Experiments and formal methods for privacy research

Sören Preibusch
University of Cambridge, Computer Laboratory
Cambridge CB3 0FD
Soeren.Preibusch@cl.cam.ac.uk, privacy-calculus.net

ABSTRACT

In the light of reportedly rising concerns about privacy and their inhibition to electronic business growth, companies could market efficient privacy as a win-win situation. But our understanding into consumers' privacy attitudes and subsequently into their privacy-related decision-making online remains limited. As academics, we have a hard time recommending concrete measures for effectively improving data protection on the Web.

Partly, this lack of insight can be attributed to a scarcity of valid data to begin with. Partly, we are lacking adequate methods to interpret existing empirical evidence and to engineer superior data protection into data processing.

This manifesto is neither about providing a definite answer to what privacy is; nor is it yet another bloodless framework. It is a call to exploit the synergies from conducting experimental research into privacy and from applying formal methods of computing to privacy artefacts, for the benefit of research, businesses, and online users.

1. PROBLEM STATEMENT

Despite a reported resurgence in public concern about privacy, we are still lacking decisive insights into whether or which consumers care about privacy, and if they do, how to quantify their privacy concerns—with a price tag or otherwise. Research into improving privacy practices online exhibits two problems: first, throughout the stakeholders, a reliance on stated practices instead of observed behaviour; and second, an inability to reason about privacy choices and eventually support them on an engineering level. We need advancing and integrating experimental and formal methods to understand and to support consumers' privacy decision-making.

On the demand side, divergence between self-reported and actual practice has been documented for online users [18, 9]. Consumers take less protective action then they claim, and their decision-making deviates from articulated preferences. The latter are also shaped by more or less accurate media coverage, resulting in paranoia about data protection [6]—vague fears rarely result in firm action.

On the supply side, online offerings are embellished by unsubstantiated privacy claims such as privacy seals or textual advertising of their privacy practices. For instance, online social network operators that promote their privacy practices actually implement privacy procedures that are significantly inferior compared to their competitors [5]. With few exceptions, such as the European Privacy Seal, certification of privacy procedures often relies on their descriptions rather than their actual implementation. A formal methods approach with code scrutiny could provide stronger technical guarantees and agile re-certification.

Opinion polling into privacy attitudes plus the naïve application of measurement instruments with limited validity result in detrimental misconceptions. These misconceptions, for instance, could falsely suggest a universally high level of privacy awareness, making further regulation unnecessary; companies could erroneously under- or over-estimate the business impact of restraining or widening their data collection.

I therefore argue for two changes: first, focus on better data and richer models thereof, second, formal underpinning to yield engineered compliance. The key to translating articulated privacy worries into more adequate policy-making, education, and marketing lies in applying experimental research rather than opinion-probing. Collected field evidence requires sophisticated models of privacy choice that cope with inconsistencies. Techniques to reason with and about policies, amongst which formal methods of computing, can provide guarantees that implementations are sound and enforce privacy agreements resulting from users' manifold decisions.

2. IMPLICATIONS FOR THE ECONOMICS OF PRIVACY

Consumers repeatedly state their adverse reaction towards excessive collection of personal information, with more than two thirds indicating they would cancel an online transaction and potentially switch to an alternative vendor. A customer-friendly privacy design is turned into a business advantage and empirical evidence for its success exists for monopolies [10], and competitive markets including online retailing and online social networking [5]. Especially for highly competitive markets, it seems straightforward to escape price competition by differentiating on privacy. When selling at higher prices but with an overall more privacy-friendly design, the latter becomes a quality parameter. Ultimately, if

submitted on 2010-07-08;
revised version on 2010-08-18:
Privacy and Usability Methods Pow-wow (PUMP) 2010

consumers abandoned online service providers with dissatisfying privacy practices or convinced other users to do so, superior practices would prevail in the market.

Figure 1 shows a typical Web order form. Potential buyers are asked to provide some personal details, most of which are mandatory. To mitigate concerns about revealing one's phone number, a brief privacy assurance has been attached to the input field, emphasising that customers will be contacted at most once per month. Note that such usage explanations are missing for other fields. The purpose of collecting email addresses may only be guessed, with the checkboxes in the lower part of the form providing some hints. It further remains unclear how a benevolent company would enforce the self-imposed limit so that accidentally texting users more often becomes impossible—an issue to be revisited in the following section. Finally, if they conducted a controlled study, the company might realise later that the well-intended privacy assurance did not calm down users but instead heightened their concerns.

The requirement to provide one's phone number may be a major deterrent to shop on the depicted Web site. Telephone numbers are amongst the data items participants in our experiments indicate to be least willing to provide. One consumer in three reports to have already provided a false phone number to evade a Web site's excessive data collection [4]. If the business value from contacting buyers on their mobile outweighs the expected loss in transactions, phone number verification seems worthwhile. Otherwise, making the phone number an optional data item, a pictographic privacy seal, or offering music and films at discount prices could be a more powerful tactic. Empirically challenging consumers' privacy decision-making in competitive markets indicates that a plurality of Web shoppers regularly subdue their privacy concerns to the promise of material gain [3]. The number of alternative privacy strategies and an apparent dichotomy between predicted and observed market equilibria calls for further investigation. The result may be that "it depends" what the most profitable privacy design is. (Spelling out on what it depends would actually be a huge step forward.)

Users' failure to take appropriate protective action could reflect technical inability, cognitive or psychological barriers rather than lack of interest, and mandates investigating usability issues and incentives. Visual cues, economic incentives, interface design, data flow guarantees, and default settings may be enablers or inhibitors for users to exercise their right to informational self-determination. Again, we require rigorous experimental research, in the laboratory and in the field, ideally sampling from a demographically varied population.

3. DIVERSITY IN PRIVACY ATTITUDES

The challenge in turning privacy into a competitive advantage lies in the heterogeneity of consumers' privacy preferences that govern what is subjectively perceived as superior. As a result, it is difficult to reduce the dimensionality of privacy preferences into a tractable privacy typology—i.e., target groups or market segments in business parlance. The concept of medium or average privacy concerns is misleading and a "pragmatic majority" provides little help for corporate planning. Even socio-demographically homogeneous consumers exhibit strongly varying willingness to provide items of personal information, the most tangible and fundamental dimension of information privacy [15]. We are lacking

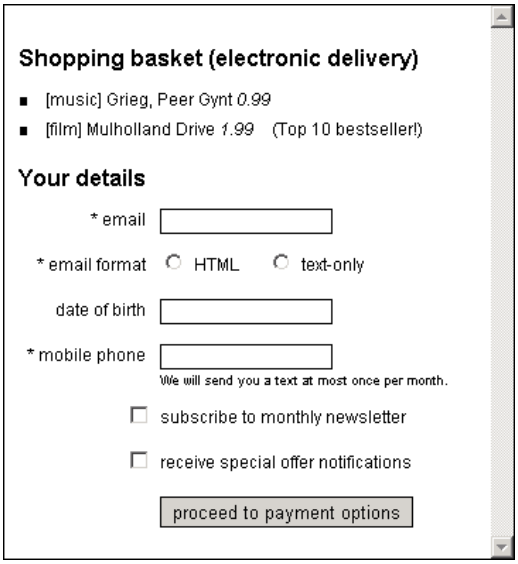

Figure 1: Idealised screenshot of a typical order form encountered during online shopping.

valid instruments to measure privacy concerns. We are also lacking sophisticated methods to cluster or otherwise structure a consumer population along privacy dimensions.

Few scales to measure privacy concerns in a valid manner exist [17, 12], with attempts to extend their applicability to an Internet economy [7]. This causes difficulties in discovering consistent and stable patterns in privacy concerns and actions, over time and across individuals. We are short on formal and statistical methods to reason about privacy preferences, their aggregates and dynamics, complicates clustering consumers. When a company is unable to sense cluster membership for a new customer, it can neither greet him with a smart privacy default to improve usability and conversion rates, nor achieve derivative marketing endeavours such as price discrimination based on privacy-related behaviour. The resulting quasi-absence of privacy default-sets to facilitate configuration of privacy options is deplorable.

Yet, even the most sophisticated menu of privacy choices has to cater for individual deviations. Regulators and companies need to embrace the complexity in customers' privacy preferences to achieve mutually beneficial deployments of privacy enhancing technologies.

Longitudinal studies, to follow the evolution of online users' privacy preferences and their manifestations, are overdue. (The study on "taste for privacy" [11] is at least a start.)

4. ENCODING AND PROGRAMMING WITH PRIVACY CHOICES

Privacy preferences eventually materialise in the acceptance of some privacy policy. Research into the latter has focused on two aspects: encoding privacy choices in policies and enforcing them programmatically.

On the one hand, there are techniques for machine-readable representation, encoding, and potentially inter-organisational circulation of often convoluted privacy policies. Examples include P3P and EPAL, targeted at company-to-consumer communication and cross-organisational data exchange respectively. Both have failed to gain momentum in the mar-

ket. More recent endeavours such as XACML and its privacy profile may prove more successful.

On the other hand, we see programming languages with built-in features for privacy-aware coding and support for automated reasoning over data flows. Existing mainstream languages have been augmented with data protection features such as annotating variables and control flow constructs; alternative languages have been designed from scratch. Examples include JIF and DEFCon, augmenting Java with information flow control and policy-aware message passing, respectively with mostly static analysis and runtime enforcement of privacy requirements. Outside production environments, new programming languages (e.g. AURA) anchor privacy enforcement as low as in a trusted computing base that guards access to storage/computing resources and peripherals.

Still, insofar as these programming languages pursue security goals rather than privacy goals, their applicability remains an open question: different from data security, where information can be classified as having a high or low confidentiality level by itself, usage restrictions in privacy policies are context-dependent. The decision whether a data item may be used, whether the value of a variable may be accessed, depends on the how it is used or accessed.

Mapping privacy requirements to variables can be tricky. Returning to the example of Figure 1, the customer's email address may certainly be used to send an order confirmation, but including it in regular newsletter mailings is subject to individual consent. Different policies apply to the same data item, depending on its usage. Similarly, a single privacy requirement may span across data items: the Web shop may notify its customers of discounts on their birthday. This practice is again subject to consent (the second tickbox in the form), and requires two data items, one of which (date of birth) is optional and may not be available for all customers—orthogonally to their present or missing consent. Whilst a programming language such as JIF [13] comes with tight compiler checks to prevent the programmer from accidentally leaking information once the original input is correctly annotated with privacy policies ("labels"), information flow control without declassification will also prevent such seemingly innocuous features as a bestseller list.

Privacy-aware database retrieval is also aimed at run-time enforcement of privacy policies. It advocates on-the-fly enforcement with statistically guaranteed bounds on information disclosure. Their focus is a population of users whose data is collectively subjected to an exogenously provided information release policy.

P3P-aware database management systems, or the representation of P3P-style policies in the JIF programming language [8] are proof-of-concept works on how to bridge between these two strands of research of encoding and enforcing.

5. ENFORCING DIVERSE, DYNAMIC POLICIES

In privacy negotiations, consumers and service providers establish, maintain, and refine privacy policies as individualised agreements through the ongoing choice amongst service alternatives [14]. Static privacy designs of products and services become flexible as mandatory data input fields are made optional, data items may be substituted by alternatives, and consumers overall gain more influence over the handling of their data. Dynamics are introduced as once-established privacy agreements can be amended subsequently. Albeit fiddly to implement in the back-end, modifiability in privacy choices can be exposed through a simple interface such as opt-out links in a newsletter, activation of personalised product recommendations or publication of a personal profile page with a single button click.

A flexible privacy design materialises in individually and temporally varied privacy policies as consumers seize the opportunity to exercise choice in releasing data items, restricting recipients as well as primary and secondary uses of those, and capping retention time. Data and their accompanying multi-dimensional policies are joint inputs to data processes, following the sticky policy paradigm. Policies as runtime inputs are a challenge for static analysis of privacy requirements. Adherence goes beyond a single policy which would be known a priori.

As a result of privacy negotiations, combinations of data items agglomerate to amorphous data records. Even similarly filled data records may be governed by different privacy policies. In database terms, privacy negotiations bring intensional and extensional heterogeneity. Functionally, software needs to cope with missing or expired data items, as consumers decide to leave blank input fields made optional or restrict data retention time. Non-functionally, the use of existing data items needs to respect purpose binding.

The resulting complexity may seem overwhelming. But it is not architecturally new: already today, Web sites are tailored to the individual in function and in presentation. Users of legacy browsers have a restricted user experience, users may switch between versions with or without Flash animations. Examples of persistent settings include the format of emails (also shown in Figure 1), or existing privacy configurability (think of Google Dashboard, to "view and manage all the data stored with [an] account", or the dozens of privacy settings on larger online social networking sites).

Unintended data leakage witnesses of the non-triviality in respecting users' choices. The propagation of complexity through the data tool-chain forbids manual inspection and requires mechanised software verification. Model checking is one of the technical approaches to evaluate software correctness for ranges of inputs, but it remains somewhat disconnected from mainstream program deployment.

Programming languages which support creating and manipulating policy requirements as first-class citizens treat the latter as inputs and not constants. Conflicts between runtime inputs (policies and data) need compile-time detection to avoid disruption. On a functional level, conditional program flow implements policy-specific data handling; in practice, and to circumvent exhaustive enumeration, it requires inequality comparators over privacy policies.

The programming language JIF supports policies noted as literals in the source code and their creation at runtime, plus some tests for ordering of policies (e.g. whether a given security label is less restrictive than another one). But despite runtime representation of security labels, programmers are confronted with an inability to serialise them, hindering the persistent storage of privacy requirements

Policy comparison traditionally assumes monotonous ordering and transitivity, or at least a universal mental model. It encodes rational decision making of economic agents—both to be challenged empirically.

6. LIMITS IN ENFORCEMENT

Universal compliance with freely negotiated privacy policies gets intricate as the degrees of freedom proliferate. Programming costs for writing case-by-case handling grow unless clever pre-processing is introduced as a further level of abstraction. Regardless of a potentially alleviated task of code-production, proof power is another limiting factor.

Run-time constructed privacy policies may be found to be incompatible with the data processing procedures; other policies may be detected as semantically unsound or self-contradictory in the light of external knowledge [2]. When excluding these error cases by design, feedback through an empirically established multidimensional density function over privacy preferences allows marketers and consumer associations to quantify how many consumers will be dissatisfied after the negotiation space is restricted.

Language-based enforcement also imposes limits on the ability to strengthen or weaken privacy policies retrospectively. To some extent, it requires attaching a new policy to an existing data item. Integrity requirements may prohibit the economically promising renegotiation of privacy policies, as the service provider and its customers build up a trust relationship and users become willing to extend their consent. Individuals could also be empowered to modify the coordinates of the privacy agreement by revoking previously granted usage rights, whilst companies may actively stimulate further information disclosure. Again, there is little longitudinal field evidence on the existence and prevalence of dynamics.

On the side of the service provider, strong enforcement may prevent disruptive changes to the Web site (and create the illusion of restricting functional innovativeness). Operators may be reluctant to grandfather old functionality and their users' corresponding policies, eventually forcing them to enrol in the new version. I do not argue it is technically infeasible, but I acknowledge the enterprise may seem too daring.

7. RECOMMENDATIONS, PERSPECTIVES

It is easy to blame consumers for "irrational" behaviour in their privacy decisions. It is more challenging to explore the motivations behind their decisions. The combined application of experimental investigation and formal calculus promises more insightful models and algorithms to understand and to support the empirically evident diversity in privacy preferences and decisions.

Achieving insightful experiments is laborious, time-consuming, and expensive. It requires rigorous design and deployment with elimination or control of potentially confounding factors. Moreover, simplistic designs outdo fancy manipulations despite their higher appeal. The inherent difficulties of empirical investigation are exacerbated in the area of privacy research by priming effects and social desirability bias.

General advice on conducting empirical research into privacy is beyond this paper. Validity and reliability are the hallmarks of empirical research; careful setup and ample pretesting are indispensable.

I would advocate three design decisions: first, we should look at competitive markets with two companies and a third option for participants not to do business with any of these, with drop-out as a dependent variable. Monopolistic and competitive markets behave differently. Users make relative judgements about privacy design. As a standalone proposition, the same privacy design is rated more positively than when there is a slightly better alternative. Participants' decisions when facing a single company are hardly a valid predictor for online users' behaviour in competitive markets. It also seems advantageous to start with two alternative companies rather than with three or more: even irrelevant alternatives may psychologically distort decision-making. This is not to say that single-company investigations are worthless. They are well-suited for manipulations in user-interfaces or when consumers face a quasi-monopoly.

Second, as we are still at the beginning, we should focus on one dimension of privacy at a time, such as the extent and sensitivity of collected data items or their retention period or the purposes for which personal information will be used. Given consumers' heterogeneity in privacy attitudes, bundling privacy dimensions or collapsing them into a one-dimensional privacy rating provides little guidance into why consumers prefer one privacy design over another.

Third, we should strive for longitudinal studies of privacy preferences and actions. They can make a strong case for causality in consumers' choices and level out spurious effects such as the influence of media coverage. The dynamics in privacy decision-making are under-explored.

Ironically, empirical studies into privacy attitudes accumulate sensitive information themselves. Advice is available on how to conduct ethical research online [1]; researchers may be required by law or by departmental guidelines to have their research design approved by an ethics committee / institutional review board. In my experience, such a review is not a dispensable burden but an opportunity to improve the intended study.

Insofar as consumers are guided by privacy assurances, these should be substantiated. If service providers are unable to provide proof of their data protection endeavours, customers may be unconvinced [16]. Mechanised analysis provides the required guarantees but has yet to reach maturity and to be useful beyond a few prototypical demonstrations. Good documentation is a key issue in facilitating adoption. When enforcing privacy policies, languages to encode them must be equipped with clear and interoperable semantics. User-defined, sticky privacy policies call for operators over policies homomorphic to data operators, and notions of policy refinement, ordering, and difference quantification: string concatenation is policy composition. Evidence from field deployments is driving these requirements. Once available, empirical data about consumers' privacy attitudes and choices can serve as a benchmark for new algorithms to cluster privacy attitudes and aggregate privacy policies. I see the latter as the most promising area for mid-term achievements in unifying rigorous methods with solid data.

In conclusion, we are facing a research challenge and an engineering challenge to make data protection a competitive advantage. We need formal methods not for the sake of it, but for privacy assurances we can count on. We need models and tools to cope with diversity and evolution in privacy attitudes. And, we need experiments to learn and look behind consumers' privacy choices online.

8. ACKNOWLEDGEMENTS

Florian Kammüller (TU Berlin), Alastair Beresford, and Joseph Bonneau (University of Cambridge) have provided helpful feedback on earlier versions of this text.

This manifesto is part of research into a "Privacy Calculus", jointly supported by the British Council and the Deutscher Akademischer Austausch Dienst (ARC Project 1351).

9. REFERENCES

[1] K. A. Barchard and J. Williams. Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, 40(4):1111–1128, 2008.

[2] B. Berendt, S. Preibusch, and M. Teltzrow. A privacy-protecting business-analytics service for on-line transactions. *Int. J. Electron. Commerce*, 12(3):115–150, 2008.

[3] A. Beresford, S. Preibusch, and D. Kübler. Unwillingness to pay for privacy: A field experiment. IZA Discussion Papers 5017, Institute for the Study of Labor (IZA), June 2010.

[4] BITKOM, Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. [Federal Association for Information Technology, Telecommunications and New Media]. 12 Millionen Deutsche machen Falschangaben im Web [12 million Germans cheat on Web forms], 2010.

[5] J. Bonneau and S. Preibusch. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *The Eighth Workshop on the Economics of Information Security (WEIS)*, 2009.

[6] D. Cvrcek, M. Kumpost, V. Matyas, and G. Danezis. A study on the value of location privacy. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 109–118, New York, NY, USA, 2006. ACM.

[7] T. Dinev and P. Hart. Internet privacy concerns and their antecedents - measurement validity and a regression model. *Behaviour & Information Technology*, 23(6):413–422, 2004.

[8] K. Hayati and M. Abadi. Language-based enforcement of privacy policies. In D. Martin and A. Serjantov, editors, *Privacy Enhancing Technologies*, volume 3424 of *Lecture Notes in Computer Science*, pages 302–313. Springer Berlin / Heidelberg, 2005.

[9] C. Jensen, C. Potts, and C. Jensen. Privacy practices of internet users: Self-reports versus observed behavior. *International Journal of Human-Computer Studies*, 63(1–2):203–227, 2005.

[10] A. Kobsa and M. Teltzrow. Contextualized communication of privacy practices and personalization benefits: Impacts on users' data sharing and purchase behavior. In *Privacy Enhancing Technologies (PET)*, 2005.

[11] K. Lewis, J. Kaufman, and N. Christakis. The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer-Mediated Communication*, 14(1):79–100, 2008.

[12] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, 15(4):336–355, 2004.

[13] A. C. Myers. Jif: Java + information flow, 1999-2009. `http://www.cs.cornell.edu/jif/`.

[14] S. Preibusch. Key facts on privacy negotiations, 2009. `http://privacy-negotiations.de/`.

[15] S. Preibusch. Privacy types revisited, 2010. `http://talks.cam.ac.uk/talk/index/22536`.

[16] H. Rossnagel. The market failure of anonymity services. In P. Samarati, M. Tunstall, J. Posegga, K. Markantonakis, and D. Sauveron, editors, *Information Security Theory and Practices. Security and Privacy of Pervasive Systems and Smart Devices*, volume 6033 of *Lecture Notes in Computer Science*, pages 340–354. Springer Berlin / Heidelberg, 2010.

[17] J. H. Smith, S. J. Milberg, and S. J. Burke. Information Privacy: Measuring Individuals' Concerns about Organizational Practices. *MIS Quarterly*, 20(2):167–196, Apr. 1996.

[18] S. Spiekermann, J. Grossklags, and B. Berendt. E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In *EC '01: 3rd ACM conference on Electronic Commerce*, pages 38–47, 2001.