Privacy Protection or Data Value: Can We Have Both?

Ken Barker(✉)

University of Calgary, Calgary, AB, Canada
kbarker@ucalgary.ca

Abstract. Efforts to derive maximum value from data have led to an expectation that this is "just the cost of living in the modern world." Ultimately this form of data exploitation will not be sustainable either due to customer dissatisfaction or government intervention to ensure private information is treated with the same level of protection that we currently find in paper-based systems. Legal, technical, and moral boundaries need to be placed on how personal information is used and how it can be combined to create inferences that are often highly accurate but not guaranteed to be correct. Agrawal's initial call-to-arms in 2002 has generated a large volume of work but the analytics and privacy communities are not truly communicating with the goal of providing high utility from the data collected but in such a way that it does not violate the intended purpose for which it was initially collected [2]. This paper describes the current state of the art and makes a call to open a true dialog between these two communities. Ultimately, this may be the only way current analytics will be allowed to continue without severe government intervention and/or without severe actions on behalf of the people from whom the data is being collected and analyzed by either refusing to work with exploitative corporations or litigation to address the harms arising from the current practices.

1 Introduction

Users provide information about themselves to either receive a value of some kind or because they are compelled to do so for legal or moral reasons. For example, a patient needing health care due to an illness or injury must disclose to their physician personal information that is considered in almost all cultures to be highly private so they can receive the health care they need. The reason the information is being disclosed is to receive the medical care they need to maintain or regain their health. There is an expectation, and indeed a high ethical standard (Hippocratic Oath: Article 8) that compels this information be kept private and not used for any other purpose.[1] This example is often cited in the literature

[1] There are societally and individually acceptable deviations from this high standard such as informing an insurance company about the costs of the service so the physician can be paid. However, this is done with the explicit informed consent of the patient and there is an expectation that this information will be kept private and will not be used for any other purpose.

N. Kumar and V. Bhatnagar (Eds.): BDA 2015, LNCS 9498, pp. 3–20, 2015.
DOI: 10.1007/978-3-319-27057-9_1

to motivate the need for privacy and the expectations implied here are often applied to other interactions that occur. However, there are interactions that occur where there is no expectation of privacy or that a "private" communication will be kept confidential. For example, a person interviewed by a press reporter does not have the same expectation that the communication will be protected. In fact, the exact opposite is the expectation and both parties entering into such a communication understand that the reason for the conversation is to capture the discussion with the purpose of writing a very public story. Analytic advocates argue that data collection is often done with both parties understanding that the information collected will be used for many forms of analysis well beyond its apparent initial purpose. For example, search query data submitted to a search engine is not considered "private" by the search engine provider and in many cases, the collection of such data is the basis of their business model. However, police investigations now regularly include searches of suspects browser search histories for evidence that might link the suspect to the crime. It is a fairly obvious claim that this was not the intention of the suspect and as a result will likely feel privacy is being violated.

A commonly held myth is that privacy cannot be protected; that people no longer care about their privacy because any interactions in the modern world, by definition, requires that we forfeit our privacy; and that we should simply accept it and live without concern about it only because we gave up our right to privacy a long time ago. Clearly this myth is easily revealed as "false" by virtue of the large number of things each individual does in their daily lives to protect their privacy. Very few people would consider cameras in the bathroom a reasonable privacy of violation even if there might be substantial public good that might accrue from such cameras. The ubiquity of public survivance cameras does not imply the public accepts being monitored everywhere so there are societal limits that restrict the extent to which cameras can be used. Why should there these limits not also be placed on other forms of survivance?

Analytics, is effectively survivance. Click stream data describes how a user interacts on a website and some would argue it reveals their intent. Mining of such data allows an analyst to predict a particular users interests and to tailor the way information is presented to the user. This tailoring, it is argued, is a "user value" because it allows the webpage to provides "opportunities of *specific* interest" to the user. The argument continues that this is a "win-win-win" because the webpage vendor can provide a marketing opportunity to third-parties by promising to deliver advertising content to users with the maximum likelihood of purchasing their products. This, in effect, creates a "third" winner and this the third-party product provider who is able to more effectively sell their product or service.

Unfortunately there are serious issues associated with this particular three way "win". First, even if we accept that analytics for the purpose of "adding value to the customer's experience" is an acceptable tradeoff, the data collected may be used in any number of unanticipated ways. A person who is interested in understanding how a disease impacts a friend could be denied insurance coverage if there was a large number of searches undertaken about a particular disease,

even if the person does not have such a disease. A teenage person searching for information about date rape drugs so they can better understand how a friend may have been subjected to such a criminal act may be brought under suspicion as a result of the police investigation into who might have committed the crime.

Second, the analytics will combine data in ways that will produce inferences that are inaccurate. Data mining will fail to produce useful inferences if accuracy is considered sacrosanct. The *accuracy paradox* [9] demonstrate that accuracy can even improve when known errors are introduced into the system. This has led researchers to develop terms and definitions that are more appropriate for analytical work such as *precision* and *recall* when evaluating the effectiveness of their algorithms. This is not problematic when data is considered in aggregate but introduces significant issues if the analytics purports to identify instances that may introduce inaccurate descriptions of individuals. The primary goal of classification systems is to place individuals within each class so they can be considered largely homogeneous. When an instance is misclassified the potential harm is substantial and it is unlikely that anyone would agree to a privacy policy statement that explicitly allowed for such errors.

Third, the analytics will lead to unintended consequences. The famous case of Target^{TM} identifying a pregnant teenager and "helpfully" providing product information is often touted as how precise and successful data analytics is [5]. However, it does not address the harm that comes from such "helpful" marketing campaigns. Although the father in this particular situation was quite polite in his response when he says, "[I]t turns out there's been some activities in my house I haven't been complete aware of," is exceptionally muted, it is clear that not all parents would react quite so calmly. If the father's trust in such analytics was so great that he assumed Target^{TM}'s analysis was indeed correct but that daughter had chosen to seek an abortion without her parent's consent, the consequences could have been even more significant. Target argues that this is providing value for their customer and admits that the goal is to capture a high value market segment but acknowledges that not everyone receives such "helpful" information in the same way [5] so they have taken steps to disguise how specific this one-to-one marketing is for individuals.[2]

2 Big Data

Big Data is a term that has been used since the inception of modern computer systems and in particular since the development of database management systems (DBMS). Unfortunately, it is a term that can only be defined relative to something and even when care is taken in its definition, gaps readily form that lead to disagreements. Absolute size does not work as demonstrated by the date of the inaugural *Very Large Data Base* conference that occurred in 1975 to

[2] One-to-One marketing "... means being willing and able to change your behaviour toward an individual customer based on what the customer tells you and what else you know about that customer." [8].

help deal with "large" databases which would be consider miniscule by today's systems. Gudiveda *et al.* [6] define big data as "data too large and complex to capture, process and analyze using current computing infrastructure" and then point to the popular characterization using five "V"s, which may not be applicable in all big data environments. The "V"s are:

- Volume - currently defined as terabytes (2^{40}) or even as large as exabytes (2^{60}).
- Velocity - data is produced at extremely high rates so streaming processes are required to help limit the volume.
- Variety - often data captures are very heterogeneous and may include structured, unstructured or semi-structured elements.
- Veracity - provenance is often crucial in determining if the source is valid, trustworthy, and of high quality.
- Value - the data should be of value either in itself or via subsequent analytics.

Each of these characteristics has a different impact on how big data is managed but since our focus here is on privacy, we consider briefly how each impacts on analytics and the nature of the data that is privacy sensitive. Volume, in and of itself, is not necessarily a privacy sensitive characteristic. For example, large volumes of radio telescope data does not contain data that is private. The challenges here are found in stream processing through sparse data which is not relevant to this paper. The amount of data collected is largely orthogonal to privacy issues but as the volume increases relative to the available compute power, there is a risk that privacy (or for that matter security) considerations will be displaced by the need for efficiency. Thus, volume introduces a risk but is not in itself a privacy threat.

Similarly the velocity with which data is received is not inherently problematic for privacy. The challenge becomes one of ensuring that any privacy requirements for the quickly arriving data is managed in a timely way. Much like volume, the strong temptation would be to sacrifice proper meta-data management (including privacy annotations) in the interest of simply collecting the data. Later we argue that data must be appropriately annotated with privacy provenance and if the collection mechanisms are too slow, this critical foundational step could be compromised.

A key aspect of big data is variety. The analytics that can be undertaken when large volumes of heterogeneous data is collected provides both an opportunities for tremendous insight and to invade privacy. Research into how to integrate and validate heterogeneous data sources has been underway for at least three decades. The new issues introduced by privacy meta-data that will likely be in conflict because they come from different sources, in different formats, will provide a rich set of research opportunities.

One of the critical features of any big data project is ensuring the veracity of the data collected. This will be critical in privacy as one of the key privacy obligations is to ensure that data collected can be checked for correctness. In addition, any data found to be incorrect must be deleted and possibly replaced with correct values. Thus, the veracity characteristic will be challenging because

the data must be validated upon collection, maintained correctly throughout its lifecyle and deleted (really deleted) at the end of its life [7].

Data analytics is fundamentally motivated by the desire to derive value from data. In privacy research, this is often expressed as the privacy-utility tradeoff. Privacy is trivially guaranteed by not providing any access to the data collected, but this sacrifices all utility. Conversely, utility is maximized by eliminating any data access but this provides no data privacy at all. Thus, the definition of "value" in big data needs to be tempered by the responsibility to facilitate appropriate privacy for the data provider.

3 Privacy

Data ownership is likely a central issue in how people think about data privacy. If you, as an individual, believe that information about yourself is ultimately your "property" and should be controlled and disseminated by you, your perspective about privacy will likely hold that the individual owns that data and should have the right to determine how, or even if, it is used by others. If you, possibly thinking as a corporation, believe that information about those with whom you interact belongs to the corporation, then you will view the artifacts arising from interactions with individuals as your property. Some would argue that these artifacts are "shared" property so both parties have the right to claim ownership but this position is difficult to realize pragmatically when considering privacy. In other words, once a piece of private data is released, it cannot be easily retracted and protected from further disclosure.[3] This suggests that there are at least three distinct schools of thought when it comes to data ownership and we consider each below.

3.1 Individuals Own Data About Themselves

Medical data, religious beliefs, sexual orientation and political viewpoints are often considered private by individuals and most jurisdictions have laws that prevent, for example, an employer asking questions about these beliefs or viewpoints. However, if a person seeks treatment for a particular disease, they must reveal detailed medical histories to ensure their physician will treat them appropriately. This highly personal data is often believed to be owned by the patient and they have the right to control how it is used and to who it might be disseminated. However, hospitals collect this data and often in the interest of the greater good argue that this data, appropriately anonymized, should be used for medical research. The Hippocratic Oath (Article 8) requires that this not be done without explicit permission so patients are often asked to sign documents

[3] This is often the argument made by those who believe that "nothing is private" any longer and we should just accept this as a reality. However, the argument is self-evidently specious since the argument's proponents are often quite protective of some aspects of themselves as discussed earlier.

that allow their data to be used for this purpose.[4] Thus, the hospital is acknowledging that ownership of the data resides with the patient and the research to be done is with explicit permission from the patient (i.e. data owner).

3.2 Corporations Own Data Collected

Users interact with websites for many reasons but their primary purpose is the acquisition of some knowledge or service. The way users interact is often a high value commodity that the corporation can use for analytics for its own operation or to resell this behavioural information to other interested parties. For example, a search engine may collect the queries posed by users to help identify what is trending or of interest to its user community. The value may be substantial for organizations that want to communicate with the public about things that may be of interest. A political party that makes a policy statement would likely want to see if their announcement is generating interest and if it is being received favourably. This kind of data is often considered the property of the website and as such can be used to derive benefits for themselves. The corporation will attempt to protect the individual users by anonymizing the data through aggregation or using some other technique but it strongly believes this data is owned by them and it is likely that most people would agree with their claim.

An online shopping site will also collect similar search queries and use it in the same way as a search engine. However, the shopping site may have an opportunity to collect additional data in the event that a user makes a purchase. The user must enter data relevant to the purchase itself including name, address, contact information, credit card information and other details directly applicable the purchase. The shopping site will use this information in the first instance for the purpose of selling and delivering the purchased product but is now able to link very specific identifying information to the individual doing the search. Often these website will also install, often with permission, an artifact on the users computer so future trips to the site will also be readily tracked. The shopping site would claim that this data is owned by them. Few people would likely agree that information such as name, address, and credit card data is now "owned" by the website but this is the current state of the art and there is not typically a need for the company to delete the information after its initial intended use.

3.3 Shared Ownership of Data

One sharing mechanism suggested is that the data remains the property of the data owner but because they have willing shared it with the corporation, they

[4] It is unclear if this permission is collected in a completely non-coercive way. A patient might feel that by failing to sign such a document they may not receive the best possible care. Clearly this would not be the case but the perception may be a critical factor in providing such permission and this would likely be considered coercive in some way by a reasonable person. However, and much more likely, the patient is simply overwhelmed with the amount of documents required as they seek treatment so they may simply sign the documents presented to them with due consideration as they seek care.

have become co-owners of this data. The argument is initially compelling because the corporation should have the right to know who their customers are and the user clearly does not want to pass exclusive ownership of their data to anyone else. To understand the implications it is important to note that the purpose the user provided the information was to complete a sale. If the corporation fully agreed with this as the sole purpose for the collection of the data, it would willingly delete the data once the sale was complete and not claim any ownership. Unfortunately, the value of the interaction is not just in the sale but also in the information derived from the consumers behaviour during the purchasing process and afterward as related items can be suggested to that individual in a highly targeted way.

Perhaps there is another way to "share" the ownership of this data without violating the purpose for which it was provided. Some aspects of the interaction could reasonably be seen as the property of the corporation in the same way the search engine from Sect. 3.2 owned the search data. Clearly this would have to be unlinked from the more specific information that was collected for the sale itself but if this is done the corporation would have a strong claim to owning the data collected. Conversely, the demographic information would be deleted (or at least fully unlinked from the search data) so the ownership of the search data could be held by the corporation. This would address the ownership issue because each party would retain the portion of the data that is uniquely theirs.

3.4 Summary – Ownership

The issue of ownership is ultimately an extremely divisive one and may need to be considered in light of the specific situation. However, the general principles that seem to apply are:

Personal Data, of any kind, remains the property of the original owner and cannot be transferred to another without explicit, easily understood, informed consent.

Behavioural Data, such as click streams and search queries, are owned by the corporation but must not be linked to the individual either explicitly (as in Sect. 3.2) or through analysis.

Our earlier work has argued that ownership is not the key issue but rather who should have control over an individual's private data, for what purpose, and who should have access to it. This immediately lead to the question of how this control can be enforced in an environment where organizations have a vested interest in maximizing the value of any data they are able to collect. "[F]or knowledge itself is a power"[5] [3] is only true if that knowledge can be used. This raises the question of what principles should be applied when determining how the data is to be used.

[5] This quote is often paraphrased as "knowledge is power" but Sir Francis Bacon is actually speaking of the limits of God and that "knowledge" is in fact only a part of God's power. It also suggests that it must be weighed against other aspects of power including, in this case, Godly judgement.

4 Privacy Principles

Building on the core principles of U.S. Privacy Act (1974) [1], which articulated six principles that (1) ensured an individual can determine what data is collected; (2) guarantees that data collected is only used for the purpose for which it is collect; (3) provides access to your own data; (4) and information is current, (5) correct and (6) only used for legal purposes, the OECD developed a list of eight principles including: limited collection, data quality assurance, purpose specification, limited use, security safeguards, openness, individual participation and accountability to guide all agents follow best practices [4]. Agrawal *et al.* [2] collected these into a set of ten principles in their development of the seminal work on Hippocratic databases. The ten principles mirror those indicated by the governmental definitions but are described so they can be operationalized as design principles for a privacy-aware database. The principles are: (1) a requirement to specify the purpose; (2) acquiring explicit consent; (3) limiting collection to only required data; (4) limiting use to only that specified; (5) limiting disclosure of data as much as possible; (6) retention of the data is limited and defined; (7) the data is accurate; (8) security safeguards are assured; (9) the data is open to the provider; and (10) the provider can verify compliance to these principles.

Although there have been serious challenges to the appropriateness of each of these principles, privacy proponent generally support them and argue that if they were implemented in full in modern systems, privacy protection would be well supported. If we assume that these principles are necessary and sufficient to provide privacy protection then the key question is: Who should be ultimately in control of ensuring that private data (or private data derived from inference tools) is protected? We will consider both extremes of a continuum before considering possible middle ground between them.

At the one extreme, data about an individual should be fully under the control of the principle. In other words, information about me and its use should only be permitted in complete conformance with my *preferences*.[6] This would require that privacy meta-data be tightly linked to all data so anyone that has access to the data can determine precisely how or if the data can be used for the proposed purpose. Systems would have to be built that could operate based on the privacy meta-data and only use it if the principle specifically allowed for its use for a particular purpose. If the meta-data does not explicit allow for its use for a purpose, the system would have to ignore the data's value as a normal part of its operations.

At the other extreme, any data collected by an organization must be protected by the organization. In some ways, this is implied by the other end point but there is an important difference. Organizations often consider anonymization as sufficient protection of individual data privacy. In other words, the organization, once it acquires a data item, even if private and containing personally identifiable information, can use it as a normal part of their analytical

[6] *Preferences* are used here when discussing the desires of an individual with respect to data about them.

operations as long as it is not possible to link the data back to the individual. The organization should, though very few do, then guarantee that any information derived from this analysis is not used for any purpose other than that specified by the individual. Even if we accepted that an organization might promise to behave in accordance with this very high standard, the temptation to derive increased value from the linkage would be difficult to ignore. Furthermore, the success of data anonymization is highly questionable. For example, *k-anonymity* was almost immediately usurped by *l-diversity*, which was quickly shown to be fragile with the advent of *t-closeness*. This anonymity arm's race has continued unabated since the first papers appeared in the area and with each incremental refinement the amount of utility derivable from the data is reduced. The literature continues to blossom with yet another anonymization technique that almost always begins by showing how its predecessor fails. *Differential privacy*, originally intended for use in a statistical database, argues that in a sufficiently large repository, the addition of an individual's record is protected provided a key parameter holds. Differential privacy works as described provided the parameters are respected and the repository is sufficiently large. Ideally the systems works in a *non-interactive* environment with a trusted curator to compute and publish the statistics. In the *interactive* environment, source data must be retained in the repository, to allow for responses to queries that are not immediately captured by the "statistics" calculated *a priori*. The trusted curator must also remain as a part of the system for the lifetime of the database. However, Dwork argues that both are equivalent although the latter requires ongoing monitoring.

Anonymity could be guaranteed if corporations were willing to follow the OECD's principles and users fully understood the implications of their privacy choices, it would be possible to have users opt out of participation in any corporate processes that could lead to their data being exposed. There are two key issues that prevent this from occurring. First, and probably the most challenging is the business model of the Internet. The highest principle, since its inception, appears to be that the user should not have to pay for it. Unfortunately, users do pay for it, if not with money, then with their privacy and possibly with their security. This philosophy of "free services" and "increased value delivery" is now prevalent in all aspects of our daily activities including loyalty cards that require we give up much demographic information for what is ultimately a fairly small reward. Unfortunately this first issue may not be resolved until users see sufficient harm in the behaviour to compel corporations to behave differently. The tradeoff will be that services that we now expect for "free" will likely require some other form of remuneration.

The second issue may well be more tractable and must be put in place before governments can move to legislate appropriate corporate behaviour. Currently, we do not have the systems necessary to support the OECD's principles in a way that allows a company to provide a requested service and capture the users' preferences about how their personal data should be handled after the service is delivered. For example, the easy solution to protecting privacy would suggest that after the transaction was completed, any related data should be

deleted immediately.[7] Corporations currently try to keep every data item collected for an undetermined period of time because of its high residual value.[8] The reasons for this are a mix of partial truths and current computing system's inability to provide sufficient information about how an individual wants their data treated. Once data is curated (either from a public or private database, from a customer transaction, or from inferencing multiple data points) there is no way to explicitly link each data item back to the customer's privacy preferences. The reason is that those preferences are not collected and even if they are, once the data has been "mined" (possibly legitimately), the privacy preferences are lost. However, this is ultimately a technical as opposed to a human behavioural challenge so techniques can be developed to provide user privacy if we want to do so.

5 A Call for a New Approach

Where do we go from here? Scott McNealy argues "you have zero privacy anyway ... get over it" and this rather self-serving statement reflects what some believe to be an intractable problem in the modern era. Some even take it further to say that we do not care about our privacy any longer because we live in such an omnipresent surveillance society. However, the statement is patently wrong. A few examples are sufficient to demonstrate the point. Very few of us would argue that peeping into a window to take photographs of us in the bathroom should be considered appropriate or acceptable. We strongly support the home, or facilities such as bathrooms, as very private places and few people live in glass houses where the exterior walls are also transparent to the world. Some might argue that this is not the same thing but consider the second example.

Probably the group that "lives online" more than any other is teenage girls. The argument often put forward, even after much training about dangers online, is that this group still insist on putting information about themselves readily on social networks. Unfortunately this misses the point. Teenage girls do not consider the kind of information they are putting online as particularly private (the potential dangers notwithstanding) so a generalization has been made that they simply do not care about their privacy. Clearly the first example would be appropriate here, as teenage girls would consider a bathroom to be an extremely private space, but the argument is made strongly if a parent looks at the young girl's cell phone. This is considered by that group to be a highly sensitive form of private communication as text messages to friends (and other data kept on the phone)

[7] We set aside legal issues associated with maintaining records of sales transaction to meet requirements such as tax regulations or service agreements. It would be easy to argue that this falls within the scope of the user's purpose anyway but we are instead concerned with the "permanent" storage of this data for unspecified or other purposes that are common practice today.

[8] Many claim that credit card information is not maintained without specific permission but recent leaks have included credit card information in addition to all information required to identify an individual.

makes this a much more personal item than the diary that may have been kept by an older generation. Any parent that invades this private space of a young person is likely to find that they have breached a trust and invaded privacy not unlike the bathroom example. Similar examples could be drawn for everyone. There are things about us and about what we do or in what we are interested that simply do not belong in the public domain. Thus, we do not get away from this issue because "we cannot do anything about it" or "we do not care about it."

We need to develop systems that simply do not allow our data (or data about us)[9] to be used without our prior approval and consent. For the sake of brevity, the core aspect is we must tie together the data we provide with the privacy preferences we have about how it can be used. The data should not be used without this associated provenance. In short, any data item d_i must be linked to the providers privacy preferences p_i in an uncouplable way $< d_i, p_i >$. This privacy tuple will be the basis that provides the privacy protection necessary to enforce preferences. Although, at the highest level, this sounds simple, it will ultimately require a change in how we treat, manage, move, use and ensure compliance for data access. The approach will provide demonstrable conformance to the privacy policies and agreements that are in place, and will ultimately meet the demanding requirements of the OECD's principles.

We can now turn to the obligations that will be necessary to realize this vision and this will act as a call-to-arms on delivering a privacy conscious data management process. Although many different players will be involved in ensuring any privacy protection system is operational and the commitments made are honoured, we will focus on only a few initial pieces. The reader will likely think of many more but will also find that the modest list provided here represents many years of interesting research and implementation.

5.1 User and Corporate Obligations

Users clearly need to engage more meaningful in protecting their own privacy. This is not a case of "blame the victim" but rather a call to support users in making appropriate decisions about when, how and to whom data about themselves is being released. The current approach of informed consent by a protracted and often barely comprehensible legal agreement, presented just as the user is about to access a system or install a piece of software, is at best disingenuous. Most users are known to not read these documents and it is unlikely anyone could identify a person who reads all such agreements before ticking the appropriate box and gaining access. Excellent work has been done for many years to make our systems user friendly and accessible. The same effort needs to be undertaken so users have equally friendly, accessible and understandable presentations of a system's privacy policies and implications. This goes beyond the initial presentation of a legal agreement to allow access but to the continuous monitoring of user

[9] A much harder problem that we can turn to later.

interactions with the system so the users are continuously informed about the impact of their decisions on their privacy. Several groups are working on usable privacy and there is some interesting work beginning to emerge but this work has not yet been picked up in sufficient numbers by the corporate community. This will ultimately be a strategic advantage for corporations who are able to meaningfully protect and inform their clients.

Corporations ultimately need to commit to only using data they collect for its intended purpose.[10] This will require that systems be built capable of capturing the privacy agreements in place and that corporations be able to demonstrate that any data in its control is conformant with those commitments.

5.2 System Obligations

Systems must be developed that capture $< d_i, p_i >$ and bind them in such a way that data is only accessible when explicit permission is given and access is provided for the intended and agreed to purpose. We refer to this agreement as the *privacy commitment* and use this term when referring to the organizations commitment to the user (data provider) and the user's agreement that they understand the agreement. To develop systems that allow data to be bound to privacy commitments will ultimately involve many aspects of the data management system. The balance of this paper will describe some of the key elements that must work together to satisfy privacy commitments and ensure that both parties (the organization and the data providers) function appropriately and in an auditable way.

Database management systems receive queries from users/applications and seek to optimize the database's organization to maximize throughput. At the highest level, this is done through a user interface, a query processor, a transaction manager, and a file system that collectively optimizes queries, ensures consistency, and returns only correct/verifiable results. Ultimately all of these components may need to be privacy-aware to provide an end-to-end solution that addresses all principles required by OECD. However, we will focus on two key elements. First, the acquisition of any data requires that privacy commitments are included and stored in the system (Sect. 5.3). Secondly, the query processor must inspect those privacy commitments to ensure that the data accesses are legitimate (Sect. 5.4).

The final preliminary element is how privacy might be compromised? This is often called the *attack model* and can be quite complex if all possible attacks are to be considered. For our purposes, we use a simple attack model where the data analyst, wishing to maximize data value, seeks to pose queries that are as

[10] The intended purpose is not defined by the company's desire to acquire as much information as possible and leverage it for maximum utility. Intended purpose is defined by an agreement between a well-informed user and the clearly stated intentions of the corporation. If the corporation desires to change their intention, this is done by returning to the user to get an updated agreement.

complete as possible.[11] This means that the analyst wants to have included all data items in the database and seeks to pose queries that will include all relevant data. The attack could be an attempt to identify a particular value such as a user's private information (e.g. salary). Alternatively the attack could be of the form of an aggregation query where the goal is to include all data elements in the calculation of the aggregation.

5.3 Data Acquisition Process

Data acquisition should include the capture of any information necessary to enforce *privacy commitments*. Privacy commitments are defined by negotiation between the data provider (user) and the data collector (organization).[12] The results of these negotiations are captured in a set of *privacy policies* that collectively define the privacy commitments. The negotiations process is most easily seen as an *a priori* process whereby each user provides their privacy *preferences* to the collector, which provides for the provider a statement of their privacy *practices*. Traditionally this has been a largely unidirectional conversation in that the organization has, through a privacy statement, indicated their practices and the users has either accepted it (and gained access) or declined (and been refused access). This approach has worked to the corporations advantage in the past but for the reasons described earlier will ultimately not be a sustainable model.

The question becomes: How can negotiations be undertaken with multiple users so they provide the maximum amount of access for the corporation while having their privacy interests protected? There are no current complete solutions to this problem. However, several elements must be brought together to realize one:

- Clear interfaces must be provides so users understand privacy choices and their impacts. Users should be able to easily set their preferences in light of the organizations practices and understand what benefits they derive from the choices they make and the risks they accept as a result.
- Organizational practices must be constructed to be "negotiable" in a clear and transparent way. The current all-or-nothing approach is unacceptable and does not allow for individual sensitivities nor for individual flexibility within their own choices.
- Users must be able to audit how their data is accessed. A clear privacy requirement is that of transparency and that can only occur with access that is reasonable, timely, and accurate.

[11] The "attack" here is a bit of a misnomer in that the analyst is probably only doing their job and not intending to attack or unethically compromise the users privacy. However, from the user's perspective, if they have not consented to participating in a particular kind of analysis, they will see the analyst job as an attack on their privacy.

[12] In the following we will use the terms "provider" and "collector" to represent more abstractly the concept of a user and organization, respectively.

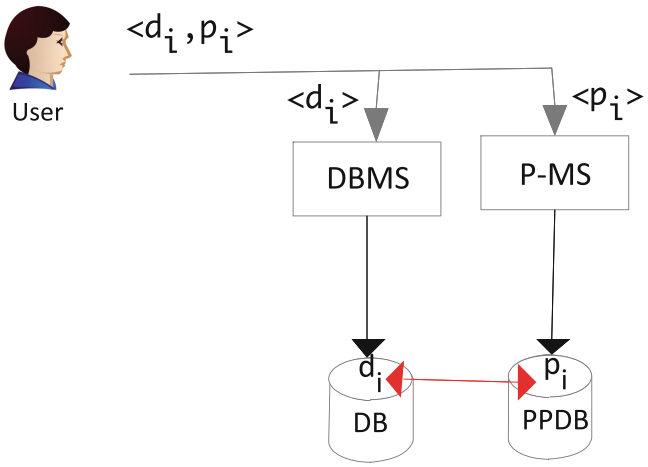

Fig. 1. Data acquisition requirements

– Users must be able to alter or withdraw their permission to access the data provided. This does not mean past access would be considered invalid but that future access is made in light of the changes made by the user. This also implies that the temporal aspects of privacy be included as a key part of the privacy commitments.

The reader will likely be able to identify many other open questions related to data acquisition and the user interface issues that must be in place to facilitate privacy commitments.

Although the interface challenge will be substantial, the output from the process should be a privacy policy that is linked to every collected data item. Figure 1 illustrates that any data item (d_i) acquired from the users is collected with its corresponding privacy policy (p_i) information. The data item itself will be stored in the database (DB) and the privacy information in the privacy-policy database (PPDB). Thus, both elements are retained in the system and used when queries are executed against the provided data. Figure 1 suggests that d_i is separated from p_i so the data can be managed using a traditional non-privacy-aware DBMS, while the privacy policy elements are maintained in a separate *privacy management system (P-MS)*. This architectural model depicts a compromise in that existing (or legacy) DBMSs will likely exist into the foreseeable future so managing privacy is likely to require a separate but tightly linked subsystem that only allows access when the privacy commitment can be honoured. Future, more secure, privacy-ware data management systems could more tightly link d_i with p_i so access is only possible to the data if the commitments are honoured. Unfortunately, current legacy systems would allow access to the DB without forcing privacy checking of the data with the PPDB. However, we do not consider this "super user" form of attack because we assume that the database administrator (DBA) is trustworthy. Recall that our attacker is an analyst who would not be given direct access to the DBMS because they are simply a system user.

5.4 Managing Private Queries

Fully supporting a privacy-aware data management system will involve many interacting architectural components. Our goal here is modest in that we consider only the processing of a high level query and introduce the key architectural components necessary to facilitate the production of appropriate results. We will begin by describing the components and then discuss how they provide privacy under the attack model presented earlier.

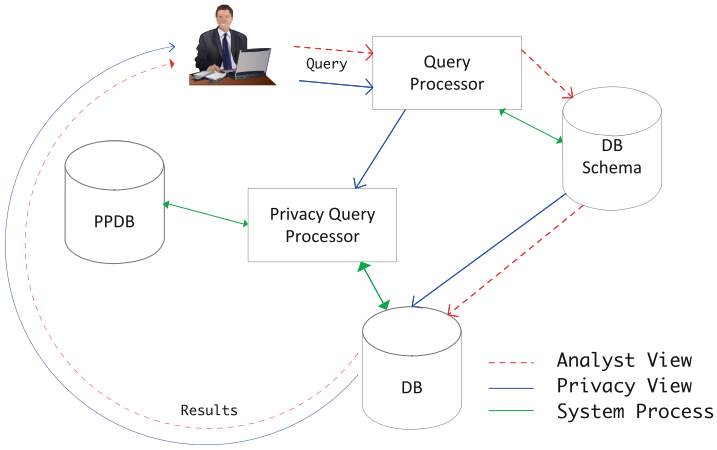

Fig. 2. Privacy preserving query processor architecture (Color figure online)

The analyst submits queries unaware of any underlying data flow and receives results back from the system in an opaque, privacy preserving way. Figure 2 depicts the analyst sending queries to the *query processor* based on view or database schema provided by the DBA. The database schema (DB Schema) is a euphemism for the schema, data directory and catalog necessary to pose queries (i.e. transactions) on the database (DB) itself. The relevant system level interactions are depicted by **green** arrows in Fig. 2 and all such interactions are opaque to the analyst. The query processor produces the data access plan in the normal way and passes its output onto the *Privacy Query Processor* (P-QP), which is responsible for modifying the query based on the privacy policy database (PPDB) that captures the privacy commitments for each data item. The P-QP modifies the access plan so the only data accessed are those conformant with the privacy commitments made between the organization and its users. Data that does exists outside of the commitments is never accessed.[13]

[13] The normal database operations, such as paging algorithms, may actually access some of the private data but since it is not seen at this level as a part of the modified query, it is not returned. There are associated exposures from paging in data that is not a part of the query *per se* because it is exposed in memory but this is not relevant to our simplified attack model.

Figure 2 also depicts the privacy protection afforded as a result of the architecture. The analyst will pose queries and the perceived flow is depicted by the **dashed red** arrows. The query is submitted to the query processor, which has provided the analyst with a view based on the DB Schema. The DB Schema facilitates access to the data itself that is stored in the database. The analyst sees *all* of the data legitimately afforded by that view so the result returned is seen as a complete answer to the query posed.

Privacy is provided by the **blue** arrows. Queries posed by the analyst, after normal processing by the DBMS's query processor are then routed to the Privacy Query Processor to ensure any data accessed in the plan is legitimate. The query that is executed based on the DB Schema, as the analyst believes, but only privacy-preserving solutions are returned to the analyst in the results. This is true of all kinds of queries:

Specific Queries: that ask for a specific record's value will either receive the value if access is legitimate or "null" if there is no value or a privacy violation would occur with its release. The analyst will not be able to discern with certainty that the value is being withheld.

Aggregate Queries: that summarize multiple data items will only include data that can be legitimately accessed. This is inherently different than most current approaches which trust in the aggregation to anonymize, and thereby protects privacy, individual data. Unfortunately, these techniques are known to be susceptible to various kinds of probing attacks or by multi-query attacks that ask slightly different questions of the same data to identify unique values in the database. The proposed architecture does not allow any data to be accessed unless it is permitted by the privacy commitments.

6 Summary

The primary purpose for this paper is to address the growing need to account for individual privacy in big data analytics. The first challenge is to understand that this issue is not restricted to "big data" but is certainly exacerbated by it. Governments, data provides, and the general public are becoming increasingly aware of the risks associated with the unrestrained use of data about us and will ultimately move to enforce limits on how it is used. Users do not currently fully understand privacy and how the data collected about them impacts on their daily lives; and this has been taken advantage of by many modern organizations. However, this unrestrained use of our personal data is fast approaching its limits.

Modern analytics will shortly be challenged in its fundamental assumptions. Current beliefs around ownership assume that data "generated" by analytics about an individual "belongs" to the analyst and can be used or resold without restraint (legal or ethical). This assumption will be challenged. An organization that is able to infer information about me with a high degree of accuracy should not be able to claim that they can use it even if it was not provided by me directly. Current mobile tracking systems and their associated analytics should not be

used to make (possibly incorrect) assumptions about the person. For example, a police organization may request travel records from an insurance company to determine who might be frequently in "attendance" at a geolocation where a racist group has its offices. The insurance company collected the geolocation information for a completely different purpose such as reduced auto insurance rates because of the customer's claim to be a "good driver." It is entirely possible that those offices are geographically indistinguishable from an HIV clinic that the person is attending with a friend who has AIDS, but the police analytics cannot make such a semantic distinction so inappropriately flags the person as a member of a racist organization (in the interest of public safety). The insurance company's analytics reveals that the individual may be good driver but attends an HIV clinic so is denied coverage when attempting to renew their life insurance policy with this organization. The model described here would disallow both kinds of queries since they would violate the privacy commitments in place.

This paper has address societal and technical motivation for the creation of a privacy aware data management system. It has provided a very high level architecture and called for the development of one of the key enabling technologies that must be developed so users can make informed decisions about their privacy preferences. However, it has only scratched the surface. Fortunately, there is much activity underway.

The ultimate future challenge is that this requires not just technological solutions but also societal, governmental, legal, and corporate engagement. However, the technologists, should not start by saying that any solution we produce is doomed to failure without the committed involvement of all of these other players. In fact, these other players can abdicate responsibility for privacy solutions because there are no technological solutions available, which would lead to the vicious cycle where all stakeholders blame the lack of action of others for our collective inability to protect privacy.

Privacy protection or data value: can we have both? Although these two goals need to be held in tension and care must be made to ensure that systems are created that honour privacy commitments made in exchange for providing data, there is no reason we cannot have both. The motivation behind extracting maximum value from data to facilitate maximal dollar value is a powerful motivator for many modern data collectors. However, as with any technology, reasonable constraints must be placed on its use to ensure that it is used appropriately. Although governments might ultimately intervene to impose those constraints if industry cannot find a way to manage itself, there are only surmountable technical challenges that will ultimately prevent us from achieving utility while protecting privacy. This paper calls for two technical developments. First, data is only collected when the privacy information describing its use is collected with it. Secondly, that systems be developed that interrogates that privacy information before allowing access to the data. The idea itself is a simply one ... its implementation will be more challenging.

References

1. The privacy act of 1974 (September 26, 2003 1974). http://www.archives.gov/about/laws/privacy-act-1974.html
2. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Hippocratic databases. In: VLDB 2002: Proceedings of the 28th International Conference on Very Large Databases, vol. 28, pp. 143–154. VLDB Endowment, Hong Kong (2002)
3. Bacon, F.: Religious Meditations, Of Heresies (1597)
4. Bennett, C.: Regulating Privacy: Data Protection and Public Policy in Europe and the United States. Cornell University Press, Ithaca (1992)
5. Duhigg, C.: How companies learn your secrets. New York Times Mag. (2012)
6. Gudivada, V., Baeza-Yates, R., Raghavan, V.: Big data: promises and problems. Computer **48**(3), 20–23 (2015)
7. Mayer-Schonberger, V.: Delete: The Virtue of Forgetting in the Digital Age. Princeton University Press, Princeton (2011)
8. Peppers, D., Rogers, M., Dorf, B.: Is your company ready for one-to-one marketing? Harvard Bus. Rev. **77**, 151–160 (1999)
9. Zhu, X., Davidson, I.: Knowledge Discovery and Data Mining: Challenges and Realities. Information Science Reference. IGI Global, Hershey (2007)