PRIVACY-PRESERVING CROWDSOURCING-BASED RECOMMENDER SYSTEMS FOR E-COMMERCE & HEALTH SERVICES

FRANCISCO JOSE CASINO CEMBELLIN

UNIVERSITAT ROVIRA i VIRGILI

Privacy-Preserving Crowdsourcing-Based Recommender Systems for E-commerce & Health Services

---

FRAN CASINO

DOCTORAL THESIS
2017

UNIVERSITAT ROVIRA I VIRGILI
PRIVACY-PRESERVING CROWDSOURCING-BASED RECOMMENDER SYSTEMS FOR E-COMMERCE & HEALTH SERVICES
FRANCISCO JOSE CASINO CEMBELLIN

Universitat Rovira i Virgili

Department of

Computer Engineering and Mathematics

Doctoral Thesis

Privacy-Preserving Crowdsourcing-Based Recommender Systems for E-Commerce & Health Services

Author:

Fran Casino

Thesis Advisor:

Dr. Agusti Solanas

Doctoral thesis submitted to the Department of Computer Engineering and Mathematics in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computer Science

UNIVERSITAT ROVIRA i VIRGILI

Tarragona
2017

Departament d'Enginyeria
[DEIM]
Informàtica i
Matemàtiques

Avinguda dels Països Catalans, 26
Campus Sescelades
43007, Tarragona

I STATE that the present doctoral thesis, presented by Fran Casino for the degree of Doctor of Philosophy in Computer Science, has been carried out under my supervision and that it fulfils all the requirements to be eligible for the European Doctorate Award.

Tarragona, 27th of June 2017

Dr. Agusti Solanas
(Advisor)

UNIVERSITAT ROVIRA I VIRGILI
PRIVACY-PRESERVING CROWDSOURCING-BASED RECOMMENDER SYSTEMS FOR E-COMMERCE & HEALTH SERVICES
FRANCISCO JOSE CASINO CEMBELLIN

Resum

En l'actualitat els sistemes de recomanació han esdevingut un mecanisme fonamental per proporcionar als usuaris informació útil i filtrada, amb l'objectiu d'optimitzar la presa de decisions, com per exemple, en el camp del comerç electrònic. En aquest context, Internet ofereix una gran quantitat d'informació sobre una extensa varietat de productes i serveis que poden ser d'utilitat per a compradors potencials. No obstant, aquesta gran quantitat d'informació pot esdevenir un problema en comptes d'una solució, ja que pot dificultar la presa de decisions. El Filtratge Col·laboratiu (FC) és un sistema de recomanació que inclou una gran família de mètodes. L'objectiu del FC és fer suggeriments sobre un conjunt d'ítems $\{I\}$ (*e.g.* llibres, música, pel·lícules o rutes), basats en les experiències d'un conjunt d'usuaris $\{U\}$ que ja han adquirit i/o valorat alguns d'aquests elements. Per tant, es requereix informació sobre les preferències dels usuaris per tal de proporcionar recomanacions de qualitat. En aquest sentit, la poca densitat de les dades, deguda, en primer lloc, a la gran quantitat d'usuaris i productes que comparteixen escenari en el context del comerç electrònic i, en segon lloc, a la manca de resposta dels usuaris, és un dels problemes més rellevants d'aquest tipus de sistemes.

L'ús generalitzat dels sistemes de FC a Internet ofereix grans oportunitats i beneficis per a les empreses i els usuaris, però hi ha un gran inconvenient: la manca de privadesa. La importància de la privadesa en els sistemes de FC s'incrementa al ritme amb què es recull i emmagatzema la informació de cada usuari. La gestió descuidada de la informació personal, a més de ser il·legal, podria donar lloc a greus conseqüències tant per als usuaris dels quals s'emmagatzema la informació, com per a les empreses.

Aquesta tesi contribueix al disseny d'algorismes i sistemes destinats a fer front als problemes explicats anteriorment. En primer lloc, introduïm els conceptes de FC, control de revelació estadística i FC amb privadesa. En segon lloc, proposem un conjunt de mètodes d'imputació estadística per tal de fer front a la manca d'informació en els sistemes de FC. A continuació, proposem dos mètodes de FC amb privadesa amb els quals podem generar recomanacions de qualitat alhora que protegim la privadesa dels usuaris participants.

A més de les aportacions indicades, també ens centrem en el procés d'urbanització que s'està produint a tot el món. Avui dia, les ciutats intel·ligents estan guanyant rellevància i la seva infraestructura es pot utilitzar per millorar els serveis de salut prestats als ciutadans. Aquesta és la filosofia del concepte de Salut Intel·ligent (de l'anglès Smart Health). En

aquest context, els sistemes de comunicació sense fils juguen un paper clau, com a facilitadors de les connexions ubiqües i en temps real, l'augment de la funcionalitat dels sistemes participants i la disminució dels costos operatius. En aquest escenari, diversos sistemes sense fils coexisteixen, fet que requereix d'una anàlisi de radiofreqüències en termes de relacions de cobertura/capacitat, amb especial consideració dels efectes de les interferències. No obstant, l'anàlisi de radiofreqüències, en termes de senyal útil rebut per un conjunt de connexions, així com per a les connexions que hi puguin interferir, pot ser una tasca difícil a causa de: la mida dels escenaris; l'existència de múltiples materials dependents de la freqüència; i la variabilitat inherent de les connexions mòbils.

Hem estudiat com les tècniques de filtratge d'informació, com ara el FC, es poden utilitzar per tal de potenciar les capacitats de les ciutats intel·ligents i la salut intel·ligent. Per tal de recolzar aquesta teoria, en aquesta tesi es proposa la idea d'utilitzar els sistemes de recomanació integrats amb la infraestructura d'informació de les ciutats intel·ligents, per tal d'oferir als ciutadans recomanacions de rutes que tinguin en compte les seves condicions de salut i preferències. D'altra banda, presentem un mètode que combina un algorisme d'anàlisi de radiofreqüències basat en 3D Ray Launching i una tècnica de FC per tal d'analitzar el rendiment dels canals sense fils, emulant escenaris sensibles al context. L'objectiu del mètode proposat és proporcionar estratègies d'implementació òptimes per a sistemes sense fils massius i xarxes de sensors sense fils. En aquest sentit, els resultats obtinguts mostren una destacable millora tant en la precisió de les mesures de radiofreqüència com en el cost computacional, en comparació amb altres métodes de l'estat de l'art.

Resumen

En la actualidad los sistemas de recomendación se han convertido en una herramienta indispensable para proporcionar a los usuarios información útil y filtrada, con el objetivo de optimizar la toma de decisiones en una gran variedad de contextos. La cantidad de datos existente en Internet es tan extensa que los usuarios necesitan sistemas automáticos para ayudarles a distinguir entre información valiosa y ruido. El Filtrado Colaborativo (FC) es un sistema de recomendación que comprende una gran familia de métodos. El objetivo del FC es realizar sugerencias sobre un conjunto de ítems $\{I\}$ (*e.g.* libros, música, películas o rutas), basadas en las experiencias de un conjunto de usuarios $\{U\}$ que ya han adquirido y/o valorado algunos de estos elementos. Por lo tanto, se requiere información sobre las preferencias de los usuarios para proporcionar recomendaciones de calidad. En este sentido, la poca densidad de los datos, debida, en primer lugar, a la gran cantidad de usuarios y productos que comparten escenario en el contexto del comercio electrónico y, en segundo lugar, a la falta de respuesta de los usuarios, es uno de los problemas más relevantes de este tipo de sistemas. El uso generalizado de los sistemas de FC en Internet ofrece grandes oportunidades y beneficios para las empresas y los usuarios, pero hay un gran inconveniente: la falta de privacidad. La importancia de la privacidad en los sistemas de FC crece al ritmo con el que se recoge y almacena la información de cada usuario. La gestión descuidada de la información personal, además de ser ilegal, podría dar lugar a graves consecuencias tanto para los usuarios, cuya información se almacena, como para las empresas.

Esta tesis contribuye al diseño de algoritmos y sistemas destinados a hacer frente a los problemas explicados anteriormente. En primer lugar, introducimos los conceptos de FC, control de revelación estadística y FC con privacidad. En segundo lugar, proponemos un conjunto de métodos de imputación estadística para hacer frente a la falta de información en los sistemas de FC. Posteriormente, proponemos dos métodos de FC con privacidad con los que podemos generar recomendaciones de calidad a la vez que protegemos la privacidad de los usuarios participantes.

Además de las aportaciones indicadas, también nos centramos en el proceso de urbanización que se está produciendo en todo el mundo. Hoy en día, las ciudades inteligentes están ganando relevancia y su infraestructura se puede utilizar para mejorar los servicios de salud prestados a los ciudadanos. Esta es la filosofía del concepto de Salud Inteligente (del inglés Smart Health). En este contexto, los sistemas de comunicación inalámbricos juegan un papel clave, como facilitadores de las conexiones ubicuas y en

tiempo real, el aumento de la funcionalidad de los sistemas participantes y la disminución de los costes operativos. En este escenario, varios sistemas inalámbricos coexisten, lo que requiere de un análisis de radiofrecuencias en términos de relaciones de cobertura/capacidad, con especial consideración de los efectos de las interferencias. Sin embargo, el análisis de radiofrecuencias, en términos de señal útil recibida por un conjunto de conexiones, así como para las conexiones que puedan interferir, puede ser una tarea difícil debido a: el tamaño de los escenarios; la existencia de múltiples materiales dependientes de la frecuencia; y la variabilidad inherente de las conexiones móviles.

Hemos estudiado como las técnicas de filtrado de información, como el FC, se pueden utilizar para potenciar las capacidades de las ciudades inteligentes y la salud inteligente. Para apoyar esta teoría, en la presente tesis proponemos utilizar los sistemas de recomendación integrados con la infraestructura de información de las ciudades inteligentes, para ofrecer a los ciudadanos recomendaciones de rutas que tengan en cuenta sus condiciones de salud y sus preferencias. Por otra parte, presentamos un método que combina un algoritmo de análisis de radiofrecuencias basado en 3D Ray Launching y una técnica de FC para analizar el rendimiento de los canales inalámbricos, emulando escenarios sensibles al contexto. El objetivo del método propuesto es proporcionar estrategias de implementación óptimas para sistemas inalámbricos masivos y redes de sensores inalámbricos. En este sentido, los resultados obtenidos muestran una destacable mejoría tanto en la precisión de las medidas de radiofrecuencia como en el coste computacional, en comparación con otros métodos del estado del arte.

Abstract

Recommender systems have become a fundamental mechanism to provide users with useful selected information, which could be effective to optimise a large amount of decisions, for instance, in the e-commerce field. In this context, the Internet provides a wealth of information on a huge variety of products and services that may be useful to potential buyers. However, this wealth of information may become a problem rather than a solution because it can hinder the decision making. Collaborative Filtering (CF) is a recommender system that comprises a large family of methods. The aim of CF is to make suggestions on a set of items $\{I\}$ (*e.g.* books, music, films or routes), based on the preferences of a set of users $\{U\}$ that have already acquired and/or rated some of those items. Therefore, information about the users' behaviours and preferences is required in order to provide profitable recommendations. In this sense, one of the most relevant problems faced by businesses is non-response. The widespread use of CF on the Internet provides great opportunities and benefits to both companies and users, but there is a major drawback: the lack of users' privacy. The importance of privacy in CF systems is emphasised by the growing pace at which information of each user is collected and stored. Careless management of personal information, besides being illegal, could lead to serious consequences for both users, whose information is stored, and companies.

This dissertation contributes to the design of algorithms and systems that address the issues explained above. First, we provide an extensive background in CF, Statistical Disclosure Control and privacy-preserving CF methods. Second, we propose a set of imputation methods in order to deal with the lack of information in CF based recommender systems. Next, we present two privacy-preserving CF methods, which protect the privacy of the users and achieve a remarkable recommendations accuracy.

In addition to the aforementioned contributions, we also focus on the urbanisation process that is taking place worldwide. Nowadays, smart cities are gaining importance and their infrastructure can be used to improve the healthcare services provided to citizens. This is the philosophy of the so-called Smart Health concept. In this context, wireless communication systems play a key role, as enablers of real time and location independent connectivity, increasing system functionality and decreasing operational costs. In this scenario, multiple wireless systems co-exist, which requires an in-depth radioplanning in terms of coverage/capacity ratios, with particular consideration of the impact of interferences. However, radiofrequency analysis, in terms of useful received signal levels for a given set of connections as

well as for potential interfering connections, can be a challenging task due to: the large size of scenarios; the existence of multiple frequency dependent materials; and inherent variability of mobile connections, given by the movement of potential scatterers.

We have studied that information filtering techniques such as CF may augment the capabilities of smart cities and smart health. In order to support our vision, we propose the use of recommender systems integrated with the sensing infrastructure of smart cities to provide citizens with routes recommendations that take into account their health conditions and preferences. Moreover, a novel approximation based on the combination of an in-house developed 3D Ray Launching code and a CF technique is used to analyse the performance of wireless channels emulating context-aware scenarios. The aim of the proposed method is to provide optimal deployment strategies for massive wireless system and wireless sensor networks. In this sense, the results show a remarkable improvement in both accuracy of radiofrequency measurements and computational cost, compared with other methods of the state-of-the-art.

Contents

UNIVERSITAT ROVIRA I VIRGILI
PRIVACY-PRESERVING CROWDSOURCING-BASED RECOMMENDER SYSTEMS FOR E-COMMERCE & HEALTH SERVICES
FRANCISCO JOSE CASINO CEMBELLIN

Chapter 1

Introduction

This chapter introduces the issues faced in this doctoral thesis. In addition, it briefly describes the solutions we propose to tackle those issues. Finally, the structure and organisation of the thesis are outlined.

Contents

1.1 Motivation

Our society lives an age where the eagerness for information has resulted in problems such as *infobesity*, especially after the arrival of *Web 2.0*. In this context, automatic systems such as recommenders are increasing their relevance, since they help to distinguish noise from useful information. Nowadays, recommender systems (RS) [149] play an active role in the Internet through the advances in data mining and artificial intelligence. Collaborative Filtering (CF) [67] appears with the aim to provide automatic recommendations in a digital environment. CF is a crowdsourcing-based recommender system, which involves a large family of recommendation methods. The aim of CF is to suggest/recommend items (*e.g.* books, films or routes) based on the preferences of users that have already acquired and/or rated those items. In order to make recommendations, CF methods rely on large databases with information on the relationships between sets of users and items. These data are organised in matrices composed by n users and m items, where each cell (i, j) stores the rating/opinion of user i about item j. In Section 2.1, we define Collaborative Filtering and recall the most relevant methods in the state-of-the-art. Moreover, in Chapter 2 we discuss the main drawbacks and shortcomings of CF and propose some ideas to overcome or mitigate them.

In the RS field, one of the most relevant issues is the lack of information (*i.e.* non-response, which is closely related with sparseness of RS datasets).

There are well-known methods in the literature [146][49] to deal with missing data in medical and statistical surveys. However, non-response in surveys is still an open problem, since there is not an optimal solution that satisfies all constraints [145][144]. Although there are many missingness scenarios in surveys with different kinds of data (*i.e.* categorical and numerical), in this dissertation we focus on the most likely situations in the CF context (*i.e.* numerical datasets). In Section 2.1.1, we introduce the statistical inference concept and discuss the issues related with the lack of information in CF. Later, in Chapter 3, we propose classical and new imputation methods to deal with missing data in CF datasets, which have specific characteristics such as high dimensionality and high percentage of unknown/null values.

The quality of a recommender system can be evaluated. According to Herlocker et al. [73], metrics evaluating recommendation systems can be classified into the following broad categories: predictive accuracy metrics, classification accuracy metrics and rank accuracy metrics. Moreover, data from live experiments are frequently used as benchmarks to evaluate the efficiency, quality and robustness of CF methods [73]. In Section 2.4, we introduce the most prevalent metrics and datasets in the CF field that are also studied in this thesis.

Despite the extensive range and amount of data available on the Internet, people are reluctant to disclose their personal information and interests. This makes privacy one of the most relevant problems of CF. One of the main limitations and risks of CF is the lack of users' privacy. Privacy is a fundamental right and privacy protection is a hot topic to which many research efforts have been devoted from a variety of fields [173, 196, 108]. Thus, among all the open problems of CF [153][40], in this dissertation we concentrate on the protection of the privacy of users involved in CF processes. With the aim to address such privacy issues, current research focuses on Privacy-Preserving Collaborative Filtering (PPCF) methods. In Section 2.3, we review the state-of-the-art of PPCF and propose a classification of PPCF methods according to how information is stored and how recommendations are computed.

Statistical disclosure control [78], a.k.a. statistical disclosure limitation, seeks to transform microdata sets (*i.e.* datasets consisting of records corresponding to individual respondents) prior to publication in such a way that avoids re-identification. Therefore, it should not be possible to: (i) re-identify the respondents corresponding to any particular record in the anonymised published microdata set —identity disclosure—; and (ii) to discover the value of a confidential attribute (*e.g.* salary) for a *specific* respondent —attribute disclosure—.

In order to publish useful data while preserving privacy, the original data is assumed to be a private table consisting of multiple records. Direct identifiers (name, passport no., etc.) need to be suppressed from the dataset prior to the anonymisation process. However, some of the attributes that remain in the anonymised dataset may be *quasi-identifiers*, that is, attributes that may facilitate indirect re-identification of respondents through external data sources.

In Section 2.2, we define Statistical Disclosure Control and we recall the most well-known techniques used to release useful information while preserving individuals' sensitive data. Moreover, in Chapter 4, we propose two PPCF microaggregation-based approaches.

In addition to the aforementioned research topics, in this dissertation we also focus in the worldwide process of urbanisation that is taking place and the need for more sustainable and liveable cities. Nowadays, the sustainable improvement of citizen's quality of life is one of the main challenges to be faced by governments, policy makers and researches throughout the world. For that purpose, one key issue is the optimisation of resources in different aspects such as healthcare systems, energy consumption, transportation systems, water resource and waste management, among many others. In order to achieve optimised resource usage and the corresponding cost reduction, context-aware scenarios enable data collection of multiple systems and allow the optimised management of those systems. In this context, citizens collaborate among them by means of their smartphones to share information with the city and improve their quality of life, clearly benefiting healthcare systems. Hence, the convergence of these factors opens the door to the consolidation of smart health (s-health) [157], and to the creation of systems that benefit from this concept and augment it with other powerful information filtering systems such as recommender systems. One of the most challenging aspects within this framework is to achieve sustainable healthcare services provisioning. Traditionally, healthcare services entailed large amounts of resources, many of which required patients to be in direct contact with health specialists for diagnostics and treatment. These s-health related issues motivate our last research line in this thesis. In Section 2.5.1.2, we recall the s-health concept. Later, in Chapter 5, we propose two s-health applications in which CF systems play a key role. More concretely, in Section 5.1 we propose a smart route recommender system, and in Section 5.2 we assess the importance of wireless channel characterisation in complex environments by means of an optimised hybrid wireless channel characterisation method.

1.2 Contributions

The main contributions of this dissertation are the following:

1. **Contributions to Collaborative Filtering:** As previously stated, CF systems have inherent problems such as non-response, which translates into sparseness. This problem is exacerbated if we consider the characteristics of CF datasets. In order to deal with the lack of information, we propose a collection of classical and new imputation methods in Chapter 3. Moreover, we propose a new metric, called behavioural precision, to perform a more robust analysis of the recommendation's quality.

2. **Contributions to Privacy-preserving Collaborative Filtering:** Privacy is one of the most relevant problems when collecting data about individuals. An important part of this dissertation is devoted to privacy. In Chapter 2, we recall the most prevalent SDC techniques and we perform a classification of the PPCF state-of-the-art methods. Moreover, in Chapter 5, we propose two new PPCF microaggregation-based techniques. These approaches achieve a remarkable trade-off between privacy and recommendation's accuracy.

3. **Recommender Systems to Enable Smart Health:** The urbanisation process and the need for sustainable and more liveable cities are among the challenges to be faced by the human race in the next decades. In this context, healthcare service provisioning and the quality of life of the citizens are two of the most relevant aspects to be considered. In Section 2.5.1, we show some numbers about the urbanisation process. Moreover, in Section 2.5.1.2, we recall the s-health concept and the need for a sustainable healthcare. In Chapter 5, we propose two ways in which recommender systems are able to enrich/augment s-health. First, we propose a smart route recommender that takes into account the information provided by the smart city and physical condition and preferences of the citizens to recommend routes that better fit their capacities. Second, we propose an enhanced wireless channel characterisation method based on a Hybrid Ray Launching and a Collaborative Filtering approach. Moreover, we analyse the performance of this method in terms of accuracy and computational cost in diverse scenarios.

1.3 Organisation

This thesis is organised as follows:

- Chapter 2 provides an extensive background on recommender systems and CF. Moreover, it introduces the Statistical Disclosure Control concept and reviews the state-of-the-art of PPCF. Subsequently, the most prevalent evaluation tools in the previously stated fields are presented. Next, the urbanisation process, the healthcare evolution and the s-health concept are discussed. Finally, technical background on Wireless Channel Characterisation and some clues about our contributions in this field are briefly pointed out.

- Chapter 3 is devoted to CF-based recommender systems. The information about users' behaviours and tastes is needed to obtain quality recommendations. However, non-response is one of the classical problems in this research field. In this chapter, new and classical imputation methods to overcome non-response are proposed. Moreover, a new set of metrics is presented so as to provide a more accurate and comprehensive study of the outcomes.

- Chapter 4 presents three PPCF methods and studies their characteristics using well-known metrics. Extensive experiments show that methods such as Maximum Distance to Average Vector (MDAV) and, overall, Variable-MDAV, are able to protect the privacy of users while achieving quality recommendations.

- Chapter 5 is devoted to smart health applications and sustainable healthcare provisioning. First, a healthy route recommender system is presented. Such method uses the citizen's profile as well as the infrastructure of the smart city to provide them with route recommendations. Next, our contributions to Wireless Channel Characterisation are exposed in chronological order. These contributions include a new hybrid method that efficiently predicts wave propagation behaviour of wireless communication technologies in context-aware scenarios.

- Finally, Chapter 6 summarises our contributions and describes possible future research lines.

UNIVERSITAT ROVIRA I VIRGILI
PRIVACY-PRESERVING CROWDSOURCING-BASED RECOMMENDER SYSTEMS FOR E-COMMERCE & HEALTH SERVICES
FRANCISCO JOSE CASINO CEMBELLIN

Chapter 2

Background and State of the Art

This Chapter provides the reader with the essential context needed to understand the topics that will be discussed in this dissertation. Section 2.1 introduces the reader to recommender systems. More specifically, it focuses on Collaborative Filtering and its main limitations, namely sparseness and privacy. In Section 2.1.1, the Statistical Inference concept is presented, and in Sections 2.2 and 2.3, the reader is provided with background on Statistical Disclosure Control and a review of the state-of-the-art on Privacy-Preserving Collaborative Filtering methods, respectively. Section 2.4 introduces the most prevalent evaluation tools in the Collaborative Filtering and Statistical Disclosure Control contexts. Next, the healthcare evolution along with the new Smart Health paradigm, are discussed in Section 2.5.1. Finally, the Wireless Channel Characterisation problem is discussed in Section 5.2.

Contents

2.1 Recommender Systems and Collaborative Filtering

From the early days of the Internet to nowadays, the amount of available information has grown at a great pace, especially after the arrival of *Web 2.0*. The continuous increase of data has led to well-known problems such as information overload or *infobesity*. This problem has attracted many researchers, whose aim is to provide users with solutions that help them to distinguish noise from relevant information. Thus, the Internet provides a wealth of information on a huge variety of products and services that may be useful to potential buyers. However, this wealth of information may become a problem rather than a solution because it can hinder the process of decision making. Recommender systems (RS) [149][143] play an active role in the Internet through the advances in data mining and artificial intelligence. Furthermore, the new way to use and understand the network based on the *Web 3.0*, which fosters the user-computer interaction, raises RS to another level, integrating them into everyday life in a transparent and efficient manner. Recommendations based on the Internet are especially relevant for certain types of industries, such as e-commerce. Therefore, RS are a useful alternative to search engines since they help users to discover items they might not have found by themselves. Internet users are increasing their participation in multiple contexts such as the gastronomic field (*e.g.* with sites like Tripadvisor or Foursquare) where users give their opinion on restaurants, the e-commerce field (*e.g.* Amazon, eBay, AliExpress) where users evaluate items previously acquired, or in the audiovisual context, with sites like Netflix or IMDb. It is worth to note that nowadays, recommendations from personal acquaintances or opinions posted by consumers online are the most trusted forms of advertising [44][106].

In this research field, crowdsourcing has an important role, since users collaborate with others for their own benefit and for the creation of communities (*e.g.* trust communities, similar-users communities or the *Web 2.0*). There are well-known examples of competition-based crowdsourcing in the state-of-the-art (*e.g.* the Netflix prize[15]).

Collaborative Filtering (CF) [67] is a crowdsourcing-based recommender system, which involves a large family of recommendation methods. The main aim of CF is to suggest/recommend items (*e.g.* books, films or routes), based

Table 2.1: Example of data matrix. Rows correspond to users and columns to items. Cells contain the ratings from users.

U \ I	i_1	i_2	...	i_j	...	i_{m-1}	i_m
u_1	5	1	...	1	...	1	5
u_2	2	3	...	5	...	5	2
$\vdots$	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
u_i	3	5	...	3	...	4	2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$	$\vdots$
u_{n-1}	2	2	...	4	...	2	2
u_n	3	4	...	5	...	1	1

on the preferences of users that have already acquired and/or rated those items.

In order to make predictions, CF methods use large databases that store information on the relationships between sets of users and items. These data are modelled as matrices composed of n users and m items, and each cell (i, j) stores the evaluation/rating of user i on item j. Therefore, a rating is assigned, which can be within a range of values (*e.g.* between 0 and 10) or simply with binary votes (positive/negative, or bought/not bought) as in market basket databases. Table 2.1 shows an example of this kind of matrices, in which users rate items with values between 1 and 5.

The recommendations provided by CF methods are based on the assumption that similar users will be interested in the same items. As a result, items well rated by a user u_a could be recommended to another user u_b, if u_a and u_b are similar.

CF methods can be classified into three main categories according to the data they use to make the recommendation [160]: memory-based methods, which use the full matrix with all ratings; model-based methods, which use statistical models and functions of the data matrix but not the complete data matrix; and hybrid methods, which combine the two previous strategies with content-based recommendation methods. In memory-based CF, recommendations are made in two steps: (i) neighbourhood search and (ii) recommendation prediction.

- **Neighbourhood search:** Given a user $u_a \in U$, we use similarity functions such as Pearson Correlation [142], the Cosine similarity [23] or the Euclidean distance, to determine the users that are most similar to u_a (*i.e.*, the neighbourhood of u_a).

- **Recommendation/Prediction computation:** Once the neighbourhood of u_a is determined, we can make a recommendation/prediction using well-known methods such as the ones proposed in [148][142]. These methods can be used to predict a vote or to recommend the top-N items for u_a (*i.e.* the N items that are expected to better please user u_a).

Model-based CF methods create a model from the full matrix, on which to make recommendations. The emergence of these methods is justified by the constraints of memory-based CF in terms of scalability, complexity of calculation and sparseness. Some well-known methods to reduce the dimensionality of a matrix [174] are Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). However, the use of dimensionality reduction methods could affect the quality of the recommendations since they reduce the data range. There exists a huge variety of model-based CF methods: dimensionality reduction methods such as SVD, Regularised SVD (RSVD) and its variants, and SVD++ [90], latent semantic methods [75] linear regression methods [93], clustering methods [19], and Bayesian network models, among others [160].

Hybrid CF methods combine memory-based and model-based methods, in such a way to preserve the advantages of the algorithms involved and neutralise their shortcomings. Examples of these methods are the Personality Diagnosis [124] and the Probabilistic memory-based model [151].

Recently, the use of trust systems has experienced an important increase on the Internet. A trust statement is defined as the explicit opinion expressed by a user on another user, regarding the perceived quality of certain features of that user [110]. The concept of trust is prevalent, for instance, in search engines such as Google, which uses global trust metrics [111], and in e-commerce (*e.g.* eBay), in which users express their level of satisfaction after purchasing a product. The trust statements expressed by every user are aggregated to produce a community or a neighbourhood [66] as seen, for instance, in social networks. Since ratings scored by users are sparse, often finding similar users becomes impossible. In order to overcome that, trust-based heuristics are applied to find users that are regarded as the most trustworthy by the rest of users. The information provided by trust networks can be combined with CF matrices to create a trust-aware recommender system [110], which better deals with problems such as data sparsity, cold-start users and fake identities. Despite such advantages, trust systems have problems like *controversial users* [1], which can be minimised with local trust

[1] A user that is trusted or appreciated by many and is distrusted or negatively rated

metrics [111].

Regardless the CF method, there are several limitations inherent to this kind of recommender system. Some of the most important limitations [160, 153, 34, 68] are *sparseness*, *scalability*, *cold start*, *shilling*, *synonymy*, *bribing*, *copy-profile attacks*, and the lack of privacy. For more on CF, we point the interested reader to [153][18][40][20][160] for a review of the state-of-the-art and the most relevant advances and trends.

2.1.1 Statistical Inference

Non-response in surveys is a historically well-known problem. Since the 70s, several methods have been proposed to deal with missing responses in medical and statistical surveys [146][86].

A common method of handling non-response in surveys is to impute a value of each missing datum under some model for non-response. For instance, in [86] and [49] the authors describe a variety of classical imputation methods that are frequently used. Notwithstanding, in many medical and survey settings, missing data can cause difficulties in estimation, precision and inference [72]. In order to minimise the introduced bias when handling non-response in surveys, several well-known methods have been developed such as Multiple Imputation, proposed by Rubin [146], and the Expectation-Maximization algorithm [49]. However, the aforementioned methods could not be the best option to handle non-response, as described in [145][144] where authors discuss that each situation is different, for instance, depending on the analysed variable and the percentage of non-response.

In order to decide the best way to handle missing data, it is helpful to know why they are missing [63]. According to the state-of-the-art, we consider the following scenarios:

Missingness completely at random (MCAR). A variable is missing completely at random if the probability of missingness is the same for all units, (*i.e.* there is not an apparent reason for a missing vote in a randomly selected item).

Missingness at random (MAR). Most missingness is not completely at random and such situation can be defined as the probability that a variable is missing, depending only on available information. For instance, usually people interested in a special kind of items do not vote these due to another measured variable/item with which is related in some way (*e.g.* if we have gender information about users, we may observe that men may be less likely

by many

to disclose their interests in love movies than women, so the probability that this votes are missing may depend on the gender type). Cases in which data are missing depending on the values of the missing/unobserved data (*e.g.* rich people are less likely to disclose the specifics of their income) are classified as missingness not at random (MNAR) and will not be discussed here.

In surveys, data may be biased if the inference does not consider direct relationships between variables, depending on the kind of missingness. Therefore, a previous study of data should be performed to select a proper imputation method, which minimises the added noise. Medical and population surveys consider different kinds of data whose relationships have to be studied, since usually non-response is MAR or MNAR. However, in this dissertation we consider the most likely situations in CF, in which relationships between users and items are measured using continuous data and non-response is MCAR or MAR. If there is no clear evidence it is hard to prove MCAR or MAR in CF because in most cases items are not strongly correlated and users have varied behaviours and tastes. For instance, there is no guarantee that users, which like comedy movies may also like drama, or not. As a result, a proper way to deal with incomplete data in CF is to use imputation methods that bias in the least possible way the matrix of statistical indicators. In order to accomplish that, we describe several well-known imputation methods in Section 3.1.1 and we also propose new imputation approaches in Section 3.2.

2.1.2 The Lack of Information in Collaborative Filtering

Although we can find users with different levels of participation, they tend to vote only a small subset of the items present in CF datasets. This situation hinders the creation of reliable neighbourhoods and, thus, affects the recommendations' quality. Therefore, as previously stated in Section2.1, data sparsity or *sparseness* is one of the most challenging problems in CF. There are other CF issues related with the lack of information such as cold start, which is defined as the problem to offer recommendations to a profile from which the system has not enough information (*e.g.* not participatory users or recently added profiles). In order to deal with cold start, CF methods rely on other sources of information [158] such as content-based data [160], social context [62], cross-domain information [115] or the users' profiles [2]. However, how to use efficiently this extra information is still an active line of research [12].

Missing data imputation in CF is not a new research line. For instance, in

[23] the authors propose the straightforward idea of assigning a default value (*i.e.* default voting technique) to missing ratings to increase the number of referrals. In [185], the authors use k-means to cluster users and then use smoothing strategies to fill unseen rating data. Another clustering approach is presented in [35], where authors first fill a dataset with default voting (using its central value) and then apply a variant of the Maximum Distance to Average Vector approach [52] to provide users with recommendations. Moreover, this k-anonymous approach guarantees the privacy of users involved. Other approaches take into consideration the correlations between users and items. For instance, in [104], the authors propose a missing data prediction algorithm that uses information of users, items or both. They take advantage of user correlations and item correlations and compute a confidence threshold to avoid predicting missing data with bad quality. Another example is showed in [25], where the authors compute the voting tendencies of users and items to predict the users' behaviours. Other approaches also consider information from incomplete instances with missing values such as the method showed in [195], which uses iterative non-parametric estimators to impute missing values. This method was refined in [197] to impute missing data in datasets with heterogeneous attributes. In [193], the author presents an imputation approach that fills incomplete instances in a given dataset by using its shell neighbours only (*i.e.* the left and right nearest neighbours with respect to each attribute). Other research such as the one presented in [140] proposed a method that can identify which missing data should be imputed automatically according to a user's or an item's own rating history by determining a key set of missing data. This approach was recently improved in [141] with a maximum imputation framework, to maximise the imputation benefit for each predicted rating. Finally, well-known approaches such as matrix factorization [136] also have been shown to face data sparsity issues by means of imputation methods in an effective way.

2.2 Statistical Disclosure Control

Data mining techniques, including statistical computations and analyses assume that the variables have specific levels of measurement. The state-of-the-art classifies these variables as follows [55]:

Continuous. Those variables with which numerical and arithmetic operations can be performed. The main advantage of continuous data is that arithmetical operations are possible, and one of the main drawbacks is that every combination of numerical values in the original dataset is likely to be

unique, which could lead to disclosure.

Categorical. A variable is considered categorical when it takes values over a finite set and standard arithmetical operations are not applicable. Therefore, methods based on arithmetical manipulation cannot be used on categorical data. There exist two main types of categorical attributes:

- *Ordinal.* An ordinal variable/attribute takes values in an ordered range of categories. Rankings, satisfaction scales, the instruction level and the political preferences (left-right) are examples of ordinal attributes. Typical operators applicable to ordinal data are *median*, max and min.
- *Nominal.* A nominal attribute takes values in an unordered range of categories. The only possible operator is comparison for equality. The eye colour, the ID number and the address of an individual are examples of nominal attributes.

Statistical Disclosure Control (SDC, [78]), also known as statistical disclosure limitation, seeks to anonymise microdata sets (*i.e.* datasets consisting of multiple records corresponding to individual respondents). Such anonymisation is performed prior publication in such a way that it is not possible to re-identify the respondent corresponding to any particular record in the published microdata set —identity disclosure— nor is it possible to discover the value of a confidential attribute (*e.g.* salary) for a *specific* respondent —attribute disclosure—.

In order to publish useful data while preserving privacy, the original data is assumed to be a private table consisting of multiple records. Each record consists of the following types of attributes [184][55]:

Identifier: Attributes that can directly and uniquely identify an individual, such as passport number and mobile number.

Quasi-identifier: Publicly known characteristics of individuals that might be used by an adversary. A Quasi-Identifier $\mathcal{QI}$ cannot be used to uniquely identify a person by itself. However, when combined with other $\mathcal{QI}$s it can lead to re-identification of a person by narrowing down the possible identities of a specific record. As a result, this improves the confidence of an adversary regarding the possible real identity behind an anonymised record. Typical examples of $\mathcal{QI}$ attributes are gender, zip code, and age.

Sensitive Attribute: Attributes that an adversary ignores in most cases and wants to ascertain. Examples of SA can be considered the salary or the disease of a person in a financial or medical dataset, respectively.

Non-sensitive Attribute: Those attributes that contain non-sensitive information on the respondent, (*i.e.* attributes other than identifiers, $\mathcal{QI}$s and sensitive attributes. Note that attributes of this kind cannot be neglected when protecting a dataset, because they can be part of a $\mathcal{QI}$. For instance, Job and Town may be considered non-confidential outcome attributes, but in small towns there are only few people or just one individual that has a specific job such as bartender, baker or doctor. Therefore, the combination of these attributes could lead to disclosure with high probability.

In the case of census or medical data records, most attributes in a $\mathcal{QI}$ can be expected to be nominal or ordinal. However, continuous attributes can also be present. In some cases, numerical outcome attributes give enough clues for re-identification and thus, an attacker can use continuous attributes as (part of) a $\mathcal{QI}$. For instance, if respondents are workers and salary is an outcome attribute, direct competitors know what is the salary paid for companies in a certain sector. In the case of recommender systems and CF, the ratings are usually stored as continuous data (integers or floats). In such scenario, we consider each rating as a $\mathcal{QI}$.

Since the objective of anonymisation methods is to prevent confidential information from being linked to specific respondents, the table is anonymised before being published to others, that is, direct identifiers (name, passport no.,etc.) are removed and $\mathcal{QI}$ are modified. As a result, individual's identity and sensitive attribute values can be hidden from adversaries.

2.2.1 Data Anonymisation Techniques

The current proposals suggest several ways to hide information. Existing methods can be classified into statistical perturbation methods and cryptographic methods [55]. A well-known methodology is the aggregation of individuals in a data file by enforcing them to share the same characteristics through data perturbation methods such as suppression, generalisation, sampling, noise addition, rounding, swapping and etc. However, such methods only hide partial information, which could not be enough to preserve the users' privacy. After removing obvious identifiers, some of the most basic methods for maintaining privacy of publicly released datasets employed by data releasing agencies include the following [134] [112] [55]:

Top/bottom coding This technique is able to reduce the disclosure risk of extreme values by limiting the largest (or smallest) value for a given variable. For example, if an individual has an extremely large salary, rather than

reporting the exact amount, an agency may simply report it as over $100,000$. Likewise, negative values of income may be reported as less than 0.

Rounding. This method replaces an observed value by a simpler or an approximate representation, according to a rounding methodology [74]. The most basic form of rounding is to replace an arbitrary value by an integer. For instance, a value 1.7 will be rounded up to 2 and a value 1.2 will be replaced by 1.

Sampling. The sampling method consists in selecting a portion or a sample of the population, with a view to draw conclusions about the entire population. It is based on the law of statistical regularity that says that a moderately large number of items chosen at random from the large group are almost sure on the average to possess the features of the large group. Moreover, following the law of inertia of large numbers, the larger the size of the sample, the more accurate the results are likely to be.

Suppression. In a contingency table, cells with too few observations cannot be released to the public, because it could lead to disclosure. The suppression of these cells is a simple procedure for controlling such disclosure. The same applies to the values of some combination of variables, if they are unique or nearly unique in the table.

Generalisation. This operation replaces values with a parent value in the taxonomy of an attribute. Typical generalisation schemes include full-domain generalisation, subtree generalisation, multidimensional generalisation, etc.

Limitation of detail. Similarly to generalisation, this technique records variables into intervals and collapses categories, in which only a small number of observations appear.

Anatomisation. This operation does not modify the $\mathcal{QI}$ or the sensitive attribute, but de-associates the relationship between the two. Anatomisation-based method releases the data on $\mathcal{QI}$s and the data on sensitive attributes in two separate tables.

Data swapping. Data swapping methods propose to locate the fields in a dataset that are close, according to a context, and exchange the values of these fields without modifying them. For instance, in recommender systems, this method manages to preserve statistical values but hides the specific preferences of users, so achieves privacy. For example, if the u_a has rated a terror film t_a with a 9 (*i.e.* assuming that the range of values goes from

0 to 10) we could swap the rating of t_a by the one of t_c, being t_c another terror film. Therefore, we could obfuscate t_a's value without compromising u_a's preference for the terror genre.

Addition of noise. Rather than releasing the actual values of the data, noise is added to the data in an attempt to prevent a linkage attack from occurring. The perturbed data can be correctly analysed by accounting for the extra variability from the added noise.

In the case of noise addition, we may perturb the values of a dataset using a Gaussian distribution with zero mean and standard deviation σ (*i.e.* $\mathcal{N}(0, \sigma)$). The higher the σ value, the greater the range of the generated values (*i.e.* it is more likely to generate values close to the boundaries of the value range). We may also use a discrete uniform distribution $\mathcal{U}(S)$, where S is the set of actual values present in the category that is being evaluated (*i.e.* we may substitute rare or values with too few observations by other real values present in the dataset, such as age or weight). Laplace distributions are also prevalent because of their interesting properties [56].

A simple way to hide a number a is to add a random number r to it. Although we cannot do anything to a since it is disguised, we can conduct certain computations if we are interested in the aggregate data, rather than in each individual data [127]. The main idea of randomised perturbation is to perturb/obfuscate the data in such a way that certain computations can be performed while preserving users' privacy. Although information from each individual user is obfuscated, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for computations that are based on aggregate information. For instance, scalar product and random sum are among such computations and prevalent in CF algorithms. For those computations, we can still generate meaningful outcome without knowing the exact values of individual data items because the needed aggregate information can be estimated from the obfuscated data.

Random Sum Let $\vec{O}$ be the original vector with n values, where $\vec{O} = (o_1, o_2, \ldots, o_n)$. $\vec{O}$ is disguised by $\vec{R} = (r_1, r_2, \ldots, r_n)$, where r_i's are values generated by a Gaussian distribution with 0 mean and standard deviation σ. Let $\vec{O'} = \vec{O} + \vec{R}$ be the disguised data that is known. Since r_i's are uniformly distributed in domain $[-\sigma, \sigma]$, the contribution of the sum of the random values to the actual sum of the values of vector $\vec{O}$ is close to zero. In the long run, the relative error will converge to zero. Therefore, we have

$$\sum_{i=1}^{n}(o_i + r_i) = \sum_{i=1}^{n} o_i + \sum_{i=1}^{n} r_i \approx \sum_{i=1}^{n} o_i \tag{2.1}$$

Scalar Product Let $\vec{A}$ and $\vec{B}$ be the original vectors, where $\vec{A} = (a_1, a_2, \ldots, a_n)$ and $\vec{B} = (b_1, b_2, \ldots, b_n)$. Let $\vec{R}$ and $\vec{V}$ be two vectors with values generated by a Gaussian distribution with 0 mean and standard deviation σ, so that $\vec{R} = (r_1, r_2, \ldots, r_n)$ and $\vec{V} = (v_1, v_2, \ldots, v_n)$ and the sum of the values of $\vec{R}$ and $\vec{V}$ are equal to 0. $\vec{A}$ is disguised by $\vec{R}$ and $\vec{B}$ is disguised by $\vec{V}$. Let $\vec{A'} = \vec{A} + \vec{R}$ and $\vec{B'} = \vec{B} + \vec{V}$ be the disguised data that are known. The scalar product of $\vec{A}$ and $\vec{B}$ can be estimated from $\vec{A'}$ and $\vec{B'}$ as follows:

$$\vec{A'} \cdot \vec{B'} = \sum_{i=1}^{n}(a_i b_i + a_i v_i + r_i b_i + r_i v_i) \tag{2.2}$$

Because $\vec{R}$ and $\vec{B}$ are independent, we have

$$\sum_{i=1}^{n} r_i b_i \approx 0 \text{ and similarly } \sum_{i=1}^{n} a_i v_i \approx 0 \text{ } and \text{ } \sum_{i=1}^{n} r_i v_i \approx 0 \tag{2.3}$$

Therefore, we have:

$$\sum_{i=1}^{n}(a_i + r_i)(b_i + v_i) = \sum_{i=1}^{n}(a_i b_i + a_i v_i + r_i b_i + r_i v_i) \approx \sum_{i=1}^{n} a_i b_i \tag{2.4}$$

As previously stated, a number of information disclosure limitation techniques have been designed for data publishing, including sampling, suppression, rounding, swapping, noise addition and etc. These techniques, however, compromised data integrity of the tables. In order to overcome that, there are techniques that guarantee that data published satisfies some privacy definitions.

Microaggregation is a family of SDC algorithms for datasets used to prevent against re-identification, which works in two stages:

1. The set of records in a dataset is clustered in such a way that: i) each cluster contains at least k records; ii) records within a cluster are as similar as possible.

2. Records within each cluster are replaced by a representative of the cluster, typically the centroid record (*i.e.* the average of the cluster).

When microaggregation is applied to the projection of records on their $\mathcal{QI}$ attributes, the resulting dataset is k-anonymous [162], that is, to an intruder each record in the dataset is indistinguishable within a cluster of k records in terms of $\mathcal{QI}$s. While k-anonymity protects against identity disclosure [162], it is insufficient to prevent attribute disclosure. To overcome this limitation of k-anonymity, Machanavajjhala et al. [105] introduced a new notion of privacy named l-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least l well represented values. One issue with l-diversity is that it is limited in its assumption of adversarial knowledge. As showed in [95], it is possible for an adversary to gain information about a sensitive attribute as long as she has information about the global distribution of this attribute. This assumption generalises the specific background and homogeneity attacks used to motivate l-diversity. Another problem with privacy-preserving methods in general is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitive. However, being close to the value is often good enough, especially with numerical attributes. Since recommender systems usually work with continuous data, this needs to be considered. In order to overcome the limitations of l-diversity, a privacy notion named t-closeness was presented in [95], which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (*i.e.* the distance between both distributions should be no more than a threshold t). One key novelty of such approach is that the authors separate the information gain that an observer can get from a released data table into two parts: that about all population in the released data and that about specific individuals. Therefore, only the second kind of information gain needs to be limited. Another privacy definition, p-sensitivity [161], focuses on preserving the number of distinct values for each sensitive attribute in each group. Therefore, to satisfy p-sensitivity, a sensitive attribute must appear at least p times in the same group/cluster. We can find methods in the state-of-the-art that guarantee one or more privacy definitions (*e.g.* k-anonymity and p-sensitivity or k-anonymity and t-closeness) [95, 175]. More recently, a new notion of privacy named differential privacy or ε-privacy appeared [56]. Differential privacy states that the risk to one's privacy should not substantially (as bounded by a parameter ε) increase as a result of participating in a statistical database. Thus, an attacker should not be able to learn any information about any participant that they could not learn if the participant had opted out of the database. However, the main drawback is that a huge amount of noise needs to be added to achieve ε-privacy. This problem is exacerbated in CF datasets,

due to their high dimensionality.

The majority of the previously stated methods are compatible with different types of variables (*e.g.* numerical and categorical) because they follow a preprocessing step [55].

2.3 Privacy-Preserving Collaborative Filtering

The widespread use of CF on the Internet provides great opportunities and benefits to both companies and users, but there is a major drawback: the lack of users' privacy. The importance of privacy in CF systems is emphasised by the growing pace at which information of each user is collected and stored. Careless management of personal information, besides being illegal, could lead to serious consequences for both users, whose information is stored, as well as businesses. One of the main problems in CF is that customers who believe that their preferences/profiles may be exposed, do not give their assessment on a particular item or, if given, they do it incorrectly or distorted [45]. This user behaviour, derived from the lack of privacy feeling, incurs in a reduction of both the number of assessments as well as their quality. Another drawback is that companies can acquire data of the preferences of many users in a given market, getting a big advantage over new competitors if they decide to expand into other markets. Moreover, the existence of large monopolies on the Internet (Google, Amazon) is another clear disadvantage, so users' data could be transferred between different entities, which are managed by large companies, without the users' awareness. In [114], the authors show how the history of purchases and opinions of user profiles in eBay can disclose their interests and the types of products they have purchased. Moreover, serious privacy issues can raise if information about users are correlated with other sources. Another example is showed in [71] where authors perform a study about how e-commerce websites characterise users and personalise their searches to perform price discrimination (*i.e.* discounts for some kind of users) or steering (*i.e.* reorder recommendations to influence our searches or our purchases).

Interestingly, although CF methods with privacy obfuscate and/or hide information of user profiles, the creation of communities of similar users, which is a very common fact in the network, can become a double-edged sword. First, users can easily find reliable recommendations on items from communities in a particular context. Second, it may incur a *homophily* problem in the network, more specifically a problem of *value homophily* [113], so that recommendations outside the context of the community would give results with little sense, precisely because of the homogeneity of the group. In

order to solve the privacy issues raised by the systematic collection of private information, which is required for the proper use of CF, current research focuses on Privacy-Preserving Collaborative Filtering (PPCF) methods.

In a dynamic market like the Internet, companies may cooperate to obtain better recommendations for their customers. Due to privacy and business concerns, data should not be disclosed between companies. In this context, data might be partitioned between various parties in different ways:

Vertical partitioning (VP): in which companies own disjoint sets of items but with the same users. This situation can be found within third parties of the same company, but usually VP is more suitable to obtain information about individuals crossing large amounts of information of different kinds of databases.

Horizontal partitioning (HP): in which different parties hold disjoint sets of users with opinions of the same items. Worldwide communities or e-commerce companies, which topics may be related, are suitable for this kind of data partition model. For instance, many companies selling books can collaborate to achieve better recommendations and predictions for its users, increasing the benefits without losses.

Arbitrary partitioning (AP): in which there is no pattern of how data are distributed. If the entire set is defined by an $m \times n$ user-item matrix, one party A holds a subset of users $m_a \leq m$ while another party B holds the rest $m_b = m - m_a$ and the same is applied for items. Notice that VP and HP are specific cases of AP. This is the most realistic scheme in e-commerce, where companies offer different products within the same context and users can be interested in an undetermined number of different companies.

As shown in the literature and in Section 2.2, there are several ways to protect privacy in databases. In the next subsections, we propose a classification of the PPCF approaches into centralised methods and decentralised methods according to where the information is stored, and we recall the most relevant ones in each category.

2.3.1 Centralised PPCF Methods

Centralised methods use a third party to make intermediate calculations among users or entities. Moreover, ratings are stored in a single server where recommendations and predictions are computed. Cases in which data are partitioned are not considered centralised, because data are distributed among different parties. The general scheme of centralised architecture is depicted in Figure 2.1.

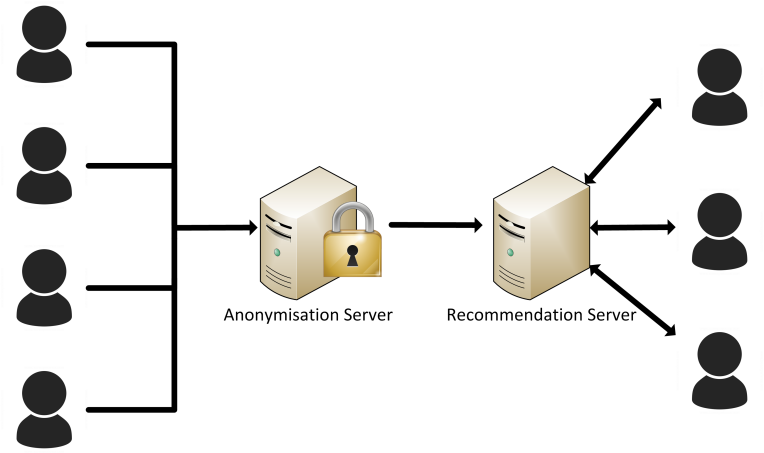


Figure 2.1: Centralised PPCF scheme. First, the information of users is collected and anonymised in a secure server. Next, such obfuscated information is stored in the recommendations server. Finally, users establish a bidirectional communication with the server (*i.e.* query & recommendation). Note that the obfuscation is performed in the anonymisation server.

Several centralised PPCF methods can be found in the literature. In [127], Polat et al. describe a technique that achieves a good balance between privacy and quality of recommendations. Due to privacy concerns, the central server should not store real users' data. To avoid the disclosure of the real users' profiles, users distort their data by adding random vectors, which are created following a Gaussian distribution, before sending them to the server. To obfuscate the data, the server decides a range $[-\alpha, \alpha]$, known by users, to truncate the random values generated by the Gaussian distribution. Next, each user u_i calculates the z-scores z_{ij} of the elements that u_i has rated to transform the original distribution in one which the mean equals to 0 and the standard deviation is equal to 1. Finally, each u_i creates n_i uniform random numbers r_{ij} in the range $[-\alpha, \alpha]$, where n_i is the number of rated items by the user. After that, each u_i aggregates the random numbers to the z-score ratings and generates the disguised z-scores $z'_{ij} = z_{ij} + r_{ij}$. Subsequently, users send the disguised z-scores to the server. After obtaining the disguised z-scores z'_{ij} of various users, the server is able to sent the aggregated information that users need to compute the predictions locally. In a more recent work [130], Polat et al. proposed the use of SVD to reduce the dimensionality of the database and alleviate the sparseness problem. In [194], Zhang et al. propose an item-based PPCF scheme,

which adds noise to the ratings to protect the users' privacy. In this case, the level of perturbation of each value is selected according to its importance in the recommendation process, instead of the classic item-invariant perturbation [127, 130]. Their results show that their method obtains a better privacy/accuracy trade-off in respect to item-invariant perturbation. Following the data obfuscation research line, Parameswaran et al. proposed the *Nearest Neighbour Data Substitution* (NeNDS) method in [121]. Such proposal uses a CF server where different sources combine their data to obtain enough information to better deal with data sparsity. The scheme assumes that entities have three types of databases: User-info, Item-info and Ratings-info. Those databases are obfuscated and sent to the central server, which performs a data aggregation and returns the databases with new values to each source. In order to preserve the properties of each database content, each field is treated separately and permutations of similar items are performed to obfuscate the data. Next, to provide a more robust data perturbation, the authors use geometric transformations (scaling, rotation and translation) because they preserve both the database subgroups and the distances between elements. In [155], Shorki et al. propose to share ratings among similar users to obfuscate the real data, so users' profiles become partially mixed. In this way, modified profiles can be sent to a central server, preserving the privacy of the users involved. More examples of centralised methods can be found in [18] and [19].

2.3.2 Decentralised PPCF Methods

Decentralised methods use the members of a distributed network to perform intermediate calculations and predictions on items ratings. The use of decentralised schemes generally ensures that information exposed is much less than in case of centralised systems, but this entails the use of expensive protocols and more complex calculations. Typically, in decentralised PPCF methods, users store their own ratings. Well-known PPCF with partitioned data schemes, which involve different parties sharing their data to perform CF with more referrals, are also considered decentralised methods. The general scheme of decentralised architecture is depicted in Figure 2.2.

A vast number of methods and secure protocols for data mining with partially distributed datasets can be found in the literature: (i) privacy-preserving outlier detection [167][133]; (ii) privacy-preserving nearest neighbour search [152][133]; (iii) privacy-preserving clustering [82] [81] [132] and (iv); privacy-preserving SVM [169]. Moreover, due to its interesting properties, public key homomorphic cryptosystems [119, 58], secure multi-party

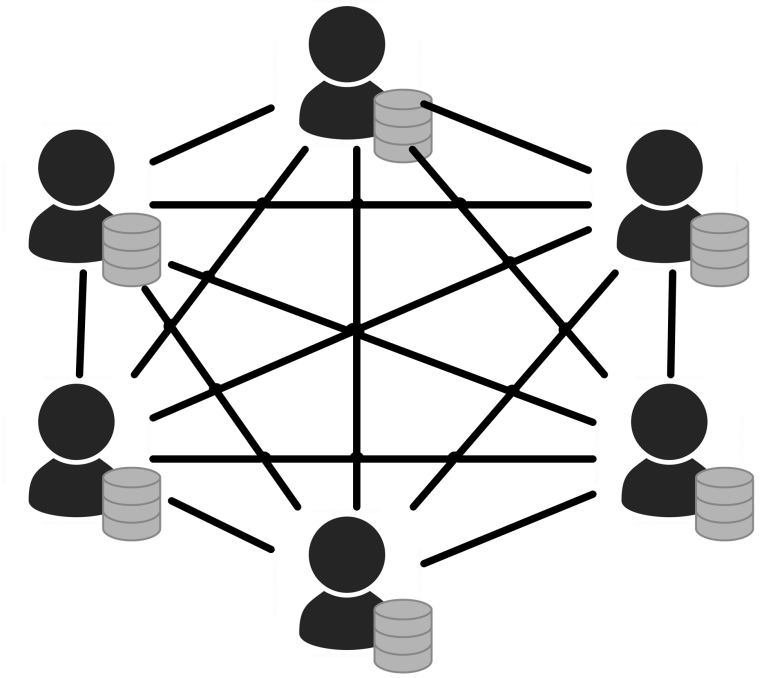


Figure 2.2: Decentralised PPCF scheme. Users store their ratings locally. Note that information is exchanged using privacy-preserving protocols between users.

computations [190] and cryptographic protocols are used on the Internet.

Several approaches with partitioned market basket databases have been proposed in the literature [129, 84, 189]. These kind of databases are suitable to perform top-N recommendations with high accuracy and low computational cost due to their binary content. For instance, Polat et al. proposed a method in [129], which performs recommendations over vertically and also horizontally partitioned market basket data. In such datasets, ratings are stored as $r_{ij} = 0$ if the item i_j has not been bought for the user u_i and $r_{ij} = 1$, otherwise. The authors argue that purchasing a product does not necessarily please the user as he may be disappointed about it, but it is still a good indicator of the preference of a user for an item. However, it is better to show the users' preferences as if they liked the item (1) or not (0). Assuming that data are partitioned between parties A and B, the authors use privacy-aware protocols and random permutations to avoid the disclosure of private information. The authors note that a user is a potential consumer of the products that have been already purchased by his neighbours. How-

ever, he may also be interested in items that have not been purchased by the most dissimilar users. Therefore, before recommending the top-N items to u_a, the values of the most dissimilar users of u_a are changed from 0 to 1. Finally, the n items that contain more 1s are recommended. Polat et al. [189] proposed the use of a naïve Bayesian classifier (NBC) to perform PPCF recommendations in APD. The proposed method has different offline and online stages to increase the overall efficiency. The authors use secure protocols in the offline stage to construct the NBC-based prediction model. This prediction model construction has two steps. The first one includes the estimation of the likelihood probabilities, which is conducted using secure protocols based on public key Pailier's [119] homomorphic encryption. The second one, utilises randomised perturbation techniques to calculate the probability of having 1 or 0 on a target item. The online stage implies the recommendation estimation, which is performed by their online recommendation estimation protocol. A similar NBC approach on binary data has been also proposed by Kaleli et al. [84], yet using a peer-to-peer (P2P) network. Authors argue that centralised storage poses several hazards to users because a single entity controls users' data. To avoid that privacy issue, they propose a P2P network with which users (acting as peers) communicate and exchange data to compute recommendations. Furthermore, NBC with privacy has been also conducted with numerical rating contexts [87]. In [187], Polat et al. propose a hybrid PPCF approach on cross distributed data (CDD), which is a simpler data partition scheme than APD and can be considered as a specific combination of HPD and VPD cases. Later, in [188], the authors proposed a more realistic APD scenario, embracing more data configurations than the VPD approach showed in [128]. In [186], Yakut et al. propose the use of SVD to conduct PPCF on both vertically and horizontally partitioned data. When data are partitioned between two companies P and Q, which have R_P and R_Q matrices respectively, the authors need to handle four principal issues to conduct the SVD [174] of the full matrix, defined as $R = USV^T$. These issues are to remove the sparsity of R, normalise the ratings, compute R^TR and, finally, find the column vectors of U. They use privacy-preserving protocols to calculate and securely exchange all data involved. Finally, both parties have the estimated U_p, S_p, and V_p matrices to construct the SVD model with which perform recommendations. There are also examples of privacy-preserving approaches with Trust-Aware systems. In [50], the authors highlight the problem of disclosing trust network data because this information may reveal the users' behaviours and communities. To overcome that issue, they present a scheme, which is considered decentralised because trust computations are made locally by each

user. Privacy-preserving Trust-Aware systems have also been proposed with VPD. The scheme proposed in [85] is composed of both an offline and an online stage. First, they perform privacy-preserving offline computations to obtain the partial trusts between users. Next, a distance-based private sorting algorithm (DPSA) is utilised by the parties to determine the neighbourhood of each user. The neighbourhood will be formed by each user's k most trusted neighbours. During the offline process, parties use DPSA protocol n times, where n is the number of users, to determine the neighbourhoods. Next, they use a private recommendation protocol, which uses Paillier's [119] homomorphic encryption, to provide online predictions. In [77], the authors use Pearson correlation (PC) to compute similarities between users on HPD. However, to compute PC in HPD the preference of the user u_i over item i_j is disclosed because the mean value of u_i needs to be shared among the entities. To avoid that disclosure, authors propose a secure protocol to compute PC between users using ElGamal [58] homomorphic encryption. Once the neighbourhoods are established, recommendations can be computed straightforward. In a similar approach [192], the authors use a secure protocol to compute the scalar product (*i.e.* because the numerator of the PC is a scalar product). Authors argue that secure scalar product protocol is more efficient than ElGamal encryption because the trust entity, which is needed in the protocol to generate random numbers, can provide these numbers before the protocol starts. This can be performed by knowing in advance how many scalar products are going to be calculated between entities. Therefore, entities can drastically reduce communication overhead exchanging information in one step and not in n. In [26], J. Canny proposes a peer-to-peer recommendation scheme using encrypted data. In this scheme, users have their own rating vector and only the aggregates are exposed. Users can decide whether to share their data for CF or not, so this scheme may encourage the formation of communities rather than the use of similarities to compute recommendations. To obtain the ratings' set, the author uses an iterative SVD method, which is computed using the conjugate gradient method (*i.e.* an iterative algorithm that solves lineal equation systems in which matrices are symmetric) [131]. Later, in [27], Canny proposed the use of Expectation-Maximization factor analysis [49], which deals better with sparse data, to compute the reduced dimensionality matrix. In this case, the author shows a gain in speed and a good quality of recommendations with respect to [26].

Berkvosky et al. [16] also propose the use of a peer-to-peer decentralised scheme. In this case, each user has its own rating vector and there is no need for a central server or trust party. When user u_a sends a request over

an item rating i_a, each user decides whether to provide her data for CF or not. If u_b decides to collaborate, u_b calculates the similarity between him and u_a using the Cosine similarity metric [160] and sends the result to u_a with r_{ba} (*i.e.* the rating of u_b for item i_a). Once data are collected, the k nearest neighbours of u_a are selected to predict the value of i_a for u_a, using weighted average according to their similarity.

In [163], the authors propose a PPCF method based on similarity between items. They argue that similarities between items can be made public without privacy leaks because users' profiles are not exposed, neither the similarities between them. Authors use public key Paillier cryptosystem [119] for its homomorphic properties and its faster decryption process (compared with ElGamal), to calculate the Cosine correlation between items with privacy.

In [88], the authors propose a scheme, which is similar to the one showed in [163]. In this case, to reduce the overhead of the proposed protocols, the authors perform item clusters and user clusters. The outcomes show a reduction in communication overhead and computational cost with respect to computations performed using the original dataset.

Basu et al. [14] propose the use of cloud computing to perform weighted Slope One-based PPCF, which computes item deviations to perform predictions. In order to decouple the users from their ratings when they are submitted, the authors propose the use of well-known anonymisation systems. The proposed scheme has two steps. First, the cardinalities and the deviations between items are computed and stored in a cardinality matrix and a deviation matrix, respectively. Since users are not related with the information they provide and any concrete rating is sent, these operations do not disclose information. Next, users must send their rating vector to the cloud to compute predictions. To prevent data disclosure, their profiles are sent encrypted using a modified version of the public key Paillier [119] cryptosystem. This modification is needed to operate with negative numbers because deviations, which are involved in the prediction process, may not be positive. The cloud receives the user profile encrypted with the public key, and conducts homomorphic operations with the data it has and the user's profile information to obtain the encrypted recommendations. Next, it sends the encrypted recommendations back to the user, who will decrypt them with her private key. The proposed scheme has one relevant privacy issue, which is that the user reveals the number of items she has rated in the prediction phase. However, this problem can be solved if users compute predictions locally, receiving the required information from the cloud, or also making requests for unnecessary items. In [99], the authors implemented

a k-means clustering approach, which uses homomorphic encryption and trapdoor information to improve the efficiency of distance computations. In [154], the author shows the drawbacks of differential privacy and distortion privacy. The author analyses approaches that optimise the trade-off between privacy and utility in order to achieve privacy and robustness against attacks while minimising the information loss. Other example of differential privacy applied to CF methods (in this case matrix factorization) is showed in [17]. The author concludes that several factors need to be considered and, depending on the data available and the latent factors of the matrix, more noise needs to be added to achieve the desired level of privacy. Therefore, it is a very data-dependent mechanism. In the work presented in [164], authors classify the-state-of-the-art into cryptographic techniques and data obfuscation methods. Next, they propose privacy-preserving recommendation protocols for context-aware recommender systems, which include social relationships as part of the inputs to compute recommendations. Moreover, they point out several issues and propose two new protocols that provide more accuracy and diversity. For more on PPCF we refer the reader to [175] [18] and [19].

2.3.3 Current Trends, Open Problems and New Issues

Current tendencies to achieve better accuracy on recommendations focuses on how to obtain more information. Approaches that use trust networks [110] made on partitioned databases, and extra information provided by content-based methods, help to deal with cold user and sparseness problems. In that sense, well-known imputation methods such as the ones discussed in Section 2.1.2 and the ones that will be discussed in Chapter 3 could also be used to fill CF matrices and alleviate sparseness. According to a previous study of the dataset, the most fitted imputation method can be selected to obtain accurate and realistic results. Shilling attacks are also receiving increasing interest [68] due to their relevance in the accuracy of the recommendations.

Large multivariate datasets have several well-known problems. We consider the proper detection of outlier users (i.e. *outliers*) as a significant issue. In this way, several approaches to deal with outliers in large multivariate datasets have been proposed [135, 48, 171]. The presence of outliers in datasets affect the quality of the data and, therefore, the applied CF algorithms may return biased results due to its presence. For instance, clustering algorithms may incorrectly select the proper users to form a group or a cluster may be wrongly divided into small pieces.

In the privacy area, the current trend is to use secure computation protocols and public key homomorphic properties to protect the users' data while CF is performed. However, one of the main issues is that most of the state-of-the-art proposed schemes are not suitable to satisfy real-time requirements. For this reason, besides accuracy and privacy, as CF methods are prevalent to recommend products to users in real time, computation and communication overheads also need to be considered. This fact adds the variable *time* in the well-known accuracy vs. privacy trade-off. Several approaches [14] propose to decouple the users from their ratings using recognised techniques such as mix-servers or anonymisation systems. Although such methods do not guarantee security on their own, their use should be generalised to enhance the privacy of already known CF & PPCF methods. As previously stated, it is possible to classify PPCF in centralised and decentralised methods. Centralised schemes offer several benefits but also privacy issues. On the one hand, the overall efficiency is increased because extra communication costs between users and third parties are avoided. On the other hand, in centralised methods, data are managed by one party, which has total control over it, with the privacy issues implied if the data have a low protection level. No one can guarantee that companies sell information to other parties due to bankruptcy [26]. However, we can use well-known privacy protection techniques as the ones discussed in Section 2.2 to anonymise sensitive and private information and prevent its disclosure. Decentralised schemes usually have a higher data protection than centralised. This happens because involved users and parties manage their own datasets, and the data sent between them are usually controlled by secure protocols. That extra privacy gain incurs in an expensive communication overhead that could affect real-time recommendations. Although time issues can be partially addressed by incremental upgrading of the ratings matrix in a centralised scheme, this solution is difficult to be applied on schemes where users manage their own ratings and any third party is used to perform intermediate calculations, because any data are stored. In addition, users who control their own profiles tend to form communities with already known users or with those having similar behaviours. As previously stated in Section 2.3, this fact may incur in a well-known homophily problem [113], which could be partially solved by the addition of different communities. Therefore, large clustered societies could be created by means of clustering techniques and users seeking for quality recommendations on a given context could have enough information.

2.4 Evaluation Tools in Recommender Systems

In this Section, we introduce the most prevalent metrics and datasets in the CF field. Moreover, we propose a new metric, called behavioural precision, to perform a more robust analysis of the recommendation's quality. Finally, we recall some metrics from the SDC context such as information loss and disclosure risk.

2.4.1 Metrics in Recommender Systems

Herlocker et al. [73] proposed a classification of recommender systems metrics according to the following categories: predictive accuracy metrics, such as Mean Absolute Error (MAE) and its variants; classification accuracy metrics, such as precision, recall, F1-measure, ROC sensitivity and behavioural precision; rank accuracy metrics, such as Pearson's product-moment correlation, Kendall's Tau, Mean Average Precision (MAP) and etc. However, in this section we focus on the most prevalent CF metrics.

The Root Mean Squared Error (RMSE) is a popular metric because it is the Netflix prize [15] metric for film recommendation performance:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - r_i)^2}{n}} \tag{2.5}$$

where n is the number of predicted elements, p_i is the predicted value over the element i, and r_i is the real value of i. RMSE amplifies the contributions of the absolute errors between the predictions and the true values. Other well-known technique is the Mean Absolute Error (MAE). We compute the error between the original dataset values and the predicted values using the MAE metric, defined as

$$MAE = \frac{\sum_{i=1}^{n}|p_i - r_i|}{n} \tag{2.6}$$

where n is the number of predicted elements, p_i is the predicted value over the element i, and r_i is the real value of i. The lower the MAE value, the better the accuracy of the predictions.

Classification accuracy metrics try to assess the successful decision making capacity of recommendation algorithms [150]. They measure the amount of correct and incorrect classifications performed by the RS as relevant or irrelevant items. Therefore, this measure is useful for user tasks such as finding good items. The classification accuracy metrics ignore the exact rating

or ranking of items as only the correct or incorrect classification is measured. The well-known precision and recall metrics are computed from a 2×2 table, such as the one shown in Table 2.2. The item set must be separated into two classes (*i.e.* relevant or not relevant). Therefore, we need to transform the rating scale into a binary scale.

Table 2.2: Table that stores the amount of true/false and positive/negative recommendations.

	Relevant	**Irrelevant**	**Total**
Recommended	tp	fp	tp+fp
Notrecommended	fn	tn	fn+tn
Total	tp+fn	fp+tn	N

Precision (also confidence in data mining) is computed as the ratio of recommended items that are relevant to the total number of recommended items:

$$precision = tp/tp + fp \tag{2.7}$$

This is the probability that a recommended item corresponds to the user's preferences. Recall (also called sensitivity in psychology) is calculated as the ratio of recommended items that are relevant to the total number of relevant items:

$$recall = tp/tp + fn \tag{2.8}$$

This is the probability that a relevant item is recommended. Both measures, precision and recall, are inversely related. This can be noticed if we vary the size of the set of recommendations. In most cases, increasing the size of the recommendation set will increase recall but decrease precision. Precision and recall describe the recommender's performance regarding the true positives. However, we may find analysis considering false positives and false negatives [150].

Concerning the recommendation quality measures, MAE gives us an overall value that does not reflect in a unitary&precise form the recommendations' quality. For instance, a low MAE could indicate high accurate or nearly perfect recommendations for lots of users while poor and practically useless recommendations for many others. Other well-known measures such as precision [160] focus on what percentage of recommended items (*i.e.* in a

ranked list) are of interest to the user. Nevertheless, we propose the use of behavioural precision metrics to characterise the profile of the user, taking into account both positive and negative assessments, hence obtaining more information about the users' interests. We present two measures, namely *binary* precision and *four-level* (4L) precision, to gauge how recommendations fit the users' interests. In order to measure such behavioural precision, the value range of each database is divided into two (*i.e.* binary) or into four (*i.e.* 4L) levels and we ascertain in which level is u_a's recommendation. With regard to the binary precision, if the recommendation belongs to the same level as the original assessment we obtain a match or, conversely, an error. Concerning the 4L precision, we manage four possibilities. First, we get a match if the recommendation is in the same level of u_a's real assessment. Second, if the recommendation is in the adjacent level of the original assessment we achieve a slight match. Third, in the event that the recommendation has a different behaviour and is not in an adjacent level we obtain a slight error and, finally, if the recommendation falls in the opposite level, we obtain a *reversal* (*i.e.* a recommendation that is opposite of the user's behaviour). Therefore, we may divide the ratings range in different levels (the wider the range, the more possible precision levels) to perform a more accurate analysis of the recommendations. Notwithstanding, we selected two and four levels for the sake of clarity. Note that the more levels, the more complicated is to find users with the same behaviours, especially if we analyse sparse data. Moreover, users have different vote tendencies, which may change in regards of the possible range of values [118]. However, a methodology to select the proper number of precision levels in regard to the sparseness and the ratings range of a dataset will be left to future work.

The interested reader can refer to [73] and [150] for more on accuracy metrics in RS.

2.4.2 Datasets in Recommender Systems

Drawing convincing conclusions from artificial data is risky, hence, data from live experiments are more desirable for CF research [160]. With the aim to assess the quality of CF methods, the methods in the state-of-the-art propose several well-known CF datasets with different characteristics. The Netflix prize dataset [15] contains over 100 million ratings of 480,000 users on 17,000 films. The ratings were collected between October 1998 and December 2005 and reflect the distribution of all ratings received during this period. Each rating has a customer id, a film id, the date of the rating, and the value of the rating. Book-Crossing [198] contains 278,858 users providing

Table 2.3: Characteristics of the most commonly used datasets in Collaborative Filtering.

Dataset	Users	Items	Ratings	Sparseness %	Content
Netflix	480,000	17,000	$100 * 10^6$	98.77 %	Films
EachMovie	72,916	1,628	$2.8 * 10^6$	97.64 %	Films
Movielens 100k	943	1,682	$1 * 10^5$	93.69 %	Films
Movielens 1M	6,000	4,000	$1 * 10^6$	95.83 %	Films
Jester	73,421	100	$7.34 * 10^6$	44 %	Jokes
Book-Crossing	278,858	271,379	$1.14 * 10^6$	99.99 % <	Books
Book-Crossing (least 20)	3,700	130,000	$2.4 * 10^5$	99.95 %	Books

1,149,780 ratings (explicit/implicit) about 271,379 books collected from the Book-Crossing community. Due to its extreme sparseness (*i.e.* more than 99.99%), we performed a reduced version considering only users that voted at least 20 items and we discarded implicit ratings. As a result, our Book-Crossing (least 20) dataset contains ratings of 3,701 users on 130,000 items. Despite such reduction, the sparseness of this dataset is still about 99.95%. Movielens was developed by Grouplens [142] and it is one of the reference sets in CF. The most prevalent of its variants is the Movielens 100k, which contains 100,000 ratings of 943 users on 1,682 films. The Movielens 100k range values are comprised between 1 and 5. This database is highly sparse, since more than 90% of the fields are empty. Once completely filled, the matrix contains a total of 1,586,126 values. Jester [70] is a joke recommendation system developed at University of California, Berkeley. The entire database has 100 jokes and ratings of 73,421 users. As a result, the matrix contains a total of 7,342,100 values. However, this database is not as sparse as Book-Crossing or Movielens 100k since Jester has approximately the 44% of empty cells. Another common dataset is the EachMovie Dataset (http://research.compaq.com/SRC/eachmovie/). This extensive dataset has over 2.8 million ratings from over 70,000 users, and it includes information such as timestamps and basic demographic data for some of the users. The main characteristics of the previously described databases are summarised in Table 2.3.

2.4.3 Metrics in SDC

In order to measure the quality of the protection provided by a perturbation method we consider two factors, namely the information loss and the disclosure risk. The information loss is generally associated to the sum of squared errors (SSE). The SSE is commonly used as a measure of the distortion introduced on the original data. Given an original dataset O represented by a matrix of $n \times m$ elements o_{ij} and a distorted/protected dataset P represented

by a matrix of $n \times m$ elements p_{ij}, the SSE is computed as follows:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{m} (o_{ij} - p_{ij})^2 \tag{2.9}$$

In the special case of microaggregation, the SSE is computed in vector notation as follows:

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mathbf{x}}_i)'(x_{ij} - \bar{\mathbf{x}}_i) \tag{2.10}$$

where g is the number of clusters generated by the algorithm, n_i the number of elements in each cluster, x_{ij} the vector of the j-th user of the i-th cluster and $\bar{\mathbf{x}}_i$ is the average vector of the i-th cluster. Subsequently, the total sum of squares is defined as

$$SST = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mathbf{x}})'(x_{ij} - \bar{\mathbf{x}}) \tag{2.11}$$

where $\bar{\mathbf{x}}$ is the mean of the entire dataset.

Once we obtain the SSE and the SST we can normalise the information loss between 0 and 1 using (2.12) :

$$IL = \frac{SSE}{SST} \tag{2.12}$$

Resulting in a high value in case of a large loss of information and a low value if the data have not been severely modified.

The disclosure risk (DR) measures the probability of correctly relating a record of the obfuscated/protected data matrix with a record of the original matrix. It is also known as the probability of re-identification, or the re-identification risk. For an attacker, the re-identification procedure consists in computing the distances (*e.g.* the Euclidean distance) between a given protected record p_i (corresponding to user i), and the target records o_j, that could be obtained from third party sources such as censuses and the like. In our case we assume the best scenario for an attacker (the oracle scenario) in which he has the original dataset O and the distorted dataset P and he tries to link each record p_i in P with the records o_j in O.

For each record $p_i \in P$ the attacker determines the closest record $o_j \in O$. If that closest record o_j is actually the original record belonging to p_i, the attacker succeeds and we say that p_i has been re-identified. To compute the disclosure risk, we try to re-identify all records and then compute the percentage of correct re-identifications.

As previously stated, two factors are considered to measure the quality of the privacy provided by a perturbation method, namely the information loss and the disclosure risk. Notwithstanding, the quality of recommender systems depends on the accuracy of the predictions, since we need to measure the real data usability when computing recommendations, being far less significant the value of the information loss. Therefore, in terms of privacy and utility of the data, both MAE and DR should be low.

Hence, the best approach will be the one yielding optimal trade-off between quality of the predictions (*i.e.* MAE) and the level of privacy (*i.e.* DR). We propose to combine such measures in a score, namely T-score, defined as:

$$\text{T-score} = \frac{MAE + DR}{2} \tag{2.13}$$

In this case, the lower the percentage the better the obtained score.

2.5 Recommender Systems to Enable Smart Health

Smart cities are progressively gaining importance due to the worldwide process of urbanisation that is taking place and the need for more sustainable and liveable cities. In the next sections we point out the relevance of providing healthcare services within the context of a smart city. Moreover, we recall the concept of smart health and show how the sensing infrastructure of a smart city provides s-health with all its power and potential.

2.5.1 Healthcare Evolution and the Urbanisation Process

Our society is getting older and chronic diseases are gaining relevance. Nowadays, millions of people suffer from chronic respiratory diseases such as asthma or Chronic Obstructive Pulmonary Disease (COPD) [123]. More concretely, over 300 millions of people suffer from asthma [65] and more than 3 million people died of COPD in 2012, which is equal to 6% of all deaths globally that year [179]. Musculoskeletal conditions such as arthritis and back pain affect near 2,000 million people worldwide and have the fourth greatest impact on the overall health of the world population, considering both death and disability [172]. This kind of disease has increased by 45% over the past 20 years and causes the 21.3% of all years lived with disability, second only to mental and behavioural disorders. However, it is worth to emphasise that the first causes of death over the past decade are hearth-related diseases [180]. The aforementioned statistics are exacerbated

if the process of urbanisation and the climatic changes are taken into account. For instance, exposure to air pollution is responsible for millions of premature deaths and diseases worldwide every year [178] and overexposure to ultraviolet light (UV) is related to different types of skin cancer [177]. Moreover, adverse weather conditions combined with pollution exacerbate the negative effects on human health, especially in children, the elderly and in patients with chronic diseases [21]. Notwithstanding, the average age of the world population has experienced a progressive increase over the last 50 years mainly due to the advances in the fields of medicine and healthcare. Although a physically active lifestyle is important to reduce/avoid physical issues such as cardiorespiratory problems, morbidity or musculoskeletal diseases, it has also positive neurocognitive effects [29]. For instance, an increase in cardiorespiratory fitness is associated with increased cortical thickness in both healthy adults and elder people diagnosed with dementia [139], and acute exercise in younger adults has shown to positively affect the performance on tasks of executive function mediated by the prefrontal cortex [13]. Therefore, the ageing of the society and the need for fostering healthy habits among the population imply a great challenge for public healthcare systems. Many efforts have been devoted to provide such systems with tools to cope with this situation, and improve them in terms of efficiency, accuracy, reliability, and sustainability. The healthcare sector has always been very active in seeking new technologies that could be integrated so as to improve it. Worldwide, healthcare models are steadily shifting towards patient-centric approaches in which patients are not only passive elements of the systems, but proactive contributors to their own health and that of the others. Behind this paradigm shift we find Information and Communication Technologies (ICT) that are playing a fundamental role to make it a reality.

The adoption of ICT within the healthcare sector led to the concept of electronic health (e-health) [60]. After the consolidation of e-health, the generalised use of mobile devices (*e.g.*, smartphones) opened the door to mobile health (m-health) to the general public. Actually, m-health has an extraordinary potential as it adds to the advantages of e-health the benefits of ubiquity of mobile devices (*i.e.*, global monitoring capabilities, immediacy, etc) [166]. ICT might also be used for a variety of health-related tasks, namely communication between patients, doctors and carers, distant provision of care, remote support to diagnostic, electronic medical records, medication adherence control, etc. Although the deployment of e-health and m-health implies investments, their adoption could contribute to lessen healthcare system costs, in the mid term, and to improve the patient's quality of life.

Along these trends towards the adoption of e-health and m-health, another important trend is taking place: urbanisation. People are moving into cities and this poses difficult challenges to cities' governments. The city management has to be adapted to a growing and very demanding population and, as a result, cities have started to incorporate ICT [69] and the concept of smart city has appeared. Currently, we can already find several examples of cities that pursue the goal of being "smart" (*e.g.*, Amsterdam, [97], Toronto, New York, London, Tokyo, Hong Kong or Barcelona, to name a few).

The definition of Caragliu et al. [28] extended in [126] says that smart cities are:

> "*cities strongly founded on ICT that invest in human and social capital to improve the quality of life of their citizens by fostering economic growth, participatory governance, wise management of resources, sustainability, and efficient mobility, while they guarantee the privacy and security of the citizens".*

In the trends that we have briefly described, namely the wide adoption of e-health and m-health and the rise of smart cities, citizens play a fundamental role: they are called to actively participate and collaborate. Citizen participation is also apparent from the new trends towards crowdsourcing. In the context of a smart city, citizens are understood as intelligent sensors that contribute with their knowledge and sensing capabilities to the city. Citizens are no longer information consumers but also producers, thus, they become *prosumers* that share information about a wide variety of things. In a much wider perspective, the sharing of information has led to the introduction of recommender systems [143], which have evolved and integrated in our daily lives in a transparent manner. Recommender systems take advantage of the collaboration among users and help them make better decisions and create more united communities.

2.5.1.1 Mobile Health

The generalisation of mobile devices with unprecedented processing and communication capabilities in the healthcare sector gave birth to mobile health (m-health). In the words of Istepanien et al. [80], m-health could be understood as *"emerging mobile communications and network technologies for healthcare systems".* m-Health helps to improve the efficiency of tasks such as the remote monitoring of patients and the communication of real-time data among the main actors of the system, namely doctors, patients, nurses and caregivers [183]. In addition, m-health reduces the time required

to gather data and it can be seen as a better way of providing effective access to health services to people without nearby hospitals.

The main contributions of m-health over e-health [91, 157] are: (i) easy access to services and data: Thanks to the ubiquity of mobile devices, services can be provided everywhere and data could be collected from everywhere regardless of the location the of patients (ii) patient-centric: m-health is user-oriented, and (iii) personalisation: Patients can receive customised services specially tailored to their needs. We point the interested reader to [138] for some of the most recent advances in m-health.

2.5.1.2 Smart Health

With the aim to use the infrastructure of modern smart cities to improve the quality of life of citizens and their healthcare system, the concept of smart health (s-health), introduced by Solanas et al., appears as a natural evolution of e-health in the context of smart cities and is defined as follows:

> *"Smart health (s-health) is the provision of health services by using the context-aware network and sensing infrastructure of smart cities."* Solanas et al. [157]

Considering the above definition, any application/service that uses the smart city infrastructure to provide healthcare or to promote healthy habits could be considered a smart health application. Clearly, s-health is a subclass of e-health because it is founded in ICT like m-health. However, it differs from m-health in that the infrastructure of the smart city is not necessary mobile and in most cases it will be static.

Nowadays, with the great efforts that are devoted to build the cities of the future, full of sensors and actuators that would help their citizens live better, the healthcare sector has turned into them to create a powerful symbiosis and create s-health. Smart cities are (or soon will be) equipped with myriads of sensors able to analyse a great variety of features that could affect our health. The following are some examples:

In big cities with diverse orography, temperature and humidity can vary significantly. It is well-known that temperature and humidity affect perspiration. Thus, this information could be used to suggest citizens the right quantity of liquids to drink during their daily activities depending on their location in the city. This is especially important for elderly and children, and people suffering from congestive heart failure (CHF) [165] and similar diseases.

Highly populated cities with dense traffic conditions, and cities with nearby polluting industries might be affected by variable degrees of pollution. The information from pollution sensors distributed in the city is highly relevant for people with respiratory problems such as COPD [123]. For instance, in 2011, 430,000 premature deaths in the EU were associated with air pollution in cities, according to the European Environment Agency [59]. Moreover, sensors that can detect allergens are especially important for people that suffer from allergic rhinitis and similar conditions [57].

Luminosity sensors: light conditions, mainly at night, could vary a lot from one street to another. Citizens with visual impairments [46] could be affected by the amount of light and might decide to choose different paths by using the information provided by luminosity sensors.

Thanks to the use of cameras and other volumetric sensors, smart cities can detect crowds. This could be a really important information for people with panic disorders and diverse forms of agoraphobia [11].

This sensing infrastructure provides s-health with all its power and potential. However, it is even more powerful because it can be used to augment the capabilities of m-health. m-Health extends the capabilities of indoor monitoring environments and it is a powerful tool that allows the advance of several lines of research [182]. The generalisation of the Internet of Things (IoT) and the great efforts that are devoted to build context-aware environments have contributed to the massive deployment of s-health solutions. IoT is distributed and ubiquitous and enables many s-health applications. In this sense, s-health applications are "by default" user-centric and oriented to a personalised improvement of patients' health. However, data generated by the context-aware sensing infrastructures and s-health applications can also be used by governments [83], healthcare authorities, and other stakeholders to reduce costs and improve efficiency [122].

2.5.2 Wireless Channel Characterisation

Nowadays, it is well-known that the sustainable improvement of citizen's quality of life is one of the main challenges to be faced by governments, policy makers and researches throughout the world. Ambient Assisted Living (AAL), e-health, m-health and the broader concept of s-health (s-health) [157], in which contextual information within a Smart City is employed, play a key role in achieving sustainable healthcare. In this context, real time monitoring and interaction of patients and users with health specialists can be performed remotely [120][94], decreasing overall costs and increasing

quality of life, reducing patients displacement, and allowing them to live in their homes [92][116]. Parameters such as biomedical signals, drug distribution, patients' behaviour, interactions, and alarm signals can be readily collected and analysed. Therefore, wireless communication systems have gained huge importance as they enable real-time connectivity, increasing system functionality and decreasing operational costs [107, 79, 24].

The implementation of context-aware environments relies on the use of a wide variety of communication systems and wireless communication solutions, which allow seamless connectivity by means of different wireless infrastructures [4]. The sustained use of wireless communication systems has led to the adoption of adaptive modulation and coding and spectrum allocation schemes, to optimise coverage/capacity requirements. In this sense, interference control plays a vital role in the performance of wireless systems, particularly in 4G and 5G communication systems. Overall, power spectral density of interference depends on network topology, spatial concentration of users, and intrinsic characteristics of network terminals and access points/base stations. When considering the implementation of AAL environments and s-health solutions, we need to face scenarios that are complex in terms of wireless propagation, due to the presence of multiple elements, such as furniture, building structure, or people, which can give rise to strong fading effects in a nonuniform manner. Moreover, user density also affects interference levels, with density values that can exhibit strong variations as a function of scenario location or time of analysis, particularly in the case of interaction with wearable devices, wireless body area networks, or device-to-device connections. Accurate estimations of these interference levels will help to avoid possible communication errors, which in some cases, like some e-health applications where medical sensors take part, could be critical [98]. In the case of healthcare systems, one of the most complex situations is present within Emergency Rooms, in which a large amount of users can be potentially located, multiple wireless systems could be under operation and there are potential interfering sources such as operating magnetic resonance imaging systems [30, 36]. Therefore, the impact of the scenario on signal levels and the location of potential wireless transceivers have a direct effect on overall system performance in terms of quality of service and required energy consumption [100].

Wireless channel characterisation, in terms of useful received signal levels for a given set of connections as well as for potential interfering connections, can be a challenging task due to the large size of scenarios, the existence of multiple frequency dependent materials, and inherent variability of mobile connections, given by the movement of potential scatterers. In this sce-

nario, several approaches can be followed, from empirical estimations, which provide results at low computational cost with low precision, to full-wave simulation techniques, which provide high accuracy at very large computational cost. As a midpoint, deterministic methods like Ray Tracing and Ray Launching, combining Geometric Optics and Uniform Theory of Diffraction can provide adequate accuracy while reducing computational cost. However, the computational cost of simulations in high definition (HD) prevents their use in complex environments and their low definition (LD) counterparts are used. In the case of very large scenarios, such as cities, this approach can still be computationally too demanding and combination with other estimation approaches is compulsory [10][9]. In Chapter 5 a novel approximation based on the combination of an in-house developed 3D Ray Launching code and a collaborative filtering technique of bi-dimensional calculation points is used to analyse the performance of wireless channels emulating AAL scenarios.

2.5.2.1 3D Ray Launching

As previously stated, the 3D ray launching code is based on Geometrical Optics (GO) and the Uniform Theory of Diffraction (UTD). The principle of the algorithm is that rays are launched in a solid angle with input angular resolution of parameters θ and Φ. Electromagnetic phenomena such as reflection, refraction, and diffraction are taken into account. We also consider the properties of all the obstacles within the environment considering the conductivity and permittivity of their different materials at the frequency of the system under analysis. It is worth noticing that a grid is defined in the space and the ray launching parameters are stored at each cuboid during the propagation of each ray. The configuration parameters of the system are frequency, power, gain, polarisation and directivity of the transceivers, bit rate, angular resolution of the launching and diffracted rays, and number of reflections. The received power is calculated with the sum of the electric field vectors inside each cuboid of the defined mesh. When a ray hits an obstacle, new reflected and transmitted rays are generated with new angles provided by Snell's law [22]. When a ray hits an edge, a new family of diffracted rays is generated, as we can see in Figure 2.3, with the diffraction coefficients $D^{\perp\parallel}$ given by the UTD, which are shown in (2.14), given by [102, 101].

$$D^{\perp\parallel} = \frac{-e^{(-j\pi/4)}}{2n\sqrt{2\pi k}} \left\{ \begin{array}{l} \cot\left(\frac{\pi+(\Phi_2-\Phi_1)}{2n}\right) F\left(kLa^+\left(\Phi_2-\Phi_1\right)\right) \\ +\cot\left(\frac{\pi-(\Phi_2-\Phi_1)}{2n}\right) F\left(kLa^-\left(\Phi_2-\Phi_1\right)\right) \\ +R_0^{\perp\parallel}\cot\left(\frac{\pi-(\Phi_2+\Phi_1)}{2n}\right) F\left(kLa^+\left(\Phi_2+\Phi_1\right)\right) \\ +R_n^{\perp\parallel}\cot\left(\frac{\pi+(\Phi_2+\Phi_1)}{2n}\right) F\left(kLa^+\left(\Phi_2+\Phi_1\right)\right) \end{array} \right\}$$

where $n\pi$ is the wedge angle, F, L and $a^{\mp}$ are defined in [102], and $R_{0,n}$ are the reflection coefficients for the appropriate polarisation for the 0 face or n face, respectively. The complete approach has been explained in detail in [9], and it has been used successfully in different complex indoor environments [147, 8, 5, 6].

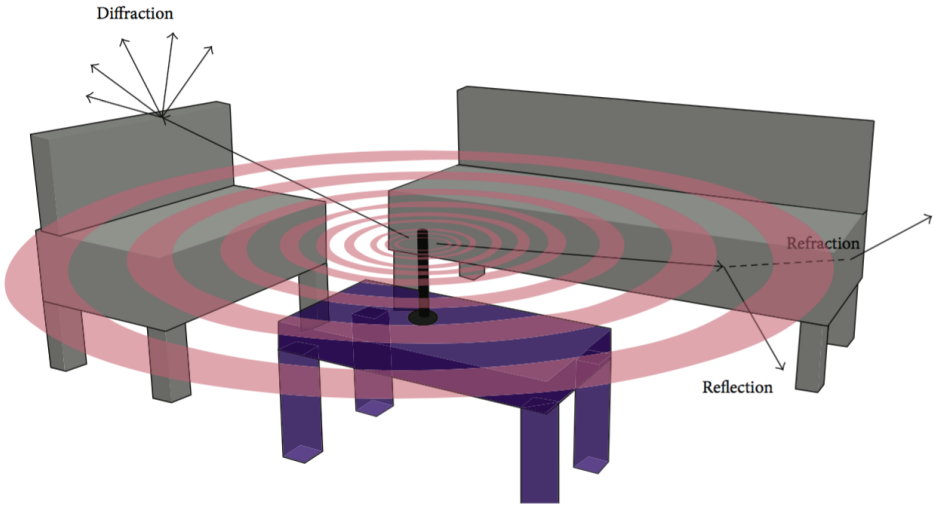


Figure 2.3: Simplified representation of the operation principle of the in-house 3D ray launching algorithm. Reprinted from [100].

CHAPTER 3

Contributions to Collaborative Filtering

The relevance of automatic recommender systems, such as Collaborative Filtering (CF), is increasing in the Internet and e-commerce due to the benefits they generate. However, to provide profitable recommendations, we require information about the users' behaviours and preferences. Therefore, to obtain such information we need to deal with non-response, which is a classical problem. In this Chapter, we introduce new imputation methods for CF-based recommender systems to deal with missing values and provide quality recommendations. Moreover, we conduct several experiments using novel metrics. The results show that our imputation approaches generate more accurate recommendations than well-known state-of-the-art methods.

Contents

3.1 Imputation Methods for Collaborative Filtering

The content of this Chapter is related with Collaborative Filtering and Statistical Inference (Chapter 2, Sections 2.1, 2.1.1 and 2.1.2). In the next sections, we propose both classical and new imputation methods suitable to be applied in CF-based recommender systems. The aim of these techniques is to deal with data sparsity while achieving accurate recommendations. Finally, we evaluate the predictions obtained with the presented approaches applied to three well-known CF databases in terms of accuracy and behavioural precision. The rest of the Chapter is organised as follows. Section 3.1.1 presents the most classical imputation methods in the recommender systems field. Section 3.2 proposes and describes new imputation approaches. Section 3.3 includes the experimental results, which are widely discussed on Section 3.4. Finally, Section 3.4.3 concludes the Chapter and provides directions for future research.

3.1.1 Classical Imputation Methods

In what follows, we describe well-known classical imputation methods that will be implemented in our system. First, let us set down some notation. The rating matrix is denoted by $\vec{R}_{n\times m}$, where n is the number of rows (users) and m is the number of columns (items) and with $r_{u,i}$ being the rating user u provided for item i. Therefore, the set of users is defined as $U = \{u_i, u_{i+1}, \cdots, u_n\}$ and the set of items $I = \{i_i, i_{i+1}, \cdots, i_m\}$. Moreover, $\vec{r_u}$ is the vector of all ratings provided by user u, $|\vec{r_u}|$ denotes its cardinality and $\overline{r}_u$ is the average of user's u ratings. The same notation will be used for items (each case will be distinguished).

Default voting. Given a value range defined by the ratings set $S = \{s_i, s_{i+1}, \cdots, s_p\}$, we select a value $s_i \in S$ and use equation (3.1) to fill the empty cells of $\vec{R}$. Therefore,

$$\forall\ r_{u,i} \in \vec{R} = \begin{cases} s_i, & \text{if } r_{u,i} = null \\ r_{u,i}, & \text{otherwise} \end{cases} \tag{3.1}$$

Item mean. First, $\forall i \in I$, we compute the mean of each item $\overline{r}_i$. Next, $\forall\ r_{u,i} \in \vec{R}$, we replace the null values of each item by their corresponding $\overline{r}_i$.

Weighted item mean. We use equation (3.2) to compute a weighted mean of the items, in which the most frequent ratings contribute more than others.

In order to ease computations, weights are normalised so that, $\sum_{j=1}^{m} w_{r_{i_j}} = 1$, $\forall\ \vec{r_i} \in \vec{R}$, where $w_{r_{i_j}}$ is the weight of the j^{th} element in $\vec{r_i}$. Thereafter, $\forall\ r_{u,i} \in \vec{R}$, we replace the null values of each item by their corresponding weighted item mean $\overline{r}_{i_w}$.

$$\overline{r}_{i_w} = \frac{\sum_{j=1}^{m} w_j r_{i_j}}{m} \tag{3.2}$$

where w_j is the weight of r_{i_j}.

Mean (item + user). The *mean (item + user)* is another well-known approach, in which an average between each corresponding user's mean $\overline{r}_u$ and each item's mean $\overline{r}_i$, is computed to fill the null values of $\vec{R}$, so that $r_{u,i} = (\overline{r}_u + \overline{r}_i)/2$.

Tendencies. The tendencies-based method [25], uses the vote tendency of users and items to weight their contribution when recommendations are computed. The tendency of a user is defined as the average difference between her ratings and the item mean:

$$t_u = \frac{\sum_{i \in \vec{r_u}} (r_{u,i} - \overline{r}_i)}{|\vec{r_u}|} \tag{3.3}$$

where $|\vec{r_u}|$ is the cardinality of the vector of all ratings provided by user u and $\overline{r}_i$ is the average rating of item i. To capture the tendency of an item, we compute it using the following equation:

$$t_i = \frac{\sum_{u \in \vec{r_i}} (r_{u,i} - \overline{r}_u)}{|\vec{r_i}|} \tag{3.4}$$

where $|\vec{r_i}|$ is the cardinality of item's i ratings vector and $\overline{r}_u$ is the average rating of user u. Therefore, the tendency of a user u_a is positive if u_a's ratings are above the item mean. In regard to items, an item has a positive tendency if it tends to be rated above the user mean. The recommendations are computed using a different equation in each of the four possible scenarios:

1. Both item and user have positive tendencies. In this scenario, we select the maximum value between $\overline{r}_u + t_i$ and $\overline{r}_i + t_u$:

$$p_{ui} = \max(\overline{r}_u + t_i, \overline{r}_i + t_u) \tag{3.5}$$

2. Both item and user have negative tendencies. In this case, we select the minimum value between both expressions:

$$p_{ui} = \min(\overline{r}_u + t_i, \overline{r}_i + t_u) \tag{3.6}$$

3. In the case that the tendencies of the user and the item are different, we compute the average of the tendencies. Hence, we modify the equation used in the original approach to ease computations:

$$p_{ui} = (\overline{r}_u + t_i + \overline{r}_i + t_u)/2 \tag{3.7}$$

4. Finally, if the tendencies are not a good indicator of user and item behaviours we use the *mean (item + user)* approach.

Gaussian noise addition. In this case, for each item $i \in I$, we compute the mean $\overline{r}_i$ and the standard deviation σ_i using equation (3.8). Next, we fill the null values of each item with values sampled from $\mathcal{N}(\overline{r}_i, \sigma_i))$. The higher the σ_i value, the greater the range of the generated values (*i.e.* it is more likely to generate values close to the boundaries of the value range). The same may be applied for users.

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{m}(r_{i_j} - \overline{r}_i)2}{m}} \tag{3.8}$$

Discrete uniform distribution. Likewise, we may use a discrete uniform distribution $\mathcal{U}(D)$, where D is the set of values existing in each item or user vector. For instance, given an item vector $\vec{r_i} = (1, 2, 2, 3, 3, 4, 5, 1)$, the set $D = \{1, 2, 2, 3, 3, 4, 5, 1\}$, so that $|\vec{r_i}| = |D|$, and the values sampled from $\mathcal{U}(D)$ will appear with the same probability distribution of D.

In addition to the previously stated imputation methods, there are more sophisticated approaches to deal with missing data, most of them related with distance/similarity calculations [7] (*e.g.* Euclidean distance, Pearson correlation or cosine similarity) and closest user imputation. The distance-based imputation procedure, detailed in Algorithm 3 is as follows: First, we compute the distance/similarity between a user u_i and the rest of users (Algorithm 3, *line 5*). Iteratively, the closest profile to u_i, namely $u_{closest}$, is selected to replace the null values of u_i (Algorithm 3, *line 6*). This process is performed from the closest user to the furthest, or until u_i's profile becomes completely filled. Note that the previous procedure is performed for each user of the database. These similarity computations may overwrite

the original matrix $\vec{R}$ or, otherwise, reflect changes in an auxiliary matrix $\vec{R}'$ (Algorithm 3, *line 7*). Therefore, in $\vec{R}'$, distances are computed without considering the updated vectors, thus using the original values. Note that although we use Euclidean distance, any other similarity metric could be used.

Algorithm 1 Similarity-based Imputation

```
1: function SIMILARITY-BASED IMPUTATION (DataSet R, SimilarityMetric SM, Boolean
   AuxiliarMatrix)
2:     Initialisations(R, SM);
3:     while (UnprocessedUsers) do
4:         u = SelectRandomUser(R);
5:         d_u = DistanceVectorUser(u, SM, R);
6:         r_u = UserImputation(r_u, d_u, R);
7:         if AuxiliarMatrix = true then
8:             UpdateData(R', r_u);
9:         else
10:            UpdateData(R, r_u);
11:        end if
12:    end while
13:    R_F = BuildFinalDataset(R,R',AuxiliarMatrix);
   return R_F                                   ▷ Completely Filled DataSet
14: end function
```

Euclidean distance between common-rated elements. In this method, for each pair of users u_a and u_b, we compute distances (Algorithm 3, *line 5*) only considering the elements in their corresponding common-rated item set, defined as $\vec{r_a} \cap \vec{r_b}$. For that reason, it is possible that many users have zero distance among them, although not being equal. This method utilises an auxiliary matrix $\vec{R}'$ to store the updated profiles.

Pearson Correlation. The Pearson correlation (PC) method [142] measures the extent to which two variables linearly relate with each other. For the user-based algorithm, the Pearson correlation between users u_a and u_b is

$$w_{a,b} = \frac{\sum_{i \in I}(r_{a,i} - \overline{r}_a)(r_{b,i} - \overline{r}_b)}{\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r}_a)^2}\sqrt{\sum_{i \in I}(r_{b,i} - \overline{r}_b)^2}} \tag{3.9}$$

where the $i \in I$ summations are over the items that both u_a and u_b have rated and $\overline{r}_a$ is the average rating of the co-rated items of the a^{th} user. This approach also reflects the updates in an auxiliary matrix $\vec{R}'$.

3.2 Our Proposal

An overview of our system is depicted in Figure 3.1. The system receives an input database with a set of parameters such as number of items, number of users, range of values, among others. Next, the system processes the data and fills the dataset according to an imputation method. Finally, the output is a filled dataset (*i.e.* a matrix that has not null values). In this article, we perform an extensive study of the recommendation's quality achieved by the output datasets (memory-based approach). Notwithstanding, such datasets could be also used as the input of model-based approaches (*e.g.* clustering-based methods, matrix factorization or other dimensionality reduction methods) to improve their outcomes.

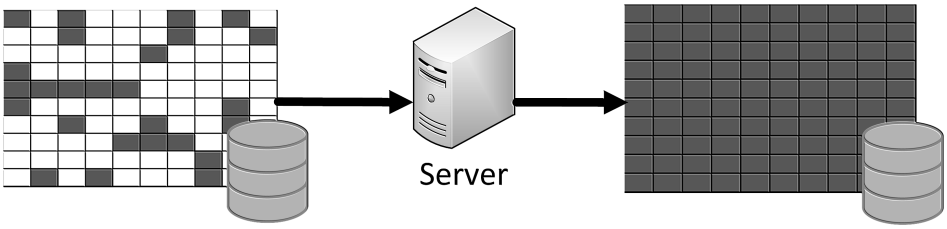


Figure 3.1: Overview of our system's architecture. The system (middle) receives an input database (left) and performs some computations. The result (right) is a dataset that has no empty cells. Adapted from [42].

In the next subsections, we propose several imputation methods to assign a value for the empty cells of a ratings matrix. If such matrix has *cold users* (*i.e.* users whose ratings vector is empty) or *cold items* (*i.e.* items, which have not been rated by any user) after applying an imputation method, we use the *default voting* imputation with the center of the value range to ensure that the matrix is completely filled.

3.2.1 Constant Euclidean Distance between Non-Rated Elements (CEDNE)

Classical Euclidean distance computation between two vectors requires them to have no empty values. Therefore, the values assigned to fill the matrix to allow distance calculations modifies the users' rating vectors. This modification may cause that users, which theoretically should not be neighbours became so, depending on the procedure. For instance, assuming that Table 3.1 has a value range $S = \{1, 2, 3, 4, 5\}$ and that we use the center value 3 to fill it, the rating $r_{c,3}$ will be closer to $r_{b,3}$ than $r_{a,3}$, the latter being a

real vote. Therefore, we may bias users' similarities depending on the imputation method. Users may also be displaced away if we use 0, negative or boundary values arbitrarily. As previously stated, one way to prevent this is to calculate the distances or similarities only between items that both users have rated. This methodology is well-known in CF (*e.g. Pearson correlation* or the *Euclidean distance between common-rated elements*). However, such methods have also drawbacks, for example, the distance between u_a and u_c, would be equal to 0 using the common-rated metric (*cf* Table 3.1). Moreover, the distance between u_a and u_b will be always higher than 0, which seems contradictory because these users are interested in more common items (*i.e.* compared with u_c), expressing similar tastes.

As a conclusion we need to face two problems. First, how to fill the empty values to compute distances without moving closer or further the users involved. Second, if we compute distances only between common rated elements, we are also being unfair. Moreover, to establish an appropriate metric for each pair of users is a difficult problem to optimise that changes according to data. Therefore, our aim is to perform distance computations between users in a way that avoids iniquities. For this purpose, we will use constant distance between the elements that a user has voted neither the other. Although this constant can be set as desired, we propose two ways to compute it:

Table 3.1: Example matrix. The values range between 1 and 5.

U \ I	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_a	-	1	2	2	2	-	-	1	2
u_b	1	2	4	2	1	-	4	1	2
u_c	2	-	-	2	-	1	-	-	2

CEDNE with constant weight. Given a ratings set S, we select the highest value, denoted as s_{max} and the lowest one, denoted as s_{min}. Next, we compute our constant weight value $w_c = (s_{max} - s_{min})/2$. Therefore, we use the half of the difference between the highest and the lowest possible rating values as the outcome, every time that we compute the distance between a rated element with a non-rated one.

CEDNE with center value. In this case, we select the center value of a ratings set S, denoted as s_{mid}. Next, the difference d between a real rating

$r_{i,j}$ and a null value is computed as $d = r_{i,j} - s_{mid}$. Therefore, all non-rated elements are temporally replaced by s_{mid} to compute distances.

3.2.2 Dynamic Imputation One by One

This method has two variants with different characteristics, with special remark to the computational cost.

Dynamic Euclidean distance. In this method, we compute the Euclidean distance between user u_i and the other users and we select the closest one, namely $u_{closest}$ that can provide new values to $\vec{r_i}$. Once $\vec{r_i}$ becomes updated, instead of selecting the next closest neighbour, we recompute distances between u_i and the rest of users. Therefore, distance computations are performed using the original matrix $\vec{R}$, in which changes are reflected. Following this scheme, each time that we update a user profile, we find the next $u_{closest}$ dynamically. However, we observed that selecting all the profitable elements of $\vec{r_{closest}}$ at each iteration resulted in biased profiles.

Dynamic Euclidean distance one by one. This approach was created to overcome the shortcomings of the *Dynamic Euclidean distance* method. The main difference between these approaches is the amount of data copied in each iteration. First, we select the item to be evaluated randomly at each iteration (Algorithm 2, *line 4*). Next, the users only copy the value of the item that is being evaluated, instead of all profitable values between them and their closest user (Algorithm 2, *line 8*). An example of the procedure is depicted in Fig. 3.2. In order to optimise the process, we only compute distances with users that have a real assessment on the evaluated item (Algorithm 2, *line 7*). Therefore, the computational cost decreases as the sparseness is increased. Due to the huge sparseness of the evaluated matrices, distances are computed only between few users and the cost of this approach is drastically reduced.

3.2.3 Binary Euclidean Distance between Common-Rated Elements

This method implements a binary scheme to compute the distance between users. For each pair of users u_a and u_b, we denote their common-rated item set as $\vec{r_a} \cap \vec{r_b}$ and their distance vector as $\vec{dist}_{a,b}$, where $|\vec{r_a} \cap \vec{r_b}| = |\vec{dist}_{a,b}|$. Given a ratings set S, we denote s_{min}, s_{mid} and s_{max} as the minimum, the center and the maximum possible values of a rating, respectively. Therefore, in our binary scheme, we define the positive interval as $pos = [s_{mid}, s_{max}]$ and

Algorithm 2 Dynamic One by One Imputation

function DYNAMIC ONE BY ONE (DataSet $\vec{R}$, SimilarityMetric SM)
  Initialisations($\vec{R}$, SM);
  while (UnprocessedItems) **do**
    i = SelectRandomItem($\vec{R}$);
    while (UnprocessedUsers) **do**
      u = SelectRandomUser($\vec{R}$);
      $\vec{d}_u$ = EfficientDistanceVectorUser(u, SM, $\vec{R}$);
      $r_{u,i}$ = ClosestUserRating(i, $\vec{d}_u$, $\vec{R}$);
      $\vec{r}_u$ = OneItemImputation($\vec{r}_u$,$r_{u,i}$);
    end while
    UpdateData($\vec{R}$, $\vec{r}_u$);
  end while
  $\vec{R}_F$ = BuildFinalDataset(R);
  return $\vec{R}_F$ ▷ Completely Filled DataSet
end function

U \ I	i_1	i_2	i_3	i_4	i_5
u_a	1	1	4	3	-
u_b	3	-	2	3	1
u_c	4	-	2	-	3
u_d	-	5	2	4	1
u_e	4	3	1	2	5

Figure 3.2: Example of the *Dynamic Euclidean distance one by one* imputation procedure. At this step, the item that is being evaluated is i_2 and the active user is u_c (both in blue). We determine that u_c's closest user with a non-zero value for i_2 is u_e and thus $r_{c,2} = r_{e,2}$. Once the matrix becomes updated, we select the next user, which has not rated i_2 and the procedure is repeated. Note that $r_{c,4}$ will be kept empty until i_4 becomes evaluated, in contrast to previously stated Euclidean imputation approaches.

the negative interval as $neg = [s_{min}, s_{mid}]$. Note that we consider the center value as both positive and negative, to increase the number of comparisons between users. Next, we are able to compute the values of $\vec{dist}_{a,b}$ as follows:

$$\forall\ i \in \vec{dist}_{a,b},\ dist_{a,b,i} = \begin{cases} 1, & \text{if } r_{a,i}, r_{b,i} \in pos \\ 1, & \text{if } r_{a,i}, r_{b,i} \in neg \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

Therefore, in the event that both users rate an item i with the same

behaviour (*i.e.* both positive or both negative), the i^{th} element of their corresponding distance vector is equal to 1 and, for the rest of cases, 0. Hence, the binary similarity between users u_a and u_b is computed as $bin_{a,b} = \sum_{i=1}^{|\vec{dist}_{a,b}|} dist_{a,b,i}$. Therefore, the higher the $bin_{a,b}$ value, the closer u_a and u_b are.

An example of this binary distance calculation is depicted in Figure 3.3. Note that this method stores the updated profiles in an auxiliary matrix $\vec{R}'$ and that the cost of computing binary distances is substantially lower than computing numerical distances. Moreover, this binary scheme could be also implemented considering more levels (*i.e.* three, four, or more depending on the value range). However, the more levels considered, the more difficult is to find coincidences if the matrix is sparse. For that reason, to determine an optimal relationship between the applicable levels and the sparseness of a dataset is an interesting study that will be left to future work.

U \ I	i_1	i_2	i_3	i_4	i_5
u_a	2	2	4	3	-
u_b	5	-	2	3	1
u_c	4	4	2	2	2
u_d	-	5	3	4	1
u_e	4	1	4	2	5

$$dist_{a,b} = (0,0,0,1,0) \longrightarrow bin_{a,b} = 1$$
$$dist_{a,c} = (0,0,0,1,0) \longrightarrow bin_{a,c} = 1$$
$$dist_{a,d} = (0,0,1,1,0) \longrightarrow bin_{a,d} = 2$$
$$dist_{a,e} = (0,1,1,1,0) \longrightarrow bin_{a,e} = 3$$

Figure 3.3: Example of binary distance calculations. The matrix range set is $S = \{1,2,3,4,5\}$. Therefore, we consider the interval [3-5] as positive votes and the interval [1-3] as negative ones. The vector $dist_{a,i}$ contains the number of common and not common behaviours between u_a and u_i. In this example, $bin_{a,e}$ achieves the highest distance and, hence, u_e is the closest user of u_a (in blue). Next, the empty cells of $\vec{r_a}$ will be filled using $\vec{r_e}$'s values. Thereafter, we will select u_a's next closest user and repeat the procedure until $\vec{r_a}$ becomes filled or all users become evaluated.

Finally, a hybrid approach that implements the *Dynamic Euclidean distance one by one* imputation method but uses binary Euclidean distance calculations as similarity metric has been also implemented. Table 3.2 sum-

Table 3.2: Imputation methods and their main characteristics. In regard to computational costs, n denotes de number of users, m is the number of items, and d, c, b and p are the cost of computing Euclidean distances, Euclidean distances only considering common-rated items, binary Euclidean distances and *Pearson correlation*, respectively, with a set of n users. Finally, $b*$ and $c*$ are the costs of computing distances only with the set of users that contain a valid value for the m^{th} item (*i.e.* the one that is being evaluated), and thus they are inversely proportional to the sparseness.

Method	Metric	Aux. Matrix	Cost
Pearson correlation	Correlation	yes	$O(np)$
Euc. dist. with common-rated items	Euc. dist. comm.	yes	$O(nc)$
Cedne with central value	Euc. distance	yes	$O(nd)$
Cedne with independent value	Euc. dist.	yes	$O(nd)$
Binary Euc. dist. with common-rated items	Binary Euc. dist.	yes	$O(nb)$
Dynamic Binary Euc. dist. one by one	Binary Euc. dist.	no	$O(nmb*)$
Dynamic Euc. dist. one by one	Euc. dist. comm.	no	$O(nmc*)$
Uniform distribution of items	Uniform dist.	no	$O(m)$
Gaussian distribution of items	Gaussian dist.	no	$O(m)$
Gaussian distribution of users	Gaussian dist.	no	$O(n)$
Weighted item mean	Weighted mean	no	$O(m)$
Item mean	Mean	no	$O(m)$
Item + User mean	Mean	no	$O(nm)$
Tendencies-based imputation method	Tendencies	no	$O(nm)$
Default voting (central value)	Default voting	no	$O(1)$

marises the approaches that have been implemented and their main characteristics.

3.3 Experimental Setup

In this Section, we describe the experimental setup of the proposed imputation methods. We use the well-known leave-one-out experiment, in which the values of the closest neighbour of the active user u_a are selected as the recommended values, to evaluate the accuracy of recommendations.

We compute the error between the original dataset values and the recommended values using the mean absolute error (MAE) defined in (5.1) and that we repeat here for the sake of clarity:

$$MAE = \frac{\sum_{i=1}^{n} |p_i - r_i|}{n} \tag{3.11}$$

where n is the number of recommended elements, p_i is the predicted value over the element i, and r_i is the real value of i. The lower the MAE value,

the better the accuracy of the recommendations. Moreover, we use the behavioural precision metrics (*i.e.* both the binary and the 4L behavioural precision) defined in Section 2.4.1 to obtain more information about the users' interests.

With the aim to assess the quality of our methods, we use three well-known CF datasets with different characteristics, namely Book-Crossing (least 20), Jester and Movielens 100k. Table 2.3 in Chapter 2 summarises their main characteristics.

We have created two sets of experiments. In the first one, we randomly select the 80% of the items for training and the remaining 20% conforms the recommendation set. The second test includes the 90% of the items in the training set, leaving the remaining 10% for recommendations. The recommendations are computed only for the original ratings of each user. The tests have been repeated 100 times and in each test, the elements of the training set and the recommendation set have been selected randomly. Table 3.3, Table 3.4 and Table 3.5 show, for each dataset, the average outcomes of these experiments. In order to ease the comparisons, we also depicted the outcomes of behavioural precision obtained for Movielens, Jester and Book-Crossing datasets in Figures 3.4, 3.5 and 3.6, respectively. For the sake of clarity, such figures represent the outcomes obtained using a training set of the 90%, since they are similar (*i.e.* although slightly better) to the ones obtained using a training set of the 80%. In the case of the *default voting* method, we have selected the central value of each dataset's ratings set. Moreover, we discarded using *default voting* with the maximum or the minimum value of the ratings set because the bias introduced resulted in bad quality recommendations.

3.4 Discussion

In this section, we discuss the outcomes of the imputation methods presented in Sections 3.1.1 and 3.2 for each evaluated dataset. In Section 3.4.1, we review the accuracy of the recommendations in regard to the MAE metric. Finally, we analyse the results of behavioural precision in Section 3.4.2.

3.4.1 Accuracy of the Recommendations

In Table 3.3 column "MAE", we can observe the accuracy obtained by each method concerning Movielens 100k dataset. The MAE values decrease when the training set is increased, with the exception of the *Gaussian distribution of users* method. In such case, the values used to fill the matrix have a ran-

Table 3.3: Results of the Movielens 100k dataset concerning recommendation accuracy and behavioural precision.

Movielens 100k

Method	Training set 80 % / Recommendation set 20%						Training set 90 % / Recommendation set 10%					
	MAE	Bin. precision		Four-level precision			MAE	Bin. precision		Four-level precision		
	%	Match	Error	Match	S. Match	S. Error	%	Match	Error	Match	S. Match	S. Error
Pearson Correlation	23.008	84.689	15.310	66.403	24.277	7.387	1.932	22.811	85.015	14.884	67.152	23.517	7.435	1.894
Euc. dist. with common-rated items	23.259	84.948	15.052	65.624	25.655	6.984	1.738	23.150	84.729	15.271	65.780	25.133	7.198	1.889
Cedne with central value	21.782	87.172	12.828	65.434	27.848	5.887	0.832	21.635	86.783	13.217	65.696	27.420	5.986	0.898
Cedne with independent value	22.760	85.612	14.388	65.941	25.514	6.707	1.837	22.671	85.410	14.590	66.168	25.277	6.738	1.817
Bin. Euc. dist. with common-rated items	21.295	86.931	13.069	66.851	25.575	6.166	1.409	21.221	86.578	13.422	67.151	25.210	6.245	1.394
Dynamic Bin. Euc. dist. one by one	17.539	89.364	10.636	73.601	20.356	5.000	1.044	17.365	89.353	10.647	73.342	20.567	4.994	1.097
Dynamic Euc. dist. one by one	17.599	88.256	11.744	73.724	19.150	5.623	1.503	17.291	88.074	11.926	73.459	19.103	5.750	1.688
Uniform distribution of items	25.209	82.370	17.630	61.907	27.126	8.516	2.450	24.738	83.293	16.707	62.890	26.795	8.118	2.197
Gaussian distribution of items	25.456	72.538	27.462	45.324	36.776	14.221	3.680	25.285	73.811	26.189	47.169	35.688	13.602	3.540
Gaussian distribution of users	23.146	75.544	24.456	38.901	47.738	12.165	1.196	23.308	76.130	23.870	39.776	47.131	11.875	1.219
Weighted item mean	20.083	81.836	18.164	48.101	43.939	7.128	0.833	20.047	81.437	18.563	47.995	43.957	7.247	0.801
Item mean	20.116	82.293	17.707	40.962	52.493	6.235	0.311	20.105	81.834	18.166	41.516	51.522	6.616	0.347
Item + User mean	19.627	85.218	14.782	38.127	57.520	4.241	0.112	19.600	85.149	14.851	37.690	57.919	4.283	0.108
Tendencies-based imputation method	18.483	81.615	18.385	46.836	46.316	6.454	0.394	18.275	81.655	18.345	47.300	46.018	6.276	0.406
Default Voting (central value)	23.673	98.320	1.680	45.749	53.287	0.824	0.140	23.605	98.361	1.639	46.298	52.771	0.797	0.134

Table 3.4: Results of the Jester dataset concerning recommendation accuracy and behavioural precision.

Jester Dataset																
	Training set 80 % / Recommendation set 20%								Training set 90 % / Recommendation set 10%							
	MAE	Bin. precision		Four-level precision			MAE	Bin. precision		Four-level precision						
Method	%	Match	Error	Match	S. Match	S. Error	Error	%	Match	Error	Match	S. Match	S. Error	Error		
Pearson Correlation	18.901	66.375	33.624	41.393	44.839	12.112	1.654	18.849	66.456	33.543	41.631	44.762	11.992	1.612		
Euc. dist. with common-rated items	18.238	69.336	30.664	44.571	42.983	10.797	1.649	18.132	69.286	30.714	44.490	43.370	10.660	1.480		
Cedne with central value	18.567	68.448	31.552	43.334	44.030	11.087	1.550	18.199	69.136	30.864	43.753	44.240	10.625	1.383		
Cedne with independent value	19.363	67.902	32.098	42.416	43.652	11.753	2.179	19.090	68.485	31.515	42.777	43.824	11.379	2.020		
Bin. Euc. dist. with common-rated items	14.034	83.991	16.009	58.728	34.152	5.958	1.163	13.333	84.495	15.505	60.388	32.965	5.608	1.038		
Dynamic Bin. Euc. dist. one by one	14.085	84.704	15.296	58.553	34.480	5.793	1.174	13.773	84.932	15.068	59.284	33.955	5.642	1.118		
Dynamic Euc. dist. one by one	16.332	73.322	26.678	49.510	40.328	8.899	1.263	16.007	73.566	26.434	49.887	40.474	8.538	1.101		
Uniform distribution of items	19.455	67.040	32.960	40.830	45.389	12.088	1.693	19.231	66.934	33.066	40.950	45.699	11.859	1.492		
Gaussian distribution of items	19.102	67.822	32.178	41.546	45.266	11.605	1.584	19.344	66.377	33.623	40.639	45.838	12.047	1.476		
Gaussian distribution of users	18.690	68.253	31.747	42.535	44.881	11.128	1.456	18.126	69.126	30.874	43.325	44.948	10.471	1.256		
Weighted item mean	18.754	68.693	31.307	43.131	43.966	11.217	1.686	18.563	67.665	32.335	42.647	43.944	11.587	1.822		
Item mean	19.019	67.520	32.480	42.514	44.079	11.672	1.735	18.701	68.131	31.869	42.995	44.053	11.380	1.573		
Item + User mean	18.706	69.005	30.995	43.126	44.252	10.897	1.725	18.366	69.138	30.862	43.526	44.490	10.501	1.482		
Tendencies-based imputation method	18.383	69.046	30.954	43.722	43.991	10.544	1.743	18.087	69.759	30.241	44.239	43.903	10.205	1.652		
Default Voting (central value)	19.200	75.294	24.706	48.019	40.759	9.424	1.798	19.012	76.516	23.484	48.493	40.692	9.083	1.732		

Table 3.5: Results of the Book-Crossing dataset concerning recommendation accuracy and behavioural precision.

Book-Crossing

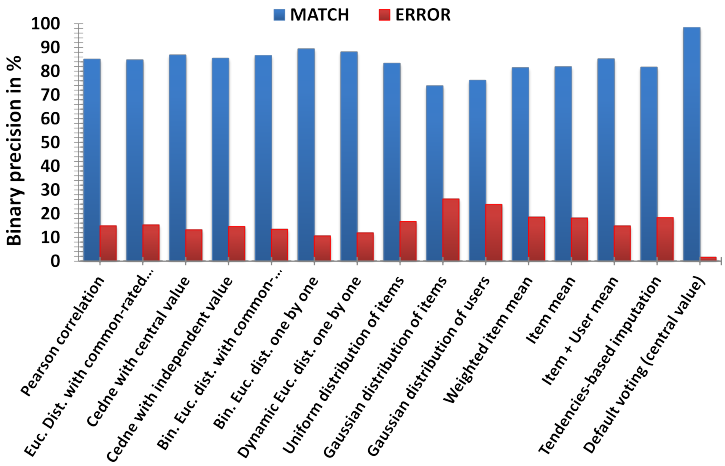
Method	Training set 80 % / Recommendation set 20%						Training set 90 % / Recommendation set 10%					
	MAE	Bin. precision		Four-level precision			MAE	Bin. precision		Four-level precision		
	%	Match	Error	Match	S. Match	S. Error	%	Match	Error	Match	S. Match	S. Error	Error
Pearson Correlation	9.826	89.339	10.660	75.261	16.810	6.966	0.960	9.786	89.346	10.653	75.369	16.801	6.873	0.956
Euc. dist. with common-rated items	9.886	87.913	12.087	74.081	17.134	7.606	1.179	9.870	88.022	11.978	74.053	17.283	7.482	1.182
Cedne with central value	9.424	89.551	10.449	75.976	16.376	6.699	0.949	9.420	89.478	10.522	75.973	16.281	6.750	0.996
Cedne with independent value	10.596	88.502	11.498	73.087	18.318	7.407	1.188	10.580	88.546	11.454	73.099	18.460	7.341	1.100
Bin. Euc. dist. with common-rated items	10.650	88.431	11.569	72.820	18.595	7.402	1.183	10.491	88.681	11.319	73.203	18.421	7.272	1.104
Dynamic Bin. Euc. dist. one by one	0.614	99.480	0.520	98.510	1.101	0.337	0.052	0.591	99.474	0.526	98.461	1.148	0.348	0.044
Dynamic Euc. dist. one by one	0.414	99.556	0.444	98.988	0.687	0.335	0.032	0.414	99.586	0.403	98.991	0.718	0.260	0.031
Uniform distribution of items	9.662	89.207	10.793	75.496	16.477	7.033	0.994	9.604	89.505	10.495	75.776	16.374	6.864	0.986
Gaussian distribution of items	10.052	90.011	9.989	72.128	22.114	5.084	0.674	10.024	90.092	9.908	72.312	21.958	5.043	0.688
Gaussian distribution of users	13.298	87.528	12.472	61.939	31.555	5.801	0.705	13.283	87.345	12.655	61.634	31.726	5.908	0.732
Weighted item mean	7.691	92.636	7.364	76.936	20.197	2.566	0.301	7.634	92.657	7.343	76.866	20.187	2.667	0.280
Item mean	7.911	92.550	7.450	75.203	22.455	2.215	0.127	7.889	92.631	7.369	75.345	22.364	2.167	0.124
Item + User mean	9.890	88.152	11.848	68.968	29.133	1.816	0.083	9.920	88.183	11.817	68.963	29.135	1.819	0.083
Tendencies-based imputation method	7.309	92.976	7.024	77.903	20.562	1.451	0.084	7.287	92.962	7.038	77.932	20.539	1.452	0.077
Default Voting (central value)	27.765	99.962	0.038	37.300	62.674	0.025	0.001	27.724	99.978	0.022	37.415	62.569	0.015	0.000

dom component and thus, the recommendations may be affected. We also observe that the worst accuracy is obtained by random distribution methods, the *default voting* method and the *Euclidean distance with common-rated items*. The *default voting* method is invariant according to the amount of information used and thus, it always gives the same value whatever the size of the training set. In regard to PC and *Euclidean distance with common-rated items*, data sparseness has a negative impact on this methods because it hinders the creation of neighbourhoods using only common assessments. Moreover, this fact affects correlation-based measurements [149][47], especially for computations over raw data. We also note that the *CEDNE with central value* performs better than *CEDNE with independent value*, meaning that such independent value increases the noise when distances are computed. Concerning mean metrics, the *item + user mean* seems to obtain better results than classical distance methods, followed by the *weighted item mean* method. Interestingly enough, the *Tendencies imputation* method achieves the third position in regard to accuracy, only outperformed by the *Dynamic Euclidean distance one by one* method and its binary variant. Therefore, the best methods are these that introduce less noise at each step, which seems to be relevant when the dataset is very sparse. It is worth to note that, although it has the same drawbacks than PC, the *binary Euclidean distance with common-rated items* method performs better than the rest of distance-based methods, closely followed by the *CEDNE with central value* approach. In regard to Jester dataset, Table 3.4 shows the outcomes obtained by the implemented approaches. The *Gaussian distribution of items* obtains less accuracy when the training set is increased, following a similar behaviour than the *Gaussian distribution of users* in the Movielens 100k dataset. However, the rest of results are better when the training set is increased. The worst results are achieved by random distribution methods, the *default voting* and the *CEDNE with independent value*. Next, mean-based methods obtain the best results, only outperformed by the *Tendencies-based* method and the methods proposed in Section 3.2. More concretely, the best methods are binary-based ones, followed by the *Dynamic Euclidean distance one by one method*. It is also worth mentioning that in this case, CEDNE approaches obtain better results compared with the ones achieved with Movielens 100k. Therefore, we may conclude that distance-based methods perform better when the matrix is not sparse. Note that metrics based in common-rated items perform better when there is enough information, especially in the case of binary approaches, which outperform the rest of methods. Moreover, such distance metric is further effective as the granularity of the value range is increased, since the preferences/behaviours are more easily iden-

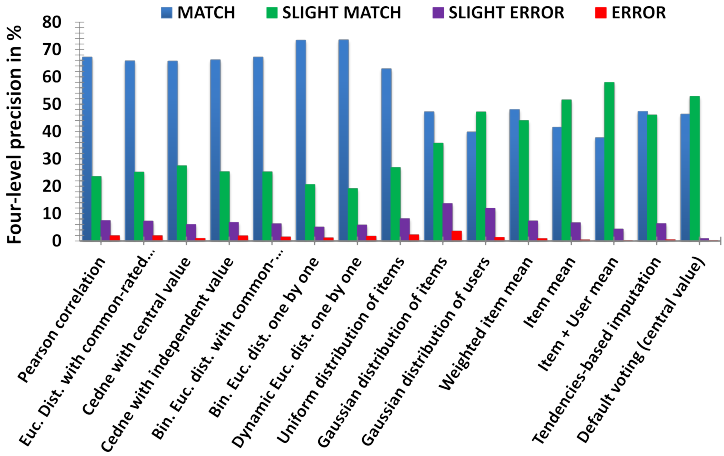
tifiable. Book-Crossing is the most sparse dataset among the ones being evaluated and its outcomes (*cf* Table 3.5) are similar than these obtained by Movielens 100k. In this case, random-based distributions and distance computations based on common-rated items obtain the worst results. Mean-based methods along with the *Tendencies-based imputation* method obtain better results than the rest of distance-based metrics except for the *Dynamic Euclidean distance one by one* method and its binary approach. However, it is worth noticing that the *item + user mean* obtains worse results when compared with the rest of mean-based approaches. This situation only occurs in the case of Book-Crossing dataset. Moreover, we observe that the *default voting* method obtains an acutely negative MAE. The reason for such bad outcomes is the percentage of extreme values present in the dataset. In other words, users tend to evaluate items using extremely high or extremely low values. Therefore, neighbourhoods consist of users with quite similar tastes and far from the average values. Finally, we observe that the results obtained by the *Dynamic Euclidean distance one by one* and its binary variant are the best (as previously occurred with Movielens 100k) methods. Moreover, the MAE obtained by such methods is approximately 20 times better than the one obtained by mean-based methods. Therefore, our *Dynamic Euclidean distance one by one method* groups similar users in such an efficient way that, along with the huge sparseness value of Book-Crossing dataset, allows the creation of extremely similar user profiles.

3.4.2 Behavioural Precision

In the case of Movielens 100k, Table 3.3 and Figures 3.4(a) and 3.4(b) show the behavioural precision obtained by the implemented methods. We may appreciate that MAE is related to the binary behavioural precision of the methods. However, when we analyse the precision using four levels, we obtain information that could not be discovered using only accuracy metrics. The random-based distributions obtain the worst binary behavioural results, followed by the mean-based methods, including the *Tendencies-based imputation* method. The distance-based methods obtain the best behavioural precision (*i.e.* both binary and 4L) than the rest of methods. This outcomes show that, although the MAE obtained is slightly worse than, for instance, mean-based methods, the behaviour of the users is captured in a more precise way (*i.e.* the recommendations fall in the same behavioural level with higher probability). We may also observe that the *uniform distribution of the items* achieves a behavioural precision similar to the one obtained by distance-based metrics. This occurs because the values used to fill the ma-



(a) Binary precision - Movielens 100k

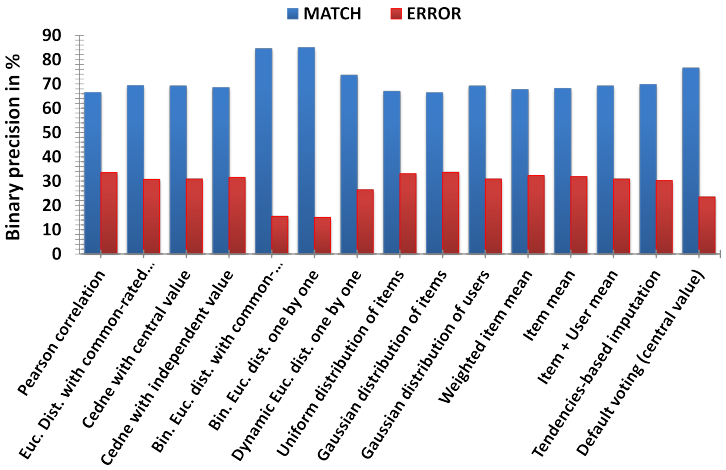


(b) Four level precision - Movielens 100k

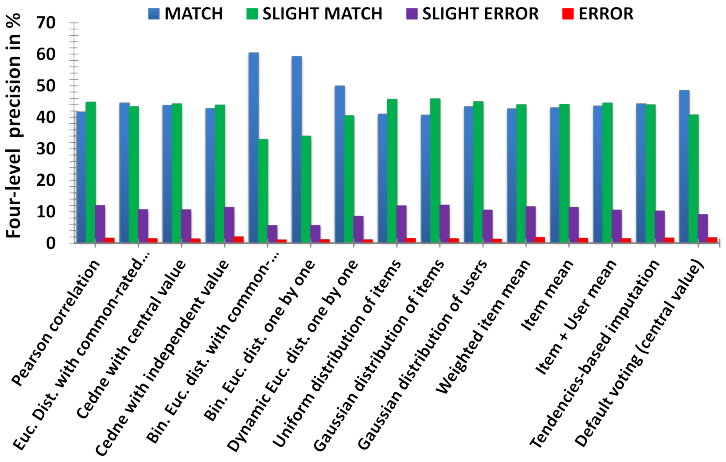
Figure 3.4: Behavioural precision obtained by the proposed methods in Movielens 100k dataset. Reprinted from [42].

trix are computed according to the number of times that an assessment is repeated in each item profile. Therefore, the more extreme values, the more probable is that the item retains its vote tendency. Note that the best binary precision is achieved by the *default voting* method because we use the central value as an "always positive" behaviour. Therefore, the percentage of binary behavioural precision in the case of our *default voting* approach reflects the amount of central value ratings in the dataset. However, the 4L precision of this method (*cf* Table 3.3) is similar to the one obtained by mean-based methods. Overall, the best methods are the *Dynamic Euclidean distance*

one by one and its binary alternative, outperforming the rest of methods in 4L precision.



(a) Binary precision - Jester



(b) Four level precision - Jester

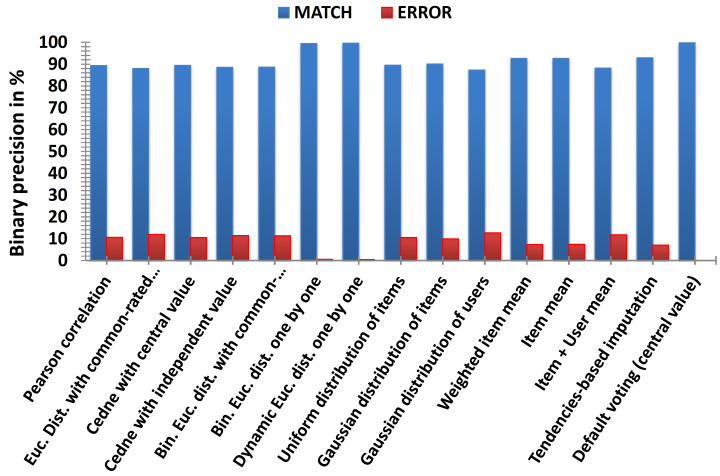
Figure 3.5: Behavioural precision obtained by the proposed methods in Jester dataset. Reprinted from [42].

We have depicted in Table 3.4 and in Figures 3.5(a) and 3.5(b) the outcomes of the Jester dataset concerning behavioural precision. In this case, the behavioural precision obtained by the methods is very similar except for the *default voting*, the binary-based distances and the *Dynamic Euclidean distance one by one* approach. As previously stated, *default voting* uses the central value, which results in an advantage over other competitors. However, we may observe that the outcomes of this approach in regard to

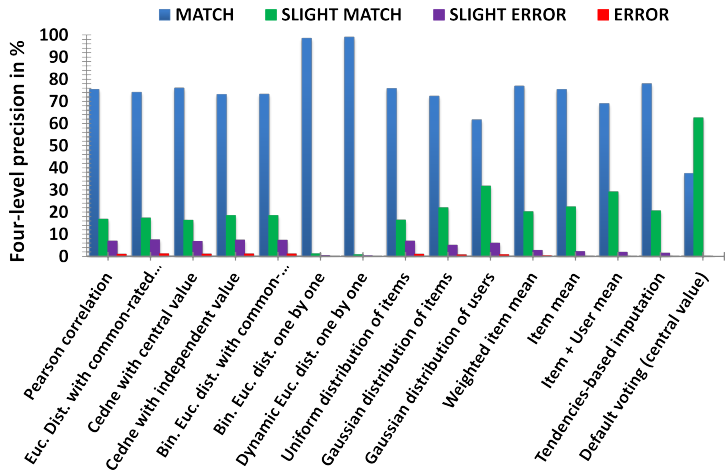
the 4L precision also show higher quality than the majority of methods. This could be explained observing the value range of Jester, which is significantly wider than the other two evaluated datasets. Therefore, we may conclude than there are less extreme values in this dataset, which implies better behavioural results for non-distance computations. Such wide-range also affects distance-based metrics, especially in the case of Euclidean-based distances, which obtain less behavioural precision in Jester dataset than in Movielens 100k. Overall, the best precision is obtained by the binary-based methods, which are approximately a 10% more precise than the next best method (*cf* Table 3.4).

Despite being the most sparse dataset, Book-Crossing obtains the best behavioural precision (*cf* Figures 3.6(a) and 3.6(b)). As previously stated in Section 3.4.1, this dataset contains a huge percentage of extreme values. Therefore, if we consider such ratings and the sparseness of the dataset, we may conclude that the methods will introduce less noise in the dataset. We may observe that mean-based methods obtain slightly better behavioural precision than the distance-based methods. Concerning the *default voting* method, the 4L precision is, as expected, the worst compared with the rest of methods. Finally, the best precision is obtained by the *Dynamic Euclidean distance one by one* method and its binary variant, which overwhelmingly outperform the rest of approaches.

After analysing the outcomes observed in Tables 3.3, 3.4 and 3.5, we may conclude that the best overall method is the *Dynamic binary Euclidean distance one by one* approach, because precisely captures the behaviour of users when there is enough data and also faces sparseness in a very efficient way. In average, binary-based methods are more precise when there is enough data. We may also note that the *Dynamic Euclidean distance one by one* method has the best performance when the dataset is sparse. In regard to the rest of distance-based methods, the *CEDNE with central value* obtains the best accuracy and behavioural precision. In contrast, methods that only consider common-rated ratings (*i.e.* PC and *Euclidean distance with common-rated items*) achieve better performance only when the dataset is not sparse. Finally, the *Tendencies-based* method achieves the best outcomes among approaches that are not based in distance computations, considering both accuracy and behavioural precision. To conclude, we have observed that the MAE results do not capture the quality of the recommendations in a precise way, since the best MAE was not always obtained by the most precise method. Therefore, our behavioural precision metrics were necessary to perform a deeper study and a qualitative analysis of the datasets and the outcomes of the methods.



(a) Binary precision - BX



(b) Four level precision - BX

Figure 3.6: Behavioural precision obtained by the proposed methods in Book-Crossing dataset. Reprinted from [42].

3.4.3 Conclusions

Although there are several well-known methods in the literature to deal with missing data in medical and statistical surveys, non-response is still an open problem, since there is not an optimal solution that satisfies all constraints [145][144]. In this chapter, we have proposed classical and new imputation methods to deal with incomplete data in CF datasets, which have specific characteristics such as a high dimensionality and a high percentage of null

values. Moreover, we have performed experiments with three well-known datasets, to evaluate the quality, behavioural precision and usefulness of the presented methods. The results show that, although the best outcomes are achieved by binary-based Euclidean distance methods and by the *Dynamic Euclidean distance one by one* approach, we may obtain different results depending on the characteristics of data.

Chapter 4

Contributions to Privacy-Preserving Collaborative Filtering

As previously stated in Chapter 2, privacy is a fundamental right and privacy protection is a hot topic to which many research efforts have been devoted from a variety of fields [181, 173, 196, 108]. Thus, among all the open problems of CF, in this chapter we concentrate on the protection of the privacy of users involved in CF processes. This chapter presents several PPCF methods whose aim is to achieve quality recommendations while preserving the user's privacy. In what follows, we describe three PPCF centralised methods, namely Gaussian noise addition, Maximum Distance to Average Vector and Variable Maximum Distance to Average Vector in Sections 4.1, 4.2 and 4.3, respectively. Finally, in Sections 4.4 and 4.5, we compare the aforementioned approaches to ascertain which one achieves the best outcomes in regard to quality of the recommendations and privacy protection.

Contents

4.1 Gaussian Noise Addition

As previously stated in Section 2.2.1, there are several ways to hide a number such as randomised perturbation. The aim of this procedure is to obfuscate data in such a way that certain computations can be performed while preserving users' privacy. In this sense, we may perturb the values of a dataset using a Gaussian distribution with zero mean and standard deviation σ (*i.e.* $\mathcal{N}(0, \sigma)$). The higher the σ value, the greater the range of the generated values (*i.e.* it is more likely to generate values close to the boundaries of the value range). The procedure is described as follows:

- Once the matrix is completely filled (*i.e.* using imputation methods such as the ones presented in Chapter 3), we compute the z-scores of each column (item) of the dataset to standardise the data (*i.e.* give the same statistical weight to each dimension), using the following expression
$$z - score = \frac{x_i - \mu}{\sigma} \tag{4.1}$$
where x_i is the i-th value of item x and μ and σ are the mean and the standard deviation of item x, respectively. In this way, the mean and the standard deviation of the transformed item are 0 and 1, respectively.
- Next, we apply Gaussian noise addition (GNA).
- Finally, we de-standardise the data to compute the recommendations.

It is worth to mention that we selected GNA because it has proven to better preserve privacy by adding much less quantity of noise in large high-dimensional datasets than other methods, such as uniform noise addition [3].

4.2 Maximum Distance to Average Vector

In this section, we propose a technique for PPCF based on microaggregation, which provides accurate recommendations estimated from perturbed data while guaranteeing user k-anonymity.

In [55] a simple microaggregation heuristic named Maximum Distance to Average Vector (MDAV) is described, in which all clusters have exactly k records, except the last one, which might have between k and $2k-1$ records. In such approach, the fact that k-anonymity only protects against identity disclosure (but not against attribute disclosure) is not a problem, because

all attributes are regarded as $\mathcal{QI}$s. Therefore, all attributes are modified by microaggregation to reach k-anonymity; hence, protection against attribute disclosure is also offered (in case preferences on some items are considered confidential). We use MDAV because of its simplicity, although it has the limitation of using clusters of fixed size k.

In fact, in our PPCF application, attributes are the preferences of users on different items, and each record collects the preferences of a particular user. We consider all attributes to be $\mathcal{QI}$s, because a large number of preferences has been shown to lead to user re-identification (*e.g.* [117] identified Netflix users based on their preferences).

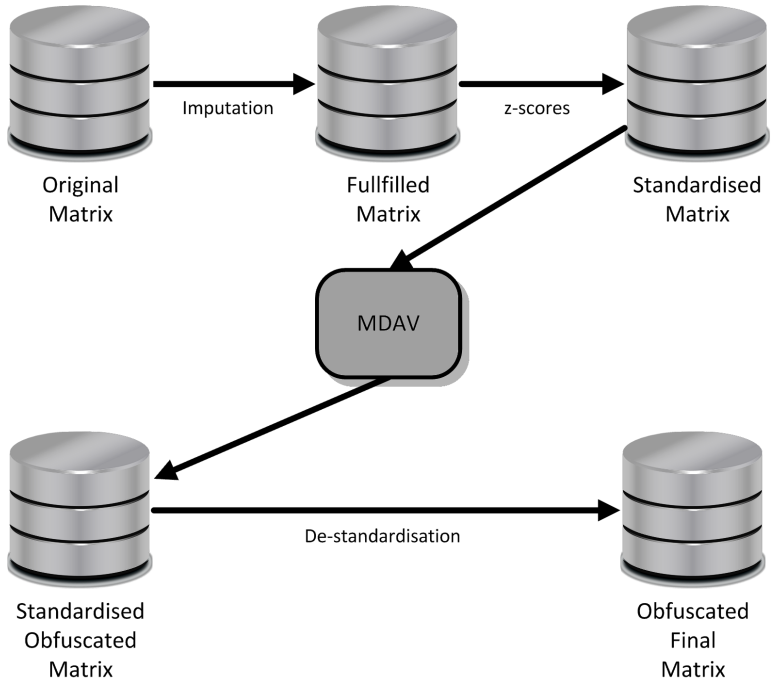


Figure 4.1: Graphical scheme of the MDAV method. Adapted from [35].

Our scheme, illustrated in Figure 4.1, works as follows:

1. Ensure that the dataset contains no missing values for any attribute in any record. This is necessary to compute the Euclidean distance between records. Imputation methods (*e.g.* the ones presented in Chapter 3) or non-personalised values can be utilised to fill the empty fields of the dataset matrix. For our experiments we have used default center value imputation (see Section 3.1.1).

2. Once the matrix is completely filled, we compute the z-scores of each column (item) of the dataset using Equation (4.1) to standardise the data.

3. Next, we are able to apply the MDAV clustering. Users will be grouped into a number of clusters, with each cluster c_i consisting of the k most similar users, according to the Euclidean distance, where k denotes the cluster cardinality. By selecting the most similar users, we maximise the cluster homogeneity and we therefore reduce the information loss. Once the cluster relationships are established, the mean values of each c_i, denoted as $\overline{c}_i$, are computed. Afterwards, each value of c_i is replaced by the corresponding $\overline{c}_i$.

4. The MDAV clustering process will result in a new dataset, in which members of the same cluster c_i will have the same profiles and become indistinguishable within their group. Therefore, after applying MDAV, this dataset will satisfy k-anonymity.

5. Finally, to make predictions, the results are de-standardised to obtain the final obfuscated dataset.

The original MDAV algorithm specifies that, if at the end of the clustering process there are p records between k and $2k-1$ ($k \leq p < 2k$) that do not belong to any cluster, they should form a final cluster c_f themselves. In our approach, we manage the unassigned records more accurately. First, we compute the distance between every record in c_f and $\overline{c}_f$. Next, we compare the distance between each member of c_f and all clusters. If more than half of the records in c_f are closer to $\overline{c}_f$ than to any other cluster, we form a final cluster with the c_f elements. Otherwise, each record is added to the closest cluster among those already formed.

4.3 Variable Maximum Distance to Average Vector

In this section, we present a PPCF approach based on variable-sized group microaggregation, which provides k-anonymity to users involved. The general scheme of our system is illustrated in Figure 4.2. First, we fill the matrix to compute Euclidean distances between records. We may use the default value or the overall mean, among others (see Chapter 3), to fill/impute the empty fields of the dataset matrix. Next, we compute the z-scores of each column (item) of the dataset to standardise the data, using Equation (4.1). Subsequently, we apply the Variable-size Maximum Distance to Average Vector (V-MDAV) method to sanitise the data. Thereafter, users will

be grouped into n clusters and each cluster c_i will be conformed by the k most similar users (*i.e.* according to Euclidean distance). The cardinality of each group, denoted by k, will be computed using a heuristic that will be discussed later. Next, each value of c_i is replaced by its corresponding $\overline{c}_i$. The V-MDAV clustering process will result in a new dataset that satisfies, at least, k-anonymity. Finally, to make predictions, the results are de-standardised to obtain the final obfuscated dataset.

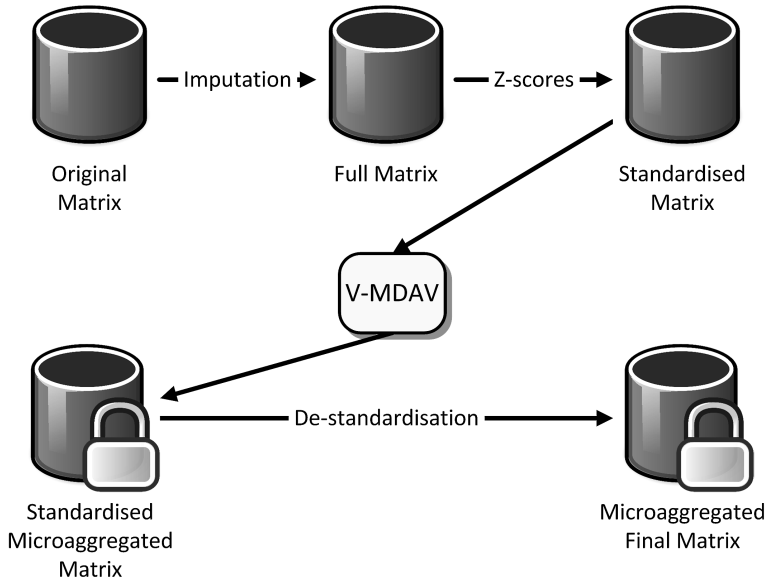


Figure 4.2: Step by step of our proposal. Adapted from [41].

As previously stated in Section 2.2.1, when microaggregation is applied to the projection of records on their $\mathcal{QI}$ attributes, the resulting dataset is k-anonymous, that is, to an intruder each record in the dataset is indistinguishable within a group of k records in terms of the $\mathcal{QI}$s. V-MDAV [156] intends to overcome the limitations of the methods which compute groups of fixed-size cardinality (*e.g.* MDAV) by computing a variable size k-partition with a similar computational cost. Note that there are situations, in which fixed-sized heuristics yield a k-partition far from the optimal one [51]. Therefore, such heuristics lack in flexibility for adapting the group size to the distribution of the records in the data set, which may result in poor within-group homogeneity.

The V-MDAV approach follows a similar strategy to MDAV and its complexity is almost the same except for two main differences:

1. MDAV computes a centroid in each iteration. V-MDAV only computes

Algorithm 3 Variable group size algorithm

```
1: function V-MDAV(DataSet D, Integer k)
2:     ComputeDistancesMatrix(D);
3:     C = ComputeCentroidOfDataSet(D);
4:     while (RecordsToAssign> k − 1) do
5:         e = SelectTheMostDistantRecordToCentroid (D,C);
6:         g_i = BuildGroupFromRecord(e,D,k);
7:         g_i = ExtendTheGroup(g_i,D,k);
8:     end while
9:     g_1 ... g_s = RemainingUnassignedRecords(D,g_1 ... g_s);
10:    M = BuildMicroaggregatedDataSet(D,g_1 ... g_s);
11: return M                                  ▷ microaggregatedSet M
12: end function
```

the dataset centroid at the beginning. This results in a computational time improvement.

2. MDAV does not build a matrix of distances; on the contrary, it computes distances as many times as needed. Thus, V-MDAV is faster.

Moreover, once a fixed-cardinality group is formed (Algorithm 3, *line 6*) our proposal applies a heuristic (Algorithm 3, *line 7*), which allows it to fit the dataset distribution better by generating variable-sized groups. The expansion of the group is computed as follows:

Given a group g with p records, the record e_{min} among unassigned records outside g nearest to g and the minimum distance d_{in} between e_{min} and g are defined by Equation (4.2) and (4.3):

$$d_{in} = \min_{j \in [1, N_{un}]} d(e_i^g, e_j), \forall_i \in [1, p] \tag{4.2}$$

$$e_{min} = \arg \min_{j \in [1, N_{un}]} d(e_i^g, e_j), \forall_i \in [1, p] \tag{4.3}$$

where e_i^g denotes the i-th record in group g, e_j denotes the j-th record in the unassigned set of records and N_{un} is the number of unassigned records, that is, the number of records that have not yet been assigned to any group. If Equation (4.3) is satisfied by more than one record, one of them is randomly selected as e_{min}. Next, the minimum distance d_{out} from the selected record e_{min} to any of the remaining unassigned records is computed using Equation (4.4):

$$d_{out} = \min_{j \in [1, N_{un}], e_{min} \neq e_j} [d(e_{min}, e_j)], \tag{4.4}$$

Finally, to decide on the inclusion of e_{min} into group g, we compare its distance d_{in} to g with its distance d_{out} to the closest unassigned neighbor. Expression (4.5) gives the decision criterion:

$$ADD\ RECORD = \begin{cases} YES & \text{if } d_{in} < \gamma\, d_{out} \\ NO & \text{otherwise} \end{cases} \tag{4.5}$$

where γ is a *gain* factor that has to be tuned to improve the adaptability of V-MDAV. However, determining the best values of γ is not straightforward [156]. The extension process is repeated until the group size equals $2k-1$ or the condition in Expression (4.5) is not satisfied, because as shown in [55], in an optimal k-partition, each group consists of k to $2k-1$ records.

V-MDAV may leave (as well as MDAV) some records unassigned at the end of the main loop. Therefore, the remaining records are assigned to their closest group (Algorithm 3, *line 9*). At the end of all these steps, a microaggregated dataset M is generated from the resulting k-partition (represented in Algorithm 3, *line 10*) as $(g_1 \dots g_s)$.

4.4 Gaussian Noise vs MDAV

In this section, we report the experimental results of the MDAV method (Section 4.2) and compare them against those obtained with the prevalent Gaussian noise addition method (GNA) described in Section 4.1. First, Section 4.5.1 shows the results related to the privacy and the utility provided by the analysed methods. Next, Section 4.5.2 assesses the quality of the recommendations.

Experiments with GNA were repeated 50 times with each evaluated σ. In order to test the quality of both methods, we use two well-known CF datasets: Movielens 100k and Jester (see Section 2.4.2 for more details). We use the MAE and behavioural precision metrics, defined in Section 2.4, to evaluate the methods in terms of privacy and recommendation's quality.

4.4.1 Protection Assessment: Information Loss and Privacy

In the following tables and figures we show both SSE and DR results for the analysed methods: MDAV and GNA. Note that in terms of privacy and utility of data, both SSE and DR should be low. Table 4.1 shows the results obtained with MDAV for different values of k, which represent the cardinality of clusters, while Table 4.2 shows the results obtained using GNA with different values of σ. Clearly, MDAV achieves a better trade-

off between SSE and DR than GNA for the both databases. Note that our method satisfies k-anonymity and its DR is upper-bounded by $1/k$ by design.

Table 4.1: Results of MDAV-based PPCF. For the sake of clarity, the SSE results of Movielens 100k and Jester databases are displayed in a 10^3 and in a 10^6 scale, respectively.

		MDAV															
	k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**25**	**50**	**75**	**100**	**150**	**200**
ML 100k	SSE	64	87	99	105	110	114	117	119	120	130	134	136	136	138	139
	DR %	40.82	26.51	19.93	15.90	12.19	12.19	9.65	7.95	7.21	2.33	0.63	0.21	0.21	0.10	0.10
Jester	SSE	25	37	44	48	52	54	57	58	60	69	73	75	76	77	78
	DR %	47.45	30.57	22.17	17.12	13.72	11.33	9.45	8.00	6.80	1.57	0.51	0.30	0.24	0.14	0.12

Table 4.2: Results of GNA-based PPCF. For the sake of clarity, the SSE results of Movielens 100k and Jester databases are displayed in a 10^3 and in a 10^6 scale, respectively.

		GNA															
	σ	**0.25**	**0.5**	**0.75**	**1**	**1.25**	**1.5**	**1.75**	**2**	**2.5**	**3**	**3.5**	**4**	**5**	**10**	**20**	**50**
ML 100k	SSE	246	248	257	275	302	336	376	418	509	592	663	727	830	1078	1221	1339
	DR%	100	100	98.51	89.28	68.50	50.58	44.53	27.99	18.76	10.49	8.58	7.21	4.24	1.40	0.42	0.10
Jester	SSE	6	25	55	93	136	181	226	268	342	404	454	495	558	698	774	821
	DR%	99.87	95.29	78.27	58.52	42.05	30.22	22.03	15.75	8.12	4.19	2.25	1.30	0.46	0.03	0.01	0

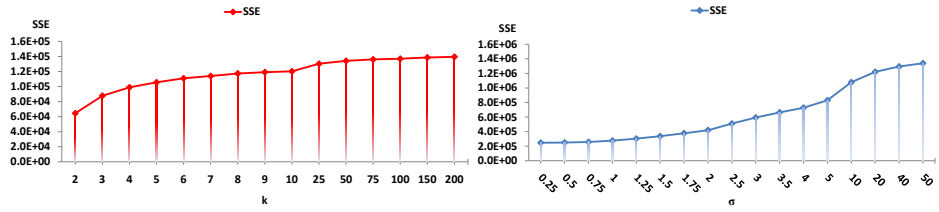
4.4.1.1 Movielens 100k

The outcomes of MDAV-based PPCF for different values of k are depicted in Figure 4.3(a) and Figure 4.3(c). Note that their behaviour is pretty antagonistic. When SSE is increased (*cf* Figure 4.3(a)), DR is reduced accordingly (*cf* Figure 4.3(c)).

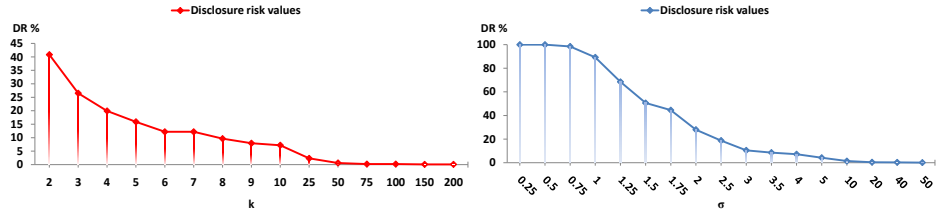
In Figure 4.3(b) and Figure 4.3(d) we can observe, respectively, the SSE and DR obtained by the GNA approach. Similarly to MDAV-based approach, when SSE grows, DR decreases. However, GNA method needs to add much more distortion to data (*i.e.* SSE is increased) than the MDAV-based approach to reach an equivalent DR.

4.4.1.2 Jester

The behaviour of the results obtained for Jester database, illustrated in Figure 4.4, is almost identical to the one obtained for Movielens 100k database. Likewise, data remain affected in the same way, but in another scale, due to



(a) SSE values of the MDAV method (b) SSE values of the GNA method



(c) DR values of the MDAV method (d) DR values of the GNA method

Figure 4.3: SSE and DR results of the implemented methods on the Movielens 100k database. Reprinted from [35].

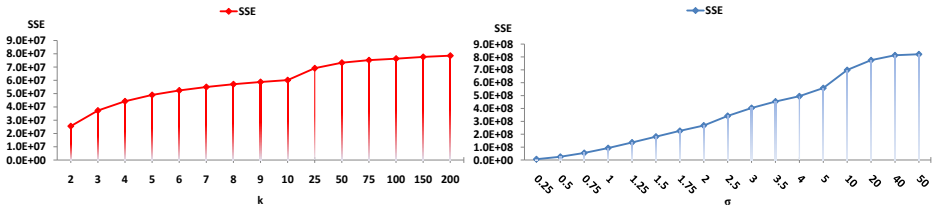
Jester's range of values. Since the σ values proposed for GNA approach are the same in both matrices, the amount of added noise has a lower impact in Jester database because the range of values is significantly wider. In the most extreme case (*i.e.* σ=50), the obtained DR is equal to 0. However, the SSE is so high (8.21×10^8) that data are practically useless.

4.4.2 Recommendation Accuracy

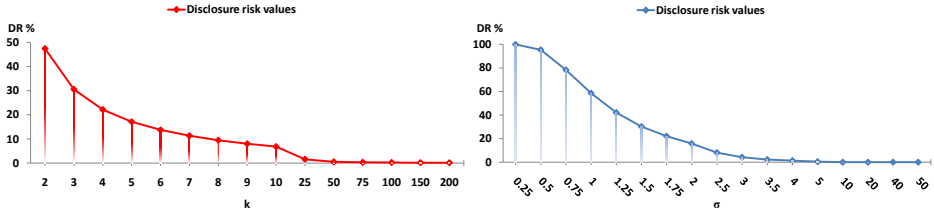
Since the protected data will be used by recommender systems, the SSE metric does not entirely capture the information loss. Therefore, we need to analyse how accurate are the recommendations. To do so, we define a training set with 80% of the item values and a test set with the remaining 20%, for both databases. We use the protected records for the users in the training set and the original records for the users in the test set. The predictions are computed as follows:

- *Find closest neighbour.* Given a user u_a in the test set, find its closest user, say $u_{closest}$, in the protected training dataset.
- *Assign prediction.* The predicted values for user u_a are those that correspond to $u_{closest}$ in the test set.

Once the prediction for all users in the test set has been done as above, we compute the error between the original values of the test set and the



(a) SSE values of our method (b) SSE values of the GNA method



(c) DR values of our method (d) DR values of the GNA method

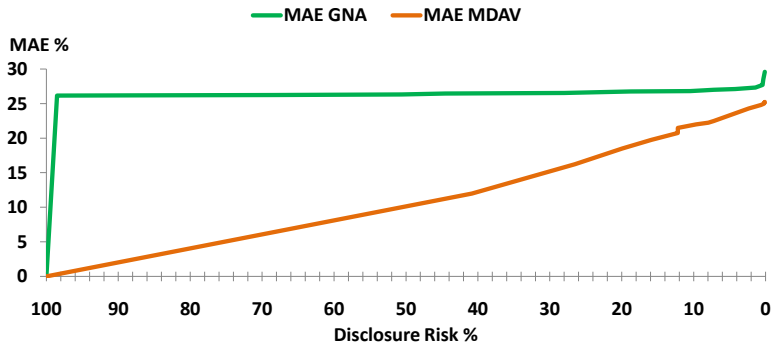
Figure 4.4: SSE and DR results of the implemented methods on Jester database. Reprinted from [35].

values assigned by the above procedure. To compute this error we apply the widely-used mean absolute error (MAE) (5.1). The MAE outcomes in respect of the DR are displayed in Figure 4.5.

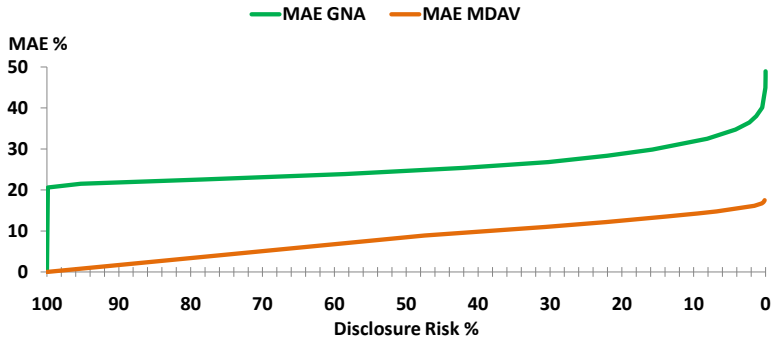
As depicted in Figure 4.5(a), the MAE outcomes of GNA method grow significantly with low values of added noise due to Movielens 100k's short range of values. Moreover, the values reach extreme values when a considerable amount of noise is added. Additionally, MAE is negatively affected by the sparseness of the matrix. Finally, the growth increases as the value of σ does. The growth of MAE in MDAV-based method is linear with respect to k and achieves significantly lower values, which means more accuracte recommendations.

In regard to Jester database (Figure 4.5(b)), the MAE achieved by the MDAV method is better than the one achieved with Movielens 100k due to Jester's lower sparseness. Moreover, the value range of Jester is wider and thus the noise introduced by GNA affects data with a slower pace.

In order to perform a clearer comparison between MDAV and GNA, we have selected obfuscated datasets with the same privacy level (*i.e.* DR). Thereafter, for the sake of simplicity, we have performed a single comparison for both databases. Note that any DR value could have been chosen as long as it is the same for both methods. The results of such comparison are displayed in Table 4.3.



(a) Movielens 100k dataset



(b) Jester dataset

Figure 4.5: Relation between MAE and DR for the analysed methods on the selected databases. (The lower the better). Reprinted from [35].

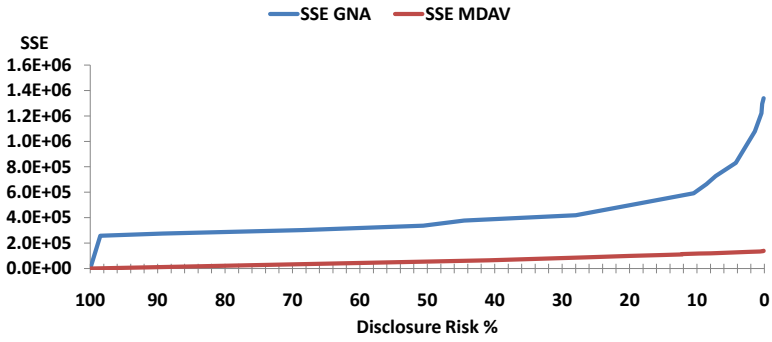
Table 4.3: Example of a comparison of MAE and %MAE results between GNA and MDAV for the analysed datasets. The lower the better.

Database	Method	MAE	%MAE
ML 100k	$MDAV,\ k = 10$	0.89	22.25
	$GNA,\ \sigma = 4$	1.08	27
Jester	$MDAV,\ k = 9$	2.91	14.55
	$GNA,\ \sigma = 2.5$	6.50	32.5

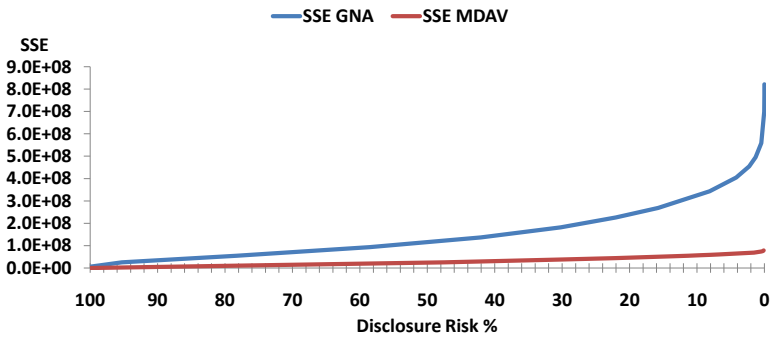
4.4.3 MDAV vs GNA: Comparison and Discussion

In the previous section we presented the results obtained by our microaggregation-based approach and the classical GNA approach. In this section we briefly compare these results and show that MDAV-approach is

superior, both in terms of privacy and recommendation's accuracy.



(a) Movielens 100k dataset



(b) Jester dataset

Figure 4.6: Relation between SSE and DR of the analysed methods. Reprinted from [35].

4.4.3.1 Movielens 100k

In Figure 4.6(a), we depict a comparison between SSE and DR for both methods. In the X-axis we represent DR and in the Y-axis we show SSE. This figure can be used to read the amount of noise, in terms of SSE, that is required by each method to achieve a given DR. For example, for a fixed value of $DR = 30\%$, MDAV roughly introduces an error in the order of $100K$ while GNA requires $400K$.

The smallest possible DR value, if data are obfuscated with MDAV method, is $\frac{1}{943} \simeq 0.1\%$ for Movielens 100k dataset. In order to obtain such value, MDAV needs to form clusters of $k = 150$ elements, which leads

to a SSE of 138,650. In contrast, GNA obtains such DR value with a SSE of 1,339,008, which is almost one order of magnitude larger.

These outcomes show that MDAV perturbs data in a much more efficient way. Moreover, as already mentioned, our method provides k-anonymity and therefore upper-bounds the disclosure risk by design.

Regarding the quality of recommendations, Figure 4.5 shows the relation between MAE and DR for both datasets. The growth resembles that of Figure 4.6, with slight differences due to the % of prediction set and rating's truncation. Moreover, Table 4.3 shows an example of the recommendations accuracy. Note that if recommendations are conducted using the data protected with MDAV, MAE is 22.25% with a DR value of 7.21%, which is a considerable privacy level. On the contrary, the values predicted using the data protected with GNA lead to a 27% MAE, which is almost 5 percentage points higher. Therefore, we may conclude that both the quality of recommendations and the quality of privacy are better in our method based on MDAV.

4.4.3.2 Jester

In Figure 4.4(d) and in Table 4.2 we notice that the DR reaches a considerable privacy level with low values of σ. Although the growth pace of SSE is nearly the same for both databases, the quantity of added noise in Movielens 100k is lower due to its shorter range, compared with Jester, especially for high values of σ.

The quality of recommendations is shown in Table 4.3. For almost the same level of privacy, GNA reaches a 32.5% MAE, which is more than twice the MAE achieved by MDAV method (*i.e.* 14.55%). The remarkable accuracy achieved by MDAV is explained by the density of Jester, which contains more than 55% of original votes.

In Figure 4.6(b) we can observe the efficiency of the applied noise in Jester. The behaviour is nearly identical to that shown in Figure 4.6(a) for Movielens 100k, except for the initial growth of SSE with GNA method. Such growth difference is produced because Jester has a wider range of values and, thus, data are less obfuscated.

Both Figures 4.6(a) and 4.6(b) confirm that MDAV is applicable to sparse and dense databases and its quality is superior than the one of GNA, regardless of data. Obviously, the quality of recommendations is highly related with the density of the database.

The experimental results presented in this section show the effectiveness of the MDAV-based technique in protecting users' privacy without compro-

mising the quality of the recommendations. In this sense, MDAV perturbs data in a more efficient way than other well-known methods such as GNA.

4.5 MDAV vs V-MDAV

In this section, we show the experimental results obtained by V-MDAV and MDAV approaches. Moreover, we conduct an alaysis of the datasets without privacy, regarding the quality of the recommendations, to show the goodness of both proposals. First, we show the outcomes related to the privacy of the analysed methods in Section 4.5.1. Next, we assess the quality of recommendations in Section 4.5.2. Moreover, we use new metrics based on the behaviour of users (see Section 2.4.1) to overcome the drawbacks of traditional well-known measures.

In order to conduct the experiments we have adjusted a test on both Movielens 100k and Jester datasets. We randomly select the 80% of the item values as the training set and the remaining 20% is considered as the prediction set. The experiments are repeated 50 times and the average of the outcomes is considered as the predicted value. The matrices were previously filled using the center of their corresponding range values (*i.e.* default voting method), to compute the Euclidean distances when V-MDAV and MDAV are applied.

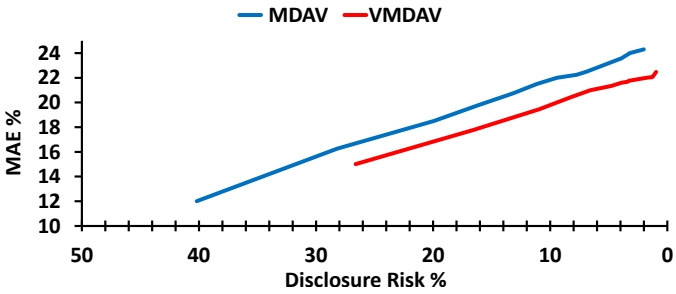
4.5.1 Privacy Analysis

Typically, two factors are considered to measure the quality of the privacy provided by a perturbation method, namely the information loss and the disclosure risk. Notwithstanding, here we focus in the quality of the predictions, since we need to measure the real data usability when computing recommendations, being far less significant the value of the information loss. Therefore, we use the mean absolute error (MAE) and the disclosure risk (DR) to assess the quality of the recommendations and the privacy level, respectively. Clearly, in terms of privacy and utility of data, both MAE and DR should be low. Hence, the best approach will be the one yielding optimal trade-off between quality of recommendations (*i.e.* MAE) and the level of privacy (*i.e.* DR). In order to exploit the combination of these measures, we propose the use of a score, which is a well-known procedure in the SDC field [54, 53]. Therefore, we compare the score achieved by different approaches to determine how efficient they are in terms of accuracy and privacy. We combine the aforementioned measures in a score, namely T-score, defined

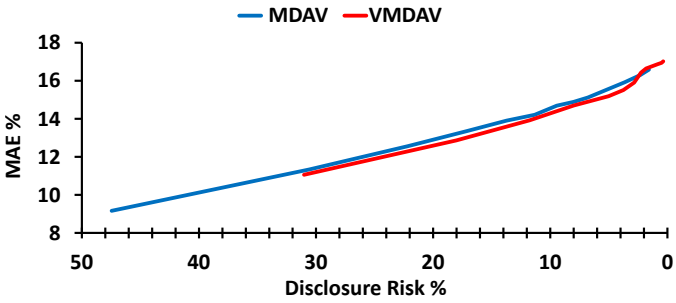
as:

$$\text{T-score} = \frac{MAE + DR}{2} \tag{4.6}$$

In this case, the lower the percentage the better the obtained score.



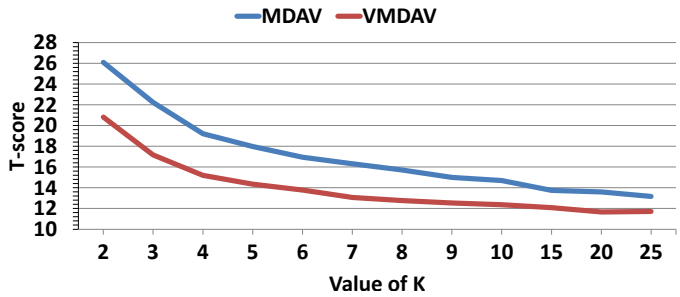
(a) MAE and DR comparison - Movielens 100k



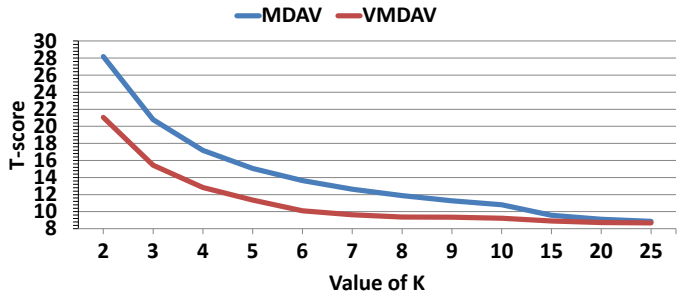
(b) MAE and DR comparison - Jester

Figure 4.7: MAE and DR comparison. The lower the better. Note that the highest DR value achieved by V-MDAV is far lower than the one obtained by MDAV. Therefore, V-MDAV heuristic tend to aggregate elements in each group, which increases the cardinality (*i.e.* always between k and $2k-1$) and decreases the DR value. Reprinted from [41].

V-MDAV and MDAV have been applied to obfuscate both datasets to determine which method achieves better results in terms of data utility and privacy. We have evaluated the V-MDAV proposal with different values of γ (*i.e.* ranging from 0.1 to 3.0, being $\gamma = 0$ a value that indicates no extension of the groups and, hence, a fixed-size cardinality V-MDAV) for every k, to generate the groups that better fit the data distribution. We compare the MAE and DR for both datasets (*cf* Figure 4.7). We may observe that the outcomes of MDAV and V-MDAV are very similar in the case of Jester (*cf* Figure 4.7(b)). This results are not surprising because MDAV is known to perform very well on scattered datasets [156]. In the case of Movielens 100k, a huge percentage of data has been filled with the central value (*i.e.* the



(a) Efficiency of the applied noise - Movielens 100k



(b) Efficiency of the applied noise - Jester

Figure 4.8: Efficiency of the applied noise in both datasets, regarding the T-score obtained by the evaluated microaggregation approaches. The lower the better. Reprinted from [41].

sparseness of this dataset is more than 90%). Such procedure transforms Movielens 100k dataset in a clustered dataset, in which V-MDAV has been shown to outperform MDAV [156]. Moreover, in Figure 4.8, we compare the best value of T-score for every k.

4.5.2 Recommendations Analysis

The well-known "leave-one-out experiment", in which the values of the closest neighbour of the active user u_a are selected as the predicted values, has been conducted to evaluate the accuracy of recommendations. The error between the original dataset values and the predicted values is computed using the MAE metric.

In order to adapt the leave-one-out metric to enable the analysis of raw databases, we have conducted several modifications. Since raw matrices contain non-rated items, the Euclidean distance between common-rated items is used to compute the closest neighbour of each user. Thus, we may need to select more than one neighbour to obtain the predicted values for the user

that is being evaluated. Moreover, the evaluated user u_a might be the only one who rated a specific item. In such case, we will not be able to assign a prediction value for that item, since nobody rated it. To address such issues, the central value is used as a prediction. The aforementioned problem does not affect microaggregation-based methods because matrices are filled before data obfuscation. Note that we use Euclidean distance along the paper instead of other similarity approaches such as Pearson Correlation, because the data sparseness makes it difficult to find users with common assessments. The latter is known to affect correlation-based measurements [149, 47], specially for computations over raw data. In addition to MAE, we use behavioural precision metrics (see Section 2.4.1) to characterise the profile of the user, taking into account both positive and negative assessments, hence obtaining more information about the users' interests.

The MAE obtained in experiments without privacy, which compare raw data with data filled using the center value, is shown in Table 4.4. In the case of behavioural precision, the outcomes are depicted in Figure 4.9.

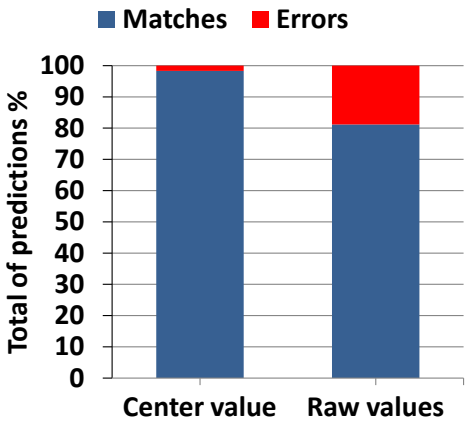
MAE %	**Movielens 100k**	**Jester**
Center value	23.519	19.022
Raw data	26.205	22.913

Table 4.4: Recommendations of the evaluated datasets, without privacy.

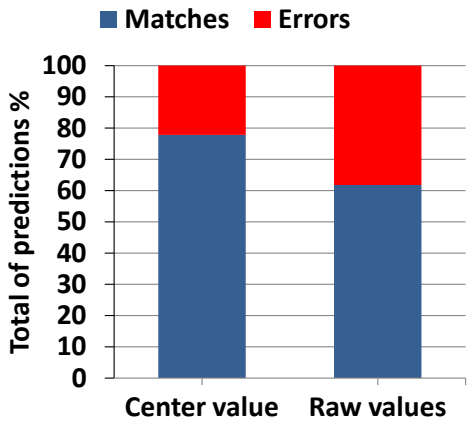
Apparently, our experiments indicate that V-MDAV outperforms the results of MDAV. For the sake of simplicity, in what follows we only focus on further analysis of V-MDAV with respect to the quality of its recommendations. Regarding the experiments with privacy, the MAE results of V-MDAV proposal for every value of k are depicted in Figure 4.10 and the measuremens about the behavioural precision are illustrated in Figure 4.11.

4.5.3 MDAV vs V-MDAV: Comparison and Discussion

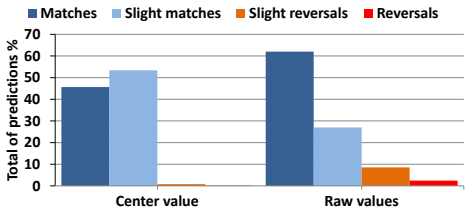
In this section we compare the efficiency of both microaggregation-based proposals in terms of privacy and recommendations quality. Moreover, we discuss the outcomes of the recommendations with raw datasets (*i.e.* without obfuscation) and the results of the recommendation's accuracy and behavioural precision obtained by the V-MDAV method.



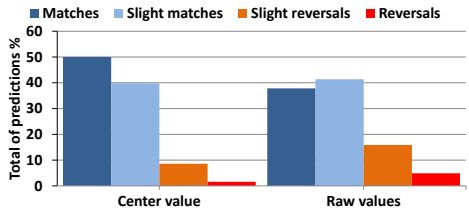
(a) Binary behavioural prediction accuracy - Movielens 100k



(b) Binary behavioural prediction accuracy - Jester



(c) 4L behavioural prediction accuracy - Movielens 100k



(d) 4L behavioural prediction accuracy - Jester

Figure 4.9: Recommendations of the datasets without privacy (*i.e.* before data ofuscation). Reprinted from [41].

4.5.3.1 Privacy

To assess the efficacy of our obfuscation methods, we have applied the T-score metric proposed in Section 4.5.1, and the results of the experiments are depicted in Figure 4.8. Although the degrowth pace is nearly the same for both approaches, the T-score values are very close for high values of k. This occurs because, as the value of k increases, the gain factor γ of the V-MDAV method reaches better T-score values when the cardinality of the group is closer to k. Therefore, when the group cardinality is drastically increased, the groups generated by both methods become almost identical. The most extreme case would be reached if $TotalUsers/2 < k \leq TotalUsers$ and, consequently, both methods would generate the same single group. Notwithstanding, as it may be observed in Figure 4.8, the V-MDAV method outperforms the MDAV approach for every value of k in Movielens 100k and for almos all in the case of Jester. Moreover, as showed in Figure 4.7 V-MDAV

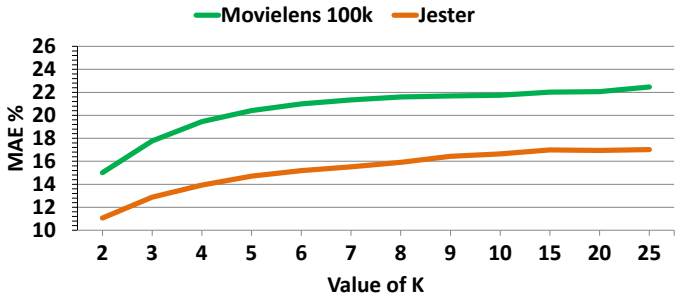


Figure 4.10: MAE% obtained after obfuscating data with V-MDAV. Reprinted from [41].

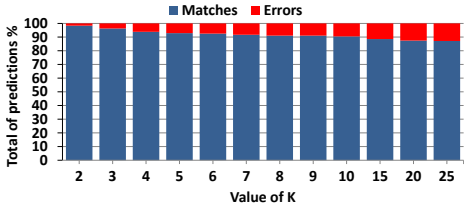
also outperforms MDAV when we compare both MAE and DR, especially for clustered datasets such as Movielens 100k. As a conclusion, V-MDAV clearly obfuscates data more efficiently than MDAV and, by extension, than Gaussian noise addition [35].

In what follows, we discuss the quality of the recommendations, regarding the results shown in Section 4.5.2. We analyse the results of recommendations without privacy in Section 4.5.3.2 and the outcomes with privacy in Section 4.5.3.3.

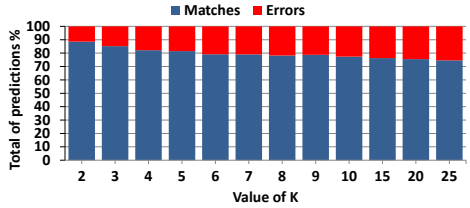
4.5.3.2 Recommendations without privacy

The MAE outcomes shown in Table 4.4, indicate that filling the databases with the center value lead to better prediction accuracy than analysing raw matrices. This occurs because, as discussed in Section 4.5.2, we need to perform a high number of distance computations for every user to obtain a prediction. Therefore, predictions can be provided by users, which are not neighbours at all and, hence, such values become random recommendations. Furthermore, recommending a center value gives less biased results for both databases. The latter phenomenon is exacerbated in Movielens 100k dataset due to its sparseness.

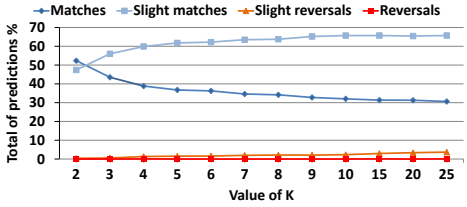
Additionally, the outcomes of behavioural precision depicted in Figure 4.9 indicate that filling the datasets with their corresponding center values, provides both higher binary and 4L precision than using raw data. More concretely, in the case of binary precision (Figures 4.9(a) and 4.9(b)), we achieve more than 95% accuracy for Movielens 100k and more than 75% accuracy precision for Jester, results which clearly outperform the ones obtained with raw databases. In the case of 4L precision analysis (Figures 4.9(c) and **??**), the vast majority of recommendations conducted using the



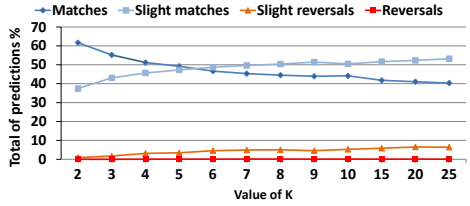
(a) Binary behavioural prediction accuracy - Movielens 100k



(b) Binary behavioural prediction accuracy - Jester



(c) 4L behavioural prediction accuracy - Movielens 100k



(d) 4L behavioural prediction accuracy - Jester

Figure 4.11: Recommendations after applying V-MDAV to the evaluated datasets. Reprinted from [41].

filled version of Movielens 100k dataset have the same behaviour of the users, thus the error is close to 0%, as observed with binary precision. In contrast, the raw Movielens 100k dataset generates approximately a 10% of slight errors, and almost a 5% of reversals. In the case of Jester, we observe that the sum of slight reversals and reversals is below 10%, while the total sum of errors obtained by raw Jester is more than 20% (cf Figure **??**). Moreover, the behavioural matches of raw Jester database are about a 15% lower than these achieved with the filled version. Hence, filling the matrices with the center value translates into more accuracy and precision when the recommendations are computed, in comparison with raw matrices.

4.5.3.3 Recommendations with privacy

Movielens 100k and Jester datasets are very different in terms of sparseness, as discussed in Section 2.4.2. As a result, Jester dataset achieves lower (*i.e.* and thus, better) MAE values due to its density, *cf* Figure 4.10. Notwithstanding, all MAE values are lower than those from the datasets without privacy. Interestingly enough, such results indicate that V-MDAV smooths data in a manner that the achieved recommendations are far less biased. In other words, from the perspective of accuracy of recommendations, it is better to obfuscate data with V-MDAV (*e.g.* especially for $k = 2$), than

use the imputed matrices without obfuscation. The measuremens about the behabioural precision of V-MDAV, depicted in Figure 4.11, also indicate better quality than the ones without privacy. Hence, such method achieves less percentage of errors and reversals for considerable privacy levels. In Figure 4.11, we observe that the outcomes of precision are better for Movielens 100k than for Jester, which initially seems to contradict the MAE values depicted in Figure 4.10. However, this precision accuracy can be explained since 90% of Movielens 100k's assessments are the respective center values, which may have both beneficial and detrimental effects on data. One the one hand, it is beneficial because the recommendations will not be biased; on the other hand, it is detrimental because even if they are not biased, they will not be high accurate. This fact is reflected in Figures 4.11(c) and 4.11(d), where Jester reaches a little higher percentage of errors than Movielens 100k (although in most cases they are lower than 5%), but also a higher percentage of matches because Jester dataset is far less sparse.

The aforementioned behavioural precision results could also be affected by the way user's perform their assessments, in regard to the range values of the evaluated data (*e.g.* the higher the range, the more precise the users can be). Moreover, such precision outcomes indicate that, although the MAE obtained by Jester's experiments is better, the binary precision of Movielens 100k dataset was higher. Additionally, although the 4L precision accuracy (see Figure 4.11(d)) achieved more percentage of matches, it also obtained more slight reversals, a fact which could not be assessed only with MAE metrics. These results fully support our precision heuristics, which indicate that measuring the prediction's error solely with the well-konwn MAE metric is not enough to test the quality of recommendations. Therefore, the proposed behavioural measurements can be considered a necessary tool to perform a more robust analysis of a recommender system and assess its quality.

4.6 Conclusions

Collaborative Filtering is a recommender system used to perform automatic recommendations to users in multiple contexts. Despite the great advantages of using CF, we have highlighted its downside regarding users' privacy. However, there is a trade-off between the privacy of users' preferences and the quality of the recommendations obtained.

In this chapter, we have proposed three PPCF methods to protect the privacy of the users involved in CF processes. We have analysed/discussed how V-MDAV obtains better results and provides both more privacy and data usability than well-known methods such as MDAV and Gaussian noise

addition [35]. In addition, we showed that the quality of predictions achieved by microaggregation-based methods was even better than these obtained without obfuscating the datasets. Moreover, both microaggregation-based proposals achieve k-anonymity, which guarantees privacy by design, a feature not offered by GNA. It is important to emphasise that our microaggregation-based methods concentrate on the protection of the data and they might be combined with other techniques such as the encryption of identifiers to provide a holistic protection of the users' privacy. The use of behavioural measures to compute the quality of predictions is also a relevant contribution of this chapter. From our point of view, the MAE metric is not enough, since it could be affected by outliers and could not give accurate information about the quality of recommendations (*e.g.* lots of users could have nearly-perfect recommendations and many others could receive reversals while obtaining an average MAE value). Therefore, the use of behavioural measures allowed us to better analyze data and increase its usability.

Chapter 5

Applications of Collaborative Filtering

This chapter describes two Smart Health applications that augment their capabilities by means of Collaborative Filtering systems. Section 5.1 describes a novel smart route recommendation system, which uses the data collected from the smart city along with other crowdsourcing-based information to provide citizens with recommendations. The second approach, described in Section 5.2, is focused on sustainable healthcare service provisioning and describes a novel hybrid wireless channel characterisation method, based on 3D ray launching and a Collaborative Filtering approach. The experimental results of both applications show that Collaborative Filtering methods can be successfully used to provide enhanced Smart Health services.

Contents

5.1 Recommender Systems with Real-time Constraints for Smart Health

In this section, we show how recommender systems could be used to provide healthcare services within the context of a smart city in which citizens collaborate with the city to improve their quality of life. We describe different scenarios so as to show the usefulness of our approach and set the foundations for further research lines that are briefly pointed out.

There are plenty of health-related mobile applications devoted to sports [170] and health monitoring [109], while, healthcare applications and Ambient Assisted Living (AAL) solutions [137] are a growing research trend. The growth of such applications and systems in conjunction with emerging wearable computing technologies [159] will extend the notion of context awareness into what is called "socially aware computing" [103]. Large-scale ensembles of such heterogeneous systems entail new social/collective intelligence [61], in which users will become a part of "massive urban superorganisms" [191]. Therefore, it is natural to move towards such next-generation trends and develop systems that exploit context-aware (*e.g.* sensing infrastructures of smart cities) and collective behaviours (*e.g.* crowdsourcing-based information).

The aim of our smart route proposal is to go a step beyond the current state-of-the-art and describe a novel system that augments m-health in the context of the recently proposed s-health concept. Moreover, we perform an analysis on how the use of recommender systems within smart cities are able to create useful services to foster healthier and more secure habits for citizens and patients. First, our proposal has been tested using simulated data to prove its usefulness and feasibility. Second, a mobile application, namely *SmartRoute*, has been implemented to collect real data to run a

pilot test of our proposal. To the best of our knowledge, this is the fist work that combines recommender systems, crowdsourcing and the smart health paradigm. Moreover, empirical results and a functional application to illustrate the efficacy of our approach in real scenarios is also provided.

5.1.1 Proposed Scheme

In this section we describe the motivation that led us to design this system and what are our expectations from it. Then, the main actors of our scheme we enumerate and, finally, the general operation of the proposal and its data structures are proposed.

5.1.1.1 Rationale and Desiderata

In the context of smart cities, that struggle to foster healthy habits among their citizens and promote s-health, many citizens perform physical activities in the city, namely walking, jogging, running, bicycling, etc. With the aim to promote these healthy habits, it would be desirable to count with a system that could dynamically adapt to the needs and tastes of the citizens. Therefore, we propose a new way of using the sensing capabilities of smart cities by means of recommender systems that allow citizens/patients to obtain recommendations about the routes that better fit their capacities. The system would consider real-time constraints and information from several sources: (i) citizens' preferences, (ii) citizens' health conditions and, (iii) real-time information provided by the smart city infrastructure.

Our aim is to develop a system that fulfils the following properties:

- Citizens can obtain recommendations of routes that best fit their needs and preferences.
- Citizens might use off-the-shelf smartphones. Any other special device should not be required.
- The system will be dynamic and adapt to the real-time changes affecting variables monitored by the city sensors.
- The system will be collaborative and react in real-time.
  - Citizens will be seen as sensors allowed to contribute with their knowledge and experience to the system.
  - Citizens will be allowed to inform about unexpected situations that could affect other citizens (*e.g.* a rockslide has blocked a route).

– Citizens are allowed to provide the system with new routes so as to enrich the system.

5.1.1.2 Main Actors and Resources of the System

Our system comprises several interacting actors and resources. The main actors are the following:

Citizens. They could suffer from health conditions (then we call them **patients**) or could be healthy, in which case we refer to them simply as citizens. It is assumed that users have mobile phones or any other device that allows them to be connected to the Internet and exchange data with other parties.

Smart city sensors. The sensors of the smart city are continuously monitoring several variables, namely temperature, humidity, luminosity, pollution, allergens, etc. It is assumed that these data are available in real-time.

Databases. Information resources that contain information from citizens. In this case, two kinds of information can be distinguished:

- *Health-related information.* Information like electronic health records (EHR). However, there is no need for a complete set of EHR. On the contrary, a general overview of the health condition of citizens is enough. This way, privacy issues are significantly lessened or overcome.
- *Users' preferences information.* Information about the preferences of users with regard to city routes where they perform their physical activities. Users are allowed to rate routes according to their preferences and those rates are stored in this database.

Smart city recommender system (SCRS). Apart from citizens, this is the most important actor of our system. The SCRS is embodied in a computer program that runs in a server. It is responsible for recommending routes to citizens. Those routes are selected by using collaborative filtering techniques and by considering the preferences of users, their health condition and the real-time information that comes from the city sensors.

Communications infrastructure. It is presumed that the smart city is equipped with the proper infrastructure to allow the exchange and communication of data among all the actors. This communication infrastructure might consist of wireless networks of different nature (*e.g.* zigbee, bluetooth, IEEE 802.11).

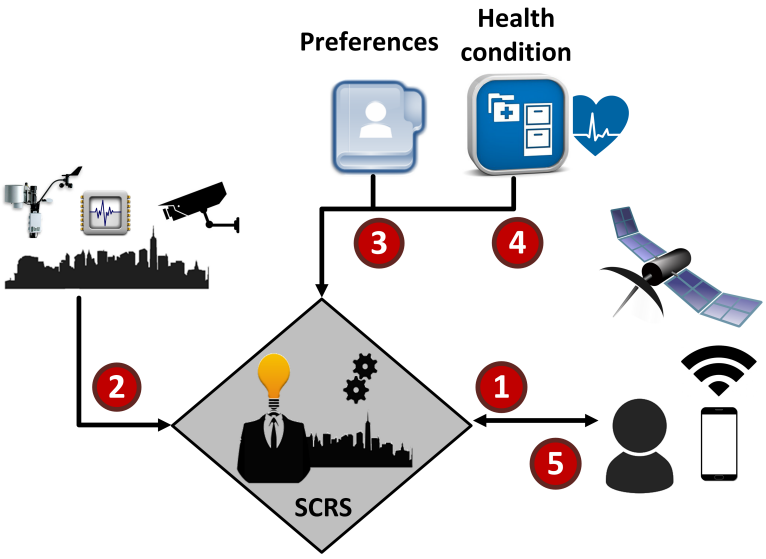


Figure 5.1: General scheme and basic operation of our proposal. Adapted from [33].

5.1.1.3 System Operation

An overview of the general scheme of our system architecture and its main actors is shown in Figure 5.1. Sensors provide real-time environmental information (*e.g.* luminosity, temperature, humidity, pollution) to the SCRS through the communication infrastructure of the smart city.

On receiving citizen queries, the SCRS checks the health information of citizens and their preferences and combines them with the real-time information of the smart city sensors to finally compute real-time recommendations that are forwarded back to the citizens.

The SCRS uses health-related information in combination with routes information to avoid recommendations that might endanger the health of citizens. Following the scheme in Figure 5.1 we can describe the main steps of the recommendation procedure as follows:

1. A citizen sends a query to the SCRS asking for a route recommendation. The citizen also sends extra information such as his/her location. Communication networks such as these based on Wi-Fi/3G/4G-LTE are utilised to enhance satellites' location services in dense urban areas, where its accuracy could be affected. Notwithstanding, continuous advances towards improving precision, reliability and continuity in global navigation satellite systems are being performed [96].

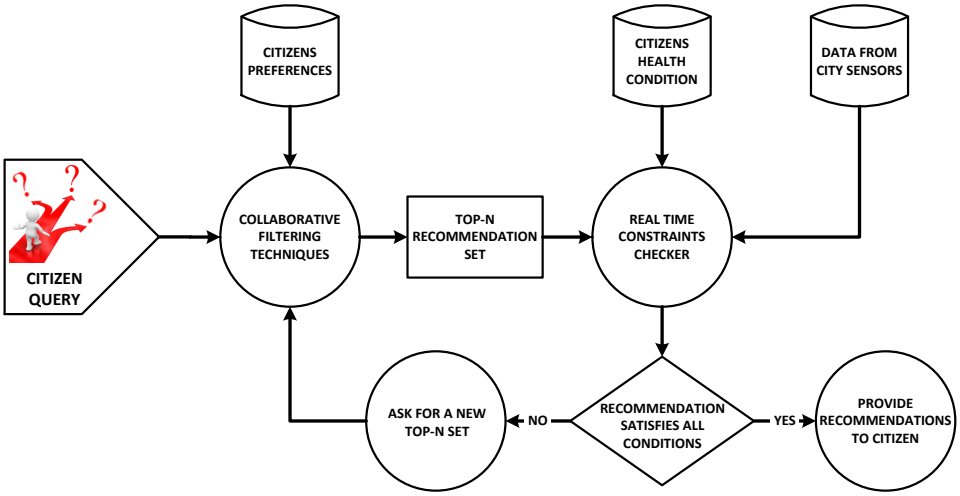


Figure 5.2: Decision flow of the SCRS upon the reception of a citizen query. Note that the procedure iterates until a proper recommendation fitting the health-related and environmental constraints of the citizen are met. Reprinted from [38].

2. The SCRS checks the real-time information coming from the city infrastructure, namely the degree of humidity, luminosity, temperature, pollution, etc. Next, our system selects the closest routes within a radius of 5 km. To avoid overwhelming the citizen with too much information and to increase the performance, the results are limited to 50 routes. Subsequently, the system determines each route status using the information gathered from each route's nearest sensor. The status is assessed by analysing the retrieved measurements using well-known metrics (*e.g.* air quality [168], ultraviolet radiation[1], wind speed and temperature and rain[2]). Depending on these measurements, the status of a route can be "danger" if one or more measures indicate that the route may be dangerous/hazardous for the user's health, "caution" if there are values that indicate possible risks, (specially for citizens, which are more sensitive to specific weather conditions or air quality) and "idle", if there is no evidence of health risk. Moreover, crowdsourcing-based information coming from other citizens about problems/issues in a route is also considered to change the route's status.

[1]World Meteorological Organization `www.wmo.int`
[2]Agencia Estatal de Meteorologia `www.aemet.es`

3. The SCRS uses the preferences of the citizen and applies a collaborative filtering method over the complete database of routes and obtains a top-N recommendation with the N most promising routes for the citizen.

4. The SCRS obtains healthcare information about the citizen. Such information is combined with the information received from the smart city to discard previously recommended routes incompatible with the health status of the citizen. Routes whose status is "danger" are discarded. Additionally, routes that could present difficulties to specific citizens are filtered. For instance, routes that are not well paved are not considered for citizens with high values of reduced/impaired mobility. Likewise, the status of routes with lots of greenery is set to "caution" for citizens with respiratory problems. After applying this health-related filter, if there is no route that satisfies the criteria, the SCRS goes back to Step 3 and looks for a new set of top-N recommendations that could satisfy the health and environmental constraints of the citizen.

5. Finally, the SCRS recommends a list of N routes that fit the preferences of the citizen and, at the same time, are adapted to his/her health condition and to the real time environmental information provided by the city sensors.

The detailed decision flow of the SCRS is depicted in Figure 5.2. The SCRS is continuously updated with environmental information from the city sensors and citizens. Such information, gathered in real time, prevents citizens from selecting routes that do not properly fit their needs due to, for instance, bad weather, crowds or high concentrations of pollutants. Note that, although we have implemented a CF method to prove the usefulness of the scheme shown in the sections below, in theory any proper CF method could be used.

5.1.1.4 Data Structures

The proposed scheme requires the utilisation of several matrices/tables in order to store the needed information. Table 5.1 is an excerpt of the database that stores the users' ratings about the utilised real routes. To compute the top-N route recommendations, the SCRS applies a CF technique based on finding the most similar/closest users. This is done using the information of Table 5.1. Therefore, the system recommends those routes that similar citizens voted with the highest values.

Table 5.1: Excerpt of citizens' ratings in $[0, 10]$. The higher the better.

	$Route_a$	$Route_b$	...	$Route_m$
u_1	2	4	...	1
⋮	⋮	⋮	⋱	⋮
u_i	3	2	...	8
⋮	⋮	⋮	⋱	⋮
u_n	6	5	...	1

Table 5.2: Example of table with health information from citizens. Values range from 0 to 1 at intervals of 0.1. Higher values indicate worse health condition.

	Age	**Visual Impairment**	**Respiratory Problems**	**Reduced Mobility**	**Heart Disease**
u_1	23	0.4	0	0.8	0
⋮	⋮	⋮	⋮	⋮	⋮
u_i	57	0.2	0. 9	0.5	0.7
⋮	⋮	⋮	⋮	⋮	⋮
u_n	45	0.2	0.8	0.1	0.2

Additionally, the SCRS uses a table that contains information about the health condition of the citizens. Table 5.2 is an example of health information, which is analysed by the SCRS to avoid inappropriate recommendations. The information about the routes is also stored. Table 5.3 is an excerpt of the routes database with information retrieved from a real scenario in the city of Tarragona. A representation of routes in the city of Barcelona and in the city of Tarragona are depicted in Figure 5.3. The routes contain checkpoints (*i.e* intermediate route points), which add dynamism and clarify possible loops.

5.1.2 Data Collection

To show the usefulness, feasibility and applicability of our approach, we have performed two experiments. First, we have simulated some data to fill two databases of preferences and health conditions. Second, we have

Table 5.3: Example of real routes data collected in the city of Tarragona.

	$Route_a$	...	$Route_k$
Start Location	41° 4'44.54"N 1°12'49.58"E	...	41° 7'45.65"N 1°14'32.90"E
End Location	41° 6'32.82"N, 1°14'58.55"E	...	41° 8'8.21"N, 1°14'59.02"E
Distance (km)	9.6	...	2.32
Elevation Gain (m)	0	...	55
Pavement Quality	Very good	...	Average
Status	Idle	...	Caution

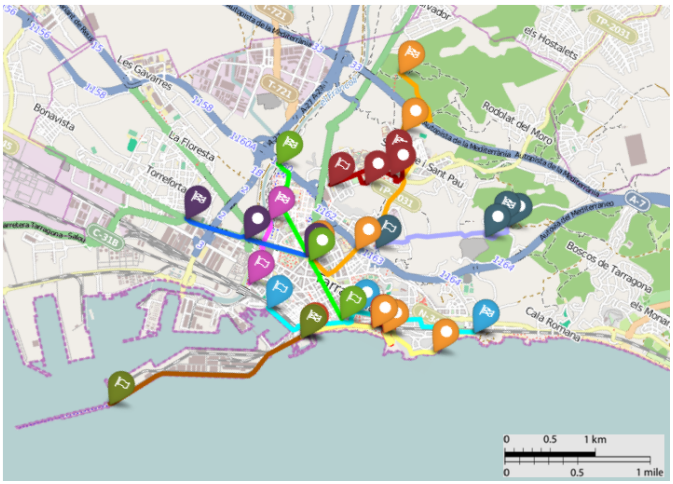
developed a mobile application, named *SmartRoute*[3], with which users may receive recommendations of routes to do sport according to their health condition and preferences. This application implements all procedures and methods described in Section 5.1.1.3. Progressively, data collected from the *SmartRoute* application are used to populate another database with real user's data. In what follows, we describe the data simulation procedure and the *SmartRoute* application functionalities. Next, we study the accuracy of the recommendations provided by our solution on the databases (*i.e* two simulated and one real) and we report the obtained experimental results.

5.1.2.1 Data Simulation

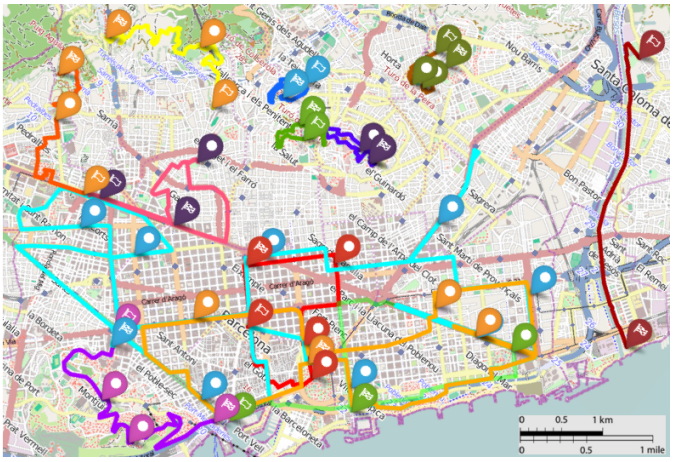
In order to test our proposal we have created two databases that store the preferences of simulated users in the cities of Tarragona and Barcelona. Tarragona database stores the preferences of 1,000 simulated users about 11 real routes, represented in Table 5.3 and depicted in Figure 5.3(a). In the case of Barcelona, the database stores the preferences of 50,000 users over 28 designed routes, graphically represented in Figure 5.3(b). Therefore, Barcelona dataset stores a total of 1,400,000 ratings and, hence, is two orders of magnitude larger than the Tarragona dataset. It is worth to emphasise that since the stored routes contain checkpoints, the real amount of routes and sub-routes is much higher. The simulated data have been generated as follows:

Citizens simulation. In order to simulate the citizen's profiles, we randomly select their age between 18 and 90 by using a distribution according with the age pyramid of the country [43]. Next, we select four main health issues, namely visual impairments, respiratory problems, reduced mobility and heart diseases. These health issues are only indicative and are not intended

[3]The application can be downloaded from the Google Play site https://play.google.com/store/apps/details?id=com.smartrouteandroid



(a) Graphic representation of routes belonging to Tarragona.



(b) Graphic representation of routes belonging to Barcelona.

Figure 5.3: Detailed maps of the routes. Here, each colour corresponds to a concrete route. Markers with a blank flag denote the start of a route, while markers with a chequered flag indicate the end. Markers with white circles inside correspond to the specific checkpoints of each route. Reprinted from [39].

for a precise characterisation of citizens. On the contrary, their goal is to illustrate the operation of our approach in a realistic and practical scenario. In a further embodiment of our approach, more health problems could be

added. To decide whether each citizen suffers from a given disease, we simulate it by using real data provided by the World Health Organisation [1] and the World Heart Federation[4]. Specifically, we consider that the probability of suffering from visual impairments is 3.4%, the probability of having respiratory problems is 3.2%, the probability of suffering from reduced/limited mobility is 2%, and the probability of suffering from some kind of cardiovascular disease is 14%. In the case of heart diseases, we considered that mainly people over 45 years are affected, since the probability of finding this kind of issues in younger people decreases drastically, and is mainly negligible.

Profile characterisation. After being simulated, citizens may be classified in different profiles depending on their health issues and age. A binary tree representing this classification is depicted in Figure 5.4. It can be observed that there exists 16 categories and 4 age intervals. This results in 64 different profiles.

Ratings simulation. It is assumed that citizens belonging to a given profile have similar needs and thus, would have similar ratings. Based on this, a relationship between the routes and the citizens profiles can be established.

- *Determine the range of routes' features.* The values of the features of each route (*i.e.* distance, elevation gain and pavement quality) are classified in a range between 0 and 5. In this case, the higher the easier (better) for the citizen. For instance, if the elevation gain is 0 meters, this feature of the route is classified as 5 (very easy). On the contrary, the highest elevation gain is classified as 0 (very difficult). The equivalences between features values and ranges have been normalised to fit in the aforementioned range. Note that features are considered independently.
- *Assigning citizens skills.* Depending on the citizens age and their health conditions, they will react differently to the routes features. For example, citizens suffering from heart diseases would be quite reluctant to follow routes with high elevation gains. To quantify this for example, we assign two negative skill points (related to the elevation gain) to every citizen suffering from heart diseases at the highest level (*i.e.* 1, according to Table 5.2). Therefore, the value of each health condition is multiplied by the values of the skills assignment heuristic represented in Table 5.4.
- *Determine the ratings.* Finally, to create the rating of a citizen for a given route, we subtract from each feature (distance, elevation and

[4] `www.world-heart-federation.org`

pavement quality) all the "modifiers" related to health conditions and age (*cf.* Table 5.4). Next, the resulting values of each feature are aggregated. Since each feature is in a rage between 0 and 5, this results in a value between 0 and 15 (where 0 is the worst value and 15 the best). Then, the result is scaled to the widely used range $[0, 10]$. Also, to introduce some variability we add Gaussian noise sampled from $\mathcal{N}(0, 1.5)$.

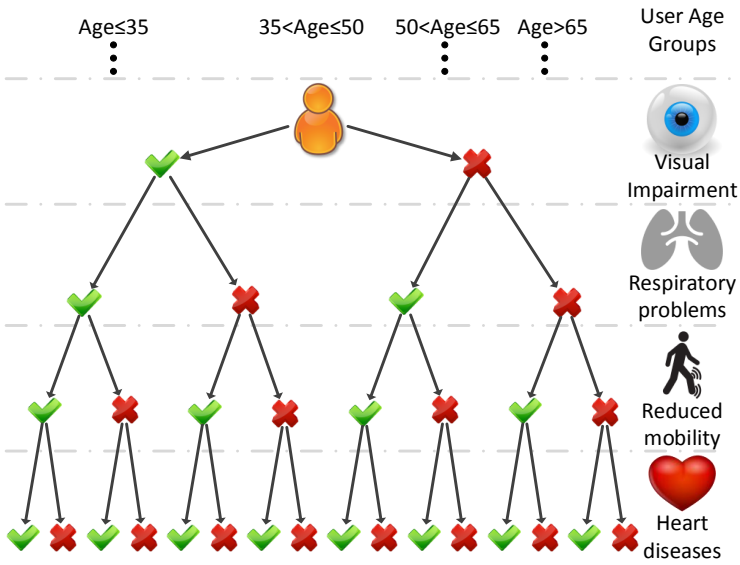


Figure 5.4: Binary classification of the possible citizens profiles. We consider 4 age intervals. For each age interval we distinguish 16 profiles that represent all the possible combinations of suffering a given disease. Reprinted from [39].

5.1.2.2 Real Data Experiments

With the aim to collect real data, we have developed the *SmartRoute* mobile application. Our application uses real weather data (*e.g.* temperature, wind speed, ultraviolet radiation) and air quality measurements [168] (*e.g.* O_3, NO_2 or PM_{10}). Such information is retrieved by the citizen's nearest sensors, which are deployed throughout the country by the Catalan air quality monitoring network (XVPCA) [64]. The sensor's data are publicly avail-

Table 5.4: Route features modifiers depending on health status and age.

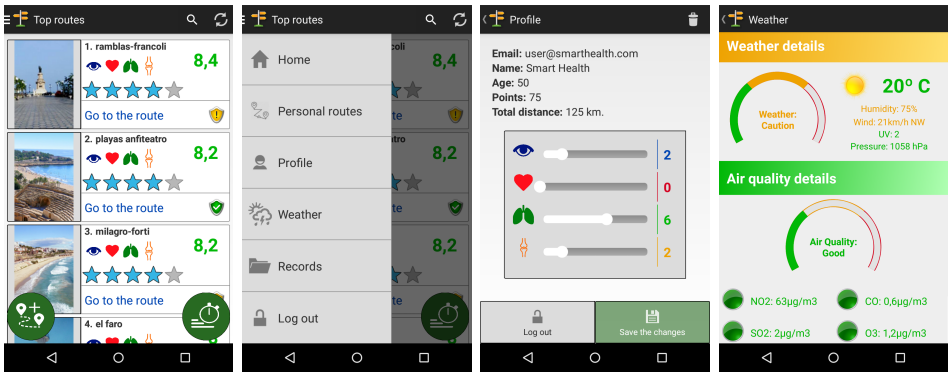
	Distance	**Elevation**	**Pavement**
18<Age≤35	0	0	0
35<Age≤50	-1	-1	0
50<Age≤65	-2	-2	0
65<Age	-3	-3	-1
Visual Impairment	0	0	-1
Breathing Problems	0	-1	0
Reduced Mobility	-1	-1	-3
Heart Diseases	-1	-2	0

able, provided by the Catalan Government[5]. Such data is updated hourly and retrieved by our system using script functions.

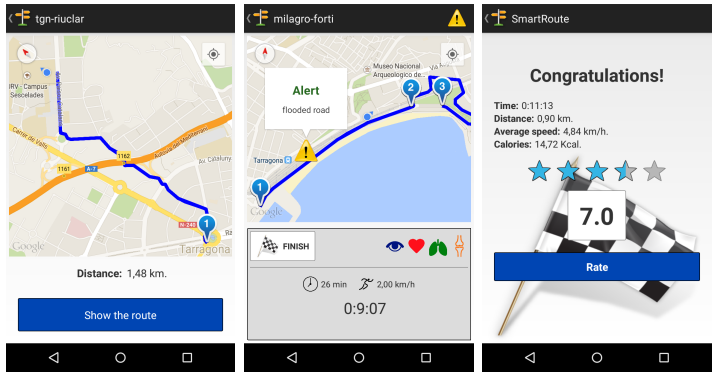
Using our app is straightforward. Citizens log in the app and, from the home screen (Figure 5.5(a)), they are able to: (i) perform a quick start (*i.e.* go to the closest route and start), (ii) look for a specific route, (iii) get the list of the most rated routes (*i.e.* top routes) or (iv) create a new route by adding checkpoints while walking. The app menu (Figure 5.5(b)) allows to check/manage records of different route sessions and update the citizen's health conditions (*i.e.* using the sliders to modify such values from 0 to 1, as showed in Table 5.2) or other data inside their profiles (Figure 5.5(c)). Additionally, citizens may check weather and air quality information retrieved from their closest stations (Figure 5.5(d)) and check the feedback provided by our app (*i.e.* the background colour will change according to the quality/dangerousness of such information). Moreover, citizens are provided with a personalised recommended route list (*i.e.* personal routes), depending on their tastes and skills. Citizens will be able to sort such recommendations by status, ratings or both according to their needs. Once a route is selected, the application shows the path to reach the starting point using the current location of the user (Figure 5.5(e)) as well as additional information (*e.g.* possible warnings in the route and a description). When a route session starts, statistics such as the length, approximate duration, speed and possible alerts or changes in the route are shown in real time (Figure 5.5(f)). Additionally, citizens may send notifications to warn others of issues in that route. Finally, the app shows the statistics collected during the session (Figure 5.5(g)) and asks for the rate of the citizens, which

[5]http://dtes.gencat.cat/icqa/start.do?lang=en

provide feedback to the SCRS.



(a) Home (b) Menu (c) Profile (d) Weather quality



(e) Route info (f) Route session (g) Finish & rate

Figure 5.5: Screenshots of the *SmartRoute* application. Reprinted from [38].

5.1.3 Experimental Results and Discussion

We performed experiments over the simulated health-related data to measure the quality of our proposal in terms of accuracy and robustness when data are sparse (*i.e.* sparseness is a well-known problem in CF methods). First, we extracted between 10% and 50% of the ratings stored in the generated datasets. Next, we predicted them using the nearest neighbour's rating (*i.e.* we applied a k nearest neighbours approach with $k = 1$, which is a well-known accepted method). Note that although larger values of k could be used, using $k = 1$ reduces the computational cost and makes our approach more practical. The neighbours have been determined by using the Euclidean distance between users, considering only the assessments which they have in common routes (which also reduces the computational cost). It

is worth to emphasise that any CF technique could be implemented without requiring specific customisations (*i.e.* the recommendation step is seen as an independent module in our system). Once the prediction for all ratings has been performed, the error between the original values and the values assigned by the above procedure is computed using the mean absolute error (MAE) (5.1), defined in Section 2.4.1.

The use of the k-nearest-neighbour approach allows real-time recommendations because most distance computations can be performed off-line. However, if there is a need for more computation efficiency further techniques may be applied [35] [19]. For instance, routes could be clustered by location or features, and users by profiles or skills, drastically reducing computational time.

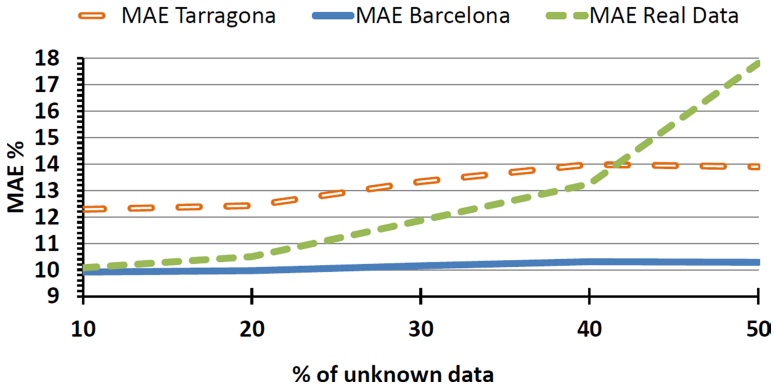


Figure 5.6: Mean absolute error of the prediction for different percentages of unknown data over the three studied databases.

Figure 5.6 shows (in red) the results with the Tarragona database. It is observed that the MAE remains low even when the percentage of unknown data (*i.e.* sparseness) grows. For 10% of unknown data we obtain a 12.28 MAE, while for 50%, the error only grows to 13.89. Clearly, when the percentage of unknown data grows, so does the error, but it remains in very low levels satisfying the user's behaviours in average (*i.e.* reversal recommendations are not provided). The results for the Barcelona database are depicted in Figure 5.6 (in blue). The growth pace of MAE values is slower than that obtained in the Tarragona Database and the values are lower. This is due to the larger number of referrals in the Barcelona database, which results in higher chances of finding similar users.

Finally, we have run a trial with 20 citizens in Tarragona and we have used the *SmartRoute* application to store their preferences and provide them with real recommendations. Figure 5.6 shows (in green) the results for the

real data database. Due to the small size of the sample, it is more difficult to find similar users when we increase the amount of missing data. However, the recommendations are still accurate (*i.e.* the highest MAE value is below 17.81%, *cf* Figure 5.6)). Thus, showing that the results over simulated data are similar to those over real data. For this trial, we selected participants with different ages and profiles and asked them to complete and rate all proposed routes in Tarragona. Next, users received recommendations based on the other's ratings, using the previously stated procedures. Overall, the recommendations pleased the participants and fitted their tastes. Moreover, after analysing the results and ratings, we observed that young and healthy users rate better challenging routes, while elderly and people with disabilities tend to rate challenging routes with lower values. Such results support the validity of our data simulation heuristics and provides new insights to create better heuristics for recommendation purposes.

5.1.4 Illustrative Case Scenarios

Beyond the aforementioned results regarding the quality of our recommender system, we highlight several realistic scenarios to show the usefulness of our proposal in practice.

Real-time Information from Multiple Sources. The SCRS receives data from multiple sources of the smart city. For example, it could be connected to the database of the police department or the fire department. Policemen and firemen update a database with real-time information about the state of streets and roads. In this example, a citizen has sent his/her query to the SCRS, which has already computed a recommendation according to his/her preferences and his/her health condition. However, the SCRS checks the database of the fire department and detects that the route, which has been recommended has been recently closed due to a rockslide. Therefore, the SCRS sends a warning to the citizen so that he/she could select a new route. In this example, the smart city infrastructure has saved valuable time to the citizen and has prevented him/her from a potential accident.

The Power of Participatory Citizens Information provided by participatory citizens is one of the most important assets of any smart city and, by extension, of our recommender system. In this example, a citizen is already doing some physical activities in a route that has been recommended by our system some time ago. Eventually, he/she notices that a section of the route is flooded and sends a notification to the SCRS. Other citizens that are nearby or in the same route receive a warning from the SCRS and, additionally, it will stop recommending this route until the problem is solved.

Smart City Sensors to Adapt Routes in Real-time Smart cities have a huge number of sensors (*e.g.* humidity sensors, pollution sensors, temperature sensors, etc.) that send continuous information to the SCRS. In a real scenario, a citizen that is doing a physical activity in a route with very dense traffic conditions might receive a warning due to, for instance, an increase in the concentration of pollutants. When this happens, the citizen could avoid this polluted area, returning to the main route when possible. This way, citizens are allowed to perform healthy activities and can be warned if the environmental conditions change and could affect them.

5.2 Sustainable Healthcare Service Provisioning through Enhanced Wireless Channel Characterisation

The advent of Context-Aware environments, mainly driven by the trend in Smart City/Smart Region development, will increase to a greater extent the deployment of 4G mobile networks and Internet of Things. This will have a decisive impact on many areas from Smart Transportation and sustainability to e-participation and Smart Healthcare [157].

One of the main considerations in this scenario is to control interference precisely, to increase coverage/capacity ratios. The wide variety of wireless systems deployed in large, dense urban scenarios require radioplanning tasks in order to account for useful server signals as well as intra-system and inter-system interference sources. Several techniques can be employed, from semi-empirical regressive methods, which exhibit large errors and measurement dependent models, to deterministic-based techniques such as full wave electromagnetic simulation. As a midpoint between precision and computational cost, Ray Launching (RL) methods offer a good trade-off between precision and computational cost.

However, when large, complex scenarios in which many potential transceivers can be located, RL exhibits high computational cost and convergence constraints [10][9]. In order to minimise computational cost for certain scenarios, in this section we propose the combination of in-house 3D RL code with a Collaborative Filtering (CF) method.

The main idea is to use the ability of CF methods to predict rates and infer the values of empty cells in matrices obtained in Low Definition (LD) simulations performed by the RL approach. Such task requires the implementation of a knowledge database employing HD simulation results. The proposed methodology has been applied to received power levels, within the

complete simulation volume represented in matrix form, although it can be extended to other parameters if required. In the last couple of years, we have developed several methodologies and optimisations to the original approach. For a better understanding, we will expose our research in this field in temporal order.

5.2.1 One-Dimensional Hybrid Simulation Technique

Our goal is to predict (recommend) values that LD simulations were unable to compute properly due to their low resolution. The context-aware scenarios analysed in the next sections have different characteristics in terms of size, obstacles and materials. Our proposal is based on memory-based CF and comprises two steps: (i) database creation and (ii) values prediction.

5.2.1.1 Database creation

The first step is to create a knowledge database that will be later used to predict missing values in LD simulations. Each scenario obtained by the 3D RL technique is modelled by a matrix $M_{n\times p}$ and is managed as follows:

The matrix $M_{n\times p}$ is serialised into a vector $V = (v_1, v_2, \cdots, v_{L_V})$, i.e. a point in $\mathbb{R}^{L_V}$, where $L_V = n \times p$. From this vector we create a set of vectors $SV = \{sv_1, sv_2, \cdots, sv_{L_V - L_{SV}+1}\}$. Each vector in SV has length L_{SV} and can be represented as a point in a subspace $\mathbb{R}^{L_{SV}}$. Note that each vector $sv_i = (v_i, v_{i+1}, \cdots, v_{i+L_{SV}-1}), \forall i \in [1, L_V - L_{SV} + 1]$. Figure 5.7 shows an example of the creation of the knowledge database consisting of vectors in a subspace $\mathbb{R}^3$, *i.e.* with $L_{SV} = 3$.

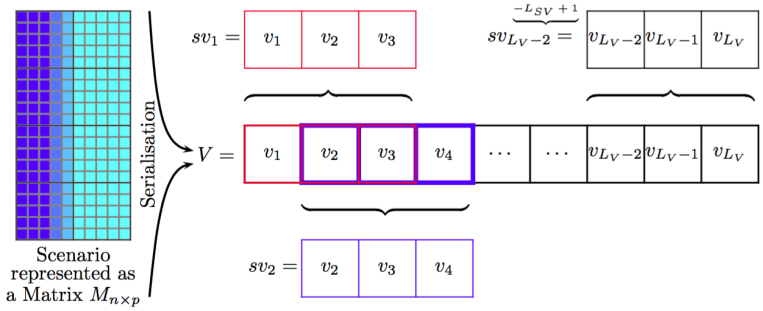


Figure 5.7: Database creation example. $L_{SV} = 3$. Reprinted from [32].

Databases might contain information from multiple scenarios. However, note that each knowledge database contains vectors $sv_i \in \mathbb{R}^{L_{SV}}$ only (*i.e.* vectors from the same subspace). Therefore, a database is created for each

subspaces. Moreover, it is important to emphasise that databases are created in pairs: DB_{LD} and DB_{HD}. DB_{LD} is the database created with scenarios simulated in LD, while DB_{HD} is the database created with the same scenarios simulated in HD. This way, it is possible to keep a relation between patterns in LD and HD.

5.2.1.2 Values prediction

In this stage, our system receives a LD simulation S with missing values (*i.e.* empty cells resulting from low angular resolution) as an input. Therefore, our aim is to predict them so that the resulting values are as similar as possible to those obtained in HD simulations.

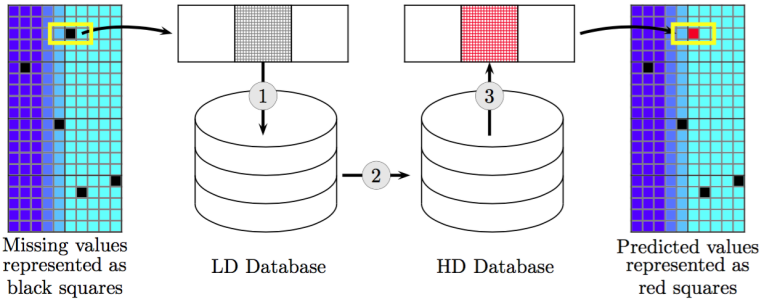


Figure 5.8: Missing values prediction: 1) The most similar patterns are found in the LD Database, 2) The corresponding HD patterns are determined, 3) The average of those values is computed and used to replace the missing value. Reprinted from [32].

First, values in S are normalised (*i.e.* to be comparable with those in the knowledge database) and a pair of knowledge databases DB_{LD} and DB_{HD} with LD and HD patterns in a given subspace $\mathbb{R}^{L_{SV}}$ are selected. Then, for every vector $v_j \in \mathbb{R}^{L_{SV}}$ from S containing missing values, we find the k closest patterns $CP^{LD} = \{cp_1, cp_2, .., cp_k\}$ in DB_{LD}. Next, we determin their corresponding HD patterns $CP^{HD} = \{cp'_1, cp'_2, .., cp'_k\}$ in DB_{HD} and we compute its average as $\bar{cp'} = \frac{1}{k}\sum_{i=1}^{k} cp'_i$. We use the values of the average pattern $\bar{cp'}$ as the prediction of the missing values. Without loss of generality, we determine the closest patterns using the Euclidean distance over non-missing values of the vectors. A graphical representation of this procedure is given in Figure 5.8. In order to increase the quality of the predictions, we only consider vectors having, at most, one missing value. This procedure is iteratively applied until all missing values are determined and, if necessary, knowledge databases with patterns in lower dimensionality

subspaces are used.

5.2.2 Knowledge Databases and Metrics

Table 5.5: Information of the selected scenarios to perform a statistically sound knowledge database. Times in seconds

	Rows	Cols	Layers	Source position (R,C,L)	Time HD	Time LD	Density %	Spars. %
Sim-1	130	70	42	95,45,6	110137	2087	2.60	13.93
Sim-2	126	182	38	82,56,35	25360	2981	3.72	43.14
Sim-3	124	273	32	70,212,23	37236	3765	3.06	43.71
Sim-4	58	62	35	18,31,11	81509	1390	3.78	0.77
Sim-5	36	60	38	35,5,30	30637	1112	4.74	1.34
Sim-6	30	30	35	3,20,11	55711	637	4.02	0
Sim-7	30	30	35	10,3,10	36112	712	8.63	0
Sim-8	50	50	35	40,20,13	43628	1088	2.98	0.24
Sim-9	40	30	35	38,15,11	90288	954	3.45	0.02
Sim-10	40	30	35	6,5,16	77479	1008	2.29	0.01
Sim-11	70	50	30	10,42,21	64139	1040	5.16	1.91
Sim-12	90	60	30	10,42,13	80004	1228	6.22	6.31
Sim-13	175	80	40	46,50,14	52583	1764	6.69	16.61
Sim-14	196	136	38	100,70,20	48910	2036	1.04	36.63
Sim-15	36	60	38	21,50,8	112404	1406	4.23	0.83
Sim-16	32	64	32	10,30,9	47577	918	0.92	0

Two knowledge datasets have been created to assess our solution. $Dataset_{12}$ includes the information of the first 12 scenarios depicted in Table 5.5. The simulated scenarios are diverse and have 30 to 40 layers (*i.e.* each one representing two-dimensional planes at different heights). Moreover, they contain a variety of features (*i.e.* corridors, columns, walls, doors and furniture). Each scenario has been simulated in LD and HD. $Dataset_{16}$ is an improved version of $Dataset_{12}$, which has 4 additional context-aware medical scenarios (*cf* Table 5.5). Depending on the desired vector length L_{SV}, we can create different personalised knowledge datasets to apply a concrete prediction strategy (*e.g.* to predict the missing values of a scenario using a knowledge database of sub-vectors with $L_{SV} = 7, 5$ or 3). In this sense, several strategies are proposed in the following sections, to ascertain which one achieves the best outcomes in each evaluated case. Although the CF approach significantly helps to improve the quality of LD simulations, it does not always predict the same value of the HD counterparts. Thus,

we compare the prediction quality of several CF strategies. In order to do so, we compute the error between the real values of the HD simulation and the values predicted by such strategies using the MAE, that was already introduced in Chapter 2, but that we are repeating here for the sake of clarity:

$$MAE = \frac{\sum_{i=1}^{n} |p_i - r_i|}{n} \tag{5.1}$$

where n is the number of missing values predicted, p_i is the predicted value for element i, and r_i is the real value of i in the HD simulation. Note that the HD simulation is only used to compute the error but it is not involved in the prediction process.

5.2.3 One-Dimensional Approach for Context-Aware Scenarios

This section shows the work presented in [31], in which the 1D approach is compared with a set of well-known methods to show its quality and usefulness. In this case, using the information of $Dataset_{12}$, we create three 1D LD knowledge databases with vector lengths $L_{SV} = 3, 5$, and 7 and their three HD counterparts. Each knowledge database contains more than half a million patterns/vectors in their corresponding subspaces. Next, we predict the values of 4 LD simulations (not included in the knowledge database) with diverse dimensions, density (*i.e.* percentage of occupied space) and sparseness (*i.e.* percentage of missing values), as depicted in Table 5.6.

Table 5.6: Test simulations features. "Time LD" and "Time HD" are the seconds needed to simulate each scenario in LD and HD respectively.

Rows	**Cols**	**Layers**	**Time (s) HD**	**Time (s) LD**	**Density %**	**Spars. %**
70	50	30	64139	1040	5,166	2,203
90	60	30	80004	1228	6,223	7,092
175	80	40	52583	1765	6,696	17,051
196	136	38	48910	2037	1,041	37,636

With the aim to analyse the accuracy and performance of our approach, we have tested three different prediction strategies. In Strategy 1 we predict the missing values with the LD database of sub-vectors with $L_{SV} = 7$. Next, if there are values that could not be predicted, we apply the LD database

with $L_{SV} = 5$. Finally, if there remain missing values we apply the LD database with $L_{SV} = 3$, which by construction guarantees the filling of all missing values in our simulations. In Strategy 2 we first use the LD database with $L_{SV} = 5$ and then that with $L_{SV} = 3$. Finally, in Strategy 3 we only use the LD database with $L_{SV} = 3$. In all cases we consider an aggregator value $k = 100$. Thus, for each missing value, we find the k most similar sub-vectors and compute their average. We compare the prediction quality of our three CF strategies with other four classic methods, namely linear interpolation, average simulation value per layer, per row, and per column.

Table 5.7: MAE (in dB). The lower the better. The row "Raw values" shows the MAE when missing values are kept empty. The column "All" is the average of all simulations weighted by their size.

	Sim-A		**Sim-B**		**Sim-C**		**Sim-D**		**All**	
Method	MAE	σ	MAE	σ	MAE	σ	MAE	σ	MAE	σ
Strategy 1	6.47	0.75	9.91	0.79	6.99	0.52	10.70	1.40	9.25	1.05
Strategy 2	6.48	0.79	10.14	0.92	7.06	0.60	11.14	1.31	9.53	1.04
Strategy 3	6.21	0.86	8.16	1.02	7.16	0.56	13.38	1.09	10.62	0.90
Linear interpolation	23.03	4.36	22.88	3.14	24.08	2.83	24.91	2.21	23.98	2.60
Layer mean value	8.84	4.11	5.31	0.46	9.31	1.46	16.37	1.27	12.76	1.42
Column mean value	7.86	3.48	5.11	0.52	9.83	1.62	16.58	1.20	13.02	1.39
Row mean value	7.59	0.93	7.95	0.99	8.65	1.55	14.05	1.28	11.51	1.32
Raw values	56.26	2.09	45.67	1.08	62.07	1.59	73.19	1.15	66.42	1.33

Table 5.8: Times (in seconds) to predict all missing values

Time (in seconds)		
Strategy 1	**Strategy 2**	**Strategy 3**
17	19	17
54	54	35
152	167	179
473	479	531

In Table 5.7 and Table 5.8 we show the MAE results and times, respectively, of the aforementioned methods for each simulation. Since our method is applied to each layer independently, the overall procedure is parallelised. Hence, Table 5.8 shows the worst layer prediction time, which is the actual total cost of our method. As an example, one of the implemented indoor test scenarios, as well as simulation results for the estimation of received power at all locations is depicted in Figure 5.9. The results correspond to differ-

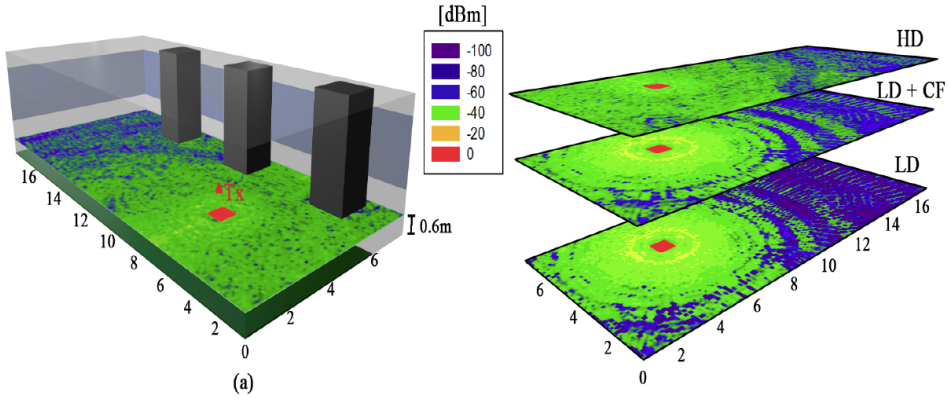


Figure 5.9: In the left, a schematic of a used test scenario. In the right, the received power level estimation when using HD (top), LD+CF (middle) and LD (bottom). Reprinted from [31].

ent simulation techniques, from HD, to LD+ CF and finally only LD. The LD+CF offers qualitatively an adequate result in terms of received power level estimation.

We may observe that our approach achieves lower/better MAE than the rest of methods for Sim-A, Sim-C and Sim-D and that Strategy 1 is the overall best method (*cf*Table 5.7). However, we note that the minimum error is not always obtained by the same strategy. Therefore, if more accurate results were to be obtained, parameters such as k, sub-vector length and its corresponding database might be tuned depending on each simulation's features. In contrast to the results of Sim-A, Sim-C and Sim-D, the MAE obtained by mean-based methods in Sim-B is lower/better than that achieved by our proposed strategies. In this case, the homogeinity of Sim-B and the short range of measurements benefit mean-based methods. However, it is important to notice that pure mean-based methods do not capture the real neighbourhood conditions of the data and might only be used in very homogeneous scenarios. Also, we observe that sparseness has an adverse effect on the prediction accuracy of all methods. This is very apparent in Sim-D, which is very sparse. This result is not surprising since with less data it is more difficult to make better decisions. Overall, as it can be seen in Table 5.7 (column: "All"), Strategy 1 outperforms Strategy 2 and Strategy 3, and obtains the best prediction accuracy, which is about 20% better that the best method not proposed by us. This result supports our main claim that CF can succeed to improve simulations accuracy (even with simple approaches).

5.2.4 Optimised One-Dimensional Hybrid Simulation Technique

In order to increase the accuracy of our method, several optimisations were performed in the database creation step as well as in the values prediction step.

5.2.4.1 Database Optimisations

In order to increase the quality of the original database (*i.e.* the one described in Section 5.2.1.1, we select LD sub-vectors that have similar/highly correlated values with their HD equivalences, this way introducing less noise in the database. Therefore, we establish a threshold to determine the maximum distance between LD and HD values of each sub-vector and consider only the ones that fall inside that threshold. Moreover, we observed that LD measurements performed close to null/error points, even falling inside the threshold, were not highly correlated with HD values. Hence, we refined the pattern selection using this procedure: (i) we compute the Manhattan distance between cells containing null/error points and cells that contain the rest of LD measurements; and (ii) we set a minimum distance (*i.e.* we selected 3 due to the minimum sub-vector size) and discard LD values whose distance is lower (*i.e.* closer to null/error points). Table 5.9 shows an example of Manhattan distance computations. Finally, we use the LD sub-vectors of the simulation that is being predicted (*i.e.* this data will be different for each simulation and will be included in the database only during the prediction process). The aim is to provide more referrals by using local values of simulations and avoid low quality predictions if our database does not contain quality samples for that scenario. After applying such optimisations, the result is a more reliable and robust database.

5.2.4.2 Values Prediction Optimisations

In the approach described in Section 5.2.3, the values prediction procedure was recursively applied until all missing values were filled and, if necessary, the sub-vector length was iteratively reduced. However, we observed that iteratively reduce the sub-vector length increased noise in each step. Hence, the predictions obtained by each sub-vector length should be applied independently. We also observed that predictions when data was extremely sparse in a vector resulted in biased outcomes. Thus, for each vector, we compute the percentage of empty cells and avoid computations if this percentage surpasses the 50%, considering that there are not enough referrals.

Table 5.9: Example of Manhattan distance computation between null cells (-1, highlighted in red) and the rest. Considering a threshold distance = 3, possible noisy cells are detected (highlighted in orange). Any sub-vector containing "orange" cells are discarded.

5	6	6	5	4	3	4
4	5	5	4	3	2	3
3	4	4	3	2	1	2
2	3	3	2	1	-1	1
1	2	3	2	1	-1	1
-1	1	2	3	2	1	2
-1	1	2	3	3	2	2

Moreover, we apply our method by rows and by columns (initially we applied it only by rows) and compute the mean of the outputs, only considering values which have been predicted both by rows and columns, discarding the rest of values. Finally, we use the mean of the obtained predictions to fill the cells that could not be predicted.

5.2.4.3 Computational Cost and other Optimisations

The complexity of the one-dimensional approach can be analysed as follows: each LD simulation has M levels with F vectors, which have to be compared with the LD database. However, each vector is decomposed in $N-L+1$ sub-vectors, where N and L are the vector and sub-vector lengths, respectively. Therefore, the final cost of this method is $D*(M*F*(N-L+1))$, where D is the number of sub-vectors existing in the LD database. Thus, the size of the database and the size of the LD simulation have strong influence on the costs. In order to optimise the search process and hence, reduce the computational cost of this approach, we have created an additional structure (optimisation structure) showed in Table 5.10. Such structure has the same number of rows and columns than the LD database and each row corresponds to the same sub-vector in both structures. For each column, we compute the sum of the sub-vector's elements after subtracting the vth element according to the column position, where $v \in [1, L]$. Given a sub-vector sv_i, we use Equation (5.2) L times (*i.e.* one for each sub-vector's value) to obtain its corresponding optimisation values Osv_i. Subsequently, each time that we analyse a new sub-vector sv_a, we sum its values and compare them with the values of the vth column in the optimisation structure, where v corresponds

to the position of the empty cell in sv_a. Next, we set a threshold and avoid distance computations if the aforementioned comparison falls outside the threshold. Notice that the database is created offline and hence, the cost of this process is negligible. By performing such optimisations, we cluster the database and reduce the cost of the search process.

$$Osv_{i_v} = \left(\sum_{j=1}^{L} sv_{i_j} \right) - sv_{i_v} \tag{5.2}$$

Table 5.10: Optimisation structure. Here, each row stores the corresponding optimisation values for each sub-vector, computed as described in Equation (5.2).

Osv \ Values of v	1	...	L
Osv_1	Osv_{1_1}	...	Osv_{1_L}
...	...	...	...
Osv_n	Osv_{n_1}	...	Osv_{n_L}

5.2.5 Optimised One-Dimensional Approach in Medical Complex Scenarios

In this section we study the usefulness of our optimised one-dimensional method in large medical complex environments [30]. We compare such optimised approach with the method and strategies presented in Section 5.2.3 in terms of accuracy and computational cost. Finally, we point out the benefits of our proposal in the s-health context.

In this case, we use $Database_{16}$ to create five LD knowledge databases with $L_{SV} = 11$, 9, 7, 5 and 3, and their five HD counterparts. Each knowledge database contains approximately one million patterns/sub-vectors. We have applied our solution on a complete floor of an existing hospital[6] (not included in the knowledge database), which has rooms with different density (*i.e.* percentage of occupied space) and sparseness (*i.e.* percentage of missing values). Moreover, it contains small doctor's offices with medical devices and special isolated rooms, typical of this type of scenarios. See

[6]We have recreated the emergency floor of the *Hospital de Navarra* (Navarra, Spain) to generate the simulations.

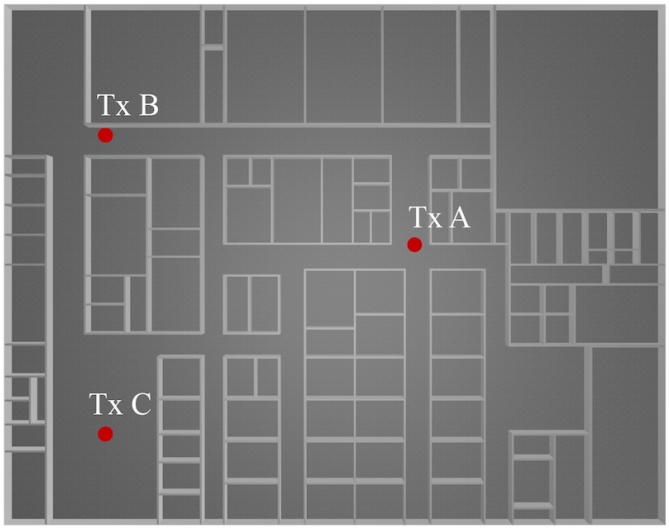


Figure 5.10: Detail of the scenario with the location of signal transmitters. Each transmitter was used separately for its corresponding simulation (*e.g.* *Tx A* was used for Sim-A). Reprinted from [30].

Table 5.11 for the exact data. From this scenario, a total of three simulations, namely Sim-A, Sim-B and Sim-C, have been performed using different configurations of the position of transmitters, showed in Figure 5.10. The simulations' obtained results in LD and HD have been used to test our algorithms.

We have tested two different types of prediction strategies, shown in Table 5.12. The first family (*i.e.* Strategies from 1.1 to 1.5) corresponds to the initial approach presented in Section 5.2.3. The second, which includes strategies from 2.1 to 2.5, corresponds to the optimised method presented in Section 5.2.4. For instance, in Strategy 1.1 we predict the missing values with the LD database of sub-vectors with $L_{SV} = 11$. If all missing values are not filled, we apply the LD database with $L_{SV} = 9$ and next we repeat the procedure with $L_{SV} = 7$ and with $L_{SV} = 5$ if needed. Finally, if there remain missing values we apply the LD database with $L_{SV} = 3$, which by construction guarantees the filling of all missing values in our simulations. The procedure is the same for strategies belonging to the first family. In contrast, Strategy 2.1 only uses the database with $L_{SV} = 11$ to compute predictions, Strategy 2.2 uses $L_{SV} = 9$ and so on. In all cases, after observing the accuracy of different aggregator values (*i.e.* $k = 1$, $k = 5$, $k = 10$, $k = 25$, $k = 50$ and $k = 100$), we selected an aggregator value $k = 25$. Hence, for

each missing value, we find the k most similar sub-vectors and compute their average.

Table 5.11: Test simulations features. 'Time LD' and 'Time HD' are the seconds needed to simulate each scenario in LD and HD respectively.

	Rows	**Cols**	**Time (s) HD**	**Time (s) LD**	**Density %**	**Sparseness %**
Sim-A	62	48	68915	463	7.86	18.11
Sim-B	62	48	29664	504	7.86	22.35
Sim-C	62	48	21094	370	7.86	30.19

Table 5.12: Summary of prediction strategies.

Strategy	**Prediction Strategy**
1.1	$L_{SV} = 11 \rightarrow L_{SV} = 9 \rightarrow L_{SV} = 7 \rightarrow L_{SV} = 5 \rightarrow L_{SV} = 3$, $(k = 25)$
1.2	$L_{SV} = 9 \rightarrow L_{SV} = 7 \rightarrow L_{SV} = 5 \rightarrow L_{SV} = 3$, $(k = 25)$
1.3	$L_{SV} = 7 \rightarrow L_{SV} = 5 \rightarrow L_{SV} = 3$, $(k = 25)$
1.4	$L_{SV} = 5 \rightarrow L_{SV} = 3$, $(k = 25)$
1.5	$L_{SV} = 3$, $(k = 25)$
2.1	$L_{SV} = 11$, $(k = 25)$
2.2	$L_{SV} = 9$, $(k = 25)$
2.3	$L_{SV} = 7$, $(k = 25)$
2.4	$L_{SV} = 5$, $(k = 25)$
2.5	$L_{SV} = 3$, $(k = 25)$

In Table 5.13 and Table 5.14 we show the MAE results and times, respectively, of the aforementioned methods for each simulation. Note that Table 5.14 shows the worst layer prediction time, which is the actual total cost of our method. As an example, one of the implemented indoor test scenarios, as well as simulation results for the estimation of received power at all locations is depicted in Figure 5.11.

The results in Table 5.13 show that strategies from 2.1 to 2.5 achieve lower/better MAE than strategies form 1.1 from 1.5, if we compare them considering their highest L_{SV} value. Moreover, the second family of strategies achieves the best overall results for each simulation in terms of MAE and σ which implies that predictions are not only better but more stable/reliable. In Table 5.13 we may observe that Strategy 2.3 is the overall best method. Note that strategies 2.4 and 2.5 (*i.e.* these corresponding to the lowest sub-vector size) have improved their results considerably in respect to the other

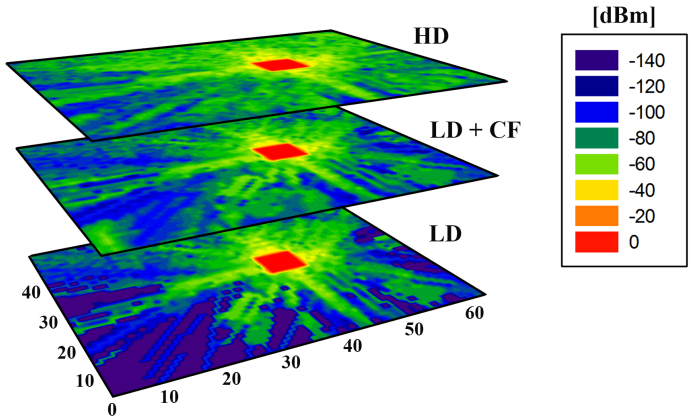


Figure 5.11: Received power level estimation when using HD (top), LD+CF (middle) and LD (bottom). Reprinted from [30].

Table 5.13: MAE (in dB). The lower the better. The row "Raw values" shows the MAE when missing values are kept empty. The column "All" is the average of all simulations' outcomes, weighted by their corresponding sparseness.

	Sim-A		**Sim-B**		**Sim-C**		**All**	
Method	MAE	σ	MAE	σ	MAE	σ	MAE	σ
Strategy 1.1	10.44	8.86	8.89	7.38	9.33	8.67	9.47	8.31
Strategy 1.2	10.26	8.75	9.08	7.15	9.26	8.58	9.46	8.28
Strategy 1.3	9.80	8.37	8.62	7.21	9.25	8.62	9.19	8.11
Strategy 1.4	10.32	9.08	9.20	7.56	9.49	8.98	9.61	8.54
Strategy 1.5	13.43	11.72	11.43	8.99	11.57	10.24	12.01	10.22
Strategy 2.1	7.48	5.98	7.01	6.01	7.60	5.99	7.38	5.99
Strategy 2.2	7.33	5.93	6.78	5.88	7.57	6.01	7.26	5.95
Strategy 2.3	7.29	5.76	6.81	5.90	7.53	5.92	7.24	5.87
Strategy 2.4	7.49	5.81	7.07	6.03	7.52	5.85	7.36	5.89
Strategy 2.5	9.91	8.14	9.12	7.33	8.01	6.32	8.84	7.10
Raw values	63.13	9.48	62.88	7.91	63.20	9.04	63.08	8.80

strategies. This occurs because the local simulation's data included in the database helps discarding more efficiently other similar samples, whose number grows as L_{SV} is reduced. However, although the results showed in Table

Table 5.14: Times (in seconds) to predict all missing values.

	Time (in seconds)									
Strategies	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**
Sim-A	74	76	84	88	101	32	33	35	39	45
Sim-B	86	90	91	99	112	38	39	41	46	52
Sim-C	85	88	90	94	109	38	37	41	44	51

5.13 correspond to average values (for the sake of clarity) we note that the minimum error is not always obtained by the same strategy. Therefore, if more accurate results were to be obtained, parameters such as k, sub-vector length and its corresponding database might be tuned depending on each simulation's features. We may also observe that, although Sim-C is the most sparse, the MAE values obtained by the first family of strategies for Sim-A are the highest. This occurs because, in addition to the sparseness value, the quality of the samples found in the database determine the accuracy of the method. The same occurs if we compare the accuracy obtained in Sim-B with the rest of simulations, taking into account their corresponding sparseness values. Overall, as showed in Table 5.13 (column: "All"), Strategy 2.3 outperforms the rest of strategies and obtains the best prediction accuracy, which is more than 20% better than the best strategy of the first family (*i.e.* Strategy 1.3).

The time required to compute the predictions for each implemented strategy is showed in Table 5.14. The HD simulation time is 3 orders of magnitude higher than the one needed by the strategies proposed in this article, as showed in Table 5.11. Moreover, the time grows as the sub-vector's size is reduced. This occurs because: (i) the number of sub-vectors generated in the database creation step is larger and (ii) we may find a higher number of similar samples, which implies more distance computations. We may also note that strategies from 2.1 to 2.5 need approximately a 50% less time to compute predictions than strategies from 1.1 to 1.5. There are mainly two reasons behind this fact: (i) the database's size has been considerably reduced due to optimisation techniques and (ii) predictions are not recursively computed for each value of L_{SV}.

The results and discussion provided in this section show that the optimised method presented in Section 5.2.4 clearly outperforms the one presented in Section 5.2.3 in terms of accuracy and computational cost.

5.2.6 Performance Analysis of ZigBee Wireless Networks for AAL

In this section, we use our optimised 1D approach to analyse the performance of ZigBee-based wireless sensor networks (WSNs) in ambient assisted living (AAL) environments with the aim of drastically reducing the computational time required to obtain accurate simulation results [100]. The measurements were performed in a common apartment located in the neighbourhood of "La Milagrosa" in the city of Pamplona, Navarre (Spain). The apartment has approximately $65m^2$ and it consists of 2 bedrooms, 1 kitchen, 1 bathroom, 1 study room, 1 living room and 1 small box room, as it can be seen in Figure 5.12. The dimensions of the scenario are $9.05m$ x $7.255m$ x $2.625m$. In order to obtain accurate results with the 3D ray launching simulation tool, the real size and material properties (dielectric constant as well as conductivity) of furniture such as chairs, tables, doors, beds, wardrobes, bath, walls, etc. have been taken into account. In Table 5.15 the properties of the materials with greater presence in the scenario are listed.

Table 5.15: Properties of the most common materials for our 3D ray launching simulations.

Material	**Permittivity** (ε_r)	**Conductivity** (σ)[S/m]
Air	1	0
Plywood	2.88	0.21
Concrete	5.66	0.142
Brick wall	4.11	0.0364
Glass	6.06	10^{-12}
Metal	4.5	4×10^7
Polycarbonate	3	0.2

Scenarios like the apartment studied in this article are expected to need a large number of wireless devices deployed in quite small areas to enable novel AAL applications. In order to emulate a dense wireless network, a 56-device ZigBee network has been distributed throughout the apartment. Those devices could be either static or mobile and wearable devices. Figure 5.12 shows the distribution of devices: ZigBee End Devices (ZED) are represented by red dots and ZigBee Routers (ZR) by green dots. To provide a more realistic setup, non-ZigBee wireless devices have been also placed within the scenario as to analyse their coexistence in terms of inter-system

Figure 5.12: (a) 3D simulated scenario with wireless devices shown in coloured dots, (b) top view of the scenario. Reprinted from [100].

interferences. For that purpose, a WiFi access point (black dot in Figure 5.12) and a WiFi device that could be a laptop or a smart phone (blue dot in Figure 5.12) have been placed in the living room and in the study room respectively. The characteristics of antennas and devices used in the scenario, such as transmission power level and radiation pattern, have been defined as those typical for real devices. Table 5.16 shows the main parameters used in the simulations for both ZigBee and WiFi devices.

Table 5.16: Parameters for the 3D ray launching simulations.

	LD	HD
Frequency	2.4GHz	2.4GHz
ZigBee Transmitted power	0dBm	0dBm
WiFi Transmitted power	20dBm	20dBm
Antenna gain	1.5dBi	1.5dBi
Diffraction enabled	No	Yes
Horizontal plane angle resolution ($\Delta\Phi$)	2º	1º
Vertical plane angle resolution ($\Delta\theta$)	2º	1º
Maximum permitted reflections	3	7
Cuboids resolution	10cm × 10cm × 10cm	10cm × 10cm × 10cm

Next, a measurement campaign was carried out to validate and compare real measurements with estimated values obtained by the 3D ray launching algorithm. More details about the validation of the results obtained by the 3D ray launching method can be found in [100]. The next step is to proceed with the performance analysis of the dense ZigBee network deployed within the scenario. For that purpose, the estimations obtained by both the HD Ray launching method and the hybrid LD + CF method are shown. Finally, we

summarise how these results have been obtained and we evaluate, in detail, the difference between the computational costs of HD simulations and the LD + CF approach.

With the information of $Database_{16}$, we have created five LD knowledge databases with L_{SV} = 11, 9, 7, 5, and 3 and their five HD counterparts. In this study, the scenario under analysis has similar characteristics to those which have been used for the creation of the database. Its dimensions and density are summarised in Table 5.17. Rows and Columns refer to the planar dimensions of the scenario (*i.e.* the dimensions (x,y) of the matrices analysed/used by the CF method). The 'Layers' row indicates the number of matrices, that is, the height (z) of the scenario. Finally, the 'Density' row shows the percentage of the volume of the scenario occupied by obstacles.

Table 5.17: Dimensions and characteristics of the scenario.

Rows	**Cols**	**Layers**	**Density %**
91	73	27	9,497

Table 5.18: Simulation strategies: subvector length L_{SV} and aggregator value k.

Strategy	**Prediction Strategy**
1.1	$L_{SV} = 3$, $(k = 25)$
1.2	$L_{SV} = 5$, $(k = 25)$
1.3	$L_{SV} = 7$, $(k = 25)$
1.4	$L_{SV} = 9$, $(k = 25)$
1.5	$L_{SV} = 11$, $(k = 25)$

The accuracy and performance of our approach has been analysed using different prediction strategies (*cf* Table 5.18). For instance, Strategy 1 uses the previously created knowledge database with L_{SV} = 11 to compute predictions; in Strategy 2, we use L_{SV} 9, and so on. In all cases, an aggregator value k= 25 is used. Hence, for each missing value of an LD simulation, the CF approach finds the k= 25 most similar subvectors and computes their average to predict the missing value.

Without loss of generality and for the sake of brevity, we have randomly selected 24 sensors from the studied scenario, depicted in Figure 5.12, and we have compared the simulation results obtained by HD simulations and our hybrid approach LD + CF. Table 5.19 reports the obtained results. More

Table 5.19: Average results (over all layers of the scenario) of the hybrid LD + CF versus the HD approach.

	Spars. %	Time HD	Time LD	Time CF	Time LD+CF	Best Strat.	MAE_R	MAE_P	σ_P
Sim-1	47,221	90650	2981	112	3116	1	74,707	11,825	9.05
Sim-2	49,933	52309	3543	187	2783	4	75,181	13,252	10.22
Sim-3	30,291	64959	2953	55	3302	1	73,091	9,796	7.47
Sim-4	49,071	53709	3168	106	2948	1	72,737	10,812	8.36
Sim-5	41,458	58763	3197	116	3053	3	75,051	10.52	8.56
Sim-6	42,722	83949	3004	88	2730	1	72,509	10,281	7.57
Sim-7	33,574	90189	2596	84	3362	1	72,439	9.88	7.12
Sim-8	34,711	76368	3247	102	3211	2	73,408	9,914	7.49
Sim-9	39,952	94287	2842	103	3153	1	74,996	9,814	7.82
Sim-10	20.03	69627	2937	79	2889	2	78,893	7,797	6.25
Sim-11	24,915	80113	3315	99	2947	3	78.38	9.72	7.67
Sim-12	35,317	79136	2882	98	3246	1	70,076	13,394	8.92
Sim-13	35,266	56627	2608	101	2982	1	72,139	12,021	8.83
Sim-14	43,428	59785	2906	115	2776	1	73,255	14,092	10.1
Sim-15	43,154	46410	2870	138	2558	4	76,506	12,005	9.11
Sim-16	38,066	67031	2642	95	2971	1	72.24	13,756	9.94
Sim-17	45,551	50729	3278	123	2509	2	76,485	13,294	9.89
Sim-18	42,902	83750	3109	120	3106	1	69,985	14,618	10.73
Sim-19	55,942	50423	3050	168	2761	2	74,522	12,073	8.91
Sim-20	54,066	58216	2810	186	3109	3	72,824	11,724	8.8
Sim-21	57,699	55179	3183	201	2985	3	74,018	10,716	8.29
Sim-22	44,755	74635	2606	99	2869	1	71,712	14,338	10.88
Sim-23	38,979	59727	2877	56	2983	1	64,586	16,998	11.45
Sim-24	48,459	75307	3245	79	2794	1	74,443	10,435	7.86

concretely, Table 5.19 reports *Sparseness* that is the % of empty values in the LD simulation, *Time HD, Time LD, Time CF*, and *Time LD + CF* which are the time in seconds of each method (for LD + CF, we report the worst time, not the average), *Best Strategy* which is the strategy that performed better, MAE_R which is the mean absolute error considering that null cells are kept empty (*i.e.* LD versus HD), MAE_P which is the mean absolute error considering that null cells are replaced by the values predicted by the CF method (*i.e.* LD + CF versus HD), and σ_P that is the standard deviation of MAE_P. We can observe that the results obtained by our hybrid method (MAE_P) clearly outperform those of the LD simulation alone (MAE_R). Moreover, this significant improvement requires very little time and, in addition, the low values of σ_P indicate that predictions are stable and reliable. In terms of computational cost, our hybrid approach requires

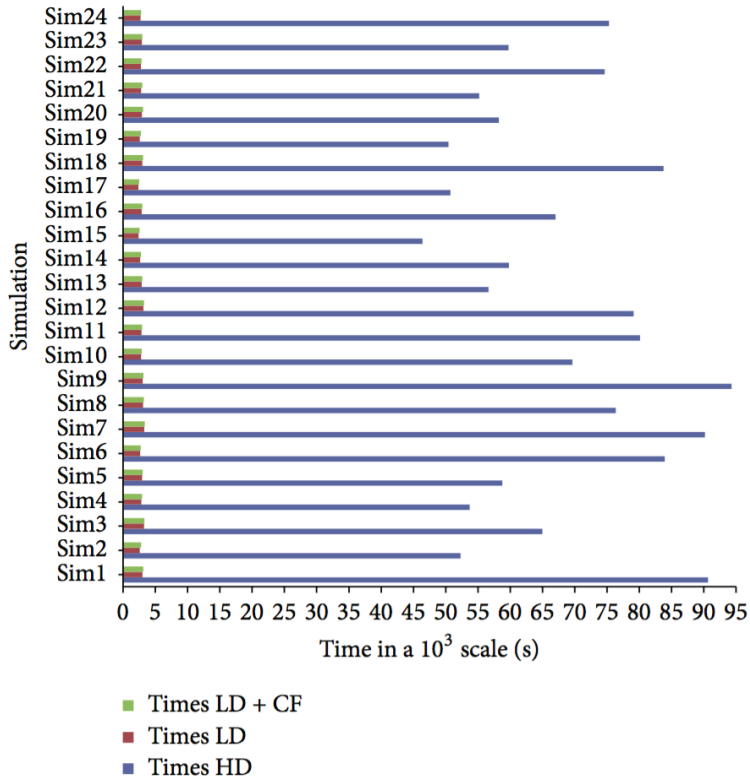


Figure 5.13: Time comparison of the different approaches (*i.e.* LD, LD + CF, and HD) for each simulation. Reprinted from [100].

10 to 20 times less time than HD simulations and the cost of the CF method is almost negligible (*cf* Table 5.19 column "Time LD + CF" and Figure 5.13). Also, we observe that sparseness has an adverse effect on the prediction accuracy of all methods. However, the knowledge database is also an important factor. For instance, if we compare Sim21 and Sim23, we observe that Sim21 has a higher sparseness value but the prediction's accuracy (*i.e.* MAE_P) is far better than in Sim23, which exhibits the opposite behaviour.

The relation between MAE and simulation/prediction time is depicted in Figure 5.14. LD predictions (in red) correspond to MAE_R values in Table 5.19. LD + CF results (in green) show that the computational time is slightly increased with respect to LD simulations. However, the MAE is reduced between 8 and 10 times. Finally, HD simulations (in blue) require is 10 to 20 times more time than the rest of approaches. Clearly, the best trade-off between MAE and time is obtained by our proposed hybrid LD + CF method. Hence, we may conclude that our method outperforms the

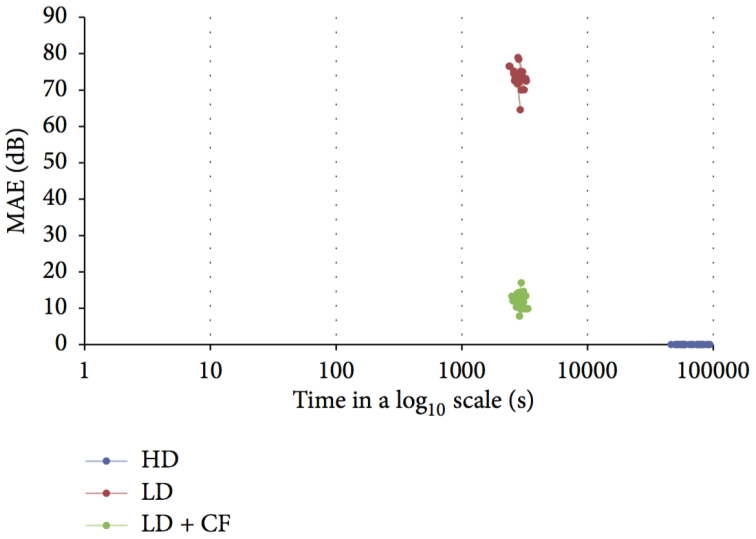


Figure 5.14: Relationship between MAE and execution time for LD, LD + CF, and HD approaches. Reprinted from [100].

others when we consider both accuracy and computational cost.

5.2.7 Optimal Parameter Estimation for Wireless Signal Analysis in Context-Aware Scenarios: A brief Study on the Number of Reflections Parameter

In this section, we study the impact of the 3D ray launching approach parameters when HD and LD simulations are performed. The aim of this work is to discover which is the best possible configuration given a particular scenario. More concretely, we study the number of reflections parameter and its impact on post-processing techniques such as hybrid methods based on CF techniques [37].

As previously stated in Chapter 2, the RL technique is a deterministic method for radio propagation analysis based on Geometrical Optics (GO) and the Geometrical Theory of Diffraction (GTD), with its extension to the Uniform Theory of Diffraction (UTD). The principle of the algorithm is that when a ray finds an object in its path, two new rays are created: a reflected ray and a refracted ray. These rays have new angles provided by Snell's law [76]. Next, the parameters of transmission T and reflection R are calculated, as well as the angle of incidence Ψ_i and Ψ_t. Thereafter, the new angles (θ_r,Φ_r) of the reflected wave and (θ_t,Φ_t) of the transmitted wave can be calculated. Finally, the new angles for the reflected and transmitted wave

(*i.e.* rays) can be computed. The number of reflections (nr) considered in the RL technique are highly relevant, as shown in [9], where the authors performed a convergence analysis of the RL approach, achieving satisfying results but at the cost of increasing the computational time [176]. In the RL technique, the standard range of values of nr goes from 1 to 7. Figure 5.15 represents an example of different cases of nr considered within an indoor scenario.

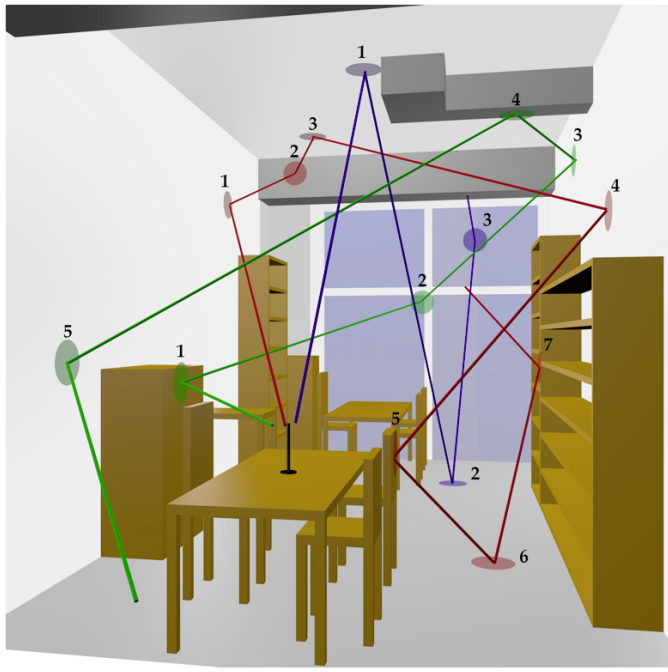


Figure 5.15: Schematic representation of the behaviour of rays with $nr = 3$ (in blue), $nr = 5$ (in green) and with $nr = 7$ (in red) within an indoor scenario. Reprinted from [37].

In this section, we study the impact of this parameter in post-processing techniques. For that purpose, we use the Optimised 1D approach presented Section 5.2.5. Simulations and measurements have been performed in a room of the Jerónimo de Ayanz Communications Research Center of the Public University of Navarre. We have simulated this scenario several times with a different nr value. The main characteristics of these simulations are summarised in Table 5.20.

In this case, we used the information of $Database_{16}$ to create five LD knowledge databases with $L_{SV} = 3$, 5, 7, 9 and 11, and their five HD counterparts. The simulations' obtained results in LD and HD have been

Table 5.20: Characteristics of the simulations performed with the RL method. Note that there is only one HD simulation and thus, one value for its computational cost.

Number of reflections	Rows	Cols	Time (s) HD	Time (s) LD	Density %	Sparseness %
3	130	70	150000	1288	1.24	14.51
4	130	70	150000	1811	1.24	5.47
5	130	70	150000	2563	1.24	1.71
6	130	70	150000	3642	1.24	0.73
7	130	70	150000	5127	1.24	0.60

used to test our algorithms. With the aim to analyse the accuracy and performance of our approach, we have created five prediction strategies, shown in Table 5.21. This time, the aggregator value is $k = 25$.

Table 5.21: Summary of prediction strategies.

Strategy	Prediction Strategy
1.1	$L_{SV} = 3$, $(k = 25)$
1.2	$L_{SV} = 5$, $(k = 25)$
1.3	$L_{SV} = 7$, $(k = 25)$
1.4	$L_{SV} = 9$, $(k = 25)$
1.5	$L_{SV} = 11$, $(k = 25)$

In Figures 5.16 and 5.17, we can observe that sparseness and MAE, respectively, decrease at a very similar pace. However, measurements performed by the RL method with $nr > 3$ seemed to introduce noise in the simulation. In order to ascertain that, for each simulation (*i.e.* from $nr = 3$ to $nr = 7$) we computed the MAE without taking into account the values measured with $nr = 3$ (*i.e.* only considering the information added due to the increase in the number of reflections). The outcome of this comparison is depicted in Figure 5.18.

Finally, we study the accuracy of the CF prediction approach concerning the nr parameter. Figure 5.19(a) shows the MAE of predictions only considering the null values of each scenario. Therefore, the percentage of values predicted for each scenario corresponds to Table 5.20, column "sparseness". In the case depicted in Figure 5.19(b), we compute the MAE considering the values introduced by the RL method as predictions. Hence, the percentage

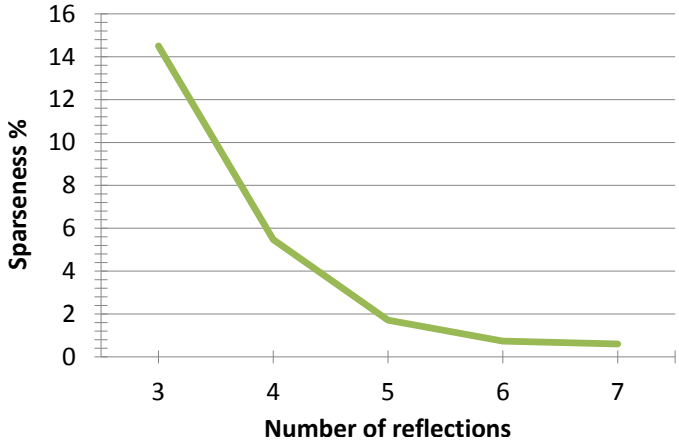


Figure 5.16: Sparseness of the LD simulations. Reprinted from [37].

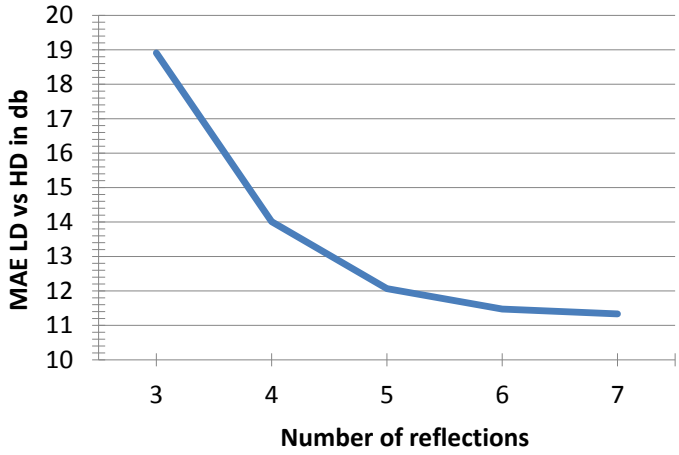


Figure 5.17: MAE comparison between LD and HD values for each scenario. Reprinted from [37].

of values predicted by the RL method in a scenario S with $nr = p$ corresponds to the difference between the sparseness of S with $nr = 3$ and with $nr = p$, where $p \geq 3$. The rest of predictions (*i.e.* for the remaining empty cells) are computed by our CF approach.

As observed in Figures 5.16 and 5.17, increasing nr resulted in more accurate (*i.e.* compared to the default null value) and dense LD simulations. Note that for $nr = 7$ the sparseness of the simulation is practically reduced to 0. On the contrary, the time required to perform a LD simulation increased with nr, as seen in Table 5.20, column "Time LD". Thereafter, we studied the impact of the nr parameter in the CF approach. However, contrary

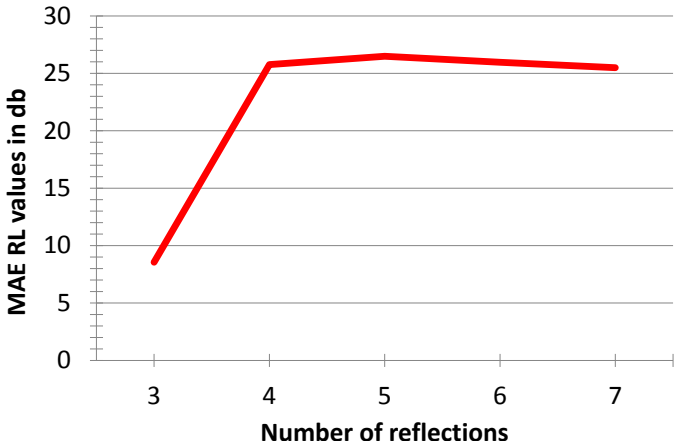


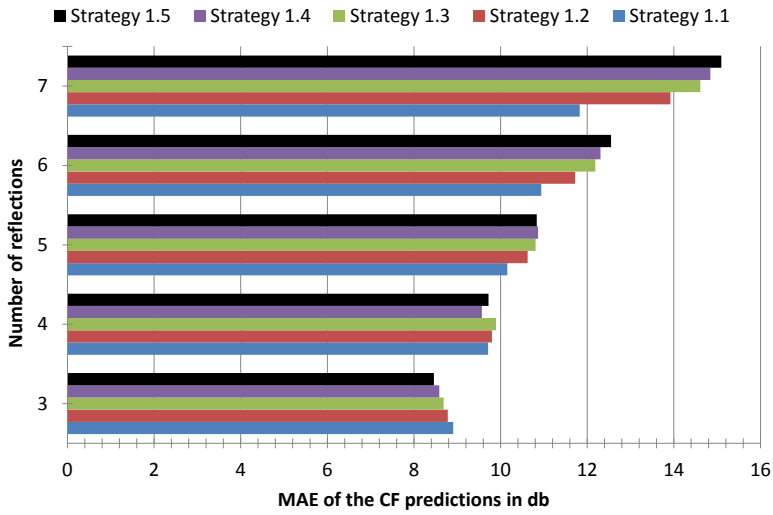
Figure 5.18: MAE of the additional values introduced by the RL simulation when nr is increased. Reprinted from [37].

to initial expectations, the accuracy of predictions decreased when nr was increased, as observed in Figure 5.19(a). Moreover, when considering both RL+CF values as predictions (*cf* Figure 5.19(b)) the MAE increased severely because the vast majority of values were predicted by the RL method. Note that for $nr \geq 5$, the percentage of sparseness (*cf* Table 5.20) is so low that the MAE depicted in Figure 5.19(b) is similar to the one showed in Figure 5.18. The loss of accuracy achieved by the RL method for high values of nr is related with the poor accuracy of the estimated LD values. The method considers each ray that reaches a new zone in the scenario (*i.e.* a zone with an error/null measurement), as a valid value. However, despite the increase of nr, there are zones reached by only few rays, so that RL measurements are less accurate than the patterns predicted by the CF method.

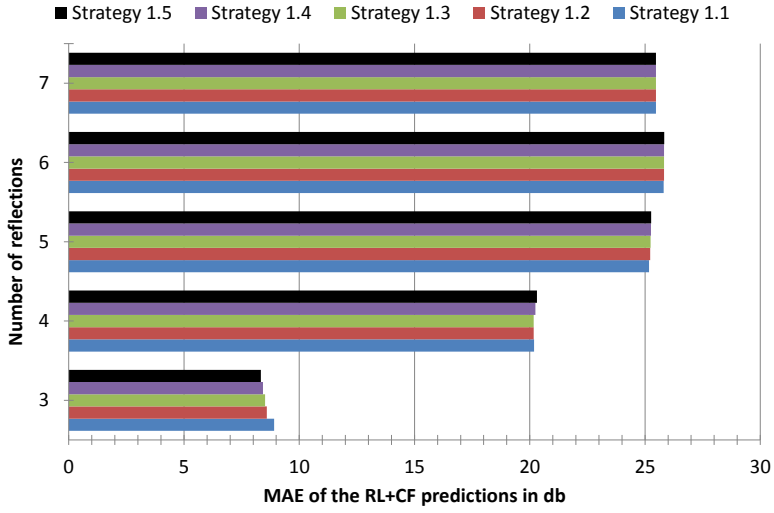
we could not find similar patterns in our database due to

Figure 5.19(a) shows that strategies with lower L_{SV} perform better than the rest, except for $nr > 4$. This behaviour indicates that the noise introduced by the RL method when $nr > 4$ prevents finding similar patterns with accuracy. Note that the higher the L_{SV}, the more difficult is to find similar patterns, especially if we consider biased ones.

After observing the outcomes of our experiments, we may conclude that the values measured by the RL method with $nr > 3$ introduced noise in the simulations. Therefore, correlations between LD and HD patterns were biased, and the quality of the predictions decreased when the number of RL values was increased.



(a) MAE considering only the null values of each simulation.



(b) MAE taking into account RL values and CF predictions. We compute predictions considering the null values of the simulation with nr=3.

Figure 5.19: MAE obtained by each strategy for the different scenarios in regards to CF and RL+CF values. The lower the better. Reprinted from [37].

5.2.8 Optimised Two-Dimensional Hybrid Simulation Technique

With the aim to better capture the behaviour of the rays simulated by a set of emitters in the 3D ray launching method we presented a two-dimensional RL+CF approach in [32]. This method has a database creation step as well as a prediction step (like in the one-dimensional approach), which are described in Sections 5.2.8.1 and 5.2.8.2, respectively.

5.2.8.1 2D Database creation

First, we create a knowledge database that will be later used to predict missing values in LD simulations. Each scenario obtained by 3D ray launching techniques is modelled by a matrix $M_{n\times p}$ and is managed as follows:

The scenario represented by the matrix

$$M_{n\times p} = \begin{bmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,p} \\ m_{2,1} & m_{2,2} & \dots & m_{2,p} \\ \vdots & \vdots & & \vdots \\ m_{n,1} & m_{n,2} & \dots & m_{n,p} \end{bmatrix}$$

is divided into a set SM of sub-matrices:

$$SM = \{sm_{1,1}, sm_{1,2}, \cdots, sm_{1,(p-q+1)}, \cdots, sm_{(n-q+1),1},$$

$$sm_{(n-q+1),2}, \cdots, sm_{(n-q+1),(p-q+1)}\}$$

where each sub-matrix $sm_{i,j}$ is a squared matrix of size $q \times q$, so that,

$$sm_{i,j} = \begin{bmatrix} m_{i,j} & \dots & m_{i,(j+q-1)} \\ \vdots & & \vdots \\ m_{(i+q-1),j} & \dots & m_{(i+q-1),(j+q-1)} \end{bmatrix}, \forall i \in [1, n-q], \forall j \in [1, p-q]$$

Figure 5.20 shows an example of the creation of the knowledge database with squared sub-matrices of size 3×3.

5.2.8.2 2D Values prediction

Given a LD simulation S with missing values (*i.e.* empty cells resulting from low angular resolution) our aim is to predict them so that the resulting values are as similar as possible to those obtained in HD simulations.

We assume that the values in S are normalised and that a pair of knowledge databases DB_{LD} and DB_{HD} with LD and HD patterns represented by

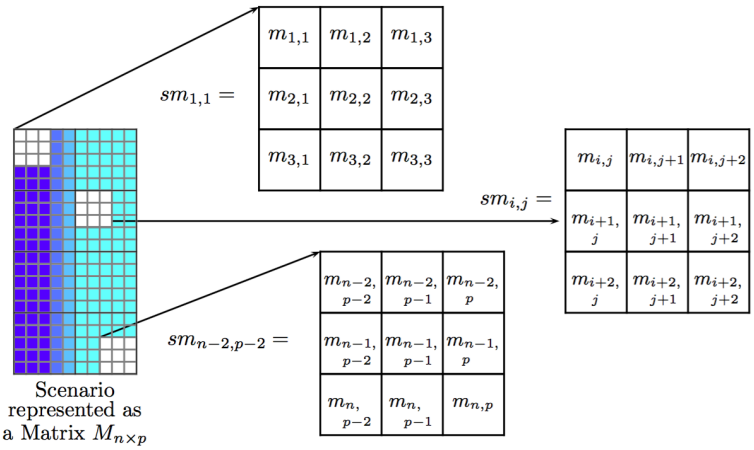


Figure 5.20: Database creation with $3x3$ sub-matrices. Reprinted from [32].

matrices of size $q \times q$ have been selected. Then, for every sub-matrix $sm_{i,j}$ of size $q \times q$ from S containing missing values, we compute the k closest sub-matrices $CSM^{LD} = \{csm_1, csm_2, .., csm_k\}$ in DB_{LD}. Next, we determine their corresponding HD patterns $CSM^{HD} = \{csm'_1, csm'_2, .., csm'_k\}$ in DB_{HD} and we compute its average as $\bar{csm'} = \frac{1}{k}\sum_{i=1}^{k} csm'_i$. We use the values of the sub-matrix $\bar{csm'}$ as the prediction of the missing values. Like in the 1D-approach, we use the Euclidean distance over non-missing values to determine the closest sub-matrices. In order to increase the quality of the predictions, we only consider sub-matrices having, at most, one missing value.

5.2.8.3 2D Optimisations

The complexity of our method can be analysed as follows: each LD simulation is represented by a $n \times m$ matrix, which is decomposed in a collection of squared sub-matrices such that $SM = \{sm_{1,1}, sm_{1,2}, \cdots, sm_{1,(p-q+1)}, \cdots, sm_{(n-q+1),1}, sm_{(n-q+1),2}, \cdots, sm_{(n-q+1),(p-q+1)}\}$, which have to be compared with the LD database. Therefore, the final cost of our method is $d*(n-(q-1)*(m-(q-1))$, where d is the number of sub-matrices existing in the LD database. We have implemented an optimisation structure similar to the one presented in Section 5.2.4, to reduce the computational cost of our approach. Such structure has the same number of rows and columns than the LD database and each row corresponds to the same sub-matrix in both structures. The sub-matrices are decomposed and

stored as one-dimensional vectors in the database to ease the process so that sm_i corresponds to the i^{th} sub-matrix of the database. Each column of the optimisation structure stores the sum of the sub-matrix's elements after subtracting the v^{th} element, corresponding to the column position, where $v \in [1, q * q]$. Given a sub-matrix sm_i, we use Equation (5.3) $q \times q$ times (*i.e.* one for each sub-matrix's value) to obtain its optimisation values O_{sm_i}. Subsequently, for each sub-matrix sm_a, we sum its values and compare them with the values of the v^{th} column in the optimisation structure, where v corresponds to the position of the empty cell in sm_a. Next, we define a threshold and we avoid distance computations if the aforementioned comparison falls outside the threshold. Note that this database is created offline and hence, the cost of this process is negligible. By performing such optimisations, we clusterise the database and reduce considerably the cost of the search process. For further details about the definition of the method and its computational cost in respect to the one-dimensional approach, see [32].

$$O_{sm_{i_v}} = \left(\sum_{j=1}^{q*q} sm_{i_j} \right) - sm_{i_v} \tag{5.3}$$

Table 5.22: Two-dimensional optimisation structure. Each row stores the corresponding optimisation values for each sub-matrix, computed as described in Equation (5.3).

O_{sm} \ Values of v	1	...	$q * q$
O_{sm_1}	$O_{sm_{1_1}}$	...	$O_{sm_{1_{q*q}}}$
...	...	...	...
O_{sm_d}	$O_{sm_{d_1}}$	...	$O_{sm_{d_{q*q}}}$

5.2.9 Two-Dimensional Approach in Context-Aware Scenarios

In this section, we have extended the experiments showed in Section 5.2.3 to include our two-dimensional approach. Therefore, we have created two sets of databases using $Database_{12}$. First, three 1D LD knowledge databases with vector lengths $L_{SV} = 3, 5$, and 7 and their three HD counterparts. Second, three two-dimensional LD databases with $q = 2, 3$ and 4 and the

corresponding HD ones. We have applied our solution on 4 LD simulations (not included in the knowledge database) with diverse dimensions, density (*i.e.* percentage of occupied space) and sparseness (*i.e.* percentage of missing values), as depicted in Table 5.23. In each one of the scenarios the transmitter is placed in different locations, to obtain generalizable results independent of antenna location. The specific location for each scenario is given in column Row/Column/Layer in Table 5.23. With the aim to analyse the versatility, accuracy and performance of our approach, we have tested two different types of prediction strategies, shown in Table 5.24. The first family (*i.e.* Strategies from 1D.1 to 1D.3) corresponds to the 1D-approach. The second, which includes strategies from 2D.1 to 2D.3, corresponds to the two-dimensional method. For instance, in Strategy 1D.1 we first predict the missing values with the LD knowledge database of vectors with $L_{SV} = 7$. If all missing values are not filled, we next apply the LD knowledge database with $L_{SV} = 5$, and if there remain missing values we apply the LD knowledge database with $L_{SV} = 3$. In Strategy 2D.1, we use sub-matrices with $q = 2$ to compute predictions. Obviously, each family of strategies uses their corresponding 1D or 2D knowledge databases. In all cases we consider an aggregator value (*i.e.* number of closest patterns that will be used to compute the prediction) $k = 100$.

Table 5.23: Test simulations features.

Rows	**Cols**	**Layers**	**Source Position R,C,L**	**Time (s) HD**	**Time (s) LD**	**Density %**	**Spars. %**
70	50	30	10,42,21	64139	1040	5,166	2,203
90	60	30	10,42,13	80004	1228	6,223	7,092
175	80	40	46,50,14	52583	1765	6,696	17,051
196	136	38	20,80,5	48910	2037	1,041	37,636

Table 5.24: Summary of prediction strategies.

Strategy	**Prediction Strategy**
1D.1	$L_{SV} = 7 \rightarrow L_{SV} = 5 \rightarrow L_{SV} = 3$, $(k = 100)$
1D.2	$L_{SV} = 5 \rightarrow L_{SV} = 3$, $(k = 100)$
1D.3	$L_{SV} = 3$, $(k = 100)$
2D.1	$q = 2$, $(k = 100)$
2D.2	$q = 3$, $(k = 100)$
2D.3	$q = 4$, $(k = 100)$

The prediction quality of our CF prediction strategies has been compared with other four well-known methods, namely linear interpolation, average simulation value per layer, average simulation value per row, and average simulation value per column. In Table 5.25 and Table 5.26 we show the MAE results and times, respectively, of the aforementioned methods for each simulation. Note that Table 5.26 shows the worst layer prediction time, which is the actual total cost of our method. As an example, the simulation results of one of the implemented indoor test is depicted in Figure 5.21. The results correspond to different simulation techniques, from High Definition, to Low Definition + CF and finally only LD.

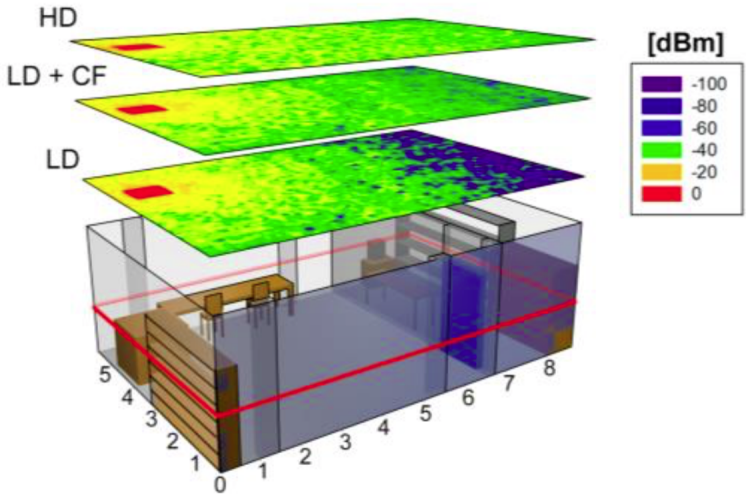


Figure 5.21: Received power level estimation when employing HD (top), LD+CF (middle) and LD (bottom). The scenario corresponds to Sim B. Reprinted from [32].

The outcomes showed in Table 5.25 indicate that our strategies achieve lower/better MAE values than the rest of methods for Sim-A, Sim-C and Sim-D and that Strategy 2D.3 is the overall best method. In opposition to the results of Sim-A, Sim-C and Sim-D, the MAE obtained by mean-based methods in Sim-B is lower/better than that achieved by our proposed strategies. This is due to the fact that Sim-B is very homogeneous and the range of measurements is very short, thus benefiting mean-based methods. The outcomes obtained by 2D strategies (*i.e.* Strategy 2D.3) outperform those obtained by 1D strategies for Sim-A, Sim-B, Sim-C and Sim-D (*cf* Table 5.25). Therefore, we may conclude that 2D strategies represent the wave propagation behaviour with more accuracy. Moreover, as observed in Table 5.26, 2D strategies require less time to perform predictions. This occurs be-

Table 5.25: MAE (in dB). The lower the better. The row "Raw values" shows the MAE when missing values are kept empty. The column "All" is the average of all simulations weighted by their size.

	Sim-A		Sim-B		Sim-C		Sim-D		All	
Method	MAE	σ	MAE	σ	MAE	σ	MAE	σ	MAE	σ
Strategy 1D.1	6.47	0.75	9.91	0.79	6.99	0.52	10.70	1.40	9.25	1.05
Strategy 1D.2	6.48	0.79	10.14	0.92	7.06	0.60	11.14	1.31	9.53	1.04
Strategy 1D.3	6.21	0.86	8.16	1.02	7.16	0.56	13.38	1.09	10.62	0.90
Strategy 2D.1	5.88	1.58	6.85	0.48	8.42	0.96	15.11	1.06	11.82	1.01
Strategy 2D.2	5.69	1.56	7.01	0.55	7.19	0.56	11.02	0.59	9.19	0.84
Strategy 2D.3	5.82	1.89	7.05	0.59	6.84	0.49	9.78	0.58	8.52	0.79
Linear Interpolation	23.03	4.36	22.88	3.14	24.08	2.83	24.91	2.21	23.98	2.60
Layer mean value	8.84	4.11	5.31	0.46	9.31	1.46	16.37	1.27	12.76	1.42
Column mean value	7.86	3.48	5.11	0.52	9.83	1.62	16.58	1.20	13.02	1.39
Row mean value	7.59	0.93	7.95	0.99	8.65	1.55	14.05	1.28	11.51	1.32
Raw values	56.26	2.09	45.67	1.08	62.07	1.59	73.19	1.15	66.42	1.33

Table 5.26: Times (in seconds) to predict all missing values

Time (in seconds)						
Simulation	**1D.1**	**1D.2**	**1D.3**	**2D.1**	**2D.2**	**2D.3**
Sim-A	17	19	17	15	16	16
Sim-B	54	54	35	42	49	56
Sim-C	152	167	179	132	141	143
Sim-D	473	479	531	361	396	416

cause the optimisation performed in the search process has greater impact in 2D databases, because they contain less patterns than 1D databases. Also, we observe that sparseness has an adverse effect on the prediction accuracy of all methods. This is especially apparent in Sim-D, which is very sparse. Overall, as it can be seen in Table 5.25 (column: "All"), Strategy 2D.3 outperforms the rest of strategies, and obtains the best prediction accuracy, which is about 25% better than the best method not proposed by us.

5.2.10 Optimised Two-Dimensional Approach in Medical Complex Scenarios

In this section, we analyse the performance of our two-dimensional approach in medical complex scenarios to predict the behaviour of signal propagation.

We use the method defined in Section 5.2.8, that considers the values of the neighbouring dimensions (*i.e.* 2D data instead of 1D). We compare 1D and 2D approaches in terms of accuracy, with the aim to ascertain which method achieves better outcomes. Two sets of databases have been created using the information of $Database_{16}$. First, five one-dimensional LD knowledge databases with vector lengths $L_{SV} = 3, 5, 7, 9$ and 11, and their five HD counterparts. Second, four two-dimensional LD databases with $q = 2, 3, 4$ and 5 and the corresponding HD ones. The emergency floor of the Hospital de Navarra (Navarra, Spain) has been selected as the simulation scenario. As previously stated in Section 5.2.5, this scenario has rooms with different materials and medical devices. See Table 5.27 for the exact data. From this scenario, a total of four simulations, namely Sim-1, Sim-2, Sim-3 and Sim-4, have been performed using different configurations of the position of transmitters, showed in Figure 5.22. The simulations' obtained results in LD and HD have been used to test our algorithms.

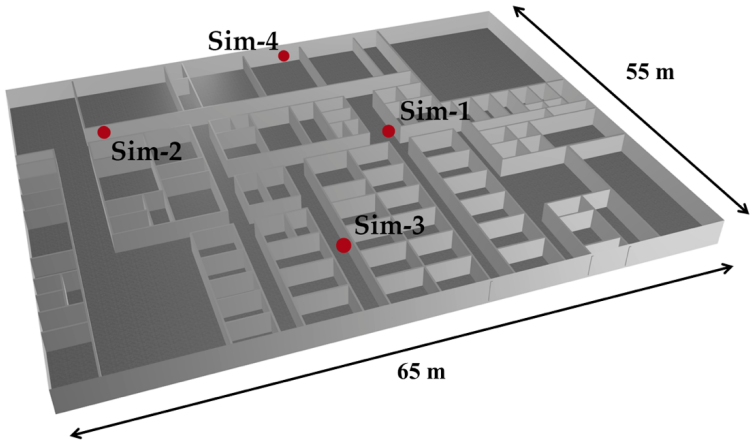


Figure 5.22: Detail of the scenario with the location of signal transmitters. Each transmitter was used separately for its corresponding simulation. Reprinted from [36].

We have created two families of strategies, shown in Table 5.28, to analyse the accuracy of our approach. The first family (*i.e.* Strategies from 1.1 to 1.5) corresponds to the one-dimensional approach presented in Section 5.2.5. The second, which includes strategies from 2.1 to 2.4, corresponds to the two-dimensional defined in Section 5.2.8. For instance, in Strategy 1.1 we predict the missing values with the LD database of one-dimensional sub-vectors with $L_{SV} = 3$. In Strategy 2.1, we use sub-matrices with $q = 2$

to compute predictions. Obviously, each family of strategies uses their corresponding one-dimensional or two-dimensional knowledge databases. We selected an aggregator value $k = 25$ for all strategies.

Table 5.27: Test simulations features. 'Time LD' and Time HD' are the seconds needed to simulate each scenario in LD and HD respectively.

	Rows	**Cols**	**Time (s) HD**	**Time (s) LD**	**Density %**	**Sparseness %**
Sim-1	62	48	68915	463	7.86	18.11
Sim-2	62	48	29664	504	7.86	17.21
Sim-3	62	48	46233	470	7.86	14.34
Sim-4	62	48	71872	534	7.86	42.14

Table 5.28: Summary of prediction strategies.

Strategy	**Prediction Strategy**
1.1	$L_{SV} = 3$, $(k = 25)$
1.2	$L_{SV} = 5$, $(k = 25)$
1.3	$L_{SV} = 7$, $(k = 25)$
1.4	$L_{SV} = 9$, $(k = 25)$
1.5	$L_{SV} = 11$, $(k = 25)$
2.1	$q = 2$, $(k = 25)$
2.2	$q = 3$, $(k = 25)$
2.3	$q = 4$, $(k = 25)$
2.4	$q = 5$, $(k = 25)$

In Table 5.29, we show the MAE results of the aforementioned methods for each simulation. Figure 5.23 shows an example of a graphical representation of simulation results for the estimation of received power at all locations of the implemented indoor test scenarios. The results correspond to different simulation techniques, from HD, to LD+CF and finally only LD. The LD+CF achieves a remarkable outcome in terms of received power level estimation.

Strategies from 2.2 to 2.4 achieve lower/better MAE than strategies form 1.1 to 1.5 (*cf* 5.29). Moreover, two-dimensional strategies obtain the best

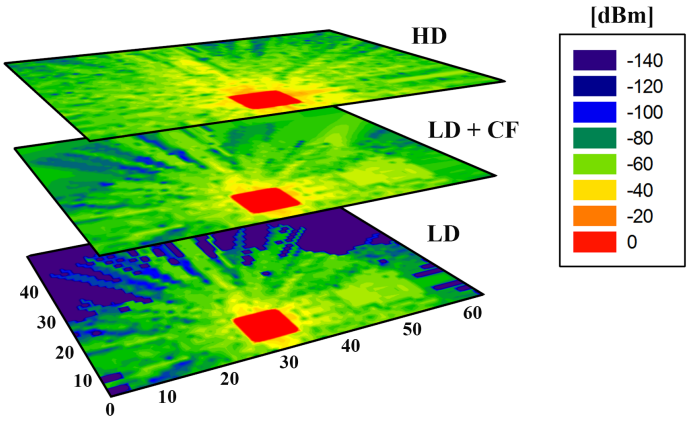


Figure 5.23: Received power level estimation when using HD (top), LD+CF (middle) and LD (bottom). Reprinted from [36].

Table 5.29: Accuracy of the predicted values. MAE is showed in dB. The lower the better. The row "Raw values" shows the MAE when missing values are kept empty. The column "All" is the average of all simulations' outcomes, weighted by their corresponding sparseness.

	Sim-1		Sim-2		Sim-3		Sim-4		All	
Method	MAE	σ	MAE	σ	MAE	σ	MAE	σ	MAE	σ
Strategy 1.1	10.07	7.84	8.42	7.11	12.01	8.61	14.90	10.11	12.28	8.86
Strategy 1.2	7.80	5.58	6.66	5.47	9.47	7.21	11.08	8.37	9.35	7.09
Strategy 1.3	7.44	5.53	6.58	5.47	8.78	6.68	10.45	7.94	8.87	6.80
Strategy 1.4	7.40	5.55	6.65	5.67	8.07	6.23	11.16	8.11	9.09	6.85
Strategy 1.5	7.41	5.50	6.77	5.56	7.85	5.99	10.88	7.52	8.95	6.51
Strategy 2.1	7.84	5.89	7.97	6.79	11.75	8.93	14.50	10.83	11.53	8.80
Strategy 2.2	7.70	5.56	7.12	6.62	7.75	6.26	8.70	6.99	8.06	6.52
Strategy 2.3	7.35	5.41	7.11	6.25	7.44	5.88	8.69	6.56	7.93	6.17
Strategy 2.4	7.26	5.30	7.92	6.66	7.27	5.79	8.39	6.43	7.90	6.15
Raw values	63.13	9.48	62.88	7.91	66.20	10.02	73.08	11.33	68.13	10.12

results for each simulation in terms of MAE and σ (except for Strategy 2.1), which implies that predictions are not only better but more stable/reliable.

However, although the results showed in Table 5.29 correspond to average values we note that the best outcomes are not always obtained by the same strategy. For instance, in Sim-2, Strategies from 1.2 to 1.5 obtain more

accurate results than Strategies from 2.1 to 2.4. In this case, the position of Sim-2's empty cells and the patterns found in the LD database increase the amount of error propagated by two-dimensional strategies. This added noise should be reduced by means of a multivariate error propagation analysis, which will be left to future research. Table 5.29 also shows that Sim-4 obtains the worst predictions due to its high sparseness. Depending on the scenario's features, parameters such as k, sub-vector length, sub-matrix size, and its corresponding database might be tuned to enhance the prediction's accuracy. Overall, as showed in Table 5.29 (column: "All"), Strategy 2.4 outperforms the rest of strategies and obtains the best prediction accuracy, which is more than 10% better than the best strategy of the first family (*i.e.* Strategy 1.3). Therefore, the outcomes and discussion provided in this section show that the 2D-approach outperforms the 1D-approach in terms of accuracy.

5.3 Conclusions

Smart cities are gaining importance and their infrastructure can be used to improve the healthcare services provided to citizens. This is the philosophy of the so-called Smart Health concept [157].

In Section 5.1, we have proposed the idea of using recommender systems integrated with the sensing infrastructure of smart cities to promote citizen's healthy habits in real time. We provided an overview of the proposed system, its main components, and its operation. Moreover, we have created a statistically sound simulated dataset with real health information to test our proposal. In addition, a mobile application has been developed. Such application uses real data provided by the sensing infrastructure of the smart cities and crowdsourcing-based information provided by citizens who tested our system in real scenarios. Further versions will include gamification techniques [125, 89] to motivate user's participation. The experimental results demonstrate that our system could be deployed in real scenarios and provide good recommendations to citizens even with a high percentage of unknown data. Finally, several real life scenarios to show the usefulness of our approach have been discussed.

Nowadays, plenty of simulation techniques based on deterministic methods like Ray Tracing and Ray Launching are used to make better decisions on the deployment of antennas for particular purposes. However, simulations' quality largely depends on the number of rays, providing sufficient angular resolution in the ray launching process. The computational cost of simulations in high definition (HD) prevents their use in complex environ-

ments and their low definition (LD) counterparts are used. In Section 5.2, we have proposed several techniques based on CF to lessen the low quality problems of LD simulations. Such techniques outperform previous works in this field in terms of accuracy, precision and computational cost. Moreover, we show that our approaches obtain results very similar to those of HD in much less time. In Section 5.2.7, we analysed the incidence of the number of reflections in medium-size context-aware scenarios recreated by a RL approach. Our experiments showed that increasing the value of nr resulted in lower sparseness and better measurements in respect of values that could not be predicted with lower nr. However, although predictions were better than null/error measurements, they introduced noise in the simulation. In Section 5.2.6, we showed that radiopropagation in indoor complex AAL environments has strong dependence on the network topology, the indoor scenario configuration, and the density of users/devices within it. Moreover, we have shown that our Rl+CF approach enables us to reduce drastically the simulation time consumption, while the error of the estimations remains low. As a conclusion, using combined deterministic-CF techniques allows the estimation of radio-planning tasks in large, context-aware scenarios with a potentially large amount of transceivers.

Chapter 6

Conclusions

This chapter summarises the contributions of the dissertation. In addition, it sketches some lines for future work that arise from either partially reached goals or expected improvements.

Contents

The topics discussed in this dissertation focus on recommender systems applied to different kinds of fields such as e-commerce and context-aware scenarios. First, we provide an extensive background on Statistical Disclosure Control, Recommender Systems, Collaborative Filtering and Smart Health. Second, we propose and test a set of methods to efficiently deal with the most relevant problems of recommender systems, such as non-response (sparseness) and privacy. Extensive experiments show that our methods achieve a remarkable accuracy while protecting the privacy of the users involved, as well as minimizing sparseness of Collaborative Filtering datasets. Moreover, we propose a new metric to perform a more robust analysis of the outcomes. Finally, we analyse the urbanisation process that is taking place around the world and provide practical examples of smart-health applications enhanced with CF methods. We believe that our contributions will have a positive impact on many areas from Smart Transportation and sustainability to e-participation and Smart Healthcare, in which many efforts are being devoted.

6.1 Publications and Research Stays

The main publications supporting the content of this thesis are the following:

Journals

2014 | Fran Casino, Constantinos Patsakis, Josep Domingo-Ferrer, Domènec Puig and Agusti Solanas, **"A k-Anonymous Approach to Privacy Preserving Collaborative Filtering"**, Journal of Computer and System Sciences. 81 (6) 1000-1011 || CS, Q2, IF: 1.138

2016 | Peio Lopez-Iturri, Fran Casino, Erik Aguirre, Leyre Azpilicueta, Francisco Falcone, and Agusti Solanas **"Performance Analysis of ZigBee Wireless Networks for AAL through Hybrid Ray Launching and Collaborative Filtering"**, Journal of Sensors, vol. 2016, Article ID 2424101 || EEES, Q2, IF: 1.129

2016 | Fran Casino, Leyre Azpilicueta, Peio Lopez-Iturri, Erik Aguirre, Francisco Falcone and Agusti Solanas **"Optimised Wireless Channel Characterisation in Large Complex Environments by Hybrid Ray Launching-Collaborative Filtering Approach"**, Antennas and Wireless Propagation Letters, 2016 || EEES, Q2, IF: 2.533

2017 | Fran Casino, Constantinos Patsakis, Edgar Batista, Frederic Borràs and Antoni Martínez-Ballestè **"Healthy Routes in the Smart City: A Context-Aware Mobile Recommender"**, IEEE Software - In Press || CS, Q1, IF: 2.190 (2016)

Conferences

2013 | Fran Casino, Constantinos Patsakis, Josep Domingo-Ferrer, Domènec Puig and Agusti Solanas, **"Privacy Preserving Collaborative Filtering with k-Anonymity through Microaggregation"**, 10th IEEE International Conference on e-Business Engineering (ICEBE 2013), September 10-13, Coventry, United Kingdom. || CORE B

2013 | Fran Casino, Constantinos Patsakis, Domènec Puig and Agusti Solanas, **"On Privacy Preserving Collaborative Filtering: Current Trends, Open Problems and New Issues"**, 10th IEEE International Conference on e-Business Engineering (ICEBE 2013), September 10-13, Coventry, United Kingdom. || CORE B

2015 | Fran Casino, Leire Azpilicueta, Peio Lopez-Iturri, Erik Aguirre, Francisco Falcone, Agusti Solanas **"Hybrid-based Optimization of Wireless Channel Characterization for Health Services in Medical Complex Environments"**, The 6th International Conference on Information, Intelligence, Systems and Applications - IISA 2015, At Corfu, Greece. **BEST PAPER AWARD**

2015 Amaia Ortiz de Lejarazu, Peio Lopez-Iturri, Erik Aguirre, Leyre Azpilicueta, Francisco Falcone, Fran Casino, Agusti Solanas, **"Challenges in the Implementation of Context-Aware Scenarios within Emergency Rooms"**, The 6th International Conference on Information, Intelligence, Systems and Applications - IISA 2015, At Corfu, Greece.

2015 Fran Casino, Peio Lopez-Iturri, Erik Aguirre, Leyre Azpilicueta, Agusti Solanas, Francisco Falcone **"Dense Wireless Sensor Network Design for the Implementation of Smart Health Environments"**, International Conference on Electromagnetics Advanced Applications (ICEAA) 2015, Torino, Italy; 09/2015

2015 Fran Casino, Edgar Batista and Agusti Solanas **"Context-Aware Recommender for Smart Health"**, The 1st IEEE International Smart Cities Conference - ISC2 2015, Guadalajara, Mexico 10/2015

2016 Fran Casino, Peio Lopez-Iturri, Leyre Azpilicueta, Erik Aguirre, Francisco Falcone and Agusti Solanas **"Optimal Parameter Estimation for Wireless Signal Analysis in Context-Aware Scenarios: A Brief Study"**, The 7th International Conference on Information, Intelligence, Systems and Applications - IISA 2016, At Porto Carras, Greece

2016 Fran Casino, Peio Lopez-Iturri, Erik Aguirre, Leyre Azpilicueta, Francisco Falcone and Agusti Solanas **"Two-dimensional Collaborative Filtering Approach to Wireless Channel Characterization in Medical Complex Environments"**, The 2nd IEEE International Smart Cities Conference - ISC2 2016, Trento, Italy 09/2016

2017 Peio Lopez-Iturri, Fran Casino, Erik Aguirre, Leyre Azpilicueta, Francisco Falcone and Agusti Solanas **"Hybrid Ray Launching-Collaborative Filtering Approach for Wireless Propagation in Indoor Environments"**, IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, San Diego, California, July 9-14, 2017

Other publications co-authored by the candidate and related to Wireless Channel Characterisation or Smart Health, but not included in this thesis, are listed below:

Conferences

2014 Antoni Martínez-Ballestè, Oriol Casanovas-Marsal, Agusti Solanas, Fran Casino and Montserrat Garcia-Martinez, M., **"An autonomous system to assess, display and communicate the pain level in newborns"**, Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on , vol., no., pp.1-5, 11-12 June 2014.

2015 | Jaume Vergés-Llahí, Hamed Habibi, Fran Casino, Domènec Puig **"GAMEABLING: Platform of Games for People with Cerebral Palsy"**, European Project Space on Computer Vision, Graphics, Optics and Photonics EPS-VISGRAPP, pp.64-87, 2015. ISBN: 978-989-758-156-4 || CORE B

2015 | Edgar Batista, Frederic Borras, Fran Casino and Agusti Solanas, **"A Study on the Detection of Wandering Patterns in Human Trajectories"**, The 6th International Conference on Information, Intelligence, Systems and Applications - IISA 2015, At Corfu, Greece.

2015 | Edgar Batista, Fran Casino and Agusti Solanas **"Wandering Detection Methods in Smart Cities: Current and New Approaches"**, The 1st IEEE International Smart Cities Conference - ISC2 2015, Guadalajara, Mexico 10/2015

2016 | Peio Lopez-Iturri, Fran Casino, Erik Aguirre, Leyre Azpilicueta, Agusti Solanas and Francisco Falcone **"Analysis of Vehicular Connectivity in Smart Health Service Provision Scenarios"**, The 7th International Conference on Information, Intelligence, Systems and Applications - IISA 2016, At Porto Carras, Greece

2016 | Edgar Batista, Fran Casino and Agusti Solanas **"On Wandering Detection Methods in Context-Aware Scenarios"**, The 7th International Conference on Information, Intelligence, Systems and Applications - IISA 2016, At Porto Carras, Greece

Research Stays and Visits

During this doctoral thesis, the candidate performed two research stays at international universities listed below:

Oct 2016 - Feb 2017 | Visiting researcher in the Department of Information Science and Technology, ISCTE-IUL, Lisbon, Portugal.

Mar 2017 | Visiting researcher in the Department of Informatics, Piraeus University, Piraeus, Greece. Cryptacus ICT Cost Action IC1403.

Ongoing Research and Submitted Articles

2017 | Fran Casino, Leyre Azpilicueta, Peio Lopez-Iturri, Erik Aguirre, Francisco Falcone and Agusti Solanas **"Optimal Parameter Estimation for Wireless Signal Analysis in Context-Aware Scenarios"**, Antennas and Wireless Propagation Letters - Submitted || EEES, Q2, IF: 2.533 (2016)

2017 | Fran Casino and Agusti Solanas **"Handling Non-response in Collaborative Filtering Based Recommender Systems"**, Expert Systems with Applications - to be submitted || CS, Q1, IF: 3.928 (2016)

2017 | Fran Casino, Constantinos Patsakis and Agusti Solanas**"Privacy-Preserving Collaborative Filtering with Variable Group-Sized Microaggregation"**, IEEE Transactions on Knowledge and Data Engineering - to be submitted || CS, Q1, IF: 3.438 (2016)

2017 | Athanasios Zigomitros, Fran Casino, Constantinos Patsakis and Agusti Solanas **"A Survey on Privacy-Preserving Data Publishing of Relational Data"**, ACM Computing Surveys - to be submitted || CS, Q1, IF: 6.748 (2016)

Contributions to URV

2014 | Fran Casino **"Privacy-Preserving Recommender Systems for E-commerce & Health Services"**, 1st URV Doctoral Workshop in Computer Science and Mathematics

2015 | Fran Casino **"Recommender Systems with Privacy for Context-Aware Services"**, 2nd URV Doctoral Workshop in Computer Science and Mathematics

2015 | Fran Casino **"Efficient Wireless Channel Characterisation for Context-Aware Scenarios"**, 3rd URV Doctoral Workshop in Computer Science and Mathematics

TEACHING | Disseny de Xarxes (Laboratories) (2015, 2016)
Sistemes d'Informació a les Organitzacions (Laboratories) (2017)
Number of credits 12 (6+3+3)

6.2 Future Work

Next, we sketch possible lines for future work in the same order in which we have presented our main contributions.

1. In Chapter 3, we proposed classical and new imputation methods to deal with incomplete data in CF datasets, which have specific characteristics such as high dimensionality and high percentage of null values. Moreover, we performed extensive experiments with three well-known datasets, showing that our proposed methods were able to obtain better quality recommendations and behavioural precision than well-known state-of-the-art methods. Future research in this topic will focus on two ways: (i) combine this imputation methods with privacy-preserving CF methods and (ii), use context-aware information to deal with CF's inherent problems such as *sparseness*, *cold user/item* and *scalability*.

2. Collaborative Filtering is a recommender system used to perform automatic recommendations to users in multiple contexts. Despite its great advantages, we highlighted its downside regarding users' privacy. In Chapter 4, we proposed three PPCF methods to protect the privacy of the users involved in CF processes. The experimental results show that V-MDAV obtains better results and provides both more privacy and data usability than well-known methods such as MDAV and Gaussian noise addition. Future work in this topic will focus on two lines. First, to overcome the dimensionality issues and outlier/malicious users to obtain less biased results. Second, to improve the efficiency of our methods by enabling their implementation in a decentralised setting.

3. In Section 5.1, we proposed the idea of using recommender systems integrated with the sensing infrastructure of smart cities to promote citizen's healthy habits in real time. In this field, future work will focus on open research topics related to s-health such as the privacy protection of citizens that use our proposal. To do so, we will concentrate on the privacy model proposed by Martínez et al. in [108], and the privacy-preserving Collaborative Filtering techniques proposed by Casino et al. in [35].

   In Section 5.2, we proposed and tested a set of hybrid methods combining RL and CF techniques in order to estimate context-aware radio-planning tasks with high accuracy in a time-efficient way. Further work in radiopropagation analysis will focus on three directions: First, to explore the use of three-dimensional structures instead of two-dimensional sub-matrices. Second, to completely transform LD simulations into HD simulations in a time-efficient way. Third, to study the performance of this technology in large complex environments, such as smart cities.

Bibliography

[1] The World Health Organization. Website.

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749, 2005.

[3] C. C. Aggarwal. On Randomization, Public Information and the Curse of Dimensionality. *2007 IEEE 23rd International Conference on Data Engineering*, pages 136–145, 2007.

[4] E. Aguirre, M. Flores, L. Azpilicueta, P. López-Iturri, F. Falcone, V. Ramos, and A. Solanas. Implementing context aware scenarios to enable smart health in complex urban environments. In *MeMeA*, pages 508–511, 2014.

[5] E. Aguirre, P.L. Iturri, L. Azpilicueta, S. De Miguel-Bilbao, V. Ramos, U. Gárate, and F. Falcone. Analysis of estimation of electromagnetic dosimetric values from non-ionizing radiofrequency fields in conventional road vehicle environments. *Electromagnetic Biology and Medicine*, 34(1):19–28, 2015. cited By 3.

[6] E. Aguirre, P. López, L. Azpilicueta, J. Arpón, and F. Falcone. Characterization and consideration of topological impact of wireless propagation in a commercial aircraft environment [wireless corner]. *IEEE Antennas and Propagation Magazine*, 55(6):240–258, Dec 2013.

[7] Y. Ar and E. Bostanci. A genetic algorithm solution to the collaborative filtering problem. *Expert Systems with Applications*, 61:122 – 128, 2016.

[8] L. Azpilicueta, F. Falcone, J.J. Astráin, J. Villadangos, A. Chertudi, I. Angulo, A. Perallos, P. Elejoste, and I.J. García Zuazola. Analysis of topological impact on wireless channel performance of intelligent street lighting system. *Radioengineering*, 23(1):412–420, 2014. cited By 3.

[9] L. Azpilicueta, M. Rawat, K. Rawat, F.M. Ghannouchi, and F. Falcone. Convergence analysis in deterministic 3d ray launching radio channel estimation in complex environments. *ACES*, 29(4):256–271, April 2014.

[10] L. Azpilicueta, M. Rawat, K. Rawat, F.M. Ghannouchi, and F. Falcone. A ray launching-neural network approach for radio wave propagation analysis in complex indoor environments. *Antennas and Propagation, IEEE Transactions on*, 62(5):2777–2786, May 2014.

[11] B. Bandelow. Assessing the efficacy of treatments for panic disorder and agoraphobia: Ii. the panic and agoraphobia scale. *International clinical psychopharmacology*, 1995.

[12] I. Barjasteh, R. Forsati, D. Ross, A. H. Esfahanian, and H. Radha. Cold-Start Recommendation with Provable Guarantees: A Decoupled Approach. *IEEE Transactions on Knowledge and Data Engineering*, PP(99):1, 2016.

[13] J. C. Basso, A. Shang, M. Elman, R. Karmouta, and W. A. Suzuki. Acute Exercise Improves Prefrontal Cortex but not Hippocampal Function in Healthy Adults. *Journal of the International Neuropsychological Society*, 21(Special Issue 10):791–801, 2015.

[14] A. Basu, J. Vaidya, H. Kikuchi, and T. Dimitrakos. Privacy-preserving Collaborative Filtering for the Cloud, 2011.

[15] J. Bennett and S. Lanning. The Netflix Prize. In *KDD Cup and Workshop*, pages 1–4, 2007.

[16] S. Berkvosky, F. Ricci, Y. Eytani, and T. Kuflik. Enhancing Privacy and Preserving Accuracy of a Distributed Collaborative Filtering. *Proceedings of the 2007 ACM conference on Recommender systems RecSys 07*, pages 9–16, 2007.

[17] A. Berlioz, A. Friedman, M. A. Kaafar, R. Boreli, and S. Berkovsky. Applying differential privacy to matrix factorization. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 107–114, New York, NY, USA, 2015. ACM.

[18] A. Bilge, C. Kaleli, I. Yakut, I. GunesS, and H. Polat. A survey of privacy-preserving collaborative filtering schemes. *International Journal of Software Engineering and Knowledge Engineering*, 23(08):1085–1108, 2013.

[19] A. Bilge and H. Polat. A comparison of clustering-based privacy-preserving collaborative filtering schemes. *Applied Soft Computing*, 13(5):2478–2489, May 2013.

[20] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109 – 132, 2013.

[21] D. Bogdanović and K. Lazarević. Early warning system and adaptation advice to reduce human health consequences of extreme weather conditions and air pollution. In *Handbook of Research on Democratic Strategies and Citizen-Centered E-Government Services*, chapter 15. IGI Global, jan 2015.

[22] R. D. Borcherdt. Reflection and refraction of type-ii s waves in elastic and anelastic media. *Bulletin of the Seismological Society of America*, 67(1):43–67, 1977.

[23] J. Breese, D. Heckerman, and C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *UAI 98*, pages 43–52, 1998.

[24] P. M. Butala, Y. Zhang, T. Little, and R. C. Wagenaar. Wireless system for monitoring and real-time classification of functional activity. In *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pages 1–5. IEEE, 2012.

[25] F. Cacheda, V. Carneiro, D. Fernández, and V. Formoso. Comparison of collaborative filtering algorithms. *ACM Transactions on the Web*, 5(1):1–33, February 2011.

[26] J. Canny. Collaborative filtering with privacy. *Security and Privacy, 2002. Proceedings. 2002 IEEE*, pages 45–57, 2002.

[27] J. Canny. Collaborative Filtering with Privacy via Factor Analysis. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 14:238—-245, 2002.

[28] A. Caragliu, C. del Bo, and P. Nijkamp. Smart cities in europe. In *CERS'09, 3rd Central European Conference in Regional Science*, pages 45 – 59, October 2009.

[29] J. Carson Smith, K. I. Erickson, and S. M. Rao. Introduction to the jins special issue: Physical activity and brain plasticity. *Journal of the International Neuropsychological Society*, 21:743–744, 11 2015.

[30] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, and A. Solanas. Hybrid-based optimization of wireless channel characterization for health services in medical complex environments. In

Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on, July 2015.

[31] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, and A. Solanas. Optimized wireless channel characterization in large complex environments by hybrid ray launching collaborative filtering approach. Technical report, Smart Health Research Group. Department of Computer Engineering and Maths, Universitat Rovira i Virgili, February 2015.

[32] F. Casino, L. Azpilicueta, P. Lopez-Iturri, E. Aguirre, F. Falcone, and A. Solanas. Optimised wireless channel characterisation in large complex environments by hybrid ray launching-collaborative filtering approach. *IEEE Antennas and Wireless Propagation Letters*, PP(99):1–4, 2016.

[33] F. Casino, E. Batista, C. Patsakis, and A. Solanas. Context-aware recommender for smart health. In *Smart Cities Conference (ISC2), 2015 IEEE First International*, pages 1–. IEEE, 2015.

[34] F. Casino, J. Domingo-Ferrer, C. Patsakis, D. Puig, and A. Solanas. Privacy preserving collaborative filtering with k-anonymity through microaggregation. In *ICEBE*, pages 490–497, 2013.

[35] F. Casino, J. Domingo-Ferrer, C. Patsakis, D. Puig, and A. Solanas. A k-anonymous approach to privacy preserving collaborative filtering. *Journal of Computer and System Sciences*, December 2014.

[36] F. Casino, P. Lopez-Iturri, E. Aguirre, L. Azpilicueta, F. Falcone, E. Batista, and A. Solanas. Two-dimensional collaborative filtering approach to wireless channel characterization in medical complex scenarios. In *Smart Cities Conference (ISC2), 2016 IEEE Second International*, pages 1–6. IEEE, 2016.

[37] F. Casino, P. Lopez-Iturri, L. Azpilicueta, E. Aguirre, F. Falcone, and A. Solanas. Optimal parameter estimation for wireless signal analysis in context-aware scenarios: A brief study. In *Information, Intelligence, Systems and Applications (IISA), 2016 7th International Conference on*, July 2016.

[38] F. Casino, C. Patsakis, E. Batista, F. Borras, and A. Martinez-Balleste. Healthy routes in the smart city: A context-aware mobile recommender. *IEEE Software*, In Press - 2017.

[39] F. Casino, C. Patsakis, A. Martínez-Ballesté, F. Borrás, and E. Batista. Technical report: Implementation and validation of a smart health application, 2017.

[40] F. Casino, C. Patsakis, D. Puig, and A. Solanas. On privacy preserving collaborative filtering: Current trends, open problems, and new issues. In *e-Business Engineering (ICEBE), 2013 IEEE 10th International Conference on*, pages 244–249, Sept 2013.

[41] F. Casino, C. Patsakis, and A. Solanas. Privacy-preserving collaborative filtering with variable-sized group microaggregation. *Transactions on Knowledge and Data Engineering*, Submitted.

[42] F. Casino and A. Solanas. Handling non-response in collaborative filtering-based recommender systems. *Expert Systems with Applications*, Submitted.

[43] Central Intelligence Agency. "Spain." The World Factbook, 2015.

[44] The Nielsen Company. Personal Recommendations and Consumer Opinions Posted Online are the most Trusted Forms of Advertising Globally, 2009.

[45] L. F. Cranor, J. Reagle, and M. S. Ackerman. Beyond concern: Understanding net users' attitudes about online privacy. Technical report, The Internet Upheaval: Raising Questions, Seeking Answers in Communications Policy, 2000.

[46] J. E. Crews and V. A. Campbell. Vision impairment and hearing loss among community-dwelling older americans: implications for health and functioning. *American Journal of Public Health*, 94(5):823, 2004.

[47] E. Q. da Silva, C. G. Camilo Junior, L. M. L. Pascoal, and T. C. Rosa. An evolutionary approach for combining results of recommender systems techniques based on Collaborative Filtering. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 959–966, 2014.

[48] P. Das and D. Mandal. Statistical Outlier Detection in Large Multivariate Datasets. *acsu.buffalo.edu*, pages 1–9, 2004.

[49] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[50] N. Dokoohaki, C. Kaleli, H. Polat, and M. Matskin. Achieving Optimal Privacy in Trust-Aware Social Recommender Systems. *Lecture Notes in Computer Science*, 6430(Social Informatics):62–79, 2010.

[51] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, and F. Sebé. Efficient multivariate data-oriented microaggregation. *The VLDB Journal*, 15(4):355–369, August 2006.

[52] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[53] J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing sdc methods for microdata on the basis of information loss and disclosure. In *Proceedings of ETK-NTTS 2001, Luxemburg: Eurostat*, pages 807–826. Eurostat, 2001.

[54] J. Domingo-Ferrer and V. Torra. *A quantitative comparison of disclosure control methods for microdata*, pages 111–133. Elsevier, 2001.

[55] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[56] C. Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. *Calibrating Noise to Sensitivity in Private Data Analysis*, pages 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[57] M. S. Dykewicz and D. L. Hamilos. Rhinitis and sinusitis. *Journal of Allergy and Clinical Immunology*, 125(2):S103–S115, 2010.

[58] T. ElGamal. A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Transactions on Information Theory*, 31(4):469–472, 1985.

[59] The European Environment Agency. Air quality in Europe, 2014, 2014.

[60] G. Eysenbach. What is e-health? *Journal of medical Internet research*, 3(2), 2001.

[61] A. Ferscha, P. Lukowicz, and F. Zambonelli. The Superorganism of Massive Collective Wearables. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing:*

Adjunct Publication, UbiComp '14 Adjunct, pages 1077–1084, New York, NY, USA, 2014. ACM.

[62] R. Forsati, M. Mahdavi, M. Shamsfard, and M. Sarwat. Matrix Factorization with Explicit Trust and Distrust Relationships. *CoRR*, abs/1408.0, 2014.

[63] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[64] Generalitat de Catalunya. La qualitat de l'aire a Catalunya–Anuari 2014, 2014.

[65] The Global Asthma Report. Global burden of disease due to asthma, 2014.

[66] J. Golbeck. Filmtrust: movie recommendations from semantic web-based social networks. In *Consumer Communications and Networking Conference*, volume 2, pages 1314–1315, 2006.

[67] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[68] I. Gunes, C. Kaleli, A. Bilge, and H. Polat. Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, 42(4):767–799, 2014.

[69] V.C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, and G.P. Hancke. Smart grid technologies: Communication technologies and standards. *Industrial Informatics, IEEE Transactions on*, 7(4):529–539, Nov 2011.

[70] D. Gupta, M. Digiovanni, H: Narita, and K. Goldberg. Jester 2.0: Evaluation of a new linear time collaborative filtering algorithm. *22nd International ACM SIGIR*, pages 291–292, 1999.

[71] A. Hannak, G. Soeller, D. Lazer, A. Mislove, and C. Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 305–318. ACM, 2014.

[72] O. Harel and X. Zhou. Multiple imputation : Review of theory , implementation and software. *Aging*, 26(January):3057–3077, 2007.

[73] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, January 2004.

[74] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2002.

[75] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems TOIS*, 22(1):89–115, 2004.

[76] H. D. Hristov. *Fresnal Zones in Wireless Links, Zone Plate Lenses and Antennas*. Artech House, Inc., 2000.

[77] C. L. A. Hsieh, J. Zhan, D. Zeng, and F. Wang. Preserving Privacy in Joining Recommender Systems, 2008.

[78] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[79] S. Hyun Kim, D. Wan Ryoo, and C. Bae. U-healthcare system using smart headband. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pages 1557–1560. IEEE, 2008.

[80] R. Istepanian, S. Laxminarayan, and C. S. Pattichis. *M-health: emerging mobile health systems*. Springer, 2006.

[81] G. Jagannathan and R. N. Wright. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. *conference on Knowledge discovery in data*, 2005.

[82] S. Jha, L. Kruger, and P. Mcdaniel. Privacy Preserving Clustering. *Proceedings of the 10th European Symposium on Research in Computer Security*, pages 397–417, 2005.

[83] C. E. Jimenez, A. Solanas, and F. Falcone. E-government interoperability: Linking open and smart government [guest editors' introduction]. *Computer*, 47(10):22–24, 2014.

[84] C. Kaleli and H. Polat. P2P collaborative filtering with privacy. *Turkish Journal of Electric Electrical Engineering and Compute Science*, 18(1):101–116, 2010.

[85] C. Kaleli and H. Polat. Privacy-Preserving Trust-Based Recommendations on Vertically Distributed Data, 2011.

[86] G. Kalton and D. Kasprzyk. Imputing for missing survey responses. *Proceedings of the Section on Survey Research ...*, pages 22–31, 1982.

[87] M. Kantarcioglu and J. Vaidya. Privacy preserving naive bayes classifier for horizontally partitioned data. *In IEEE ICDM workshop on privacy preserving data mining*, pages 3–9, 2003.

[88] H. Kikuchi, H. Kizawa, and M. Tada. Privacy-Preserving Collaborative Filtering Schemes, 2009.

[89] D. King, F. Greaves, C. Exeter, and A. Darzi. 'gamification': Influencing health behaviours with games. *Journal of the Royal Society of Medicine*, 106(3):76–78, 2013.

[90] Y. Koren and R. Bell. Advances in collaborative filtering. *Recommender Systems Handbook*, pages 43–52, 2011.

[91] S. Kumar et al. Mobile health technology evaluation: The mHealth evidence workshop. *American Journal of Preventive Medicine*, 45(2):228–236, 2013.

[92] S. Led, L. Azpilicueta, E. Aguirre, M. M. de Espronceda, L. Serrano, and F. Falcone. Analysis and description of holtin service provision for aecg monitoring in complex indoor environments. *Sensors*, 13(4):4947–4960, 2013.

[93] D. Lemire and A. Maclachlan. Slope one predictors for online rating-based collaborative filtering. *Society for Industrial Mathematics*, 05(12):471–475, 2005.

[94] A. Leone, G. Diraco, and P. Siciliano. Context-aware aal services through a 3d sensor-based platform. *Journal of Sensors*, 2013, 2013. cited By 2.

[95] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, April 2007.

[96] X. Li, X. Zhang, X. Ren, M. Fritsche, J. Wickert, and H. Schuh. Precise positioning with current multi-constellation global navigation satellite systems: Gps, glonass, galileo and beidou. *Sci. Rep.*, 5, 02 2015.

[97] Liander and AIM. Amsterdam smart city. Website, 2012.

[98] D. Lin, X. Wu, F. Labeau, and A. Vasilakos. Internet of vehicles for e-health applications in view of emi on medical sensors. *Journal of Sensors*, 2015, 2015.

[99] D. Liu, E. Bertino, and X. Yi. Privacy of outsourced k-means clustering. In *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security*, ASIA CCS '14, pages 123–134, New York, NY, USA, 2014. ACM.

[100] P. Lopez-Iturri, F. Casino, E. Aguirre, L. Azpilicueta, F. Falcone, and A. Solanas. Performance analysis of zigbee wireless networks for aal through hybrid ray launching and collaborative filtering. *Journal of Sensors*, 2016(2424101):1–16, 2016.

[101] R. J. Luebbers. Comparison of lossy wedge diffraction coefficients with application to mixed path propagation loss prediction. *IEEE Transactions on Antennas and Propagation*, 36(7):1031–1034, 1988. cited By 36.

[102] R. J. Luebbers. A heuristic utd slope diffraction coefficient for rough lossy wedges. *IEEE Transactions on Antennas and Propagation*, 37(2):206–211, 1989. cited By 85.

[103] P. Lukowicz, S. Pentland, and A. Ferscha. From Context Awareness to Socially Aware Computing. *IEEE Pervasive Computing*, 11(1):32–41, 2012.

[104] H. Ma, I. King, and M. R. Lyu. Effective Missing Data Prediction for Collaborative Filtering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 39–46, New York, NY, USA, 2007. ACM.

[105] A. Machanavajjhala, D. Kifer, J. Gehrke, and M Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[106] K. A. Mackinnon. User Generated Content vs. Advertising: Do Consumers Trust the Word of Others Over Advertisers? *The Elon Journal of Undergraduate Research in Communications*, 3(1):14–22, 2012.

[107] I. Martínez, J. Escayola, M. Martínez-Espronceda, L. Serrano, J. Trigo, S. Led, and J. García. Standard-based middleware platform for medical sensor networks and u-health. In *Computer Communications and Networks, 2008. ICCCN'08. Proceedings of 17th International Conference on*, pages 1–6. IEEE, 2008.

[108] A. Martínez-Ballesté, P. A. Pérez-Martínez, and A. Solanas. The pursuit of citizens' privacy: A privacy-aware smart city is possible. *IEEE Communications Magazine*, 51(6), 2013.

[109] B. Martínez-Pérez, I. De La Torre-Díez, M. López-Coronado, and J. Herreros-González. Mobile apps in cardiology: Review, 2013.

[110] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, 492-508.*, 3290(8):492–508, 2004.

[111] P. Massa and P. Avesani. Trust metrics on controversial users: balancing between tyranny of the majority. *International Journal on Semantic Web and Information Systems*, pages 1–21, 2007.

[112] G. J. Matthews and O. Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statist. Surv.*, 5:1–29, 2011.

[113] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[114] T. Minkus and K. W. Ross. I know what you?re buying: Privacy breaches on ebay. In *Privacy Enhancing Technologies*, pages 164–183. Springer, 2014.

[115] N. Mirbakhsh and C. X. Ling. Improving Top-N Recommendation for Cold-Start Users via Cross-Domain Information. *ACM Trans. Knowl. Discov. Data*, 9(4):33:1—-33:19, 2015.

[116] A. Moreno, I. Angulo, A. Perallos, H. Landaluce, I.J.G. Zuazola, L. Azpilicueta, J.J. Astrain, F. Falcone, and J. Villadangos. Ivan: Intelligent van for the distribution of pharmaceutical drugs. *Sensors (Switzerland)*, 12(5):6587–6609, 2012. cited By 13.

[117] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy sp 2008*, 0:111–125, 2008.

[118] M. P. O'Mahony, N. J. Hurley, and G. Silvestre. Detecting noise in recommender system databases. *Proceedings of the 11th international conference on Intelligent user interfaces - IUI '06*, page 109, 2006.

[119] P. Paillier. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. *Advances in Cryptology EUROCRYPT 99*, 1592:223–238, 1999.

[120] F. Palumbo, P. Barsocchi, F. Furfari, and E. Ferro. Aal middleware infrastructure for green bed activity monitoring. *Journal of Sensors*, 2013, 2013. cited By 9.

[121] R. Parameswaran and D. M. Blough. Privacy Preserving Collaborative Filtering Using Data Obfuscation. *2007 IEEE International Conference on Granular Computing (GRC 2007)*, pages 380–380, November 2007.

[122] C. Patsakis, P. Laird, M. Clear, M. Bouroche, and A. Solanas. Interoperable privacy-aware e-participation within smart cities. *Computer*, 48(1):52–58, 2015.

[123] R. A. Pauwels, A. S. Buist, P. M. Calverley, C. R. Jenkins, and S. S. Hurd. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 163(5), 2012.

[124] D. Y. Pavlov and D. M. Pennock. A Maximum Entropy Approach to Collaborative Filtering in Dynamic, Sparse, High-Dimensional Domains. *Advances in Neural Information Processing Systems 15*, 15(2/3):1441–1448, 2003.

[125] P. Pereira, E. Duarte, F. Rebelo, and P. Noriega. A review of gamification for health-related contexts. In Aaron Marcus, editor, *Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments*, volume 8518 of *Lecture Notes in Computer Science*, pages 742–753. Springer International Publishing, 2014.

[126] P. A. Pérez-Martínez, A. Martínez-Ballesté, and A. Solanas. Privacy in smart cities - a case study of smart public parking. In *PECCS*, pages 55–59, 2013.

[127] H. Polat. Privacy-preserving collaborative filtering using randomized perturbation techniques. *Third IEEE International Conference on Data Mining*, pages 625–628, 2003.

[128] H. Polat and W. Du. Privacy-preserving collaborative filtering on vertically partitioned data. *Knowledge Discovery in Databases: PKDD 2005*, pages 651–658, 2005.

[129] H. Polat and W. Du. Privacy-preserving top-N recommendation on distributed data. *Journal of the American Society for Information Science*, 59(7):1093–1108, 2008.

[130] H. Polat and L. Hall. SVD-based Collaborative Filtering with Privacy. *Proceedings of the 2005 ACM symposium on Applied computing SAC 05*, pages 791–795, 2005.

[131] M. James D. Powell. Restart procedures for the conjugate gradient method. *Mathematical programming*, 12(1):241–254, 1977.

[132] P. K. Prasad and C. P. Rangan. Privacy preserving birch algorithm for clustering over arbitrarily partitioned databases. In *International Conference on Advanced Data Mining and Applications*, pages 146–157. Springer, 2007.

[133] Y. Qi and M. J. Atallah. Efficient Privacy-Preserving k-Nearest Neighbor Search. *2008 The 28th International Conference on Distributed Computing Systems*, pages 311–319, June 2008.

[134] C. Quantin, F. Allaert, and L. Dusserre. Anonymous statistical methods versus cryptographic methods in epidemiology. *International Journal of Medical Informatics*, 60(2):177 – 183, 2000.

[135] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. *ACM SIGMOD Record*, pages 1–20, 2000.

[136] M. Ranjbar, P. Moradi, M. Azami, and M. Jalili. An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems. *Engineering Applications of Artificial Intelligence*, 46, Part A:58–66, 2015.

[137] P. Rashidi and A. Mihailidis. A survey on ambient-assisted living tools for older adults. *IEEE Journal of Biomedical and Health Informatics*, 17(3):579–590, 2013.

[138] P. Ray. Special issue on emerging technologies in communications - Area 1 m-health. *IEEE Journal on Selected Areas in Communications*, 2013.

[139] K. Reiter, K. A. Nielson, T. J. Smith, L. R. Weiss, A. J. Alfini, and J. C. Smith. Improved Cardiorespiratory Fitness Is Associated with Increased Cortical Thickness in Mild Cognitive Impairment. *Journal of the International Neuropsychological Society*, 21(Special Issue 10):757–767, 2015.

[140] Y. Ren, G. Li, J. Zhang, and W. Zhou. The Efficient Imputation Method for Neighborhood-based Collaborative Filtering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 684–693, New York, NY, USA, 2012. ACM.

[141] Y. Ren, G. Li, J. Zhang, and W. Zhou. The maximum imputation framework for neighborhood-based collaborative filtering. *Social Network Analysis and Mining*, 4(1):1–15, 2014.

[142] P. Resnick, N. Iacovou, and M. Suchak. GroupLens: an open architecture for collaborative filtering of netnews. *Proceedings of the ACM conference on Computer supported cooperative work CSCW*, pp(3):175–186, 1994.

[143] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[144] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 2013.

[145] J. A. Roderick. Survey Nonresponse Adjustments. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 1–10, 1984.

[146] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, June 1987.

[147] I. Salaberria, A. Perallos, L. Azpilicueta, F. Falcone, R. Carballedo, I. Angulo, P. Elejoste, A. Bahillo, J.J. Astrain, and J. Villadangos. Ubiquitous connected train based on train-to-ground and intra-wagon communications capable of providing on trip customized digital services for passengers. *Sensors (Switzerland)*, 14(5):8003–8025, 2014. cited By 4.

[148] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[149] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[150] G. Schröder, M. Thiele, and W. Lehner. Setting goals and choosing metrics for recommender system evaluations. In *CEUR Workshop Proceedings*, volume 811, pages 78–85, 2011.

[151] A. Schwaighofer, V. Tresp, and H. Kriegel. Probabilistic memory-based collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):56–69, January 2004.

[152] M. Shaneck, Y. Kim, and V. Kumar. Privacy Preserving Nearest Neighbor Search, 2006.

[153] Y. Shi, M. Larson, and A. Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 47(1):1–45, 2014.

[154] R. Shokri. Privacy games: Optimal user-centric data obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2015(2):1–17, 2015.

[155] R. Shokri, P. Pedarsani, G. Theodorakopoulos, and J. Hubaux. Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. *Proceedings of the third ACM conference on Recommender systems RecSys 09*, page 157, 2009.

[156] A. Solanas and A. Martínez-Ballesté. V-MDAV : A Multivariate Microaggregation With Variable Group Size. *Seventh COMPSTAT Symposium of the IASC*, 2006.

[157] A. Solanas, C. Patsakis, M. Conti, I. Vlachos, V. Ramos, F. Falcone, O. Postolache, P.A. Pérez-Martínez, R. Di Pietro, D. Perrea, and A. Martínez-Ballesté. Smart health: A context-aware health paradigm within smart cities. *IEEE Communications Magazine*, 52(8):74–81, August 2014.

[158] L. H. Son. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Information Systems*, pages –, 2014.

[159] T. Starner. How Wearables Worked their Way into the Mainstream. *Pervasive Computing, IEEE*, 13(4):10–15, October 2014.

[160] X. Su and T. M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(Section 3):1–19, 2009.

[161] X. Sun, H. Wang, J. Li, and T. M. Truta. Enhanced p-sensitive k-anonymity models for privacy preserving data publishing. *Trans. Data Privacy*, 1(2):53–66, August 2008.

[162] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and ...*, 10(5):557–570, 2002.

[163] M. Tada, H. Kikuchi, and S. Puntheeranurak. Privacy-Preserving Collaborative Filtering Protocol Based on Similarity between Items, 2010.

[164] Q. Tang and J. Wang. Privacy-preserving context-aware recommender systems: Analysis and new solutions. *IACR Cryptology ePrint Archive*, 2015:364, 2015.

[165] J. Thom and W. B. Kannel. Congestive heart failure. *Disease Management & Health Outcomes*, 1(2):75–83, 1997.

[166] M. Tomlinson, M. J. Rotheram-Borus, L. Swartz, and A. C. Tsai. Scaling Up mHealth: Where Is the Evidence? *PLoS Medicine*, 10(2), 2013.

[167] J. Vaidya and C. Clifton. Privacy-preserving outlier detection, 2004.

[168] S. van den Elshout, K. Léger, and F. Nussio. Comparing urban air quality in Europe in real time. In *Environment International*, volume 34, pages 720–726. 2008.

[169] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 1995.

[170] S. K. Vashist, E. M. Schneider, and J. H. T. Luong. Commercial Smartphone-Based Devices and Smart Applications for Personalized Healthcare Monitoring and Management. *Diagnostics*, 4(3):104–128, 2014.

[171] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *Proceedings of the 23rd USENIX Security Symposium (USENIX Security)*, 2014.

[172] T. Vos et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380(9859):2163–96, December 2012.

[173] A. Wahid, C. Leckie, and C. Zhou. Estimating the number of hosts corresponding to an intrusion alert while preserving privacy. *Journal of Computer and System Sciences*, 80(3):502–519, 2014.

[174] M. E. Wall, A. Rechtsteiner, and L. M. Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[175] J. Wang and Q. Tang. Recommender systems and their security concerns. Cryptology ePrint Archive, Report 2015/1108, 2015. `http://eprint.iacr.org/`.

[176] F. Weinmann. Ray tracing with po/ptd for rcs modeling of large complex objects. *IEEE Transactions on Antennas and Propagation*, 54(6):1797–1806, June 2006.

[177] The World Health Organization. Solar ultraviolet radiation: Global burden of disease from solar ultraviolet radiation, 2006.

[178] The World Health Organization. Exposure to air pollution: A major public health concern, 2010.

[179] The World Health Organization. Prevention and control of noncommunicable diseases: guidelines for primary health care in low resource settings, 2012.

[180] The World Health Organization. The top 10 causes of death, 2012.

[181] W. Wu, J. Zhou, Y. Xiang, and L. Xu. How to achieve non-repudiation of origin with privacy protection in cloud computing. *Journal of Computer and System Sciences*, 79(8):1200–1213, 2013.

[182] F. Xhafa. *Advanced Technological Solutions for E-Health and Dementia Patient Monitoring*. Advances in Medical Technologies and Clinical Practice:. IGI Global, 2015.

[183] B. Xu, L. D. Xu, H. Cai, C. Xie, J. Hu, and F. Bu. Ubiquitous data accessing method in iot-based information system for emergency medical services. *Industrial Informatics, IEEE Transactions on*, 10(2):1578–1586, May 2014.

[184] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren. Information security in big data: Privacy and data mining. *IEEE Access*, 2:1149–1176, 2014.

[185] G. Xue, C. Lin, Q. Yang, W. Xi, H. Zeng, Y. Yu, and Z. Chen. Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 114–121, New York, NY, USA, 2005. ACM.

[186] I. Yakut and H. Polat. Privacy-Preserving Svd-Based Collaborative Filtering on Partitioned Data. *International Journal of Information Technology & Decision Making*, 09(03):473–502, May 2010.

[187] I. Yakut and H. Polat. Privacy-preserving hybrid collaborative filtering on cross distributed data. *Knowledge and Information Systems*, 30(2):405–433, April 2011.

[188] I. Yakut and H. Polat. Arbitrarily distributed data-based recommendations with privacy. *Data & Knowledge Engineering*, 72:239–256, February 2012.

[189] I. Yakut and H. Polat. Estimating NBC-based recommendations on arbitrarily partitioned data with privacy. *Knowledge-Based Systems*, 36:353–362, December 2012.

[190] A. C. Yao. Protocols for secure computations. *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pages 160–164, November 1982.

[191] F. Zambonelli. Toward Sociotechnical Urban Superorganisms. *Computer*, 45(8):76–78, August 2012.

[192] J. Zhan, I. C. Wang, and C. L. Hsieh. Towards efficient privacy-preserving collaborative recommender systems. *GrC 2008.*, pages 778–783, 2008.

[193] S. Zhang. Shell-neighbor method and its application in missing data imputation. *Applied Intelligence*, 35(1):123–133, 2010.

[194] S. Zhang, J. Ford, and F. Makedon. A privacy-preserving collaborative filtering scheme with two-way communication. *Proceedings of the 7th ACM conference on Electronic commerce - EC '06*, pages 316–323, 2006.

[195] S. Zhang, Z. Jin, and X. Zhu. Missing data imputation by utilizing information within incomplete instances. *Journal of Systems and Software*, 84(3):452–459, 2011.

[196] X. Zhang, C. Liu, S. Nepal, and J. Chen. An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud. *Journal of Computer and System Sciences*, 79(5):542–555, 2013.

[197] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu. Missing Value Estimation for Mixed-Attribute Data Sets. *IEEE Transactions on Knowledge and Data Engineering*, 23(1):110–121, 2011.

[198] C. N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.

UNIVERSITAT ROVIRA I VIRGILI
PRIVACY-PRESERVING CROWDSOURCING-BASED RECOMMENDER SYSTEMS FOR E-COMMERCE & HEALTH SERVICES
FRANCISCO JOSE CASINO CEMBELLIN

UNIVERSITAT
ROVIRA i VIRGILI