Privacy in Recommender Systems

Arjan J.P. Jeckmans, Michael Beye, Zekeriya Erkin, Pieter Hartel, Reginald L. Lagendijk, and Qiang Tang

Abstract In many online applications, the range of content that is offered to users is so wide that a need for automated recommender systems arises. Such systems can provide a personalized selection of relevant items to users. In practice, this can help people find entertaining movies, boost sales through targeted advertisements, or help social network users meet new friends.

To generate accurate personalized recommendations, recommender systems rely on detailed personal data on the preferences of users. Examples are ratings, consumption histories, and personal profiles. Recommender systems are useful; however, the privacy risks associated to gathering and processing personal data are often underestimated or ignored. Many users are not sufficiently aware if and how much of their data is collected, if such data is sold to third parties, or how securely it is stored and for how long.

This chapter aims to provide insight into privacy in recommender systems. First, we discuss different types of existing recommender systems. Second, we give an overview of the data that is used in recommender systems. Third, we examine the associated risks to data privacy. Fourth, relevant research areas for privacy-protection techniques and their applicability to recommender systems are discussed. Finally, we conclude with a discussion on applying and combining different privacy-protection techniques in real-world settings, making clear mappings to reflect typical relations between recommender system types, information types, particular privacy risks, and privacy-protection techniques.

A.J.P Jeckmans (✉) • P. Hartel • Q. Tang
Distributed and Embedded Security, Faculty of EEMCS, University of Twente, Enschede, The Netherlands
e-mail: a.j.p.jeckmans@utwente.nl; q.tang@utwente.nl; pieter.hartel@utwente.nl

M. Beye • Z. Erkin • R.L. Lagendijk
Information Security and Privacy Lab, Faculty of EEMCS, Delft University of Technology, Delft, The Netherlands
e-mail: m.r.t.beye@tudelft.nl; z.erkin@tudelft.nl; r.l.lagendijk@tudelft.nl

N. Ramzan et al. (eds.), *Social Media Retrieval*, Computer Communications and Networks, DOI 10.1007/978-1-4471-4555-4_12, © Springer-Verlag London 2013

1 Introduction

In recent years, online applications have become an important part of daily life for millions of users. People consume media (Youtube, Flickr, LastFM), do their shopping (Amazon, Ebay), and interact (Facebook, Gmail) online. Because the range and amount of content that is offered to users is often huge, automated recommender systems are employed. By providing personalized suggestions, these systems can help people find interesting media, boost sales through targeted advertisements, or help people meet new friends. Because of their automated nature, recommender systems can meet the demands of large online applications that operate on a global scale.

All recommender systems share a common trait: in order to generate personalized recommendations, they require information on the attributes, demands, or preferences of the user. Typically, the more detailed the information related to the user is, the more accurate the recommendations for the user are. Service providers running the recommender systems collect information where possible to ensure accurate recommendations. The information supplied can either be automatically collected or specifically provided by the user. Automatically collected information is the result of users interacting with the recommender systems and making choices based on recommendations. For example, page views on Ebay are used to automatically present a selection of recommended similar items (*recommendations for you*). Similarly, recommended videos on Youtube are influenced by recently viewed videos. Based on purchases by other users, items on Amazon are accompanied by package deals (*frequently bought together*) or related items (*customers who bought this item also bought*). Based on sites visited, Google serves personalized advertisements. Based on your friends and social interactions, Facebook suggests new friends to make. LinkedIn, based on a user's cv and connections, recommends interesting companies, job offers, and news. Vice versa, LinkedIn also recommends people to recruiters posting new job openings. Many dating sites, such as Match.com or PAIQ, recommend partner matches to its users. Many more examples of such systems exist and they will continue to exist, in the future. Users can also specifically provide information. In this way, users build their own profile specifying their likes and dislikes or containing general information (such as age and gender) about themselves. For example, LastFM and Youtube allow users to specify their favorites. Facebook allows listing profile information as well as interests.

However, potential threats to user privacy are often underestimated. The more detailed the information related to the user is, the larger the threat to the user's privacy is. In order to enhance their recommender systems, service providers are collecting and consolidating more and more information. For example, in recent privacy policy updates, Google stated that they consolidate information from all their services to a single profile. Facebook continues to expand its reach around the Internet, giving the ability to share more and *like* almost everything. Information might be abused by the service provider, sold to a third party, or leaked by a hacker. There is an inherent trade-off between utility (getting accurate personalized recommendations) and privacy. Research into regulations, anonymization, and privacy-preserving algorithms aims to improve privacy, while maintaining utility. In this chapter, we will analyze the privacy risks associated with recommender systems and the research that helps to minimize these risks.

We first look at the state of the art of various types of recommender systems in use today (Sect. 2). Second, we give our categorization of the types of information generally involved in recommender systems, operation and how this is mapped to the various types of recommender systems. Third, we identify the privacy concerns in recommender systems and give our own classification of them. To see how the privacy concerns affect the recommender systems, the privacy concerns are mapped to the different types of information (Sect. 3). Fourth, we give an overview of existing research into state-of-the-art privacy-protection techniques (Sect. 4). The relationship between the research and privacy concerns is also given. Finally, we conclude and discuss (Sect. 5).

2 Recommender Systems

In this section, we give an overview of the different recommender system types. We then list the information present in recommender systems. Finally, we show what information is typically used in which recommender type. This relationship will serve as a basis, to describe the privacy concerns in the next section.

A recommender system provides a set of *items* (e.g., content, solutions, or other users) that is most relevant to a particular user of the system. Typically, recommender systems achieve this by predicting *relevance scores* for all items that the user has not seen yet. Items that receive the highest score get recommended (typically the top-N items or all items above a threshold t). The prediction is made by considering both the traits of the item and user. Typically, systems look at similarities between items, similarities between users, or relations between particular types of items and particular types of users. The performance of a recommender system is determined by looking at the recommendation accuracy, that is, the error between given and expected results.

2.1 Recommender System Types

Adomavicius and Tuzhilin [1] gave an overview of the state of the art in recommender systems and possible extensions. They list only the three popular types of that time: collaborative filtering, content-based, and hybrid. We make a distinction between basic recommender types and improved recommender types. Improved recommender types build upon basic recommender types by combining them or adding new information. Any improved recommender type can be combined with any basic recommender type. The following basic recommender system types have been around for quite some time:

Collaborative Filtering. One of the first collaborative filtering recommender systems is Tapestry, by Goldberg et al. [16]. This system was designed to retrieve email messages from Usenet mailing lists, relevant to a user's particular interests. Goldberg et al. observed that conventional mailing lists are too static and rarely form a perfect match to a user's demands. Tapestry relies on what the authors termed *collaborative filtering techniques*, which are still widely used today. In collaborative filtering, each user rates content items. These ratings determine similarity between either users (similar users like similar items) or items (users like items similar to highly rated items). Different metrics exist to compute similarity. Recommended for the current user are those items that are rated highest by his most similar peers or contain those items that are rated most similar to his favorite items.

Content-based. Content-based recommender systems use item similarity to determine recommendations. Unlike the collaborative filtering method, item similarity is computed by item metadata. Examples of metadata are kitchen for restaurants, genre for movies, and artist for music. Recommended are those items that are most similar to the user's favorite items. An example of a content-based recommender system is Newsweeder, by Lang [24].

Demographic. When detailed information about the user's preferences is not available, demographic information can lead to somewhat personalized recommendations. Grundy, by Rich [32], is an example of this. Demographic information may include age, gender, country of residence, education level, etc. The demographic information is matched to a stereotype, and the items attached to this stereotype are recommended. Personalization for the user is limited due to the generalization to a stereotype.

Knowledge-based. When requiring a recommendation, the user enters his preferences in the recommender system. The system then outputs a (number of) potential recommendations based on (expert) knowledge contained in the system. Possibly, the user can give feedback, and the recommendation is refined. After a few iterations, the recommendation is tailored to the user. Entree [7] is an example of such a system, built to help diners find a suitable restaurant. In learning knowledge-based recommender systems, feedback from the user is fed back into the system to add to the knowledge [25].

The following improvements have been proposed to the basic recommender systems mentioned above:

Context-aware. In many application domains, contextual information is available, which can be used to improve recommendations. Common examples of contextual information are location, group dynamic, time, date, and purpose. While user preferences and domain knowledge are relatively static, context is highly dynamic in nature. Every recommendation, even for the same user, may have a completely new context. Adomavicius and Tuzhilin [2] provided a discussion on contextual information in recommender systems. They showed three ways in which such contextual information can be added to existing recommender systems: (1) Use a prefilter to remove content items (and information associated with them) that do not fit the context from the system. (2) Use a postfilter to remove recommendations that do not fit the context. (3) Add the context to the model of the recommender system and use the contextual information during the recommendation process.

Ensemble. Ensembles of recommender systems combine several of the same type of recommender system to improve performance. The idea behind ensembles is to get multiple opinions before making a decision. Schclar et al. [34] detailed the use of ensembles on collaborative filtering.

Hybrid. Hybrid recommender systems, like ensembles, combine multiple recommender systems. However, in a hybrid system, multiple different types of recommender systems are combined. A comprehensive overview of different hybridization techniques is given by Burke [8]. As concluded by Burke, given a certain hybridization, not all basic recommender systems can be (straightforwardly) combined.

Social. The rise of online social networks increased the availability of a user's social information (e.g., the friendship network). Because friends typically share interests, the information they supply to the recommender system is more likely to fit with the user. Or alternatively, the social information can be used to infer communities of similar users. As an example, Konstas et al. [22] utilized the social information in LastFM to improve collaborative filtering.

2.2 Information in Recommender Systems

We now discuss the different types of information typically used in recommender systems. We do not aim to give a complete categorization of information used but instead to explore the diversity of information used in recommender systems.

Behavioral information is the implicit information that the recommender system can gather while the user interacts with the broader system. For example, product views in a webshop or not fully watching a movie on a video on demand site.

Contextual information describes to the context in which a recommendation query is made. Common examples of contextual information are location, social group, time, date, and purpose.

Domain knowledge specifies the relationship between a user stereotype and content items. Domain knowledge is usually static but can change over time.

Item metadata is descriptive information about content items. Examples of metadata are kitchen for restaurants, genre for movies, and artist for music.

Purchase or consumption history is the list of content that has previously been purchased or consumed by the user.

Table 1 Information that is present in different recommender system types

↓ Rec. sys./info. →	Behavioral	Contextual	Domain knowledge	Item meta-data	History	Recommendation	Feedback	Social	User attributes	User preferences
Collaborative filtering	•	·	·	·	•	●	·	·	·	●
Content-based	·	·	·	●	•	●	·	·	·	●
Demographic	·	·	●	·	·	●	·	·	●	·
Knowledge-based	•	·	●	•	·	●	•	·	•	●
Context-aware	•	●	•	•	·	·	·	·	•	·
Ensemble	·	·	·	·	·	●	·	·	·	·
Hybrid	·	·	·	·	·	●	·	·	·	·
Social	·	·	·	·	·	·	·	●	·	·

Recommendations are the output of a recommender system, typically a ranked list of items. In some systems, the relevance score for each content item is also given to the user.

Recommendation feedback is information about the recommendation provided by the user. Feedback can be expressed as positive, negative, or something more nuanced (stating a reason as well).

Social information describes the relationship between different users. Many sites allow users to specify a friendship relation (or similar) to other users, community membership, or both.

User attributes describe the user. Examples of user attributes are demographic information, income, and marital status.

User preferences are explicitly stated opinions about items or groups of items. Preferences are expressed by either a scalar measure (rating items on a scale of 1–5 stars), a binary indicator (keeping a list of favorites), or text (tags and comments).

2.3 Summary

Table 1 shows the type of information used by the type of recommender system. Information is ● almost always present, • sometimes present, or · almost never present in a recommender system. For improved recommender types, we mark what information is added to the basic recommender types. There is a clear distinction between the information used in the basic recommender types. The improved recommender types either add new types of information or combine multiple basic recommender systems.

3 Privacy Concerns in Recommender Systems

Because users need to reveal information in order to make use of the desired functionality of a recommender system, a trade-off exists between utility and user privacy. Obtaining accurate recommendations is one thing, but sharing personal information may also lead to privacy breaches. In this section, we will look into privacy in recommender systems and potential privacy concerns with a focus on user privacy.

3.1 Privacy and Confidentiality

The word privacy has many subtly different meanings, each with their own definition. Privacy on the Internet revolves mainly around *information privacy*. Kang [20] used the wording of the Information Infrastructure Task Force (IITF), as cited below:

> Information privacy is “an individual’s claim to control the terms under which personal information – information identifiable to the individual – is acquired, disclosed or used.”

This concept of information privacy is strongly related to the notion of *confidentiality*, from the field of information security, but not to be used interchangeably. Confidentiality is concerned with the secrecy of individual pieces of information. Information privacy focusses on the individual who is the subject of said information, the effects that disclosure have on this person, and his or her control and consent. In our overview of privacy-protection technologies, the focus will lie on preventing unwanted disclosure and usage of information, but not on the effects on the person.

When using online applications, users generally share a lot of (personal) information. Whether it is uploading ratings or comments, posting personal information on a profile, or making purchases, information is always shared within a particular *scope* [28]. Privacy involves keeping a piece of information in its intended scope. This scope is defined by the size of the audience (breadth), by extent of usage allowed (depth), and duration (lifetime). When a piece of information is moved beyond its intended scope in any of these dimensions (be it accidentally or maliciously), a privacy breach occurs. So, a breach may occur when information is disclosed to a party for whom it was not intended, when information is abused for a different purpose than was intended, or when information is stored beyond its intended lifetime.

Weiss [39] stated that on the traditional Web, privacy is maintained by limiting data collection, hiding users’ identities, and restricting access to authorized parties only. Often, in practice, information and identity become closely linked and visible to large groups of people. Profiles may be publicly visible, comments can be seen by all viewers of a content item, and some sites list the last users to visit a particular page. It becomes harder for a user to monitor and control his personal information, as more of it becomes available online. This problem mainly applies to systems where the user logs in to an account and where tools are available to express a user's preferences.

Often, users are not very aware of their (lack of) privacy. In a study on social network users in particular, Gross and Acquisti [18] showed that most users do not change the default privacy settings, while sharing a large amount of information on their profile. Tufekci [38] concluded in his case study that privacy-aware users are actually more reluctant to join social networks, but once they do join, they still disclose a lot of information. As opposed to social networks, in most recommender systems, privacy toward other users is probably not the largest issue. Users place a lot of implicit trust in service providers, expecting them to handle user information in a fair and conscientious way and continue to do so in the future. By using the system, users enter into a relationship with the service provider, who can generally view *all* information in the system, including private uploads, browsing and purchase behavior, and IP addresses. It is also the service provider who decides which information is stored, how long it is kept, and how it is used or distributed. Usually, privacy statements are offered to display the position the service provider takes and to acquire the user's consent. However, this leaves users little choice: they can either agree to the terms or will not benefit from using the system. The power balance is clearly in favor of the service provider.

3.2 Privacy Concerns

Privacy breaches can involve a variety of parties (fellow users, the service provider, or outsiders) and may be a deliberate act (snooping, hacking), or accidental (mismanagement, lingering data). Depending on the sensitivity of information involved, such incidents may have serious consequences. Lam et al. [23] already identified some threats to privacy in recommender systems. Their concern is the amount of (personal) information that is collected by the service provider and the potential leakage of this information. Independent of their work, we explicitly identify the privacy concerns in recommender systems and classify them as follows:

Data Collection. Many users are not aware of the amount and extent of information that a service provider is able to collect and what can be derived from this information. This may be due to the fact that privacy statements are seldomly read, and people have become used to pursuing online activities. Usually there is no way to opt-out of such data gathering, other than not using the system at all. As collection practices do not match with the users' expectations, this concern relates to the extent of information usage.

Data Retention. Online information is often difficult to remove, the service provider may even intentionally prevent or hinder removal of data. This is because there is commercial value in user information, for both competitive advantage through analysis and/or data sales. Furthermore, information that is apparently erased from one place may still reside somewhere else in the system, for example, in backups, to be found by others. The data retention concern relates to the intended lifetime, as information can be available longer than intended.

Data Sales. The wealth of information that is stored in online systems is likely to be of value to third parties and may be sold in some cases. Users' ratings, preferences, and purchase histories are all potentially interesting for marketing purposes. Data sales usually conflicts with the privacy expectations of users. Even though data is often anonymized before being sold to protect user privacy, re-identification is a threat that is often overlooked or ignored. For example, the information published by Netflix as part of their recommender systems prize, though anonymized, allowed for re-identification [27]. Narayanan and Shmatikov linked the anonymized records to publicly available records (such as IMDb) based on rating similarity and time of rating. If two records give a similar rating to a movie around the same time, they are likely to be from the same person. A higher number of similar movie ratings (in rating and in time) increases the confidence of the link between the records. This concern relates mainly to the extent of information usage.

Employee Browsing Private Information. The service provider as an entity has full access to the information, and its employees might take advantage of this. This is in conflict with the intended breadth of the audience, and the privacy that the service provider has promised its users.

Recommendations Revealing Information. Recommendations inherently are based on the information contained in the recommender system. For example, in collaborative filtering that information is the ratings of all the users, or in knowledge-based recommender systems it is the expert knowledge. Each recommendation reveals a tiny piece of information about the private information. It is unclear how a large number of recommendations impact the disclosure of information. This could be used to reveal information about other users (compromising their privacy) or information about the recommender system itself (potentially leading to reverse engineering of the system). Here, we focus on the privacy of the user, not the security of the system. Ramakrishnan et al. [31] looked at the privacy of eccentric users (users with unusual ratings) from a graph perspective. When looking at recommendation results, these users are at a higher risk than average users. As eccentric users cannot hide in crowds of other users, when their data is used for making recommendation, other data is often not. The recommendations output by the system are then based on only a few users, with a strong correlation between the input of the eccentric users and the recommendation output. This is in conflict with the intended breadth of the audience.

Shared Device or Service. Privacy at home can be just as important as privacy online. When sharing a device like a set-top box or computer, or a login to an online service, controlling privacy toward family and friends may be difficult. For example, a wife who wants to hide from her husband the fact that she purchased

Table 2 Privacy concerns for user information in recommender systems

↓ Concern/info. →	Behavioral	Contextual	Domain knowledge	Item meta-data	History	Recommendation	Feedback	Social	User attributes	User preferences
Data collection	●	•	·	·	●	•	●	•	•	•
Data retention	•	•	·	·	•	•	•	•	•	•
Data sales	●	●	·	·	●	●	●	●	●	●
Employee	●	●	·	·	●	●	●	●	●	●
Recommendations	·	·	·	·	●	·	•	•	·	●
Shared service	•	·	·	·	●	●	•	·	·	●
Stranger views	·	•	·	·	●	•	•	●	●	●

a gift for him. Unless she has a private account, her husband might inadvertently see her purchase or receive recommendations based on it. Many would want to keep some purchases private from their kids or their viewing behavior from their housemates. While some services allow for separate accounts, this is not always possible. For example, targeted advertising works with cookies that are stored in the browser, which is implicitly shared on a computer. This is related to the intended breadth of the audience.

Stranger Views Private Information. Users can falsely assume some information to be kept restricted to the service provider or a limited audience, when in reality it is not. This can be due to design flaws on the part of the service provider or a lack of the user's own understanding or attention to his privacy. When a stranger views such private information, there is a conflict with regard to the intended breadth of the audience. Rosenblum [33] showed, for example, that information in social networks is far more accessible to a widespread audience than perceived by its owners.

3.3 Summary

Because recommender systems typically contain a large amount of information, often about its users, they form an interesting target for attack. Information could end up in the wrong hands or be misused by legitimate data holders. Given the amount and detail of information within recommender systems, the privacy concerns should be taken seriously. Table 2 gives an overview of how different concerns impact different information within recommender systems. In this table, impact is either high (●), medium (•), or low (·). We can see that the impact on domain knowledge and item metadata is low for all privacy concerns. This is due to the fact that this data is not about the user but about the content items. The user's history and preferences have the highest privacy concerns. This is mainly due to their accurate representation of the user's opinions about items.

4 Research into Privacy-Protection Technologies

We have seen a wide variety of privacy issues associated with recommender systems. Research from many areas could be applied to alleviate some of the aforementioned concerns. We will provide an overview of research areas and briefly discuss their mechanisms, advantages, and limitations.

4.1 Awareness

Research in this mainly social field aims to enhance user awareness of the privacy issues that exist within online systems. It can aid users in specifying their privacy boundaries. The Platform for Privacy Preferences (P3P) [12] is an initiative that aims to provide websites with a standardized format in which they can define their privacy policy. Visitors of the website can then, through client-side *user agents* (e.g. plug-ins for their browser or applets), easily check the details of a privacy policy and see what will happen to information they submit. This system can help to increase user awareness but only for users that employ agents and if websites properly define their privacy policies and adhere to them.

Tsai et al. [37] showed that when privacy information is shown *more prominent*, and users are made more aware of the privacy consequences, privacy is taken into account when shopping online. Tsai et al.'s study also shows that some users are even willing to pay more for the product, if it means getting more privacy. Offering users the ability to opt-out for or opt-in to data collection would in many cases level the playing field between users and service providers.

4.2 Laws and Regulations

This legal field of research aims to find proper and broad laws and regulations that protect the users' privacy, while not greatly hindering businesses. It also focusses on compliance of both users and service providers to established laws and social conducts. Laws and regulations form an important and much needed tool. For example, the Article 29 Working Party has been working toward regulations for online behavioral advertising [21].

This legal approach runs after the technology, as specific laws dealing with personal information as related to the Internet often take long to be developed. Also, laws are generally used to solve matters *after* things go wrong, whereas most technical solutions attempt to *prevent* violations.

4.3 Anonymization

As pointed out in Sects. 2 and 3.2, sales of information can be a major source of revenue for service providers. If this were to be done without any further consideration for privacy, users might take offense and stop using the system (thus hurting revenue) or take justified legal action. Service providers may try to remove the privacy issues associated with data sales, by *obscuring the link between users and data sold* [36].

This can be done through anonymization, which involves removing any identifying (or identifiable) information from the data, while preserving other structures of interest in the data. As mentioned before, the information published by Netflix as part of their recommender systems' prize, though anonymized, allowed for re-identification [27]. This mainly stems from the fact that information can only be *partially* removed or obfuscated, while other parts *must be kept intact* for the dataset to remain useful. In the real world, it is difficult to predict which external sources of information may become available, allowing pieces of data to be combined into identifiable information.

When looking at anonymization during recommendation, Cissée and Albayrak [11] utilized trusted agents (essentially moving the trust around) to act as a relay and filter the information that is sent. This way, the user can interact (through the agent) with the recommender system in an anonymous way. The user hides his personal information from the service provider and is safe from the service provider linking his rating information to a person. However, the user still needs to trust that the agents (either hardware or software based) and the service provider do not collude.

4.4 Randomization and Differential Privacy

Similar to anonymization is randomization. In randomization (sometimes referred to as perturbation), the information fed into the system is altered to add a degree of uncertainty. Polat and Du [29] proposed a singular value decomposition predictor based on random perturbation of data. The user's data is perturbed by adding a random value (from a fixed distribution) to each of the ratings; unknown ratings are filled in with the mean rating. They go on to show the impact on privacy and accuracy and their inherent trade-off due to perturbation. In later work [30], their setting is different. A user wants two companies to collaboratively compute recommendations for him. This user acts as a relay for the two companies. The user's privacy is based on randomizing values. Berkovsky et al. [6] proposed to combine random perturbation with a peer-to-peer structure to create a form of dynamic random perturbation. For each request, the user can decide what data to reveal and how much protection is put on the data. Different perturbation strategies are compared based on accuracy and perceived privacy. Shokri et al. [35] added privacy by aggregating user information instead of perturbing. Aggregation occurs between users, without interaction with the recommender system. Thus, the recommender system cannot identify which information is part of the original user information and what is added by aggregation. A degree of uncertainty is added to the user's information similar to randomization.

Recently the field of randomization is shifting toward differential privacy [13], which aims to obscure the link between single users' information in the input (the user's information) and output (the recommendation). This is accomplished by making users in released data computationally indistinguishable from most of the other users in that dataset. This is typically accomplished by adding noise to the inputs or output, to hide small changes that arise from a single user's contribution. The required level of noise depends on how and how often the data will be used and typically involves a balancing act between accuracy of the output and privacy of the input. Such indistinguishability also applies strongly to collaborative recommender systems, where a user should be unable to identify individual peers' ratings in the output he receives. As each recommendation leaks a little bit of information about the input (even with noise), with a larger number of recommendations, the added noise should be greater to provide the same level of privacy. McSherry and Mironov [26] proposed collaborative filtering algorithms in the differential privacy framework. Noise is added to the item covariance matrix (for item similarity). Since the item covariance matrix is smaller than the user covariance matrix, less noise needs to be added and more accuracy is preserved.

The drawback of these techniques is that the security of these methods is hard to be *formally proven*, as is done in classical cryptography. The noise levels in differential privacy techniques must not overwhelm the initial output data and thus remove utility of the results completely. At the same time, enough noise must be added in order to hide the contribution of a user. When combined with multiple computational results and external information, even more noise is needed to protect the privacy of a user.

4.5 Privacy-Preserving Cryptographic Protocols

We first give an overview of some of the tools used in privacy-preserving cryptographic protocols, before addressing the protocols themselves. Among the tools [17] are secure multiparty computations, secret sharing, homomorphic encryption, and zero-knowledge proofs.

Secure multiparty computations are a class of protocols that allow two or more parties to collaboratively compute a function based on input held by each of them. The output of this function can be given to one of the parties or all of them. Any function can be computed, but the complexity of the protocol depends on the function. For example, multiplication and integer comparison.

Secret sharing distributes a number of shares of a value among different parties. The shares of a fixed number of parties need to be combined in order to reconstruct the original value. With less than the fixed number of shares, no information about the value can be obtained. Some secret sharing schemes allow basic operations (such as addition) to be performed.

Homomorphic encryption allows one (or sometimes more) operation (e.g., addition or multiplication) on the encrypted values, by performing a corresponding operation on the ciphertexts. This allows anyone to compute a (basic) function on the encrypted values, without knowledge of the actual values. Decryption is then required to get the result of the function.

Zero-knowledge proofs allow a user to prove a property about a value, without revealing that value. For example, that a value is in a given range of possible values. To do this, the user first sends a commitment to the verifier. Then the verifier ask the user to open the commitment in a certain way. The commitment can only be opened correctly when the property of the value holds. With a certain probability, the user can correctly open the commitment even if the property does not hold. However, by running multiple zero-knowledge proofs, this percentage can be reduced.

Privacy-Preserving Cryptographic Protocols Without Server

Privacy-preserving cryptographic protocols without a central server aim to remove the trust that is placed in service providers by removing them from the picture. Secure multiparty computations protect the privacy of users against each other. Canny [9, 10] used a combination of secure multiparty computation, homomorphic encryption, and zero-knowledge proofs to create a privacy-preserving recommender protocols without a central server. The users collaborate to privately compute intermediate values of the collaborative filtering process. These intermediate values (based on all users) are then made public. In the next step, the users perform singular value decomposition and factor analysis, which leads to a model for recommendations. This model is made publicly available and can be used by each user independently to compute recommendations for themselves.

The system proposed by Hoens et al. [19] allowed trusted friends to collaboratively compute recommendations with each other. They rely on Facebook for retrieving friendship information and a server to facilitate asynchronous messaging. Homomorphic cryptography and secure multiparty protocols are used to compute the actual recommendations for a given item.

Because a decentralized structure works strongly toward taking power away from the service provider, it is contrary to existing business models. This means that existing companies are not likely to adopt such a structure or aid its development. Another drawback is the involvement of many users that is required to make (the model for) the recommendations. These users need to interact with each other, but not all users will be available at the same time. This can lead to considerable delays or a loss of accuracy.

Privacy-Preserving Cryptographic Protocols with Server

Privacy-preserving cryptographic protocols with a central server aim to make use of the centralization offered by the service provider, while using secure two-party computation and encryption to ensure the privacy of the users. Good motivations for the service provider would be a reduced liability for the data collected, an increased perception of security among users (and thus, a competitive advantage), and adherence to possible stricter future laws.

Aïmeur et al. [3] provided a framework for collaborative filtering, where user information is separately stored over two parties. An agent has access to ratings, and the company has access to the items, so that they together can generate recommendations for the user. The centralized structure is preserved, but neither the agent nor the company can link the user's ratings to the items. Erkin et al. [14, 15] proposed a collaborative filtering algorithm based on homomorphic cryptosystems. In their framework, a central server acts as a mediator between the users and is in charge of combining the results given by different users. When desiring a recommendation, a user sends an encrypted request to the central server. The server distributes this request to other users that can work on the request by using the homomorphic properties of the cryptosystem. A secure two-party computation then determines for each user if their information should be included in the recommendation or not. The central server then combines the (still encrypted) results to generate the recommendation.

Basu et al. [4,5] proposed a privacy-preserving version of the slope one predictor for collaborative filtering. The assumption is that different parties hold different parts of the information, this essentially allows multiple companies to collaborate. They precompute the deviation and cardinality matrices under encryption and make the cardinality matrix public. Then the prediction for a single item can be computed under encryption and all parties collaborate to decrypt the result.

The drawback of these schemes (that add a layer of encryption) is efficiency. The homomorphic operations and secure two-party computations are always more expensive than their unprotected counterparts. In fact, the discrepancy is often huge. This results in poor efficiency and scalability for these protocols, an issue that the research tries to address.

4.6 Summary

Table 3 shows which research areas contribute to address which privacy concern: a ● indicates that the area is helpful to address a particular concern, a • indicates that the area is somewhat helpful, while a · indicates that the area does not seem applicable. The majority of the research areas focusses on protecting the user's information from the service provider. As can been seen in Sect. 3.3, the privacy concerns related to the service provider have a high privacy impact.

None of the research areas mentioned in this section can offer complete user privacy for all recommender systems. Privacy is multifaceted, as are the domains in which recommender systems are applied. Several areas will likely need to be combined to develop proper privacy-protection techniques for a given application. In addition, service providers should be encouraged or required to implement such solutions, and users need to be made aware of the benefits of using them.

Table 3 Privacy concerns and relevant research areas

↓ Concern/research →	Awareness	Law	Anonymization	Randomization	Protocols w/o server	Protocols w/ server
Data collection	•	●	·	•	●	●
Data retention	·	●	·	·	●	●
Data sales	·	●	●	•	●	●
Employee	·	●	·	•	●	●
Recommendations	·	·	·	●	·	·
Shared service	•	·	·	·	·	·
Stranger views	•	·	·	•	•	•

5 Conclusion

We have seen that recommender systems play an important role in the online experience of millions of people. While accuracy has been the focus in recommender system development, we argue that privacy should not be overlooked. We have seen that depending on the type of information utilized by a recommender system, various privacy concerns exist. The fact that trust in the service provider is not always justified further complicates matters. With increased information sharing, users must weigh the advantages of getting (more accurate) recommendations against the privacy risks and should more often be given the choice to opt-in or opt-out of data collection.

Many areas of research can help to protect user privacy, ranging from technical (e.g., system design and cryptography) to nontechnical (e.g., sociology and law). However, we must realize that one single research area cannot address all privacy concerns. Furthermore, we notice a trend in the different research areas. The areas of awareness and law do not focus on any single specific type of recommender system. However, the areas that provide technical solutions mainly focus on collaborative filtering recommender systems. The other types of recommender systems are barely (if at all) represented.

As commonly known, in the technical solutions, there is an inherent trade-off between privacy, accuracy, and efficiency. Randomization techniques increase privacy by lowering accuracy and leaving efficiency the same. Cryptographic and secure multiparty computation protocols increase privacy by lowering efficiency and leaving accuracy the same. However, when aiming for a specific trade-off in a certain scenario and goal, it is difficult to choose the right solution. Comparison is difficult because researchers use different datasets and different measures for accuracy. It is an open question how different privacy-protection techniques compare to each other when applied to the same dataset, with the same accuracy measure, and the same programming language and hardware.

Our conclusion is that in order to develop a full solution to protect user privacy, the strengths of several research areas will need to be brought together. Ideally, privacy-protection techniques are built into the system design. These privacy-protection techniques should not harm the operations of the recommender system. Therefore, the users and the service provider should not be overburdened, and the functionality and accuracy of the recommender system should not be hampered.

Acknowledgements The research for this work was carried out within the Kindred Spirits project, part of the STW Sentinels research program.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
2. Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 217–253. Springer, New York (2011)
3. Aïmeur, E., Brassard, G., Fernandez, J., Mani Onana, F.: Alambic: a privacy-preserving recommender system for electronic commerce. Int. J. Inf. Secur. **7**(5), 307–334 (2008)
4. Basu, A., Kikuchi, H., Vaidya, J.: Privacy-preserving weighted slope one predictor for item-based collaborative filtering. In: Proceedings of the International Workshop on Trust and Privacy in Distributed Information Processing. Springer, Berlin/Heidelberg (2011)
5. Basu, A., Vaidya, J., Kikuchi, H.: Efficient privacy-preserving collaborative filtering based on the weighted slope one predictor. J. Internet Serv. Inf. Secur. (JISIS) **1**(4), 26–46 (2011)
6. Berkovsky, S., Eytani, Y., Kuflik, T., Ricci, F.: Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In: Proceedings of the 2007 ACM Conference on Recommender Systems, Minneapolis, pp. 9–16 (2007)
7. Burke, R.: Knowledge-based recommender systems. In: Encyclopedia of Library and Information Systems, vol. 69, pp. 180–200. Marcel Dekker, New York (2000)
8. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User Adapt. Interact. **12**, 331–370 (2002)
9. Canny, J.: Collaborative filtering with privacy. In: IEEE Symposium on Security and Privacy, Oakland, pp. 45–57 (2002)
10. Canny, J.: Collaborative filtering with privacy via factor analysis. In: Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval, Tampere, pp. 238–245 (2002)
11. Cissée, R., Albayrak, S.: An agent-based approach for privacy-preserving recommender systems. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '07, pp. 182:1–182:8. ACM, New York (2007)
12. Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J.: The platform for privacy preferences 1.0 (p3p1.0) specification. http://www.w3.org/TR/P3P/
13. Dwork, C.: Differential privacy. In: Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, 10–14 July 2006, Proceedings, Part II, pp. 1–12 (2006)
14. Erkin, Z., Beye, M., Veugen, T., Lagendijk, R.L.: Privacy enhanced recommender system. In: Thirty-First Symposium on Information Theory in the Benelux, Rotterdam, pp. 35–42 (2010)

15. Erkin, Z., Beye, M., Veugen, T., Lagendijk, R.L.: Efficiently computing private recommendations. In: International Conference on Acoustic, Speech and Signal Processing-ICASSP, Prague, pp. 5864–5867 (2011)
16. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM **35**, 61–70 (1992)
17. Goldreich, O.: Foundations of cryptography: a primer. Found. Trend Theor. Comput. Sci. **1**, 1–116 (2005)
18. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: WPES '05: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, pp. 71–80. ACM, New York (2005)
19. Hoens, T.R., Blanton, M., Chawla, N.V.: A private and reliable recommendation system for social networks. In: 2010 IEEE Second International Conference on Social Computing (SocialCom), Minneapolis, pp. 816–825 (2010)
20. Kang. J.: Information privacy in cyberspace transactions. Stanf. Law Rev. **50**(4), 1193–1294 (1998)
21. Kohnstamm, J.: Opinion 2/2010 on online behavioural advertising. Technical report 00909/10/EN WP 171, article 29 data protection working party, 6 2010. http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp171_en.pdf
22. Konstas, I., Stathopoulos, V., Jose, J.M.: On social networks and collaborative recommendation. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, pp. 195–202. ACM, New York (2009)
23. Lam, S., Frankowski, D., Riedl, J.: Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. In: Müller, G. (ed.) Emerging Trends in Information and Communication Security. Lecture Notes in Computer Science, vol. 3995, pp. 14–29. Springer, Berlin/Heidelberg (2006)
24. Lang, K., Newsweeder: learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339. Morgan Kaufmann, San Francisco (1995)
25. Lorenzi, F., Ricci, F.: Case-based recommender systems: a unifying view. In: Intelligent Techniques for Web Personalization. Lecture Notes in Computer Science, vol. 3169, pp. 89–113. Springer, Berlin/Heidelberg (2005)
26. McSherry, F., Mironov, I.: Differentially private recommender systems: building privacy into the netflix prize contenders. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, pp. 627–636 (2009)
27. Narayanan, A., Shmatikov, V.: How to break anonymity of the netflix prize dataset. In: CoRR: Computing Research Repository, pp. 1–24. Cornell University, Ithaca (2006)
28. Palen, L., Dourish, P.: Unpacking "privacy" for a networked world. In: CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 129–136. ACM, New York (2003)
29. Polat, H., Du, W.: Svd-based collaborative filtering with privacy. In: Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, pp. 791–795 (2005)
30. Polat, H., Du, W.: Privacy-preserving top-n recommendation on distributed data. J. Am. Soc. Inf. Sci. Technol. **59**, 1093–1108 (2008)
31. Ramakrishnan, N., Keller, B.J., Mirza, B.J., Grama, A.Y., Karypis, G.: Privacy risks in recommender systems. IEEE Internet Comput. **5**(6), 54–62 (2001)
32. Rich, E.: User modeling via stereotypes. Cognit. Sci. **3**(4), 329–354 (1979)
33. Rosenblum, D.: What anyone can know: the privacy risks of social networking sites. IEEE Secur. Priv. **5**(3), 40–49 (2007)
34. Schclar, A., Tsikinovsky, A., Rokach, L., Meisels, A., Antwarg, L.: Ensemble methods for improving the performance of neighborhood-based collaborative filtering. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 261–264. ACM, New York (2009)

35. Shokri, R., Pedarsani, P., Theodorakopoulos, G., Hubaux, J.-P.: Preserving privacy in collaborative filtering through distributed aggregation of offline profiles. In: Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pp. 157–164. ACM, New York (2009)
36. Sweeney, L.: K-anonymity: a model for protecting privacy. IEEE Secur. Priv. **10**(5), 557–570 (2002)
37. Tsai, J.Y., Egelman, S., Cranor, L., Acquisti, A.: The effect of online privacy information on purchasing behavior: an experimental study. Inf. Syst. Res. **22**, 254–268 (2011)
38. Tufekci, Z.: Can you see me now? audience and disclosure regulation in online social network sites. Bull. Sci. Technol. Soc. **28**(1), 20–36 (2008)
39. Weiss, S.: The need for a paradigm shift in addressing privacy risks in social networking applications. In: The Future of Identity in the Information Society, Brno, vol. 262, pp. 161–171. IFIP International Federation for Information Processing. Karlstad (2008)