DOI:10.1145/3535335

To view the accompanying Technical Perspective, visit doi.acm.org/10.1145/3535334 tp

On Sampled Metrics for Item Recommendation

By Walid Krichene and Steffen Rendle

Abstract

Recommender systems personalize content by recommending items to users. Item recommendation algorithms are evaluated by metrics that compare the positions of truly relevant items among the recommended items. To speed up the computation of metrics, recent work often uses sampled metrics where only a smaller set of random items and the relevant items are ranked. This paper investigates such sampled metrics in more detail and shows that they are inconsistent with their exact counterpart, in the sense that they do not persist relative statements, for example, *recommender A is better than B*, not even in expectation. Moreover, the smaller the sample size, the less difference there is between metrics, and for very small sample size, all metrics collapse to the AUC metric. We show that it is possible to improve the quality of the sampled metrics by applying a correction, obtained by minimizing different criteria. We conclude with an empirical evaluation of the naive sampled metrics and their corrected variants. To summarize, our work suggests that sampling should be avoided for metric calculation, however if an experimental study needs to sample, the proposed corrections can improve the quality of the estimate.

1. INTRODUCTION

Recommender systems are a key technology in online platforms for personalizing the selection of *items* that are shown to a user. Examples include recommending which products to buy, which videos to watch or which songs to play. Recommendations are typically user-dependent and often context-dependent. A key operation of recommender systems is to retrieve a ranked list of the best items for a user in a particular context. This task is called *item recommendation*. Usually, the catalogue of items to retrieve from is large: tens of thousands in academic studies and often millions or more in industrial applications. Finding matching items from this large pool is challenging as the user will usually only explore a few of the highest ranked ones. When building new algorithms for item recommendation, it is crucial to understand how well a particular item recommendation algorithm performs. For evaluating item recommenders, usually sharp metrics such as precision or recall over the few highest scoring items (e.g., top 10), are favored. Other popular choices include average precision or normalized discounted cumulative gain (NDCG), which place a strong emphasis on the top ranked items.

Item recommendation can be very costly for large catalogues because without additional indexing it requires scoring all items. Recently, it has become common in research papers to speed up evaluation by sampling a small set of irrelevant items and ranking the relevant items only among this smaller set.[2,4–6,8–10] Although sampling the loss during training is well-studied,[11] to the best of our knowledge, the implications of sampling during *evaluation* have not been explored, and this work attempts to shed light on the topic. In particular, we show that findings from sampled metrics (even in expectation) can be inconsistent with exact metrics. This means that if a recommender A outperforms a recommender B on a sampled metric, this does not imply that A has a better metric than B when the metric is computed exactly. This problem occurs even in expectation; that is, with unlimited repetitions of the measurement. Moreover, a sampled metric has different characteristics than its exact counterpart. In general, the smaller the sample size, the less difference there is between different metrics, and in the small sample limit, all metrics collapse to the area under the ROC curve (AUC), which discounts positions linearly. This is particularly problematic because many ranking metrics are designed to focus on the top positions, which is not the case for AUC.

As we will show, the sampled metrics can be viewed as high-bias, low-variance estimators of the exact metrics. Their low variance can be particularly misleading if one does not recognize that they are biased, as repeated measurements may indicate a low variance, and yet no meaningful conclusion can be drawn because the bias is *recommender-dependent*, that is, the bias depends on the recommender algorithm being evaluated. We also show that this issue can be alleviated if one applies a point-wise correction to the sampled metric, by minimizing criteria that trade-off bias and variance. Empirical performance of the sampled metrics and their corrections is illustrated on a movie recommendation problem.

This analysis suggests that if a study is really interested in metrics that emphasize the top-ranked items, sampling candidates should be avoided for the purposes of evaluation, and if the size of the problem is such that sampling is necessary, corrected metrics can provide a more accurate evaluation. Lastly, if sampling is used, the reader should be aware that the reported metric has different characteristics than its name implies.

2. EVALUATING ITEM RECOMMENDATION

This section starts by formalizing the most common evaluation scheme for item recommendation. Let there be a pool of

The original version of this paper was published in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

n items to recommend from. For a given instance[a] $\mathbf{x}$, a recommendation algorithm, A, returns a ranked list of the n items. In an evaluation, the positions, $R(A, \mathbf{x}) \subseteq \{1, \ldots, n\}$, of the withheld relevant items within this ranking are computed—R will also be referred to as the *predicted ranks*. For example, $R(A, \mathbf{x}) = \{3, 5\}$ means for an instance $\mathbf{x}$ recommender A ranked two relevant items at positions 3 and 5. Then, a metric M is used to translate the positions into a single number measuring the quality of the ranking. This process is repeated for a set of instances, $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$, and an average metric is reported:

$$\frac{1}{|D|}\sum_{\mathbf{x}\in D} M\left(R\left(A, \mathbf{x}\right)\right). \tag{1}$$

This problem definition assumes that in the ground truth, all relevant items are equally preferred by the user, that is, the relevant items are a *set*. This is the most commonly used evaluation scheme in recommender systems. In more complex cases, the ground truth includes preferences among the relevant items. For example, the ground truth can be a ranked list or weighted set. Our work shows issues with sampling in the simpler setup, which implies that the issues carry over to the more complex case.

3. METRICS

This section recaps commonly used metrics for measuring the quality of a ranking. For convenience, the arguments, A, $\mathbf{x}$, from $R(A, \mathbf{x})$ are omitted whenever the particular recommender, A, or instance, $\mathbf{x}$, is clear from context. Instead, the shorter form R is used.

Area under the ROC curve (AUC) measures the likelihood that a random relevant item is ranked higher than a random irrelevant item.

$$\begin{aligned}\mathrm{AUC}(R)_n &= \frac{1}{|R|\left(n-|R|\right)}\sum_{r\in R}\sum_{r'\in\left(\{1,\ldots,n\}\setminus R\right)}\delta\left(r<r'\right) \\ &= \frac{n-\frac{|R|-1}{2}-\frac{1}{|R|}\sum_{r\in R} r}{n-|R|},\end{aligned} \tag{2}$$

with the indicator function $\delta(b) = 1$ if b is true and 0 otherwise. Precision at position k measures the fraction of relevant items among the top k predicted items:

$$\mathrm{Prec}(R)_k = \frac{\left|\left\{r\in R : r\le k\right\}\right|}{k}. \tag{3}$$

Recall at position k measures the fraction of all relevant items that were recovered in the top k:

$$\mathrm{Recall}(R)_k = \frac{\left|\left\{r\in R : r\le k\right\}\right|}{|R|}. \tag{4}$$

Average Precision at k measures the precision at all ranks that hold a relevant item:

$$\mathrm{AP}(R)_k = \frac{1}{\min\left(|R|,k\right)}\sum_{i=1}^{k}\delta\left(i\in R\right)\mathrm{Prec}(R)_i. \tag{5}$$

Normalized discounted cumulative gain (NDCG) at k places an inverse log reward on all positions that hold a relevant item:

$$\mathrm{NDCG}(R)_k = \frac{1}{\sum_{i=1}^{\min(|R|,k)}\frac{1}{\log_2\left(i+1\right)}}\sum_{i=1}^{k}\delta\left(i\in R\right)\frac{1}{\log_2\left(i+1\right)}. \tag{6}$$

3.1. Simplified metrics

The remainder of the paper analyzes these metrics for $|R| = 1$, that is, when there is exactly one relevant item. We will denote its rank by r. This will simplify the analysis and give a better understanding of the differences between these metrics. The metrics of the previous section simplify to the following:

$$\mathrm{AUC}(r)_n = \frac{n-r}{n-1}, \tag{7}$$

$$\mathrm{Prec}(r)_k = \delta\left(r\le k\right)\frac{1}{k}, \tag{8}$$

$$\mathrm{Recall}(r)_k = \delta\left(r\le k\right), \tag{9}$$

$$\mathrm{AP}(r)_k = \delta\left(r\le k\right)\frac{1}{r}, \tag{10}$$

$$\mathrm{NDCG}(r)_k = \delta\left(r\le k\right)\frac{1}{\log_2\left(r+1\right)}. \tag{11}$$

For metrics such as Average Precision and NDCG, it makes sense to also define their untruncated counterpart, that is, for $k = n$:

$$\mathrm{AP}(r) = \frac{1}{r}, \tag{12}$$

$$\mathrm{NDCG}(r) = \frac{1}{\log_2\left(r+1\right)}. \tag{13}$$

Some other popular metrics can be reduced to these definitions: For $|R| = 1$, *Reciprocal Rank* is equivalent to Average Precision, *Hit Ratio* is equivalent to Recall and *Accuracy* is equivalent to Recall at 1, and Precision at 1.

Figure 1 visualizes how the different ranking metrics trade-off the position versus quality score. Average precision has the sharpest score decay, for example, rank 1 is twice as valuable as rank 2, whereas for NDCG, rank 1 is 1.58 more valuable than rank 2. The least position-aware metric is AUC, which places a linear decay on the rank; for example, improving the ranking of a relevant item from position 101 to 100 is as valuable as an improvement from position 2 to 1.

3.2. Example

This section concludes with a short example that will be used throughout this work. Let there be three recommenders A, B, C and a set of n = 10, 000 items. Each recommender is evaluated on five instances (i.e., $|D| = 5$) with one relevant item each. For each instance, each recommender creates a ranking and the position at which the relevant item appears is recorded. Assume that recommender C manages to rank the relevant item in one of the evaluation instances on position 2; besides this, it never achieves a good rank for the other four instances. Assume recommender B ranks relevant items in two evaluation instances at position 40. And recommender A is never good nor terrible and the relevant items

[a] For example, a user, context, or query.

Figure 1. Visualization of metric versus predicted rank for $n = 10,000$. The left side shows the metrics over the whole set of 10, 000 items. The right side zooms into the contributions of the top 100 ranks. All metrics besides AUC are top heavy and almost completely ignore the tail. This is usually a desirable property for evaluating ranking because users are unlikely to explore items further down the result list.

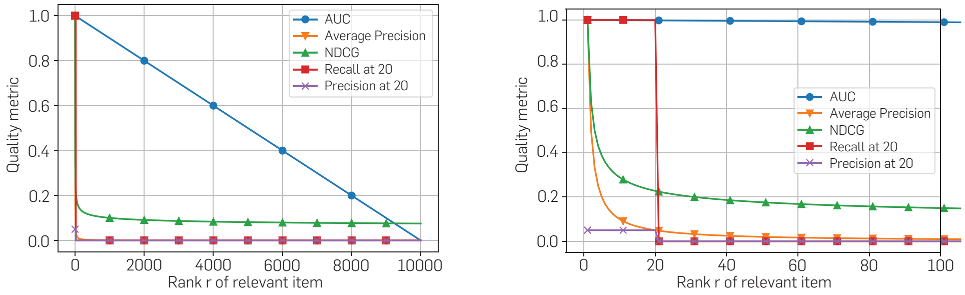


Table 1. Toy example of evaluating three recommenders A, B, and C on five instances.

	Predicted ranks	AUC	AP	NDCG	Recall@10
A	100, 100, 100, 100, 100	**0.990**	0.010	0.150	0.000
B	40, 40, 8437, 9266, 4482	0.555	0.010	0.122	0.000
C	212, 2, 743, 5342, 1548	0.843	**0.101**	**0.208**	**0.200**

are ranked at position 100 in each of the five instances. Table 1 shows more details about the predicted ranks and the corresponding evaluation metrics. On AUC, recommender *A* is the best as it cares about all ranks equally. For top heavy metrics (AP, NDCG, and Recall), recommender *C* scores the highest. This example will be revisited in Section 4.2 when sampled metrics are discussed.

4. SAMPLED METRICS

Ranking all items is expensive when the number of items, *n*, is large. Recently, it has become common to rank only a small set, consisting of the relevant items together with a random sample of *m* irrelevant ones. The metric is then computed on the ranking generated by this subset.[2,4–6,8–10] It is common to pick the number of sampled irrelevant items, *m*, to be orders of magnitude smaller than the number of items *n*, for example, $m = 100$ samples for datasets with $n = \{4k, 10k, 17k, 140k, 2M\}$ items,[2,4,8] $m = 50$ samples for $n \in \{2k, 18k, 14k\}$ items,[5] or $m = 200$ samples for $n \in \{17k, 450k\}$ items.[10] This section will highlight that this approach is problematic. In particular, results can become inconsistent with the exact metrics.

Let $\tilde{R}$ be the ranks of the relevant items among the union of relevant items and the *m* randomly sampled irrelevant ones. It is important to note that $\tilde{R}$ is a random variable, that is, it depends on the random sample of irrelevant items. The properties of $\tilde{R}$ will be analyzed in Section 4.3.

4.1. Inconsistency of sampled metrics

A central goal of evaluation metrics is to make comparisons between recommenders, such as, *recommender A has a higher value than B on metric M*. When comparing recommenders among sampled metrics, we would hope that at least the relative order is preserved in expectation. This property can be formalized as follows.

Definition 1. Let the evaluation data *D* be fixed. A metric *M* is consistent under sampling if the relative order of any two recommenders *A* and *B* is preserved in expectation. That is, for all *A*, *B*,

$$\frac{1}{|D|}\sum_{\mathbf{x}\in D} M\big(R(A,\mathbf{x})\big) > \frac{1}{|D|}\sum_{\mathbf{x}\in D} M\big(R(B,\mathbf{x})\big)$$
$$\Leftrightarrow E\left[\frac{1}{|D|}\sum_{\mathbf{x}\in D} M\big(\tilde{R}(A,\mathbf{x})\big)\right] > E\left[\frac{1}{|D|}\sum_{\mathbf{x}\in D} M\big(\tilde{R}(A,\mathbf{x})\big)\right]. \qquad (14)$$

If a metric is inconsistent, then measuring *M* on a subsample is not a good indicator of the true performance of *M*.

4.2. Example

Now, the example from Section 3.2 is revisited and the same measures are computed using sampling. Specifically, $m = 99$ random irrelevant items are sampled, the position $\tilde{R}$ of the relevant item among this sampled subset is found, and then the metrics are computed for the rank $\tilde{R}$ within the subsample. This procedure with a comparable sample size is commonly used in recent work.[2,4,5,8,10]

Table 2 shows the sampled metrics for the example from Section 3.2. As this is a random process, for better understanding of its outcome, it is repeated 1000 times and the average and standard deviation is computed.[b]

Compared to the exact metrics in Table 1, the relative ordering of metrics completely changed. On the exact metrics, C is clearly the best with a 10x higher average precision than B and A. But it has the lowest average precision when sampled measurements are used. A and B perform the same on the exact metrics, but A has a 2x better average precision on the sampled metrics. Sampled average precision does not give any indication of the true ordering among the methods. Similarly, sampled NDCG and sampled Recall at 10 do not agree with the exact metrics. Only AUC is consistent between sampled and exact computation. The other metrics are inconsistent.

Figure 2 shows the same study as in the previous table, as we vary the number of samples, *m*. The relative ordering of recommenders changes with an increasing sample size. For example, for average precision, depending on the number

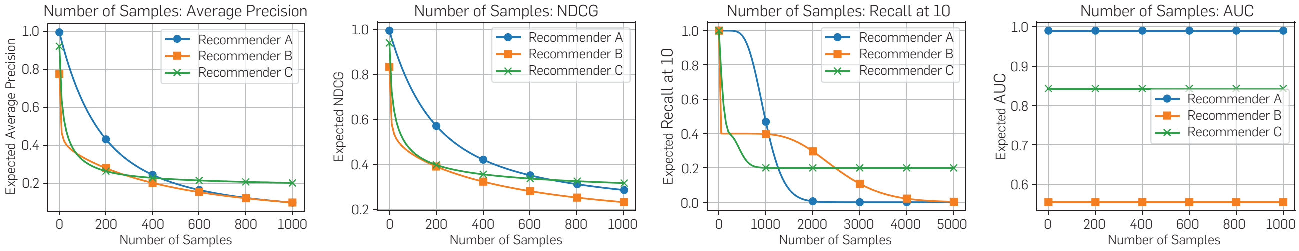
[b] In a real evaluation, the process would not be repeated because this would defeat the purpose of sampling to reduce computational cost.

Table 2. Sampled evaluation for the recommenders from Table 1.

	Predicted ranks	AUC	AP	NDCG	Recall@10
A	100, 100, 100, 100, 100	**0.990** ± 0.004	**0.630** ± 0.129	**0.724** ± 0.097	**1.000** ± 0.000
B	40, 40, 8437, 9266, 4482	0.555 ± 0.014	0.336 ± 0.073	0.444 ± 0.054	0.400 ± 0.000
C	212, 2, 743, 5342, 1548	0.843 ± 0.014	0.325 ± 0.050	0.460 ± 0.039	0.567 ± 0.092

On sampled metrics, the relative ordering of A, B, and C is not preserved, except for AUC.

Figure 2. Expected sampled metrics for the running example (Section 3.2 and 4.2) as the sample size is increased. For Average Precision, NDCG, and Recall, even the relative order of recommender performance changes with the number of samples. That means, conclusions drawn from a subsample are not consistent with the true performance of the recommender.



of samples, any conclusion could be drawn: A better than C better than B (for sample size < 50), A better than B better than C (for sample size ≈ 200), C better than A better than B (for sample size ≈ 500), and finally C better than A equal B (for large sample sizes). This example shows that the bias of sampled average precision is recommender-dependent and sample-size dependent. This is why the relative ordering of recommenders changes as we change the sample size. Similar observations can be made for NDCG. Recall is even more sensitive to the sample size, and it takes about $m = 5000$ samples out of $n = 10{,}000$ items for the metric to become consistent. Only AUC is consistent for all m, and the expected metric is independent of sample size.

4.3. Rank distribution under sampling

This section takes a closer look at the sampling process and derives the distribution of ranks, $\tilde{R}$ and the expected metrics. For simplicity, the analysis is restricted to rankings with exactly one relevant item, that is, $|\tilde{R}| = 1$, so we can use the simplified metrics from Section 3.1. Let r denote the true rank of the unique relevant item, and $\tilde{r}$ denote its measured rank on the sample.

When an irrelevant item is sampled uniformly, it can either rank higher or lower than the relevant item. If the number of all items is n, then the probability that the sampled item j is ranked above r is:

$$p(j<r)=\frac{r-1}{n-1}. \tag{15}$$

For example, if r is at position 1, the likelihood of a random irrelevant being ranked higher is 0. If $r = n$, then the likelihood is 1. Note that the pool of all possible sampled items excludes the truly relevant item and thus has size $n - 1$.

Repeating the sampling procedure m times with replacement and counting how often an item is ranked higher, corresponds to a Binomial distribution. In other words, the rank $\tilde{r}$ obtained from the sampling process follows $\tilde{r}\sim B\left(m,\frac{r-1}{n-1}\right)+1$. If there are no successes in getting a higher ranked item, the rank remains 1, if all m samples are successful, the rank is $m + 1$. The expected value of the metrics under this distribution is

$$E\left[M(\tilde{r})\right]=\sum_{i=1}^{m+1}p(\tilde{r}=i)M(i). \tag{16}$$

Note that this is implicitly a function of r, m, and n, which appear as parameters of the Binomial distribution. Figure 3 visualizes the expected metrics $E(M(\tilde{r}))$ as we vary r. The figure highlights the weight that the sampled metric assigns to different ranks. Metrics such as Average Precision or NDCG are much less top heavy. Even sharp metrics such as recall become smooth. Only AUC remains unchanged. In general, all metrics converge to a linear function in the small sample limit, similar to AUC behavior.

4.4. Expected metrics

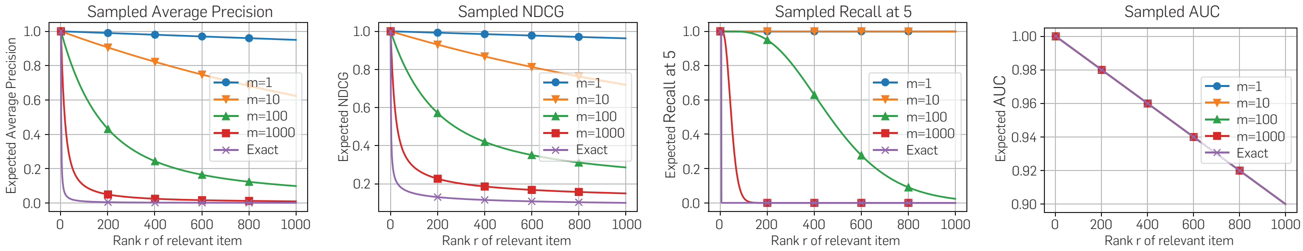
This section analyzes sampled metrics more formally by applying Equation (16) to particular metrics. The discussion focuses on uniform sampling with replacement, that is, Binomial distributed ranks. Similar results hold for uniform sampling without replacement. In this case, the distribution is hyper-geometric, with population size $n - 1$, where a pool of $r - 1$ items can be potential successes. Where appropriate, this variation will be discussed as well.

Expected AUC. First, AUC is a linear function of the rank:

$$\mathrm{AUC}_n(r)=\frac{n-r}{n-1}=-\frac{1}{n-1}r+\frac{n}{n-1}=\mathrm{const}_1 r+\mathrm{const}_2. \tag{17}$$

Thus by linearity of the expectation, and the fact that $\tilde{r}$ follows a Binomial distribution, we have

Figure 3. Characteristics of sampled metrics with a varying number of samples, *m*. Sampled Average Precision, NDCG, and Recall change their characteristics substantially compared to exact computation of the metric. Even large sample sizes (*m* = 1000 samples of *n* = 10000 items) show large bias. Note this plot zooms into the top 1000 ranks out of *n* = 10000 items.



$$E\left[\mathrm{AUC}_{m+1}(\tilde r)\right] = \mathrm{AUC}_{m+1}\left(E\left[\tilde r\right]\right) = \mathrm{AUC}_{m+1}\left(1+m\frac{r-1}{n-1}\right)$$
$$= \frac{m+1-1-m\dfrac{r-1}{n-1}}{m+1-1} = \frac{n-r}{n-1} = \mathrm{AUC}_n(r).$$

That means AUC measurements created by sampling are unbiased estimators of the exact AUC. This result is not surprising because the AUC can alternatively be defined as the expectation that a random relevant item is ranked over a random irrelevant one. Consequently, AUC is a consistent metric under sampling.

This result also holds for any sampling distributions where the expected value of the sampled rank is $1+m\frac{r-1}{n-1}$. For example, this is also true for sampling from a hypergeometric distribution—that is, uniform sampling without replacement.

Cut-off metrics. For a cutoff metric such as recall or precision:

$$E\left[\mathrm{Recall}_k(\tilde r)\right] = \sum_{i=1}^{m+1} p(\tilde r = i)\mathrm{Recall}_k(i) = \sum_{i=1}^{m+1} p(\tilde r = i)\delta(i \le k)$$
$$= \sum_{i=1}^{k} p(\tilde r = i) = \mathrm{CDF}\left(k-1; m, \frac{r-1}{n-1}\right), \tag{18}$$

where CDF denotes the cumulative distribution of the Binomial distribution. This analysis carries over to any sampling distribution, such as the hypergeometric distribution.

Average precision. For the expected value of sampled average precision, we distinguish two cases. If $r = 1$, then $\tilde r = 1$ and the sampled metric is always equal to 1. If $r > 1$, then $p(j < r) > 0$ and

$$E\left[\mathrm{AP}(\tilde r)\right] = \sum_{i=1}^{m+1} p(\tilde r = i)\mathrm{AP}(i) = \sum_{i=1}^{m+1} p(\tilde r = i)\frac{1}{i}$$
$$= \frac{1-\left(1-p(j<r)\right)^{m+1}}{p(j<r)(m+1)} = \frac{1-\left(\dfrac{n-r}{n-1}\right)^{m+1}}{(r-1)\dfrac{m+1}{n-1}}. \tag{19}$$

Interestingly, this can be written as:

$$\frac{1-\mathrm{AUC}_n(r)^{m+1}}{r-1}\left(\frac{n-1}{m+1}\right) = \frac{1-\mathrm{AUC}_n(r)^{m+1}}{r-1}\mathrm{const}.$$

If $AUC_n(r)^{m+1} \approx 0.0$, this would be $\frac{1}{r-1}$ and would be similar to the unsampled average precision metric. However, as soon as the relevant item is reasonably highly ranked (i.e., AUC is close to 1.0), it takes many samples m for this term to approach 0.

Small sample size. This section investigates the behavior of sampled metrics in the limit, where $m = 1$. In this case, $\tilde r \in \{1, 2\}$, and for any metric M and any sampling distribution:

$$E\left[M(\tilde r)\right] = p(\tilde r = 1)M(1) + \left(1 - p(\tilde r = 1)\right)M(2).$$

For uniform sampling[c] of items, $p(\tilde r = 1)$ is the probability to sample an item that is ranked after r, that is, $\frac{n-r}{n-1}$. Now,

$$E\left[M(\tilde r)\right] = \frac{n-r}{n-1}\left(M(1) - M(2)\right) + M(2)$$
$$= r\frac{M(2)-M(1)}{n-1} + \frac{nM(1)-M(2)}{n-1} = r\,\mathrm{const}_1 + \mathrm{const}_2,$$

which is a linear function of the true rank r, regardless of the metric. If we only care about the ordering produced by two different metrics on a set of rankings (Equation (14)), we can ignore const_2. Similarly, for const_1, only the sign matters when comparing two sets of ranking. This sign of $M(2) - M(1)$ depends on how much ranking a relevant item at position 1 is preferred over ranking it at position 2. For metrics that cannot distinguish between the first and second position, such as precision and recall at $k \ge 2$, the sampled metric is always constant and not useful at all. For any reasonable metric, const_1 should be negative, that is, ranking at position 1 gives a higher metric than position 2. To summarize, for $m = 1$, all metrics give the same qualitative result in expectation. There is no reason to choose one metric over the other if we are only interested in relative statements such as "metric of *A* is higher than metric of *B*." Furthermore, the qualitative result with $m = 1$ coincides with exhaustive AUC because (i) all sampled metrics, such as sampled AUC, are indistinguishable for $m = 1$ as shown in this section, and (ii) sampled AUC is consistent with exhaustive AUC as shown in Subsection *Expected AUC*.

The discussion above shows that it does not make sense to choose different metrics for $m = 1$; any sensible metric gives the same qualitative statement. A similar observation can be found in Figures 2 and 3 where all metrics behave similarly for small samples sizes.

[c] Here $m = 1$, so it does not matter whether sampling is with or without replacement.

5. CORRECTED METRICS

So far, we have shown that sampled metrics have different characteristics than the same metric on the full set of items. This section investigates whether we can design a sampled metric $\hat{M}$, a function from $\{1, \ldots, m + 1\}$ to $\mathbb{R}$, such that $\hat{M}(\tilde{r})$ provides a good estimate of $M(r)$. We will consider different definitions of what a "good" estimate is.

5.1. Unbiased estimator of the rank

Our first approach is motivated by a simple observation. The sampled metrics that are commonly used are obtained by applying the exact metric M to the observed rank $\tilde{r}$, that is, $\hat{M}(\tilde{r}) = M(\tilde{r})$. But $\tilde{r}$ is a poor estimate of the true rank r, in fact it always under-estimates it. Instead, one can measure the metric not on the observed rank $\tilde{r}$, but on an unbiased estimator of r. Recall from Section 4.3 that $\tilde{r}|r \sim B\left(m, \frac{r-1}{n-1}\right)+1$. If we let $p := \frac{r-1}{n-1}$, then an unbiased estimator of p is given by $\frac{\tilde{r}-1}{m}$. Thus an unbiased estimator of $r = 1 + (n - 1)p$ is given by $\hat{r} := 1 + \frac{(n-1)(\tilde{r}-1)}{m}$. This motivates using the following corrected metric:

$$\hat{M}(\tilde{r}) = M\left(1 + \frac{(n-1)(\tilde{r}-1)}{m}\right). \tag{20}$$

Because the rank estimate is a real number in $[1, n]$, and the original metric M is only defined on natural numbers, we can either round the rank estimate or extend M using, for example, linear interpolation. In our experiments, we round using floor $\lfloor . \rfloor$.

5.2. Minimal bias estimator

The first correction used an unbiased estimator of the rank. However, whenever M is nonlinear, $\hat{M}(\tilde{r}) = M(\hat{r})$ is biased in general. A criterion one may seek to optimize is the average bias of $\hat{M}(\tilde{r})$, that is, $\sum_r p(r)(E[\hat{M}(\tilde{r})|r] - M(r))^2$, where $p(r)$ is a prior on the distribution of ranks, if available,[d] or the uniform distribution otherwise. Because $\hat{M}$ is a function from $\{1, \ldots, m + 1\}$ to $\mathbb{R}$, $\hat{M}$ can equivalently be viewed as a vector in $\mathbb{R}^{m+1}$. Thus we seek to find a vector $\hat{M}$ that minimizes the following problem:

$$\begin{aligned} &\underset{\hat{M} \in \mathbb{R}^{m+1}}{\arg\min} \sum_{r=1}^{n} p(r)(E[\hat{M}_{\tilde{r}}|r] - M(r))^2 \\ &= \underset{\hat{M} \in \mathbb{R}^{m+1}}{\arg\min} \sum_{r=1}^{n} p(r)\left(\sum_{\tilde{r}} p(\tilde{r}|r)\hat{M}_{\tilde{r}} - M(r)\right)^2. \end{aligned} \tag{21}$$

This is a least-squares problem, and its solution is given by

$$\hat{M} = (A^T A)^{-1} A^T \mathbf{b}, \tag{22}$$

where

$$\begin{aligned} A \in \mathbb{R}^{n \times m+1}, \quad & A_{r,\tilde{r}} = \sqrt{p(r)}p(\tilde{r}|r), \\ \mathbf{b} \in \mathbb{R}^n, \quad & b_r = \sqrt{p(r)}M(r). \end{aligned} \tag{23}$$

Note that the problem is underdetermined when $m+1 < n$, that is, in general, one cannot obtain an unbiased estimator for all r. This is consistent with the observation made in Section 4.4 about small sample sizes, that for the limit case $m = 1$, any metric coincides with (an affine transformation of) AUC.

It may also be desirable for the solution $\hat{M}$ to be monotone nonincreasing, so that on any given evaluation point, a higher rank $\tilde{r}$ results in a lower estimated metric $\hat{M}_{\tilde{r}}$, although this constraint is not essential when averaging over a large number of evaluation points. The monotonic constraint corresponds to the linear inequalities $\hat{M}_{\tilde{r}+1} \geq \hat{M}_{\tilde{r}}$ for all $\tilde{r}$. In this case, problem (21) becomes an isotonic regression problem.[1] We will refer to this as *Constrained Least Squares* in the experiments.

5.3. Bias-variance trade-off

One potential issue with the minimal bias estimator is that it could have high variance, which we observe numerically in Section 6. In order to alleviate this problem, we can regularize problem (21) by introducing a variance term:

$$\underset{\hat{M} \in \mathbb{R}^{m+1}}{\arg\min} \sum_{r=1}^{n} p(r)((E[\hat{M}_{\tilde{r}}|r] - M(r))^2 + \gamma \operatorname{Var}[\hat{M}_{\tilde{r}}|r]), \tag{24}$$

where γ is a positive constant. This is a regularized least squares problem and its solution is given by:

$$\hat{M} = \left((1.0 - \gamma)A^T A + \gamma \operatorname{diag}(\mathbf{c})\right)^{-1} A^T \mathbf{b}, \tag{25}$$

with A and $\mathbf{b}$ are from Equation (23) and $c_{\tilde{r}} = \sum_{r=1}^{n} p(r)p(\tilde{r}|r)$. When $\gamma = 0$, this reduces to problem (21). When $\gamma = 1$, this reduces to the least squares estimator and the solution is

$$\hat{M}_{\tilde{r}} = \frac{\sum_{r=1}^{n} p(\tilde{r}|r)p(r)M(r)}{\sum_{r=1}^{n} p(\tilde{r}|r)p(r)} = \sum_{r=1} p(r|\tilde{r})M(r). \tag{26}$$

In a real study, measurements are aggregated over many evaluation points, which reduce the overall variance, so a lower value $\gamma < 1$ is preferable.

5.4. Example

Figure 4 shows an example of a corrected average precision metric $\widehat{\mathrm{AP}}$, for several choices of the parameter γ, and for a uniform prior $p(r)$. The sample size is $m = 100$ and the full item set is $n = 10000$, that is, a sampling rate of 1%. As can be seen, when no order constraint is applied, lower values of γ give oscillating solutions on the sample (left figure). This is not a problem in aggregate over the full evaluation set (right figure). All corrected sampled metrics are closer, in expectation, to the true metric.

5.5. Effect of the sample size and dataset size

Increasing the sample size m reduces the bias of the sampled metrics, as seen in Figure 3, as well as the corrected metrics: for example, the solution in Equation (21) has a lower bias when optimizing over a higher dimensional vector $\hat{M} \in \mathbb{R}^{m+1}$. Increasing the size of the item set, n, has the opposite effect. Increasing the number of evaluation points, $|D|$, decreases the variance of the average estimates. This mostly benefits the corrected metrics introduced in this section; the uncorrected metrics are high-bias estimators which will have a large error even in the limit of zero variance.

[d] Note that the true distribution of ranks $p(r)$ is algorithm dependent and typically unknown. We use a uniform prior in our experiments.

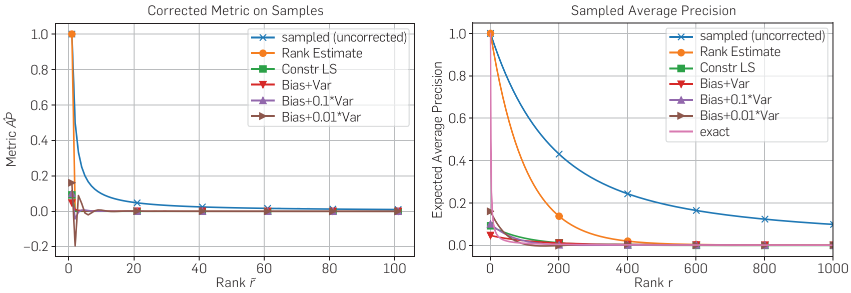


Figure 4. Evaluating the corrected metric $\hat{AP}$ on a sample of $m = 100$ items (left) is equivalent to measuring the metric on the full item set of $m = 10,000$ (right). Different choices of correction algorithms are plotted.

6. EXPERIMENTS

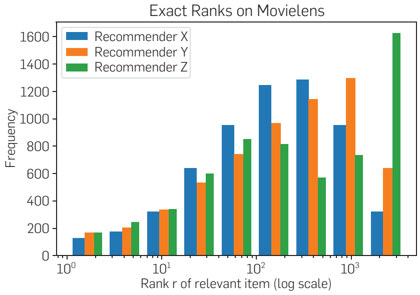
In this section, we study sampled metrics on real recommender algorithms and a real dataset. We investigate the following: (1) Do recommender algorithms create different ranking distributions, for example, some are better in the top, some are better overall? (2) Are results from sampled metrics and exact metrics inconsistent, for example, a given recommender is better on the sampled metric but worse on the true metric? (3) Can corrections help to get more reliable results?

Our study uses the item recommendation evaluation dataset and setup from[4]: The data comes from the movie recommender Movielens[3] where users rate movies. Following the item recommendation protocol, the ratings are ignored, and the task is to recommend movies that a user is likely to rate next. The evaluation protocol removes the most recently rated item from each user's timeline and hides it for evaluation purposes—this item becomes the *relevant* item when measuring the metrics. We use a sampling size of $m = 100$, and we use the metrics Recall@10=HR@10, NDCG@10, AP and AUC. We study the behavior of sampled metrics on three popular recommender system algorithms: matrix factorization and two variations of item-based collaborative filtering (see the appendix in the conference version of this work[7] for details). We want to emphasize that the purpose of this study is not to make a statement about which particular recommender algorithm is good. The purpose is rather to assess the behavior of metrics and correction methods on different algorithms. To deemphasize the particular recommender method and hyper-parameter choice, we will refer to matrix factorization as "recommender X," to the two item-based collaborative filtering variations as "recommender Y" and "recommender Z."

6.1. Rank distributions

For each of the 6040 test users, we rank all items (leaving out the user's training items) and record at which position the withheld relevant item appears. In total, we get 6040 ranks. Figure 5 shows the distribution of these ranks. The plot indicates the different characteristics of the three recommenders. Z is the best in the top 10 but has very poor performance at higher ranks as it puts the relevant items of over 1600 users in the worst bucket. X is more balanced and puts only few items at poor ranks; 2310 items are in the top 100 and less than 300 are in the bottom half. Y is in the middle, with a better top 10 performance than X, but tends to put the relevant item at a worse rank overall.

Figure 5. Distribution of predicted ranks for three recommender algorithms on the Movielens 1M dataset.



6.2. Sampled metrics

The leftmost block in Table 3 reports the exact metric and the sampled metric with standard deviation.[e] As expected from the rank distributions, for Recall, NDCG and AP, recommender Z is better than Y better than X on the exact metric. However, on the sampled metric this does not hold. For sampled Recall, the order is reversed and recommender X is much better than Y which is better than Z. All the measures have low standard deviation, so the issue is not that of variance, but is due to the bias in the sampled metrics. Also for NDCG and AP, the worst recommender on the exact metric (X) appears to be the best according to sampled metrics. The relative ordering of the two better recommenders is correct. For AUC, all sampled results are consistent with the exact metrics.

These results indicate that sampled metrics can be inconsistent in real experiments. In particular, if a study would have compared the recommenders only on the sampled metrics, the study would have drawn the wrong conclusion about the performance of the recommender with respect to top heavy metrics such as Recall, NDCG, and AP. The worst recommender (X) would have been found to be the best one.

6.3. Corrected metrics

We finally investigate if correction methods can help. We consider the three correction strategies proposed in Section 5: *rank estimate* (Equation (20)), *constrained least squares (CLS)* (Equation (21)) with the constraint $\hat{M}_{\tilde{r}} \geq \hat{M}_{\tilde{r}+1}$, and *bias-variance trade-off (BV-γ)* with $\gamma \in \{1.0, 0.1, 0.01, 0.001\}$ and a uniform rank distribution $p(r) = 1/n$.

[e] We repeated the sampling experiment 100 times to measure the variance.

Table 3. Evaluation of three recommenders (X, Y, and Z) on the Movielens dataset.

				Sampled with correction					
	Recommender	Exact	Sampled (uncorrected)	Rank estimate	CLS	BV 1	BV 0.1	BV 0.01	BV 0.001
Recall	X	7.60	66.19 ± 0.25	17.46 ± 0.32	8.52 ± 0.16	4.71 ± 0.07	6.49 ± 0.25	7.18 ± 0.59	7.32 ± 1.17
	Y	8.84	56.51 ± 0.22	17.26 ± 0.28	8.42 ± 0.14	4.60 ± 0.07	6.95 ± 0.21	8.18 ± 0.51	8.54 ± 1.07
	Z	9.42	54.20 ± 0.22	18.67 ± 0.32	9.10 ± 0.15	4.97 ± 0.07	7.44 ± 0.24	8.63 ± 0.59	9.08 ± 1.20
NDCG	X	3.76	39.21 ± 0.20	17.46 ± 0.32	3.99 ± 0.07	2.16 ± 0.04	3.00 ± 0.12	3.34 ± 0.32	3.41 ± 0.71
	Y	4.59	34.82 ± 0.16	17.26 ± 0.28	3.94 ± 0.06	2.12 ± 0.03	3.24 ± 0.10	3.85 ± 0.28	4.03 ± 0.66
	Z	4.79	35.34 ± 0.16	18.67 ± 0.32	4.27 ± 0.07	2.29 ± 0.04	3.46 ± 0.12	4.05 ± 0.32	4.31 ± 0.74
AP	X	3.75	32.55 ± 0.21	18.12 ± 0.31	3.58 ± 0.06	2.44 ± 0.03	3.13 ± 0.09	3.37 ± 0.21	3.42 ± 0.49
	Y	4.32	30.01 ± 0.20	17.81 ± 0.28	3.54 ± 0.06	2.32 ± 0.03	3.19 ± 0.07	3.62 ± 0.18	3.73 ± 0.45
	Z	4.44	30.71 ± 0.21	19.20 ± 0.31	3.82 ± 0.06	2.45 ± 0.03	3.38 ± 0.08	3.79 ± 0.21	3.97 ± 0.51
AUC	X	89.13	89.12 ± 0.04	89.24 ± 0.04	89.12 ± 0.04	88.36 ± 0.04	89.04 ± 0.04	89.11 ± 0.04	89.12 ± 0.04
	Y	85.33	85.33 ± 0.04	85.48 ± 0.04	85.32 ± 0.04	84.63 ± 0.04	85.26 ± 0.04	85.32 ± 0.04	85.32 ± 0.04
	Z	74.73	75.04 ± 0.23	75.24 ± 0.20	75.04 ± 0.21	74.51 ± 0.23	75.02 ± 0.20	75.02 ± 0.24	75.02 ± 0.23

Sampled metrics are inconsistent with the exact metrics. Corrected metrics, especially Bias^2 + γ* Variance with $\gamma \leq 0.1$ produce the correct relative ordering in expectation.

The right block of Table 3 shows the expected metrics under correction. All methods are closer to the exact results than sampling without correction. In particular, CLS and BV with low γ have values close to the exact metric—which indicates a low bias. All identify the order better, for example, all of them place recommender Z as the best performing method for Recall, NDCG and AP. Some of them (BV with low γ) also get the order of recommenders X and Y right. We also conducted experiments with different choices of the sampling size m and observed that uncorrected metrics need more than $m = 1000$ samples (equivalent to $1/3^{rd}$ sampling rate) to correctly order recommenders X and Y, whereas the corrected metric using a bias-variance trade-off with $\gamma = 0.1$ already has the correct ordering with less than $m = 60$ samples.

Although the corrections seem to be effective in expectation, one also needs to consider the variance of these measurements. In addition to the variances reported in Table 3, we ran for each metric pairwise comparisons between any two recommenders and counted how often (over 100 repetitions of the experiment) a sampled metric correctly assigns a higher value to the recommender that performs better on the exact metric (see Table 4 in the conference version of this work[7] for details). For example, for Recall and "X versus Y" we count in how many of 100 repetitions, the sampled Recall of recommender X is worse than Y. As expected, we observe that the corrected metrics with both low variance and low bias, in particular BV 0.1, have a high success rate (>90%) on all but the challenging AP metric for the Y versus Z comparison. Also the simple 'rank estimate' correction is surprisingly effective and is strictly better than the uncorrected metric in all comparisons. It is worth mentioning that under the rank estimate correction, Recall@10 and NDCG@10 are identical (see Table 3). However, it still represents an improvement over the uncorrected metrics which are much more biased and lead to the wrong conclusion. The rank estimate method is trivial to implement (that is, upscaling the rank before applying the metric). In a study with sampled evaluation, this should be preferred over uncorrected metrics. Other corrections such as the adjusted bias-variance can get higher gains but are more difficult to implement.

7. SUGGESTIONS

Our results have shown a sampled metric can be a poor indicator of the true performance of recommender algorithms under this metric. For uncorrected metrics this is mostly due to the large bias introduced by sampling. Using correction methods, this bias can be reduced but at the cost of higher variance. If a study needs to use sampled metrics and is still interested in the true performance of the metrics, we suggest to use a correction method as proposed in this work. In this case, it is important to rerun the experiment with different samples (for example, different random seeds). It is already common, in most evaluations, to repeat an experiment N times—usually by varying the dataset (for example, N-fold cross validation). In this case, variance is introduced by the differences in the dataset split and potentially by the initialization of the recommender algorithm. In a sampled evaluation, adding a different seed for negative sampling will add another source of variance. This means that it may be more difficult to find "statistically significant" differences between two recommenders. If even under the increased variance a difference is found, then this is a stronger indication the recommender is truly better under the exact metric. The lower the bias in the corrected metric (for example, the lower γ), the stronger the indication. Although this evaluation is preferable over uncorrected metrics, it is still prone to either not identifying differences (due to variance) or drawing false conclusions because of the bias. This bias can only be eliminated by avoiding sampling altogether.

8. CONCLUSION

This work seeks to bring attention to some issues with sampling of evaluation metrics for item recommendation. It has shown that most metrics are inconsistent under sampling and can lead to false discoveries. Moreover, metrics are usually motivated by applications, for example, does the top 10 list contain a relevant item? Sampled metrics do not measure the intended quantities—not even in expectation. For this reason, sampling should be avoided as much as possible during evaluation. If an experimental study needs to sample, we propose correction methods that give a better estimate of the true metric, however at the cost of increased variance. Our analysis focused on the case of a single relevant item. The general case may be treated by making the approximation that observed ranks are independent, in which case similar correction methods can be applied. Deriving correction methods without independence is an interesting direction for future research. C

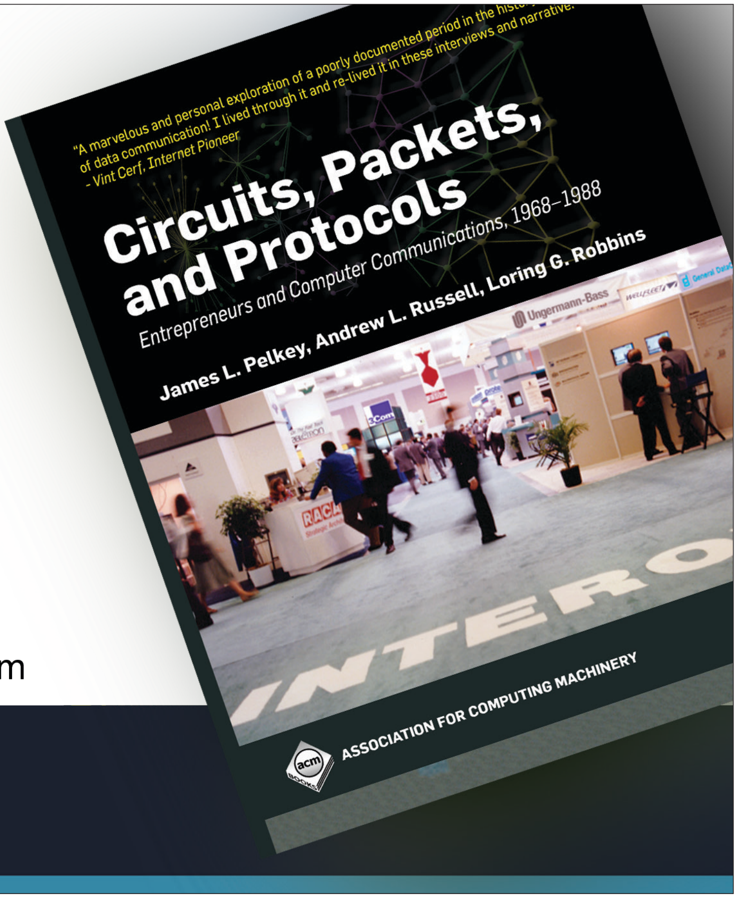
Acknowledgment

We would like to thank Nicolas Mayoraz and Li Zhang for their helpful comments and suggestions.

References

1. Barlow, R., Bartholomew, D., Bremner, J.M., Brenner, H.D. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression.* John Wiley, 1972.
2. Ebesu, T., Shen, B., Fang, Y. Collaborative memory network for recommendation systems. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18* (Ann Arbor, MI, USA, 2018), ACM, New York, NY, USA, 515–524.
3. Harper, F.M., Konstan, J.A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst. 5*, 4 (2015).
4. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.-S. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW '17* (Perth, Australia, 2017), International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 173–182.
5. Hu, B., Shi, C., Zhao, W.X., Yu, P.S. Leveraging meta-path based context for top- n recommendation with a neural co-attention model. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18* (London, U.K, 2018), ACM, New York, NY, USA, 1531–1540.
6. Krichene, W., Mayoraz, N., Rendle, S., Zhang, L., Yi, X., Hong, L., Chi, E., Anderson, J. Efficient training on very large corpora via gramian estimation. In *International Conference on Learning Representations* (2019) OpenReview.net, New Orleans, LA, USA.
7. Krichene, W., Rendle, S. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20* (New York, NY, USA, 2020). Association for Computing Machinery, New York, NY, USA, 1748–1757.
8. Wang, X., Wang, Xu, C., He, X., Cao, Y., Chua, T.-S. Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI), AAAI '19* (2019), 5329–5336.
9. Yang, L., Bagdasaryan, E., Gruenstein, J., Hsieh, C.-K., Estrin, D. Openrec: A modular framework for extensible and adaptable recommendation algorithms. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18* (Marina Del Rey, CA, USA, 2018), ACM, New York, NY, USA, 664–672.
10. Yang, L., Cui, Y., Xuan, Y., Wang, C., Belongie, S., Estrin, D. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18* (Vancouver, British Columbia, Canada, 2018), ACM, New York, NY, USA, 279–287.
11. Yu, H.-F., Bilenko, M., Lin, C.-J. Selection of negative samples for one-class matrix factorization. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (2017), 363–371.

Walid Krichene and Steffen Rendle ([walidk, srendle]@google.com), Google Research, Mountain View, CA, USA.


"A marvelous and personal exploration of a poorly documented period in the history of data communication! I lived through it and re-lived it in these interviews and narrative."
- Vint Cerf, Internet Pioneer
Circuits, Packets, and Protocols
Entrepreneurs and Computer Communications, 1968–1988
James L. Pelkey, Andrew L. Russell, Loring G. Robbins
INTERO
ASSOCIATION FOR COMPUTING MACHINERY