The Privacy API: Facilitating Insights In How One's Own User Data Is Shared

Bram Bonné, Peter Quax and Wim Lamotte

Hasselt University - tUL - imec
Wetenschapspark 2
3590 Diepenbeek, Belgium
Email: {bram.bonne,peter.quax,wim.lamotte}@uhasselt.be

***Abstract*—This work describes a proposal to increase transparency and legibility for users of services where the service provider is handing over user data to marketeers, advertisers and other third parties in order to cover the costs of providing the service or to increase revenue. The proposed solution takes the form of an API, offered by the service provider to the user, which provides the user with the same data that is passed on to third parties, but limited to data related to this particular user. In this work, it is discussed how such an API should be implemented, how it should be enforced, and which important considerations should be made when implementing it. The API maps directly on already existing channels used by the service provider, and should require minimal implementation effort while substantially increasing transparency. It provides benefits to users by providing better insight in how their data is used and what it is worth, to service providers by making it easier to comply with regulations and by increasing (potential) user trust, and to regulators by providing a consistent framework for assessing compliance. Furthermore, the proposed approach allows the inclusion of consumer organizations or other trusted third parties as part of the information flow.**

***Index Terms*—privacy; user data; pii; data broker;**

I. Introduction and context

Over the past few years, software applications have increasingly been offered in the form of services, often without requiring a direct payment. Instead of – or in addition to – directly asking users to pay for the service, many internet services have shifted to using advertisements or selling user data to third parties as their source of revenue. Historically, these services have been limited to applications such as social networks, though recently non-free service providers (such as mobile operators or hardware manufacturers) have also been using these methods [1], [2] as a way to increase their revenue.

A problem with paying for services with data instead of money is that, while it is relatively easy for a user to assess the value of the service itself, it can be difficult to assess the value and scope of the provided data. This value is not only determined by how much the service provider profits from it directly (e.g. because it allows the provider to train machine learning models with it, because it allows to profile the user, or because it allows to provide a better service or generate customer value), but also by what third parties can do with it. Moreover, implications of sharing this data with either the service provider or third parties are often unclear. While there have been initiatives and court cases that try to estimate the value of personally identifiable information or users' interests [3], this information is often only a guess at best, and is based on what the average user shares instead of on the data of a specific person.

Furthermore, service providers often provide some form of information inferred from the data (using data mining techniques) to third parties. The user providing the data to the service provider more often than not does not possess the same means (in the form of technology or infrastructure) as the service provider, leaving the user with no idea about which information can be inferred from their data [4], [5]. This has led to users distrusting service providers, opting instead for more 'guerilla' techniques of thwarting data mining algorithms by providing falsified or obfuscated data [6]. Similar to users, it can be difficult for consumer organizations to assess in what way service providers are handling user data, or to provide metrics for comparing different service providers in the area of privacy.

We suggest that the only way a user can decide whether a service is worth the data he/she has to hand over in order to use it, is by providing him/her with some way of knowing exactly which data is used, and in what way it is used, to generate revenue for the company providing the service. We propose an API, offered by the service provider to every user, allowing to query the exact same data third party marketing companies and advertisers have access to (but limited to information about the specific user accessing the API). The aim is to tackle the Legibility principle of Human-Data Interaction [7].

This work does not aim to provide a catch-all solution to all privacy issues, but rather proposes a technique that can be easily implemented for existing systems, providing a valid intermediary step on the path to a complete privacy framework such as the one proposed by Su et al. [8].

II. Definitions

In order to be able to provide a clear explanation of the information flow in the next sections, the different actors are defined first:

Service provider The entity providing an (internet) service, requiring the user to provide some form of personal data in order to use the service. The provided service

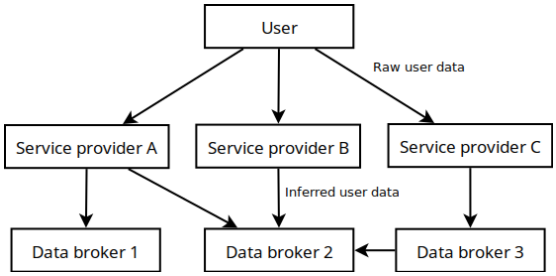

Fig. 1. Current information flow

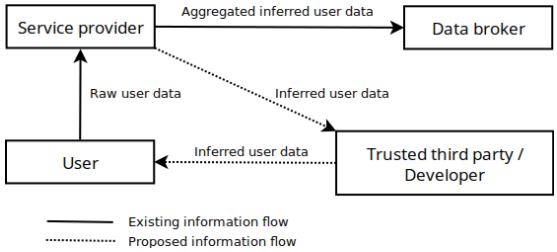

Fig. 2. Proposed information flow

can be (but is not limited to) a social network (such as Facebook), a webmail service (such as Google's GMail), an app (such as Snapchat) or an Internet-of-Things device or platform (such as a Samsung Smart TV or Google's Brillo).

User The person using the (internet) service, whose data is being monetized by the service provider, e.g. in order to keep the service free or cheap, or to provide extra functionality.

Data broker The entity paying or otherwise offering incentives to the service provider in exchange for users' data. This can be for example a third-party analytics or marketing company, an advertising agency or a direct data consumer such as an insurance company gathering information about its clients.

Trusted third party (TTP) A third party (organization) that is trusted by the user, with different users being able to trust different third parties. This can be an institution founded with the purpose of defending consumer rights, such as the Electronic Frontier Foundation, Consumer Watchdog, the American Civil Liberties Union or the Consumer Federation of America. This third party can be either non-profit or for-profit. For most cases, the TTP can be considered to be a consumer organization. However, the term 'trusted third party' (or TTP) is used in this text to include other organizations not strictly fitting the definition of a consumer organization. For example, an organization raising awareness on online privacy could also be considered to fit the definition of a TTP.

A visualization of how information flows between these different parties is available in Figure 1. Note that this visualization does not yet include the "trusted third party" as described above; this party will be introduced as part of the new information flow in the next section.

III. The API in practice

The implementation of the privacy API aligns with existing API's offered to third party data brokers. These data brokers currently already access data inferred about the users of the service. The privacy API would expose exactly the same interface to the user as is currently exposed to data brokers, with the only difference being that instead of aggregated information, the API only exposes information about the currently logged in user.

The actual responsibility of visualizing the data and its implications to the user is in the hands of trusted third parties. These organizations are able to provide a service to users, allowing them to see exactly which data is available to data brokers through the services they are using. TTP's effectively act in the same way as data brokers, with the exception that users give them permission to handle their data beforehand. These TTP's are then able to aggregate user data from a variety of services and inform the user about which data is shared by which service provider. As will be discussed in Section V, the "trusted third party" is not strictly required: a user could also access the privacy API directly instead of relying on a TTP to visualize their data. However, we envision that most less technically inclined users will rely on a third party organization to visualize and interpret the data. This information flow is visualized in Figure 2.

Access control – defining which data will be available through the API – is handled in the same way as other APIs offered by the service provider to third parties. Users can log in directly at the service provider, generating a token that can be passed on to the TTP allowing access to only this user's data via the API.

Note that offering the privacy API does not expose any proprietary information about the service provider or its algorithms: it essentially provides information as a black box and on a need-to-know basis to the user, without revealing how inferences on the data were made. Furthermore, the API does not offer any more information to competitors than is already available to them if they would buy user data from the first party as a data broker, or if they would buy this data through a third party.

The existing API (and thus, the privacy API) can take the form of a regular programming interface, allowing data brokers to query the data directly, but it can also take the form of a more informal 'data dump' containing aggregated user information or statistics, delivered regularly to the data brokers. An API for the second method would work in much the same way, allowing the user (or a TTP acting on behalf of the user) to get an up-to-date data dump of the part of their personal data that is used to generate the information sent to data brokers.

Implementing the privacy API in this way shifts the need for user trust from the service provider and its third party data brokers to a "trusted third party". To see why this shift is needed, consider the motives of both these parties. In the

case of the service provider or the data broker, the main motive is to keep as many people using the service as possible, which might conflict with the goal of providing transparency. However, for a TTP such as consumer organization, the main added value to the consumer is exactly this providing of information. Moreover, since providing transparency is now effectively decoupled from providing the service, the user is now able to choose between different organizations analyzing the data from privacy APIs. Similarly, it is easier to vet well-known consumer organizations than it is to vet unknown third party data brokers. As a last benefit of this approach, more technically inclined users are now able to build their own application upon the privacy API endpoint, not relying on any third party, and allowing for a diverse ecosystem.

IV. Methods of enforcement

A crucial part of the implementation of a privacy API is that it should operate in exactly the same way as the API that is provided to data brokers (providing the exact same interface). This requirement should be enforced to ensure that service providers do not stray from the original goal of this API.

The European Union already has strict privacy regulations in place for companies that manage user data and that want to operate within the EU [9], [10], requiring these companies to provide users with an interface to get a complete overview of all collected user data, and enforcing these requirements by starting lawsuits against non-compliant companies [11]. Different companies have been implementing this requirement in different ways, with e.g. Google providing a privacy dashboard containing controls for managing which data is saved in a Google account and an overview of all data generated for that account[1]. While we applaud the efforts made by different companies to be more transparent about their data collection and retention practices, there is no standardized way of presenting this information, requiring the general public to rely on the goodwill of the service providers to provide a clear and complete overview.

A limitation shared by most of these approaches that implement the EU's regulations is that the user can only see which data the service provider has about them, without knowing which part of that data is shared with third party data brokers. Moreover, the service provider may be using data mining techniques to infer extra knowledge from this data, which could also be shared with data brokers. Consider for example the case of a user who only provided the service provider with their purchasing habits (e.g. by using a loyalty card at a store), from which the service provider inferred that this user is a parent. Existing solutions would only show the user an overview of their purchasing habits, while the privacy API would show exactly which inferred information was passed on to the data brokers.

A privacy API could be made mandatory by lawmakers in much the same way as the discussed regulations, but can also provide a more efficient alternative to the service providers themselves. Indeed, implementing a privacy API might require substantially less effort than providing a privacy dashboard, giving service providers a more low-cost way of complying with these regulations. Moreover, this could prove to be a valid alternative to more involved regulations aiming to limit collection of user data by service providers, as the privacy API shows exactly the amount and type of data that is provided to third parties, eliminating guesswork about which data is used only to improve the service itself.

An alternative to enforcement by law is to approach this problem in a similar way to how security audits work. In this case, third party auditing companies could offer certificates of compliance to companies that pass an audit standardized by the industry. This maps directly to standards such as ISO/IEC 27001, where companies are audited to make sure they adhere to security best practices.

Even without being enforced by lawmakers, or without being incentivized by auditing companies, service providers can implement the privacy API to grow trust and goodwill among their (potential) users, showing that transparency about data collection and monetization is one of the company's values.

V. Considerations and limitations

Note that the main difference between the privacy API and the API that is being offered to data brokers is that the privacy API might disclose more identifiable or personal data about the specific user than the aggregated data that is available through the data broker API. This is a desired side effect of the proposed approach: as data brokers tend to apply deanonymization techniques to the gathered data [12], in which anonymized or aggregated data is tied back to the original user from whom the data originated, having the most specific information available is important to see which data could in principle be inferred by the third party data broker itself. However, this also means that user data is now managed by an extra party (the TTP), effectively creating a larger attack surface for malicious actors trying to obtain this data and requiring the TTP to be 'trusted' not only with the data itself, but also with securely managing it.

One limitation of the proposed approach is that it is retroactive: users need to be using the service before they can assess which data is shared. As such, this approach can not be used directly to help users decide beforehand whether they want to use a service (in contrast to proactive approaches like P3P). However, since data aggregators and marketeers care mostly about recent data, this could be a valid trade-off for the user to make. Similarly, (anonymized) data about other users of the service could be used by consumer organizations (acting as the trusted third party for multiple users) to paint a general picture about which data is provided and to compare different service providers in the area of privacy.

Secondly, the proposed approach only works for service providers that pass on user data to third parties. Service providers can still use all user data internally (either to improve

[1]Google's activity controls and activity overview are available at https://myaccount.google.com/activitycontrols and https://myactivity.google.com/myactivity, respectively.

their service to users, or to mine this data for interesting patterns). For this, we count on security regulations and audits instead of providing an API (which would put the responsibility in the hands of the user instead of the service provider).

VI. Related work

Previous approaches to solving this problem have tried to create ontologies or formalized languages for describing privacy-sensitive data, like P3P [13] which never saw widespread adoption despite having been implemented on 15% of the top 5,000 websites [14] and despite being supported in Microsoft's Internet Explorer and Edge browsers. Common criticisms to these approaches is that they often fail to capture the semantics and relationships of the data, that they are difficult to understand for less technical users, that they make it difficult to arrive at an agreed-upon vocabulary and that the user requirement of defining privacy preferences beforehand can lead to misdirected settings [15].

Techniques like TaintDroid [16] inspect the information flow going from a user's (Android) device to the internet, specifically tagging personally identifiable information like IMEI numbers and location data being sent by apps to advertisement servers. This provides great insights in how information is shared by mobile apps to third party data brokers. However, it only shows data shared by the app directly, and cannot distinguish between data that is sent to the service provider for internal use (e.g. for improving the service) and data that is sent to the service provider to be shared with data brokers afterwards.

Similarly, the privacy API differs from the 'right of access' requirement in the EU's General Data Protection Regulation (GDPR) [10] mentioned in Section IV in that instead of offering the data that has been provided by the user to the provider, it offers the data that is available to data brokers. This may exclude certain data that will not be passed on, and may also include any extra inferred data that is passed on to data brokers.

Very recently, a related proposal has been made by Su et al. [8], where the (privacy-sensitive) flow of health data is formalized. They propose to separate the data operator from the data source, mediating between the actual data collector and the data sinks. Our approach differs from the aforementioned one in that it does not require any modifications to current implementations and contracts between data collectors and third party data sinks, pushing for more transparency within the existing process instead. Indeed, our work rather proposes an intermediate step towards a complete privacy framework, offering a solution which can be easily implemented on top of existing systems without requiring large modifications.

VII. Conclusion

In this work, we formulated an initial proposal for a privacy API which allows complete transparency about the user data that is passed on by service providers to third parties. We showed that implementing such an API can have benefits for all parties involved: service providers (in the form of easier compliance and increased user trust), consumers (in the form of more transparency and comfort, and improved information for making decisions), legislators (in the form of regulations that are straightforward to implement and audit) and consumer organizations (in the form of better metrics).

We propose to perform further research based on this concept, implementing a proof-of-concept privacy API for a small set of service providers handling user data. Thus, this paper can be considered as a call to action for service providers interested to cooperate with researchers in the fields of computer science and privacy. This gives innovative companies a way to show they are serious about privacy, while allowing researchers to assess the feasibility of a privacy API.

References

[1] M.-J. Lee, "Samsung Adds More Ads to Its TVs," 2016. [Online]. Available: http://www.wsj.com/articles/samsung-adds-more-ads-to-its-tvs-1464600977

[2] A. Troianovski, "Phone Firms Sell Data on Customers," 2013. [Online]. Available: http://www.wsj.com/articles/SB10001424127887323463704578497153556847658

[3] P. Gliman and N. Glady, "What's The Value Of Your Data?" 2015. [Online]. Available: https://techcrunch.com/2015/10/13/whats-the-value-of-your-data/

[4] J. H. Reiman, "Driving to the panopticon: A philosophical exploration of the risks to privacy posed by the highway technology of the future," *Santa Clara Computer & High Tech. LJ*, vol. 11, p. 27, 1995.

[5] D. J. Solove, "Data mining and the security-liberty debate," *The University of Chicago Law Review*, vol. 75, no. 1, pp. 343–362, 2008.

[6] F. Brunton and H. Nissenbaum, "Political and ethical perspectives on data obfuscation," *Privacy, due process and the computational turn: The philosophy of law meets the philosophy of technology*, pp. 164–188, 2013.

[7] R. Mortier, H. Haddadi, T. Henderson, D. McAuley, and J. Crowcroft, "Human-data interaction: the human face of the data-driven society," *Available at SSRN 2508051*, 2014.

[8] X. Su, J. Hyysalo, M. Rautiainen, J. Riekki, J. Sauvola, A. I. Maarala, H. Hirvonsalo, P. Li, and H. Honko, "Privacy as a service: Protecting the individual in healthcare data processing," *Computer*, vol. 49, no. 11, pp. 49–59, 2016.

[9] Council of European Union, "Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *L281*, pp. 31–50, 1995.

[10] ——, "Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," *L119*, pp. 1–88, 2016.

[11] Belgian Commission for the Protection of Privacy, "The judgment in the facebook case," 2015. [Online]. Available: https://www.privacycommission.be/en/news/judgment-facebook-case

[12] B. Schneier, *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company, 2015, ISBN: 978-0393244816.

[13] L. F. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle, "The platform for privacy preferences 1.0 (p3p1.0) specification," *W3C recommendation*, vol. 16, 2002.

[14] L. F. Cranor, A. M. McDonald, S. Egelman, and S. Sheng, "2006 privacy policy trends report," *CyLab Privacy Internet Group*, 2007.

[15] C. Perera, C. Liu, R. Ranjan, L. Wang, and A. Y. Zomaya, "Privacy-knowledge modeling for the internet of things: A look back," *Computer*, vol. 49, no. 12, pp. 60–68, 2016.

[16] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, "TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones," *ACM Transactions on Computer Systems (TOCS)*, vol. 32, no. 2, p. 5, 2014.