http://dx.doi.org/10.14400/JDPM.2013.11.11.143

Big data, how to balance privacy and social values

Joo-Seong Hwang[*]
Graduate School of Public Policy and Information Technology, SEOULTECH[*]

빅데이터, 프라이버시와 사회적 가치의 조화방안

황주성
서울과학기술대학교 IT정책전문대학원[*]

Abstract Big data is expected to bring forth enormous public good as well as economic opportunity. However there is ongoing concern about privacy not only from public authorities but also from private enterprises. Big data is suspected to aggravate the existing privacy battle ground by introducing new types of privacy risks such as privacy risk of behavioral pattern. On the other hand, big data is asserted to become a new way to by-pass tradition behavioral tracking such as cookies, DPIs, finger printing… and etc. For it is not based on a targeted person. This paper is to find out if big data could contribute to catching out behavioral patterns of consumers without threatening or damaging their privacy. The difference between traditional behavioral tracking and big data analysis from the perspective of privacy will be discerned.

Key Words : big data, privacy, inferential analytics, big data commons, big data ecosystem

요 약 빅데이터는 막대한 경제적 기회뿐만 아니라 공적 가치를 낳을 것으로 예상된다. 하지만 공공기관은 물론 민간기업의 빅데이터 사용은 프라이버시 침해에 대한 우려를 지속적으로 제기하고 있다. 행위패턴의 프라이버시 등 기존에는 없었던 새로운 위험을 유발함으로써 빅데이터는 프라이버시에 대한 기존 논의의 틀을 와해시킬 우려가 크다는 것이다. 반면, 빅데이터는 쿠키 등 행위추적에 근거한 개인정보의 부작용을 불식시키는 대안으로 인식되기도 한다. 본 논문은 빅데이터가 행위정보를 기반으로 하는 개인정보와는 어떻게 다른지를 밝히는데 초점을 둔다. 나아가, 개인정보로부터 파행되는 기존의 프라이버시 문제를 해결하기 위해 빅데이터에 대한 정책이 어떠한 대안을 가질 수 있는지도 제시할 것이다.

주제어 : 빅데이터, 프라이버시, 추론적 분석, 빅데이터 생태계

1. Big data, a big problem to privacy?

Big data is based on advances in data mining, analytics, massive increase in computing power, and expansion of data storage capacity. In addition, the increasing number of people, devices, and sensors that are now connected by digital networks has

* 이 연구는 서울과학기술대학교 신임교수 연구정착비의 지원으로 수행되었습니다.

Received 4 October 2013, Revised 25 October 2013
Accepted 20 November 2013
Corresponding Author: Joo-Seong Hwang(SEOULTECH)
Email: jameshwang9@gmail.com

ISSN: 1738-1916

revolutionized the ability to generate, communicate, share, and access data. Big data means ′datasets whose size is beyond the ability of typical data base software tools to capture, store, manage, and analyze(McKinsey Global Institute, 2011)′. It is different from traditional data source in several features. First, it is often automatically generated by a machine such as sensor, search engines, and cameras without person′s involvement . Second, it comes from an entirely new sources like browsing behavior, car navigation and SNS postings. Third, it is atypical and non-friendly for analysis as in text streams. Fourth, it is not originally designed for valuable results, therefore most of big data are mixed with messy and ugly data(Franks, 2012, pp. 7-8).

Big data creates enormous value for innovation, productivity, efficiency, and growth. At the same time, the ‘data deluge’ presents privacy concerns that could stir a regulatory backlash. The inferential techniques being used on big data can offer great insight into many complicated issues such as security, safety, and consumer needs. Potential abuses of inferential data, however could imperil personal privacy, civil liberties and consumer freedoms (Bollier, 2010, p.2). In order to craft a balance between beneficial uses of data and the protection of individual privacy, policymakers must address some of the most fundamental concepts of privacy law, including the definition of ‘personally identifiable information’, the role of consent, and the principles of purpose limitation and data minimization (Tene & Polonetsky, 2012b, p.63).

Recent concerns on privacy harms caused by big data, however, seem to be too much exaggerated. They do not discern the increased privacy harms by big data itself, and those aggravated by big data techniques but originated from personal data. This poses a problem because big data is accredited to the very source of privacy disaster, though the key is still up to the personal data. To the contrary of common understanding, big data could be an alternative to harmonize the tension between social values of personal data and privacy risks. Big data is not always to be personal. It is rather based on and oriented to the non-personal, de-identified data. It needs to be clarified what the privacy harms and risks specifically added by big data is. And, how is it different from those privacy harms caused by personal data? It is not easy to separate those harms from big data and those from personal data. It may be not meaningful to do so because big data and personal data are too much intermingled to be divided. This paper is to look into what is the real influence of big data on privacy. In order to fully utilize the benefits of big data for security, safety and innovation, we should scrutinize which is the real problem, big data or personal data.

2. Where comes the real value of big data?

The term of big data was first used in science to refer to large data sets requiring the processing capacity of supercomputers (Boyd and Crawford 2011). Today, the term refers to the aggregation of massive stacks of data originating from different sources, produced by humans or machines. The advantage over processing different data sets separately is that it becomes possible to find correlations and infer additional information by aggregating, comparing, or otherwise analyzing data combined into one single large data set (De Filippi and Pordedda, 2012).

Values of big data can be easily recognized in such public fields as security, crime - prevention, healthcare, traffic management, insurance and employment. Big data analysis on a large medical dataset maintained by the FDA for more than 30 years, helps researchers to make a groundbreaking discovery. Russ Altman and his colleges found that Paxil and Pravachol have a unknown dreadful side effect when taken together. They could confirm their hunch by examining de-identified search engine logs. Those users who searched both Paxil and Pravachol together turned out to type in more words related to diabetes than those who searched for just one of them (Tene & Polonetsky, 2012a). The finding of Altman and his team is not the sole major medical breakthrough based on big data analysis. The discovery of Vioxx's adverse effects was possible owing to the dataset collected by Kaiser Performance, the California-based care consortium. Google flu trends, which predicts and locates outbreaks of the flu was able to do so by analyzing aggregated search queries.

Another case is found in traffic management. Using data from toll pricing or bus payment system, public authorities can decide where to construct new roads or mass-transit systems. Not to mention, individual drivers can also benefit from smart routing based on real-time traffic information. Big data also could play a role in predicting and preventing crime as experimented in Santa Cruz' PredPol. PredPol's system features a map of a city marked with red squares to show zones where crimes are likely to occur. It calculates its forecasts based on times and locations of previous crimes, combined with sociological information about criminal behavior and patterns.

From several cases mentioned above we can easily notice that big data does not always mean to be personal data. It is certain that most of them come from data generated by individual persons. But, this does not automatically mean that the data should be linked to a specific person or personally identifiable information (PII) such as name, social security number or address. The value of big data comes from finding out patterns of behavior from a huge amount of non-identified or anonymous data. The key virtues of the big data are in relevancy and sincerity, not in personality. For example, Dr. Alexander Dix, a privacy chief of Berlin, argues that whether big data needs to be personal data. In the research field, personal data are not necessary. In order to find out valuable information, personal data is not prerequisite. The power of big data is in what is done with that data, in how it is analyzed, and in what actions are taken based on the findings(Franks, 2012, p. 6). After finding out a relation, we need some personal data such as sex, age, or location to do an inferential recommendation. Even in this stage, however, the PII is not essential to achieve its value. When Amazon and Apple use their databases of consumer purchases to make recommendations to prospective customers, most people welcome the advice. It may help them identify just the book or music that they may want. On the other hand, people start getting very uneasy when buying suggestions are made based on how much the seller know about a particular person. This accrues to the realm of behavioral targeting (Bollier, 2010, p. 23).

Remaining issues are the reliability and accuracy of the inferential analytics. The value of big data analysis comes from its ability to find correlations and make inference on relations. There are severe debates if correlation is enough to suggest inferential decision making. Critics suggest that big data rely not on causation but on correlations. The newly discovered information, therefore is not only unintuitive and unpredictable, but also results from a fairly opaque process. Inference from big data may be thought as data mining on steroids (Rubinstein , 2012, p.3). The real problem is that big data enables authorities to make inferences that amount to "probabilistic cause." But this cannot replace "probable cause". Probabilistic cause remains a less reliable and more abstract predictive standard (Boiller, 2010, p. 34). Extrapolating from correlations can yield specious results even if large data sets are used. The classic example may be "My TiVO Thinks I'm Gay." The Wall Street Journal once described a TiVO customer who gradually came to realize that his TiVO recommendation system thought he was gay because it kept recommending gay-theme films.

However, others say that "correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. It's time to ask: What can science learn from Google?" (Boiller, 2010, p. 5). They strongly argue that probabilistic cause could be a legitimate tool for persuading a judge. The idea that you can't use probabilistic data to get a probable cause standard is suggested to be silly. If there is relevant data for relations for making a judgment that is important to society, the goal should not be to ban the use of correlations and data analysis. The goal should be to monitor it properly.

3. Challenges of big data to privacy

3.1 Privacy harms by big data

Many scholars as well as professional stakeholders have worried about additional privacy concerns caused by big data. This paper has reviewed several of them and summarized those suspected risks in Table 1. Though different terms are used by authors, major concerns are summarized into four items; re-identification, secondary use(consent, purpose, time limitation), new privacy reveal, increased accessibility.

〈Table 1〉 Privacy risks by big data

Boiller(2010)	Goldberg & Miller(2012)	Tene & Polonetsky (2012b)	Kuner et als. (2012)
-re-identification -secondary use	-re-idenfication -new privacy reveal -increased accessibility	-re-identification -inferential prediction -auto-decisionmaking -access & exclusion -chilling effect	-collection limitation -consent -right to to be forgotten -surveillance

First, re-identification is the new technology that disrupts the whole privacy policy landscape by undermining the faith of anonymization(Ohm, 2010). Anonymization is the main tool allowing data collectors to keep personal data in storage or hand them over to a third party without compromising the privacy of the people tracked. Over the past few years, however, computer scientists have repeatedly shown that even anonymized data can often be re-identified and attributed to specific individuals. De-identification could be possible because of big data analytics. Big data offers greater opportunities for re-identifying the data, linking a given set of medical information to a specific person (Bollier, 2010, p.31). Owing to big data anonymization can no longer remain effective in protecting users against tracking and profiling. (Rubinstein, 2012, p.5)

Second, inhibition of secondary use clash with the very premise of big data insisting that more is better (Bollier, 2010 p.36). Secondary use is to use information originally collected for one purpose in order to fulfill a different purpose without consent of the data subject. This has been strongly prohibited in various privacy principles and laws such as fair information practices of U.S. and DPR of EU. This principle has also been known as the purpose specification principle. And it is strongly coupled with notice and consent principle and data minimization principle. The very reason d'etre of looking to big data is to identify patterns that create answers to questions we didn't even know to ask. So limiting data-collection in the first place could undercut the potential benefits that big data might deliver (Bollier, 2010, p.36). In the environment of cloud computing, the use of big data for secondary processing and profiling is amongst the most threatening for privacy and data protection (De Filippi & Pordedda, 2012).

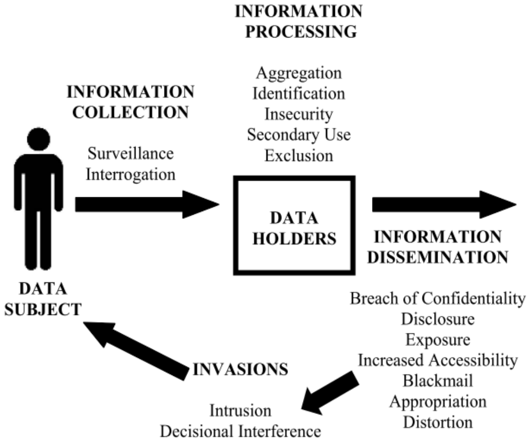
Third, analysis of large data sets sometimes reveals new information that is not just a summation of the individual underlying information (Goldberger & Miller, 2012). Big data systems make it relatively easy for companies and governments aggregate any number of data sets from any number of data markets, which, in turn, makes it that much easier to derive private information. The added value of big data is obtained by aggregating different types of data extracted from different sources, connecting them together with other pieces of data about the same users, different users, users they are in connection with, or the whole community of users to which they belong (De Filippi & Pordedda, 2012). The more data you are able to collect and connect to other data sets, the easier it is to obtain what was thought to be private information and then tie that information to a specific individual (Craig & Ludloff, 2012, p.56). The case of Target is one of the best examples of new privacy revealed by big data. Target has collected and built a history dataset of buying records together with demographic information. Mining the data, Target was able to identify about 25 products that allow the assign each shopper a 'pregnancy prediction score'. Take a fictional Target shopper named Jenny Ward, who is 23. She lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There is say, an 87 percent chance that she is pregnant and that her delivery date is sometime in late August (Forbes, 2012.2.16). Does the girl in Target case argue for privacy rights in the new information obtained by analysis? Maynard case even in a slightly different situation told that she could. According to the judge, one can have a reasonable expectation of privacy in the sum of his/her behaviors even though he had no expectation of privacy in his/her individual behaviors in public space (Goldberger & Miller, 2012).

Fourth, increased accessibility also impedes the risk of privacy disclosure. When publicly available information becomes more easily accessible as a result of big data, do the data providers retain privacy rights allowing them control over the continued use of the data? If there is a risk of increased harm by the republication, the answer is likely to be yes. In Ostergren v. Cuccinelli case, the court verdict that an individual's interest in controlling the dissemination of information regarding personal matters does not dissolve simply because that information may be available to the public in some forms. Because there is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country, and a computerized summary located in a single clearinghouse of information (Goldberger & Miller, 2012). Even though the data is publicly accessible, that does not mean it is allowed to be widely distributed (Boyd, 2012)

We can accept privacy risks mentioned above as the increased harms caused by big data. But, it is not essentially created by big data, but instead by personal data. Take the Target case back to this issue. Privacy braking happened when Target attached the pregnancy prediction score to a specific person with their personally identifiable data such as guest ID, credit card, name and address. If they did not violate the privacy principles such as consent, secondary use, purpose limitation, there would not have been such a big fuss. What if they just send similar targeted advertisements to everybody over a certain amount of score like the Amazon's book recommendations? The essence of this incident lies in the non-proper use of personal data rather than inferential advertisements enabled by big data analysis.

3.2 Direct and indirect privacy harms by big data

Solove proposed a list of privacy taxonomy in an attempt to identify privacy problems in a comprehensive and concrete manner (Solove, 2006). His goal was simply to define the activities which cause privacy harms, and explain why and how they can cause trouble. His taxonomy is worth introducing here as a framework to discern what is the real privacy harms which is purely exacerbated by big data. There are four basic groups of harmful activities: (1) information collection, (2) information processing, (3) information dissemination, and (4) invasion. Each of these groups consists of different related subgroups of harmful activities (Fig 1).



[Fig. 1] Taxonomy of privacy by Solove(2006)

Big data as an environment should be responsible for all the increased privacy harms directly or indirectly incurred. But, when you look into the process and mechanism how big data is causing new or expanded privacy harms, we can easily notice most of them occur in close link with personal data(PII: personally identifiable data). Therefore, we need to divide the privacy harms of big data into two categories. The first is those harms aggravated by big data but originated from personal data (indirect harms), and the latter is those caused directly by big data analysis (direct harms). The argument here is that only the latter harms which are related to inferential analytics should be accrued to big data. The first is an issue originally related to personal data management. I am not contending that big data is not a threat to personal data. Its impact is quite huge and detrimental because it fundamentally transforms privacy landscape regulated by existing norms and principles. It has shaken the core principles of personal data. In this regards, big data should be in charge of all the four increased privacy risks listed above. But, it is not the big data analysis itself to be fixed, rather personal data.

Borrowing Solove's taxonomy on privacy harms, the indirect privacy harms aggravated through personal data is quite wide-spread (see Table 2.).

〈Table 2〉 Indirect and direct privacy harms by big data

Taxonomy	Privacy Harm	indirect	direct
Information Collection			
- surveillance	- chilling effect, self-censorship & inhibition	√	
- interrogation	- against self-incrimination	√	√
Information Processing			
- aggregation	- unexpected privacy risk, decision against oneself	√	√
- identification	- information baggage, inhibit ability to change, fear of reprisal	√	
- insecurity	- potential future harm		
- secondary use	- not expected nor desired, misunderstood	√	
- exclusion	- vulnerability, frustration		
Information Dissemination			
- breach of confidentiality	- betrayal of trust		
- disclosure	- reputation demage, inhibit autonomy & anonymity		
- exposure	- loss of self esteem		
- increased accessibility	- more risk of disclosure		
- blackmail	- controlled by someone		
- appropriation	- affront of dignity, property		
- distortion	- defamation, false reveal		
Invasion			
- intrusion	- solitude,	√	√
- decisional interference	- information disclosure, harm to autonomy		

Most of them are related with four privacy risks mentioned in the previous section. The direct harms however are more subtle and limited. It can increase some risks about interrogation, aggregation and decisional interference. As for interrogation, every person has a privilege against self-incrimination. He or she has a right not to be forced to testify against themselves. Big data analysis using SNS postings against one-self could be new privacy harm to self-incrimination. New privacy information from aggregation is outstanding harms from big data as mentioned earlier. Decisional interference is governmental interference with people's decisions regarding certain matters of their lives such as getting pregnancy. Big data related with patient's healthcare could be a concerning source to this kinds of new risks.

Pattern privacy based on group profiling could induce severe direct harms such as inaccurate grouping and discrimination. While it allows service providers to recommend a more desirable product with high customization and personalization, categorizing users as a result of inferred information also allows providers to discriminate amongst customers according to the category in which they have fallen (De Filippi & Pordedda, 2012). Newly discovered information derived from personal data can be unreliable and inaccurate, especially when that data has been anonymized or generalized by being transformed into group profiles. Group profiles are traits that apply to individuals as members of a reference group, even though a given individual may not actually exhibit the property in question.(Rubinstein, 2012, p.5). This often cause problems such as sending irrelevant targeted ads to miss-grouped people. Inferential predictions, however, does not automatically impede privacy.

4. How to deal with personal data in the big data environment

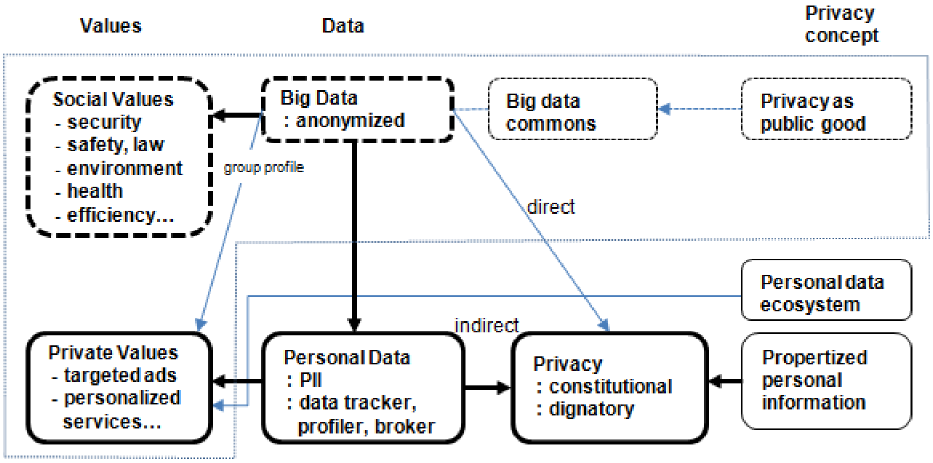
With all the direct and indirect risks (or harms) of big data, it can also lead us to replacing personal data with anonymized big data. This means big data could lessen the dependency on personal data for targeted advertizing. With all the risk of re-identification, anonymization (de-identification) is an excellent approach. It is still an effective way to reconcile the public benefit while minimizing private harm (Boiller, 2010, p.31). Some scholar insists that no useful database can ever be perfectly anonymous, and as the utility of data increases, the privacy decreases. However, we should not deny the efficacy of anonymization, and question the effectiveness of personal vs. non-personal information. One possible conclusion, apparently supported by Ohm, is that all data should be treated as personally identifiable and subjected to the regulatory framework. Yet such a result would create perverse incentives for organizations to abandon de-identification and therefore increase, rather than alleviate, privacy and data security risks. A further pitfall is that with a vastly expanded definition of personally identifiable information, the privacy and data protection framework would become all but unworkable. The current framework, which is difficult enough to comply with and enforce in its existing scope, may well become unmanageable if it extends to any piece of information (Tene & Polonetsky, 2012b, p.66).

The separation of direct privacy harms of big data from indirect ones has a great implication to policy directions on big data as well as personal data. It is not a panacea to privacy to suppress the utilization of big data. Instead, we may need to build a data ecosystem where various anonymized data commons are collected and shared for public as well as private purposes. This might be the most feasible and possible way to harmonize tensions between personal data and privacy in the big data environment. In this regards, alternative policy directions on personal data should be reviewed and evaluated. Policy directions suggested hitherto can be boiled down to three categories – privacy concept as public good, remedy on existing scheme of propertized personal information, and personal data ecosystem.

Privacy concept as public good comes from constitutional privacy, meaning that privacy is constitutive of society (Solove, 2006). Constitutive privacy understands privacy harms as extending beyond the “mental pain and distress” caused to particular individuals; privacy harms affect the nature of society and impede individual activities that contribute to the greater social good. Privacy considerations no longer arise out of particular individual problems; rather, they express conflicts affecting everyone. More structural problems, called as architectural problems by Solove belong to this scope. They involve the creation of the risk that a person might be harmed in the future. They are akin, in many ways, to environmental harms or pollution. In the taxonomy, two kinds of architectural issues emerge most often. First is the enhancement of the risk that a harm will occur. Activities involving a person's information, for example, might create a greater risk of that person being victimized by identity theft or fraud and suffering dignitary harms as well as monetary or physical harms. Second, a particular activity can upset the balance of social or institutional power in undesirable ways. A particular individual may not be harmed directly, but this balance of power can affect that person's life. People's behavior might be chilled, making them less likely to attend political rallies or criticize popular views. Surveillance can also have these effects. This kind of harm is often referred to as a "chilling effect". Privacy as public good means that it is better to keep it as stringent as possible. Proponents of this approach call for limits on the propertization of personal information. Schwartz sees information privacy as a public good like clean air or national defense (Schwartz, 2004). He refers to the public good at stake as 'the privacy commons—a space for anonymous and semi-anonymous interactions' (Rubinstein, 2012, p.12). Privacy commons resemble creative commons in that individual can decide the position to his own privacy.

Remedy on existing scheme is the perspective of EU's General Data Protection Regulation proposed in December 2012. In developing this new set of data protection rules, the Commission of EU was well aware of the 'dramatic technological changes' that have occurred since the DPD was first proposed in 1996. And also she was very concerned with problems raised by profiling and data mining. Despite this realization, the Commission held firm in its belief that 'the current framework remains sound as far as its objectives and principles are concerned'. They are still firmed based on the propertized personal information. The ownership and control are up to the collectors. While the Regulation introduces several new privacy rights – notably the right to be forgotten and data portability, the other changes it makes are incremental at best. For example, it proposes stricter transparency obligations and a tighter definition of consent. It strengthens the existing provision concerning 'automated individual decisions' by including a new provision on profiling, but the changes are limited and focus mainly on enhancing transparency. It also imposes new responsibilities on data controllers including data protection by design and default (Rubinstein, 2012, p.12). This is almost in same vein as 'privacy by design' prospoed by Caboukian (De Filippi & Porcedda, 2012). The seven principles porposed by Cavoukian are: a proactive or preventative approach; privacy by default; privacy embedded in design; positive sum game; end-to-end security; respect for users; visibility and transparency.

The last approach, personal data ecosystem is a kind of 'sharing the wealth' strategy premised on data controllers providing individuals with access to their data in a usable format and allowing them 'to take advantage of applications to analyze their own data and draw useful conclusions from it'. New business model should give individuals the capacity to benefit from big data and hence the motivation to learn about and control how their data are collected and used(Rubinstein, 2012, p.12). The pivot point of this ecosystem is the concept of 'user-centricity', which seeks to integrate diverse types of personal data while putting end users at the centre of data collection and use, subject to a set of global data principles that include transparency, trust, control, and value creation.(Rubinstein, 2012, p.9). Personal data services (PDSes) will be actual services provided by personal data 'guardians' who owe fiduciary duties to their individual. These agencies will provide both a secure data store for a wide variety of personal information(including official records like birth and



[Fig. 2] Social values and privacy concept of big data

marriage certificates, licences, and passports, transaction records, online profiles, and social media content, and user names and passwords) as well as a new class of user-driven services ranging from personal RFPs to more participatory forms of healthcare, to ‘FixMyStreet’ and similar grass-roots citizenship efforts.

Comparing these three alternative policy approaches, we can conjecture each will have different remifications on the big data policy([Fig. 2]).

The first approach (privacy as public good) would turn the dependency on the personal information to big data. Privacy is considered to be a public good to be protected from private interests and propertization. Instead, group profiles from anonymous big data commons will fulfill not only private needs such as targeted advertizing or personalized services but also social and public values such as national security, safety, environment ... etcs. This approach will support a favorable formation of anonymous big data commons. Big data commons means 'a space in which actors across sectors develop ways and rules for the safe and beneficial management and use of anonymized personal data'(WEF, 2013, p.1). This commons require market environment such as incentives to contribute anonymized data to a data commons. It also needs infrastructure like public and open data as well as regulations such as transparent rules on data sharing and privacy. The second approach of GDPR is still lingering around the existing scheme of the propertized personal information. Big data is recognized to be a great threat to the data protection principles. The most important issues, therefore is to fix the existing principles in response to the direct and indirect challenges of big data environment. In this vein, GDPR has paved a road toward personal data ecosystem with the adoption of data portability. The right to be forgotten preserves the way out of the ecosystem. There does not seem to be much room for big data commons. Because big data is taken for a strong technical tool to support the new derivation and accumulation of propertized personal data. The last approach(personal data ecosystem) has the concept of privacy as private good in common with the existing scheme. Difference is that right for the personal information is up to individuals. In this scheme, personal data will not be as much monetized as in the existing scheme. Instead its main values will be in promoting personalized services. Every personal data will belong to a personal domain, and there might be a small room for big data analysis for public purposes.

From the perspective of privacy as private good, big data looks like a big threat to privacy as well as personal data. Big data, however, could provide a new alternative to the existing regime of personal data. The original values of big data come from finding out probablistic relations from inferential analytics. It actually does not need personally identifiable data to get to the valuable results. This means that big data could lead to 'personal data - free' environment where big data and privacy could coexist.

5. Conclusion

This paper aims to find out how to balance privacy and social values in big data environment. It is true that big data could be a big threat to the existing personal data regime. Its direct harms on privacy, however should be seperated from those indirect harms aggravated through personal data. Rather, big data could lessen the dependency on personal data for targeted advertizing as well as social purposes. The author argues that big data should be recognized as a new opportunity for 'big data commons'. This will lessen privacy problems from personal data collected and propertized by private companies. We should discern the problems and harms caused by big data from those by personal data. Policy efforts should focus on how to build a sustainable big data commons and big data ecosystem that are free from privacy harms. Further research should be done about how to classify big data from its contributions to social and public values. Like interstate highways and industrial complexes in the industrial age, some kind of big data such as search queries could be considered to be public infra with social consensus.

ACKNOWLEDGMENTS

This study was financially supported by Seoul National University of Science & Technology as to a newly appointed professor.

REFERENCES

[1] Boiller, D., 2010, ***The Promise and Peril of Big Data***, Washing D.C. : The Aspen Institute.

[2] Boyd, D. and Crawford, K., 2011, Six Provocations for Big Data. Paper presented at the Oxford Internet Institute's "***A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society***" (September 21, 2011).

[3] Boyd, D., 2010, Privacy and Publicity in the Context of Big Data, ***WWW***, Raleigh, North Carolina, April 29.

[4] Chester, J., 2012, Cookie wars: How new data profiling and targeting techniques threaten citizens and consumers in the "Big Data" era, In Gutwirth et al. (eds.), ***European Data Protection: In Good Health?***, B.V: Springer Science + Business Media

[5] De Filippi, P. and Porcedda, M.G., 2012, Privacy Belts on the Innovation Highway

[6] Forbes, 2012. 2. 16, How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

[7] Franks, B., 2012, ***Taming the Big Data Tidal Wave: finding opportunities in huge data streams with advanced analytics***, New Jersey: John Wiley & Sons

[8] Goldberg, N. and Miller, M., 2011, The practice of law in the age of 'Big Data'

[9] Kuner, C. et als, 2012, The challenge of 'big data' for data protection, ***International Data Privacy Law***, 2012, Vol. 2, No. 2 ,47-49

[10] Mckinsey Global Institute, 2011, ***Big Data: The Nest Frontier for Innovation, Competition, and Productivity***

[11] Ohm, P., 2010, Broken promises of privacy:

responding to the surprising failure of anonymization, *CLA LAW REVIEW* 57, 1701-1776

[12] Rubinstein, 2012, Big data: the end of privacy or a new beginning?, Public law & legal theory research paper series, working paper No. 12-56

[13] Solove, 2006, A Taxanomy of privacy, *University of Pennsylvania Law Review*, vol. 154, no. 3. 477~560.

[14] Schwartz, P., 2004, ‘Property, Privacy, and Personal Data’, *Harvard Law Review*, vol. 117, no. 7. 2055~2128.

[15] Tene, O. and Polonetsky, J., 2012a, Privacy in the age of big data: a time for big decision, *Stanford Law Review Online*, vol.64, 63~69.

[16] Tene, O. and Polonetsky, J., 2012b, Big data for all: privacy and user control in the age of analytics, *Northwestern Journal of Technology and Intellectual Property*, Available at http://ssrn.com/abstract=2149364

황 주 성(Joo-Seong Hwang)

- 1985년 2월 : 서울대 사회과학대학 지리학과(문학사)
- 1987년 2월 : 서울대 대학원 지리학과(문학석사)
- 1996년 2월 : 서울대 대학원 지리학과(문학박사)
- 1989년 7월 ~ 2012년 2월 : 정보통신정책연구원 연구위원
- 2012년 3월 ~ 현재 : 서울과학기술대학교 IT정책전문대학원 부교수
- 관심분야 : IT정책, 방통융합, 인터넷 이용행태
- E-Mail : jameshwang9@gmail.com