The privacy landscape: product differentiation on data collection – working draft –

Sören Preibusch
Computer Laboratory
University of Cambridge
sdp36@cl.cam.ac.uk

Joseph Bonneau
Computer Laboratory
University of Cambridge
jcb82@cl.cam.ac.uk

Abstract

Whilst the majority of online consumers do not seem to take the privacy characteristics of goods and services into account with their consumption choices, a sizeable proportion consider differences in data collection and processing amongst alternative suppliers when deciding where to buy. Meeting their heterogeneous privacy preferences would require varied privacy regimes between different suppliers. Based on an empirical evaluation of 140 Web sites across five industries, we consider two questions: (1) can privacy-conscious consumers find a privacy-friendly seller/provider? (2) is this alternative associated with higher prices? We interpret the empirical evidence using the economic model of horizontal differentiation. As an overarching conclusion, differentiation on privacy is more prevalent in markets where consumption is priced—an observation that confirms the prediction from theory. Surprisingly, sellers that collect less data charge lower prices, with high significance. Implications for regulation and for further study are discussed.

WEIS 2011
The Tenth Workshop on the Economics of Information Security
http://preibusch.de/publ/privacy_landscape/

Contents

1 Competition on privacy

The policy of self-regulation on consumer privacy, implemented prominently in the United States more than a decade ago as well as in other jurisdictions, assumes “market resolution of privacy concerns” [14]: a firm whose data collection and use is incompatible with consumers’ privacy preferences will see its demand diminish. To some extent, ideally to the socially optimal extent, privacy-friendliness should be maximally profitable. Indeed, three out of four consumers state they would cancel their online shopping transaction if asked for personal information they are unwilling to provide. The majority of them indicate they would switch to an alternative Web site [1].

Consumers exhibit heterogeneity in their privacy preferences, expressing different levels of discomfort in revealing different data items to commercial Web sites [11]. Simply reducing the amount of personal information collected online cannot satisfy all user’s preferences, though. Only fifteen percent of Web users would want to give up the personalisation benefits they receive in exchange for not revealing details about themselves [4]. Therefore, consumers’ ability to choose a company whose privacy practices are aligned with their individual demand for data protection requires differentiation on the supply side. If failure to provide desirable levels of data protection becomes universal across firms, consumers will be unable to divert their demand due to lack of alternatives.

Dissatisfying security and privacy practices have been identified across many Web industries: deployment of tracking technologies, data sharing with affiliates, and vague textual privacy notices are ubiquitous [7]. Three out of four Web sites across popularity strata and audience use persistent cookies that allow re-identifying users over multiple visits [13]. Few studies, however, have looked at privacy differences within a given industry. For instance, whilst P3P adoption has been reported to vary strongly across industries, the within-industry variance has not been studied [5].

A distinction between accessibility of privacy policies and actual privacy practices was found in online social networks by Bonneau and Preibusch [2]. Amongst online social networking sites competing for members worldwide, privacy practices varied considerably, but were rarely used as a promotional argument. There was nonetheless a trend that social networking sites offering premium, that is priced accounts, collected less personal information from their members. At the national level, a similar comparison was performed by Fraunhofer SIT, concluding that platform operators over-collected personal information compared to the technical minimum for service provision. None of the market players was found to be leading on privacy in general; instead, different platforms excelled at different aspects of data protection [6]. Privacy practices of competing firms have also been assessed against one another by consumer watch organisations in areas including online social networks, online dating platforms, music downloads, online travel agents or webmail providers—all of these studies being aimed at guiding consumption choices [18].

Empirical evidence regarding consumers’ reactions towards differentiated supply in privacy is scarce. In laboratory experiments, users of a product search engine were found to choose companies with good privacy reviews despite higher prices when shopping for sensitive items [16]. However, when shopping for DVDs, consumers’ consumption choices were dominated by price, rather than privacy concerns [1].

We investigate privacy practices at competing Web sites within specific industries. To the best of our knowledge, this is the first endeavour to search for such evidence of privacy competition on the Web. We studied five industries that supply goods and services over the Web, some of which are offered for a fee or price, others at a zero price. In total, 130 competing Web sites were analysed (Table 1), plus 10 non-competing Web sites. While merchants differentiated on data collection, little variation was found for zero-price services. For e-commerce sites, higher prices were associated with more abundant data collection. Web sites facing little competition tend to collect significantly more personal details than other services offered for free.

Industry	product	pricing	n/o sites
Consumer electronics retail (cameras)	good	priced	29
Entertainment retail (DVDs)	good	priced	22
Social networking	service	zero-price	23
Web search	service	zero-price	18
Blogging	service	zero-price	42

Table 1: Overview of studied Web sites and industries

2 Methodology

This study is driven by one over-arching research question: do competing online firms differentiate themselves on privacy? As a corollary, we also investigate whether privacy differentiation, if any, is related to different pricing behaviour. We further study what impact the absence of competition has on data collection.

2.1 Hypotheses

Web operators have been observed to exhibit differing privacy regimes. Social networking sites, for instance, exhibit pronounced variation in the amount of personal information collected from their users. Some sites promote themselves as privacy-friendly, and although this promotional claim is often not met in practice, some sites do seem to occupy a privacy niche [2]. Several German social networking sites tried to turn data protection into a positively distinguishing feature in early 2010, at a time when Facebook was overtaking the former national incumbents [12]. Differences have also been observed regarding the care for user passwords amongst Web sites offering free sign-up. However, better security was not used as an argument to entice users, and practices differ more strongly between industries rather than within industries [3].

Economic theory predicts that firms differentiate their products to evade price competition that would drive their profits towards zero as prices approach marginal costs.

Horizontal differentiation, introduced by Hotelling, describes the concept of products differentiated along a characteristic for which there is no universal better or worse, but for which consumers develop a taste. The taste characteristic of privacy has been established in the context of online social networking [9]; our own research indicates that a vertical ordering cannot be established even for data items of seemingly obvious sensitivity levels: in comparing the data items 'home address' and 'hobbies', both of which might be used alternatively for personalising entertainment recommendations, 58% of consumers are more willing to provide the latter, but 20% are more comfortable with revealing the former [1].

Heterogeneous preferences are conceptualised as a distribution of consumers along a taste interval of length 1. The position of a consumer on the interval is determined by her preference for something on the left over something on the right. For instance, someone at position 0 would strongly prefer revealing one's home address over revealing one's hobbies, and vice versa for a consumer located at position 1. The strategic behaviour of competing firms facing such a demand is well studied in economics ([15, ch. 7]). Two antagonistic effects determine their choices of prices and positions on the interval, meaning their practices within the privacy space.

Given a structure of positive prices, firms want to increase their market share by moving closer to each other, towards the centre of the interval: a firm on the 'left' side of the interval serves all customers located between itself and the left edge, and by symmetry, the firm on the 'right' side is the preferential supplier for those located between it and the right edge. In expanding these monopolistically served regions, firms have a tendency to move towards the middle. However, as firms are close to one another, price competition becomes stronger. In the extreme case, firms located at the same position are no longer differentiated and compete solely on price. In order to relax competition, maximum differentiation is sought by the firms when they charge prices [15]. A differentiated product can establish a profitable niche. Our first research hypothesis is therefore:

Hypothesis 1: *Web sites offering the same product at a positive price will differentiate on privacy.*

A principle of minimum differentiation applies instead if price competition disappears, typically due to exogenous effects. Many services on the Web are offered at zero price, which is effectively fixed by consumer expectations. With no price competition to avoid, firms must compete for the same consumers at this given price. Web operators would group around the consumer with average privacy concerns. The entire market is then split proportionally amongst the alternative vendors, yielding a higher demand than a fully occupied small privacy niche. Application of this principle would also explain why competing Web sites offering free registration exhibit similarly good or bad security practices without making it an aspect of differentiation. Our second research hypothesis is therefore:

Hypothesis 2: *Web sites offering the same product at zero price will differentiate less on privacy.*

If firms have successfully escaped price competition through discrimination, they may charge an above-market price without losing all customers. A priori, as data protection is a horizontally rather than a vertically differing characteristic, it is unclear whether a company collecting more data will be cheaper or more expensive. Experience suggests that additional personal information can be used for price discrimination, enabling firms to better extract rent from consumers. In comparing an industries with and without the ability to price discrimination, one often observes that in the presence of personalised pricing, firms sell to more customers some of which are charged lower prices. Additionally, personal information can be monetised directly by increasing sales with targeted advertising or indirectly through selling or renting this data to third parties. Consumers buying from a company collecting more information would thus subsidise the retail price through the monetary value of the data they provide. Moreover, empirical evidence suggests that privacy awareness and social status increase concordantly amongst consumers. Privacy concerns and disposable income would have a common antecedent and thus be positively associated. A more privacy-friendly firm would attract buyers with a higher willingness to pay [17]. In summary, our third research hypothesis is therefore:

Hypothesis 3: *If the same product or service, differentiated on privacy, is offered by two competing Web sites, it will be offered more cheaply at the less privacy-friendly firm.*

Further, the absence of competition would imply that consumers have to consume from a given Web site regardless their tastes. If this applies to a service offered for free, the markup would be in the amount of collected and monetisable data, rather than in the price. In analogy to the under-supply of quality by monopolists, our fourth research hypothesis is therefore:

Hypothesis 4: *Web sites facing little competition are less privacy-friendly.*

2.2 Operationalisation

The *relevant market* of a firm is delimited geographically, temporally, by product, and by the position in the value chain. For a homogeneous consumer product, we limit the market to substitute products available online at the same time, and in the same currency if the product is priced.[1]

Competition between two firms is assumed if they operate in the same relevant market and are perceived by consumers as substitutes. We do not adopt a metric of competition based on technological similarity or cross-price elasticity. For physical products, we take a conservative approach in only considering identical offerings (the same DVD title or digital camera model) to be substitutes. We intentionally avoid comparing potentially similar products (such as DVD box sets and any of the component DVDs, or cameras of the same model but differing in colour or minor accessories). Two metrics are used as a proxy to determine the competitors of a given Web site. First, co-occurrence in the Web site category describing the relevant market on the Open Directory Project (ODP); second, co-occurrence within search results for a given product.

The *privacy regime* of a company is measured by the amount of personal information the consumer is asked to provide explicitly during a registration prior to consuming a Web site's product. We do not include technical data collected implicitly such as a users' IP address or stored third-party cookies. If no registration is required prior to consumption, the amount of explicitly collected personal information is nil. For a given data item, such as 'first name', 'email address' and so on, a company may collect this information on a mandatory basis, on an optional basis, or not at all. Added to the privacy regime are login options, the provision of a privacy notice, a visual distinction between mandatory and optional fields, and the means of seeking consent to data collection.

Differentiation in privacy regimes is assessed by the (existence or lack of) similarity between the privacy regimes across the entire industry. Testing for differentiation is distinct from establishing an ordering over the privacy regimes. In the absence of a definite statistic, we assess the disparity in privacy practices within a given market using two metrics. First, measuring sample *diversity* for nominal or ordinal data is not typically considered in basic statistics, although a number of indices of qualitative variation (IQV) exist. Amongst these, the 'Variance Analog' (VA) metric is the only one that continuously extends to multivariate applications whilst offering computational ease [19]. The VA positions a population along a continuum ranging from homogeneity ($\mathrm{VA} = 0$) to heterogeneity ($\mathrm{VA} = 1$). It can be calculated from vector-based variables (such as those describing data collection regimes), any entries of which can be polytomous or dichotomous [10]. Despite its advantages, the VA suffers from not attaining its theoretical maximum of 1 even for orthogonal variables. Further, it is rather unstable when adding categories with low frequencies [19]. We therefore consider a second metric, Cronbach's alpha, which is also suitable for ordinal data, to measure agreement between all competing firms. For well-behaved data, alpha indicates how consistently the scores for each characteristic contribute to a global score for a latent concept: if one is unable to assess a concept of interest directly, a battery of scores is used instead. Alpha gives an indication how well those proxies, taken together, capture the original concept. A value of one would mean that assessing a Web site by its collection/non-collection behaviour for a selection of data items is a perfect proxy for this Website's overall data collection regime. A value of zero would mean that the individual scores cannot be combined into an aggregate score as they would not vary concordantly. However, Cronbach's alpha can reach negative values which cannot be given a sensible interpretation. This happens as the dimensions of assessment vary systematically discordantly. This may happen for bad choices of the individual scores, but also if the data is not drawn from a single population. In light of the critiques this statistic has received, we interpret alpha as a measure of *internal consistency* rather than measurement reliability.

[1] While modern payment cards usually can transparently make payments in foreign currencies, shipping costs often limit consumers to Web merchants aimed at their own country. In addition, for both DVDs and digital cameras, which we studied, manufacturers often price products differently in different regions, in the case of DVDs using a specifically designed DRM system (region codes) to facilitate price discrimination.

	merchant A	merchant B	ordering	merchant C	merchant D	ordering
product 1	$7	$8	$<$	$3	$5	$<$
product 2	$6	$8	$<$	$3	$5	$<$
product 3	$12	$10	$>$	$6	$7	$<$
product 4	$4	$4	$=$	$6	$7	$<$
product 5	$10	–	n/a	$8	$4	$>$
median	$6.50	$8.00	2 ‘$<$’, 1 ‘$>$’	$6.00	$5.00	4 ‘$<$’, 1 ‘$>$’
significance in pairwise ordering counts			$p_{\mathrm{G}} = 0.68$			$p_{\mathrm{G}} = 0.31$

Table 2: Illustration of our procedure to compare price levels between two companies, using fictitious data. Two pairs of companies are compared, A and B on the one hand and C and D on the other hand. Merchant B does not sell product 4. Comparison is done using the median price and the consistency in pairwise ordering counts. Both metrics typically agree, although one may encounter data where this does not apply.

Ordering privacy regimes is achieved pairwise by comparing the amount of personal information collected by two companies. A *strict subset relation* is used: a company is said to collect less data if and only if the other company collects all plus at least one additional data item. Given a pair of companies, privacy regimes can only be ordered if all of the data items one of the companies collects are also collected by the other company. The order of privacy regimes may therefore be undefined using the subset relation. We deliberately make no attempts to compare the data collection regimes of companies when both of them have unique data items. Although average sensitivity scores are available for commonly asked personal information, this approach would be incompatible with our assumption of horizontally differentiated consumers. When using the subset test, optional and mandatory data items are treated alike.

Ordering companies by price is done by first ordering all products they both sell by price, which is a metric variable. We use quoted prices for a given product in new condition and excluding bundles. Prices are only compared if they are in the same currency. A company is then said to be cheaper than another if it sells common products consistently cheaper. Consistency is desirable since the proliferation of personal information when buying from a newly chosen retailer each time is typically more privacy-invasive than committing to one seller. We use two complementary approaches to assess this consistency. Table 2 illustrates the procedure.

First, across all studied products, for a given pair of companies, we compare the number of cases in which the product is sold at a strictly lower price to the number of occurrences of a strictly higher price. Equal prices (ties) are discarded. Products not sold by either or both companies are removed from the price comparison. A 2×2 G-test is used to assess whether there is a consistent difference in prices, at $p = 0.10$, benchmarked against a half-split (note that the G-test allows non-integer counts). Second, for each company, the median price across the products it offers is computed. Again using strict inequalities, a company is said to be cheaper if its median price is lower than the one of a competing firm. This implements the rationale that customers may themselves compare multiple prices during explorative browsing to get a feel for the prices offered at a site.

2.3 Selection of industries and Web sites

Industries were selected which provide goods and services typically consumed online. We operationalised this typicality as popularity according to Alexa “top sites”, where the US country ranking was used to guide our choice of industries: Google was ranked first, Facebook second, then Yahoo!, YouTube, Amazon.com fifth and Blogger sixth, eventually leading to the inclusion of search engines, social networking sites, retailing, and Web blogging into our sample. Top Web sites were used as seeds to discover other Web sites in the same industry. The mapping from Web sites to industries is not functional, however. The most visited Web sites in particular often supply multiple services and which sites they compete with depends on the service under consideration. For instance, we consider Bing as a competitor for Google in the market of search engines, but Microsoft Advertising is not included since they compete in different markets.

Sites we considered not to face substantial competition were excluded, as described in Section 2.3.3, since their practices are explicitly outside the scope of our research agenda of examining the existence of privacy-friendly alternatives.

2.3.1 Sampling positive-price Web sites

Online retailing was found to be the only industry where products were not offered for free. The ability to buy from an alternative vendor is determined by the availability of the same product. An online clothing store and an online book store are not direct competitors. The sampling of online merchants was thus realised by the sampling of products, yielding a list of electronic retailers selling an acceptably high proportion of them.

For our study, we desired products which are distinctly branded independently of individual merchants so that they can be easily compared across sites. We chose to consider sellers of DVDs and digital cameras, both of which have a large number of online sellers and come in a relatively small number of distinctly labelled varieties. This choice of products is also appropriate for reasons of external validity and pragmatic consideration: DVDs and cameras are typically bought online, they are homogeneous so that the choice of the retailer does not impact on perceived quality, they are not price-regulated (unlike, for instance, books or train tickets), they are, as far as we know, not price-discriminated (unlike air travel or hotels), and the title or product identifier are unique enough to identify the same product across different stores.

For digital cameras, we obtained a list of 31 best-selling cameras for the year 2010. The complete list is provided in Appendix A.2. Using Google Product Search (`products.google.com`), we then compiled the list of all shops listed as selling any of our top cameras, and kept the 29 sites which offered the largest proportion of our cameras. We consider the sampling bias introduced by Google to be small if not desirable. Price information was then recorded automatically using the Google Product Search interface. We recorded only the "base price" for a new product with no accessories, excluding shipping and handling costs. We also recorded a seller rating for each of these merchants, as well as the availability of Google Checkout as an alternative to registration with the site.

For DVDs, we selected a random sample of 20 film titles from the list of 500 highest-grossing films of all time provided by `imdb.com`. The complete list is provided in Appendix A.1. The set of competing online sellers was compiled via `dmoz.org` and Alexa, using the same methods as for non-priced products (Section 2.3.2). For each film, we recorded the price for each merchant, in a semi-automated manner. For films with multiple versions (extended versions, collector's editions, etc.), the cheapest version offered was chosen, excluding Blu-ray discs. Four sites were excluded because their selection turned out to be too limited (`shop.abc.net.au`, `shopto.net`, `game.co.uk`, `mymemory.co.uk`, `homeshopping.24ace.co.uk`), leading to a sample of 22 sites.

2.3.2 Sampling zero-price Web sites

Sampling through Google Product Search is unavailable for products which are not sold, but offered at a zero price. As an alternative, a manual inspection of categories under which a top Web site was listed in the Open Directory Project, accessed via `dmoz.org`, was used to determine the most relevant market for a Web site. By region/language, categories are organised in trees, which are similar in structure. For top Web sites, hundreds of categories are returned. The most sensible category can be identified quite easily though.[2]

As a guiding principle, the English-language tree was used, and 'Computers' was expected to show up as the first sub-category. Once the representative category of a top site was determined, all other Web sites in this category were manually inspected to decide whether they should be considered as competitors. Sites listed in the same category needed to be excluded if they were inaccessible, discontinued or temporarily 'under construction', not functioning properly, not directed towards consumers or documentation about a service rather than offering the service itself.

The Web site sampling via `dmoz.org` was complemented by sampling using a seed Web site's 'Related Links', as provided by Alexa. According to the rather vague definition given by Alexa, construction of related links takes into account characteristics of the sites, but also click paths by users, resulting in not necessarily symmetric relations. Listings typically include alternative Web sites, which can be interpreted as competitors. Highly popular sites, however, are often found clustered together, as by definition the Web population is likely to visit several of the top sites (for instance, Google-Facebook-Yahoo!-YouTube, ranked 1 to 4). Non-competing sites, which match the interests of the audience are also included and must not be counted as competitors. Taking the example of the University of Cambridge, Stanford, Oxford etc. are competitors on the global market for higher education, but the railway is rather a means of getting to the University and `cambridge.ac.uk` is just a redirect to the institution's site (see Table 3).

[2] For flickr, for instance, the first two categories are "Recreation: Outdoors: Urban Exploration: Image Galleries" and "Society: Death: Death Care: Cemeteries: Image Galleries". Links therein include `http://www.flickr.com/groups/abandonded_gas_stations/` or `http://www.flickr.com/groups/cemeterygates/`. A link to the homepage `http://www.flickr.com/` is found in category "Computers: Internet: On the Web: Web Applications: Photo Sharing", listed 73th and one of only three categories listed under "Computers".

2.3.3 Excluded Web sites

From the outset, we excluded industries for which we had currently no ability to determine accurate pricing behaviour or for which we concluded that the sites were not offering a homogeneous product. This notably excluded airlines or car rental companies. Also excluded were markets for which there was an exogenously given consumption constraint, such as banking sites, ISPs or mobile phone operators.

Amongst those Web sites identified as being part of a chosen relevant market, we excluded all that had discontinued their service, locked registration, technically failed during registration after retries, required an upfront payment or were identified as duplicates of other Web sites in the sample.

For the top 25 according to Alexa, we set aside Web sites and thus industries for which we were unable to identify a set of competitors that sufficed our expectations of face validity. As an indicator, we used the absence of same-industry Web sites in Alexa's "related links", leading to the exclusion of YouTube, Twitter, or eBay amongst other. These Web sites were recorded to form a 'monopolies' sub-sample.[3] This would include flickr.com, for instance, ranked 22$^{\text{nd}}$ amongst US top sites: Web sites such as Blogger.com, Fotolog.com, Photobucket, Webshots or Twitter could all be considered competing sites, in markets for photo viewing, uploading or sharing, or photo-centric or other self-presentation, but based on the related-links indicator and our judgement, we concluded otherwise (similarly for aol.com, despite it facing competition as a search engine).

2.4 Data collection procedure

Assessment of all Web sites in our sample was manual, although supported by tools. Data was collected using a fresh profile in the Firefox browser, version 3.6.10, using the extensions 'Autofill Forms', 'CipherFox', 'Ghostery', and 'View Cookies', all in the latest version as of March 10th, 2011. A UK Internet connection was used; forced redirects to regional sites (e.g. `aol.co.uk` instead of `aol.com`) were followed. Data collection was assessed by completing the sign-up forms of each Web site, sometimes spanning multiple Web pages. All fields marked as mandatory were filled in before submitting; fields not marked as mandatory were left blank. A constant set of personal information was used representing a male in his late thirties with an imaginary residence in California. Changes to this standard sign-up profile were made only to adhere to site-specific requirements, such as (un-) availability of the test username, addresses from United Kingdom or lookup-based postcode checks for California. Valid email addresses were provided at all times, tailored to the specific site. Emails received as part of creating an account were opened and links for account 'verification' were clicked.

2.5 Data standardisation and analysis procedures

Conservatively, data items collected by a given Web site were coded '1', regardless of whether disclosure was mandatory or optional. Data items not collected during initial sign-up were coded '0'. Potentially, optional data items could be assigned any score between 0 and 1, denoting the extremes of absent respectively present collection. We treated them no different from mandatory items because visual indicators that would discriminate between optional and mandatory items were often missing (as discussed below). We certainly noticed a tendency amongst Web sites not to indicate a certain data item was optional despite it not being required for successful registration. We further observed the practice of placing a heading "all field are required" above the form and adding "optional" in fine print to some fields. If present, such visual indicators are often overlooked, in particular for standard forms. Also, automatic form fillers are typically oblivious to compulsiveness so that consumers making use of them will reveal even optional items. We further note that the classification of an input field as 'mandatory' or 'optional' may not be clear-cut: for gender enquiry, we encountered drop-down lists with a third option of "unspecified" and "decline to state".

Cronbach's alpha, a statistic used to assess the internal consistency of characteristics exhibited by several entities, was calculated for all five markets. The coefficient alpha is sensitive to the number of characteristics; we therefore took a constant set of data items across all industries. Name details (first name, surname, full name) were lumped into one synthetic data item 'name' if at least one of them was collected. Across all Web sites in our sample, the collection of first name and last name is highly positively correlated ($\rho = 0.93$). The cases when both are collected in separate fields largely outnumber the cases of collection as "full name" (68 vs. 15).

We similarly lumped into one characteristic all address data, that is street address, city, state, and postal code. This also follows the rationale that in countries such as the United Kingdom, there is almost a functional relationship between the postal code and the street address. Country was not included in this list, as we felt that country information alone was not of comparable sensitivity to the other fine-grained address requirements. For date of birth, we collapsed the day, month, and year components (which were again often collected in conjunction,

[3] We do not claim that Web sites listed in this sub-sample are monopolists, but rather that we were unable to find competing Web sites.

cam.ac.uk	facebook.com	amazon.com
University of Cambridge	Facebook	Amazon.com
Stanford University	Xanga	Buy.com, Inc.
University of Oxford	Myspace	Amazon UK
Massachusetts Institute of Technology	LinkedIn	eBay
Cambridge University Press	Google	Barnes & noble.com
The Gates Cambridge Scholarships	Friendster	ABC Shop Online
Ucas : Universities & Colleges Admissions Service	University Life	Alibris
QS Top Universities: study abroad guides, THE QS World University Rankings, Bach	Student Life and Culture Archival Program	CD Universe
National Rail Enquiries	Zynga Inc.	CBC Boutique
London School of Economics and Political Science (LSE)	YouTube	Buy Sell Rent Media
www.cambridge.ac.uk/	Yahoo!	AOL
		google.com/

Table 3: Examples of 'Related Links', as provided by Alexa. Whilst the listings typically exhibit convincing face validity, they require a manual inspection. Shown is the complete list per site, in its original formatting of the Alexa output, including inconsistent capitalisation and trailing slashes. (The site cam.ac.uk is shown for illustrative purposes only and not part of the sample.)

$\rho = 0.85$); for telephone number, the collection of any or specifically mobile or landline was equally collapsed. We thereby avoided inflating the measured consistency of data collection practices, by removing items that vary concordantly per se.

As a measure of outlier correction, we did not include in our analysis data items only collected by very few sites, such as "I identify myself as" (Blackplanet only), years of attending and rating of college and university (Perfspot only), tax identification number (Howard only), species (NuTang only), boat name (SailBlogs only), and "interested in" (once amongst social networking sites, thrice amongst camera retailers, thrice amongst weblog hosts; note the differing semantics).

We systematically recorded but excluded from the analysis the occurrence of input fields to type twice one's password, username, email or to enter a CAPTCHA.

Both our measures for statistical dispersion, Cronbach's alpha and the Variance Analog are sensitive towards changes in the number of characteristics. For the latter, standardisation is expressively discouraged and considered inappropriate [10], making comparisons between datasets with varying numbers of characteristics misleading. To ease cross-market analyses, we therefore excluded items just appearing in a single industry. In particular, this meant excluding blog title (only appearing amongst weblogs) and company name (mainly for camera retailers, otherwise once amongst weblogs and DVD retailers each).

For retailers, an adjusted coefficient alpha was calculated, $\alpha^{\dagger}$, which will be used subsequently: name and address details were removed, as those are exogenously given minimum data requirements for an online shop that ships physical goods. Moreover, we noticed that sites fell into two regimes, varying on whether a shipping address was collected during sign-up or during the first checkout. Differentiating on this behaviour would not provide a valid assessment of a company's data practices.

Eight characteristics were used to calculate VA and α, six for $\alpha^{\dagger}$. For coefficient alpha, the commonly used cut-off value of 0.80 was used to ascertain that an industry does not have differentiated data collection practices.

Price comparisons were evaluated using median prices and by pairs of companies within each market, subdivided between currencies were applicable, as described in Section 2.2. Given the symmetry of the evaluation, $N \times (N-1)/2$ pairs of companies were thus compared given a market of N suppliers.

For comparisons of prices and privacy regimes, only strict inequalities are taken into account. Cases of equal prices or equal data collection schemes are ignored. When privacy and price orderings are considered together across the sample, only cases for which both comparisons are decidable and significant are counted. Again, this is a conservative approach.

3 Results

We collected statistics for 140 sites spread across the five categories. The overall frequency of collection for different data items is shown in Table 5. In addition to these items, we notice variance in other privacy-related presentation choices. For example, only 11% of sites included a checkbox indicated agreement to the site's privacy policy, and 23% a textual reference to the agreement. Less than one third of the Web sites highlight mandatory form fields. A fair share of 19% of sites displayed privacy seals during the signup process (most of them commercially issued), counting also more general seals such as 'Verisign Secured' or 'McAfee secure'. We also noticed several compelling differences between industries in the type of data collected. For example, only 19% of weblog hosts required an address, while 50% of online merchants did prior to any purchase. However, 87% of social networking sites required a date of birth, while only 26% of the other sites across all industries in our sample did.

Amongst DVD retailers, one retailer, dvdstreet.co.uk was selling through the Amazon.com platform; the data collection scheme from the perspective of the consumer, is thus the same as for Amazon itself, and the Web site was excluded when assessing the industry-wide variance in data collection. Five Web sites were found to operate as front-ends to the same shopping system 'elysium': `sendit.com`, `whsmithentertainment.co.uk`, `zavvi.com`, `thehut.com`, and `asda-entertainment.co.uk`. WHSmith Entertainment was chosen as the representative in assessing market-wide differentiation on privacy and the other four sites were discarded. These duplicate sites were again taken into consideration when assessing the correlation of pricing and data collection practices.

Search engines are the only market for which data collection is consistent, with a large proportion of Web sites collecting no data. The market for online social networking exhibits a negative coefficient alpha, with a low variance analog, suggesting that sites in this market may fall into multiple, internally consistent camps by collecting broadly non-overlapping data items—and may thus be offering services which cannot be regarded as competing. Further analysis of this phenomenon would require an expanded sample size and is thus left for future work. Weblogs have a moderate variance analog and coefficient alpha, which may stem from the strongly

Camera retailers: www.abesofmaine.com, www.adorama.com, www.amazon.com, www.antarespro.com, www.aztekcomputers.com, www.beachaudio.com, www.bestbuy.com, www.buydig.com, www.capitolsupply.com, www.cdw.com, clickfrom.buy.com, www.compnation.com, www.compsource.com, www.compuplus.com, www.daxmart.com, www.ecost.com, www.futurepowerpc.com, www.govgroup.com, www.homemylife.com, www.howardcomputers.com, www.neobits.com, www.nextdaypc.com, www.nextwarehouse.com, www.pcrush.com, www.ritzcamera.com, www.tekmentum.com, underbid.com, www.valleyseek.com, www.walmart.com;	29 sites $\alpha = 0.80$ $\alpha^\dagger = 0.63$	VA $= 0.27$
DVD retailers: www.101cd.com, www.alibris.com, amazon.com, www.borders.com, www.buy.com, www.cdplus.com, cdquest.com, www.cduniverse.com, www.chapters.indigo.ca, www.dvd.co.uk, www.fye.com, www.gohastings.com, hmv.com, www.moviesandgamesonline.co.uk, www.play.com, www.tesco.com, www.wherehouse.com, www.whsmithentertainment.co.uk; excluded: www.asda-entertainment.co.uk, www.sendit.com, www.thehut.com, www.zavvi.com;	18 sites $\alpha = 0.64$ $\alpha^\dagger = 0.44$	VA $= 0.28$
Social networking sites: www.2befriends.net, badoo.com, www.bebo.com, www.blackplanet.com, facebook.com, fropper.com, hi5.com, www.linkedin.com, www.livejournal.com, www.meinvz.net, www.mocospace.com, multiply.com, www.myspace.com, www.myyearbook.com, www.netlog.com, www.orkut.com, www.perfspot.com, www.plaxo.com, signup.live.com, www.skyrock.com, www.sonico.com, www.tagged.com, www.xanga.com;	23 sites neg. α	VA $= 0.23$
Search engines: www.amfibi.com, www.aol.co.uk, uk.ask.com, www.bing.com, www.chacha.com, cluuz.com, www.entireweb.com, www.google.com, www.hakia.com, kalooga.com, www.mahalo.com, middlespot.com, www.mozdex.com, www.searchhippo.com, www.spiderline.net, www.ulysseek.com, www.wotbox.com, www.yahoo.com;	18 sites $\alpha = 0.89$	VA $= 0.38$
Weblog hosts: www.aeonity.com, www.blog.com, www.blog-city.infopages.php, www.blogdrive.com, www.blogger.com, bloghi.com, www.blogigo.com, blogmyway.com, www.blogomonster.com, blogs.scriptologist.com, blogs.trhonline.com, en.blogspirit.com, www.blogster.com, www.blogstudio.com, blogtext.org, www.efx2blogs.com, www.fotopages.com, www2.globbo.org, hubpages.com, www.inube.com, www.kitehost.com, lifewithchrist.org, www.ohblog.com, ohlog.com, moblog.net, www.mycookingblog.com, www.nutang.com, www.problogs.com, www.sailblogs.com; www.squarespace.com, sweetcircles.com, tabulas.com, www.tblog.com, theblogs.net, www.thoughts.com, www.tripsailor.com, www.tumblr.com, www.typepad.com, www.upsaid.com, weblogs.us, wordpress.com, www.xanga.com;	42 sites $\alpha = 0.55$	VA $= 0.26$
Non-competing sites: aol.com, cnn.com, craigslist.org, ebay.com, flickr.com, msn.com, paypal.com, twitter.com, youtube.com, wikipedia.org;	10 sites n/a	n/a

Table 4: Summary of assessed Web sites, per industry. Cronbach's alpha is indicated as a statistic of internal consistency in data collection; In the case of online retailers, $\alpha^\dagger$ indicates the value adjusted for base data collection (name and address), and with companies operating on the same Web platform collapsed into one subject (DVD retailers only). The non-standardised Variance Analog (VA) is given.

Market	Manda-	Privacy	some	email	username	password	some	some	some	gender	avg.	signifi-
	tory	policy	name	address			tele-	address	DOB			cance
	marked	linked					phone					p_t
Camera retailers	66%	72%	79%	86%	7%	86%	59%	59%	10%	3%	49%	
DVD retailers	28%	78%	83%	100%	17%	100%	39%	44%	33%	17%	54%	
Social networking sites	22%	100%	74%	100%	35%	96%	0%	43%	87%	91%	66%	0.9
Search engines	17%	17%	17%	44%	33%	50%	6%	28%	28%	28%	26%	0.004
Weblog hosts	33%	67%	50%	98%	88%	88%	2%	19%	29%	24%	50%	0.03
Non-competing sites	30%	90%	60%	100%	80%	100%	20%	60%	50%	50%	65%	

Table 5: Privacy communication and data collection practices: for each sub-sample, i.e. market, the proportion of sites are given that indicate mandatory input fields in their registration forms, that link to their privacy policy from there, and respectively, that collect certain items of personal information. The average for a market (pen-ultimate column) gives the proportion of data items collected across the Web sites in that sub-sample. The last column gives the significance at which non-competing Web sites collect more data items than other Web sites that offer services for free.

	UK DVD retailers		US DVD retailers		Camera retailers		
	data collected:		data collected:		data collected:		
prices:	more	less	more	less	more	less	prices:
higher	8	2	9	0	76	48	higher
lower	5	9	0	8	74	102	lower
	$p_{\mathrm{G}} < 0.028$		$p_{\mathrm{G}} < 0.0000012$		$p_{\mathrm{G}} < 0.0010$		

Table 6: Trend and significance of associations between prices and the amount of data collection. For each sub-market, the number of co-occurrences of higher/lower prices and more/less data collected is given. Strict inequalities are used for comparison, using median prices across assortments and subset relations, as described in Sections 2.2 and 2.5. Significance levels as determined by a G-test.

diversified market in terms of general purpose and niche blogging sites. Both electronic commerce markets, for DVDs and for cameras, exhibit differentiated data collection practices, when compared to the other markets we studied.

4 Analysis and interpretation

All studied markets on which homogeneous products are sold at a positive price, that is, sales of DVDs and cameras, exhibit differentiated data collection practices. In the light of aforementioned shortcomings with both dispersion metrics, we consider a market to be differentiated on privacy if it is amongst the three out five most dispersed Web site populations according to both metrics. A G-test reveals that these results are significant despite the small sample size ($p < 0.01$). We also note that Web sites selling goods rather than providing a service for free are very significantly more likely to highlight which input fields are mandatory (two-tailed t-test for difference in arithmetic means, $p < 0.01$). Unlike the existence of a privacy policy, typically mandated by regulations, the open communication of data collection practices is a voluntary best practice. Thus, Hypothesis 1 is supported; Web sites selling goods at a positive price do differentiate on privacy.

Web search engines, which offer an online service at zero-price, exhibit higher consistency in their data collection practices than any of the industries selling products at a positive price (Table 4). Results are inconclusive for social networking sites, as discussed in Section 3 and borderline for Web logs, for which VA is moderately high. We conclude that Hypothesis 2 is supported; Web sites offering free services differentiate less on privacy than those selling goods for a positive price.

Hypothesis 3 states a negative association of the amount of personal information collected by a firm and the prices it charges. We test this hypothesis for the markets of retailing cameras and DVDs only, as Web search and online social networking are not priced. For UK DVD retailers, 24 pairs of strict inequalities could be established, 17 for US DVD retailers, and 300 for camera retailers, using the median price for comparisons of price levels (Table 4). There is a highly significant, positive association between the amount of data collection and price level. Companies charging higher prices collect more personal information. These results are confirmed using the median method for comparing prices across assortments, as described in Section 2.5. Hypothesis 3 is rejected; priced goods are not offered more cheaply by more privacy-friendly firms.

We note that the empirical evidence regarding prices and privacy regimes is in fact the opposite of the hypothesised relationship. Further investigation reveals that this result cannot be attributed to differences between top sites and lower tier sites or between online-only and multichannel retailers. Any attempt to explain this phenomenon would be post-hoc and thus speculative. Plausible, although not tested intra-sample effects include brand effects or a positive relationship between data collection and perceived service quality through personalisation. Price discrimination does not seem to be a plausible explanation, since there is no evidence that online retailers in our study charged individualised prices.

All Web sites operating without major competition in our sample are offering free services. We therefore relate their data collection practices to other markets that provide free services. With the exception of online social networking, for which no difference was found, non-competing Web sites are collecting more personal information than search engines and weblog hosts with high significance (both $p < 0.05$ in a two-tailed t-test). Hypothesis 4 is, therefore, supported.

We note that the non-competing Web sites in our sample also enjoy high popularity, which, however, we do not consider a convincing explanation for over-collection of personal information: previous research has indicated a positive relationship between good privacy practices and high popularity [2]. Plus, one can assume that more popular sites are also under stronger privacy scrutiny from the media and other interest groups.

5 Conclusions and critical review

5.1 Summary

When shopping online, consumers are faced with a supply that is differentiated in privacy. They may choose to buy from a company whose data protection practices are compatible with their own preferences. Our empirical evidence suggests that electronic retailers compete on privacy. However, at high levels of significance, consumers do not face a trade-off between privacy and price. In choosing a privacy-friendly retailer for DVDs or digital cameras, consumers are also likely to get a better deal more than half of the time and to pay less across the assortment of the seller.

Consumers may choose from a broad variety of hosts for their weblog, and they have fair chance of finding a provider whose privacy regime matches their preferences. They have less choice when using Web search engines. Although they all offer a basic service without registration requirements, there is little variance in the data one needs to provide when signing up for a personalised service.

Web sites which do not face strong competition are significantly more likely to ask for more personal information than other services provided for free, such as Web search or blogging. Social networking sites, however, collect data to an extent otherwise seen only for sites for which the risk of loosing customers to competitors is low.

Our findings on the variety in and the amount of data collection depending on price and market structures are in line with the predictions economic theory makes. The co-occurrence of more privacy and lower prices, however, comes as a surprise and mandates further study.

5.2 Limitations

We address several limitations. First, our operationalisation of privacy as the extent of data collection ignores the importance of use and sharing of personal information—which may be even more important than data collection. However, these facets of data protection are typically unobservable to the user. Second, our economic model may have been too simplistic, although it did explain the observed phenomena reasonably well. Neither lock-in effects, particularly prevalent in service consumption, nor differences in quality between service alternatives are considered. We only consider one-time consumption without accounting for the effect of repeated purchases. Third, data collection was largely resistant against automation, in particular as determining the relevant market and competitors therein requires judgement. Inevitably, this also introduces sources of human error. Fourth, we have only studied five industries so far. Our sample does not include paid-for services, physical goods offered for free, or services delivered beyond the Web. Finally, regarding our conclusions, online merchants may differentiate on privacy for other reasons than competing on privacy.

5.3 Managerial and regulatory implications

In the case of search engines or online social networking, one may conjecture that service providers are too close to one another from a social welfare perspective. Welfare could be increased if there were more variance in privacy regimes. A social planner would mandate that search engines and online social networks were more spread out over the continuum of privacy preferences. Higher dispersion would also mean that more Web users start using social network sites, who, at the time being, would incur prohibitively high ‘transportation costs’ in signing up for the service. Given the difficulty—or “impossibility” [8]—to regulate sites like Facebook, Google and its global competitors, transforming them into paid-for services could be a viable approach, to incentivise differentiation on privacy in the operators’ own interests.

From a managerial perspective, there is an incentive to introduce new products and occupy several positions in the product space as a single company. Several markets with dominating firms exist on the Web, including online social networking, Web search, online auctions, or video sharing. Current quasi-monopolists might therefore increase their market shares and attract new customers by differentiating into a multi-brand monopoly.

References

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

[2] Joseph Bonneau and Sören 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.

[3] Joseph Bonneau and Sören Preibusch. The password thicket: technical and market failures in human authentication on the web. In *The Ninth Workshop on the Economics of Information Security (WEIS)*, 2010.

[4] Personalization Consortium. Personalization & privacy survey, 2000, 2005. via Internet Archive.

[5] Lorrie Faith Cranor, Serge Egelman, Steve Sheng, Aleecia M. McDonald, and Abdur Chowdhury. P3P deployment on websites. *Electronic Commerce Research and Applications*, 7(3):274–293, 2008.

[6] Fraunhofer Institut für Sichere Informationstechnologie SIT. Privatsphärenschutz in Soziale-Netzwerke-Plattformen. Technical report, Fraunhofer SIT, August 2008.

[7] Joshua Gomez, Travis Pinnick, and Ashkan Soltani. KnowPrivacy. Technical report, UC Berkeley, School of Information, June 2009.

[8] Sabine Leutheusser-Schnarrenberger. Regulierung im Netz: „Ihr Reflex greift zu kurz“, 2011. Interviewers: Tina Hildebrandt and Heinrich Wefing.

[9] Kevin Lewis, Jason Kaufman, and Nicholas 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.

[10] Stanley Lieberson. Measuring population diversity. *American Sociological Review*, 34(6):850–862, 1969.

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

[12] Sören Preibusch. Datenschutz-Wettbewerb unter Social Network Sites [Privacy competition amongst social networking sites]. In *Freundschaft und Gemeinschaft im Social Web. Bildbezogenes Handeln und Peergroup-Kommunikation auf Facebook & Co.*, pages 269–284. Nomos Verlag, 2011.

[13] Andrew F. Tappenden and James Miller. Cookies: A deployment study and the testing implications. *ACM Trans. Web*, 3:9:1–9:49, July 2009.

[14] The White House. The framework for global electronic commerce, 1997.

[15] Jean Tirole. *The Theory of Industrial Organization*. The MIT Press, 1988.

[16] Janice Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. The effect of online privacy information on purchasing behavior: An experimental study. In *The Sixth Workshop on the Economics of Information Security (WEIS)*, 2007.

[17] Hal Varian, Fredrik Wallenberg, and Glenn Woroch. Who signed up for the do-not-call list? In *Workshop on Economics and Information Security*, 2004.

[18] Stiftung Warentest. Suche „Datenschutz“ [Search “Privacy”], 2011. `http://www.test.de/suche/?q=Datenschutz`.

[19] Allen R. Wilcox. Indices of qualitative variation and political measurement. *The Western Political Quarterly*, 26(2):325–343, 1973.

A.1 DVDs used for price comparison

50 First Dates
An Officer and a Gentleman
Beverly Hills Cop II
Charlie's Angels: Full Throttle
Enemy of the State
Fun with Dick and Jane
Ghostbusters II
Jumanji
Jurassic Park III
My Big Fat Greek Wedding
Saving Private Ryan
Shrek
Sleeping with the Enemy
The DaVinci Code
The Nutty Professor
The Sixth Sense
The Sting
Top Gun
True Lies
Tropic Thunder

Table 7: The list of films used to compare prices of entertainment retailers. It represents a random sample from the top 500 highest-grossing films worldwide, provided by `imdb.com` as of March 2011.

A.2 Digital cameras used for price comparison

Canon PowerShot G12
Canon PowerShot S95
Canon PowerShot SD1300
Canon PowerShot SD1400
Canon PowerShot SX20
Canon PowerShot SX30
Casio EXILIM G EX-G1
Casio EXILIM EX-Z35PE
Fujifilm FinePix JV100
Fujifilm FinePix W3
Fujifilm FinePix S1800
Fujifilm FinePix XP10
Kodak EASYSHARE C143
Kodak EASYSHARE M590
Nikon Coolpix L22
Nikon Coolpix P100
Nikon Coolpix S8100
Olympus E-PL1
Olympus FE-47
Olympus SP-800UZ
Olympus Stylus Tough 6020
Panasonic Lumix DMC-FZ35K
Panasonic Lumix DMC-ZS7
Pentax Optio W90
Pentax X90
Sony Cyber-shot DSC-H55
Sony Cyber-shot DSC-HX1
Sony Cyber-shot DSC-HX5V
Sony Cyber-shot DSC-TX9

Table 8: The list of cameras used to compare prices of electronics retailers. It was taken from a 2010 holiday camera shopping guide published at `cnet.com`.