Critique on Natural Noise in Recommender Systems

Wissam Al Jurdi, Jacques Bou Abdo, Jacques Demerjian, Abdallah Makhoul

Critique on Natural Noise in Recommender Systems

WISSAM AL JURDI, Univ. Bourgogne Franche-Comté, FEMTO-ST Institute, CNRS, 1 cours Leprince-Ringuet, 25200, Montbéliard, France | LaRRIS, Faculty of Sciences, Lebanese University, Fanar, Lebanon
JACQUES BOU ABDO, University of Nebraska at Kearney, USA
JACQUES DEMERJIAN, LaRRIS, Faculty of Sciences, Lebanese University, Fanar, Lebanon
ABDALLAH MAKHOUL, Univ. Bourgogne Franche-Comté, FEMTO-ST Institute, CNRS, 1 cours Leprince-Ringuet, 25200, Montbéliard, France

Recommender systems have been upgraded, tested and applied in many, often incomparable ways. In attempts to diligently understand user behavior in certain environments, those systems have been frequently utilized in domains like e-commerce, e-learning, and tourism. Their increasing need and popularity have allowed the existence of numerous research paths on major issues like data sparsity, cold start, malicious noise and natural noise, which immensely limit their performance. It is typical that the quality of the data that fuel those systems should be extremely reliable. Inconsistent user information in datasets can alter the performance of recommenders, albeit running advanced personalizing algorithms. The consequences of this can be costly as such systems are employed in abundant online businesses. Successfully managing these inconsistencies results in more personalized user experiences. In this article, the previous works conducted on natural noise management in recommender datasets are thoroughly analyzed. We adequately explore the ways in which the proposed methods measure improved performances and touch on the different natural noise management techniques and the attributes of the solutions. Additionally, we test the evaluation methods employed to assess the approaches and discuss several key gaps and other improvements the field should realize in the future. Our work considers the likelihood of a modern research branch on natural noise management and recommender assessment.

Additional Key Words and Phrases: recommender systems, natural noise management, evaluation metrics

ACM Reference Format:
Wissam Al Jurdi, Jacques Bou Abdo, Jacques Demerjian, and Abdallah Makhoul. 2021. Critique on Natural Noise in Recommender Systems. 1, 1 (January 2021), 28 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Throughout the years, recommender systems (RS) are becoming more and more crucial to almost all online businesses worldwide [1]. With various methods ranging from prominent collaborative filtering (CF) techniques to advanced latent factor models, they portray a significant role in most top-ranked commercial platforms like Amazon, Netflix, Spotify, and Last.fm [2]. This emerges from the substantial problem such approaches try to efficiently tackle through highly personalized services, the information overload. The underlying power of the

Authors' addresses: Wissam Al Jurdi, wissam.aljurdi@st.ul.edu.lb, Univ. Bourgogne Franche-Comté, FEMTO-ST Institute, CNRS, 1 cours Leprince-Ringuet, 25200, Montbéliard, France | LaRRIS, Faculty of Sciences, Lebanese University, Fanar, Lebanon; Jacques Bou Abdo, University of Nebraska at Kearney, USA, bouabdoj@unk.edu; Jacques Demerjian, LaRRIS, Faculty of Sciences, Lebanese University, Fanar, Lebanon, jaques.demerjian@ul.edu.lb; Abdallah Makhoul, Univ. Bourgogne Franche-Comté, FEMTO-ST Institute, CNRS, 1 cours Leprince-Ringuet, 25200, Montbéliard, France, abdallah.makhoul@univ-fcomte.fr.

XXXX-XXXX/2021/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

personalized recommendations generated by various types of RSs is primarily dependent on the presence of generous user contributions in the forms of ratings, reviews, tags, etc. Researchers studying and enhancing RSs and their algorithms have tremendously focused on algorithmic improvements paying nominal attention to the quality of the data. The involvement of the human factor in the rating elicitation process is immensely prone to errors. Ratings, reviews, and other details recommender algorithms rely on hold critical information that might not always be sincere or consistent. This is recognized as noise in the datasets used by RSs to personalize information to users. Naturally, if recommenders employ inaccurate data to learn user behavior, they will inevitably output inconsistent and unsatisfactory results.

There are two types of noise in RSs, malicious noise, and natural noise. Simply put, noise is the rating feedback that does not reflect a user's true preference or intention. This might be purposely arranged by attackers for certain reasons like biasing a recommender's output (malicious noise) [3], or it could occur naturally because of a user's inconsistent or negligent rating behavior (natural noise) [4][5]. Malicious noise results from numerous attacks carried on online applications that are typically powered by diverse types of RSs. This field has witnessed much attention in the past years [3], conversely, the natural noise domain hasn't yet received the full focus of researchers. Natural noise occurs inherently due to user behaviors, and that's what makes it unique. As emphasized by the very first work [6] and described through the publications embodying it at later stages, natural noise solely occurs due to human error that lead to data inconsistencies. It does not produce any pattern, and consequently it's unusually complex to model. Significant improvements are required to develop a generic noise-aware recommendation algorithm capable of overcoming natural and malicious inconsistencies that might be present in the datasets of RSs.

The performance of recommenders that's predominantly measured with conventional yet renowned offline tests employing accuracy metrics, such as MAE, RMSE and F1-Score, almost always records scant improvements. This poses a critical issue in the testing mechanism since evaluating RSs is inherently difficult for many reasons [7]. First, different algorithms appear to perform better or worse on varied datasets. Second, the goals for which recommenders are evaluated may differ; current works like the natural noise field primarily focus on accuracy improvements while the proper aim of a recommender is to provide a substantial personalized experience. Accuracy falls short on measuring the most fundamental aspect of an algorithm and that is how much personalized the results are for a user [8]. Researchers tend to focus on amplifying the accuracy tests of the system, always offline, while very few target other properties that have notable effect on user personalization. Ultimately, commercial systems measure user satisfaction by the number of products actually purchased from recommendations and not by the score of a recommender's MAE or RMSE. Third, an authentic comparative evaluation of recommender algorithms poses a significant challenge in deciding what combination of accuracy metrics to use. The first and second reasons can be partly attributed to the fact that the quality of user interactions in the datasets used by RSs is frequently overlooked.

Throughout this work, the natural noise studies were grouped into three main paths (Figure 1), the Magic Barrier path, the classical natural noise management path, and the preference-dependent natural noise management path. The term *natural noise* was introduced by O'Mahony et al. [6] as the inconsistencies in user data that occur without malicious intent. Subsequently, it was demonstrated by Amatriain et al. [4] that indeed many users are actually inconsistent in the rating elicitation process. The most significant studies that influenced the natural noise research topic and the three paths were Herlocker et al. [7] and O'Mahony et al. [6]. A threshold termed *Magic Barrier* was speculated [7] in which the authors argued that there seems to be a certain point were recommenders fail to get more accurate. They attributed this discovery to "inherent variability" in datasets - inconsistent user profiles. The pivotal outlook to point out in this case is that the authors only analyzed the evaluation methods of RSs and did not refer in any way to algorithm enhancements. This essential viewpoint was missed by the first path on natural noise that originated from [7] where the authors debated that other types of evaluation metrics are to be engineered. They also emphasized that algorithms should be measured in accordance

with how well they can communicate their reasoning to users, or with how little data they can yield accurate recommendations; if this is valid, researchers require new metrics to evaluate those new algorithms. Therefore, the study in [7] did not discuss nor prove that noise reduction induces better recommender performance, it merely proposed the concept of curating new algorithms with distinct evaluation techniques. The path that originated from [7] in the natural noise field bore an alternative interpretation of the matter and tried to quantify the Magic Barrier limit in hopes for better accuracy results, completely missing the point of the important study in [7]. Detached from the concept of the Magic Barrier, the second path targeted dealing with natural noise through several techniques mainly tested on CF algorithms. Some of the proposals typically employed classic clustering methods to identify variations in user profiles while others resorted to more complicated fuzzy profiling techniques and matrix factorization modeling. In the third path of natural noise management, few proposals joined typical datasets with each other for secondary data as a natural noise management solution.

An intriguing point to note about natural noise management proposals is that throughout the three paths, the difference between identifying noise at the level of ratings and dealing with it at the level of users (noisy ratings vs. noisy users) was never technically analyzed. Toledo et al. [9] explicitly state that Li et al. [10] cover natural noise at the user level (identify if a user is inconsistent in his rating or not) and that it is necessary to provide a ratings-based solution. Unfortunately, no supporting evidence was provided to demonstrate how this would actually benefit a recommendation system in terms of performance after natural noise management.

Personalizing algorithms that cater for the main aim of recommenders might be missing key algorithmic improvements that ought to be measured by means beyond accuracy, however, the results of those algorithms is radically dependant on the quality of the underlying dataset. Thus, accounting for natural noise in the datasets is of paramount importance and an area that requires deeper investigation. Further, if researchers plan on working towards achieving improved means for measuring recommenders, it is necessary for them to re-evaluate the current protocols (natural noise algorithms or any other recommender approach) that previously relied on conventional evaluation metrics to judge performances and benchmark results.

The discussion on the validity of the evaluation methods used on all the natural noise approaches will be done through the functional analysis of the following two hypotheses:

(1) With the same recommender configuration and on various datasets, random rating removal cannot produce better performance results compared to a natural noise management method.
(2) The accuracy metric results of the above experiments always result in consistent measurements.

This article presents the following contributions to the natural noise management field:

(1) A survey that classifies natural noise management techniques that have been proposed since 2006 and conceptually analyzes their strengths and weaknesses.
(2) Analysis and critique on evaluation metrics, benchmark datasets and recommender types that have been used in the natural noise management proposals.
(3) A comparison through statistical analysis of the natural noise management mechanisms and their underlying attributes providing insight as to how the natural noise path ought to sustain its development; highlights on the significant gaps in the field are also presented.
(4) An evaluation of the two hypotheses and a demonstration on how the uncorrelated results adversely affect the natural noise proposals path.

The remainder of the article is arranged as follows:

- Section 2. A presentation of some state-of-the-art works on the natural noise management field.
- Section 3. A discussion of all the natural noise management proposals since the field initiation in 2006. The algorithms are grouped into three major paths.
- Section 4. Statistics and analysis of the main attributes used in the natural noise paths like the evaluation metrics, benchmark datasets and recommender types.

- Section 5. An investigation of the accuracy metrics that have been used in evaluating the accuracy of collaborative filtering predictions and recommendations after natural noise management. This section presents the gaps in the natural noise management paths.
- Section 6. Final conclusions, including a list of areas were we feel future work is particularly warranted.

Table 7 of the appendix summarizes all the notations used in this paper.

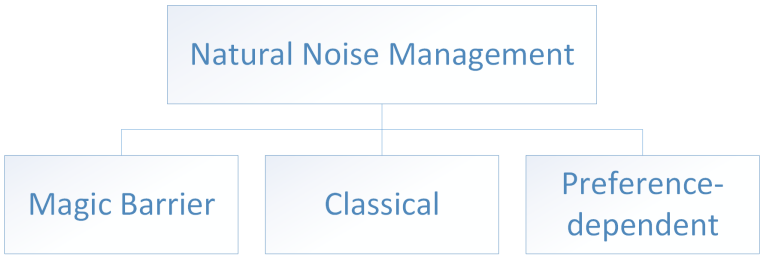

Fig. 1. Natural noise management paths

2 RELATED WORK

After the subject of malicious noise in RSs [3] has been extensively addressed, NN has lately started to deeply spark the interest of researchers. Ever since the path's introduction in [6], it has taken on several forms as proposals approached the problem in unique assorted ways. Till now, there has been no deep analysis of the diverse proposals on NNM introduced in the literature. Approaches such as those that directly deal with NNM or discuss it in terms of the concept of the Magic Barrier touched upon in [7]; moreover, there are no direct surveys on the topic. Practically all proposals in the NNM path followed the same discussion strategy throughout their idea development before introducing the approach. Summarized, they basically state that Amatriain et al. [4] deployed a re-rating method on users and 40% of them displayed inconsistent results with their previous ratings. This confirmed that users' ratings can be irregular in the sense that they may rate the same item differently at diverse points in time, and proved the primary idea of Hill et al. [11]. Their research fundamentally influenced the speculations about various types of evaluation techniques for a sophisticated level of personalizing algorithms in [7], by showing users provide inconsistent ratings when asked to rate the same movie at separate points in time.

Castro et al. [12] and Toledo et al. [9] in the classical NNM path mentioned a few algorithms and previous work. In those connected studies that represent one type of the notable approaches to NNM, the authors categorized a few previous NNM works into two primary classes, the first that target individual recommendations and the other that target recommenders for groups of users [13]. The approaches are then split into two groups, those that are based on crisp functions and those that introduce fuzzy profiling. These works do not mention any research from the Magic Barrier path. Subsequently, Martínez et al. [14] discussed NNM in RSs and summarized their previous approaches in [9] and [12]. However, the study does not provide an attribute analysis of all the previous NNM approaches in the literature nor direct comparisons in terms of datasets or algorithm complexities. One very recent study on NNM in RSs by Bag et al. [15] introduced a sparsity-aware model by slightly amending the previous approach of Toledo et al. in [9]. This study mentions few researches from the literature, however, it was very brief and lacked technical analysis. Most major research fields were unmentioned and dismissed from the implementations.

NN in recommender systems appears to lack a well-defined track of research approaches. It is rather pursued from many viewpoints (e.g. [6],[7],[16]) as clearly seen in every publication/article in the divergent paths. NNM is in need of a well-defined path for addressing inconsistencies in any recommender dataset. To efficiently

overcome the issue of NN in RSs, researchers must develop an algorithmic program that works on any type of data, is ergonomic, has a reasonable execution time, and significantly impacts key personalization metrics (Beyond accuracy metrics [8]). In this work, we extensively cover all the NNM approaches, the techniques used in developing the algorithms and statistical analysis of the attributes deployed with them. We categorize all studies based on the paths they took, the complexity of the algorithms and the dependence on supplemental data that might be unavailable in regular datasets. We start from the point were NN was first introduced in [6] and targeted in the path that was influenced by [7]. Furthermore, we provide strategic directions about the NN field and discuss numerous gaps and essential critical points towards it taking into consideration the notion of serendipity in recommenders.

3 APPROACHES TO NATURAL NOISE MANAGEMENT

There were several attempts to change the research evaluation process of RSs from the offline accuracy-focused studies to the online user-personalized tests [7]. Nevertheless, up until now, the evaluation retains the initial accuracy-based solutions. It is the case in all the NNM paths to be discussed in this section. This is likely attributed to the Netflix prize competition that evaluated the performance of the winning algorithm based on better accuracy results (RMSE) [1]. Netflix ended up not using the winner algorithm since they have realized that better RMSE does not strictly mean superior personalized recommendations [2]. Netflix's reason to disregard the winning algorithm was exactly what was discussed by Herlocker et al. [7]. The authors introduced the concept of the Magic Barrier limit and speculated how the evaluation process should be re-visited from a peculiar angle. One major feature they stressed on is discarding accuracy metrics and focusing on engineering new metrics for user-oriented approaches such as how well an RS communicates it's reasoning to users, or with how little data it can yield accurate recommendations.

The first study on NN emerged side by side with [7] in 2016 by O'Mahony et al. [6], and since then has attracted the attention of researchers gaining much popularity very recently. It was backed up by a study conducted in 1995 [11] and took on several definitions such as, the noise that results from user's preference change over time or, the inherent rating inconsistencies of users. However, the studies that resulted from [7] and [6] were not consistent and held various approaches to the problem. The first path that started out from [7] does not actually deal with NN in RSs datasets. Simply put, it proposes an attempt to merely assess the quality of RSs based on the concept of inherent noise and variations in user-rating over time. There were no solutions as to how improvements to RSs beyond this calculated limit (Magic Barrier) can be achieved, but only how to approach calculating the limit based on accuracy standards such as RMSE. The following sections will introduce the NNM researches categorized into three major paths based on several criteria especially the way the researches approached the problem.

The first path that emerged from the Magic Barrier study in [7] mainly focused on calculating the Magic Barrier of RSs. This wasn't an approach to deal with NN, but an attempt to open up a way for further improving RSs from an accuracy outlook. The second path dealt directly with NN and most papers throughout the path proposed methods to either eliminate noise, or correct it. This path was named the classical NNM as most of the algorithms were based on discrete formulas and work on datasets that contains users and their ratings only. The last approach was termed the preference-dependent path. With more sophisticated algorithms, it mainly focuses on information that further extend the basic data in most widely used RSs datasets, such as reviews, director information (in the case of movie datasets) etc. All the published studies in this first and second paths are laid out in timelines with brief summaries in tables 1 and 3 respectively.

[1] https://www.thrillist.com/entertainment/nation/the-netflix-prize, accessed: 12/01/2019
[2] https://www.wired.com/2012/04/netflix-prize-costs, accessed: 12/01/2019

3.1 The Magic Barrier - Logic vs. Accuracy

The study of natural noise all started out with the term Magic Barrier that was speculated in [7] and which presented a highly significant challenge that is present in deciding what combination of measures to use in comparatively evaluating recommenders in general. All enhancements and tuning on the algorithms that constitute RSs appeared to produce similar output qualities in terms of the MAE accuracy metric – "many researchers find their newest algorithms yield an MAE of 0.73 (on a five-point rating scale) on movie rating datasets. Though the new algorithms often appear to do better than the older algorithms they are compared to, we find that when each algorithm is tuned to its optimum, they all produce similar measures of quality". Hence, the expression Magic Barrier was introduced as the point where natural variability may prevent us from getting any more accurate.

From then forward, the NN research has taken on several paths, the first being a series of publications by the same authors where they tried to measure and quantify the Magic Barrier of [7]. Their path that originated from [17] appears to have taken its own approach under the NNM route and defined the Magic Barrier from their own perspective and terms. It exhibited little correlation and very few comparisons with other approaches that explicitly dealt with NNM in recommenders' datasets. On top of that, it can be observed that the researchers tried tackling the Magic Barrier of [7] by defining it as the point at which the performance and accuracy of a recommender algorithm cannot be further enhanced due to inherent noise in the data, and every improvement in accuracy (exclusively measured by MAE or RMSE) might denote an over-fitting and not a more competent performance. In addition, they quantified this definition by the notion that a mathematical characterization of the Magic Barrier that was speculated in [7] is missing. They presented this characterization of the Magic Barrier based on RMSE and claimed it allows us to assess the authentic performance of a recommender as well as compute the actual room for improvement. The research path that branched from this study is explained and analyzed below and will be referred to as the *Accuracy Barrier* while the concept that [7] introduced will be referred to as the *logic barrier* in order to separate the two and avoid confusion for future research on the topic as they are different at their core.

Before continuing with presenting the Accuracy Barrier path, it is substantial to note that the authors explicitly state that this approach represents a mere attempt to estimate the logic barrier. It is impossible to directly determine the logic barrier because it involves an optimal rating function which is usually unavailable [18][19].

3.1.1 Major Path on the Subject. In their early model [17], the authors conducted an experiment with a user study scenario in an attempt to assess and quantify the Accuracy Barrier of a recommendation system. The experiment included designing an online form to gather opinions from users on items that they have previously rated using the MoviePilot recommender and dataset [3]. The difference between the opinions and ratings was defined as the Accuracy Barrier of the dataset powered by the RMSE metric. The ultimate assumption was that the Accuracy Barrier of RSs can be better assessed by noise estimation. They presented a preliminary model for the Accuracy Barrier and the level of accuracy a recommender system can achieve without over-fitting to the noise in the data. The authors assumed the existence of additional transactions for $r_{u,i}$ given at different points in time and called them $o_{u,i}$ (opinion of user u on item i). After that, the error between those ratings was defined as $\epsilon_{u,i} = o_{u,i} - r_{u,i}$ and the first attempt towards the Accuracy Barrier estimation was proposed in equation 1.

$$E(f * |R) = \sqrt{\frac{1}{|R|} \sum_{(u,i)\epsilon R} (o_{u,i} - r_{u,i})^2} \tag{1}$$

where $f*$ is an unknown rating function that knows the true opinions $o_{u,i}$ of each user u about any item i. Equation 1 refers to the estimate RMSE of function $f*$. The authors continue to stress on the idea that there might

[3] https://www.moviepilot.de/

be a rating function f that results in a lower RMSE on R, however, those tend to over-fit the given rating set R and are likely to degrade the recommendation performance and that is why equation 1 defines their Accuracy Barrier point.

Their idea was further developed and backed-up in another work [19]. In it, they expanded the analysis and the case-study with a commercial movie recommender and investigated the inconsistencies of the user ratings. In addition, they provided an estimate of the Accuracy Barrier to attain their goal of assessing the genuine quality of a recommender. The same mathematical characterization of the Accuracy Barrier was further developed and expanded yet still based solely on RMSE as in equation 1; according to the authors, that allows the assessment of the authentic performance of a recommender as well as the amount of room for improvement. They reveal how the Accuracy Barrier represents the standard deviation of inherent rating inconsistencies in user ratings and present a noise model before deriving it. After the estimation of the Accuracy Barrier for MoviePilot, the authors concluded that said estimate is useful for assessing the quality of a recommendation method and revealing room for improvements. Recommenders with a prediction accuracy close to the estimated Accuracy Barrier can be regarded as *optimal*. They continue to state that further improvements on such recommenders are meaningless. The mathematical representation of their estimate of the Accuracy Barrier was further developed in a procedure [19] and took the following final form based on the average: $B_x = \sqrt{\frac{1}{|X|}\sum_{(u,i)\epsilon X}\epsilon^2_{u,i}}$; where X is a randomly generated subset of user-item pairs and $\epsilon^2_{u,i}$ is the variance of the ratings. With a similar logic as before, the authors added there might be a rating function $f \epsilon F$ that results in better RMSE scores, however, this is considered over-fitting and meaningless improvements. The results of the experiment show that the recommender system of MoviePilot can be better enhanced since the Accuracy Barrier yielded a value of 0.61 (close to the numerical step of the rating scale of MoviePilot) while the RMSE of MoviePilot's recommendation engine is about 1.8.

Subsequently, Bellogin et al. [18] continued approaching the problem in an alternative way. They defined an experimental method to calculate the coherence of users in a dataset and revealed how the results are correlated with the Accuracy Barrier of [19] in RSs. They utilized an external source to achieve this goal with which one can measure the inconsistencies in the ratings by describing them in terms of specific features like genres (the authors adopted movie datasets like MovieLens). The formulation of the coherence of a user u based on a set of item features F was formulated according to equation 2.

$$c(u) = -\sum_{f \epsilon F} \sigma_f(u) \tag{2}$$

The authors adopted the standard deviation for calculating the coherence of user profiles where $\sigma_f(u)$ is defines as:

$$\sigma_f(u) = \sqrt{\sum_{i \epsilon I(u,f)} (r(u,i) - \bar{r}_f(u))^2} \tag{3}$$

Where $\bar{r}_f(u)$ corresponds to the average rating within the set of items rated by user u that belong to feature f. It is obvious here that $\sigma_f(u)$ is the standard deviation used by the authors to represent the variation between the user's rating and a specific feature f, and $c(u)$ basically measures the variance of an individual's rating relative to the feature space by which items are defined. Based on the formulation of equation 2, the users are clustered into two groups, easy and difficult. This will then constitute the training set of groups formed to train a recommender algorithm. Employing a UB-CF approach with 5-fold cross-validation, the authors evaluated their method using RMSE because it is related to the concept of the Accuracy Barrier in [17][19]. The results of their experiment revealed that:

- The user coherence of equation2 provides good predictions of the Accuracy Barrier for a recommender.
- It is possible to utilize the user coherence groups to build different training and test models in such a way that the error decreases for every user (accuracy error).

In their latest study [20], the authors provided a more explicit representation of the Accuracy Barrier expressed in [19], along with a correlation with [18]. There are no other contributions to their Accuracy Barrier approach yet further experiments demonstrated that being statistically coherent in terms of rating deviation within an item's attribute space (genres in this case) can convey enough information to predict the users' inconsistencies. The study also concluded how an RS can be trained differently depending on the users' inconsistencies predicated by their rating coherence [19] (equation 2); this allowed cheaper (less computation power, time and tuning) recommendation cycles for the easy users' group (those with high coherence). Furthermore, the experiments also revealed that the prediction performance can be improved by 10% to 40% when only training with easy users while the group labeled as difficult will receive worse recommendations in general.

3.1.2 Path Influenced by the Magic Barrier. Amatriain et al. [4], backed up by a small proposal [5] done in 2009, addressed the problem of analyzing and characterizing the noise in user ratings. They presented a user-study aimed at quantifying the noise that originates from inconsistencies in those ratings. The research tried to answer the following important queries on the subject of NN:

- Are users inconsistent when providing ratings?
- How large is the error due to such inconsistencies?
- What are the factors that have an impact on user inconsistencies?

This study performed three trials and involved 118 users who were asked to rate items from a calculated subset of the Netflix Prize dataset in an attempt to analyze the user inconsistencies in items they had rated. They came up with three primary variables that produced a significant impact on the user inconsistencies:

(1) The rating scale. Ratings are more consistent at the ends of the scale and significantly less consistent in the middle of it.
(2) Item order. A rating interface that groups movies that are likely to receive similar ratings should help minimize user inconsistencies.
(3) User rating speed. This might sound counter-intuitive, however, the smaller the time interval between ratings in a row the fewer the user inconsistencies are in a dataset.

It is unclear how the authors related their study of rating inconsistencies and user stability metrics through RMSE to the logic barrier of [7]; however, what's clear is that this study influenced the path of the Accuracy Barrier that started with [17][19] especially the RMSE approach for calculating user rating inconsistencies.

The study by Yu et al. [21] was moderately influenced by the Accuracy Barrier, and the authors used the same clustering method of [19] in their approach to overcoming the issue of NN datasets. Unlike the previously discussed studies in the path, this research provides a broader solution to RSs. It explored directly dealing with noise in datasets, generating recommendations, and measuring the performance improvements. The authors proposed a generic framework to seamlessly harness different pre-processing, and recommendation approaches for ratings of unique users. The users in a dataset are classified into several groups based on the quantity and quality of their ratings by several data pre-processing strategies. After that, the authors suggest a transfer latent factor model to convey trained models between groups in the training phase. Additionally, it was argued how recommenders that take all user information as input suffer from two major challenges, data quantity, a computational challenge, and data quality, an NN challenge. The primary idea of the approach is to variably process diverse types of users when training RSs. This is due to the established fact that users possess dissimilar rating quantities and the of those inherently varies with behavior. Moreover, some users maintain consistent rating behavior while others suffer from plenty of inconsistencies. The key steps of the approach are shown in Figure 2 and summarized as follows:

(1) Classifying user groups.

Users were split into six groups based on two major criteria, the number of ratings a user has (quantity) and the coherence measure of a user (quality). The authors adopted the coherence approached proposed in [18] (equation 2). The user groups generated from this approach are shown in Table 2.

(2) Processing noisy ratings.
The noise detection method was also inspired by the proposal of Bellogin et al. in [18]. The authors adopted from [18] the idea of item features (based on genres) and implemented the following equation which calculates the rating noise degree: $RND_{(r_{ui})} = \sum I\left(\sum_{fi} \frac{|r_{ui}-\bar{r}_{uf}|}{\bar{r}_{uf}} > \vartheta\right) \Big/ \|f_i\|$, where f_i represents the item features similar to equation 3 from [18][19]. RND expresses the relationship between the features that have a significant relative deviation (more than the threshold ϑ compared with its same item feature set) and the aggregate number of features. The processing of the ratings identified as noisy was dealt with differently taking into account that removing ratings from the light users' group would worsen the sparsity problem. Accordingly, the authors adopted three options. First, no noise processing was done for medium and easy users. Second, the noise was removed for heavy users only and third, noise correction for was implemented on the light users' group. The correction method was done based on the average rating of items that had the same features according to equation : $r'_{ui}(corrected) = \frac{\sum_{fi \epsilon F} \bar{r}_{uf_i}}{\|F_i\|}$. Ultimately, a sampling phase was implemented for heavy users because their ratings contain redundant and repetitive information. The authors adopted the harmonic mean of entropy, replacing entropy with variance and inverse frequency to account for items in the long tail.

(3) Transferring models between user groups.
In the final step in their proposed approach, the authors observed the data quality and quantity varied sharply between the groups of users. As a result, they proposed to transfer the trained item latent factor models between those groups. The results of the protocol (Figure 2), measured by RMSE and precision, conveyed how the recommendation performance was significantly enhanced.

In an attempt to improve the recommendation output of RSs, Saia et al. [22] introduced a new approach based on a previous proposal in [23]. They measured the similarity between two items from a user profile and discarded those that appear as highly dissimilar. The authors argue that by eliminating those incoherent items from a user profile, the metrics' (RMSE and Average Difference) accuracy improvements will be genuine and not over-fitting or useless; however, their approach requires item text descriptions in the datasets. It requires four particular steps, data pre-processing, semantic similarity evaluation, dynamic coherence-based modeling and finally, item recommendations.

3.2 The Classical Natural Noise Path

Arguably the most famous NN approach was proposed in 2006 by O'Mahony et al.[6], in parallel with the prominent [7] that had influenced the Accuracy Barrier path. The authors were first to introduce the term *Natural Noise* in RSs' datasets and defined it as the noise that arises from imperfect user behavior when rating and reviewing items. Predominantly, the noise in datasets was grouped into two significant categories:

(1) Natural: That which results from human activity errors when rating items they view/purchase.
(2) Malicious: That which results from deliberately biasing reviews in a system mainly to increase the recommendation frequency.

Their core objective, in this case, was to develop techniques to identify NN and discard it to improve the accuracy performance of a recommender. In addition, their solution also accounted for one type of malicious noise as well, but in this study, we are only focusing on NNM. Broadly, the authors measured the consistency of a certain rating $r_{u,v}$ as the MAE between the actual rating of a user and the predicted rating ($p_{u,i}$) of said user. The predicted rating can be identified using a certain recommendation algorithm (G) that is trained with a trusted set of user

Table 1. Timeline of the Accuracy Barrier path after the speculation on the evaluation of recommenders in [7]

2006	**[7]** The concept of a Magic Barrier is speculated. Ways beyond accuracy that should be implemented for the sake of superior evaluation of recommender systems are discussed. The approach is referred to as the *logic barrier* in this study
*2009	[5][4] The problem of analyzing and characterizing the noise in user feedback through ratings of movies is introduced. The noise that originates from inconsistencies in ratings is quantified
2012	**[17][19]** The authors introduce an extension to the idea presented in [7]. A measure using RMSE is estimated and referred to as the Magic Barrier. In this study, their approach is dubbed as the *Accuracy Barrier*. It requires re-rating items by the same users
2014	**[18]** The authors propose a user classification approach that predicts the Accuracy Barrier of datasets in [17][19]. It analyzes user ratings using other factors in datasets such as genres. The study complements the ideas of the Accuracy Barrier approach and uses RMSE as a metric for recommender evaluation
*2016	[21][22] Influenced by the user classification method of [18], a noise management algorithm that requires feature availability in datasets (such as genres) is proposed by [21]. The study [22] introduces a coherence-based method to identify inconsistent items
2018	**[20]** The authors introduce an extension of works [19], in terms of the Accuracy Barrier experiment, and [18], in terms of the user coherence formulation

* These publications are not directly linked to the Magic Barrier path, however, they are directly/indirectly influenced by it.

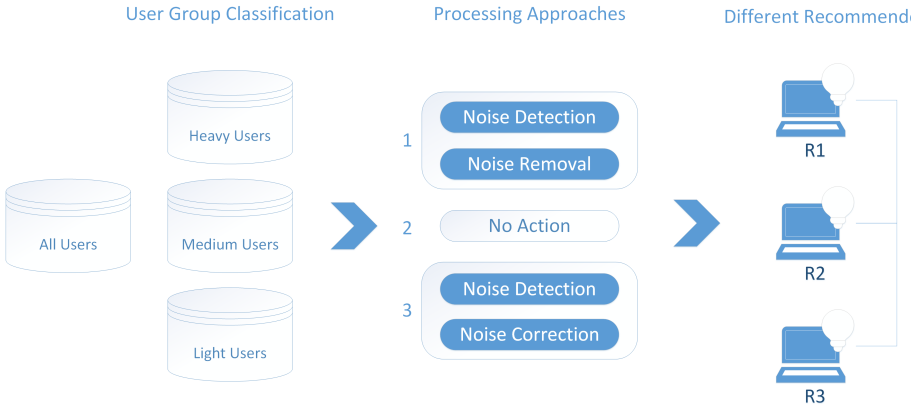

Fig. 2. The proposed framework of [21] which includes a natural noise management mechanism

Table 2. Different user-item classification groups adopted in the classical natural noise path such as in studies [9] (left) and [21] (right)

Group	Description
Critical user	$\|W_u\| \geq \|A_u\| + \|S_u\|$
Average user	$\|A_u\| \geq \|W_u\| + \|S_u\|$
Benevolent user	$\|S_u\| \geq \|W_u\| + \|A_u\|$
Weakly-pref item	$\|W_i\| \geq \|A_i\| + \|S_i\|$
Averagely-pref item	$\|A_i\| \geq \|W_i\| + \|S_i\|$
Strongly-pref item	$\|S_i\| \geq \|W_i\| + \|A_i\|$

Group	Description
HEUG	high ratings, high consistency
HDUG	high ratings, low consistency
MEUG	medium ratings, high consistency
MDUG	medium ratings, low consistency
LEUG	few ratings, high consistency
LDUG	few ratings, low consistency

data, meticulously selected by a system administrator. This consistency was formulated as follows:

$$c(G,T)_{u,i} = \frac{|r_{u,i} - p_{u,i}|}{r_{max} - r_{min}} \tag{4}$$

A rating was considered noise if $c(G,T)_{u,i}$ was greater than some threshold th. The authors argued their approach allowed the possibility of analyzing the ratings of a neighborhood (k) of a certain user we wish to recommend items for. The experiment results were conducted on several selected training sets and revealed how MAE was improved with minor development in coverage when NN was completely eliminated according to equation 4.

Li et al. [10] target NN in social RSs and refer to it as noisy but non-malicious users (NNMU). Their idea was based on the assumption that the ratings provided by the same user on closely correlated items should produce similar scores. The authors proposed a method for NNMU detection by capturing and accumulating individuals' self-contradictions. Formulating it as a constrained quadratic optimization problem, they defined those self-contradictions as the cases where a unique user provides very different rating scores on closely correlated items. Unlike the previous approach [6], this method identified noisy users. If a certain user appears to have made too many self-contradictions, the noise degree of his profile will rise and he will automatically be classified as an NNMU. The optimization problem had the following input and output:

- Input
  - G - item-item correlation graph

  - y_L - all ratings in the test user profile
- Output
  - $\rho \epsilon [0, 1]$ - amount of noise in y_L (high $\rho \implies$ more likely to be a NNMU). Where $\rho = \frac{1}{K} \sum_{k=1}^{K} |\hat{\xi}_k|$ and $R_{min} - y_L \leq \xi \leq R_{max} - y_L$

Toledo et al. [24] with an extended publication by the same authors in [9] propose an alternative approach to deal with NN on the rating level in recommenders' datasets. The proposed framework includes two phases:

(1) Noise detection: Verifying if a rating is considered as noise based on a user-item profile classification scheme
(2) Noise correction: Employing a classic CF method to predict a new rating to replace the noisy ratings when necessary

Based on a group of ratings ($r_{u,i}$) classes (weak, mean and strong) for both items and users, the authors define three particular sets for each group that constitute the preferences for each user/item: W_u, A_u and S_u for users and W_i, A_i and S_i for items. The thresholds used to group the ratings and users into the three sets are defined in equations 5 and 6 which are applied for both the item and the user sets.

$$\begin{aligned} W &= |\{r_{u,i} < k\}| \\ A &= |\{k \leq r_{u,i} < v\}| \\ S &= |\{r_{u,i} \geq v\}| \end{aligned} \tag{5}$$

$$\begin{aligned} k &= r_{min} + \{\tfrac{1}{3}(r_{max} - r_{min})\} \\ v &= r_{max} - \{\tfrac{1}{3}(r_{max} - r_{min})\} \end{aligned} \tag{6}$$

Subsequently, the classification for each group is performed based on Table 2, and after that, the possible noisy ratings ($r_{u,i}$) are corrected ($r^*_{u,i}$) using a traditional UB-CF algorithm with PCC, $k = 60$ neighbors and the original training set. The rating will be replaced if $|r_{u,i} - r^*_{u,i}| > \delta$. Experiments were applied on several parameter variation options which were named global-pv, user-based-pv, and item-based-pv. The results revealed improvement in MAE and F1 and the results were compared with [6] and [10] which are also NNM protocols, and two other algorithms that target malicious noise.

Afterward, Castro et al. [25] in a study in 2016 targeted dealing with NN under a different recommendation approach known as the group recommendation systems (GRSs). GRSs represent variations of the normal recommender strategies where individual recommendations or preferences are aggregated to form personalized recommendations for a group of users (grouping strategies) [26]. The authors argued how GRSs employ explicit ratings nevertheless possess varying levels of information in their datasets and therefore are susceptible to NN that biases the recommendations. In this work, the core algorithm of [9] was used and modified to account for group preferences as part of the variations introduced for local data (the preferences belonging to the group members) and global data (the preferences belonging to all the users in the entire dataset). The results showed how NNM of the group ratings provides slight improvements to the group recommendation performance, while when applied to the entire dataset, it increases the performance of the GRS. Furthermore, the authors demonstrated how their hybrid approach that aggregated a cascade of both the global and local approaches that manage NN (first at the global level - entire dataset - and then at the local level - group ratings) had superior performance results.

In a parallel study by Yera et al. [27], the same authors argued how all the currently NNM solutions are unable to properly manage the inherent uncertainty and vagueness of customers' preferences. Accordingly, they proposed a novel fuzzy method to address this issue and improve, yet as well, the recommendation accuracy of recommenders. They added that the problem with previous approaches of NNM was that they solely represented and managed inherent rating uncertainties by means of crisp values which implied obvious lack in robustness. The authors added that the previously proposed approaches like their own works in [12], [9] etc. are not flexible

and robust enough to deal with the uncertainty and vagueness of both the ratings and the NN. In this proposal, the same workflow was adapted for the users in a recommender's dataset which is: profiling (instead, using fuzzy sets this time), noise detection, and noise correction. The steps of the approach are summarized as follows:

(1) Fuzzy profiling: obtain the fuzzy profiles of users, items, and ratings.
(2) Noise detection: apply a noise classification process on the previous profiles.
(3) Noise correction: noisy ratings are processed if needed.

It is comparatively explicit this approach is the exact replica of their previous technique [9] in terms of flow and logic, however, the particular difference, in this case, is that the recent method was implemented using fuzzy tools. The authors compared it to [6],[10] and [9], and in almost all cases, the MAE and F1 measures revealed better results with fuzzy profiling.

Latha et al. [28] proposed an approach that assigns lesser popularity scores to users who are not providing good ratings for exceedingly desired items. Users with a popularity score of less than a certain threshold are identified as noisy users. The steps of the approach are summarized as follows:

(1) Identify popular items in a dataset using the random walk approach.
(2) Assign popularity scores to the users based on their ratings of popular items.
(3) Identify noisy users and discard them from the training set.

The popularity score of a user is calculated based on equation: $popularity\_score_u = \frac{|PI_u|}{|PI|} \times log\frac{|I_u|}{|PI_u|}$, where PI is the set of popular items, PI_u is the set of popular items rated by user u and I_u is the set of items rated by user u. The first component of the equitation considers how the user rates popular items while the second checks whether said user also rates unpopular items. Compared with only [10] and [6], the results of this approach showed better MAE, RMSE and F1 metrics results.

In a more recent study in 2017 by Castro et al. [12], preceded by a survey on fuzzy tools in RSs [29], the authors combined their ideas of NNM that were presented in [25] and [27] and proposed an NNM for GRSs based on fuzzy tools. The approach follows the exact same approach in [27] were they compared the new NNM (for GRSs) using fuzzy tools with NNM (also for GRs) using crisp values of [25]. They used only MAE which ultimately resulted in improvements in the majority of evaluation scenarios and while few groups exhibited decay in the recommendation quality.

Choudhary et al. [30] aimed to handle the issue of NNM in multi-criteria recommendation systems (MCRS) with nothing but the ratings of users in recommenders' datasets. A MCRS is a technique that provides recommendations by modeling a user's utility for an item as a vector of ratings along with several criteria [31]. The authors asserted how all the previous works on NNM up until this point had been done on overall ratings based on a sole criterion. The approach they followed to deal with noise in the datasets was the exact same approach proposed in [9]:

(1) Classification of ratings users and items: The authors used the approach of [9] (Figure 2).
(2) Detection of noisy ratings.
(3) Noisy rating correction.

The authors did not contribute anything to NNM; they merely used the approach in [9] and supplied the noise-free dataset to a multi-criteria recommender approach.

The study by Bag et al. [15] came as an improved attempt to the series of proposals [9][27] and [12]. It was the first to approach *sparsity* [32] as a major challenge in the whole NNM paths, and the authors asserted that removing NN can significantly amplify the issue of data sparsity. Formerly, the issue of sparsity was touched upon very briefly in [21], however, there were no systematic approaches that addressed the problem when dealing with NN. They merely stated the aggregate number of noisy ratings that the noise correction algorithm eliminated from the training set. The method they proposed used the same approach that was presented in [9]. They grouped ratings according to Table 2 with a slight modification on the noise correction methodology.

Rather than predicting a new rating in the correction phase, they utilized the concept of self-contradiction which replaced the user's rating with the classified group threshold (Weak, Average, or Strong) when it was spotted to be self-contradicting. As for the sparsity issue, they integrated the Bhattacharyya similarity measure in the UB-CF approach [33]. The results show improved MAE and RMSE values and a better time complexity compared to [9] since they eliminated the re-prediction step that was used to correct noisy ratings.

In a very recent proposal [34], Yera et. al presented extended research on two of their previous ideas. In it, they connected the two fuzzy models for NNM previously proposed in [27] (RRs for individuals) and [12] (RSs for groups) that guaranteed robust modeling for the uncertainty associated to the user profiles (i.e. NN in datasets). In their experiments, they compared NN-Crisp [9] with NN-FT and NNMG-Crisp [25] with NNMG-FT. Put differently, the authors provided a deeper study on their previously proposed classical NNM approaches for individual and group recommenders [9] and [25], and their proposed fuzzy approaches [27] and [12].

3.3 The Preference-dependent Path

All the researches that depend on external data that aren't typically available in recommender systems' dataset fall under this path. Till now, there are only two interconnected proposals that directly address NNM. In researches [35] and [36], Pham et al. proposed a matching method between the user preferences and the dataset items in an attempt to determine whether the ratings of a certain item are actually reliable. Through two manually constructed small datasets (portions of MovieLens and Netflix joined with IMDB), the authors used item attributes to detect inconsistencies in tastes by comparing the actual preference value provided by the user with the rating predicted by a model. The inconsistencies are then corrected using preferences provided by expert users that exhibit overlapping tastes with the target user.

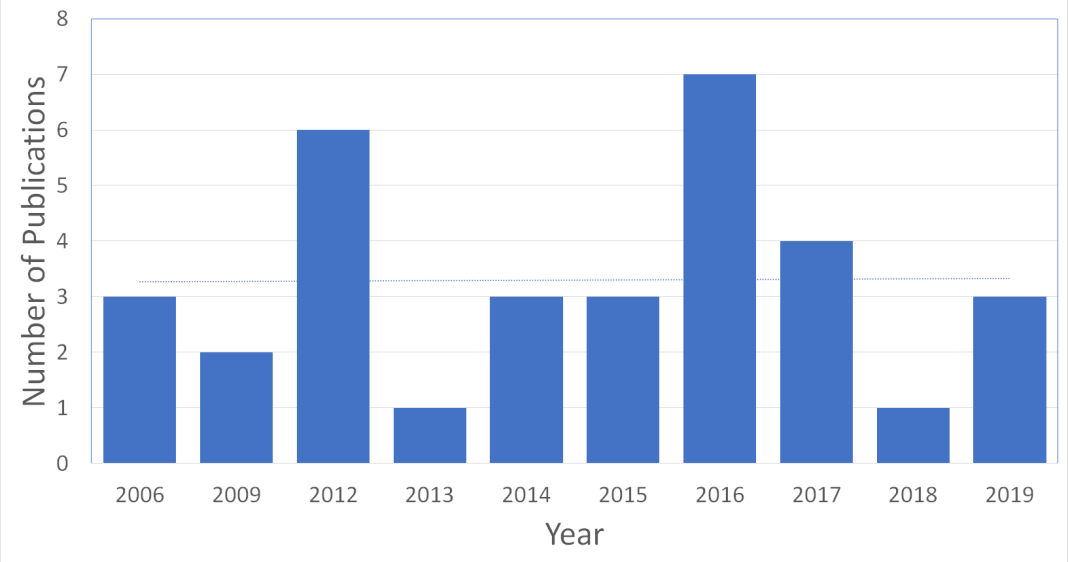

Fig. 3. Number of publications on natural noise since it's inception in 2006

Table 3. Timeline of the classical natural noise management path after the introduction of the concept in [6]

2006	**[6]** The first work that tackles natural noise in the recommender system datasets is proposed. Noise is defined as the consistency of a certain rating compared to a predicted value for it using a certain recommender algorithm
2012	**[10]** A comparison with [6] is put forward. Unlike [6], the authors define the noise in terms of a user-profile as a quadratic optimization problem
2014/15	**[24][9][27][28]** In [24] and [9], the authors introduce a noise detection and correction method. It is based on user profile classification without any additional information to the dataset. Research [27] introduces the same strategy as in [9] but with fuzzy profiling. Study [28] introduces a noisy user detection approach by a popularity score method
2016/17	**[25][12][30]** Study [9] is managed in proposal [25] to implement a natural noise management (NNM) algorithm for group recommender systems (GRSs). It is improved more significantly in [12] following a similar logic as [27] and [25] to propose an NNM based on fuzzy tools fuzzy for GRSs. Research [30] adopts the approach in [9] to employ it to a multi-criteria recommender
2019	**[15][34]** A modified user-classification and noise correction approach that takes dataset sparsity into consideration is proposed by [15] based on [9]. Research [34] combines [27] and [12] in an extended study for fuzzy tools with NNM

Table 4. Specification details of natural noise management approaches

Study	Target	NNM	Target	Category	Recommenders	Datasets	Evaluation Method
[12]	G	D & C	Ratings	Classical	UB-CB	Ml-100k, Nf-Tiny	MAE
[10]	I	D & R	Users	Classical	UB-CF	Ml-100k, BC, EM	Prec., Rec.
[9]	I	D & C	Ratings	Classical	IB-CF, MF	Ml-100k, MT	MAE, F1
[6]	I	D & R	Ratings	Classical	UB-CF	Ml-100k, EM	MAE, Cov.
[21]	I	D & C	Ratings	Pref-dependent	IB, MF	Ml-Latest-Full	RMSE, Prec.
[27]	I	D & C	Ratings	Classical	UB-CF, IB-CF, SO	Ml-100k, MT, Nf-Tiny	MAE, F1
[25]	G	D & C	Ratings	Classical	UB-CF, IB-CF	Ml-100k, Nf-Tiny	MAE
[34]	I	D & C	Ratings	Classical	UB-CF, IB-CF, SO	Ml-100k, Nf-Tiny	MAE, F1
[28]	I	D & R	Users	Classical	IB-CF	Ml-100k, Jester	MAE, RMSE, F1
[15]	I	D & C	Ratings	Classical	UB-CF	Ml-1m	MAE, RMSE, F1, Prec., Rec.
[35]	I	D & C	Ratings	Pref-dependent	None	ml*, nf*	RMSE

Fig. 4. The studies distributed over the datasets and the metrics used

3.4 Natural Noise vs. Malicious Noise

The preceding delineated works include all the proposals that targeted NNM in recommender datasets. What makes natural noise special is that it results from a user attitude and even in its exceptional cases can go unnoticed by a recommender [37]. When inconsistent and misleading behavioral patterns stack up in the dataset, they cause the recommender to learn from those anomalies, rather than side-stepping them, and provide recommendations that are inspired by them. Since it's challenging for a recommender to detect natural noise - mainly since it does not portray any suspicious or defined pattern - the results that are laid out to users end up being biased and, in many cases, highly inaccurate. This can be detrimental to systems that heavily rely on recommenders for sales, especially in commercial online stores for example.

Contrary to NN, malicious noise results from numerous forms of attacks carried out on online applications that are typically powered by diverse types of RSs, and it has witnessed much of the research attention in the past few years [12]. Attack patterns are most of the times defined, and the system can be trained to filter out the anomaly signatures. Further, malicious noise is primarily the result of an external adversary that aims at carving the recommender's output rather than it being generated in the dataset and by the system itself like in the case of natural noise.

Due to this vast difference between natural noise and malicious noise, the methods [3] that had successfully worked on managing and eliminating malicious attacks (Probe, Bandwagon, Segment, Crawling, etc.) proved ineffective and inapplicable in the natural noise case [9][18]. As demonstrated in the previous sections, the peculiar form of NN - mainly composed of inconsistencies resulting from user behavior - makes it hard to define a pattern for it as opposed to the specific outlined nature of malicious attacks. That's why in most cases, the studies throughout the path resorted to classification method in attempts to categorize both users and their interactions.

4 STATISTICAL ANALYSIS OF THE PATHS

In this section, we curate substantial information about the previously discussed studies by providing some statistical figures. This allows a more thorough understanding of how NN was gradually approached by researchers. First, the number of publications over the years is presented in Figure 3 where it can be noticed that the pace of publications was almost steady overall between 2006 and 2009, with peaks arising in the years 2012 and 2016. The first spike mainly included proposals that attempted to reinforce the initial arguments on the concept of NN (see Tables 1 and 3), while the second included a series of researches that introduced solving the NN issue with a marginally distinctive form of recommenders (GRSs), and predominantly employed fuzzy tools for NNM. This steady pace can be somewhat attributed to the intricate nature of NN compared to malicious noise, and the fact that it is more recent than malicious noise, which in turn averaged a different publication pattern over the years [3] compared to its counterpart. The study of NN was influenced by a number of works such as [11], that conferred about rating inconsistencies in datasets, and [7], that deeply impacted the entire logic barrier path. All researches are still hinting that the study of NN is fairly new and it is explicit that the field is yet open to many advances and proposals to address the missing gaps that will be discussed in the subsequent section. The NNM approaches, accompanied by their comprehensive specifications, are summarized in Table 4.

Next, we plot the percentage of issues in each of the three categorized research tracks on NNM in Figure 5 (left). Classic NNM has the lead with 37% of the total publications registering the highest publication rate, followed by the Magic Barrier with 33% of the total publications. The classic path includes more discrete formulations that defined easy and effective ways to deal with natural noise in the datasets, and further, both the Classic NNM and the Magic Barrier proposals chiefly depend on the exceedingly common recommender datasets (user-item matrix type) that typically do not require external features. It's reasonably expected that the preference-dependant NNM would obtain a more reduced rate as the dependency of the proposed algorithms and solutions was specific to certain features of datasets that aren't universally available nor conveniently accessible by all researchers in

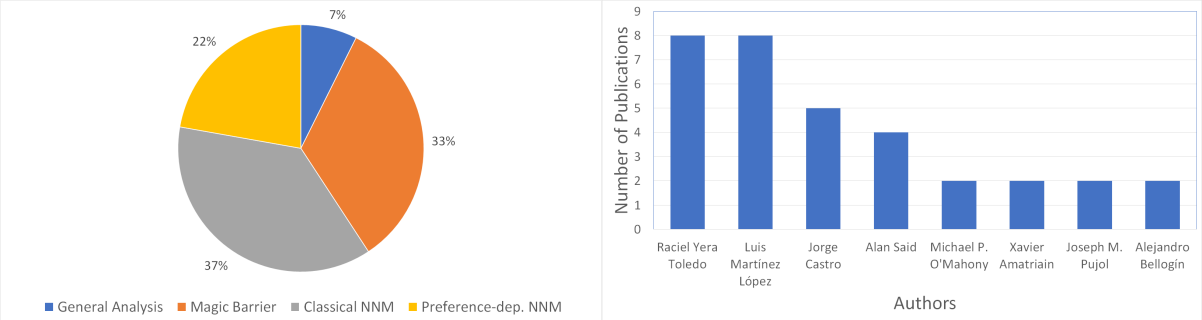

Fig. 5. Percentages of the natural noise research directions (left) and the major researchers in the natural noise management field (right)

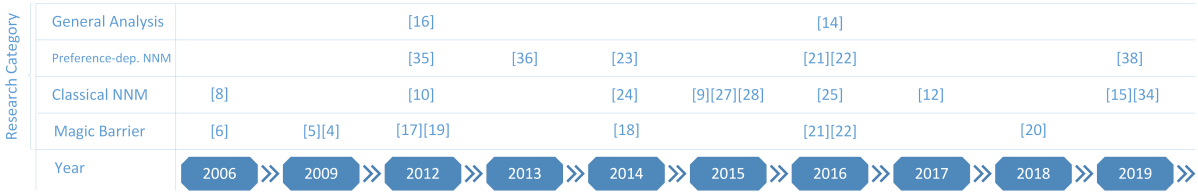

Fig. 6. A detailed timeline of the fundamental studies in the natural noise field

the field. The general analysis, amounting to 7% of the total publications, included studies that briefly discussed the problems of NN in recommenders' datasets without proposing any solution or method to deal with it. For conciseness, the most significant publications across the three tracks are laid out in the comprehensive timeline of Figure 6, starting with the inauguration of the concepts of NN and the Magic Barrier in year 2006. This timeline clearly presents the specific and most influential publications that contributed to the research peaks of 2012 and 2016 and offers a general idea of the NN track flow over the years. More specifications for the most famous NNM algorithms have been gathered and grouped together in an even more comprehensive manner in Table 4. The data in this table indicates how every algorithm targets recommender datasets (whether it's for group or standard recommenders), the NNM method used and whether it's built for correcting the noise after detection or merely removing it (D & C, D & R), and the path class of the approach. Further, the recommender type, datasets used, and evaluation metrics implemented to measure the corresponding study have also been appended to the table. Additionally, It is apparent from Table 4 that the most used recommender type in the NNM paths was CF in both its forms, UB and IB, registering an appearance in almost all the proposals. Lastly, for researchers who aspire to reinforce the works on NNM, it is always beneficial to have an idea about the individuals that were contributing generously to the field up until this point; Figure 5 (right) plots the authors versus their respective number of publications.

On a more specific level, the datasets used in the studies across the NN track are detailed in Table 5. They basically constitute the famous open-source recommender datasets that were utilized in the Classic NNM and most of the Magic Barrier proposals, with Netflix Tiny now being pulled out from the public and EachMovie retired. As previously mentioned, the data that was used in the Preference-dependent track is not publically available and can be exceedingly delicate to reproduce in many cases. We also preview the datasets for each corresponding publication in Table 4 and present the metrics and datasets used in Figure 4. The Ml-100k dataset, arguably one of the most famous datasets in the overall recommender system research field [38], registered the

Table 5. The details of the datasets used in the natural noise approaches

Name	Category	User × Item	Scale Range	Step	Total Ratings	Status
MovieLens Latest Full	Movies	23, 0000 × 30, 000	[1,5]	0.5	21, 000, 000	Available
Jester	Jokes	73, 421 × 100	[-10,10]	1	4, 100, 000	Available
EachMovie	Movies	72, 916 × 1, 628	[1,6]	1	2, 811, 983	Retired (04)
Book-Crossing	Books	278, 858 × 271, 3790	[1,10]	1	1, 149, 780	Available
MovieLens 1M	Movies	6040 × 3952	[1,5]	1	1, 000, 209	Available
MovieTweetings(2013)	Movies	21, 018 × 12, 569	[0,10]	1	140, 000	Available
MovieLens 100k	Movies	943 × 1682	[1,5]	1	100, 000	Available
Netflix Tiny	Movies	4, 427 × 10, 000	[1,5]	1	56, 136	NA

* NA or Retired means that the datasets are not officially available anymore.

Table 6. Accuracy results for each natural noise mechanism across four different datasets

		Natural Noise Management Mechanism					
				Random			
Dataset	N/Metric	Original	NN Filter	Random-N	Highest-N	Middle-N	Lowest-N
Ml-Latest-Small	N	0	10655	10655	10000	10000	10000
	MAE	0.6937	0.6161	0.6880~0.7030	0.6543	0.6926	**0.5640**
	RMSE	0.9077	0.8216	0.8960~0.9214	0.8537	0.9205	**0.7077**
Ml-100k	N	0	12071	12071	12000	12000	12000
	MAE	0.7575	0.6791	0.7522~0.7744	0.7238	0.7732	**0.6285**
	RMSE	0.9607	0.8726	0.9516~0.9789	0.9180	0.9892	**0.7766**
Ml-1m	N	0	128916	128916	120000	120000	120000
	MAE	0.7473	0.6580	0.7475~0.7529	0.7057	0.7576	**0.6133**
	RMSE	0.9392	0.8401	0.9394~0.9454	0.8843	0.9647	**0.7521**
Hetrec-Ml	N	0	92634	92634	92000	92000	92000
	MAE	0.6254	0.5626	0.6230~0.6280	0.5799	0.6332	**0.4699**
	RMSE	0.8160	0.7465	0.8140~0.8202	0.7521	0.8382	**0.5858**

highest usage percentage (36%), followed by Nf-Tiny (23%). Interestingly, those two datasets are considerably small (100k vs. 56k ratings) compared to their peers in the list of Table 5. It's also surprising how the choice of datasets for testing NNM algorithms (through the metrics indicated in Table 4) in the Classic NNM and Magic Barrier was not done in a pre-calculated manner where it seemed that the authors merely chose the most famous recommender datasets to test novel approaches; this begs the question, are the post-NNM recorded accuracy improvements dataset-dependent? (More on the gaps and issues in Section 5.3)

5 ANALYSIS AND HYPOTHESES TESTING

Throughout the NNM paths that were detailed in Section 3, it has become apparent now that the common accuracy metrics, especially MAE and RMSE (Table 4), portrayed a significant role in deciding whether an NNM method

improved a recommender's performance or not. In this section, we evaluate the two hypotheses introduced in the beginning of this work by conducting two experiments. The first one introduces the concept of randomness and helps us interpret the relationship between the noise predicted by NNM measures and the metrics used to test the general performance after said noise has been managed (discarding it or correcting it). The second experiment allows us to extend on the first notion and checks whether the accuracy metrics used in the NNM publications provide consistent results. The recommender we employed in the two tests is a user-based CF algorithm with a cosine similarity measure, one out of many measured for KNN-based algorithms [39].

5.1 A Randomness-based Natural Noise Method

This experiment was uniquely designed to test the first proposed hypothesis, i.e., whether a recommender's accuracy performance can be positively affected by arbitrary ratings removal from a target recommender dataset; should this claim hold, it would prove that the NNM approaches that traditionally presented accuracy improvement outcomes to show how a method is more effective than the other require radical revisions. That said, If a random-based straightforward process such as this could indeed achieve a similar performance to an NNM technique, then the evaluation approaches used in the NN field need to be adequately addressed; this does not mean that the NNM proposals are wholly wrong, but it evidently would signify that the foundation that the NNM path is basing on might not be totally correct. To investigate this, our experiment ran on four varied datasets, Ml-Latest-Small, Ml-100k, Ml-1m and Hetrec-Ml, that were selected based on the popularity in the NNM field (Table 4). For each simulation round, we implemented the following data removal schemes which are chiefly based on randomness and logical intuition:

(1) $Random - N$: Remove random N ratings.
(2) $Lowest - N$: Remove the lowest N ratings based on the corresponding rating scale.
(3) $Highest - N$: Remove the highest N ratings.
(4) $Middle - N$: Remove the ratings that are in the middle of the rating scale.

The aggregate number of ratings (N) to be removed in every mechanism differs between datasets and was determined through the use of the most commonly used NNM algorithm (Toledo et al. [9]) which we employed in this experiment. In the case of the first scheme ($Random - N$), we typically had to re-train the data every time the arbitrary N ratings were removed since unlike in the other random methods, the N ratings will vary and we need to measure the accuracy each time a different proportion from the dataset was removed to prevent biased results (in the other methods, N is typically constant which implies that the accuracy results will be constant under the given conditions of the recommender). Accordingly, the simulation was conducted for a total of 150 iterations in the case of $Random - N$. The final accuracy outputs of the aforementioned mechanisms were compared to the results from the NNM protocol and the original dataset without any noise management or ratings removal. The results of the test are presented in Table 6 which details the the variations of MAE and RMSE in each round for every dataset. The table also shows the value of N for every method and presents the minimum and maximum of the 150 $Random - N$ iterations. For a clearer presentation, Figure 7 depicts the $Random - N$ plot of the iterations where in each round a value of $N = 10,655$ (Table 6) arbitrary ratings are being eliminated from the dataset. What's intriguing in the results is that the $Lowest - N$ scheme showed the best accuracy outcome across all four datasets. Surprisingly, $Random - N$ showed very acceptable scores that fluctuated between the $Original$ and $Middle - N$ schemes, sometimes achieving better output than both such as in Ml-Latest-Small (0.6880 - 0.8960), Ml-100k (0.7522, 0.9516) and Hetrec-Ml (0.6230, 0.8140). The $Middle - N$ almost always had the worst MAE and RMSE results across all the datasets. The NNM scheme of [9] always came in the middle, registering slightly better results than the $Highest - N$ in all the datasets.

It is evident now that utilizing the same accuracy metrics that were employed to evaluate the effectiveness of NNM algorithms on RSs in the previously discussed paths is controversial. Our purely random-based trial

resulted in comparable significant MAE and RMSE improvements especially with the $Lowest - N$ scheme with relatively acceptable results for the others. This disproves the first hypothesis and clearly validates the concept that the evaluation methods for NNM that are used to assess the performance and show that one is better than the other are flawed and require radical revisions. To a great extent, this also validates the fundamental opinion examined by Herlock et al. [7] which was touched upon in the introduction of this work: algorithms should be measured in accordance with how well they can communicate their reasoning to users, or with how little data they can yield authentic recommendations, we require new metrics to evaluate those new algorithms and not merely rely on improvements that show scant enhancements in accuracy (such as MAE and RMSE) and label one better than the other.

5.2 Accuracy Consistency Test

This experiment was designed specifically to target the second proposed hypothesis and it extends on the idea of the previous randomness notion and investigates the metrics used to test the effectiveness of the NNM methods (Table 4). The collective results from the earlier $Random - N$ simulation conducted are presented in Figure 8 for two datasets, Ml-Latest-Small and Ml-100k. Figure 8 (right) depicts the values of MAE and RMSE that were previously shown in a different format in Figure 7. Analyzing those accuracy plots, it is apparent that there are plenty of cases throughout the 150 iterations (from the $Random - N$ scheme experiment) were the measurements portrayed conflicting results and this is the case for all of the four datasets used in the our experiments. One notable example (marked in Figure 8) shows how in two consecutive runs there was a 1.5% decrease in MAE versus a 2.5% increase in RMSE. This ultimately refutes the second hypothesis and signifies how the metrics adopted to test the success of NNM measure are not a reliable performance measure, and definitely should not be the sole tests upon which the success of an NNM proposal is being evaluated.

5.3 Gaps in the NNM Paths

The outcome of the two previous experiments and the analysis throughout the sections of this work have shown that the NNM field definitely lacks a well-defined consistent approach, and further, it introduces many potential enhancement opportunities to ultimately become effective on an RS. From the randomness approach that triggered oddly comparable MAE and RMSE results to one of the best performing NNM algorithms (Section 5.1) to the inconsistency of the used evaluation methods on NNM (Section 5.2), this section summarizes all the possible gaps and weakness in the previously discussed publications in the NNM path. Those gaps are grouped into five primary categories.

5.3.1 The Natural Noise Misconception and Inconsistency. Some proposals confuse the definition of NN in datasets and provide distinct explanations and implementation approaches such as that of Tong et al. [40]. In their introduction, the authors discussed that what one user considers as a mediocre rating (e.g. 1 out of 5 stars) compared to another who rates 3 out of 5 as bad, is NN. In essence, this is not the appropriate definition of NN in [6] and which was thoroughly discussed in the introduction of this study, however, it is merely an interpretation of users' unique standards and rating baselines. This is an effect that does not occur solely across individuals but across varied cultures as well, some countries for instance are more harsh with their ratings than others. Primitive CF methods have hitherto attempted to normalize these differences; one example would be the adjusted cosine similarity measure (ACOS) [13].

5.3.2 The Magic Barrier - Accuracy Barrier Conflict. One of the most significant gaps in the Accuracy Barrier path is conceivably the RS evaluation methodology and the erroneous interpretation of the Magic Barrier of Herlocker et al. [7]. In [7], the authors concluded that the accuracy metrics results of various algorithms clearly suggest that algorithm improvements in CF systems may come from divergent directions rather than just continued

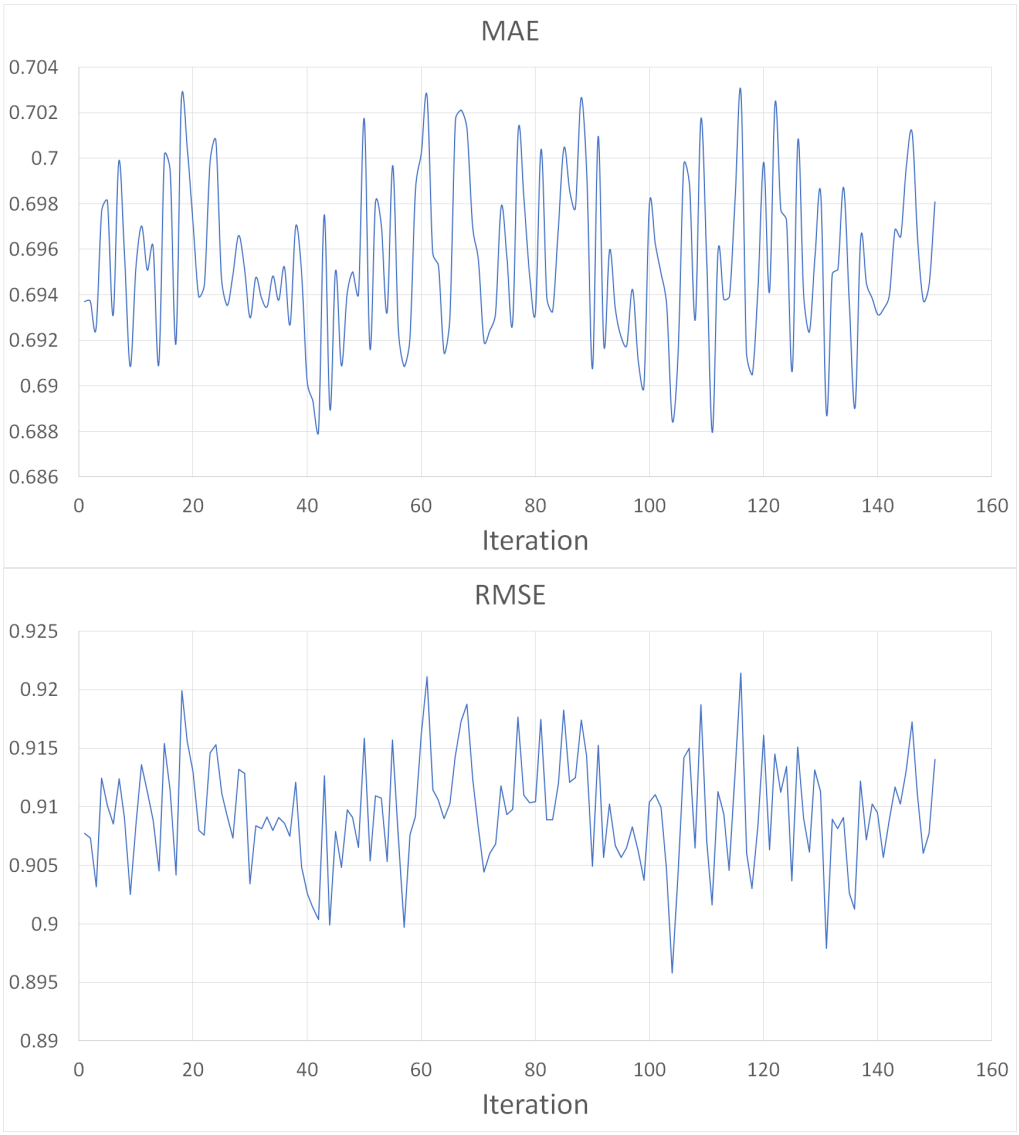

Fig. 7. Example accuracy results of the Random-N mechanism applied to Ml-Latest-Small

improvements in MAE or RMSE; it is possible that the most efficient algorithms should be measured in accordance with how well they can communicate their reasoning to users, or with how little data they can yield accurate recommendations. Lastly, the work hypothesized that modern metrics need to be developed to evaluate those new algorithms. The authors, in this case, are basically debating the effectiveness of the remarkably minute variations in MAE or RMSE (sometimes in the order of 0.01) when various RSs are evaluated against each other in terms of accuracy. Therefore, it is evident that the purpose of the revision in [7] was never intended to introduce the Magic Barrier as the definition that was taken on by the first path (Accuracy Barrier) of the NNM approaches

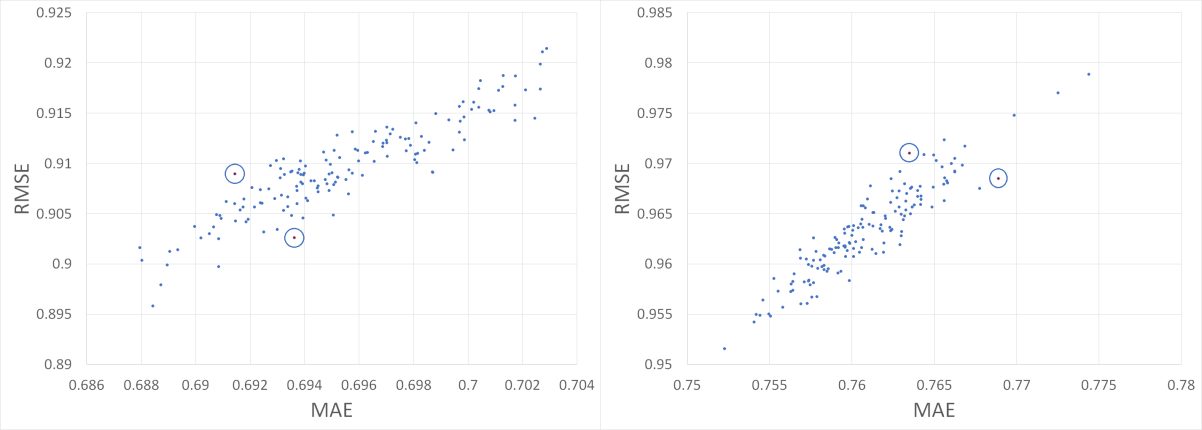

Fig. 8. MAE vs RMSE for Ml-Latest-Small (left) and Ml-100k (right)

(the point at which the performance and accuracy of an algorithm cannot be enhanced due to noise in the data), but rather, it was abundantly evident that new metrics should be introduced in order to better understand users in recommender system datasets.

The Accuracy Barrier is based on the item opinions gathered from users who have already rated them at least once. Said et al. [17] and [19] explicitly state how use views on previously rated items and their initial ratings by the same users differ conceptually, however, they were essentially handled as being the same when the authors estimated the Accuracy Barrier. This proves that the Accuracy Barrier path is relying on re-ratings from the users and on an accuracy approach (mainly RMSE) to estimate the limit beyond which there can be no added improvement to the recommender's accuracy. There are two key weaknesses in this case; the first would be having to provide other opinions from users in a dataset to calculate the maximum performance of the system, knowing that those opinions and their previous values differ conceptually. The second concern is the RMSE function that was used to estimate the ultimate performance of a recommender. Is the Accuracy Barrier a proper measure of a recommender's maximum effectiveness knowing that RMSE might not be measuring the true accuracy of a recommender (Section 5.2)?

The Accuracy Barrier track ultimately claims that a good "Magic Barrier" estimate is useful for assessing the quality of recommendations and for revealing room for improvements, however, the studies throughout the path do not propose or implement any methods that experimentally disclose those improvements nor how to improve a system after calculating the Accuracy Barrier. Markedly, Said et al. [19] concluded that recommenders with prediction accuracy close to the estimated Accuracy Barrier can be regarded as "optimal systems". This conclusion also presents similar issues to what was raised in our previous point. An "optimal recommender output" basically represents a very general term especially when evaluating the RS's output is done through employing MAE and RMSE metrics only.

5.3.3 The Accuracy Barrier Weaknesses. As previously mentioned, calculating the Accuracy Barrier strictly depends on a primary phase which is gathering viewpoints from users on items they have formerly seen and rated. This proposal is eminent in all the works throughout the path such as in Said et al. [19], where the authors proposed that a real-world recommender system should regularly interact with users by polling opinions about items they have previously graded allowing them the opportunity to audit their own performance and take measures to improve the recommendation engine where appropriate. This act poses a drawback in real-world

applications since users have to provide a second opinion on many items. Those users might have forgotten why a certain item received their dear appreciation in the first place for instance, which would cause their second rating to become inaccurate; this begs the question, are the second user views provided in the MoviePilot experiments [17][19] reliable and noise-free? Further, will those users genuinely care about enhancing the accuracy of an RS for a certain platform so much as to spend valuable time providing an accurate second opinion on products?

The authors pointed out in a later study [20] that one major drawback of measuring and comparing the performance using only static, previously collected test data, is that user behavior in the data is not always reliable. That said, how will the second round of ratings collected from some users differ from the first? It seems that the Accuracy Barrier does not open up doors for performance improvements but introduces yet another rating process that is equally (if not more) prone to NN. The rating elicitation process is intrinsically susceptible to noise due to several reasons touched upon in the introduction of this journal. In addition, subsequent studies should tackle the following missing points:

(1) A concrete definition of "re-rating movies again after a certain period of time".
(2) A measure for an adequate amount of time to identify consistencies of ratings assuming this method is effective, and accuracy remains our target.

5.3.4 Accuracy Evaluation. As touched upon earlier in the introduction and throughout the experiment discussion there is a certain flawed assumption about the performance of RSs that requires further analysis in the Accuracy Barrier. All the studies [18][20][19][17] define Herlocker's version of the Magic Barrier [7] as the level of prediction accuracy that an RS can attain with the lowest possible error. Their version (Accuracy Barrier) reveals whether there is room for additional meaningful accuracy improvement or that any further enhancement is typically meaningless. This approach contradicts with the primary purpose of RSs which was revealed with the introduction of the notion of the Magic Barrier in [7]. The outcomes of our experiment in Sections 5.1 and 5.2 revealed how the metrics applied to evaluate NNM algorithms of RSs result in conflicting outputs, and attempting to tackle the concept of an Accuracy Barrier for performance evaluation and improvement through the use of MAE and RMSE is counter-intuitive. Those two following points summarize the problems with accuracy and evaluation, and should be considered in the future proposals on the subject:

(1) Accuracy metrics that were employed to assess the performance of NNM methods (Table 4) should be re-visited to align with [7].
(2) Other factors like diversity or serendipity [8][41] should be accounted for. The accuracy evaluations are still pre-dominantly used and relied on, even in very recent proposals [34; 42] on NNM. Serendipity and accuracy are significantly different in their nature and increasing one leads to the decrease in the other [8]. Said et al. [22] propose a method to calculate the similarity between items and remove the ratings of those from a user profile that are dissimilar to each other. This can result in a sheer elimination of any form of essential serendipitous results that might occur in the recommender's output.

Throughout the fuzzy-inspired proposal that emerged in 2015 on NNM [27] [34], the authors suggested that the issue of natural noise in RSs is similar to that in fields independent from RSs such as [43] and [44]. Consequently, they were inspired to propose a user clustering mechanism based on fuzzy profiling. The accuracy problems discussed above remain in this case since the performance of NNM methods is still being assessed by the level of MAE and RMSE scores. Further to that, the fields that inspired the fuzzy NNM proposals [43] (Noise Reduction Methods for Brain MRI Images) and [44] (scenario of imbalanced datasets) are intrinsically contrasting compared to RSs and their datasets. Recommendation engines' datasets are unique and in most cases contain explicit ratings of users while those of MRI images maintain an unrelated format and a peculiar application altogether compared to recommendation engines.

5.3.5 General Problems with the Approaches. Some general essential factors that must be taken into consideration in further proposals that might enhance the NNM paths are:

(1) A measure of frequency of NN correction on datasets? Is there a certain limit beyond which applying NN becomes useless or counter-efficient?
(2) Almost all studies across the three paths create subsets from the datasets in that are shown in the Table 5. This shows that there are high time complexities of the algorithms with the noise detection and correction methodologies that need to be properly addresses and eventually benchmarked.
(3) As seen in the beginning of this section, accuracy might not be the most suitable measure, therefore, an NNM approach that is computationally demanding and produces a superior MAE or RMSE result in the end should be reviewed on different levels before ranking it as an optimal method for NNM compared to the others in the path.
(4) Most published studies in the NNM paths use CF recommender algorithms to evaluate their approaches on NN. Will a good-performing NNM method still be effective when the recommender algorithm is inevitably modified?
(5) There are other problems with CF approaches that researchers usually overlook. For example, Bag et al. [15] argue that removing noise from the dataset amplifies the sparsity issue in them, and this is conceptually true. They continue adding that Toledo et al. [9] who used PCC to predict ratings as replacement for noise have increased the issue through adopting a flawed correction measure as the environment might be sparse to start with, and PCC performs poorly in scarce environments [15]. PCC does indeed perform poorly in terms of accuracy for users who maintain a limited amount of ratings, however, the authors never discussed critical factors like the sum of ratings the users that were given re-ratings had, the neighborhood size, the method to decide upon a correct neighborhood size, etc. In [15], the authors chose to employ another similarity measure for noise correction, ruling out PCC as an option, and still not providing any information about the critical variables of an RS, namely the neighborhood size, the number of items those noisy individuals had, their respective contribution to their neighborhood, etc.

6 CONCLUSION

Implementing an effective and agile natural noise management algorithm for recommender systems' datasets is challenging due to various parameters that ought to be taken into consideration, especially in the evaluation process. Evidently, there has been no attempt to synthesize what is traditionally known about the performance evaluation of recommender systems and natural noise management, nor to systematically recognize the implications of evaluating them for numerous tasks and diverse contexts while testing the performance of a natural noise technique. Throughout this comprehensive study, we surveyed and categorized all the natural noise handling algorithms starting from their inauguration in 2006. In addition, we carefully introduced empirical results from two hypotheses that provided critical insight on the consistency of the evaluation methods used in the proposed noise management techniques. The first experiment illustrated how randomness could in fact achieved comparable outcomes to one of the most conventional mechanisms while the second proved that the metrics that were employed to test those techniques and rank one better than the other are typically displaying inconsistent and unreliable results. We hope this article will naturally increase the awareness of the evaluation of recommenders especially in the natural noise management field and encourage the development of more standardized natural noise methods evaluated by measures beyond traditional accuracy.

As seen in the previous section, there are many gaps paths that constitute the natural noise management field and which undoubtedly require considerable attention in the future. The potential problems we set to address include the development of proper evaluation methods that can be broadly used with any recommender technique and serve us to better assess the true effectiveness of a recommender devoid of inconsistencies. In addition to

that, natural noise lacks proper development in terms of the type of datasets it is applied to. An effective noise management approach needs to be scalable and properly work with diverse kinds of datasets irrespective of the recommendation algorithm employed. External data that administrators must retrieve from customers to implement a certain noise management approach remains an inadequate solution to the problem.

REFERENCES

[1] Charu C. Aggarwal. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition, 2016.

[2] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to Recommender Systems Handbook*, pages 1–35. Springer US, Boston, MA, 2011.

[3] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Polat. Shilling attacks against recommender systems: A comprehensive survey. *Artif. Intell. Rev.*, 42(4):767–799, December 2014.

[4] Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. I like it... i like it not: Evaluating user ratings noise in recommender systems. In Geert-Jan Houben, Gord McCalla, Fabio Pianesi, and Massimo Zancanaro, editors, *User Modeling, Adaptation, and Personalization*, pages 247–258, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[5] Xavier Amatriain, Josep Pujol, Nava Tintarev, and Nuria Oliver. Rate it again: Increasing recommendation accuracy by user re-rating. In *Book*, pages 173–180, 01 2009.

[6] Michael O'Mahony, Neil Hurley, and Guenole Silvestre. Detecting noise in recommender system databases. In *Collaborative Recommendations: Algorithms, Practical Challenges And Applications*, volume 2006, pages 109–115, 01 2006.

[7] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[8] Wissam Al Jurdi, Miriam El Khoury Badran, Chady Abou Jaoude, Jacques Bou Abdo, Jacques Demerjian, and Abdallah Makhoul. Serendipity-aware noise detection system for recommender systems. *IKE'19*, 2018.

[9] Raciel Yera Toledo, Yailé Mota, and Luis Martinez. Correcting noisy ratings in collaborative recommender systems. *Knowledge-Based Systems*, 76, 03 2015.

[10] Bin Li, Ling Chen, Xingquan Zhu, and Chengqi Zhang. Noisy but non-malicious user detection in social recommender systems. *World Wide Web*, 16, 11 2013.

[11] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 194–201, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[12] Jorge Castro Gallardo, Raciel Yera Toledo, and Luis Martinez. A fuzzy approach for natural noise management in group recommender systems. *Expert Systems with Applications*, 94, 11 2017.

[13] J. Bobadilla, F. Ortega, A. Hernando, and A. GutiéRrez. Recommender systems survey. *Know.-Based Syst.*, 46:109–132, July 2013.

[14] Luis Martinez, Jorge Castro Gallardo, and Raciel Yera Toledo. Managing natural noise in recommender systems. In *Theory and Practice of Natural Computing: 5th International Conference, TPNC 2016, Sendai, Japan, December 12-13, 2016, Proceedings*, pages 3–17, 12 2016.

[15] Sujoy Bag, Susanta Kumar, Anjali Awasthi, and Manoj Tiwari. A noise correction-based approach to support a recommender system in a highly sparse rating environment. *Decision Support Systems*, 118:46–57, 03 2019.

[16] Daniel Kluver, Tien T. Nguyen, Michael Ekstrand, Shilad Sen, and John Riedl. How many bits per rating? In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 99–106, New York, NY, USA, 2012. ACM.

[17] Alan Said, Brijnesh J. Jain, Sascha Narr, Till Plumbaum, Sahin Albayrak, and Christian Scheel. Estimating the magic barrier of recommender systems: A user study. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1061–1062, New York, NY, USA, 2012. ACM.

[18] Alejandro Bellogín, Alan Said, and Arjen P. de Vries. The magic barrier of recommender systems – no magic, just ratings. In Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben, editors, *User Modeling, Adaptation, and Personalization*, pages 25–36, Cham, 2014. Springer International Publishing.

[19] Alan Said, Brijnesh Jain, Sascha Narr, and Till Plumbaum. Users and noise: The magic barrier of recommender systems. In *User Modeling, Adaptation, and Personalization*, volume 7379, 07 2012.

[20] Alan Said and Alejandro Bellogín. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction*, 04 2018.

[21] Penghua Yu, Lanfen Lin, and Yuangang Yao. A novel framework to process the quantity and quality of user behavior data in recommender systems. In *Web-Age Information Management: 17th International Conference, WAIM 2016, Nanchang, China, June 3-5, 2016, Proceedings, Part I*, pages 231–243, 06 2016.

[22] Roberto Saia, Ludovico Boratto, and Salvatore Carta. A semantic approach to remove incoherent items from a user profile and improve the accuracy of a recommender system. *J. Intell. Inf. Syst.*, 47(1):111–134, August 2016.

[23] Roberto Saia, Ludovico Boratto, and Salvatore Carta. Semantic coherence-based user profile modeling in the recommender systems context. In *Book*, 10 2014.
[24] Raciel Yera Toledo, Luis Martinez, and Yaile Mota. Managing natural noise in collaborative recommender systems. In *Book*, pages 872–877, 06 2013.
[25] Jorge Castro Gallardo, Raciel Yera Toledo, and Luis Martinez. An empirical study of natural noise management in group recommendation systems. *Decision Support Systems*, 94:1–11, 02 2017.
[26] Toon Pessemier, Simon Dooms, and Luc Martens. Comparison of group recommendation algorithms. *Multimedia Tools Appl.*, 72(3):2497–2541, October 2014.
[27] Raciel Yera, Jorge Castro, and Luis Martínez. A fuzzy model for managing natural noise in recommender systems. *Appl. Soft Comput.*, 40(C):187–198, March 2016.
[28] R. Latha and R. Nadarajan. Ranking based approach for noise handling in recommender systems. In *Multimedia Communications, Services and Security: 8th International Conference, MCSS 2015, Kraków, Poland, November 24, 2015. Proceedings*, pages 46–58, 11 2015.
[29] Raciel Yera Toledo and Luis Martinez. Fuzzy tools in recommender systems: A survey. *International Journal of Computational Intelligence Systems*, 10:776 – 803, 03 2017.
[30] Priyankar Choudhary, Vibhor Kant, and Pragya Dwivedi. Handling natural noise in multi criteria recommender system utilizing effective similarity measure and particle swarm optimization. *Procedia Comput. Sci.*, 115(C):853–862, November 2017.
[31] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Recommender Systems Handbook*, volume 1-35, pages 1–35. Publisher, 10 2010.
[32] Zan Huang, Hsinchun Chen, and Daniel Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):116–142, January 2004.
[33] Bidyut Kr. Patra, Raimo Launonen, Ville Ollikainen, and Sukumar Nandi. A new similarity measure using bhattacharyya coefficient for collaborative filtering in sparse data. *Know.-Based Syst.*, 82(C):163–177, July 2015.
[34] Raciel Yera Toledo, Jorge Castro Gallardo, and Luis Martinez. *Natural Noise Management in Recommender Systems Using Fuzzy Tools*, pages 1–24. Publisher, 01 2020.
[35] Hau Xuan Pham and Jason J. Jung. Preference-based user rating correction process for interactive recommendation systems. *Multimedia Tools Appl.*, 65(1):119–132, July 2013.
[36] Xuan Hau Pham, Jason J. Jung, and Ngoc-Thanh Nguyen. Integrating multiple experts for correction process in interactive recommendation systems. In Ngoc-Thanh Nguyen, Kiem Hoang, and Piotr Jedrzejowicz, editors, *Computational Collective Intelligence. Technologies and Applications*, pages 31–40, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
[37] Wei Zhou, Junhao Wen, Yun Sing Koh, Qingyu Xiong, Min Gao, Gillian Dobbie, and Shafiq Alam. Shilling attacks detection in recommender systems based on target item analysis. *PloS one*, 10:e0130968, 07 2015.
[38] Joeran Beel. And the winner is... movielens — on the popularity of recommender-system datasets. *Journal*, August 2019.
[39] Lamis Hassanieh, Chady Abou Jaoude, Jacques Bou abdo, and Jacques Demerjian. Similarity measures for collaborative filtering recommender systems. In *MENACOMM*, pages 1–5, 04 2018.
[40] Chao Tong, Yu Lian, Jianwei Niu, and Xiang Long. A novel rating prediction method based on user relationship and natural noise. *Multimedia Tools and Applications*, 77, 03 2017.
[41] Miriam Badran, Jacques Bou abdo, Wissam Al Jurdi, and Jacques Demerjian. Adaptive serendipity for recommender systems: Let it find you. In *ICAART 2019*, pages 739–745, 01 2019.
[42] Dongsheng Li, Chao Chen, Zhilin Gong, Tun Lu, Stephen Chu, and Ning Gu. *Collaborative Filtering with Noisy Ratings*, pages 747–755. Publisher, 05 2019.
[43] S K, Kishan Kalitkar, G Subba, and Rao. A review on noise reduction methods for brain mri images. *SPACES-2015, Dept of ECE, K L UNIVERSITY*, 01 2015.
[44] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 11 2013.

A NOTATIONS

Notations that were used in this article are shown in the table below.

Table 7. Notations mentioned in this study and that are very common in the recommender system field.

Notation	Definition
$r_{(u,i)}$	rating of user u on an item i
r_{min}	minimum rating in a rating scale
r_{max}	maximum rating in a rating scale
CF	collaborative filtering
UB-CF	user-based collaborative filtering
IB-CF	item-based collaborative filtering
MF	Matrix factorization
SO	SlopeOne recommender
MAE	mean absolute error
PCC	pearson correlation coefficient
Prec.	precision
Rec.	recall
NN	natural noise in datasets
NNM	natural noise management
I	individual recommendations
G	groups recommendations
D & R	noise detection and removal
D & C	noise detection and correction
HEUG	Heavy Easy User Group, users with high ratings and high consistency
HDUG	Heavy Difficult User Group, users with high ratings and low consistency
MEUG	Medium Easy User Group, users with medium ratings and high consistency
MDUG	Medium Difficult User Group, users with medium ratings and low consistency
LEUG	Light Easy User Group, users with few ratings and high consistency
LDUG	Light Difficult User Group, users with few ratings and low consistency