Massachusetts Institute of Technology
Engineering Systems Division

Working Paper Series

ESD-WP-2006-03

POLICY FOR THE PROTECTION AND REUSE OF NON-COPYRIGHTABLE DATABASE CONTENTS

Hongwei Zhu[1], Stuart Madnick[2], and Michael Siegel[3]

**[1]MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142**

**[2]MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142**

**[3]MIT Sloan School of Management
30 Wadsworth Street, Cambridge, MA 02142**

February 2006

Policy for the Protection and Reuse of Non-Copyrightable Database Contents

Hongwei Zhu, Stuart Madnick, Michael Siegel

Working Paper CISL# 2005-12

November 2005

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

Policy for the Protection and Reuse of Non-Copyrightable Database Contents

Hongwei Zhu, Stuart Madnick, Michael Siegel
MIT Sloan School of Management
30 Wadsworth Street, MA 02142, USA
{mrzhu, smadnick, msiegel}@mit.edu

Abstract

With the increasing use of the Internet, many of us feel strongly about the free and unfettered exchange and use of information. But the actual situation is not that simple. After the European Union adopted the Database Directive to provide legal protection for non-copyrightable database contents, the U.S. has introduced six legislative proposals, all of which failed to become a law. One of the major difficulties of formulating a socially beneficial database law is in finding the right balance between protecting the incentives of creating publicly accessible databases (including semi-structured web sites) and preserving adequate access to factual data for value creating activities. We address the problem by developing an extended spatial competition model that explicitly considers the inefficiencies in policy administration. With the model, we can determine various conditions and the corresponding socially beneficial policy choices. 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, the socially beneficial policy choice can be protecting a legal monopoly, encouraging competition via compulsory licensing, discouraging voluntary licensing, or even allowing free riding. The results provide useful insights to the formulation of a socially beneficial database protection policy.

Keywords: database protection, 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 repository of data sources. The accessibility of the Web and a variety of technologies allow someone to easily create new databases by systematically extracting and combining contents of other sources. In fact, we demonstrated in Zhu et al. (2002) that using the Cameleon (Firat et al., 2000) data extraction technology and the Context Interchange (COIN) mediation technology[1] (Goh et al., 1999), we can easily create a database using data of online vendors to provide a global price comparison service.

While many technology-enabled data reuse activities create value for society, these activities may be against the interests (e.g., financial interests) of the source owners 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.

In formulating public policy on this issue, one should consider various stakeholders and different factors related to the value of data and the value created from data reuse. There can be many stakeholders, among 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. For example, a creator who invested in creating a database is interested in recouping the investment using the revenues the database helps to generate. The revenues can be reduced when a reuser creates a

[1] The extraction technology allows us to extract and reuse price data from other web sources; the COIN mediation technology subsequently reconciles semantic differences (e.g., prices quoted in different currencies) amongst disparate sources and diverse users.

[2] A database can contain copyrightable contents, e.g., a database containing MP3 songs. In this cause, 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 copyright protection. In the U.S., data records about certain facts, e.g., phone number listings in white pages, are not copyrightable.

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 database could diminish. There may be other reasons for having restrictions on data reuse. For example, a database creator may want to restrict reuses as a means of ensuring data quality because certain reuses can potentially introduce inaccuracies in data. Not all reuses are for financial purposes only, in which case, a reuser may view restrictions on reusing publicly accessible data as a violation of "freedom of speech" right, an essential element of human rights protected by international law. Besides, privacy concerns often arise when the data contains personal information. Furthermore, people from different cultures and jurisdictions often hold different views, and thus attach different values (not necessarily financial values), to these various factors.

While all factors involved are worthwhile for study, it is beyond the scope of this paper to provide a comprehensive analysis on all of them. Rather, 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, i.e., 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 achieve this objective by developing an economic model, using the model to identify various conditions, and determining policy choices under these various conditions.

2 Background on Legal Challenges and Protection of Database Content

2.1 Legal Challenges to Data Reuse

As mentioned earlier, technologies such as web data extraction and context mediation have made it much easier to create new databases by reusing contents from other existing databases. New business practices consequently 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 the largest online auction site *eBay*. Similarly, *mySimon* 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, e.g., *MaxMiles* allows one to mange various rewards program accounts and *Yodlee* aggregates both financial and rewards program accounts. Common to these aggregated databases is that they add value by providing ease of use of existing data, either publicly available or accessible on behalf of users (e.g., through the use of their user IDs and passwords). Various types of data reuse and the business strategies for data reuse can be found in Madnick and Siegel (2002).

Unfortunately, these value added data reusers have often faced legal challenges for the data they extracted. For example, EBay won a preliminary injunction against Bidder's Edge and the two firms later settled the case. mySimon sued Priceman and the latter ceased to operate for fear of legal consequences. There have been other cases[3]. The legal principles commonly used in the plaintiff claims include copyright infringement, trespass to chattels, misappropriation, violation

[3] E.g., *HomeStore.com v. Bargain Network* (S.D. Cal, 2002), *TicketMaster v. Tickets.com* (C.D. Cal., 2000), *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 Hugenholtz (2001).

of federal Computer Fraud and Abuse Act, false advertisement, and breach of contract[4]. Since none of the cases reached a definite conclusion, it is still a question whether it is legal to reuse publicly available or otherwise accessible factual data in value creating activities. Although the issue of reusing facts existed long before the Web became pervasive, the difficulty in applying the laws that predate the Web and the ease of data reuse in recent years has given lawmakers a certain sense of urgency to resolve the issue by creating a new database protection law.

One of the purposes of such a law is 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 (Samuelson, 1996). 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.

Debate in the past and discussions in existing literature (Samuelson, 1996; Richman and Samuelson, 1997; Sanks, 1998; Maurer and Scotchmer, 1999; Reichman and Uhlir, 1999; O'Rourke, 2000; Lipton, 2003) have identified this major issue but fall short in finding this delicate balance. In this paper, we develop an economic model to identify various conditions for setting a reasonable balance. Before delving into the model, we first briefly describe the landscape of legal protection for databases. After a formal presentation of the model, we relate it with legal proposals and discuss several useful insights developed from our analytic results.

2.2 A Brief History of Database Legislation

Non-applicability of Copyright Law. The impetus for database protection started in 1991 after the Supreme Court in the U.S. decided the *Feist v. Rural*[5] case. In compiling its phone book covering the service area of *Rural Telephone Co.*, *Feist Publications* copied about 8,000 records of

[4] A legal analysis of these claims can be found in court documents, e.g., the eBay case in 100 F. Supp. 2d 1058. ND Cal., May 24, 2000.

[5] 499 US 340, 1991.

Rural's White Pages. In the appeal case, the Supreme Court decided that Feist did not infringe Rural's copyright in that white pages lack the minimal originality to warrant copyright protection. It is the original selection and arrangement of data, not the investment in creating the database or the contents in the database, that is protected by copyright in the U.S. Thus, under current case law, copyright law has not been found to restrict the reuse of the contents in the type of database concerned in this paper.

New Database Legislation. While the database creators in the U.S. were pushing for new database legislation, the European Union (EU) introduced the Database Directive[6] in 1996 to provide legal protection for database contents. Under its reciprocity provision, databases from countries that do not offer similar protection to databases created by EU nationals are not protected by the Directive within the EU. This created a situation where U.S. database creators felt they were neither adequately protected at home, nor abroad. In response, the database industry pushed the Congress to create a new law to provide similar protection to database contents. Since then, the U.S. has attempted six proposals, all of which already failed to pass into law. The most recent two bills are HR 3261 and HR 3872. Figure 1 briefly summarizes these legislative proposals.

[6] "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.

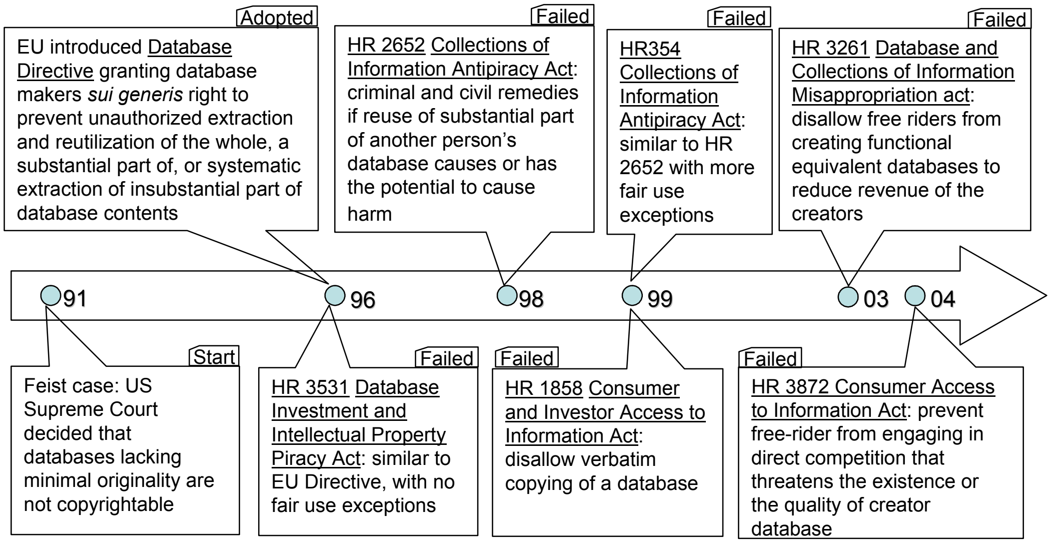
Adopted
EU introduced Database Directive granting database makers *sui generis* right to prevent unauthorized extraction and reutilization of the whole, a substantial part of, or systematic extraction of insubstantial part of database contents
Failed
HR 2652 Collections of Information Antipiracy Act: criminal and civil remedies if reuse of substantial part of another person's database causes or has the potential to cause harm
Failed
HR354 Collections of Information Antipiracy Act: similar to HR 2652 with more fair use exceptions
Failed
HR 3261 Database and Collections of Information Misappropriation act: disallow free riders from creating functional equivalent databases to reduce revenue of the creators
91
96
98
99
03
04
Start
Feist case: US Supreme Court decided that databases lacking minimal originality are not copyrightable
Failed
HR 3531 Database Investment and Intellectual Property Piracy Act: similar to EU Directive, with no fair use exceptions
Failed
HR 1858 Consumer and Investor Access to Information Act: disallow verbatim copying of a database
Failed
HR 3872 Consumer Access to Information Act: prevent free-rider from engaging in direct competition that threatens the existence or the quality of creator database

Figure 1. History of Database Protection Legislation.

The *sui generis*[7] 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. Lawful users are restricted not to "perform acts which conflict with normal exploitation of the database or unreasonably prejudice the legitimate interests of the maker of the database". Here "the legitimate interests" can be broadly interpreted and may not be limited to commercial interests.

HR 3531 of 1996 closely followed this approach with even more stringent restrictions on data reuse. Although the Database Directive has been adopted by the EU, HR 3531 failed to pass in the U.S. One of the main concerns is the constitutionality of the scope and strength of the kind of protection in the EU Database Directive (Reichman and Samuelson, 1997; Colsten, 2001). Other issues in the EU Database Directive include the ambiguity about the minimal level of investment required to qualify for protection (Ruse, 2001; Hugenholtz, 2003), its lack of compulsory license provisions (Colsten, 2001), the potential of providing perpetual protection under its provision of automatic right renewal after substantial database update, the ambiguity in

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

what constitutes a “substantial” update, and several other issues which we discuss in the next subsection.

All subsequent U.S. proposals took a *misappropriation* approach where the commercial value of databases is explicitly considered. 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.

The proposal 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).

The purpose of HR 3872 is to prevent misappropriation while ensuring adequate access to factual information. It disallows only the free-riding that endangers the existence or the quality of the creator database. Unlike in HR 3261, injury in the form of decreased revenue alone is not an offence. Another difference from HR 3261 is that it suggests the Federal Trade Commission be the enforcing authority.

These legislative initiatives demonstrate the substantial difficulties in formulating a database protection and data reuse policy that strikes the right balance, which is a prevailing issue in dealing with other kinds of intellectual property (Besen and Raskind, 1991). An economic understanding of the problem can help address other issues discussed in the next section.

2.3 Other Major Issues

Extensive legal discussions have raised many concerns about a new database law. These include but are not limited to the issues discussed below.

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, no one else but eBay itself can generate the bidding data of items auctioned on eBay. A law that prevents others from using the factual data from a sole source in effect legalizes a data monopoly. Downstream value creating reutilizations of the data will be endangered by a legal monopoly.

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 over invest at an inefficient level to qualify (Samuelson, 1996).

Update distortion and eternal protection. This is an issue in EU law, which allows for automatic renewal of *sui generis* right once the database is substantially updated. Such a provision can induce socially inefficient updates and make possible eternal right through frequent updates (Koboldt, 1997).

Constitutionality. Although the Congress in the U.S. is empowered by the Constitution to regulate interstate commerce under the Commerce Clause[8] and the misappropriation approach often gives a database law a commercial guise, the restrictions of the Intellectual Property Clause[9] often apply to any grant of exclusive rights in intangibles that diminishes access to public domain and imposes significant costs on consumers (Heald, 2001). Most database contents are facts in the public domain; disallowing mere extraction 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". Since little extra value for the society as whole is being created by simply duplicating a database in its entirety, preventing verbatim copying of a database is clearly constitutional. Extracting all contents of a database is very much like duplicating the database. Unlike in copyright law where there is a reasonably clear idea-expression dichotomy (i.e., copyright protects the expression, not the idea conveyed by the expression), extraction-duplication in data reuse is much like a continuum, not a dichotomy (Heald, 2001). Thus a constitutional database law needs to determine up to how much one is allowed to extract database contents.

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 (Tabuchi, 2002) also reveals different opinions from other countries and regions. Enforcement will be a big problem without international harmonization.

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

[9] 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".

We believe the solution to these challenges hinges upon our capability of 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; a database policy so formulated should survive the scrutiny of constitutionality; other inefficiencies can be avoided or at least better understood. We will take on this challenge in the rest of the paper by developing an economic model for database protection and data reuse policy.

2.4 Policy Instruments

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 and the market fails. Policy should intervene to restore 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.

The recent U.S. proposal HR 3261 contains several useful aspects, underlined in the previous section, which 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. It is common that information goods can be differentiated via temporal versioning (Shapiro and Varian, 1998). 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 are simply (1) the grant of legal protection when all criteria are met, and (2) the specification of penalties to violators. They focus on specifying what types of reuse constitute a violation and completely ignore the concern of the creator becoming a legal monopoly (except for HR 1858). Provisions on what the creator is supposed to do should not be ignored, e.g., under certain conditions the creator should be asked to license its data under reasonable terms. Such provisions are often found in other intellectual property laws, e.g., compulsory license provisions in patent laws in various jurisdictions. 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 Literature Review

There have been extensive legal studies on database protection policy[10] since 1996. Recently, Lipton (2003) 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 (2001) 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 (2001) 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

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

from the final version. Hugenholtz (2003) 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 economics and information systems research that directly addresses the issues of database protection policy. We are aware of only one paper by Koboldt (1997), who 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.

Several possible economic theories can be applied in analyzing the issues. Cumulative innovation theory (Scotchmer, 1991) from the patent literature has been used to informally explain the importance of ensuring adequate access to data for knowledge and value creation (Maurer, 2001). The notion of the tragedy of anticommons (Heller, 1998) is also useful because information aggregators rely on the access to multiple information sources. As is shown in Buchanan and Yoon (2000), 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 can be viewed as the “commons”, thus, providing more than necessary protection to databases is analogous to anticommons and will lead to underutilization of protected databases.

4 A Model of Differentiated Data Reuse

As the legal discussions suggest, the reuser is sometimes a competitor of the creator in the database market[11]. 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[12] 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.

As to the Priceman and mySimon databases, both provide price comparison, but the Priceman database has a wider coverage of online vendors, which may suggest a vertical differentiation between the two databases. But the reality can be more complicated, e.g., while mySimon is less comprehensive, it may be more reliable and responsive than PriceMan. Consumers often have different preferences over this set of features. Thus the two databases are largely horizontally differentiated.

There may be cases where a reuser creates a database that is either superior or inferior in every feature relative to the creator database (to target a different market). However, we are interested in 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 database are horizontally differentiated, with competing products located at different locations in the characteristics space. We will base our analysis on an extended spatial competition model, which

[11] 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 model focuses on "information goods" only, thus it does not capture such effect.

[12] 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, e.g., at equal price, all prefer high quality to low quality. See Tirole (1988) for detail.

was introduced by Hotelling (1929) and has been widely used in competitive product selection and marketing research (Salop, 1979; Schmalensee and Thisse, 1988).

4.1 Model Setup

We consider a duopoly case where there are two suppliers of database: (1) a database creator who creates one database product, and (2) a data reuser who produces a different database by reusing a portion of the contents from the creator's database. Both databases are for sale in the market. For example, the database creator could be a marketing firm who compiles a database of New England business directory that includes all business categories. A firm specializing in colleges in Greater Boston area may compile an entertainment guide by reusing a portion of the business directory. The two databases are different in terms of scope, organization, and purpose. In other words, they are differentiated in the product characteristics space. We can understand the databases in *eBay v. Bidder's Edge* case as well as in other data reuse cases similarly. Although many creator and reuser databases are free to individual consumers to view and search, a consumer still pays a price in an economic sense, e.g., time spent and certain private information revealed (e.g., search habit).

In a spatial competition model, we index database features onto a straight line of unit length, with the creator's database at point 0 and the reuser's database at point 1; the prices they set for their databases are p_0 and p_1, respectively. Consumers have heterogeneous preferences over the database features. For simplicity, we assume a unit mass of consumers uniformly distributed along the line $[0,1]$. We assume each database is worth a value v to a consumer with exact preference match. A customer at $x \in [0,1]$ consumes either none or exactly one database. When he does consume a database, he enjoys value v, pays a price, and also incurs a preference mismatch cost determined by the distance and a penalty rate *t*. This is summarized by the following utility function:

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

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 and data reuse market. Many databases from which reusers extract contents are byproducts of business processes. The eBay database is the byproduct of its online auction business. Data in various accounts are generated by transactions of business activities. The cost in creating and maintaining these databases is not a decision variable to be optimized by calculating expected returns on the databases *per se*. MySimon itself is a reuser of vendor data; it is also a database creator when its database contents were extracted by Priceman. The reuse by Priceman is rather serendipitous in that mySimon made its investment decision without ever imagining its data could have been reused by another reuser. 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 reusers. Therefore, the database locations in the feature space are not decision variables, either.

In this model, parameter *t* measures the degree of differentiation of the two databases with respect to consumer preferences; differentiation increases with *t*. When *t* is large, the two products are highly differentiated and the two firms can be two local monopolies. When *t* is small, the two products 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. For the purpose of analyzing if the creator is willing to allow reuse of its data, we also analyze the monopoly case where the creator is the only firm in the market.

In the rest of the paper, unless otherwise noted, profit and social welfare are gross without counting the fixed cost or transaction cost. Utilitarian social welfare is used, which is the sum of firm profit and consumer surplus.

LEMMA 1. In the duopoly case, the market is covered if $t \le v$, and is not fully covered otherwise. In the monopoly case, the market is covered by creator's database if $t \le v/2$, not fully covered otherwise. Best price, maximum profit, and social welfare vary with the differentiation parameter t in both cases as summarized in Table 1. ■

Table 1. Price, profit, and social welfare at different differentiation levels

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

PROOF. Duopoly, little differentiation (*t≤2v/3*). In the case of full market coverage, there exists a location $\tilde{x} \in [0,\ 1]$, such at $u_{\tilde{x},0} = u_{\tilde{x},1} \ge 0$. Then, the demand for database 0 is $\tilde{x}$ and the demand for database 1 is $(1 - \tilde{x})$. Solving profit maximization for both firms with respect to 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 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 deviates 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*.

Duoploy, high differentiation (*v<t*). Each firm's demand is up to the location of the marginal consumer whose utility of purchasing a database is 0. Take the creator, this marginal consumer is 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 ($t \leq v/2$). Similar to 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 highly differentiated duopoly case, it is better for the monopoly to cover a fraction of the market. Straightforward optimization yields the results. ■

We graph the result of Lemma 1 in Figure 2 to help make useful observations; values for both axes are the factors of the product valuation v.

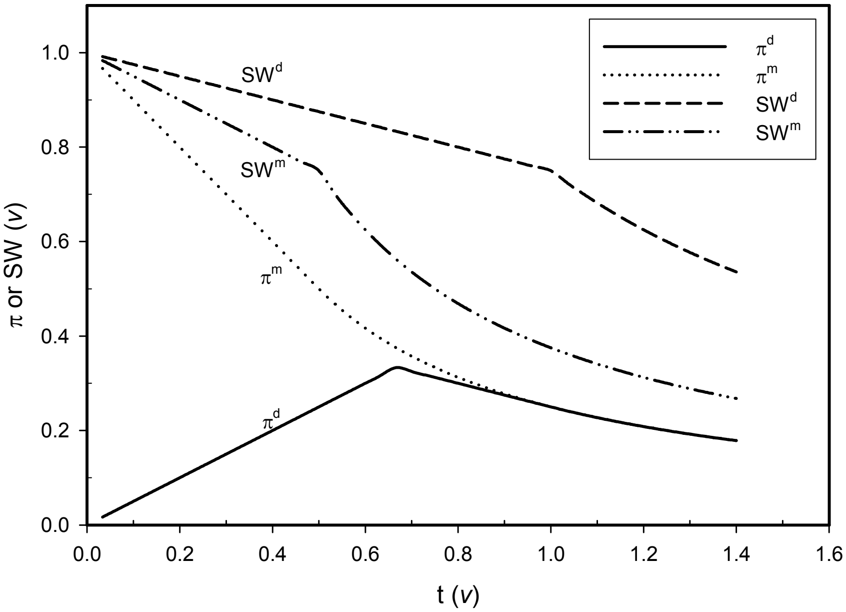

Figure 2. Change of profit and social welfare with differentiation factor *t*.

COROLLARY 1. $\pi^m > \pi^d$ *if* $t < v$, *and* $\pi^m = \pi^d$ *otherwise.* ■

Corollary 1 says that when the reuser's database is not sufficiently differentiated from the creator's, the creator makes less profit because of competition from the reuser's database. When the two databases are highly differentiated, 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 access to the data, it will deny access if a reuser intends to free ride and make a database without sufficient differentiation.

COROLLARY 2. *$SW^d > SW^m$ for all $t>0$.* ■

Corollary 2 says that from the social welfare perspective, two differentiated databases are better than one database. The implicit assumption for this to hold is that the creator makes a positive net profit.

4.2 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., $\pi^m \geq F$.

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 $t \leq v$. Therefore, free riding will cause market failure if $\pi^d < F \leq \pi^m$ and $t \leq v$. That is, if F falls in the region above the π^d line and below the π^m line in Figure 2, free riding causes market failure, in which case an intervention is necessary. 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 rival's costs (Salop and Scheffman, 1987), 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 investment level T, with which the marginal cost of the reuser becomes $C_1(T)$. The cost of installing such anti-extraction technologies is often small enough to be negligible. When

T is such that $0 \le C_1(T) \le \min\left\{\frac{3}{2}(v-t), 3t, 2v-3t\right\}$, the creator profit becomes $\pi^d_{0,T} = {(t+\frac{C_1}{3})^2}\big/{2t} > \pi^d = \frac{t}{2}$, i.e., the creator profit is higher than when the technology is not used. The reuser profit is $\pi^d_{1,T} = {(t+\frac{2C_1}{3})(t-\frac{C_1}{3})}\big/{2t}$, which could be greater or less than π^d, depending on the level of $C_1(T)$. The first two items in the constraint for $C_1(T)$ ensure that the reuser is not deterred; the third item ensures full coverage of the market. Obviously, if $C_1(T)$ is very high, the reuser will be deterred. However, anti-extraction techniques were not very effective in practice[13]; we suspect that $C_1(T)$ has been too small to have 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 wasteful investment because they merely help transfer consumer surplus and reuser profit to the creator. A database law that grants the creator the right to license its data to reusers can reduce or eliminate this social inefficiency. When database creators are also reusers, the cost-raising problem may not arise at all[14].

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[15] and sometimes are sued by the creators. The uncertainty of various proposed database bills creates significant 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

[13] 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.

[14] 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.

[15] For instance, a few online travel agencies recently 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.

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.

4.3 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 pay a data reuse 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 profit. Negotiating the fee schedule *r* and administrating data reuse policy often incur some cost. To model this reality, let us suppose that when the creator asks for *r*, it actually gets αr, where $\alpha \in [0,\ 1]$ and it measures transaction efficiency. Thus, in the duopoly case with data reuse policy in place, ($\pi^d + \pi^d$) is the best the creator can get to offset its fixed cost *F* if it ever allows someone to reuse its data[16]. Before we develop the formal analysis, we describe the intuitions by plotting this upper bound condition along with profit curves in Figure 3.

[16] We assume there will be no collusive joint profit maximization. A Nash bargaining outcome will be 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.

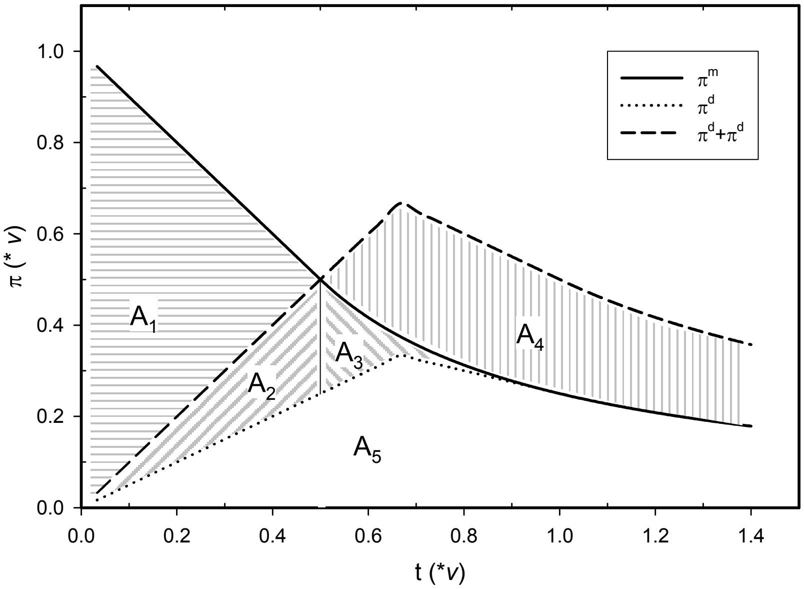

Figure 3. Change of Profits with Differentiation Factor *t*.

We mark five areas A_1 through A_5 in Figure 3. The upper-bound ($\pi^d + \pi^d$) curve will 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* comparing to profits (the Y-axis) are such that *F* falls in one of the areas. If *F* falls in A_1 (i.e., *t* is between 0 and 0.5**v*, and *F* is below the solid line for monopoly profit and above the dashed line for 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.

For $F \in A_2$, 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 $F \in A_3$, 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. When

$F \in A_4$, 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. In this case, the reuser can be considered as a database creator who incurs a fixed cost F_1, where $F_1<F$. When a third party reuses data from either or both jointly developed databases, the model can be used to analyze various conditions between this third party (i.e. the reuser) and the creator(s).

Finally, when $F \in A_5$, 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. To simplify analysis, we let $r = \pi^d$, i.e., we assume that the creator has the negotiation skills or legal power to ask the reuser to disgorge all profits from reusing the data. 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.

THEOREM 0. (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. $\hat{\alpha} = 0.5$ 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 $t > 2v/3$. ■

PROOF. With a fee paying reuser, the social welfare is $SW^{dl} = SW^d - (1-\alpha)\pi^d$. Licensing is socially beneficial only if $SW^{dl} \ge SW^m$. Using the results in Lemma 1, we can solve the inequality and obtain $\hat{\alpha}$. ■

This is a refinement to Corollary 2. 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.

Theorem 1. (A_1: Little differentiation, high cost) When $t \leq v/2$ and $(1+\alpha)t/2 < F \leq v - t = \pi^m$, legal protection to the creator's database should be granted. The existence of a reuser database causes a market failure even if the reuser pays a licensing fee; it is socially beneficial to let the creator be a monopoly in the market by disallowing the creation of the reuser database. ■

Proof. In the presence of a reuser database, the creator's profit is *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. In both cases, the creator cannot make a positive net profit, 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 $SW^m - F = v - t/2 - F \geq t/2 > 0$. ■

Theorem 2. (A_2: Little differentiation, moderate cost) When $t \leq v/2$ and $t/2 \leq F < (1+\alpha)t/2$, legal protection to the creator's database should be granted. 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 $\alpha > \hat{\alpha} = 0.5$. If $\alpha \leq 0.5$, it is better to let the creator be a monopoly. ■

Proof. This can be easily proved with Lemma 1 and Theorem 0. ■

Theorem 3. (A_3: Moderate differentiation, moderate cost) When $v/2 < t \leq v$ and $\pi^d < F \leq \min\{(1+\alpha)\pi^d, v^2/4t = \pi^m\}$, legal protection of the creator's database should be granted. The creator is willing to license its database if $\alpha \geq (\pi^m - \pi^d)/\pi^d = \tilde{\alpha}$. Within the range of differentiation, $\tilde{\alpha}$

can be less than or greater than $\hat{\alpha}$. If $\hat{\alpha} < \alpha < \tilde{\alpha}$, compulsory licensing is necessary; if $\tilde{\alpha} < \alpha < \hat{\alpha}$, licensing should be disallowed even though the creator prefers.

PROOF. Legal protection is necessary because with free riding the creator cannot make enough profit to cover its fixed cost. When $\alpha \geq \tilde{\alpha}$, $\pi^d + \alpha\pi^d \geq \pi^m$, i.e., the creator is better off licensing its database to the reuser. When $\hat{\alpha} < \alpha < \tilde{\alpha}$, it is socially beneficial to license but the creator makes less than monopoly profit, 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. The values of $\hat{\alpha}$ and $\tilde{\alpha}$ are presented graphically in Figure 4. ■

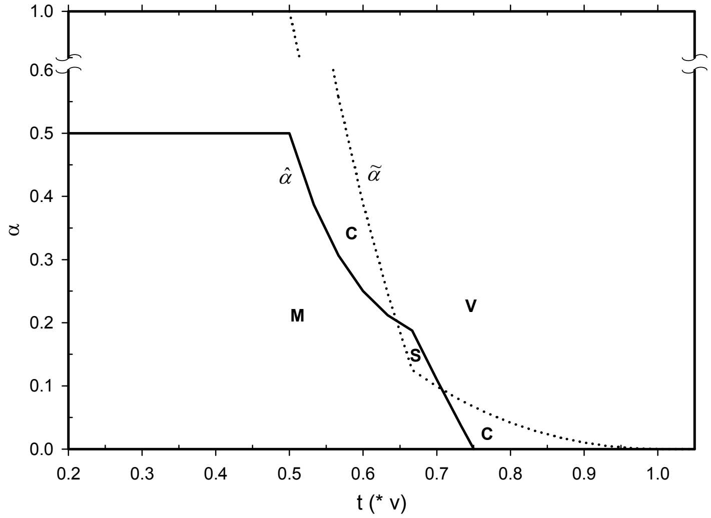

Figure 4. Change of minimal transaction efficiency with differentiation factor *t*.

In Figure 4, we see in most cases $\tilde{\alpha} > \hat{\alpha}$, meaning that generally transaction efficiency requirement is higher for voluntary licensing than for compulsory licensing. We label each region with a symbol denoting the socially beneficial policy choice, i.e., C indicates the necessity of compulsory licensing, V means voluntary licensing by the creator, M represents the case where a monopoly is better, and S is a special case where the creator wants to license but it is socially wasteful to license and the creator should be a monopoly.

THEOREM 4. (A$_4$: Moderate to high differentiation, high cost) 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 with cost sharing agreement. ■

PROOF. It is straightforward that the creator will not create the database because the development cost is greater than monopoly profit. If the cost is below joint profit discounted with transaction cost, the creator is willing to participate in joint development to make a positive profit. ■

THEOREM 5. (A$_5$: Trivial databases) It is socially beneficial NOT to grant legal protection to databases when $F \le \pi^d \le v/_3$ and $\alpha < 1$. ■

PROOF. With low cost of creating the database, there is no market failure even without legal protection. For $\alpha < 1$, $SW^d > SW^{dl} = SW^d - (1-\alpha)\pi^d$. Recall from corollary 2 that $SW^d > SW^m$. Social welfare without protection is SW^d, it is SW^{dl} or SW^m with protection. Therefore, no protection is socially beneficial. For all *t*, it is straightforward to show $\max(\pi^d) = v/_3$ using Lemma 1. ■

COROLLARY 3. (A5: Highly differentiated databases) When $t > v$ and $\alpha < 1$, it is socially beneficial NOT to grant legal protection to databases. ■

PROOF. This is a special case of Theorem 5. 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. ■

Most enacted and proposed database protection bills grant legal protection only to databases that require substantial expenditure to create and maintain. Theorem 5 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 5. Thus, we should not specify an absolute dollar amount threshold for a database to qualify for legal protection.

From Theorem 5, it seems desirable to set a fixed cost threshold $\hat{F}$ equal to duopoly profit, i.e., $\hat{F} = \pi^d \leq v/3$. As we will see next, this can induce excessive investment when an efficient firm can create the database at a cost slightly lower than $\hat{F}$.

4.4 Over Investment Distortion

THEOREM 6. (Over investment distortion) Suppose $\hat{F} = \pi^d$, when $\alpha > \hat{\alpha}$ and $t \leq v$, a creator with $\underline{F} < F < \hat{F}$ has incentives to over invest to qualify for legal protection, where F is the cost when the creator produces the database efficiently. The value of $\underline{F}$ depends on α and t. When $t \leq \frac{2v}{3+2\alpha}$ and $\underline{F} = \max\{(4+\alpha)t/2 - v, 0\}$, or when $\frac{v}{2} < t \leq \frac{v}{\sqrt{2(1+2\alpha)}}$ and $\underline{F} = \max\{\frac{2(2+\alpha)t^2 - v^2}{4t}, 0\}$, the creator aggressively over-invest $\frac{(1+\alpha)t}{2}$ to become a monopoly; when $\frac{2v}{3+2\alpha} < t \leq \frac{v}{2}$ or $\frac{v}{\sqrt{2(1+2\alpha)}} < t \leq \frac{2v}{3}$, and $\underline{F} = (1-\alpha)t/2$, the creator only moderately over-invest $\pi^d = t/2$ to become eligible for licensing fee; when $\frac{2v}{3} < t \leq v$, $\alpha > \frac{(v-t)^2}{2vt-t^2}$, and $\underline{F} = (1-\alpha)\frac{2v-t}{4}$, the creator only over-invest $\pi^d = \frac{2v-t}{4}$ to qualify for receiving licensing fee. ■

PROOF. When $F < \hat{F}$ (i.e., F is in A$_5$ area in Figure 3), the reuser can legally free ride, so the creator's net profit is $\pi^d - F$. The creator has incentive to over-invest to a level in A$_1$, A$_2$, or A$_3$ areas as long as it can make a higher net profit.

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

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

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 $F \geq \frac{(4+\alpha)t}{2} - v$, 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 $t \leq \frac{2v}{3+2\alpha}$. With the assumption of $\alpha > \hat{\alpha} = 0.5$, we know that $\frac{2v}{3+2\alpha} < \frac{v}{2}$. Similarly, the incentive compatible conditions for over-investing to $\hat{F}$ to only qualify for receiving licensing fee are:

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

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 $F \geq (1-\alpha)\pi^d = \underline{F}$, where $\pi^d = \frac{t}{2}$; solving (4) yields $t > \frac{2v}{3+2\alpha}$.

When $\frac{v}{2} < t \leq \frac{2v}{3}$, 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 $\alpha > \frac{(v-t)^2}{2vt-t^2}$. There is no incentive to become a lawful monopoly because the creator earns a bigger profit when there is a fee paying reuser. ■

COROLLARY 4. Over investment can also occur even if the creator already qualifies for protection but is subject to compulsory licensing as specified in Theorem 2. Specifically, the creator over invests at the level of ${}^{(1+\alpha)t}\!/_2$ when $(2+\alpha)t - v = \underline{F} < F < {}^{(1+\alpha)t}\!/_2$, $t \leq {}^{v}\!/_2$, and $\alpha > \hat{\alpha} = 0.5$.

PROOF. The creator over invests if $\pi^m - \pi^{dl} > \pi^{dl} - F$, which gives $F > \underline{F} = (2+\alpha)t - v$. The lower bound for α is necessary from Theorem 2. ■

Theorem 6 shows that when $F \in A_5$, the unprotected database creator wants to spend more at F', where $F' \in A_2$ or $F' \in A_3$, so that the database now becomes qualified and the creator can earn a bigger profit. The creator can more aggressively invest at F'' such that $F'' \in A_1$, in which case the creator becomes a legal monopoly. Corollary 4 shows that a creator with $F \in A_2$ wants to move to A_1 by spending more. These distortions benefit the creators but are socially wasteful.

5 Discussion

5.1 Summary of Findings

In this spatial competition model, we consider the creator, the reuser, and the consumers. Depending on the condition, the reuser can be a free-rider or a fee paying data reuser, or reuse is disallowed. As a unique feature of the model, we explicitly consider inefficiencies of policy administration and abstract it as the transaction efficiency parameter α . This is an improvement over previous policy research that often ignores this factor, which in effect assumes perfect efficiency of policy implementation and enforcement.

The model also allows us to clarify several important notions in policy design. 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 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 database creation 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; any revenue reduction due to competition from the reuser is an offence. Thus the purpose of the proposal is about fairness to creators, not about social welfare maximization. In the model, the incentives of creating the database do not completely disappear as long as the creator makes a positive profit.

With this model, we are able to specify socially beneficial policy choices under various conditions determined by the magnitude of fixed cost of database creation (*F*), the degree of differentiation between the reuser database and the creator database (*t*), and the transaction efficiency (α). 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 license 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 (2002) finds that depending on cost distribution no copyright protection can be socially beneficial. In the presence of demand network externalities, Takeyama (1994) finds that unauthorized reproduction of any intellectual property is Pareto improving, i.e., 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 (2003) for their discussion on the adverse effect of anti-piracy investment.

We also discover the possibility of excessive investment distortions that come with this design of database policy. 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 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.

5.2 Implementation and Implications

The model and the results provide helpful guidelines to specifying and implementing a socially beneficial database protection policy. This can be illustrated perhaps with critiques to the recent U.S. proposals, particularly HR 3261.

HR 3261 of 2003 is generally in line with the results here. It takes an appropriate approach with a focus on competition and the commercial value of databases. 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 will become a lawful monopoly under the proposal, which is harmful to society. These shortcomings will likely raise constitutionality concerns.

In addition, the three conditions in HR 3261 are essentially a paraphrasing of the misappropriation doctrine established in *INS v. AP*[17] of 1918 and more recently in *NBA v. Motorola*[18] of 1997. In the former appeal case, International New Service (INS) took factual stories from the Associated Press's (AP) bulletins and East Coast newspapers and wired them to its member newspapers on the West Coast. In the later case, Motorola transcribed NBA playoff scores from broadcast and sent them to its pager subscribers. Both INS and Motorola were found guilty under the misappropriation doctrine. The practical question is if it makes sense to codify a well established doctrine and make it into a statute to regulate data misappropriation only. With its substantial similarity to the misappropriation doctrine, HR 3261 would have little impact because without it any data reuse case can be decided using the doctrine.

With 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. Our model clarifies this vague criterion.

HR 1858 prevents duplication of a database, which is an extreme case where t is nearly zero. With no differentiation, the reuser (now a duplicator) adds little value to the society, i.e., $SW^d \approx SW^m$, and market failure is highly likely with even a moderate creation fixed cost, thus

[17] 248 US 215 (1918).
[18] 105 F.3d 841 (1997).

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.

Overall, the results from the economic model provide useful insights for formulating database protection policy. By revisiting the two undecided cases described in the introduction section, we can briefly discuss the implications of a policy suggested by the model.

eBay v. Bidder's Edge. In the eBay case, the computing resource is not the subject matter of such a policy, which concerns the data, not the resources that deliver the data. Thus trespass on the Internet is a different issue out of the scope of this discussion. 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 has a much broader coverage; thus, there is competition from the reuser database. In terms of functionality, eBay's database allows one to buy and sell items; the reuser database does not provide any actual auction service. Thus the two databases exhibit significant differentiation. Searching alone does not, in general, reduce eBay's revenue from its auction service. In addition, searching and actual auction are two different markets. If we subscribe to the spin-off theory (Hugenholtz, 2003), the eBay database will not meet the cost criterion. Therefore, free reuse by Bidder's Edge should be allowed under our model.

mySimon v. Priceman. In mySimon case, the reuser database is a superset of the creator's. Both are in the searchable comparison shopping database market. Free riding by the reuser certainly reduces 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 or it is penalized for violating of the database protection policy.

5.3 Concluding Remarks

We address a pressing issue in database legislation that needs to find the right balance between protecting incentives of database creation and preserving enough 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. From these results, we derive several guidelines that are useful to law makers when they consider economic factors in database policy formulation. A better understanding of the economic issues in database legislation provided by our analysis will be helpful in developing consensus towards international harmonization in database regulation.

There are also a number of limitations in our analysis. As discussed earlier, we focus on financial interests in database contents; other factors concerning societal values of data and data reuse are not considered. Our economic model considers the competition between the creator and reuser databases. The model does not capture other effects of data reuse, e.g., network effects of database products. In addition, the model also ignores factors that are specific to the kind of data being reused, e.g., privacy concerns when the reused data is about personal information, and increased price competition concerns when price data is reused.

In addition to relaxing the limitations identified above, we plan to look into a few other areas in future research. Out 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 Landes and Posner (2003), intellectual property is also the input to intellectual property creation. With strong protection for database contents, the cost of database creation will likely rise. In addition, many online databases have characteristics of two-sided markets (Rocket and Tirole, 2003; Parker and Van Alstyne, 2005), e.g., they 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. Nevertheless, the current model captures many of the major issues in database legislation and should be helpful to the formulation of a socially beneficial data reuse policy.

References

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

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

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

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

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

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

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

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

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

Hugenholtz, P.B. 2003. 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.

——. 2001. 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.

Koboldt, C. 1997. 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) 127-138.

Landes, W.M., R. A. Posner. 2003. *The Economic Structure of Intellectual Property Law* Belknap Press.

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

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

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

——. 2001. Intellectual Property Law and Policy Issues in Interdisciplinary and Intersectoral Data Applications. Data for Science and Society, National Research Council.

O'Rourke, M. A. 2000. Shaping Competition on the Internet: Who Owns Product and Pricing Information? *Vanderbilt Law Review,* **53**(6)**,** 1965-2006.

Parker, G. G., M., Van Alstyne. 2005. Two-Sided Network Effects: A Theory of Information Product Design. *Management Science,* (forthcoming).

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

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

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

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

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

——, D.T. Scheffman. 1987. Cost-Raising Strategies. *J. Industrial Economics* **36**(1) 19-34.

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

Schmalensee, R., J. F. Thisse. 1998. Perceptual Maps and the Optimal Location of New Products: An Integrative Essay. *International Journal of Research in Marketing* **5**(4) 225-249.

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

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

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

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

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

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

Zhu, H., Madnick, S. and Siegel, M. 2002. Global Comparison Aggregation Services. *1st Workshop on E-Business*, Barcelona, Spain.