SMU Data Science Review

Volume 2 | Number 3 Article 13

2019

The Data Market: A Proposal to Control Data About You

David Shaw
Southern Methodist University, shawd@smu.edu

Daniel W. Engels
Southern Methodist University, dwe@smu.edu

The Data Market: A Proposal to Control Data About You

David Shaw and Daniel W. Engels

Master of Science in Data Science,
Southern Methodist University,
Dallas TX 75275 USA
{shawd, dwe}@smu.edu

Abstract. The current legal and economic infrastructure facilitating data collection practices and data analysis has led to extreme over-collection of data and the overall loss of personal privacy. Data over-collection has led to a secondary market for consumer data that is invisible to the consumer and results in a person's data being distributed far beyond their knowledge or control. In this paper, we propose a Data Market framework and design for personal data management and privacy protection in which the individual controls and profits from the dissemination of their data. Our proposed Data Market uses a market-based approach utilizing blockchain distributed ledgers for data distribution transparency and control and digital rights management technologies to provide for data confidentiality and secure distribution of the data. Our market framework and design provides an economic, legal, and advanced technology infrastructure that protects an individuals' right to privacy while nurturing a flourishing information economy.

1 Introduction

The rise of big data analysis techniques and technologies have provided us with an unprecedented insight into both ourselves and the world around us. While this insight can be powerfully educational, we are also beginning to experience some of the ramifications of vast data collection and its usage to gain these insights. One of the key casualties in the rise of data is informational privacy; as our ability to process wider arrays and larger volumes of data increases, so does the demand for data that reveals the intimate personal details of everyday life. The vast collection and storage of this data in Internet-connected databases has dramatically increased the risk of privacy breaches while providing for the easy dissemination of this data worldwide. Data that may have previously been thought of as innocuous or difficult to obtain, such as an individual's Web browsing history or aggregated medical data, can now be leveraged to exploit more pointed vulnerabilities and gain private insights into individuals even when those individuals are not explicitly identified in the data itself. Infamously, Massachusetts Governor William Weld had his medical data re-identified from a supposedly de-identified insurance data set in 1997, purportedly by cross-referencing this medical data set with a publicly available Cambridge, Massachusetts voter registry list [4].

A consequence of the increased demand for data about individuals is the existence of flourishing secondary information markets. Companies like Google and Facebook have become multi-billion dollar firms from a business model that involves attracting consumers to a free, useful service, collecting the consumers' data as they use these services, and selling this raw or analyzed data to other businesses. On a surface level, it's easy to point to the resultant economic boom arising from this business model as a net societal benefit. After all, the usual outcome of a successful industry is job creation and economic stimulus, like with American manufacturing in the post World War II era. However, in the current information market, this economic benefit is not equitably distributed to all participants. As our data, that is, data collected about us as individuals, becomes more vast and, therefore, more powerful and more intimate, we must in turn be more mindful of how our data is gathered and propagated in order to ensure our personal privacy.

The first stage in this model of data brokerage, involving consumers receiving a free or discounted service in exchange for personal information, has not evolved much since the early days of grocery store surveys. Where the value of data has increased significantly is in the secondary data market, where primary data collectors sell their users' data. Whereas the primary data collection is usually done with the consumer's consent and compensates the data originator, i.e., the consumer, for the sacrifice of his or her privacy, transactions on the secondary information market are done without the consumers' knowledge, consent, or immediate benefit since the profits are collected entirely by the primary collector selling the data. The ability to gain value from the analysis of huge data sets combined with the ease with which such data sets may be created and distributed with modern networks and computing systems has resulted in a data economy rife with over-collection and over-distribution of private data.

We propose a public Data Market as one mechanism that can be used to address the over-sharing of private information. In the Data Market, users can exert control over their own data, including control over their data distribution to secure their information from rampant re-brokerage on the secondary market. Rather than individuals selling their unrestricted, raw private data to primary collectors who solely profit from its distribution, individuals would instead sell licenses for use of their data. Data collectors can still analyze an individual's data under terms governed by the license, but they would not be able to sell and distribute that data unless authorized under the license.

Fundamentally, the use of licenses is necessary because primary data collectors have shown to be poor faith actors when given total access to private information. Before the spread of personal computers and growth in Internet users, data collectors were often limited to collecting information about individuals in their local area, and had limited distribution networks to propagate that information. As a result, data collectors like local banks and retail stores typically used collected data internally rather than sharing it externally [24]. Only a few primary data brokers existed that would aggregate data from multiple primary data collectors. Internet infrastructure, along with relaxation of

data collection regulations, changed the data collection landscape - sharing and redistributing data became far cheaper and accordingly far more profitable than just internal use. Therefore, methods to curb the redistribution of data about individuals is necessary.

The digital rights management framework and technologies used to safeguard against copyright information in digital media can be applied in the same way to private data and information. Digital Rights Management (DRM) systems have been utilized in the past to protect digital media like video games and music by only allowing certain approved playback devices access the protected media, as well as, by digitally watermarking the distributed media. Historically, DRM systems have proven to be notoriously simple to break. Good actors and caretakers of data will faithfully use DRM systems and abide by their data license agreements. Bad actors will bypass the DRM safeguards; however, compromised and stripped DRM systems are easily detected during an audit. Thus, good actors and corporations are incentivized to use only DRM protected data.

Previous concerns over a data market have generally hinged on the administrative costs of policing a market. Kenneth Laudon [15] proposed a National Information Market in 1997, but his model relied on extensive government regulation and an army of investigators to ensure compliance with the marketplace rules. To address compliance, we use blockchain technology to publicly monitor transactions in our Data Market. The development of blockchain distributed ledgers allows accounting and policing of a market to be carried out democratically, rather than by a central bureaucracy. Each data transaction in our Data Market will result in a transaction block added to an immutable chain of data blocks that is distributed across multiple servers. In this way, we can easily differentiate legally obtained data from illegally obtained data by whether the according transaction is registered in the public ledger.

Our Data Market uses blockchain distributed ledgers so that all data transactions are logged in a publicly accessible distributed ledger. Blockchains have proven very successful in managing transactions of cryptocurrencies like Bitcoin, and a similar implementation can be added in our data market. A blockchain stores a record of each data transaction. These records can be used to determine authorized licensees and to ensure that data licensees are not violating the terms of their license. Legitimately obtained private data can be tracked by locating the associated data blocks in the block chain, verifying its authenticity. By contrast, illegally obtained data will have no such addition to the block chain in addition to compromised or no DRM protection on the data itself. The blockchain's immutability property allows identification of illegally propagated data, and the offending parties can be penalized accordingly.

In the remainder of this paper, we present a more detailed view of our proposed Data Market. In Section 2 we present a brief overview of modern large scale data collection practices. We examine the legal and market landscape for data privacy and ownership in Section 3. In Section 4 we examine in more detail three of the most common data collection practices. We present our Data Market concept details in Section 5. We summarize our concept in Section 6.

2 Modern Data Collection

The economic infrastructure of data collection is a chief cause of the resultant over-collection of our data. Simply put, at present, it is profitable for data collectors to obtain as much data about individuals, that is, *our data,* as possible. A primary objective of our proposed Data Market is to correct the unaddressed externalities, in particular the lack of oversight on collection and distribution of our data, that has led to over-collection of data, and, through this oversight, to reduce data collection to an acceptable, or at least observable, level.

The concept of supply and demand is intrinsically tied to the distribution of scarce resources. However, a key distinction separating our data from other tangible goods is the lack of any concept of scarcity. While there are physical limitations on how much data can be collected, like a small user base for an application or a lack of appropriate back-end storage systems, once collected, data can be copied and redistributed an infinite number of times at essentially zero cost. Figure 1 illustrates this unchecked distribution of our data.

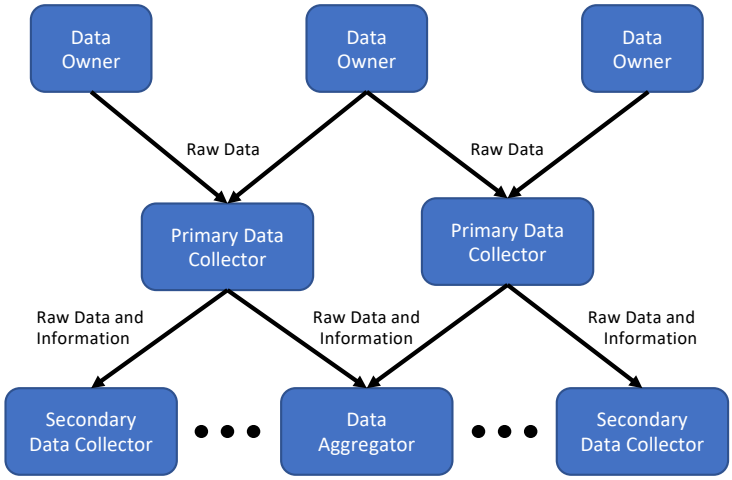

Fig. 1. The current data transaction model. Primary data collectors are allowed to freely re-distribute collected data to any foreign parties, without any approval necessary by the original data owner.

Privacy is intimately related to the unchecked distribution of our data. Privacy is generally defined in the data context as the ability to control the distribution and flow of one's personal data [21]. In the case of consumers volunteering personal information about themselves, we can discuss the actual transaction in terms of the brokered privacy rather than brokered data. Individual privacy is a finite good unlike personal data. With unchecked data distribution, once an individual surrenders personal information to a data collector, the individual no

longer has control over the distribution and flow of that data - or rather, they have ceded a portion of their control to the data collector. And, the data owner must trust the data collector to not further distribute their data or simply accept that they, as the data owner, have ceded their data to not just the data collector, but to many unknown data collectors and data aggregators.

The market value of privacy rather than the market value of data has been explored in multiple articles and books [10][20][11]. The market for privacy from secondary data brokerage suffers from a more classic economic problem: that of externalities. While data collection firms are reaping the full benefit from data brokerage, they are not bearing the full cost of collection and transmission of data. Data breaches can cost companies millions of dollars in restitution and fines, but this amount pales in comparison to the value lost in undiscovered data breaches [19]. Proliferation of private data bears a significant privacy cost on the original data owner, where the risk of foreign data breaches or re-identification of aggregated data is taken by the data owner, i.e., the individual the data corresponds to, rather than the data collector. As a natural consequence, data collectors are encouraged to aggressively over-collect individual data as they are internalizing the full benefit of collection, but externalizing some of the costs [22].

Fundamentally, negative externalities like over collection arise from ill-defined property rights. Coase's Theorem [6] states that if property rights can be negotiated costlessly, the socially optimal economic outcome will be achieved in the presence of a negative externality. In the context of a data market, this theorem states that as long as data ownership rights can be costlessly negotiated between data owners, purchasers, and collectors, the equilibrium level of collection will settle at some socially optimal level as illustrated in the black lines of Figure 2 while subsidized demand will result in some higher price as illustrated in the red line of Figure 2.

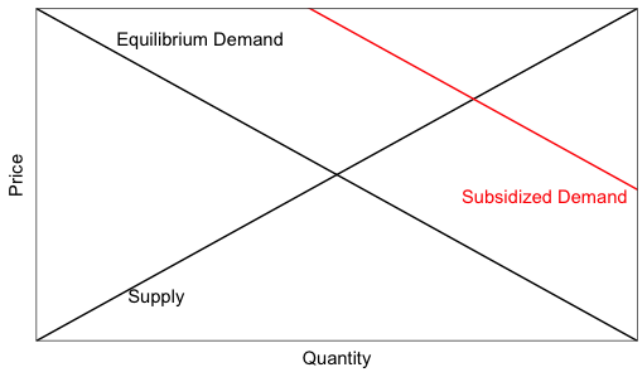

Fig. 2. Supply-demand model of current data market. When the costs of collecting data are externalized by the collecting firms, the optimal level of data collection is exceeded.

In practice, the assumption of costless negotiation is rarely met. Asymmetry of information in contract and technology structure between data-generating individuals and the collecting firms means that an extremely small minority of the population is actually aware of what data they are generating and the true value of that data. Additionally, as more and more industries have shifted to the Internet, the need for individuals to maintain an online profile, by which they generate significant amounts of data seamlessly and with every interaction, has become increasingly necessary. Opting out of online data transactions is simply not a possibility for many businesses and individuals, resulting in a concurrent increase in the number of primary data collectors and in the potential data breaches. Negotiating data rights between every single entity privy to one's data would undoubtedly be an extremely costly negotiation process. For something as simple as purchasing an item online, a socially optimal level of data collection can be obtained only if the purchaser can negotiate data property rights separately with his or her bank, credit card company, market intermediary, seller, ISP, and OS proprietor at the bare minimum. In practice, any number of third party software and hardware vendors would also need to be involved depending on the configurations of the purchaser's home network and the various websites and payment services involved. It's clear that there is really no practical notion of costless negotiation of data rights in the modern implementation of the Internet. Thus, the issue of externalized costs must be addressed in a different way to preserve a sustainable information economy that protects data rights.

3 Data Privacy and Ownership

Free market economics often elicit the need for legislation to correct market inefficiencies or imbalances, such as the Internet's cost-benefit asymmetry of data collection that significantly favors the data collectors over the privacy rights of the individual. One solution to this imbalance is the recognition of property rights that individuals have over their data. Data is recognized to have significant value and to pose violations of privacy rights when distributed beyond the data owner's control. Therefore, data is property in an era of the Internet and Artificial Intelligence (AI), and the existing property and privacy legislation should be applied to protect individuals and their data. Additionally, data specific legislation may be passed that would grant individuals autonomy over their personal data, prohibit sales and brokerage of their data, and prosecute parties illegitimately trading in owner's data thereby explicitly recognizing that personal data is just like any other physical form of property. Currently in the United States, data is not recognized as being owned by any one entity, let alone the individual without whom the data doesn't exist. Protections like the Fourth Amendment exist to guard against certain unauthorized access to private data, but the data itself is not considered to be privately owned [21]. In the context of online privacy, this means that while individuals are legally protected against involuntary access to their data, there are no mechanisms for ensuring that voluntarily surrendered information about themselves will not propagate

beyond their initially surrendered intentions. In fact, nearly all of the privacy statements that all users are forced to agree to in order to use an Internet service or application, both essential and non-essential services and applications alike, provide for the service or application provider, and even some third parties, to use any and all collected data in a manner of their choosing. Simply visiting a website can result in your data being stolen by third parties without your knowledge or actual consent.

Notable exceptions to the lack of property recognition afforded to data are intellectual property rights and trade secret rights. Intellectual property is essentially legally protected information, with rights officially granted via patents. Patents give the holder sole proprietorship over the information as stipulated in the patent. More specifically, patent rights allow a patent holder to exclude others from using the invention claimed in the patent. While the domain of what is considered intellectual property remains fairly narrow, the definition is easily expanded to include private information although the process to claim those rights under the current patent system may prove to be cumbersome and unworkable. Trade secret rights, however, protect proprietary data and information against commercial use by others. Data about an individual is clearly a trade secret, and misappropriation of trade secrets is forbidden by the Uniform Trade Secrets Act (UTSA) [26] and the Economic Espionage Act of 1996 [3]. While these Acts target corporate trade secrets, their wording is broad enough to cover the huge data captured about individual's in a world that did not exist when these Acts were passed. Attaching intellectual property rights to private data would at least move to balance the incentive structure of the current market - firms can no longer externalize the costs of data collection. A 1998 study showed that between a third to a quarter of Internet users would be willing to reveal demographic information in exchange for some benefit, indicating that there are profitable transactions to be made [13]. And, it is this trading of personal data for some benefit that is at the foundation of some of the largest companies such as Alphabet and Facebook. However, beyond the immediate benefit of the service or application, users obtain no direct benefit or compensation from the subsequent use, sale, or distribution of their data.

Trading personal data and information need not be completely illegal, but to ensure sustainable transaction volumes, the original owners must be compensated fairly and not be forced to relinquish all rights to data that is about them. Currently, the law grants intellectual property rights to "promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries" [2]. While personal data is not noted directly in this list, it does fall under this purview as a writing, and recognizing patent protections to private data would allow clear avenues to trace back data origination and ensure the proper parties are compensated for sacrificing their privacy.

An early application of intellectual property rights towards online data arose in the form of digital rights management systems in the early 2000s. Digital Rights Management (DRM) systems arose in response to the proliferation of

peer-to-peer file sharing popularized by Napster [9][5]. Prior to digital record transmission, recorded entertainment such as music and video games were distributed via physical media, on game discs and compact discs (CDs). With the advent of digital media and file sharing, the supply of multimedia became virtually infinite. To combat legitimately purchased music from circulating illegally on the secondary file sharing network, the entertainment industries implemented digital rights software into their products in order to limit what users could do with purchased data.

Under a digital rights management system, rather than selling the actual data, vendors instead sell licenses to access the data. Any consumer wishing to access the data must do so through a clearing house, which reimburses the original vendor appropriately[16]. The licenses usually come with safeguards against digital copying, as well as, time and machine limits. Theoretically, DRM systems would ensure that only legitimate purchasers of that data would have access to it. Registered playback devices would be required to "unlock" the encrypted data, and non-registered devices would be identified and barred from access by the DRM system. Early implementations of digital rights management mainly applied to digital sales of music and video games, aided by the Digital Millennium Copyright Act of 1996. The Digital Millenium Copyright Act (DMCA) criminalized attempts to circumvent digital rights management, attaching penalties to illegal cracking and redistribution of protected media.

4 Common Data Collection Practices

Awareness of data collection mechanisms is central to any meaningful discussion of data privacy and data management. Enforcement of a data market model is not possible without some reasonable awareness of how our digital footprint is being monitored and tracked, both with and without our knowledge and/or consent. In this section, we present an overview of some of the most common data collection practices today; affinity cards and credit cards, web cookies, and mobile device location tracking.

4.1 Affinity Cards and Credit Cards

Credit cards were an early example of corporate data collection, used to analyze consumer behavior long before individuals started generating digital footprints on the Internet. In 1978, the Supreme Court ruled that state anti-usury laws could not be applied to national banks operating in that state [1]. As a result of this ruling, operating costs for national banks lowered significantly. Banks took advantage of this opportunity by widening their credit card marketing campaigns to gain customers nationwide [24].

This nationalization of credit markets resulted in significant increases in demand for consumer information, not just in one's local area but from corporations all over the country. The banks were shortly followed by retailers and

manufacturers, who offered their own co-branded affinity cards as a service, enticing consumers by offering low fees on corporate credit cards and cash rebates for affinity cards. All of these benefits were provided in exchange for purchasing and credit history information [24]. As a result, corporations could gain nearly full access into an individual's purchasing history and current consumer behavior with that corporation simply by tracking purchases made with the corporate card or tied into an affinity card. Affinity cards are now commonplace at retailers, and physical cards are often replaced by simply entering an associated phone number.

4.2 Cookies

A powerful tool used by data collectors is the web cookie. Cookies were originally implemented to mitigate the disadvantages of stateless Hypertext Transfer Protocol (HTTP). Under this Internet communication protocol, the client-side browser and the web server interact for only as long as it takes for a request and a response to be received, before disconnecting. In this manner, HTTP requests are essentially "stateless" - the state of the request is not maintained when a new request is instantiated [14]. While statelessness makes client-server interactions easier to facilitate, it may cause issues in sites that do want to preserve state between requests, such as an online shopping site that wants to remember what a customer is putting in their shopping cart [14]. Cookies mitigate this problem by storing required state information, and passing this state information back and forth between the web client and web server.

Cookies become a privacy concern when the information they collect and store becomes shared outside of the user's knowledge and control. As websites have become more complex, they commonly request other third party servers for certain functionality, particularly advertisements, potentially transmitting the client's information to the third party site that then installs its own cookie on the user's machine. When the user is unaware of the cookies being installed on their machines, they are no longer in control of that information and have thus lost privacy [14]. In modern applications, flash cookies have been implemented in conjunction with the traditional HTTP cookies, making cookie removal even harder [23]. Additionally, data collectors such as Facebook and other marketing companies now commonly embed cookies in wholly unrelated sites for the sole purpose of monitoring consumers' online behavior. This is most certainly done without the consumers' knowledge and approval.

Oftentimes the user has no ability to consent to or refuse certain cookies without losing access to the desired webpage - at least in the United States. As cookies have evolved from information to enable user-friendly functionality into essentially spyware, the ability to opt in or out of cookie usage has become a larger privacy issue. This is magnified in the case of third party cookies, in which a foreign site inserts cookies into a wholly unrelated site in order to monitor user behavior. When discussing third and tertiary party data brokerage, cookies are a prime example of how data is actually transferred between entities. Advertisers often make deals with individual webpages in order to allow the advertiser to

embed their cookies in the child page and monitor the site's users' behavior from afar.

To curb the proliferation of cookies, the European Union passed the General Data Protection Regulation (GDPR) in 2018. This legislation set restrictions on which cookies were allowed, limiting data controllers to personal information processing only with the user's consent or with a proven legal basis [8]. Effectively, this meant web cookies needed to either be necessary to the functioning of the website and model the initial concept of a state-preserving storage device, or be opt-in from the user, allowing the user to select which cookies are desired for functionality and which cookies to forgo. American websites, by contrast, either make no mention of the cookies supported or present cookies as an opt-out service thereby rendering paid or otherwise state-needing operations unworkable. This all or nothing cookie approach greatly increases the number of site visitors that end up with cookies from unwanted and unauthorized data collectors.

A 2019 study by Dabrowski, Merzdovnik, Ullrich, Sendera, and Weippl [8] measured the GDPR's effect on cookie load faced by visitors to the top 100,000 websites, as ranked by Alexa. The authors found that, when compared to cookie load in 2016, almost half of the top 1,000 websites used some form of geographic discrimination to limit the persistent cookies installed on European user's machines while installing significantly more cookies on machines from other jurisdictions. The study also found that cookie load was reduced by 46.7% for US consumers as well when compared to 2016, suggesting that GDPR regulations have been successful in limiting unnecessary cookie usage in Internet traffic.

4.3 Mobile Devices

An even more recent and omnipresent development in information collection is the ubiquity of cellular phones. In a privacy context, each phone's ability to store and report to a plethora of different applications its owner's location information, in real time, is worrisome. Modern smartphones not only have the capability to record the holder's location at all times, but are actually mandated to do so by the U.S. Federal Communication Commission's (FCC's) Enhanced 911 Phase II, put into effect in October 2001. The E911 mandates that the caller's longitude and latitude be broadcast in the event of a 911 call. Cell phones are constantly tracking their owner's location via WiFi access points and GPS signals, and the resultant location information can be used in many ways, from innocuous advertising to tracking fugitives, as was the case in the arrest of John McAfee (he was located using embedded location information from a selfie posted by a fan [12]).

Mobile phones are not the only device that tracks our location information. The increasing popularity of radio frequency identification (RFID) enabled devices also contributes to the constant mapping of people's physical location. Bus passes, vehicle toll systems, credit cards, and building access cards are all examples of RFID-enabled devices that help facilitate device verification. RFID devices' popularity stems from the low cost and small size needed to perform these automated operations, but also raise security concerns over the devices'

tendencies to communicate promiscuously with all frequency-synchronized readers [17]. Leech attacks on unsecured RFID devices can be carried out by surreptitiously reading the RFID tags, and replicating or storing the tag data on a "ghost" device [17]. Users of all forms of mobile devices must also place a lot of trust in the collecting entity to secure their data against data breaches and also against secondary propagation. A 2013 study [17] showed that 95% of individuals could be identified given only 4 spatio-temporal data points. This study showed that de-identification of anonymized location data points can be done almost trivially, meaning that any secondary dissemination of location data will allow the secondary collector to immediately identify users even if the data is intended to be anonymous.

5 The Data Market

To address the issues with the current state of data brokerage and distribution, we propose a voluntary primary data market, the *Data Market,* in which individuals have control over the distribution of their data. Users participating in this market would have property rights over their personal data and information, and they would be compensated for all transactions concerning their data and information. The main goal of this market is to reduce the informational and control asymmetry present in the current market, wherein secondary data brokerage leads to user data being propagated to third parties that the original owner knows nothing about. Therefore, our new market needs to be able to both fluidly process primary data exchanges while also limiting re-use and re-distribution of that data.

The primary benefit of this Data Market would be equalizing the asymmetry in the cost/benefit structure of the existing data brokerage system. Rather than allow data collectors to internalize the full benefit of personal data and externalizing some of the costs, we propose that licenses for access to raw data and information replace the primary collection of raw data itself. In this manner, collected data can be traced back to the original owner and he/she may be compensated adequately in the event of data redistribution, thereby minimizing some of the existing asymmetry. If the incentive structure is balanced, market forces will allow for a socially optimal level of data collection and compensation, where we can both reap the benefits of a thriving Data Market while still preserving some level of personal privacy. As mentioned in Coase's theorem, a key to solving market inefficiencies is the ability to negotiate costlessly. This principle is not achieved in any practical manner in the current data infrastructure, nor is it likely to be achieved in our vision of a market. However, if a set of central standards is developed that dictates the terms of data brokerage, we can vastly reduce the cost of negotiation and narrow the inequalities present in current data collection practices.

Authentication is essential to this economy, as we must be able to differentiate parties that have legitimately purchased access to private data from those who obtain the data illegally. To this end, we propose that sales of data would

consist of selling licenses to data, as opposed to selling ownership of the data itself. Under a similar approach to digital rights management, users can control the transmission of their personal information by licensing their data rather than ceding full control to the primary purchaser. In this way, an owner's data is treated just like software and creative works of art. Instead of requiring individual negotiations, the license terms will be set centrally by a legislative body, allowing users and collectors alike to circumvent an expensive negotiation process. Although removing negotiating power from both parties disallows pareto efficiency in our Data Market, the present informational asymmetry in data collection legislation and technology precludes a meaningful negotiation process anyways.

Our Data Market will make use of modern iterations of digital rights management systems as well as blockchain distributed ledgers in order to facilitate privacy regulations and data transmissions in the Marketplace. DRM systems provide a base level of security protection, protecting licensed data from being exploited outside the terms of the license. And, blockchain ledgers will immutably record transactions and ensure the veracity of data transactions within the market.

In the remainder of this section, we further expound upon various aspects of our Data Market proposal.

5.1 Technology

In our Data Market implementation, modern digital rights management technology will facilitate the initial transaction of data, provide safeguards against illegal secondary data sales, and provide a secondary auditing mechanism. Traditional cryptography runs on the assumption that in a two-party communique, the sender trusts the receiver with the raw, unencrypted data. Most common symmetric key cryptographic protocols follow this model, and they are designed to make the transmissions inaccessible to parties without the required key while allowing the key holder free access to the transmitted data. However, DRM systems evolved because the receiver in a DRM transaction cannot be trusted with the unencrypted data, but they need access to the data for their personal use.

To achieve this asymmetry in usability and visibility, content under DRM is generally encrypted using a modified public/private key schema. Public key cryptography uses a pair of mathematically related keys (typically referred to as a public key and a private key) for authentication, encryption, and decryption. Data encrypted by a user's public key can be decrypted only by the corresponding private key, and vice versa. Both parties in a public key cryptography schema have access to the other's public key but not the other's private key. By keeping one's private key a secret, we ensure that only the entity with the right public-private key pair can access the unencrypted data [25]. By using a modified public/private key scheme, digital rights management adds a layer of authentication and confidentiality to a data transaction. As the Data Market is not transmitting raw data but licenses to access data instead, DRM affords a level of protection against unwanted copying and redistribution of data.

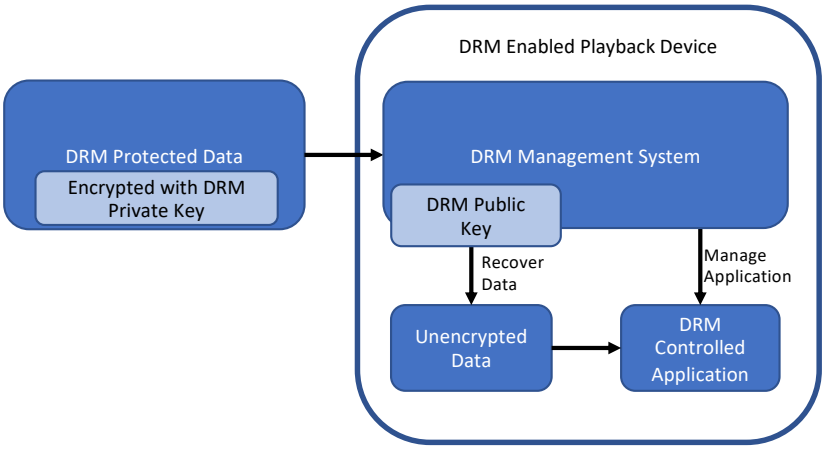

Fig. 3. The basic digital rights management model of usability.

Figure 3 illustrates the basic operation and use of a DRM protected block of data. The DRM protected data is encrypted with a private key that affords both confidentiality and origin authentication protection. The DRM protected data is sent to a DRM enabled playback device. A DRM enabled playback device may be an application specific device such as an MP3 player, or it may be a general purpose computer or server running a DRM enabled application. A DRM enabled playback device contains the corresponding public key that enables decryption of the DRM protected data. Note that even though the public key is referred to as "public" the public key is, in fact, kept secret within the DRM enabled playback device. Multiple DRM enabled playback devices may be authorized to access a certain block of DRM protected data. Each of these devices would contain the public key for that block of DRM protected data.

In the extreme, each block of data is protected by a unique public/private key pair. And, all authorized DRM enabled playback devices would need to obtain the corresponding public key from the Data Market in order to access the DRM protected data. This follows the traditional DRM architecture and operations [25].

In addition to authentication, the data license governs usage of the data. To combat the open propagation of data between tertiary sources, the license sets the terms on what can be done with the data, notably that resale without the original owner's knowledge and consent is prohibited. To aid enforcement of the license terms, data distributed with DRM contains a digital watermark, further verifying its authenticity. Any illegitimately shared data will not contain the watermark, so fraudulent data can be easily identified. This is frequently done using cryptographically secure hash functions as well as digital certificates [16]. Additional monitoring software can be distributed along with the license, in order to make sure that all uses of purchased data is in accordance with the license.

Some early implementations of DRM suffered from the overhead required for asymmetric authentication; however, the exponentially more powerful processors and larger volume of active memory allotted to modern computing devices means the overhead from public key cryptography, for DRM purposes at least, is negligible.

DRM as a standalone system can be broken relatively easily. Consequently, simply distributing data licenses governed by DRM will not ensure full security against redistribution of data. To lend some transparency to all data sales, we utilize the blockchain as a public distributed ledger of all data sales. All data sales are logged in the blockchain, and transactions are verified in the same manner bitcoin transactions are verified - democratically, by checking the hash values of each transaction and adding the transactions to the blockchain. In that manner, every transaction can be tracked, and users can audit transmission of their own data. In contrast to proposals such as the National Information Market [15], there would be no need to impose significant government resources into policing our Data Market - it can be policed publicly instead, while still maintaining anonymity and privacy. In our proposed Market, a publicly viewable blockchain will be set up that will record all data transactions. Each block will have the parties' public keys available, which provides both anonymity in hiding the parties' true identities and concurrently provides security by validating the identities of the parties [18]. Using a centralized ledger as such would allow public monitoring of data propagation. If a company owns or has access to data that does not correspond to a block in the blockchain, the entire market will know that they obtained that data either illegally or through some private secondary market.

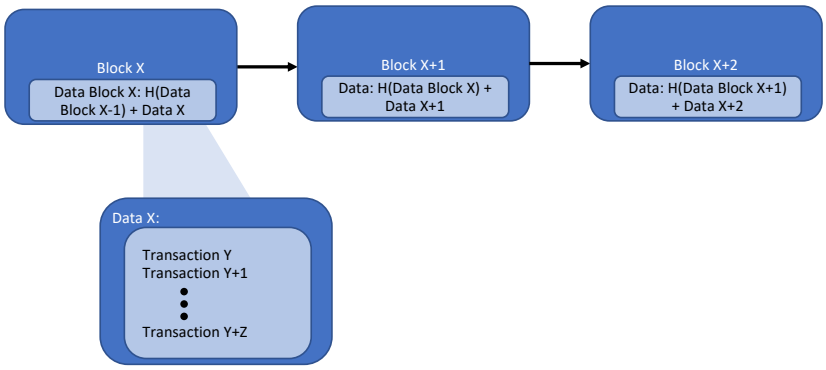

Fig. 4. A diagram of the block chain. Each block contains the hash of the previous block prepended onto the incoming data, ensuring the block chain's immutability. Incoming data blocks must solve a mathematical equation in order to be verified and added to the chain.

Analogous to bitcoin mining, before adding a transaction to the blockchain, the transaction must be verified. Transactions are aggregated into blocks as shown in Figure 4 prior to their addition to the blockchain. Adding data blocks to the blockchain is usually done by providing a proof of work involving finding a nonce that, when appended onto both the hash of the previous block and the current block, produces a hash value lower than a certain target [7]. The target value is set by the network in an attempt to enforce a certain transaction rate - for bitcoin, this rate is set globally and adjusted to match the desired transaction rate. Our market can implement a similar system in order to prevent block verification crossover and ensure the integrity of the blockchain.

Another advantage of blockchain technology is the reduction in overhead due to the decentralized system. Examples of centralized systems are notaries or other document certification systems, where that authority is responsible for providing proof of integrity, ownership, and existence [7]. In a data rich market, forcing all transactions to be approved by one entity could lead to bottlenecks and significant delays. Decentralizing the verification of each block can essentially parallelize the verification checks and lead to cheaper and faster transactions.

5.2 Behavior in the Market

The Data Market provides an open and visible marketplace where data owners and even primary data collectors can sell their data. Effectively, the Data Market operates as the secondary market clearinghouse for data owners and primary data collectors. By moving the secondary market to the Data Market, individuals will gain visibility and potential monetary gains from the secondary sale of their data. This market is achievable even if a subset of data collectors and subset of data owners agree to this model. However, there is little doubt data collectors will have their business model changed from this implementation, as they will be restricted from resale without violating license agreements. This limits the economic incentive for data collectors to switch. Public shaming and/or legal means to ensure cooperation may be necessary. If the population, and, therefore, the state, deems diminishing privacy a large enough threat to social welfare and national security, legislation can be passed to ensure that a democratically governed public Data Market is the most profitable option for all participants in data brokerage.

Although digital rights management and blockchain provide a base level of data security in our Market, there still need to be enforceable legal consequences to ensure cooperation in the Data Market. The blockchain allows identification of illegal data holders, but without proper legal consequences there are no incentives to refrain from violating the license terms and redistributing data as occurs now. Existing regulations like the DMCA and intellectual property laws must be applied to licensed data, with the according penalties for violations. In accord, licensed data in the Market would be subject to Fair Use doctrines. Although we are essentially copyrighting all data collected by Market participants, certain use cases will allow delicensing of this data for secondary propagation. For example, data collected for academic purposes that is then published will

likely fall under fair use. Authors of a study would not then need to compensate the study participants every time their study is cited. However, consumer data that is sold privately for marketing purposes would not have the same fair use circumvention. Ultimately, the goal of our Data Market is not to completely curb secondary data propagation, but rather to set limits on the volume of such transactions.

6 Summary

We present a novel implementation framework for a primary Data Market in which individuals have control over the outflow of their own personal information. Our Data Market framework and design provides an economic, legal, and advanced technology infrastructure that protects an individuals' right to privacy while preserving a flourishing information economy. This Data Market allows socially responsible companies to collect data in an ethical and transparent manner and maintain a healthy information economy without overly jeopardizing our personal privacy. Once we finally legally recognize individuals' rights to control their own information flow, our proposed Data Market mechanisms will manage data transactions and information outflow, ensuring participants a higher level of privacy than at present.

In this market, rather than data collectors amassing raw data at will, raw data must be licensed by its original owner, i.e., the individual about whom the data relates, at an agreed upon price that can be set through market dynamics. The license will restrict redistribution of licensed data, disallowing collectors from freely propagating their customers' data for their own profit. All secondary sales of licensed data must be approved by the original owner, and must redistribute a portion of the transaction back to the original owner. The preference, of course, is that secondary sales transact through the Data Market with one entity receiving a finder's fee or brokerage fee for supporting the transaction. In this manner, the problem of cost externalities can be mitigated. Since the data owner is the one at risk in the event of a security breach or data loss, they must be compensated for taking that risk. The Data Market utilizes digital rights management and blockchain technology to enforce the privacy regulations set by our market, in order to protect licensed data and monitor data transactions respectively.

Primary data collection mechanisms are unlikely to change with or without a Data Market. We discuss the roles of cookies, affinity cards, and location data in reducing our privacy, but those devices also serve many functional purposes. Eliminating or penalizing primary data collection that the user benefits from is not the goal of our Data Market. Our concerns revolve around the redistribution of private data that benefits only the data collectors and not the users at all. Rampant secondary and tertiary redistribution of collected data is the primary threat to our privacy, as we have no knowledge and no control over the flow of our data once we release it to a primary collector. The Data Market will re-establish participants' control over their data even when released to a primary collector, ensuring that secondary propagation must be allowed by the original owner

and that sales must compensate the original owner. Conversely, data collectors participating in the Market will still have the opportunity to profitably collect and analyze data for their own purposes, and may still benefit from propagating that data. We expect that a stable equilibrium rate of data transactions will be achieved in this Data Market, allowing everyone to benefit fully from the knowledge in our data without overly sacrificing our fundamental privacy.

References

1. Marquette nat'l bank of minneapolis v. first of omaha serv. corp., 439 u.s. 299, 310 (1978)
2. United states constitution, article i 8 cl. 8
3. Economic Espionage Act of 1996. Public Law 104-294, H.R. 3723 (Oct 11 1996)
4. Barth-Jones, D.: The 're-identification' of governor william weld's medical information: A critical re-examination of health data identification risks and privacy protections, then and now. SSRN Electronic Journal (June 2012)
5. Bridy, A.: Why Pirates (Still) Won't Behave: Regulating P2P in the Decade after Napster. Rutgers Law Journal **40**(3) (Spring 2009)
6. Coase, R.: The problem of social cost. The Journal of Law and Economics **3** (1960), https://www.law.uchicago.edu/files/file/coase-problem.pdf
7. Crosby, M., Nachiappan, Pattanayak, P., Verma, S., Kalyanaraman, V.: Blockchain technology: Beyond bitcoin. Applied Innovation Review (2016)
8. Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G., Weippl, E.: Measuring cookies and web privacy in a post-gdpr world: Methods and protocols. In: Choffnes, D., Barcellos, M. (eds.) Passive and Active Measurement. pp. 258–270. Springer International Publishing (March 2019)
9. Garnett, N.: Digital Rights Management, Copyright, and Napster. SIGecom Exch. **2**(2), 15 (Mar 2001)
10. Glenn Woroch, Hal Varian, F.W.: The demographics of the do-not-call list. IEEE Security and Privacy (2005)
11. Huberman, B.A., Adar, E., Fine, L.R.: Valuating privacy. IEEE Security & Privacy **3**(5), 22–25 (2005)
12. Karanja, A., Engels, D.W., Zerouali, G., Francisco, A.: Unintended Consequences of Location Information: Privacy Implications of Location Information Used in Advertising and Social Media. SMU Data Science Review **1**(3) (December 2018)
13. Kehoe, C., Pitkow, J., Sutton, K., Aggarwal, G., Rogers, J.D.: Gvu's tenth world wide web user survey. Graphic, Visualization, and Usability Center (1998)
14. Kristol, D.M.: Http cookies: Standards, privacy, and politics. ACM Trans. Internet Technol. **1**(2), 151–198 (November 2001)
15. Laudon, K.: Markets and privacy. Information Systems Working Papers (July 1993)
16. Liu, E., Liu, Z., Shao, F.: Digital rights management and access control in multimedia social networks. In: Pan, J.S., Krömer, P., Snášel, V. (eds.) Genetic and Evolutionary Computing. pp. 257–266. Springer International Publishing, Cham (2014)
17. Ma, D., Saxena, N., Xiang, T., Zhu, Y.: Location-aware and safer cards: enhancing rfid security and privacy via location sensing. IEEE transactions on dependable and secure computing **10**(2), 57–69 (2012)
18. Pilkington, M.: Blockchain technology: Principles and applications. In: Olleros, F.X., Zhegu, M. (eds.) Handbook of Research on Digital Transformations, chap. 11, pp. 225–253. Edward Elgar Publishing, Inc., Northampton, MA, USA (2016)

19. Romanosky, S., Acquisti, A.: Privacy costs and personal data protection: Economic and legal perspectives. Berkeley Tech. LJ **24**, 1061 (2009)
20. Rössler, B.: The Value of Privacy. Polity Press (2005)
21. Samuelson, P.: Privacy as intellectual property? Stanford Law Review **52** (2000)
22. Samuelson, P.A.: Diagrammatic exposition of a theory of public expenditure. The Review of Economics and Statistics **37**(4), 350–356 (1955), http://www.jstor.org/stable/1925849
23. Soltani, A., Canty, S., Mayo, Q., Thomas, L., Hoofnagle, C.: Flash cookies and privacy. SSRN Electronic Journal (August 2009)
24. Staten, Michael E. Cate, F.H.: The impact of opt-in privacy rules on retail credit markets: A case study of mbna. Duke Law Journal **52**, 745 (2002-2003)
25. Taban, G., Cárdenas, A.A., Gligor, V.D.: Towards a secure and interoperable drm architecture. In: Proceedings of the ACM Workshop on Digital Rights Management. pp. 69–78. DRM '06, ACM, New York, NY, USA (2006)
26. of Commissioners on Uniform State Laws, N.C.: Uniform Trade Secrets Act with 1985 Amendments (1985)