Recommending items with attributes

Anonymous Author(s)

ABSTRACT

Recommendation models suggest items to users. We study a variant of this task where each item has a set of attributes, such as tags on an image, user reactions to a post, or foods in a meal. We propose RANKFROMSETS, a flexible and scalable model for recommending items with attributes. rfs treats item attributes as side information and learns embeddings to discriminate items a user will consume from items a user is unlikely to consume. We develop theory connecting the rfs loss to optimal recall and show that the learnable class of models for rfs is a superset of several previously-proposed models. To scale rfs to large datasets, we propose a stochastic optimization framework based on negative sampling. Within this framework, we suggest two specific approaches that trade off computation for accuracy. In experiments on two real-world datasets, rfs outperforms competing approaches while also learning embeddings that reveal interpretable structure in user behavior.

1 INTRODUCTION

Classical recommender system datasets contain a matrix where each row is a user and each column is an item. Each entry in the matrix indicates whether or not the user consumed the item. Modern applications often gather rich side information about items in the form of a set of attributes or tags. Intuitively, item attributes provide valuable side information for recommender systems. When the number of items is large or the user-item matrix is sparse, leveraging the attribute information is crucial to achieving good performance.

In practice, modeling item attributes in a recommender system is not straightforward. Standard multiple-matrix factorization methods may fail if the attribute vocabulary or item set is too large. At the same time, oversimplifying the problem risks sacrificing the ability to capture nonlinear statistical dependencies between attributes. For instance, a user may enjoy items tagged with attributes A and B, B and C, or A and C, but not all three. Finding the right balance between scalability and flexibility is therefore a primary goal.

Even when a model can be scaled, it is not always clear how the training procedure connects to the recommender system evaluation metric. It may be that a model with optimal loss leads to suboptimal recommender system performance. For example, a matrix factorization method may minimize mean squared error but the recommender system may be evaluated on precision. While it is plausible that minimizing error will improve precision, the connection between the two is implicitly assumed in many methods. Ideally, a well-designed recommender system should have an optimization loss that matches its evaluation metric.

This paper proposes RANKFROMSETS (RFS), a scalable model for recommending items with attributes. RFS casts the recommendation problem as binary classification. Given an item and a user, RFS treats attributes as features and classifies whether or not the item is likely to be consumed by the user. RFS learns embeddings for each user and attribute; each item is represented as the mean of its attribute embeddings. To scale to large datasets, RFS is trained using a stochastic optimization procedure that randomly samples items that are unlikely to be consumed.

RFS enjoys two benefits from framing the recommendation problem as classification. First, the RFS classification objective function is directly tied to recommender recall, and we show that a classifier with zero worst-case error achieves maximum recall. Second, the model is provably flexible enough to learn any class of recommendation model based on set-valued side information. Thus, the RFS model generalizes other models while retaining scalability.

We study RFS on two datasets. We consider a dataset of 65k users clicking on 637k papers posted to the arXiv; the attributes of a paper are the unique words in its abstract. RFS outperforms a state-of-the-art content-based matrix factorization model [7] on this dataset, and the learned document embeddings capture the structure of human-labeled topics. Second, we consider a large dataset of 55k users logging 16M meals using the LoseIt! diet tracking app. Again, RFS outperforms other baseline models, and the learned meal embeddings reveal a diverse set of recommendations that capture the notion of taste.

2 BACKGROUND

We highlight two themes in research on recommendation models. First, we describe several recommendation models that incorporate side information and note why they do not scale to datasets with many items. We also describe models that optimize proxies of ranking metrics and discuss why these proxies are not directly tied to optimal recommendation evaluation metrics.

2.1 Recommendation with side information

Side information is included in recommendation models in several ways; we focus on deep learning and matrix factorization approaches. Item side information can be modeled with deep representations, or can be included in content-based matrix factorization models as an additional matrix.

Deep representations of side information. Deep learning-based recommendation models incorporate side information in multiple ways [34]. For example, items that have words as attributes can be represented using recurrent neural networks [1]. Lian et al. [14] use

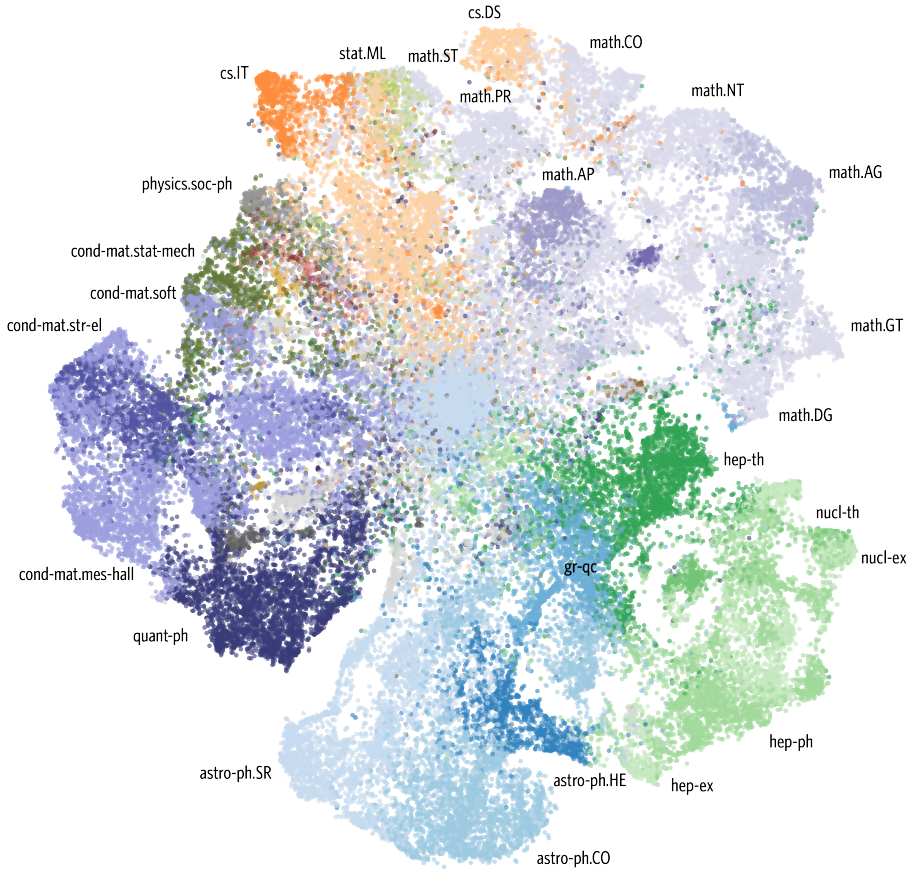


Figure 1: RANKFROMSETS trained on arXiv reading behavior clusters researchers by their most frequently-read arXiv category (best viewed in color). On this data, RANKFROMSETS is trained to recommend items using their attributes as described in Section 4.2; the set of attributes for a paper is the set of words in the abstract. A visualization of the user embeddings in the inner product regression function in Equation (4) yields an interpretable map of science. We use the t-SNE algorithm [18] for visualizing the high-dimensional embeddings in two dimensions. In this map of science, fields of study are related according to patterns in how people read papers in neighboring fields. Each marker represents a user embedding; the color assigned to a user is determined the user's most-read arXiv category. The color assigned to a category is determined by the most-read categories across the arXiv, with similar colors assigned to similar fields according to the arXiv ontology. For an interactive version of this map, please visit the anonymous URL https://app.box.com/v/rankfromsets. The interactive version enables zoom and display of all 143 arXiv category labels to explore the relationships between different fields of science.

an attention mechanism to weight recommendations according to available item and user side information, and Dong et al. [6] use denoising autoencoders to model side information in a deep recommendation model, but requires fitting parameters for every item. An example of a more efficient approach is the method in Chen et al. [5], where embeddings are jointly learned for users, items, and item text for recommendation, but this method focuses on unsupervised pre-training of text representations. There are several examples of 'tag-aware' or 'tag-based' deep recommendation models [16, 36], such as Xu et al. [29], which focuses on data where users and items have different attributes and use autoencoders to learn user, item, and attribute representations. They use a cosine similarity-based objective function which is not tied to a metric used to evaluate recommendation performance.

Matrix factorization with side information. Shi et al. [23] survey several matrix factorization methods that incorporate side information. Gopalan et al. [7] develop a Bayesian matrix factorization model for recommending items based on side information in the form of words in documents. Wang and Blei [28] develop a regression model that uses a topic model to incorporate side information into recommendations. There are also several 'tag-based' or 'tag-aware' content-based matrix factorization models [2, 17, 35]. Such

content-based matrix factorization methods maximize the conditional log-likelihood of the data (or a bound on the log-likelihood); optimizing these objective functions does not optimize an evaluation metric. Furthermore, all of these methods are not scalable to large numbers of items as they require learning unique parameters for every item. Specifically, such content-based matrix factorization methods require learning a matrix that has a row for every item. For items with attributes, it is often infeasible to store this matrix in memory or exploit efficient coordinate ascent optimization schemes that require processing this entire matrix.

2.2 Recommendation via ranking

Recommendation models can be trained on loss functions that approximate ranking-based evaluation metrics. We describe recommendation models trained on such loss functions and extensions that include side information.

Learning to rank. The literature on learning to rank includes models that optimize proxies of evaluation metrics, such as mean average precision, mean reciprocal rank, or discounted cumulative gain [15, 31]. For example, Bayesian personalized ranking models optimize a pairwise ranking objective function [21] that trains the model to rank items a user consumed higher than items a user did not consume. This objective is a heuristic motivated by an analogy to the receiver operating characteristic; a model trained on this objective does not provably maximize this metric. Song et al. [24] extend Bayesian personalized ranking using deep neural networks, but do not model side information.

Learning to rank with side information. Models that optimize proxies of ranking metrics that use side information include Shi et al. [22], where a smoothed approximation of mean average precision is used as a loss function. Yuan et al. [32] use a proxy of a ranking loss to fit a polynomial that models predictions of item consumption using item and side information features. Ying et al. [30] uses denoising autoencoders to represent item information in a model trained with a pairwise ranking loss. Cao et al. [4] use a ranking loss to jointly learn embeddings of items and attributes; they focus on the case where users interact directly with both attributes and items with said attributes. All of these models require learning unique parameters for every item, and do not scale to large numbers of items. An example of a scalable method that uses the Bayesian personalized ranking criterion is in Okura et al. [20], but this approach requires data with timestamps and negative item feedback.

3 THE RANKFROMSETS MODEL

RANKFROMSETS is a recommendation model that recommends items with attributes to users. Let $u \in \{1, \ldots, N\}$ be a user, $m \in \{1, \ldots, M\}$ be an item, and $y_{um} \in \{0, 1\}$ be a binary indicator where 1 indicates user u consumed item m. For each item m, there is an associated set of attributes $x_m \in \{0, 1\}^{|V|}$ from a vocabulary of V attributes.

We assume that a recommendation model is given a budget of K recommendations to be made for each user. In response, the recommender system produces a list of K distinct recommendations $\mathbf{r}_u = (r_{u1}, \ldots, r_{uK})$ for each user. The goal of the recommendation task in this paper is to maximize the expected recall,

$$\text{Recall@}K = \mathbb{E}_u \left[\frac{\sum_{r \in \mathbf{r}_u} y_{ur}}{\sum_m y_{um}} \right], \tag{1}$$

where the expectation is over all users in the empirical data distribution $\mathcal{D}$. RANKFROMSETS (RFS) combines three techniques to maximize Equation (1). First, we cast recommendation as a classification task. Second, we learn user- and attribute-level embeddings. Statistical strength is shared between items with similar attributes by representing items as the mean of their attribute embeddings. Third, we scale RFS to large datasets by using a stochastic optimization-based negative sampling training procedure to fit the model.

3.1 Recommendation through classification

RANKFROMSETS (RFS) casts the recommendation problem as a classification task. Given a user-item pair (u, m), RFS learns to predict the probability that item m will be consumed by user u,

$$p(y_{um} = 1 | u, m) = \sigma\left(f\left(u, x_m\right)\right), \tag{2}$$

where x_m is the set of attributes associated with item m and σ is the sigmoid function. The recommendations made by RFS are the maximum likelihood set,

$$\mathbf{r}_u(K) = \operatorname*{argmax}_{\mathbf{r} \in \mathbb{N}^K} \sum_{m \in \mathbf{r}} f\left(u, x_m\right). \tag{3}$$

To motivate treating recommendation as classification, we make the following observation.

PROPOSITION 1. *Let $u \in \mathcal{U}$ be a user, $x \in \mathcal{X}$ be an item, and $y(u, x) \in \{0, 1\}$ be an indicator of whether user u logged item x. Let $\mathcal{E}$ be the worst-case error for binary classifier $\hat{y}(u, x)$ on any (u, x) pair drawn from the data $\mathcal{D}$,*

$$\mathcal{E} = \max_{(u,x) \in \mathcal{D}} 1\left[\hat{y}(u, x) \neq y(u, x)\right].$$

A binary classifier with zero worst-case error ($\mathcal{E} = 0$) maximizes recommendation recall.

PROOF. A model with zero worst-case error is a perfect classifier: it assigns greater probability to positively-labeled datapoints than to negatively-labeled datapoints. In other words, it ranks positive examples above negative examples. Recall in Equation (1) is measured by the fraction of positively-labeled items in a ranking returned by the model. In a classifier that achieves zero worst-case error, positively-labeled datapoints must be ranked higher than other datapoints, maximizing recall. □

Proposition 1 is simple, but conceptually important. Under the assumption that a perfect classifier exists, a consistent method for learning a classifier will be a consistent method for learning a recommendation system that targets expected recall. In practice, as

with any regression method, a perfect classifier is unachievable. Proposition 1 is a guiding principle rather than a finite-sample guarantee of maximal performance. As we show in Section 4, the classification approach of RFS performs well in practice.

3.2 Embedding users and items

For recommending items with attributes, Proposition 1 says that building a classifier such as RANKFROMSETS is optimal if we measure recommendation performance with recall. To parameterize the RFS classifier, we need to choose a regression function $f(u, x_m)$. A straightforward parameterization is an inner product,

$$f(u, x_m) = \theta_u^\top \left(\frac{1}{|x_m^+|} \sum_{j \in x_m^+} \beta_j + g(x_m) \right) + h(x_m) . \quad (4)$$

Each element in the inner product regression function in Equation (4) has an intuitive interpretation:

- The user embedding $\theta_u \in \mathbb{R}^d$ captures the latent preferences for user u. This captures the individual-level tastes of a user and is analogous to the user preference vector in classical collaborative filtering or the row embedding in matrix factorization.
- The attribute embedding $\beta_j \in \mathbb{R}^d$ is the latent quality conveyed through item m having attribute j. The set x_m^+ contains only attributes with $x_{mj} = 1$. Attributes that are not associated with item m are ignored.
- The item embedding function $g(x_m) \in \mathbb{R}^d$ captures qualities not conveyed through the set of item attributes. This term in the regression function enables collaborative filtering by capturing unobserved patterns in item consumption such as popularity. We describe how to construct this function below.
- The item intercept function $h(x_m) \in \mathbb{R}$ makes an item more or less likely simply due to availability.

Scalable item embedding and item intercept functions. The parameterization of the item embedding function $g(x_m)$ depends on the size of the data. If the number of items is small, g can function as a lookup for unique intercepts for every item. However, if the number of items is so large that unique item intercepts lead to overfitting, we need a scalable parameterizations of item embeddings g using additional information about every item. For example, if the data consists of foods in meals, we can define a meal intercept as the mean of food intercepts, yielding a scalable item intercept function. The item intercept function $h(x_m)$ that maps item attributes to scalars is constructed in the same way. We study both of these choices in Section 4.

The inner product regression function in Equation (4) has several benefits. It requires computing a sum over only the attributes each item is associated with. This enables RFS to scale to large attribute vocabularies where traditional matrix factorization methods are intractable. Second, the embed-and-sum approach to set modeling is provably flexible. We describe deep variants of RFS and detail how RFS can approximate other recommendation models in the next section.

3.3 Deep extensions

The RFS inner product regression function in Equation (4) is a log-bilinear model. But there are several other choices of regression function, and we draw on the deep learning toolkit for classification to build two other example architectures. With finite data and finite compute, one architecture will outperform another, or prove insufficiently flexible to capture patterns in user consumption. The optimal choice of architecture depends on computational tradeoffs. We conclude this section by showing that in the regime of infinite data and compute, all the RFS architectures we propose can approximate other recommendation models that operate on set-valued input such as matrix factorization.

Parameterizing RFS using a neural network. As an alternative to the log-bilinear model in Equation (4), we can use a deep neural network as a regression function:

$$f(u, x_m) = \phi \left(\theta_u, \frac{1}{|x_m^+|} \sum_{j \in x_m^+} \beta_j, g(x_m) \right) + h(x_m) , \quad (5)$$

where the deep network ϕ has weights and biases and takes as inputs the user embedding, sum of attribute embeddings, and item intercept. Such a neural network can represent functions that may or may not include the inner product in Equation (4) (*ex ante*, it is unclear whether a finite-depth, finite-width neural network can represent the inner product).

Parameterizing RFS with a residual network. Another possibility of regression function for RFS is a combination of Equations (4) and (5), using an idea borrowed from deep residual networks for image classification [9]. In this architecture a neural network ϕ with the same inputs as in Equation (5) is used to learn the residual in the inner product model:

$$f(u, x_m) = \theta_u^\top \left(\frac{1}{|x_m^+|} \sum_{j \in x_m^+} \beta_j + g(x_m) \right) + \phi + h(x_m) . \quad (6)$$

The choice of regression function in RFS depends on the data. On finite data, with finite compute, one parameterization of RFS will outperform another. To demonstrate this, we simulated synthetic data from the same generative process RFS employs with a ground-truth regression function (a square kernel) where the residual and deep parameterizations outperformed the inner product architecture. These results are included in Appendix D, and motivate exploring other architectures than the three examples we give.

3.4 Generalization property

If we take a step back from the setting of finite data and compute, a bigger picture emerges, which reveals the choice of regression function in RFS does not matter. Any RFS architecture is sufficiently flexible to approximate other recommendation models that operate on set-valued input. Before applying the theoretical results derived in Zaheer et al. [33] to RFS, we first describe the class of recommendation models that operate on set-valued input.

RFS and content-based matrix factorization are examples of permutation–invariant models. The regression function f in RFS operates on set-valued input: the unordered collection of item attributes x_m^+. A set is, by definition, permutation–invariant: it remains the same if we permute its elements. Functions that operate on set-valued inputs must also be permutation–invariant. RFS is permutation–invariant as the set of attributes associated with an item enter into Equations (4) to (6) via summation. Other examples of permutation–invariant models are matrix factorization, recommendation models based on word embeddings, and permutation-marginalized recurrent neural networks. We show that these models are permutation–invariant in Section 4.1, and evaluate them in Section 4.

RFS can approximate other permutation-invariant models such as matrix factorization. We apply a theorem from Zaheer et al. [33] to show that RFS can approximate other recommendation models.

PROPOSITION 2. *Assume the vocabulary of attributes (set elements) is countable, $|V| < |\mathbb{N}_0|$. RFS can approximate any permutation-invariant recommendation model.*

The proof follows directly from Theorem 2 in Zaheer et al. [33] and we will not restate it here. (The only change to the proof is the mapping from set elements to one-hot vectors, $c : V \to \{0,1\}^{|V|}$ to yield a unique representation of every object in the powerset.)

Proposition 2 means that any of the parameterizations in Equations (4) to (6) are flexible enough to approximate other principled recommendation models that leverage item attributes, such as content-based matrix factorization [7]. This proposition also supports exploring other parameterizations of RFS that may have better computational or statistical properties.

3.5 Stochastic optimization

Denote the parameters of the RANKFROMSETS model by θ, and let $\mathcal{D}$ be the empirical data distribution. The model parameters are the user embeddings, attribute embeddings, parameters of the item embedding g and item intercept h functions, and any weights and biases of the neural networks in Equations (5) and (6). The maximum likelihood objective function is

$$\mathcal{L}(\theta, \mathcal{D}, \lambda_u) = \mathbb{E}_u \Big[\mathbb{E}_{(x_m, y_{um}) \sim \mathcal{D} \mid u, y_{um}=1} \left[\log p(y_{um} = 1 \mid x_m; \theta) \right] \tag{7}$$

$$- \lambda_u \mathbb{E}_{(x_k, y_{uk}) \sim \mathcal{D} \mid u, y_{uk}=0} [\log p(y_{uk} = 0 \mid x_k; \theta)] \Big] \tag{8}$$

$$\tag{9}$$

The parameter λ_u is a reweighting parameter for negatively-labeled datapoints, and it is derived as follows. This classification objective function derives from an assumption about the data-generating process. We assume that every user-item interaction is unique and every item is uniquely determined by its attributes. A user u does or does not consume an item m; there is no noise in this process. This means there is a single label for every datapoint (x_m, y_{um}). This assumption means that a classifier trained on Equation (7) is consistent: in the infinite data limit, any reweighting of λ_u will preserve the ordering of items, as there is only a single label.

Stochastic optimization and per-user negative sample balance. We use stochastic optimization to maximize Equation (7). In practice, the number of negatively-labeled datapoints is large, and we find that balancing the positive and negative labels per-user helps optimization. We describe two ways to fix λ_u below, using two negative sampling schemes that are dependent on the choice of evaluation metric.

User balance via corpus sampling. Negative samples can be drawn uniformly over the entire corpus of items. If the item set is large, this can be an expensive procedure. However, in related applications [27], corpus sampling was shown to substantially outperform other sampling strategies. This negative sampling scheme and resulting approximation of Equation (7) is similar to the objective functions in other recommender systems [10, 24].

On large datasets, it is infeasible to calculate recall for evaluation, as Equation (1) requires ranking every item for every user (e.g. in Section 4.3 we study a dataset with over 10M items). We define a scalable evaluation metric based on recall, and describe how it leads to a natural choice of negative sampling distribution.

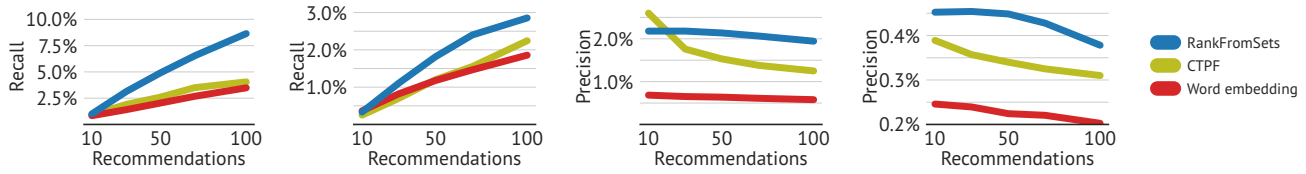
Sampled recall. Start with held-out datapoints with positive labels, $(x_m, y_{um} = 1)$. Sampled recall is calculated as follows. For every held-out datapoint, $K-1$ datapoints with negative labels $(x_k, y_{uk} = 0)$ are sampled from the rest of the held-out data, which together yield a set of K datapoints. A recommendation model is used to rank the K datapoints $r_{u1}, \ldots, r_{uK}$. Sampled recall is the fraction of the held-out datapoints that the model ranked in the top k:

$$\text{SampledRecall@}k = \mathbb{E}_{um} \left[\frac{\sum_{r \in \{r_{u1}, \ldots r_{uk}\}} y_{ur}}{K} \right], \tag{10}$$

where the expectation is over users and items in the held-out set of datapoints. This evaluation metric is scalable: instead of using a model to rank every item, it requires ranking K items. Sampled recall is 1 if $k = K$, as the held-out datapoints with $y_{um} = 1$ are in the list of K datapoints to be ranked.

User balance via batch sampling. When sampled recall is used as an evaluation metric, batch sampling is a natural way to draw negative samples. Sampled recall is calculated on items drawn from other user's data. We define batch sampling as generating negative samples by permuting mini-batch items. Besides corresponding to the sampled recall metric, this technique is memory-efficient, as it requires that only the current mini-batch be in memory.

In addition to scalability, both of the negative sampling procedures above have the advantage of implicitly balancing the classifier. As Veitch et al. [27] note, using stochastic gradient descent with negative sampling is equivalent to a Monte Carlo approximation of the reweighted, or balanced, classification loss.



(a) In-matrix recall **(b) Out-matrix recall** **(c) In-matrix precision** **(d) Out-matrix precision**

Figure 2: RANKFROMSETS with the inner product regression function in Equation (4) outperforms collaborative topic Poisson factorization (CTPF, described in Gopalan et al. [7]) and a word embedding model on recommending arXiv papers to scientists. In this dataset the items are documents and the attributes are the unique words in the abstracts. Recommendation performance is evaluated using precision and recall to match the evaluation in Gopalan et al. [7], and these metrics are reported on training (in-matrix) documents and cold-start (out-matrix) documents with no clicks in the training set.

4 EXPERIMENTS

We study the performance of RANKFROMSETS on two datasets. The first data consists of researcher reading behavior from the arXiv; the task is to recommend documents to scientists. The second is crowdsourced food consumption data from a diet tracking app, and the task is meal recommendation. On both benchmarks, RFS outperforms several baseline methods. An example implementation is in Appendix A and anonymized code for running experiments is at TODO (data cannot be shared, as releasing diet or arXiv data is associated with several privacy concerns).

4.1 Baselines

We compare RFS to a suite of recommendation models that rank items using their sets of attributes.

Word embedding model. We fit a word embedding model [19] on the training data using the implementation of fastText from Bojanowski et al. [3]. The dimensionality of the embeddings is set to match models being compared to. We use a context window in the model that is slightly different than in the original implementation. For an item with attributes x_m, the context for attribute $j \in x_m^+$ is the set of other attributes of the same item $j' \in x_m^+ : j' \neq j$. After learning the attribute embeddings β_j for $j \in V$, item embeddings are computed as the average of their attribute embeddings. Users are represented as the average of the embeddings of the items they consume. Recommendation is performed by cosine similarity of an item embedding and a user embedding.

Collaborative topic Poisson factorization. This is a probabilistic matrix factorization model of user consumption data. The recommendation model follows a latent variable scheme,

1. **Document model:**
   (a) Draw topics $\beta_{vk} \sim \text{Gamma}(a, b)$
   (b) Draw document topic intensities $\theta_{dk} \sim \text{Gamma}(c, d)$
   (c) Draw word count $w_{dv} \sim \text{Poisson}(\theta_d^T \beta_v)$.
2. **Recommendation model:**
   (a) Draw user preferences $\eta_{uk} \sim \text{Gamma}(e, f)$
   (b) Draw document topic offsets $\epsilon_{dk} \sim \text{Gamma}(g, h)$
   (c) Draw $r_{ud} \sim \text{Poisson}(\eta_u^T(\theta_d + \epsilon_d))$.

Gopalan et al. [7] develop a variational inference algorithm for posterior inference in this model; we use their implementation.

Collaborative topic Poisson factorization is a permutation-invariant recommendation model. To show that collaborative topic Poisson factorization [7] is permutation-invariant, consider the Poisson likelihood function over words w_{dv}. Conditional on the latent item representation θ_d and latent word representation β_v, every word in the document w_{dv} is independent; the joint probability of words in a document factorizes:

$$p(w_d \mid \theta_d, \beta_v) = \prod_{w_{dv} \in w_d} p(w_{dv} \mid \theta_d, \beta_v). \tag{11}$$

In this model, predictions are made using expectations under the posterior. The posterior is proportional to the the log joint of the model, and the attributes of items (words in documents) enter into the model only via the above product. The product of the probability of words in a document is invariant to a reordering of the words in the document. Therefore, collaborative topic Poisson factorization is permutation-invariant.

Permutation-marginalized recurrent neural network. The Bernoulli distribution used in Equation (4) can be parameterized using a recurrent neural network. We treat attributes as a sequence and marginalize over all permutations of orderings,

$$p(y_{um} = 1 \mid x_m) = \frac{1}{|\pi(x_m)|} \sum_{\pi \in \pi(x_m)} \sigma\left(\phi(\theta_u, \{\beta_{\pi(1)}, \ldots, \beta_{\pi(J)}\})\right). \tag{12}$$

Here β_j are attribute embeddings, $\pi(x_m)$ denotes the set of all permutations of the set x_m, and ϕ is output of an LSTM recurrent neural network [11] projected to one dimension using a linear layer. This model is permutation-invariant as there is an explicit sum over all permutations.

4.2 Recommending research papers

We benchmark RFS on data of scientists reading research papers on the arXiv, where the goal is to recommend papers to scientists. The arXiv usage data represents one year of usage (2012) and consists of 65K users, 637K preprints, and 7.6M clicks.

Evaluation. We follow Gopalan et al. [7] and use the same test and validation splits, and the same set of held-out 10K users. To match Gopalan et al. [7], we compute precision in addition to recall. The held-out validation and test splits each consist of 20% of the ratings and 1% of the documents. In-matrix documents refer to documents that have clicks in the training data, while out-matrix or cold-start documents have no previous clicks.

Hyperparameters for RFS. We test the stochastic gradient descent algorithm with and without momentum [25]. We use a linear learning rate decay that decays to zero in the maximum number of iterations, 200k. We perform a grid search over learning rates of $1, 5, 10, 15, 25$ and momenta of $0.5, 0.9, 0.95, 0.99$. The minibatch size is set to $\{2^{16}\}$. We use a single negative sample per datapoint, sampled uniformly over the entire dataset (corpus sampling is defined in Section 3.5). To match the hyperparameters in Gopalan et al. [7], we set the dimensionality of embeddings and intercepts to 100. Evaluation is performed every 20k iterations.

Hyperparameters for the permutation-marginalized recurrent neural network. The embedding and hidden state sizes are fixed to 100. Evaluation is performed every 20k iterations. We grid search over learning rates of $10^{-1}, 10^{-2}, 10^{-3}$ with the Adam optimizer [12]. If validation performance does not improve, we reload the best parameters and optimizer states, and decay the learning rate by 0.9. We subsample single permutations during training and testing.

Results. Figure 2 shows that RFS with the inner product parameterization outperforms collaborative topic Poisson factorization in terms of in-matrix recall by over 90%. It also improves over collaborative topic Poisson factorization (CTPF) in terms of out-matrix recall, out-matrix precision, and in-matrix precision (for the latter, only when the number of recommendations is greater than 30). The word embedding model performs about as well as CTPF in terms of recall, but performs worse in terms of precision. The permutation-marginalized recurrent neural network model performance was an order of magnitude worse than the other methods; we omit its performance.

Qualitatively, RFS reveals patterns in usage of the arXiv. Figure 1 is a dimensionality-reduced plot of the user embeddings that reveals connections between fields of study. Scientists who focus on high energy physics, `hep`, neighbor specialists in differential geometry, `math.DG`; these areas share techniques. Machine learning researchers, `stat.ML` readers, neighbor statisticians or `math.ST` readers, highlighting the close connection between these fields. Plots for document embeddings show similar patterns. This illustrates how RFS captures rich patterns of interaction between users and items, while benefitting from information in the item attributes.

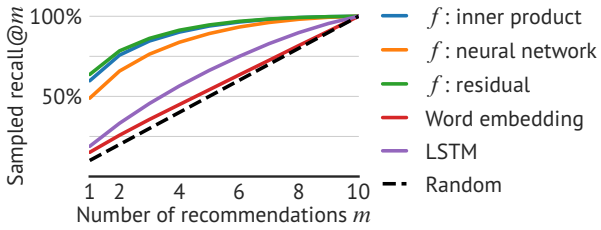


Figure 3: RANKFROMSETS outperforms models based on word embeddings permutation-marginalized recurrent neural networks (denoted by LSTM) on meal recommendation. The recommendation models are trained on data from a food tracking app as described in Section 4.3 and are evaluated using the sampled recall metric, Equation (10). The inner product, neural network, and residual regression functions for RANKFROMSETS are in Equations (4) to (6).

4.3 Recommending meals

We evaluate RFS on data collected from the LoseIt! food tracking app. This app enables users to track their food intake to eat healthy. We use a year's worth of data from 55k active users. This corresponds to 16M meals, where each meal is comprised of a subset of 3M foods.

Preprocessing. Each meal is a subset of 3M foods. We filter the vocabulary by keeping words that occur at least 20 times in the food names, resulting in 9963 words. A meal is represented as the union of the sets of words occurring in the food names.

Evaluation. We hold out 1% of the items (meals) for validation and hold out 1% of the items for evaluating test performance. We use the sampled recall metric in Equation (10) with

Hyperparameters for RFS. The embedding size is set to 128. For the neural network and residual models in Equations (5) and (6) the number of hidden layers is two, and the number of hidden units is set to 256. The item embeddings $g(x_m)$ (and item intercepts $h(x_m)$) are computed as the mean of learned food embeddings (intercepts). We use the RMSProp optimizer in Graves [8] and grid search over learning rates in $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. We use a batch size of 64 and a single negative sample for every datapoint in a minibatch (batch sampling as defined in Section 3.5). Evaluation is performed every 50k iterations.

Hyperparameters for the permutation-marginalized recurrent neural network. We grid search over batch sizes of $\{32, 64, 128\}$. For every item in a minibatch, we sample a single permutation of attributes to approximate the sum in Equation (12). We use a single negative sample per datapoint, and set the embedding and hidden state sizes to 128. We use the Adam optimizer [12] and grid search over learning rates of $\{10^{-2}, 10^{-3}, 10^{-4}\}$. Evaluation is performed every 1k iterations. If the validation performance does not improve, we reload the best parameters and optimizer states, and decay the learning rate by 0.9.

Query meal	Nearest meal by cosine similarity
Two scoops of Raisin Bran cereal, organic Moroccan green tea, almond milk, light honey, tap water, large banana, large strawberries	Vita Bee bread, salted butter, fresh medium tomatoes, large fried whole egg, small banana
Iceberg lettuce, cantaloupe cubes, diced honeydew melon, cherry tomatoes, olives, dry-cooked unsalted hulled sunflower seed kernels, chopped hard-boiled egg, cucumbers, dried cranberries, fat-free ranch dressing	Green leaf lettuce, chopped sweet red bell peppers, crumbled feta cheese, large hard-boiled egg, chopped cucumber, oil-roasted salted sunflower seeds, sliced radishes, sliced strawberries, pitted Calamata olives, fat-free balsamic vinegar
Boston roast pork, mackerel, artichoke hearts, spinach, pimiento-stuffed Manzanilla olives, carrots, mushrooms, peppercorn ranch dressing	Broiled top round steak, tomatoes, cucumber, baby yellow squash, zucchini, black olives, extra virgin olive oil
Meatloaf with tomato sauce, chopped sweet red bell peppers, extra virgin olive oil, cooked asparagus spears, sweet potatoes, orange, cantaloupe cubes	Chicken breast, breadcrumbs, fresh tomatoes, shredded green leaf lettuce, extra virgin olive oil, spinach, chopped yellow onion, sweet large yellow bell peppers, whole mushrooms, chili peppers, vinaigrette
Ciabatta bun, cooked skinless chicken breast, fresh baby spinach, shredded iceberg lettuce, shredded mozzarella cheese, ketchup, frozen yogurt bar	Small whole wheat submarine roll, broiled round roast beef, roasted light turkey meat without skin, fresh medium tomatoes, honey smoked ham, shredded iceberg lettuce, sliced mozzarella cheese

Table 1: RANKFROMSETS, trained on food consumption data, provides diverse recommendations. We fit the model to data from a food tracking app as described in Section 4.3; items are meals and attributes are the ingredients in the meal. We represent meals as the mean of their attribute embeddings, and use cosine similarity to compute the nearest neighbors of meals. This reveals that RANKFROMSETS uncovers latent patterns of consumption in the attribute embeddings that can be leveraged to improve recommendations. For example, the second-last query meal is a mix of meat, vegetables, and fruit, and the nearest neighbor meal is a different meat and a side of salad. The last query meal is a sandwich, and its nearest neighbor is also a sandwich, but with different ingredients.

Results. Figure 3 plots the sampled recall. This shows that RFS outperforms the permutation-marginalized recurrent neural network and the word embedding model. We tried to fit the collaborative topic Poisson factorization model in [7] to this dataset, but the code released by the authors was unable to scale to this size of data.

Qualitatively, RFS learns an interpretable representation of items, as shown by nearest neighbors of meals in Table 1. In this table, we display breakfast, lunch, and dinner meals, alongside their nearest neighbors. We find that the nearest neighbors are also breakfast, lunch, and dinner meals respectively, showing that the attribute embeddings learned by the model can be used to explore qualitative patterns in the learned latent space.

5 DISCUSSION

The task of recommending items with attributes is difficult for several reasons. It is unclear how to incorporate set-valued side information into models that scale to large numbers of items and attributes. In addition, existing recommendation models that leverage item attributes are not directly tied to evaluation metrics. We developed RANKFROMSETS (RFS), a scalable recommendation model for items with attributes. Theoretically, we showed that optimizing its loss function optimizes recall, and that it can approximate other well-founded recommendation models such as content-based matrix factorization. Empirically, RFS outperformed competitors and scaled large datasets, and was very simple to implement as shown in Appendix A.

Simple theory and implementation. It is surprising that RFS outperformed the collaborative topic Poisson factorization model in Gopalan et al. [7], as RFS is a much simpler model. RFS is simpler than CTPF conceptually; it does not require approximate posterior inference, but mere binary classification. RFS is also simpler in terms of implementation: a scalable example is a few dozen lines of python in Appendix A, compared to several thousand lines of C++ released by the authors of CTPF.

Architectures. There is a wealth of deep learning models that can be used to parameterize RFS. For example, attention-based transformer networks can operate on set-valued input [13] and are an interesting choice of architecture for RFS.

Side information. The RFS can easily accommodate various metadata available about users, items, or attributes. Attribute-level or user-level side information is particularly interesting, and when available, should lead to improved recommendation performance.

Regularization. Further performance gains from regularizing RFS are also interesting avenues for future work. It would be interesting to test L2 regularization for sparsity and dropout for parsimony.

Pretrained embeddings. We note that the word embedding baseline performed almost as well as collaborative topic Poisson factorization, which is surprising given that the word embedding model is much simpler. Further performance gains in RFS should be possible by initializing user and item attribute embeddings to pretrained word embeddings as in Chen et al. [5].

Theory (approximation). Proposition 2 means that RFS can approximate the output of any order-invariant model such as matrix factorization. However, there is no guarantee that fit with maximum likelihood estimation using the objective in Equation (7), RFS *will* approximate a specific model. We leave the development of such approximation techniques to future work.

Theory (ranking metrics). How well does binary classification perform for other ranking-based recommendation metrics, such as non-discounted cumulative gain? Analyzing this question is more difficult, and we leave this to future work. We conjecture that a different loss function should allow a similar proof to Proposition 1.

REFERENCES

[1] Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the GRU: Multi-task Learning for Deep Text Recommendations. *ACM Recommender Systems.*

[2] Toine Bogers. 2018. *Tag-Based Recommendation.* Springer International Publishing, 441–479.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Association for Computational Linguistics.*

[4] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. 2017. Embedding Factorization Models for Jointly Recommending Items and User Generated Lists. In *ACM SIGIR (SIGIR '17).* ACM, New York, NY, USA, 10.

[5] Ting Chen, Liangjie Hong, Yue Shi, and Yizhou Sun. 2017. Joint Text Embedding for Personalized Content-based Recommendation. *arXiv:1706.01084.*

[6] Xin Dong, Lei Yu, Zhonghuo Wu, Yuxia Sun, Lingfeng Yuan, and Fangxi Zhang. 2017. A Hybrid Collaborative Filtering Model with Deep Structure for Recommender Systems. *AAAI Conference on Artificial Intelligence.*

[7] Prem Gopalan, Laurent Charlin, and David M. Blei. 2014. Content-based Recommendations with Poisson Factorization. *Advances in Neural Information Processing Systems.*

[8] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. *arXiv:1308.0850.*

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition.*

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, 173–182.

[11] Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8, 1–32.

[12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations.*

[13] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. 2018. Set Transformer. *arXiv:1810.00825.*

[14] Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2018. Towards Better Representation Learning for Personalized News Recommendation: a Multi-Channel Deep Fusion Approach. *International Joint Conference on Artificial Intelligence.*

[15] Junjie Liang, Jinlong Hu, Shoubin Dong, and Vasant G. Honavar. 2018. Top-N-Rank: A Scalable List-wise Ranking Method for Recommender Systems. *arXiv:1812.04109.*

[16] Nan Liang, Hai-Tao Zheng, Jin-Yuan Chen, Arun Sangaiah, and Cong-Zhi Zhao. 2018. TRSDL: Tag-Aware Recommender System Based on Deep Learning–Intelligent Computing Systems. *Applied Sciences.*

[17] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive recommending with Tag-Enhanced Matrix Factorization (TagMF). *International Journal of Human-Computer Studies.* Advances in Computer-Human Interaction for Recommender Systems.

[18] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems.*

[20] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based News Recommendation for Millions of Users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17).* ACM, New York, NY, USA, 1933–1942.

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. *Uncertainty in Artificial Intelligence.*

[22] Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, Martha Larson, Alan Hanjalic, and Nuria Oliver. 2012. TFMAP: Optimizing MAP for Top-N Context-Aware Recommendation. *ACM Special Interest Group on Information Retrieval.*

[23] Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative Filtering Beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. *Comput. Surveys* 47, 1, 1–45.

[24] Bo Song, Xin Yang, Yi Cao, and Congfu Xu. 2018. Neural Collaborative Ranking. *ACM Conference on Information and Knowledge Management*, 10.

[25] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the Importance of Initialization and Momentum in Deep Learning. *International Conference on Machine Learning.*

[26] Adrianus Willem van der Vaart. 1998. *Asymptotic Statistics.* Cambridge University Press.

[27] Victor Veitch, Morgane Austern, Wenda Zhou, David M. Blei, and Peter Orbanz. 2019. Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data. *International Conference on Artificial Intelligence and Statistics.*

[28] Chong Wang and David M. Blei. 2011. Collaborative Topic Modeling for Recommending Scientific Articles. *ACM Knowledge Discovery and Data Mining.*

[29] Zhenghua Xu, Thomas Lukasiewicz, Cheng Chen, Yishu Miao, and Xiangwu Meng. 2017. Tag-Aware Personalized Recommendation Using a Hybrid Deep Model. *International Joint Conference on Artificial Intelligence*, 3196–3202.

[30] Haochao Ying, Liang Chen, Yuwen Xiong, and Jian Wu. 2016. Collaborative Deep Ranking: A Hybrid Pair-Wise Recommendation Algorithm with Implicit Feedback. *Pacific-Asia Conference on Knowledge Discovery and Data Mining.*

[31] Lu Yu, Chuxu Zhang, Shichao Pei, Guolei Sun, and Xiangliang Zhang. 2018. WalkRanker: A Unified Pairwise Ranking Model With Multiple Relations for Item Recommendation. *AAAI Conference on Artificial Intelligence.*

[32] Fajie Yuan, Guibing Guo, Joemon M. Jose, Long Chen, Haitao Yu, and Weinan Zhang. 2016. Optimizing Factorization Machines for Top-N Context-Aware Recommendations. *Web Information Systems Engineering.*

[33] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. 2017. Deep Sets. *Neural Information Processing Systems.*

[34] Shuai Zhang, Lina Yao, and Aixin Sun. 2017. Deep Learning Based Recommender System: A Survey and New Perspectives. *arXiv:1707.07435.*

[35] Yi Zhen, Wu-Jun Li, and Dit-Yan Yeung. 2009. TagiCoFi: Tag Informed Collaborative Filtering. *ACM Recommender Systems*, 8.

[36] Yi Zuo, Jiulin Zeng, Maoguo Gong, and Licheng Jiao. 2016. Tag-aware recommender systems based on deep neural networks. *Neurocomputing.* Big Learning in Social Media Analytics.

A REPRODUCIBILITY SUPPLEMENT: CODE

We give an example implementation of RANKFROMSETS with the inner product regression function in Equation (4) in python with the PyTorch package (version 1.0.0). This shows how easy it is to implement RFS; we used the same implementation to achieve state-of-the-art results in Section 4.2.

# recommendations	10	30	50	70	100
Recall (percent)	0.02	0.06	0.11	0.14	0.21

Table 2: The permutation-marginalized LSTM model performs poorly on the arXiv dataset of researcher reading behavior in terms of in-matrix recall. We report the held-out recall averaged over 100 users due to computational constraints of the model. It is much lower than that of other methods, but better than random.

B REPRODUCIBILITY SUPPLEMENT: LSTM

C REPRODUCIBILITY SUPPLEMENT: PERMUTATION-INVARIANT MODELS

Matrix factorization permutation-invariant. We show that collaborative topic Poisson factorization [7] is permutation-invariant. The likelihood of the attributes (words) associated with items (documents) in this model factorizes, and multiplication is permutation-invariant. Conditional on the latent item representation θ_d and latent word representation β_v, every word in the document w_{dv} is independent. The joint probability of the words in a document factorizes:

$$p(w_d \mid \theta_d, \beta_v) = \prod_{w_{dv} \in w_d} p(w_{dv} \mid \theta_d, \beta_v). \tag{13}$$

In this model, predictions are made using expectations under the posterior. The posterior is proportional to the the log joint of the model, and the attributes of items (words in documents) enter into the model only via the above product. The product of the probability of words in a document is invariant to a reordering of the words in the document. Therefore, collaborative topic Poisson factorization is permutation-invariant.

D REPRODUCIBILITY SUPPLEMENT: SIMULATION STUDY

With finite data and a finite number of paramaters, the optimal parameterization of RFS is dependent on the data-generating distribution. Recall that observations of user-item interactions are generated by a Bernoulli distribution with logit function f. We describe a choice of logit function f that leads to the residual and deep architectures in Equations (5) and (6) outperforming the inner product architecture in Equation (4) in terms of predictive performance. Note that the number of parameters across architectures must be equal for a fair comparison.

We simulate data from the following generative process:

1. **For every user** u**:**
   (a) Draw user embedding $\theta_u \sim \text{Normal}(0, \text{I})$.
2. **For every item attribute** j**:**
   (a) Draw attribute embedding $\beta_j \sim \text{Normal}(0, \text{I})$.
3. **For every item** m**:**
   (a) Draw item topics $\theta_m \sim \text{Dirichlet}(\alpha)$
   (b) Draw number of item attributes $M \sim \text{Poisson}(\lambda)$
   (c) Draw nonzero item attributes $x_m^+ \sim \text{Multinomial}(M, \theta_m)$.
4. **For every user, item observation:**
   (a) $y_{um} \sim \text{Bernoulli}\left(y_{um}; \sigma(f(\theta_u, x_m^+)\right)$

The logit function f is the square kernel:

$$f(\theta_u, x_m^+) = \left(\theta_u^\top \sum_{j \in x_m^+} \beta_j\right)^2.$$

Before sampling from the Bernoulli, we standardize the logits output by f across users and subtract 7 to achieve sparse user-item observations.

Simulation setup. We set the Dirichlet parameter to be $\alpha = 0.01$ and the Poisson rate to be $\lambda = 20$. We generate data for 1k users, 5k item attributes, 30k items, and hold out 100 users for each of the validation and test sets. We fix the momentum to 0.9 [25] and grid search over stochastic gradient descent learning rates of $10, 1, 0.1, 0.01$ and over two learning rate decay schedules. The first linear learning rate decay goes to zero over 100k iterations, while the second divides the learning rate by 10 if the validation in-matrix recall does not improve (evaluation is performed every 500 iterations). We run the grid search on one instance of data generated from this model, then for the best performing hyperparameters for each model trained on this instance, we regenerate data 30 times and average results over these synthetic datasets.

Regression function	Inner product	Deep	Residual
Recall	0.29 ± 0.15	0.32 ± 0.14	0.33 ± 0.18

Table 3: A simulation study demonstrating that the choice of parameterization of RANKFROMSETS is data-dependent. We report the in-matrix recall averaged over 100 users and 30 replications of the simulation where data is regenerated. The residual model in Equation (6) outperforms the deep model, Equation (5), and the inner product model in Equation (4).

Simulation results. The results in Table 3 demonstrate that the residual model outperforms both the deep and inner product architectures for data generated by the above generative process.

Generalization. The above example shows that the choice of architecture in RFS is data-dependent. To ensure that the model does not overfit as new users or items are included in the training data, we need to compare the number of parameters to the number of datapoints. A model with parameters the size of the training data can overfit by memorizing the training data. For generalization to be possible, overfitting can be avoided if the number of parameters grows slower than the size of the data. The technical backing for this comes from asymptotic statistics and the concept of sieved likelihoods. Specifically, the maximum likelihood estimation procedure with the objective function in Equation (7) can be replaced by maximization of a sieved likelihood function. The 'sieve' refers to filtering information as the number of parameters (in this case, user and item representations) grows with the number of observations. The sieved likelihood function enables the analysis of asymptotic behavior as the number of users grows $U \to \infty$ and the number of items grows $I \to \infty$, An example of a technique to grow the number of parameters in a way that supports generalization is given in Chapter 25 of Vaart [26].