Open access • Journal Article • DOI:10.2753/JEC1086-4415120306

A Privacy-Protecting Business-Analytics Service for On-Line Transactions

Bettina Berendt, Sören Preibusch, Maximilian Teltzrow

Institutions: Katholieke Universiteit Leuven, German Institute for Economic Research

Published on: 01 Apr 2008 - International Journal of Electronic Commerce (M. E. Sharpe, Inc.)

Topics: Web analytics, Privacy software, Analytics, Web service and Web modeling

A Privacy-Protecting Business-Analytics Service for On-Line Transactions

Bettina Berendt, Sören Preibusch, and Maximilian Teltzrow

ABSTRACT: Analysis of consumer-related and consumer-generated data is a very important way to measure the success of on-line retailing. The software packages for data analysis have two major shortcomings: (1) solutions are not offered as a service reachable by standard procedures over the Internet, but as isolated standalone applications or ERP system modules; (2) privacy restrictions need to be integrated into a framework of business analytics for Web retailers. The first aspect can be addressed with standardized developer software for Web services, but the second must consider privacy legislation, privacy specifications on Web sites (P3P), and data reidentification problems. These shortcomings are addressed by a proposed formal model of these problems and an implementation of the model as a declarative specification of privacy constraints, expressed as an extension of P3P. The constraints are complemented by a logic identifying the elements in a given set of Web analytics that might lead to data reidentification and therefore violate implicit privacy constraints. A Web-based service is presented that uses these components to automatically adapt the set of available Web analytics to an on-line retailer's P3P policy. The system was tested on a large data set from a major European multichannel retailer.

KEY WORDS AND PHRASES: Data mining, electronic business, P3P, privacy, Web analytics.

Data mining for business planning, forecasting, monitoring, and control has become a common routine, particularly in e-business, where companies can learn more about their customers through on-line interactions. Mining techniques can help companies to increase customer loyalty, increase sales, understand consumer behavior, enhance site navigation, and personalize Web sites [15, 32, 39, 51, 59]. Customer data can be analyzed by means of a variety of tools often known as CRM (customer relationship management) or BI (business intelligence) solutions [43].

Although analysis of customer data is now commonplace, companies need to be aware that the use of such data may conflict with mandatory or self-imposed privacy requirements. A considerable amount of work exists on how to formalize textual privacy policies and implement privacy choices in current Internet technology [6, 8, 13]. However, privacy conflicts arise when privacy policies have to be integrated into a company's analysis framework. In particular, problems arising from data reidentification (often referred to as triangulation or inference problems [55]) and conflicts between P3P and privacy legislation pose a challenge to the analysis of consumer data.

This paper presents a formal model of these problems, together with an implementation of the model as a declarative specification of privacy constraints, expressed as an extension of P3P. The constraints are complemented by a logic that identifies those elements in a given set of business analytics that might lead to data reidentification and therefore violate implicit privacy constraints. The authors propose to offer this privacy-constrained analytics

International Journal of Electronic Commerce / Spring 2008, Vol. 12, No. 3, pp. 115–150.

1086-4415/2008 $9.50 + 0.00.
DOI 10.2753/JEC1086-4415120306

system as a Web-based service for on-line retailers. The service offers a set of business analytics that the client would like to compute. It selects the analytics in accord with privacy restrictions and returns the result to the client. As a hosted solution, the service not only helps companies avoid in-house hardware, software, and training expenditure for analytics systems, but also ensures that they can rely on up-to-date knowledge of privacy-protection necessities and newly emerged privacy threats. The system is designed as a Web service and in its current implementation also offers a browser front-end. Thus it leverages state-of-the-art standards for interoperability and user-friendliness. The system has been applied and tested on Web log, transaction, and external data from a large multichannel retailer (i.e., a retailer that operates and e-shop *and* physical stores).

Background and Related Work

Related work and motivating background come from the areas of Web analytics and data protection/privacy.

Business Analytics for Web Retailers

On-line analytical processing (OLAP) and data-mining techniques are typically used to analyze data from consumer interactions. The data analyses lead to a set of measures that are commonly termed (business) analytics, indicators, or metrics. The discussion in this paper will use the terms *analytics* and *indicators* interchangeably. Examples are "micro-conversion rates" [34], "e-metrics" [15], "Web traffic measurements" [38], "operational metrics" [52], "visit related measures" [40], "CRM analytics" [49], and "Web log metrics" [32]. New Web usage analytics for multichannel retailers have been presented by several researchers [32, 57, 58]. Web usage analysis has been extended by adding demographic and purchase data [11]. Further Web measurements have been introduced [9, 61]. An increasing number of vendors offer Web analytics software. Many of these systems are installed and operated by and at the company operating the Web site (e.g., as parts of systems with different core functionalities, such as ERP), but hosted solutions are becoming increasingly popular. Hosted solutions allow companies to offload non-core activities rather than incur a large overhead on specialist hardware, software, and personnel, and to write off costs as operating expenses [33]. Examples of hosted analytics solutions are Google Analytics, ClickTracks, WebSideStory, Omniture, Coremetrics, Fireclick, IBM SurfAid, and WebTrends.

As a summary of these proposals, three categories of business indicators are defined that represent the analysis framework of an on-line retailer. The list of indicators, their definitions, and the required data attributes are shown in List 1.

Low-level Web analytics have not been included in the proposed service. Basic statistical aggregations of Web logs, such as visits per day, distribution of user agents, entry and exit pages, most frequently visited Web pages, request

```
customer (customer_id, geo_id, credit_rating, first_name, surname,
title, gender, date_of_birth)
address (address_id, customer_id, geo_id, country_code, street,
street_number, street_number_supplement, customer_zip_code, town,
recipient_address, post_office_box, phone_number, e-mail_address)
order (order_id, customer_id, session_id, store_id, product_id,
status, invoice_value, currency, order_date, order_time, delivery_
type, payment_method, credit_card_no, customer_card_no, status_change)
product (product_id, category_id, product_name, product_weight,
product_size, price, cost)
product_category (category_id, category_name)
return (return_id, order_id, store_id, return_date, return_value,
return_address)
store (store_id, geo_id, store_country_code, store_street_name, store_
street_number, store_zip_code, store_town)
session (session_id, order_id, ip_location, access_time, browser_type,
status_code, referrer)
page (page_id, concept_id, session_id, page_name, page_content)
page_concept (concept_id, concept_name, concept_content)
belongs_to (page_id, concept_id)
contains (session_id, page_id)
location_zip* (geo_id, micro_id, zip_code, longitude_zip_code,
latitude_zip_code)
microgeography* (micro_id, detail_type, detail_value)
characterizes (micro_id, geo_id)
```

List 1. Exemplary Data Schema of an Online Retailer

Notes: Underlined: primary key, dashed: foreign key. Log data in the table session could be linked to attributes in the table customer via a unique order_id when a user made an online purchase. If a site uses cookies, the attribute cookie_id can be stored in the table session. * Third-party data sources can be added to extend a retailer's database with additional consumer profile information. We acquired demographic data from Deutsche Post Direkt that matches zip codes and geographic coordinates. Thus, the table location_zip could be added. The column detail_type in the table microgeography includes further microgeographic attributes such as household size, creditworthiness, age, product affinity, spending capacity, preferred order medium or preferred communication media that could be added via the geo_id (e.g., zip code).

errors, and server load are popular. However, while they may identify system-administration problems or usability challenges, they are of less concern to the strategy of customer relationship management. These analytics can be calculated by standard tools, including shareware software [29]. Product analyses [e.g., 32] are not present in the framework but can be added to it using plug-ins. A simplified view of an on-line retailer's data schema consisting of typical data attributes that can be collected on-line is given in Table 1.

- A first group of Web analytics relates to orders. The analytics in this group segment customers according to transaction and demographic characteristics, and can be used for customer scoring, product assortment, or site personalization [15, 46].
- A second set of analytics measures consumer preferences among a Web site's service offerings. These analytics measure how often customers use certain payment, delivery, and return options. The indicators are particularly useful for multichannel retailers to optimize their distribution strategies [34].

Analytics (examples)	Explanation	Required data entities (see List 1)	Personal data needed?*
Order analytics			
Number of orders	Number of orders in given time frame	order_id, order_date	N
Mean number of purchases per customer	Arithmetic mean of number of orders per customer	customer_id, order_id, order_date	P
Mean transaction amount per customer	Arithmetic mean of generated revenue per customer	customer_id, invoice_value, order_date	P
Mean transaction amount per order	Arithmetic mean of generated revenue per order	invoice_value, order_id, order_date	N
Mean margin per customer	Arithmetic mean of margin per order	customer_id, invoice_value, order_id, cost, product_id, order_date	P
Mean margin per order	Arithmetic mean of margin per order	invoice_value, order_id, cost, product_id, order_date	N
Mean interpurchase time	Arithmetic mean of time between two successive purchases by same customer	customer_id, order_id, order_date	P
Gini coefficient	Concentration coefficient of generated revenue per customer	customer_id, invoice_value, order_id, order_date	P
Recency distribution	Classes of number of customers that repeatedly purchased within same time frame from their most recent visit and present time	customer_id, order_id, order_date	P
Frequency distribution	Classes of number of customers that incurred the same number of orders in a time frame	customer_id, order_id, order_date	P
Monetary value distribution	Classes of number of customers that generated the same range of order value in a time frame	customer_id, invoice_value, order_id, order_date	P
Margin distribution	Classes of the number of customers that generated the same range of profit margin in a time frame	customer_id, invoice_value, cost, product_id, order_id, order_date	P
Recency for specific customer	Time between an individual's subsequent orders (individual specified by name and address)	customer_id, first_name, surname, address, order_id, order_date	I
Frequency for specific customer	Number of orders an individual incurred in a time frame (individual specified by name and address)	customer_id, first_name, surname, address, order_id, order_date	I

Monetary value for specific customer	Monetary value an individual incurred in a time frame (individual specified by name and address)	customer_id, first_name, surname, address, invoice_value, order_id, order_date	I
Margin for specific customer	Margin an individual incurred in a time frame (individual specified by name and address)	customer_id, first_name, surname, address, invoice_value, cost, product_id, order_id, order_date	I
Revenue contribution	Revenue contribution of classes of customers (Lorenz curve)	customer_id, invoice_value, order_id, order_date	P
Demographic analytics			
Gender split	Ratio of female and male customers	customer_id, gender, order_date	P
Mean revenue/gender	Arithmetic mean of revenue generated by female and male customers	customer_id, invoice_value, gender, order_date	P
Mean margin/gender	Arithmetic mean of margin generated by female and male customers	customer_id, invoice_value, cost, product_id, gender, order_date	P
Customer-distance correlation	Pearson correlation between number of customers normalized with population density in their ZIP code with the ZIP code's distance to next physical store	customer_id, customer_zip_code, geo_id, store_zip_code, longitude_zip_code, latitude_zip_code, order_date	P
Mean revenue by age distribution	Classes of age and corresponding mean revenue per order	customer_id, date_of_birth, order_id, invoice_value, order_date	P
Number of customers per location (Zip code)	Classes of locations and corresponding number of customers	order_id, customer_id, customer_zip_code, order_date	P
Number of transactions per location (Zip code)	Classes of locations and corresponding number of transactions	order_id, customer_id, customer_zip_code, order_date	N
Revenue per location	Classes of locations and corresponding revenue	order_id, customer_id, invoice_value, customer_zip_code, order_date	N
Margin per location	Classes of locations and corresponding margin	order_id, customer_id, invoice_value, cost, product_id, customer_zip_code, order_date	N
Microgeographic details of customers in Zip code area	Analytics describing microgeographic details of customers in given ZIP code area	customer_id, order_id, address_id, customer_zip_code, geo_id, micro_id, detail_type, detail_value, order_date	P
Microgeographic details for specific customer	Analytics describing microgeographic details of individual customers	customer_id, first_name, surname, address, geo_id, micro_id, detail_type, detail_value, order_date	I

(continues)

(*Continued*)

Service analytics

In-store payment ratio	Number of orders paid in-store per number of all transactions	order_id, payment_method, order_date	N
On-line payment ratio	Number of orders paid on-line per number of all transactions	order_id, payment_method, order_date	N
Cash-on-delivery payment rate	Number of orders paid cash on delivery per number of all transactions	order_id, payment_method, order_date	N
In-store payment migration ratio	Number of repeat customers who changed payment preferences from on-line to in-store in at least one subsequent transaction per number of all customers	customer_id, order_id, payment_method, order_date	P
On-line payment migration ratio	Number of repeat customers who changed payment preferences from in-store to on-line in at least subsequent transaction per number of all customers	customer_id, order_id, payment_method, order_date	P
Pickup in-store ratio	Number of orders picked up in store per number of all transactions	order_id, delivery_type, order_date	N
Direct delivery ratio	Number of orders delivered directly per number of all transactions	order_id, delivery_type, order_date	N
In-store delivery migration ratio	Number of repeat customers who changed delivery preferences from on-line to in-store in at least one of the following transactions per number of all customers	customer_id, order_id, delivery_type, order_date	P
Direct delivery migration ratio	Number of repeat customers who changed delivery preferences from in-store to on-line in at least one subsequent transaction per number of all customers	customer_id, order_id, delivery_type, order_date	P
Returns to stores ratio	Number of orders that were returned to physical stores per number of all transactions	customer_id, order_id, store_id, order_date	P
Product weight and pickup distribution	Classes of number of orders consisting of products within the same weight range compared with number of pickups	order_id, product_id, product_weight, delivery_type, order_date	N
Product size and pickup distribution	Classes of orders consisting of products within same size range compared with number of pickups	customer_id, order_id, product_id, product_size, delivery_type, order_date	N
Revenues and pickup distribution	Classes of number of orders consisting of products within same revenue range and compared with number of pickups	customer_id, order_id, invoice_value, delivery_type, order_date	N

Returns from location distribution	Distribution of returns from locations (e.g., ZIP codes)	customer_id, order_id, customer_zip_code, store_id, order_date	N
Returns/name and address	Distribution of individuals and respective number of returns	customer_id, order_id, first_name, surname, address, store_id, order_date	I
Conversion analytics			
Look-to-click	Visitors who performed product click-through/visitors who saw product impression	session_id, page_id, concept_name, access_time	N
Click-to-basket	Visitors who effected basket placement/visitors who performed product click-through	session_id, page_id, concept_name, access_time	N
Basket-to-buy	Visitors who made product purchase/visitors who effected basket placement	session_id, page_id, concept_name, access_time	N
Look-to-buy	Visitors who made product purchase/visitors who saw product impression	session_id, page_id, concept_name, access_time	N
Reach	Suspects/whole population	session_id, page_id, concept_name, access_time	N
Acquisition	Visitors who become active site investigators (prospects/suspects)	session_id, page_id, concept_name, access_time	N
Conversion	Visitors who purchase (customers/prospects)	session_id, page_id, order_id, concept_name, access_time	N
Retention	Repeat customers/customers	customer_id, order_id, order_date	P
Life-cycle interruption analytics			
Abandonment	Visitors who filled shopping cart and abandoned it/active site investigators	session_id, page_id, concept_name, access_time	N
Attrition	Customers who subsequently became customers elsewhere/customers	customer_id, first_name, surname, address, order_date	I
Churn	Attrited customers/customers minus attrited customers	customer_id, first_name, surname, address, order_date	I

Table 1. Analytics.

* I: requires at least personal identification data, P: requires at least pseudonymous data, N: no modification required; all analyses are time-framed.

- A third set of analytics proposes fine-grained measures of conversion success in on-line retailing. They measure how well a site transforms Web site visitors into on-line buyers, or "lookers to bookers," and are useful for site optimization and navigation improvement [15, 34].

Some of these analytics build on personal data and may thus raise privacy concerns. The following section, therefore, investigates the legal and contractual requirements on privacy and data protection.

Privacy Requirements

Companies' data-analysis practices, such as those described above, have significantly increased users' privacy concerns. User perceptions of insufficient privacy protection are not only a societal problem but are harmful to companies. Privacy concerns have a negative effect on intention to transact on-line and make users avoid Web sites or discontinue transactions with them [17, 59]. Some of these concerns have been alleviated by the enactment of privacy legislation in many countries. Transnational and supranational agreements, such as the "Safe Harbor" framework concluded between the European Commission and the U.S. Department of Commerce in 2000, complement privacy legislation. Moreover, site owners are increasingly adopting the Platform for Privacy Preferences (P3P), an industry standard for privacy protection designed to give users more control over their personal information when visiting Web sites. The discussion in this section describes privacy requirements and the restrictions they impose on the calculation of business indicators in the analysis framework.

Legal Restrictions

Laws protecting the privacy of individuals are in effect in more than 30 countries [31]. For comprehensive resource collections, see www.epic.org, www.privacy.org, www.privacyinternational.org, www.privacyexchange.org, and [3, 4, 48]. The way privacy laws can restrict the use of consumer data will be illustrated by describing the privacy legislation in effect in Europe, where national and federal-state laws govern data-handling practices. Similar ideas are expressed in the Fair Information Practices (FIP), which were set down as a recommendation in the United States and updated by the Federal Trade Commission, and in the Guidelines on the Protection of Privacy and Transborder Flows of Personal Data of the Organisation for Economic Cooperation and Development (OECD) [23, 25, 41]. The P3P standard (discussed further below) was modeled on these principles.

Directive 2002/58/EC of the European Parliament and Council concerning the processing of personal data and the protection of privacy in the electronic communications sector and its extension for electronic data have been adopted in national laws in most European Union (EU) member states [21, 22]. *Personal data* comprises "any information relating to an identified or identifiable natural person . . .; an identifiable person is one who can be identified, directly

or indirectly, in particular by reference to an identification number or to one or more factors specific to his [*sic*] physical, physiological, mental, economic, cultural or social identity" (EU Directive 95/46/EC, Art. 2 (a)). When users can be identified with reasonable effort based on the data collected, privacy laws already apply. P3P therefore uses the term *identifiable data* to denote personal data. In addition, it distinguishes whether the data collector can perform the (re)identification only with assistance from other parties ("identifiable data") or also without them ("identified data"). For details, see Section 1.3 of the P3P specification [13].

With regard to their personal data, individuals must grant a data collector the right to collect, for a given purpose, certain *kinds of data* (those that are relevant to the purpose), and these data may be processed only for that *purpose* and that *recipient* (the data collector). All further uses again require the explicit consent of the individual. In other words, every combination of (*kind of data, purpose, recipient*) processing that has not been allowed by the concerned individual is forbidden (cf. Directive 95/46/EC, Article 7). Further details can be found in Appendix 2.

An important consequence of the EU privacy legislation for an analysis framework is that certain analytics are only allowed based on pseudonymous data. Indicators that require identified data must be blocked. In the listing of analytics in Table 1, each indicator is annotated by whether it requires identified or pseudonymous data.

The scope of privacy legislation often extends beyond the Web and also encompasses physical and organizational requirements. Thorough privacy and security can only be achieved if technical measures are supported by an organizational background and physical protection measures, and vice versa.

Contractual Restrictions: P3P Specifications

In addition to the mandatory legal restrictions, companies can impose further restrictions on their data-collection and data-usage practices. A company's privacy practices are typically posted as on-line privacy statements (also known as "privacy policies" or "privacy disclosures"). A more advanced approach to codifying a company's Web privacy practices is the Platform for Privacy Preferences (P3P). P3P is an XML (Extensible Markup Language) application with a twofold goal. First, P3P enables Web sites to express their privacy practices in a standardized and machine-readable format. P3P policy reference files and P3P policies can be retrieved automatically and interpreted easily by user agents. Second, P3P gives users control over the disclosure of their personal data to Web sites they visit, promoting trust and confidence in the Web. Web browsers support users in formulating their preferences.

P3P is an industry-supported self-regulation approach to privacy protection. It is a W3C recommendation as a protocol to communicate how a site intends to collect, use, and share personal information about its visitors. In 2004, P3P adoption was 33 percent for the top 100 Web sites and 22 percent for the top 500 Web sites [20]. A 2006 study indicated an increase in adoption rates [19]. The current development status of P3P is the Working Group Note of the P3P 1.1 Specification, published in November 2006 [13].

P3P-enabled browsers parse a site's privacy policy automatically and compare it to the privacy preferences of the visitor, who can then decide whether or not to use the service.[1] Once a P3P policy is set up on a Web site, it becomes a legally binding agreement predicated on notice and consent between the Web site and the user. Like the terms and conditions, privacy policies become part of the concluded contract irrespective of the way they were published (i.e., as plain text or as a P3P policy). In the United States, the Federal Trade Commission and several states have sued companies that did not adhere to their privacy policies for unfair and deceptive business practices.

To formulate a privacy policy in P3P, the Web site publishes a P3P policy containing different STATEMENTS. Each statement is the description of one scenario of how the collected data will be used. It comprises information on the type of DATA to be collected, the PURPOSEs for which the data will be used, how long the RETENTION time will be, and the RECIPIENTs to which the data will be disclosed. For each of the four privacy dimensions (data, purpose, retention time, and recipient), P3P offers predefined sets of values to describe the scenario. The type of DATA can be expressed on different granularity levels using a hierarchically organized base data schema. By defining new data schemas, sites can precisely specify the data they collect beyond the base data set. However, the base data set provides sufficient detail for the vast majority of current applications. In particular, if user agents are unfamiliar with issuer-defined elements in these schemas, they will be able to provide only minimal information to the user about the new elements. A STATEMENT thus formalizes the (kind of data, purpose, recipient) combination permission required by legal requirements or principles (see above). The language is designed to be extensible by further elements. An example (JURISDICTION) will be described below.

P3P requires companies to specify precisely what data are used for what purpose and who is allowed to use the data. However, P3P describes different combinations of DATA, PURPOSE, and RECIPIENT separately. This may lead to problems when data are combined.

P3P cannot constrain or modify existing privacy legislation. Thus, the use of P3P by itself does not constitute compliance with the EU Data Protection Directive, although it can be an important part of an overall compliance strategy [13]. The latest version of the P3P specification (Version 1.1, November 2006) includes a JURISDICTION extension element whereby a known URI as a unique identifier of a body of legislation can be inserted and then can be recognized and displayed by user agents. This element was introduced as a result of the W3C Workshop on the Future of P3P held in November 2002, opening the expressiveness of P3P toward the European privacy legislation. However, the semantics of JURISDICTION elements remain unclear.

Inference Problems

A problem not yet directly addressed in privacy policies is that data recombination after collection may allow new inferences. Inferences exploit the possibility of combining separate sets of identified and unidentified data. Even

if identity keys are not known, attributes from secondary data sources may uniquely identify a single person [16, 55]. Inference problems are also known as "reidentification" or "triangulation" problems.

Inference problems also occur when customer attributes are exchanged and matched with secondary demographic data. This matching is legal in the online domain if the user profile remains pseudonymous. An analysis becomes privacy-critical, however, if a pseudonymous ID is linked to secondary data that identify an individual person. For example, researchers who order specialized books in their fields of interest are likely to be identified by the ZIP code that indicates the location of their university or research institution. This may enable the data miner to find out who the customers are, especially in sparsely populated ZIP codes.

Inferences also depend on the secondary data available for data linkage. For example, Sweeney used officially available voter registration lists containing the attributes *name*, *ZIP code*, and *date of birth* to reidentify hospital patients [55]. Such data are not publicly available in every country. German federal law, for instance, prohibits access of third parties to voter registries (§17 I of the German Federal Electoral Law).

Besides the inference problems that have already been identified, there is an inherent risk that future inference problems may affect a company's analysis framework. Procedures for inferring countries and regions from IP addresses (e.g., www.maxmind.com) show that it is possible to infer higher-order data from seemingly arbitrary numerical IDs. Probabilistic inferences may be discovered from large amounts of data using data-mining methods.

The increasing public awareness of the privacy problems associated with Web data analysis has two consequences for companies (which are legally responsible for ensuring that the data they collect are processed only in permitted ways). On the one hand, they are becoming more and more aware of the need to avoid harmful data disclosures; on the other hand, they are also becoming aware that privacy protection is not trivial, even with the best intentions, as illustrated by the recent events around the AOL search-data disclosure [7, 36]. The first realization constitutes a certain incentive to compute analytics in-house. Hosted-analytics vendors must therefore ensure that their security measures and behavior are trusted. The second realization, in contrast, constitutes a clear incentive to use hosted solutions—if they have and can demonstrate expertise not only on the technical aspects of analytics computation, but also on current privacy threats, inference possibilities, and legal restrictions. Certification could help to establish trust in reliable vendors (analogous to [12]).

It should be pointed out that the present focus on a "safe" way of interpreting a company's P3P policy is only one piece of a comprehensive company policy for preserving customers' privacy. Equally important are policies that govern data-handling practices within a company and between a company and its business partners. Data-handling practices play a dual role in an analysis of possible privacy threats. They are necessary complements to P3P-focused frameworks (improper handling may subvert even a thoroughly P3P-compliant analytics framework). An in-depth analysis of this question is beyond the scope of the present paper, but it will be touched upon in the final section of this paper.

P3P Integration in Databases and Data-Mining Systems

There are several proposals for integrating P3P into database and data-mining systems. One work on limiting disclosure in "Hippocratic databases" proposes a database-centric way of selectively blocking access to data [35]. It offers a "table-semantics model" that conceptually defines a view of each data table for each purpose-recipient pair, based on the constraints specified in a privacy policy. Operationally, this approach rewrites SQL queries to respect disclosure limitations. In another paper the authors give an overview of an implementation that integrates the checking of users' APPEL preferences at the server [2]. A third work outlines an extension of P3P and APPEL that allows the data owner to specify restrictions on whether and how mining results should be delivered to specified recipients [26].

The present contribution is related to but also differs from these approaches. The filter defining executable indicators that is described further below can be regarded as an extended version of the table-semantics model described above [35], except that (a) it integrates the inference problem to block the access to data, and (b) it allows for non-privacy-infringing aggregates. The latter means that evaluations of otherwise private attributes are allowed under anonymization or pseudonymization, which is not supported in the approach of [35]. The present proposal aims to ensure that *all* users are indeed protected at the level promised by the company's P3P policy (with additional blocking according to *individual* users' preferences being an extension for future work). Finally, [26] only considers the *results* (rather than the *input*) of an analysis. The present proposed approach takes the same perspective in its focus on whether the computation of an indicator is allowed, but it also considers the path of this computation and the data used in it.

To the best of the authors' knowledge, no currently existing formalism or tool addresses the consequences of inference problems and legal restrictions on the input data for P3P STATEMENTs, integrates the considerations described above into an analytics-computation framework that uses and extends P3P, and comprises a software architecture, a business model, and an implementation. The problem analysis and solution approach proposed in the following two sections are intended to close this gap.

Problem Statement: The Purpose/Inference Problem

In summary, certain purposes are allowed in the analysis of certain data, and the data may be used for this purpose by certain recipients. These relationships can be expressed in P3P STATEMENTs. However, the basic relational framework of P3P cannot account for inferences that may *substitute* certain data.

Regardless of whether their use is permitted, data are or are not *technically available*, and for each indicator, certain data are *required.* Legal regulations may *restrict* data usage. These relations, summarized in Figure 1, constitute the problem specification for the development of a privacy-protecting analysis framework.

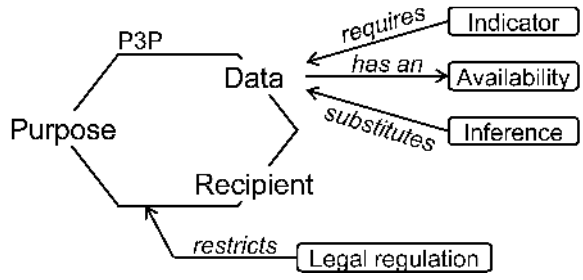

Figure 1. The Purpose/Inference Problem: Overview

Data-retention time is omitted in the problem statement, on the assumption that the data are present when the analysis is carried out. Data purged from the data collector's databases to comply with expired retention times are *no longer available*, so this issue is covered by the technical availability. However, problems may arise in longitudinal analyses that span a timeframe longer than the retention period. In consequence, a given attribute will not be available for all records.

The discussion in the following two sections formalizes the purpose/inference problem.

Data Types and Relations

It is necessary to distinguish between input data (data worked on by the indicator computation) and process data (data the process works with). The input data are Web log and Web user data (e.g., purchase, socioeconomic, geographic, and other data) together with the enterprise's privacy policy. Physically, this policy consists of the P3P file. The process data embody the business logic that defines the whole analysis process. Operationally, the process data consist of indicator definitions and additional technical information.

Input Data

The input data describe the *data items*, *purposes*, and *recipients* as well as the relations between them. The input data consist of three sets: the set of basic data elements D, the set of purposes P, and the set of recipients R. These are the same entities as are defined in P3P. Note that all these sets are enumerable:

D = {dynamic.clickstream.uri.authority, . . . , business.contact-info.telecom.online.uri}
consists of all leaves of P3P's hierarchical data schema structure. The elements of D are atomic data elements. Hence, "user.name" is not part of D, because it has descendants like "user.name.prefix" and "user.name.given." Note that D can be extended by the

issuer of the policy. Furthermore, D_{set} is defined as the subset of D stated by the P3P policy. Hence, D_{set} is an element of the power set over D: $D_{\text{set}} \in \mathcal{P}D$.[2]

P = {current, admin, . . . , other-purpose} is the set of the 12 relevant purposes as defined for the PURPOSE element.

R = {ours, delivery, . . . , other-recipient} is the set of the six possible values of the RECIPIENT element.

These two sets are not extensible. $P_{set} \subseteq P$ and $R_{set} \subseteq R$ are defined analogously to D_{set}.

A P3P STATEMENT establishes a relation between elements belonging to these three groups by assembling the DATA, PURPOSE, and RECIPIENT elements.

Process Data

The analysis framework introduces two new data entities for the process data. The first is the set of business indicators, which is formed by all the indicators that can be calculated from the present data. Currently, this set is fixed, but it may become extensible, because users can register third-party plug-ins for new indicators, such as industry-specific metrics. The second entity is the technical availability $A = \{true, false\}$. A indicates whether an instance of data is physically stored in the enterprise and can be made available to the analysis process. Note that this availability is defined purely technically. No privacy aspects are considered at this point.

Functional Data Relations

The functional data describe the relations between the *availability*, the *indicators*, and the *data items*. Before analyzing the functional relations between the different data, it is necessary to introduce the notation of the functional relationship [42]. A function f is a triple $(\mathfrak{D}_f, \mathfrak{W}_f, \mathfrak{R}_f)$, formed by a domain $\mathfrak{D}_f$, a range $\mathfrak{W}_f$, and a relation $\mathfrak{R}_f$, the function graph. This function graph has to be injective; that is, there are no two pairs $(a, b_1) \in \mathfrak{R}_f$ and $(a, b_2) \in \mathfrak{R}_f$ with $b_1 \neq b_2$. The function f maps the argument value x to the result value $y = f(x)$ if the pair $(x, y) = (x, f(x))$ is part of the function graph: $(x, y) \in \mathfrak{R}_f$. A given function $f = (\mathfrak{D}_f, \mathfrak{W}_f, \mathfrak{R}_f)$ is called partial if $\pi_1(\mathfrak{R}_f) \subset \mathfrak{D}_f$, where π_1 is the projection defined as $\pi_1(A \times B) = A$. Otherwise—that is, if $\pi_1(\mathfrak{R}_f) = \mathfrak{D}_f$—$f$ is called total.

By combining sets of data, recipient, and purpose, P3P statements can be interpreted as a triple of $(D_{set}, R_{set}, P_{set})$. For ease of expression, this triple of sets is expanded to a set of triples $(d \in D, r \in R, p \in P)$ and the function h is defined with $h: D \times R \times P \rightarrow \{allowed\}$.[3] This function is (usually) partial, because not all purposes are allowed to everyone for all the data. In the following, the function k is used with $k(x) = h(x)$ if $x \in \mathfrak{R}_h$ and $k(x) = not\ allowed$ otherwise. k is total; k: $D \times R \times P \rightarrow \{allowed, not\ allowed\}$.

Example: consider a statement as a fragment of a P3P file, such as the following excerpt from Example 4.1 in the work by Cranor, Wenning, and Shunter [13]:

```
. . .
<STATEMENT>
  <PURPOSE><individual-decision required="optout"/></PURPOSE>
  <RECIPIENT><ours/></RECIPIENT>
  <RETENTION><stated-purpose/></RETENTION>
  <DATA-GROUP>
    <DATA ref="#user.name.given"/>
    <DATA ref="#dynamic.cookies"> . . . </DATA>
  </DATA-GROUP>
</STATEMENT>
. . .
```

This fragment defines the following elements of $\mathcal{R}_k$: ((user.name.given, ours, individual-decision), allowed), ((dynamic.cookies, ours, individual-decision), allowed).

There are two more functions that establish relations: the function *requiredFor:* $D \times I \rightarrow$ {*true, false*} defined on the data D, and the indicator I that states whether a data item is used within the calculation of an indicator. The function *isAvailable:* $D \rightarrow A$ indicates whether a given data item is available. By definition, *isAvailable*(<>) = *true* where <> indicates "no data."

As all the sets are enumerable, the functions *k, requiredFor*, and *isAvailable* can be defined pointwise for all elements. They are deterministic. Extensions of D require an extension of all three function graphs.

Functions and Work Processes

Given the set of all possible indicators, the subset "executable indicators" is the set of all the indicators that can be executed. Each business analysis in the framework is an indicator I. So, $I \supseteq I_{executables} = t(I)$, where the function t selects all executable indicators from I. This section will provide the definition of t. In other words, t acts as a filter on I.

Whether a given indicator $i \in I$ is executable or not depends on two requirements—its execution has to be feasible, and its execution must be allowed. With respect to an implementation of the framework, it is reasonable to check in this order because the check for technical requirements is usually simpler.

The first technical requirement is the presence of the definition of this indicator (the implementation has to know how to calculate it). We presume that this is always guaranteed. The second technical requirement is the presence of the data needed for its calculation, that is, $isAvailable(d_j)=true\ \forall d_j$: $requiredFor(d_j, i)=true$. In summary, both the process data and the input data have to be available for the computation to be successful.

The restrictions imposed by the privacy policy are expressed by k. The computation of an indicator i is allowed if $k(d_j, r, p)=allowed\ \forall d_j$: $requiredFor(d_j,$

i)=*true*, where $r \in R$ and $p \in P$ have to be specified by the analyst. In practice, *r* has to include "ours." There is no fixed relation between purpose and indicator, because the calculation of a given indicator can have multiple purposes. Moreover, purposes in P3P are defined as *secondary purposes.* The last working draft introduced *primary purposes*, but there was no link to business processes or business objectives. The lack of such relations is a serious problem for the privacy-compliant analysis of consumer data.

For the purposes of the present discussion, $t_{r,p}$, the filter for the indicators that are executable given a recipient and a purpose, is defined as a composition of functions already known (<> is "no analysis"):

$$t_{r,p}(i) = \begin{cases} & \text{if } \forall dj: requiredFor\left(d_j, i\right) = true: \\ i & isAvailable\left(d_j\right) = true \wedge \\ & k\left(d_j, r, p\right) = allowed \\ <> & \text{otherwise.} \end{cases}$$

The discussion so far has assumed that a company stores its data with attribute names and level of aggregation as defined by the P3P base data schema. In real systems, this assumption is usually not fulfilled. Additional matching and transformation of data (including renaming, aggregation, or disaggregation) are necessary. But as this is only a question of naming and storing, it has no impact on the theoretical process of decision-making.

Impact of Data Inference on Decision-Making

An inference is defined as a function $s: \mathcal{P}D \rightarrow D$. If there is an inference, then we can write $s(\{d_1, d_2, \ldots, d_n\}) = d_{n+1}$ with $d_i \neq d_j \Leftrightarrow i \neq j$. Trivial inferences like the projection $\pi_1(d_1, d_2) = d_1$ are thereby excluded. The existence of inferences is a problem for the decision on whether an indicator can be calculated or not.

As an example, consider two data items for which the same restrictions on purpose and recipient apply: (d_1, r_1, p_1) and (d_2, r_1, p_1). Moreover, there is an inference such that $s(\{d_1,d_2\}) = d_3$. For d_3, the following purpose limitation applies: (d_3, r_1, p_3) *with* $p_3 \neq p_1$, in other words, all purposes other than p_3 are disallowed. Note that all intended purposes have to be indicated in a P3P policy file, which means that any purposes not contained in the file are disallowed. Consider an indicator *i* that the recipient r_1 wants to use for the purpose p_1 which requires the data d_3. Calculating this analysis by d_3 directly is prohibited by the P3P policy. However, it is possible to calculate the indicator from d_1 and d_2. Thus, inferences may bypass privacy restrictions.

Site users who accept the policy are not protected against this violation of their privacy preferences unless they employ a user agent that (1) is aware of this inference possibility and (2) extends the usage restriction to also cover inferred data.

How Are Inferences Established?

The possibility of inferences can arise in two ways: out of definitional/functional dependencies and out of statistical dependencies. The first type of dependency concerns data that by definition, or by the factual relations in a database, depend on one another (the intensional and extensional facets of functional dependence are discussed in [50]). For example, one can infer the country from the international telephone prefix of a telephone number. Such dependencies can generally be obtained from publicly available data sources, and they are comparatively stable over time and over different populations.

The second type of dependency arises from statistical relations within a data set, so inferences generally have a certain error bound. (Extensional functional dependency can be seen as a limiting case.) For example, the popularity of first names changes with time. This knowledge is exploited by data analysis companies like microm (microm Micromarketing-Systeme und Consult, www.microm-online.de), which claims to be able to predict a person's age from his or her first name with high accuracy. Such inferences can be inferred from a classifier learned from a database containing both names and birth dates or from public sources [27]; the accuracy of the classifier (or some other suitable measure of prediction quality) is the certainty with which the inference can be made. Such dependencies often need access to non–publicly available data, and they may not be stable across different populations. For example, age prediction is likely to be less accurate in a population born in a time when first-name popularities were relatively stable than in a population born in a time when first names went through short-lived fashions.

Probabilistic inferences are an important current research topic in data mining. However, their legal meaning and implications, and thus their role for statements made in P3P, are only beginning to be understood. Thus they will not be considered any further in the present work.

Impact of Legal Restrictions

As has already been pointed out, laws impose restrictions on using data. These restrictions are usually independent of recipient and purpose and may allow two uses only if separate and not simultaneous [22]. For instance, the combined usage of personal data and clickstream data is prohibited by German legislation. The Data Retention Directive 2006/24/EC of the European Union envisages long-term storage of telecommunication usage data (e.g., e-mail sending times, telephone call durations), *separated* from content data (e.g., the e-mail body or the messages exchanged by telephone) and user data (e.g., the name of the e-mail sender or the caller).

As an example, consider two data items d_1 and d_2 that can be used by a given recipient r_1 for a given purpose p_1. The separate usages of d_1 and d_2 are in accordance with the law and can be coded by two statements in P3P: (d_1, r_1, p_1) and (d_2, r_1, p_1). The indicators based on purpose p_1 will therefore be in $t(I)$.

However, the combined usage of the data elements d_1 and d_2 may conflict with legal requirements. Simultaneous use is generally subject to higher hurdles.

Technically, this imposes a relation on data. In the simplest case, it will be a binary relation such that $notSimultaneous(d_1,d_2)$ means that if d_1 is used, d_2 may not be used simultaneously.

Analysis Tool Development

The formalized problems can be addressed through a specially developed tool that allows data analysts to calculate a set of business analytics as described in the foregoing analysis framework without violating the given privacy restrictions. The description of the tool's development is structured by the usual phases of software development processes, proceeding from the business model via design to implementation.

Business Model

The main function of the proposed privacy-protecting analysis service is to calculate those Web indicators in the framework that are not restricted, given a set of privacy constraints and data elements. If a site is P3P-enabled, the analysis service automatically parses the specifications about collected data, purpose, and jurisdiction, and indicates potential restrictions when indicators are calculated. Although the P3P policy only indicates (probably) available data on the intensional level, parsing the database scheme provides information on the actual intensional level. The availability of a specific attribute for a set of records has to be checked on the extensional level.

If the site is not P3P-enabled, manual specifications are required. The service's business model consists of three actors and three core transmissions of data:

1. The original data holder is a visitor to a Web site. The data holder transmits data ($(d_1, \ldots d_n)$ in the framework above) to the Web site, usually a retailer, who is the data collector.
2. The data collector requests the analysis of these data subject to privacy restrictions from the analysis service provider.
3. The analysis service provider returns the analysis results $I_{executables}(d_1, \ldots d_n)$ to the data collector and indicates the privacy requirements.

Note that the tool does not protect the consumer from deliberate privacy violations by the retailer or the service provider. It only supports the data analyst in calculating allowed indicators and in recognizing potential privacy conflicts and possible usage purposes that must be respected. Thus, the business model requires that all parties must be trusted.

Standards for secure communication, such as a Secure Socket Layer (SSL) [53], are integrated in the framework. Further security questions, such as at-

tacks by a malevolent hacker or employee, are not within the scope of this paper.

The service could be offered for a per-service fee or as a renewable or permanent license or subscription. The business model could be enhanced by comparing analysis results between participating companies to create and sell benchmark reports for specific industries. Naturally, calculating these benchmarks is also subject to applying privacy restrictions but nonetheless is possible on nonidentifiable data (see sections 1.3.2 and 3.3.4 in [13]).

While this third-party architecture is advantageous, it could be modified such that the analyses run on the data collector's side, with update installations offered when necessary. This would avoid the necessity of transferring data, which under certain circumstances may be a problem for privacy protection.

Design

An extension of P3P based on the purpose/inference problem and the main data types and relations, functions, and work processes already defined will now be proposed. The P3P extension addresses inferences and legal restrictions.

Additional elements can be included in a policy by the element EXTENSION [13].

An unordered list of inference statements is here suggested. Each INFERENCE statement consists of the data that can be inferred if a given set of data is present. A human-readable explanation can be added within the CONSEQUENCE element.

From a given premise, it may be possible to conclude n consequences. This is expressed as n separate INFERENCE statements, each with an atomic consequence. In addition, one may want to express an inference possibility, such as "if d_1 and (d_2 or d_3) are given, then it is possible to infer d_4." This may be split into two statements: "if d_1 and d_2, then d_4" and "if d_1 and d_3, then d_4." However, the introduction of the connector OR in addition to AND makes the formulation and reading of inferences easier for human users and allows compact expression of complex triangulation patterns.

DATA-GROUPs can be placed within one of these elements to express the logical relations between them. The following fragment shows an example.

```
. . .
<EXTENSION optional="no">
<INFERENCES xmlns="http://preibusch.de/namespaces/SIMT/
inferences">
  <INFERENCE>
    <CONSEQUENCE> If the zip code and the birth date are
                  known, the home address can be reconstruct-
ed.
    </CONSEQUENCE>
    <GIVEN>
      <AND>
```

```
      <DATA-GROUP>
        <DATA ref="#user.home-info.postal.country"/>
        <DATA ref="#user.home-info.postal.stateprov"/>
        <DATA ref="#user.home-info.postal.postalcode"/>
        <DATA ref="#user.bdate"/>
      </DATA-GROUP>
    </AND>
  </GIVEN>

  <INDUCED>
    <DATA-GROUP>
      <DATA ref="#user.home-info.postal.street"/>
    </DATA-GROUP>
  </INDUCED>
 </INFERENCE>
 <INFERENCE>
  <CONSEQUENCE> The international telephone code can be
                reconstructed from the name of the
                country, and vice versa.
  </CONSEQUENCE>
  <GIVEN>
    <OR>
      <DATA-GROUP>
        <DATA ref="#user.home-info.postal.country"/>
      </DATA-GROUP>
      <DATA-GROUP>
        <DATA ref="#user.home-info.telecom.telephone.
          intcode"/>
      </DATA-GROUP>
    </OR>
  </GIVEN>
  <INDUCED>
    <DATA-GROUP>
      <DATA ref="#user.home-info.postal.country"/>
      <DATA ref="#user.home-info.telecom.telephone.intcode"/>
    </DATA-GROUP>
  </INDUCED>
 </INFERENCE>
</INFERENCES>
</EXTENSION>
. . .
```

As this extension adds further restrictions to the policy, it is mandatory.

In accordance with the W3C specification of P3P, an INFERENCES extension is defined using the Augmented Backus-Naur Form (ABNF) notation [14], which can be found in Appendix 1.

Coding Legal Restrictions in a P3P Policy

Whereas the STATEMENTs in a policy file allow using the data within the specified borders, legal specifications always restrict uses. *A priori*, one can

say that any legal restriction can be coded in a P3P policy by listing all allowed uses. Thus, the missing uses are prohibited. But this realization does not respect the simultaneity restriction that the STATEMENTs in P3P have no temporal semantics. Two statements can be merged into a single statement only on the basis of the comprised details on recipient, purpose, data-group, and retention time [13].

Consider once again the example given earlier—two data items d_1 and d_2 that can be used by a given recipient r_1 for a given purpose p_1. The separate usages of d_1 and d_2 are in accordance with the law and can be coded by two statements in P3P: (d_1, r_1, p_1) and (d_2, r_1, p_1). According to the statement-merging rules of P3P, these two statements are semantically equivalent to the single statement $(\{d_1, d_2\}\ r_1, p_1)$. However, the combined usage of the data elements d_1 and d_2 conflicts with legal requirements.

Hence the introduction of a new element LEGAL is suggested that restricts combined usage in order to remedy this lack of P3P. The proposed extension of the JURISDICTION element in the P3P Working Draft in February 2006 [13] does not address the deficiencies mentioned above. P3P's JURISDICTION element can be added to annotate the regulatory environment in which a certain recipient is placed. The URI of jurisdictions can be indicated, but the regulation's semantics are not included in the policy. Hence, a machine-based evaluation in the context of the given policy is hampered.

Within the proposed LEGAL element, several RESTRICTION elements can be specified. Each RESTRICTION can have three attributes; the introduction of additional attributes or values needs to be discussed. The ISSUER attribute specifies the name of the legal authority that codified the restriction. The LAW attribute contains the name (possibly shortened) of the legal norm that is the origin of this restriction. The values of both attributes are free-form texts. The FOR attribute indicates the region the site user must belong to for this restriction to be applied. Possible values are space-separated combinations of the ISO country abbreviations, such as "US" for the United States of America, "GB" for the United Kingdom, and "DE" for Germany, as specified in the ISO 3166–1 norm [28], also allowing for the abbreviation "EU" for all member states of the European Union. If the attribute is omitted or its value is left blank, the restriction is assumed to hold regardless of national boundaries.

Finally, the empty tag VICEVERSA summarizes the repetition of the same restriction with reversed WHILE and DONT elements:

```
<RESTRICTION>
  <VICEVERSA />
  <WHILE> A </WHILE>
  <DONT> B </DONT>
</RESTRICTION>
```

is equivalent to:

```
<RESTRICTION>
  <WHILE> A </WHILE>
  <DONT> B </DONT>
</RESTRICTION>
```

```
<RESTRICTION>
  <WHILE> B </WHILE>
  <DONT> A </DONT>
</RESTRICTION>
```

The main elements are WHILE and DONT. Both of them contain a single DATA-GROUP with one or more DATA elements. The use of all the DATA in the DONT element concurrently with the DATA in the WHILE element is not allowed. As this extension adds further restrictions to the policy that cannot be ignored, it is a mandatory extension. Backward compatibility of extended P3P policies is assured by using an extension mechanism of the P3P policy reference file and the referenced P3P privacy policies [42].

Consider the following fragment of an extended P3P policy as an example:

```
<RESTRICTION>
  <WHILE>
    <DATA-GROUP>
      <DATA ref="#dynamic.clickstream"/>
    </DATA-GROUP>
  </WHILE>
  <DONT>
    <DATA-GROUP>
      <DATA ref="#user.home-info.postal"/>
      <DATA ref="#user.business-info.postal"/>
    </DATA-GROUP>
  </DONT>
</RESTRICTION>
```

The fragment states that the simultaneous use of either home address, business address, or both is not allowed while using clickstream data.

The ABNF definition of the LEGAL element is shown in Appendix 1.

Within the RESTRICTION element, a CONSEQUENCE element can be defined, as it is defined in the P3P specification and also used for the extension by INFERENCE.

The following fragment shows an example of the P3P extension using the LEGAL element:

```
<LEGAL>
  <RESTRICTION
   issuer="European Commission"
   law="EU Privacy Directive"
   for="EU">
    <VICEVERSA/>
    <CONSEQUENCE> Information about site usage is not allowed
                  to be combined with identifiable personal
                  user data.
    </CONSEQUENCE>
     <WHILE>
      <DATA-GROUP>
```

```
      <DATA ref="#user.name"/>
    </DATA-GROUP>
  </WHILE>
  <DONT>
    <DATA-GROUP>
      <DATA ref="#dynamic.clickstream"/>
      <DATA ref="#dynamic.http"/>
    </DATA-GROUP>
  </DONT>
 </RESTRICTION>
</LEGAL>
```

As the same legal restrictions and inference rules apply for a large variety of Web sites, mechanisms that include a set of referenced legal restrictions hosted by a trusted provider (e.g., governmental authorities) should also be developed.

Workflow

Figure 2 summarizes the processes within the framework, including the successive data exchanges and actions between the involved participants. Inter-unit exchanges rely on standardized protocols and data-description formats (see below for details). Note that the framework includes the extensions for legal restrictions and inference problems.

User Interface

This section gives a (nonexhaustive) technical description of the user interface.

The analysis service has three specification phases. Future releases will support automated data retrieval and policy parsing. In each of the three phases, the analyst is given a specific task. Input errors are directly reported.

The first phase is the specification of the data the enterprise has stored—data availability is defined here. The second phase is the specification of the P3P privacy policy that applies to the data specified in the first step. The third phase is the selection of the analysis time frame and the desired business indicators. The list of grouped indicators is presented. The user interface only shows the indicators that are allowed given the available data and legal privacy restrictions. Other indicators are disabled and displayed in gray. The time frame (time interval of analysis) can be typed directly or chosen from a calendar control.

Once an indicator has been chosen, a set of three output formats are proposed depending on the type of analysis: output as HTML, as XML, or as an image. Images are generated dynamically using standard classes of the .NET framework. The analyst can handle this image like all other images, saving it, copying it, and so on. Image formats (PNG, GIF, JPEG, BMP, TIFF, etc.), colors, and fonts can be freely configured. Direct streaming avoids problems with

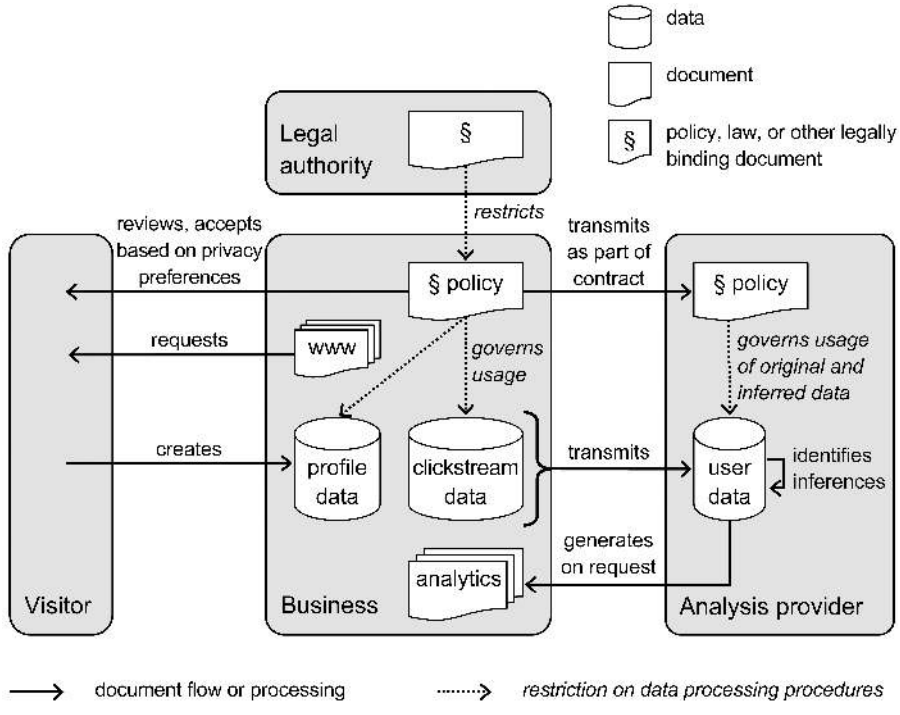

Figure 2. Dependencies and Workflow (High-Level View). Note That Visitors (Should) Create Data Only *After* They Have Reviewed and Accepted the Business' P3P Policy

asynchronous page requests, image generation, and image requests. Moreover, there are no problems with temporary files, such as manual administration of naming, creation, and deletion issues. No time lags were detected during the analyses based on the data of a large on-line retailer. The image generation "on the fly" does not slow down the output flush. Other implementations relying on the same technologies have experienced similar results.

Figure 3 shows a screenshot of the analysis tool user interface (phase 3 of the specification process).

The screenshot illustrates the main features of the architecture and user interaction:

- **Operation:** The Web-based interface is accessible on every computer. No local software installation is needed. (The screenshot is a browser window.)
- **Customization:** Individual accounts provide customization. In each account, the analyses portfolio assembles chosen indicators for future sessions (*top right*). The user is assisted by guided tasks (*top left*). Several analysis processing options can be configured to fit specific business needs (*bottom right*).

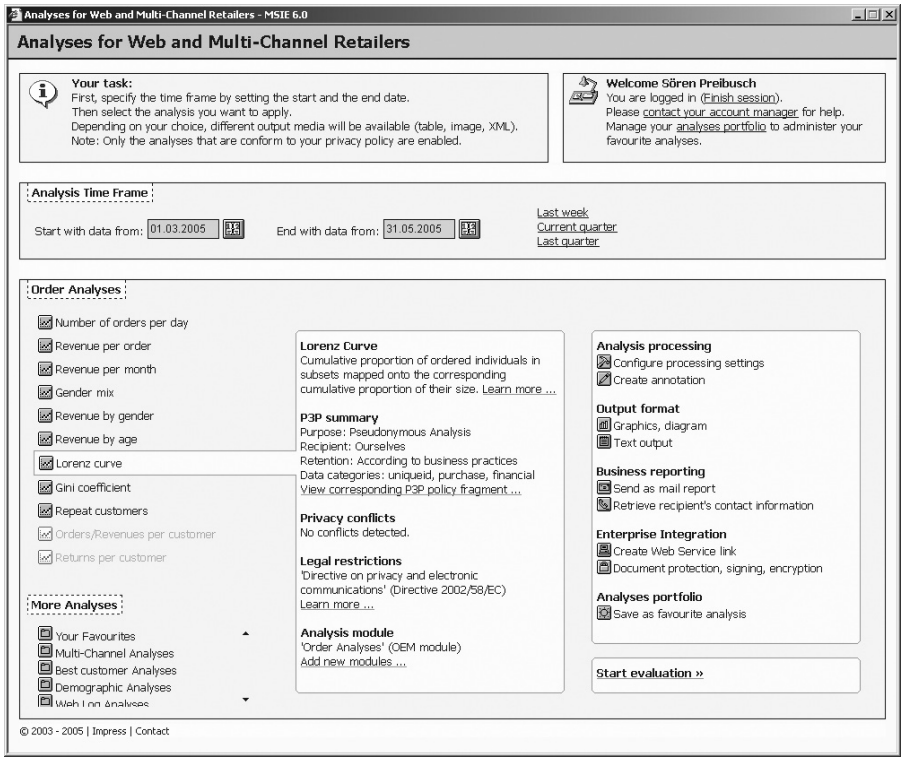

Figure 3. Analytics Management Interface

- **Data:** The time frame can be set. It indicates which data are taken into account (*middle*).
- **Indicators:** The group constituting the account's chosen indicators is opened by default. Available indicators can be selected by clicking. More than 80 indicators are grouped by topic, and third-party add-ons can be integrated (*bottom left*). The user is provided with a short description of the selected indicators, with more detailed information available on request (*bottom middle*).
- **Privacy information and preservation:** The P3P policy fragment that applies in the system's evaluation of a given intended analysis is automatically detected, and binding restrictions are listed. Applicable legal restrictions are listed for the user's notice (*bottom middle*). In case of privacy conflicts or missing data, the computation is prohibited by the system, shown by disabled options (*bottom left*), and the user gets a detailed summary (*bottom middle*).
- **Integration into other processes and other features:** Processing results can be presented in different formats, including XML for post-processing and integration into other applications. Collaboration features are available. Enterprise integration features enable data analyses in a service-oriented environment. Electronic signatures and

content encryption prevent unauthorized report manipulation and protect the results (*bottom right*).

Implementation Notes

The tool is a Web-based application written in C# in Microsoft .NET. The Web server dynamically generates Web pages to interact with the analyst, who is not required to install additional client software. This architectural design is one requirement for the analysis to be offered as a Web-based service, as described in the business model. All browser types are supported as long as they adhere to the ECMAScript standard (by implementing JScript or JavaScript).

Two databases are involved in the analysis process. The first is a Microsoft (MS) Access database providing the complete preprocessed Web data to be analyzed. The second is an MS SQL Server database that holds the process data.

Data can be exchanged by electronic transmission (e.g., download) or by transfer of physical storage media. If the data collector trusts the service provider and legislation does not restrict the use of certain data for analysis purposes, the complete data set can be transmitted without modifications. The Web service standards SOAP and WSDL provide sophisticated constructs for coding complex types and arrays of values for transmission. In the unlikely case that the analysis service is mistrusted or if legal policies restrict the use of certain data for analysis purposes, the retailer must protect confidential information before the data are transmitted.

Encryption techniques for protecting sensitive data in a two-party business relationship have been proposed in the literature. The basic idea is to leave the data on the data collector's server and transfer only encrypted data to the service provider [18, 47]. A hardware-based approach to encryption has been described that uses a secure coprocessor [5]. Encryption functions are useful for a limited number of algorithmic operations, such as addition, subtraction, multiplication, and inverse multiplication, and for basic database queries, such as selection, projection, and join [10]. However, current encryption techniques are not suited for more complex mining queries, such as those in the present analysis framework.

According to the P3P Guiding Principles, measures have been taken to implement mechanisms for protecting any information that is transferred from the analyst to the tool, and vice versa [13]. HTTP over a high-SSL encryption is used as a trusted protocol for the secure transmission of data. Restrictive session timeouts prevent the abuse of foreign sessions. Analysts have to log on with a personal password, and temporary session cookies are used to prevent other analysts from "stealing" a session.

Case Study

The analysis framework was tested on data from a large European multichannel retailer. The company operates an e-shop and a network of 700 special-

ized physical retail shops in more than 10 European countries and more than 6,000 associated independent retailers. Transaction information from 13,000 customers in Germany who bought over a time period of eight months was analyzed. Web usage data from the same time period were imported into the tool's database. Demographic data based on ZIP codes was purchased and also joined to the database. The data model is depicted in List 1.

The results were valuable for the company in many ways. The conversion analytics showed the site's success at transforming visitors to buying customers [57]. Recommendations for Web site design have been derived based on the results. The conversion analytics imposed no privacy threats under Germany's privacy legislation because no data revealing a person's identity are required—not even an IP address (annotated as "N" in Table 1).

The service analytics gave the company valuable insight into customers' payment, distribution, and return preferences [57]. The results confirm the retailer's strategic focus on multichannel retailing.[4] Some service analytics could be calculated without using identifiable or pseudonymous data (e.g., in-store pickup ratio, on-line payment ratio). However, when repeat customers' behavior was of interest, then at least pseudonymous information was required to match the transactions of the same customer over time (annotated as "P" in Table 1). Finally, some analytics required person-related information ("I") (e.g., to measure the recency of an individual's subsequent orders).

The tool automatically matches all analytics chosen by the retailer against the privacy restrictions set forth in the retailer's legislative environment and the P3P specifications. Analytics that are not allowed cannot be calculated with the tool. For example, the multichannel retailer was not allowed to calculate all analytics requiring individual data because of the restrictions set forth in European privacy legislation. Potential inference conflicts and data retention were also checked.

The application of the analysis framework to an on-line retailer's consumer data revealed privacy problems in a real-world context. Potential inference problems and legislative privacy implications were identified and could be addressed within the analysis framework.

Summary and Outlook

A framework for deploying Web analytics has been set up and tested on data from a multichannel retailer. The different data types involved in the data-analysis process and the functional relations between them have been identified. An automated way of filtering business indicators according to privacy restrictions has been presented. The proposed extensions of the Platform for Privacy Preferences specification allow for the coding of both data inferences and legal usage restrictions.

Further research will have to be done before the proposed extensions can be included in the P3P specification. It is envisioned that service providers will propose different privacy policies for their customers from which they can choose one that accords with their desired level of privacy. Individually negotiated privacy policies constitute a new challenge for large-scale data

analyses [45]. A matching and combination of service provider's policies, site-user-defined P3P preferences, and legal restrictions will help to protect customers' informational self-determination [8].

The integration of time will be an important extension of the framework. Dynamics enter the framework when the company using the analysis provider (the "retailer") can change its P3P policy such that certain uses now become *allowed* or disallowed. Users must be notified, and they must accept the changes just as they accepted the previous policy. (It is assumed here, as throughout the paper, that all this has happened in a legally correct way.) This presents no problem for the legitimacy of old analysis results, and no problem for any indicator that uses data collected after the change of privacy policy. For new analyses, the new P3P policy immediately applies. This policy is fetched from the retailer's server each time analyses are computed, thereby ensuring that it is up-to-date. The new data are used in accordance with the current rules.

A problem can arise only in the (empirically infrequent) case in which the user has granted permission to use certain data "as of now" for a new purpose, that these data have already been collected for a different purpose and are therefore technically available, and that the new purpose is an indicator whose computation also involves past data (e.g., an average value over the last six months). In this case, the indicator should not be computed immediately (in the example, starting only at the end of month 6 after the change in policy). In the current formalism, this (temporary) problem can be solved by renaming the data something like "turnover_jan" or "turnover_feb," and defining different (D, R, P) tuples on them. In a longer-term perspective, however, the formalism should be extended to include time.

Another important future step in the development of the framework will be a more comprehensive view of data analysis and data-handling practices. Privacy policies coded in P3P cover interactions between the service provider and its customers. Data flows between the service provider and its business partners (e.g., logistics) and intra-organizational data processing must be governed by enterprise policies. For these purposes, IBM has proposed EPAL, a formal language for writing enterprise privacy policies to govern data-handling practices in IT systems according to positive and negative authorization rights [44]. EPAL concentrates on the core privacy authorization while abstracting from all deployment details, such as data model or user-authentication. Compatibility with P3P is currently only described as a draft on a conceptual level.

EPAL targets intra- and inter-organizational data-handling practices, but enforcement of policies across multiple applications and over organizational boundaries is still a challenge, mainly due to incompatible deployment descriptions. The present framework envisions a supplementary implementation of EPAL policies by the data collector and the analysis service provider to govern internal data-processing processes and data transfers.

Another interesting extension is the combination of the present approach with privacy-preserving data mining (PPDM) [37, 56]. PPDM develops algorithms that modify the original data such that the private data and private knowledge remain private even after the mining process [60]. It takes a different approach to privacy protection: Whereas PPDM focuses on how to process

data, the present approach focuses on whether certain data may be processed (for a certain purpose) at all. PPDM addresses the inference problem: It modifies algorithms and/or architectures such that inferences become impossible or imprecise enough to avoid the reidentification of individuals. Thus, the success of PPDM hinges on two things: first, the formulation of algorithms for the relevant data-mining tasks, and second, the possibility of enforcing that everyone who has access to the data effects the required modification of algorithms and/or architectures. The present approach requires certain trust relationships. Some forms of PPDM do this too, but other (distributed) forms do not need this requirement. However, unlike the present approach, PPDM does not consider the requirement, made by data subjects or laws, that certain data may not be used for certain purposes. Therefore, the two approaches are complementary. The authors aim to explore this complementarity in future work.

NOTES

1. Several languages exist for expressing these preferences [1, 62].

2. $\mathcal{PS}$ denotes the power set over the set S. The power set is the set of all possible subsets of S. Sometimes $\mathcal{PS}$ is also noted as 2^S.

3. It is assumed that privacy policies are consistent and thus that a certain use is either allowed or not. The assumption is based on the fact that an inconsistent privacy policy must be interpreted in favor of the customer (EU Directive 93/13/EEC, April 5, 1993, on Unfair Terms in Consumer Contracts applies analogously). In the interest of the company, therefore, a privacy policy should not be inconsistent; and it can be assumed that the company lawyers have drawn up a consistent policy. Note that in P3P, only statements of the kind "It is possible that we will use these data to do this" can be formulated. In this sense, P3P helps to avoid certain inconsistencies.

4. The advantages of multichannel business strategies have been investigated in a number of articles [30, 54]; see also [24] for a study of consumer motivations for switching channels.

REFERENCES

1. Agrawal, R.; Kiernan, J.; Srikant, R.; and Xu, Y. An XPath-based preference language for P3P. In G. Hencsey and B. White (eds.), *Proceedings of the Twelfth International Conference on World Wide Web.* New York: ACM Press, 2003, pp. 629–639.

2. Agrawal, R.; Kiernan, J.; Srikant, R.; and Xu, Y. An implementation of P3P using database technology. In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, and E. Ferrari (eds.), *Proceedings of Advances in Database Technology—EDBT 2004.* Lecture Notes in Computer Science, 2992. Berlin: Springer, 2004, pp. 845–847.

3. Agre, P.E., and Rotenberg, M. *Technology and Privacy: The New Landscape.* Cambridge, MA: MIT Press, 1997.

4. Andrews, S. *Privacy and Human Rights 2002.* London: Electronic Privacy Information Center, 2002.

5. Asonov, D., and Freytag, J.C. Almost optimal private information retrieval. In R. Dingledine and P.F. Syverson (eds.), *Proceedings of the 2nd*

Workshop on Privacy Enhancing Technologies. San Lecture Notes in Computer Science, 2482. Berlin: Springer, 2003, pp. 239–243.
6. AT&T. AT&T Privacybird. www.privacybird.com.
7. Barbaro, M., and Zeller, T. A face is exposed for AOL Searcher No. 4417749. *New York Times,* August 9, 2006 (www.nytimes.com/2006/08/09/technology/09aol.html?_r=1&oref=slogin).
8. Barth, A., and Mitchell, J.C. Enterprise privacy promises and enforcement. In C. Meadows (ed.), *Proceedings of the 2005 Workshop on Issues in the Theory of Security.* Long Beach, NY: ACM Press, 2005, pp. 58–66.
9. Beal, B. Analyzing the CRM analytics race. SearchCRM.com, 2003, (http://searchcrm. techtarget.com/originalContent/0,289142,sid11_gci929770,00.html).
10. Boyens, C. Privacy trade-offs in Web-based services. Ph.D. dissertation, Humboldt-Universität zu Berlin, Institute of Information Systems, 2005 (http://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=25242).
11. Chevalier, K.; Bothorell, C.; and Corruble, V. Discovering rich navigation patterns on a Web site. In G. Grieser, Y. Tanaka, and A. Yamamoto (eds.), *The 6th International Conference on Discovery Science.* Lecture Notes in Artificial Intelligence/Lecture Notes in Computer Science, 2843. Berlin: Springer, 2003, pp. 62–75.
12. Cranor, L.F., and Reidenberg, J.R. Can user agents accurately represent privacy notices? In TPRC (ed.), *Proceedings of the Telecommunications Policy Research Conference 2002*. Caret, VA: TPRC (http://intel.si.umich.edu/tprc/archive-search-abstract.cfm?PaperID=65).
13. Cranor, L.F.; Wenning, R.; and Schunter, M. (eds.). *The Platform for Privacy Preferences 1.1 (P3P1.1) Specification: W3C Working Group Note 13 November 2006 (follow-up to the W3C Recommendation P3P1.0 of 16 April 2002).* Cambridge: World Wide Web Consortium (W3C), 2006 (www.w3.org/TR/P3P11).
14. Crocker, D., and P. Overel. Augmented BNF for syntax specifications: ABNF, RFC2234, IETF, 1997 (www.ietf.org/rfc/rfc2234.txt).
15. Cutler, M., and J. Sterne. *E-Metrics—Business Metrics for the New Economy.* Cambridge: NetGenesis Corp., 2000 (www.emetrics.org/articles/whitepaper.html).
16. Denning, D.E. *Cryptography and Data Security,* Reading, MA: Addison-Wesley, 1982.
17. Dinev, T., and Hart, P. Internet privacy concerns and social awareness as determinants of intention to transact. *International Journal of Electronic Commerce, 10*, 2 (Winter 2005/6), 7–29.
18. Domingo-Ferrer, J., and Herrera-Joancomarti, J. A privacy homomorphism allowing field operations on encrypted data. *Actas de las I Jornadas de Matemática Discreta i Algorítmica.* Barcelona: Universitat Politècnica de Catalunya, 1998 (http://citeseer.ist.psu.edu/245679.html).
19. Egelman, S.; Cranor, L.F.; and Chowdhury, A. An analysis of P3P-enabled Web sites among top-20 search results. In M.S. Fox and B. Spencer (eds.), *Proceedings of the Eighth International Conference on Electronic Commerce.* New York: ACM Press, 2006, pp. 197–207.
20. Ernst & Young. P3P dashboard report. May 2004 (www.privacyassociation.org/docs/sum04/1–4Tretick3.pdf).

21. EU. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data. *Official Journal of the European Communities, 1995* (November 23, 1995, No. L 281), pp. 31–55 (http://europa.eu.int/eur-lex/en/consleg/main/1995/en_1995L0046_index.html).
22. EU. Directive 2002/58/EC of the European Parliament and of the Council Concerning the Processing of Personal Data and the Protection of Privacy in the Electronic Communications Sector. *Official Journal of the European Communities, 2002* (July 31, 2002, No. L 201), pp. 37–47 (http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_201/l_20120020731en00370047.pdf).
23. Federal Trade Commission (FTC). Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress (May 2000) (www.ftc.gov/reports/privacy2000/privacy2000.pdf).
24. Gupta, A.; Su, B.; and Walter, Z. An empirical study of consumer switching from traditional to electronic channels: A purchase-decision process perspective. *International Journal of Electronic Commerce, 8*, 3 (Spring 2004), 131–161.
25. HEW. U.S. Department of Health, Education and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, Records, Computers, and the Rights of Citizens (1973), viii (www.epic.org/privacy/consumer/code_fair_info.html).
26. Hintoglu, A.A.; Saygin, Y.; Benbernou, S.; and Hacid, M.S. Privacy preserving data mining services on the Web. In S.K. Katsikas, J. Lopez, and G. Pernul (eds.), *Proceedings of TrustBus 2005.* Lecture Notes in Computer Science, 3592. Berlin: Springer, 2005, pp. 246–255.
27. Huschka, D.; Gerhards, J.; and Wagner, G.G. *Naming Differences in Divided Germany.* Berlin: Deutsches Institut für Wirtschaftsforschung. DIW Berlin: Research Notes 8 (www.diw.de/deutsch/produkte/publikationen/researchnotes/docs/papers/rn8.pdf).
28. International Organization for Standardization. *ISO 3166 Code Lists,* www.iso.org/iso/en/prods-services/is03166ma/02iso-3166-code-lists/index.html.
29. KDNuggets. Association for Knowledge Discovery and Data Mining forum: Data mining, knowledge discovery, genomic mining, Web mining, www.kdnuggets.com.
30. King, R.C.; Sen, R.; and Xia, M. Impact of Web-based e-commerce on channel strategy in retailing. *International Journal of Electronic Commerce, 8*, 3 (Spring 2004), 103–130.
31. Kobsa, A. Personalized hypermedia and international privacy. *Communications of the ACM, 45*, 5 (2002), 64–67.
32. Kohavi, R., and Parekh, R. Ten supplementary analyses to improve e-commerce Web sites. In T. Senator, P. Domingos, C. Faloutsos, and L. Getoor (eds.), *Proceedings of the 5th ACM WebKDD 2003 Web Mining for E-Commerce Workshop "Webmining as a Premise to Effective and Intelligent Web Applications" at SIGKDD'03.* New York: ACM Press, 2003, pp. 29–36 (www.acm.org/sigs/sigkdd/kdd2003/workshops/webkdd03/wkdd03-paper4.pdf).

33. LeClaire, J. Web analytics: Client-side, server-side or hosted? *E-Commerce Times,* March 21, 2006 (www.ecommercetimes.com/story/49264.html).
34. Lee, J.; Podlaseck, M.; Schonberg, E.; and Hoch, R. Visualization and analysis of clickstream data of online stores for understanding Web merchandizing. *Data Mining and Knowledge Discovery, 5*, 1/2 (2001), 59–84.
35. LeFevre, K.; Agrawal, R.; Ercegovac, V.; Ramakrishnan, R.; Xu, Y.; and DeWitt, D.J. Limiting disclosure in Hippocratic databases. In M.A. Nascimento, M.T. Özsu, D. Kossmann, R.J. Miller, J.A. Blakeley, and K.B. Schiefer (eds.), *Proceedings of the Thirtieth International Conference on Very Large Data Bases.* San Francisco: Morgan Kaufmann, 2004, pp. 108–119.
36. Li, K. AOL chief technology officer resigns: Sources. Reuters, August 21, 2006 (www.kdnuggets.com/news/2006/n16/36i.html).
37. Machanavajjhala, A.; Gehrke, J.; Kifer, D.; and Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. In IEEE (ed.), *Proceedings of the International Conference on Data Engineering (ICDE) 2006.* Los Alamitos, CA: IEEE Computer Society, 2006, pp. 24–35.
38. Malacinski, A.; Dominick, S.; and Hartrick, T. *Measuring Web Traffic.* Armonk, NY: IBM, 2001 (www-128.ibm.com/developerworks/web/library/wa-mwt1/ and www.ibm.com/developerworks/web/library/wa-mwt2).
39. Moe, W., Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream, *Journal of Consumer Psychology, 13*, 1 & 2 (2003), 29–39.
40. Moe, W., and Fader, P. Capturing evolving visit behavior in clickstream data. Working Paper, University of Pennsylvania, Wharton School, 2000 (www-marketing.wharton.upenn.edu/ideas/pdf/00–003.pdf).
41. OECD. Organization for Economic Cooperation and Development, Directorate for Science, Technology and Industry. *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* (www.oecd.org/document/18/0,2340,en_2649_34255_1815186_1_1_1_1,00.html).
42. Pepper, P. *Funktionale Programmierung in OPAL, ML, HASKELL und Gofer* [Functional Programming in OPAL, ML, HASKELL and Gofer], 2d ed. Berlin: Springer, 2003.
43. Piatetsky-Shapiro, G. Software for data mining and knowledge discovery. Association for Knowledge Discovery and Data Mining, 2006 (www.kdnuggets.com/software/index.html).
44. Powers, C., and Schunter, M. (eds.). Enterprise privacy authorization language (EPAL 1.2). W3C Member Submission, November 10, 2003 (www.w3.org/Submission/EPAL).
45. Preibusch, S., Implementing privacy negotiations in e-commerce, In X. Zhou, J. Li, H.T. Shen, M. Kitsuregawa, and Y. Zhang (eds.), *Frontiers of WWW Research and Development—APWeb 2006: 8th Asia-Pacific Web Conference.* Lecture Notes in Computer Science, 3841. Berlin: Springer, 2006, pp. 604–615.
46. Reinartz, W.J., and Kumar, V. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, 67, 1 (2003), 77–99.
47. Rivest, R.; Adleman, L.; and Dertouzos, M. On data banks and privacy homomorphisms, In R.A. DeMillo (ed.), *Foundations of Secure Computation.* New York: Academic Press, 1978, pp. 169–177.

48. Rotenberg, M. *The Privacy Law Sourcebook 2001: United States Law, International Law, and Recent Developments.* Washington, DC: EPIC, 2001.
49. SAP AG. *CRM Analytics.* Walldorf, Germany: SAP AG, 2001.
50. Schultz, R.A. Understanding functional dependency, In S.A. Becker (ed.), *Effective Databases for Text & Document Management,* Hershey, PA: Idea Group, 2003, pp. 278–287.
51. Spiliopoulou, M., and Pohle, C. Data mining for measuring and improving the success of Web sites. *Data Mining and Knowledge Discovery, 5*, 1/2 (2001), 85–114.
52. Srivastava, J.; Desikan, P.; and Kumar, V. Web mining—concepts, applications and research directions. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.), *Data Mining: Next Generation Challenges and Future Directions.* Menlo Park, CA: AAAI/MIT Press, 2004, pp. 405–423.
53. Stallings, W. *Cryptography and Network Security: Principles and Practice.* Upper Saddle River, NJ: Prentice Hall, 1999.
54. Steinfield, C.; Bouwman, H.; and Adelaar, Th. The dynamics of click-and-mortar electronic commerce: Opportunities and management strategies. *International Journal of Electronic Commerce*, 7, 1 (Fall 2002), 93–119.
55. Sweeney, L., Computational disclosure control: A primer on data privacy protection. Ph.D. dissertation, Massachusetts Institute of Technology, 2001 (www.swiss.ai.mit.edu/6805/articles/privacy/sweeney-thesis-draft.pdf).
56. Sweeney, L. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*, 5 (2002), 557–570.
57. Teltzrow, M., and Berendt, B. Web-usage-based success metrics for multi-channel businesses. In R. Kohavi, B. Liu, B. Masand, J. Srivastava, and O.R. Zaïane (eds.), *Proceedings of the 5th ACM WebKDD 2003 Web Mining for E-Commerce Workshop "Webmining as a Premise to Effective and Intelligent Web Applications."* New York: ACM Press, 2003, pp. 17–27 (www.acm.org/sigs/sigkdd/kdd2003/workshops/webkdd03/wkdd03-paper3.pdf).
58. Teltzrow, M., and Günther, O. Web usage metrics for multi-channel retailers. In K. Bauknecht, A.M. Tjoa, and G. Quirchmayr (eds.), *Proceedings of the Fourth International Conference, EC-Web 2003.* Lecture Notes in Computer Science, 2738. Berlin: Springer, 2003, pp. 328–338.
59. Teltzrow, M., and Kobsa, A. Impacts of user privacy preferences on personalized systems—a comparative study. In C.-M. Karat, J. Blom, and J. Karat (eds.), *Designing Personalized User Experiences for eCommerce.* Dordrecht: Kluwer Academic, 2004, pp. 315–332.
60. Verykios, V.S.; Bertino, E.; Fovino, I.N.; Provenza, L.P.; Saygin, Y.; and Theodoridis, Y. State-of-the-art in privacy preserving data mining. *SIGMOD Record, 33,* 1 (2004), 50–57.
61. Weigend, A. Analyzing customer behavior at Amazon.com. In T. Senator, P. Domingos, C. Faloutsos, and L. Getoor (eds.), *ACM Conference on Knowledge Discovery and Data Mining (SIGKK'03).* New York: ACM Press, 2003, p. 5.
62. W3C. A P3P Preference Exchange Language 1.0 (APPEL1.0). W3C Working Draft 15 April 2002 (www.w3.org/TR/P3P-preferences).

Appendix 1. ABNF Definitions

Augmented Backus-Naur form of the proposed extension INFERENCES (note: the elements CONSEQUENCE and DATA-GROUP are defined in P3P):

```
inferences = "<INFERENCES>" 1*inference "</INFERENCES>"
inference =  "<INFERENCE>"
             consequence
             given
             induced
             "</INFERENCE>"
given =      "<GIVEN>" logical "</GIVEN>"
induced =    "<INDUCED>" data-group "</INDUCED>"
logical =    or_set | and_set
or_set =     "<OR>" ((1*data-group)|logical) "</OR>"
and_set =    "<AND>" ((1*data-group) | logical) "</AND>"
```

ABNF definition for the proposed extension LEGAL (note: the element DATA-GROUP is defined in P3P):

```
legal =        "<LEGAL>" 1*restriction "</LEGAL>"
restriction =  "<RESTRICTION
                 issuer =" quotedstring "
                 law =" quotedstring "
                 for=" quotedstring ">"
                 [viceversa]
                 while
                 dont
               "</RESTRICTION>"
viceversa =    "<VICEVERSA/>"
while =        "<WHILE>" data-group "</WHILE>"
dont =         "<DONT>" data-group "</DONT>"
```

At http://preibusch.de/namespaces/SIMT/inferences and http://preibusch.de/namespaces/SIMT/legal, the reader will find XML schema definitions and non-normative definitions based on a Document Type Definition (DTD). The latter is for notice only— P3P's language definition no longer relies on DTD, and the DTD definitions have been removed from the specification.

Appendix 2. Data-Protection Principles and Their Legal Bases

EU Directive 2002/58/EC is based on Directive 95/48/EC on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, which describes the EU's general data-protection principles. They rely on much the same principles as the OECD's Guidelines on the Protection of Privacy and Transborder Flows of Personal Data [41]. Essentially, personal data should not be collected secretly and, if possible,

only with the informed consent of the individual concerned (OECD "collection limitation" principle). The purpose for which they are collected must be made known, and they may not be used later for other purposes ("purpose specification" principle / "finality" principle) or by third parties ("use limitation" principle). Data collectors must ensure that data are relevant to those purposes and are correct. Data collectors must also take precautions to prevent data misuse and unauthorized access that might compromise the data's integrity ("security safeguards" principle). There must be ways for individuals to learn about the data being kept about them and if necessary to demand correction ("openness" and "individual participation" principles). Finally, a data controller should be accountable for complying with measures that give effect to these principles ("accountability" principle). Similar ideas are expressed in the Fair Information Practices (FIP), set down as a recommendation in the United States and updated by the Federal Trade Commission [22, 23]: notice, choice, access, and security. Both P3P and the extension proposed in this paper aim at capturing these principles, but (like any other technology) they cannot prevent the necessity of judgmental decisions and therefore raise legal concerns [13].

These principles have been transformed into legal requirements in a number of countries, in particular in the EU countries. For example, in Germany the use and processing of personal data ("person-related data"; see §3 I BDSG) is governed by the Federal Data Protection Act BDSG (*Bundesdatenschutzgesetz)*, the privacy laws of the states, the Teleservices Data Protection Act TDDSG (*Teledienstedatenschutzgesetz*), and the Media Services State Contract MDStV (*Mediendienstestaatsvertrag*). An important example of purpose limitation is the differentiation between "stock" data that are necessary for the reasons, contextual form, and changes of a contractual relationship and "usage" data that are required for the usage of services (§5,6 TDDSG). Data collected for billing purposes in electronic retailing are person-related and cannot be used for purposes other than transaction fulfillment. However, legislation explicitly allows the creation of pseudonymous user profiles for statistical analysis purposes (§6 III TDDSG; Art. 4 I e, Art. 20 II Regulation (EC) No 45/2001).

BETTINA BERENDT (bettina.berendt@cs.kuleuven.be) is an associate professor in the department of computer science at Katholieke Universiteit Leuven, Belgium. She obtained her habilitation in information systems from Humboldt University Berlin, Germany, and her Ph.D. in computer science/cognitive science from the University of Hamburg. Her research interests include Web and Social-Web mining, digital libraries, personalization and privacy, and information visualization. Her work has been published in the *VLDB Journal*, *Communications of the ACM*, *INFORMS Journal on Computing*, and *Data Mining and Knowledge Discovery*.

SÖREN PREIBUSCH (spreibusch@diw.de) works at the German Institute for Economic Research in Berlin, and completed his studies in industrial engineering at the Berlin Institute of Technology in 2008. His research interests include information privacy with emphasis on individually negotiated privacy policies in electronic transactions. He also investigates the applicability of formal methods and verification techniques in the industrial software production process, especially for safety-critical large scale installations. Preibusch has been working as an assistant reviewer for e-learning materials and has participated in consulting projects for mobile services in public transport and the financial sector. He is co-founder of the Internet portal IT-Productivity and

a member of the Deutsche Physikalische Gesellschaft, and has been a scholar of the German National Academic Foundation since 2003.

MAXIMILIAN TELTZROW (teltzrow@wiwi.hu-berlin.de) is a consultant with Arthur D. Little in Munich, Germany. He studied management and industrial engineering at the Berlin Institute of Technology, where he graduated in 2000. He received his Ph.D. in information systems from Humboldt University Berlin and has worked as a scholar of the German Research Foundation and the German Academic Exchange Service at the University of California, Berkeley, San Diego, and Irvine. His current research interests include privacy-preserving mining of user data, the development of e-metrics, and Web-based services. He has published in the *Journal of Electronic Commerce Research*, *Lecture Notes in Computer Science*, and the *IEEE Conference on Electronic Commerce*.