Purchase details leaked to PayPal (Short Paper)[1]

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

[#]formerly Microsoft Research
preibusch.de

[*]KU Leuven
www.kuleuven.be

Abstract. We describe 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 Web tracking, online shops make it impossible for their customers to avoid this proliferation of their data. We record and analyse leakage patterns for *N*=881 US Web shops sampled from Web users' actual online purchase sessions. More than half of the sites shared product names and details with PayPal, allowing the payment provider to build up comprehensive consumption profiles across the sites consumers buy from, subscribe to, or donate to. In addition, PayPal forwards customers' shopping details to Omniture, a third-party data aggregator with an even larger tracking reach. 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 and mobile retailing, and a growing business opportunity. 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 method. This corresponds to data items traditionally collected during credit card transactions. However, a much richer set of data items becomes available for online purchases, including an itemised bill or information about the buyer, allowing for value-added services. These data are valuable for payment providers and merchants who can benefit from lower fees.

1.2 Privacy concerns and the principle of data minimisation

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. Of particular interest is shopping data, whose value is demonstrated through myriads of loyalty card schemes.

[1] Online companion at: http://preibusch.de/publ/paypal_privacy

Purchase tracking now happens across channels (online / offline) and even if users are not enrolled in a loyalty scheme [1], [2].

Our research motivation is the ability of payment providers to collect purchase details at scale. Similar to 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). 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 the linkage of multiple transactions even across different logins or accounts.
- 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 merchants 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. Although the principle of data minimisation has long been codified in national law and international privacy guidelines (e.g., by the OECD [3]), it is only with the European Union's upcoming General Data Protection Regulation, that data minimisation is becoming an enforceable principle [4].

1.3 Research questions and our contribution

Ahead of tightening regulation regarding data minimisation, recognising that online payment handling is a growing market, we set out to explore the tracking capabilities of online payment providers.

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 [5]. We investigate which personal and transactions details merchants are sharing with PayPal above pure order totals (Fig. 1). 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 and found unsatisfactory privacy practices across the industry [6]. They also investigated data protection practices across different industries [7] 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 [8].

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 [9]. Other researchers surveyed the use of more advanced and resilient tracking mechanisms such as evercookies [10], [11], [12], browser fingerprinting [12], [13], [14] and cookie syncing [12], commonly reporting on questionable practices and unexpected prevalence of such technologies.

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 [15]. 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 [16]. When privacy comes for free, more than 80% of consumers choose the company that collects less personal information [16]. Earlier results indicated that price discounts override online shoppers' privacy preferences [17].

3 Methodology

3.1 Background: PayPal integration, information flows, privacy agreements

PayPal has been a pioneer to offer payment acceptance to electronic retailers, albeit 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 [18]. On a technical level, there are two different integration routes depending on how the session data is 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 buyers' browsers.

Integration via GET is simple and readily available for hosted Websites, as no server-side communication is required ("buttons" in PayPal parlance). 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 [19]. 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.

Payment sessions referenced via an EC token are very common. 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 is transmitted (Section 3.3).

The "Legal Agreements for PayPal Services" [20] outline a number of requirements for merchants. All information submitted to the API must be "true, correct, and complete" [21]. Whereas all fields containing personal information are optional [22], 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 [22].

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 from real online shopping sessions, seeded from Internet Explorer users who opted in to share their browsing history. Practices at these popular online destinations impact a large consumer population. Stores are identified by their URL, as occurring before the PayPal checkout page in browser sessions. For each URL, we selected a single product for purchase, following a strict procedure.

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.

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).

3.3 Experimental protocol

For reliable results, a strict data collection protocol was followed during the main data collection, after a pilot study on 40 Websites. The details of the experimental setup and procedures are laid out in the Online Companion. 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 while navigating from the product page to PayPal's checkout screen. Browsing was done in Firefox and all HTTP and HTTPS traffic was captured by mitmproxy [23] and stored. This includes GET and POST requests and the parameters submitted with them. Web forms

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

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: □ = sometimes, ■ = always, blank = never

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. Although data collection was tool-supported, there was always a human in the loop.

4 Data analysis

4.1 Data description

Dataset. From an initial list of 1200 extracted from browsing sessions, we successfully collected data for N=881 merchant Websites: HTTP(S) traffic traces until reaching the PayPal login page, and screenshot upon arrival. The parsed logs and transcribed screenshots constitute all evidence of personal identifiable information (PII) leakage a customer can capture. More than 86% of all Websites use a token implementation; we rely on the screenshots for those as PII leakage cannot be inferred from the client logs.

To verify our screenshot-based approach, we checked whether the PayPal screen always displays all PII received over the GET query-string. We were able to confirm that whenever customer or product data was 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.

Clustering of leakage patterns. The leakage patterns form the backbone of our work. To analyse the data more deeply, we reduce the number of distinct patterns by clustering all 881 URLs into only few classes (Table 1). We use EM clustering [24], which automatically determines the appropriate number of clusters.

A natural question is whether a particular combination of endpoint and token usage enforces or prevents leakage. Analysing the clusters with association rule mining indicates 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 conclude that PayPal's available API methods do not bias Web shops to treat customers' privacy in a specific way.

4.2 Adding Alexa metadata: Website popularity and quality

We investigated whether Website popularity and technical quality had an influence on privacy-friendliness. We use the Alexa Web Information Service (AWIS) features 'speed percentile' and 'traffic rank' as proxies. Speed percentile 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 0.4m for C_1, 1.0m for C_3 and 1.4m 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.

4.3 Third-party tracking facilitated by PayPal, and internal persistent cookies

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 were sent to paypal.d1.sc.omtrdc.net and paypal.112.2o7.net, which both belong to Omniture [25]. 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 [26]. Remarkably, PayPal also shares the Referer URL of the checkout page, which reveals the URL of the Web shop, and potentially the product to be purchased.

The transfer of these details enables Adobe to build a better profile of 152 million PayPal users [5], by combining payment details with other online activities recorded on more than 300,000 Omniture-tracked Websites [27], 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 Referer headers as studied in [28], since the PayPal checkout page actively collects and sends the Referer 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 make it possible for Omniture to track users by their browser fingerprints even if they block cookies or use private browsing mode [13].

According to its privacy policy, PayPal may share customers' personal information with third-party service providers [29] who are limited to use 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.

PayPal still deploys two questionable, internal tracking mechanisms: evercookies and browser fingerprinting. Although these techniques may be helpful in preventing account hijacking or similar fraudulent activities, their use is not mentioned explicitly in PayPal's privacy policy. These tracking techniques are difficult to avoid for users and have led to lawsuits and multi-million dollar settlements in the past [30].

5 Limitations

As outlined in Section 2, 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.

This comes at the price of diversity of goods that are sold. It easily observed that there are more Web shops selling physical goods than there are commercial dating Websites, for instance. This makes statistically significant statements about differing privacy practices hard, if not impossible.

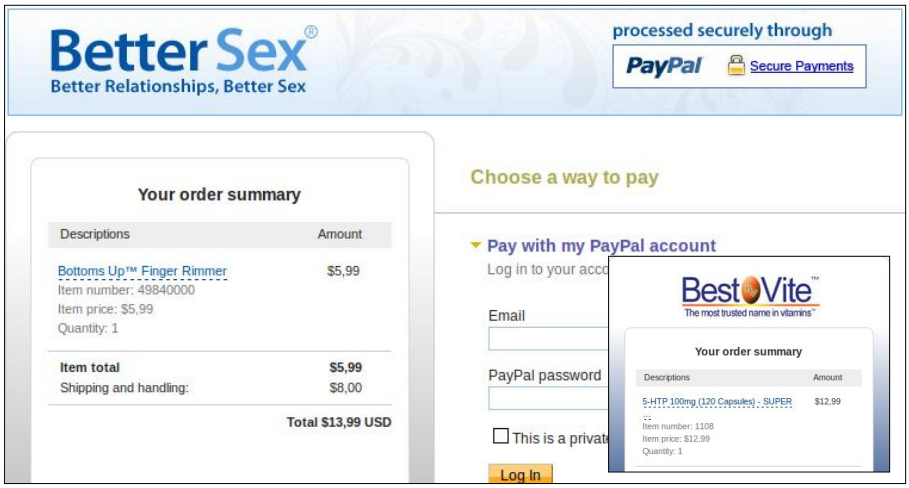

Fig. 1. Sites selling sensitive products also leak product details to PayPal: adult toys and medication (5-HTP addresses depression, anxiety, sleep disorders). Also see the online companion.

For obvious reasons, our data collection setup could not cover server-to-server communication, which, according to PayPal documentation [18], 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 is not covered in our analysis.

6 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 our analysis, 52% of the Web shops in our study shared product names, item numbers and descriptions with PayPal. 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.

By exploring the alternative privacy preserving practices that can be followed by Web shops, we distilled the following suggestions: (1) apply 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; (4) use a payment gateway to prevent leakage of product URL via Referer header.

Better privacy practices for handling online payments is not only desirable for end users, but also for the merchants and payment providers whose bussinesses 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.

References

1. J. Valentino-DeVries und J. Singer-Vine, „They Know What You're Shopping For," 7 December 2012. [Online]. Available: http://on.wsj.com/TQ8Dbi.
2. C. Duhigg, „How Companies Learn Your Secrets," 16 February 2012. [Online]. Available: http://nyti.ms/QbbTyS.
3. OECD, „The OECD Privacy Framework," 2013.
4. European Commission, „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)," 2012.
5. PayPal, „About PayPal," 2014. [Online]. Available: https://www.paypal-media.com/about.
6. J. Bonneau und S. Preibusch, „The Privacy Jungle: On the Market for Data Protection in Social Networks," in *Eighth Workshop on the Economics of Information Security (WEIS)*, 2009.
7. J. Bonneau und S. Preibusch, „The password thicket: technical and market failures in human authentication on the web," in *Ninth Workshop on the Economics of Information Security (WEIS)*, 2010.
8. S. Preibusch und J. Bonneau, „The privacy landscape: product differentiation on data collection," in *Economics of Information Security and Privacy III*, Springer, 2013, pp. 263--283.
9. B. Krishnamurthy und C. E. Wills, „On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks (WOSN)*, 2009.
10. A. Soltani, S. Canty, Q. Mayo, L. Thomas und C. J. Hoofnagle, „Flash Cookies and Privacy," in *Intelligent Information Privacy Management, Papers from the 2010 AAAI Spring Symposium, Technical Report SS-10-05*, 2010.

11. M. Ayenson, D. J. Wambach, A. Soltani, N. Good und C. J. Hoofnagle, „Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning," SSRN, 2011.
12. G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan und C. Diaz, „The Web never forgets: Persistent tracking mechanisms in the wild," in *Proceedings of CCS 2014*, 2014.
13. P. Eckersley, „How unique is your web browser?," in *Proceedings of the 10th international conference on Privacy enhancing technologies (PETS)*, 2010.
14. G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses und F. a. P. B. Piessens, „FPDetective: Dusting the web for fingerprinters," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013.
15. J. Y. Tsai, S. Egelman, L. Cranor und A. Acquisti, „The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study," *Information Systems Research*, 22(2), pp. 254--268, 2011.
16. N. Jentzsch, S. Preibusch und A. Harasser, „Study on monetising privacy. An economic model for pricing personal information," European Network and information Security Agency (ENISA), 2012.
17. S. Preibusch, D. Kübler und A. R. Beresford, „Price versus privacy: an experiment into the competitive advantage of collecting less personal information," *Electronic Commerce Research*, 13(4), pp. 423--455, 2013.
18. PayPal, „How would you like to integrate with PayPal?," 2013. [Online]. Available: https://developer.paypal.com/webapps/developer/docs/.
19. PayPal, „Getting Started With Express Checkout," 2013. [Online]. Available: https://developer.paypal.com/webapps/developer/docs/classic/express-checkout/integration-guide/ECGettingStarted/.
20. PayPal, „Legal Agreements for PayPal Services," 2014. [Online]. Available: https://www.paypal.com/us/webapps/mpp/ua/legalhub-full.
21. PayPal, „PayPal Developer Agreement," 2013. [Online]. Available: https://www.paypal.com/us/webapps/mpp/ua/xdeveloper-full.
22. PayPal, „SetExpressCheckout API Operation (NVP)," 2014. [Online]. Available: https://developer.paypal.com/docs/classic/api/merchant/SetExpressCheckout_API_Operation_NVP.
23. mitmproxy project, „mitmproxy 0.9 - Introduction," 2013. [Online]. Available: http://mitmproxy.org/doc/index.html.
24. A. P. Dempster, N. M. Laird und D. B. Rubin, „Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), pp. 1--38, 1977.
25. Adobe Systems Incorporated, „Digital marketing | Adobe Marketing Cloud," 2014. [Online]. Available: http://www.adobe.com/solutions/digital-marketing.html.
26. Adobe Systems Incorporated, „SiteCatalyst variables and query string parameters," 2014. [Online]. Available: http://helpx.adobe.com/analytics/using/digitalpulse-debugger.html#id_1298.
27. BuiltWith Pty Ltd, „Websites using Omniture SiteCatalyst," 2014. [Online]. Available: http://trends.builtwith.com/websitelist/Omniture-SiteCatalyst.
28. B. Krishnamurthy und C. Wills, „Privacy diffusion on the web: a longitudinal perspective," in *Proceedings of the 18th international conference on World wide web (WWW)*, 2009.
29. PayPal, „Privacy Policy," 20 February 2013. [Online]. Available: https://www.paypal.com/webapps/mpp/ua/privacy-full.
30. R. Singel, „Online Tracking Firm Settles Suit Over Undeletable Cookies," 12 May 2010. [Online]. Available: http://www.wired.com/2010/12/zombie-cookie-settlement/.