2018

Profiling with Big Data: Identifying Privacy Implication for Individuals, Groups and Society

Paola Mavriki
University of the Aegean, Department of Information and Communication Systems Engineering, Greece, pmavriki@aegean.gr

Maria Karyda
Dept. of Information and Communication Systems Engineering, University of the Aegean, mka@aegean.gr

PROFILING WITH BIG DATA: IDENTIFYING PRIVACY IMPLICATIONS FOR INDIVIDUALS, GROUPS AND SOCIETY

Research full-length paper

Mavriki, Paola, University of the Aegean, Department of Information and Communication Systems Engineering, Greece, pmavriki@aegean.gr

Karyda, Maria, University of the Aegean, Department of Information and Communication Systems Engineering, Greece, mka@aegean.gr

Abstract

User profiling using big data raises critical issues regarding personal data and privacy. Until recently, privacy studies were focused on the control of personal data; due to big data analysis, however, new privacy issues have emerged with unidentified implications. This paper identifies and investigates privacy threats that stem from data-driven profiling using a multi-level approach: individual, group and society, to analyze the privacy implications stemming from the generation of new knowledge used for automated predictions and decisions. We also argue that mechanisms are required to protect the privacy interests of groups as entities, independently of the interests of their individual members. Finally, this paper discusses privacy threats resulting from the cumulative effect of big data profiling.

Keywords: Profiling, Privacy Implications, Big Data Analytics, Group Privacy.

1 Introduction

A growing part of human activities is mediated today by digital services and devices (Kosinski et al., 2016) and people are continuously generating vast amounts of data by revealing personal data willingly or unwillingly, as they perform everyday online activities such as shopping, communicating with family members, paying taxes, reading the news etc. (Mai, 2016). This data can be used to assist users in their internet experience, by providing them personalized services (Vemou and Karyda, 2015) but it also may be used through business intelligence and analysis tools for opinion mining, e-marketing, political e-marketing, social network analysis, eGovernment services, etc. (H. Chen et al., 2012).

The term "big data" has been used to describe the data sets and analytical techniques that are so large and complex that they require advanced and unique data storage, management, analysis, and visualisation technologies (H. Chen et al., 2012). Volume, Velocity, Variety (Laney, 2001), Veracity (Storey and Song, 2017) are key attributes used in literature to describe big data while some authors refer also to the Value of big data which derives from the new knowledge gained from the applications of advanced analytics technologies to big data sets (Nguyen et al., 2013; Rubinstein, 2014).

Big data analytics is increasingly attracting the interest of the academic world, industry and governments and personal data has become a raw material of production as a new source of great economic, scientific and social value (Tene and Polonetsky, 2012). Businesses focus on acquiring relevant data about customers as part of their marketing and sales strategies. Simultaneously, they try to identify individuals who may be persuaded to become their new customers and under which conditions (Hildebrandt, 2012). Marketing campaigns can achieve an efficient segmentation of the electorate, identify targets and communicate effectively. Focus group data, polling and advanced statistical analysis provide them with critical information they need to create increasingly refined voter segments (Conley,

2018). In political or commercial e-marketing, companies, parties or interest groups, are not just searching for the attributes of predefined classes of (potential) customers, voters or followers, but they invest in finding out which classes they should distinguish in the first place (Hildebrandt, 2012).

User profiling is one of the more controversial technologies regarding personal data and privacy (Pandit and Lewis, 2018). Until recently, privacy studies were focused mainly on the control of personal data. Due to big data analysis, however, new privacy issues have emerged with unidentified implications: for instance, knowledge about an individual may be revealed even if her/his personal data is protected (Hildebrandt, 2008a). Furthermore, accurate personality profiles may be constructed by processing digital footprints such as Facebook likes for example (Kosinski et al., 2016).

Hildebrandt, (2012, 2009, 2008b, 2008a, 2006) discusses profiling as pattern recognition and introduces key distinctions between personalised and group profiling. Recently, some authors (Floridi, 2014; Mittelstadt, 2017; Taylor et al., 2016) argue that profiling and machine learning technologies are employed at group level as are used to formulate types, not tokens as they work to scale, enabling their users to target collective entities as well as individuals. They also suggest that, since most people are not targeted by profiling and machine learning technologies as individuals, but as members of specific groups, the privacy of these groups needs to be examined further and considered in the context of data protection. Jaquet-Chiffelle (2008), employing the terminology provided by Hildebrandt (2008b), conceptualize group profiling as an indirect process of profiling but, new applications of data technologies are reshaping the definition of targeting. New types of profiling by new applications of data technologies cannot be explained from the indirect profiling conception perspective: groups of people may be targeted today only by discovering their presence in a particular place at a particular time (Taylor, 2017).

Furthermore, it seems that using profiling technologies for personalization may have long-term effects on society, as in the case, for example, of political marketing, where targeted content can make elections less fair as potential voters are only exposed to specific information (Goodman et al., 2017). According to Baruh and Popescu (2017), '..as long as regulatory efforts center on individual privacy literacy and self-management but fail to recognize the nature of privacy as both a collective value and a collective social phenomenon, these efforts are destined to fail'. They also suggest that there is a need to consider the collective aspects of privacy, to develop 'new ways of calculating privacy risks, as well as new ways to frame risk information to expand the benchmark used by individual decision-makers.'

Addressing this gap, this paper discusses the implications of recent big data-driven forms of profiling and identifies new privacy threats that emerge, for individuals, groups and society. Whereas related literature explores privacy threats and implications for individuals, this paper extends the scope adopting a multi-level approach that considers threats and implications for collective entities (groups) and society as well.

The remaining paper is structured as follows: in the following section, we discuss intelligent profiling technologies raising new challenges for privacy. Section 3 analyzes privacy risks for individuals, groups and society that stem from, these applications, while Section 4 describes privacy implications for individuals, groups and society. Section 5 concludes the paper with a discussion and further research.

2 Intelligent profiling with big data and new privacy challenges: The current landscape

In the age of big data, the competitive advantage of data collection is no longer dependent on the volume of data but on the differentiation of information from the noise and the simultaneous retention of data that is noise today but can become information tomorrow. This goal is achieved through profiling: 'the process of discovering correlations between data in databases that can be used to identify and represent a human or nonhuman subject (individual or group) and/or the application of profiles (sets of correlated data) to individuate and represent a subject or to identify a subject as a member of a group

or category (Hildebrandt, 2008b)'. Discovered correlations and patterns may be indicative of expected future behaviour. In this context, a profile may be characterized as knowledge that allows to differentiate the relevant from the irrelevant data (Gutwirth et al., 2012; Hildebrandt, 2009, 2006).

Big data techniques employ a variety of tools used to discover knowledge in high volume, highly dynamic, and highly heterogeneous data. Network analysis, sentiment analysis, trust and reputation management, machine learning, cluster analysis are some examples of big data techniques which may result in highly comprehensive user profiles (Hasan et al., 2013). Profiling applications pursue cover a wide variety of purposes ranging from anti-terrorism to direct marketing (Schermer, 2011), while the of knowledge included in a user profile varies accordingly to the purposes and the domain it is used.

Schermer (2011), identifies two distinct approaches in the use of data mining for profiling: "descriptive data mining" and "predictive data mining". While the goal of descriptive data mining is to discover unknown relations between different data objects in a database, the goal of predictive data mining is to make a prediction based on patterns that were determined using known information. Related to profiling, this means that information about an individual is mined in order to determine whether she fits the previously established profile (Schermer, 2011).

Similarly, according to Mittelstadt (2017), groupings can occur in two senses: either in the description of subjects or actions taken on the basis of probabilities and predictive analytics. For the former, data subjects must always be considered along a limited set of dimensions. For predictive analytics, dimensions can also refer to non-descriptive choices, for instance the choice to deliver a certain advertisement due to observations of prior actions.

Profiling application can facilitate users through providing personalized services and can also be useful in cases such as fraud detection; at the same time however, privacy issues emerge. In the age of big data, as Marx (1998), predicted twenty years ago, 'new technologies have the potential to reveal the unseen, unknown, forgotten, or withheld'. In the following we investigate automated profiling for individuals, groups and communities with big data analysis.

2.1 Profiling individuals

Information contained in a user profile can be either provided explicitly by the user or inferred by the service that manages the profile. Common information contents of user profiles include: user interests; user knowledge, user background and skills; user goals; user behaviour; user individual characteristics; and user context. This content can be considered as private information and it may be harmful for a person to reveal it (Hasan et al., 2013). However, in the case of big data analysis, one major new challenge for privacy protection is the blurring of boundaries between public and private information: 'massive amounts of data are publicly available and can be freely scraped from online platforms and environments' (Kosinski et al., 2016).

Today, personal knowledge about an individual may be revealed even if her explicitly provided personal data is protected: digital footprints can be used to infer personal details. Kosinski et al. (2013) model of psychological profiling for example, uses Facebook Likes to predict a range of highly sensitive personal attributes including: sexual orientation, intelligence, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.

As big data techniques can process high volume data from multiple sources user data from different sources can be linked and aggregated into a single user profile. User information from different sources can also be correlated to validate the information discovered from one source (Hasan et al., 2013). For example, Buraya et al. (2017), use Twitter and Instagram data to increase the performance of personality profiling and Farseev and Chua (2017), integrate data from sensors and multiple social media sources.

In addition, big data techniques can process unstructured as well as structured data. Unstructured data of different varieties generated by users is growing in volume with high velocity and contains lots of

useful information about the users (Hasan et al., 2013). Farnadi et al. (2018), for example, incorporate multiple types of data (e.g. textual, visual and relational content) which users generate on social media platforms. Advertisers have recently started assembling shopping profiles of individuals based on compilations of publicly available metadata such as the geographic locations of social media posts (Ekbia et al., 2015).

The fact that big data gathering is based on many different sources entails loss of control of their information for individuals, as in many cases they are not even aware of the processing or cannot trace how data are flowing from one system to another (D'Acquisto et al., 2015).

Predictive algorithms are used by the finance industry for credit scoring and trading and are increasingly rating people in countless aspects of their lives, to assess whether they are good credit risks, desirable employees, reliable tenants, valuable customers (Provost and Fawcett, 2013). One example of using predictive algorithms is the case of self-help health and fitness approaches, such as "Quantified Self" and "PatientsLikeMe," generate data sets that help identify or predict health attributes. When these data sets are cross-referenced with traditional health information it is possible to generate a detailed picture about a person's health, including information the person may never have disclosed to a health care provider. By combining the use of these data sets with predictive analytics, big data can dramatically increase the amount of related data that may be considered private (Crawford and Schultz, 2014).

Finally, an individual profile may be built by using behavioural biometrics either by extracting profiling information from the measured biometric such as the determination of gender from a person's voice for instance, or by performing identification or verification of the individual and using the person's identity to link current information to other data related to this person (such as identifying a person's signature in a financial transaction and using the result of this identification to monitor all of the person's financial transactions together (Yannopoulos et al., 2008). The next generation of biometrics includes the measurement. and analysis of new biometric traits, such as behavioural or soft biometrics (i.e. biometrics which may change over time, such as gait analysis) and physiological biometrics (including heartbeat detection, pheromone detection) (Finn et al., 2013). Simultaneously with their wider deployment, use of biometrics raises new privacy and data protection issues.

2.2 Profiling groups and communities

Often, big data profiling is employed for identifying target groups. In many cases, group profiling is preferable to individual profiling, because it is more cost efficient than considering each individual profile (e.g. individuals need to be approached by letter or by phone while groups may be approached by an advertisement or a news item) (Custers, 2013). Group profiling is used either to find shared features between members of a predefined community or to define categories of individuals sharing some properties (Jaquet-Chiffelle, 2008). Group profiles may contain known information (e.g. 'people who smoke live, on average, a few years less than people who do not'). Yet, group profiles may contain also new facts (e.g. 'people living in zip code area 8391 may have a larger than average chance of having asthma') (Custers, 2013).

Classification and clustering are mainly used to identify groups (Custers, 2013). People are grouped together by their qualitative attributes and habits (e.g. "low-income people", "working class mom", "metro parents") and predict the future behaviour of these clusters of individuals (Mantelero, 2016). Group profiling is also used for political marketing, as segmentation allows a more efficient allocation of communication resources and helps campaigns to identify target audiences and then get them out to vote (Davidson and Binstock, 2011). Traditionally, voters are grouped on the basis of religion, ethnicity, race, income, education, profession, party identification etc. More recent groupings however, are based on combinations of age, ideology, and lifestyle (Russmann, 2016).

There is also an increasing demand for a precise identification of social networks communities and groups (e.g. members of a family, colleagues, fans of a brand, political groups etc.) coming from brand monitoring, business intelligence and e-reputation management (Gadek et al., 2017).

Xie et al. (2012), identify user communities from folksonomy data, while Abdelsadek et al. (2018), approach of social media analytics allows the visual revealing of the community structure and the related characteristics on Twitter. Also, groups can be identified and classified for recommendation systems from analysing and processing group photos using metadata embedded in images (the time stamp and GPS coordinates of the place where the photo was taken) (Y.-Y. Chen et al., 2012).

A group of individuals may also be identified by monitoring activities in SNS and looking for profiles that check in at the same location, within the same interval of time and are connected through social ties. Group identification can be improved by monitoring at the spatio-temporal features and, in particular, by comparing the user trajectories (Gasparetti, 2017). The group's future behaviour may also be predicted, through based on patterns of past behavior predictive analytics, while researchers can use it to observe the entire population (Spencer, 2015). Finally, group profiling can also be performed through surveillance technologies and monitoring technologies such as RFID-enabled travel documents, unmanned aircraft systems (Finn et al., 2013), GPS technologies included in mobile phones (Asakura and Iryo, 2007), CCTV etc.

Profiling with big data analytics may raise problems that are different from the problems that may be raised by forms of statistical profiling. For example, in data mining, depending on the technique that is used, every possible relation can be investigated, while, in empirical statistical research, usually only causal relationships are considered. The relations found using data mining are not necessarily causal or they may be causal without being understood. In this way, the scope of profiles that are discovered may be much broader (only a small minority of all statistical relations is directly causal) with unexpected profiles in unexpected areas (Custers, 2013).

Moreover, analytics allows for a new type of group to be formed that is not protected by anti-discrimination provisions, because the groups need not align with existing protected classes or attributes (Mittelstadt, 2017). According to Taylor (2017), people may be profiled with big data without being personally identified. Therefore, 'the way that current understandings of privacy and data protection focus on individual identifiability becomes problematic when the aim of an adversary is not to identify individuals, but to locate a group of interest – for example an ethnic minority, a political network or a group engaged in particular economic activities'.

Concluding, sophisticated applications are capable today to capture, store, use and analyse automatically countless facets of people's lives at individual, group and community level. At the same time, as more information on personality, finances, location, relations and online activity is collected and processed, protecting privacy is becoming more and more complex and difficult. New technologies and their applications bring new challenges for privacy which need to be examined. In the next section we address this issue by analysing privacy threats for individuals, groups and society stemming from profiling with big data analysis.

3 Identifying privacy threats

Westin (1967), defined the right to privacy as the right 'to determine how much of their personal information is disclosed and to whom, how it should be maintained and how disseminated', arguing that privacy operates at the individual, group, and organisational/ institutional level. Then years later, Altman (1977), conceptualized privacy 'as the selective control of access to the self, involving dialectic, optimization, and multimodal processes', seeing privacy as 'a boundary control process whereby people sometimes make themselves open and accessible to others and sometimes close themselves off from others'. Since then, many privacy definitions have been formulated (Smith et al., 2011). Relatively recent privacy conceptualizations analyse privacy focusing on the flow of personal information

in terms of access to and control over personal information and many authors use the expression ‘‘informational privacy” (Tavani, 2007).

However, according to Mai (2016), in the age of big data, the “datafication” of personal information shapes a new type of information society hence the issue of privacy has to be approached from different perspectives. Through quantification (datafication), data (numbers) and information (text, music, movies etc.) are transformed in elements that can be analysed in more sophisticated ways and across large data sets for patterns and correlations. Today, most digital devices are connected to the Internet and many of people’s daily activities are digitally mediated and connected. In the near future, datafication might include “everything”. Therefore, the traditional conceptions of informational privacy are challenged as in the age of big data “controlling personal data” is losing its meaning. Introducing the datafication model of privacy, Mai (2016), changes the focus from data collection to data processing and analysis and argues that privacy in the age of big data concerns about the new insights that others can generate based on the already available data. Similarly, Tene and Polonetsky (2012), argue that in the age of big data privacy is not only about controlling individual fragments of personal data that the individual has an interest to control, but privacy concerns also arise as the information is produced about individuals as they are sorted and classified for specific purposes.

Related to group privacy, Mittelstadt (2017), recognizes three types of groups as potential rightsholders: collectives (labour unions, political groups etc.), ascriptive groups (ethnic groups, patient cohorts) and Ad hoc groups (examples market segments, profiling groups). In some context collectives and ascriptive groups are already legally recognized as legitimate rightsholders. Collective rights include the right to self-determination held by nations, or legal rights granted to corporations (List and Pettit 2011, in Mittelstadt, 2017). Ad hoc groups however, lack both collective identity and agency.

According to Floridi (2017), profiling is not a descriptive practice, ‘it is a designing one, and it comes with the consequence of creating the condition of possibility of the profiled individuals, now constituted as a group by the very act of profiling, to act as a group in order to claim respect for its own privacy’. Also, Floridi claims that privacy requires a radical re-interpretation in the age of information: ‘a re- interpretation is achieved by considering each individual person or group as constituted by his, her or its information, and hence by understanding a breach of an individual’s informational privacy as a form of aggression towards that individual’s identity’. Considering this perspective, a group right to privacy may understood as the right to privacy as a right to immunity from unknown, undesired, or unintentional changes in one’s own identity as an informational entity, both actively (because collecting, storing, reproducing, manipulating) and passively (Floridi, 2017).

Moreover, Barocas and Nissenbaum (2014), (in Taylor, 2017), have pointed out that the use of big data poses new questions to do with privacy on the group level, in contrast to the individual level on which it has traditionally been conceptualized. They point out the difference between single digital databases from big data: as the last it is used in aggregated form, where harm is less likely to be caused by access to personally identifiable information on individuals and more likely to occur where authorities or corporations draw inferences about people on the group level. Their conceptualization of the problem suggests that if it is to remain relevant, the idea of privacy must be stretched and reshaped to help us think about the group as well as the individual.

Related research has identified significant new privacy threats stemming from profiling with big data analytics, which are related to predictive analytics and to data-driven automated decision-making (Ekbia et al., 2015; Mantelero, 2016; Tene and Polonetsky, 2012). In the following we discuss privacy threats related to profiling based on big data analysis, extending the scope of current research to include not only privacy treats and implications for individuals, but also for collective entities (groups of individuals) and society.

3.1 Threats for individual privacy

Predictive analytics poses new privacy threats for individuals, as new information, generated by predictive algorithms is beyond the individuals' control. Also, when new personal information is produced, it is not clear who owns this information or has the right to it (Mai, 2016). However, further knowledge created about individuals may not necessarily be related to the initial purposes of data collection (Mantelero, 2016).

Another stream of individual privacy threats related to automated profiling, stem from the accumulation of personal data (incremental effect). Fragments of data regarding an individual user may be linked peace by peace until an individual profile is entirely exposed (Tene and Polonetsky, 2012). According to Solove (2005), aggregation (gathering together of information about a person) can cause dignitary harms because of how it unsettles expectations. Aggregation can also lead to power asymmetries, as it can increase the power that others have over individuals. Moreover, data compilations may be both telling and incomplete. Aggregated data may reveal facets of people's lives, but the data is often reductive and disconnected from the original context in which it was gathered and this leads to distortion (Solove, 2005).

A well- known, but aggravated by big data analysis threat for individual privacy is lack of access and exclusion which deepens the information asymmetries (Tene and Polonetsky, 2012). Exclusion reduces accountability on the part of government agencies and businesses that maintain records about individuals. Harms related to this type of privacy threat are explored by Solove (2005), who, among others argues that 'exclusion it is a harm created by being shut out from participating in the use of one's personal data, by not being informed about how that data is used, and by not being able to do anything to affect how it is used.'

Furthermore, personal information of location-based services users may be deduced from temporal and spatial dimensions exposing users to several privacy threats. For example, workplace, home, information about personal preferences, shopping habits, dating habits, driving speeds etc. may be deduced (Benoist, 2008). Generally, in Location- Based Services (LBSs), two major categories of threats are involved: the release of location information (when the user's identity is known) and re-identification through location referring to an adversary's ability to reduce a user's degree of anonymity by considering location information. For instance, by knowing that an anonymous user of a geosocial network was in a given place at a given time, can reveal information such as health problems, affiliations, and habits. If the user considers his involvement in the geosocial network to be sensitive, re-identification is a privacy violation (Vicente et al., 2011).

Finally, the threat of identification is further aggravated by big data analysis. In some online networks, user profiles and relationship data are public, but many users maintain pseudonymous profiles. Narayanan and Shmatikov (2009), show that anonymity is not sufficient for privacy in social networks. They developed a generic re-identification algorithm and showed that it can successfully de-anonymize several thousand users in the anonymous graph of a popular microblogging service (Twitter), using a completely different social network (Flickr) as the source of auxiliary information. Because it connects people to data, identification attaches informational baggage to people. Identification can also inhibit one's ability to be anonymous or pseudonymous while anonymity and pseudonymity protect people from bias based on their identities and enable people to vote, speak, and associate more freely by protecting them from the danger of reprisal (Solove, 2005).

3.2 Privacy threats for collective entities (groups and communities)

Big data analytics group people together by their qualitative attributes and habits (e.g. low-income people, "working class mom", "metro parents") and predict the future behaviour of these clusters of individuals. According to Hildebrandt (2008a), the identified groups can consist of people that think of themselves as a community such as members of an association (collectives and ascriptive groups), but they also can consist from people who are not necessarily aware that they belong to it (Ad-hoc

groups). The re-identification of groups may be considered a group privacy threat as according to Floridi (2014), '…re- identifiable groups are ipso facto targetable groups'. Also, problems may occur from the hundreds of different variables used by big data analysis to infer predictive information about groups. In many cases, these variables concern aspects that are not clearly related to the final profiles created by analytics (Mantelero, 2016).

One category of privacy threats stemming from group profiling involves data-driven decision making. In this context, the main target of the collective dimension of data processing is not the data subject, but the clusters of people created by big data gatherers (Mantelero, 2016). For example, groups may be sorted and classified into categories based on their deduced economic and political value (Clarke, 1999). In the case which predictive data mining is focussed on characteristics such as ethnicity, gender, religion or sexual preferences, discrimination of groups may occur (Schermer, 2011). When decisions are taken on the basis of this kind of generalization, groups may also be discriminated (Hildebrandt, 2008a). Moreover, 'discrimination might be transferred to new forms of population segments, dispersed throughout society and only connected by one or more attributes they have in common. Such groups will lack political force to defend their interests and might not even know what is happening' (Custers, 2013). By identifying patterns in the behaviour of groups of individuals and taking decisions may also affect the internal dynamics of groups, with consequences for the collective issues of the people involved (Mantelero, 2016).

Further privacy issues for groups are related to location-based services. Ashouri-Talouki et al. (2012), differentiate three major privacy issues for groups: the safeguard of the location privacy of each group member inside the group (intragroup location privacy), the preservation of the location privacy of each group member from anyone outside the group, and the protection of the location privacy of the meeting place in the case of a secret meeting.

Moreover, there are also privacy issues stemming from being member of a group which may involve one or more members of the group, but not necessarily the entire group. In the case of algorithmically grouping for instance, as previously mentioned, a group's members are not necessarily aware that they belong to it while the profiles generated by mining other people's data are often applied to individuals whose data match the profile. Therefore, since the implications of profiling for the autonomy of individuals are not related with their personal data, an individual can be discriminated on the basis of behaviours of other people in his group (Hildebrandt, 2012; Taylor et al., 2016).

Finally, one more category of privacy threats is stemming from surveillance. Because of its inhibitory effects, surveillance can be a tool of social control, enhancing the power of social norms, which work more effectively when people are being observed by others in the community (Solove, 2005). Surveillance may also violate privacy of association (Finn et al., 2013) while freedom of speech and religion are largely collective rights, requiring association for their full expression. According to Fisher, (2004), 'Participation in intermediate associations enhances democracy in a number of ways: it reduces alienation by cementing bonds between people, it trains citizens for democratic participation, and it gives them influence over group expression and action, thereby inculcating civic virtue. It also enhances popular sovereignty by amplifying the individual voice, joining it to that of the larger group. In addition, associations provide a buffer between individuals and the State, and help prevent the State from exerting overweening power against individuals'.
Investigating Facebook as a prototypical example of web 2.0 surveillance Fuchs (2011), refers among others to digital inequality, lack of democracy, the attempted manipulation of needs, limitation of the freedom to choose and intransparency.

Concluding, new privacy threats stemming from intelligent profiling are related mainly with the predictive nature of the generated through big data analysis new knowledge. Also, previously identified privacy threats may be aggravated. In addition, we argue that privacy threats may have multilevel dimensions. Through the cumulative effect, they may affect beside the individual, groups and society leading to complex and various implications which we address in the next section.

4 Discussion: Implications for individuals, groups and society

Traditionally, profiling was based mainly on a few standard variables (e.g. sex, age, family income, place of residence) their predictive ability was limited. Today, big data analytics use hundreds of different variables to infer predictive information about people (Mantelero, 2016) giving a high degree of complexity and variety to the privacy implications stemming from profiling. A deep and detailed investigation of this issue is possible by taking into consideration every domain of its application. Due to space limitation however, we refer broadly to privacy implications stemming from profiling in the commercial, political and security area.

In a commercial and political marketing context, big data-driven profiling is used to a great extent for scoring systems which through predictive algorithms are rating people in countless aspects of their lives. Citron and Pasquale (2014), study the implications related to scoring systems and argue that 'the realm of management and business more often features powerful entities who turn individuals into ranked and rated objects'. For instance, the case of predicting the pregnancy of a teenager has been widely discussed: the retail chain Target's predictive analytics "guessed" that a customer was pregnant and disclosed her name to their marketing department, building personally identifiable information about her without collecting it directly (Crawford and Schultz, 2014). Users whose data are being mined do not have the means to anticipate what the algorithms will come up with, they do not know how they will be categorised or the consequences (Hildebrandt and Gutwirth, 2008). As Tene and Polonetsky (2012) point out, 'the online company knows the preferences of the transacting individual inside and out, perhaps better than the individual knows him or herself'. Yet, the information asymmetries between costumers and online companies is 'like a game of poker where one of the players has his hand open and the other keeps his cards close'.

Nickerson and Rogers (2014), investigate the same issue in a political marketing context where predictive models (behavior scores, support scores, and responsiveness scores) are used to make targeting campaign communications more efficient and to support broader campaign strategies. The "products" to market in political marketing are policies, ideas, programmes, principles and beliefs (Baines et al., 2003) and the implications related to targeting citizens aiming to influence them may have long-term effects on society. For example, by targeting users with information that appeals to them, the groups established perceptions and beliefs are accentuated. Also, targeted content can make elections less fair as potential voters are only exposed to limited information. Moreover, targeted messaging can increase the focus on divisive issues as message targeting to the individual concerns of citizens as part of a group (Goodman et al., 2017).

Even more, in areas of limited statehood, political targeting of people may be potentially life-threatening. In environments where the state lacks the capability and accountability mechanisms necessary to protect against physical and digital privacy violations, identification and association with groups facing demographic-based discrimination may result in aggression against both actual and perceived group members (e.g. in the election-related violence in Kenya in 2007–2008, in the Rwandan genocide of 1994 and in the conflict in the Central African Republic in 2013–2014) (Kammourieh et al., 2017; Taylor, 2017). Moreover, targets of political surveillance are chilled in the exercise of their rights to engage in free speech and the free exercise of religion while, suffering actual or potential damage to their reputations, they change their behaviour (Fisher, 2004).

Knowledge- based policing also comes with a number of implications for individuals and groups. According to Leese (2014), in this case, data-driven profiling differs considerably from the traditional profiling practices as 'profiling is enacted in a confirmatory or hypothesis-testing way to explore whether certain patterns of characteristics are represented in the analysed population data and, if so, to put the identified individuals under scrutiny'. Predefined profiles variables may include gender, nationality, religious beliefs raising critique in terms of social sorting or racial profiling. According to Tsoukala (2010: 47–48 in Leese, 2014), 'in risk-based policing, it has been shown that certain societal subgroups have been identified as high-risk sections of populations and have been repeatedly discrim-

inated against for instance, North African youths in French suburbs, football supporters in the UK, or Roma people in Italy'.

In the security and social policies field, big data analysis can lead to a transparency paradox: citizens become increasingly transparent to government, while the profiles, algorithms and methods used by government organisations are obscure to citizens. This may result a shift in the balance of power between state and citizen in favour of the former. The secret nature of activities in the field of security reinforces this transparency paradox (Broeders et al., 2017). Surveillance in particular, may be performed by governments and it may have profound implications for freedom and democracy. Because of its inhibitory effects, surveillance can be a tool of social control (Solove, 2003).

Generally, big data analysis and profiling may reinforce social stratification by reproducing and reinforcing the bias that is present in every dataset as data are extracted through observations, computations, experiments and record keeping (Broeders et al., 2017). The 'data cleaning' process involves decisions about what attributes and variables will be used, and which will be ignored (Boyd and Crawford, 2011). Thus, data may be 'inherently partial, selective and representative', while the criteria used in their capture can distort the results of data analyses (Broeders et al., 2017) raising concerns related to the reliability and accuracy of the resulting from big data processing profiles (Boyd and Crawford, 2011). If uncorrected, the bias that potentially characterises every dataset may lead to discrimination and unfair treatment of particular groups in society. When used on a large scale, the results of big data analyses may increase social and economic inequalities. Furthermore, through data-driven profiling individuals may be judged based correlations and inferences of what they might do, rather than what they actually have done. This it is contradictory with the cornerstone of criminal law: the presumption of innocence (Broeders et al., 2017).

5 Conclusions

Big data analytics are reshaping the digital environment elevating profiling technologies to a higher level of performance. The complexity and obscurity of the data processing and the predictive character of the inferred from data new knowledge, are some of the new features of profiling which complicate and even confuse under some circumstances the privacy implications stemming from it. As Calo (2011), argues, privacy harms are characterized 'by an absence of understanding, a vague discomfort punctuated by the occasional act of disruption, unfairness,...'.

When discussing privacy, there is a tendency of focusing on data collection. This paper argues that in contrast with traditionally profiling, one of the biggest changes and challenges related to privacy is that protecting personal data it is not sufficient anymore: big data analytics may infer personally identifying information only by using digital footprints left behind by any user of any digital service. Profiling with big data analysis aggravates existing privacy concerns: e.g. identification of individuals or group is much easier now. It also poses new privacy issues, such as the need to shift attention from controlling personal data to controlling the new knowledge generated, which in many cases may involve sensitive issues such as health, race, or sexuality.

But another characteristic of profiling with big data analytics is categorization. Generalization and categorization raise privacy issues at many levels: we show in this paper that the new forms of profiling are focused beside the individual on groups and communities challenging their collective privacy interests. In some cases, such as political marketing for instance, political parties are interested in discovering and targeting groups of potential supporters and voters rather than individuals. According to Mantelero (2016), issues related to privacy that arise from this new situation are different from the traditional issues of individual privacy and group privacy. 'We are neither in the presence of forms of analysis that involve only individuals, nor in the presence of groups in the traditional sociological meaning of the term, given group members' lack of awareness of themselves as part of a group and the lack of interactions among people grouped into various clusters by data gatherers'.

However, as Mittelstadt (2017), argues, advances in data analytics necessitate new controls for the privacy interests of ad hoc groups formed by algorithmic classification. Mechanisms are required to protect the privacy interests of groups independent of the interests of their individual members. 'Anti-discrimination mechanisms built around offline identifiers are thus insufficient to protect groups constructed by algorithmic systems from harmful decisions based on attributes that are not merely proxies for legally protected attributes (e.g. ethnicity, gender)'.

Even more, as Calo (2011) points out, 'privacy harm is not merely individual, …but can lead to societal harms..'. We claim in this work that in particular, profiling for political purposes may have implications on the society level. Furthermore, we argue that privacy harms may have a cumulative effect as privacy implications are taking a high degree of complexity and variety and also extend from individuals to groups and from groups to the whole society comprising even those individuals and groups whose privacy are not threaten directly.

Finally, this paper has identified and analyzed privacy threats stemming from data-driven profiling at three levels: individual, group and society; we show that new privacy threats are related mainly with the generation of new knowledge which may be used for automated predictions and decisions. For a deeper understanding there is a need to investigate privacy threats and implications in each domain of profiling application focusing especially on new knowledge generation. Also, there is a need to examine privacy from a multilevel perspective as it seems that under specific circumstances, privacy of groups need also to be considered in the context of data protection.

References

Abdelsadek, Y., Chelghoum, K., Herrmann, F., Kacem, I., Otjacques, B., 2018. "Community extraction and visualization in social networks applied to Twitter". *Information Sciences* 424, 204–223. https://doi.org/10.1016/j.ins.2017.09.022

Altman, I., 1977. "Privacy Regulation: Culturally Universal or Culturally Specific?" *Journal of Social Issues* 33, 66–84. https://doi.org/10.1111/j.1540-4560.1977.tb01883.x

Asakura, Y., Iryo, T., 2007. "Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument". *Transportation Research Part A: Policy and Practice, Success and Failure of Travel Demand Management: Is Congestion Charging the Way Forward?* 41, 684–690. https://doi.org/10.1016/j.tra.2006.07.003

Ashouri-Talouki, M., Baraani-Dastjerdi, A., SelçUk, A.A., 2012. GLP: "A cryptographic approach for group location privacy". *Computer Communications* 35, 1527–1533.

Baines, P.R., Worcester, R.M., Jarrett, D., Mortimore, R., 2003. "Market Segmentation and Product Differentiation in Political Campaigns: A Technical Feature Perspective". *Journal of Marketing Management* 19, 225–249. https://doi.org/10.1080/0267257X.2003.9728208

Baruh, L., Popescu, M., 2017. "Big data analytics and the limits of privacy self-management". *New media & society* 19, 579–596.

Benoist, E., 2008. "Collecting data for the profiling of web users." in: *Profiling the European Citizen. Springer, pp. 169–184.*

Boyd, D., Crawford, K., 2011. "Six Provocations for Big Data". (SSRN Scholarly Paper No. ID 1926431). *Social Science Research Network, Rochester, NY.*

Broeders, D., Schrijvers, E., van der Sloot, B., van Brakel, R., de Hoog, J., Hirsch Ballin, E., 2017. "Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data". *Computer Law & Security Review* 33, 309–323. https://doi.org/10.1016/j.clsr.2017.03.002

Buraya, K., Farseev, A., Filchenkov, A., Chua, T.-S., 2017. "Towards User Personality Profiling from Multiple Social Networks". in: *AAAI.* pp. 4909–4910.

Calo, R., 2011. "The boundaries of privacy harm". *Ind. LJ* 86, 1131.

Chen, H., Chiang, R.H., Storey, V.C., 2012. "Business intelligence and analytics: From big data to big impact". *MIS quarterly* 36.

Chen, Y.-Y., Hsu, W.H., Liao, H.-Y.M., 2012. "Discovering Informative Social Subgraphs and Predicting Pairwise Relationships from Group Photos." in: *Proceedings of the 20th ACM International Conference on Multimedia, MM '12. ACM, New York, NY, USA,* pp. 669–678. https://doi.org/10.1145/2393347.2393439

Citron, D.K., Pasquale, F., 2014. "The scored society: due process for automated predictions". *Wash. L.* Rev. 89, 1.

Clarke, R., 1999. "Introduction to dataveillance and information privacy, and definitions of terms." Roger Clarke's Dataveillance and Information Privacy Pages.

Conley, B., 2018. "Thinking What He Says: Market Research and the Making of Donald Trump's 2016 Presidential Campaign." in: *Political Marketing in the 2016 U.S. Presidential Election, Palgrave Studies in Political Marketing and Management. Palgrave Macmillan*, Cham, pp. 29–48. https://doi.org/10.1007/978-3-319-59345-6_3

Crawford, K., Schultz, J., 2014. "Big data and due process: Toward a framework to redress predictive privacy harms." *BCL Rev*. 55, 93.

Custers, B., 2013. "Data dilemmas in the information society: Introduction and overview" in: *Discrimination and Privacy in the Information Society. Springer*, pp. 3–26.

D'Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y.-A., Bourka, A., 2015. "Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics". arXiv preprint arXiv:1512.06000.

Davidson, S., Binstock, R.H., 2011. "Political marketing and segmentation in aging democracies". in *Routledge handbook of political marketing* (pp.36-49). Routledge.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V.R., Tsou, A., Weingart, S., Sugimoto, C.R., 2015. "Big data, bigger dilemmas: A critical review." *Journal of the Association for Information Science and Technology* 66, 1523–1545. https://doi.org/10.1002/asi.23294

Farnadi, G., Tang, J., De Cock, M., Moens, M.-F., 2018. "User Profiling through Deep Multimodal Fusion" in: *Proceedings of the 11th ACM International Conference on Web Search and Data Mining. ACM.*

Farseev, A., Chua, T.-S., 2017. "TweetFit: Fusing Multiple Social Media and Sensor Data for Wellness Profile Learning.", in: *AAAI*. pp. 95–101.

Finn, R.L., Wright, D., Friedewald, M., 2013. "Seven types of privacy", in: *European Data Protection: Coming of Age. Springer*, pp. 3–32.

Fisher, L.E., 2004. Guilt by expressive association: "Political profiling, surveillance and the privacy of groups". *Ariz. L. Rev.* 46, 621.

Floridi, L., 2017. "Group privacy: a defence and an interpretation." in: *Group Privacy. Springer,* pp. 83–100.

Floridi, L., 2014. "Open data, data protection, and group privacy." *Philosophy & Technology* 27, 1–3.

Fuchs, C., 2011. New media, web 2.0 and surveillance. *Sociology compass* 5, 134–147.

Gadek, G., Pauchet, A., Malandain, N., Khelif, K., Vercouter, L., Brunessaux, S., 2017. "Topical cohesion of communities on Twitter." *Procedia Computer Science, Knowledge-Based and Intelligent Information & Engineering Systems*: *Proceedings of the 21st International Conference, KES-*20176-8 September 2017, Marseille, France 112, 584–593. https://doi.org/10.1016/j.procs.2017.08.171

Gasparetti, F., 2017. "Personalization and Context-awareness in Social Local Search: State-of-the-art and Future Research Challenges. Pervasive and Mobile Computing", *Special Issue IEEE International Conference on Pervasive Computing and Communications (PerCom)* 2016 38, 446–473. https://doi.org/10.1016/j.pmcj.2016.04.004

Goodman, E., Labo, S., Tambini, D., Moore, M., 2017. "The new political campaigning" [WWW Document]. URL http://blogs.lse.ac.uk/mediapolicyproject/policy-briefs/ (accessed 1.29.18).

Gutwirth, S., Leenes, R., De Hert, P., Poullet, Y., 2012. "European data protection: coming of age." *Springer Science & Business Media*.

Hasan, O., Habegger, B., Brunie, L., Bennani, N., Damiani, E., 2013. "A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case", in: *2013 IEEE International Congress on Big Data. Presented at the 2013 IEEE International Congress on Big Data,* pp. 25–30. https://doi.org/10.1109/BigData.Congress.2013.13

Hildebrandt, M., 2012. "The Dawn of a Critical Transparency Right for the Profiling Era". *Digital enlightenment yearbook 2012*, 41-56.

Hildebrandt, M., 2009. "Who is Profiling Who? Invisible Visibility", in: *Reinventing Data Protection? Springer, Dordrecht*, pp. 239–252. https://doi.org/10.1007/978-1-4020-9498-9_14

Hildebrandt, M., 2008a. "Profiling and the rule of law". *Identity in the Information Society* 1, 55–70.

Hildebrandt, M., 2008b. "Defining profiling: a new type of knowledge?", in: *Profiling the European Citizen. Springer*, pp. 17–45.

Hildebrandt, M., 2006. "Profiling: From data to knowledge." *DuD* 30, 548–552. https://doi.org/10.1007/s11623-006-0140-3

Hildebrandt, M., Gutwirth, S., 2008. *Profiling the European citizen. Springer.*

Jaquet-Chiffelle, D.O., 2008. "Reply: Direct and Indirect Profiling in the Light of Virtual Persons." *Profiling the European Citizen: Cross Disciplinary Perspectives. Springer* 55–63.

Kammourieh, L., Baar, T., Berens, J., Letouzé, E., Manske, J., Palmer, J., Sangokoya, D., Vinck, P., 2017. "Group Privacy in the Age of Big Data", in: *Group Privacy. Springer*, pp. 37–66.

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

Kosinski, M., Wang, Y., Lakkaraju, H., Leskovec, J., 2016. "Mining big data to extract patterns and predict real-life outcomes". *Psychological methods 2*1, 493.

Laney, D., 2001. "3D data management: Controlling data volume, velocity and variety". *META Group Research Note* 6.

Leese, M., 2014. "The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union". *Security Dialogue* 45, 494–511.

Mai, J.-E., 2016. "Big data privacy: The datafication of personal information". *The Information Society* 32, 192–199. https://doi.org/10.1080/01972243.2016.1153010

Mantelero, A., 2016. "Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection." *Computer law & security review* 32, 238–255.

Marx, G.T., 1998. "Ethics for the New Surveillance." *The Information Society* 14, 171–185. https://doi.org/10.1080/019722498128809

Mittelstadt, B., 2017. "From Individual to Group Privacy in Big Data Analytics". *Philos. Technol.* 30, 475–494. https://doi.org/10.1007/s13347-017-0253-7

Narayanan, A., Shmatikov, V., 2009. "De-anonymizing social networks", in: *Security and Privacy, 2009 30th IEEE Symposium On. IEEE*, pp. 173–187.

Nguyen, M.-H.C., Haynes, P., Maguire, S., Friedberg, J., 2013. "A user-centred approach to the data dilemma: Context, architecture, and policy". *Digital Enlightenment Yearbook 2013: The Value of Personal Data* 227.

Nickerson, D.W., Rogers, T., 2014. "Political Campaigns and Big Data". *The Journal of Economic Perspectives* 28, 51–73.

Pandit, H.J., Lewis, D., 2018. "Ease and Ethics of User Profiling in Black Mirror" in: *Companion Proceedings of the The Web Conference 2018, WWW '18. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland,* pp. 1577–1583. https://doi.org/10.1145/3184558.3191614

Provost, F., Fawcett, T., 2013. "Data Science and its Relationship to Big Data and Data-Driven Decision Making". *Big Data* 1, 51–59. https://doi.org/10.1089/big.2013.1508

Rubinstein, I.S., 2014. "Voter privacy in the age of big data". *Wis. L. Rev*. 861.

Russmann, U., 2016. "Voter Targeting Online in Comparative Perspectives: Political Party Websites in the 2008/2009 and 2013 Austrian and German Election Campaigns". *Journal of Political Marketing* 0, 1–24. https://doi.org/10.1080/15377857.2016.1179241

Schermer, B.W., 2011. "The limits of privacy in automated profiling and data mining". *Computer Law & Security Review* 27, 45–52. https://doi.org/10.1016/j.clsr.2010.11.009

Smith, H.J., Dinev, T., Xu, H., 2011. "Information privacy research: an interdisciplinary review". *MIS quarterly* 35, 989–1016.

Solove, D.J., 2005. "A Taxonomy of Privacy (SSRN Scholarly Paper No. ID 667622)". *Social Science Research Network, Rochester, NY*.

Solove, D.J., 2003. "Reconstructing electronic surveillance law". *Geo. Wash. L. Rev*. 72, 1264.

Spencer, S.B., 2015. "Privacy, Predictive Analytics, and Electronic Commerce Regulation" (SSRN Scholarly Paper No. ID 2678325). *Social Science Research Network, Rochester, NY*.

Storey, V.C., Song, I.-Y., 2017. "Big data technologies and Management: What conceptual modeling can do." *Data & Knowledge Engineering* 108, 50–67. https://doi.org/10.1016/j.datak.2017.01.001

Tavani, H.T., 2007. "Philosophical theories of privacy: Implications for an adequate online privacy policy." *Metaphilosophy* 38, 1–22.

Taylor, L., 2017. "Safety in numbers? Group privacy and big data analytics in the developing world, in: *Group Privacy. Springer,* pp. 13–36.

Taylor, L., Floridi, L., van der Sloot, B., 2016. "Group privacy: New challenges of data technologies. Springer".

Tene, O., Polonetsky, J., 2012. "Big data for all: Privacy and user control in the age of analytics". Nw. J. *Tech. & Intell. Prop.* 11, xxvii.

Vemou, K., Karyda, M., 2015. "Evaluating Privacy Practices in Web 2.0 Services", in: *MCIS.* p. 7.

Vicente, C.R., Freni, D., Bettini, C., Jensen, C.S., 2011. "Location-related privacy in geo-social networks". *IEEE Internet Computing* 15, 20–27.

Westin, A.F., 1967. Privacy and freedom, atheneum. New York 7.

Yannopoulos, A., Andronikou, V., Varvarigou, T., 2008. "Behavioural biometric profiling and ambient intelligence" in: *Profiling the European Citizen. Springer*, pp. 89–109.