Efficacy of *Right to be Forgotten* in Social Networks

Keywords: identity, linkability, privacy, social networks, data governance, GDPR

Abstract: Online social networks like Facebook have become silent witness to our online activities either by our consent or by our desire to avail services for free. Being witness to our activities they construct so much knowledge about us that they can predict our intentions, priorities, personal traits. Realizing the potential privacy implications of such an aggregation of user data, European Union devised GDPR. One of the core tenets of this regulation is *right to be forgotten* and we argue that it will have little or no impact when invoked on Facebook platform. We investigate the reasons behind this shortcoming of GDPR *w.r.t.* Facebook by providing a detailed understanding of Facebook's platform and how it handles user data internally. With this understanding of the platform we argue that Facebook's business model will fail if this tenet of GDPR is enforced in its *spirit.*

1 INTRODUCTION

What if trust is broken? It will have silent costs to our online lives. Trust is the cornerstone of our online ecosystem. Trust is the confident relationship to the unknown and we rely on trusted third parties to take this risk. For example, we rely on DNS servers to reach intended places online. We rely on Google to search information online. We rely on Apps to obtain specific functionalities like banking, news. With repeated positive experiences, users develop trust with these intermediaries and start sharing information that is sensitive. With any doubt of breach of trust, the users move from one trusted third party to the other, if available. By definition, the trusted third party knows the details of transactions it is facilitating. A scheme to monetize this knowledge through advertising was tried and became hugely successful so much so that collecting data in itself became a specialization. Myriad of online services collect, analyze, use, and exchange users' personal information either by consent or convoluted opt-in/opt-out schemes with a sole motivation of harvesting user data. Data has indeed become new oil/gold and is been siphoned off wherever possible, with little respect for privacy. More than a thousand companies are now involved in a digital information value chain that harvest data from any online activity and delivers targeted content to online or mobile users within roughly 36 seconds of their entry into the digital realm. The sophistication and ease of targeting users has become so widely available and rampant that there is a sense of anxious helplessness. GDPR is devised to specifically address the misuse of personal data.

Credit bureaus have long been gathering information about our earnings, spending habits and loan-repayment histories to determine our creditworthiness. Tech companies have taken this one step further, collecting data on our web-surfing habits and which numbers we call. Over the period, it has become cheaper to collect and process data. Via social media, we have volunteered information on our friends and our likes and dislikes, and shared family photographs. Our smartphones can know everywhere we go and can keep track of our health[1].

The techniques that Cambridge Analytica uses to produce its psychometric profiles are the cutting edge of data-driven methodologies first devised 100 years ago. The science of personality research was born in 1917. That year, in the midst of America's fevered entry into war, Robert Sessions Woodworth of Columbia University created the Personal Data Sheet, a questionnaire that promised to assess the personalities of Army recruits. The war ended before Woodworth's psychological instrument was ready for deployment, but the Army had envisioned its use according to the precedent set by the intelligence tests it had been administering to new recruits under the direction of Robert Yerkes, a professor of psychology at Harvard at the time. The data these tests could produce would help decide who should go to the fronts, who was fit to lead and who should stay well behind the lines[2]. The benefits and threats of data-oriented decision making process are limitless in the age of AI/ML/DL techniques. GDPR is a first exhaustive effort to contain

[1] washingtonpost.com/news/innovations/wp/2017/09/29/why-we-need-new-regulations-to-protect-us-from-companies-like-equifax

[2] https://www.nytimes.com/2018/03/22/opinion/democracy-survive-data.html

the menace of connived data control and sharing.

Organization: In the following section, we elucidate Facebook's data platform, its business model and architectural components. In Section 3, transformation of personal data through the information value chain on Facebook is explained. In Section 4, we argue about the effectiveness of GDPR tenets, and in Section 5 we discuss we broadly discuss the containment of privacy. We conclude in Section 6.

2 ONLINE SOCIAL NETWORKS

OSN is an online platform that allows users to form social connections with others on the platform. Connected users interact with each other and also with each others' content, connections. The platform provides privacy settings to its users with which the users inform the platform who else on the platform can *access* users' data and connections. The users explicitly trust the platform therefore it can access and *observe* its users' data and *interactions.*

OSNs have phenomenally transformed the way we reach, engage, express ourselves with our social surrounding. And it being *online*, gets a granular level of insight about its users' time of activity, location, and frequency. This insight about user behavior attracted businesses and organizations to the platform to engage with their community of users. User insight evolved as an attractive proposition for platform owners since such insights[3] found compelling usefulness in many applications.

Facebook is one such platform that epitomizes all other social networks in terms of features, followers, and loose data governance. There are various reasons for it being such that we shall list in the course of this paper. And we shall use it to put forward our analysis about privacy in social networks, in general.

2.1 The OSN business model

Majority of the social networks have evolved into platforms that provide more than the name suggests. Their evolution from vanilla social network to a data platform that encircles almost all of the user activities is due to the following: i) the cost of collecting, storing, and processing of data is reducing, ii) volume of online services is increasing, iii) competitive pressure to micro-target users for advertisement requires an in-depth knowledge about the users, which can be obtained by observing users' online activities over a long period. They innovate in designing algorithms and models that extract value from the large unstructured information users generate while online. Their revenue model relies on collecting and converting unstructured sets of user information into structured, meaningful, formatted data (see Figure 1). This formatted data is the valuable resource the platform generates. Users contribute in generation of this resource and in return get improved online experience, convenience, apart from the *free* social communication service. The platform owner sufficiently de-sensitizes the formatted data and allows advertisers to identify users as potential customers. The business model has following deliverables to its stakeholders:

Social communication: This is the core deliverable and generates users' social behavioral profile when users interact with other users on the platform. Users are nudged to interact under different contexts so that new contextual insights are collected and the old ones are reenforced, corrected.

Convenience: The platform provides single-sign-on (SSO (Sun and Beznosov, 2012; Fett et al., 2015)) facility to online services that want to authenticate their users via the platform. Thus the online service saves on maintaining its own authentication infrastructure. The platform can observe user-service authentication interaction and entices the service to use its APIs to interact with its users' connections. A user of the platform get authentication convenience in exchange of the platform witnessing her interactions with the service. Specific functional conveniences (like: finance, sports) are provided through Apps and Groups. Users association to such functional conveniences allows the platform to confidently categorize them and personalize their services.

Personalization: It is paramount for the platform to keep its users engaged while achieving its objective to gather increasing amounts of insights about users by enticing them to interact socially. NewsFeed (newsfeed.fb.com) is one such tool that allows the Facebook platform to introduce its users to updates from their social sphere. NewsFeed prioritizes the updates based on user's past interactions as observed by the platform, inferred interests, and also predicts potential categories of advertisements that user might react to. The goal is to use the actionable intelligence (see Figure 1) generated by the platform to fill every inch of user screen with purposeful information.

Advertisement: The platform, along with an App ecosystem that relies on social connections on the platform generate their revenue by compiling an audience type requested by its advertisers. The

[3] facebook.com/iq/tools-resources/audience-insights

revenue generated from advertisers is used to sustain the costs incurred to provide free services.

Measurement: Advertisers, App developers, Group admins are given access to analytics information so that they can measure impact of their engagement with users.

The value proposition in collecting and converting unformatted user data into meaningful formatted data is so high that almost all entities on the platform (like apps) and off the platform (like ISPs, DNSs, webservers) treat the user data in a meaningful way. Different entities engage users with their respective privacy policies for data they collect and may sell the data in market as an additional revenue stream. In the following we shall concentrate on Facebook platform.

2.2 The OSN platform

Facebook has evolved from a social network to an intermediary with a bouquet of overarching services that supplement its core business of collecting, interpreting, and monetizing user data. So far, it has been successful in pushing its users to accept evolving privacy norms by offering them compelling user experience through convenience[4], personalization. In this section we explore the stakeholders of this platform and how they collaborate in its business model.

Users. Users are the largest part of the platform. They are represented as nodes on a graph, called social graph, in which *labelled* edges represent relationships between nodes. Users' *interactions* within their reachable neighborhood and beyond, with the nodes introduced by the NewsFeed, builds behavioral profiles of users. NewsFeed is Facebook's intelligent algorithm that prioritizes social updates from a user's neighborhood. If we assume that each object (e.g., content like photos, also represented as a node) on the social graph has a category type associated with it, like: education, finance, food, sarcasm, celebrity, etc., then a subject's interaction with these objects determine the probability of interest the subject may have in such categories. Each interaction of a subject with its neighborhood node improves the confidence level of subject-category mapping. The objective of NewsFeed algorithm is to increase subjects' interaction with variety of content categories (International Personality Item Pool, 2018) so that a rich user profile can be built. Such a user profile is pivotal in determining relevancy of updates to the user and also to match the user with an advertiser advertising for a subset of such categories (ProPublica Data Store, 2016). Higher the engagement of the user, higher the confidence value in categorizing the user.

Apps. The platform gives a general purpose connectivity and interaction mechanism to the users, whereas the Apps give a context to user profile. App serves a specific functionality (e.g., finance, education, dating, et al.) to its users and that functionality is a stronger measure of categorization. Apps can opt for monetization by serving advertisements to the users via the App. Apps obtain analytics over their users interactions (see Figure 1). The analytics information contains attributes (like mobile advertisement ID, Facebook UID, email, phone, device info, location, etc.) that can uniquely measure interactions of App users. To advertise itself, or to persuade its existing users the App shares its analytics with advertisers to reach the existing and new users.

Analytics and trackers (Pixel). Pixel is a micro-targeting framework (`fb.com/business/learn/facebook-ads-pixel`) that uniquely identifies users of the platform and also users off-the-platform (Portokalidis et al., 2012; Acar et al., 2014). This is a script that generates a unique tracking number each time a defined event occurs. The events could be as simple as loading a website or a user selecting a product in her cart. The unique number concatenated with cookie at user side tracks the user event by event. These user behavior analytics are availed by the platform to the advertisers so that advertisers can measure the impact of their advertising campaigns.

Advertisers. The platform's ability to find relevant audiences for a specific category/issue brings advertisers to the platform[5] . Advertisers build advertisement campaigns by requesting specific audience type from the platform against a fee. To compose an audience request, advertiser uploads[6] data fields that get compared against the user profiles built by the platform. Upon evaluating the scope of the campaign being submitted by an advertiser, the platform may refuse to execute the campaign; if i) it could not find any users for the requested audience type, or ii) the requested audience size is too small. However, the advertisers are allowed to micro-target a specific audience that is already engaged with a campaign. Advertisers do so by defining events inside

[4]The tyranny of convenience, Tim Wu, NYT, 16/2/18. https://www.facebook.com/zuck/posts/10104899855107881

[5]https://developers.facebook.com/docs/marketing-api/reference/custom-audience

[6]https://www.facebook.com/business

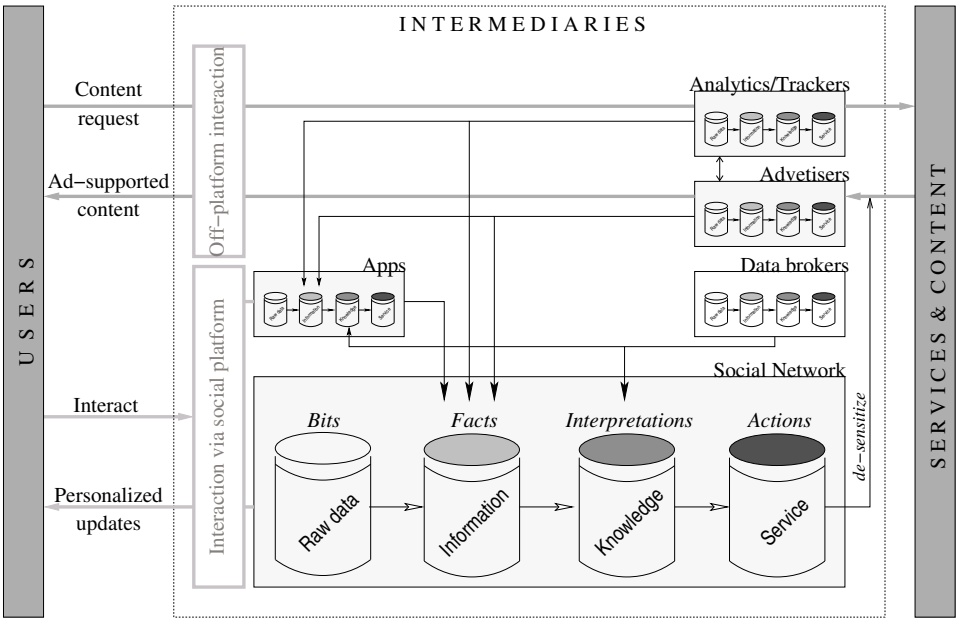

Figure 1: Business model of social networks (*record* → *interpret* → *monetize*).

the Apps/services and triggering corresponding actions upon event actuation. For example, list of users who have browsed a product inside the App but did not checkout, visit a particular website.

Platform owner (FBAN). Facebook Audience Network (`fb.com/audiencenetwork`) is the core component of the platform and has complete access to users' profiles generated to date. It has its own data-set and models that are built by tracking users (analytics) and their meta-data from other sister platforms (like WhatsApp, Messenger, Instagram). It accepts audience requests and based on corroboration of uploaded data by advertiser with its proprietary data-sets, it generates a target audience. It incorporates data from data-exchanges (like acxiom, datalogix, datamarket, datastreamx, equifax, bluekai, epsilon) and data from government departments (like land/property records, census, electoral roll) for its users, which help enrich users' existing profiles.

Data platforms like Facebook whose business model is based on advertisement have a similar schematic design that we saw above. The platform provides a compelling service (like email, chat) in return of a consent from its users to access their data and observe their interactions. The platforms compete with each other (Schneier, 2015) to attract advertisers by innovating ways for advertisers to track and target existing and new customers. Innovations include features providing convenience, personalization either by the platform or via the App ecosystem. The core characteristic of these platforms is *labelled data collection* where users voluntarily label themselves either through the generic features of the platform or through the specific features of the Apps they authorize. Users are given a set of options to control the access to their data by other users, Apps, and to some extent advertisers. In the following we shall see the access control provided by Facebook platform.

2.3 Access control

The Facebook data platform appears to have a somewhat hybrid, ad-hoc access control system in place. It is primarily designed for speed and ease of feature integration. To do so, it organizes its users, their content, and their relationships as a social graph (Bronson et al., 2013) where nodes represent subjects & objects like users, Apps, content and edges represent relationships among the nodes. Nodes have types[7] and can be queried by other nodes (of type user and App) directly or indirectly. Graph queries[8] return data only if the access policy specified at the node is conformed by the querying node. Reachability is the primary condition for access. There is a logical hierarchy among the node types on the platform. It is designed in a way that higher a node in the hierarchy, larger is the access to data. In other words, an App type of node can make far richer types of queries on the social graph than a

[7] developers.facebook.com/docs/graph-api/reference
[8] developers.facebook.com/docs/graph-api/explorer

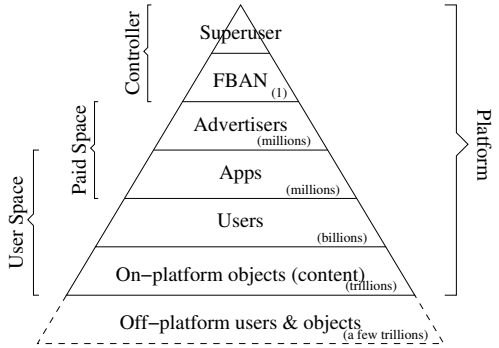

Figure 2: Facebook platform: hierarchy of access.

User type of node – an thus far can collect higher volumes of user data. Figure 2 depicts the volume of data on the platform and the access hierarchies among the stakeholders of the platform. Since the evolution of data begins with the users creating, sharing, and interacting with the data, users are said to be the primary owners of the data and therefore we list out the access control methods available to them.

User to user. User nodes on the social graph can query other user nodes and their objects (nodes depicted at the bottom of the pyramid in Figure 2) in accordance with the policy specified at the target node. Facebook allows its users to specify access policies over their objects using a mix of intensional and extensional labels like "Friends", "Family", "Friend of friends", et al. This type of natural labeling nomenclature allows users to organize their social relationships as they do in offline world. This communicates the affinity/strength of the edge between nodes; thus, inversely a node's ability to influence the node at the other end of the edge. Therefore the list of "Friends" of a user becomes a valuable information and Facebook provides means to protect such objects from unauthorized access. However, in (Patil and Shyamasundar, 2017b) the authors show that the access policies on the Facebook platform do leak user data despite correct policy specification by the users on their objects.

User to app. The user nodes cannot query the App nodes above in the hierarchy whereas the App nodes can query the User nodes directly and their objects indirectly. An App's access to user nodes is not controlled by the natural labels, as provided for user-to-user access control, but through an explicit list of permissions (total 48 such permissions are available in v2.12 of Facebook's graph API) that user confers on the App at the time of installation. The access control for user data by Apps is designed for facilitating functionalities of the Apps. Users give a set of permissions as requested by an App to obtain the functionality. Whatever transactional/observational data is generated, during the course of App functionality usage, is not controlled by the user. Facebook prompts users that by installing an App, a user abides to the privacy policy of the App. There are 3 broad categories of Apps: i) Apps that rely on FB for authentication (SSO), ii) Apps that modify the social graph with consent from user, and iii) Apps that tailor user experience based on the social graph of its users. The first two categories, by design, shares the user activity with the platform; whereas in the third category the user activity is recorded outside the platform. User's permissions to Apps are perpetual and the data availed by Apps are not governed by FB's privacy policy. Furthermore, the permissions to Apps override the access controls expressed by users in user-to-user layer. For example, a post by a user with policy "Only Me" is accessible by an App. In (Patil and Shyamasundar, 2018) the authors list out scenarios in which Apps either breach users' stated policies or simply undermine user's sensitive information for which the platforms does not provide any measure of protection. For example, an App can find out what other Apps the user has installed.

User to advertiser. Advertisers too are provided with UIDs but these are not query-able as the UIDs of users and Apps. Neither the advertising nodes can query the nodes in "User Space" as shown in Figure 2. However, advertisers[9] get indirect access to users' behavior data through the analytics available to Apps and off-platform services (websites, and Apps). The only control users get over advertisers[10], regarding advertisers' access to user data, is by disassociating with them.

User to platform. By virtue of being the owner of the platform, the superuser node can query any other node, on the social graph, without any access restrictions[11].

Users implicitly trust the platform and therefore are comfortable with being observed by the platform. Users may not invoke similar level of trust with the Apps that are supported by the platform and therefore the platform provides its users a choice to control the set of permissions an App can obtain to access user data. *The notion of privacy of a subject comes to*

[9] https://lifehacker.com/5994380/how-facebook-uses-your-data-to-target-ads-even-offline

[10] https://www.facebook.com/ads/preferences

[11] motherboard.vice.com/en_us/article/kzxdny/facebook-investigating-employee-stalking-women-online

fore only in the presence of an observer/violator who knows the subject. And therefore, the user-to-user access control is relatively expressive than the user-to-App or user-to-advertiser access control.

In light of the recent Cambridge Analytica (Lee Edwards, 2018) revelations, this falsely perceived *status quo* about privacy, that existed for past decade, is being questioned widely. This is not a one off scenario of users' privacy breach but it is due to lack of a uniform platform-wide access control model. The access controls are implemented layer-wise, where policies in one layer may contradict with policies in another. As a response to Cambridge Analytica incident, Facebook maintains that it will review and limit Apps' access to user data and also highlights that users are owners of their content and can control who can access these content. What it is curiously missing is who can access the meta-data and behavioral data about user interactions. This position is tenable, for now, because there is an absence of Internet-wide agreement about *who owns the meta-data – data about data* and scope of data-usage, which is a challenge to verify, though defined.

Facebook allows its users to delete their accounts. Through this paper we would like to bring the reader's attention to the following questions: i) How effective a user's decision is to delete her account in order to avoid her being profiled? ii) Post deletion, is it technically possible for the platform to uniquely identify the user? iii) What are the technical limitations to enforce GDPR's "right to be forgotten" intent?

Before we get into the analysis of these questions, it would be appropriate to understand the characteristics of information, digital transactions, the life-cycle of personal data, its mutation into other types in presence of other auxiliary data, and the risks of de-identification of de-sensitized/anonymized data (Ohm, 2009).

3 LIFE-CYCLE OF PII IN OSN

In this section we take the reader through a brief journey of a user's data on OSN platform – Facebook.

PII (McCallister et al., 2010) is "any information about an individual maintained by an agency, including i) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and ii) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information." To distinguish an individual is to identify an individual.

Identifiers distinguish a user (or a group of users, or a passive object) from another. Each unique entity that needs to be interacted with has an identifier, e.g., name. Identifiers may have attributes like postal address, city name, date of birth. Attributes are a generic class of identifiers, which do not identify a subject on its own but in presence of its association with a subject improves the uniqueness of subject identification. Observer is an entity that has knowledge of identifiers of subjects and it may assign private attributes to subjects based on their activities under observation. For example, an ISP may legitimately assign attributes like gamers, bankers, student, etc. based on the online activities of its customers. Observer may develop data models based on its customers' online behavior and may devise a method to predict attribute/category for an unknown subject when that subject's log of online activities is fed to the model. Therefore, the potency of an observer is proportional to the volume of data it has access to. When an attribute is unique to a subject then the attribute is equivalent to PII in the given context. For example, if there is only one person with a specific DoB in a database (knowledge-base) then that attribute uniquely identifies the subject it is assigned to. If there are more than two subjects that have same DoB, then the probability of correctly associating a subject to an action reduces to half, and likewise. Whereas instead of one attribute, two attributes of subjects are considered under the same observation model, the probability of subject identification greatly increases. Users neither have knowledge about potency of their observers nor sufficient motivation to judiciously reveal their attributes while online. *Level of privacy is a loose measure of asymmetry of motivation, ability between an observer and the subjects being observed.* An observer (advertiser or its collaborator) is financially motivated to identify its audience and has technical ability (via the platform) to do so; whereas, the users are motivated to get functional benefits of the free service being provided, until and unless adversely affected.

Coarse classification of PII. In (Gurevich et al., 2016), the authors classify personal data (infons) of a subject into 4 flat categories as shown in Figure 3. Examples of these types of infons are: *Public* – name, email, phone; *Directly Private* – passwords; *Partially Private* – salary, blood group; *Inversely Private* – mobile location logs. We term this classification coarse because in absence of context the above examples may fall in multiple classes. For example, passwords can be categorized as partially private either, because the validator retains a copy of the password. Likewise, an individual's credit rating can be categorized

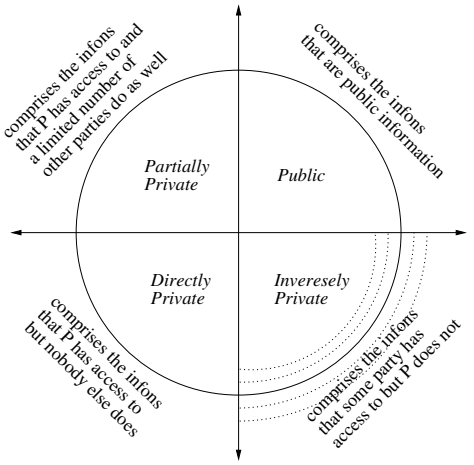

Figure 3: Classification of PII (infon) of subject P.

as inversely private until that individual obtains a copy upon payment. Therefore, context is an important aspect in accurate categorization of PII – this is what Facebook's social graph does – it records all the previous interactions of a subject and changes the state of the social graph so that new state is available to all other nodes of the graph.

In the following, we walk through the process of how users' PII on an OSN platform gets transformed from governable verbatim strings to ungovernable diffused data. The process is abstracted out in 4-steps and depicted in Figure 4.

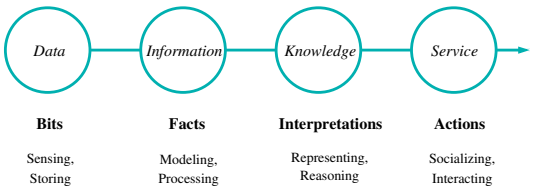

Figure 4: Abstractions along the information value chain.

Voluntary labeling (sensing, storing of data). When a user signs-up with the platform by providing her personal details, the platform assigns a 64-bit unique ID (FBID) to the user and is represented as a node on the social graph. FBID acts as a primary identifier on the platform and user fills out various personal details (DoB, affiliation, city, languages) as attributes to the FBID node. User node's edges to other nodes represent relationship of a specific type: social affinity (friends, family, acquaintance, groups, et al), object ownership (photos, post, video, et al), actions (check-in, like, comment, tag, events, et al), and installation of an App. All these possible edge formations by a node is used for labeling the node according to the type of the peer node. For example, user installing a sports App will label the user in "sports" category. The user "like" a post of category "sport" by other user will reinforce the labeling. Thus, a node's category influences the categories of the nodes interacting with it. The platform labels the object nodes (content, location, Apps, groups) to determine interests of its subjects when they voluntarily interact. Apps may have their own private labels, which they may or may not share with the platform.

Observational labeling (modeling, processing of information). The platform observes and records its users activities on-platform and off-platform (through Pixel, for example). These observations include facts like IP address, type of mobile OS, type of browser (DeKoven et al., 2017), active time on platform, active time on other platforms[12], call logs, browser logs, location history, etc (Chaabane et al., 2012). All this factual information along with the voluntary labels/categories form rich profiles about users and also group of users. Further enrichment and fortification of information is done by correlating data from external sources[13] like credit-rating agencies, census data, electoral rolls. Both the Apps and platform track user behavior by modeling events in the interactions of users on/off the platform. This analytics is processed and compiled as formatted information such that it can readily be used to re-target the users towards attending incomplete events; for example, a user added items to shopping basket but did not check-out, so re-targeting such users to complete payment event is an objective. A user who has history of adding items to basket but not checking-out is not the audience advertisers are interested in.

Mutation & diffusion of data (representing, reasoning of knowledge). Advertisers are interested in getting a high conversion rate for their advertising budget. Personality of a user is a strong measure to anticipate her behavior. The user profiles containing rich sets of labels are then synthesized (Youyou et al., 2015; Kosinski et al., 2013) into valuable individual personality traits[14] – which in its abstract form is represented by OCEAN (Costa and McCrae, 2012) or Big Five[15] score. Thus the verbatim user profile tuples <FBID, labels> get mutated into <FBID, labels, OCEAN>. Then the users can be dB-identified and organized according to their personality traits so

[12]https://www.wsj.com/articles/facebooks-onavo-gives-social-media-firm-inside-peek-at-rivals-users-1502622003

[13]https://www.facebook.com/help/494750870625830

[14]https://ipip.ori.org/AlphabeticalItemList.htm

[15]http://www.outofservice.com/bigfive/

that advertisers can rent it out to construct their tailored audience for a specific campaign. Each campaign has a context and FB-AN does the placement of advertisement since it has complete knowledge – user profiles, traits, and the context. Expertise from other well-known psychometric models is used to further synthesize the knowledgebase of FBAN to build new reasonings (Kristensen et al., 2017) about users' behavior prediction in presence of certain events that are triggered either on the platform or elsewhere. This is how verbatim a user's PII data gets mutated and be diffused in information value chain, as knowledge, as shown in Figure 4. It is possible that same knowledgebase can be constructed using two different datasets. In other words, exclusion or inclusion of a small quantity of input data does not change the knowledgebase substantially.

Users are made available the logs of their actions on the platform but not within the Apps. The inferred categories[16] that are labelled against users' respective FBIDs are neither made available to users.

Monetization of data (actions: putting knowledge into use). Facebook's overarching *sensing, recording* platform (Section 2.2) continuously does 3 tasks in tandem to consolidate its hold on users' information value chain: i) collect & classify user PII data/actions, ii) correlate data/actions with other facts to build/improve profiles, and iii) dynamically build audiences as per contexts specified by its customers. In the process the users PII data and actions are continuously transforming across the information value chain and getting diffused into actionable knowledge. The omnipresent, overarching ecosystem of Facebook, through its platform components and analytical feedback loop from internet-wide content collaborators, is a real-time context delivery service. FBAN's knowledgebase along with platform's real-time context prediction capability helps Facebook to attract advertisers, governments, and persuaders to build/identify tailored audiences. However, Facebook takes precaution to prevent its customers from tailoring audience that has size smaller than 100. The same knowledgebase is also used in Facebook's NewsFeed algorithm, which is famous for its prioritization of relevant updates to a user.

This brings us to the most interesting question in PII's life-cycle: is it possible to identify a user with the help of a knowledgebase/model that is trained on the data involving the user's PII? In the following section, we shall argue about it in affirmation while explaining the efficacy of soon to be enacted GDPR for EU subjects.

[16] In 2016, ProPublica collected more than 52,000 unique attributes that Facebook has used to classify users (ProPublica Data Store, 2016)

4 EFFICACY OF GDPR IN OSN

GDPR's tenets will certainly change how users PII is collected, processed, and disposed off. Users will have to be explicitly consented, informed, and allowed to access/delete their PII. Any entity that legally interacts with subjects across EU are covered by GDPR with certain exceptions[17], which does not cover OSNs. Handlers of PII are categorized as: i) *data processors* – entities processing data on behalf of the controller, and ii) *data controllers* – entities deciding what personal data must be processed and how processing will occur. Facebook acts as both[18] the controller (for its users) and the processor (for its millions of Apps; where Apps act as the controller of user data) of PII. Facebook Apps in their data controller role need to get explicit consent from their users to collect and use the user data, which might directly/indirectly be available to underlying platform of Facebook and thus gets mutated and diffused into FBAN's knowledgebase. Keeping this in mind, the efficacy of *right to be forgotten*[19] is analyzed below.

Right to be forgotten. When a user invokes this right, the data controller and processor have to delete all the data they have collected (not only the PII) that can directly or indirectly uniquely identify the user. In case of Facebook, the *spirit* of this GDPR article is that the user should not be remembered for any of the services offered by Facebook. In *letter*, Facebook will delete all the directly/indirectly identifiable data about the user from its information value chain. But as described in previous section, the identifiers (FBID, attributes, labels) that are associated with a user exist in verbatim form only till the modeling/processing stage of platform's information value chain. Beyond that stage it exists in a diffused form, as knowledge, which would have been the same even if the user had not joined the platform. Figure 5 shows the effective boundary of GDPR's delete operation on an OSN platform like Facebook. In other words, right to erasure will delete only the PII, voluntary labels, observational labels but it cannot force the platform/Apps to reverse the knowledge created using machine-learning models trained using user's data.

[17] https://gdpr-info.eu/recitals/no-52/

[18] https://www.facebook.com/business/gdpr

[19] https://gdpr-info.eu/art-17-gdpr/

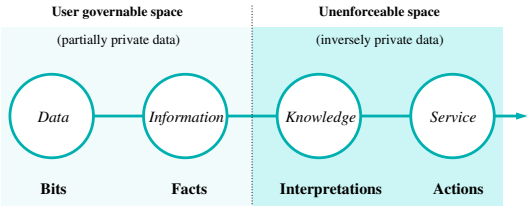

Figure 5: Information value chain: scope of regulation

Post invocation of right to erasure by an user, the user continues to get identified by her metadata (IP, locale, time-zone, behavioral fingerprints, hardware fingerprints, et al) and platform's capability to track[20] non-users (Portokalidis et al., 2012; Acar et al., 2014) across the affiliate services/websites that are associated with the platform. This tracking data allows Facebook to determine context of the non-user, which is sufficient to match the non-user's attention to an advertiser with as much relevance as when the user was associated with the platform. And it will be technically not feasible to prove that Facebook has used PII to match the user with an advertiser. In fact, the user will be better off by staying associated with the platform instead of being forgotten by the platform because while associated with the platform the user can at least control her ad preferences.

Even if Facebook wants to honor user's right to be forgotten in *spirit*, it will have to track such users in order to instruct FBAN to do so. This could be easily achieved by issuing such users a "do not track" cookie. But this will break the business model of Facebook as losing the context of user's attention will reduce the click-through rate for advertisements. In other words, *right to be forgotten* is the least effective tenet of GDPR. We are all witness to the rampant disregard for "Do Not Track" setting in modern browsers. What this tenet of GDPR may achieve with limited success is: suspension of perpetual compilation of user's personality and behavioral traits.

5 DISCUSSION

There is a technological imbalance among the stakeholders of our digital economy. The platforms have attained technical superiority in data collection, processing and actively influence the design of fundamental interfaces to Internet (browser, DNS, mobile OS, Apps) to further consolidate[21] their data-driven business model: *record everything* $\rightarrow$ *interpret* $\rightarrow$ *monetize and persuade.*

Advertisers' primary motivation/aim is to reach their intended audience with minimum expenditure. The expenditure is optimum when there is an exact match, in other words, precision targeting is inevitable. OSNs have created the ability to identify their users for specific criterions that advertisers are interested in. Through OSNs, the *motivation* of advertising industry is easily achieved due to the technical *ability* of OSNs to accurately find the users. On the other hand, users typically lack the motivation and ability to make elaborate privacy decisions. Furthermore, for the platform/advertiser, the cost of making a wrong probabilistic guess about intended target is negligible whereas, for the target/user, it is costly to get identified.

GDPR has stiff financial penalties for laxity in personal data handling. However, provenance is difficult given the fact that a data controller (an App) on Facebook platform, by design interacts with the underlying social graph. Depending on the design of the App, partial or full data of App user is recorded on Facebook's social graph. The data then propagates further in platform's information value chain. It is easy to create an App and start collecting user data without much of practical liabilities. This is because any user can create an App. Facebook has introduced a concept of Scope_ID that issues local identifier to App users such that the identifier can only be valid within the scope of the App. Users cannot be tracked for their activities across the Apps. However, FBAN can resolve the Scope_IDs of all Apps. It will be interesting to see what changes Facebook will usher in to its platform to be compliant with GDPR while acting as a data processor for its Apps. We contend that the relationship is not so distinguishing between Apps and Facebook platform as controller and processor.

In presence of ubiquitous tools like big-data analytics and deep neural networks, preserving privacy appears to be a herculean task (de Montjoye et al., 2015; De Montjoye et al., 2013). To address this challenge coherently, we need to undertake a SoK for identifiers. Because it is their usage beyond the perceived scope of their utility that leads to potential privacy breaches. The SoK should put forward a framework for the use of identifiers in terms of their scope, contextual availability, temporal validity, linkability – and the effects of these parameters on each other. Architects of online services have service functionality and user convenience as primary design criteria. Providing them a methodology to judiciously use the above mentioned parameters with an understanding of their costs and benefits to the system they design.

[20] theregister.co.uk/2018/04/17/facebook_admits_to_tracking_non_users/

[21] Net neutrality blocked ISPs from providing services subsidized with advertising.

6 CONCLUSIONS

To pursue is an innate desire and manifests in many ways. Advertisement is one of such external manifestations of persuasion. OSNs with their close proximity to user's day-to-day activities have become a real-time gauge of user's mindset – necessary for effective and quick persuasion. A business model has evolved in which users trade their private data for convenience and ad-supported services. The trade-offs continue to favor the platform owners and their techniques to track user activities have brought in a privacy anxiety and helplessness to users – primary stakeholders of online ecosystem. A lack of transparency in collection, processing, and usage of of personally identifiable data by OSNs is undermining trust in online services. As the facade of self-regulation has failed spectacularly, GDPR is the first right step to address this crisis. However, an important tenet of this regulation, *right to be forgotten*, has very little efficacy in OSNs with their prevalent omnipresent *data sensing* platform. We have argued that to implement this tenet in its *spirit*, platform's business model breaks and we have little evidence that self-regulation has worked. To bring in accountability, transparency, and competitive innovation; OSNs should either treat the Apps as their own extensions or relinquish its role of data processor when a data controller brings in data of users (to the platform for processing) who have invoked their right to be forgotten on the platform. But, will it be naive to expect voluntary declaration of *conflict-of-interest* from the class of entities who have failed to self-regulate for decades?

REFERENCES

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., and Diaz, C. (2014). The web never forgets: Persistent tracking mechanisms in the wild. In *ACM CCS'14*, pages 674–689.

Ann Cavoukian (2011). Privacy by design: The 7 foundational principles. `online`.

Athey, S., Catalini, C., and Tucker, C. (2017). The digital privacy paradox: Small money, small costs, small talk. Working Paper 23488, Nat. Bureau of Econ. Research.

Barth, A., Datta, A., Mitchell, J. C., and Nissenbaum, H. (2006). Privacy and contextual integrity: Framework and applications. In *IEEE S&P'06*, pages 184–198.

Bronson, N., Amsden, Z., Cabrera, G., Chakka, P., Dimov, P., Ding, H., Ferris, J., Giardullo, A., Kulkarni, S., Li, H., Marchukov, M., Petrov, D., Puzar, L., Song, Y. J., and Venkataramani, V. (2013). TAO: Facebook's Distributed Data Store for the Social Graph. In *USENIX ATC 13*, pages 49–60.

Camenisch, J. and Lehmann, A. (2015). (Un)Linkable Pseudonyms for Governmental Databases. In *ACM CCS'15*, pages 1467–1479.

Chaabane, A., Kaafar, M. A., and Boreli, R. (2012). Big friend is watching you: Analyzing online social networks tracking capabilities. In *Proc. of ACM Workshop on Online Social Networks*, pages 7–12. ACM.

Costa, P. T. and McCrae, R. R. (2012). *The Five-Factor Model, Five-Factor Theory, and Interpersonal Psychology*, chapter 6, pages 91–104. Wiley-Blackwell.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., and Pentland, A. ". (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539.

DeKoven, L. F., Savage, S., Voelker, G. M., and Leontiadis, N. (2017). Malicious Browser Extensions at Scale: Bridging the Observability Gap between Web Site and Browser. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*.

Dixon, P. (2017). A Failure to "Do No Harm" – India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the U.S. *Health and Technology*, 7(4):539–567.

Esteve, A. (2017). The business of personal data: Google, Facebook, and privacy issues in the EU and the USA. *International Data Privacy Law*, 7(1):36–47.

European Union (2018). 2018 reform of EU data protection rules. `online`.

Fett, D., Küsters, R., and Schmitz, G. (2015). SPRESSO: A Secure, Privacy-Respecting Single Sign-On System for the Web. In *ACM CCS'15*, pages 1358–1369.

FTC (2012). Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers. `online`.

Gurevich, Y., Hudis, E., and Wing, J. M. (2016). Inverse Privacy. *Communications of ACM*, 59(7):38–42.

International Personality Item Pool (2018). The 3,320 IPIP Items in Alphabetical Order. `online`.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Kristensen, J. B., lbrechtsen, T., Dahl-Nielsen, E., Jensen, M., Skovrind, M., and Bornakke, T. (2017). Parsimonious data: How a single facebook like predicts voting behavior in multiparty systems. *PLOS ONE*, 12(9):1–12.

Lee Edwards (2018). Cambridge Analytica and the deeper malaise in the persuasion industry. `online`.

Leon, P. G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., Bauer, L., Christodorescu, M., and Cranor, L. F. (2013). What Matters to Users?: Factors That Affect Users' Willingness to Share Information with Online Advertisers. In *SOUPS*, pages 7:1–7:12. ACM.

McCallister, E., Grance, T., and Scarfone, K. A. (2010). SP 800-122. Guide to Protecting the Confidentiality

of Personally Identifiable Information (PII). Technical report, National Institute of Standards & Technology.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, Vol. 57, p. 1701, 2010*.

Patil, V. T. and Shyamasundar, R. K. (2017a). Privacy as a currency: Un-regulated? In *Int. Conf. on Security & Crypto., SECRYPT 2017*, volume 4, pages 586–595.

Patil, V. T. and Shyamasundar, R. K. (2017b). Undoing of Privacy Policies on Facebook. In *Data and Applications Security and Privacy XXXI - 31st Annual IFIP WG 11.3 Conference, DBSec 2017*, pages 239–255.

Patil, V. T. and Shyamasundar, R. K. (2018). Role of Apps in Undoing of Privacy Policies on Facebook. Technical report, IIT Bombay. `online`.

Portokalidis, G., Polychronakis, M., Keromytis, A. D., and Markatos, E. P. (2012). Privacy-preserving social plugins. In *USENIX Security Symp.*, pages 631–646.

ProPublica Data Store (2016). Facebook ad categories. `online`.

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., and Crowcroft, J. (2012). The Personality of Popular Facebook Users. In *Proc. of the ACM 2012 Conf. on Computer Supported Cooperative Work*, pages 955–964.

Schneier, B. (March 2015). *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World*. W. W. Norton & Company.

Sun, S.-T. and Beznosov, K. (2012). The Devil is in the (Implementation) Details: An Empirical Analysis of OAuth SSO Systems. In *ACM CCS*, pages 378–390.

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.