Evaluation and Assessment of Recommenders Using Monte Carlo Simulation

Renato A. C. Capuruço
PhD Candidate
University of Western Ontario, Department of Electrical and Computer Engineering
London, Ontario, CANADA, N6A 5B9
r.capu@uwo.ca

Luiz F. Capretz
Associate Professor
University of Western Ontario, Department of Electrical and Computer Engineering
London, Ontario, CANADA, N6A 5B9
lcapretz@uwo.ca

ABSTRACT
This paper presents a stochastic model based on Monte Carlo simulation techniques for measuring the performance of recommenders. A general procedure to assess the accuracy of recommendation predictions is presented and implemented in a typical case study where input parameters are treated as random values and recommender errors are estimated using sensitive analysis. The results obtained are presented and a new perspective to the evaluation and assessment of recommender systems is discussed.

Author Keywords
Collaborative Filtering, Recommender Evaluation, Monte Carlo Simulation, Sensitive Analysis, Stochastic.

ACM Classification Keywords
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Collaborative Filtering*; G.3 [**Mathematics of Computing**]: Probability and Statistics – *Probabilistic algorithms (including Monte Carlo).*

INTRODUCTION
In the literature, recent investigations have shown that recommender algorithms have a number of performance complications for worse or better, depending on several factors such as on the dataset chosen for testing, and data sparseness due to new users or few ratings (cold start) [1]. Another major challenge in recommenders is the fact that the user similarity computation is particularly susceptible to additional ratings that are added to, or changed in the database, at which point the similarity values should be recalculated over time [2]. Incorporating the different sources of uncertainty that affect the overall performance into the recommender effectiveness analysis complicates the evaluation method and renders traditional deterministic statistical approaches used for evaluation as insufficient to deal with the random formulation involved, particularly with random predictive behavior due to unwarranted input parameters. The novelty of this work is in the development of an evaluation model for efficiently representing the direct impact of the various recommender parameters on performance, quantifying the variability and reliability of prediction errors, and facilitating the understanding of different sources of uncertainty.

In this paper, recommendation quality is evaluated according to a stochastic-based model that is established with the help of a sensitivity analysis scheme built upon multiple simulation scenarios. These scenarios represent the possible effects of particular combinations of input parameters to the prediction error through the recommender prediction algorithm associated with each run. By aggregating all of these individual performance indicators of each scenario, key summary statistics can be inferred to enable a more complete assessment, measurement, and representation of the recommender robustness. Lastly, reports on significant findings are outlined.

RELATED WORK
Approaches to empirical research incorporate both quantitative and qualitative methods for collecting and analyzing data [3]. *Quantitative methods* collect numerical data and analyze it using statistical methods, relying on precise measurement outcome to yield conclusions. There are a number of evaluation metrics have been available to evaluate the recommender systems performance [4]. These include statistical coverage and accuracy metrics. Coverage metrics such as *precision*, r*ecall* and *F1-measure* are widely used metrics to evaluate the quality of recommendations [5]. According to Palanival and Sivakumar [6], while "Precision" is defined as the ratio of the selected relevant items to the selected items, "Recall" is calculated as the ratio of the elected relevant items to the relevant items. The "F1-measure" is a combination metric that gives equal weight to both "Precision" and "Recall". Accuracy metrics, on the other hand, are standard statistical calculations to compare the numerical deviation of the predicted ratings from the respective actual user ratings. The *mean absolute error* (MAE) and *root mean square error* (RMSE) are computed on result data where lower values indicate more accurate predictions. Relevant to recommenders, all of these efforts are deterministic in nature, that is, given a particular set of initial user-item rating conditions, the evaluation performs the same way.

Based on the preceding discussions, we argue in this work that recommender evaluation is a continuous, on-going process much more than determining the precise error outcome at a given moment. It is rather a way of gauging the performance of predictions over time, which in the context of this work, is achieved by simulating those

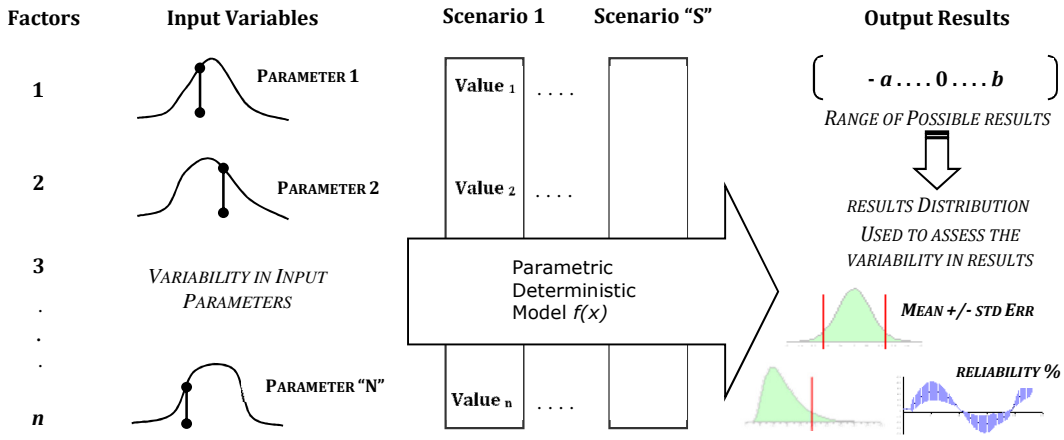


Figure 1. Monte Carlo Simulation Principle

conditions using Monte Carlo simulation techniques for uncertainty modeling. In a stochastic model, randomness is present, and input variable ratings are not described by unique values, but rather by their probability distributions. The Monte Carlo method has been reported as appropriate when the final outcomes to a decision problem depend on the effects of a number of different uncertain events (i.e., rating activity) and on the manner in which they might combine (i.e., proposed recommendation strategy) [7]. Another motivation for this work is the lack of experimentation with stochastic modeling in the context of recommenders.

MONTE-CARLO METHOD

A Monte Carlo method is a stochastic technique used to assess uncertainty in the performance of systems [8]. The word "stochastic" means that it uses random numbers and probability analysis in its formulation. The term "Monte Carlo" comes from the name of the city of Monte-Carlo in the principality of Monaco, Europe. The city's main attractions are casinos, which run activities such as roulette wheels, dice and slot machines. These games provide entertainment by exploiting the random behavior of each game. Similarly, Monte Carlo methods consider the situation when the parameters or factors affecting a problem are not deterministic.

The beginning of real use of Monte Carlo methods as research tools remotes to the development of the atomic bomb as part of the Manhattan Project during World War II due to the experimental mathematics-nature of the problems being tackled. Physicist Nicholas Metropolis, inspired by his colleague Stanislaw Ulam's interest in poker, coined the term for the experimentations that were conducted soon after the project was over [9]. However, they are now applied to a wide range of multivariable problems, from nuclear reactor design, econometrics and stellar evolution to stock and market forecasting, just to name a few.

Problems handled by Monte Carlo methods are of two types called probabilistic or deterministic according to whether or not they are directly concerned with the behavior and outcome of random processes. In the case of a probabilistic problem, the simplest Monte Carlo approach is to observe random numbers, chosen in such a way that they directly simulate the physical random processes of the original problem, and to infer the desired solution from the behavior of these random numbers [10].

Monte Carlo Simulation

In the case of a deterministic problem, the idea behind the Monte Carlo approach is to exploit the strength of theoretical mathematics where one can write down symbolic expressions or formal equations, which abstract the essence of a problem and reveal its underlying structure by replacing theory by experiment whenever the former falters [11]. More specifically, a Monte Carlo simulation is a derived method for iteratively evaluating a deterministic model using sets of random numbers as inputs.

In a Monte Carlo simulation, as presented in Figure 1, a random selection process is used to create multiple scenarios, in which the parameters of the known factors that affect the process take one of their possible values. As such, each scenario provides one possible solution to the problem. Together, these scenarios give a range of possible solutions, some of which are more probable and some less probable. When the process is repeated for hundreds or thousands of scenarios, the average solution will give an approximate answer, considering all of the variability among the scenarios. The data generated from the simulation can be represented as probability distributions (or histograms) or converted to error bars, reliability predictions, tolerance zones, and confidence intervals. Accuracy of this answer can be improved by increasing the number of scenarios.

In this approach, the effects of a particular combination of factors can also be closely examined by analyzing the uncertainty propagation, where the goal is to determine how random variation, lack of knowledge, or error affects the sensitivity, performance, or reliability of the system that is

being modeled [12].

Summary Statistics

In order to effectively communicate the evaluation results when performing a data analysis by using Monte Carlo simulation techniques, it is necessary to summarize the set of observations due to the large amount of observations. There are four basic measures that do that, as below:

Measure of Location

Relates to the tendency of data to be clustered around a central value, that is, the measure of central tendency is an average of a set of measurements. However, it should be noted that depending on the context, the word average can be interpreted as mean, median, mode, or other measure of location. The *arithmetic mean* is the most commonly used measure, and it is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

where, n is the number of observations and, x represents an observation.

Measure of Dispersion

Expresses the amount of variability or spread there is from the "average" (mean). The Standard deviation is a widely used measure of variability or diversity used in statistics, and can be estimated by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

where, $\{ x_1, x_2, \ldots, x_n \}$ are the n observed values, and $\bar{x}$ is the mean value of these observations.

A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values. Together, *location* and *dispersion* are the two mostly used properties of distributions. The standard error can be used to calculate confidence intervals for the true population mean [13], for instance, for a 95% 2-sided confidence interval, the Upper Confidence Limit and Lower Confidence Limit are calculated as:

$$\begin{aligned} 0.95 = 1 - \alpha &= P(-z \le Z \le z) \\ &= P\left(-1.96 \le \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \le 1.96\right) \end{aligned} \tag{3}$$

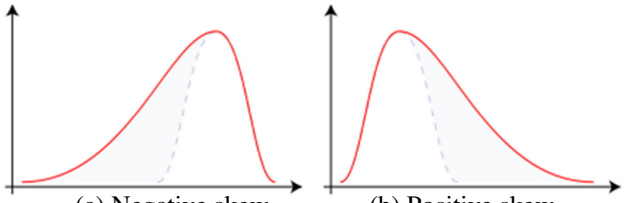
where, the number z follows from the cumulative (normal) distribution function $P(Z)$, α is the significance level, n is the number of observed values, $\bar{x}$ is the mean value of these observations.

Measure of Shape

Common measures of the shape of a distribution are *Skewness* and *Kurtosis*. Whereas the first relates to the asymmetry of the probability distribution, the second measure quantifies the "peakedness" of the distribution and the heaviness of its tail [14]. *Skewness* values can be positive or negative, or even undefined, as shown in Figure 2. In case the left tail of a distribution is longer (Figure 2.a) that implies that the mass of the distribution is concentrated on the right of the distribution and in this case it is said that the distribution has a *negative skew*, or *left-skewed*, *left-tailed*, or *skewed to the left*; likewise, a positive skew (Figure 2.b) means that the mass of the distribution is concentrated on the left of the figure (the right tail is longer) which is said to be *right-skewed*, *right-tailed*, or *skewed to the right*. In case of the distribution is symmetric, then the mean is equal to the median and the distribution will have close to zero *skewness*. For a sample of n values the *skewness* is equal to

$$skew = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}} \tag{4}$$

where, $\{ x_1, x_2, \ldots, x_n \}$ are the n observed values, and $\bar{x}$ is the mean value of these observations.



(a) Negative skew (b) Positive skew

Figure 2. Example of Skewed Distributions

The *Kurtosis*, as specifically measuring the heaviness of the tail, can also be interpreted as the extent to which the distribution of the variable falls off relatively slowly or rapidly near the extremes [15]. As such, longer fatter tails, and often (but not always) a sharper peak are *high kurtosis* distributions; similarly, a *low kurtosis* distribution has shorter, thinner tails, and often (but not always) a more rounded peak. For a sample of n values the *Kurtosis* is equal to

$$kurtosis = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 \tag{5}$$

where, $\{ x_1, x_2, \ldots, x_n \}$ are the n observed values, and $\bar{x}$ is the mean value of these observations. A perfectly normally distributed probability density function has kurtosis equal to zero.

Measure of Order

Relates to Percentile-Rank functions which can be used to describe the probability that a real-valued random variable x with a given probability distribution will be found at a value less than or equal to X [16]. *Percentiles* represent the area under the normal curve; the 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third

quartile (Q3). It can be computed as an integral of the probability density function as follows:

$$F(x;\ \mu,\sigma^2) = \ \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \quad (6)$$

where, *erf* is the special function of sigmoid shape related to the integral of the standard normal distribution.

EMPIRICAL STUDY

This section presents the experimental evaluation procedure that was derived in order to compare the algorithms and the results of the evaluation are discussed.

Dataset

The experimental data comes from an in-house movie recommendation system built for research purposes. The database currently consists of 27 users who provided 46 ratings in the range of 1(min) to 5(max) to 25 movies. The lowest sparsity level is therefore $(27 \times 25) - 46 / (27 \times 25) \approx 0.93$. The prediction algorithms are tested over a pre-selected 26-ratings set. The actual dataset to the case study was kept small for simplicity and expediency once this paper focuses on the evaluation method, not specific results attained.

Simulation Model

The simulation model is accomplished by generating numerous runs with random input rating values (step 1) in the range of 1 to 5 and, for each run, determining the error and improvement associated in predicting the results (steps 2 and 3), to finally compute the complete summary statistics of all runs to report on the outcome variability.

Step 1 – Input parameter

The computation of similarity metric takes as input a user-to-item matrix of size $m \times n$ in which the i-th row of m total number of users contains the rating values of the i-th user against every other item of n total number of items.

Step 2 – Parametric Prediction Model

The baseline prediction is computed using Pearson's correlation coefficient:

$$pSim(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\ .\ (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 .\ \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (7)$$

where, n is the total number of commonly rated items, x_i and y_i represent the current rate of a pair of items of two individuals x and y ($i = 1$ to m), and $\bar{x}$ and $\bar{y}$ represent the average of all of those rates. The second similarity metric, which influences standard recommendation accuracy, is a compound weighting that combines baseline similarity (Eq. (7) with a modifier metric in an aggregation function. For practicality, the modifier metric formulation $m(x,y)$ was based on a previous study [17], and aggregated as a harmonic function, as follows:

$$wSim(x,y) = \frac{2\big(pSim(x,y).m(x,y)\big)}{pSim(x,y) + m(x,y)} \quad (8)$$

Next, the classic last step of Collaborative Filtering computes the final prediction, as follows:

$$Pred(x,y) = \bar{x} + \frac{\sum_{i=1}^{k} wSim(x,y)\ .\ (y_i - \bar{y})}{\sum_{i=1}^{k} wSim(x,y)} \quad (9)$$

where, the predicted rating of item i for the current individual x is the weighted sum of the ratings given to item i by k neighbours y of x; in the proposed algorithm, all y neighbours of individual x are considered, that is, $k = n$.

Step 3 – Output parameter

The simulation considers two response variables. The computation of the numerical deviation of the predicted ratings from the respective actual individual rating is given by the Mean Absolute Error (MAE), as follows:

$$MAE_{baseline\ |\ modified} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

where, n is the number of observations, y_i is the prediction/calculated and $\hat{y}_i$ is the true/observed value. Predictions' overall perceived benefits (gain or loss) between the two strategies are given by:

$$Benefit = \frac{MAE_{modified} - MAE_{baseline}}{MAE_{baseline}}\ (\%) \quad (11)$$

RESULTS AND DISCUSSIONS

Figure 3 shows a visual representation of the simulation data results as histograms. Figure 3.a represents the prediction error response variable, and Figure 3.b depicts the improvement outcome variable. For each of the two variables, an array of N (=25 and =17, respectively) evenly spaced numbers was created as bins. The number of times a particular result occurred on each bin was recorded. To fit the histogram with a cumulative probability distribution, it was necessary to scale the histogram so that the area under the curve is equal to 1. To scale the histogram, the following method was employed: *Scaled = (Count/Points) / (BinSize)*. Once the scaled histogram is plotted, it is possible to glean a lot of good information from it. For instance, Figure 3.a suggests that there are about 50-50% chances that the modified prediction strategy outperforms the classical approach; the uncertainty in MAE is quite large, varying between 0.600 to 0.830; the distribution does not look like a perfect Normal distribution (right-skewed). Likewise, Figure 3.b indicates that the modified strategy may outperforms the traditional approach most of the time but the uncertainty associated with the performance gain/loss seems to be very large to make such assumption. Nevertheless, the benefit distribution does not look like a perfect Normal distribution either. Moreover, the distribution is skewed to the left, suggesting that the horizontal axis data are in reverse order, as some shape similarity between both charts was expected. This issue is confirmed when observing Table 2 and Table 4 calculations where kurtosis and positive performance figures are

inverted, respectively. Both histograms do not appear to have outliers, truncation, multiple modes, etc.

Even though the histograms tell a very good story about the models' behavior, a more pragmatic approach is to estimate the probability of being below or above some values, or between a set of specification limits. Table 5 to Table 5 show the summary statistics of the simulation results that were derived for that purposes.

	MAE	Benefit
Sample Size (runs)	350	350
Mean	0.694	0.1 %
Median	0.690	0.8 %
Min	0.613	11.8 %
Max	0.830	- 19.5 %

Table 1. Central Tendency (Location)

	MAE	Benefit
StDev	0.035	5.0 %
Skewness	0.592	- 0.592
Kurtosis	0.621	0.621

Table 2. Spread and Shape

		MAE	Benefit
Q (.025)	k = 0.05	0.634	-10.2 %
Q (.975)		0.766	8.8 %
Q (.475)	k = 0.95	0.687	0.4 %
Q (.525)		0.692	1.1 %

Table 3. Quantiles, Percentiles, Intervals

	MAE	Benefit
Pr (y > Traditional)	46%	54 %
Pr (min < y < Traditional)	54%	46 %

Table 4. Probabilities

	MAE	Benefit
Lower Conf. Limit	0.691	- 0.5 %
Upper Conf. Limit	0.698	0.6 %
(Significance Level α = 95%)		

Table 5. Confidence Interval for the Means

This case study has focused on the effects of uncertainties of ratings alone. However, the recommendation quality of recommenders depends on several factors. Because of that, there are a number of possible extensions to the simulation methodology currently being pursued by us. This includes extending the study to account for the effects of:

- Different data representation schemes such as categorized, normalized or as-collected inputs.
- Data sparsity when all possible ratings are considered in the simulation, and not only "given" ratings.
- Different similarity calculation algorithms such as cosine.
- Different aggregator methods such as addition, subtraction and multiplication as transformation functions to the original recommender formulation.
- Different evaluation metrics such as RMSE, Precision, Recall and F-1 measure.

For this study, the number of simulation scenarios (runs) was determined based on practicality and experience. For the future, we proposed that the simulation continues until a stopping criterion is reached. This can be achieved by establishing a desired precision for the calculations. Since the iterative Monte Carlo simulation technique computes

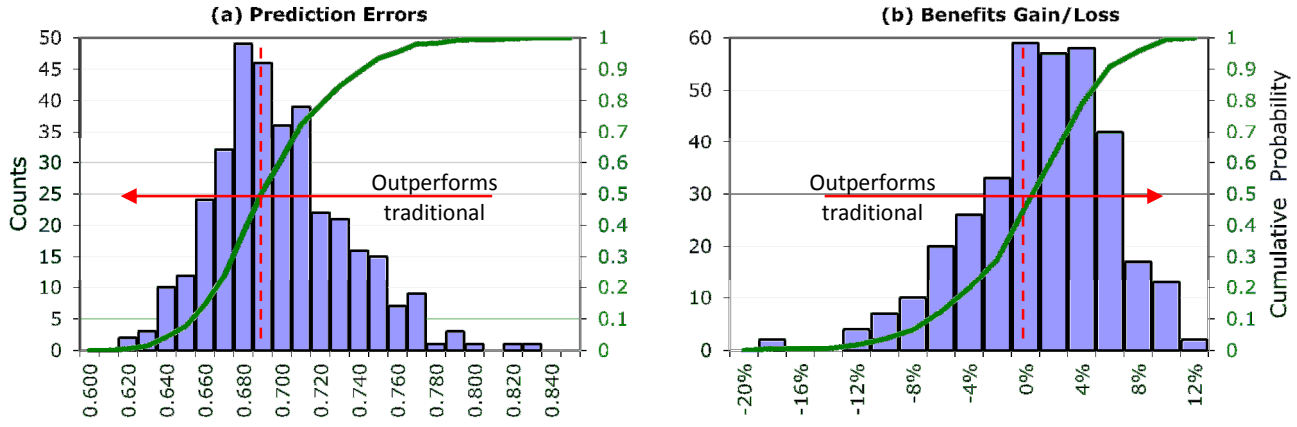


Figure 3. Histograms of Monte Carlo Simulation Results

successive approximations to the solutions, a percentage difference between a computed iterate and all previously computed interactions could limit the maximum amount of time spent iterating.

CONCLUSION

The main purpose of the paper is to suggest a new method to evaluate recommenders using stochastic rather than deterministic approach. It is in this regard that we consider our method to be different and more refined to deal with the complexity associated with the uncertainties in input parameters of recommenders. In addition to providing an estimate of the likely improvement decision of a particular strategy and its variance, the advantage of applying the Monte Carlo simulation technique is that it can provide a more complete assessment of the probability of (under) outperforming at a given level under different conditions. The proposed evaluation model has been successfully applied to a real-world case study project to demonstrate the usefulness of the model and its capabilities over current practice. This work is expected to help researches and practitioners to gain many insights into the performance of recommenders. More specifically, the perceived benefits of the developed model are expected to be improved understanding, higher confidence, longer lasting value, and better depiction of performance indicators of recommender predictions. While this work is focused mainly on the input parameters problem, it can be adapted to any number of parameters that ultimately affect the performance of all particular implementations of recommender solutions.

REFERENCES

[1] M. Papagelis, D. Plexousakis and T. Kutsuras, "Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences," in *Third International Conference in Trust Management (iTrust 2005), LNCS 3477, pp. 224 – 239*, Paris, France, 2005.

[2] K. Mori, "Trust-Networks in Recommender Systems," Master Thesis - Dept. of Computer Science, San Jose State University, 2008.

[3] D. I. K. Sjoberg, T. Dyba and M. Jorgensen, "The future of empirical methods in software engineering research," in *FOSE '07 - Future of Software Engineering*, IEEE Computer Society - Washington, DC, USA, 2007.

[4] J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems,* vol. 22, no. 1, pp. 5-53, 2004.

[5] Y. Cao and Y. Li, "An Intelligent fuzzy-based recommendation system for consumer electronic products," *Expert Systems with Applications,* vol. 87, no. 1, pp. 139-153, 2007.

[6] K. Palavinel and R. Sivakumar, "Fuzzy Multicriteria Decision-Making Approach for Collaborative Recommender Systems," *International Journal of Computer Theory and Engineering,* vol. 2, no. 1, pp. 57-63, 2010.

[7] J. Birnie and A. Yates, "Cost Prediction Using Decision/Risk Analysis Methodologies," *Construction Management and Economics,* vol. 9, no. 2, pp. 171-186, 1991.

[8] S. Weinzierl, "Introduction to Monte Carlo Methods," 23 June 2000. [Online]. Available: http://arxiv.org/abs/hep-ph/0006269.. [Accessed 16 March 2012].

[9] N. Metropolis and S. Ulam, "The Monte Carlo Method," *Journal of the American Statistical Association,* no. 44, pp. 335-341, 1949.

[10] J. M. Hammersley, "Monte Carlo Methods for Solving Multivariable Problems," *Annals of the New York Academy of Sciences,* vol. 86, no. 3, pp. 844-874, 1960.

[11] E. W. Weisstein, "Monte Carlo Method," 01 June 2005. [Online]. Available: http://mathworld.wolfram.com/MonteCarloMethod.html. [Accessed 16 March 2012].

[12] J. W. Wittwer, "Monte Carlo Simulation Basics," 1 June 2004. [Online]. Available: http://www.vertex42.com/ExcelArticles/mc/MonteCarloSimulation.html. [Accessed 16 March 2012].

[13] M. Smithson, Confidence intervals (Quantitative Applications in the Social Sciences Series), vol. No. 140, Belmont, CA: SAGE Publications, 2003.

[14] P. T. Von Hippel, "Mean, Median, and Skew: Correcting a Textbook Rule," *Journal of Statistics Education,* vol. 13, no. 2, 2005.

[15] K. P. Balanda and H. MacGillivray, "Kurtosis: A Critical Review," *The American Statistician,* vol. 42, no. 2, p. 111–119, 1988.

[16] H. A. David and H. N. Nagaraja, Order Statistics, 3rd ed., Hoboken, NJ: John Wiley & Sons, Inc., 2003, p. 458.

[17] R. Capuruço and L. Capretz, "A Fuzzy-based Inference Mechanism of Trust for Improved Social Recommenders," in *Proceedings of the 20th conference on User Modeling, Adaptation, and Personalization: 3rd International Workshop on Social Recommender Systems*, Montreal, Canada, 2012.