inequality)**.** Given a path between X_0 and X_n of variables $X_0, ..., X_n$, with no other connected variables. It must be the case that

$$I(X_1, X_n) \leq \min_i I(X_i, X_{i+1}).$$

Proof. This proof mirrors the proof of Theorem 65.
For any i, consider $I(X_0, ...X_i; X_{i+1}, ..., X_n)$.
By the logic of Equation B.13, $I(X_0, ...X_i; X_{i+1}, ..., X_n) = I(X_i, X_{i+1})$.
By the logic of Equation B.14, $I(X_0, ...X_i; X_{i+1}, ..., X_n) \geq I(X_0, X_n)$.
Therefore, $\forall i, I(X_0; X_n) \leq I(X_i, X_{i+1})$ and $I(X_0; X_n) \leq \min_i I(X_i, X_{i+1})$. □

Theorem 68 applies to any paths on the condition that none of the variables are observed. Its proof is identical to the proof for Markov chains because isolated, unobserved paths are Markov equivalent to Markov chains.

Some proofs extending this result follow from theory of Bayesian networks. Recall that there are two conditions under which a path betweent two variables is blocked. First, an unobserved head-to-head connection on the path blocks the path and makes the terminal nodes conditionally independent. Second, an observation of a head-to-tail or tail-to-tail node blocks the path and makes the terminal nodes conditionally independent. If the only paths between two variables are blocked, then they are d-separated and therefore independent, with zero mutual information.

Theorem 69 (Blocked Path Mutual Information)**.** For any blocked paths between X_0 and X_n of variables $X_0, ..., X_n$ with no other connected variables, $I(X_0, X_n) = 0$.

Proof. If the only path between X_0 and X_n is blocked, then X_0 and X_n are d-separated and conditionally independent. If X_0 and X_n are conditionally independent, then $I(X_0, X_n) = 0$. □

The difficult case for determining the mutual information of a path is the case where there are observed common effects on the paths. This breaks the conditions for the proof of Theorem 68. It is possible for $I(X_i, X_{i+1}) = 0$ but $I(X_{i-1}, X_{i+1} | X_i) > 0$. As a simple example, consider again the case where X_{i-1} and X_{i+1} are fair coin tosses and $X_i = X_{i-1} \oplus X_{x+1}$.

If there are many common effect nodes on the path and only some of them are observed, then the path is blocked and the mutual information is solved using Theorem 69; the mutual information of the path is zero. Similarly, if there are common cause or chain triplets on the path and the central node of the triplet is observed, the mutual information of the path is trivially ze So we need consider only the case where there's a path where *all and only* the common effect nodes are observed.

Theorem 70 (Path Mutual Information Theorem (PMIT))**.** Given a path between X_0 and X_n of variables $\{X_0, ..., X_n\} = \mathcal{X}$ with no other connected variables. Let $\mathcal{X}_E$ be the common effect nodes, meaning only those nodes X_i such that the edge structure of the path is $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$. The mutual information of the path when all the common effects are observed is is:

$$I(X_0; X_n|\mathcal{X}_E) \leq min_i \begin{cases} I(X_i, X_{i+1}) & \text{if} X_i, X_{i+1} \notin \mathcal{X}_E \\ I(X_{i-1}, X_{i+1}) & \text{if} X_i \in \mathcal{X}_E \end{cases}$$

Proof. For any i, consider

$$I(X_0, ..., X_i; X_{i+1}, ..., X_n|\mathcal{X}_E)$$

By the Chain Rule for mutual information, this can be expanded as

$$\sum_{j=0}^{i} I(X_j; X_{i+1}, ..., X_n|X_0, ..., X_{j-1}, \mathcal{X}_E)$$

Consider two cases.

In the first case, $X_i \notin \mathcal{X}_E$ and $X_{i+1} \notin \mathcal{X}_E$.

By logic similar to Equation B.13 and Equation B.14, then as before $I(X_0; X_n) \leq I(X_i, X_{i+1})$.

In the second case, X_i is a common effect node, i.e $X_i \in \mathcal{X}_E$. It is not possible to have two common efffect nodes adjacent on a path. So in any case where either X_{i-1} or X_{i+1} is in the conditioning set, the path is blocked. We can therefore compute the mutual information and its Chain Rule expansion as:

$$\begin{aligned} & I(X_0, ..., X_{i-1}; X_{i+1}, ..., X_n|\mathcal{X}_E) \\ = & \sum_{j=i-1}^{0} I(X_j; X_{i+1}, ..., X_n|X_{i-1}, ..., X_{j-1}, \mathcal{X}_E) \\ & = I(X_{i-1}; X_{i+1}, ..., X_n|\mathcal{X}_E) \\ & = I(X_{i-1}; X_{i+1}|\mathcal{X}_E) \\ & = I(X_{i-1}; X_{i+1}|X + i) \end{aligned} \tag{B.17}$$

Since once again by the logic of Equation B.14 this value is greater than or equal to the mutual information of the path, we have

$$I(X_0; X_n) \leq I(X_{i-1}, X_{i+1} | X_i)$$

for the cases when $X_i \in \mathcal{X}_E$.

Combining these results, we get the bound on the mutual information of the path. □

Note that Theorem 68 is a special case of PMIT, or Theorem 70, where the set of common effects on the path $\mathcal{X}_E$ is empty.

Appendix C

Multi-Agent Influence Diagrams (MAIDs) and Data Games

This appendix contains formal specifics of Multi-Agent Influence Diagrams and data games.

C.1 Multi-Agent Influence Diagrams

Multi-Agent Influence Diagrams (MAIDs) are a game-theoretic extension of Bayesian networks developed by Koller and Milch [76]. A MAID is defined by:

1. A set $\mathcal{A}$ of agents
2. A set $\mathcal{X}$ of chance variables
3. A set $\mathcal{D}_a$ of decision variables for each agent $a \in \mathcal{A}$, with $\mathcal{D} = \bigcup_{a \in \mathcal{A}} \mathcal{D}_a$
4. A set $\mathcal{U}_a$ of utility variables for each agent $a \in \mathcal{A}$, with $\mathcal{U} = \bigcup_{a \in \mathcal{A}} \mathcal{U}_a$
5. A directed acyclic graph $\mathcal{G}$ that defines the parent function Pa over $\mathcal{V} = \mathcal{X} \cup \mathcal{D} \cup U$
6. For each chance variable $X \in \mathcal{X}$, a CPD $Pr(X|Pa(X))$
7. For each utility variable $U \in \mathcal{U}$, a CPD $Pr(U|Pa(U))$

The decision variables represent moments where agents can make decisions about how to act given only the information provided by the variable's parents.

Definition 71 (Decision rules)**.** A *decision rule* δ is a function that maps each instantiation $\mathbf{pa}$ of $Pa(D)$ to a probability distribution over $dom(D)$.

Definition 72 (Strategy)**.** An assignment of decision rules to every decision $D \in \mathcal{D}_a$ for a particular agent $a \in \mathcal{D}_a$ for a particular agent $a \in \mathcal{A}$ is called a *strategy*.

Definition 73 (Strategy profile)**.** An assignment σ of decision rules to every decision $D \in \mathcal{D}$ is called a *strategy profile*. A *partial strategy profile* $\sigma_{\mathcal{E}}$ is an assignment of decision rules to a subset $\mathcal{E} \subset \mathcal{D}$. $\sigma_{-\mathcal{E}}$ refers to a restriction of σ to variables not in $\mathcal{E}$.

Decision rules are of the same form as CPDs, and so a MAID can be transformed into a Bayes network by replacing every decision variable with a random variable with the CPD of the decision rule of a strategy profile.

Definition 74. If $\mathcal{M}$ is a MAID and σ is a strategy profile for $\mathcal{M}$, then the *joint distribution for $\mathcal{M}$ induced by σ*, denoted $P_{\mathcal{M}[\sigma]}$, is the joint distribution over $\mathcal{V}$ defined by the Bayes net where:

- the set of variables is $\mathcal{V}$;
- for $X, Y \in \mathcal{V}$, there is an edge $X \to Y$ if and only if $X \in Pa(Y)$;
- for all $X \in \mathcal{X} \cup \mathcal{U}$, the CPD for X is $Pr(X)$;
- for all $D \in \mathcal{D}$, the CPD for D is $\sigma(D)$.

Definition 75. Let $\mathcal{E}$ be a subset of $\mathcal{D}_a$ and let σ be a strategy profile. We say that $\sigma *_{\mathcal{E}}$ is *optimal for the strategy profile* σ if, in the induced MAID $\mathcal{M}[\sigma_{-\mathcal{E}}]$, where the only remaining decisions are those in $\mathcal{E}$, the strategy $\sigma *_{\mathcal{E}}$ is optimal, i.e., for all strategies $\sigma'_{\mathcal{E}}$:

$$EU_a((\sigma_{-\mathcal{E}}, \sigma *_{\mathcal{E}})) \geq EU_a((\sigma_{\mathcal{E}}, \sigma'_{\mathcal{E}}))$$

A major contribution of Koller and Milch [76] is their analysis of how to efficiently discover Nash Equilibrium strategy profiles for MAIDs. Their method involves analyzing the qualitative graphical structure of the MAID to discover the *strategic reliance* of decision variables. When a decision variable D strategically relies on D', then in principle the choice of the optimal decisionr rule for D depends on the choice of the decision rule for D'.

Definition 76 (Strategic reliance)**.** Let D and D' be decision nodes in a MAID $\mathcal{M}$. D *strategically relies on* D' if there exist two strategy profiles σ and σ' and a decision rule δ for D such that:

- δ is optimal for σ;
- σ' differs from σ only at D';

but no decision rule $\delta*$ that agrees with δ on all parent instantiations $\mathbf{pa} \in dom(Pa(D))$ where $P_{\mathcal{M}[\sigma]}(\mathbf{pa}) > 0$ is optimal for σ'.

Definition 77 (s-reachable)**.** A node D' is *s-reachable* from a node D in a MAID $\mathcal{M}$ if there is some utility node $U \in \mathcal{U}_D$ such that if a new parent $\widehat{D'}$ were added to D', there would be an active path in $\mathcal{M}$ from $\widehat{D'}$ to U given $Pa(D) \cup \{D\}$, where a path is active in a MAID if it is active in the same graph, viewed as a BN.

Theorem 78. If D and D' are two decision nodes in a MAID $\mathcal{M}$ and D' is not s-reachable from D in $\mathcal{M}$, then D does not strategically rely on D'.

C.1.1 Tactical independence

This dissertation introduces a new concept related to Multi-Agent Influence Diagrams: tactical independence.

Definition 79 (Tactical independence)**.** For decision variables D and D' in MAID $\mathcal{M}$, D and D' are *tactically independent* for conditioning set $\mathcal{C}$ iff for all strategy profiles σ on $\mathcal{M}$, in $P_{\mathcal{M}[\sigma]}$, the joint distribution for $\mathcal{M}$ induced by σ,

$$D \perp\!\!\!\perp D' | C$$

Because tactical independence depends on the independence of variables on an induced probability distribution that is representable by a Bayesian network, the d-separation tests for independence apply readily.

Theorem 80. For decision variables D and D' in MAID $\mathcal{M}$, and for conditioning set $\mathcal{C}$, if D and D' are d-separated given $\mathcal{C}$ on $\mathcal{M}$ considered as a Bayesian network, then D and D' are tactically independent given $\mathcal{C}$.

Proof. Suppose D and D' are d-separated given $\mathcal{C}$ on $\mathcal{M}$ considered as a Bayesian network.

For any strategy profile σ, the joint distribution for $\mathcal{M}$ induced by σ, $P_{\mathcal{M}[\sigma]}$ has the same graphical structure as $\mathcal{M}$ considered as a Bayesian network.

Therefore, D and D' are d-separated given $\mathcal{C}$ in the graph corresponding to $P_{\mathcal{M}[\sigma]}$ for all σ.

Because D and D' are d-separated given $\mathcal{C}$ in the Bayesian network, $D \perp\!\!\!\perp D'|C$. □

C.1.2 Notation

We will use a slightly different graphical notation than that used by Koller and Milch [76].

In the models in this paper, we will denote random variables with undecorated capital letters, e.g. A, B, C. I will denote strategic nodes with a tilde over a capital letter, e.g. $\tilde{A}, \tilde{B}, \tilde{C}$. The random variable defined by the optimal strategy at a decision node, when such a variable is well-defined, will be denoted with a hat, e.g. $\hat{A}, \hat{B}, \hat{C}$. Nodes that represent the payoff or utility to an agent will be denoted with a breve, e.g. $\breve{A}, \breve{B}, \breve{C}$. Particular agents will be identified by a lower case letter and the assignment of strategic and utility nodes to them will be denoted by subscript. E.g., $\tilde{A}_q$ and $\breve{U}_q$ denote an action taken by agent q and a payoff awarded to q, respectively.

C.2 Data Games

What distinguishes a data game from a MAID is the use of optional arrows to support mechanism design. A dotted arrow in a data game an optional arrow. The diagram defines two separate models, one including the arrow and one without. When considering an instantiation of the model with the dotted edge present, we will say the model or edge is *open*. When the edge is absent, we'll say it's *closed*.

As we have distinguished between strategic reliance and tactical independence, we can distinguish between the strategic and tactical value of information.

The strategic value of an information flow to an agent is the difference in utility to that agent in the open and closed conditions of the game, given each game is at strategic equilibrium for all players.

Definition 81 (Strategic value of information)**.** Given two MAID diagrams $\mathcal{M}_o$ and $\mathcal{M}_c$ that differ only by a single edge, e, and a strategic profile solution for each diagram, $\hat{\sigma}_o$ and $\hat{\sigma}_c$, the *strategic value of* e *to* a is the difference in expected utility to a under the two respective induced joint distributions:

$$E(P_{\mathcal{M}_o[\hat{\sigma}_o]}(U_a)) - E(P_{\mathcal{M}_c[\hat{\sigma}_c]}(U_a))$$

Definition 81 is an incomplete definition because it leaves open what *solution concept* is used to determine the strategic profile solutions. For the purpose of the results in this paper, we use Nash Equilibrium as the solution concept for determining strategic value of information.

In contrast with the strategic value of information, the tactical value of information is the value of the information to an agent given an otherwise fixed strategy profile. We allow the agent receiving the data to make a tactical adjustment to their strategy at the decision variable at the head of the new information flow.

Definition 82 (Best tactical response to information)**.** Given two MAID diagrams $\mathcal{M}_o$ and $\mathcal{M}_c$ differing only in optional edge e with head in decision variable D_a, the *best tactical response to* e given strategy profile solution $\hat{\sigma}$, $\hat{\delta}_{\sigma,e}$ is the decision rule δ for D such that δ is optimal for $\hat{\sigma}$ for player a.

Definition 83 (Tactical value of information)**.** Given two MAID diagrams $\mathcal{M}_o$ and $\mathcal{M}_c$ differing only in optional edge e with head in decision variable D, the *tactical value of* e to agent a given strategy profile solution $\hat{\sigma}$ is the difference in expected utility of the open condition with the best tactical response to e and the closed condition using the original strategy:

$$EU_a((\hat{\sigma}_{-D}, \hat{\delta}_{\hat{\sigma},e}) - EU_a(\hat{\sigma})$$

Note that the uniqueness of a best tactical response has not yet been proven. However, if the best tactical response is not unique, then the tactical value of the information will be the same for any best tactical response. This definition, like Definition 81, depends on an implicit solution concept.