Electronic Commerce Research and Applications 15 (2016) 52–64

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

journal homepage: www.elsevier.com/locate/ecra

Shopping for privacy: Purchase details leaked to PayPal ☆

Sören Preibusch [a],*, Thomas Peetz [b], Gunes Acar [b], Bettina Berendt [b]

[a] *Microsoft Research, UK*
[b] *KU Leuven, Belgium*

ARTICLE INFO

Article history:
Received 25 January 2015
Received in revised form 1 November 2015
Accepted 23 November 2015
Available online 7 December 2015

Keywords:
Privacy
Online payments
Payment providers
PayPal
Tracking
Electronic commerce
Electronic retailing
Data minimisation
Data leakage

ABSTRACT

We present a new form of online tracking: explicit, yet unnecessary leakage of personal information and detailed shopping habits from online merchants to payment providers. In contrast to the widely debated tracking of Web browsing, online shops make it impossible for their customers to avoid this dissemination of their data. We record and analyse leakage patterns for the 881 most popular US Web shops sampled from actual Web users' online purchase sessions. More than half of the sites we analysed shared product names and details with PayPal, allowing the payment provider to build up fine-grained and comprehensive consumption profiles about its clients across the sites they buy from, subscribe to, or donate to. In addition, PayPal forwards customers' shopping details to Omniture, a third-party data aggregator with even larger tracking reach than PayPal itself. Leakage to PayPal is commonplace across product categories and includes details of medication or sex toys. We provide recommendations for merchants.

1. Introduction

1.1. Online payment providers process rich transaction data

Online payment handling is a key enabler for electronic retailing and a growing business opportunity as mobile commerce takes off. Contactless payments have been pioneered in successful yet isolated applications, such as public transport (e.g., Oyster in London, touch & travel in Germany) or entertainment (e.g., Disneyland Finextra Research 2009, Starbucks (Hamblen 2012). General-purpose digital wallets and near-field payment capabilities are now integrated in all major mobile phone operating systems (Google 2014, Microsoft 2014, Apple Inc. 2014) and promise wider adoption across verticals.

Payment providers are intermediaries between merchants and their customers who buy and then pay for goods and services. As intermediaries, payment providers necessarily gain insight into the transaction as they process personal information, just like the delivery company will need the customer's postal address. The minimum data requirements for payment handling are the order total, the receiving merchant and an authenticated payment instrument. This corresponds to data items traditionally collected during credit card transactions. However, a much richer set of data items becomes available for online, mobile and in-app purchases, including an itemised statement of the goods purchased or information about the buyer, allowing value-added services. Amongst credit card issuers, these data are known as Level II and III but have been rarely available for point-of-sale or transactions (Software Inc. 2014).

The move towards richer transaction details is driven and enabled by three factors: first, the extended role of payment providers as shopping cart solutions, so that itemised data availability becomes a necessity; second, technically enabled by the lack of data length restrictions found in legacy payment processing; third, the mining of detailed transaction data for fraud detection and prevention (Klarna 2013). For instance, MasterCard reported acceptance by over 19 million merchants worldwide back in 2001, but only 1% would be able to "capture and transmit Level II and Level III data". These include itemised product descriptions, quantities and prices (MasterCard 2001), but still fewer details than what new online payment providers collect.

☆ Preliminary results were presented as a short paper at the Financial Cryptography and Data Security 2015 conference. This article contains the full analysis.

* Corresponding author.

E-mail address: mail@soeren-preibusch.de (S. Preibusch).

URLs: http://preibusch.de (S. Preibusch), http://www.kuleuven.be (T. Peetz), http://www.kuleuven.be (G. Acar), http://www.kuleuven.be (B. Berendt).

http://dx.doi.org/10.1016/j.elerap.2015.11.004
1567-4223/© 2015 Elsevier B.V. All rights reserved.

1.2. Potential benefits of data collection by payment providers

Fraud detection and prevention is the most-publicised benefit of collecting and inspecting purchase details. The rise of riskier card-not-present transactions over the Web or on mobile has mandated new efforts in fighting crime. Between 2002 and 2012, the most recent year for which data are available, the annual fraud losses on UK-issued payment cards has decreased from £427 million to £388 million. Whereas counterfeit, lost or stolen card fraud has decreased from £257m to £97m (−62%) during that period, card-not-present fraud for electronic commerce alone has quintupled from £28m to £140m and now accounts for the majority of losses (Financial Fraud Action UK 2013). Despite continued e-commerce growth, fraud volumes have been decreasing since their peak in 2008. The industry attributes these accomplishments to automated cardholder address verification and card security codes, to initiatives like MasterCard's SecureCode and Verified by Visa, and to the "effectiveness and sophistication of customer-profiling neural networks that can identify unusual spending patterns" (Financial Fraud Action UK 2013). The required collection of details about buyers and their purchases is therefore attractive for payment providers and merchants who can benefit from lower fees. As another example, the payment provider Klarna allows customers to pay after order placement and shipping. At the same time, it absorbs the credit risk for merchants and controls losses through risk assessment based on diverse factors, including purchase details (Gustafsson and Magnusson 2014).

Fighting payment fraud is only one of many more applications for purchase information. Payment providers have a twofold incentive to collect details for the transactions they process. One the one hand, they can use the additional data for operational efficiency in a broad sense; on the other hand, they can offer convenience features to consumers.

1.2.1. Operational efficiency

Payment providers operate in a highly regulated environment and some obligations cannot be fulfilled efficiently unless purchase details are known. They must comply with tax and legal requirements, such as products prohibited in certain regions (e.g., gambling, alcohol sales) or money laundering. They must also detect and prevent crime, such as fraud and policy violations. As an example of transaction monitoring, PayPal has "hundreds of highly trained specialists working around the clock to prevent fraudulent activity and identify suspicious transactions" (PayPal 2015). Details from past transactions are also a shared secret between the provider and its customers, and can be used for additional authentication or account recovery. Purchase details can be monetised for product innovation, as market research, and through direct marketing on an individualised basis. Insofar as payment providers provide escrow services and help buyers who have been defrauded by the merchant, transaction details can be used for risk screening. For instance, PayPal's buyer protection only covers certain physical goods. Whilst mainly in the self-interest of the provider, operational efficiency enables payment services for consumers and merchants at acceptable fees in the long run.

1.2.2. Convenience features

Buyers can enjoy peace of mind when their purchase details are displayed back to them in the very moment when making the payment. They can also inspect the transaction history in their account and get a detailed statement of previous purchases. When payment providers collect purchase details, they can offer sought-after spending reports and financial self-analysis.

1.3. Privacy concerns

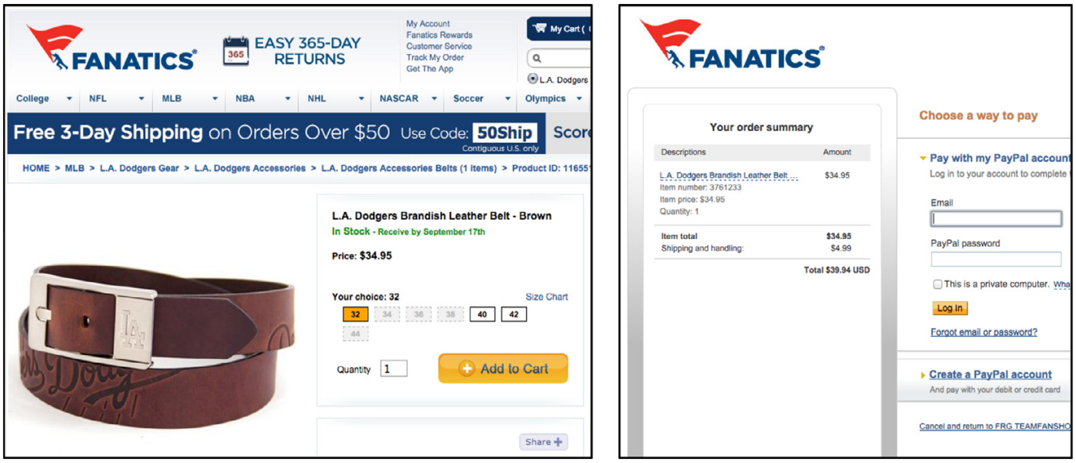
The large-scale collection and processing of personal details causes privacy concerns. Concern is no longer limited to traditional items of personal information like address or demographics, but increasingly about consumption behaviour. Despite the quantified-self movement and although Web users volunteer personal information with high prevalence (e.g., 55% knowingly entered their weekly spending behaviour into a Web form where this item was optional, Malheiros et al. 2013), extended records of usage data are problematic. Widespread tracking of browsing patterns by Websites and aggregators has raised attention in mainstream media (WSJ Online 2013). Browsing history leaked to advertisers (TRUSTe 2009), electricity consumption recorded by smart meters (McDaniel and McLaughlin 2009), or mobility trajectories in pay-as-you-drive insurance policies (Scism 2013) have all been found to be associated with elevated privacy concerns. Of particular interest is shopping data, whose value is demonstrated through myriads of loyalty card schemes. Purchase tracking now happens across merchants and channels (online/offline) and even if users are not enrolled in a loyalty scheme (Valentino-DeVries and Singer-Vine 2012, Duhigg 2012).

Our research looks at the tracking capabilities of payment providers, namely PayPal. An illustrative example is provided in Figs. 1 and 4.

Our research motivation is the ability of payment providers to collect purchase details at scale. As in the domains of Web tracking and analytics, a small number of providers cover multiple Websites (merchants) and can link transactions across those. Compared to cookie-like tracking, the privacy issues are exacerbated:

- Embedded tracking code is—in principle—ancillary to the core functionality of the Web page and can safely be filtered out (e.g., with ad-blockers or Tracking Protection in Internet Explorer). Payment handling is however essential to shopping, and users cannot complete the transaction without interacting with the payment provider.
- Unlike browsing patterns linked to a cookie identifier, consumption patterns linked to a payment method are not pseudonymous but identifiable through offline details such as credit card numbers or bank account details, which often include full name.
- Payment cards or account information serve as persistent identifiers, allowing longitudinal linkage of multiple transactions even across different logins or accounts with the payment provider.
- Consumers are typically unable to evade such data collection unless they refrain from shopping with the given merchant. The collection of shoppers' details is a negative externality of the contract between the merchant and the payment provider.
- Payment handling is universal across sellers and sectors. Consumer details are collected and merged across transactions even for sensitive products and merchants. This includes pharmacies or adult entertainment, for instance, where shoppers deliberately moved out of the high street and onto the Web in a pursuit of privacy.

Privacy threats arise from detailed purchase patterns when more than the minimum data required are collected. The principle of data minimisation has long been codified in national law and international privacy guidelines, such as the "collection limitation principle" in the OECD privacy framework (OECD 2013) or the Madrid Privacy Declaration (The Public Voice 2009). The principle of data minimisation as such is now contained in the text of the European Union's upcoming General Data Protection Regulation (European Commission 2012; Council of the European Union



L.A. Dodgers Brandish Leather Belt - Brown
In Stock - Receive by September 17th
Price: $34.95
Your choice: 32
Quantity 1

Your order summary

Descriptions	Amount
L.A. Dodgers Brandish Leather Belt ... Item number: 3761233 Item price: $34.95 Quantity: 1	$34.95
Item total Shipping and handling:	$34.95 $4.99
	Total $39.94 USD

Choose a way to pay

Pay with my PayPal account
Log in to your account to complete

Email

PayPal password

□ This is a private computer. Wha

And pay with your debit or credit card

Fig. 1. The Web shop passes the product name and item number on to PayPal.

2015), which will be directly applicable law in the EU Member States.

1.4. Theoretical background: privacy and e-commerce payment intermediaries

The benefits of data collection by payment providers, but also the associated privacy concerns discussed above, can be interpreted in the general framework of e-commerce intermediaries and their roles.

Different streams of research, including information systems, have extensively discussed the privacy aspects of payment intermediaries between retailers and their customers since the advent of electronic commerce in the early 2000s. Much less effort has been devoted to examining the technical reality and evolving business practices of this tri-party relationship.

Intermediation is a technique to overcome a trust deficit that the seller may not adequately secure the customer's payment details. Using a payment provider rather than processing card details directly thus leads to more conversions for the potentially untrusted merchant (Heck and Vervest 1998). An intermediary can overcome security and privacy concerns that can inhibit online shopping. (Hoffman et al. 1999) Whilst the trust boost is the payment providers' most important consumer-facing role, it only applies to sellers of inferior trust than the payment provider. Websites with a longer history on the Internet develop their own brand and exhibit less prominent use of trusted third parties (Palmer et al. 2000). Even then, merchants continue to benefit from aggregation features and cost savings for set-up and ongoing transactions offered by intermediaries (Bailey and Bakos 1997, Arnab and Hutchison 2007).

The technical expertise required to interface with a payment provider, albeit smaller than directly integrating with a credit card acquirer, still depends on which services of the provider are used (Section 3.1). Smaller merchants with fewer resources are often encouraged by the providers themselves to start with simple integration methods. Efforts to ensure the security of customer accounts vary across online industries (Bonneau and Preibusch 2010), larger, more popular, and more mature Websites have significantly better overall privacy protection (Bonneau and Preibusch 2009).

Insofar as payment providers bridge trust gaps, they overcome consumers' privacy concerns vis-à-vis the merchant. New privacy issues are introduced, though, which have been explained above. On an institutional level, intermediaries may be a threat to privacy when shaping the ability to transact or adopt privacy-enhancing solutions: Examples include PayPal blocking customers who use Tor, a privacy-enhancing technology to hide their IP address (Lewman 2010). PayPal has also withheld donated funds and prevented further donations to WikiLeaks, an activist organisation concerned about transparency and privacy (Poulsen 2010). Guidance issued by European data protection authorities unanimously classifies payment providers as data controllers rather than data processors, acknowledging the control they exercise in re-purposing customer data (Information Commissioner's Office (ICO) 2014, Isle of Man Information Commissioner 2015).

1.5. Contribution and research questions

Ahead of tightening regulation regarding data minimisation, recognising that online payment handling is a growing market, noting that information privacy is becoming a positive competitive differentiator, we set out to explore the tracking capabilities of online payment providers.

As the first kind of such investigation, the focus is on exploring and describing current practices. We conducted the first industry-wide, empirical survey that quantifies the flows of customer data from $N = 881$ merchants to PayPal. We describe current practices of data proliferation which can soon be deemed privacy leaks.

PayPal is chosen as the most pervasive online payment provider, covering Websites across strata of popularity (PayPal 2014a). We investigate which items of personal data and which transaction details merchants are sharing with PayPal as customers complete their checkout (Figs. 1 and 4). Our goal is to quantify the prevalence of data flows towards PayPal and to measure the amount of data shared above pure order totals. Our survey of the ecosystem also looks for per-sector differences in data sharing with payment providers or whether more popular Websites leak more or less personal details.

2. Related work

Our investigation complements and expands an existing body of literature that has empirically examined privacy and tracking practices at large. Bonneau and Preibusch studied privacy practices across the entire online social networking ecosystem (Bonneau and Preibusch 2009). They found unsatisfactory privacy practices throughout the industry, which were still better for more popular and mature sites. They also investigated data protection practices among different industries (Bonneau and Preibusch 2010) and found that poor practices were commonplace regarding password security, although merchant sites did better than newspaper sites. Specifically for Web shops, more expensive shops were found to collect significantly more personal details than their cheaper competitors (Preibusch and Bonneau 2013).

A number of Web privacy surveys studied the private information leakage, different tracking mechanisms and their prevalence on the Web. Krishnamurthy and Wills show how personally identifiable information leaks via online social networks, including the leakage by HTTP Referer header (Krishnamurthy and Wills 2009a). Roesner et al. presented a taxonomy of third-party tracking and developed tools for defending against tracking by social sharing buttons (Roesner et al. 2012). Multiple researchers surveyed the use of more advanced and resilient tracking mechanisms such as evercookies (Soltani et al. 2010, Ayenson et al. 2011, McDonald and Cranor 2011, Acar et al. 2014), browser fingerprinting (Acar et al. 2014, Eckersley 2010, Nikiforakis et al. 2013, Acar et al. 2013) and cookie syncing (Acar et al. 2014, Olejnik et al. 2014), commonly reporting on questionable practices and unexpected prevalence of such technologies.

Researchers studying tracking on mobile platforms found that many apps leak private information to third-party servers including precise location, personal data and unique identifiers (Enck et al. 2014, Hornyack et al. 2011, Gibler et al. 2012, Egele et al. 2011). A study intersecting the interface used for embedding mobile ad libraries found that apps share highly sensitive information such as ethnicity along with postal code, gender, age and income (Book and Wallach 2013). The same study found a positive correlation between app popularity and the privacy leakage. More recently, by analysing the unprecedented amount of 1.1 million Android apps, Viennot et al. showed how apps mishandling of authentication tokens may lead to unauthorized access to user data and resources on Amazon Web Services and Facebook (Viennot et al. 2014).

Another line of research has looked into users' reaction to online tracking and behavioural advertising. A 2013 Pew Research study found, motivated by the concerns about online tracking, that 86% of Internet users have tried to be anonymous online and took some effort to avoid tracking (Rainie et al. 2013). Ur et al. found a majority of users in their study were either fully or partially opposed to online behavioural advertising, finding the idea smart but creepy (Ur et al. 2012). Leon et al. studied the factors affecting users' willingness to share information with the advertisers and found that perceived sensitivity of information, data-retention policies and the scope of data use are the prominent factors (Leon et al. 2013).

Finally, researchers looked into consumers' privacy choices in online shopping. Buyers of sensitive products (vibrators) were found to pay a premium to shop with a retailer whose privacy practices were labelled as superior by a product search engine (Tsai et al. 2011). In the largest ever lab and field experiment in privacy economics, almost one in three Web shoppers paid one euro extra for keeping their mobile phone number private (Jentzsch et al. 2012). When privacy comes for free, more than 80% of consumers choose the company that collects less personal information (Jentzsch et al. 2012). Earlier results indicated that price discounts override online shoppers' privacy preferences (Preibusch et al. 2013).

3. Methodology

We conducted a blind field experiment with 1200 shopping Websites, by observing their inbound and outbound data flows during checkout. The Websites did not know they were subjected to data capture, which followed a strict experimental protocol. The data collection setup fleshes out the integration with PayPal, which is described first. We then provide details on the sampling, the experimental protocol, and describe additional data sources for data enrichment.

3.1. Background: PayPal integration and information flows

PayPal provides payment processing to merchants and has been a pioneer to offer payment acceptance to electronic retailers, although its product range now covers a plenitude of card and card-less payment and identity services for online, offline, and mobile transactions. Similar to a cloud service, PayPal's offerings are characterised by their ease of set-up, pay per use, and self-service.

PayPal offers multiple ways to be embedded in the shopping workflow, traditionally depending on the type of payment, for instance (e.g., donations, recurring subscriptions, one-time checkouts) (PayPal 2013a). On a technical level, there are two different integration routes depending on how the session data are transmitted from the merchant to PayPal: (1) server-to-server integration, where SOAP Web services or REST APIs are used to communicate transaction details from the merchant to PayPal; (2) integration via the client, where transaction parameters are passed exclusively through the query string (GET) by means of consumers' browsers.

Integration via GET is simple and readily available for hosted Websites, as no server-side communication is required. In PayPal parlance, this integration method is called "buttons". More sophisticated methods use server-to-server communication between the application server and the payment provider: the merchant creates a session with the payment provider when submitting all relevant transaction data. This session is then referenced through a session identifier or token ("EC token"), which is the only information that the client needs to pass on (PayPal 2013b). This method requires more technical expertise, but is less susceptible to manipulation by the client. However, server-to-server communication cannot be observed in a study like ours, where the client is instrumented. It would require server logs from PayPal or the merchant (or broad-coverage network sniffing capabilities). Also, when integrating with PayPal through payment buttons, merchants can still hide submitted information and prevent tampering by encrypting the transaction parameters (PayPal 2013c).

A variant of server-to-server payment integration is the use of a further intermediary that calls and processes the PayPal workflows on behalf of the merchant. Such an intermediary is typically found for more complex integration, to mediate between multiple payment methods.

Whereas encrypted buttons are encountered rarely, payment sessions referenced via an EC token on the client side are very common (Table 3). The unobservable flow of personal information between servers is a challenge for our research. We therefore use personal data that PayPal displays back to the user to establish a lower bound for the privacy invasion by the data that are transmitted; this method was confirmed to be accurate during data analysis (Section 4.5).

The "Legal Agreements for PayPal Services" (PayPal 2014b) outline a number of requirements for merchants. All information submitted to the API must be "true, correct, and complete" (PayPal 2013d). Whereas all fields containing personal information are optional (PayPal 2014c,d), a "description field to identify the goods" and a URL linking back to the original product page must be provided for the popular Express Checkout method (PayPal 2014c).

3.2. Sampling

We sample online shops that target US consumers and provide checkout in US Dollar via PayPal. The US market is chosen for its size and for being the home market of PayPal. We sample popular Web shops. Practices at these online destinations matter most as they impact a large consumer population. Stores are identified by their URL, as occurring before the PayPal checkout page in browser sessions. Data are collected from a sample of Internet Explorer users who opted into share their browsing history.

Sampling originally yielded Website domains rather than product pages for each potential shop. These domains were ordered and processed in decreasing order of popularity and inspected manually as Websites may have ceased to offer PayPal, may have shut down, may be unreachable or otherwise no longer qualify by our sampling criteria. In particular, sampling by referrer produced false positives, such as Webmail providers or search engines. These sites were noted and excluded. All domains were inspected manually and if the site matched the sampling criteria, a single product was selected to measure data leakage during checkout. Product selection followed a simple protocol, choosing the first available product in the first product category except sale or seasonal categories.

We excluded Websites offering business services (B2B such as email marketing campaigns), banks and insurances, and restricted Websites which required a prior customer relationship such as utility companies. Airline Websites were often excluded for we were unable to complete the purchase according to our data collection protocol. EBay, PayPal internal and duplicate Websites were excluded (e.g., homedepot.ca as a duplicate of homedepot.com).

We deliberately included Websites selling non-material goods such as in-game purchases for extras or virtual currencies, or online account top-ups. In line with our research agenda, we also refrained from filtering out adult Websites.

Hosting sites (e.g., Yahoo! shops or Google Sites) were excluded and separated from the sample for future analysis. Such sites host multiple shops with differing implementation practices under a single domain. A few representative sub-shops were chosen for affiliate shops (e.g., spreadshirt.com) and shop-in-shop solutions (e.g., atgstores.com).

4. Experimental protocol

Before the main data collection, we ran a pilot study with separate 40 Websites sampled from the DMOZ/Open Directory Project in the (English) electronics Web shops category. This seeding sample covers a broad range of lesser known online retailers, which we inspected manually whether they offer checkout via PayPal or not. Based on this pilot, we then established the following data collection infrastructure and process (Fig. 2).

For reliable results, a strict data collection protocol was followed during the main data collection. The details of the experimental setup and procedures are laid out in the Appendix. To avoid contamination of the results by residual cookies or other re-identification methods, a virtual machine was used and reset for every recording anew. Transaction data were recorded by navigating in the Web shop and to PayPal to the point of checkout; browsing was done in Firefox and all HTTP and HTTPS traffic was captured by mitmproxy (Mitmproxy Project 2013) and stored. This includes GET and POST requests and the parameters submitted with them. Web forms were completed by using the same fictitious profile data on every site, a woman in her 40s living in a major US city. A unique email address was used for each Website.

The entry point for each Website was the first available product in a product category ('department'). We excluded volatile categories such as featured items or sale promotions. Recording started at the product page identified during sampling and finished with the PayPal checkout screen. An example of this screen is given in Fig. 1. We captured and archived this PayPal screen (screenshot image and page mark-up) and manually tallied the presence of personal and product details displayed on the page. We later tested and confirmed the accuracy of inferring data transfer from the screenshot (Section 4.5).

Problems that we encountered during the data collection were recorded and dealt with consistently. For broken or unavailable Websites, we progressed as far into the purchase as possible. Unavailable items were substituted by the next available product according to the product sampling procedure. Websites that refused the existing profile data for any reason, were recorded but later excluded from the dataset. Some Websites redirected to a non-US version of the shop based on IP address geo-location, in which case we tried to navigate back to the US store. When possible, we completed the checkout without registering with the shops. Although data collection is tool-supported, there is always a human in the loop.

After the full data analysis was completed, we reached out to a sub-sample of online retailers to explain their point of view.

4.1. Data enrichment: adding metadata

To analyse privacy friendliness by industry, we manually and automatically annotated the Websites in our sample with the kind of service they offer and the products they sell. We also added metrics for popularity and technical quality.

4.2. Manual Categorisation of Web shops

We manually categorised the Websites as specialised in *retail*, *commercial services*, *donations*, *dating*, *events*, *airlines*, or *other* if they did not fit in any of these categories. The *commercial services* Websites were further subcategorised as *educational*, *software*, *Website*, or *other*.

Retail Web shops sell physical goods, while *commercial services* are selling primarily nonphysical goods such as courses (*educational*), online-backups (*software*), or access to a Website (*Website*). *Donation* Websites do not provide anything in return for money, and technically need not even register the donor's name. Although small in number, the categories *events* and *airlines* were included because they by nature will require more personal information than a retail store.

4.3. Automated Categorisation of Web shops

For a more fine-grained, product-driven categorisation of Web shops, we turned to the curated ontologies of DMOZ and AWIS (Alexa Web Information Service). They both turned out to only have data for the most popular Websites (35% coverage from AWIS keywords, 48% from AWIS categories, 52% from DMOZ—53% when combining DMOZ and AWIS), so we looked for an alternative solution.

Exploiting Amazon.com's status as one of the biggest warehouses of the Internet, we used the Amazon.com Product Advertising API (Amazon Web Services 2013) to obtain a better classification. This is a novel approach which has been applied for the first time in our investigation. For every URL in our dataset, we queried the API with the HTML Meta keywords describing the merchant Website. This yielded a list of up to ten matching products, for each of which one or more product categories were listed. Amazon product categories are a graph structure, which can be

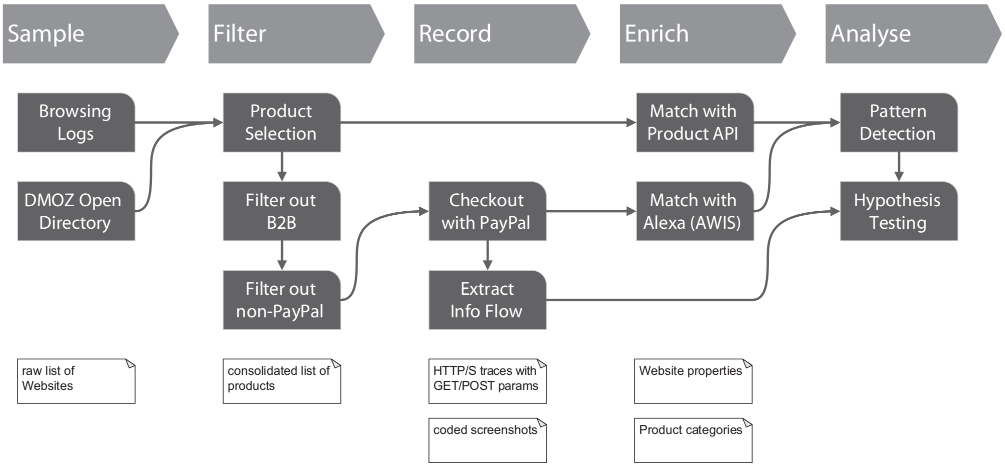


Fig. 2. Steps of the data collection process; data collections are given in the lower portion.

browsed like a tree, similarly to DMOZ. We applied a voting algorithm over the returned items for each Website; the items were sorted in descending relevance by the Amazon API.

The API also assigns exactly one node of Amazon's product type and product group type categories to each product. However, not much useful information could be extracted from these, and we largely discarded them from further investigation.

Although the Amazon product ontology was used for describing Website categories, we retained AWIS to assess popularity and technical quality, indicated by the 'traffic rank' and 'speed percentile' respectively.

4.4. Data sources

In summary, we use the following data sources for our analysis:

- A list of online retailers offering PayPal checkout, sampled from actually visited Web shops in real browsing sessions, as per the method described above.
- Website metadata from external sources: (1) Website popularity, maturity and audience demographics licensed from the Alexa Web Information Service (AWIS) API; (2) manual and automated merchant categorisation via Website keywords and Amazon Product API.
- From captured HTTP traffic to the server: a trace of all personal information sent from the merchant to PayPal via query string or the presence of a server-initiated transaction session, as detected by the submission of a session token. Transcribed screen-shots of the PayPal login page, showing personal data that PayPal displays back to the customer, which was confirmed to equal data received from the merchant.
- From captured HTTP traffic from the server and from saved HTML pages: data on third-party trackers (Omniture) deployed on PayPal's checkout pages.
- Written responses received from the merchants we asked for clarification on their data sharing practices with PayPal.

4.5. Data analysis

From an initial list of 1200 shopping domains, we successfully collected the data for 881 merchant Websites: HTTP(S) traffic traces until reaching the PayPal login page, and screenshot upon arrival. These parsed logs and transcribed screenshots were confirmed to be accurate evidence of personal identifiable information (PII) leakage.

First, we describe our data set in terms of PayPal API implementations and predominant patterns of PII leakage. We also explain how we used machine learning techniques to reduce the multitude of these patterns to a manageable number. Second, we explain the data enrichment we performed. We looked for predominant practices on the Internet by adding metadata to our data set, allowing us to slice the data set by Website and product categories.

4.6. Descriptive statistics and cluster analysis

4.6.1. Endpoints and tokens

As described in the Background section, there are two basic methods for a Web shop to communicate with PayPal: using a token or using GET parameter transmission. For both methods, the logs indicate there exist a number of different PayPal API endpoints for the Web shops to use. Table 3 shows their distribution over our dataset.

The most predominant endpoint (1) is the only one that is currently mentioned in the PayPal API help documents (PayPal 2014e). The second-most used endpoint (2) appears to be an older endpoint; although we have no PayPal document confirming this, there is exactly one mention in the online help. From the context, it stands to reason that this was merely an oversight when updating the documentation. Endpoint 3 seems to exist to catch typos; there is no mention of it in any PayPal documents. Endpoint 4 has a more distinctive name, but is likewise undocumented. Finally, endpoint 5 is a localized version ("/bg"). We see multiple calls to such localized endpoints in the logs, but they generally forward all data to endpoint 1. This Bulgarian endpoint did not, so we included it as a special case.

Token usage is widespread: For endpoint 1, more than 84% of all Websites employ one, and 86% across all endpoints. For such Websites, we have to rely on the screenshots because PII leakage cannot be inferred from the HTTP GET logs.

4.6.2. Data accuracy

The PayPal API accepts various parameters, a subset of which is reserved for the customer PII. While some of this is readily displayed on the PayPal login screen before payment, we also investigated whether any PII was sent to PayPal by parsing the log files. Based on the PayPal API help documents (PayPal 2014e), we determined all relevant parameters and parsed the logs for occurrences.

Table 1
Leaked data by clusters ranked from good to bad privacy practices. The common leakage of product details is more worrying than the seeming absence of customer data: PayPal collects identity details directly during payment. Leaked by: □ = some sites, ■ = all sites, blank = no sites in that cluster.

	Site count	Cluster description	Address	Email	Phone	Shipping	Quantity	Prices	Description	Prod. Name	Leak min	Leak max
C_1☺	391 (44%)	Leaks nothing or at most one item.	□				□		□		0	1
C_2😐	34 (4%)	Leaks two of names, item numbers, and prices.				□	□	□	□		1	3
C_3😐	292 (33%)	Leaks at least names, item numbers, and prices.					□	□	□	□	3	4
C_4☹	155 (18%)	Leaks at least most product details and always shipping costs.				■	□	■	□	□	4	5
C_5☠	9 (1%)	Leaks name and address in addition to product details.	■	□		□	□	■	□	□	6	7

To verify our screenshot-based approach, we investigated whether the PayPal login screen always displays all PII that the API receives over the GET query-string. We were able to confirm that whenever customer or product data were leaked via GET, it showed up on the PayPal login screen. The only exception was for shipping costs of USD 0.00, which was forwarded but hidden in 36 cases. This means the transcribed screenshots give an accurate account of transmitted customer data.

4.6.3. Pure leakage patterns

After aggregating the screenshot and log data, three pure leakage patterns emerge. These pure patterns account for a total of 805 URLs (91%). The remaining 76 URLs form a long tail of 25 patterns, none of which occurs more than 13 times, and 15 patterns occur exactly once. With 338 out of 881 sites, the predominant pure pattern leaks no data at all. The other two both leak the item names, descriptions, numbers and the customer's name to PayPal, with the third also informing PayPal about the shipping costs.

4.6.4. Clustering of all leakage patterns

The leakage patterns form the backbone of our work. In order to analyse the data more deeply, we shorten the long tail of these patterns and reduce the number of distinct patterns by clustering all 881 URLs into only a few classes.

While the three pure leakage patterns identified above have very intuitive descriptions, a *k*-Means clustering (Filkov and Skiena 2004) failed to identify similarly meaningful patterns for any value of *k*. We thus resorted to EM clustering (Dempster et al. 1977), which automatically determines the appropriate number of clusters. The result is shown in Table 1.

4.6.5. Integration patterns by cluster

A natural question is whether a particular combination of endpoint and token usage enforces or prevents leakage. Analysing the clusters, it becomes obvious that there is no such relationship: None of the clusters are homogeneous with respect to endpoints and tokens, except for C_2, which does not contain any token implementations.

Privacy-friendly Websites tend to use a token more often: 98% of all Websites in Cluster C_1 were using a token, compared to 86% and 85% for C_3 and C_4, respectively ($p < 0.0001$, two-tailed Fisher's exact test).

We observe that no Websites leaking customer addresses rely on a token implementation. With a sample size of nine this holds little statistical significance, but we found no indication in the API documentation that this is a requirement on PayPal's side.

We used the association rule mining algorithm Apriori (Hipp et al. 2000) to see whether the cluster membership of a Web shop correlates with its implementation. Requiring a confidence of at least 0.4, the resulting rules did not consistently link the usage of a token or an endpoint to any degree of privacy-friendliness. We conclude that PayPal's available API methods do not bias Web shops to treat customers' privacy in a specific way.

4.6.6. Leakage patterns by Website category

Table 2 shows a breakdown of the Website categories. The distribution of Website categories per cluster largely reflects the overall distribution over the data set. No clear trends can be identified, although a large majority of *donation* sites sit unnecessarily in the privacy-unfriendly C_3. Commercial services are also mainly found in this cluster. While the leaked PII is necessary for this category of Websites, we do not see an immediate need to forward them to PayPal.

Prediction from the category of the Website showed no promise. With the exceptions outlined above, all clusters proved to be too mixed, and Apriori failed to produce rules with a confidence of even 0.4. Hinting towards a variety of attitudes towards privacy among Web shops, this is a positive result for customers.

Looking at the product categories associated with the Websites found in each cluster, C_3 contains categories from all over the category tree. C_5 and C_2 generally have too few entries to make

Table 2
Manual Website categorisation, including sub-types for commercial services.

Type	Site count		Subtype	Site count
Retail	764			
Commercial services	66	→	Website	20
Donations	25		Software	19
Dating	9		Educational	11
Events	8		Other	16
Airlines	6			
Other	3			

Table 3
Distribution of PayPal endpoints used by merchant sites.

	PayPal endpoint URL	Site count (with/ without token)
1	https://www.paypal.com/cgi-bin/webscr	847 (731/116)
2	https://www.paypal.com/webscr	25 (18/17)
3	https://www.paypal.com/cgibin/webscr	4 (4/0)
4	https://www.paypal.com/checkoutnow	4 (4/0)
5	https://www.paypal.com/bg/cgi-bin/webscr	1 (1/0)
	Total	881 (758/133: 86% with token)

substantiated claims, but books are clearly predominant. C_1 and C_4 both contain a lot of clothing and related items. While C_1 has some home appliances in it, C_4 has a tendency towards personal health products, including sports products.

Looking at Website popularity and quality, we found that technical implementation quality has no immediate bearing on cluster membership. Rather, we see that the number of sites from a certain cluster scale with the overall number of sites in the speed percentile. We further see that the distribution of sites from the clusters over the percentile bins follow no specific pattern. It can thus not be said that the speed of a Website has a positive correlation with its privacy-friendliness.

Less popular sites are found significantly more often in clusters that exhibit more leakage. More popular sites tend to leak less. For illustrative purposes, the average traffic rank is 0.4 million for C_1, 1.0 million for C_3 and 1.4 million for C_4. A Mann–Whitney U test indicates a highly significant difference in the traffic ranks per cluster ($p = 0.001$ for both pairwise comparisons). Sites in the worst leakage C_5 do not appear among the 50 highest ranked in our sample (Fig. 3).

4.6.7. Third-party tracking facilitated by PayPal

Analysis of the HTTP traffic observed during the experiments revealed the use of Adobe's Omniture tracking software on PayPal checkout pages. When a user lands on the PayPal checkout page, two HTTP requests are sent to paypal.d1.sc.omtrdc.net and paypal.112.2o7.net subdomains, both of which belong to Adobe's Omniture tracking software (Adobe Systems Incorporated, Digital marketing|Adobe Marketing Cloud 2014). The requests contain metadata about the payment to be made, such as currency and transaction token, along with the user's browser characteristics such as plugins, screen dimensions and software versions (Adobe Systems Incorporated 2014). Remarkably, PayPal also shares the referrer URL of the checkout page, which reveals the URL of the Web shop, and potentially the product to be purchased. This leakage enables Adobe to build a better profile of 152 million PayPal users (PayPal 2014a), by combining payment details with other online activities recorded on more than 300,000 Omniture-tracked Websites (BuiltWith Pty Ltd 2014), which notably includes 50 of the Web shops analysed in this study.

Note that the leakage described here is different from the indirect information leakage via referrer headers as studied in Krishnamurthy and Wills (2009b), since the PayPal checkout page actively collects and sends the referrer of the checkout page, which would not be shared otherwise with the Omniture domains. Furthermore, by sending high-entropy browser properties such as plugins and screen dimensions, PayPal makes it possible for Omniture to track users by their browser fingerprints even if they block or delete their cookies (Eckersley 2010).

According to its privacy policy, PayPal may share customers' personal information with third-party service providers (PayPal 2013e) who are limited to using PayPal customers' information "in connection with the services they perform for [PayPal]." Assuming the information shared with Omniture is subject to a similar agreement, it is hard to make sure whether payment information, product URL or browser characteristics are interpreted as personal information or not, given the possible interpretations of the policy and lack of transparency around PayPal's contracts with third-parties.

As of September 14th, 2014, long after we finished with the experiments, the PayPal checkout page no longer references a third-party tracker, though Omniture is still used on the PayPal homepage (see Appendix for reproduction).

4.6.8. Flash evercookies and browser fingerprinting for internal tracking

We also found the use of two questionable tracking mechanisms by PayPal internally, namely, evercookies and browser fingerprinting. Upon registering for a new account, PayPal places a Flash cookie named paypalLSO.sol, which includes a 70 character long identifier string. On each visit to PayPal checkout page, this Flash cookie is read by an invisible Flash object (mid.swf) and appended to the payment form as a hidden element.

Evercookies (also known as supercookies or zombie cookies) make use of obscure browser storage mechanisms to store tracking identifiers. Being a resilient tracking technology, evercookies can be used to restore standard (HTTP) cookies intentionally removed or blocked by users. In the past, use of such techniques led to lawsuits and multi-million dollar settlements (Singel 2010).

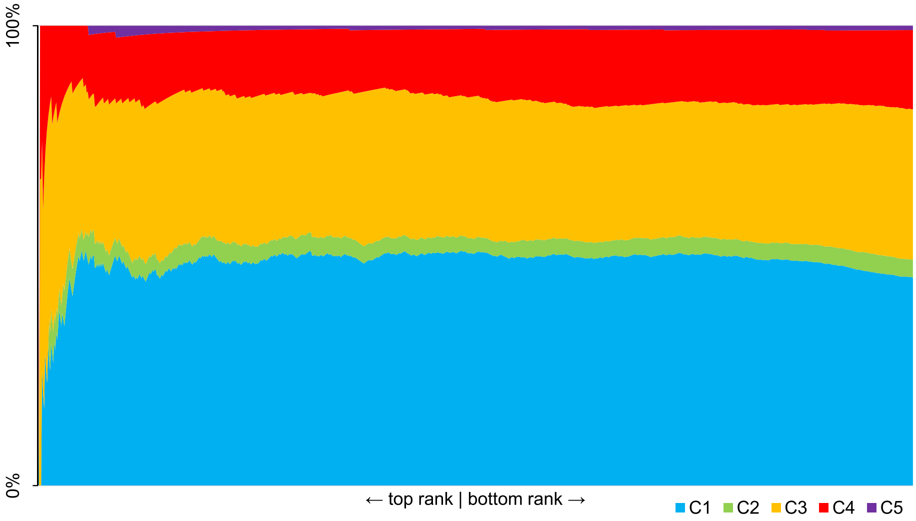
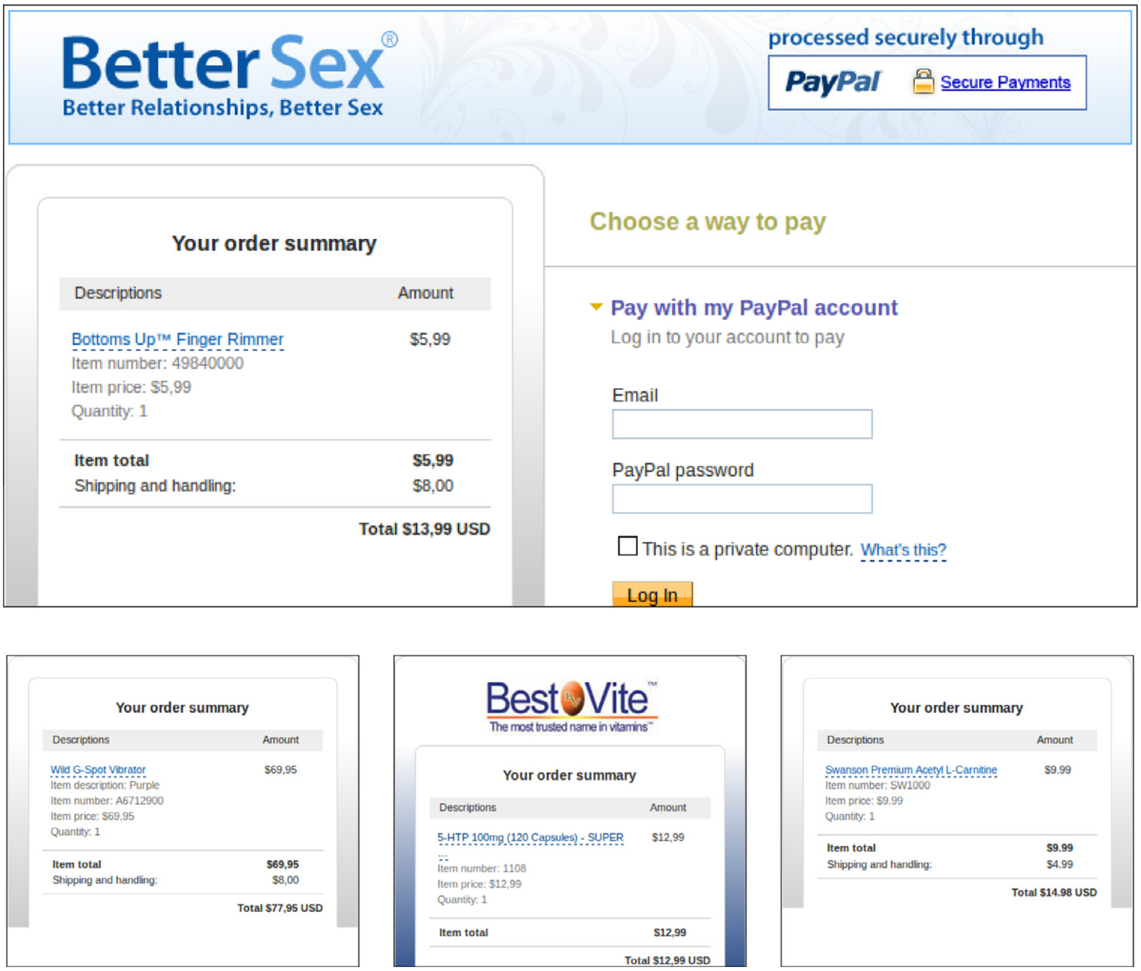


Fig. 3. Cluster membership over the sites' popularity (traffic ranks) found in the sample.



Better Sex®
Better Relationships, Better Sex

processed securely through
PayPal Secure Payments

Your order summary

Descriptions	Amount
Bottoms Up™ Finger Rimmer Item number: 49840000 Item price: $5,99 Quantity: 1	$5,99
Item total	**$5,99**
Shipping and handling:	$8,00
	Total $13,99 USD

Choose a way to pay

Pay with my PayPal account
Log in to your account to pay

Email

PayPal password

□ This is a private computer. What's this?

Log In

Your order summary

Descriptions	Amount
Wild G-Spot Vibrator Item description: Purple Item number: A6712900 Item price: $69,95 Quantity: 1	$69,95
Item total	**$69,95**
Shipping and handling:	$8,00
	Total $77,95 USD

BestVite™
The most trusted name in vitamins™

Your order summary

Descriptions	Amount
5-HTP 100mg (120 Capsules) - SUPER ... Item number: 1108 Item price: $12,99 Quantity: 1	$12,99
Item total	**$12,99**
	Total $12,99 USD

Your order summary

Descriptions	Amount
Swanson Premium Acetyl L-Carnitine Item number: SW1000 Item price: $9.99 Quantity: 1	$9.99
Item total	**$9.99**
Shipping and handling:	$4.99
	Total $14.98 USD

Fig. 4. Also sites that sell sensitive products leak product details to PayPal. Two examples of adult toys (finger rimmer, vibrator) and medication: 5-HTP addresses depression, anxiety, sleep disorders, and MDMA hangover; acetyl carnitine is used for many indications including Alzheimer's disease and depression.

4.6.9. Browser fingerprinting is another advanced tracking mechanism employed by PayPal

Through a field study, EFF's Panopticlick study showed that combining multiple browser properties, one can extract a fingerprinting that can be used to track users without relying on the stored identifiers such as cookies (Eckersley 2010).

When a user visits the PayPal checkout page, a script (pa.js) collects multiple browser properties including browser plugins, screen dimensions and 36 different Windows ActiveX components to get information about the installed software and language packages. The Appendix lists the properties collected by PayPal's script. Unfortunately, PayPal's privacy policy is not as explicit about fingerprinting as it is for Flash cookies.

Despite using the same technology, the assessment of PayPal's user re-identification through persistent cookies has to be more favourable than third-party online tracking. In a payment scenario, these advanced tracking techniques can be used to prevent account hijacking or similar fraudulent activities. PayPal's privacy policy explains that "cookies, pixel tags, 'Flash cookies,' or other local storage [...]" are used to "recognize you as a customer; customize PayPal Services, content, and advertising; measure promotional effectiveness; [...] mitigate risk and prevent fraud [...]." (PayPal 2013e). Nonetheless, for a company that manages millions of payments each day, there are many improvements to make, beginning from preventing information leakage to third parties, being more transparent about in-house tracking mechanisms and strictly isolating the use of advanced tracking tools to combat fraud.

4.6.10. Responses received from the merchants

After the full data analysis was completed, we tried to contact a sample of 59 online retailers for which the proliferation of purchase details could be particularly embarrassing, including adult entertainment, pest control, shape-wear, vitamins and medication. We used the same text across all shops in our request, and asked: "What data do you transfer to PayPal when a customer of your shop decides to use that payment option for a purchase?" We clearly mentioned "the context of a scientific study on online payment providers".

Most retailers could be contacted by email (38 shops, 64%); for 17 shops (29%), we had to use an online contact form. Four shops (7%) provided no means of getting in touch. A single merchant replied the next day and explained that they "transfer the absolute minimal data that is required by Paypal" (sic!), which however was not consistent with our records. No other shops replied to our enquiry within four weeks (or any time after for that matter); twelve auto-replies were received but never followed by an actual response.

The overall lack of responsiveness from the retailers is bad news for consumers who are left alone with their privacy concerns. Furthermore, the only retailer who made an effort to reply seemed unaware of more far-reaching data proliferation. This might suggest that merchants need tools to identify data flows, and better guidance on how to implement a privacy-friendly checkout procedure. Our article aims to help with both issues.

4.6.11. Limitations

As outlined above, our sampling strategy combined Web shop URLs from different sources to cover both larger and smaller merchants. We expect our dataset to contain an equal distribution over more and less professional Websites, as well as more and less frequented ones. The Websites we analysed are sites that are actually visited by Web users, as the sampling was guided by browsing session logs.

This comes at the price of diversity of goods that are sold. As expected and confirmed by our manual categorisation of Websites, there are more Web shops selling physical goods than there are commercial dating Websites. As a result, our categories are not evenly distributed over the dataset at all. This makes statistically significant statements about privacy practices hard, if not impossible. The non-existence of association rules with high confidence is an immediate result of the skewed distribution of categories.

Further breaking down the manually assigned categories into a more fine-grained version was done in a semi-automated way. This is necessarily prone to errors, stemming from incomplete or faulty inferences from the information provided by the Amazon Product Advertising API. Naturally, this API can only deliver information based on Amazon's product range. Labelling our dataset using this information will again be biased: Our method for assigning labels to Websites discarded relatively infrequent labels. The biggest problem here is not primarily that we may not have enough shops per label in our dataset, but rather that the labels may be infrequent because Amazon has few goods from a certain category. Again, under a skewed distribution, statistical significance is difficult to obtain.

For obvious reasons, our data collection setup could not cover server-to-server communication, which, according to PayPal documentation (PayPal 2014e), can be used by merchants to communicate with PayPal. Also, in our experiments we did not go beyond the PayPal checkout page to complete the payments. As a result, the data collected and leaked after the PayPal checkout page are not covered in our analysis.

5. Conclusion and discussion

We presented a new species in the zoo of online tracking systems: explicit leakage of personal information and detailed shopping habits from online merchants to payment providers. In contrast to the widely debated tracking of Web browsing, online shops make it impossible for their customers to avoid this proliferation of their data.

By mediating online payments between merchants and buyers, payment providers are in a position to access sensitive payment details that can be used to build a detailed profile of shopping habits. Being the most popular payment provider, PayPal learns how much money its 152 million customers are spending and where. These customers are identified by name, email and postal address and through their bank details. We have demonstrated that merchant Websites are unnecessarily forwarding product details to PayPal that give a detailed view on consumers' purchases.

According to the 881 sites studied in our analysis, 52% of the most popular US Web shops shared product names, item numbers and descriptions with PayPal. Besides the negative privacy impact, consumers whose data are proliferating could suffer from less favourable payment terms (e.g., unavailable payment methods of higher interest rates on consumer loans based on their purchase patterns). On the other hand, the remaining 388 sites did not share any purchase details except the amount to be paid, confirming that sharing sensitive details is not necessary for electronic retailers.

Further, we reported on the PayPal's use of the tracking service Omniture, which amplifies the privacy concerns by exposing transaction details to a widely deployed third-party tracker. A third-party tracker that has access to general Web tracking information, as well as to the details of successfully completed transactions, is in a particularly privileged situation to monitor consumption choices at large.

Web shops that use the technically more advanced token-based integration are often more privacy-friendly. Also, less popular sites are significantly more often among those that leak more personal information. There are no systematic differences across product categories, meaning that all kinds of shoppers are exposed.

To the extent that PayPal, as an example of payment providers in general, collects personal information at scale, it becomes a constituent part of the online shopping experience: neither researchers nor enforcement authorities can reduce its role to a passive intermediary when assessing the privacy impact of e-commerce transactions.

By exploring the alternative privacy preserving practices that can be followed by Web shops, we distilled the following suggestions for merchants: (1) apply the data minimization principle—do not leak information that is not required for processing the transaction; (2) inform customers about the data sharing in your privacy policy; (3) offer alternative, privacy-friendly payment methods, such as direct debit or pre-payment; (4) use a payment gateway to prevent leakage of product URL via referrer header.

Future research through qualitative interviews with decision-makers and engineers at merchants should look at the drivers and motives behind PayPal integration choices and their privacy consequences. On the technical side, expanding the scope to mobile and in-app payments promises valuable for these growing, yet opaque transactions. Better privacy practices for handling online payments are not only desirable for end users, but also for the merchants and payment providers whose businesses depend on the users' trust.

At times when personal information is said to be new currency on the Web, it seems unfair that consumers are charged twice during checkout.

Appendix

A sample HTTP request collected during the experiments

(a) Request URL https://paypal.d1.sc.omtrdc.net/b/ss/paypal-global/1/H.25.3/s68449009894746?AQB=1&ndh=1 [trimmed – see below].

(b) Request parameters (URL decoded)

```
AQB: 1
ndh: 1
t: 16/5/2014 3:37:43 1
-120
fid: 09476854BACDB25F-2
A6965C4F340BABA
vmt: 51437A79
vmf: paypal.112.2o7.net
ce: UTF-8
ns: paypal
pageName: main:ec:::sta
rt
g: https://www.paypal.c
om/cgi-bin/webscr?cmd=_
express-checkout&token=
EC-81F4649270038960T
r: https://www.poolpart
sonline.com/ShoppingCar
t.aspx?add=true&ReturnU
rl=%2fp-59779-covers-um
brella-furniture.aspx
cc: USD
c: 8
j: 1.8.5
v: N
k: Y
bw: 1024
bh: 613
ch: ec
server: main
products: ;ec
c1: xpt/Checkout/ec/Log
in
c5: 2P234932RV746033J
c7: none
v7: none:none:none
c8: none
c9: none
c17: Pay with a PayPal
account - PayPal
c19: main:ec:::start
v19: D=c7
c20: 1402882661
c21: EC-81F464927003896
0T
c25: main:ec:::start:me
mber:1: v25: main:ec:::
start:member:1: v28: tn
c-a-ecg-cntl
c30: glb
p: Shockwave Flash;iTun
es Application Detector
;QuickTime Plug-in 7.6.
6;VLC Multimedia Plugin
(compatible Totem 3.0.1
);Windows Media Player
v31: main:ec:::start
c35: out
c36: paypal.com/cgi-bin
/webscr?cmd=_express-ch
eckout
v36: US
c37: member:1:
c39: D=pageName
c40: c97f8013a3fba
c47: D=pageName
c50: en_us
v50: 0nehgENrmdgxT5Wkli
OJGPcyFf6%2bLQoeXAFWsge
tQyDTdJbVzxWcfz6BH%2bVC
LRtF1d9zERVjOXU%3d_146a
25299b6
c53: h.25.3|01.17.2013
c56: yes
c64: 2294ec411f0df
c72: UTF-8
h1: main_ec__
s: 1024x768
Plug-in 10 (compatible;
Totem);DivX® Web Player
;
AQE: 1
```

Listing 1. The request parameters sent to Adobe's Omniture tracking suite domain (omtrdc.net) leaks the product name and ID as a part of the referrer URL (parameter `r`). The plugin details, screen dimensions, browser window size and software versions are collected and sent to the tracking endpoint.

Browser properties collected by the PayPal analytics script pa.js

Script URL: https://www.paypalobjects.com/pa/js/pa.js

- User Agent string (`navigator.userAgent`)
- File name, description and version of installed browser plugins
- Browser plug-ins (`navigator.plugins`)
- All content types that the browser can handle (`navigator.mimeTypes.type`)
- Screen resolution and colour depth, browser window width & height
- JavaScript version
- Version numbers of 36 different Windows ActiveX components such as Outlook Express, MSN Messenger Service and Microsoft virtual machine

Data collection setup

We used a consistent setup to capture the information that merchants share with PayPal when their customers proceed to checkout.

Virtual machine: We used a clean virtual machine for each session in a best effort to prevent profiling by cookies or other client side data.

Proxy: We used mitmproxy (Mitmproxy Project 2013) for intercepting and recording Web traffic. By adding mitmproxy's certificates to the browser, we recorded all HTTP(S) requests and responses in decrypted form, including message bodies. We stored the network dump (.dmp) generated by mitmproxy to enable playback of the exact Web traffic. This helped us to ensure the reproducibility of our analysis.

Browser: We used a Firefox browser (version 25.0), configured to relay all communications through the intercepting proxy. The browser configuration was slightly modified from the default, as explained below; however, all privacy-related settings were left unchanged, simulating a default user.

Browser add-ons: Data collection was supported by browser add-ons for auto-filling forms and capturing the screen. 'Autofill Forms' (https://addons.mozilla.org/en-us/firefox/addon/autofill-forms/) helped us by quickly providing various information we are expected to fill in to forms on shopping sites. The use of the form-filler also ensured the same profile data was used on every site. We used 'Screengrab' (https://addons.mozilla.org/en-US/firefox/addon/screengrab-fix-version/) to take the screen capture of the PayPal page we were redirected to complete the payment.

Browser configuration: The browser configuration was kept closely to the default to mimic the privacy exposure of a mainstream user. We therefore kept all privacy-related settings unchanged and used the default non-private browsing mode. We continued to allow popups, Flash, Silverlight and Java (if available) and did not actively block script execution. We allowed and recorded requests to phishing/malware databases, which are enabled by the original settings.

At the same time, the following options were turned off to prevent cluttering of the recorded Web traffic: auto-update search engines, spell checking, crash report, Firefox health report and OCSP certificate verification.

Proxifying script: We used a Python script to launch mitmproxy and the browser with the product URL. The same script was used for parsing the network dump captured by mitmproxy and for outputting the captured HTTP(S) requests and responses for further analysis. The logs generated by the script were fed to a parser script that mined the data in the Web traffic and detected information flows.

Reproducing the original Omniture tracking on PayPal's checkout pages

Note that the PayPal checkout pages from the Internet Archive can be used to validate our findings about Omniture tracking:

https://web.archive.org/web/20140228095312/https://www.paypal.com/cgi-bin/webscr?cmd=_flow&SESSION=_4T7tr0uKMzmNjcRwv_KSFh0Dminf5Qm11xcKqg33aOA_Q80mcRJXTVjTxK&dispatch=50a222a57771920b6a3d7b606239e4d529b525e0b7e69bf0224adecfb0124e9b61f737ba21b0819827f0298a8d8382cff5df9729c4c3c2b2.

References

Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F.A.P.B., 2013. FPDetective: Dusting the web for fingerprinters. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security.

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., Diaz, C., 2014. The Web never forgets: Persistent tracking mechanisms in the wild. In: Proceedings of CCS.

Adobe Systems Incorporated, 2014. SiteCatalyst variables and query string parameters. [Online]. Available: <http://helpx.adobe.com/analytics/using/digitalpulse-debugger.html#id_1298>.

Adobe Systems Incorporated, Digital marketing|Adobe Marketing Cloud, 2014. [Online]. Available: <http://www.adobe.com/solutions/digital-marketing.html>.

Amazon Web Services, Inc., 2013. Understanding BrowseNode Results When Drilling Down," 1 August 2013. [Online] Available: <http://docs.aws.amazon.com/AWSECommerceService/latest/DG/UnderstandingBrowseNodeResultsWhenDrillingDown.html>.

Apple Inc., 2014. Apple – iPhone 6 – Apple Pay, 2014. [Online]. Available: <http://www.apple.com/iphone-6/apple-pay/>.

Arnab, A., Hutchison, A., 2007. Using payment gateways to maintain privacy in secure electronic transactions. In: New Approaches for Security, Privacy and Trust in Complex Environments, Boston.

Ayenson, M., Wambach, D.J., Soltani, A., Good, N., Hoofnagle, C.J., 2011. Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning, SSRN.

Bailey, J.P., Bakos, Y., 1997. An exploratory study of the emerging role of electronic intermediaries. Int. J. Electronic Commerce 1 (3), 7–20.

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

Bonneau, J., Preibusch, S., 2010. The password thicket: technical and market failures in human authentication on the web. In: Ninth Workshop on the Economics of Information Security (WEIS).

Book, T., Wallach, D.S., 2013. A Case of Collusion: A study of the interface between ad libraries and their apps. In: Proceedings of the Third ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM).

BuiltWith Pty Ltd, 2014. Websites using Omniture SiteCatalyst. [Online] Available: <http://trends.builtwith.com/websitelist/Omniture-SiteCatalyst>.

Council of the European Union, 2015. Proposal for a Regulation of the European Parliament and the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data – Analysis of the final compromise text with a view to agreement. Presidency to Permanent Representatives Committee, 15 December 2015. Available: <http://www.statewatch.org/news/2015/dec/eu-council-dp-reg-draft-final-compromise-15039-15.pdf>.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39 (1), 1–38.

Duhigg, C., 2012. How Companies Learn Your Secrets, 16 February 2012. [Online] Available: <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=2&pagewanted=all>.

Eckersley, P., 2010. How unique is your web browser? In: Proceedings of the 10th International Conference on Privacy Enhancing Technologies (PETS).

Egele, M., Kruegel, C., Kirda, E., Vigna, G., 2011. PiOS: detecting privacy leaks in iOS applications. In: NDSS.

Enck, W., Gilbert, P., Chun, B.-G., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.N., 2014. TaintDroid: an information flow tracking system for real-time privacy monitoring on smartphones. Commun. ACM 57 (3), 99–106.

European Commission, 2012. Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).

Filkov, V., Skiena, S., 2004. Integrating microarray data by consensus clustering. Int J Artificial Intelligence Tools 13 (4), 863–880.

Financial Fraud Action UK, 2013. Fraud the facts 2013.

Finextra Research, 2009. Disneyland Paris to test contactless payments, 29 July 2009. [Online]. Available: <http://www.finextra.com/news/fullstory.aspx?newsitemid=20321>.

Gibler, C., Crussell, J., Erickson, J., Chen, H., 2012. AndroidLeaks: automatically detecting potential privacy leaks in android applications on a large scale. In: Trust and Trustworthy Computing. <http://link.springer.com/chapter/10.1007/978-3-642-30921-2_17>.

Google, 2014. Google Wallet – Shop in Stores. [Online] Available: <http://www.google.com/wallet/shop-in-stores/>.

Gustafsson, K., Magnusson, N., 2014. Risk Algorithm Paves Global Expansion for Klarna Payment System, Bloomberg, 2 February 2014. [Online]. Available: <http://www.bloomberg.com/news/articles/2014-02-02/risk-algorithm-paves-global-expansion-for-klarna-payment-system>.

Hamblen, M., 2012. Starbucks invests £16m in US mobile payment venture, 9 August 2012. [Online]. Available: <http://www.computerworlduk.com/news/mobile-wireless/3374970/starbucks-invests-25m-mobile-payment-venture-in-us/>.

Heck, E.V., Vervest, P., 1998. Web-based auctions: how should the chief information officer deal with them. Commun. ACM 41 (7), 99–100.

Hipp, J., Güntzer, U., Nakhaeizadeh, G., 2000. Algorithms for association rule mining—a general survey and comparison. ACM SIGKDD Explorations Newsletter 2 (1), 58–64.

Hoffman, D.L., Novak, T.P., Peralta, M., 1999. Building consumer trust online. Commun. ACM 42 (4), 80–85.

Hornyack, P., Han, S., Jung, J., Schechter, S., Wetherall, D., 2011. These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. In: Proceedings of the 18th ACM Conference on Computer and Communications Security.

Information Commissioner's Office (ICO), 2014. Data controllers and data processors: what the difference is and what the governance implications are.

Isle of Man Information Commissioner, 2015. Data Protection Act – Data Controller or Data Processor?

Jentzsch, N., Preibusch, S., Harasser, A., 2012. Study on Monetising Privacy. An Economic Model For Pricing Personal Information. European Network and information Security Agency (ENISA).

Klarna, 2013. Klarna Checkout. [Online] Available: <https://klarna.com/sell-klarna/our-services/klarna-checkout>.

Krishnamurthy, B., Wills, C.E., 2009. On the leakage of personally identifiable information via online social networks. In: Proceedings of the 2nd ACM Workshop on Online Social Networks (WOSN).

Krishnamurthy, B., Wills, C., 2009. Privacy diffusion on the web: a longitudinal perspective. In: Proceedings of the 18th International Conference on World Wide Web (WWW).

Leon, P.G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., Bauer, L., Christodorescu, M., Cranor, L.F., 2013. What matters to users?: factors that affect users' willingness to share information with online advertisers. In: Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS).

Lewman, A., 2010 The team of PayPal is a band of pigs and cads!, 23 August 2010. [Online] Available: <https://lists.torproject.org/pipermail/tor-talk/2010-August/002978.html>.

Malheiros, M., Preibusch, S., Sasse, M.A., 2013. "Fairly truthful": The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In: Trust and Trustworthy Computing.

MasterCard, 2001. MasterCard Corporate Purchasing Card Implementation Guide.

McDaniel, P., McLaughlin, S., 2009. Security and privacy challenges in the smart grid. IEEE Security Privacy 7 (3), 75–77.

McDonald, A.M., Cranor, L.F., 2011. Survey of the Use of Adobe Flash Local Shared Objects to Respawn HTTP Cookies, CMU-CyLab-11-001.

Microsoft, 2014. Wallet FAQ for Windows Phone | Windows Phone How-to (United States). [Online] Available: <http://www.windowsphone.com/en-us/how-to/wp8/apps/wallet-faq>.

mitmproxy project, mitmproxy 0.9 – Introduction," 2013. [Online]. Available: <http://mitmproxy.org/doc/index.html>.

Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., Vigna, G., 2013. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In: IEEE Symposium on Security and Privacy (SP).

OECD, 2013. The OECD Privacy Framework.

Olejnik, L., Minh-Dung, T., Castelluccia, C., 2014. Selling off privacy at auction. In: Annual Network and Distributed System Security Symposium (NDSS).

Palmer, J.W., Bailey, J.P., Faraj, S., 2000. The role of intermediaries in the development of trust on the WWW: the use and prominence of trusted third parties and privacy statements. J. Computer-Mediated Commun. 5 (3).

PayPal, 2013. How would you like to integrate with PayPal? [Online] Available: <https://developer.paypal.com/webapps/developer/docs/>.

PayPal, 2013. Getting Started With Express Checkout. [Online] Available: <https://developer.paypal.com/webapps/developer/docs/classic/express-checkout/integration-guide/ECGettingStarted/>.

PayPal, 2013. Encrypted Website Payments – Technical Overview. [Online] Available: <https://www.paypal.com/us/cgi-bin/webscr?cmd=p/xcl/rec/ewp-techview-outside>.

PayPal, 2013. PayPal Developer Agreement. [Online]. Available: <https://www.paypal.com/us/webapps/mpp/ua/xdeveloper-full>.

PayPal, 2013. Privacy Policy, 20 February 2013. [Online] Available: <https://www.paypal.com/webapps/mpp/ua/privacy-full>.

PayPal, 2014. About PayPal. [Online] Available: <https://www.paypal-media.com/about>.

PayPal, 2014. Legal Agreements for PayPal Services. [Online] Available: <https://www.paypal.com/us/webapps/mpp/ua/legalhub-full>.

PayPal, 2014. SetExpressCheckout API Operation (NVP). [Online] Available: <https://developer.paypal.com/docs/classic/api/merchant/SetExpressCheckout_API_Operation_NVP/>.

PayPal, 2014. REST API Reference – PayPal Developer. [Online] Available: <https://developer.paypal.com/docs/api/>.

PayPal, 2014. How would you like to integrate with PayPal? [Online] Available: <https://developer.paypal.com/docs/>.

PayPal, 2015. Purchase Protection – How to Stay Safe and Sound with PayPal. [Online] Available: <https://cms.paypal.com/cgi-bin/marketingweb?cmd=_render-content&content_ID=security/buyer_protection>.

Poulsen, K., 2010. PayPal Freezes WikiLeaks Account, 04 12 2010. [Online] Available: <http://www.wired.com/2010/12/paypal-wikileaks/>.

Preibusch, S., Bonneau, J., 2013. The privacy landscape: product differentiation on data collection. In: Economics of Information Security and Privacy III. Springer, pp. 263–283.

Preibusch, S., Kübler, D., Beresford, A.R., 2013. Price versus privacy: an experiment into the competitive advantage of collecting less personal information. Electronic Commerce Res. 13 (4), 423–455.

Rainie, L., Kiesler, S., Kang, R., Madden, M., Duggan, M., Brown, S., Dabbish, L., 2013. Anonymity, Privacy, and Security Online. Pew Research Center.

Roesner, F., Kohno, T., Wetherall, D., 2012. Detecting and defending against third-party tracking on the web. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI).

Scism, L., 2013. State Farm Is There: As You Drive, 4 August 2013. [Online] Available: <http://online.wsj.com/news/articles/SB10001424127887323420604578647950497541958>.

Singel, R., 2010. Online Tracking Firm Settles Suit Over Undeletable Cookies, 12 May 2010. [Online] Available: <http://www.wired.com/2010/12/zombie-cookie-settlement/>.

Sage Software Inc., 2014. Level 3 processing data. Enhanced credit card processing.

Soltani, A., Canty, S., Mayo, Q., Thomas, L., Hoofnagle, C.J., 2010. Flash Cookies and Privacy. In: Intelligent Information Privacy Management, Papers from the 2010 AAAI Spring Symposium, Technical Report SS-10-05.

The Public Voice, 2009. The Madrid Privacy Declaration: Global Privacy Standards for a Global World, 3 November 2009. [Online] Available: <http://thepublicvoice.org/madrid-declaration/>.

TRUSTe, 2009. Behavioral Targeting: Not that Bad?! TRUSTe Survey Shows Decline in Concern for Behavioral Targeting, 4 March 2009. [Online] Available: <http://www.truste.com/about-TRUSTe/press-room/news_truste_behavioral_targeting_survey>.

Tsai, J.Y., Egelman, S., Cranor, L., Acquisti, A., 2011. The effect of online privacy information on purchasing behavior: an experimental study. Inf. Syst. Res. 22 (2), 254–268.

Ur, B., Leon, P.G., Cranor, L.F., Shay, R., Wang, Y., 2012. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In: Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS).

Valentino-DeVries, J., Singer-Vine, J., 2012. They Know What You're Shopping For, 7 December 2012. [Online]. Available: <http://online.wsj.com/news/articles/SB10001424127887324784404578143144132736214>.

Viennot, N., Garcia, E., Nieh, J., 2014. A measurement study of google play. In: ACM International Conference on Measurement and Modeling of Computer Systems.

WSJ Online, 2013. [Online]. Available: <http://online.wsj.com/public/page/what-they-know-digital-privacy.html>.