ANALYTIC SURVEILLANCE: BIG DATA BUSINESS MODELS IN THE TIME OF PRIVACY AWARENESS

Vigilancia analítica: modelos comerciales de datos masivos y concienciación sobre la privacidad

Eva-Patricia Fernández-Manzano and María-Isabel González-Vasco

Eva-Patricia Fernández-Manzano graduated with a PhD in Media from *Complutense University of Madrid*. She has a master's degree in media business management from *Instituto de Empresa* (*IE Business School*), and a master's in big data and business intelligence management from *Escuela de Organización Industrial* (*EOI*). She is a professor at the *Rey Juan Carlos University* in media management. Her research focuses on how media companies work with technology. She has published several papers and books about media and technology, including *Big data: Eje estratégico en la industria audiovisual* (*UOC*, 2016) and *Data management in audiovisual business: Netflix as a case study* (*EPI*, 2016).
https://orcid.org/0000-0001-7655-872X

Universidad Rey Juan Carlos
Edificio Departamental, Despacho 43
Camino del Molino, s/n. 28943 Fuenlabrada (Madrid), Spain
eva.fernandez@urjc.es

María-Isabel González-Vasco is an associate professor at *Universidad Rey Juan Carlos* in Madrid, Spain. Her main research interests include design and analysis of cryptographic tools arising from group theoretical problems, as well as probable security of cryptographic tools (signature schemes, key establishment protocols). She has published more than 40 journal and conference papers, a book, and a number of technical reports on group theory and cryptography. She is a member of the *International Association for Cryptologic Research* (*IACR*) and of the board of the *Spanish Royal Mathematical Society*.
http://orcid.org/0000-0002-7452-9121

Universidad Rey Juan Carlos
Departamental II, Despacho 34
C/ Tulipán, s/n. 28933 Móstoles (Madrid), Spain
mariaisabel.vasco@urjc.es

Abstract

Massive data collection and analysis is at the heart of many business models today. New technologies allow for fine-grained recommendation systems that help companies make accurate market predictions while also providing clients with highly personalized services. Because of this, extreme care must be taken when it comes to storing and managing personal (often highly sensitive) information. In this paper we focus on the influence of big data management in media business content platforms, mainly in well-known OTT (Over the Top) services. In addition, we comment on the implications of data management in social networks. We discuss the privacy and security risks associated with this novel scenario, and briefly comment on tools that aid in securing the privacy of business intelligence within this context.

Keywords

OTT; Social networks; Security; Big data; Machine learning; Privacy; Cryptography.

Resumen

La gestión y análisis de datos masivos es la base de muchos modelos de negocio tecnológicos. Algunos de ellos ofrecen recomendaciones detalladas que ayudan a las empresas a identificar predicciones de consumo, lo que se traduce en una oferta de servicios altamente personalizados hacia los clientes. En este proceso resulta vital extremar el cuidado tanto en el almacenaje, como en la gestión de datos personales (en ocasiones información altamente sensible). Este artículo pone el foco en la influencia del uso del big data en el marco de la producción y distribución de contenidos audiovisuales, concretamente en las plataformas denominadas OTT, así como en las redes sociales. Para ello se analizan los riesgos que surgen respecto a la privacidad y seguridad del usuario, identificando además determinadas aplicaciones útiles que preservan y protegen a los usuarios.

Palabras clave

OTT; Redes sociales; Seguridad; Datos masivos; Aprendizaje automático; Privacidad; Criptografía.

Manuscript received on 09-09-2017
Accepted on: 12-02-2018

Fernández-Manzano, Eva-Patricia; **González-Vasco, María-Isabel** (2018). "Analytic surveillance: Big data business models in the time of privacy awareness". *El profesional de la información*, v. 27, n. 2, pp. 402-409.

https://doi.org/10.3145/epi.2018.mar.19

1. Introduction

The management of massive amounts of data allows businesses to accurately match content production to demand. Thus, so called OTT (over the top) services, enclosing products from music to e-sports, are increasingly popular. However, the rise of new models for managing big data digital platforms, such as social networks, has revealed a lack of effective protocols for dealing with sensitive information.

Digital corporations handle huge amounts of data, some of which are collected from internal audits, but most of which is gathered from intense surveillance of users' behavior and service demands. These data are anonymized, aggregated, and later used for business decision making through a plethora of methods and techniques in what is commonly referred to as *big data management.*

Soon after the breakout of technology-based business models, it became clear that this new scenario called for new technical and legal tools to thwart their intrinsic privacy and security risks. Not only are tech-giants concerned, but smaller companies have become acutely aware of the need to address information security and privacy threats. This of course includes media agencies, digital content distribution platforms, social networks, on-line retailers, e-health companies, etc. Users, in particular those defined as technically-illiterate, are the most vulnerable actors in this bursting data-driven play. Whether this is understood by potential clients or not, it has already provoked significant changes in the way companies are perceived and, most importantly, in the success of their technological bets. Many users assume that the price for enjoying certain services is paid through the acceptance of opaque privacy policies. These policies are often not fully understood by the client, who obliviously permits access to their consumption practices, daily routines, political prejudices, or even sexual inclinations to remote third parties. The idea of such unavoidable compromise has of course been fed by many service providers, much to their own benefit.

While many tech-based businesses, like on-line retailers, social networks, or OTT companies often improve their services through massive data collection, their invasive techniques are, in most cases, not justified and could be avoided. Yet, these techniques have a prize that companies are not willing to pay unless enforced; either by law or by client demand. It is fair to say the Snowden revelations (**Preibusch**, 2015) were a major catalyst for this awareness. The so-called Snowden-effect is often defined as an increase in public concern about information security and privacy resulting from the Snowden reports that detailed *NSA* (*National Security Agency*; intelligence agency of the United States *Department of Defense*) surveillance activities.
Beyond institutional espionage, regularly publicized security breaches give users and corporations food for thought, whether it is the *Heartbleed* security breach or the spread of *Wannacry* ransomware. Obviously, this growing concern has had a significant influence on the design of business strategies today.

In this paper we explore the different business environments in which massive data collection from users is at the core of decision-making, with a focus on digital content distribution platforms and social networks. This article can be viewed as a follow up work to our article (**Fernández-Manzano**; **Clares-Gavilán**; **Neira**, 2016), in which big data management is considered an essential key in tech-business. Thus, here we follow that same line of inquiry by connecting users' data management to business decision making.

2. Aims and motivation

This article studies the current situation of big data business models, focusing on how information from users is collected and analyzed, and to what extent this influences business decisions. Further, we discuss privacy and security risks that arise in this context, and comment on a few research lines that are providing promising tools to minimize these risks.

For this reason, we give a brief description of how social networks and OTT manage issues related to data collection, privacy, or user habits. With the goal of stimulating academic debate, we recommend some best practices.

3. Massive data collection: Wherefrom and what for?

Individual users of new technologies and on demand media distribution have witnessed the industry evolving towards an exchange model, in which data is traded for services. In particular, users who consume content through an Internet-connected device are often aware of privacy abuses, and thus question the liability of the service in terms of protecting their personal sphere.

As a general rule, all information put online by an individual user forms a digital footprint or trail. This trail is comprised by cookies, device IDs, time and location stamps, IP addresses, etc. Starting from the theoretical framework established by **Fernández-Manzano** (2016), we can see different ways in which customer data is exchanged. Following a generic classification born from on-demand content distribution services, we identify the following data sources:

- *Web and social media*: data derived from the use of web pages or social networks, often sensitive by nature, structured (for instance, through *tags*) or unstructured, derived from relationships to other users (*social graphs)*, and so forth.
- *VOD*: data generated (more or less deliberately) by users of OTT platforms. This information is sensitive and must be managed with care in business environments (for instance, using anonymization techniques).

On top of these data sources, in the intricate interactive ecosystem arising from the internet outgrowth, new information retrieval techniques, such as those derived from so-called *machine to machine* (M2M) communication, are used. M2M communication allows companies to collect raw data from different devices which are later sent through the Internet to the next data processing layer. This information can, of course, be highly sensitive; that is the case when IoT (Internet of things) devices, such as wearables, are used for remote patient monitoring. However, biometric information like heart or brain patterns, fingerprints or iris scans are highly sensitive data and should be treated with extreme care.

The above means (and many others) allow for the collection of huge amounts of data and, as a result, business strategies are being transformed through big data analytics, business intelligence, and machine learning; pinpointing client behavior patterns which improve key marketing predictions. Nonetheless, handling these techniques with care is a must, in view of the growing concern of individuals and institutions, aware of the amount of information collected on citizens and held by private firms.

3.1. The case of OTT

On demand, digital, content consumption is implemented through an Internet connection allowing users to view and/or download audiovisual products. From an international perspective, it seems that the main OTT actors coincide in those countries with high rates of accessing/downloading these sort of goods. In Spain, users can gain access to very well-known platforms such as *Netflix*, *Movistar+*, *HBO*, *Amazon Prime,* or *Sky* to name a few. In all these cases, platforms are subscription-based and, as a result, their strategy focuses in client satisfaction and the platforms do not rely on advertising income. From an international perspective, other streaming platforms, such as *Hulu*, are supported by advertisements. A global vision of OTT growth is welcoming new players from already established companies such as *Disney*, *ESPN,* and *Apple*. All in all, knowing user preferences is crucial, and so is analyzing consumption habits. According to Spanish *CNMC* (*Comisión Nacional de los Mercados y la Competencia*) (2017), the top two OTTs in 2017 were *Movistar+* and *Netflix*. Because of the popularity of these two OTTs we will focus on them for our study; in addition, they have also published a lot of information about their data management practices.

The ultimate goal of gathering user information is to give a personalized experience for each client based on big data analytics. OTT platforms make use of machine learning techniques, a set of artificial intelligence techniques, which, when combined with suitable data mining strategies, result in extracted value. The first step after data collection is to organize information in a (huge) database that allows for efficient metrics and data crossing. For each user, data requested at the time of registration is stored; thereafter, the catalog of services offered to the user and ultimately purchased by the user is also stored. While the user's personal information when stored for commercial and business use must be previously anonymized, the audiovisual content itself is ordered and tagged using associated metadata. With the aim of obtaining granular tagging of these goods, companies like *Netflix* hire professional *taggers*, who are specialized in viewing and classifying digital content, linking different attributes such as production date, cast, or country of origin. Through this indexing, up to 80,000 microgenres can be identified (**Madrigal**, 2014), yielding a valuable, structured database that can be later scrutinized through algorithmic techniques towards its final goal: client satisfaction. As **Wolk** (2015) points out, this action makes the video on demand service

> "able to parse those tags to spot trends and patterns in consumption that help information both their content creation and content acquisition choices".

On the other hand, machine learning techniques in these business models have different effects. First, they allow established predictive models that, in collusion with recommender systems, give users suggestions for new purchases. The more information at hand from a single user, the more accurate his tailored recommendation would be. Second, classifying behavioral and consumption patterns turns data into value, especially when features are found that could hardly have been identified without specially designed big data techniques. *Telefónica España* is the corporate brand of the digital platform *Movistar+* and, as pointed out by Elena Gil in **Prieto** (2017), their philosophy could be summarized as *more data, more value.* As a result, Telefónica is aiming to build a database that is even more exhaustive than those belonging to digital native players like *Facebook* or *Google*, including data coming from mobile communication through their network, client interaction, or television consumption. In this spirit, the head of *Telefónica* in **Del-Castillo** (2017) revealed the company has been working on a novel recommender system that would combine behavioral data from television consumption through *Movistar+* and *Telefónica* mobile and broadband service, so that, when using a mobile device abroad one could get audiovisual recommendations linked to the visited country.

Furthermore, combining big data and machine learning techniques allows for grouping or clustering of information; for the OTT case, this results in clustering content in genres or subgenres. This allows for client segmentation, which subsequently provides hyper-personalization of content offers.

The logic behind massive data collection and analysis are the foundation of business models monitoring user behaviors and crossing their stored data with this continuous information flow. In their *Netflix* study, **Gómez-Uribe** and **Hunt** (2016) show a selection of algorithms used for this aim. For instance, *Personal video ranker* offers content hyper-personalized through linking metadata to the platform content. Similarly, *Trending now* establishes categories/trends from previous clustering derived from customer segmentation techniques.

3.2. The case of social networks

Similarly, as for the OTT case, social network platforms are able to collect, store, and manage huge amounts of client data. To some extent, it is fair to say that this information is

the main value these companies make profit from, as a majority of clients are not using any of the so-called premium (not free) services they offer. Thus, social network platforms are able to get personal data from users, and less sensitive information regarding their habits and tastes. Furthermore, according to **Liu** *et al.* (2016), the data collected is used to learn about relationships underlying the social network, which can in turn be used to characterize the social behavior of individuals and groups. This analysis is useful, for example, for identifying social leaders who may influence the behavior or consumption habits of others in the network and such identification is crucial for the design of advanced targeted marketing strategies. Behavioral advertising (**Ortiz-López**, 2017) strategies are designed from behavior or consumption data collected from the Internet. This marketing model has evolved into so-called *real time bidding* (RTB), through which service providers bid to present ads to consumers once they have been identified by their digital footprint. Similarly, **Pelrich** (2016) suggests that combining RTB with the information collected from the Internet and social networks could influence the value of television ads—for example, during the *Super Bowl* broadcast it may be found that users are surfing the Internet during the broadcast.

One of the most revealing pieces of information about individuals is the way in which they establish and maintain relationships through social networks. Mathematically, social networks can be represented by graphs. A graph is a tuple G=(V, E), where V is a finite set of vertices and E is a subset of VxV, that is, each element in E is a pair {a,b} of vertices. Typically, {a,b} is called an edge connecting vertices a and b. Edges can be directed or undirected, thus representing one or two-way links between vertices. If a and b represent users of a social network, these two users will be connected (for instance, friends in *Facebook*) if {a,b} is in E. Asymmetric relations are represented by directed edges, i.e., edges (a, b) with a prescribed start vertex –a- and end vertex –b- .

For instance, in a graph representing the following relationships in *Twitter*, the edge (a,b) stands for "a is a follower of b", thus it may be that (a,b) ∈ E but (b, a) ∉ E – that is, b need not be a follower of a. Understanding the graph topology is very useful for analyzing the underlying group of network users; indeed, connected components (sets of nodes that can be connected chaining consecutive edges) represent groups of users that are akin, and thus are likely to share information about goods. On the other hand, if it is not possible to reach a vertex b from a certain vertex a using the edges in E, the users linked to these vertices are unable to share information through the social network (and thus may be classified in different target groups).

Let us consider for instance the social networks represented by the graphs 1 and 2. On the left (Figure 1) we have a "toy" network where arrows represent information flow; that is, a directed arrow from A to B means all information posted by A will be seen by B. Thus, all data shared by the individual represented by node "L" is accessible by five other users (Ui, for i=1,2,3,4,5) while the node tagged "S" represents a user with no influence at all, as no other user has access to his posted information. Furthermore, there are two connected components in the graphs, i.e., the group of users can be split in two sets {L, U1, U2, U3, U4, U5} and {S, U6} which do not influence each other. Figure 2 represents a more realistic scenario, where edges are undirected (thus users symmetrically share their information when connected). There is a large connected component in this graph, while (on the right) we can see quite a few secluded nodes, representing users that are isolated.

In fact, social networks' knowledge of its users allows for the targeting of specific communities, so tech-companies such as *Facebook*, *Twitter,* or *Google* bet for launching audiovisual content into market. Furthermore, as **Dodard** (2014) points out, working with machine learning techniques and social networks metadata let companies obtain even more information:

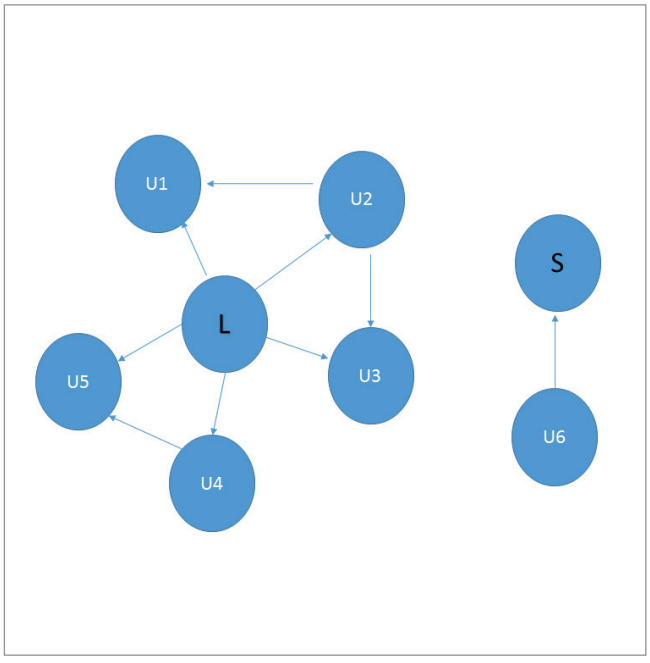

Figure 1. Simple graph representing a social network

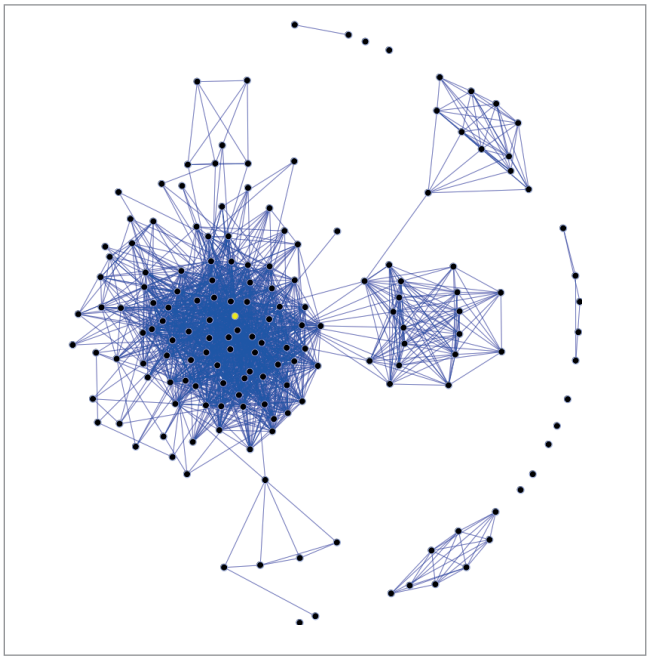

Figure 2. Screenshot taken by user Darwin Peakock in *http://graphexploration.cond.org*

> "The photo upload interface on the *Facebook* mobile app (...) immediately after selecting a photo, the app scan the image for faces and little dialogs are positioned below every face that has been detected, prompting the user to tag the person whose face appears on the picture. This creates more user engagement and allows Facebook to gently remind its users that their experience is more fun if they tag friends on the picture".

4. Privacy and security threats

According to **Brookman** and **Hans** (2013), many threats to the privacy of individuals arises from the mere act of data collection, regardless of the subsequent perusal of these data:

- Data breach. Big collections of data are indeed an inviting target for hackers. The consequences of data breaches can be measured both in direct financial damage (as the leaked personal information is often sufficient for successful fraud) and (more subjectively) in terms of moral harm (as it could reveal private, embarrassing information that a consumer did not want publicly disclosed).
- Internal misuse by employees. Indeed, employees may eavesdrop on clients' information, either aiming at bribe or industrial espionage.
- While not being such a relevant threat as a data breach, this is indeed a concern for cloud storage services.
- Unwanted secondary use. Once individuals have lost control of their data, it is hard to monitor the way (and purpose!) it may subsequently be used and analyzed.
- Changes in company practices. The policies and regulations that prevent a company from engaging in uses of the data that harm the subjects' interests may change.
- Government access without due legal guarantees. Again, through the Snowden revelations we are now aware of the continuous monitoring of information carried out by both our own government and foreign countries.

All in all, as elaborated by **Soria-Comas** and **Domingo-Ferrer** (2016), the main principles typically observed in regulations for the protection of personally identifiable information severely collided with the purpose of big data. For instance, the principles of necessity and data-minimization and purpose limitation are inherently antagonistic to big data techniques, which rely on the accumulation of data for potential (often initially undecided) use. Necessity and data-minimization means that only the data needed for the specific purpose should be collected and it can only be kept for as long as necessary, while purpose limitation imposes that the concrete goal for which information is gathered should be specified before the actual collection.

Companies are fully aware of the risk of ignoring the above threats; clearly, their reputation may be damaged by careless data storage or poorly designed access control policies. As noted by **Herold** (2010), the legal impact of feeble practices is a growing concern for organizations. Indeed, as **Warwik** (2016) points out, 90% of large organizations and 74% of SMEs reported a security breach in 2015 (according to a UK government survey), leading to an estimated total of £1.4bn in regulatory fines. Furthermore, in 2018, the European Union's *General Data Protection Regulation* (*GDPR*) will introduce fines for groups of companies far exceeding the current maximum of £500,000. The Spanish *AEDP* (*Agencia Española de Protección de Datos*) has published, together with *ISMS Forum Spain,* a specific good practice manual for big data management (**Sáiz**, 2017).

Security and privacy preserving goals

According to **Danezis** *et al.* (2015), there are six main goals we should pursue: the first three (confidentiality, integrity, and availability) are linked to general security and most o them achieved through cryptographic techniques. The latter, unlinkability, transparency, and intervenability are privacy-specific protection goals and often call for dedicated privacy preserving techniques. Confidential communication is achieved when only the intended receiver is able to access the information sent by a sender. This is typically achieved through encryption techniques. On the other hand, integrity is attained when messages cannot be tampered with by adversaries; that is, whatever the legitimate receiver gets from a communication channel was actually sent and constructed in that precise form. Digital signatures, message authentication codes (MACs), and hash functions are cryptographic tools often used for that purpose. Informally, availability is the assurance that authorized parties are able to access the information entitled to them when needed. Typical violations of this availability are so called DoS (denial of service) attacks, through which adversaries block a service by denying access to its legitimate users.

As defined in **Danezis** *et al.* (2015), "Unlinkability ensures that privacy-relevant data cannot be linked across domains that are constituted by a common purpose and context". Formally, one should require that data, but also events or actions that are somehow related, cannot be identified as such by an attacker (or at least, an attacker cannot learn about any relationship after interacting with the system that he could not identify a priori). Early erasure of data, physical separation of contexts, obfuscation, encryption, or use of pseudonyms are different cryptographic/information privacy techniques at hand for this goal.

On the other hand, transparency is essential for the auditing of information systems. A transparent system allows for the reconstruction of all data processed at any time. This is a crucial goal in applications such as electronic voting or cloud computing, and can only be achieved through strict reporting policies, storing, and delivering to clients or data owners accurate and reliable information that can stand up to regulatory scrutiny.

Last, intervenability ensures that end-users have control over how their data is stored (**Meis**; **Heisel**, 2016), processed, and transmitted by information systems. To achieve or support intervenability it is crucial to allow users to influence the data management, so that they can change their initial privacy settings at a later time.

5. Technical solutions: Cryptology and privacy preserving techniques to the rescue

As defined on the web site for the *International Association for Cryptologic Research* (*IACR*),

> "Cryptology is the science and practice of designing

computation and communication systems which are secure in the presence of adversaries".

Cryptology has two sides, a constructive one (called cryptography) and a destructive one (cryptanalysis). Typically, cryptographic designs evolve over time due to cryptanalytic attacks. Information privacy, on the other hand, is concerned with the right of individuals to control the way their data is accessed and disseminated. These two disciplines are sometimes perceived as conflicting, while they should be understood as complementary. Sure, cryptographic techniques often force users to identify themselves or to link certain files or data blocks to their identity in a publicly verifiable way (using, for instance, digital signatures). However, as we will see, there are cryptographic constructions especially designed to protect our personal information from uncontrolled exposure, while allowing us to interact and perform cooperative computations with untrusted parties.

5.1. Multiparty computation (MPC)

Multiparty computation (**Cramer**; **Damgaard**; **Nielsen**, 2015), is the branch of cryptology concerned with the problems fitting the following (general, oversimplified) setting: assume a set of n users U_1, U_2,... U_n, holding each a private input x_i, for i=1,..., n, and wishing to jointly evaluate a multivariate function f on these input. However, each user wants to keep its private input secret, that is, after the interaction, each user U_i should have only learnt the computed evaluation $f(x_1,..., x_n)$, and nothing else. The fact that no information on other user´s input is gained by any user (or by any external observer) is precisely stated through information theoretic measures, such as mutual information and entropy.

A typical (academic) example of a MPC problem is the so-called *millionaires' problem* posed by **Yao** (1982) where two millionaires are interested in knowing which of them is richer without revealing their actual wealth. At this, the private inputs x_1 and x_2 reflect their respective fortunes, while the function f is defined as

$$f(x_1, x_2) = i \text{ such that } x_{i=} \max(x_1, x_2).$$

Other relevant problems in the field deal with private computations over data sets. For instance, assume a setting in which two entities, Alice and Bob, hold respective datasets A and B and are interested in knowing their intersection. Alice and Bob may thus be retail companies who want to identify shared clients or competing banks who want to cooperate in locating bad payers. Cryptographic techniques provide different solutions to that problem (referred to as the *private set intersection problem*) giving strong guarantees that Alice and Bob will, after the interaction learn nothing but the intersection A∩B, see **D'Arco** *et al.* (2017).

Even though they provide very strong (and provable) security guarantees, cryptographic techniques are often inefficient when it comes to processing large amounts of data in real time, or need unrealistic trust assumptions (like the existence of fully trusted resources). For the case of recommender systems, there are robust cryptographic solutions which rely on inefficient techniques like homomorphic encryption schemes and zero-knowledge proof protocols (**Tang**; **Wang**, 2017), and the references herein. The same problem can be solved through privacy-preserving techniques, such as data-obfuscation which relies on blurring the original information through noise addition to restrict the information leakage from recommender outputs.

5.2. K-anonymity

K-anonymity is a privacy preserving technique (**Soria-Comas**; **Domingo-Ferrer**, 2016). Its main goal is to limit the disclosure risk of a data set by restricting the capability of outsiders to re-identify a record in an anonymized public database. Informally k-anonymity assumes that in order to re-identify a record it is required to examine a fixed set of attributes linked to it. Such a set of attributes defines a so-called *quasi-identifier*. The trick in k-anonymity is to make the combination of values of a quasi-identifier in the anonymized data set to refer to at least *k* individuals. For instance, in the figures below we achieve, through deletion, 2-anonymity with respect to the quasi-identifier gender/religion since for any combination of these two attributes found in any row of the table there are always at least 2 rows with those exact attributes.

A major risk when using k-anonymity techniques arises when independent anonymized versions of the same database are disclosed; outsiders may gain advantage if there is enough overlap across the independent data releases.

In conclusion, even when cryptographic and privacy preserving techniques are evolving towards secure and efficient solutions for private data analysis, to avoid any risk it is crucial to make explicit the exact security level of the imple-

Name	Age	Gender	Zip code	Religion
María	8	F	27883	Hindu
Carlos	25	M	27338	Catholic
Stella	56	F	29554	Buddhist
Candela	21	F	31007	Lutheran
Raphael	59	M	45002	Hindu
Ignatius	84	M	32991	Lutheran
Eva	22	F	38221	Lutheran
Marius	35	M	23998	Lutheran

Non-anonymized database

Name	Age	Gender	Zip code	Religion
	8	F		Hindu
	25	M		Catholic
	56	F		Buddhist
	21	F		Lutheran
	59	M		Hindu
	84	M		Lutheran
	22	F		Lutheran
	35	M		Lutheran

Database; anonymized by suppression: 2-anonymous with respect to quasi-identifier "Gender-Religion"

Figure 3.

mented tools, and to what extent they are compatible (i.e., when implementing jointly different techniques, individual security properties are preserved).

Unfortunately, there are few examples of real-life long-scale deployment of cryptographic and privacy preserving techniques in the big data context. Little consumer awareness and poor international regulations are the main reasons for that; however, tables are turning expeditiously. As illustrative examples, we mention the *Sharemind* tool which presents a solution to securely collect and analyze financial data for a consortium of ICT (information, computing, and telecommunications) companies using MPC.
https://sharemind.cyber.ee

Further, privacy preserving solutions for pay-TV are investigated in **Biesmans** *et al.* (2017).

6. Conclusion. Final remarks

As a consequence of the widespread usage of business models based on big data, we are facing the permanent scrutiny of user consumption habits. The way these platforms are designed makes it hard to implement effective privacy preserving techniques, as some services are indeed designed through data surveillance. However, service providers are becoming aware of the increasing demand for more privacy-concerned services; along this line, for instance, is the Spanish telecom company *Telefónica*, which has launched a cognitive intelligence platform called *Aura*, which includes a fine-grained privacy policy allowing clients to decide which of their owned data should be shared and how. Furthermore, this platform will interact with other data businesses such as *Facebook*, *Google*, *and Microsoft*.

Some data-driven companies, like *Google*, base their business model on the idea that in order to improve their services they need full access to users' information, both obtained directly and indirectly, i.e., from usage as *Google* policy reports. All in all, they also offer different options so that clients can, to some extent, control their data, such as cookie management, the ability to forbid targeted ads or the option to regularly delete the search history. In conclusion:

- Many OTT companies and social networks collect data from their own users and this, despite making them more competitive, can easily turn into a bad practice.
- Users, in particular those that could be defined as technically-illiterate, are indeed the most vulnerable actors in this growing data-driven situation.
- Cryptography and privacy preserving techniques are likely to provide useful solutions towards secure and private deployment of business intelligence models.
- Citizens are increasingly demanding higher security and privacy guarantees; as a result, regulators and institutions must work to provide these guarantees.

7. References

Biesmans, Wouter; **Balash, Josep**; **Rial, Alfredo**; **Preneel, Bart**; **Verbauwhede, Ingrid** (in press). "Private mobile pay-TV from priced oblivious transfer". In: *IEEE transactions on information forensics and security*.
https://doi.org/10.1109/TIFS.2017.2746058

Brookman, Justin; **Hans, Gautam S.** (2013) "Why collection matters: surveillance as a de facto privacy harm". In: *Big data and privacy: making ends meet. The Center for Internet and Society. Stanford Law School*, September 10.
https://fpf.org/wp-content/uploads/Brookman-Why-Collection-Matters.pdf

CNMC (2017). "El teléfono móvil, el dispositivo más utilizado para conectarse a internet por los españoles". *Estudio panel de hogares CNMC.*
https://www.cnmc.es/node/365629

Cramer, Ronald; **Damgaard, Ivan-Bjerre**; **Nielsen, Jesper-Buus** (2015). *Secure multiparty computation and secret sharing.* Cambridge University Press. ISBN: 978 1316371404

Danezis, George; **Domingo-Ferrer, Josep**; **Hansen, Marit**; **Hoepman, Jaap-Henk**; **Le-Métayer, Daniel**; **Tirtea, Rodica**; **Schiffner, Stefan** (2015). *Privacy and data protection by design—from policy to engineering.* Technical report, Enisa 7.
https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design

D'Arco, Paolo; **González-Vasco, María-Isabel**; **Pérez-del-Pozo, Ángel**; **Soriente, Claudio**; **Steinwandt, Rainer** (2017). "Private set intersection: New generic constructions and feasibility results". *Advances in mathematics of communications*, v. 11, n. 4, pp. 481-502.
https://doi.org/10.3934/amc.2017040

Del-Castillo, Ignacio (2017). "Telefónica prevé lanzar 14 series propias en dos años". *Expansión*, 17 January.
https://goo.gl/qzfHBb

Dorard, Louis (2014). *Bootstrapping machine learning*. Create Space Independent Publishing Platform. ISBN: 1500789240

Fernández-Manzano, Eva-Patricia (coord.) (2016). *Big data. Eje estratégico en la industria audiovisual*. Barcelona: Editorial UOC. ISBN: 978 84 9116 380 0

Fernández-Manzano, Eva-Patricia; **Clares-Gavilán, Judith**; **Neira, Elena** (2016). "Data management in audiovisual business: *Netflix* as a case study". *El profesional de la información*, v. 25, n. 4, pp. 568-576.
https://doi.org/10.3145/epi.2016.jul.06

Gómez-Uribe, Carlos A.; **Hunt, Neil** (2016). "The Netflix recommender system: Algorithms, business value, and innovation". *AMC Transactions on management information systems (TMIS)*, v. 6, n. 4.
https://doi.org/10.1145/2843948

Herold, Rebecca (2010). *Managing an information security and privacy awareness training program*, 2nd ed., CRC Press. ISBN: 978 1 420031256

Liu, O.; **Man, K. L.**; **Chong, W.**; **Chan C. O.** (2016). "Social network analysis using big data". In: *Procs of the Intl multiconf of engineers and computer scientists* 2016, v. II, Imecs 2016, March, 16-18, Hong Kong. ISBN: 978 988 14047 6 3
http://www.iaeng.org/WCE2014/publications.html

Madrigal, Alexis (2014). "How Netflix reverse engineered Hollywood". *The Atlantic*, 2 January.
https://goo.gl/Gy6miE

Meis, Rene; **Heisel, Maritta** (2016). "Understanding the privacy goal intervenability". *Lecture notes in computer science*, v. 9830, pp. 79-94.
https://goo.gl/HofxL3
https://doi.org/10.1007/978-3-319-44341-6_6

Ortiz-López, Paula (2016). "Aspectos legales de la gestión de datos en la publicidad digital". En: Martínez-Pastor, Esther; Nicolás-Ojeda, Miguel-Ángel. *Publicidad digital*. Madrid: Esic Editorial, pp. 187-202. ISBN: 978 84 16701 13 1

Pelrich, Claudia (2016). "The Super Bowl ads and moments that best caught -and kept- viewers' attention". *Dstillery.com*, 18 February.
https://goo.gl/P1izP6

Preibusch, Sören (2015). "Privacy behaviors after Snowden". *Communications of the ACM*, v. 58, n. 5, pp. 48-55.
https://doi.org/10.1145/2663341

Prieto, M. (2017) "En Telefónica tenemos datos más valiosos que los que atesoran los actores digitales". *Expansión*, 22 June.
https://goo.gl/HWb53W

Sáiz, Carlos-Alberto (coord.) (2017*). Código de buenas prácticas en protección de datos para proyectos big data*. Agencia Española de Protección de Datos (AEPD); Asociación Española para el Fomento de la Seguridad de la Información, ISMS Forum Spain.
https://goo.gl/k4vR5W

Soria-Comas, Jordi; **Domingo-Ferrer, Josep** (2016). "Big data privacy: Challenges to privacy principles and models". *Data science and engineering*, v. 1, n. 1, pp. 21-28.
https://doi.org/10.1007/s41019-015-0001-x

Tang, Qiang; **Wang, Husen** (2017). "Privacy preserving hybrid recommender system". In: *Procs of the 5th ACM Intl workshop on security in cloud computing*, pp. 59-66.
https://doi.org/10.1145/3055259.3055268

Warwik, Ashford (2016). "UK firms could face £122bn in data breach fines in 2018". *ComputerWeekly.com*, 17 October.
https://goo.gl/i63FSP

Wolk, Alan (2015). *Over the top, how the internet is (slowly but surely) changing the television industry*. CreateSpace Independent Publishing Platform. ISBN: 978 1 514139011

Yao, Andrew C. (1982). "Protocols for secure computations" In: *23rd Annual symposium on foundations of computer science* (FOCS 1982), pp. 160-164.
https://doi.org/10.1109/SFCS.1982.88