Artificial Intelligence Review (2020) 53:291–319
https://doi.org/10.1007/s10462-018-9655-x

Shilling attacks against collaborative recommender systems: a review

Mingdan Si[1] · Qingshan Li[1]

Published online: 19 September 2018

Abstract
Collaborative filtering recommender systems (CFRSs) have already been proved effective to cope with the information overload problem since they merged in the past two decades. However, CFRSs are highly vulnerable to shilling or profile injection attacks since their openness. Ratings injected by malicious users seriously affect the authenticity of the recommendations as well as users' trustiness in the recommendation systems. In the past two decades, various studies have been conducted to scrutinize different profile injection attack strategies, shilling attack detection schemes, robust recommendation algorithms, and to evaluate them with respect to accuracy and robustness. Due to their popularity and importance, we survey about shilling attacks in CFRSs. We first briefly discuss the related survey papers about shilling attacks and analyze their deficiencies to illustrate the necessity of this paper. Next we give an overall picture about various shilling attack types and their deployment modes. Then we explain profile injection attack strategies, shilling attack detection schemes and robust recommendation algorithms proposed so far in detail. Moreover, we briefly explain evaluation metrics of the proposed schemes. Last, we discuss some research directions to improve shilling attack detection rates robustness of collaborative recommendation, and conclude this paper.

Keywords Profile injection attack · Shilling attack · Collaborative filtering · Robustness · Attack detection

1 Introduction

Recommender systems have emerged in the past two decades as an effective way to cope with the information overload problem by suggesting information that is of potential interest to online users (Adomavicius and Tuzhilin 2005; Koren 2010; Sarwar et al. 2001). They are not only helping customers find relevant information buried in a great deal of irrelevant

✉ Qingshan Li
qshli@mail.xidian.edu.cn

Mingdan Si
mingdansi@163.com

[1] School of Computer Science and Technology, Xidian University, Xi'an 710000, China

information, but are also beneficial to the companies producing merchandise by increasing both selling rate and cross-sales and improving consumers' loyalty, because consumers tend to return to the sites which best serve their needs (Schafer et al. 2001). Recommendation systems, especially the collaborative filtering (CF)-based systems, have been successfully introduced to filter out irrelevant resources (Adomavicius and Tuzhilin 2005; Cho and Kim 2004; Gao et al. 2010; Montaner et al. 2003; Ronen et al. 2013; Silva et al. 2015; Vozalis and Margaritis 2007; Wang and Zhang 2013; Yuan et al. 2013). We have witnessed recommender systems having been widely accepted in many different domains, such as online learning and teaching (Cechinel et al. 2013; Chen et al. 2014; Cobos et al. 2013; Elbadrawy and Karypis 2016), digital library (Serrano-Guerrero et al. 2011; Tejeda-Lorente et al. 2014), online social networks (Agarwal and Bharadwaj 2013; Wang et al. 2013; Wei et al. 2013), movie and TV programs (Barragáns-Martínez et al. 2010; Shi et al. 2016), health (Abbas et al. 2015; Rivero-Rodriguez et al. 2013; Wiesner and Pfeifer 2014) and tourism (Borràs et al. 2014; Gavalas et al. 2014).

Currently, recommendation methods are usually classified into the following three main categories: content-based recommendation, collaborative filtering, and hybrid recommendation approaches (Adomavicius and Tuzhilin 2005). Collaborative filtering recommender systems (CFRSs) operate on the basis that similar users have similar tastes, and is one of the most popular and successful techniques in recommender systems (Bobadilla et al. 2013; Jia and Liu 2015; Koren 2008; Sarwar et al. 2001). However, the open and interactive nature of collaborative filtering are both a source of strength and vulnerability for recommender systems (Lam and Riedl 2004; Mobasher et al. 2006b; O'Mahony et al. 2004). It is easy to see why CFRSs are vulnerable to shilling attacks (O'Mahony et al. 2004) or profile injection attacks (Mobasher et al. 2006b; Williams and Mobasher 2012). A user-based collaborative filtering algorithm makes recommendations by finding neighbors with similar user profiles, which are assumed to represent the preferences of many different individuals. If the profile database contains biased data, these biased profiles might be considered neighbors for genuine users and eventually result in biased recommendations.

Unscrupulous producers may opt to take a more deceitful route to influence recommender systems that their items are recommended to users more often, whether or not they are of high quality. This is precisely the bad negative impacts found in Lam and Riedl (2004) and O'Mahony et al. (2004). For example, an instance of a company generating false "recommendations" to consumers arose in June 2001 when Sony Pictures admitted that it had used fake quotes from non-existent movie critics to promote a number of newly released films.[1] The online retailer Amazon.com has found that their online retailer pulls a link to a sex manual that appeared next to a spiritual guide by well-known Christian televangelist Pat Robertson.[2] Amazon conducted an investigation and determined these results were not that of hundreds of customers going to the same items while they were shopping on the site. On the contrary, it is the unscrupulous producers taking a more deceitful artifice has led to the absurd recommendation results. Also, eBay, which uses a recommender system as a reputation mechanism, has found itself continually dealing with users who subvert the system in various ways, including purchasing good ratings (feedback) from other members in order to bolster their own reputations.[3] To protect such personal preferences, privacy-preserving collaborative filtering (PPCF) systems have been developed (Bilge and Polat 2013; Casino et al. 2013, 2015; Jeckmans et al. 2013; Ozturk and Polat 2015). However, researchers have

[1] http://news.bbc.co.uk/2/hi/entertainment/1368666.stm.

[2] https://www.cnet.com/news/amazon-blushes-over-sex-link-gaffe/.

[3] http://www.auctionbytes.com/cab/abn/y03/m09/i17/s01.

proved that various PPCF systems are also vulnerable to such shilling attacks (Bilge et al. 2014a; Gunes et al. 2013a, b).

It is imperative to handle shilling attacks or profile injection attacks for the sake of the overall success of CF or PPCF algorithms. Due to its importance, researchers have been giving increasing attention to such attacks. In the literature, some researchers focus on shilling attack detection schemes (Chirita et al. 2005; Mehta 2007; O'Mahony et al. 2003; Yang et al. 2016b; Zhang and Zhou 2014), some of them study shilling attacks and their types (Bhaumik et al. 2011; O'Mahony 2004; O'Mahony and Smyth 2007a, b; Williams et al. 2006), and while others scrutinize how to develop robust CF algorithms or enhance the robustness of CF systems against profile injection attacks (Bilge et al. 2014a; Gunes et al. 2013a, b; Gunes and Polat 2015). The researchers also evaluate various shilling attacks using different metrics on benchmark data sets (Lam and Riedl 2004; Mobasher et al. 2006b; O'Mahony et al. 2004).

Profile injection attacks can be categorized based on the knowledge required by the attackers to mount the attack, the intent of a particular attack, and the size of the attack (Williams and Mobasher 2012). According to intent, shilling attacks can be grouped as push attacks, nuke attacks and random vandalism (Burke et al. 2015; Mobasher et al. 2007b). The attacks might be classified as low-knowledge attacks and high-knowledge attacks (Burke et al. 2015). In addition, they can also be categorized based on the size of the attack (Williams and Mobasher 2012). Although there are some studies surveying shilling attacks against CF or PPCF algorithms, they come short discussing such attacks and detection schemes. There is no comprehensive survey discussing shilling attack strategies, attack detection schemes, robust CF recommendation algorithms, and their evaluation in detail for the last 5 years.

In this study, we present a review about shilling attacks against various CF algorithms. We investigate major research directions, such as employing attack modes, detecting shilling attacks, and developing robust CF algorithms with respect to shilling attacks. What's more, we describe evaluation metrics and future research directions which may be effective to improve shilling attack detection rates and robustness of CFRSs.

The paper is structured, as follows: in Sect. 2, we briefly discuss related studies focused on surveying about shilling attacks. After discussing profile injection attacks in detail in Sect. 3, we study major research directions about profile injection attacks in Sect. 4. We then introduce evaluation metrics of the proposed schemes in Sect. 5. In Sect. 6, we summarize the shortcomings of existing attack detection methods, and discuss some future research directions which may improve shilling attack detection and robustness of CFRSs. Finally, we conclude our paper in Sect. 7.

2 Related work

Collaborative filtering schemes are deployed commonly by e-commerce sites to entice customers and they are publicly available. However, they are not strictly robust enough to resist shilling attacks (O'Mahony et al. 2004) or profile injection attacks (Mobasher et al. 2006b; Williams and Mobasher 2012) since their openness (Bilge et al. 2014a; Chirita et al. 2005; Gunes et al. 2013b). Generally, such shilling attacks are applied to either push/nuke popularity of specific merchandise or just damage overall performance of recommendation systems by injecting biased profiles.

One of the earliest instances of a collaborative filtering based recommender system is Tapestry (Goldberg et al. 1992), a mail filtering software developed in the early nineties for the intranet at the Xerox Palo Alto Research Center. In 1994, Resnick et al. (1994) automated the collaborative filtering process and introduced an automated collaborative filtering (ACF) algorithm based on k-Nearest-Neighbor. Since then, a number of improvements to *k*NN have been proposed (Good et al. 1999; Herlocker et al. 1999; Sarwar et al. 2001), and the research about CFRSs has been growing.

Fraudulent behavior, such as bogus ratings, is first discussed in Dellarocas (2000). Dellarocas (2000) discusses fraudulent behaviors against reputation reporting systems and proposes a set of mechanisms, which can reduce even eliminate the negative effects of such fraudulent behavior for the sake of constructing more robust online reputation reporting systems. To the best of our knowledge, attacks on CFRSs is first discussed in 2002 (O'Mahony et al. 2002a, b). O'Mahony et al. (2002a, b) give the definition of system robustness and argue vulnerabilities of various CFRSs against shilling attacks to promote specific recommendations. Since then, researchers have been studying on defining such shilling attacks, detecting profile injection attacks and developing robust CF algorithms against known attack types. Moreover, there are a number of studies compiling up-to-date developments in this field. In other words, some researchers focus on surveying about shilling attacks and their effects on recommendation systems.

Mehta and Hofmann (2008) survey about robust CF recommendation algorithm only. They describe several state-of-the-art algorithms for robust collaborative filtering systems, including intelligent neighbor selection, association rules, Probabilistic Latent Semantic Analysis (PLSA), and Singular Value Decomposition (SVD), and summarize characteristics that a robust recommender algorithm should have.

In another survey paper, Zhang (2009) presents a survey of existing research on the shilling attack types, algorithm dependence, attack detection schemes and evaluation metrics for shilling attacks. They describe shilling attack types such as random attack, average attack, bandwagon attack, segment attack, and reverse bandwagon attack. Moreover, they explain some attributes such as generic attributes, model-specific attributes and intra-profile attributes to detect shilling attacks. They then propose three metrics, including prediction shift, hit ratio and ExpTopN, to evaluate the efficiency of different shilling attack types.

Similarly, Gunes et al. (2014) discuss the up-to-date research about shilling attacks covering the studies focusing on shilling attack strategies, profile injection attack types, attack detection algorithms, robustness analysis, robust algorithms, and cost–benefit analysis. To the best of our knowledge, this is the most latest and comprehensive survey about shilling attacks so far. Moreover, they have also discussed some missing works which should be completed for further research.

In addition to the abovementioned survey papers, some other researchers have summarized the shilling attack detection schemes (Burke et al. 2005c; Mobasher et al. 2007a, b).

As stated previously, the survey conducted by Mehta and Hofmann (2008) describe robust CF approaches only. However, major researches in this field fall into four major categories: describing shilling attack strategies, studying shilling attack detection schemes, developing robust algorithms or enhancing robustness of CF recommendation, and evaluating proposed schemes. Thus, their survey focuses only on one of the major trends and leaves more works to be done. Similarly, Zhang (2009) discusses limited number of attack types, attack detection strategies, and evaluation metrics; thus, falls short leaving more works to be done. To the best of our knowledge, Gunes et al. (2014) have made the most comprehensive survey about shilling attacks so far. However, they introduce information about manipulating recommender systems until 2011 only. Furthermore, there are several new works about attack types, shilling attack detection schemes and robustness of recommender systems presented since then. Hence, in this paper, we extensively discuss latest developments of attack types, attack detection schemes, robust recommendation algorithms or techniques to improve robustness. We also give some future research directions which may improve shilling attack detection rates and robustness of CFRSs.

3 Profile injection attack

In this subsection, we will give a detailed description about shilling attacks (O'Mahony et al. 2004) or profile injection attacks (Williams and Mobasher 2012), such as attack intent, attack/filler size, attack cost, target item, attack model and attack types.

3.1 Attack intent

Different shilling attacks may have very different intents, but, the eventual goal of an attacker may be one of several alternatives (Burke et al. 2015; Lam and Riedl 2004). Two specific intents are "push" and "nuke". An attacker may insert bogus profiles into recommender systems to make a merchandise more likely ("push") or less likely ("nuke") to be recommended. Another possible aim of shilling attack is random vandalism (Burke et al. 2015), which is just to interfere with the recommendation algorithm with the goal of causing users to stop trusting the system and eventually to stop using it.

3.2 Attack/filler size

The strength of shilling attacks is specified using two metrics: filler size and attack size.

Filler size is the number of ratings assigned in a given attack profile (Williams and Mobasher 2012). The addition of ratings has a relatively lower cost for attackers when compared with the cost of creating additional attack profiles. Moreover, common users usually rate more than a small fraction of all the merchandise in a large recommendation space and no one can read every book that is published or view every movie. Thus, the filler size is usually set to 1–20%.

Attack size is the number of bogus profiles injected into the system by the attackers (Williams and Mobasher 2012). The number of profiles injected into a well-developed recommender system is usually set to 1–15%, because a shilling attack has an associated cost value which depends on the level of effort and information needed to successfully execute the attack.

3.3 Attack cost

From the perspective of the attackers, the best attack against a system is one that yields the biggest impact for the least amount of effort (Burke et al. 2005d). There are two types of effort involved in mounting an attack.

The first is the effort involved in crafting profiles, which is called *knowledge-cost* (Burke et al. 2005d). Knowledge-cost can be divided into high-knowledge attack and low-knowledge attack. A high-knowledge attack requires the attackers to know detailed knowledge of the ratings distribution in a recommender system's database. For example, some shilling attacks that require the attackers to know the mean rating and standard deviation for every item. A low-knowledge attack is one that requires system-independent knowledge such as knowledge might be obtained by consulting public information sources.

The second aspect of the effort is the number of profiles that must be added to the system for the sake of the attack being effective, which is called *deployment-cost* (Burke et al. 2005d). The ratings are less important since the insertion of ratings can be easily automated. Most sites employ online registration schemes requiring human intervention, and by this means, the recommender system operator can impose a cost on the creation of new user profiles, which explains why shilling attacks require only a smaller number of profiles are particularly attractive from an attacker's perspective.

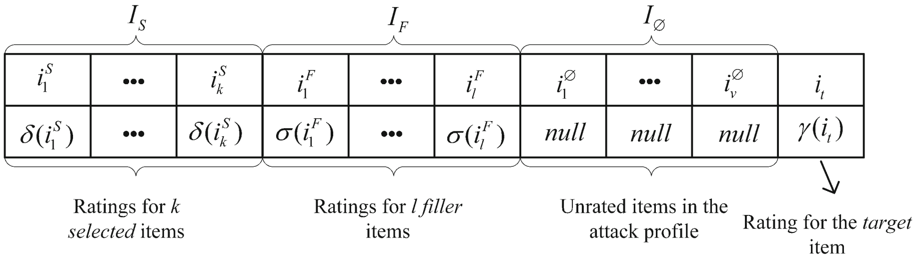
3.4 Target item

Item popularity (O'Mahony et al. 2005) refers to the number of ratings received by each item. Item likeability (O'Mahony et al. 2005) refers to the average rating received by each item. Items with low popularity and low likeability or "New" items that have few ratings are usually easier to be attacked because their average rating can be easily manipulated by a group of attackers. The target item will be rated as either the maximum rating ("push") or the minimum rating ("nuke") depending on attack intent.

3.5 Attack model

Attackers usually conduct shilling attacks by injecting an attack profile as shown in Fig. 1, which is first defined in Bhaumik et al. (2006) and Mobasher et al. (2007a) to mislead the CFRSs. However, in real life, an attacker may attack multiple target items simultaneously. Yang et al. (2016a) and Chung et al. (2013) have proposed creating a multi-target attack profile as shown in Fig. 2. An attack model (Bhaumik et al. 2006; Mobasher et al. 2007a) is an approach for constructing attack profiles, based on the knowledge about recommender system, including rating database, products, and/or users.

The basic attack profile consists of an n-dimensional vector of ratings, where n is the total number of items in the system. The set of target items, I_T, together with a rating function γ assigning it a rating value, and will be rated as either r_{max} ("push") or r_{min} ("nuke") depending on the attack intent. The set of selected items, I_S, with particular characteristics determined by the attacker, which is usually used to perform group attacks. The set of filler items, I_F, usually chosen randomly, together with a rating function σ mapping I_F to rating values, to make the profile look normal and harder to detect. The set of unrated set, I_Φ, contains


I_S
I_F
$I_\varnothing$
Ratings for *k* *selected* items
Ratings for *l filler* items
Unrated items in the attack profile
Rating for the *target* item

i_1^S	•••	i_k^S	i_1^F	•••	i_l^F	$i_1^\varnothing$	•••	$i_v^\varnothing$	i_t
$\delta(i_1^S)$	•••	$\delta(i_k^S)$	$\sigma(i_1^F)$	•••	$\sigma(i_l^F)$	*null*	*null*	*null*	$\gamma(i_t)$

Fig. 1 The general form of a single-target attack profile

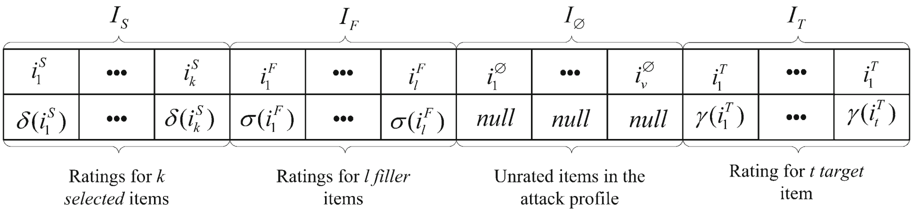

I_S
I_F
$I_\varnothing$
I_T
i_1^S ••• i_k^S
i_1^F ••• i_l^F
$i_1^\varnothing$ ••• $i_v^\varnothing$
i_1^T ••• i_1^T
$\delta(i_1^S)$ ••• $\delta(i_k^S)$
$\sigma(i_1^F)$ ••• $\sigma(i_l^F)$
null null null
$\gamma(i_1^T)$ ••• $\gamma(i_t^T)$
Ratings for *k* *selected* items
Ratings for *l filler* items
Unrated items in the attack profile
Rating for *t target* item

Fig. 2 The general form of a multi-target attack profile

those items not rated in the profile. The main differences between the different shilling attack models are embodied in the selection of filler items, selected items, and rating strategies.

3.6 Attack types

In this subsection, we will introduce most well-known attack models against CFRSs and their attack deployment strategies.

3.6.1 Basic attacks

Two basic attack models, introduced originally in Lam and Riedl (2004), are the random and average attack models. Both of these attacks involve the generation of profiles using randomly assigned ratings to the filler items in the profile.

Random attack In the random attack, the set of selected items I_S are empty. The set of target items I_T, are assigned the maximum rating (r_{max}) or the minimum rating (r_{min}) in the case of push or nuke attack intent, respectively. The set of filler items I_F are assigned to random ratings with a normal distribution around the mean rating value across the whole database. The random attack profile is depicted in Table 1. This type of attack is alternatively called RandomBot attack (Chirita et al. 2005; Lam and Riedl 2004).

The knowledge required to mount random attack is quite minimal, especially since the overall rating mean in many systems can be determined by an outsider empirically (or, indeed, may be available directly from the system). However, this attack is not particularly effective (Burke et al. 2005c; Lam and Riedl 2004).

Average attack Average attack is a more powerful attack described in Lam and Riedl (2004), which uses the individual mean for each item rather than the overall rating mean (except for

Table 1 The form of random and average attack profile

Attack model	I_S	I_F	I_Φ	I_T
Random	Null	Random ratings with a normal distribution around the mean rating value across the whole database	Null	r_{max}/r_{min}
Average	Null	Random ratings with a normal distribution around the mean rating for item i in I_F	Null	r_{max}/r_{min}

the target item). In the average attack, each assigned rating for a filler item corresponds to the mean rating for that item, across all users in the database who have rated it. The average attack profile is depicted in Table 1. The average attack is similar to random attack, and the only difference between the average attack and the random attack is the manner in which ratings are assigned to the filler items in the profile. Alternatively, it is called AverageBot attack (Hurley et al. 2007; Lam and Riedl 2004).

The average attack might be considered to have considerable knowledge cost of order $|I_F|$ because the mean and standard deviation of these items must be known. Experiments have shown that the average attack can be just as successful even when using a small filler item set (Burke et al. 2005d).

3.6.2 Low-knowledge attacks

The average attack requires a relatively high degree of system-specific knowledge on the part of attackers. The next set of attack types are those for which the knowledge requirements are much lower.

Bandwagon attack Bandwagon attack, also known as a popular attack (O'Mahony et al. 2006b), is similar to the random attack, and can be viewed as an extension of the random attack. Bandwagon attack takes advantage of the *Zipf*'s distribution of popularity in consumer markets: a small number of items, for example, bestseller books, will receive the lion's share of attention and also ratings.

There are two variant bandwagon attack models, which could be called bandwagon (average) attack and bandwagon (random) attack (Wu et al. 2012; Yang et al. 2016a).The bandwagon (average) and bandwagon (random) attack use selected items which have high item popularity in the database. These items are assigned the maximum rating value r_{max} together with the target items. In bandwagon (average) attack, the ratings for the filler items are determined randomly in a similar manner as in the average attack. While in bandwagon (random) attack, the ratings for the filler items are determined randomly in a similar manner as in the random attack. The form of bandwagon attack profile is shown in Table 2.

Bandwagon (random) attack is easy to implement because it requires public knowledge rather than domain specific knowledge and as effective as the average attack (Cheng and Hurley 2010; O'Mahony et al. 2005, 2006b). Thus it is more practical to mount.

Segment attack Mobasher et al. (2005) introduced the segment attack and demonstrated its effectiveness against the item-based algorithm. Segment attack is designed to target a specific group of users who are likely to buy a particular product (Mobasher et al. 2005, 2006b). In other words, it is likely that an attacker wishing to promote a particular product will be interested not in how often it is recommended to all users, but how often it is recommended to likely buyers. For example, the producer of a horror movie might want to get the movie recommended to film viewers who have seen and liked other horror movies. In attack profiles, attackers insert the maximum rating value r_{max} for items which users in the segment probably like, together with the target items, and the minimum rating value r_{min} for the filler items, thus maximising the variations of item similarities (Burke et al. 2005a, b; Hurley et al. 2007; Mobasher et al. 2007a).

Probe attack Probe attack (Mobasher et al. 2007a; O'Mahony et al. 2005) generates profiles by response to recommender itself. To perform this strategy, attackers assign authentic ratings

Table 2 The form of (reverse) bandwagon attack profile

Attack model	I_S	I_F	I_Φ	I_T
Bandwagon (average)	Popular items rated r_{max}	Random ratings with a normal distribution around the mean rating for item i in I_F	Null	r_{max}
Bandwagon (random)	Popular items rated r_{max}	Random ratings with a normal distribution around the mean rating value across the whole database	Null	r_{max}
Reverse bandwagon (average)	Least popular items rated r_{min}	Random ratings with a normal distribution around the mean rating for item i in I_F	Null	r_{min}
Reverse bandwagon (random)	Least popular items rated r_{min}	Random ratings with a normal distribution around the mean rating value across the whole database	Null	r_{min}

to create a seed profile and then use it to generate recommendations from the system. These recommendations are generated by the neighboring users (Mobasher et al. 2007a; O'Mahony et al. 2005). This way, a target item can be biased in a neighborhood based recommender system easily. In a sense, the probe attack provides a way for the attackers to incrementally learn about the system's rating distribution. Another advantage of this attack besides being simple is that it requires less domain knowledge (Burke et al. 2005c; Hurley et al. 2007; Mobasher et al. 2007a; O'Mahony et al. 2005, 2006a).

3.6.3 Nuke attack models

All of the attack models described above can also be used for nuking a target item. For example, in the case of the random and average attack models, this can be accomplished by associating the minimum rating r_{min} with the target items instead of the maximum rating r_{max}. However, attack models that are effective for pushing items, are not necessarily as effective for nuke attacks (Williams and Mobasher 2012). Thus, researchers have designed additional attack models particularly for nuking items.

Love/hate attack The love/hate attack is a very simple attack, with no knowledge requirements, but nonetheless effective attack against both user-based and item-based recommendation algorithm (Williams and Mobasher 2012). In attack profiles, randomly chosen filler items are rated with the minimum rating value r_{min} while the target items are given the maximum rating value r_{max}. Mobasher et al. (2007b) extended love/hate attacks to push items by switching the roles of r_{min} and r_{max}. The form of love/hate attack profile is described in Table 3.

Reverse bandwagon attack Reverse bandwagon attack (Mobasher et al. 2007a) is a variation of bandwagon attack discussed above to nuke particular products. In this attack, profiles are generated based on giving the minimum rating value to least popular items together with the set of target items. The ratings for the filler items are determined randomly in a similar manner as in the random attack. Similarly, reverse bandwagon attack is relatively easy to implement (Mobasher et al. 2007a). There are two variant reverse bandwagon attack models, which could be called reverse bandwagon (average) attack and reverse bandwagon (random)

Table 3 The form of love/hate attack profile

Attack model	I_S	I_F	I_Φ	I_T
Love/hate (push)	Null	Ratings assigned with r_{min}	Null	r_{max}
Love/hate (nuke)	Null	Ratings assigned with r_{max}	Null	r_{min}

attack (Wu et al. 2012; Yang et al. 2016a). The variant reverse bandwagon attack profile are also shown in Table 2. Experiments have shown that it is a very effective nuke attack against item-based recommender systems (Cheng and Hurley 2010).

3.6.4 High-knowledge attacks

Perfect knowledge attack Perfect knowledge attack (Williams and Mobasher 2012) is one such attack in which the attackers reproduce the precise details of the data distribution within the profile database. In a perfect knowledge attack, the biased profiles injected by the attackers match exactly with the profiles already in the system except that some particular items that they exhibit bias for (r_{min}) or against (r_{min}) (O'Mahony et al. 2004). However, for an attacker, it is not very realistic to be able to obtain so accurate information on data source.

Sampling attack Sampling attack (Burke et al. 2005d) is one such attack in which profiles are constructed from entire user profiles sampled from the actual profile database, augmented by a positive rating for the pushed item. Although it shows instability of CF algorithms, it is relatively impractical to realize this type of attack because original user profiles are not easily accessible (Burke et al. 2005d; Mobasher et al. 2005; O'Mahony et al. 2004).

Favorite item attack Favorite item attack, is also called the "consistency attack" in Burke et al. (2005c). Favorite item attack looks at knowledge of actual user preferences rather than knowledge about items. Such an attack is mounted not against the system as a whole, but by targeting a given user (or a group of users), and produces attack profiles consistent with popular or unpopular items of corresponding user for push and nuke attacks, respectively (Burke et al. 2005c, d).

3.6.5 Obfuscated attacks

From a practical view, attackers may attempt to conceal their injected attack profiles so that they could more effectively masquerade as genuine profiles, while still biasing the system. In order to achieve such objective, researchers have proposed strategies to make the signatures of attackers less prominent (Williams and Mobasher 2012; Williams et al. 2006). The deployment strategies of different obfuscated attacks are depicted in Table 4.

Noise injection Noise injection (Williams et al. 2006) is designed to mask the signature of common attack models. In profile, all the filler and selected items ($I_F \cup I_S$) will be added a noise according to a Gaussian distribution random number multiplied by a constant α. The deployment strategy of noise injection is described in Table 4. $R_{u,i}$ is the original assigned rating given to item i by attack profile u, $r_{u,i}$ is the rating assigned to item i by the obfuscated attack profile u.

Table 4 The form of obfuscated attack profile

Attack model	Deployment strategies
Noise injection	$\boldsymbol{I_F} \cup \boldsymbol{I_S} : R_{u,i} = r_{u,i}$ + random number × α; $\boldsymbol{I_T} : r_{max}$ or r_{min}
Target shifting	$\boldsymbol{I_F} \cup \boldsymbol{I_S} : R_{u,i} = r_{u,i}$; $\boldsymbol{I_T} : r_{max-1}$ or r_{min+1}
User shifting	$\boldsymbol{I_F} \cup \boldsymbol{I_S} : R_{u,i} = r_{u,i}$ + shift(u, O_s); $\boldsymbol{I_T} : r_{max}$ or r_{min}
Mixed attack	Injecting various shilling attack profiles simultaneously
Average over popular items	$\boldsymbol{I_F}$ is chosen with equal probability from the top X% of most popular items

Target shifting For a push attack, is simply shifting the rating given to the target item from the maximum rating to a rating one step lower, or in the case of nuke attacks increasing the target rating to one step above the lowest rating (Williams et al. 2006). The deployment strategy of target shifting is also depicted in Table 4.

User shifting It involves increasing or decreasing (shifting) all ratings for a subset of items per attack profile by a constant amount in order to reduce the similarity between attack users (Williams et al. 2006). Shifts can take the positive or negative form, and the amount of shift for each profile is governed by a standard normally distributed random number. The deployment strategy of user shifting is shown in Table 4. Where O_s is any subset of $I_F \cup I_S$ to be obfuscated, shift(u, O_s) is a function governing the amount to either increment or decrement all ratings within set O_s for profile u.

Mixed attack It involves attacking the same target item and producing from different attack models and attacking the different target item with different attack models (Bhaumik et al. 2011). The attack profiles contain the same amount of average, random, bandwagon and segment attack profiles.

The mixed attack is similar to the hybrid shilling attack described in Cao et al. (2013), which injects various types of shilling attack profiles into the recommender system simultaneously. The hybrid shilling attack also cannot be detected easily. The deployment strategy of mixed attack is shown in Table 4.

***Average over popular items* (AOP)** The average over popular items (AOP) attack (Hurley et al. 2009) is a simple and effective strategy to obfuscate the average attack. The deployment strategy of AOP attack is depicted in Table 4. The difference between average attack and AOP attack is depicted in Table 5. In average attack, the filler item is chosen from the entire collection of items except the target items. While in AOP attack profile, the filler item is chosen with equal probability from the top X% of most popular items rather than from the entire collection of items, where X is selected to ensure non-detectability.

Table 5 The form of average attack and average over popular items attack profile

Attack model	I_S	I_F	I_Φ	I_T
Average attack	Null	Chosen from entire collection of items except for the target items	Null	r_{max}/r_{min}
AOP attack	Null	Chosen with equal probability from the top X% of most popular items	Null	r_{max}/r_{min}

3.6.6 Other attacks

Power users, are those who can exert considerable influence over the recommendation outcomes presented to other users in CFRSs. Recommender system operators encourage the existence of power user communities and leverage them to help fellow users make informed purchase decisions. For example, amazon vine (https://www.amazon.com/gp/vine/help) invites the most trusted reviewers on Amazon to post opinions about new and pre-release items to help their fellow customers make informed purchase decisions. But, what happens when power users provide biased ratings for new items?

To address above issue, Wilson and Seminario (2013) and Seminario (2013) define and study a novel Power User Attack (PUA). Different from previous attack types, which has targeted the use of similarity-focused attack models that generate synthetic attack user profiles using random or average item ratings or a variant of these two approaches (Lam and Riedl 2004; Mobasher et al. 2007b), power user attack uses influential users to successfully attack CFRSs. Seminario and Wilson (2014b, 2016) also investigate a new complementary attack model, the Power Item Attack (PIA), which uses influential items to successfully attack recommender systems. The general form of Power User Attack and Power Item Attack profile is structured in Table 6.

Power user attack analyses rely critically on power user selection and power item attack rely critically on power item selection. There are three approaches to select power users or power items (Wilson and Seminario 2013; Seminario 2013; Seminario and Wilson 2014a, b, 2016), which include InDegree (or ID), Aggregated Similarity (or AS) and Number of Ratings (or NR).

PUA model In the PUA model, the set of filler items I_F is empty. The set of target items, I_T, are assigned the maximum rating (r_{max}) or the minimum rating (r_{min}) in the case of push or nuke attacks, respectively. The set of selected items, I_S, are combined by items and ratings copied from power users' profiles. The form of PUA profile is depicted in Table 6. There are three PUA attack types according to the ways to select power users.

(1) *PUA-AS attack*
The top 50 users with the highest aggregate similarity scores become the selected set of power users. The attack model requires at least 5 co-rated items between user u and v and does not use significance weighting (Wilson and Seminario 2013; Seminario 2013; Seminario and Wilson 2014a).

(2) *PUA-ID attack*
Based on the in-degree centrality concept from social network analysis (Wasserman and Faust 1994), power users are those who participate in the highest number of neighborhoods. For each user u, calculate his/her similarity with every other user v using significance weighting, then discard all but the top 50 neighbors for each user u. Count the number of similarity scores for each user v and select the top 50 users of v's (Wilson and Seminario 2013; Seminario 2013; Seminario and Wilson 2014a).

Table 6 The general form of PUA and PIA profile

Attack model	I_S	I_F	I_Φ	I_T
PUA model	Copy ratings and items from power user profiles	Null	Null	r_{max}/r_{min}
PIA model	Power items, ratings set with normal distributed around item mean	Null	Null	r_{max}/r_{min}

(3) *PUA-NR attack*
In PUA-NR attack, power users are those with the highest number of ratings, which is called the most active heuristic in Goyal and Lakshmanan (2012). In Wilson and Seminario (2013), Seminario (2013) and Seminario and Wilson (2014a), the authors choose the top 50 users based on the total number of ratings they have in their user profiles.

PIA model In the PIA model, the set of filler items I_F are empty. The set of target items, I_T, are assigned the maximum rating (r_{max}) or the minimum rating (r_{min}) in the case of push or nuke attacks, respectively. The set of selected items, I_S, are power items and each assigned rating for a selected item to the mean rating for that power item, across the users in the database who have rated it. The form of PIA profile is depicted in Table 6. There are also three PIA attack types according to the ways to select power items.

(1) *PIA-AS attack*
The top-N items with the highest aggregate similarity (AS) scores become the selected set of power items. The attack requires at least 5 users who have rated the same item *i* and item *j* (Seminario and Wilson 2014b, 2016). Different from the PUA-AS attack, PIA-AS attack does not use significance weighting.
(2) *PIA-ID attack*
Based on the in-degree centrality concept from social network analysis (Wasserman and Faust 1994), power items are those who participate in the highest number of similarity neighborhoods. For each item *i*, calculate its similarity with every other item *j* using significance weighting, then discard all but the top-N neighbors for each item *i*. Count the number of similarity scores for each item *j* and select the top-N items of *j*'s (Seminario and Wilson 2014b, 2016).
(3) *PIA-NR attack*
Power items are those items with the highest number of user ratings like popular items. We select the top-N items based on the total number of user ratings they have in their profiles (Seminario and Wilson 2014b, 2016).

Bandwagon and average hybrid attack The bandwagon and average hybrid attack model (Zhang 2011) combines average attack model with bandwagon attack model. In bandwagon and average hybrid attack model, the ratings for the filler items are determined randomly in a similar manner as in the random attack. The set of target items, I_T, are assigned the maximum rating (r_{max}) or the minimum rating (r_{min}) in the case of push or nuke attack intents, respectively. The set of selected items, I_F, include bandwagon items and average items. Firstly, the set of the selected bandwagon items are those have high item popularity in the database, which are assigned the maximum rating within the attack profile. Secondly, the selected average items partition denotes the items the attacker can knows the mean rating of the item, and their ratings will be given a rating according to the average attack model strategy. The form of bandwagon and average hybrid attack profile is shown in Table 7.

Random vandalism Three possible aims of shilling attacks are push, nuke and random vandalism (Burke et al. 2015), which is just to interfere with the recommendation algorithm with the goal of causing users to stop trusting the system and eventually to stop using it. Often, a lot of spam is purely junk, with no specific pattern, but random insertion of data. This phenomenon has been observed both with email spam and web spam (Mehta and Nejdl 2008). The deployment strategy of random vandalism profile is shown in Table 8.

Table 7 The form of bandwagon and average hybrid attack profile

Item	I_S	I_F	I_Φ	I_T
Rating	Selected bandwagon items rated r_{max}, average items deploy ratings with normal distributed around item mean	Random rated items with normal distributed around system mean	Null	r_{max}/r_{min}

Table 8 The general form of random vandalism profile

Attack model	$I_F \cup I_S$	I_Φ	I_T
Random vandalism	Random number in $[r_{min}, r_{max}]$	Null	r_{max}/r_{min}

4 Major research directions

Studies about profile injection attacks against CFRSs mainly focus on shilling attack strategies or generating profile injection attacks, shilling attack detection algorithms and robust CF algorithms against shilling attacks.

4.1 Shilling attack strategies

Initially, attack strategies towards existing CFRSs are discussed in Lam and Riedl (2004). Lam and Riedl (2004) explore four open questions that may affect the effectiveness of such shilling attacks and propose two attack strategies referred to as AverageBot and RandomBot. In each case, all available items are used in the attack profiles, and the issue of attack cost is not considered (Lam and Riedl 2004). In addition, Lam and Riedl (2005) also discuss attack effectiveness, difficulty, and detection concepts in general. What's more, Lam et al. (2006) extend their previous work by further discussing open researches about possible attacks on privacy-preserving prediction schemes. Later, O'Mahony (2004) developed an attack strategy which is called "Product Push Attack" in his Ph.D. dissertation. The basic strategy for the attackers is to create an attack profile which correlates strongly and positively/negatively with the ratings of a set of genuine users in the system, and to add to this the maximum/minimum rating for the target item (O'Mahony 2004). Product push attack include three different attack strategies such as random push attack, focused push attack and large attack profiles. Random push attack creates attack profiles by selecting a number randomly. Focused push attack is an attack strategy focusing on the correlation similarity metrics. Large attack profiles is developed to against a greater number of users. Then in 2005, extent of such knowledge is analyzed and it is presented that if even little such knowledge is known, effective attacks can be mounted on recommender systems (O'Mahony et al. 2005).

Also, Williams et al. (2006) have examined various obfuscate strategies to make the signatures of attackers less prominent, including noise injection, user shifting and target shifting. Ray and Mahanti (2009) explore the importance of target item and filler items in mounting effective shilling attacks. They proposed attack strategies are based on intelligent selection of filler items and the filler items are selected on the basis of the target item rating distribution. Bhaumik et al. (2011) propose an attack strategy which injects various types of shilling attack profiles into the recommender system simultaneously. Different from above attack strategies, which has targeted the use of similarity-focused attack models that generate synthetic attack user profiles using random or average item ratings or a variant of these two approaches (Lam and Riedl 2004; Mobasher et al. 2007b), Wilson and Seminario (2013), Seminario and Wilson (2014a, b) propose power user attack (PUA), which uses influential users to successfully attack CFRSs and power item attack (PIA), that uses influential items to successfully attack recommender systems.

4.2 Shilling attack detection

Various CF or PPCF recommender systems are proved vulnerable to shilling attacks (Bilge et al. 2014a; Burke et al. 2006a, b; Chirita et al. 2005; Gunes et al. 2013a, b). There are two common ways to reduce the influence of shilling attacks in CFRSs. One way is to perform shilling attack detection and remove the attack profiles from the rating database before running CF algorithms, and a number of schemas have been proposed to detect such shilling attacks. Another alternative way is to develop attack-resistant collaborative filtering algorithms, i.e., robust recommendation algorithms. In first subsection, we will introduce shilling attack detection algorithms, which can mainly be thought as supervised classification techniques, unsupervised clustering techniques, semi-supervised techniques and other techniques.

4.2.1 Supervised classification

To detect shilling attacks, several algorithms have been proposed. The earliest detection algorithm was proposed by Chirita et al. (2005). Chirita et al. (2005) select some factors introduced in Rashid et al. (2005), which may be useful in analyzing patterns for the fake profiles introduced by various shilling attack types, including Number of Prediction-Differences (NPD), Standard Deviation in User's Ratings, Degree of Agreement with Other Users and Degree of Similarity with Top Neighbors. In addition, Chirita et al. (2005) propose two new attributes to recognize attack profiles, namely, Rating Deviation from Mean Agreement (RDMA) and Degree of Similarity with Top Neighbors (DegSim). This algorithm can successfully detect random, average and bandwagon attack profiles, but it is unsuccessful when detecting segment attack and Love/Hate attack. Burke et al. (2006a, b) alternatively, consider these metrics as a classification attribute and derive two new attributes based on RDMA (Chirita et al. 2005). These reproduced attributes are weighted deviation from mean agreement (WDMA) and weighted degree of agreement (WDA). Burke et al. (2006b) also propose one more generic attribute called length variance (lengthVar), which measures how much the length of a given user profile varies from the average length in the database, where length is the number of ratings of a user. In addition, Burke et al. (2006a) propose some attribute such as DegSim', which consider co-rate factor when calculating DegSim (Chirita et al. 2005), Filler Mean Target Difference (FMTD), Profile Variance and Target Model Focus attribute (TMF) to detect shilling attacks.

Many studies employed kNN, C4.5, and SVM classifiers in supervised classification to detect attackers using mentioned classification attributes, such as rating deviation from mean agreement (RDMA), weighted degree of agreement (WDA), filler mean target difference (FMTD), and target model focus (TMF) (Burke et al. 2006a, b; Mobasher et al. 2007b; Williams et al. 2007). Then, the useful attributes are extended to three types including generic attributes, model attributes and intra-profile attributes (Gunes et al. 2014; Mobasher et al. 2007b), to detect more attack types. Alternatively, Mobasher et al. (2007b) propose a hybrid attack detection model employing both statistical techniques and classification using model-specific attributes. Bryan et al. (2008) used a variance-adjusted H_v score to find the bogus profiles which are correlated across a subset of items, whose objective is that bogus profiles will have a higher H_v score. Zhou et al. (2014) propose a novel technique for identifying group attack profiles which uses RDMA (Chirita et al. 2005) and an improved metric DegSim' based on DegSim (Chirita et al. 2005), which measures the difference in similarity between genuine profiles and attack profiles better. Without operating in batch mode, Zhang and Zhou (2014) present an online method called HHT–SVM to detect profile injection attacks by combining Hilbert–Huang transform (HHT) and support vector machine (SVM), whose main idea is the feature extraction method based on user profiles and can operate incrementally. Firstly, they construct rating series for each user profile based on the novelty and popularity of items. They then use the empirical mode decomposition (EMD) approach to decompose each rating series and extract Hilbert spectrum based features to characterize the profile injection attacks by introducing HHT. Finally, they exploit SVM to detect profile injection attacks based on the proposed features.

In order to deal with imbalanced classification problem, Yang et al. (2016b) propose three new features which focus on the number of specific ratings (such as the maximum score, minimum score or average score) on filler or selected items to distinguish attack profiles from all user profiles. The features are Filler Size with Maximum Rating in Itself (FSMAXRI), Filler Size with Minimum Rating in Itself (FSMINRI) and Filler Size with Average Rating in Itself (FSARI). What's more, Yang et al. (2016b) extract well-designed features from user profiles based on the statistical properties of the diverse attack models, and propose a variant of AdaBoost called the rescale AdaBoost (RAdaBoost) based on extracted features to detect shilling attacks, which can deal with imbalanced classification problem well. Considering the mentioned supervised approaches cannot update incrementally with the increase of user profiles, Yu (2014) propose an attack detection algorithm based on incremental learning to solve this conundrum by introducing the rough set theory. Firstly, an algorithm for building the most informative training sets is used to train decision rules by choosing the best label samples. Secondly, an incremental update scheme is used to efficiently re-train decision rules in order to make it more reasonable to cover attack profiles. Finally, statistical features based attack detection algorithm is used to distinguish attack profiles from genuine user profiles.

4.2.2 Unsupervised clustering

Clustering approach is first used to detect shilling attack by O'Mahony et al. (2003). They modify the clustering approach which was used in reputation reporting systems, and utilize it to detect malicious profiles whose aim is to nuke targeted items' popularity. Many unsupervised algorithms have been used in attack detection such as k-means clustering approach (Bhaumik et al. 2011) and principal component analysis (PCA)-based variable selection clustering method (Cheng and Hurley 2009b; Hurley et al. 2009; Mehta 2007; Mehta and Nejdl 2009), which has been proven to be preferable to probabilistic latent semantic analysis (PLSA)-based clustering method (Mehta and Nejdl 2009).

An early unsupervised shilling attack detector is PCASelectUsers (Mehta and Nejdl 2009), which employs the principal component analysis on the original rating data. Mehta (2007) proves that clustering based on Principal Component Analysis (PCA) performs very well against standard attacks when evaluated on MovieLens dataset. The motivation behind this approach is that attacks consist of multiple profiles which are highly correlated with each other, as well as having high similarity with a large number of authentic profiles. Hurley et al. (2009) present the Neyman–Pearson statistical detector, with both supervised and unsupervised versions. Lee and Zhu (2012) propose a new detector using a clustering method and the Group RDMA (GRDMA) metric. Yang et al. (2016a) propose a novel Beta-Protection (βP) based on Beta distribution to detect and exclude attackers, which is immune to missing values and does not need prior knowledge of rating distribution on each item. Bhaumik et al. (2011) detect attack profiles based on several classification attributes and accordingly produce user profiles relying on those attributes. They then apply *k*-means clustering on produced profiles to locate relatively small clusters as suspicious user groups. Similarly, Bilge et al. (2014b) propose a novel shilling attack detection method for particularly specific attacks based on bisecting *k*-means clustering approach, which can detect specific attack profiles such as bandwagon, segment and average attack effectively.

Different from the above unsupervised methods, Yang et al. (2016a) propose an unsupervised method based on graph mining to detect attack profiles. They first construct an undirected user–user graph from original user profiles and estimate the similarity between users according to the constructed graph. Then they distinguish the difference between genuine users and mendacious users in order to rule out a part of genuine profiles by utilizing similarity calculated above. Finally, they further filter out the remained genuine profiles by analyzing target items.

4.2.3 Semi-supervised method

Though there are often a small number of labeled users but a large number of unlabeled users in most of the practical recommender systems, little attention has been paid to modeling both labeled and unlabeled user profiles. Wu et al. (2012) present a Hybrid Shilling Attack Detector, or HySAD for short, to detect more complicated shilling attack types. Firstly, the HySAD algorithm collects popular shilling attack detection metrics for the purpose of feature selection via a wrapper called MC-Relief. Then the HySAD algorithm employs the semi-supervised naive Bayes (SNB_{λ}) classifier for the categorization of both labeled and unlabeled user profiles. Wu et al. (2015) propose a semi-supervised hybrid learning model called hPSD to combine both user features and user-product relations for shilling detection and gain high detection rates. Cao et al. (2013) and Wu et al. (2011) propose a new semi-supervised learning based shilling attack detection, or Semi-SAD for short, to take advantage of both labeled and unlabeled user profiles. The Semi-SAD algorithm combines naïve Bayes classifier and EM-λ (Nigam et al. 2000), an augmented Expectation Maximization (EM). Since the labeled data is valuable and quite rare compared to the unlabeled data, the algorithm first trains a naïve Bayes classifier on a small set of labeled user profiles and then utilizes EM to improve the initial naïve Bayes classifier.

4.2.4 Other detection techniques

In addition to mentioned approaches above, researchers also propose some other techniques to detect shilling attacks in CFRSs. O'Mahony et al. (2006a) propose a signal detection approach to discover natural and malicious noise patterns in a recommender system database by estimating and interpreting probability distributions of user profiles. Li and Luo (2011) propose to utilize probabilistic Bayesian network models to test whether a new profile is malicious or authentic. Similarly, Zhang et al. (2006a) propose a probabilistic approach using SVD-based data reduction method, where a compacted model of observed ratings (including real and biased ones) is generated maximizing the log-likelihood of all ratings. Then the degree of belief in a rating can be estimated as log-likelihood of this rating given the compacted model. Tang and Tang (2011) analyze rating time intervals of users to detect suspicious behavior to bias top-N lists in recommender systems. They generate an attribute to measure time behavior of users consisting of span, frequency, and mount properties. As a similar approach, Zhang et al. (2006b) propose to construct a time series of ratings for an item. They use a window to group consecutive ratings for the item, compute the sample average and entropy in each window, and interpret results to detect suspicious behavior. Xia et al. (2015) propose a novel dynamic time interval segmentation technique based item anomaly detection approach to detect shilling attacks, which could confirm the size of the time interval dynamically to group as many consecutive attack ratings together as possible.

Different from the existing algorithms, which focused extensively on rating patterns of attack profiles, Zou and Fekri (2013) develop a probabilistic inference framework that further exploits the target items for attack detection and utilize the Belief Propagation (BP) algorithm (Kschischang et al. 2001) to perform inference efficiently. Instead of focusing on the differences between genuine profiles and attack profiles, Zhou et al. (2016) study the use of SVM based method and group characteristics in attack profiles, and propose a two phase method called SVM-TIA to detect shilling attack profiles. In the first phase, Borderline-SMOTE method is used to alleviate the class unbalance problem because class unbalance problem exists in SVM classifiers. The second phase is a fine-tuning phase, and the target item analysis method is used to reduce false positive rate of the detection result. The same authors (Zhou et al. 2015) also propose another two-phase detection structure called RD-TIA for detecting shilling attacks using a statistical approach. In the first phase, they extract profile attributes which may determine the suspicious profiles by using two statistical metrics, DegSim and RDMA. In the second phase, they use Target Item Analysis (TIA) method to filter out genuine profiles. Chakraborty and Karforma (2013) consider the profiles injected into the system as outliers and detect attack profiles by using outlier analysis. Firstly, they use Partition around Medoid (PAM) clustering algorithm to detect the attack profiles. Secondly, they apply a PAM-based outlier detection algorithm to find these attack profiles in large clusters. Finally, an angle-based outlier detection strategy (Kriegel and Zimek 2008) is used for finding attack profiles in the database under attack.

Various PPCF recommender systems are also vulnerable to such attacks (Bilge et al. 2014a; Gunes et al. 2013a, b). Gunes and Polat (2016) examine the detection of shilling attacks in privacy-preserving collaborative filtering systems. O'Mahony (2004) utilizes four attack detection methods to filter out fake profiles produced by six well-known shilling attacks on perturbed data in his doctoral thesis. Gunes and Polat (2015) propose a hierarchical clustering-based shilling attack detection method in private environments. They scrutinized the ratings of target items to improve the overall performance of their scheme.

4.3 Robust algorithms

There are another ways to reduce the influence of shilling attacks on CFRSs except for performing shilling attack detection. Another alternative way is to develop attack-resistant collaborative filtering algorithms, i.e., robust recommendation algorithms. In this subsection, we will briefly discuss the studies proposing robust approaches against shilling attacks for CFRSs.

In order to reduce the impact of shilling attacks on the recommendation results, Yu et al. (2017) use kernel mapping of the rating matrix and kernel distance to construct a robust kernel matrix factorization model, which can also improve the credibility of user similarity. Then they devise a robust collaborative recommendation algorithm by combining the proposed robust kernel matrix factorization model and neighborhood model. Yi and Zhang

(2016) propose a robust recommendation method based on suspicious users measurement and multi-dimensional trust (RRM-SUMMT). Firstly, they establish the relevance vector machine classifier relying on the user profile features to identify the suspicious users in the user rating database; then they mine the implicit trust relation among users based on the user-item rating data, and construct a reliable multidimensional trust model by integrating the user suspicion information; finally they combine the reliable multidimensional trust model, the neighbor model and matrix factorization model to devise a robust recommendation algorithm. Mehta and Nejdl (2008) describe a new collaborative algorithm based on SVD which is accurate as well as highly stable to shilling. Zhang et al. (2017) propose a robust collaborative filtering method based on non-negative matrix factorization (NMF) and R1-norm. What's more, Zhang and Sun (2014) also propose a robust collaborative recommendation algorithm called LMedSMF based on least median squares estimator. Mehta et al. (2007) propose Robust Matrix Factorization (RMF) based on M-estimators algorithm as a resistant scheme against shilling attacks. As least squares is known to be sensitive to outliers, Cheng and Hurley (2010) propose a least trimmed squares based MF (LTSMF) to help improve the robustness of the least squares based MF (LSMF) models. Least trimmed squares is shown to be more robust than least squares and another popular robust method M-estimator (Mehta et al. 2007). Bilge et al. (2014a) investigate robustness of four well-known privacy-preserving model-based recommendation methods against six shilling attack types. Their empirical analysis show that DWT-based PPCF (Bilge and Polat 2012) and k-means clustering-based PPCF (Bilge and Polat 2013) algorithms can be vulnerable to shilling attack manipulations. While SVD-based PPCF and item-based PPCF (Polat and Du 2005) are, on the other hand, quite robust against the applied attack models. In addition to the abovementioned robust schemes, Mobasher et al. (2006a) suggest applying PLSA to CFRSs for robustness. The authors indicate that PLSA is a very robust CF algorithm and it is stable in the face of shilling attacks (Mobasher et al. 2006a).

Trust-aware recommender systems are more robust prediction schemes than traditional ones against shilling attacks (Lacoste-Julien et al. 2013; Ray and Mahanti 2010). For the defense of shilling attacks, there has been an increasing focus on the researches incorporating trust models into recommender systems in recent years, which can relieve some traditional limitations of collaborative filtering which including shilling attacks (O'Donovan and Smyth 2005, 2006). Zhang and Xu (2007) propose to utilize topic-level trust-based prediction algorithm. First, compute degree of trust for a user based on his/her neighbors; then, the trust is described as an item-level trust; topic-level trust then can be estimated by calculating the average score of item-level trust for a producer on the items belonging to the same topic and have been rated by the producer; finally, trust is incorporated into traditional CF scheme for producing predictions. To improve the robustness of item-based CF, Gao et al. (2014a) propose a novel CF approach by analyzing the three most commonly used relationships between users, such as rating similarity, interest similarity and linear dependence between users. Similarly, Jia et al. (2013) and Jia and Zhang (2014) present a robust collaborative filtering recommendation algorithm based on multidimensional trust model, which measures the credibility of user's ratings from the following three aspects: the reliability of item recommendation, the rating similarity and the user's trustworthiness.

5 Evaluation metrics

As discussed previously, some researchers focus on detecting shilling attacks and propose some detection algorithms. Some researchers study how to enhance the robustness of CF schemes or they develop robust CF schemes against shilling attacks. Another group of researchers evaluate the effects of shilling attacks on recommender systems' accuracy. O'Mahony et al. (2004) propose two measures to evaluate the effectiveness of CFRSs which are robustness and stability. Robustness measures the performance of the system before and after an attack to determine how the attack affects the system as a whole; stability looks at the shift in system's ratings for the attacked item induced by the attack profiles.

Since different studies focus on different aspects of shilling attacks, there are various measures are used for evaluating the proposed CF schemes. We group such metrics and listed the most widely used ones in Table 9.

For the sake of evaluating the effects of shilling attacks on recommender systems' accuracy, various metrics are used. The prediction shift (Burke et al. 2005a; Ji et al. 2007) is the average change in the predicted rating for the attacked item before and after the attack. Average precision shift (Mobasher et al. 2006b) is the average change in the predicted rating for all of items before and after the attack. Absolute prediction shift (Ji et al. 2007) measures the distortion of prediction occurring due to an attack. Hit ratio (Bhaumik et al. 2007a, b; Mehta and Nejdl 2008) measures the effect of attack profiles on top-N recommendations, which is the ratio of the number of hits across all users to the number of users in the test set. Likewise, average hit ratio (Mobasher et al. 2006b) can then calculated as the sum of the hit ratio for each item *i* following an attack on *i* across all items divided by the number of items. High rating ratio (Cheng and Hurley 2009a, b, c) shows how much predictions are pushed to high values for an attacked item. ExptopN (expected top-N occupancy) (Lam and Riedl 2004; Zhang et al. 2006a) is defined as the expected number of occurrences of all target items in a top-N recommendation list measured over all users.

In order to evaluate how effective the shilling attack detection schemes are, various metrics have been proposed. The precision (Bhaumik et al. 2006, 2011; Mehta and Nejdl 2009) is quantified by the percentage of truly detected attack profiles divided by all profiles which are classified as attack profiles. The recall (Bhaumik et al. 2006, 2011; Mehta and Nejdl 2009) is quantified by the percentage of truly detected attack profiles divided by all attack profiles. The F1-measure (Bhaumik et al. 2006; Burke et al. 2006a, b; Mobasher et al. 2007b) integrate the precision and recall rate together. Correctness (He et al. 2010) or classification accuracy is the ratio of the number of profiles correctly classified as fake profiles over the number of all profiles. The detection rate or classification accuracy is defined as the number of detected attack profiles divided by the number of attack profiles (Gao et al. 2014b; Hurley

Table 9 Accuracy metrics

	Evaluation metrics
Assessing shilling effects	Precision shift; average prediction shift; absolute prediction shift; hit ratio; average hit ratio; high rating ratio; ExptopN;
Evaluating detection methods	Precision; recall; F1-measure; correctness; detection rate; false alarm rate; ROC; specificity; NPV;
Evaluating robust algorithm	MAE; NMAE; RMSE; Coverage; MAS; RMSS;

et al. 2009; Zhang et al. 2006b). The false alarm rate (Bhaumik et al. 2011; Gao et al. 2014b; Hurley et al. 2009; Zhang et al. 2006b) measures the percentage of genuine profiles who are detected as attack profiles. The ROC curve (Zhang et al. 2006b) measures the extent to which the detection algorithm can successfully distinguish between attack profiles and genuine profiles. Specificity (Burke et al. 2015) presents the count of authentic profiles correctly classified as a fraction of the total number of authentic profiles in the system. NPV (Negative predictive value) (Burke et al. 2015) presents the count as a fraction of the total number of profiles labelled Authentic.

$$\text{Specificity} = \#true\ negatives/(\#true\ negatives\ +\ \#false\ positives)$$
$$\text{NPV} = \#true\ negatives/(\#true\ negatives\ +\ \#false\ negatives)$$

Where, the true positives and true negatives consist of profiles that were correctly classified as attack or authentic, respectively; the false positives and false negatives—consist of profiles that were incorrectly classified as attack or authentic, respectively.

To evaluate the robust CF algorithms with respect to accuracy, mean absolute error (MAE), root mean squared error (RMSE) and so on are utilized. MAE measures how close the estimated predictions to their observed ones and can be used for capturing the deviation from actual values (Mehta et al. 2007). Normalized mean absolute error (NMAE) is used to facilitate a comparison between the robustness of CFRSs operating on different rating scales and it can be computed by dividing the MAE to the difference between the maximum and the minimum ratings (Gunes et al. 2014). The root mean squared error (RMSE) can also be used to measure the accuracy of CF algorithms (Cheng and Hurley 2010). Coverage (O'Mahony et al. 2006a) is the ratio of the number of predictions that an algorithm can estimate to the total number of predictions that are requested. Xia et al. (2015) measure the stability of CFRSs by computing the mean absolute difference (MAS, mean absolute shift) or root mean squared difference (RMSS, root mean squared shift).

6 Discussion and future directions

With the development and popularity of the Internet, traditional commerce has been increasingly replaced by electronic commerce. As long as customers trade over the Internet via online vendors, the information overload problem is still exists. Then, to overcome the information overload problem, recommender systems will be widely used by many online vendors. Also, shilling attacks are still serious threats to the collaborative filtering recommender systems. Although researchers have proposed a lot of robust recommender algorithms and attack detection schemes, these attack detection methods are still deficient.

The existing attack detection schemes are designed based on one or several specific type of shilling attack models. Although the proposed detection schemes could be applicable to specific type of attack models, they may not be effective to all types of attack models. What's more, we cannot know in advance whether the recommender systems are attacked or not, and the type of attack model, in the real world. Thus, the attack detection schemes should consider the similarities of all attack models exactly, rather than simply consider certain types of attack models. Also, some attackers inject bogus ratings is just to interfere with the recommendation algorithm with the goal of causing users to stop trusting the recommender system and eventually to stop using it, instead of recommending a merchandise more likely or less likely. However, the existing attack detection algorithms could not detect these attackers effectively, which is called random vandalism in this paper.

There are several research directions worth discussing, which could potentially improve the robustness of recommender systems and the detection rates of shilling detection schemes.

(1) *Using trust relationships between users*
Certain studies have also indicated that a user is much more likely to believe a trusted user's rather than a stranger's statements. Because a trusted user will also trust his friend's opinions, in a recursive manner, trust may propagate through the relationship network (Guha et al. 2004; Guo et al. 2015b; Ziegler and Golbeck 2015). Trust-aware recommender systems have been introduced as an effective approach to overcome the data sparsity and cold-start problems (Guo et al. 2014, 2015a; Massa and Avesani 2007; Moradi et al. 2015; Shambour and Lu 2015). Researchers have also begun to improve the robustness of recommender systems by utilizing trust relationships between users (Gao et al. 2014a; Jia et al. 2013; Jia and Zhang 2014). Thus, it is likely to be effective to improve the detection rates of various detection schemas by combining user trust model.

(2) *Utilize distrust information*
Since users' trust information could help to improve the robustness of recommendation systems and overcome the data sparsity and cold-start problems in recommender systems, users' distrust information may be helpful too. Lee and Ma (2016) present a hybrid approach that combines user ratings, user trust information and user distrust information for making better recommendations. Their experimental results indicate that the distrust information is beneficial in rating prediction. The distrust and dislike information between users may also be helpful to improve the detection rates of detection algorithms.

(3) *Cross media data*
We are facing an era of online to offline (O2O)—almost everyone is using online social networks to connect friends or, more generally, to satisfy social needs at different levels. In fact, many users participate in more than one social network, such as public networks and private networks, as well as business networks and family networks. A user who has an account in Epinions might also have an account in Twitter. A new user on one website might have existed on another website for a long time. For example, a user has already specified her/his interests in Epinions and has also written many reviews about items. When the user registers at twitter for the first time as a cold-start user, data about the user in Epinions can help twitter to learn the user's preference and friendships, which can produce more accurate recommendations. Many researchers have studied the integration of cross-media data (Kong et al. 2013; Lacoste-Julien et al. 2013; Zafarani and Liu 2009; Zhang et al. 2015). Integrating cross-media data may be helpful to establish robust or anti-attack recommender systems.

7 Conclusion

As one of the most effective ways to deal with the information overload problem, collaborative filtering recommender systems (CFRSs) have been proved vulnerable to various profile injection attacks where a number of fake user profiles are inserted into the system to influence the recommendations made to the users. It is necessary to keep an eye on the current situation of shilling attacks and propose more robust recommendation algorithm or develop more accurate attack detection schemas. However, there is no comprehensive survey discussing shilling attacks, detection schemes, robust CF algorithms, and their evaluation in detail for the last 5 years. We present a review about shilling attacks against CF algorithms.

In this survey, we first briefly discussed the related survey papers about shilling attacks. Secondly, we covered possible various shilling attacks and explained their deployment modes. Thirdly, we discussed the related research about shilling attacks covering the studies focusing on shilling attack strategies, shilling detection schemes and robust recommendation algorithms. Fourthly, we discussed evaluation metrics utilized to assess the effects of shilling attacks, the accuracy of attack detection, and the performance of robust recommendation algorithms. We finally analyzed the shortcomings of existing attack detection schemes, and discussed some future research directions which could be help to improve shilling attack detection rates and robustness of collaborative recommendation.

Our survey has summarized various shilling attacks and has had a detailed description of their deployment strategies. What's more, we have conducted detailed separate surveys about shilling attacks strategies, attack detection schemes, and robustness analysis and robust recommendation algorithms, respectively. Moreover, we analyzed the shortcomings of existing detection schemes and gave some proposals to improve shilling attack detection rates and robustness of collaborative recommendation, such as by crossing media data to raise the robustness of CFRSs, by taking advantage of both users' trust relationships and distrust information to strengthen the discrimination of both genuine users and bogus users. Also, the research directions proposed may provide reference ideas for researchers, who research shilling attacks against CFRSs.

Acknowledgements This work is supported by the Projects (61672401, 61373045) supported by the National Natural Science Foundation of China; Project (315***10101) supported by the Pre-Research Project of the "Thirteenth Five-Year-Plan" of China; Projects (JBG161002) supported by the Fundamental Research Funds for the Central Universities of China.

References

Abbas A, Bilal K, Zhang L, Khan SU (2015) A cloud based health insurance plan recommendation system: a user centered approach. Futur Gener Comput Syst 43:99–109

Adomavicius G, Tuzhilin A (2005) 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

Agarwal V, Bharadwaj KK (2013) A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. Soc Netw Anal Min 3(3):359–379

Barragáns-Martínez AB, Costa-Montenegro E, Burguillo JC, Rey-López M, Mikic-Fonte FA, Peleteiro A (2010) A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. Inf Sci 180(22):4290–4311

Bhaumik R, Williams CA, Mobasher B, Burke RD (2006) Securing collaborative filtering against malicious attacks through anomaly detection. In: Proceedings of the 4th workshop on intelligent techniques for web personalization, Boston, MA

Bhaumik R, Burke RD, Mobasher B (2007a) Effectiveness of crawling attacks against web-based recommender systems. In: AAAI workshop on intelligent techniques for web personalization, Vancouver, BC, Canada, pp 17–26

Bhaumik R, Burke RD, Mobasher B (2007b) Crawling attacks against web-based recommender systems. In: International conference on data mining, Las Vegas, Nevada, USA, pp 183–189

Bhaumik R, Mobasher B, Burke R (2011) A clustering approach to unsupervised attack detection in collaborative recommender systems. In: 7th IEEE international conference on data mining, Las Vegas, NV, USA, pp 181–187

Bilge A, Polat H (2012) A improved privacy-preserving DWT-based collaborative filtering scheme. Expert Syst Appl 39(3):3841–3854

Bilge A, Polat H (2013) A comparison of clustering-based privacy-preserving collaborative filtering schemes. Appl Soft Comput 13(5):2478–2489

Bilge A, Gunes I, Polat H (2014a) Robustness analysis of privacy-preserving model-based recommendation schemes. Expert Syst Appl 41(8):3671–3681

Bilge A, Ozdemir Z, Polat H (2014b) A novel shilling attack detection method. Procedia Comput Sci 31:165–174

Bobadilla J, Ortega F, Hernando A, Gutiérrez A (2013) Recommender systems survey. Knowl Based Syst 46:109–132

Borràs J, Moreno A, Valls A (2014) Intelligent tourism recommender systems: a survey. Expert Syst Appl 41(16):7370–7389

Bryan K, O'Mahony MP, Cunningham P (2008) Unsupervised retrieval of attack profiles in collaborative recommender systems. In: ACM conference on recommender systems, Lausanne, Switzerland, pp 155–162

Burke RD, Mobasher B, Bhaumik R, Williams CA (2005a) Segment-based injection attacks against collaborative filtering recommender systems. In: IEEE international conference on data mining, Houston, TX, USA, pp 577–580

Burke RD, Mobasher B, Bhaumik R, Williams CA (2005b) Collaborative recommendation vulnerability to focused bias injection attacks. In: International conference on data mining: workshop on privacy and security aspects of data mining (ICDM), Houston, TX, USA, pp 35–43

Burke RD, Mobasher B, Zabicki R, Bhaumik R (2005c) Identifying attack models for secure recommendation. In: Proceedings of the WebKDD: workshop on the next generation of recommender systems research, San Diego, CA, USA, pp 19–25

Burke RD, Mobasher B, Bhaumik R (2005d) Limited knowledge shilling attacks in collaborative filtering systems. In: International joint conference on artificial intelligence (IJCAI 2005): workshop on intelligent techniques for web personalization (ITWP 2005), pp 17–24

Burke RD, Mobasher B, Williams CA, Bhaumik R (2006a) Classification features for attack detection in collaborative recommender systems. In: 12th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA, pp 542–547

Burke RD, Mobasher B, Williams CA, Bhaumik R (2006b) Detecting profile injection attacks in collaborative recommender systems. In: 8th IEEE conference on e-commerce technology, San Francisco, CA, USA, pp 23–30

Burke R, O'Mahony MP, Hurley NJ (2015) Robust collaborative recommendation. In: Recommender systems handbook. Springer, Boston, pp 961–995

Cao J, Wu Z, Mao B, Zhang Y (2013) Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system. World Wide Web 16(5–6):729–748

Casino F, Patsakis C, Puig D, Solanas A (2013) On privacy preserving collaborative filtering: current trends, open problems, and new issues. In: 2013 IEEE 10th international conference on e-business engineering (ICEBE). IEEE, pp 244–249

Casino F, Domingo-Ferrer J, Patsakis C, Puig D, Solanas A (2015) A k-anonymous approach to privacy preserving collaborative filtering. J Comput Syst Sci 81(6):1000–1011

Cechinel C, Sicilia MÁ, SáNchez-Alonso S, GarcíA-Barriocanal E (2013) Evaluating collaborative filtering recommendations inside large learning object repositories. Inf Process Manag 49(1):34–50

Chakraborty P, Karforma S (2013) Detection of profile-injection attacks in recommender systems using outlier analysis. Procedia Technol 10:963–969

Chen W, Niu Z, Zhao X, Li Y (2014) A hybrid recommendation algorithm adapted in e-learning environments. World Wide Web 17(2):271–284

Cheng Z, Hurley NJ (2009a) Robustness analysis of model-based collaborative filtering systems. Lect Notes Comput Sci 6206:3–15

Cheng Z, Hurley NJ (2009b) Effective diverse and obfuscated attacks on model-based recommender systems. In: 3rd ACM international conference on recommender systems, New York, NY, USA, pp 141–148

Cheng Z, Hurley NJ (2009c) Trading robustness for privacy in decentralized recommender systems. In: 31st conference on innovative applications of artificial intelligence, Pasadena, CA, USA, pp 79–84

Cheng Z, Hurley NJ (2010) Robust collaborative recommendation by least trimmed squares matrix factorization. In: 22nd IEEE international conference on tools with artificial intelligence, Arras, France, pp 105–112

Chirita PA, Nejdl W, Zamfir C (2005) Preventing shilling attacks in online recommender systems. In: 7th Annual ACM international workshop on web information and data management, Bremen, Germany, pp 67–74

Cho YH, Kim JK (2004) Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. Expert Syst Appl 26(2):233–246

Chung CY, Hsu PY, Huang SH (2013) βP: a novel approach to filter out malicious rating profiles from recommender systems. Decis Support Syst 55(1):314–325

Cobos C, Rodriguez O, Rivera J, Betancourt J, Mendoza M, LeóN E, Herrera-Viedma E (2013) A hybrid system of pedagogical pattern recommendations based on singular value decomposition and variable data attributes. Inf Process Manag 49(3):607–625

Dellarocas C (2000) Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In: 2nd ACM conference on electronic commerce, Minneapolis, MN, USA, pp 150–157

Elbadrawy A, Karypis G (2016) Domain-aware grade prediction and top-n course recommendation. In: 10th ACM conference on recommender systems, pp 183–190

Gao M, Liu K, Wu Z (2010) Personalisation in web computing and informatics: theories, techniques, applications, and future research. Inf Syst Front 12(5):607–629

Gao M, Ling B, Yuan Q, Xiong Q, Yang L (2014a) A robust collaborative filtering approach based on user relationships for recommendation systems. Math Probl Eng 2014:1–8

Gao M, Yuan Q, Ling B, Xiong Q (2014b) Detection of abnormal item based on time intervals for recommender systems. Sci World J 2014:845–897

Gavalas D, Konstantopoulos C, Mastakas K, Pantziou G (2014) Mobile recommender systems in tourism. J Netw Comput Appl 39:319–333

Goldberg D, Nichols D, Oki BM (1992) Using collaborative filtering to weave an information tapestry. Commun ACM 35(12):61–70

Good N, Schafer JB, Konstan JA, Borchers A, Sarwar B, Herlocker J, Riedl J (1999) Combining collaborative filtering with personal agents for better recommendations. In: AAAI/IAAI, pp 439–446

Goyal A, Lakshmanan LV (2012) Recmax: exploiting recommender systems for fun and profit. In: 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1294–1302

Guha R, Kumar R, Raghavan P, Tomkins A (2004) Propagation of trust and distrust. In: 13th International conference on World Wide Web, pp 403–412

Gunes I, Polat H (2015) Hierarchical clustering-based shilling attack detection in private environments. In: 3rd International symposium on digital forensics and security, pp 1–7

Gunes I, Polat H (2016) Detecting shilling attacks in private environments. Inf Retr J 19(6):547–572

Gunes I, Bilge A, Kaleli C, Polat H (2013a) Shilling attacks against privacy-preserving collaborative filtering. J Adv Manag Sci 1(1):54–60

Gunes I, Bilge A, Polat H (2013b) Shilling attacks against memory-based privacy-preserving recommendation algorithms. KSII Trans Internet Inf Syst (TIIS) 7(5):1272–1290

Gunes I, Kaleli C, Bilge A, Polat H (2014) Shilling attacks against recommender systems: a comprehensive survey. Artif Intell Rev 42:767–799

Guo G, Zhang J, Thalmann D (2014) Merging trust in collaborative filtering to alleviate data sparsity and cold start. Knowl Based Syst 57:57–68

Guo G, Zhang J, Yorke-Smith N (2015a) Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems. Knowl Based Syst 74:14–27

Guo G, Zhang J, Yorke-Smith N (2015b) TrustSVD: collaborative filtering with both the explicit and implicit influence of user trust and of item ratings. In: 29th AAAI conference on artificial intelligence, pp 123–129

He F, Wang X, Liu B (2010) Attack detection by rough set theory in recommendation system. In: IEEE international conference on granular computing, pp 692–695

Herlocker JL, Konstan JA, Borchers A, Riedl J (1999) An algorithmic framework for performing collaborative filtering. In 22nd Annual international ACM SIGIR conference on research and development in information retrieval, pp 230–237

Hurley NJ, O'Mahony MP, Silvestre GCM (2007) Attacking recommender systems: a cost-benefit analysis. IEEE Intell Syst 22(3):64–68

Hurley NJ, Cheng Z, Zhang M (2009) Statistical attack detection. In: 3rd ACM international conference on recommender systems, New York, NY, USA, pp 149–156

Jeckmans AJ, Beye M, Erkin Z, Hartel P, Lagendijk RL, Tang Q (2013) Privacy in recommender systems. In: Social media retrieval, pp 263–281

Ji AT, Yeon C, Kim HN, Jo GS (2007) Distributed collaborative filtering for robust recommendations against shilling attacks. Lect Notes Comput Sci 4509:14–25

Jia CX, Liu RR (2015) Improve the algorithmic performance of collaborative filtering by using the interevent time distribution of human behaviors. Phys A Stat Mech Appl 436:236–245

Jia D, Zhang F (2014) A robust collaborative recommendation algorithm incorporating trustworthy neighborhood model. J Comput 9(10):2329

Jia D, Zhang F, Liu S (2013) A robust collaborative filtering recommendation algorithm based on multidimensional trust model. JSW 8(1):11–18

Kong X, Zhang J, Yu PS (2013) Inferring anchor links across multiple heterogeneous social networks. In: 22nd ACM international conference on information and knowledge management, pp 179–188

Koren Y (2008) Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 426–434

Koren Y (2010) Collaborative filtering with temporal dynamics. Commun ACM 53(4):89–97

Kriegel HP, Zimek A (2008) Angle-based outlier detection in high-dimensional data. In: 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 444–452

Kschischang FR, Frey BJ, Loeliger HA (2001) Factor graphs and the sum-product algorithm. IEEE Trans Inf Theory 47(2):498–519

Lacoste-Julien S, Palla K, Davies A, Kasneci G, Graepel T, Ghahramani Z (2013) Sigma: simple greedy matching for aligning large knowledge bases. In: 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp 572–580

Lam SK, Riedl JT (2004) Shilling recommender systems for fun and profit. In: 13th International conference on World Wide Web, New York, NY, USA, pp 393–402

Lam SK, Riedl JT (2005) Privacy, shilling, and the value of information in recommender systems. In: Proceedings of user modeling workshop on privacy-enhanced personalization, Edinburgh, UK, pp 85–92

Lam SK, Frankowski D, Riedl JT (2006) Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. Lect Notes Comput Sci 3995:14–29

Lee WP, Ma CY (2016) Enhancing collaborative recommendation performance by combining user preference and trust-distrust propagation in social networks. Knowl Based Syst 106:125–134

Lee JS, Zhu D (2012) Shilling attack detection: a new approach for a trustworthy recommender system. INFORMS J Comput 24(1):117–131

Li C, Luo Z (2011) Detection of shilling attacks in collaborative filtering recommender systems. In: Proceedings of the international conference of soft computing and pattern recognition, Dalian, China, pp 190–193

Massa P, Avesani P (2007) Trust-aware recommender systems. In: Proceedings of the 2007 ACM conference on recommender systems, pp 17–24

Mehta B (2007) Unsupervised shilling detection for collaborative filtering. In: 22nd International conference on artificial intelligence, Vancouver, BC, Canada, pp 1402–1407

Mehta B, Hofmann T (2008) A survey of attack-resistant collaborative filtering algorithms. EEE Data Eng Bull 31(2):14–22

Mehta B, Nejdl W (2008) Attack resistant collaborative filtering. In: 31st Annual international ACM SIGIR conference on research and development in information retrieval, Singapore, pp 75–82

Mehta B, Nejdl W (2009) Unsupervised strategies for shilling detection and robust collaborative filtering. User Model User Adapt Interact 19(1–2):65–97

Mehta B, Hofmann T, Nejdl W (2007) Robust collaborative filtering. In: 1st ACM international conference on recommender systems, Minneapolis, MN, USA, pp 49–56

Mobasher B, Burke RD, Bhaumik R, Williams CA (2005) Effective attack models for shilling item-based collaborative filtering systems. In: Proceedings of the 2005 WebKDD workshop, Chicago, IL, USA

Mobasher B, Burke RD, Sandvig JJ (2006a) Model-based collaborative filtering as a defense against profile injection attacks. In: 21st National conference on artificial intelligence, Boston, MA, USA, pp 1388–1393

Mobasher B, Burke RD, Williams CA, Bhaumik R (2006b) Analysis and detection of segment-focused attacks against collaborative recommendation. Lect Notes Comput Sci 4198:96–118

Mobasher B, Burke RD, Bhaumik R, Sandvig JJ (2007a) Attacks and remedies in collaborative recommendation. IEEE Intell Syst 22(3):56–63

Mobasher B, Burke RD, Bhaumik R, Williams C (2007b) Toward trustworthy recommender systems: an analysis of attack models and algorithm robustness. ACM Trans Internet Technol (TOIT) 7(4):23–60

Montaner M, López B, De La Rosa JL (2003) A taxonomy of recommender agents on the internet. Artif Intell Rev 19(4):285–330

Moradi P, Ahmadian S, Akhlaghian F (2015) An effective trust-based recommendation method using a novel graph clustering algorithm. Phys A Stat Mech Appl 436:462–481

Nigam K, McCallum AK, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39(2):103–134

O'Donovan J, Smyth B (2005) Trust in recommender systems. In: 10th International conference on intelligent user interfaces, pp 167–174

O'Donovan J, Smyth B (2006) Mining trust values from recommendation errors. Int J Artif Intell Tools 15(06):945–962

O'Mahony MP (2004) Towards robust and efficient automated collaborative filtering. Ph.D. dissertation, University College Dublin

O'Mahony MP, Smyth B (2007a) Evaluating the robustness of collaborative web search. In: 18th Irish conference on artificial intelligence and cognitive science, Dublin, Ireland

O'Mahony MP, Smyth B (2007b) Collaborative web search: a robustness analysis. Artif Intell Rev 28(1):69–86

O'Mahony MP, Hurley NJ, Silvestre GCM (2002a) Towards robust collaborative filtering. Lect Notes Comput Sci 2464:87–94

O'Mahony MP, Hurley NJ, Silvestre GCM (2002b) Promoting recommendations: an attack on collaborative filtering. In: 13th International conference on database and expert systems applications, Aix-en-Provence, France, pp 494–503

O'Mahony MP, Hurley NJ, Silvestre GCM (2003) Collaborative filtering-safe and sound. Lect Notes Comput Sci 2871:506–510

O'Mahony MP, Hurley NJ, Kushmerick N, Silvestre GCM (2004) Collaborative recommendation: a robustness analysis. ACM Trans Internet Technol 4(4):344–377

O'Mahony MP, Hurley NJ, Silvestre GCM (2005) Recommender systems: attack types and strategies. In: 20th National conference on artificial intelligence, Pittsburgh, PA, USA, pp 334–339

O'Mahony MP, Hurley NJ, Silvestre GCM (2006a) Detecting noise in recommender system databases. In: 11th International conference on intelligent user interfaces, Sydney, Australia, pp 109–115

O'Mahony MP, Hurley NJ, Silvestre GCM (2006b) Attacking recommender systems: the cost of promotion. In: Proceedings of the workshop on recommender systems, in conjunction with the 17th European conference on artificial intelligence, Riva del Garda, Trentino, Italy, pp 24–28

Ozturk A, Polat H (2015) From existing trends to future trends in privacy-preserving collaborative filtering. Wiley Interdiscipl Rev Data Min Knowl Discov 5(6):276–291

Polat H, Du W (2005) Privacy-preserving collaborative filtering. Int J Electron Commerc 9(4):9–35

Rashid AM, Karypis G, Riedl J (2005) Influence in ratings-based recommender systems: an algorithm-independent approach. In: 2005 SIAM international conference on data mining, pp 556–560

Ray S, Mahanti A (2009) Filler item strategies for shilling attacks against recommender systems. In: 42nd Hawaii international conference on system sciences (HICSS 09), pp 1–10

Ray S, Mahanti A (2010) Improving prediction accuracy in trust-aware recommender systems. In: 43rd Hawaii international conference on system sciences, Kauai, HI, USA, pp 1–9

Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) GroupLens: an open architecture for collaborative filtering of netnews. In: ACM conference on computer supported cooperative work, pp 175–186

Rivero-Rodriguez A, Konstantinidis ST, Sánchez-Bocanegra CL, Fernández-Luque L (2013) A health information recommender system: enriching YouTube health videos with Medline Plus information by the use of SnomedCT terms. In: 26th International symposium on computer-based medical systems (CBMS), pp 257–261

Ronen R, Koenigstein N, Ziklik E, Nice N (2013) Selecting content-based features for collaborative filtering recommenders. In: 7th ACM conference on recommender systems, pp 407–410

Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. In: 10th International conference on World Wide Web, pp 285–295

Schafer JB, Konstan JA, Riedl J (2001) E-commerce recommendation applications. In: Applications of data mining to electronic commerce, pp 115–153

Seminario CE (2013) Accuracy and robustness impacts of power user attacks on collaborative recommender systems. In: 7th ACM conference on recommender systems, pp 447–450

Seminario CE, Wilson DC (2014a) Assessing impacts of a power user attack on a matrix factorization collaborative recommender system. In: 27th International Florida artificial intelligence research society conference, pp 81–86

Seminario CE, Wilson DC (2014b) Attacking item-based recommender systems with power items. In: 8th ACM conference on recommender systems, pp 57–64

Seminario CE, Wilson DC (2016) Nuking item-based collaborative recommenders with power items and multiple targets. In: International Florida artificial intelligence research society conference, pp 560–565

Serrano-Guerrero J, Herrera-Viedma E, Olivas JA, Cerezo A, Romero FP (2011) A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0. Inf Sci 181(9):1503–1516

Shambour Q, Lu J (2015) An effective recommender system by unifying user and item trust information for B2B applications. J Comput Syst Sci 81(7):1110–1126

Shi C, Liu J, Zhuang F, Philip SY, Wu B (2016) Integrating heterogeneous information via flexible regularization framework for recommendation. Knowl Inf Syst 49(3):835–859

Silva T, Ma J, Yang C, Liang H (2015) A profile-boosted research analytics framework to recommend journals for manuscripts. J Assoc Inf Sci Technol 66(1):180–200

Tang T, Tang Y (2011) An effective recommender attack detection method based on time SFM factors. In: 3rd International conference on communication software and networks, Xi'an, China, pp 78–81

Tejeda-Lorente Á, Porcel C, Peis E, Sanz R, Herrera-Viedma E (2014) A quality based recommender system to disseminate information in a university digital library. Inf Sci 261:52–69

Vozalis MG, Margaritis KG (2007) Using SVD and demographic data for the enhancement of generalized collaborative filtering. Inf Sci 177(15):3017–3037

Wang J, Zhang Y (2013) Opportunity model for e-commerce recommendation: right product; right time. In: 36th International ACM SIGIR conference on research and development in information retrieval, pp 303–312

Wang Z, Sun L, Zhu W, Yang S, Li H, Wu D (2013) Joint social and content recommendation for user-generated videos in online social network. IEEE Trans Multimedia 15(3):698–709

Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge

Wei C, Khoury R, Fong S (2013) Web 2.0 Recommendation service by multi-collaborative filtering trust network algorithm. Inf Syst Front 15(4):533–551

Wiesner M, Pfeifer D (2014) Health recommender systems: concepts, requirements, technical basics and challenges. Int J Environ Res Public Health 11(3):2580–2607

Williams CA, Mobasher B (2012) Thesis: Profile injection attack detection for securing collaborative recommender systems. Serv Oriented Comput Appl 1(3):157–170

Williams CA, Mobasher B, Burke RD, Bhaumik R, Sandvig JJ (2006) Detection of obfuscated attacks in collaborative recommender systems. In: Proceedings of the workshop on recommender systems, in conjunction with the 17th European conference on artificial intelligence, Riva del Garda, Trentino, Italy, pp 19–23

Williams CA, Mobasher B, Burke RD (2007) Defending recommender systems: detection of profile injection attacks. Serv Oriented Comput Appl 1(3):157–170

Wilson DC, Seminario CE (2013) When power users attack: assessing impacts in collaborative recommender systems. In: 7th ACM conference on recommender systems, pp 427–430

Wu Z, Cao J, Mao B, Wang Y (2011) Semi-SAD: applying semi-supervised learning to shilling attack detection. In: 5th ACM conference on recommender systems, Chicago, IL, USA, pp 289–292

Wu Z, Wu J, Cao J, Tao D (2012) HySAD: a semi-supervised hybrid shilling attack detector for trustworthy product recommendation. In: 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 985–993

Wu Z, Wang Y, Wang Y, Wu J, Cao J, Zhang L (2015) Spammers detection from product reviews: a hybrid model. In: IEEE international conference on data mining (ICDM 2015), pp 1039–1044

Xia H, Fang B, Gao M, Ma H, Tang Y, Wen J (2015) A novel item anomaly detection approach against shilling attacks in collaborative recommendation systems using the dynamic time interval segmentation technique. Inf Sci 306:150–165

Yang Z, Cai Z, Guan X (2016a) Estimating user behavior toward detecting anomalous ratings in rating systems. Knowl Based Syst 111:144–158

Yang Z, Xu L, Cai Z, Xu Z (2016b) Re-scale AdaBoost for attack detection in collaborative filtering recommender systems. Knowl Based Syst 100:74–88

Yi H, Zhang F (2016) Robust recommendation method based on suspicious users measurement and multidimensional trust. J Intell Inf Syst 46(2):349–367

Yu H (2014) An algorithm for detecting recommendation attack based on incremental learning. J Inf Comput Sci 11(7):2365–2373

Yu H, Gao R, Wang K, Zhang F (2017) A novel robust recommendation method based on kernel matrix factorization. J Intell Fuzzy Syst 32(3):2101–2109

Yuan NJ, Zheng Y, Zhang L, Xie X (2013) T-finder: a recommender system for finding passengers and vacant taxis. IEEE Trans Knowl Data Eng 25(10):2390–2403

Zafarani R, Liu H (2009) Connecting corresponding identities across communities. In: International conference on weblogs and social media (ICWSM 2009), vol 9, pp 354–357

Zhang FG (2009) A survey of shilling attacks in collaborative filtering recommender systems. In: Proceedings of the international conference on computational intelligence and software engineering, Wuhan, China, pp 1–4

Zhang FG (2011) Analysis of bandwagon and average hybrid attack model against trust-based recommender systems. In: 5th International conference on management of e-commerce and e-government, Hubei, China, pp 269–273

Zhang F, Sun S (2014) A robust collaborative recommendation algorithm based on least median squares estimator. JCP 9(2):308–314

Zhang FG, Xu SH (2007) Analysis of trust-based e-commerce recommender systems under recommendation attacks. In: 1st International symposium on data, privacy, and e-commerce, Chengdu, China, pp 385–390

Zhang F, Zhou Q (2014) HHT–SVM: an online method for detecting profile injection attacks in collaborative recommender systems. Knowl Based Syst 65:96–105

Zhang S, Ouyang Y, Ford J, Makedon F (2006a) Analysis of a low-dimensional linear model under recommendation attacks. In: 29th Annual international ACM SIGIR conference on research and development in information retrieval, Seattle, WA, USA, 2006:517–524

Zhang S, Chakrabarti A, Ford J, Makedon F (2006b) Attack detection in time series for recommender systems. In: 20th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA, pp 809–814

Zhang Y, Tang J, Yang Z, Pei J, Yu PS (2015) COSNET: connecting heterogeneous social networks with local and global consistency. In 21st ACM SIGKDD international conference on knowledge discovery and data mining, pp 1485–1494

Zhang F, Lu Y, Chen J, Liu S, Ling Z (2017) Robust collaborative filtering based on non-negative matrix factorization and R 1-norm. Knowl Based Syst 118:177–190

Zhou W, Koh YS, Wen J, Alam S, Dobbie G (2014) Detection of abnormal profiles on group attacks in recommender systems. In 37th International ACM SIGIR conference on research and development in information retrieval, pp 955–958

Zhou W, Wen J, Koh YS, Xiong Q, Gao M, Dobbie G, Alam S (2015) Shilling attacks detection in recommender systems based on target item analysis. PLoS ONE 10(7):e0130968

Zhou W, Wen J, Xiong Q, Gao M, Zeng J (2016) SVM-TIA a shilling attack detection method based on SVM and target item analysis in recommender systems. Neurocomputing 210:197–205

Ziegler CN, Golbeck J (2015) Models for trust inference in social networks. In: Propagation phenomena in real world networks, pp 53–89

Zou J, Fekri F (2013) A belief propagation approach for detecting shilling attacks in collaborative filtering. In: 22nd ACM international conference on Conference on information and knowledge management, pp 1837–1840