An Economic Analysis of Policies for the Protection and Reuse of Non-Copyrightable Database Contents

Hongwei Zhu
Stuart Madnick
Michael Siegel

Working Paper CISL# 2008-08

February 2008

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

An Economic Analysis of Policies for the Protection and Reuse of Non-Copyrightable Database Contents

Hongwei Zhu
Assistant Professor of Information Technology
Department of Information Technology and Decision Sciences
College of Business and Public Administration
Old Dominion University
2079 Constant Hall
Norfolk, VA 23529
hzhu@odu.edu

Stuart Madnick
John Norris Maguire Professor of Information Technology
Sloan School of Management and
Professor of Engineering Systems
School of Engineering
Massachusetts of Institute of Technology
Room E53-321
30 Wadsworth Street
Cambridge, MA 02142
smadnick@mit.edu

Michael Siegel
Principal Research Scientist
Sloan School of Management
Massachusetts of Institute of Technology
Room E53-323
30 Wadsworth Street
Cambridge, MA 02142
msiegel@mit.edu

An Economic Analysis of Policies for the Protection and Reuse of Non-Copyrightable Database Contents

Abstract

The availability of data on the Web and new data extraction technologies have made it increasingly easy to reuse existing data to create new databases and provide value-added services. Meanwhile, database creators have been seeking legal protection for their data, such as the European Union's Database Directive. The legislative development shows that there is significant difficulty in finding the right balance between protecting the incentives of creating publicly accessible databases (including semi-structured websites) and preserving adequate access to factual data for value-creating activities. We address this issue using an extended spatial competition model that explicitly considers licensing provisions and inefficiencies in policy administration. The results show that, depending on the cost level of database creation, the degree of differentiation of the reuser database, and the efficiency of policy administration, there are different socially beneficial policy choices, such as protecting a legal monopoly, encouraging competition via compulsory licensing, discouraging voluntary licensing, or even allowing free-riding. With the appropriate policy in place, both the creators and the reusers should focus on innovation that can increase the variety of databases and create value from database contents.

Keywords: database protection, non-copyrightable data, data reuse, policy, intellectual property

1 Introduction

There is an ever increasing amount of electronically accessible data, especially on the Internet and the Web. To a certain extent, the Web has become the world's largest data repository. The accessibility of the Web and the availability of new technologies (such as data extraction [9, 15], Web mashups [28, 56], and semantic data integration [17, 63]) allow someone to easily create new databases by systematically extracting and combining contents of other sources. As Tim Berners-Lee, inventor of the Web, said[1], "the exciting thing is serendipitous reuse of data: one person puts data up there for one thing, and another person uses it another way." Such serendipitous data reuse is extremely valuable. Through reuse, new knowledge, innovation and value-added services become possible.

While many technology-enabled data reuse activities create value for society, these activities may be against the interests (e.g., financial interests) of the database creators whose data has been reused. This conflict has infused debate about providing legal protection to non-copyrightable database contents[2] and regulating data reuse activities.

New database regulation will impact all stakeholders in the information economy, in which database creators, data reusers, and the consumers of the creator and/or reuser database products are the primary ones. One of the important factors to consider in policy formulation is the financial interests in database contents. A creator who invested in creating a database is interested in recouping the investment using the revenues that the database helps to generate. The revenues may be reduced when a free-riding reuser creates a competing database by extracting the contents from the creator's database. Thus, creators would like to have certain means of protecting the contents in their databases. Without adequate protection, the incentives of creating such databases could diminish. On the other hand, over-protection can cause under-utilization of

[1] An interview by Mark Frauenfelder of *Technology Review*, October, 2004, p44.

[2] A database can contain copyrightable contents (e.g., a database containing MP3 songs). In this case, the reuse of the contents is regulated by copyright law. Copyright laws in different jurisdictions may differ in the minimal requirements for database contents to deserve copyright protection. In the U.S., data records about facts (e.g., phone number listings in white pages) are generally not copyrightable.

information and make downstream value-added data reuse costly or even impossible. It is important, and often rather difficult, to formulate a policy that reasonably balances the two interests.

In this paper, we focus on the financial interests in non-copyrightable database contents, and analyze the case where the database is publicly accessible and no enforceable contract exists to restrict data reuse. We mainly address the issue of finding a reasonable balance between incentive protection and value creation through data reuse, that is, determining appropriate protection to database contents so that the creators still have sufficient incentives to create databases, and at the same time, value-added data reuse activities are accommodated. We frame and analyze this complex issue by developing an economic model, using the model to identify various conditions, and evaluating the social impacts and managerial implications of policy choices under these various conditions.

The paper makes several important contributions to the understanding of the ongoing debate about database protection. It provides an informative and succinct introduction on the issues and legal development of database protection. The economic model and the results provide a useful reference frame for discussing database protection policies and make an initial step towards identifying the right balance needed. This approach has allowed us to analyze legal issues from an economic and managerial perspective, bridge the gap between legal and managerial research, and derive insights meaningful to managers. This research is also timely. The results can be useful to policy makers as they continue to search for a balanced policy for the protection of non-copyrightable database contents. They are also useful to managers as they develop content creation or reuse strategies with anticipations of upcoming database legislation.

2 Background on Legal Challenges and Protection of Database Contents

2.1 Legal Challenges to Data Reuse

As mentioned earlier, technologies such as Web data extraction have made it much easier to create new databases by reusing contents from other existing databases. New business practices consequently have emerged to take advantage of these capabilities. For example, Bidder's Edge created a large online auction database by gathering bidding data of over five million items being auctioned on more than 100 online auction sites, including eBay (www.ebay.com), the largest online auction site. Similarly, mySimon (www.mySimon.com) built an online comparison shopping database by extracting data from online vendors. Priceman provided an improved comparison shopping service by aggregating data from over a dozen comparison databases including mySimon. There are also account aggregators that gather data from multiple online accounts on behalf of a user and perform useful analyses, for example, MaxMiles (www.maxmiles.com) allows one to manage various rewards program accounts and Yodlee (www.yodlee.com) aggregates both financial and rewards program accounts. In the U.K., William Hill Organization, one of the largest bookmakers in the UK, created a database by combining its own data (e.g., betting odds) with horseracing data obtained from the data feeds it licensed and the list of horses in upcoming races (called the fixture list) published in the newspaper. It displayed the contents of the database on its website to facilitate its betting business. The data in the feeds and in the newspaper were originally created by the British Horseracing Board (BHB), which is the governing authority for the British horseracing industry and is responsible for creating the fixture list for each year's races.

Common to these reuser databases is that they add value by providing ease of use of existing data, either publicly available, or accessible via licenses or on behalf of users (e.g., through the use of their user IDs and passwords). Various types of data reuse and corresponding business strategies can be found in [39].

Unfortunately, these value-added data reusers have often faced legal challenges regarding the data they extracted. For example, eBay won a preliminary injunction against Bidder's Edge in 2000 and the two firms later settled the case. mySimon sued Priceman in 1999 and the latter ceased to operate for fear of legal consequences. In the U.K., BHB won a suit against William Hill in 2001 (as discussed later, the case was reversed in 2005). There have been other cases[3]. The legal principles commonly used in the plaintiff claims in the U.S. include copyright infringement, trespass to chattels, misappropriation, violation of the federal Computer Fraud and Abuse Act, false advertisement, and breach of contract, all of which predate the Web and the pervasive use of information technology (IT). To apply them to Web-related and IT-enabled data reuse cases, these principles need to be extended and reinterpreted. It can be challenging to develop appropriate extensions and reinterpretations. For example, the court issued an injunction in the eBay case based on trespass to chattels[4]. There has been debate about the applicability of trespass law to cyberspace [7, 8, 25, 34, 43]. The main concern is that its application may threaten the fundamental functionality of the Internet and electronic commerce. Whether we need a new law for database protection has been debated. Although this is still an open question, the reality is that new laws have been enacted or proposed.

[3] For example, *HomeStore.com v. Bargain Network* (S.D. Cal, 2002), *TicketMaster v. Tickets.com* (C.D. Cal., 2000 and 2003), *First Union v. Secure Commerce Services, In.* (W.D. N.C, 1999), etc. Numerous cases in Europe can be found at http://www.ivir.nl/files/database/index.html and in [27].

[4] Trespass to chattels is a violation of the civil law when the infringing party has intentionally (or in some jurisdictions negligently) interfered with another person's personal property (which is called chattel). In the eBay case, the court considered Bidder's Edge's programs as computer robots that repeatedly visited eBay's web servers. Bidder's Edge generated approximately 1.53% of Web traffic on eBay's servers. The reasoning of the court was that such activity, if allowed, can use up computer resources, interfere with eBay's service to its customers, and cause harm to eBay. Details of the court analysis can be found in 100 F. Supp. 2d 1058. ND Cal., May 24, 2000. The current trespass law in the U.S. requires that a remedy is given only if the interference was substantial to cause dispossession of the property or there had been an injury related to the property.

Let us suppose a new law for database protection is the choice. One of the purposes of such a law would be to preserve the incentives of creating databases by providing legal protection to the investment in databases. This will inevitably run afoul of the societal interests in advancing knowledge by allowing reuse of facts in databases [51]. To resolve this conflict, the new law has to strike the right balance between preserving the incentives of database creation and ensuring adequate access for value-creating data reuse. Finding the right balance is a prevailing issue in dealing with other kinds of intellectual property [5].

Debate in the past and discussions in existing literature [16, 18, 38, 40, 42, 45, 46, 51, 52] have identified this major issue and addressed it with legal and informal economic analysis. However, this issue has not been framed and formally analyzed using an economic theory. In this paper, we develop an economic model as an initial step towards identifying various conditions and choices for setting a reasonable balance.

2.2 A Brief History of Database Legislation

Non-applicability of Copyright Law in the U.S. The impetus for database protection started in 1991 after the U.S. Supreme Court decided the *Feist v. Rural*[5] case. In compiling its phone book covering the service area of Rural Telephone Co., Feist Publications reused 1,309 of the approximately 7,700 records of Rural's White Pages. The Supreme Court decided that Feist did not infringe Rural's copyright in that white pages lack the minimal originality to warrant copyright protection[6]. Copyright in the U.S. protects the original selection and arrangement of data, not the investment in creating the database nor the contents in the database. Copyright law may evolve

[5] 499 US 340, 1991, obtainable from http://www.law.cornell.edu/copyright/cases/499_US_340.htm.

[6] This was because arranging entries alphabetically does not require any creativity, but the intention of the Supreme Court was not to set the bar particularly high in terms of creativity required to attract copyright protection.

and play an important role in database protection in the future[7], but currently it does not restrict the reuse of the contents in the type of database concerned in this paper.

New Database Legislation. The European Union (EU) introduced the Database Directive[8] in 1996 to harmonize copyright laws and to provide legal protection for database contents and "safeguard the investment of database makers" [13]. Following the EU's adoption of the Database Directive, the U.S. has attempted six proposals, all of which failed to pass into law. Table 1 briefly summarizes these legislative proposals.

<< Table 1>>

The *sui generis*[9] right approach taken by the EU creates a new type of right in database contents; unauthorized extraction and reutilization of the data is an infringement of this right. The EU Database Directive has raised several issues, which include the ambiguity about the minimal level of investment required to qualify for protection [26, 48], its lack of compulsory license[10] provisions [12], the potential of providing perpetual protection under its provision of automatic right renewal after substantial database update, and the ambiguity in what constitutes a "substantial" update. The Commission of European Communities [13] issued its first evaluation of the Database Directive in 2005. The evaluation shows that the scope of the *sui generis* right has proved to be difficult to interpret and its related provisions have "caused considerable legal uncertainty, both at the EU and national level".

[7] The originality requirement differs in different jurisdictions around the world.

[8] "Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases", a copy of the Directive can be found at http://europa.eu.int/ISPO/infosoc/legreg/docs/969ec.html.

[9] In Latin, meaning "of its own kind", "unique".

[10] A compulsory license is a mechanism to force the holder of a patent, copyright, or other exclusive right to grant use to others. The right holder often receives compensation either set by law or determined through negotiation or arbitration.

HR 3531 of 1996 closely followed the *sui generis* right approach with even more stringent restrictions on data reuse. It failed to pass into law primarily because of constitutionality concerns [12, 45].

All subsequent U.S. proposals explicitly considered the commercial value of databases. HR 2562 of 1998 and its successor HR 354 of 1999 penalize the commercial reutilization of a substantial part of a database if the reutilization causes harm in the primary or any intended market of the database creator. The protection afforded by these proposals can be expansive when “intended market” is interpreted broadly by the creator. At the other end of the spectrum, HR 1858 of 1999 only prevents someone from duplicating a database and selling the duplicate in competition.

HR 3261 of 2003 has provisions that lie in between the extremes of previous proposals. It makes a data reuser liable for “making available in commerce” a substantial part of another person’s database if “(1) the database was generated, gathered, or maintained through a substantial expenditure of financial resources or time; (2) the unauthorized making available in commerce occurs in a time sensitive manner and inflicts injury on the database or a product or service offering access to multiple databases; and (3) the ability of other parties to free ride on the efforts of the plaintiff would so reduce the incentive to produce the product or service that its existence or quality would be substantially threatened”. The term ‘‘inflicts an injury’’ means “serving as a functional equivalent in the same market as the database in a manner that causes the displacement, or the disruption of the sources, of sales, licenses, advertising, or other revenue” (emphasis added by the authors).

HR 3872 is to prevent misappropriation while ensuring adequate access to factual information. Unlike in HR 3261, injury in the form of decreased revenue alone is not an offence.

2.3 Related Work

There have been extensive legal studies on database protection policy[11] since 1996. Building on Lessig's view [35, 36] that non-legal means, such as technology, can introduce imbalance to intellectual property rights, Gibson [16] argues for the need of a database law that requires the re-reification of databases, a paradigm where database creators should deposit a technologically unfettered copy of their databases to a depository managed by the regulator. When not regulated, a creator can use technology to gain too much control over data and cause harm to public interests. However, the costs of operating the depository and overall social welfare impacts are not analyzed. Lipton [38] suggests a database registration system similar to that for trademark to allow database creators to claim the markets within which their databases are protected from free-riding. But social welfare analysis is not performed in this study to take account of the cost of maintaining such a system. After reviewing a number of data reuse cases in the EU and the U.S., Ruse [48] suggests reusers negotiate licenses from database creators and conform to the licensing terms. The paper also criticizes the ambiguity in the Database Directive and recommends that the EU should consider the U.S. proposals that contain more broadly defined fair uses and provisions dealing with sole source databases. Colston [12] provides a comparison of EU and U.S. approaches and suggests that the EU should reconsider the compulsory license provision that was in the early draft of the Database Directive, but removed from the final version. Hugenholtz [26] introduces an emerging spin-off theory for databases that are created as a by-product of other business activities, in which case the cost of the business process should not be counted as cost of creating the database.

There has been little economic and information systems research that directly addresses the issues of database protection policy. We are aware of only one paper by Koboldt [31], who

[11] See http://www.umuc.edu/distance/odell/cip/links_database.html for references to published legal reviews.

studies various distortions of database update for *sui generis* right renewal under the EU Database Directive. From the social welfare point of view, the provision can induce inadequate update or excessive update of the database. He points out that the problem comes from the substantial change requirement for an update to renew the *sui generis* right. He shows that setting up an upper limit for updating cost can eliminate the distortion of excessive update; no suggestion is made for eliminating the distortion of inadequate update.

The kind of data reuse considered in this paper is different from the so called small scale information sharing [3] and other related works cited thereof. We have focused on reuses by a firm that produces a competing database with varying degrees of differentiation from the creator's database. Small scale information sharing refers to the sharing of purchased information goods by a consumer with members within a small community (e.g., family members, friends, etc). The shared information goods are usually perfect substitutes of the goods from the original producer. Nonetheless, research has shown that the producer profit can go up or down in the presence of small scale sharing.

When the database creator has not released its contents to the public and thus still has full control of access to the contents, it can adapt its pricing strategies to respond to technological and market changes. West [61] studied several strategies used by the online database industry to increase revenue and discourage reuse of downloaded data (mainly through raising the cost reuse-oriented downloads). This is a different scenario from our research. We focus on the situation where the database creator has made its contents publicly availably and yet it still wishes to assert certain control over the reuses of the contents mainly via legal instruments.

This research identifies certain conditions under which the reuser should pay a licensing fee to the creator. Licensing is examined from a social welfare-enhancing perspective. In practice, the licensor will develop an optimal licensing schedule to maximize its profit. Licensing has been studied in the context of technology innovation [29] and intangible property [30] in general. Recently, Lin and Kulatilaka [37] studied the different licensing schedules for technology standard and investigated the impact of network effects on licensing choices. The results in licensing literature can be useful to database creators when they devise their licensing strategies. For example, the database contents can be licensed at a fixed fee, with a royalty determined by a unit price multiplied by the amount of content reused or the output of the licensee, or with a two-part tariff that combines a fixed fee and a royalty.

2.4 Summary of Other Legal and Economic Issues

Below we summarize other legal and economic issues identified in the literature.

Data monopoly. There are situations where data can only come from a sole source due to economy of scale in database creation or impossibility of duplicating the event that generates the data set. For example, as the governing authority for the horseracing industry, no one else but BHB itself can create the horseracing data. Another example is the trading activity data at the New York Stock Exchange (NYSE). Downstream value-creating reutilizations of the data will be endangered when a sole source creator engages in monopolistic practices [45, 51].

Cost distortion. Both the EU database directive and the latest U.S. proposals require substantial expenditure in creating the database for it to be qualified for protection. Database creators thus may overinvest at an inefficient level in order to qualify [51].

Update distortion and eternal protection. This is an issue in the EU law, which allows for automatic renewal of *sui generis* right if the database has been substantially updated. Such a provision can induce socially inefficient updates and make possible eternal right through frequent updates [31].

Constitutionality. Although the Congress in the U.S. is empowered by the Constitution to regulate interstate commerce under the Commerce Clause[12], the Intellectual Property Clause[13] restricts the grant of exclusive rights in intangibles that diminishes access to public domain and imposes significant costs on consumers [22]. Certain database contents are factual data in the public domain; disallowing mere extraction of such data for value-creating activities runs afoul of the very purpose of the Intellectual Property Clause that is to "promote the Progress of Science and useful Arts". It is still an open issue for Congress to decide which constitutional power could be used to enact a database law.

International harmonization. Given the global reach of the Web and increasing international trade, it is desirable to have a harmonized data reuse policy across jurisdictions worldwide. The EU and the U.S. are diverging in their approaches to formulating data reuse policies. A World Intellectual Property Organization (WIPO) study [57] also reveals different opinions from other countries and regions.

We believe the solution to these challenges hinges upon finding a reasonable balance between protection of incentives and promotion of value creation through data reuse. With this balance, value creation through data reuse is maximally allowed to the extent that the creators still have enough incentives to create the databases. Consensus can develop for international harmonization if we can determine the policy choices that maximize social welfare[14]; a database policy so formulated should survive the scrutiny of constitutionality and other inefficiencies can be avoided or mitigated.

[12] Constitution 1.8.3, "To regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes".

[13] Constitution 1.8.8, "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".

[14] Social welfare refers to the overall benefit to society.

2.5 Managerial and Policy Considerations

Firms create databases to serve various purposes. It is useful to distinguish two main purposes: (1) to sell the database as a good and (2) to use the database to facilitate the core business of the firm (e.g., online auction service or organizing horse races). In the latter case, the database may be created as a byproduct of business transactions or as a necessity of conducting business (e.g., eBay's bidding price, horseracing fixture list). As profit-maximizing agents, firms will exploit their databases in as many ways as they can. Thus a database created for the latter purpose sometimes can also be sold as a good when a firm can find the right customers who are willing to buy. For example, eBay now sells its data via licenses directly as well as through resellers such as DataUnison. Similarly, horseracing data is a business necessity of BHB and yet BHB also licenses the data to generate revenue. The NYSE also licenses its market activity data.

The pace of innovation has been such that database-creating firms cannot foresee all possible ways that the data can be extracted and reused. Innovative downstream firms often possess the knowledge and technical capability to allow them to reuse (without paying for) the data in ways the data-creating firms may never have thought of before. Such serendipitous reuse of data has been deemed valuable; this is true especially when the reuse has little negative impacts to the database-creating firms. Sometimes certain reuse can even bring benefits to the creators, in which case the creators are willing to have their data reused. For example, certain online vendors are wiling to allow comparison shopping service providers such mySimon to use their price data as such reuse can increase their visibility among potential buyers. But when the benefit is unclear or the impact of reuse can be negative to the creators, they desire to fend off the reusers.

Policymakers tend to approach the issue from a different perspective. While they consider fairness to the database creators, they are also concerned with social welfare which is the value of all stakeholders. When the EU introduced the Database Directives, one of the main objectives was to stimulate the database production industry by providing protection to the investment in database creation. When a database is created mainly for the purpose of supporting the firm's core business, increased legal protection has little impact on the incentives of creating the database. Thus we will focus our analysis on databases created for sale.

When database creation requires substantial expenditure and competition from free-riding reusers reduces the creator's revenue to a level that does not offset the cost, the creator would have no incentives to create the database. Policy should intervene to prevent this destruction of the database market (assuming the database was worth creating). On the other hand, data reuse is often value-creating; from a social welfare point of view, it is not necessary to intervene if the creator can remain profitable even though its revenue may decline because of competition. It is conceivable that there exist different conditions under which policy choices differ.

HR 3261 contains several useful aspects that are often considered in policy formulation. "Substantial expenditure" corresponds to the *fixed cost* in creating the database; "functional equivalent" measures the *substitutability* of the reuser database for the creator database, which is determined by the degree of *differentiation* of the two databases; "injury" or incentive reduction can be measured by *decrease of revenue*. "Time sensitive manner" is redundant with differentiation when information goods can be differentiated via temporal versioning[15] [55]. For example, real time stock quotes and 20-minute delayed stock quotes are two differentiated economic goods.

Policy instruments in most proposals on database protection focus on specifying what types of reuse constitute a violation (or a fair use), and often ignore what the creator is supposed to do

[15] A firm might choose not to time version its information goods if it cannot get a separating equilibrium between users with different time sensitivity requirements.

(e.g., under certain conditions the creator should be asked to license its data under reasonable terms). Such compulsory license provisions are often found in other intellectual property laws (e.g., copyright and patent laws) to force the right holder to grant use to others under reasonable terms, usually with a fee paid to the right holder. Thus, appropriate policy instruments should be a specification of conditions and the corresponding socially beneficial actions of the reuser as well as the creator.

3 A Model of Differentiated Data Reuse

As the legal discussions suggest, the reuser is sometimes a competitor of the creator in the database market[16]. Arguably, the intensity of competition depends on how differentiated the reuser database is from the creator database. The differentiation can be either horizontal or vertical[17] or both. Most aggregator databases are horizontally differentiated from the databases being extracted because they often have different features, over which the consumers have heterogeneous preferences. For example, while certain consumers value the extensive information about the auctioned items from eBay's database, other consumers value the searchability and ease of comparison at Bidder's Edge. Therefore the two databases are horizontally differentiated in product characteristics space.

In the ensuing discussion, we focus on horizontally differentiated cases where the creator database is better in some features, whereas the reuser database is better in the other features. In such cases, the creator and reuser databases are at different locations in the characteristics space. We will base our analysis on an extended spatial competition model, which was introduced by

[16] There are other reasons a creator does not want his data to be reused. For example, an online store may be afraid that a comparison aggregator can potentially have the effect of increasing price competition and lowering profit on sales of products. Our current model focuses on "information goods" only, thus it does not capture such effect.

[17] Product characteristics are horizontally differentiated when optimal choice at equal prices depends on consumer tastes (e.g., different consumer tastes in color). Product characteristics are vertically differentiated when at same prices all consumers agree on the preference ordering of different mixes of these characteristics, for example, at equal price, all prefer high quality to low quality. See [59] for details.

Hotelling [24] and has been widely used in competitive product selection, marketing and MIS research [1, 49, 53].

3.1 Model Setup

We consider a duopoly case where there are two database suppliers: (1) a database creator who creates one database product (e.g., BHB or eBay database), and (2) a data reuser who produces a different database by reusing a portion of the contents from the creator's database (e.g., William Hill or Bidder's Edge database). As another example, the database creator could be a marketing firm who compiles a New England business directory database that includes all business categories. A firm specializing in colleges in the Greater Boston area may compile an entertainment guide by reusing a portion of the business directory. Both databases are for sale in the market. The two databases are different in terms of scope, organization, and purpose. In other words, they are differentiated in the product characteristics space. The space is represented by a straight line of unit length, with the creator's database at point 0 and the reuser's database at point 1. Consumers have heterogeneous preferences and their ideal databases are uniformly distributed in the space between 0 and 1. There are N such consumers; to simplify notation, we normalize it to 1.

The creator and the reuser choose a price for their databases, p_0 and p_1, respectively, to maximize their profits. Given the two databases and their prices, each consumer decides whether to buy a database, and which one to buy, depending on the consumer's utility function given below. We assume a consumer consumes either none or exactly one database. A database is worth a value v to a consumer with exact preference match. When a customer whose ideal database is at $x \in [0,1]$ consumes the creator database, he enjoys value v, pays the price p_0, and also incurs a preference mismatch cost tx, where x is the distance between his ideal database and the creator database, and *t* is the preference mismatch cost per unit distance in the characteristics space. If the consumer consumes the reuser database instead, the mismatch cost is *t*(1-*x*) because the distance between the ideal database (located at *x*) and the reuser database (located at 1) is 1-*x*. This consumer's utility function is:

$$u_x = \begin{cases} 0, & \text{if buys none;} \\ v - p_0 - tx = u_{x,0}, & \text{if buys from the creator;} \\ v - p_1 - t(1-x) = u_{x,1}, & \text{if buys from the reuser.} \end{cases}$$

We further assume that both the creator and the reuser have the same marginal cost, which is normalized to 0. The creator's investment in creating the database is modeled as a fixed cost *F*. The reuser incurs a fixed cost *f*, where $F >> f$, so we normalize *f* to 0. This assumption reflects the fact that the innovative reuser possesses complimentary skills to efficiently create the second database that the creator cannot preemptively develop. Firms simultaneously choose prices to maximize their profits; consumers make purchasing decisions that maximize utility u_x.

This setup reflects the uniqueness of the database creation and serendipitous data reuse scenario. The cost of creating and maintaining the creator database is not a decision variable to be optimized by calculating expected returns on the database *per se*. Thus, for the purpose of data reuse analysis, the cost of creating the original dataset is a sunk fixed cost instead of an investment in the sense in Research and Development literature. Similarly, the database features are often designed without ever thinking of various possible reuses. In other cases, the creator may be constrained by its own business or its lack of skills so that it cannot preemptively develop databases at other locations in the feature space. Therefore, the database location is not a decision variable, either.

Table 2 provides a summary of the parameters and symbols used in the economic model.

<< Table 2 >>

In this model, preference mismatch cost parameter *t* also indicates the degree of differentiation of the creator and reuser databases. This is because the distance of the two databases is fixed to be 1, so that the maximum possible mismatch cost is *t*. When *t* is large, the two databases are highly differentiated and the two firms can be two local monopolies. When *t* is small, the two databases are close substitutes and fierce competition can lower profits to a level where the creator cannot recover its fixed cost. Our further analysis will be based on this intuition.

In the rest of the paper, unless otherwise noted, profit and social welfare are gross without counting the fixed cost or transaction cost. Since we normalized marginal cost to zero, gross profit is equivalent to revenue. We will consider two cases: (1) the monopoly case, where there is no reuser database; (2) a duopoly case, where there is a competing database from either a free-riding or a fee-paying reuser. The consideration of both cases allows us to analyze the creator's incentives and disincentives of having its data reused by a reuser. We use superscripts *m* and *d* to indicate the monopoly and duopoly cases, respectively. Gross social welfare is the sum of firm profits and consumer surplus.

With the above setup, we can solve the firm profit maximization problem, where the demand is determined by the consumer's utility function. The results are summarized in Lemma 1 below. Proofs of this lemma and certain other propositions in the form of theorems and corollaries are furnished in the Appendices.

Lemma 1 (Duopoly and Monopoly Market Coverage). In the duopoly case, the market is covered if $t \leq v$*, and is not fully covered otherwise. In the monopoly case, the market is covered by creator's database if* $t \leq v/2$*, not fully covered otherwise. Best price, maximum profit, and social welfare vary with t in both cases.*

The formulae for best price, maximum profit, and social welfare are given in Table 3 and will be referred to as part of the Lemma.

<< Table 3>>

We graph the result of Lemma 1 in Figure 1 to help make useful observations. The values of profit, social welfare, and unit mismatch cost *t* are measured as factors of the value *v*.

<< Figure 1 >>

Corollary 1 (Database Differentiation). $\pi^m > \pi^d$ *if* $t \le v$*, and* $\pi^m = \pi^d$ *otherwise.*

We know from Corollary 1 that when the reuser's database is not sufficiently differentiated from the creator's ($t \le v$), the creator makes less profit because of the competition from the reuser's database. When the two databases are highly differentiated ($t > v$), the creator is not harmed by the reuser. The corollary also implies that if the creator is the sole data source and can fully control the data, it will deny access if a reuser intends to free-ride and make a database without sufficient differentiation.

Corollary 2 (Preference for Differentiated Databases). $SW^d > SW^m$ *for all* $t>0$.

Corollary 2 says that from the social welfare perspective, two differentiated databases are better than one database, subject to individual rationality constraint (i.e., the creator makes a positive net profit).

3.2 Necessity of Database Law

When the creator's profit is less than its fixed cost (i.e., $\pi^m < F$), the database will not be created to begin with. For market failure analysis, we focus on the case where the creator is self sustainable (i.e., $\pi^m \ge F$).

In the presence of a free-riding reuser, the creator makes a duopoly profit π^d, which is smaller than monopoly profit π^m when $t \leq v$. When $\pi^d < F \leq \pi^m$, the creator will not recoup its cost and the database market will fail. Policy intervention is necessary to prevent market failure. This is often the argument for having a new database law.

Without a database law, other means that database creators can use to protect their databases seem to be ineffective in most cases. For example, a creator can use certain non-price predation strategies, namely by raising rivals' costs [36], to deter entry or at least to soften competition from the reuser. In the past, creators attempted cost raising strategies such as blocking the IP addresses used by reuser computers and frequently changing output format to make data extraction more difficult. This can be modeled by letting the creator choose a technology *I*, with which the marginal cost of the reuser becomes $C_1(I)$. The cost of installing such anti-extraction technologies is often small enough to be negligible. Using techniques similar to those in the poof of Lemma 1, we can solve for firm profit maximization. When *I* is such that $0 \leq C_1(I) \leq \min\left\{\frac{3}{2}(v-t), 3t, 2v-3t\right\}$, the creator profit becomes $\pi_0^d = {(t+\frac{C_1(I)}{3})^2}\big/{2t} > \pi^d = \frac{t}{2}$, that is, the creator profit is higher than when the technology is not used. The reuser profit is $\pi_1^d = {(t+\frac{2C_1(I)}{3})(t-\frac{C_1(I)}{3})}\big/{2t}$, which could be greater or less than π^d, depending on the level of $C_1(I)$. Obviously, if $C_1(I)$ is very high, the reuser will be deterred. However, anti-extraction techniques have not been very effective in practice[18]; we suspect that $C_1(I)$ has been too small to have a substantial effect. Therefore, we will assume no anti-extraction is in place in the rest of the analysis. Regardless of the effectiveness of anti-extraction techniques, they are socially

[18] eBay tried blocking the IP addresses used by Bidder's Edge, Bidder's Edge circumvented this obstacle by using a pool of IP addresses dynamically. Also some strategies (e.g., frequent changes to output format) might negatively impact legitimate users of the creator's database.

wasteful investment because they merely help transfer consumer surplus and reuser profit to the creator. When database creators are also reusers, the cost-raising problem may not arise at all[19].

There can be a need for a database law from the reuser's point of view. Database reusers often face legal challenges from database creators. For example, reusers often receive legal threat notices[20] and sometimes are sued by the creators. The uncertainty of various proposed database bills exacerbates the legal risks for the reusers, who are often small but innovative firms. As a result, some reusers have to exit the market, and certain value-added data reuses cannot occur. In this case, having a database law that clearly specifies the kinds of legal reuses will help to create and sustain a market of socially beneficial reuser databases.

3.3 Conditions and Choices of Data Reuse Policy

A socially beneficial data reuse policy can correct market failure by restricting certain free-riding in data reuse; the legal certainties it provides also help eliminate or reduce wasteful cost-raising investment by incumbent database creators. This can be done either by requiring the reuser to pay the creator for the data or by disallowing data reuse all together. The creator can ask the reuser to license the data by paying a fee, *r*, which can be up to the reuser's profit π^d; asking a fee $r > \pi^d$ is equivalent to disallowing reuse because the reuser would make a negative net profit. Like any other transaction, data reuse licenses inevitably incur a transaction cost (e.g., negotiating the fee schedule *r*, monitoring and enforcing the license, and the other activities of administrating data reuse policy often incur certain costs). This can be modeled with a transaction efficiency coefficient α; when the creator asks for *r*, it actually gets αr, and the transaction

[19] In the financial sector, many banks started offering account aggregation service shortly after account aggregators emerged. That is, banks as database creators, became data reusers, so they had incentives to lower data reuse cost. As a result, they initiated a standardization project to facilitate aggregation, see "FSTC to Prototype Next Generation Account Aggregation Framework" at http://www.fstc.org/press/020313.cfm. In this case, legal intervention is unnecessary.

[20] For instance, a few online travel agencies sent warning letters to data reusers that allow consumers to compare prices. See "Cheap-Tickets Sites Try New Tactics" by A. Johnson, Wall Street J., October 26, 2004.

cost is $(1-\alpha)r$,where $\alpha \in [0,\ 1]$. To simplify the analysis, we let $r = \pi^d$, that is, we assume that the creator has the negotiation skills or legal power to ask the reuser to disgorge all profits from reusing the data[21]. Thus, in the duopoly case with a fee-paying reuser enforced by a data reuse policy, ($\pi^d + \pi^d$) is the best the creator can get (assuming α=1) to offset its fixed cost *F* if it ever allows someone to reuse its data. Before we develop the formal analysis, we describe the intuition by plotting this upper bound condition along with the profit curves in Figure 2.

<< Figure 2 >>

We mark five regions A_1 through A_5 in Figure 2. The upper-bound ($\pi^d + \pi^d$) curve will be lower when α decreases, which enlarges A_1 and reduces A_2, A_3, and A_4. There are different implications when *t* (the X-axis) and the fixed cost *F* (measured along the Y-axis) fall in one of the areas. If they fall in A_1 (i.e., *t* is between 0 and 0.5*v*, and *F* is below the monopoly profit and above the sum of duopoly profits), the upper bound ($\pi^d + \pi^d$) curve is below π^m curve, meaning that even if the creator can reap all the profit made by the reuser, it still cannot cover all its fixed cost. In this case, the existence of a reuser causes an uncorrectable market failure, so it is better to let the creator be a lawful monopoly.

If *t* and *F* fall in A_2, market failure can be corrected by asking the reuser to pay for the reuse of the data. But the creator prefers to be a monopoly. To maximize social welfare, the policy should insist that the creator license its data to the reuser. When *t* and *F* fall in A_3, the creator can make more than it would as a monopoly, thus it is willing to license its data.

For the database to be created to begin with, we have assumed that a monopoly profit is greater than the required fixed cost ($\pi^m > F$). Area A_4 shows an interesting scenario. In this case,

[21] We assume there will be no collusive joint profit maximization. There are other possible bargaining outcomes, such as a 50/50 split of the reuser profit. For purpose of market failure correction, this outcome can be simulated by setting α to 0.5, although welfare analysis will be somewhat different.

a monopolist creator cannot afford to create the database but the database and a variant of it still can be created so long as the creator and the reuser can share the fixed cost.

Finally, when t and F fall in A_5, the cost of creating the database is low and free-riding would not cause market failure. It actually enhances social welfare when $\alpha < 1$ because transferring r to the creator costs the society $(1-\alpha)r$.

Next, we will formalize the above intuitive explanations. For notational simplicity, we let $\pi^{dl} = (1+\alpha)\pi^{d}$ and $SW^{dl} = SW^{d} - (1-\alpha)\pi^{d}$, which respectively denote the gross profit of creator and gross social welfare when the creator licenses its database to the reuser. Theorems 2-7 correspond to the five regions in Figure 2.

Theorem 1 (Minimal transaction efficiency). There exists a minimal transaction efficiency $\hat{\alpha}$, below which having a monopoly is welfare-enhancing compared to having a duopoly with a fee-paying reuser. Conversely, it is welfare-enhancing to have a duopoly with a fee-paying reuser when $\alpha > \hat{\alpha}$. $\hat{\alpha} = 0.5$ when $t \le v/2$; $\hat{\alpha} = \frac{3v^2+6t^2-8vt}{4t^2}$ when $v/2 < t \le 2v/3$; and $\hat{\alpha} = \max\{0, \frac{3v^2-4vt}{4vt-2t^2}\}$ when $t > 2v/3$.

Having a monopoly is social welfare-enhancing only if $SW^{dl} < SW^{m}$. Solving this using Lemma 1 will yield $\hat{\alpha}$ in Theorem 1. When free-riding causes market failure, data reuse policy must choose between asking the reuser to pay and disallowing data reuse all together. High transaction costs may out weigh the welfare gain from having a reuser database. When transaction efficiency is below this threshold, it is better that the creator not license data to the reuser; conversely, when transaction efficiency is above this threshold, the creator should license its database to the reuser, subject to the constraint that the creator can make a positive profit with licensing fee from the reuser.

The managerial implications to creators are that if policymakers deem that transaction costs of maintaining data licensees are too high to generate any social value and a free-riding reuser will cause market failure, they will make reuse illegal all together. However, this scenario is highly unlikely. As we have shown, $\hat{\alpha} \leq 0.5$, which means that the transaction efficiency needs to be small (less than 50%) for this scenario to occur. Thus data creators should expect to be negotiating data licenses with many potential reusers of their data when a new data reuse law is introduced.

Theorem 2 (Region of Low Differentiation and High Cost: A$_1$*). When* $t \leq v/2$ *and* $(1+\alpha)t/2 < F \leq v - t = \pi^m$, *it is socially beneficial to grant legal protection to database and let the creator be a monopoly in the market by disallowing the potential reuser to recreate the database.*

This is good news to the creators: a reuser will not be allowed to create a similar database by reusing the creator's database contents. An extreme case of such disallowed reuse is when the reuser database is a duplicate of the creator's database so that the consumers are indifferent between the two databases (*t*=0 in this case). HR 1858 of 1999 attempted to prohibit such reuse. The lower bound of *F* is obtained from $F>\pi^{\mathrm{dl}}$, a condition under which the creator cannot offset its fixed cost even with the highest possible license fee from the reuser.

Theorem 3 (Region of Low differentiation and Moderate Cost: A$_2$*). When* $t \leq v/2$ *and* $t/2 \leq F < (1+\alpha)t/2$, *it is socially beneficial to grant legal protection to database. The creator is not willing to license its database to the reuser, but it is socially beneficial to require a compulsory license so long as* $\alpha > \hat{\alpha} = 0.5$.

The theorem shows the necessity of a compulsory license provision, which is missing in the EU Database Directive and HR 3261. A compulsory license will ensure that valued-added data reuse is maximally allowed without causing market failure. Under the conditions specified in the theorem, the creator must allow the reuser to use the data, and the reuser must pay a license fee to the creator.

When compulsory license provision applies, the creator should focus on devising licensing schemes to maximize its profit. Because the creator and the reuser will divide a “pie” of duopoly profits ($2\pi^d = t$), a bigger pie (i.e., when t is larger) can benefit both parties. With our model setup, t is larger when the two databases are more differentiated. In the process of developing the licensing terms, the creator can influence the reuser to reuse the data with as much differentiation as possible; at the same time, the creator can adjust its own database characteristics to be as different from those of the reuser as possible.

If through licensing the creator can make a profit bigger than monopoly profit, the creator will prefer having a fee-paying reuser to being a monopoly. We call such licenses *voluntary* licenses to distinguish it from *compulsory* licenses. The following theorem explores the conditions and social welfare implications of the two forms of licensing.

Theorem 4 (Region of Moderate Differentiation and Moderate Cost: A$_3$*). When* ${}^{v}\!/_{2} < t \le v$ *and* $\pi^d < F \le \min\{(1+\alpha)\pi^d, {}^{v^2}\!/_{4t} = \pi^m\}$*, it is socially beneficial to grant legal protection to database. The creator is willing to license its database if* $\alpha \ge {}^{(\pi^m - \pi^d)}\!/_{\pi^d} = \tilde{\alpha}$*. Within the range of differentiation,* $\tilde{\alpha}$ *can be less than or greater than* $\hat{\alpha}$*. It is socially beneficial to enforce compulsory licensing when* $\hat{\alpha} < \alpha < \tilde{\alpha}$*, to disallow voluntary licensing when* $\tilde{\alpha} < \alpha < \hat{\alpha}$*, , and to allow voluntary licensing when* $\hat{\alpha} < \tilde{\alpha} < \alpha$.

Theorem 4 identifies conditions under which compulsory license is needed or voluntary license should be disallowed to enhance social welfare. The conditions depend on transaction efficiency in relation to the two critical values, $\hat{\alpha}$ and $\tilde{\alpha}$, which are presented graphically in Figure 3.

<< Figure 3 >>

In Figure 3, we see in most cases $\tilde{\alpha} > \hat{\alpha}$, meaning that generally transaction efficiency requirement is higher for voluntary licensing than for compulsory licensing. There are different socially beneficial policy choices as indicated in different regions in the figure. Note there is a special case where the creator wants to license but it is socially wasteful to license and the creator should be a monopoly.

The managerial implications to the firms are essentially the same as those for Theorem 2. The main difference here is that the creator may be willing to license its data.

Theorem 5 (Region of Moderate to High Differentiation and High Cost: A$_4$). When $v/2 < t$ *and* $v^2/4t = \pi^m < F \le (1+\alpha)\pi^d$ *, the creator will not create the database, not because of the threat of free-riding, but because of the high cost. The databases can only be jointly developed by the creator and the reuser.*

Theorem 6 (Region of Low Cost Databases: Left Portion of A$_5$) It is socially beneficial not to grant legal protection to database when $F \le \pi^d \le v/3$ *and* $\alpha < 1$.

When the fixed cost is less than the duopoly profit, free-riding will not cause market failure. When transaction cost is greater than 0 (i.e., $\alpha < 1$), social welfare with a free-riding reuser is higher than having a fee-paying reuser (due to transaction cost of licensing) or having the creator as a monopoly (Corollary 2).

Most enacted and proposed database protection bills grant legal protection only to databases that require substantial expenditure to create and maintain. Theorem 6 shows that such provisions are necessary for enhancing social welfare. We have been using the magnitudes of fixed cost to determine the socially beneficial policy choices. These magnitudes are not measured in absolute dollar amount, rather, they are relative to the market value of the database as explicitly shown in Theorem 6. Thus, we should not specify an absolute dollar amount threshold for a database to qualify for legal protection.

Theorem 7 (Region of Highly Differentiated Databases: Right Portion of A_5). When $t > v$ and $\alpha < 1$, it is socially beneficial NOT to grant legal protection to database.

When $t > v$, $\pi^d = \pi^m$, thus free-riding of the reuser has no impact to the creator. When $\alpha < 1$, it is socially better not to collect a license fee to avoid transaction cost.

3.4 Overinvestment Distortion

From Theorem 6, it seems desirable to set a fixed cost threshold $\hat{F}$ equal to duopoly profit (i.e., $\hat{F} = \pi^d \le v/3$). As we will see next, this can induce excessive investment (or overstatement of cost) when an efficient firm can create the database at a cost (or the true cost is) slightly lower than $\hat{F}$. In our model setup, we treat *F* as sunk cost, thus only overstatement of cost is possible. The factors that induce overstatement of cost and overinvestment are the same. Their main differences are in the gains to the perpetrating firm and the welfare losses to the society (e.g., excessive investment leads to more welfare loss than overstatement). In the following analysis, we will focus on the factors that induce these distortions, and will use overinvestment to refer to either distortion.

Theorem 8. (Overinvestment Distortion). Suppose $\hat{F} = \pi^d$, when $\alpha > \hat{\alpha}$ and $t \le v$, a creator who can produces the database efficiently at a fixed cost $F \in (\underline{F}, \hat{F})$ has incentives to overinvest to qualify for legal protection. The value of $\underline{F}$ depends on α and t. The creator may aggressively overinvest $\overline{F} = \frac{(1+\alpha)t}{2}$ to become a monopoly when (1) $t \le \frac{2v}{3+2\alpha}$ and $\underline{F} = \max\{(4+\alpha)t/2 - v, 0\}$; or (2) $\frac{v}{2} < t \le \frac{v}{\sqrt{2(1+2\alpha)}}$ and $\underline{F} = \max\{\frac{2(2+\alpha)t^2 - v^2}{4t}, 0\}$. The creator may moderately overinvest $\widetilde{F} = \frac{t}{2}$ to be-

come eligible for licensing fee when $\frac{2v}{3+2\alpha} < t \le \frac{v}{2}$ *or* $\frac{v}{\sqrt{2(1+2\alpha)}} < t \le \frac{2v}{3}$*, in both cases* $\underline{F} = {}^{(1-\alpha)t}\!/_{2}$*. The creator may overinvest* $\tilde{F} = \frac{2v-t}{4}$ *also to qualify for receiving licensing fee when* $\frac{2v}{3} < t \le v$*,* $\alpha > \max(\frac{(v-t)^2}{2vt-t^2}, \hat{\alpha})$*, and* $\underline{F} = (1-\alpha)\frac{2v-t}{4}$.

Corollary 3. Overinvestment can also occur even if the creator already qualifies for protection but is subject to compulsory licensing as specified for the region of low differentiation and moderate cost (Theorem 3). Specifically, the creator overinvests $\overline{F} = {}^{(1+\alpha)t}\!/_{2}$ *when* $\max\{(2+\alpha)t - v, \frac{t}{2}\} = \underline{F} < F < {}^{(1+\alpha)t}\!/_{2}$, $t \le {}^{v}\!/_{2}$*, and* $\alpha > \hat{\alpha} = 0.5$.

Theorem 8 shows that when *t* and *F* fall in the low cost database region (A_5), the unprotected database creator may spend more at $\tilde{F}$ to move to the regions of moderate cost with low differentiation (A_2) or moderate differentiation (A_3) to qualify for protection with a fee-paying reuser. The unprotected creator may even spend at $\overline{F}$ to move to the region of low differentiation and high cost (A_1) to qualify for the legal monopoly status. Corollary 3 shows that a creator with *t* and *F* in A_2 wants to move to A_1 by spending more. These distortions benefit the creator but are socially wasteful.

4 Discussion

4.1 Summary of Findings

Using an extended spatial competition model, we have shown that depending on the condition, the reuser can be a free-rider or a fee-paying data reuser, or reuse is disallowed. A unique feature of the model is the use of a transaction efficiency coefficient to explicitly consider the inefficiencies of licensing and policy administration. This is an improvement over previous approaches that ignore this factor.

The model also allows us to clarify, and provide an economic interpretation of, two notions in the existing EU law and U.S. proposals. The “substantial expenditure” requirement is not clearly defined in the EU Database Directive and the current U.S. proposals. We can see from this model that it should not be an absolute value; rather, it should be the fixed cost relative to the overall market value of the database product. The minimal cost for qualification also depends on the degree of differentiation of the reuser database. Another notion is the reduction of creator incentives by the free-riding of reusers. HR 3261 regards reduced revenue as an injury which in turn reduces the incentives of creating the database, but it is ambiguous about the threshold that triggers legal action. Our model facilitates the discussion and determination of this threshold. For example, if the goal of the law is about fairness to the creator, any revenue reduction due to competition of a free-riding reuser is an offence and should be avoided. If, on the other hand, the law is about social welfare maximization, the creator’s net profit should be used to assess if legal intervention is necessary. We adopted the latter approach in our analysis.

With this model, we are able to identify socially beneficial policy choices under various conditions that are determined by the magnitude of fixed cost of database creation, the degree of differentiation between the reuser database and the creator database, and the transaction efficiency. Roughly speaking, under the assumptions of this model, no protection should be given if the database can be created with trivial expenditure or the reuser database is highly differentiated. When legal protection is granted, it may take various forms (e.g., no reuse with the creator being a legal monopoly, reuse with compulsory license, and discouragement of voluntary licensing). Reuse should be disallowed if the reuser database is a close substitute of the creator database and the cost of creating the database is high. In other words, a legal monopoly is socially desirable in this case. In the other cases, the transaction efficiency plays an important role of determining if compulsory licensing is required, or if licensing is beneficial to the creator but wasteful to the society, thus voluntary licensing should be discouraged.

There are two reasons why allowing free data reuse under certain circumstances can be social welfare-enhancing. First, technology has been such that the fixed cost incurred by the reuser is negligible compared with that incurred by the database creator. Thus, the reuser database is a "free" product to the society and social welfare is generally higher when there are two databases. Similar results exist in other intellectual property studies where the costs of producing copies are negligible. For example, Yoon [62] finds that depending on cost distribution, no copyright protection can be socially beneficial. In the presence of demand network externalities, Takeyama [58] finds that unauthorized reproduction of any intellectual property is Pareto improving, that is, both consumers and the infringed producer, thus the society as a whole, benefit from unauthorized reproduction. Second, expenditure on preventing reuse can be socially wasteful when reuse does not cause market failure. We informally discussed the social welfare effect of investment that raises the reuser cost; similarly, the expenditure on monitoring data reuse is also wasteful. This is also true in copyright enforcement; see Chen and Png [11] for their discussion on the adverse effect of anti-piracy investment.

We also discover the possibility of overinvestment distortions when a minimal cost is set to qualify for legal protection. Creators with unqualified databases have incentives to over-spend in database creation to become qualified; creators who are asked to license their databases may want to invest excessively to become a legal monopoly. These distortions occur only when the reuser database has little or moderate differentiation with the creator database. We have not yet found a mechanism to eliminate the distortions at this point. Thus, the court is expected to scrutinize cases carefully to identify and penalize those who purposefully over-spend in database creation.

The policy choices will impact both database creators and data reusers. The two factors important to managers are: (1) the degree of differentiation between the databases; and (2) the transaction efficiency when licensing is required.

Both the creator and the reuser can benefit from increasing degrees of differentiation. The reuser should avoid duplicating the creator's contents for purposes similar to those of the creators. Such reuse will be prohibited. Instead, the reuser should leverage its special skills to create its database with features and purposes as different from the creator's as possible. The creator should actively explore other possible ways of using its database contents, which when successful can preemptively place multiple differentiated databases in the product space to deter entry of reusers. In summary, both the creator and the reuser should focus on innovation to develop more varieties of databases to soften competition and better serve the diverse needs of the consumers. Innovation can go beyond just database products. For example, in the online retailing setting, a vendor (whose price data has been reused by competitors) can offer other value-added services (e.g., product comparison) to differentiate it from competing vendors [10]; a vendor who reuses competitors' pricing and inventory data can use dynamic pricing to potentially increase short-term profit [14].

The creator should also monitor the reuses of its contents. The cost involved is reflected in the transaction efficiency coefficient. Server log files and certain analytical tools can be used to automate usage monitoring and reduce the cost. In addition, there is ongoing research [60] to develop policy-aware architecture and related technologies with which the cost of monitoring reusers and enforcing them to comply with specified policy parameters can be further reduced.

4.2 Applications

The model and the results provide a useful reference frame for discussing database protection policies and make an initial step towards identifying the right balance needed. We will illustrate the applications of the model and the analytical results by commenting on the recent U.S. proposals and several cases mentioned earlier.

HR 3261 of 2003 is generally in line with the results here. The scope of the proposal is confined by the term of "functional equivalent", which means that the proposal concerns reuse that produces a close substitute of the creator's database. Although it is a bit vague, it does intend to protect non-trivial databases only. However, HR 3261 is obviously crude and lacks important compulsory licensing provisions. Our model has roughly three levels in both the degree of differentiation and the cost of database creation. This allows for fine tuning of policy choices. HR 3261 takes a more or less binary approach. It thus misses several opportunities of social welfare maximization. Without compulsory license, sole source creators can become a lawful monopoly under the proposal, which is harmful to society. These shortcomings will likely raise constitutionality concerns.

In HR 3872 of 2004, injury alone is not an offense that triggers government intervention, which only comes in when the injury reaches the point where the creator would not create the database or maintain its quality. This criterion corresponds to the zero net profit threshold used in our analysis.

HR 1858 prevents duplication of a database, which is an extreme case of no differentiation. With no differentiation, the reuser (now a duplicator) adds little value to the society and the creator will not create the database even at a moderate creation fixed cost, thus database duplication should be disallowed. The proposal also clearly specifies a compulsory licensing requirement for sole source creators. Although HR 1858 would very likely pass constitutional scrutiny, it has certain drawbacks (e.g., its scope is deemed to be too narrow because it only covers one extreme case of data reuse).

Below we further illustrate the applications of the model by commenting on three previously mentioned cases.

eBay v. Bidder's Edge. In the eBay case, the computing resource[22] is not the subject matter of such a policy, which concerns the data, not the resources that deliver the data. As discussed later, our model can be extended to model the impact of reuse on the creator's cost. Let us focus on the data for now. According to the model, we need to at least examine the degree of differentiation of the database developed by the reuser Bidder's Edge. In terms of searching of bidding data, the reuser database had a much broader coverage; thus, there was competition from the reuser database. In terms of functionality, eBay's database allowed one to buy and sell items; the reuser database did not provide any actual auction service. In addition, the eBay database also contained the actual transaction data (price and quantity of items sold), which was not available in the database of Bidder's Edge. Thus the two databases exhibited significant differentiation. Searching alone does not, in general, reduce eBay's revenue from its auction service. In addition, conducting a search and participating in an actual auction involve two different markets. If we subscribe to the spin-off theory [26], the eBay database will not meet the cost criterion[23]. Therefore, free reuse by Bidder's Edge should be allowed under our model (Theorems 6 and 7).

[22] Use of eBay's computing resources was eBay's argument against Bidder's Edge in its claim of Trespass to Chattels.

[23] A number of E.U. cases (http://www.ivir.nl/files/database/index.html) support the spin-off theory. For example, in the appeal case between *Zoekallehuizen.nl* (a site that searches and lists houses for sale in the Netherlands) and *NVM* (the Dutch Association of Real Estate Agents, and two real estate agents), the court decided the real estate agents' data is not protected by database right because the agents' websites did not show substantial investment. The data in their sites are the results of their investment in their main activities which are not related to database creation.

mySimon v. Priceman. In the mySimon case, the reuser database was a superset of the creator's. Both were in the searchable comparison shopping database market. Free-riding by the reuser would certainly reduce creator's revenue. If the reduction reaches a level that the creator cannot make a positive profit, which is likely in this case, then the reuser should be asked to pay a fee for using the data (Theorems 3 or 4).

BHB v. William Hill. The court referred the questions raised by William Hill in its appeal to the European Court of Justice (ECJ). The ECJ issued its ruling[24] in late 2004 which favored William Hill. Our model can be used to interpret this important ruling (the first one on the EU Database Directive). The ECJ made a distinction between making a database by creating data and making a database by gathering existing data. The ECJ ruled that the investment in the obtaining and verification of the contents of a database protected by the Database directive "does not cover the resources used for the creating of materials which make up the contents of a database", nor does it cover the "resources used for verification during the stage of creation of materials which are subsequently collected in a database". As the horseracing authority, BHB created the "official" horserace list. Even though it spends approximately £4 million to maintain its database, most of this cost belongs to the two categories not covered by the Directive. In the model, the cost of creating the database is an important factor. We identified that a creator may overinvest or overstate its cost in database creation, thus cost statement should be carefully scrutinized. This ruling provides a guideline for determining the cost of creating the database.

In the ruling, the ECJ also clarified that the prohibited extraction and reuse refers to "unauthorised actions for the purpose of reconstituting, through the cumulative effect of acts of extraction, the whole or a substantial part of the contents of a database" and "thereby seriously

[24] See ECJ case C-203/02, retrievable using the case number at http://curia.europa.eu/jurisp/cgi-bin/form.pl?lang=en. Other similar cases include C-338/02, C-444/02, and C-46/02.

prejudice the investment by the maker". This ruling can also be explained by our model. When a reuser reconstitutes the creator's database contents, the reuser database will not be sufficiently differentiated from the creator's. In this case, our model shows that the creator will not have enough revenue to recoup the cost incurred in creating the database, and such reuse should be prohibited.

4.3 Limitations and Future Work

There is a pressing need for database legislation to balance between protecting incentives of database creation and preserving sufficient access to data for value-creating activities. With an extended spatial competition model, we are able to identify a range of conditions and determine different policy choices under these conditions. The model and the analytic results provide a useful framework for discussing and understanding the economic factors that need to be considered in database policy formulation.

To make the model more useful, the key parameters (t, α, and F relative to overall market value of database) of the model need to be operationalized so that their values can be appropriately assessed in practice. Future research needs to develop a systematic assessment method that considers all relevant factors. The method should address specific situations, such as, whether t would be low or high, (1) if someone reuses all the data from the creator database but adds a significant amount of additional data to make the database useful for a great many more purposes; or (2) if someone reuses only the most valuable and costly 10% of the creator database.

There are a number of other limitations in our analysis. We have focused on financial interests in database contents without considering other factors concerning societal values of data and data reuse. Our economic model considers the competition between the creator and reuser databases. The model does not capture other effects of data reuse, such as network externalities of database products (i.e., the creator database becomes more or less valuable when there are more consumers using the reuser database, or vice versa). In addition, the model also ignores factors that are specific to the kind of data being reused, for example, privacy concerns when the reused data is about personal information (see [21] for possible ways of overcoming privacy concerns), increased price competition concerns when the reuser compiles a price comparison database (see [2] for an analysis of the effects of reduced search cost, [32] for pricing strategies in the presence of price comparison, [4] for observations of price change frequencies at two competing online bookstores), and the reuse may affect the cost of the creator (e.g., Bidder's Edge's repeated queries used eBay's computing and network resources). It is possible to extend the model to include a cost term to the creator's profit function and let the term be dependent on certain characteristics of reuse. Another limitation is that we have assumed that a consumer consumes maximally one database. This assumption needs to be relaxed in future research because it is possible that a consumer consumes both databases.

In addition, our current analysis is based on a horizontal differentiation model; in the future, we plan to examine data reuse that is vertically differentiated (e.g., the reuser may produce a database of inferior or superior quality to target a different market). We also need to look at dynamic characteristics. As stressed in [33], intellectual property is also the input to intellectual property creation. With strong protection for database contents, the cost of database creation will likely rise. This effect can be modeled using the cumulative innovation theory [54] from the patent literature. The theory has been used to informally explain the importance of ensuring adequate access to data for knowledge and value creation [41]. Furthermore, data reusers are often aggregators that draw data from multiple sources. The implications of the increased database protection can be examined using an emerging theory of *anticommons*, which is the opposite of the *commons* problem where free-riding and overuse of a public good causes its depletion (see [19, 20] for discussions about the commons problem of the Internet). With anti-commons, there are multiple right holders to a resource. As is shown in [6], when there exist multiple rights to exclude, a valuable resource will be underutilized due to increased prices. Databases that hold factual information as a whole (e.g., the Web) are a valuable resource, thus, providing more than necessary protection to databases is analogous to anticommons and will lead to underutilization of protected databases. Lastly, many online databases have characteristics of two-sided markets [44, 47], such as databases that target both information seekers as well as advertisers. Therefore, the modeling techniques for two-sided markets and their interlinked network effects are worth exploring to derive new insights for policy formulation purposes.

Although there are areas for future research, as identified above, the current model captures many of the major issues in database legislation. We believe it is an important step in formalizing the discussion and formulation of a socially beneficial data reuse policy.

Acknowledgements: The authors thank the editor and reviewers for their helpful comments. Especially we thank one of the reviewers, whose extensive comments have helped us to improve the clarity and organization of the paper. We also thank Jacqueline Lipton for her inputs on legal issues, Lorin Hitt for several useful suggestions, including the one about non-separating equilibrium of temporal versioning, and several colleagues who read the earlier versions of the paper and provided us with helpful feedback.

References

1. Aron, R., and Clemons, E.K. Achieving the Optimal Balance Between Investment in Quality and Investment in Self-Promotion for Information Products. *Journal of Management Information Systems*, 18, 2 (Fall 2001), 65-88.

2. Bakos, Y. Reducing buyer search costs: Implications for electronic marketplaces. *Management Science. 43*, 12 (Dec. 1997), 1676-1692.

3. Bakos, Y., Brynjolfsson, E., and Lichtman, D. Shared Information Goods. *Journal of Law and Economics,* 42, 1 (1999) 117-115.

4. Bergen, M.E., Kauffman, R.J., and Lee, D. Beyond the Hype of Frictionless Markets: Evidence of Heterogeneity in Price Rigidity on the Internet. *Journal of Management Information Systems*, 22, 2 (2005), 57-89.

5. Besen, S.M., and Raskind, L.J. An Introduction to the Law and Economics of Intellectual Property. *Journal of Economic Perspectives,* 5, 1 (1991), 3-27.

6. Buchanan, J. M., and Yoon, Y. J. Symmetric Tragedies: Commons and Anticommons. *Journal of Law and Economics,* 43, 1 (2000), 1-43.

7. Burk, D.L. The Trouble with Trespass. *Journal of Small & Emerging Business Law*, 4, 1 (2000), 27-56.

8. Chang, E.W. Bidding on Trespass: eBay, Inc. v. Bidder's Edge, Inc. and the Abuse of Trespass Theory in Cyberspace-Law. *AIPLA Quarterly Journal,* 29, 4 (2001), 455-470.

9. Chang, C.H., Kaye, M., Girgis, M.R., and Shaalan, K.F. A Survey of Web Information Extraction System. *IEEE Transactions on Knowledge and Data Engineering*, 18, 10 (2006), 1411-1428.

10. Chellappa, R. K., and Kumar, K.R. Examining the Role of "Free" Product-Augmenting Online Services in Pricing and Customer Retention Strategies. *Journal of Management Information Systems*, 22, 1 (2005), 355-377.

11. Chen, Y., and Png, I. Information Goods Pricing and Copyright Enforcement: Welfare Analysis. *Information Systems Research*, 14, 1 (2003), 107-123.

12. Colsten, C. Sui Generis Database Right: Ripe for Review? *The Journal of Information, Law and Technology*. 3, (2001).

13. Commission of the European Communities (CEC). First Evaluation of Directive 96/9/EC on the Legal Protection of Databases. 12 December 2005, Brussels.

14. Dewan, R. M., Freimer, M.L., and Jiang, Y. A Temporary Monopolist: Taking Advantage of Information Transparency on the Web. *Journal of Management Information Systems*, 24, 2 (2007), 167-194.

15. Firat, A., Madnick, S.E., and Siegel, M.D. The Cameleon Web Wrapper Engine. *Workshop on Technologies for E-Services (TES'00),* Cairo, Egypt, (2000).

16. Gibson, J. Re-reifying Data. *Notre Dame Law Review,* 80, 1 (2004), 163-242.

17. Goh, C. H., Bressan, S., Madnick, S., and Siegel, M. Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. *ACM TOIS,* 17, 3 (1999), 270-293.

18. Grove, J. Wanted: Public Policies That Foster Creation of Knowledge. *Communications of the ACM* 47, 5 (2004), 23-25

19. Gupta, A., Jukic, B., Parameswaran, M., Stahl, D.O., and Whinston, A.B. "Streamlining the Digital Economy: How to Avert a Tragedy of the Commons," *IEEE Internet Computing*, December 1997, 38-46.

20. Gupta, A., Stahl, D.O., and Whinston, A.B. The Internet: A Future Tragedy of the Commons? in *Computational Approaches to Economic Problems*, H. Amman, B. Rustem, and A. B. Whinston eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997, 347 – 361.

21. Hann, I. H., Hui, K.L., Lee, S.Y., and Png, I.P.L. Overcoming Online Information Privacy Concerns: An Information-Processing Theory Approach. *Journal of Management Information Systems*, 24, 2 (2007), 13-42.

22. Heald, P.J. The Extraction/Duplication Dichotomy: Constitutional Line Drawing in the Database Debate. *Ohio State Law Journal*, 62, 2 (2001), 933-944.

23. Heller, M.A. 1998. The Tragedy of the Anticommons: Property in the Transition from Marx to Markets. *Harvard Law Review*, 111, 3, 621-688.

24. Hotelling, H. Stability in Competition. *Economic Journal*, 39, 153 (1929), 41-57.

25. Hunter, D. Cyberspace as Place and the Tragedy of the Digital Anticommons. *California Law Review,* 91, 2 (2003), 439-520.

26. Hugenholtz, P.B. Program Schedules, Event Data and Telephone Subscriber Listings under the Database Directive: The "Spin-Off" Doctrine in the Netherlands and elsewhere in Europe. 11th Annual Conference on International Law & Policy, New York, (2003).

27. Hugenholtz, P.B. The New Database Right: Early Case Law from Europe. Ninth Annual Conference on International IP Law & Policy, Fordham University School of Law, New York,

April 19-20, (2001).

28. Jhingran, A. Enterprise Information Mashups: Integrating Information, Simply. The 32nd VLDB Conference, Seoul, Korea (2006), 3-4.

29. Katz, M.L. and Shapiro, C. Network externalities, competition and compatibility. *American Economic Review*, 75, 3 (1985), 424-440.

30. Katz, M.L. and Shapiro, C. How to license intangible property. *Quarterly Journal of Economics*, 101, 3 (1986), 567-589.

31. Koboldt, C. The EU-Directive on the legal protection of databases and the incentives to update: An economic analysis. *International Review of Law and Economics*, 17, 1 (1997), 127-138.

32. Koças, C. A Model of Internet Pricing Under Price-Comparison Shopping. *International Journal of Electronic Commerce*, 10, 1 (2005), 111-134.

33. Landes, W.M., and Posner, R.A. *The Economic Structure of Intellectual Property Law*, Cambridge, MA: Belknap Press, 2003.

34. Lemley, M.A. Place and Cyberspace. *California Law Review*, 91, 2 (March 2003), 521-542.

35. Lessig, L. The New Chicago School. *Journal of Legal Studies*, 27, 2 (1998), 661-691.

36. Lessig, L. The Law of the Horse: What Cyberlaw Might Teach. *Harvard Law Review*, 113, 2 (1999), 501-549.

37. Lin, L., and Kulatilaka, N. Network Effects and Technology Licensing with Fixed Fee, Royalty, and Hybrid Contracts. *Journal of Management Information Systems*, 23, 2 (Fall 2006), 91-118.

38. Lipton, J. Private Rights and Public Policies: Reconceptualizing Property in Databases. *Berkeley Technology Law Journal*, 18, 3 (2003), 773-852.

39. Madnick, S.E., and Siegel, M.D. Seize the Opportunity: Exploiting Web Aggregation. *MISQ Executive*, 1, 1 (2002), 35-46.

40. Maurer, S.M., and Scotchmer, S. Database Protection: Is It Broken and Should We Fix it? *Science*, 284 (May 1999), 1129-1130.

41. Maurer, S.M. Intellectual Property Law and Policy Issues in Interdisciplinary and Intersectoral Data Applications. Data for Science and Society, National Research Council, (2001).

42. O'Rourke, M.A. Shaping Competition on the Internet: Who Owns Product and Pricing In-

formation? *Vanderbilt Law Review*, 53, 6 (2000), 1965-2006.

43. O'Rourke, M.A. Is Virtual Trespass an Apt Analogy? *Communications of the ACM*, 44, 2, 98-103.

44. Parker, G.G., and Van Alstyne, M. Two-Sided Network Effects: A Theory of Information Product Design. *Management Science*, 51, 10 (2005), 1494-1504.

45. Reichman, J.H., and Samuelson, P. Intellectual Property Rights in Data? *Vanderbilt Law Review*, 50, 1 (1997), 52-166.

46. Reichman, J.H., and Uhlir, P.F. Database Protection at the Crossroads: Recent Developments and Their Impact on Science and Technology. *Berkeley Technology Law Journal*, 14, (Spring 1999), 793-838.

47. Rochet, J.-C., and Tirole, J. Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1, 4 (2003), 990-1029.

48. Ruse, H.G. Electronic Agents and the Legal Protection of Non-creative Databases. *International Journal of Law and Information Technology*, 3, 3 (2001), 295-326.

49. Salop, S.C. Monopolistic Competition with Outside Goods. *The Bell Journal of Economics*, 10, 1 (1979), 141-156.

50. Salop, S.C., and Scheffman, D.T. Cost-Raising Strategies. *J. Industrial Economics*, 36, 1 (1987), 19-34.

51. Samuelson, P. Legal Protection of Database Contents. *Communications of the ACM*, 39, 12 (1996), 17-23.

52. Sanks, T.M. Database Protection: National and International Attempts to Provide Legal Protection for Databases. *Florida State University Law Review,* 25, 4 (Summer 1998), 991-1016.

53. Schmalensee, R., and Thisse, J.F. Perceptual Maps and the Optimal Location of New Products: An Integrative Essay. *International Journal of Research in Marketing*, 5, 4 (1988), 225-249.

54. Scotchmer, S. Standing on the Shoulders of Giants: Cumulative Research and the Patent Law. *Journal of Economic Perspectives*, 5, 1 (1991), 29-41.

55. Shapiro, C., and Varian, H.R. *Information Rules: A Strategic Guide to the Network Economy*. Cambridge, MA: Harvard Business School Press, 1998.

56. Sheth, A,P., Gomadam, K., and Lathem, J. SA-REST: Semantically Interoperable and Easier-to-Use Services and Mashups. *IEEE Internet Computing*, 11, 6 (2007), 91-94.

57. Tabuchi, H. International Protection of Non-Original Databases: Studies on the Economic Impact of the Intellectual Property Protection of Non-Original Databases. CODATA, Montreal, Canada (2002).

58. Takeyama, L.N. The Welfare Implications of Unauthorized Reproduction of Intellectual Property in the Presence of Demand Network Externalities. *The Journal of Industrial Economics*, 22, 2 (1994), 155-166.

59. Tirole, J. *The Theory of Industrial Organization*. Cambridge, MA: The MIT Press, 1988.

60. Weitzner, D.J., Hendler, J., Berners-Lee, T., Connolly, D. Creating a Policy-Aware Web: Discretionary, Rule-based Access for the World Wide Web. In Web and Information Security, Ferrari, E. and Thuraisingham, B. (Eds), IRM Press, 2006.

61. West, L.A. Private Markets for Public Goods: Pricing Strategies of Online Database Vendors. *Journal of Management Information Systems*, 17, 1 (Summer 2000), 59-85.

62. Yoon, K. The Optimal Level of Copyright Protection. *Information Economics and Policy*, 14, 3 (2002), 327-348.

63. Zhu, H., Madnick, S.E. Enabling Global Price Comparison through Semantic Integration of Web Data. *International Journal of Electronic Business*. Forthcoming.

Appendices

A1. Proof of Lemma 1 (Duopoly and Monopoly Market Coverage).

Duopoly, little differentiation (*t≤2v/3*). In the case of full market coverage, there exist customers whose ideal database is located at $\tilde{x} \in [0,\ 1]$, such that $u_{\tilde{x},0} = u_{\tilde{x},1} \geq 0$. That is, these customers are indifferent between the creator and the reuser database. The demand for database 0 is $\tilde{x}$ and the demand for database 1 is $(1-\tilde{x})$. Maximizing profit for both firms with respect to respective database prices p_0 and p_1, we obtain $p_0^* = p_1^* = t$ and $\pi_0^* = \pi_1^* = \pi^d = t/2$. Positive utility constraints at $\tilde{x}$ require *t≤2v/3.* By symmetry, the social welfare is $2\int_0^{0.5}(v-tx)dx = v - t/4$.

Duopoly, moderate differentiation (*2v/3<t≤v*). This is the case that requires careful examination of corner solutions. To see that $p_0^* = p_1^* = v - t/2$ is the equilibrium, we show that given $p_1 = v - t/2$, the profit-maximizing price for the creator is also $v - t/2$, and vice versa. When $p_1 = v - t/2$, $u_{1/2,1} = 0$. If the creator charges the same price, then each firm takes up one half of the market and makes a gross profit of $v/2 - t/4$. We only need to show that any deviation by the creator yields a lower profit. For any infinitesimal positive value $\delta \in R^+$, let us first suppose the creator wants to capture more than a half of the market by choosing a lower price $p_0 = p_1 - \delta$. With $u_{\tilde{x},0} = u_{\tilde{x},1}$ we can find the creator's demand $\tilde{x} = {(t+\delta)}/{2t}$. Therefore, the creator's profit is $\pi_0 = p_0\tilde{x} = (p_1 - \delta){(t-\delta)}/{2t}$. It is easily shown that $\frac{\partial \pi_0}{\partial \delta} = \frac{v - 3t/2}{2t} - \frac{\delta}{4t} < 0$ when $2v/3 < t$ because both terms are negative. Now let us suppose that the creator wants to deviate by charging a higher price $p_0 = p_1 + \delta$; as a result, it will cover less than a half of the market. We can derive $\frac{\partial \pi_0}{\partial \delta} = \frac{t-v}{t} - \frac{2\delta}{t} < 0$ because *t≤v*.

Duopoly, high differentiation (*v<t*). Each firm's demand is determined by the marginal consumers whose utility of purchasing a database is 0. Let us consider the creator, whose marginal consumers are located at $\tilde{x} = {(v-p_0)}/{t}$. Maximizing profit yields $p_0^* = v/2$. Therefore, $\tilde{x} = {(v - \frac{v}{2})}/{t} = v/2t < 1/2$, and $\pi_0^* = v^2/4t$. By symmetry we obtain the reuser's price and profit. Social welfare is $2\int_0^{v/2t}(v - tx)dx = 3v^2/4t$.

Monopoly, moderate preference heterogeneity (*t≤v/2*). Similar to the moderately differentiated duopoly case, it is better for the monopoly to cover the entire market. Letting $u_{1,0} = 0$, we derive the price. Demand is 1. It is straightforward to derive social welfare.

Monopoly, high preference heterogeneity (*v/2<t*). Similar to the highly differentiated duopoly case, it is better for the monopoly to cover a fraction of the market. Straightforward optimization yields the results.

A2. Proof of Theorem 2 (Region of Low Differentiation and High Cost).

Suppose there is a reuser database, the creator's profit would be either *t/2* (see Lemma 1) if the reuser is a free-rider, or ${}^{(1+\alpha)t}\!/_{2}$ if the reuser pays a fee equal to its profit. Given the condition ${}^{(1+\alpha)t}\!/_{2} < F \le v - t = \pi^m$, we know the creator would make a negative net profit in both cases, thus the database will not be created and social welfare is 0. Without the reuser database, the creator earns a monopoly profit $\pi^m = v - t$, which has been assumed to be greater than or equal to *F*; net social welfare is $SW^m - F = v - {}^{t}\!/_{2} - F \ge {}^{t}\!/_{2} > 0$. Therefore, the creator should be a legal monopoly.

A3. Proof of Theorem 4 (Region of Moderate Differentiation and Moderate Cost).

Given the range of fixed cost *F*, we know that the creator can make a positive profit only if it receives a licensing fee or it is a monopoly. The creator incentives and policy choices depend on the value of *α*. When $\alpha \ge \tilde{\alpha} = {}^{(\pi^m - \pi^d)}\!/_{\pi^d}$ we derive $\pi^d + \alpha\pi^d \ge \pi^m$, where the creator's profit with a fee-paying reuser is greater than or equal to the monopoly profit. Therefore he creator will be willing to license its database (voluntary licensing). When $\hat{\alpha} < \alpha < \tilde{\alpha}$, it is straightforward to see that the creator's monopoly profit is greater than the profit with a fee-paying reuser. Thus the creator is not willing to license its data. Because $\hat{\alpha} < \alpha$, it is socially beneficial to have a fee-paying reuser (Theorem 1). Therefore, compulsory licensing is required. When $\tilde{\alpha} < \alpha < \hat{\alpha}$ the creator prefers to license its database but it is socially wasteful, thus licensing should be disallowed.

A4. Proof of Theorem 8

When $F < \hat{F}$ (i.e., F is in A$_5$ area in Figure 2), the reuser can legally free-ride, so the creator's net profit is $\pi^d - F$. The creator has incentive to overinvest to a level in A$_1$, A$_2$, or A$_3$ areas as long as it can make a higher net profit.

When $t \le \frac{v}{2}$, the creator has an incentive to overinvest to the minimal level of legal monopoly, π^{dl}, if the following conditions hold:

$$\begin{cases} \pi^m - \pi^{dl} \ge \pi^d - F & (1) \\ \pi^m - \pi^{dl} \ge \pi^{dl} - \hat{F} & (2) \end{cases}$$

where (1) is the condition under which being a lawful monopoly is better than having a free-rider; (2) ensures that being a lawful monopoly is better than having a fee-paying reuser. Solving (1) yields $F \ge \frac{(4+\alpha)t}{2} - v$, whose right-hand side can be greater than or less than 0; therefore, we have $\underline{F} = \max\{{}^{(4+\alpha)t}\!/_{2} - v, 0\}$. Solving (2) gives $t \le \frac{2v}{3+2\alpha}$. With the assumption $\alpha > \hat{\alpha} = 0.5$, we know that $\frac{2v}{3+2\alpha} < \frac{v}{2}$.

Similarly (when $t \le \frac{v}{2}$ also), the incentive-compatibility conditions for overinvesting to $\hat{F}$ to only qualify for receiving licensing fee are:

$$\begin{cases} \pi^{dl} - \hat{F} \ge \pi^d - F & (3) \\ \pi^m - \pi^{dl} < \pi^{dl} - \hat{F} & (4) \end{cases}$$

Here, (3) ensures having a fee-paying reuser is better than having a free-rider; (4) ensures having a fee-paying reuser is better than being a monopoly. Solving (3) gives $F \ge (1-\alpha)\pi^d = \underline{F}$, where $\pi^d = \frac{t}{2}$; solving (4) yields $t > \frac{2v}{3+2\alpha}$.

When $\frac{v}{2} < t \le \frac{2v}{3}$, these constraints can be solved by plugging in appropriate profit functions.

When $t > \frac{2v}{3}$, the monopoly profit is only slightly higher than the duopoly profit. When α is not too small, $\pi^{dl} > \pi^m$, which gives $\alpha > \frac{(v-t)^2}{2vt-t^2}$. From Theorem 1, $\hat{\alpha} = \max\{0, \frac{3v^2-4vt}{4vt-2t^2}\}$. When

$t > \frac{2v}{3}$, $\frac{(v-t)^2}{2vt-t^2}$ can be greater or less than $\hat{\alpha}$. Thus $\alpha > \max(\frac{(v-t)^2}{2vt-t^2}, \hat{\alpha})$ is necessary to be consistent with Theorem 4.

A5. Proof of Corollary 3

The creator overinvests at π^{dl} to qualify for being a monopoly if the net profit in this case is greater than duopoly profit with a fee-paying reuser. This is the case when $\pi^m - \pi^{dl} > \pi^{dl} - F$, which gives $F > \underline{F} = (2+\alpha)t - v$. The lower bound for α is necessary from Theorem 3.

Table 1. History of Database Protection Legislation

Year	Jurisdiction	Legislation	Outcome
1996	European Union	Database Directive. It grants database makers copyright protection for the creative selection and arrangement of the database. It also grants *sui generis* right to prevent unauthorized extraction and reutilization of the whole, a substantial part of, or systematic extraction of insubstantial part of, database contents.	Adopted
1996	USA	HR 3531: Database Investment and Intellectual Property Piracy Act. Similar to EU Database Directive.	Failed
1998	USA	HR 2652: Collections of Information Antipiracy Act. It offers the database creators criminal or civil remedies if the reuser causes or has the potential to cause harm to the creator.	Failed
1999	USA	HR 354: Collections of Information Antipiracy Act. Similar to HR 2652.	Failed
1999	USA	HR 1858: Consumer and Investor Access to Information Act. It disallows verbatim copying of a database.	Failed
2003	USA	HR3261: Database and Collection of Information Misappropriation Act. It disallows free-riders from creating functional equivalent databases to reduce the creator's revenue.	Failed
2004	USA	HR 3872: Consumer Access to Information Act. It prevents a free-rider from engaging in direct competition that threatens the existence or the quality of the creator database.	Failed

Table 2. Parameters and Symbols used in the Economic Model

Parameter/Symbol	**Definition**	**Comments**
x	Distance between ideal database and the creator's database in the product feature space	$x \in [0,1]$
v	Utility of an ideal database	$v > 0$
t	Preference mismatch cost per unit distance in database characteristics space	$t \geq 0$. Because the distance between the creator and reuser database is 1, t also indicates the differentiation between the two databases.
p	Database price	p_0 is creator database price, p_1 is reuser database price
π	Gross profit without accounting for fixed cost or licensing income/fee	See below for superscripts to indicate profits of different scenarios
F	Fixed cost incurred by the creator to create the database	
$\hat{F}$	A critical value of fixed cost	A policy may attempt to choose it as the minimal fixed cost to qualify for legal protection
$\underline{F}$	A critical value of fixed cost	The creator may choose to overinvest when fixed cost of efficient production is above this value
$\tilde{F}$	A critical value of fixed cost, above which the creator may qualify for protection with a fee-paying reuser	The creator may choose to overinvest above this level to gain legal protection
$\overline{F}$	A critical value of fixed cost, above which the creator may qualify for legal monopoly protection	The creator may choose to overinvest above this level to gain legal monopoly protection
α	Transaction efficiency coefficient	$\alpha \in [0,\ 1]$
$\hat{\alpha}$	A critical value of transaction efficiency	For licensing to be welfare-enhancing, α should be greater than $\hat{\alpha}$
$\tilde{\alpha}$	A critical value of transaction efficiency	When $\alpha > \tilde{\alpha}$, the creator may be willing to license its data
SW	Gross social welfare without accounting for fixed cost or transaction cost	It is the sum of firm profits and consumer surplus. See below for superscripts to indicate gross social welfare in different scenarios
Superscript		
*	Value maximizes gross profit	
m	The case of a monopoly	Reuser database does not exist
d	The case of duopoly with a free-riding reuser	There is a competing reuser database and the reuser does not pay a fee to the creator
dl	The case of duopoly with a fee-paying reuser	The reuser licenses data with a fee

Table 3. Price, Profit, and Social Welfare at Different Differentiation Levels

	t	**Best price**	**Maximum profit**	**Social welfare**
Duopoly	$t \le 2v/3$	$p_0^* = p_1^* = t$	$\pi_0^* = \pi_1^* = \pi^d = t/2$	$SW^d = v - t/4$
	$2v/3 < t \le v$	$p_0^* = p_1^* = v - t/2$	$\pi_0^* = \pi_1^* = \pi^d = v/2 - t/4$	$SW^d = v - t/4$
	$v < t$	$p_0^* = p_1^* = v/2$	$\pi_0^* = \pi_1^* = \pi^d = v^2/4t$	$SW^d = 3v^2/4t$
Monopoly	$t \le v/2$	$p^m = v - t$	$\pi^m = v - t$	$SW^m = v - t/2$
	$v/2 < t$	$p^m = v/2$	$\pi^m = v^2/4t$	$SW^m = 3v^2/8t$

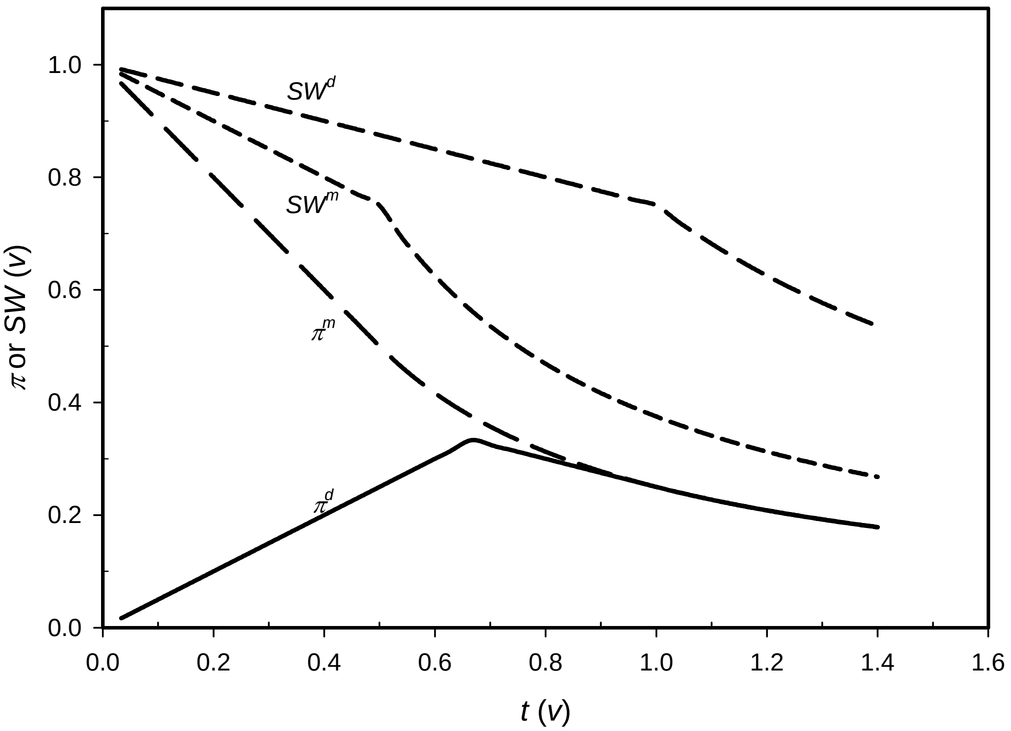

Figure 1. Change of profit (π^d, π^m) and social welfare (SW^d, SW^m) with *t*, all measured as factors of the value, *v*, of an ideal database. Unit mismatch cost *t* also indicates the degree of differentiation between creator and reuser databases.

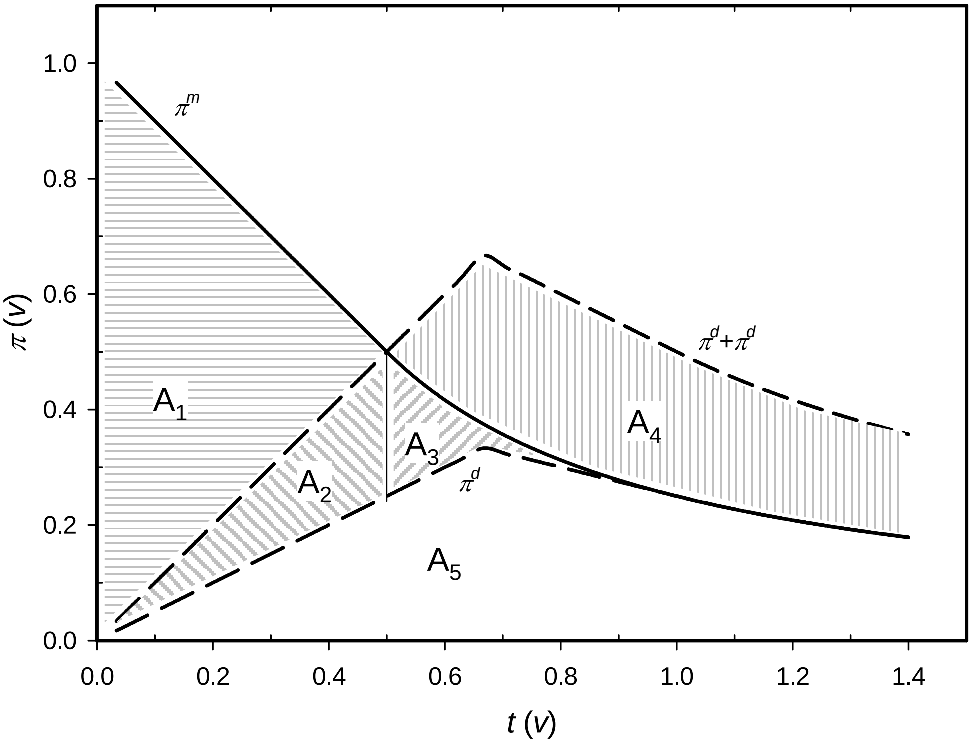

Figure 2. Changes of profits with *t*. Regions of different policy choices are marked by A_1 to A_5.

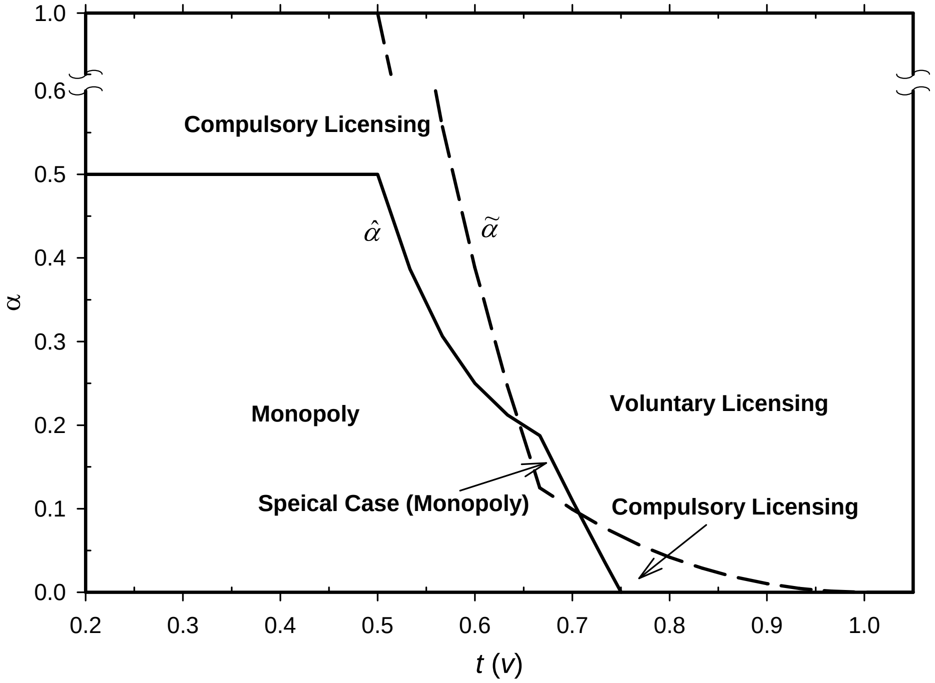

Figure 3. Changes of critical values of transaction efficiency with *t*. Different policy choices are indicated in different regions.

Biography of Authors

HONGWEI ZHU is an Assistant Professor of Information Technology at the College of Business and Public Administration, Old Dominion University. He holds a Ph.D. in Technology, Management and Policy from MIT. His research interests include data integration and reuse technologies, data quality management, data mining, information policy analysis, and information economics. His research has appeared in such journals as Data and Knowledge Engineering, Communications of the ACM, and International Journal of Electronic Business.

STUART E. MADNICK is the John Norris Maguire Professor of Information Technology, Sloan School of Management and Professor of Engineering Systems, School of Engineering at the Massachusetts Institute of Technology. He has been a faculty member at MIT since 1972. He has served as the head of MIT's Information Technologies Group for more than twenty years. He has also been a member of MIT's Laboratory for Computer Science, International Financial Services Research Center, and Center for Information Systems Research. Dr. Madnick is the author or co-author of over 250 books, articles, or reports including the classic textbook, *Operating Systems*, and the book, *The Dynamics of Software Development*. His current research interests include connectivity among disparate distributed information systems, database technology, software project management, and the strategic use of information technology. He is presently co-Director of the PROductivity From Information Technology Initiative and co-Heads the Total Data Quality Management research program. He has been active in industry, as a key designer and developer of projects such as IBM's VM/370 operating system and Lockheed's DIALOG information retrieval system. He has served as a consultant to corporations, such as IBM, AT&T, and Citicorp. He has also been the founder or co-founder of high-tech firms, including Intercomp, Mitrol, and Cambridge Institute for Information Systems, iAggregate.com and currently operates a hotel in the 14th century Langley Castle in England. Dr. Madnick has degrees in Electrical Engineering (B.S. and M.S.), Management (M.S.), and Computer Science (Ph.D.) from MIT. He has been a Visiting Professor at Harvard University, Nanyang Technological University (Singapore), University of Newcastle (England), Technion (Israel), and Victoria University (New Zealand).

MICHAEL D. SIEGEL is a Principal Research Scientist at the MIT Sloan School of Management. He is currently the Director of the Financial Services Special Interest Group at the MIT Center For eBusiness. Dr. Siegel's research interests include the use of information technology in financial risk management and global financial systems, eBusiness and financial services, global ebusiness opportunities, financial account aggregation, ROI analysis for online financial applications, heterogeneous database systems, managing data semantics, query optimization, intelligent database systems, and learning in database systems. He has taught a range of courses including Database Systems and Information Technology for Financial Services. He currently leads a research team looking at issues in strategy, technology and application for eBusiness in Financial Services.