ESPOO 2007

VTT PUBLICATIONS 647

Teemu Mutanen

Consumer Data and Privacy in Ubiquitous Computing

VTT PUBLICATIONS 647

Consumer Data and Privacy in Ubiquitous Computing

Teemu Mutanen

ISBN 978-951-38-7031-7 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)

JULKAISIJA – UTGIVARE – PUBLISHER

VTT, Vuorimiehentie 3, PL 1000, 02044 VTT
puh. vaihde 020 722 111, faksi 020 722 4374

VTT, Bergsmansvägen 3, PB 1000, 02044 VTT
tel. växel 020 722 111, fax 020 722 4374

VTT Technical Research Centre of Finland, Vuorimiehentie 3, P.O. Box 1000, FI-02044 VTT, Finland
phone internat. +358 20 722 111, fax + 358 20 722 4374

VTT, Vuorimiehentie 3, PL 1000, 02044 VTT
puh. vaihde 020 722 111, faksi 020 722 6027

VTT, Bergsmansvägen 3, PB 1000, 02044 VTT
tel. växel 020 722 111, fax 020 722 6027

VTT Technical Research Centre of Finland, Vuorimiehentie 3, P.O. Box 1000, FI-02044 VTT, Finland
phone internat. +358 20 722 111, fax +358 20 722 6027

Technical editing Maini Manninen

Mutanen, Teemu. Consumer Data and Privacy in Ubiquitous Computing [Asiakastieto ja yksityisyys jokapaikan tietotekniikassa]. Espoo 2007. VTT Publications 647. 82 p. + app. 3 p.

Keywords consumer data, privacy-preserving data mining, ubiquitous computing

Abstract

The emergence of ubiquitous computing means new devices, sensors, and protocols throughout society, and thus new sources of consumer data. The new data sources, along with new means of individual identification, constitute a personal privacy concern: what should and should not be done with personal data. The personal-privacy issue is accompanied by corporate privacy when data mining tasks are applied to consumer databases. The ubiquitous-computing environment will provide various data sources, and these databases will be distributed among various agents.

The privacy-preserving perspective on data mining is a relatively young area. The research in this area is mainly theoretical; to the best of our understanding, no real-world applications exist. In this work, we have tried to fill this gap. The current trend in the growing amount of personalization in online services has also created applications for personalized marketing. Personalized marketing services use detailed information about the context and personal history of a customer. This needs sophisticated individual identification methods, which themselves raise privacy concerns. The novelty in privacy-preserving methods is that sensitive and distributed data could be used for data mining tasks while the privacy of individuals is still preserved.

This thesis has two objectives: the first is to use consumer data from distributed sources and study how customer segmentation is possible while preserving privacy. The idea is to conduct the customer segmentation in a way that the data need not leave the agent holding the data. The other objective is the value of the knowledge acquired from collectively conducted segmentation. We believe that collectively conducted segmentation produces knowledge that cannot be acquired otherwise. The results of this work show that privacy-preserving customer segmentation is possible and that collectively conducted segmentation produces new knowledge.

Mutanen, Teemu. Consumer Data and Privacy in Ubiquitous Computing [Asiakastieto ja yksityisyys jokapaikan tietotekniikassa]. Espoo 2007. VTT Publications 647. 82 s. + liitt. 3 s.

Avainsanat consumer data, privacy-preserving data mining, ubiquitous computing

Tiivistelmä

Jokapaikan tietotekniikan ilmaantuminen tarkoittaa uusia päätelaitteita, ilmaisimia ja yhteyksiä, siten myös uusia asiakastietolähteitä. Uudet tietolähteet ja tavat yksilön tunnistamiseen nostavat esille huolen yksityisyydestä: mitä voidaan ja mitä ei pidä tehdä yksilön tiedoilla. Yksilön yksityisyyden lisäksi yritystietojen yksityisyys nousee esiin, kun tiedon louhinnan menetelmiä sovelletaan asiakastietoihin. Jokapaikan tietotekniikka tuo mukanaan lukuisia tietolähteitä, jotka sijaitsevat hajautetusti.

Yksityisyyden säilyttävä tiedon louhinta on suhteellisen nuori ala. Alalla tehty tutkimus on pääosin teoreettista, käsittääkseni yhtään tosielämän sovellusta ei ole olemassa. Tämä vajetta on yritetty paikata. Vallitseva suuntaus yksilöllistämisen kasvavasta määrästä verkkopalveluissa on luonut myös markkinoita yksilölliselle markkinoinnille. Yksilölliset markkinointipalvelut hyödyntävät yksityiskohtaisia tietoja asiayhteydestä ja yksilön käytöshistoriasta. Tämä vaatii kehittyneitä menetelmiä yksilön tunnistamiseen, ja menetelmät nostavat huolen yksityisyydestä. Yksityisyyden säilyttävien menetelmien uutuusarvo on mahdollisuus käyttää arkaluontoista ja hajautettua tietoa tiedon louhintaan, kuitenkin säilyttäen yksityisyys.

Tässä työssä on kaksi tavoitetta: ensiksi käytetään asiakastietoa hajautetuista lähteistä ja tarkastellaan yksityisyyden säilyttävän asiakassegmentoinnin mahdollisuutta. Ideana on segmentoida asiakaskanta siten, ettei arka tieto missään vaiheessa lähde sitä hallitsevalta pelurilta. Toinen tavoite liittyy tietämyksen arvoon, joka saadaan yhteisesti suoritetusta segmentoinnista. Uskomme, että yhteisesti suoritettu segmentointi tuottaa tietoa, jota ei muuten pystytä tuottamaan. Työn tulokset osoittavat yksityisyyden säilyttävän asiakaskannan segmentoinnin olevan mahdollista ja yhteisesti suoritetun segmentoinnin tuottavan uutta tietoa.

Contents

1. Introduction

The growing concern over *big brother* watching after us is a national obsession. Every day of each month, every one of us leaves some individually identifiable information behind. All the information we receive via mobile phones or Web browsers are traceable, credit or customer loyalty card use leaves marks about what we consume, and when we travel by bus or train the fare is paid by smart card. Two trends neatly synthesize this. The first trend is the growing amount of consumer data being gathered. The widespread use of computers, networks, and tags make it possible to record consumer behavior in more detail. The second trend is the use of sophisticated analysis techniques and the adoption of these into sectors with sensitive data. Together, these two raise acute privacy concerns. Who has access to data, who can use it, and for what purpose? This is what this work is all about.

In this work, we study consumer data and the possibilities of its use in the ubiquitous-computing environment. We study the benefits of how the knowledge from consumer data can be used for marketing and how new means can be applied. The idea of the use of analytical tools as the basis of marketing decisions is not entirely new. Virtually every major company uses some form of statistical or mathematical analytics as the basis of their decision making. Without taking any credit away from the actual mathematical analysis, the most important factor in being prepared for sophisticated analytics is the availability of sufficient volumes of high-quality data. One of the questions we study in relation to consumer data concerns the amount of data that can be collected by one agent. Is this enough to gain a competitive edge?

Despite the importance of the advanced use of consumer analysis, the privacy of an individual remains a central issue. The emergence of ubiquitous computing means new ways of gathering, applying, integrating, and displaying information received from various sources. Some of these sources might include sensitive information about individuals, while others might provide information that can be linked to other data sources and, in doing so, might constitute a privacy intrusion. One example of this is the location: one location might not give rise to much concern, but location trajectories can be combined from a dynamic source. The services of ubiquitous computing take advantage of a variety of sources of information and thus privacy issues are very relevant in context where computing is ubiquitous.

Although privacy is a centrally important value for people in today’s society, we believe that privacy in today’s society depends on our cultural beliefs, how people think and feel about privacy and how they value it. Privacy no longer means preventing organizations and other people from knowing about us. Privacy refers to a concern as to how personal data is used and by whom. In this work we address privacy issues by evaluating the question of sharing personal data, what can or cannot be done with it. We also study the computational methods in a privacy-preserving sense. Is it possible to use personal data for customer segmentation and still preserve the privacy of individuals?

Our goal is to use consumer data and apportion the customer base into various segments. We study the difference between regular customer segmentation and privacy-preserving segmentation. Our hypothesis is that the segmentation results are not any different; the privacy-preserving clustering protocol should produce the same clusters as those resulting from the regular clustering protocol. The novelty in the privacy-preserving protocol is that the data will never have to leave the agent holding the data, and still the customer segmentation remains possible.

Another aspect we study with consumer data is the difference the sharing of data produces. How does the customer segmentation differ from the data set that is used? We analyze the customer segments as agents conduct segmentation alone. And we compare the results from individual segmentation to the segments produced by regular clustering protocol when the agents share their data. Although there are reasons for agents not to share consumer data with other agents, we believe that each company holding consumer data can add a perspective. Our hypothesis is that the collectively conducted segmentation produces knowledge that cannot be acquired otherwise.

Following the introduction, this work has five sections. The second chapter provides all the background information for our work. First, we present a few future scenarios of ubiquitous computing to show how the emergence of this might change today’s society. We present some enabling technologies and discuss a few interesting services that provide insights into the ubiquitous computing. We also present a state-of-the-art view on targeted and personalized marketing services. These services will form one of the application areas that take full advantage of the privacy-preserving analysis of consumer data.

Chapter 2 includes also the background of a privacy-preserving data mining sector. Data mining methods have been around for two decades now. Today, these methods are being applied in various business sectors. The privacy aspect has been addressed by the data mining researchers over the past few years. Although privacy-preserving methods are not very mature, we present the taxonomy of these methods. Privacy issues are discussed from a computational perspective and a few examples of how a small number of pilot projects have failed in the task of preserving privacy are taken from the ubiquitous-computing environment.

In Chapter 3 we present the data we use for the case study. The description of the data is presented along with some preprocessing actions that are carried out. In the case study we use customer transactional data with some demographic variables. Chapter 4 then presents the methods we use for the customer segmentation. We review basic *k*-means clustering algorithm and describe our privacy-preserving clustering protocol in detail.

In Chapter 5 the results of the segmentation are presented. The analysis of the computation complexity is also given in this chapter. The last chapter includes the findings and results of this work: some discussion of future research is also provided.

2. Background

2.1 Ubiquity

The current trend toward a universal presence of mobile computing, computer networks, and wireless communications in everyday life is increasing the importance of the debate on ubiquitous computing. “Ubiquitous computing” is a term describing a seamless interaction between various devices, systems, and people without restrictions on location or time. It encompasses many different technologies; the integration of technologies is the enabler for many applications in e-commerce and connected homes. For the technological aspects, this means for example, the emergence of new short-range connection protocols, embedded software, and various intelligent networks.

This type of future visions has been referred to in the literature by a variety of names, such as calm technology, pervasive computing, ambient intelligence, “everyware” and tangible media. These terms all are used interchangeably, although they are conceptually different. For example ubiquitous computing uses the advances of mobile technologies, while pervasive computing describes the global computing environment. These terms all describe the one emerging phenomenon first described by Mark Weiser in 1991 [63]. Mark Weiser, from Xerox Palo Alto Research Center defined ubiquitous computing in 1991 as follows. The goal is to create a computing environment in which each person is continually interacting with hundreds of nearby wirelessly interconnected computers [63]. As Mark Weiser originally described the term ubiquitous not only meant “in every place” but also “in every thing”. Although he pointed out the possibility of “in every thing” computing, his main point was the idea that the most profound technologies are those that disappear.

The European Commission’s Information Society Technologies Advisory Group (ISTAG) defines the emerging nature of computing as ambient intelligence. According to this vision, people will be surrounded by intelligent and intuitive interfaces embedded in everyday objects around us, such as furniture, clothes, vehicles, roads and smart materials, even particles of decorative substances like paint [29]. It also states that Ambient Intelligence implies a seamless environment of computing, networks and specific interfaces. Ambient Intelligence stems from the convergence of three key technologies: ubiquitous computing, ubiquitous communication and intelligent user-friendly interfaces [29].

Overall, the ubiquitous-computing paradigm shift means that technology becomes virtually invisible in normal life. Instead of having a normal laptop or a desktop computer, the computing devices will be embedded into the environment. While all the visions of ubiquitous computing seem to be very far off, the limitations of present computing can already be seen. A study shows that existing technology – that is, the PC – has significant limitations because of its inability to take advantage of the importance of context and location [17]. While the study showed the limitations of present technology in relation to the ubiquitous aspect of computing, home information was found to provide an opportunity for truly ubiquitous computing.

Ubiquitous computing at the present stage lacks of sufficient standards of ubiquity. The various technologies currently thought to be part of ubiquity include: GPS, WLAN, WI-FI, RFID, UWB, etc. But the current development of mobile applications face challenges such as high diversity of devices, limited memory capabilities, small screens, etc. When ubiquitous computing in the future begins to include also other devices than PCs and mobile phones, the established standards for ubiquity will help to guarantee more advanced functionality. Table 1 shows some examples pilot projects of ubiquitous computing already established today. It can be seen clearly that there is a gap between visions and reality.

Table 1. Examples of ubiquitous computing today.

Theme	**Description**
Ubiquitous-computing city	CITY OF SONGDO – a man-made island being privately developed from the ground up, completion of all the five phases is scheduled to be finished in 2014. The plan includes, for example electronic money (“u-wallets”), RFID tags everywhere (ID cards, subway tickets, debit cards), sensors in streetlights to help steer traffic flows. The developers envision it as a “ubiquitous-computing city”.
Targeted audio	HOLOSONIC Research Lab has created a system for delivering highly directional audible sound beams to chosen targets. Another similar system on the market is HYPERSONIC SOUND. An example of possible use is to deliver personalized audio information to single persons or groups in a

	lecture or meeting room without disturbing others, or to each seat in a car. Sonic City (a project in Sweden) – the soundscape and background music will change according to the user's behavior and the urban atmosphere.
RFID chips around the society	City of Kobe has a research project where RFID tags were put into the sidewalks, lamp posts, street signs and bus stops. People would be carrying handheld devices and receiving signals "you are right now in place Y.y. If you turn right from next corner there will be store ZZ". In Britain the government is testing a new license plate with a RFID chip. An RFID chip embedded in a license plate will likely hold information about license plate number, type of vehicle, etc. Mitsukoshi has been testing an item-level RFID at its flagship store in Nihonbashi, Tokyo. E.g. Five thousand pairs of jeans, for example, will be RFID-tagged for the purpose of inventory management and improvement of services and operations. PRADA epicenter use RFID tags on clothes and Wi-Fi-based tablet PCs to scan them. Also a clothes retailer GAP has been conducting field tests with RFID tags on inventory purposes.
Wireless networking and position system	Carpetlan – indoor wireless-like networking and positioning system, every object in a room should be connected to the network. Sensacell – These non-contact sensors can detect people and objects within 6″ through materials such as glass, plastics, wood, tile, etc. Human Locator – a service aimed at retailers to control frames of animation to grab people's attention. Their software tracks multiple targets, direction, speed, size of person etc. by using movement blobs. Trackstick – records its own location, time, date, speed, heading and altitude at preset intervals. Gaudi System – the purpose of this is to assist visitors in navigating their way around a campus. The system consists of set of autonomous wireless displays and a navigation server.
Traction elevators	The idea behind the Miconic 10 elevator system is that of eliminating push buttons inside the elevator cab. Users register the destination on a keypad before entering the elevator.
Narrative spaces	The Sensing Places company produces people-driven narrative spaces for clients such as museums, theatres, shops, hotels, restaurants. The idea in short is that people's presence and movement drives the presentation of digital media. My Washington Square is an experience-mapping project exploring contemporary and historical stories submitted by online visitors; this project explores the ways in which physical spaces contain many layers of memory and experience

An example of one emergent service, Fluid Time, gives a hint of how ubiquitous computing will change our behavior. The project investigates the ability for users to flexibly arrange appointments by coordinating their own schedules with the availability of the services. These arrangements are made by using wireless networks and are thus made available to remote locations in real-time [21]. The service supports flexible time planning and delivers information about when and where a desired service might be available. This system relies on real-time data access, thus the system should be able to deliver accurate information about dynamic services.

Although the users may not want their devices to send their location, location-aware mobile devices may use the location information on the user's behalf. This is the case, for example, if one is visiting an unfamiliar area and has a device that helps to determine whether a certain type of store is near the route one will travel on, but, for privacy reasons, one does not want to reveal his position. The idea of sending location information to the user's device has been pilot tested in, for example, the city of Kobe in Japan [46].

The idea of ubiquitous computing excites many people, e.g. [23], [55], and [63]. However it is not clear how it will turn out and what solutions it will bring to ordinary life. To this date, the enabling technologies needed to fulfill many of the visions are still at the idea level or used mainly by tracking express air-mail packages. The research field seems to be divided around two central topics. The first focus is on the design of digital environments for 'digital' people to gather and participate in art, learning or entertainment. The other is on thinking of ways to augment everyday living, instead of just bringing virtual life next to it. As Mark Weiser foreshadowed in 1991, the most important benefit could be that the ubiquitous computer will help overcome the problem of information overload [63].

HP's Gene Becker describes the talk about ubiquitous computing as follows:

The potential uses and benefits of ubicomp often seem 'obvious'; most of us in the field have spun variations of the same futuristic scenarios, to the point where it seems like a familiar and tired genre of joke. 'You walk into the [conference room, living room, museum gallery, hospital ward], the contextual intention system recognizes you by your [beacon, tag, badge, face, gait], and the [lights, music, temperature, privacy settings, security permissions] adjust smoothly to *your preferences. Your new location is announced to the [room, building, global buddy list service, homeland security department], and your [videoconference, favorite TV show, appointment calendar, breakfast order] is automatically started.' And so on. Of course, what real people need/want in any given situation is far from obvious. [7]*

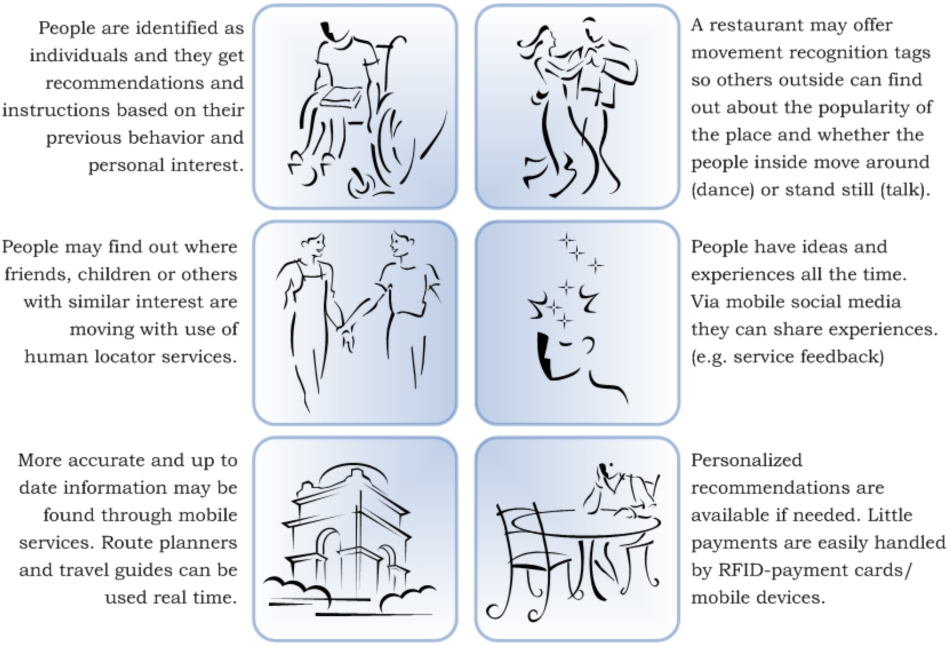
This opinion underlines the point that in the visions of ubiquitous computing the whole concept tries to understand people, context, and the world at the same time, while trying to get computers to handle everything in everyday situations.

Part of this ubiquitous-computing problem for the creation of real-world applications is that the means are not yet ready. The appropriate design and documents and conventions do not yet exist [23]. For example mobile phone maker Nokia introduced a new technology for handheld devices to communicate with each other and with accessories [1]. This Wibree technology is a new radio frequency technology that can work along with or complement Bluetooth, but only one tenth of the power is needed during use. The first devices using the Wibree technology are expected to hit the market in 2008.

This Wibree technology is just one example that shows ubiquitous-computing solutions are not ready yet, but when those will be part of real life is just a question of how strictly the term is defined. Ubiquitous computing seems to be an immediate issue or a hundred-year problem, depending on the topic. As it was presented in the Table 1, if ubiquitous computing means just wireless positioning systems or intelligent tags inside products, these things already exist. On the other hand, it is also a long-term problem that includes technical, social, ethical and political challenges [29], [36]. One of these challenges that need to be addressed now relates to the consumer and individual privacy. This will be discussed in more detail in Chapter 2.4.

2.2 Services

In the previous chapter, the idea of ubiquitous computing was explained. The visions of ubiquitous computing may seem too far reaching; the implemented services will show how it is experienced by users. One of the key concepts in ubiquitous-customer services is efficient gathering of consumer-behavior information. Figure 1 shows an example of how ubiquitous-computing services might be used in city life.



People are identified as individuals and they get recommendations and instructions based on their previous behavior and personal interest.

A restaurant may offer movement recognition tags so others outside can find out about the popularity of the place and whether the people inside move around (dance) or stand still (talk).

People may find out where friends, children or others with similar interest are moving with use of human locator services.

People have ideas and experiences all the time. Via mobile social media they can share experiences. (e.g. service feedback)

More accurate and up to date information may be found through mobile services. Route planners and travel guides can be used real time.

Personalized recommendations are available if needed. Little payments are easily handled by RFID-payment cards/ mobile devices.

Figure 1. Example of how ubiquitous services are used in the city life.

The services can be categorized roughly into three classes: marketing services, information services, and entertainment services. The marketing services are offered mainly to companies: information and entertainment services are used by consumers. Some examples of e-commerce and mobile services already implemented today are presented in Table 2. While some of the ubiquitous services can be implemented without a large amount of detailed consumer information, the majority of these services need some type of consumer profile information and/or location and time-dependent information.

An example of how ubiquitous computing modifies existing services, and enables more and wider aspects on them, can be seen in social media services. Some examples of social media services are presented in the Table 2. Instead of just Internet-based services, in the experiments called wireless graffiti, augmented reality technologies connect locations, people, media and objects to a place [27], [55]. The idea behind wireless graffiti is the belief that ordinary people hold a huge amount of knowledge about places in their head, but this is not accessible knowledge. Furthermore, the service called SNiF, for example, is a social media service for dogs! A tag keeps track of your dog, her pals, and all of your dog's activities.

Table 2. Examples of electronic- and mobile-commerce services already implemented.

Theme	**Services**
RFID payment service	Hong Kong's Octopus card, Accelite, QUICPay, mobile phone integrated RFID (e.g. Nokia, NTT docomo, KDDI), Mobile Suica (Japanese train pass)
Product information	Prada shopper, Barcode reader in mobile phones / Unicode(VTT), Weatherman 3G
Mobile and online payments through e-wallet	Google Checkout, PayPal and PayPal mobile, Mobiiliraha (Finland), Digiraha (Finland), Banking Grade SMS
Entertainment and social networking	Dodgeball, RatesYourDate, Jabberwocky, Jigsaw Proxidating, Facebook, Bluejacking
Social media	Digg, MySpace, DigitalEarth, globalideasbank, Craft manifesto, Flickr, Plazes, Hiptop Nation
Mobile advertising	Add2Phone, mobliss, AvantGo, Third Screen Media, Mophap's MobiStitch, Enpocket, Driwe iJack, Velti
Person locating service	Kidspotter, SafeTzone, Child spotter, Phonesitter
Wearable computing	Nike+ipod, Adidas+Polar, Burton+Motorola
Targeted marketing	AdSense, AdWords, Amazon Omakase, AzoogleAds, ContextWeb, and Yahoo Publisher Network
Web Storage	Web-based storage for consumers e.g. Amazon S3, Xdrive, Files Anywhere, MyDocs

While the whole ubiquitous service sector is just starting to evolve, some of the truly ubiquitous services could emerge as an emergent phenomenon. One emergent service could for example be a future data space, presented in [23], where reality is enhanced with digital information gathered from ad hoc sensor networks, short-range wireless communication, and fine-grain location information.

Or maybe ubiquitous computing will emerge just to help people with their regular routines, so people will have more time for each other. An example of this is already seen in New York. In New York City's Republic Restaurant, the servers place the food and beverage orders by using handheld devices [39]. As a result, the average table stay of the restaurants customers has come down to about 35 minutes and the servers minimize the foot traffic between the dining hall and the kitchen, so spending more time with the customers.

2.2.1 Shopping Centre Example

The idea of ubiquitous computing, e-commerce services and advanced data management, and how these converge, can be understood from a shopping centre example. Various agents operate in a shopping centre. Each of these agents collect some data from the customers, if the whole customer base inside the centre area is C so each agent has as its customer base c_i a subgroup of customers from the whole base C. All the consumer data is thus distributed among various agents. In addition to this consumer data, the shopping centre might also include agents (e.g. media agents) who collect information about traffic or monitor the activity inside the shopping centre.

The idea of ubiquitous computing in the shopping centre context means new means of collecting consumer information. The knowledge extracted from this information can be used for, for example, targeted marketing. Targeted marketing is discussed in the next chapter, but the idea of context-aware knowledge is truly ubiquitous and cannot be acquired without collective data sharing. Context-awareness will provide new means for marketers, but it also means privacy concern over the identification protocol and analysis method. The example of information sharing between the agents in the shopping centre is discussed in more detail in the segmentation chapter, 5.3.

2.3 Targeted marketing

"Targeted marketing" is a phrase used for marketing that is delivered to the market with a specific segment in mind. In fact any marketing that is made with customer return contact in mind is defined as targeted marketing. This marketing has three different phases: market segmentation, target selection and product positioning. The customers can be segmented by using different variables such as:

> GEOGRAPHICAL segmentation based on variables such as region or country, neighbourhood, density or city size.
>
> DEMOGRAPHIC segmentation based on variables such as age, gender, education, family life cycle, occupation, and income.
>
> PSYCHOGRAPHIC segmentation based on variables such as lifestyle, social class, personality, and personal interests.
>
> BEHAVIORAL segmentation based on variables such as user rates, user status, attitudes toward the product and the company.

Figure 2 shows market segmentation using two variables. Usually customers are segmented by using numerous variables and the segments are, as a result, meant to include people who are as homogenous as possible in their needs and attitudes. These segments can then be analyzed from the company's perspective to ascertain which the strong areas are and which segments have potential growth. Some segments are highlighted in Figure 2 as an example to point out the important segments. Furthermore, depending on the product at hand or the marketing goal, some segments are selected as targets for the campaign. The customers, or a small subset of them, in a particular segment are then contacted with marketing material.

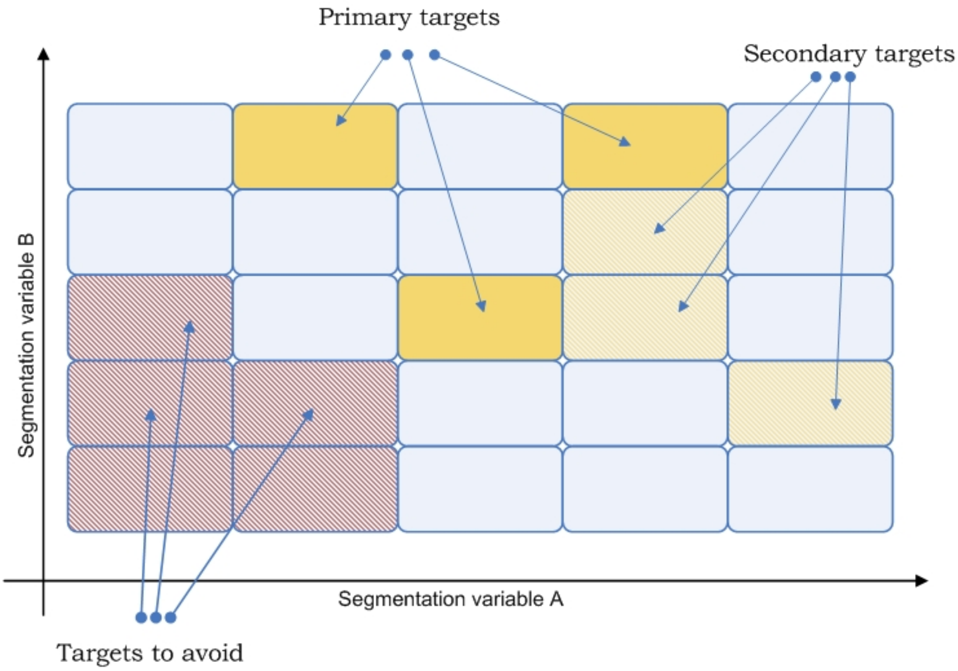


Figure 2. An example of the segmentation result using two variables.

The basic idea of targeted marketing campaigns is that the campaigns are conducted by maximizing the attention in specific customer segment. Once the customers are identified in more detail, the mass marketing approaches are not needed for customer contacts. The benefit of this kind of marketing is an increased response rate from the customers. There are also legislative restrictions for targeted marketing: for example, in Finland, health services are not allowed to be targeted at people under the age of 15. These legislative restrictions are not discussed in this work.

2.3.1 Personalized Marketing

The short introduction to targeted marketing above briefly described the idea of the marketing campaigns without the use of the online marketing methods. The phrase ‘online marketing methods’ is used here because, with the rise of the Internet and e-commerce, the targeted marketing has been revolutionized by some new services (e.g. Sponsored Listings, AdWords, Omakase, AzoogleAds, ContextWeb, and YPN). The online services have referred to their targeted marketing as “personalized marketing”. The main idea of these services is similar to the previously described targeted marketing service: a customer is contacted based on his interest and attitudes. The biggest difference, and also the competitive edge, is their sophisticated use of information covering the history of the user’s behavior. The consumer is identified based on some information (e.g. search query, previous behavior, location) which will then be used on the basis of marketing efforts. Simply, the trick is to get ads that target the relevant context and history well.

The key point is that before the advertisements could not be tied to the context as well as it can be by online services. For example, online store Amazon uses consumer data on cross-selling growth [38]: the information about buying patterns is transformed into recommendations. Chris Anderson argues [3] that this combination of good-quality recommendations with huge inventory of items is a real business advantage. The advantage is gained only if the customer can be targeted with relevant recommendations, the variety of items is not sufficient. A study [52] shows that while it is important to notice the customers and talk to them, it is also important to handle the message delivered. Consumers are highly irritated when the personalized marketing is irrelevant to their current situation in life.

In Chapters 2.1 and 2.2 some examples of ubiquitous and e-commerce services were presented. The number of such ubiquitous services that use the individual information is growing, and some of the technologies for the identification are still in the beta phase [23]. There is therefore a growing concern over privacy issues with the use of these services [18], [44], [61], [64]. Because of the individual identification the providers of these new marketing services do have to take privacy issues into account. The privacy issue and aspects are discussed in more detail in the next chapter, but before that, the applications of today’s targeted marketing services are presented.

2.3.2 Long Tail

"Long tail"[i] is a term used to describe certain business and economic models [3]. The companies taking advantage of this business model are, for example, Amazon (www.amazon.com) and Netflix (www.netflix.com). The former started in 1995 as an online bookstore, and has grown so that it sells today items in a variety of categories; the latter company is an online DVD rental service. Both of these companies operate only over the Internet and the items sold/rented are delivered via regular mail (or UPS, FedEx etc.). The basic idea of how the term "long tail" describes this business model is presented in the Figure 3.

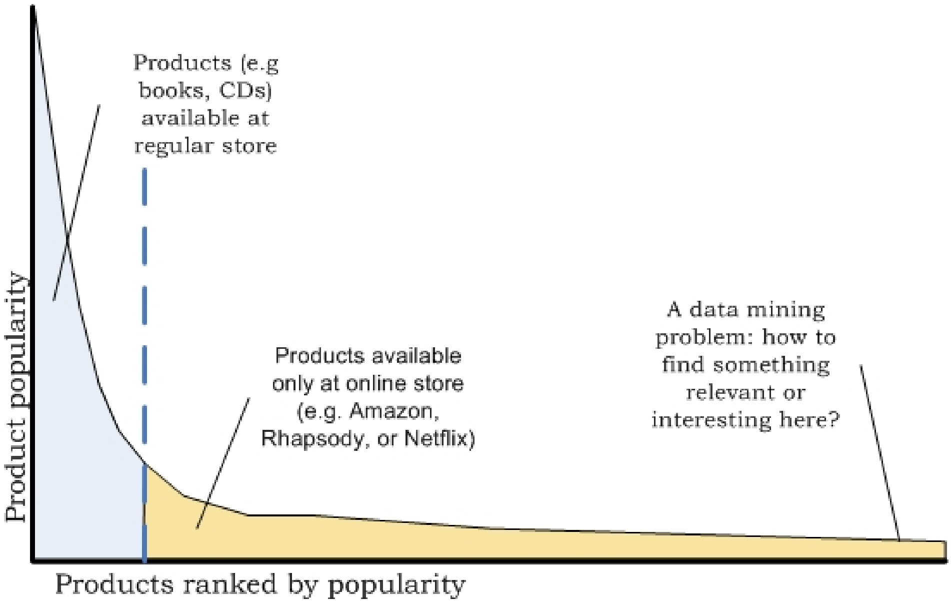


Figure 3. The basic idea of the long-tail business model. The inventory of obscure products (charted above in yellow) makes up the so-called "long tail".

The long-tail business model makes use of low demand or low sales volume products. Chris Anderson argues that these products can collectively make up a competitive market share in certain business areas [3]. In case of books/DVDs,

[i] The term "long tail" is also generally used in statistics, often applied in relation to wealth distributions or vocabulary use [1].

the relatively few bestsellers/blockbusters are sold in various places from supermarkets to small bookshops and clubs. For the long-tail business model to be competitive the store or distribution channel should be large enough. Both of the examples above have created a large inventory of items. Because of the mail delivery system the customer does not need to know the physical location from which the order has been dispatched. Therefore both of the companies can have various storage locations with various item inventories. The customer will only have online access to the combined inventory.

The combined inventory, in these cases with online service, is usually ten or more times larger than the physical store inventory. Amazon has right now an estimated inventory of about 20 million products, including nearly seven million books[ii]. This may be compared to regular Barnes & Noble bookstores in the US, which have an average of 130 000 titles on the shelves.[iii] A typical Blockbuster store has a selection of 3 000 movies. The Netflix online rental service has an inventory of 40 000 movies.

The statistical theory behind the long-tail model stems from experiments. The behavior of the long tail is formulated by Zipf’s law. Originally, this stated that, in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table [1]. Zipf's law may be stated mathematically as:

$$f(k;s,N)=\frac{1/k^s}{\sum_{n=1}^{N}1/n^s} \qquad (1)$$

where N is the number of elements, k is their rank, and s is the exponent characterizing the distribution. So, if a large collection is ordered by popularity, the second item is roughly half as popular as the first one, and third is one-third. The kth ranked element will be about 1/k of the first one in popularity. The

[ii] A search made in the Amazon US store (9.10.2006) with a keyword ‘AND’ and the category restricted only to ‘books’ produced a result of 6 939 034 titles. The keyword ‘AND’ was removed by the search algorithm, so the whole catalogue was crawled through and produced as a result.

[iii] The rest of the numbers are from C. Anderson’s book The Long Tail [3].

Zipf’s law is experimental, not a theoretical one. It is believed that, for example, the Amazon book catalog’s popularity will also follow Zipf’s law [9]. This line of thinking leads to the possibility of analyzing the value of, for example, Amazon’s sales, taking popularity as a rough measure of value.[iv]

A company operating with a large inventory is not the only advantage in the long-tail business model, although people will definitely come looking for obscure titles. The real advantage for these companies comes from use of a large product inventory combined with a powerful product recommendation system. A product recommendation engine and other application aspects are discussed in the following chapter.

2.3.3 Applications

Recommendation algorithms are best known for their use in online marketplaces, where the users input about a customer’s interest is used to generate product recommendations [47]. Usually the environment is challenging for various reasons [38]:

SCALE An online retailer can have a database with millions of customer accounts and millions of different product items.

SPEED The algorithm should produce high-quality recommendations almost in real time.

DIVERSITY New customers usually have very limited amount of information, while some existing customers might have a glut of information.

ALTERATION Each interaction provides valuable information and the algorithm should respond immediately to new information.

[iv] That is, if the catalogue has million items, then the most popular 100 will contribute a third of the total value, the next 10 000 another third, and the remaining 989 000 the final third [9].

A generic input-process-output model of a recommendation system is presented in Figure 4. Most of the recommendation-system algorithms start by finding a set of customers whose purchased products overlap with the products purchased by the user. The system then aggregates the products from similar customer, eliminates those already purchased by the user, and recommends the remaining to the user [38]. The type of algorithm used and the context defines the input variables and the goal the algorithm tries to achieve.

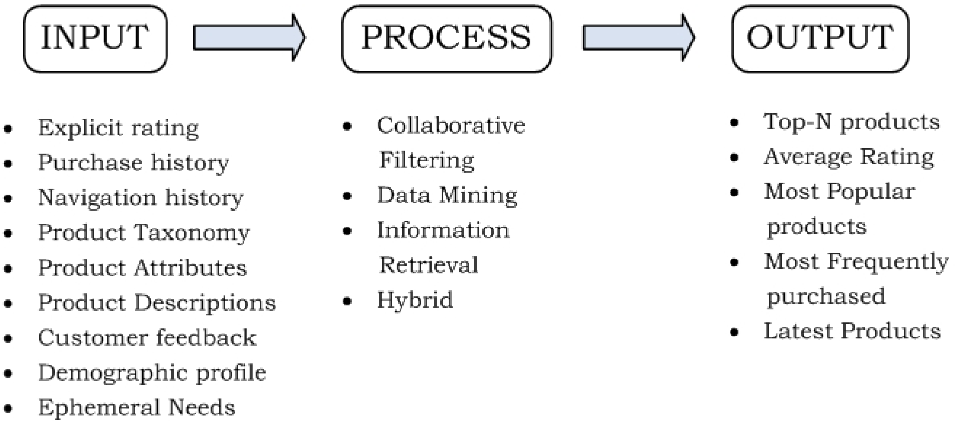


Figure 4. Generic input-process-output model of recommendation systems [51].

In the previous chapter, the long-tail business concept was presented. The mass marketing approach cannot succeed in the long-tail concept. Consider the marketing extremes: at the one end are those massive TV-advertisements for making the brand known (e.g. telecom operator ads during sport events), while at the other end are expensive high-tech products sold only to the top-100 companies by person-to-person marketing. The targeted marketing discussed here can be found in the middle of these two extremes, i.e. between products with a low margin and small target group. For example, a book publisher in Finland might have a title which will sell around 1 000 copies, a fairly successful title and also profitable. Finding those 1 000 customers could be very expensive by using a mass marketing approach.

The rise of the Internet has provided to targeted/personalized marketing with new dimensions. One of the benefits that the Internet has provided to the companies operating online (e.g. Amazon) is the lower customer cross-selling cost. “Cross-selling” is a term used for the practice of selling other products to the customers who already have purchased products from the seller [1]. This has an effect on the sales of the products found further along the long tail. So, for example, in the book publisher example in Finland, finding those 1 000 customers who are interested of buying the book is a lot cheaper online than it is otherwise. Also, when the product inventory is huge, cross-selling volumes are higher if the recommendation system is able to find and recommend obscure titles. As mentioned in Figure 3, the problem of finding something interesting from the huge inventory becomes a data mining problem, simply listing the relevant products is not sufficient.

As mentioned earlier, another type of service that revolutionized the targeted marketing sector are the text advertisements seen in the search engine result page or in other pages that provide marketing spaces for search engines. These text advertisements are unique because, while they usually target the context very well, the cost of them is based on actual clicks instead of impressions. At the beginning, these advertisements could only include text ads, but now the ads can also include sound and video content. The privacy issue relating to these services is controversial ([15]); the only official privacy statements are those the companies themselves provide ([1], [24]).

Other applications could be found from the ubiquitous-computing environment. A study shows that 74 percent of UK consumers were interested in receiving information and promotional location-based services on their mobile phones [11]. This shows that there is potential market for relevant real-time location-based marketing: people do not want more meaningless marketing or automatic help but they definitely welcome the advice of a free parking space when in need of one or a way to spend less time with change at the ticket machine. As was discussed in Chapter 2.1, the technologies to make these services work are still in beta phase. And as mentioned before, many of the truly ubiquitous services will be in the form of emergent phenomena when the ubiquitous infrastructure is ready.

All the applications discussed here produce some privacy risks, for example, what type of data is used as input to product recommendations or what elements or actions are being monitored? The potential privacy risk arises also when new types of source data are being collected; these are, for example, camera phones, high-quality audio/video feeds, movement censors, etc. Always when detailed information about consumers is collected the privacy issues should also be addressed. Whether the information collected is static (address, transactions) or dynamic (location, duration of service) in nature, it is not the information that is good or bad. It is the misuse that produces the privacy risk. The privacy issue overall, how the privacy is seen from a data mining perspective and how some methods are used to preserve privacy are presented in the next chapter.

2.4 Preserving Privacy

In the previous chapter, the targeted marketing sector and the advantages of the online business model was discussed. The targeted marketing and recommendation engines have been part of the growth of the Internet, which has triggered new opportunities for cooperative computation tasks. In cooperative computation, the answer depends on the inputs of separate entities. Some of these computation tasks can be carried out independently by a trusted entity who is allowed to access all of the inputs. However, some of these computations could occur between distrusted parties, or even between competitors, in which case the context disallows the existence of a trusted member. The privacy issue becomes relevant from the algorithm perspective when the computations are carried out.

This chapter explains the privacy issue from the privacy-preserving computation perspective, and some methods and techniques for preserving privacy. Some of the privacy issues are raised, even in public debate [18], when the question is about who has the access to which data. Although this has more to do with information security and protection than preserving privacy, this is one of the privacy aspects discussed in the next section. It is rarely the case that data themselves provide value; rather it is the knowledge that can be extracted from the data. This sensitive knowledge discovery is the main characteristic of the privacy-preserving data mining. The main objective in privacy-preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [1]. This approach requires definition what is *private*, and, because this has many different definitions (examples can be seen in Table 3), the approaches on privacy-preserving data mining are therefore also numerous.

As previously mentioned the privacy issue is very relevant in data mining tasks and there is growing public concern over the issue (e.g. [8], [18], [35], [44], [61]). The privacy issue in general is a qualitative problem, usually based on personal beliefs and values. The term “privacy” is used in a wide range of social domains [1][v]. However, our understanding of privacy is conceptually fragile. In this work, the philosophical and legal discussions on privacy are left out[vi]. The focus will be on the privacy preservation from the computational perspective.

2.4.1 Privacy

There is rich literature on data mining and knowledge discovery. However, the literature considering privacy issues in data mining area is relatively new. R. Agrawal and Srikant were the first to address privacy issues in data mining in 2000 [2]. Their work focused on formalizing privacy in terms of confidence intervals; they also showed how to reconstruct an original distribution from distorted samples. The confidence interval was measured based on how closely the original values of modified attributes could be estimated. This first work addressed the core idea of privacy by analyzing the threshold when privacy is violated.

Privacy can be divided into two classes:

INDIVIDUAL PRIVACY

CORPORATE PRIVACY.

[v] The meaning of privacy is changing. In the emerging information economy, privacy no longer means preventing organizations and other people from knowing about us. Instead, privacy now refers to concerns about the use and sharing of information –what shall and, crucially, shall not be done with personal data [16].

[vi] A detailed study of how privacy has been seen in the academic literature throughout history, see [41]. US Supreme Court Justice Louis Brandeis wrote in 1890 that the “right to be left alone” is one of the fundamental rights of a democracy. This is considered to be the first definition of privacy [62]. Privacy is also very closely related to individuality – a human being regarded as a unique personality [1]. Individuality is a very strong western value and has therefore an impact on every discussion on privacy issue. Some European legislative documents on privacy issues can be found in, for example, [1].

While the individually sensitive and corporate sensitive information link and overlap, the privacy-preserving problem in each case is different. The corporate privacy issue is usually about business secrets, and thus about sharing the data with other agents. From the individual privacy perspective, the privacy issue is not an issue if users have given authorization to use the data for the data mining task. However, if such a data mining task has not clear authorization, then what use produces privacy violation? This is the question that Agrawal et al. [2] tried to address at the beginning. Another aspect of individual privacy in data mining is individual identification. If the data at hand is anonymous, can individuals still be identified from the results?

Combining both intrusion and individually identifiable leads to a standard definition of privacy-preserving data mining [58]: A privacy-preserving data mining technique must ensure that any information disclosed

1. cannot be traced to an individual
2. does not constitute an intrusion.

Formal definitions for both of these are challenging. In Table 3, some examples of how privacy is said to be preserved in the literature are presented. To the best of my understanding, only Agrawal et al. [1] provides a metrics explanation for the privacy measures. The key question in the quantification of privacy is how to measure it.

Table 3. Examples of how the privacy is preserved in the data mining context in literature.

Agrawal [2]	A basic approach to preserving privacy is to let users provide a modified value for sensitive attributes. Two methods to generate this modified value are presented: value-class membership and value distortion.
Evfimievski [19]	A randomization method is used with a distributed database scenario. A server agent can recover aggregated properties of each of the data sets while individual entries are distorted.
Frikken [22]	Compute distance functions of routes in a private manner. Authors present a secure multiparty computation method for calculating the distance between two objects in a secure way. These objects can take several forms: points, points moving in space, and line segments. The area of points to be considered by the protocol that determines if a specific point and segment are within the threshold of each other is illustrated in Figure 5. It is enough to focus on this area because this can be run on all pairs of points and segments.
Kissner [33]	A set of operators that preserve privacy are presented. A situation occurs when two or more parties perform computation and no party learns more information about other parties' private input. Operators presented are union, intersection and element reduction.
Tasoulis [53]	Privacy concerns focus on how clustering is performed over vertically partitioned data. Clustering algorithms need to be applicable without sharing the data between parties or with a third party. Second condition in preserving privacy in clustering data is that no private information can be deduced from the results. A *k*-window algorithm was used for the clustering of the data.
Thuraisingham [57]	Privacy constraints processing. Privacy constraints are specific rules that determine the privacy level of the data. For example, all medical records are private, and all financial records are private except for those who work in public office. Constraints can assign privacy values to attributes, relations or even a database.

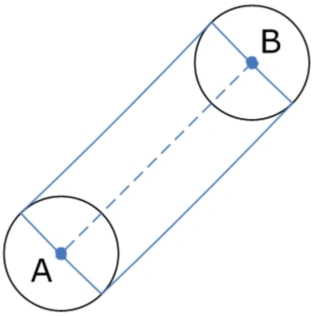


Figure 5. Area of all points within distance T from segment AB.

The problem with individual identification can be understood clearly in the next example. Consider the case where a data provider has sensitive data that the provider wants to share, with, for example, a research company. The data could be made anonymous by removing person-specific information. Although the data itself will not enable individual identification, the problem still exists. Joining the data with other sources must also not enable identification. An example of this is presented by Sweeney [50]. She studied medical records that included person-specific data with all explicit identifiers removed (name, address, telephone number etc.). The common assumption with this type of practice is that the data is anonymous because the data looks anonymous. However Sweeney's study shows that 87% of the population in the US could be identified based on only three different variables (ZIP-code, gender, and date of birth). By combining information from 'anonymous' medical records with voter registration lists Sweeney could thereby find out names and addresses for the corresponding records. The left circle in Figure 6 shows some of the attributes from the medical records. This information was linked using ZIP, birth data and gender to the voter information shown in the rightmost circle.

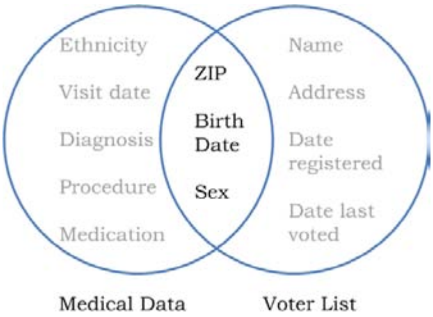


Figure 6. A combination of the medical records with the voter list information [50].

As shown by Sweeney, removing all explicit identifiers prior to release of the data is not enough to preserve the confidentiality. This same problem has been studied by Chiang et al. [13] in the health-care sector. The question of protecting privacy arose in their case because the office of national health insurance collected and maintained large amounts of health-related data. Despite the sensitive nature of the data, the office wanted to utilize the database for research purposes because of the possibly significant benefits to people, e.g. track certain diseases or a patient's response to a certain drug. The basic idea in data confidentiality is that, in the released data, no field value could be connected to an individual. This was accomplished by, for example, grouping people by heights of within 5 cm.[vii]

The User Selected Pseudonyms method allows users to control, either directly or through privacy preferences, what information about themselves they share with an environment to manage the risks to their privacy [12]. This method allows delivery of personalized services while allowing users to maintain their desired level of anonymity.

The individual privacy can be handled by asking for authorization to use the data for data mining tasks. This still does not solve the corporate privacy issues. The individuals may trust the company holding the data, but the trust might not include third parties. Or the company might not want to share data because of the leak of business insights. So, the data collection holder may be willing to take part in a data mining project, but only if its data fields are not revealed. In the corporate model, often it is the body of data that must be protected [14]. Usually privacy concerns just restrict the free flow of information. Companies and organizations do not want to reveal their private databases for various legal and commercial reasons.

2.4.1.1 Ubiquitous Privacy

Chapter 2.1 presented the idea of ubiquitous computing and some examples of possible services were shown. Table 1 showed also some techniques that might

[vii] This technique is known as *k*-anonymity and is discussed in more detail in Chapter 2.4.3.

be used to collect personalized data, e.g. location information. Teltzrow et al. [54] have conducted a study on user privacy preferences on personalized systems. They looked at thirty surveys or summaries of survey results. The main idea overall is that consumers are more concerned about privacy in online interactions than they are on regular interactions (e.g. online shopping vs. supermarket). The Sun Microsystems CEO had already declared in 1999 that the privacy as we now know it will be gone and that people should change their attitudes on the privacy matter [64]. The same idea can be seen in, for example, the Demos project on privacy [16].

Table 4. Examples of the privacy risks produced by the use of services.

Case	**Actions**	**Privacy concern**
InClass [28]	RFID-tags embedded in the students nametags in elementary school.	A number of parents of students protested the use of RFID in the school.
MetroGroup [1]	German supermarket operator. Has been trialing RFID chips in loyalty cards at its “future store”.	Canned the trial following massive privacy concerns fronted by consumer organizations.
AOL search [18]	19 million search queries publicly available.	Some people were identified by the search queries they made.
Facebook (www.facebook.com)	A social media service implemented a feature News Feeds that displays every action you take on the site to your friends.	Upset many Facebook members who responded with outrage, groups left out of protest.

Whatever the surveys and speeches say about privacy, people’s behavior could still be different. Some indication of this may be found in the examples presented in Table 4. An example of the high level of privacy concerns among people in general is presented in a pilot project called InClass. A company InCom [28] has created a system called InClass for teachers in high schools. The InClass system helps the teacher in taking, recording and reporting attendance in schools. This attendance recording is performed by the system as students wear RFID -equipped nametags. The system was created with the aim of cutting down the teachers’ administrative time. Despite the high savings of teachers’ time, parents protested against this system due to privacy concerns during the pilot project. The InClass system was called off by the company within a month [44].

Also some new services, e.g. Jigsaw and Docusearch, have raised new aspects of privacy into public debate. Some interesting discourses on the situation where one’s personal information is passed on to a third party can be found in [31], [56]. A service called Jigsaw is encouraging people to enter business contacts into an easily accessible web database. The idea is similar to the idea behind most of the social media services: an advanced use of personalization and user-generated content. Users can log in and enter contact information from other people, e.g. information acquired from business cards or email signatures. All information is entered anonymously and members get two contacts for each one they enter. Although the Jigsaw service wants only business information, the debate has centered on the question of whether this service constitutes an intrusion. On the other hand, the Docusearch service uses online and proven investigative techniques to obtain information on people and companies. This obtained information that could all be found online, and thus could be passed to a party who wants to see it.

You may think these examples presented here are just individual situations. If so, you may think about it this way: next time you are at a party, tell a stranger your salary, checking account balance, mortgage payment and social security number. If this makes you feel uneasy, you have an answer why privacy is controversial [56]. This just shows how the use of combined database information can be valuable knowledge, while information about a single variable alone is not. In a previously presented example, Jigsaw pointed out an important aspect of data sharing: users trade data to get data. This is one way to make use of combined databases.

One big problem with the privacy and integrity issue is the correct use of ubiquitous services and technologies. Consider, for example, the RFID technology for tracking products. RFID is very useful where it can be embedded within an object and thus used for purposes such as inventory management. While used for tracking objects, and if tracking humans (e.g. by RFID embedded id-cards) is forbidden, RFID can, in fact, still be used for monitoring human behavior (via the products such as medicine containers). This line of thinking produces an outcome that confirms that single items with embedded RFID-tags produce privacy risks. It is worth remembering that it is not the information itself which is good or bad, but rather the use of the information and the knowledge extracted from it that produces privacy risks.

2.4.2 Data Distribution

The ubiquitous-computing environment described in Chapter 2.1 will create huge amounts of data collected from various sources. Although the future applications will bring new types of collected data, e.g. from user locations, this ubiquitous data collection can be analyzed in the current-day setting. In this section, the different models of how the source data is organized are presented. One way to distinguish models is to analyze whether the data sources are centralized or distributed.

The techniques of the privacy-preserving data mining focus on the data mining tasks when we are not allowed to see the data. If all the data is collected and the mining tasks are executed at a single central aggregator, the actions will not pose a privacy risk if this aggregator is trusted by the agents providing data. The following data distribution section starts with the assumption that the sources and mining of the data are not located at the same site.

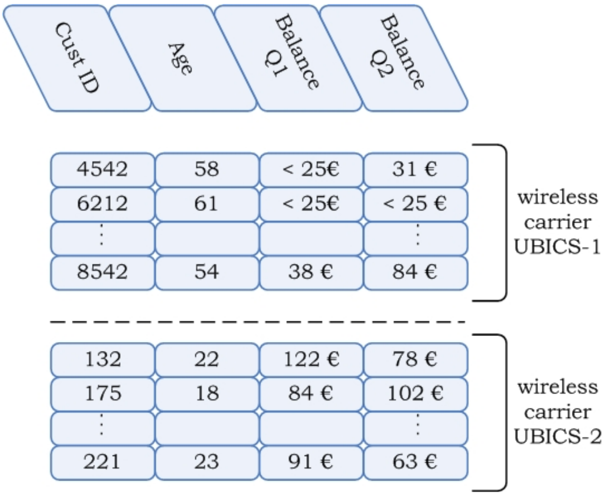
There are two basic formulations of the data distribution/partitioning models:

HORIZONTAL DISTRIBUTION

VERTICAL DISTRIBUTION.

"Horizontal distribution" refers to cases where different records with the same attributes are located in different places. It is possible that several parties collect similar data from different people, e.g. all insurance companies collect similar information while the customer base tends to be quite different. An example of horizontal distribution is shown in the Figure 7. Vaidya et al. [58] formally defines the two models as follows:

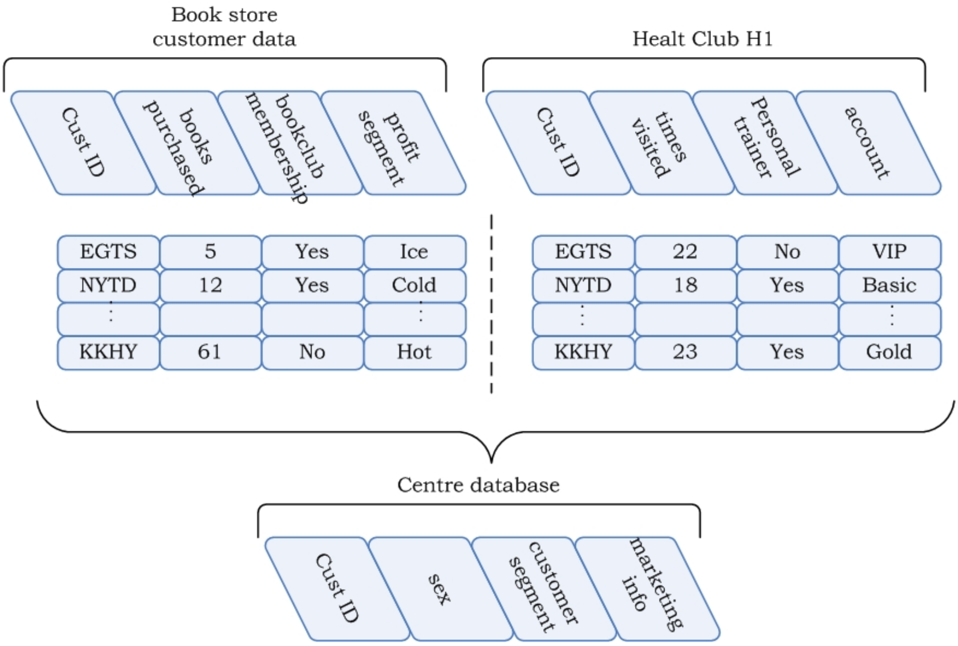
Dataset *D* is defined in terms of entities for whom the data is collected and the information that is collected for each entity. $D \equiv (E, I)$, where *E* is the entity set for whom information is collected and I is the feature set that is collected. If there are k different sites then $D_1 \equiv (E_1, I_1), \ldots, D_k \equiv (E_k, I_k)$. Therefore, in horizontal partitioning, $E_G = \cup_i E_i = E_1 \cup \ldots \cup E_k$ and $I_G = \cap_i = I_1 \cap \ldots \cap I_k$.



Cust ID	Age	Balance Q1	Balance Q2	
4542	58	< 25€	31 €	wireless carrier UBICS-1
6212	61	< 25€	< 25 €	
⋮			⋮	
8542	54	38 €	84 €	
132	22	122 €	78 €	wireless carrier UBICS-2
175	18	84 €	102 €	
⋮			⋮	
221	23	91 €	63 €	

Figure 7. Horizontal partitioning / homogeneous distribution of data.

"Vertical data distribution" refers to a situation in which the values of different attributes are collected by different agents. For example, with the same set of consumers, a book store might collect information on the customer's reading habits while a health club has the knowledge of the customer's exercise habits. These databases could then be jointly linked and used for research analysis to gain a better understanding of the customer's interests. An example of vertical partitioning is shown in Figure 8. Formally the vertical partitioning can be represented as $E_G = \cap_i E_i = E_1 \cap \ldots \cap E_k$ and $I_G = \cup_i = I_1 \cup \ldots \cup I_k$.



Cust ID	books purchased	bookclub membership	profit segment
EGTS	5	Yes	Ice
NYTD	12	Yes	Cold
⋮			⋮
KKHY	61	No	Hot

Cust ID	times visited	Personal trainer	account
EGTS	22	No	VIP
NYTD	18	Yes	Basic
⋮			⋮
KKHY	23	Yes	Gold

Figure 8. Vertical partitioning / heterogeneous distribution of data.

2.4.3 Techniques

The privacy-preserving data mining methods do not actually differ from the regular data mining methods. The goal for using each algorithm is the same, whether it is clustering, segmentation, pattern discovery, or something else. The novelty arises from addressing the privacy issue related to either to the data or the possible results. The methods are not presented here in a detailed manner. The overview of the privacy-preserving techniques is given. In this work, four approaches to privacy-preserving data mining are identified, based on the two questions [8]:

1. How the data is organized?
2. Which kind of privacy? (What is hidden vs. published/shared?)

The four privacy-preserving approaches are knowledge hiding, data perturbation and obfuscation, distributed privacy-preserving data mining, and privacy-aware knowledge sharing. The distribution of data was discussed in Chapter 2.4.2; the two ways presented were horizontal and vertical distribution. Here, those two are combined together as a distributed model, and it is the question of whether the source data is distributed among various parties or is located in a centralized party. The taxonomy based on how the data is organized is presented in Figure 9.

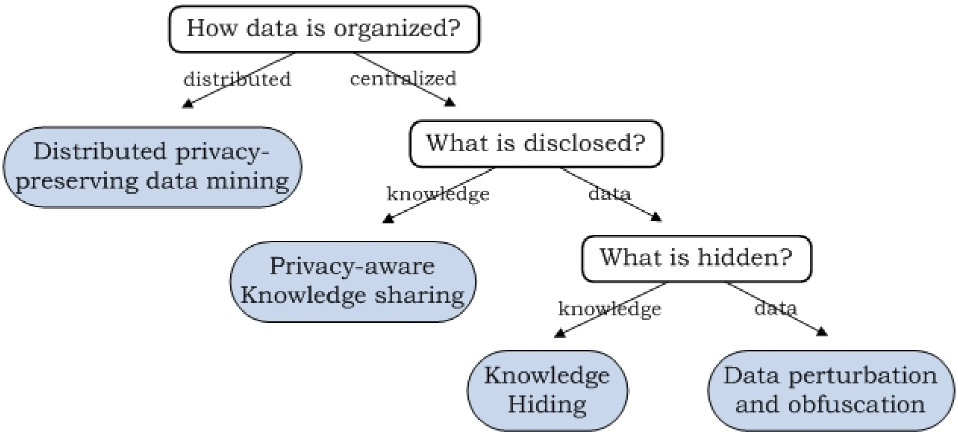


Figure 9. A taxonomy tree of privacy-preserving techniques based on how the data is organized [8].

The privacy issue, whether the privacy problem is related to individual information or corporate secrets, can be used on the basis of the second taxonomy. The privacy-based taxonomy is presented in Figure 10. Concerning the privacy issue, it is worth remembering here that most of the privacy-preserving methods, especially in the CRM sector, are based on the following two facts [1]:

1. Users are not equally protective of all the values in their record.
2. Data mining problems do not necessarily require individual records; statistical distributions of the data set might be sufficient.

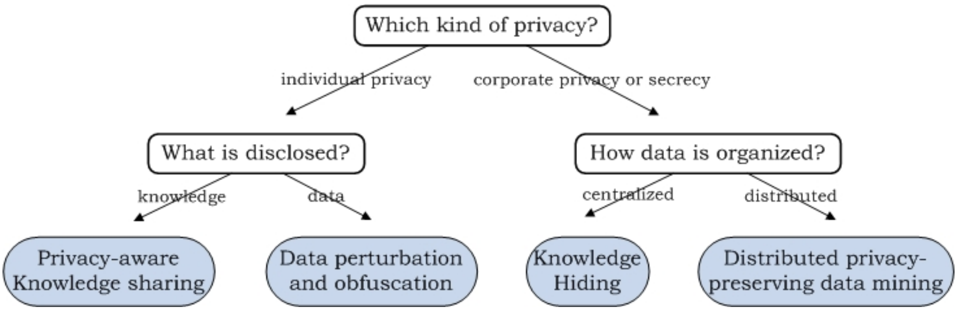

Which kind of privacy?
individual privacy
corporate privacy or secrecy
What is disclosed?
How data is organized?
knowledge
data
centralized
distributed
Privacy-aware Knowledge sharing
Data perturbation and obfuscation
Knowledge Hiding
Distributed privacy-preserving data mining

Figure 10. A privacy issue based taxonomy tree of privacy-preserving techniques [8].

Each approach should also include methods that can handle both types of data:

QUANTITATIVE DATA

BINARY AND CATEGORICAL DATA.

A common problem with privacy is how to release a version of private data so that the individuals cannot be identified. One method to prevent the individual identification of person-specific data is *k*-anonymity [50]. The definition of *k*-anonymity states that there must be at least *k* records with the same quasi-identifier, no record must be unique. As such, a *k*-anonymity protection in data release is provided if the information for each person included in the release cannot be distinguished from at least *k*-1 individuals, also included in the release.

2.4.3.1 Knowledge Hiding

In knowledge hiding, the real data is disclosed, although modified somehow. The data is modified in order to prevent the interpretation from sensitive knowledge. While the purpose is to modify the database in such way that the sensitive knowledge can no longer be accessed, the original database should still be modified as little as possible.

The knowledge-hiding approach can be explained easily by association rules. Let us suppose that we have database *D*. The association rules *R* can be mined from this database *D*. The privacy problem here is that some part of the rules in *R* includes sensitive knowledge. The database *D* should be modified to a database *D’* in such way that $R \backslash R_h$ can be mined from it [59].

$$\begin{tikzcd} D' \arrow[d] \arrow[r, phantom, "\subset"] & D \arrow[d] \\ R \setminus \mathrm{R}_{\mathrm{h}} & \mathrm{R} \end{tikzcd} \qquad (2)$$

2.4.3.2 Data Perturbation

In contrast to knowledge hiding, the data perturbation approach refers to a practice where the real data is hidden. The privacy of an individual record is protected by modifying the original values of the database. The modification is made so that it is not possible to identify the original value of individual records. Although the original values are not identifiable, it is still possible to extract valid information from the data (e.g. statistical distributions that describe the original data). This approach is sometimes also called as “distribution reconstruction”.

Usually perturbation techniques fall into two categories: probability-distribution category and fixed-data perturbation category. In its simplest form, in fixed-data perturbation methods some noise e (e.g. drawn from some probability distribution) is added to a confidential attribute X to result in a perturbed attribute Y $(Y = X + e)$. With a multi-attribute database situation, the perturbation method is applied in each of the attributes independently of the other attributes.

Obfuscation techniques include:

PERTURBATION accomplished by the alternation of a value by a new value (e.g. adding noise)

BLOCKING the replacement of an existing attribute value with a ‘?’

SWAPPING interchanging values of individual records

AGGREGATION merging several values into a coarser category

GENERALIZATION the target attribute value is generalized by, for example, grouping the heights of people within 5 cm

SAMPLING release of data only on a sample of population.

An example of data perturbation methods can be stated with association rules as follows. Let us suppose that we have database *D*. The association rules *R* can be mined from this database *D*. Now the problem is to define to algorithms *P* and *M* such that *P(D) = D'* and *M(D') = R*. *D'* form a database that does not disclose any information on singular rows of *D* [8].

The individual's privacy is protected because the reconstruction method is only able to construct distributions, the method cannot reconstruct individual values accurately. While randomized or noisy data preserves individual privacy, it still poses a challenge to data mining. Two crucial questions are how to mine the randomized data and how good the results based on randomized data are compared to the possible results from the original data [58].

2.4.3.3 Distributed Privacy-Preserving Data Mining

Distributed privacy-preserving data mining addresses the problem of several datasets where each agent owning a dataset does not communicate with other dataset owners. The privacy issue arises from both individual and corporate privacy sectors. Whether the agents hold corporate secrets that cannot be shared among the members in computation, or whether the agents hold sensitive individual information. Thus, distributed privacy-preserving data mining is also known as secure multiparty computation (SMC). The definition of secure multiparty computation states that a computation is secure if at the end of the computation, no agent knows anything except its own input and the results [1].

The history of multiparty computation started when Yao introduced a two-party secure computation problem in 1982 [66]. It has been extensively studied since the introduction (e.g. [1]) and the original problem is now also called as "Yao's millionaire problem".[viii] The Yao's millionaire problem is introduced as two millionaires (e.g. Alice and Bob) who want to know which one of them is richer without revealing their specific wealth [1]. Yao proposed a cryptographic

[viii] Another typical example is known as 'Dining cryptographers': There are N cryptographers having dinner, when it is time to pay, the waiter tells them that someone has already paid. The cryptographers want to find out whether one of them has paid the bill or whether somebody from outside paid it. This should be done in private because the person who has paid obviously wants to stay quiet about it.

solution to this two-party problem in his original paper. The secure computation methodology is similar in all the works in the literature: the computational problem is first presented as a combinatorial circuit, and each of the parties then progress a short protocol for every gate in the circuit. Examples of other secure multiparty problems in the literature can also be found under the headings private-information retrieval, privacy-preserving statistical database, and privacy-preserving data mining.

The standard method for distributed data mining is to have a central data warehouse where each agent delivers their data. This central warehouse is then mined for the combined results. This method is not valid in the privacy-preserving data mining because of the possible sensitivity of data. The distributed privacy-preserving data mining requires some local mining with each of the datasets and some mining with the combined datasets. This method is presented in Figure 11. This distributed data mining method is based on the idea that individual data need not leave the agent, solving the privacy problem with disclosure of consumer data.

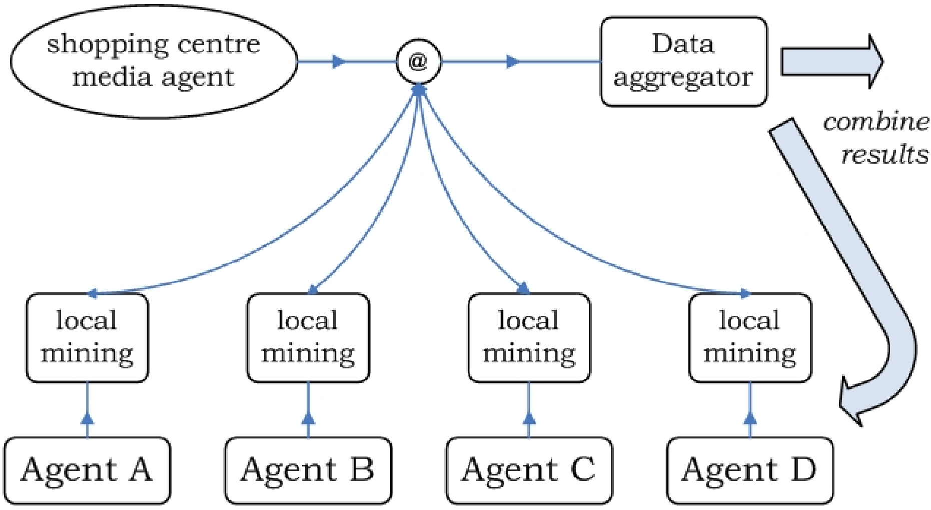


Figure 11. Distributed privacy-preserving data mining approach in shopping centre context that includes many sensitive sources of data [58].

The distributed privacy-preserving data mining approach varies from the distribution of data. The two different ways correspond to the partitioning of data, i.e. to whether the distribution of data is horizontal or vertical, or maybe even to a more complex partitioning of data. In vertical partitioning, the key element for computing is the secure scalar product of vectors representing the items, whereas in horizontal partitioning the key elements for computing are the secure union and secure sum operations [8].

2.4.3.4 Privacy-aware Knowledge Sharing

In privacy-aware knowledge sharing the important question is whether the data mining results itself violate privacy or not. The source data is hidden or left alone to the agent who holds the data. The disclosed information is the extracted knowledge from the data. The goal is to remove specific rules from the output rather than protect the input data. The two approaches achieving this are presented in Figure 12.

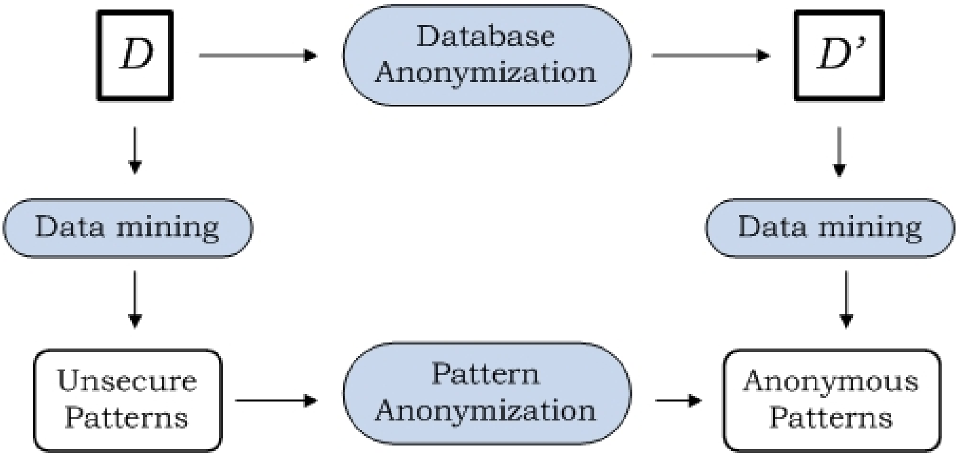


Figure 12. The different ways to achieve privacy-aware knowledge sharing [8].

In privacy the issue, the focus of knowledge sharing is on individual privacy instead of corporate secrets. This is the information about those individuals whose data is stored in the database that is mined. As far as the practitioner side of data mining is concerned, the question here is that, if the risk of privacy outweighs the reward of data mining, does it eliminate our ability to mine the data? For example, of the two approaches presented in Figure 12, the path below produces much more information than the other if the data itself is not sensitive, but the extracted knowledge is.

2.4.3.5 Applications

The privacy-preserving approaches that were discussed above all have various algorithms to consider, depending, of course, on the goal of the data mining task. These algorithms are presented in more detail in, for example, the book by Vaidya et al. [58]. Although the methods are familiar from the data mining sector, the research is fairly young in most areas of preserving privacy. For example, while there have been umpteen different regression models looked at by statisticians, all of these have assumed that the basic data is freely available at a central site. As far as preserving privacy is concerned, the only work in this area has been on linear regression [57].

The problem with a real-world application is that these privacy-preserving data mining techniques have not yet been adopted on a large scale. Table 5 presents some examples found in the literature in which the methods were evaluated with real data. It can be clearly seen that four of the seven studies presented used synthetic data instead of real data. The use of real data does not produce better results; the idea that almost none of the methods is really adapted to real-world applications can be seen from Table 5. As can also be seen in Table 5, various data mining algorithms have been considered in isolation from each other. These are, for example, decision-tree inducers, association rule mining, clustering, rough sets, and so on. An open question for future research is how these different data mining methods work together in the privacy-preserving aspect.

Although many different algorithms have been studied, the privacy aspect has still not been completely solved. This is demonstrated in, for example, the use of classifiers. The method itself might preserve privacy during the calculations, but the classifier could return sensitive data as results. Another problem is that, if the classifier takes both public and private data as input, what are the privacy issues then during and after the calculation? While some open questions remains, Table 5 clearly points out that privacy and data mining can coexist in real-world applications.

Table 5. Examples of the privacy-preserving methods evaluated with real data in the literature.

Author	**Methods / privacy techniques**	**Training and test data**
Polat et al. [1]	Collaborative filtering, perturbation techniques	Movie rating data, 1 million ratings on 3 500 movies by 7 463 users. 3 000 for training and 300 for testing.
Aggrawal et al. [2]	Decision tree classification, perturbation techniques	Two synthetic datasets: 100 000 records for training and 5 000 records for testing. Equally split between two classes (original and randomized).
da Silva et al. [49]	Distributed clustering (kernel based)	Two synthetic datasets: 500 points and 200 + 200
Merugu et al. [40]	Privacy-preserving distributed clustering	Four sets of synthetic data (5 000, 5 000, 1 000, 600)
Oliveira et al. [42]	Clustering, value distortion (geometric data transformation methods)	5 synthetic datasets (each with 6 000 points in 2D discrete space)
Atzori et al. [6]	Anonymity preserving pattern discovery, value distortion	Three sets of transaction data (8 124, 990 002, and 88 162)
Sweeney [50]	*k*-anonymity, aggregation and perturbation	Medical data and voter list

3. Material

The following two chapters cover the case study of our privacy-preserving clustering model. The goal is to use consumer data and segment the customer base into various segments. We study the difference between regular customer segmentation and privacy-preserving segmentation. The large amount of good-quality data forms the basis of the customer segmentation. The hypothesis is that the segmentation results are not any different; the privacy-preserving clustering should result in the same clusters as the normal clustering protocol. The novelty in the privacy-preserving protocol is that the data will never have to leave the agent holding the data, while still the customer segmentation remains possible.

While we study the privacy-preserving segmentation, we also study the benefits that each agent gains when taking part in this collective customer base segmentation. In this case study, we have three different agents holding the data. We study the difference of the segmentation results when the agents conduct segmentation alone using only the data they hold compared to the results of the segmentation based on the data of all the agents.

3.1 Data

When data mining tasks require available data, how can an agent get access to high-quality data that can be mined? One of the two ways to achieve this is to collect very detailed and comprehensive records from individuals and then use this database on the basis of the analysis (e.g. data mining tasks). Another way is to use various sources (e.g. different companies operating in different sectors) of customer data. Data mining techniques can thus be used on each of the databases while producing comprehensive results.

As indicated in the discussion of ubiquitous computing and services in Chapters 2.1 and 2.2, new sources of sensitive data will become relevant in the near future (e.g. real-time location from mobile devices or customer-behavior data tied to context). Also, some of the promising technologies (e.g. tags, sensors, and web services) will create new ways of sharing customer data. From the marketing perspective, the need is to have relevant data about the context and near-time behavior history of the customer. The important question is whether it can be possible for one agent to collect all the information. Or is it more realistic to use various data sources? In this case, we study this latter perspective, when the data is distributed into various locations.

In this study, we use one customer database, that of a company operating in the mail order sector [43]. Although it would be relevant to study the actual meaning of the attributes, in this work we do not analyze them in detail. The data is divided into three partitions by vertical partitioning. We use only subset of attributes. This means that each agent holds some information about each of the customers. This distribution of data simulates the scenario of a shopping centre or the downtown area of a city where the customers only visit some stores or restaurants. The restriction that this division model faces is the combination of the variables (e.g. the same customer visits two stores, which then might hold the same information). We treat each attribute as being individually owned by an agent.

The database includes 20 146 customers. The three agents each hold a subset of variables (A: 8 attributes, B: 8 attributes, and C: 10 attributes). In addition, each agent holds the information about the customer number. Customers are identified based on this arbitrarily generated number. All the other variables are unique among the agents. Figure 13 shows an example of the histograms of three variables.

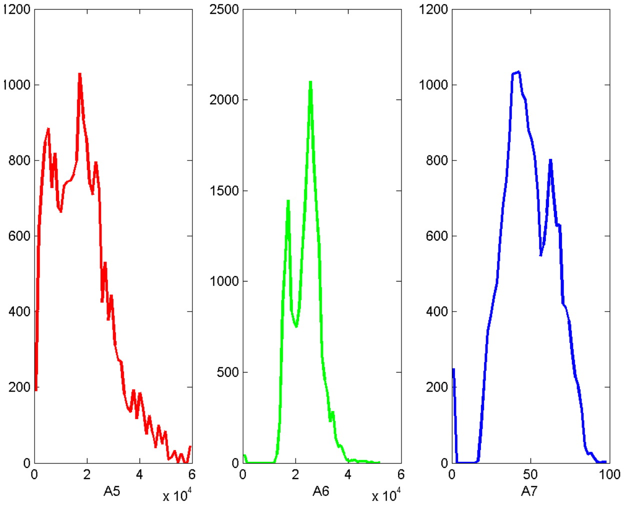


Figure 13. Histograms of three variables (A5: Number of residents in postcode area, A6: purchasing power/resident in postcode area, and A7: age of addressee in years).

As can be seen from Figure 13, the scale varies among the variables. The scales of some of the variables are close to each other; if a value is revealed, it is not perfectly known which attribute value it is. In contrast, some variables are unique in the scale of their values. This value comparison is important because we use the secure sum method and, while it is secure in the sense that the computation will not reveal the actual value, if the scale of a value is known, it might reveal sensitive information. To deal with this problem, we scale all the variables to a zero mean and a variance of one.

In addition to this privacy principle in the secure sum protocol, we also face the question of scale of the attribute values in closest cluster determination. We use the Euclidean distance as our distance measure and need the attribute values in uniform scale. Figure 14 shows the histograms of the same A5, A6, and A7 variables as were presented in Figure 14. This time the variable values are in the scale of (0, 1). Each agent conducts the scaling of its own attributes locally, thus holding all the information about the particular attributes.

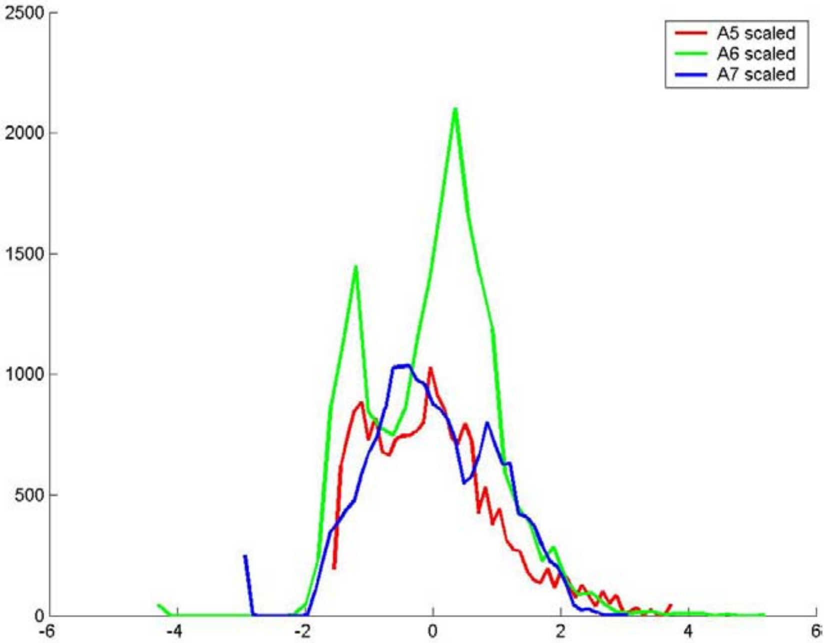


Figure 14. Histograms of three variables when scaled (A5: Number of residents in postcode area, A6: purchasing power/resident in postcode area, and A7: age of addressee in years).

The data in this case is acquired from one player, and are then partitioned into three. In real life, the simulated situation would include various agents each with a different database version. Although the data will describe the same set of customers, compatibility is a problem. How will two different databases work together? How can differences be managed? While our protocol only needs one numeric value passed to another agent at one time, the protocol should be autonomous in order to be efficient. In this work, we do not study these restrictions (such as protocols used to pass the values). Neither do we study the quality of source data in order to ascertain how reliable the database values are, for example, to see whether part of the data has been collected from a survey, or whether part of the data is from online behavior where pseudonyms are sometimes used.

4. Methods

4.1 Clustering

In this work, the privacy-preserving clustering method is used for the data presented above. Clustering is an ideal method for finding regularities in data. One way of expressing regularities is to put a set of objects into groups that are similar to each other. This is basically what clustering is all about, grouping things together. As discussed earlier in Chapter 2.3, the idea of targeted marketing is to find a small subset of the whole customer base at which to target the advertisement. What is the right subset, how large and what kind of marketing is relevant in each case is not discussed here. In this chapter, we discuss just one way to take a set of *n* objects and group them into *k* clusters, and do all this while preserving the privacy of each of the n objects.

There are several reasons for choosing clustering. First, a good clustering has predictive power. When a customer, for example, enters a movie rental store, she may look through the top 20 movies of the past week, but she can also look through the available movies from the genre point of view. If she does not know of anything particularly interesting about this, she might rent a movie based on the leading actor or director. All of these predictions, while uncertain, are useful and helpful to the customer when choosing a movie. In this movie rental case, by clustering the data, we believe that the cluster labels are meaningful, that they describe the data in more detail, and will help us choose better movies.

Second, the clustering can be a useful aid to our communication. The movie rental store owner might, for example, say to a co-worker something like ‘we will put all the cartoons here’, rather than ‘we will put all those movies that have characters that are drawn on paper here…’ The brief category name ‘cartoon’ is helpful because it is enough to describe a movie in this situation. The third reason for using clustering is that failures of cluster models may point out interesting behavior or special objects that deserve attention. This is actually better known as outlier detection, a field in data mining where the goal is to find special objects, outliers that do not fit in with the normal behavior. The idea is that the cluster model can help one focus to the objects that really deserve attention. However, outlier detection is not seen as relevant from the privacy-preserving perspective because items/customers that are highlighted lose all privacy at the individual level. This is necessary for the true positives in any case; the problem remains for the false positives – entities identified as outliers without really being so [58].

Clustering models may also serve as models of the learning process in neural systems [26]. Clustering method can also be used to find if there are any specific subgroups that are similar to each other. This idea is actually the main characteristic that separates clustering from other data mining tasks such as classification, regression, associations. These tasks have clearly defined research question and thus the 'right answer'. Of course, we may not develop optimal classifier, but still the task follows a clearly defined path. Clustering is more of an exploratory process. In clustering, we do not know the cluster mean in advance as we may not know even the proper number of clusters.

In this work, we study *k*-means clustering from the privacy-preserving perspective. *K*-means clustering [26] is a simple algorithm to group items into *k* clusters. In *k*-means clustering, each item is placed in its closest cluster. The cluster centers are calculated as the mean of the cluster members. This procedure is repeated until the center positions stabilize. The basic idea of the *k*-means clustering algorithm is as follows:

```
Input:      Database $D$, integer $k$

Output:     Cluster centers $\mu_1 \dots \mu_k$ and assignments to clusters

1. Arbitrarily select $k$ objects from $D$ as initial cluster centers $\mu'_1 \dots \mu'_k$.

2. Repeat   (a) $(\mu_1 \dots \mu_k) = (\mu'_1 \dots \mu'_k)$

            (b) Assign each $d_i \in D$ to the cluster whose center is closest.

            (c) Compute the centers of the $k$ clusters as $\mu'_1 \dots \mu'_k$.

Until $(\mu_1 \dots \mu_k)$ is close enough to $(\mu'_1 \dots \mu'_k)$.
```

The results come in two parts: first the cluster centers are produced as output and then the assignment of each object to a certain cluster. The closest cluster is usually determined by Euclidean distance, but other distance measures are also used (e.g. Manhattan distance, Minkowski distance) [26]. The *k*-means proceeds iteratively until the specified termination condition is reached (e.g. the centers move less than 1e-10). The *k*-means clustering can easily find a local minimum for the cluster centers; the global minimum is not necessarily reached. One way to avoid this is to run the *k*-means protocol for the data numerous times (50 times, for example) and then choose the best.

The iterative process of the *k*-means algorithm poses some challenges for preserving individual privacy. Simply tracking the cluster membership as cluster centers move will reveal more information about a customer moving between clusters than simply cluster membership information. Also applying data perturbation techniques, as presented in Chapter 2.4.3.2, to the clustering protocol have some challenges. If the goal is to determine the cluster that an individual belongs to (or if an individual is an outlier), perturbation-based techniques will give completely distorted results – even though the general clusters may be okay, the indication of which individual is in which cluster would be completely altered [58].

The privacy-preserving *k*-means algorithm used in this work is presented in Chapter 4.3. In order to make it fully work, first we need to analyze some security issues during the cluster computation.

4.2 Security

As discussed earlier in Chapter 2.4.3.3, the method with distributed inputs used in privacy-preserving data mining tasks is also known as secure multiparty computation (SMC). As the definition of secure multiparty computation states that a computation is secure if, at the end of the computation, no party knows anything except its own input and the results [1]. In this work, we use a secure sum method that is frequently used in distributed data mining algorithms (e.g. [32], [1]). The overall idea of a secure sum is simple, but very useful in cluster computation.

The problem is defined as follows [58]: with j agents, each agent P_i has a private value x_i. Together, the agents want to compute the sum $S = \sum x_i$ securely. One assumption is that the upper bound of the sum is known. Thus, it is assumed that the sum is in the field $|F|$. The following steps are taken in the secure sum of at least three agents:

- P_1 generates random number r from uniform random distribution over the field $|F|$ and sends it to P_2.
- For agent $P_2,\ldots,P_{j-1}$
  - P_i receives S_{i-1}
  - P_i computes $S_i = S_{i-1}+x_i$ mod $|F|$ and sends it to site P_{i+1}.
- P_j computes $S_j = S_{j-1}+x_j$ mod $|F|$ and sends it to site P_1.
- P_1 computes $S = (S_j-r)$ mod $|F|$ and sends it to all other agents as well.

The above protocol is secure in the SMC sense. Figure 15 shows the secure sum method in an example with five agents. In the example, one of the agents act as the media agent in a shopping centre and four others as basic agents (e.g. store owners).

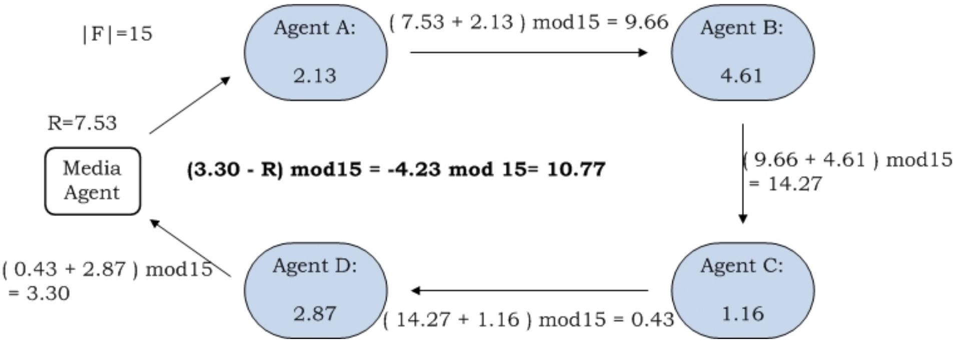

|F|=15
R=7.53
Media Agent
Agent A: 2.13
(7.53 + 2.13) mod15 = 9.66
Agent B: 4.61
(9.66 + 4.61) mod15 = 14.27
Agent C: 1.16
(14.27 + 1.16) mod15 = 0.43
Agent D: 2.87
(0.43 + 2.87) mod15 = 3.30
(3.30 - R) mod15 = -4.23 mod 15= 10.77

Figure 15. Secure computation of a sum [69].

In our case, each of these agents holds data that is necessary to cluster computation. The data is sensitive in nature, others cannot have access to it, but in order for the clustering task to work the information from each of the databases is needed. This is achieved by secure sum computation. The biggest assumption to make in this work is that each agent is semi-honest. That is, agents faithfully follow their specified protocol tasks. How this secure sum is used in the *k*-means algorithm is discussed next.

4.3 Privacy-Preserving *k*-means

We now describe the privacy-preserving *k*-means clustering protocol. *K*-means clustering has been studied briefly from the privacy-preserving perspective by Vaidya and Clifton ([58], [1]). Their version of privacy-preserving *k*-means is performed over vertically partitioned data, and has very strict privacy restrictions. Their protocol is very close to what we present in this work, but our protocol is not as strict as theirs with information sharing. The Vaidya and Clifton protocol also requires three non-colluding sites. Although these non-colluding sites may be among the parties holding the data, the need of the existence of these agents makes us modify the algorithm. Our protocol is more suitable for customer segmentation than Vaidya's and Clifton's protocol because of the smaller communication cost. The privacy issues of our protocol are discussed in more detail in the next chapter. Also Jagannathan and Wright [30] have studied the *k*-means clustering from the privacy-preserving perspective. Their protocol is not restricted to certain types of input data, but rather the work is based on arbitrarily partitioned data. Their protocol's communication cost renders it impracticable for our case.

As already previously stated, our focus is on the customer segmentation case with three agents and a media agent. This method is applicable to situations where there are more agents involved than just three. In our case, three agents (plus a media agent) are sufficient to demonstrate the use of *k*-means clustering while preserving privacy. The agents wish to cluster their joint data using *k*-means clustering, and as a result they will receive the characteristics of each cluster.

The final output of the protocol should be an assignment of a cluster number between 1 and *k* to each customer. Each member then has the information about the final centers. If desired, this information can be shared by all the agents involved. We use the Euclidean distance to determine the closest cluster center. The Euclidean distance between two points $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$ in Euclidean n-space is defined as [1]:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (3)$$

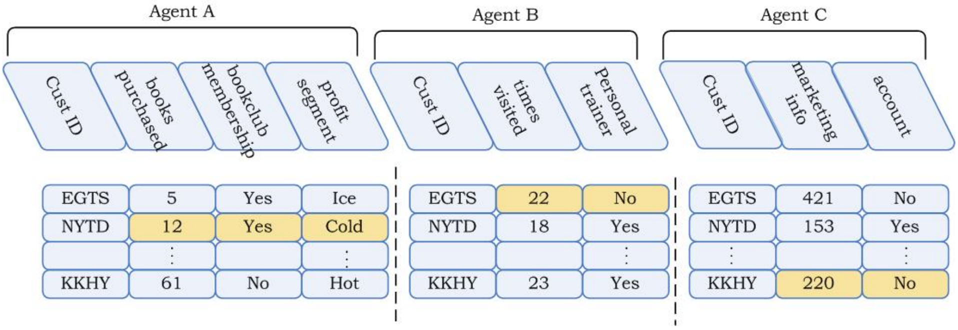
Because of the use of the Euclidean distance measure, the data must be preprocessed as presented in Chapter 3.1. In our case, the each attribute was locally scaled to a zero mean and a variance of one.

Our version of privacy-preserving *k*-means clustering is as follows:

```
Input:    Databases $D_A$, $D_B$, and $D_C$ (number of parties here n = 3), and
          integer $k$ denoting the number of clusters.

Output:   Assignment of the cluster number to the objects, and the final
          cluster centers.

1. Randomly (local) select $k$ objects from each $D_i$ as initial cluster centers
   $\hat{\mu}_k \ldots \hat{\mu}_k$. Each party knows only the information about the local variables.

2. Repeat   (a)   $\left(\mu_1^A \ldots \mu_i^A\right) = \left(\hat{\mu}_1^A \ldots \hat{\mu}_i^A\right)$,
                  $\left(\mu_i^B \ldots \mu_j^B\right) = \left(\hat{\mu}_i^B \ldots \hat{\mu}_j^B\right)$, and
                  $\left(\mu_j^C \ldots \mu_k^C\right) = \left(\hat{\mu}_j^C \ldots \hat{\mu}_k^C\right)$.

            (b) For each $d_i$ in every $D_i$
                i. Calculate locally the distance to each cluster center
                ii. Run the secure computation for closest cluster
                iii. Assign to $d_i$ the closest cluster

            (c) Compute locally new centers $\hat{\mu}_1 \ldots \hat{\mu}_k$ to the $k$ clusters.

            (d) Securely compute the amount of cluster center movement.

Until $\left(\mu_1 \ldots \mu_k\right)$ is close enough to $\left(\hat{\mu}_1 \ldots \hat{\mu}_k\right)$.
```

4.3.1 Initialization

The preparations include the scaling of the data, as shown in Chapter 3.1. The data will be scaled to a zero mean and to a variance of one. The scaling is performed locally by each of the agents. This scaling will help us later on when measuring distances between points and attributes. Each agent should save the scaling parameters as the final centers have to be scaled again to obtain understandable results.



Cust ID	books purchased	bookclub membership	profit segment
EGTS	5	Yes	Ice
NYTD	12	Yes	Cold
	⋮		⋮
KKHY	61	No	Hot

Cust ID	times visited	Personal trainer
EGTS	22	No
NYTD	18	Yes
		⋮
KKHY	23	Yes

Cust ID	marketing info	account
EGTS	421	No
NYTD	153	Yes
⋮		⋮
KKHY	220	No

Figure 16. An example of the initial assessment by the three-agent model. The points chosen as initial centers are highlighted in yellow.

The *k*-means algorithm needs an initial assessment for the positions of the cluster centers. Although the protocol is exploratory, the choice of the initial positions determines the final solutions. There are some methods developed for good initial assessment (e.g. [1]). For simplicity, in this work we select the initial points arbitrarily. Each agent randomly selects from its data *k* points, which will be the initial center values. In Figure 16, an example of the initial assessment is shown. This means that the initial cluster center assessment is not real points among the combined database. This is not a problem because, despite arbitrarily selected initial assessment, the *k*-means algorithm will eventually terminate to a fixed point. Figure 17 shows how the initialization steps are related to the whole segmentation protocol. The example presented in Figure 17 will only work with the assumption that every agent participating knows the number of customers and has a proper id number for each of them.

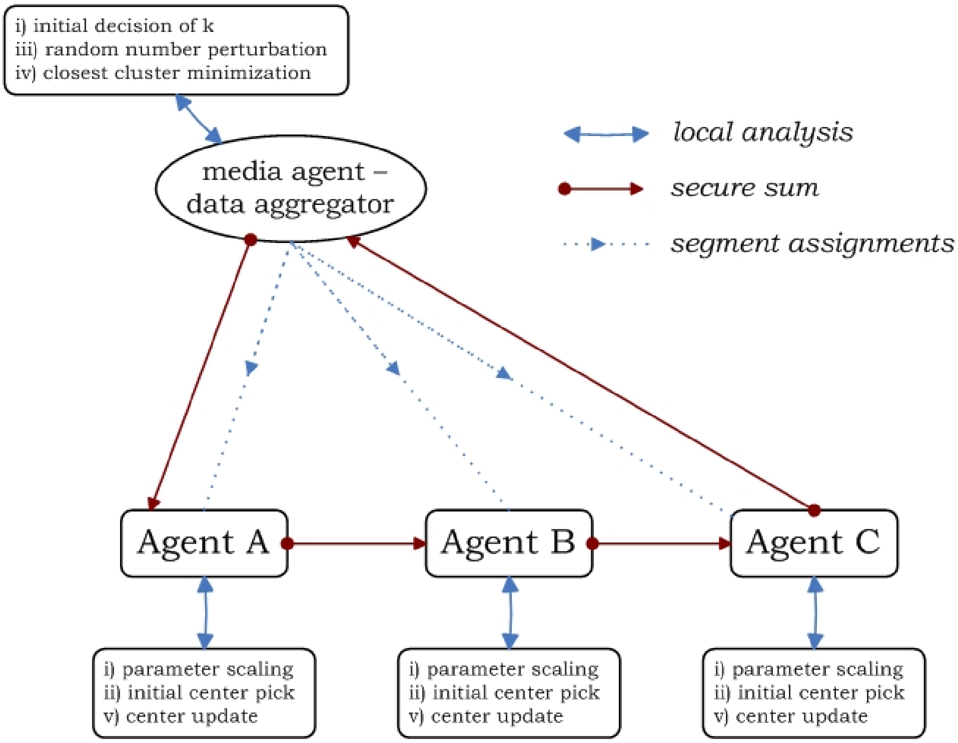


Figure 17. An example of how the information is shared between the agents in our case. Steps i) and ii) are conducted in the initialization phase, steps iii) and iv) are related to the closest cluster computation protocol, and step v) is conducted if the improvement of the iteration step is not sufficient.

4.3.2 Closest Cluster Computation

This subroutine [58] under the *k*-means algorithm will determine securely the cluster closest to each customer. This means that the routine will be invoked for every single customer in each of the iterations. Each agent has for each customer as input the component of the distance corresponding to each of the *k* clusters. This is equivalent to having a matrix of distances of dimension $k*d$ for each customer, where d is the number of attributes. For our case with Euclidean distance, this means that our job is to find the cluster where the sum of the local distances is the minimum among all the clusters. Formally stated: each attribute of every customer has its own *k*-element vector X_i:

$$A_1 \text{ has } X_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{k1} \end{bmatrix}, A_2 \text{ has} \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{k2} \end{bmatrix}, \ldots, A_d \text{ has} \begin{bmatrix} x_{1d} \\ x_{2d} \\ \vdots \\ x_{kd} \end{bmatrix} , \tag{4}$$

where the row with a minimum sum is the closest cluster, so formally we must find

$$\arg\min_{i=1\ldots k}\left(\sum_{j=1}^{d} x_{ij}\right) \tag{5}$$

Here each of the x_{ih} represents the Euclidean distance between point p_j and the center, $x_{ij} = (\mu_i - p_j)^2$.

4.3.3 Termination

The privacy-preserving *k*-means algorithm is an iterative algorithm. At the end of each of the iterations new values for cluster centers are recalculated. This calculation is conducted locally by each of the agents holding data. The new center value for each attribute in each cluster is the mean among the attribute values of the cluster members.

After the new center values are calculated the overall improvement of the iteration step is determined. Every agent can locally determine the improvement of the centers on local attributes. The media agent will determine the overall improvement by using the secure sum method on every cluster attribute. The key point is that only the difference between the old and new center values is passed on in the secure sum calculation. The real values of the cluster centers will never leave the agents holding the data. The media agent will learn only the overall improvement of the iteration step.

If the improvement of the iteration step is sufficient, then the iteration will go on. The algorithm terminates when the Euclidean distance between the cluster centers between two consecutive iterations is less than a specified value ε. Another common termination condition is the absence of any change in cluster composition. In this work, we use the specified value condition. The media agent does the comparison and will share the results among the agents. When the iteration terminates, the agents know already the final cluster assignments of all the customers and the cluster centers on the attributes that corresponds to the data they hold.

4.3.4 Cluster Validity

The term “cluster validity” is used for the procedure of evaluating the results of the clustering algorithm. There are three approaches to investigate cluster validity [25]: external criteria, internal criteria and relative criteria. Usually these validity measures are used for deciding the optimal number of clusters that fit in the data set. In this work, we use only the Sum of Squares as our validity measure, because we only want to find out which one of the trials is the best. The term “Sum of Squares” within each group is defined as:

$$SS = \sum \left(X_i - \overline{X} \right)^2 \tag{6}$$

4.3.5 Privacy Properties

The privacy-preserving protocol presented above preserves privacy in customer segmentation. As presented in Chapter 2.4.1, there are numerous definitions of privacy; our *k*-means clustering is not suitable for all of them. For example, our protocol reveals some information about the process of the computation and also some information beyond just the cluster numbers.

The first assumption to make this algorithm work is that the agents taking part are assumed to be semi-honest. A semi-honest agent follows the rules of the protocol by using its correct input; the agent is free to use what it sees during the computation protocol later on. The definition of secure computation in a semi-honest model states that the view of each agent during the execution can be simulated knowing only the input and the output of that agent [1]. This means that there is difference between saying the computation is secure and the private information is protected.

In our case the intermediate cluster assignments are shared with each agent. This contradicts the secure computation principle, but we see that the cluster assignment information is not too sensitive. All the cluster centers are held by different agents, so each agent knows only the details of its attributes. The algorithm does not necessarily have to share the attribute characteristics among the agents. The media agent in our case holds the information about the cluster assignments in the first place. This means that the media agent could be a trusted third party member (TTP) and the intermediate results could be protected. But when the intermediate results are public knowledge it makes it easier to implement the clustering algorithm in the real world. With the existence of a trusted third party member, the trust issues become very relevant.

The final results of our protocol are the final cluster center assignments. The final cluster centers are not necessarily shared and the protocol does not need them in order to work. But because it is the other reason behind conducting customer segmentation, we see that it is necessary to share the final centers with all the agents. In our case, we actually share the characteristics of each of the customer segments. One of the questions that our protocol does not address is which of the policies and actions are permitted and which are not. For example, we pre-processed the data by scaling each attribute to the zero mean and a variance of one. We did not formulate the actual standards in which to represent the data. This might provide some privacy leaks in such a case where agents share their data in a different format (e.g. in the case of qualitative attribute).

5. Segmentation

In this chapter, our privacy-preserving *k*-means protocol is applied to the data presented in Chapter 3.1. First we study the final values in each of the centers. We address the question of how the privacy-preserving *k*-means algorithm differs from the regular *k*-means algorithm. Will both algorithms find the same centers and produce the same assignments to each of the customers? And how will the different algorithms work with a different number of initial centers?

We also address the problem that an agent is able to conduct a customer base segmentation alone by using the data it holds. Why share the data with others? We look at the differences between the segmentation assignments from individual database segmentations and segmentations performed using the combined database. Although we do not analyze the characteristics of each of the final segments and their attribute values, we discuss the benefits that the sharing of data will produce. We end this chapter with a discussion about the complexity of our privacy-preserving protocol.

5.1 Segments

5.1.1 Centers

The first objective in our case was to study the privacy-preserving *k*-means algorithm compared to the normal *k*-means algorithm. The iterative *k*-means algorithm will always terminate on a fixed point. The *k*-means algorithm will not always find the optimal solution; the local minimum is also sometimes found. We avoided the possibility of local minimums by running the *k*-means protocol 50 times, after which we chose the best. We made 50 trials using 5, 8 and 10 initial centers.

Our privacy-preserving *k*-means algorithm produced almost the same results as the normal *k*-means. Figure 18 and Figure 19 present the cluster center values on each of the attributes from both the privacy-preserving algorithm and normal *k*-means. As it can be seen from the figures, the cluster centers are almost identical. When we want to find out the characteristics of customer segments, the privacy-preserving *k*-means algorithm does not differ from the normal *k*-means algorithm. The comparisons between the privacy-preserving clustering model and normal *k*-means of the $k = 5$ and $k = 10$ models is presented in Appendix A. The number of clusters does not change the outcome. Both of the clustering algorithms produce similar results.

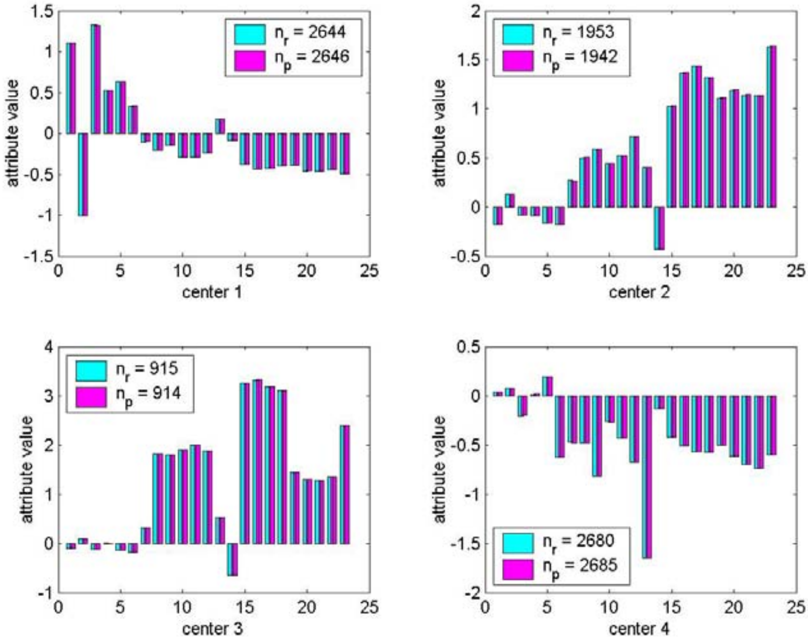


Figure 18. Comparison of the centre (1–4) values from each of the attributes in the k = 8 model.

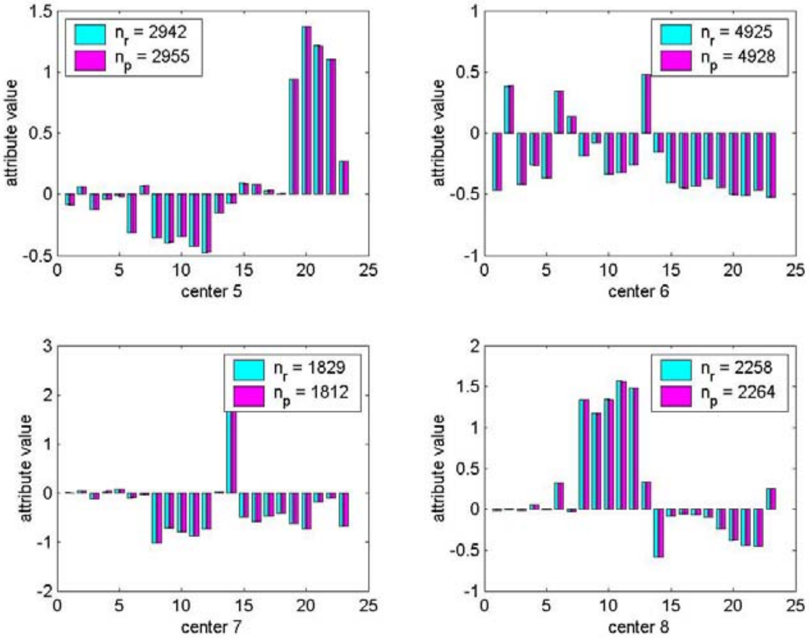


Figure 19. Comparison of the centre (5–8) values from each of the attributes in the k = 8 model.

5.1.2 Assignments

The second objective in our study was to examine the cluster assignments. In the previous chapter, the results of the cluster centers were compared. The cluster centers will be the information source for segment characteristics and thus will work only as descriptive statistics. Although the descriptive information might provide interesting insights about the data, the actual use of the clustering method for predictions will be the cluster assignments The cluster assignments tell us which segment each customer belongs to and, in case of a new customer, the potential segment.

Table 6 shows the assignments of the models, how the assignments of the protocols differ. The assignments of the normal *k*-means are taken as the baseline to which the assignments of our privacy-preserving *k*-means are being compared. The results in Table 6 show that misassignments grow as the number of cluster grow. The number of misassignments is very small compared to the whole customer base. The misassignment rate is less than 0.5% in total. Both of the clustering algorithms produce almost identical assignments.

The term “misclassification” is a somewhat misleading term because the goal of a *k*-means clustering is not to find any right answers, only some regularities, from the data. Two separate runs of *k*-means and our privacy-preserving *k*-means protocol find almost identical centers, as shown in Figure 18 and Figure 19. The customers assigned to different clusters by different algorithms are those that are situated somewhere in between the cluster centers. However, the *k*-means algorithm does not leave anything out. The possible outliers are treated the same as everything else.

Table 6. The cluster assignments of the privacy-preserving k-means algorithm compared to the regular k-means clustering.

	Same class k = 5	**Different class k = 5**	**Same class k = 8**	**Different class k = 8**	**Same class k = 10**	**Different class k = 10**
i = 1	20 140	6	20 089	57	20 078	68
i = 2	20 146	0	20 082	64	20 022	124
i = 3	20 139	7	20 083	63	20 091	55
i = 4	20 141	5	20 102	44	20 028	118
i = 5	20 140	6	20 144	2	20 015	131
mean	20 141	5	20 100	46	20 047	99

5.1.3 Individual Segmentation

One of the objectives of this work was to study the difference between the clustering results when agents share their data and clustering is collectively conducted and the results when clustering is conducted by each agent alone. The clustering is carried out by the normal *k*-means clustering protocol. Table 7 presents the results of the cluster sizes in the *k* = 8 case. The clusters are ordered in random order, but Table 7 clearly shows that the clusters cannot be mapped to each other because of the large variation in the assignments.

Although the cluster assignment sizes vary, the cluster assignments are plotted in bar-plots in Figure 20, Figure 21, and Figure 22. The cluster assignment sizes indicate that the clusters may not be mapped to each other directly. In the following figures (19, 20, and 21), the clusters assignments are compared when different datasets are used. The cluster sizes from full data are plotted on the horizontal axis, and each cluster bar is divided into those segments which are received from using datasets A, B, and C. Figure 20 represents the case where full data segments are divided into the data A segments, while Figure 21, and Figure 22 represent the case of datasets B and C, respectively. The large dispersion also highlights the fact that the clusters from different models cannot be mapped to each other.

Table 7. Different cluster sizes when different datasets are used, clusters are in random order.

	Full data	**Part A**	**Part B**	**Part C**
I	2 644	1 853	4 601	1 205
II	1 953	2 315	2 378	2 341
III	915	3 548	753	761
IV	2 680	3 355	5 279	8 774
V	2 942	3 223	3 156	1 388
VI	4 925	1 234	1 250	970
VII	1 829	3 500	1 097	792
VIII	2 258	1 118	1 632	3 915

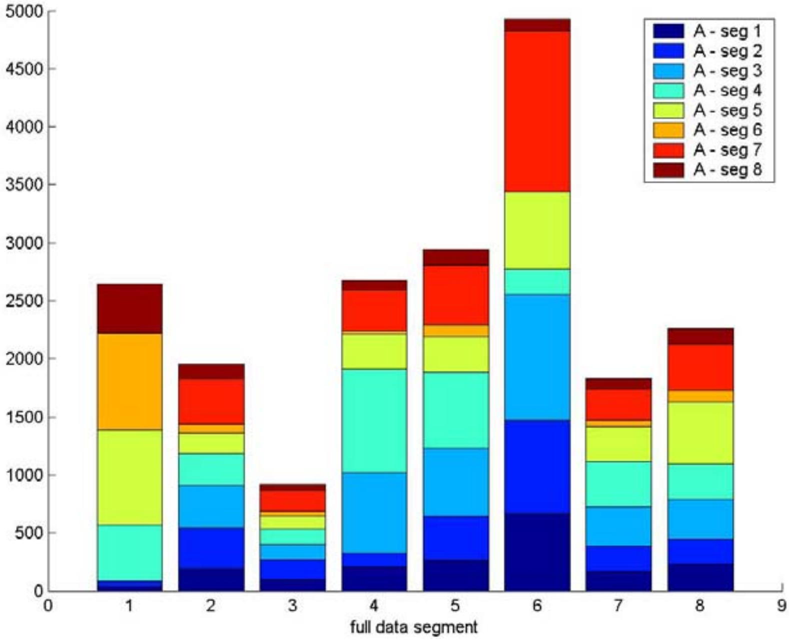


Figure 20. The segments resulting from full data segmentation are divided into segments received from the data A individual segmentation.

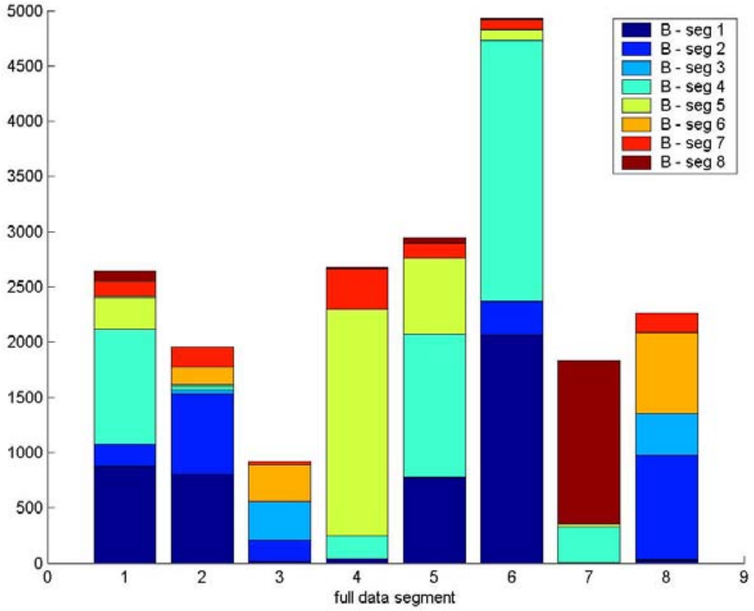


Figure 21. The segments resulting from full data segmentation are divided into segments received from the data B individual segmentation.

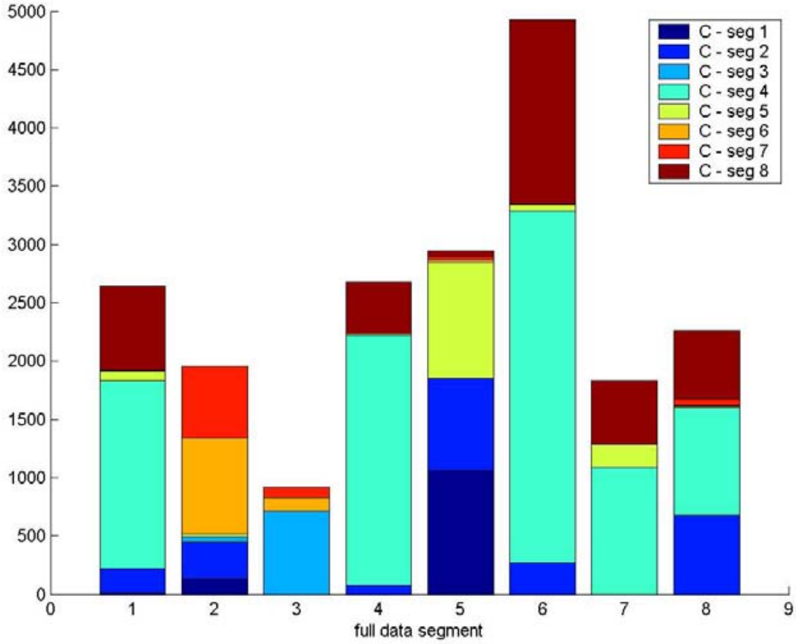


Figure 22. The segments resulting from full data segmentation are divided into segments received from the data C individual segmentation.

5.2 Complexity

The privacy-preserving *k*-means protocol presented in this work is very similar to the one Vaidya and Clifton presents in [58], so our computational complexity analysis follows a similar path. The first restriction to the analysis is that, as the algorithm is iterative in nature, the total computational complexity is dependent on the number of iterations needed to converge. The number of iterations required is dependent on the data used. Because of this, we study only the computational complexity of a single iteration of our algorithm.

Let us first refresh ourselves of the notation: *d* is the number of attributes in total, *n* is the number of customers, and k the number of clusters. The first step in our algorithm is the data preprocessing carried out locally by each party; this has the computational complexity of *O(dn)*. This initialization step is not actually part of the iterative algorithm. The determination of the closest cluster for each customer n, with d attributes, requires n*d*k steps. In addition to this, the secure sum protocol requires *n*k*(d+1)* steps. The closest cluster determination also requires of each customer the minimization over the distances between the centers. This has the computational complexity of *O(kn)*. The total computational complexity in closest cluster determination is *O(dkn)*. The calculation of new center values is again with the computational cost of *O(dkn)*.

So the total computational cost for our algorithm during one iteration step is *O(dkn),* where *k* is usually a very small constant (e.g. 5–15), and the number of attributes is usually held constant. The analysis does not take into account the communication cost between agents when passing values.

5.3 Information Sharing

The second hypothesis of this work is formulated on the idea of whether the agents should work collectively on their customer segmentation. In Figure 20, Figure 21, and Figure 22 the dispersion of full data segmentation versus individual data segmentation was presented. The relevant question for each agent is how much is the collectively conducted segmentation worth, and thus should they share their data for the segmentation task? Because in our example each agent has interesting characteristics in their segmentation, we take a closer look at each individual agent and analyze the segmentation results from the agents' point of view. Now in the figures (Figure 23, Figure 24, and Figure 25), the segmentation results from each of the individual segmentations are presented in the horizontal axis, which are then divided into the full data segments.

In the following analysis, we look at the shopping centre scenario with three agents. Assume here that agent A represents a bookstore, agent B a sports apparel store, and agent C a florist.

5.3.1 Agent A

Figure 23 presents the segmentation results from the segmentation based on data set A. It is clearly seen that four of the segments are fairly large and four are smaller. More important than the sizes of each segment is the finding that each of the segments from data A has customers from almost every full data segment.

Figure 23 shows each of the segments from data A segmentation has almost constant portions from each full data segment, with a few exceptions of course. Full data segments 3, 7, and 8 each are combined from very similar portions of data A segments. The full data segments 1 and 6 have more variation: data A segments 4, 5, 6, and 8, for example, have almost all the customers from full data segment 1. The numbers of each segment size might be confusing, so let us look at this from the shopping centre scenario point of view.

Various companies operate in a shopping centre, so assume now that agent A represents a bookstore. Some people are very enthusiastic readers and thus purchase, or borrow from the library, books from the store. The bookselling business also has an interesting characteristic: those people who are not enthusiastic rarely buy books for themselves, but rather buy books as presents. In our segmentation case, agent A might have been collecting only demographic information based on the delivery address and paying habit. This would have resulting attributes such as age, sex, address, income level, etc. The segmentation would have then produced only segments that do not make a difference between different behaviors. This is the behavior seen in the segments based on data A, and this type of attributes explains why the additional data (behavioral information) disperses the segments. How much would it be worth for the bookstore to join collectively conducted segmentation? If the store carries out marketing, the obvious benefit would be cross-selling, by focusing on relevant customers around holiday seasons and relevant customers during the rest of the year. By adding some information about its customers' physical activity behavior to the segmentation it receives, it may offer them sports-related books.

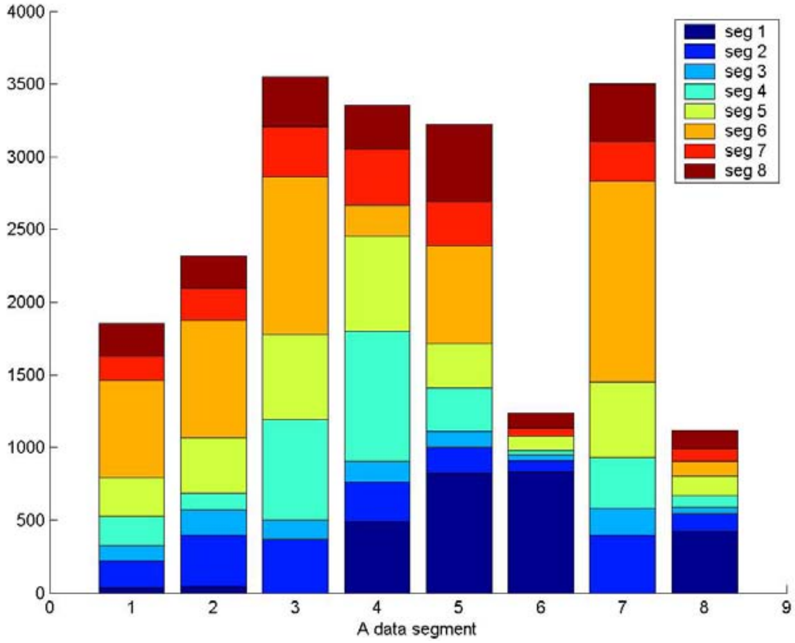


Figure 23. The segments from data A segmentation are divided into segments received from the full data segmentation.

5.3.2 Agent B

Figure 24 presents the segmentation results from agent B's individual segmentation. Three of the segments (1, 4, and 5) are relatively large, while four segments stand out as minority segments. Data B segment 8 has almost identical mapping into the full data segment 7, and data B segment 5 includes almost all the customers from full data segment 4. Data B segments 1, 2, 4, and 7 are then in contrast mixtures of various full data segments. These segment numbers indicate that agent B has a few customer segments that behave in a very similar way, and additional data from other agents do not break the segment.

In the shopping centre scenario, assume now that agent B represents a sports apparel store. Segments 5 and 8 are clear segments. The inclusion of bookstore and florist information does not affect these customers, although there is small part of full data segment 7 customers found in data B segment 4 instead of core segment 8. Segments 5 and 8 might include, for example, mainly males highly interested in physical exercising. It may seem that the store owner might not benefit at all from collective segmentation because individual segmentation already produces core segments and additional data will not provide value.

If segments 1, 2, and 4 are the core segments for the sports apparel store, the reason for taking part in collective segmentation is obvious. The full data segmentation provides additional information about the customers in these core segments. The apparel store may consider segment 5 based in data B as a core segment or a niche segment. This actually does not matter if we think about the additional information. If the apparel store wants to focus on segment 5 customers, it means that the focus is actually on full data segments 4 and 5. The additional information provides information necessary for new customer acquisition to the apparel store. On the other hand, if the store wants to focus on segment 1 based on their segmentation, they have a small portion of customers in segments 2 and 5, and huge portion in segment 4, that they may ignore without collective segmentation.

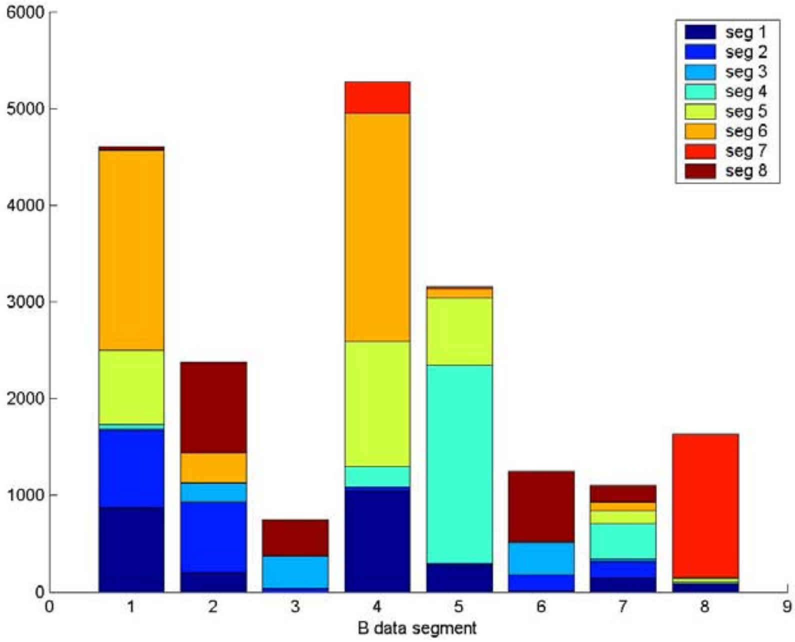


Figure 24. The segments from data B segmentation are divided into segments received from the full data segmentation.

5.3.3 Agent C

Figure 25 presents the segmentation results from agent C's individual segmentation. Two of the segments (4 and 8) are relatively large, while other segments stand out as minority segments. Data C segment 3 has almost identical mapping into the full data segment 3, and data C segments 6 and 7 are mapped into the full data segmentation as one (seg. 2). If we ignore the small portions in data C segment 2, the data C segments 1, 2, and 5 are represented in the full data segmentation by one segment: segment 5.

Let us now consider the shopping centre scenario again, and assume agent C represents a florist. Without special business intelligence about the florist's business, it is probably not a very mainstream store in the shopping centre. The florist attracts only a small portion of the whole customer base visiting the shopping centre. If segments 4 and 8 from the florist's data are combined, they represent almost 2/3 of the whole customer base, and thus these two segments (4 and 8) include customers who will not visit the florist. These are the segments for the florists to avoid. This kind of niche store may collect detailed information about its customers and segment their customer base in more detail than the collectively conducted segmentation would result in. However, the florist may benefit from the collective segmentation by analyzing the full data segment 5, which may be the best segment for the florist, but reaching for those customers (full data 5 = 2 500 customers) he would have to choose segments 1, 2 and 5 as a core (= 5 000 customers), and thus he may find savings or find new cross-selling opportunities by taking part in the collective segmentation.

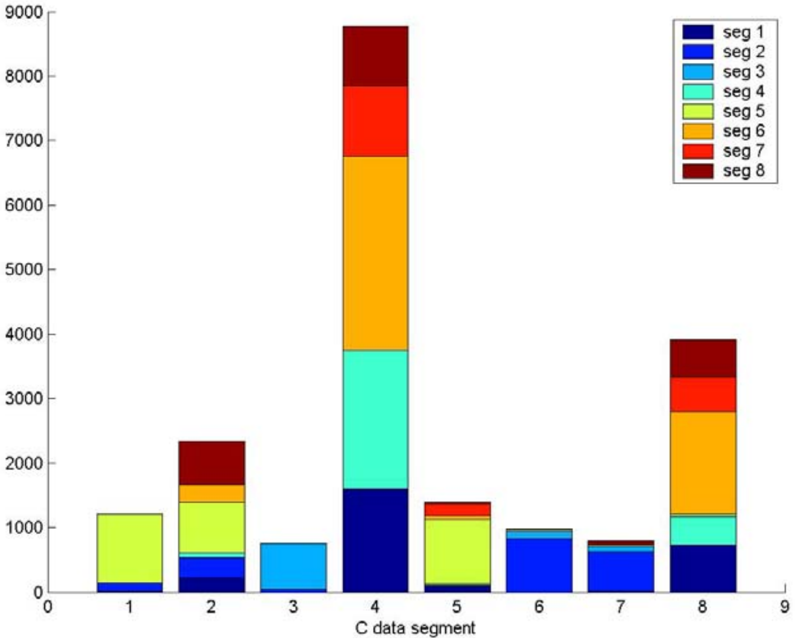


Figure 25. The segments from data C segmentation are divided into segments received from the full data segmentation.

As the scenario description for each player shows there are reasons for every agent to participate or not. The big question for every agent is how they may know before hand what the benefits will be. And how much are they worth? In the case of the florist, the collective segmentation just merges the segments, while the additional knowledge is mainly irrelevant from his business point of view. And even if the florist were up to find some knowledge from the full data segments, e.g. full data segment 5, helpful, the additional knowledge might not provide enough value for participation.

This scenario discussion illustrates the point that every agent receives additional information from collectively conducted segmentation. Each of the agents in the shopping centre has its special information about the customers along with some basic information. One of the critical tasks that remain for future studies is to define which information is critical, and which is possibly sensitive. When each agent receives some knowledge, and holds some information, who can decide the appropriate policies? Especially if the agents have conflicting goals! In our scenario, the florist had information that the bookstore, for example, found very helpful, but the florist kept all the data he needed to himself. In this case, the florist might sell the consumer data it holds for the segmentation purpose, because in the shopping centre scenario these other agents participating in the segmentation are not competitors and business secrets are not revealed. How can the differences between database values be addressed before hand? How can it be guaranteed in the real-world situation that the florists will receive sufficient compensation, and the ‘type A’ free-rider agents will not be allowed to take part?

6. Conclusions

In this work, we focused on the use of consumer data in a ubiquitous-computing environment. The novelty of ubiquitous computing using consumer data is the various sources of data. According to numerous visions, there will be new kinds of sensors, devices, panels, etc. that all produce usage information; thus it will be possible to record consumer behavior in more detail than is currently possible. The challenge for ubiquitous computing is how all the different technologies will work together. Moreover, how will those various technologies be implemented into commercial services (e.g. to benefit personalized marketing services)?

The personalized marketing services will have to address the individual privacy issue. Individuals are identified based on their previous behavior and the marketing efforts target into the context of this behavior. There has been work done on privacy, with some studies focusing on overall ethical views (e.g. [61]), while a few others try to formulate privacy within a mathematical framework (see, for example, Table 3). The privacy issue is controversial and, as the discussion in Chapter 2.4.1 indicates, in the final case for privacy there will be so many different human values in question that the mathematical framework will prove inadequate. Perhaps the most challenging task for future research is to formulate and gain a better understanding of privacy. From the data mining perspective, the relevant question is the impact of the data mining task on privacy.

We use a set of consumer data for customer segmentation. We focused on the privacy preservation in the use of the segmentation protocol. This work had two research hypotheses. The first was the privacy-preservation issue in the segmentation protocol. Our clustering protocol uses data from distributed sources and, during the calculation, the data never leaves the original site. This could imply that the privacy is preserved. However, our protocol potentially leaks some sensitive information by revealing the intermediate cluster assignments and by revealing the final cluster centers. This idea was discussed previously: the knowledge extracted from the data can produce privacy intrusion itself. The issue with privacy, and how sensitive the information of intermediate results is, is actually case dependent. Although, the intermediate cluster centers are not revealed, many would think it desirable that the algorithm be formulated the algorithm in a way that is fully secure, but still efficient.

The computational complexity of our algorithm is almost the same as it is with a normal *k*-means algorithm. Although we conducted more computational assignments, the scale factor is linear with the normal *k*-means. The communication cost that our analysis did not take into account is the communication cost between the agents. This communication might turn out to be very expensive and time consuming, but it depends on the implementation of the application. If the other option was not to perform data mining at all, from this perspective the communication complexity would seem reasonable. Future research will have to address the technology side of this communication between the agents, how the actual implementation will work.

The second of our research hypotheses related to the benefits of collective segmentation. The results in our work show that collectively conducted segmentation produces knowledge that cannot be acquired otherwise. This indicates that all agents holding customer data should share their data with others. There are basically two reasons for companies not to do this. One is that they may not want to (e.g. because they are afraid of misuse, they might reveal business intelligence/incompetence) and other is that they may not be allowed to share data (e.g. because of legislation). Our collectively conducted customer segmentation has potential to flourish if the reason for not sharing data is only the secrecy of corporate data. One interesting research question for the future is how the segmentation would differ if the parties shared only *k*-anonym data.

In our case, the data was acquired from one agent and then divided into three partitions. In real life, a similar situation would include various agents each with a different database version. We will also face a similar question with ubiquitous services that use various technologies. Each of the new technologies will produce its own application and thus data. How can different data sources be applied together into a single application? Or, can this approach work realistically in any real-world setting or is it just too simple to address the database differences in-depth?

Another challenge is the trusted third party member. These parties are hard to find, so trust issues become relevant. Our protocol would work without a trusted third party (media agent), but then one of the agents must fulfill the role. Even without having a trusted third party around our protocol still has one trust issue assumption: the honest but curious model (or semi-honest agents). The agents are honest enough to follow the protocol, but curious enough to perform calculations of their own to gain information than they are supposed to learn. The existence of a trusted third party member may also help in the initialization phase of our protocol. Our protocol was based on the assumption that each party knows the number of customers and has correct identification tags for them. The trusted third party member may help the agents synchronize their customer identification information.

One interesting research question for future research is that of how the segmentation results will work as group characteristics if the privacy issue is dealt with by grouping customers into subgroups? The information sharing issue was discussed in the previous chapter and the acquisition of knowledge was demonstrated. This value addition to agents is probably the most significant single enabler for this segmentation to be implemented. Without the value analysis and clear benefits to agents there is no reason for agents to participate. Thus, the data mining tasks are not needed if the data is not shared in the first place.

References

[1] Agrawal, D. & Aggarwal, C. On the design and quantification of privacy-preserving data mining algorithms. Proc. of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 2001. Pp. 247–255.

[2] Agrawal, R. & Srikant, R. Privacy-Preserving Data Mining. ACM SIGMOD Record, Vol. 29, 2000, pp. 439–450.

[3] A9 privacy policy, http://www.a9.com/-/company/privacypolicy.jsp.

[4] Anderson, C. The Long Tail. Random House. 2006. ISBN 184413850X.

[5] Atallah, M. & Du, W. Secure Multi-Party Computational Geometry. Lecture Notes in Computer Science, Vol. 2125, 2001, p. 165.

[6] Atzori, M., Bonchi, F., Giannotti, F. & Pedreschi, D. Towards Low-Perturbation Anonymity Preserving Pattern Discovery. In: Proceedings of the 2006 ACM Symposium on Applied Computing, SAC '06. ACM Press, New York, 2006.

[7] Becker, G. A blog writing in April 5, 2004. (www.fredshouse.net)

[8] Bonchi, F. KDD Lab Pisa. An interview on 18.10.2006.

[9] Bradley, P. & Fayyad, U. Refining initial points for *K*-means clustering. In: Proc. 15th International Conf. on Machine Learning, 1998. Pp. 91–99.

[10] Briscoe, B. Odlyzko, A. & Tilly, B. Melcalfe’s Law is Wrong. IEEE Spectrum July 06.

[11] Broersma, M. Bluejacking’ seen as marketing opportunity. ZDNet UK, 2003.

[12] Chatfield, C. & Hexel, R. User Identity and Ubiquitous Computing: User Selected Pseudonyms. Proc. of UbiComp 2005.

[13] Chiang, Y., Hsu, T., Kuo, S., Liau, C. & Wang, D. Preserving confidentiality when sharing medical database with the Cellsecu system. International Journal of Medical Informatics, 71, 2003, pp. 17–23.

[14] Clifton, C., Kantarcioglu, M. & Vaidya, J. Defining Privacy for Data Mining. In: National Science Foundation Workshop on Next Generation Data Mining, Baltimore MD, 2002. Pp. 126–133.

[15] Davis, H. Google Advertising Tools. O'Reilly Media, Inc. ISBN 0596101082.

[16] Demos.co.uk, document of the project: Private lives?

[17] Elliot, K., Neustaedter, C. & Greenberg, S. Time, Meaning and Ownership: the Value of Location in the Home. Lecture notes in computer science, Proc. of UbiComp 2005.

[18] Enter Search Term Here, Forewer. Nytimes, editorial 21.8.2006.
http://www.nytimes.com/2006/08/21/opinion/21mon2.html.

[19] European Commission, Data protection – legislative documents:
http://ec.europa.eu/justice_home/fsj/privacy/law/index_en.htm.

[20] Evfimievski, A. Randomization in privacy-preserving data mining ACM SIGKDD Explorations Newsletter, 4, 2002, pp. 43–48.

[21] Fluid time website www.fluidtime.net.

[22] Frikken, K. & Atallah, M. Privacy-preserving Route Planning. In: Proc. of the 2004 ACM works on Privacy in the Electronic Society, 2004. ISBN 1581139683. Pp. 8–15.

[23] Greenfield, A. Everyware: The dawning age of ubiquitous computing. New Riders, 2006. ISBN 0321384016.

[24] Google privacy policy, www.google.com/privacy.html.

[25] Halkidi, M. & Vazirgiannis, M. An Introduction to Quality Assessment in Data Mining. ECML/PKDD-2002 Tutorial Notes.

[26] Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001. ISBN 0387952845.

[27] Imielinski, T. & Nath, B. Wireless Graffiti – Data, data everywhere. Proceeding of the 28th VLDB Conference, Hong Kong, China, 2002.

[28] InCom www.incomcorporation.com.

[29] ISTAG – Scenarios for Ambient Intelligence in 2010. EU report 2001.

[30] Jagannathan, G. & Wright, R. Privacy-Preserving Distributed *k*-means Clustering over Arbitrarily Partitioned Data. Proc. of the 11th ACM SIGKDD int. conference on Knowledge discovery in data mining, 2005.

[31] Jigsaw Data not a company that follows standards. San Francisco Chronicle 5.9.2006.

[32] Kantarcioglu, M. & Clifton, C. Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, Sept., 2004, pp. 1026–1037.

[33] Kauppalehti-online. Nokia introduces Wibree technology as open industry initiative. 3.10.2006. http://www.kauppalehti.fi.

[34] Kissner, L. & Song, D. Privacy-Preserving Set of Operations. Lecture Notes in Computer Science, 3621, 2005, pp. 241–257.

[35] Lahlou, S. Living in a goldfish bowl: lessons learned about privacy issues in a privacy-challenged environment. Proc. of UbiComp 2005.

[36] Liikenne ja viestintäministeriö. Uusi arjen tietoyhteiskunta, taustaselvitys. http://www.mintc.fi/oliver/upl236-Taustaselvitys.pdf. (In Finnish.)

[37] Lin, X., Clifton, C. & Zhu, M. Privacy-preserving clustering with distributed EM mixture modelling. Knowledge and Information Systems 8, 2005, pp. 68–81.

[38] Linden, G., Smith, B. & York, J. Amazon.com Recommendations. IEEE Internet computing, Vol. 7, No.1, Jan 2003, pp. 76–80.

[39] Mathieson, R. Branding Unbound, American Management Association, ISBN 0814472877.

[40] Merugu, S. & Ghosh, J. Privacy-preserving Distributed Clustering using Generative Models. The 3rd IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL, 2003.

[41] Oliveira, S. & Zaiane, O. Toward Standardization in Privacy-Preserving Data Mining. ACM SIGKDD 3rd Workshop on Data Mining Standards, 2004.

[42] Oliveira, S. & Zaane, O. Privacy-preserving Clustering by Data Transformation. In: Proc. of the 18th Brazilian Symposium on Databases, Manaus, Brazil, 2003. Pp. 304–318.

[43] Polat, H. & Du, W. Privacy-Preserving Collaborative Filtering Using Randomized Perturbation Techniques. Proc. of the 3rd IEEE Int. Conf. on Data Mining (ICDM'03) 2003.

[44] Quelle AG Company data, acquired via Data Mining Cup 2004.
http://www.data-mining-cup.com.

[45] RFID Takes Attendance – and Heat
http://www.rfidjournal.com/article/articleprint/1408/.

[46] RFID in Japan. see the website:
http://www.ubiks.net/local/blog/jmt/archives3/003524.html.

[47] RFIDBuzz.com website:
http://www.rfidbuzz.com/wiki/Actors/MetroGroup.

[48] Schafer, J., Konstan, J. & Reidl, J. E-commerce Recommendation Applications. Data Mining and Knowledge Discovery, Kluwer Academic, 2001, pp. 115–153.

[49] Costa da Silva, J. & Klusch, M. Inference in distributed data clustering. Engineering Applications of Artificial Intelligence 19, 2006, pp. 363–369.

[50] Sweeney, L. *k*-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, 2002, pp. 557–570.

[51] Srikumar, K. & Bhasker, B. Personalized recommendations in e-commerce. Int. J. electronic Business, Vol. 3, 2005, pp. 4–26.

[52] Talouselämä 25.10.2006. Tiedä, älä oleta! (In Finnish.)

[53] Tasoulis, D., Laskari, E., Meletiou, G. & Vrahatis, M. Privacy-preserving unsupervised clustering over vertically partitioned data. ICCSA 2006, PT5 Lecture Notes in Computer Science 3984, 2006, pp. 635–643.

[54] Teltrow, M. & Kobsa, A. Impacts of user privacy preferences on personalized systems: a comparative study. Designing personalized user experiences in eCommerce 2004. Pp. 315–332. ISBN 140202147X.

[55] Thackara, J. In the bubble: designing in a complex world, MIT press 2005.

[56] The End of Privacy. Forbes 29.11.1999.

[57] Thuraisingham, B. Privacy-Preserving Data Mining: developments and directions. Journal of Database Management, 16, 2005, pp. 75–87.

[58] Vaidya, J., Clifton, C. & Zhu, Y. Privacy-preserving Data Mining, Springer Science+Business Media, Inc. 2006. ISBN 0387258868.

[59] Vaidya, J. & Clifton, C. Privacy-preserving *k*-means clustering over vertically partitioned data. Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.

[60] Verykios, V., Bertino, E., Fovino, I., Provenza, L., Saygin, Y. & Theodoridis, Y. State-of-the-art in Privacy-preserving Data Mining. SIGMOD Record, 2004.

[61] Van Wel, L. & Royakkers, L. Ethical issues in web data mining. Ethics and Information Technology 6, 2004, pp. 129–140.

[62] Warren, S. & Brandeis, L. The Right to Privacy. Harvard Law Rev. Vol. 4, 1890, pp. 193–200.

[63] Weiser, M. The Computers for the 21st Century. Scientific American, Vol. 265, No. 3, 1991, pp. 94–104.

[64] Wikipedia, online dictionary. http://en.wikipedia.org/.

[65] Sun on Privacy: Get over it! Wired article 26.1.1999.
http://www.wired.com/news/politics/0,1283,17538,00.html.

[66] Yao, A. Protocols for secure computations. Proc. of the 23rd Annual IEEE Symposium on foundations of Computer Science, 1982.

Appendix A

The two other clustering models (k = 5, k = 10) are compared in the Figure A1, Figure A2, and Figure A3.

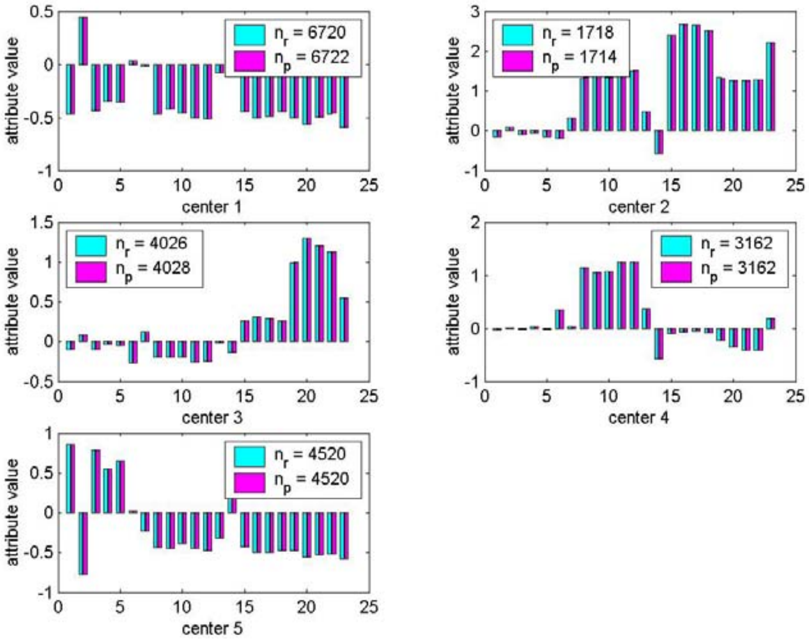


Figure A1. Comparison between the centre (1–5) values from each of the attributes in the k = 5 model.

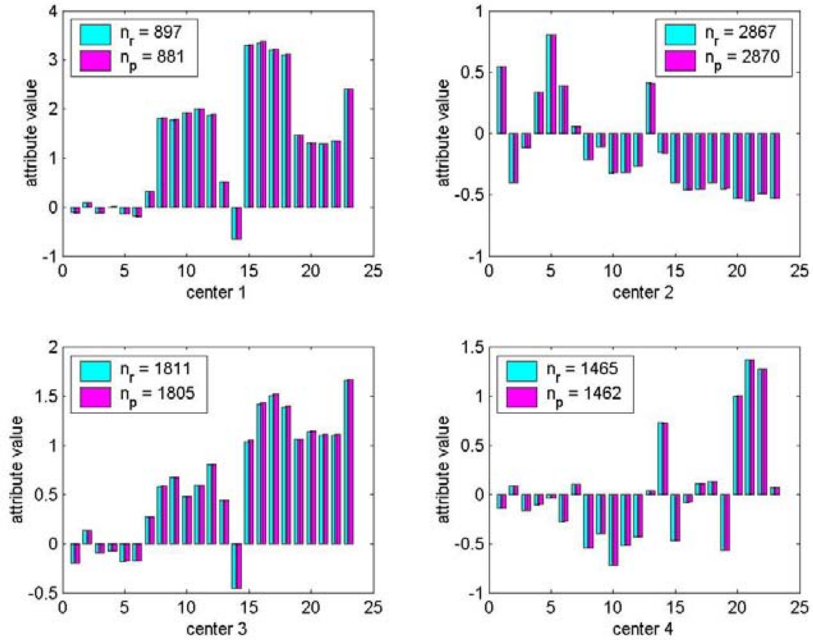


Figure A2. Comparison between the centre (1–4) values from each of the attributes in the k = 10 model.

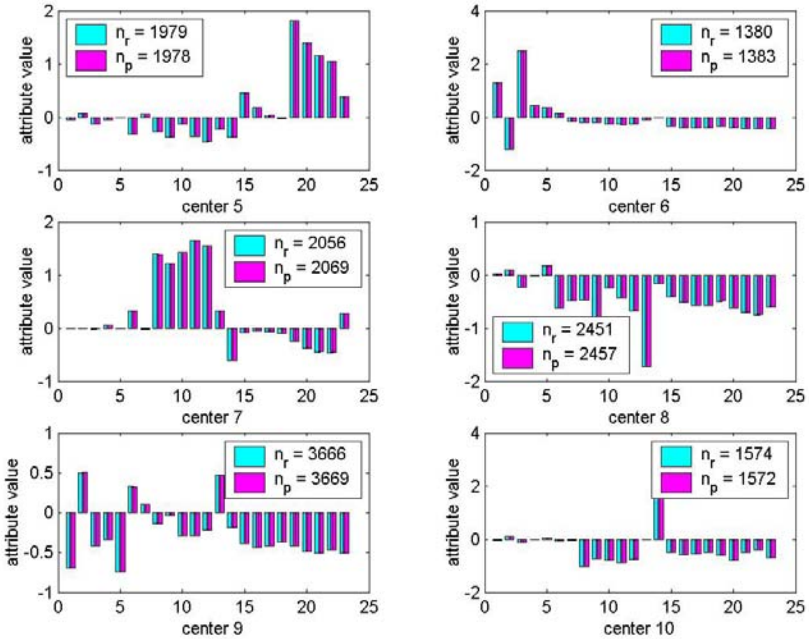


Figure A3. Comparison between the centre (5–10) values from each of the attributes in the k = 10 model.

Series title, number and
report code of publication

VTT Publications 647
VTT-PUBS-647

Author(s)
Mutanen, Teemu

Title

Consumer Data and Privacy in Ubiquitous Computing

Abstract

The privacy-preserving perspective on data mining is relatively young area. The research done in the area is mainly theoretical. The current trend with growing amount of personalization in online services has created also applications for personalized marketing. Personalized marketing services use detailed information about the context and personal history of a customer. This needs sophisticated individual identification methods, which raise privacy concern. The novelty in privacy-preserving methods is that sensitive and distributed data could be used for data mining task and the privacy of individuals is preserved.

This work has two objectives: first is to use consumer data from distributed sources and study how customer segmentation is possible while preserving the privacy. The idea is to conduct the customer segmentation in a way that the data need not leave the agent holding the data. The other objective is the value of the knowledge acquired from collectively conducted segmentation. We believe that collectively conducted segmentation produces knowledge that cannot be acquired otherwise. The results of this work show that privacypreserving customer segmentation is possible and the collectively conducted segmentation produces new knowledge.

ISBN
978-951-38-7031-7 (URL: http://www.vtt.fi/publications/index.jsp)

Series title and ISSN	Project number
VTT Publications 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)	11881

Date	Language	Pages
June 2007	English, Finnish abstr.	82 p. + app. 3 p.

Name of project	Commissioned by
UBICS	

Keywords	Publisher
consumer data, privacy-preserving data mining, ubiquitous computing	VTT Technical Research Centre of Finland P.O. Box 1000, FI-02044 VTT, Finland Phone internat. +358 20 722 4404 Fax +358 20 722 4374

Julkaisun sarja, numero ja raporttikoodi

VTT Publications 647
VTT-PUBS-647

Tekijä(t)
Mutanen, Teemu

Nimeke

Asiakastieto ja yksityisyys jokapaikan tietotekniikassa

Tiivistelmä

Jokapaikan tietotekniikan ilmaantuminen tarkoittaa uusia päätelaitteita, ilmaisimia ja yhteyksiä, siten myös uusia asiakastietolähteitä. Uudet tietolähteet ja tavat yksilön tunnistamiseen nostavat esille huolen yksityisyydestä: mitä voidaan ja mitä ei pidä tehdä yksilön tiedoilla. Yksilön yksityisyyden lisäksi yritystietojen yksityisyys nousee esiin, kun tiedon louhinnan menetelmiä sovelletaan asiakastietoihin. Jokapaikan tietotekniikka tuo mukanaan lukuisia tietolähteitä, jotka sijaitsevat hajautetusti.

Yksityisyyden säilyttävä tiedon louhinta on suhteellisen nuori ala. Alalla tehty tutkimus on pääosin teoreettista, käsittääkseni yhtään tosielämän sovellusta ei ole olemassa. Tämä vajetta on yritetty paikata. Vallitseva suuntaus yksilöllistämisen kasvavasta määrästä verkkopalveluissa on luonut myös markkinoita yksilölliselle markkinoinnille. Yksilölliset markkinointipalvelut hyödyntävät yksityiskohtaisia tietoja asiayhteydestä ja yksilön käytöshistoriasta. Tämä vaatii kehittyneitä menetelmiä yksilön tunnistamiseen, ja menetelmät nostavat huolen yksityisyydestä. Yksityisyyden säilyttävien menetelmien uutuusarvo on mahdollisuus käyttää arkaluontoista ja hajautettua tietoa tiedon louhintaan, kuitenkin säilyttäen yksityisyys.

Tässä työssä on kaksi tavoitetta: ensiksi käytetään asiakastietoa hajautetuista lähteistä ja tarkastellaan yksityisyyden säilyttävän asiakassegmentoinnin mahdollisuutta. Ideana on segmentoida asiakaskanta siten, ettei arka tieto missään vaiheessa lähde sitä hallitsevalta pelurilta. Toinen tavoite liittyy tietämyksen arvoon, joka saadaan yhteisesti suoritetusta segmentoinnista. Uskomme, että yhteisesti suoritettu segmentointi tuottaa tietoa, jota ei muuten pystytä tuottamaan. Työn tulokset osoittavat yksityisyyden säilyttävän asiakaskannan segmentoinnin olevan mahdollista ja yhteisesti suoritetun segmentoinnin tuottavan uutta tietoa.

ISBN
978-951-38-7031-7 (URL: http://www.vtt.fi/publications/index.jsp)

Avainnimeke ja ISSN VTT Publications 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)	Projektinumero 11881

Julkaisuaika Kesäkuu 2007	Kieli englanti, suom.kiel. tiiv.	Sivuja 82 s. + liitt. 3 s.

Projektin nimi UBICS	Toimeksiantaja(t)
Avainsanat consumer data, privacy-preserving data mining, ubiquitous computing	Julkaisija VTT PL 1000, 02044 VTT Puh. 020 722 4404 Faksi 020 722 4374

VTT PUBLICATIONS

629 Communications Technologies. VTT's Research Programme 2002–2006. Final Report. Ed. by Markku Sipilä. 2007. 354 p.

630 Solehmainen, Kimmo. Fabrication of microphotonic waveguide components on silicon. 2007. 68 p. + app. 35 p.

631 Törrö, Maaretta. Global intellectual capital brokering. Facilitating the emergence of innovations through network mediation. 106 p. + app. 2 p.

632 Lanne, Marinka. Yhteistyö yritysturvallisuuden hallinnassa. Tutkimus sisäisen yhteistyön tarpeesta ja roolista suurten organisaatioiden turvallisuustoiminnassa. 2007. 118 s. + liitt. 81 s.

633 Oedewald, Pia & Reiman, Teemu. Special characteristics of safety critical organizations. Work psychological perspective. 2007. 114 p. + app. 9 p.

634 Tammi, Kari. Active control of radial rotor vibrations. Identification, feedback, feedforward, and repetitive control methods. 2007. 151 p. + app. 5 p.

635 Intelligent Products and Systems. Technology theme – Final report. Ventä, Olli (ed.). 2007. 304 p.

636 Evesti, Antti. Quality-oriented software architecture development. 2007. 79 p.

637 Paananen, Arja. On the interactions and interfacial behaviour of biopolymers. An AFM study. 2007. 107 p. + app. 66 p.

638 Alakomi, Hanna-Leena. Weakening of the Gram-negative bacterial outer membrane. A tool for increasing microbiological safety. 2007. 95 p. + app. 37 p.

639 Kotiluoto, Petri. Adaptive tree multigrids and simplified spherical harmonics approximation in deterministic neutral and charged particle transport. 2007. 106 p. + app. 46 p.

640 Leppänen, Jaakko. Development of a New Monte Carlo Reactor Physics Code. 2007. 228 p. + app. 8 p.

641 Toivari, Mervi. Engineering the pentose phosphate pathway of *Saccharomyces cerevisiae* for production of ethanol and xylitol. 2007. 74 p. + app. 69 p.

642 Lantto, Raija. Protein cross-linking with oxidative enzymes and transglutaminase. Effects in meat protein systems. 2007. 114 p. + app. 49 p.

643 Trends and Indicators for Monitoring the EU Thematic Strategy on Sustainable Development of Urban Environment. Final report summary and recommendations. Häkkinen, Tarja (ed.). 2007. 240 p. + app. 50 p.

644 Saijets, Jan. MOSFET RF Characterization Using Bulk and SOI CMOS Technologies. 2007. 171 p. + app. 4 p.

645 Laitila, Arja. Microbes in the tailoring of barley malt properties. 2007. 107 p. + app. 79 p.

646 Mäkinen, Iiro. To patent or not to patent? An innovation-level investigation of the propensity to patent. 2007. 95 p. + app. 13 p.

647 Mutanen, Teemu. Consumer Data and Privacy in Ubiquitous Computing. 2007. 82 p. + app. 3 p.

VTT
PL 1000
02044 VTT
Puh. 020 722 4404
Faksi 020 722 4374

VTT
PB 1000
02044 VTT
Tel. 020 722 4404
Fax 020 722 4374

VTT
P.O. Box 1000
FI-02044 VTT, Finland
Phone internat. + 358 20 722 4404
Fax + 358 20 722 4374

ISBN 978-951-38-7031-7 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)

VTT PUBLICATIONS 647 Consumer Data and Privacy in Ubiquitous Computing Teemu Mutanen