PSDM 2002

Privacy Conflicts in CRM Services for Online Shops – A Case Study

Claus Boyens, Oliver Günther, Maximilian Teltzrow

Humboldt Universität, Berlin
Institute of Information Systems
Spandauer Str. 1
D-10178 Berlin

AGENDA

1 Privacy Conflicts

>> Consumer Privacy Concerns

- => Conflict between companies rightful interest to mine data and consumer privacy protection, (most companies want to handle private data securely due to loosing reputation, problem: legislation, consumer concerns; external service providers)

- → For 93% of consumers data privacy is very important in E-Commerce (Princeton Survey Research Associates 2002, N=1,500)
- → 68% of consumers do not shop online because of fears that their personal details will be misused (EU survey 2002, N= 9,156)
- → 37% of online consumers said they would buy more online if they were not worried about privacy issues (Forrester Research 2001)
- → Only 3% feel comfortable with giving away credit card number (Ackerman 2001)

1 Privacy Conflicts

>> Legislation

- US and EU have different approaches to data privacy protection, Safe Harbor Agreement (companies agree to adhere to specific privacy measures, e.g. customers may opt-out of data collection), Japan: (Personal Data Protection Bill 1999)
- The basic situation, governed by the 1997 EU Directive on privacy in telecommunications (which binds legislation in EU countries):

  1. In general, personally identifiable information must not be given to third parties.

  2. Data can only be retained for billing purposes and must then be erased.

  => In the European Union (EU), critical profile aggregation in shops starts when data about individual customers are used for other purposes than transaction fulfilment or if the customer has not explicitly agreed to the usage of her data for other purposes
- Currently, there exists a proposal at the EU level to make data retention compulsory for 12-24 months.
- Comprehensive Sources: www.epic.org, www.privacy.org, www.privacyexchange.org, www.privacyinternational.org (400 pages report of privacy status quo in different countries, September 2002)

1 Privacy Conflicts

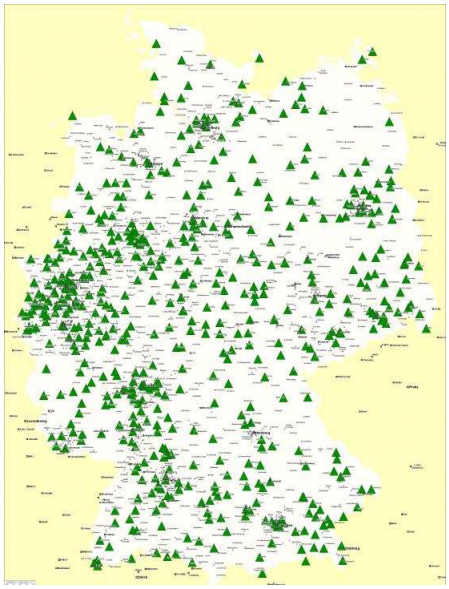
>> Available Data Basis

European e-Shop with affiliated branch network:

- **Log Data** (Browsing Behavior)
- **Purchase Data** (*name*, *billing address*, *shipping address*, *payment and delivery preferences*)
- **External data** (matches 7000 Zip Codes with longitude/latitude coordinates)
- 400 Shops with zip codes
- 14.000 customers with zip codes

Delivery	Payment	Return
Direct Delivery	Cash (offline at store)	Through Local Shop
Delivery to Store	Online Payment	Direct Delivery
	Payment on Delivery	

Shop distribution of cooperation partner



1 Privacy Conflicts

>> The Company‘s Interest

- Company wants to…

  … analyze consumer behavior for Marketing purposes
  … evaluate key metrics of online performance
  … measure the impact of its recently launched e-Shop on offline branch network

- → asked Institute of Information Systems at HU Berlin to help with data mining tasks, problems: comply with EU regulations and with reluctance of shop (company not allowed to forward private data for Marketing purposes)

- Goal of this talk: Reveal potential privacy threats in mining at an e-shop’s customer data and match privacy-protection techniques

1 Privacy Conflicts

>> Levels of Criticality

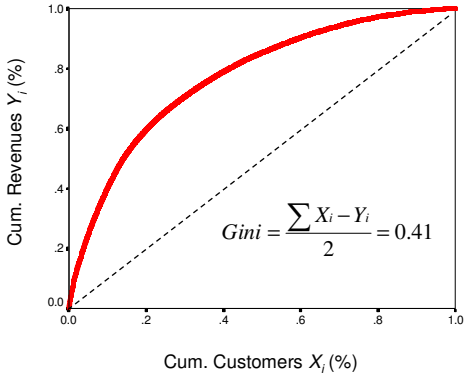
Privacy Criticality	Metrics Used	Tables Used	Attributes Used	Data Type	Privacy Method
1	Revenue Concentration	customer, order, position	customer_ID, order_id, revenue, order_date	Identity	Access limitation
2	Revenue partition according to customer attributes	customer, order, position	customer_ID, order_ID, order_Date, revenue, date_of_birth, gender, delivery_type	Identity	Access limitation, Aggregation
3a	Customers geographical distance from local shop	Customer, external data	Customer_ID, Zip_Code, Zip_Code_Shop, Latitude/ Longitude coordinates	Identity, Profile	Access limitation, Aggregation, Distortion, Randomizing, Swapping
3b	Identification of street-related product preferences	Customer, Order, Order_Position, External data	Customer ID, Street Name, Zip Code, Transaction ID, Product Name	Identity, Profile	Access limitation, Aggregation, Distortion, Randomizing, Swapping

AGENDA

1.1 Little Privacy Impact

>> Example for Little Privacy Impact

- *Lorentz Curve and Gini Coefficient*



Required Data:

`customer` (*`customer_id`*)

`order` (`order_id`, *`customer_id`*)

`order_position` (`position_id`, `order_id`, `revenue`)

1.1 Little Privacy Impact

>> Threats and Solutions

Threats:

- Criticality of the Lorenz curve depended on the other queries employed. Protected revenue figures could be exposed with combination of queries.

Solutions:

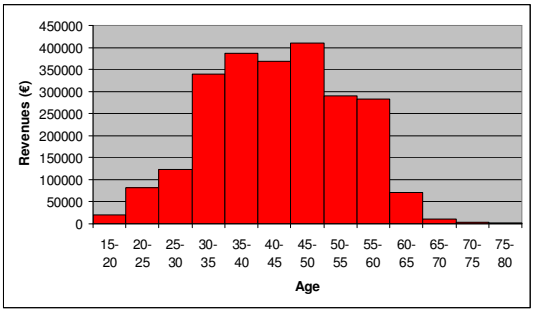
- Limited the access to all other critical customer attributes such as `name`, `gender`, `date_of_birth`, `address`, `products`, `credit_rating` should be a solid way to prevent privacy violations
- Assigned unique identifiers for `customer_id's` randomly
- Statistical Privacy Preservation Techniques: Query Restriction: Fellegi 1972, Denning and Schwartz 1979, control overlap amongst successive queries: Dobkin, Jones and Lipton 1979, clustering entities into mutually exclusive atomic populations: Yu and Chin 1979; Perturbation family (described later)

AGENDA

1.2 Significant Privacy Impact

>> Example for Significant Privacy Impact (A)

- Revenues/Age, Revenues/Gender
- Further relevant metrics: Credit_Rating, Revenues/Product, Payment Losses/Age, Revenues/Title



Required Data:

customer (*customer_id, date_of_birth*)

order (order_id, *customer_id*)

order_position (position_id, order_id, revenue)

1.2 Significant Privacy Impact

>> Threats and Solutions

Threats:

- Sweeney (2001) was able to identify 29% of all persons by combining `gender` and `date_of_birth`

Solutions:

- Limited the access to attributes such as `name, address, products, credit_rating`
- Employed the method of aggregation: For the chart Revenues/Age, it is sufficient to know the age of a customer, but the entire `date_of_birth` is not necessary. Either the age or the year of birth fully meet the requirements for this metric
- Altered the denomination of the attribute `gender` to premium_customer and regular customer, in order to prevent misuse.

1.2 Significant Privacy Impact

>> Example for Significant Privacy Impact (B)

- *Web Log Mining, Browsing Behavior of Customer Groups*

Concepts / SessionIDs	Home	Cate-gory	Product	Order Success
725738:100011	1	3	12	0
842121:332894	2	14	1	1
993242:345321	1	9	0	0
...	0	4	7	0

Table of cleaned, sessionized log data stored for analysis with OLAP techniques (Zaiane 1998)

Required Data:

`Log data` `(session_id, customer_id, pages_visited, proxy_id, geography, browser type, access_time, status_code, user_input)`

1.2 Significant Privacy Impact

>> Threats and Solutions

Threats:

- `Pages_visited`, `access_time` and institution's `proxy_id` could indicate specific individuals in a company
- `Session_ID's` could be matched with `customer_id's`

Solutions:

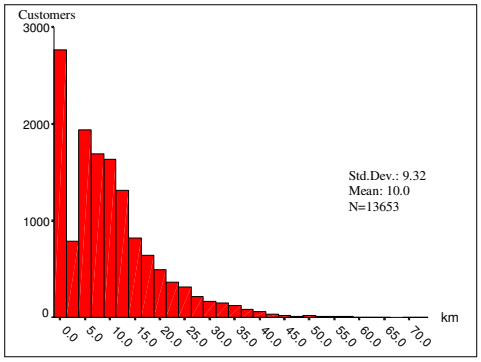
- Limited the access to browsing attributes such as `proxy_id`
- Alternated `Session_ID's`
- Aggregated browsing data to monthly reports

AGENDA

1.3 Major Privacy Impact

>> Example for Major Privacy Impact

- Geographical Analysis for marketing, customer profiling, and optimal shop placements



Required Data:

address (*zip_code*)

external data
(*longitude_customer*,
latitude_customer,
longitude_shop,
latitude_shop)

1.3 Major Privacy Impact

>> Results

→ A significant, negative Correlation exists between the number of online customers per zip code area and the distance to the next shop (Corr=-0.30). The closer an online customer lives to a physical shop, the higher is the purchase probability.

In comparison, no correlation between the customer density per zip code area and the distance to the next shop has been identified.

(Premise: Purchase Behavior is equally distributed among the population.)

→ Model can also be used to determine optimal logistical shop placements. (e.g. Study on the Spatial Distribution of Convenience Stores in the Tokyo Metropolitan Area, Compstat 2002, includes also demographic factors)

1.3 Major Privacy Impact

>> Threats

- **Trade-off** between the preciseness of results and the potential privacy violation of users: Aggregation error of ε_{av}= 7.4 km because we allocated customers to the center of a zip code area. Larger zip code areas of 200 km^2 could already induce inaccuracy of ε_{max}=16 km.
- => geographical values as granular as street coordinates are desireable to achieve better results for shop placements and customer segmentation
- **Inference Problem**: For 69% of all data records, given `date of birth` and `zip_code` are unique identifiers pointing to a specific person (Sweeney 2002)
- `Products` and `zip_code` could point to customer's preferences and residence. E.g. a researcher who orders specialized books in his field of interest is likely to be identified by the zip code that indicates the location of his university or research institution.
- **Small Data Cells**: Especially for sparsely populated zip code areas - the smallest data cell included 12 residents - the data miner could possibly find out who the customers are.

1.3 Major Privacy Impact

>> Solutions

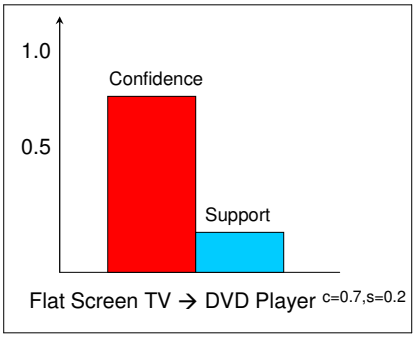
- Aggregation not the best solution (lack of precision)
- **Perturbation methods** such as *discretization*, *value distortion* and *value dissociation* are a solution for coding on zip code level. Apply randomization function based on Gaussian or Uniform perturbation. (Agrawal, Srikant 2000).
- **Suppression of data cells** of small size (Cox 1980). E.g. exclude zip code areas with small sizes below a threshold of 100 residents
- **Geographical mining remains a problem** with geo-codes as granular as street and household levels. Simple perturbation of longitude/latitude values could easily reveal original customer. More sophisticated algorithms needed.
- Practical Solution: **Partition customers into micro-cells** containing threshold number of households (e.g. 80.000 micro cells with each 50 households). Cells are characterized by attributes such as geographical coordinates, income or purchase preferences. Privacy Threat is reduced moderately.

  => Solutions for Geo-Coding remain an interesting research problem

1.3 Major Privacy Impact

>> Example of Major Privacy Impact (B)

- Street-Marketing with association rules (Product offerings for direct marketing in specific zip code areas)



Required Data:

customer customer_id

address address_id, customer_id, country_code, *street_name*, *zip_code*, town)

order (order_id, customer_id)

position (position_id, order_id, *product_name*, quantity, price)

1.3 Major Privacy Impact

>> Threats and Solutions

Threats:

- `Street_name` and `product_name` may again uniquely identify individual

Solutions:

- **Probabilistic distortion** (Rizvi, Haritsa 2002) for association rule mining providing both a high degree of privacy to the user as well as a high level of accuracy in the mining results.
- The database model in our case would consist of columns containing the online shop's products. Each row would contain a sequence of boolean operators representing a customer who purchased the product or not.

AGENDA

2 Related Work

>> Sample Publications at HU Berlin

- Spiekermann, S., Großklags, J., Berendt, B., E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior. In Proceedings of the 3rd ACM conference on Electronic Commerce , Tampa, Florida, 2001
- Boyens, C., Günther, O., Trust is not enough: Privacy and Security in ASP and Web Service Environments. In Proceedings of Advances in Database and Information Systems (ADBIS 2002), Bratislava, 2002
- Spiliopoulou, M., Web usage mining for site evaluation: Making a site better fit its users. Special Section of the Communications of ACM on ``Personalization Technologies with Data Mining'', 43(8):127-134, Aug 2000.
- Berendt, B., Mobasher, B., Nakagawa, M., & Spiliopoulou, M. (2002). The impact of site structure and user environment on session reconstruction in Web usage analysis. In B. Masand, M. Spiliopoulou, J. Srivastava, & O. Zaiane (Eds.), Working Notes of the Fourth WebKDD Web Mining for Usage Patterns & User Profiles Workshop at KDD 2002 (pp. 115-129). July 23rd, 2002, Edmonton, Alberta, CA. (PS)