I/S: A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY

A Privacy and Security Policy Infrastructure for Big Data

RAHUL TELANG*

I. INTRODUCTION

The growth of Internet and broadband shook the business and policy world. Many firms and industries (retailers, publishers, and so on) went out of business or substantially altered their business models due to the Internet. It also created new businesses, rapid innovation, and spurred entry of new firms, such as Amazon, Google, and Facebook. However, the growth of the Internet highlighted the challenges of existing laws and regulations. Individuals, firms, and even nation states had new openings to take on challenges they could not previously have undertaken. But they were also exposed to new vulnerabilities. Information Security and privacy is the most prominent challenge which can be directly attributed to the growth of computing and network infrastructure. As firms and individuals opened their networks by connecting to the outside world via a mostly public Internet, the ability of malicious actors to intrude into these networks and access restricted data and resources also increased exponentially. These actors routinely exploited weaknesses in firms' networks, widely used Internet protocols, and various software and operating systems. Even ignorance or negligence on the part of non-malicious actors had significantly larger consequences than before. The challenges of managing Information Security are well documented, and they are not just limited to firms and individuals. Even nation states are actively participating in attacking and defending their respective infrastructures. The lack of security led to the growth of a billion dollar industry with a variety of firms providing

* Professor of Information Systems and PhD Program Chair, Heinz College, Carnegie Mellon University.

security products and services. It also highlighted the limitations of our policy-making and laws. A variety of new laws concerning cyberspace have either been enacted (regarding, for example, Internet fraud, spam, data breaches, and intellectual property)[1] or are being widely debated on Capitol Hill (pertaining to such subjects as liability for software designers and Internet service providers, vulnerability trades and disclosures, and information sharing).[2]

A second development which followed the growth of the Internet and computing is what is currently termed as Big Data. As users increasingly rely on the Internet for most daily activities from search, to commerce, to social interactions, firms found opportunities to monitor and collect user navigation patterns on the Internet. Firms now can store details of online user behavior on increasingly cheap media and apply sophisticated algorithms on powerful servers to create products, services, and customer experiences which are highly personalized. Even more importantly, firms can now *target* and reach a customer by inferring their preferences from the Internet traces they are leaving behind and offer them precise products, services, and advertisements. Firms like Amazon, Google, and Netflix have demonstrated the value of these technologies by offering efficient and precise recommendations and doing sophisticated price discrimination. Other firms can use similar data to prevent fraud, predict users' health needs, or anticipate (correctly or otherwise) their future behavior. The Big Data revolution ensures that this can be done in real time, on a large scale, and on a larger variety of information than ever before. As firms collect more data, store it more efficiently, and run algorithms on this data ever more effectively, the race for storing all sorts of customer (and non-customer) data is already on. Every visit to a website is now a race between advertisers to offer even more targeted ads. The ad networks, such as Google, Baidu, and Inmobi, auction the inventory in real time by inviting various advertisers to buy a spot. They, in turn, look up their database to calculate how "precisely" they can infer user intentions and bid on the spot to show the relevant ads. All these transactions happen in real time.

This ability to transact and monetize user attention is how Facebook and Yahoo! have become multi-billion dollar firms. The business model for a variety of firms depends partly or fully on their

---

1 Eric A. Fischer, "Federal Laws Relating to Cybersecurity: Overview and Discussion of Proposed Revisions," Congressional Research Service, last modified June 20, 2013. http://fas.org/sgp/crs/natsec/R42114.pdf.

2 "Cyber Regulation, Legislation and Policy," http://www.staysafeonline.org/re-cyber/cyber-regulation-legislation-policy.

ability to predict their users' actions and offer personalized ads, products, and services. Mobile apps are another prime example where a majority of apps are ad supported. These apps have additional advantage of user location and mobility. Now the ads can be served based on where a user is located besides other network traces. Without the ability to monetize user information, it is unlikely that these app developers can bring in a variety of apps in the market.

Offering such products and services is not a new idea. the television and newspaper industries have been doing this for decades. The key difference is that these industries could (and even now) only show ads based on aggregate user characteristics or product types. For example, a sports channel might show more beer ads and a food network or a Wall Street Journal reader might be approached differently from a Pittsburgh Post-Gazette reader. Old media neither had an ability to monitor their users' behavior, nor the technology to tailor ads on the fly. The Internet changed this, and Big Data has accelerated this trend. The ability to monitor users on the Internet highway in real time and to classify them into categories that might be considered privacy-invasive, while not assuring either the accuracy or security of the relevant data, has generated much angst amongst users and policy makers.

The use of Big Data is not just limited to targeting for advertisements. A more intricate issue has been on using these customer profiles for discrimination. There are valid concerns that employers, hospitals, or insurance firms could mine Big Data (public or private) and use that information to deny products and services. They may selectively offer some services to particular users. They may even discriminate in invidious ways based on information that is inaccurate or irrelevant. A recent European ruling on the right to be forgotten is based on the notion that users might be entitled to stop such personal information that may be used against them. [3]

The two significant challenges for our society, which grow directly from the Big Data phenomenon and directly affect user security and privacy are as follows:

1. Growth in Big Data is creating incentives for firms to collect and store even more data. Invariably, this has led and will continue to lead

---

[3] "Factsheet on the 'Right to be Forgotten' ruling (C-131/12)," European Commission, last modified May 13, 2014, _http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf; Google Spain v. Agencia Espanola de Proteccion de Datos, C 131/12, May 13 2014, http://curia.europa.eu/juris/document/document_print.jsf?doclang=EN&docid=152065.

to data breaches, data loss, and data security issues. Moreover, this has also led to data sharing issues among firms as well as between firms and the government.

2. Growth in Big data is creating privacy challenges. Firms have incentives to mine customers' data and intrude on their perceived privacy.

In simplest terms, the question boils down to trading off the benefits to individuals, firms, and society, as we get better at collecting and analyzing large amounts of data in real time; versus the security and privacy threats the very same data and analysis impose. The goal of this article is not just to outline these tradeoffs, which are mostly obvious, but to discuss the variety of policy options and potential implications of those options and what might lie ahead. In this process, I will outline various stakeholders and their incentives and hope to bring in a more nuanced view to this discussion. The goal is not to advocate a particular policy but to highlight the rationales for different types of regulations and suggest their associated consequences.

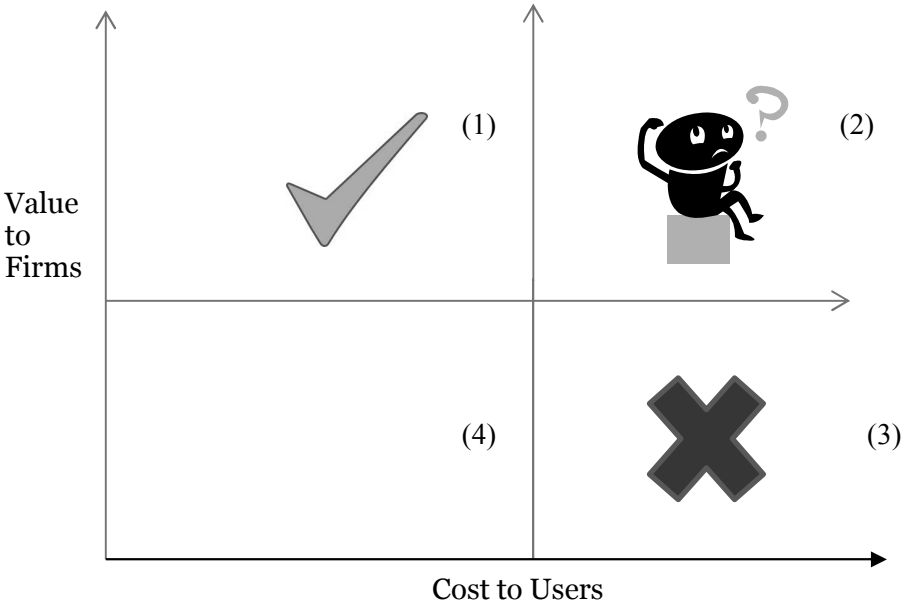
II. CONCEPTUAL FRAMEWORK

Let's just start with the larger question of firms mining customer data to create behavioral profiles, predicting their next actions, and targeting them with products, services, and ads. In that process, they will undermine users' perceived privacy.[4] There is a clear transaction taking place. Users are parting with their data – though they may be doing it unknowingly – and in return they are getting products and services. In many cases, they are essentially getting free access to a product (Facebook networking or Google search, for example) or potentially cheaper products (for example, through Amazon or over a variety of mobile apps) because the relevant firms are able to monetize user information. So what is the friction?

---

[4] In a widely noted and dramatic example, Target applied its data mining algorithm to determine whether women shoppers were likely pregnant. The store sent coupons for baby products to a pregnant high school student whose family did not yet know she was pregnant. "How Companies Learn Your Secrets," *The New York Times*, last modified February 19, 2012, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all.

There are two issues. One, the firm may be excessively harvesting customer data. In short, even if the firm offers some benefit to end users, those benefits are significantly lower than what the firm gains from such harvesting. More importantly, in some cases, the harvesting leads to privacy loss for end users. Now this is not just rent seeking by the firms but a negative externality the firm imposes on users. These privacy losses cost users monetarily but do not necessarily affect the firms (at-least in the short run). The firm does not lose much (if anything at all) by harvesting this information compared to what the user might gain by targeted products and services, or from cheaper prices. In some cases, the user may lose significantly and would not have any recourse to get any compensation. This trade-off of customer data for firm value may be summed up in the following figure.



The x-axis captures the cost to users due to firms mining their data. The y-axis captures the value to firms in collecting, analyzing, and mining this data. The left top quadrant (1) indicates that the potential value of data mining is large to the firms, but the privacy costs may not be too high. This is a potential win-win situation for all. An example might be data which is somewhat aggregated or anonymized but still very useful to the firm. The firm can still tailor its offerings, or make some predictions, but these cannot be traced back to an individual. The right lower quadrant (3) suggests that the potential value to the firm from mining this data is low but the cost to the user is potentially high. For example, many firms collect detailed user-

specific data (sometimes for record keeping) but not use it for any productive purposes. Or, the marginal benefit to the firm of such information may be small. So, while firms may find it marginally beneficial to collect information, the loss of this information can be calamitous to users. Finally, the top right quadrant (2) poses the greatest challenge. While the privacy costs are high, the benefits to firms are large, too. Our challenge as a society is to find an acceptable balance. In particular, we do not want to be in the bottom right quadrant (3) and we want a balance when in quadrant (2). How do we accomplish this?

First, this schema is conceptual at best. While the trade-off seems intuitive, the costs and benefits of Big Data are not readily quantified in any quadrant, although, at an aggregate level, one might be able to measure the costs and benefits of such data harvesting. By "aggregate level," I mean to include the costs and benefits across all four quadrants in the figure above. The questions to ask are: "How much does society (both firms and users) gain by harvesting this data?" and "How much do users lose due to their potential loss of privacy?" There are efforts already in the economics and marketing literature to measure the value of behavioral targeting.[5] Generally, it is believed that users are more likely to click targeted ads, purchase when exposed to such ads, and are made better off through personalized products and recommendations. Firms or even governments may also be able to use Big Data to spot frauds or detect negative trends, (such as rising unemployment or outbreaks of disease), and take timely corrective actions. Of course, any account of benefits must recognize that many firms like Facebook might not even exist without their ability to mine data. Measuring the cost to users due to loss of privacy is much harder. Privacy entails more than the threat of monetary losses, but the moral and emotional components due to a perceived loss of control over private information is harder to quantify.

However, even if we were to measure these costs and benefits and conclude that notwithstanding some adverse privacy incidences, in aggregate, the benefits outweigh the costs, we still have a problem. The firm is still collecting and analyzing data from the bottom right quadrant (3) where user privacy costs are too high. We would prefer that firms stop those practices and stay on the top left quadrant (1) or possibly (2). In other words, aggregate measurements still do not solve the allocation problem. A second problem in this barter of user

---

[5] Dan Breznitz and Vincenzo Palermo, "Life Is But an Online Shopping Journey? Exploring the Dynamic Interactions Between Targeted and Paid Search Advertisement Mix," Northwestern Law, last modified May 23, 2013, http://www.law.northwestern.edu/research-faculty /searlecenter/events/internet/documents/ Breznitz-Palermo.pdf.

data with products and services is that users are heterogeneous. Some users may value their privacy much more than others while some may value cheaper prices. In fact, a key stumbling block in privacy debates has always been that, unlike other commercial transactions, data transactions do not allow users to have any say. This lack of control, in my view, sums up the policymaker's dilemma.

One blunt policy instrument would be to ban the harvesting of such information. But firms and society at large will be deprived of potential benefits that accrue from being in quadrant (1) or (2). Some users indeed may be willing to trade their information and in return gain some benefits. They will be shut out from executing this trade. Users would have privacy by default, and it would be up to the users to find ways to trade their information in exchange for personalized products or services. The other extreme would be to allow firms full freedom to harvest this information and give users a chance to "buy" privacy. The default would be "no privacy," but users could pay to get their privacy back.

On the face of it, both options are sub-optimal unless we firmly believe (a) that any sort of data mining and targeting is socially detrimental and should be banned, or (b) that privacy is a fundamental right that cannot (and should not) be breached even if leads to potential benefits. Otherwise, we need a policy or other mechanism that allows for heterogeneous privacy provision and more efficient allocation.

In the U.S., the hope is usually that the market mechanism will achieve this goal. This is especially true for firms directly dealing with end users. In fact, we expect that, in a competitive market, firms should be willing to limit data harvesting, at least with regard to data of high value to customers, but little value to firms. And, if a user demands more privacy, then firms should be willing to provide it. But for the competition to work, market power cannot be concentrated within a small number of firms. More importantly, users need to be informed about firms' data usage policies and need to have some control over how their data might be used. Much of the U.S. policy making thus has focused on transparency and disclosure. The legislative efforts and recommendations (for example, the FTC's privacy report[6] or the White House report on Big Data[7]) have focused

---

[6] Federal Trade Commission. "Protecting Consumer Privacy in an Era of Rapid Change. Recommendations for Businesses and Policymakers," last modified March, 2012, http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era -rapid-change-recommendations/120326privacyreport.pdf.

primarily on forcing firms to disclose their practices so users can make informed choices leading to more competitive markets.

Optimally, we would like the firms to move from quadrant (3) to (1). It remains to be seen, in the medium or long run, if markets alone can deliver such move. Critics already argue that markets have failed and have no chance of working. Technology is progressing too fast, and firms are finding more ways to intrude on user privacy faster than markets can ever adjust. Anecdotal evidence, however, probably paints a less pessimistic picture. Some of the larger firms that directly transact with users have been responding to user concerns more aggressively. Media, Congress, and government regulatory bodies are increasingly paying attention. Google, Microsoft, and Facebook all display some responsiveness to user concerns. There is some evidence that firms are also taking steps to define their data use policies more clearly, drawing boundaries between various data, storing data less and for shorter periods, and giving users more choices.[8] Recently, for example, Facebook announced that it will let users know "why" they are being shown a particular advertisement.[9] Thus, users would have access to information that is being collected and used in targeting them. In turn, they would also have access to potentially disallow or limit the harvesting of their data. This goes back to my earlier point. A well-functioning market will promote efficient allocation. We still have to wait and see whether the momentum for market responsiveness continues; at least for consumers dealing with large reputable firms, there is reason for optimism. It is likely that firms will move away from quadrant (3) to hopefully quadrant (1) and possibly (2). In short, the hope is that the firms will stop harvesting excessive personal data. Of course, this still does put the onus on users. It is probably fair to say that some users will still face significant cognitive hurdles in controlling their data even if more options are made available to them.

---

[7] Executive Office of the President. "Big Data: Seizing Opportunities, Preserving Values," *The White House*, last modified May, 2014, http:// www.whitehouse.gov/ sites/default/files/docs/ big_data_privacy_report_5.1.14_final_print.pdf.

8 "How You Can Stay Safe and Secure Online," *Google.com*, last accessed October 9, 2014, https://www.google.com/intl/en-US/goodtoknow/online-safety. Elinor Mills, "Search engines race to update privacy policies," *CNET News*, July 22, 2007, http://news.cnet.com/Search-engines-race-to-update-privacy-policies/2100-1030_3-6198053.html.

9 Keith Wagstaff, "Facebook Lets Users Opt Out of Targeted Ads," *NBC News*, June 12, 2014, http://www.nbcnews.com/tech/social-media/facebook-lets-users-opt-out-targeted-ads-n129566.

There is another element to the market mechanism which is often overlooked. As firms increasingly deploy technology to intrude upon users and collect user data, there are a growing number of middleware firms that are developing technologies to protect privacy. Many products are available in the market – from browsers, to encryption tools, to even our computers and handheld devices – that allow users to "buy" or "protect" their privacy (popularly known as "privacy enhancing technologies" or PET).[10] Increasingly, these products will either be available for free or at low costs and may even be embedded in our standard software (browsers, emails and operating systems). There also are services like "Trusted ID" that promise to protect user data for a price. This is somewhat akin to a market where privacy becomes a luxury good and people with resources (both education and money) can afford to reduce privacy infringement. Economically, it may be efficient in that users who have a higher preference for privacy are able to get it. Socially, it may be unfair. It may put what is still an undue burden on users to protect their privacy.

However, a combination of proliferating PET products and services, plus firms responding more robustly to user concerns, offers a hope for an equilibrium where firms and users can co-exist without policy makers putting strong restrictions on data use. Of course, this would require that firms provide an option for users to limit their data collection. There is fair bit of agreement that transparency will and should be the cornerstone of a reasonable policy in this space – and that users will be educated enough to deploy some of the PET to limit their exposure should they so choose. Thus, some users will be able to prevent firms from using their information and be willing to forfeit potential benefits. In short, even if we are in quadrant (2), there is a possibility of efficient allocation where users who value privacy more than the potential benefits of information sharing will be able to exercise that option.

Markets are much less likely to work when dealing with firms that do not face users directly. An example would be third party data brokers. There are a large number of such data brokers such as Acxicom, or Datalogix. These firms collect data about users from various sources. They gather it from participating clients, such as retailers, insurance firms, and web crawlers, from public data, such as social media and public web sites, and from government records, such as court proceedings and tax records, to create customer profiles which they sell to various clients who are interested in learning about

---

[10] Julia Angwin and Emily Steel, "Web's Hot New Commodity: Privacy," *The Wall Street Journal*, last modified February 28, 2011, http://online.wsj.com/news/articles/SB10001424052748703529004576160764037920274.

potential customers. Client firms use this data mostly for marketing purposes. But in other cases, this data is also potentially used to make decisions regarding employment, insurance, and other sensitive purposes. Sometimes, the information can be highly privacy invasive even if used only for marketing purposes -- for example, sending unsolicited flyers on erectile dysfunction. Because data brokers do not transact directly with consumers, consumers have little or no leverage. Unlike the consumer-facing firms like Facebook or Google or even regular retailers who have to respond to consumer concerns and media pressure more directly, the data brokers do not face direct scrutiny with end users. Even though there are many data brokers, the precise role of competition is unclear. Since many users will not even recognize these firms, it is hard to imagine that competition from the user side will force any action. Even the client firms are probably not likely to be overly concerned about the accuracy or security of data brokers' data beyond a threshold. After all, the cost of inaccuracy or data breach or privacy invasion will be borne mostly by end users and not by the client firms.

In my view, the markets are less likely to be effective when it comes to third party data brokers. Even forcing some transparency (which the FTC has been recommending) may prove insufficient because most users do not directly deal with these firms. Transparency would require the brokers to let users review their information or even potentially opt-out. While these may be reasonably effective strategies when it comes to firms like Facebook, it is not clear if they will work for data brokers. Except for very attentive users, it is unlikely that average users can discern their privacy options and take appropriate action at these sites. Thus, it is not entirely clear that markets can readily discipline the data brokers from excessively harvesting data. In fact, it is likely that, without direct regulations, the data brokers might be content to sit in quadrant (3), which is clearly sub-optimal socially.

As noted above, Big Data can readily be used as the basis for discrimination. Firms may be able to predict a user's sexual preferences, her medical history, or her political views and, on some such basis, discriminate in providing jobs, housing, or even particular products, such as insurance. Discrimination, of course, has always existed in some form. Discriminatory practices persist despite the many laws on books that outlaw them. However, with Big Data, both the existence of discrimination and the underlying reasons for discrimination might be especially hidden from users and policymakers. Many users may disclose their political views or sexual orientation on the Internet. Algorithmic predictions about a user's sexual preference might deny her employment and the user might

never know what was the cause. Markets are much less likely to be helpful in overcoming such covert discrimination.

Although researchers have done an excellent job in highlighting the possibility of potential discrimination -- for example, an Oxford study shows how user preferences can be identified from "likes" on Facebook.[11] – There is little empirical work to suggest that firms actively pursue invidious discrimination, on a large scale, based on Big Data. We can perhaps be optimistic that data-driven discrimination of this kind will not be widespread. The potential public fallout (and legal liability) facing a firm discovered to have carried on such practices would be substantial. Discrimination is likely to remain a significant policy challenge for policymakers because, to the extent it exists, it is unlikely to be overcome by market-based mechanisms alone, and it potentially serves as a significant justification for strong restrictions on data use (even if we have to forfeit potential benefits).

Besides privacy, the other major policy issues provoked by Big Data relates to Information Security. With firms collecting and storing large volume of data, the security of this data becomes imperative. The data can be hacked, breached, or lost due to negligence. Data loss can lead to credit card misuse, identify theft, or other types of financial fraud for users. At least one foundational premise for public policy on data security is somewhat settled, however, namely, there is general agreement that firms need to bear some responsibility for data security and data breaches. Data breach notification laws passed by various states have become a default regulatory device. Firms have to inform users when they lose their data (at least if the breach reaches a certain threshold). Firms also have to provide some remediation services in cases of potential harm. The FTC can penalize firms for unfair or deceptive trade practices in connection with data breaches. Moreover, recent efforts by the SEC to encourage firms to disclose cybersecurity risks in their annual reports is an example of using transparency to encourage responsible firm behavior.

No one has actually assessed the effectiveness of these regulations in any definitive way, but the notification laws are here to stay. [12] The only question is whether, instead of state-specific laws, an overarching federal law for breach notification would work better. Many firms might prefer federal regulation as opposed to compliance with dozens of differing state regulatory regimes. Unfortunately, breach

11 Michal Kosinski, David Stillwell, and T Graepel, "Privacy Traits and Attributes are Predictable from Digital Records of Human Behavior," *PNAS* 110 (2013), 5802-5805.

12 Sasha Romanoski, Rahul Telang, and Alessandro Acquisti, "Do Data Breach Disclosure Laws Reduce Identity Theft?, Journal of Policy Analysis and Management," *JPAM* 30 (2011), 256-286.

notification laws cannot protect users completely. Hacking losses may occur much later than a breach, and users would probably find it hard to prove that a firm suffering a breach was directly responsible for a particular user loss.

Many of the data breaches involve the misuse of credit debit cards. Here, the user is somewhat better protected as banks or credit card issuers generally bear the losses. In fact, the imposition of this liability on banks and large credit card firms (Visa, American Express and MasterCard) has led to the Personal Card Industry ("PCI") security standards. PCI is a set of industry developed standards with which most retailers who handle customer cards have to comply. In short, Big Data security has become a private firm issue, though some significant some policy challenges do remain.[13] The policy judgments embodied in forced transparency and remediation through data breach notification laws has encouraged significant self-regulation by the industry.

The privacy policy issues attending consumer data are more fluid. Some hope that market forces will force the firms to move away from quadrant (3). However, the situation in quadrant (2) – where information is both valuable to firms, but sensitive for users -- is likely to remain vexing. For any regulation to be sensible, policymakers must be able to measure some defined outcomes, measure deviances of firms from these outcomes, and define appropriate punishment. Generally, regulations proceed in two ways. Either we define some rules "ex ante," or we penalize firms "ex post" (after the incidence). Most industries confront a combination of both. For example, car manufacturers have to comply with a variety of ex ante standards and safety requirements, but also are held liable should their negligence in manufacturing lead to actual injuries after purchase.

In the case of data privacy, firms that mine privacy-sensitive end user data for profit, even when beneficial to users, impose a negative externality on the market. Privacy intrusion carries a cost that only end users bear, but are not compensated for it. A standard answer to such an externality is a Pigovian tax.[14] Unfortunately, measuring the extent of privacy violations on a case-by-case basis is very difficult for both firms and users, and would entail large transaction costs. Even if users feel violated by a firm's practices, it will virtually be impossible to (i) prove that there is indeed a violation of law, and (ii) quantify the harm to the end user. Proof is complicated by the fact the end users

---

[13] A significant, unsettled issue in this space is sensitive security related data sharing between firms or between firms and government.

[14] "Pigovian tax," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org /wiki/Pigovian_tax.

may not even notice for a time that their data is being misused. Quantification is yet more complex because different users will have different tolerances for privacy intrusion. Yet, unless regulators can clearly define some measurable and quantifiable violations, ex post rules will be hard to implement. Due to this uncertainty, privacy advocates prefer a significant ex ante restriction in broad data collection and data use rather than allowing for nuanced heterogeneity and ex post punishments.

Given the difficulty in defining ex post rules, regulators have naturally gravitated towards ex ante rules that rely on compliance. Many of the consent and notice rules follow this rationale. Firms are expected to provide details of their data use to end users, clearly define their policies, and get user consent before firms can use user data. For example, HIPAA and other laws in the health care industry are geared towards consent. Within a broad definition of compliance, some rules can be stricter than others. For example, there is always a big debate on "opt-in" versus "opt-out" policies. By default, most firms would prefer a default user opt-in to their data collection and use agreements, while privacy advocates would prefer an opt-out default that would require affirmative user consent to enable a firm's data access. Due to user inertia, there are major differences in outcomes when opt-in is a default versus opt-out.

In short, much of the policy-making is focused on ex ante compliance-based rules. Of course, in cases of egregious privacy violations, users and policymakers can always go for ex post liability, but, for all the reasons stated, the overall policy push has been on ex ante compliance and transparency.

III. REGULATIONS AND INNOVATION

One the biggest objections to regulation is that it will hurt innovative activities in the economy. Many firms offer innovative products and services based on Big Data analysis. Many of these firms are small businesses, and strict compliance requirements or the threat of crushing liability could make these firms unviable.

The role of regulation, though, is quite nuanced. In some cases, regulatory compliance itself requires innovation. For example, in the case of automobiles, forcing vehicles to comply with better fuel mileage led to innovation ultimately benefitting firms, as well as consumers. It is possible that a regulation banning certain uses of personal data might lead to firms innovating in ways that enable similar products and services to be offered based more on aggregate data that is less privacy intrusive, or for firms to innovate on privacy enhancing technologies which simultaneously allow data use.

In other cases, however, regulations place compliance burdens on firms that force them to divert time and money from innovative activities to compliance efforts. A breach notification law requires firms to notify users; reissue credit cards, if needed; and provide other services—all of which could entail substantial effort and expense. Such burdens may be even more acute for small firms.

The type of innovation is also relevant to consider. Some, perhaps most, innovations are marginal. It is believed that regulatory compliance generally leads to marginal innovations.[15] Other innovations are more radical, leading to new products and services that not only replace the existing ones, but which can also be hugely beneficial to society at large. Big Data has the potential to generate both radical, as well as incremental innovations. Many firms and businesses may be able to harness Big Data, not only to make their existing lines of business more efficient and productive, but also to usher in more far-reaching improvements. A significant restriction on data use could adversely affect the revenue potential for both large (Google, for example) and small firms, robbing them of the ability to invest in those innovative activities that go beyond merely marginal improvement.

An often overlooked negative consequence of regulation is deterrence to market entry. Smaller firms are less likely to enter into markets operating with stringent regulatory environments. Thus, regulations may impede competition and reduce market effectiveness. These issues are particularly salient in fluid markets like mobile apps which are dominated by small app developers.

Stewart provides a comprehensive literature review of the effects of regulation on innovative activities in various industries like manufacturing, pharmaceuticals, and telecommunications. His review shows that regulations that do not require innovation for compliance stifle innovations. Even when they require innovation for compliance, their impact on innovation is marginal at best. However, regulatory design can play an important role in determining economic impact. Regulations that are incentive-based and rely on performance standards tend to perform the best. Regulations that allow industry and markets to find a path to implementation are also most effective in spurring innovations. An analogy in this space might be to allow firms to innovate on privacy-preserving algorithms which would allow them to use user data to a degree. Government agencies in particular

---

[15] Luke Stewart, "The Impact of Regulation on Innovation in the United States: A Cross-Industry Literature Review," *Institute of Medicine Committee on Patient Safety*, (June 2010, http://www.iom.edu/~/media/Files/Report%20Files/2011/Health-IT/Commissioned-paper-Impact-of-Regulation-on-Innovation.pdf.

have incentives to release data to the public, and better technology innovation in the privacy-preserving domain would lead to better dissemination of such data.

Unfortunately, the debate on the potential tradeoffs between regulation and innovation in the space of Big Data and privacy lacks any significant empirical foundation. It is to be expected. We are still in the nascent stages of a large phenomenon. Some work by Goldfarb and Tucker [16] suggests that firms' ability to target individuals plays an important role in effectiveness of the ads. So if the information use is restricted, the ability to target and monetize will decline sharply. An ECRI study (2003)[17] suggests that wide sharing of credit reports in the US leads to broader access to credit in the US relative to the European countries. Thus, more restrictions to data sharing have deleterious effects. However, it will take a while to gather large scale scientific evidence that provides clear pointers to better policy.

However, an undesirable consequence of our continued disagreement on a policy roadmap is that it creates an environment of uncertainty. When firms are unsure of what the next policy regulation will be, they are less willing to invest. The same is probably true for consumers. At least a subset of consumers could become wary of Internet and mobile platforms and refuse to adopt those platforms even though their participation would carry social, as well as private, benefits. This is pointed out by Stewart (2010) and by a study by Idris et al.[18] Idris et al finds that, in the context of health information exchanges ("HIE"), states with clearly defined privacy policies (even if they are more strict) are more likely to see investments in HIE than states with undefined policies.

In conclusion, the Big Data phenomenon has generated strong opinions on what policymakers should and should not do. Many prefer strong restrictions on data use and still penalties for privacy violations. Others prefer a more market-based approach. In this article, I have tried to outline the conditions where market-based approaches might work better and where we need a stronger public policy push. I feel cautiously optimistic that, in markets where users deal directly with firms, market-based mechanisms with reasonable ex

---

[16] Avi Goldfarb and Catherine Tucker, "Privacy Regulation and Online Advertising," *Management Science* 57, 1 (2011), 57-71.

[17] Nicola Jentzsch, "The Regulation of Financial Privacy: The United States vs Europe," ECRI Research Report, June 1 2003.

[18] Idris Adjerid and Rema Padman, "Impact of Health Disclosure Laws on Health Information Exchanges," *AMIA Annu Symp Proc.*, October 22, 2011, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243116.

ante compliance rules will work effectively in the medium-to-long run. On the other hand, when firms are only indirectly responsible for user data, I feel less optimistic and expect that public policy will need to play a stronger role. Finally, we have opportunities for smarter regulations which are more incentive-based and allow more freedom to firms and markets for implementation. Eventually, we need more data and research to assess the effectiveness of existing policies and chart a smarter course for the future.