The Magic Barrier Revisited: Accessing Natural Limitations of Recommender Assessment

Kevin Jasberg
Web Science Group
Heinrich-Heine-University Duesseldorf
Duesseldorf, Germany 45225
kevin.jasberg@uni-duesseldorf.de

Sergej Sizov
Web Science Group
Heinrich-Heine-University Duesseldorf
Duesseldorf, Germany 45225
sizov@hhu.de

ABSTRACT

Recommender systems nowadays have many applications and are of great economic benefit. Hence, it is imperative for success-oriented companies to compare various of such systems and select the better one for their purposes. To this end, various metrics of predictive accuracy are commonly used, such as the Root Mean Square Error (RMSE), or precision and recall. All these metrics more or less measure how well a recommender system can predict human behaviour. Unfortunately, human behaviour is always associated with some degree of uncertainty, making the evaluation difficult, since it is not clear whether a deviation is system-induced or just originates from the natural variability of human decision making. At this point, some authors speculated that we may be reaching some Magic Barrier where this variability prevents us from getting much more accurate [12, 13, 24]. In this article, we will extend the existing theory of the Magic Barrier [24] into a new probabilistic but a yet pragmatic model. In particular, we will use methods from metrology and physics to develop easy-to-handle quantities for computation to describe the Magic Barrier for different accuracy metrics and provide suggestions for common application. This discussion is substantiated by comprehensive experiments with real users and large-scale simulations on a high-performance cluster.

CCS CONCEPTS

•Human-centred computing →HCI theory, concepts and models; User models; User studies; **•Mathematics of computing →Distribution functions;**

KEYWORDS

Magic Barrier; Noise; Human Uncertainty; Distribution-Paradigm; Point-Paradigm; RMSE; Ranking Error

1 INTRODUCTION

Recommender systems have become quite essential for our modern information society. Applied within a variety of engines, they predict human behaviour (e.g. ratings a user might give to a specific item) and thus models a user's preferences. In doing so, those algorithms use specific machine learning techniques to learn about one's personal interests and to develop empathy for multiple as well as variable human aspects.

Unfortunately, human beings can not be deemed as constant functions. It has recently been shown, that users provide inconsistent ratings when requested to rate same films at different times [13]. This **Human Uncertainty**, as we understand it in this contribution, appears to be a characteristic feature of the cognitive process of decision making which influences its outcome, making it circumstantial and temporally unstable; the outcome appears to be more or less fluctuating randomly when repeating a decision making. Consequently, we may assume that observed decisions are drawn from individual distributions [10].

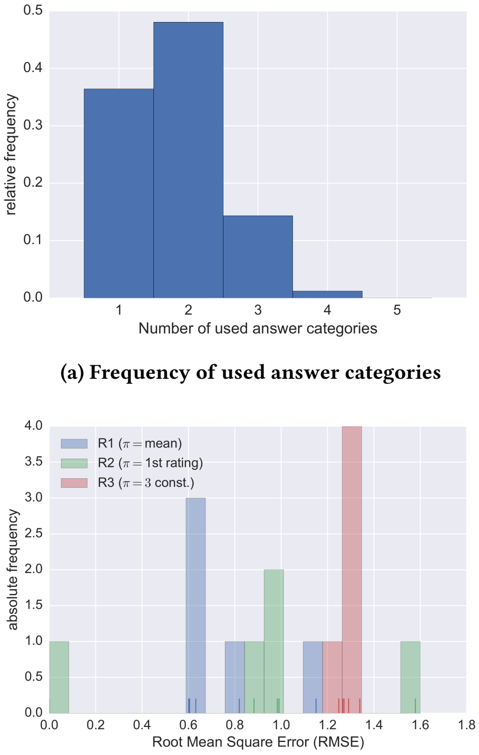
Accordingly, this complicates the evaluation of recommender systems, since it is not clear whether the difference between a given rating and the prediction is induced by the system or just a matter of Human Uncertainty. If we are able to improve the system-induced prediction quality to such an extent that only the factor of human uncertainty is left, then all visible differences within a quality metric would only exist due to this uncertainty and may vary with each repeated rating trial. This implies that rankings of different (well improved) recommender systems would shuffle with each repetition as well, i.e sound rankings do no longer exist for excellent systems but there is an equivalence class of indistinguishable optimal systems. This leads to the assumption of some Magic Barrier where natural variability may prevent us from getting much more accurate [12].

Motivating Example. As a motivating example, we consider the task of rating prediction, along with the Root Mean Square Error (RMSE) as a widely used metric for prediction quality. In a systematic experiment with real users (described in more detail in forthcoming sections), individuals rated theatrical trailers multiple times. Figure 1a shows that only 35% of all users show constant rating behaviour, whereas about 50% use two different answer categories and 15% of all users make use of three or more categories. Based on these observations, we compute the RMSE for three recommender systems (designed by definition of their predictors π) for each rating trial. Figure 1b depicts the RMSE outcomes and their frequency. It becomes apparent at once that the RMSE itself yields a particular degree of uncertainty, emerged from uncertain user feedback. When ranking these recommender systems, Figure 1b allows for three possible results

$$(R1 \prec R2 \prec R3) \ \vee \ (R2 \prec R1 \prec R3) \ \vee \ (R1 \prec R3 \prec R2), \qquad (1)$$

where the relation $\prec$ denotes "better than". The ranking problem is most obvious for recommender $R1$ as it could be both, the best

RecSys '17, August 27–31, 2017, Como, Italy.
© 2017 ACM. 978-1-4503-4652-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3109859.3109898



(a) Frequency of used answer categories

(b) Distribution of RMSE outcomes

Figure 1: Uncertain user ratings and impact on the RMSE

or the worst recommender, although it operates for the same users rating the same items. In addition, it may be possible that further repetitions of ratings would lead to even more ranking possibilities. This naturally implies to deem those RMSE scores as single draws from distributions that are strongly overlapping. As will be revealed later, recommender $R1$ is the Magic Barrier itself. Therefore, our considerations above - the indistinguishability of excellent systems close to the Magic Barrier - hold even for straightforward investigations.

The Problem. The problem of Human Uncertainty - if not explicitly considered - is that any improvement to an existing system or even the assessment of different systems might not be statistically sound. This, in particular, has financial implications when money is invested in the further development of a system but as a result, there is merely an overfitting instead of real improvements. Therefore, the crux is to recognise whether the prediction quality has really improved or is just some random artefact. So there is a need for a decision criterion whether a system still has room for improvements.

For the RMSE in particular, a criterion has recently been developed which allows for a dichotomous consideration (yes or no)[24]. But while the uncertainty of users is considered, its influence on the precise localisation of the Magic Barrier is negated. However, in our example (Fig 1b) we have seen that the RMSE (esp. the Magic Barrier) itself follows a distribution due to Human Uncertainty. As a consequence, systems with an RMSE near the "old" Magic Barrier might already be interfered by this Human Uncertainty and respectively, achieving an RMSE less than the "old" Magic Barrier does not always mean that this system is already interfered. So the question changes from "Is the prediction quality interfered by Human Uncertainty?" to "How likely is it that the prediction quality is interfered by Human Uncertainty?", which allows for more differentiated evaluation of recommender systems.

Our Objective. In this contribution, we present a method by which the Magic Barrier can be estimated for any quality assessment metric. For this purpose, we will embed the Magic Barrier into a complete probabilistic framework and deduce a pragmatic theory through complexity reduction. We aim to generate concrete and action-oriented quantities that can easily be embedded in existing approaches to recommender assessment. We also provide our data records for modelling Human Uncertainty and demonstrate its transferability using the example of Netflix Prize.

2 RELATED WORK

Recommender Systems and Assessment. The central role of recommender systems led to a lot of research and produced a variety of techniques and approaches. A good introduction and overview are given by [16, 23]. For the comparative assessment, different metrics are used to determine the prediction quality, such as the root mean squared error (RMSE), the mean absolute error (MAE), the mean average precision (MAP) along with many others [1, 7, 12]. Although we exemplify our methodology in accordance with the RMSE, the main results of this contribution can be easily adapted for alternative assessment metrics without substantial loss of generality, insofar they require for (uncertain) human input.

Dealing with Uncertainties. The relevance of our contribution arises from the fact that the unavoidable human uncertainty sometimes has a vast influence on the evaluation of different prediction algorithms [3, 5]. The idea of uncertainty is not only related to predictive data mining but also to measuring sciences such as metrology. Recently, a paradigm shift was initiated on the basis of a so far incomplete theory of error [8, 11]. In consequence, measured properties are currently modelled by probability density functions and quantities calculated therefrom are then assigned a distribution by means of a convolution of their argument densities. This model is described in [17]. A feasible framework for computing these convolutions via Monte-Carlo-Simulation is given by [18]. We take this as a basis for our own modelling of uncertainty for addressing similar issues in the field of computer science. To derive a pragmatic and easy to handle theory, we will refer to the Gaussian Error Propagation which is commonly used in physics as well [6, 19, 25].

The Magic Barrier. One of the first works addressing Human Uncertainty and its impact on recommender systems was presented in [13], where users have been proven to give inconsistent ratings on movies. The authors claim that it will never be possible to perfectly predict ratings and that there must exist an upper bound on rating prediction accuracy. Later, this upper bound was mentioned once again in [12] and received the name Magic Barrier, which is still in use nowadays. A first calculation of the Magic Barrier can be found in [24]. Derived by risk function minimisation, the authors defined the Magic Barrier as the square root of the averaged user variances (gathered from repeated ratings). Even though this approach accounts for Human Uncertainty, its influence - namely the uncertainty of the Magic Barrier itself - remains unconsidered. In our contribution, we complete this theory and therefore allow a more differentiated analysis of recommender assessment.

Experimental Designs. The complexity of human perception and cognition can be addressed by means of latent distributions [10]. This idea is widely used in cognitive science and in statistical modelling of ordinal data [14]. We adopt the idea of modelling user uncertainty by means of individual Gaussians following the argumentation in [26] for constructing our individual response models. The methodology applied in our experiments is adopted from experimental psychology [15] and works on repeating rating scenarios for same users-items-pairs as done before in [2].

3 MODELLING A MAGIC BARRIER

In this section, we embed human uncertainty into a mathematical construct and introduce an approach for estimating a Magic Barrier for a given evaluation metric. Although the term “Magic Barrier” is related to the RMSE in particular, such a barrier does basically exist for any metric comparing (uncertain) user inputs with predicted scores. Therefore, we first develop a general framework which will then be illustrated for the RMSE as a prominent example.

3.1 Changing Paradigms

As mentioned above, various experiments [2, 13] along with our own have shown that users are scattering around their true value of preference. Consequently, we may assume that observed decisions are drawn from individual distributions, as a result of complex cognition processes, and influenced by multiple factors (e.g. mood, media literacy, etc.) [10]. Therefrom, a paradigm shift has to be carried out, which is similar to the recent change of perspectives on measurement errors in metrology [8]: Every measurable quantity that is somehow related to human cognition is no longer considered as a single point (point-paradigm) but rather as a whole interval of possible values (set-paradigm) that is somehow distributed (distribution-paradigm). In the context of this paper, we will, therefore, consider user ratings as random variables. On this basis, we develop statistical methodologies that are to be explored hereinafter.

3.2 Composed Quantities

Composed quantities, in this contribution, are quantities Z that compute from a continuous function $Z = g(X_1, \dots, X_n)$ of large amounts of uncertain arguments X_i (random variables). Hence, Z becomes a random variable itself. This reasoning can be understood heuristically: For each draw, there is a variety of possibilities for a single outcome x_i of a random variable X_i. The outcomes $x_1, \dots, x_n$ of all random variables altogether result into a single outcome for the composed quantity Z by means of $z = g(x_1, \dots, x_n)$. Accounting for all the possibilities for $x_1, \dots, x_n$ (e.g. when repeating draws infinitely) will then result in a variety of possible outcomes z. Thus, the distribution of Z emerges as a convolution of n density functions with respect to the mapping g [17, 18].

3.3 Magic Barrier Estimation

The Magic Barrier is defined as the minimum of an evaluation metric when explicitly accounting for Human Uncertainty. Therefore, we must first specify an optimal recommender by defining its predictors. Then we have to compute the probability density function of the evaluation metric which arises for this optimal recommender.

What is an optimal recommender? The choice of predictors depends on the evaluation metric and the underlying data model. We will demonstrate this by using an example. In the case of the Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\tfrac{1}{N}\textstyle\sum_{\nu}(X_\nu - \pi_\nu)^2}, \tag{2}$$

the comparison of a rating X_ν and a prediction $\pi_\nu \in \mathbb{R}$ is done via $c(X_\nu) = (X_\nu - \pi_\nu)^2$, whose expectation reaches its minimum when

$$\tfrac{d}{d\pi}\textstyle\sum_{i=0}^{N}(x_i - \pi)^2 = 2 \cdot \sum_{i=0}^{N}(\pi - x_i) = 0 \;\Leftrightarrow\; \pi = \tfrac{1}{N}\sum_{i=0}^{N} x_i \tag{3}$$

where x_i denote the realisations of the random variable X_ν. Hence, the optimal recommender system with respect to the RMSE is defined by $\pi_\nu := \mathbb{E}[X_\nu]$ for each user-item-pair ν.

This might be totally different when considering the Mean Absolute Error (MAE), whose primary comparison is based on the function $c(X_\nu) = |X_\nu - \pi_\nu|$, reaching a minimum for its expectation when π_ν is the median of X_ν. The median corresponds to the expected value, only if a symmetrical distribution is chosen as the underlying data model. Consequently, when assuming all $X_\nu \sim \mathcal{N}(\mu_\nu, \sigma_\nu)$ to be normally distributed (symmetric density function), the optimal recommender system does not differ for the RMSE and the MAE respectively. Having $X_\nu \sim \Gamma(\alpha_\nu, \beta_\nu)$ being gamma-distributed instead, the optimal recommender may be different for both metrics, depending on the extent of asymmetry.

Monte-Carlo-Simulation. Now having the definition of an optimal recommender system, we need to deduce the probability density function of the evaluation metric for this optimum. In theory, this is done by a convolution of all density functions f_i of X_i, but what sounds simple at first, turns out to be quite laborious and inapplicable as demonstrated in [9]. For this reason, metrologists typically apply statistical simulations. In this paper we use **Monte-Carlo-Simulations** as described in [18]: For each of our ratings X_ν, we compute a sample $\mathcal{S}(X_\nu) := \{x_\nu^1, \dots, x_\nu^\tau\}$ of τ pseudo-random numbers (trials) that are drawn from a distribution (underlying data model). Then, we yield a sample for the evaluation metric $Z = g(X_1, \dots, X_N)$ via

$$\mathcal{S}(Z) = \left\{z_k = g(x_1^k, \dots, x_N^k) \colon k = 1, \dots, \tau\right\}. \tag{4}$$

Post hoc illustration of this sample by a normed histogram with b bins leads to an approximation for the density of Z.

Although the statistical simulation of convolutions produces excellent results while also being easy to realise, we are facing a blatant run-time problem as soon as we are entering the realm of big data. For example, for $N = 80\,000$ ratings, the simulation already takes up to an hour of runtime using a state-of-the-art computing node. To compute the Magic Barrier on the Netflix test record ($N = 2.8 \cdot 10^6$), we need about 35 hours.

In the following sections, we will derive a pragmatic estimate for the desired density function of the Magic Barrier for arbitrary metrics. With this, we get same results but need only a mere fraction of the simulation runtime. For example, the probability density for the Magic Barrier on the Netflix test record can be computed in less than 80 milliseconds.

Estimation Analytics. Even before the technical possibilities of statistical simulations existed, metrologists had estimated the expected value and the variance of quantities $Z = g(X)$. The core these estimations is to expand $g \in C^\infty(\mathbb{R})$ into its Taylor series

$$g(X) = \sum_{k=0}^{\infty} \frac{g^{(k)}(\mu)}{k!}\,(X-\mu)^k \tag{5}$$

where $g^{(k)}(\mu)$ denotes the k^{th} derivative of g evaluated at the expectation of X. Due to the linearity of the expectation[1], we yield

$$\begin{aligned}\mathbb{E}[g(X)] &= \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{g^{(k)}(\mu)}{k!}\,(X-\mu)^k\right] = \sum_{k=0}^{\infty}\frac{g^{(k)}(\mu)}{k!}\mathbb{E}\left[(X-\mu)^k\right]\\ &= \sum_{k=0}^{\infty}\frac{g^{(k)}(\mu)}{k!}m_k \end{aligned}\tag{6}$$

where m_k is the k-th central moment. For the variance and its quasi-linearity[2], we yield

$$\begin{aligned}\mathbb{V}[g(X)] &= \mathbb{V}\left[\sum_{k=0}^{\infty}\frac{f^{(k)}(\mu)}{k!}\,(X-\mu)^k\right] = \sum_{k=0}^{\infty}\left(\frac{f^{(k)}(\mu)}{k!}\right)^2\mathbb{V}\left[(X-\mu)^k\right]\\ &= \sum_{k=0}^{\infty}\left(\frac{f^{(k)}(\mu)}{k!}\right)^2(m_{2k}-m_k^2)\end{aligned}\tag{7}$$

where the last line has been simplified by using the common identity $\mathbb{V}[(X-\mu)^k] = \mathbb{E}[(X-\mu)^{2k}] - \mathbb{E}[(X-\mu)^k]^2 = m_{2k} - {m_k}^2$. The usual approximation is to omit terms of higher orders, like

$$\begin{aligned}\mathbb{E}[g(X)] &= g(\mu) + g'(\mu)\cdot m_1 + \ldots \approx g(\mu)\\ \mathbb{V}[g(X)] &= g'(\mu)^2 m_1 + g''(\mu)^2(m_4 - m_2^2)/4 + \ldots \approx g'(\mu)^2 m_1.\end{aligned}$$

We have so far only considered a smooth function with just one argument in order to guarantee an easy understanding of the methodology. When considering n arguments, we use a Taylor series in more dimensions and yield equivalent results which, together with the assumption of normality, form the Gaussian Error Propagation [6, 19, 25].

4 MAGIC BARRIER FOR THE RMSE

4.1 Application of Gaussian Error Propagation

In this section we will derive closed form approximations for the RMSE and therefore define

$$\mathcal{MB} = g(X_1,\ldots,X_N) := \sqrt{\tfrac{1}{N}\textstyle\sum_v (X_v - \mathbb{E}[X_v])^2}. \tag{8}$$

Since we have to face multiple arguments, we would usually need a Taylor series in several variables, which is quite ugly for demonstration purposes. Therefore, we first condense all ratings $X_1,\ldots,X_N$ into a single random variable and then use the one-dimensional Taylor approximation. In doing so, we choose Gaussians as the underlying data model for our ratings. By this means, every rating $X_v \sim \mathcal{N}(\mu_v,\sigma_v)$ can be written as $X_v = \sigma_v\mathbb{I}+\mu_v$ where $\mathbb{I}\sim\mathcal{N}(0,1)$. Hence, $Y_v := (X_v - \mathbb{E}[X_v])^2$ receives the expectation

$$\begin{aligned}\mathbb{E}[Y_v] &= \mathbb{E}[(\sigma_v\mathbb{I}+\mu_v-\mu_v)^2] = \mathbb{E}[(\sigma_v\mathbb{I})^2]\\ &= \mathbb{E}[\sigma_v^2\mathbb{I}^2] = \sigma_v^2\mathbb{E}[\mathbb{I}^2] = \sigma_v^2\mathbb{V}[\mathbb{I}] = \sigma_v^2\end{aligned}\tag{9}$$

as well as the variance

$$\begin{aligned}\mathbb{V}[Y_v] &= \mathbb{V}[(\sigma_v\mathbb{I}+\mu_v-\mu_v)^2] = \mathbb{V}[(\sigma_v\mathbb{I})^2]\\ &= \mathbb{V}[\sigma_v^2\mathbb{I}^2] = \sigma_v^4\mathbb{V}[\mathbb{I}^2] = \sigma_v^4\left(\mathbb{E}[\mathbb{I}^4]-\mathbb{E}[\mathbb{I}^2]^2\right)\\ &= \sigma_v^4\left(3\mathbb{V}[\mathbb{I}]^2 - \mathbb{V}[\mathbb{I}]^2\right) = 2\sigma_v^4\end{aligned}\tag{10}$$

We thus obtain a χ^2-distribution for $Z := \frac{1}{N}\sum_v Y_v$ which converges into a Gaussian for a large number N of ratings by means of the central limit theorem. The parameters of this Gaussian are

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{1}{N}\sum_v Y_v\right] = \frac{1}{N}\sum_v\mathbb{E}[Y_v] = \frac{1}{N}\sum_v\sigma_v^2 \tag{11}$$

$$\mathbb{V}[Z] = \mathbb{V}\left[\frac{1}{N}\sum_v Y_v\right] = \frac{1}{N^2}\sum_v\mathbb{V}[Y_v] = \frac{2}{N^2}\sum_v\sigma_v^4. \tag{12}$$

Now we can consider the Magic Barrier to be the image of the root function of a single random variable, i.e. $\mathcal{MB} = g(X_1,\ldots,X_N) \equiv h(Z) := \sqrt{Z}$ where $Z \sim \mathcal{N}(\frac{1}{N}\sum_v\sigma_v^2\,,\,\frac{2}{N^2}\sum_v\sigma_v^4)$. Applying the one-dimensional Taylor approximation from equations 6 and 7 leads to

$$\mathbb{E}[\mathcal{MB}] = \sqrt{\mathbb{E}[Z]} - \frac{\mathbb{V}[Z]}{8\mathbb{E}[Z]^{3/2}} - \ldots \approx \sqrt{\mathbb{E}[Z]} \tag{13}$$

$$\mathbb{V}[\mathcal{MB}] = \frac{\mathbb{V}[Z]}{4\mathbb{E}[Z]} + \frac{\mathbb{V}[Z]^2}{32\mathbb{E}[Z]^3} + \ldots \approx \frac{\mathbb{V}[Z]}{4\mathbb{E}[Z]}. \tag{14}$$

With additional assumption of normality (which is indeed a suitable model, as we will confirm soon), the approximated distribution of the Magic Barrier for the RMSE is

$$\mathcal{MB} \sim \mathcal{N}\left(\sqrt{\frac{1}{N}\textstyle\sum_v\sigma_v^2}\,,\ \frac{1}{2N}\frac{\sum_v\sigma_v^4}{\sum_v\sigma_v^2}\right) \tag{15}$$

where $\mathbb{E}[\mathcal{MB}] \approx (\sum_v\sigma_v^2/N)^{1/2}$ exactly meets the traditional Magic Barrier as defined in [24] and $\mathbb{V}[\mathcal{MB}] \approx (\sum_v\sigma_v^4)/(2N\sum_v\sigma_v^2)$ represents the traditionally neglected uncertainty of this Magic Barrier, emerged from uncertain user ratings.

4.2 Goodness of Approximation

As mentioned above, the method presented here is merely an approximation, since we omit terms of higher orders. At this point, one may wonder how well this estimate actually matches the true state. To answer this question, we first compare the simulated expectations and variances with the calculated ones in a regression analysis. Concerning the distribution model, we investigate the degree similarity using the Jensen–Shannon-Divergence.

Regression analysis. We keep the following simulations as general as possible. To this end we gradually fix a particular number N of ratings from the set $\{50, 100, 150, 200, 500, 1000\}$ and sample N expectations μ_v uniformly from the interval $[1, 5]$ as well as N variances σ_v^2 uniformly from $[\sigma_{min}^2, \sigma_{max}^2]$. These intervals result from the assumption of five repeated ratings (as happened in our

[1] $\mathbb{E}[aX+b] = a\mathbb{E}[X]+b$ holds for $a,\,b \in \mathbb{R}$ and arbitrary random variable X
[2] $\mathbb{V}[aX+b] = a^2\mathbb{V}[X]$ holds for $a,\,b \in \mathbb{R}$ and arbitrary random variable X

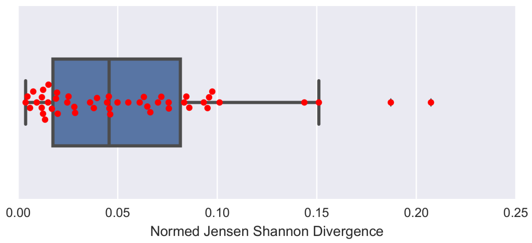


Figure 2: Jensen–Shannon-Divergence for comparing the simulated distribution with a predetermined Gaussian

experiments) with the commonly used 5-star scale. Under these conditions, the positive variance yields limitations[3]

$$\sigma^2_{min} = \text{var}(\{1,1,1,1,2\}) = 0.16 \quad (16)$$

$$\sigma^2_{max} = \text{var}(\{1,1,1,5,5\}) = 3.86 \quad (17)$$

For each pair (μ_v, σ^2_v) we then compute a sample $\mathcal{S}(X_v)$ with $\tau = 10^7$ random numbers drawn from the specified Gaussian to perform the convolution via equation 4. For many repetitions, we receive a lot of simulated expectations/variances to be plotted against the approximated ones by means of linear regression. A perfect match between simulation and approximation would lead to the regression $y = 1 \cdot x + 0$ with correlation coefficient $R^2 = 1$. The results

$$\text{Sim}(\mathbb{E}) = 0.999 \cdot \text{Apr}(\mathbb{E}) - 0.003 \qquad (R^2 = 0.99) \quad (18)$$

$$\text{Sim}(\mathbb{V}) = 0.981 \cdot \text{Apr}(\mathbb{V}) + 0.000 \qquad (R^2 = 1.00) \quad (19)$$

show that this condition is almost fully achieved and hence we may consider these approximations as appropriate.

Jensen–Shannon-Divergence. When modelling the Magic Barrier, not only the expectation and the variance are of great importance, but rather the entire probability density. While the simulated distribution arises naturally from convolution, it is predetermined for the approximation. Therefore, it is necessary to evaluate the degree of deviation of both distributions. In doing so, we proceed as done in the regression analysis above, but instead of computing means and variances, we transform our samples into discrete probability distributions P_{sim} and P_{apr} and analyse the Jensen–Shannon-Divergence (JSD)

$$\text{JSD}(P_{sim}|P_{apr}) = \frac{1}{2}D_{\text{KL}}(P_{sim}|M) + \frac{1}{2}D_{\text{KL}}(P_{apr}|M) \quad (20)$$

where $D_{\text{KL}}(P_1|P_2) = \sum_i P_1(i)\ \log_2(P_1(i)/P_2(i))$ denotes the Kullback-Leibler-Divergence and $M = \frac{1}{2}(P_{sim} + P_{apr})$. Since we use the base 2 logarithm, the JSD yields the boundaries

$$0 \leq \text{JSD} \leq 2\log(2) \quad \text{or} \quad 0 \leq \frac{\text{JSD}}{2\log(2)} \leq 1 \quad (21)$$

The outcomes for the normed JSD is shown in Figure 2. We observe that the mid-range of all outcomes is located between 0.01 and 0.08 confirming high similarity of the simulated distribution and the assumed Gaussian. There are, however, some outliers which only occur for $N = 50$ ratings. This can be explained by the fact that the RMSE contains the sum of squared normal distributions, which is χ^2-distributed, but quickly converges to the normal distribution for $N > 100$. Thus, the more ratings we have, the more adequate is a Gaussian as the assumed density. For a visual comparison of both distributions, Figure 6 depicts the simulated density as well as the approximation for our experiment with $N = 213$.

4.3 Understanding the Magic Barrier

In this section, we will take a closer look at the properties of the Magic Barrier. For this purpose, the individual dependencies of the Magic Barrier and their effects are analysed in a sensitivity analysis. In addition, we will generalise the dichotomous decision criterion from [24] and develop a pragmatic rule of thumb to ascertain whether a deeper consideration of the Magic Barrier seems worthwhile.

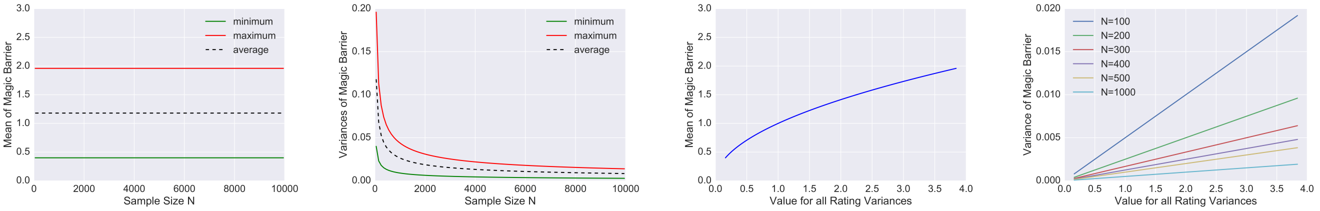
Sensitivity analysis. A sensitivity analysis is used to determine how a quantity responds to the variation of its arguments. Therefore, we vary one argument within reasonable boundaries while fixing all the other arguments at the same time.

In Figure 3a and 3b, one can observe the Magic Barrier's reaction to an increasing number N of ratings. It is seen that the expectation remains unaffected by the number of uncertain ratings. Only the extent of the uncertainty raises or lowers the mean value. On a 5-star scale together with five re-ratings, the expected value yields limitations (green and red) due to the minimum and maximum variance possible. The growth behaviour of the expectation under rating uncertainty is asymptotic. However, Figure 3b reveals that the Magic Barrier's variance is heavily impacted by the number of ratings, i.e. the precision of the Magic Barrier even gains when more uncertain ratings are added. The extent of rating uncertainty also leads to boundaries, but this influence gradually disappears for increasing N. A comparison of Figure 3b and Figure 3d reveals that the number of ratings significantly affects the first two decimal places of the variance, whereas the influence of the rating uncertainty affects only the third and fourth decimal places at most. In summary, it can be said that the extent of Human Uncertainty alone is responsible for the location of the Magic Barrier, whilst its spread can be reduced by adding ratings. However, the degree of this improvement decreases very rapidly.

What we have omitted here is the influence of the underlying data model and the applied rating scale. The rating scale limits the variance of a user and thus has a great impact of the possible location of the Magic Barrier. The underlying data model has also a great impact on the Magic Barrier but will be the discussed separately in further research.

Do we need a Magic Distribution? Now having in mind that the variance of the Magic Barrier decreases for large N, one may ask if we really need the Magic Barrier to be a distribution rather that a single score. The answer depends on many factors. First of all, the world of recommender assessment does not entirely consist of large-scale experiments, so that the variance can not be deemed to equal zero. In the case of large-scale experiments, the predefined accuracy of computed scores does matter quite a lot. For example, all RMSE scores were given to the fourth decimal place in Netflix Prize [22]. As shown in the following sections, the standard deviation of the Magic Barrier for the Netflix data set can be assumed to be

[3] Samples are only examples producing the minimum/maximum variance



(a) $\mathbb{E}[\mathcal{MB}]$ **with respect to** N **(b)** $\mathbb{V}[\mathcal{MB}]$ **with respect to** N **(c)** $\mathbb{V}[\mathcal{MB}]$ **with respect to** σ^2 **(d)** $\mathbb{V}[\mathcal{MB}]$ **with respect to** σ^2

Figure 3: Sensitivity analysis of the Magic Barrier varying the number of ratings and the extent of rating variances

$\sigma = 0.0007$, which is still seven times larger than the specified rounding accuracy of four decimal places. In this example, we see that even for large data records the effort of considering the Magic Barrier as a distribution is quite meaningful.

Furthermore, we need a non-vanishing variance for a statistically sound decision whether a system can still be improved. Following [24], any improvement of a recommender system is pointless, if the RMSE score is below the Magic Barrier, i.e. $\mathbb{E}[\text{RMSE}] < \mathbb{E}[\mathcal{MB}]$. But since both quantities are distributed, their density functions may nevertheless overlap. Figure 4 illustrates the interference of the Magic Barrier with a recommender system used in our experiments. Although the decision criterion from [24] holds, there is a significant probability that the RMSE outcome is already affected by the Magic Barrier. This probability is given by

$$P(\mathcal{MB} > \text{RMSE}) = \int_{-\infty}^{\infty} f_{\text{RMSE}}(x) \cdot \left(1 - F_{\mathcal{MB}}(x)\right) \mathrm{d}x \qquad (22)$$

where $F_{\mathcal{MB}}(x)$ denotes the cumulative distribution function. In our example from Figure 4, this probability is around 0.33, i.e. the RMSE is interfering with the Magic Barrier in one of three outcomes. For this reason, an analysis of possible improvements can not be answered by a dichotomous decision criterion (yes or no), but has to be answered by means of probabilities (How likely is it that my system can still be improved and what risk am I willing to accept?).

When is a differentiated consideration needed? However, such a differentiated approach is not always worth it. Therefore, it would be useful to have a rule of thumb to find out whether a differentiated consideration is fruitful or not. For example, a possible criterion might be the intersection of the 99%-confidence intervals of the RMSE and the Magic Barrier. Due to normality, further analysis should be taken into consideration, when

$$\mathbb{E}[\mathcal{MB}] + 3\sqrt{\mathbb{V}[\mathcal{MB}]} > \mathbb{E}[\text{RMSE}] - 3\sqrt{\mathbb{V}[\text{RMSE}]}. \qquad (23)$$

By assuming $\mathbb{V}[\mathcal{MB}] \approx \mathbb{V}[\text{RMSE}]$, which usually holds when both quantities are computed on the same data record, this criterion can be simplified to $\mathbb{E}[\text{RMSE}] - \mathbb{E}[\mathcal{MB}] < 6\mathbb{V}[\mathcal{MB}]^{1/2}$.

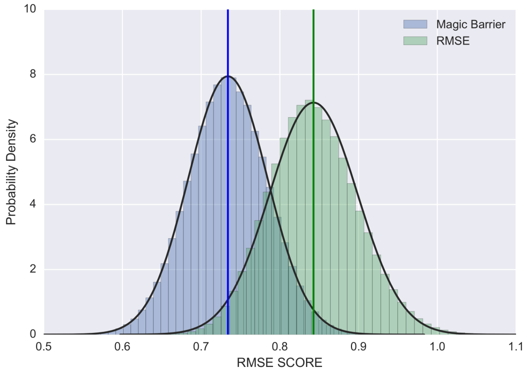


Figure 4: Interference of RMSE with the Magic Barrier

5 EXPERIMENTS

In this section, we examine our theoretical considerations in reality. To this end, we conducted a controlled experiment with real users and measured their uncertainty. We are thus able to support the chosen data model and verify our approximation on a real data set. On this basis, possible applications can be illustrated (e.g. transferring our variances to other situations where no Human Uncertainty was explicitly measured).

5.1 The Experiment

Our experiment is set up with Unipark's[4] survey engine while our participants were committed from the crowdsourcing platform Clickworker[5]. To derive a user's rating distributions, we use the method of re-rating, which was successfully used in [2, 13] before. For this purpose, participants watched theatrical trailers of popular movies and television shows and provided ratings in five repetition trials[6]. User ratings have been recorded for five out of ten fixed trailers so that remaining trailers act as distractors triggering the misinformation effect, i.e. memory is becoming less accurate due to interference from post-event information.

We received a rating tensor $R_{u,i,t}$ with $\dim(R) = (67, 5, 5)$, having $N = 1\,675$ ratings in total, where the coordinates (u, i, t) encode the rating that has been given to item i by user u in the t-th trial. From this record, we derive a unique rating distribution for each user-item-pair by considering tensor-slices in trial-dimension $R_{u,i} := \{R_{u,i,t} | t = 1, \dots, 5\}$ for which we compute Maximum-Likelihood-Parameters given a predetermined data model (e.g. Gaussians, CUB-Models, etc.). Altogether 67 people from Germany, Austria and Switzerland participated in this experiment. This group can be parted into 57% females and 43% males whose ages range

[4] http://www.unipark.com/de/

[5] https://www.clickworker.de/

[6] A full description can be found on https://anonymity/doubleblind

from 20 to 60 years while over 60% of our participants were aged between 20 and 40. This group also includes a good average of lower, medium and higher educational levels. The rating frequency habits range from “rarely” to “often” in a uniform distribution.

5.2 Data Model and Uncertainties

Proving the data model. In this contribution we opt for Gaussians since they are strongly associated with human characteristics [4] and have also been proven to be appropriate user models in [26]. Additionally, Gaussians exhibit maximum entropy among all distributions with finite mean/variance and support on $\mathbb{R}$. For each recorded item, all tensor slices having a non-vanishing variance are checked for normality by means of a one-sample KS-test [20] with confidence level $\alpha = 0.05$. The null hypothesis was never rejected, allowing to keep the Gaussian distribution as a possible model.

Proving Human Uncertainty. For each of the user-item-pairs $R_{u,i}$, we compute the Gaussian ML-Parameters and consider the variances $\mathbb{V}(R_{u,i})$ as representations of the Human Uncertainty. In our experiment, only few tensor slices contain constant ratings and hence lead to a vanishing variance. Performing an item-wise analysis, the fraction of tensor slices with non-zero variance ranges from 50 to 90% that is, only every second participant is able to reproduce its own decisions for the best case. For the worst case, only one out of ten participants is able to precisely reproduce a rating. Figure 5 depicts the distribution of variances emerged from repeated ratings within our experiment. We observe that the overall variance follows an exponential distribution $\mathbb{V} \sim \text{Exp}(\lambda)$ with parameter $\lambda = 2.11$. This power-law distribution literally means that many users have a low degree of uncertainty while only a few users have a very high degree of uncertainty.

5.3 The Magic Barrier

Figure 3a shows that the expected value of the Magic Barrier depends solely on the Human Uncertainty. For our five-star scale as well as its minimum and maximum variances, the expectation should - when equation 15 holds - be located in the interval $[0.40\,;\,1.55]$. In the case of our experiment, we have $N = 213$ rating distributions with non-vanishing variance. It is clear from Figure 3b that for this sample size, the distribution of the Human Uncertainty has a large impact on the variance of the Magic Barrier. According to equation15, the variance of the Magic Barriers should be found in the interval $[0.0008\,;\,0.0113]$.

On the basis of our data record, the simulation and approximation lead to well matching expectations (ca. 0.733) and variances (ca. 0.003) for the Magic Barrier. It is apparent, that the true values are located near the lower bound of the previously estimated intervals. This can be explained by the power-law distribution, i.e. a lot of variances are near the minimum and only a few have got higher extents. The difference of expectations is about 0.2% while the difference of variances is about 1.2%. The matching between the simulated and the assumed data model of a Gaussian can be clearly confirmed in Figure 6. The corresponding normed JSD is 0.05.

5.4 Application

Implicit Impact on Recommender Assessment. So far we have only discussed the explicit impact on the assessment of recommender systems, that is: How likely is it that a system can still be improved, just before the RMSE solely depends on Human Uncertainty itself. Now we want to investigate the implicit influence, which affects any recommender comparison, even if the corresponding RMSE distributions are not directly overlapping with the Magic Barrier.

In doing so, we generate two copies of the Magic Barrier (as the optimal recommender). Each of these copies is gradually distorted by adding artificial noise to their predictors in such a way that the relative noise difference of both copies remains constant. By increasing the noise for both copies whilst keeping their relative difference constant, we generate an offset (distance from the Magic Barrier). This offset is plotted against the probabilities of error when using the traditional point-paradigm ranking, which is given by the generalisation of equation 22. Figure 7 depicts the family of curves, mapping the distance from the Magic Barrier to the corresponding error probabilities. This distance (x-axis) represents the overall quality of a system, i.e. the larger this quantity, the worse the prediction quality. The colours encode the relative difference Δ of two recommender systems among each other. For the green curve (representing 10% noise of difference), an x-value of 0.15 means

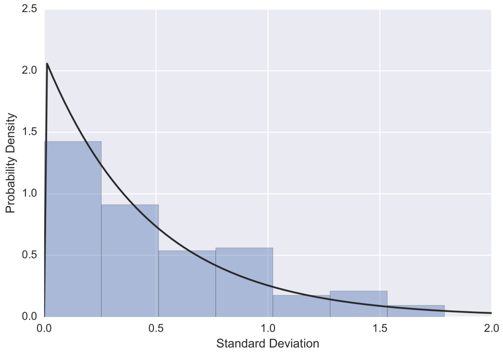


Figure 5: Distribution of variances emerged from repeated ratings within our experiment

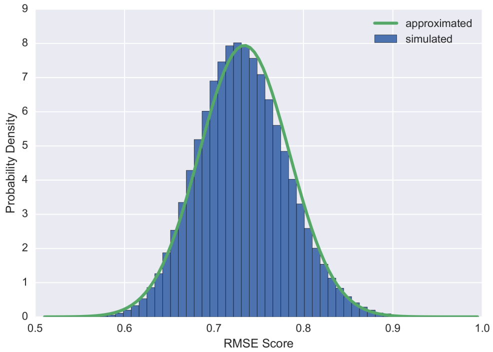


Figure 6: Visual comparison of simulated and approximation Magic Barrier based on experimental data records

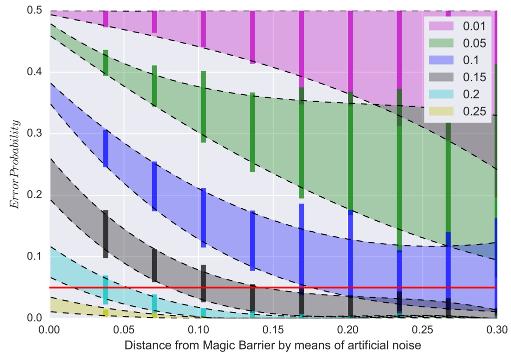


Figure 7: Error Probabilities for ranking systems with constant RMSE difference according to their distance to the Magic Barrier

that system 1 has a noise of 15% whereas system 2 has a noise of 25%. The corresponding y-value indicates the error probability for ranking both systems using the traditional point-paradigm. It is apparent, that two recommender systems can not be brought into a ranking order without considerable error probability if their relative difference is less than 15%, regardless of their basic prediction quality. As a result, we recognise the following: The distance from the Magic Barrier has a great influence on the overlaps in two constantly different recommender systems, i.e. for a fixed difference in prediction quality, they can be distinguished much better if they are bad systems, rather than good ones. On the contrary, the better a system becomes, the more improvement does a revision need, in order to be detected with statistical evidence. This basically means that a recommender system within a repeated process of improvement will certainly reach a prediction quality so that there is probably no sufficiently large amount of optimisation left, in order to distinguish further improvements from the old system with statistical evidence. This convergence is actually the true nature of the Magic Barrier, which could not have been shown without switching perspectives to the distribution paradigm.

Transferability: The Netflix Prize. Unfortunately, existing records have not gathered Human Uncertainty. Therefore, we examine the possibility of applying the findings of our experiment to such data records. To this end, we assume the distribution of Human Uncertainty, emerged from our experiment, to be valid for a larger number of ratings. Under this condition, we will examine possible consequences on Netflix Prize as an example.

The Netflix test record consists of $N = 2.8 \cdot 10^6$ ratings in total. For each of these ratings, we randomly assign a variance drawn from the Pareto distribution in Figure 5. According to this data, the Magic Barrier can be estimated to $\mathcal{MB} \sim \mathcal{N}(0.6687, 0.0007)$. Even though the standard deviation seems small, it is still in the range of Netflix's rounding accuracy of four decimal places. To estimate whether the contest winner [21] might interfere with the Magic Barrier, we use the simplification of Equation 23. Since $\mathbb{E}[\text{RMSE}_{best}] - \mathbb{E}[\mathcal{MB}] = 0.8567 - 0.6687 = 0.1880$ is greater than $6\mathbb{V}[\text{MB}]^{1/2} = 0.1587$, it can be assumed that the Magic Barrier has not yet been reached. In fact, there is still the potential for about 20% of improvement when taking the winner as the reference.

6 DISCUSSION AND CONCLUSIONS

Discussion. In our experiment, the existence of Human Uncertainty is proven and it has been shown that it corresponds to a power-law distribution, i.e. there are many users having a small variance and there are only a few users having a large variance. This implies the existence of an offset within every prediction quality metric that emerges from Human Uncertainty, the so-called Magic Barrier. Having several recommender systems whose RMSEs, for example, are lower than this Magic Barrier, every repetition of the rating proceeding would very likely result into rearrangements of the ranking order, i.e. a reliable ranking can not be built.

In this article, we have lifted an existing theory of this Magic Barrier into a completely probabilistic methodology, providing a generalisation for any quality related metric. Our estimation provides processing of big data in little time while additionally being very precise. With our probabilistic approach, the true nature of the Magic Barrier can be demonstrated: When approaching the Magic Barrier, the distinguishability of many recommender systems automatically decreases, supporting the idea of one equivalence class of optimal systems. Likewise, essential properties of the Magic Barrier have been revealed, for example, the expectation does not change for a higher number of ratings. In contrast, the variance even decreases for an additional number of uncertain ratings and allows to locate the Magic Barrier more precisely. Finally, we have demonstrated the possibility to transfer our results onto other data records in order to make careful predictions of possible interference.

Conclusion. What are the consequences for the assessment of recommender systems in general? The essence of our contribution is the revelation of the following problems:

1. User feedback always comes along with a particular degree of volatility, denoted Human Uncertainty.
2. Human Uncertainty creates a barrier from which below any assessment results are just random.
3. This barrier also implicitly influences recommender assessments; the better our systems become, the more indistinguishable they become.

At this point, it must be said that these problems are not grounded in this new perspective presented here, but have always been present in data analysis. The approach used in this contribution is just able to make these problems visible. Furthermore, these problems do not only occur within our experiments but have also been proven by other authors in different situations of user feedback. This may have far-reaching consequences, especially in the area of the recommender systems, when the selection of a supposedly better system is a monetary decision. For example, financial resources may be invested in improving a system but the improvements achieved are purely random, which remains unnoticed.

For this reason, it becomes crucial to further examine the extent of impact of Human Uncertainty within this field of research. It is also necessary to find proper solutions to these problems, e.g. designing sophisticated mechanisms to identify uncertainty and developing novel strategies to efficiently deal with it. This naturally involves research that connects the fields of behavioural decision making, cognitive psychology and recommender systems to create interdisciplinary synergy effects. We will continue to address these issues in further research.

REFERENCES

[1] Xavier Amatriain (Ed.). 2012. *Workshop on Recommendation Utitlity Evaluation: Beyond RMSE September.* ACM, Dublin, Ireland.

[2] Xavier Amatriain and Josep Pujol. 2009. Rate It Again: Increasing Recommendation Accuracy by User Re-rating. In *Proceedings of the Third ACM Conference on Recommender Systems*. ACM, 173–180.

[3] Xavier Amatriain, Josep M Pujol, and Nuria Oliver. 2009. I like it... i like it not: Evaluating user ratings noise in recommender systems. *International Conference on User Modeling, Adaptation, and Personalization* (2009), 247–258.

[4] F. Gregory Ashby. 1992. *Multidimensional Models of Perception and Cognition (Scientific Psychology Series.* Psychology Press.

[5] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RecSys '13).* 7–14.

[6] Philip Bevington and D. Keith Robinson. 2002. *Data Reduction and Error Analysis for the Physical Sciences* (3rd ed.). McGraw-Hill Education.

[7] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.

[8] Andy Buffler, Saalih Allie, and Fred Lubben. 2001. The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education* 23, 11 (2001), 1137–1156.

[9] F Kenneth Chan. 2011. Miss Distance–Generalized Variance Non-Central Chi Distribution. In *AAS/AIAA Space Flight Mechanics Meeting*. 11–175.

[10] Angela D'Elia and Domenico Piccolo. 2005. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis* 49, 3 (2005), 917–934.

[11] Michael Grabe. 2011. *Grundriss der Generalisierten Gauß'schen Fehlerrechnung*. Springer Berlin Heidelberg.

[12] Herlocker. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 1 (2004), 5–53.

[13] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. 1995. Recommending and Evaluating Choices in a Virtual Community of Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95).* 194–201.

[14] Maria Iannario. 2014. Modelling Uncertainty and Overdispersion in Ordinal Data. *Communications in Statistics - Theory and Methods* 43 (2014), 771–786. Issue 14.

[15] Helene Intraub. 1990. Presentation Rate and the Representation of Briefly Glimpsed Pictures in Memory. *Journal of Experimental Psychology* 6, 1 (1990), 1–11.

[16] Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction.* Cambridge University Press.

[17] JCGM. 2008. *Guide to the Expression of Uncertainty in Measurement.* Technical Report. BIPM.

[18] JCGM. 2008. *Supplement 1 to the GUM - Propagation of distributions using a Monte Carlo method.* Technical Report. BIPM.

[19] HH Ku. 1966. Notes on the use of propagation of error formulas. *J. Res. Nat. Bur. Standards* 70, 4 (1966).

[20] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[21] Netflix. (Oct 25 2016). Netflix Leaderboard. *http://www.netflixprize.com/leaderboard.html* ((Oct 25 2016)).

[22] Netflix. (Oct 25 2016). The Netflix Prize Rules. *http://www.netflixprize.com/rules.html* ((Oct 25 2016)).

[23] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. *Recommender Systems Handbook.* Springer.

[24] Alan Said, Brijnesh Jain, Sascha Narr, and Till Plumbaum. 2012. Users and Noise: The Magic Barrier of Recommender Systems. In *User Modeling, Adaptation, and Personalization.* Vol. 7379. Springer Berlin / Heidelberg, 237–248.

[25] John Taylor. 1997. *Introduction to error analysis, the study of uncertainties in physical measurements.* University Science Books.

[26] Yi Zhang and Jonathan Koren. 2007. Efficient bayesian hierarchical user modeling for recommendation system. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), 47–54.