Consequential ranking algorithms and long-term welfare

Behzad Tabibian
MPI-IS & MPI-SWS
me@btabibian.com

Vicenç Gomez
Universitat Pompeu Fabra
vicen.gomez@upf.edu

Abir De
MPI-SWS
ade@mpi-sws.org

Bernhard Schölkopf
MPI-IS
bs@tuebingen.mpg.de

Manuel Gomez Rodriguez
MPI-SWS
manuelgr@mpi-sws.org

Abstract

Ranking models are typically designed to provide rankings that optimize some measure of immediate utility to the users. As a result, they have been unable to anticipate an increasing number of undesirable long-term consequences of their proposed rankings, from fueling the spread of misinformation and increasing polarization to degrading social discourse. Can we design ranking models that understand the consequences of their proposed rankings and, more importantly, are able to avoid the undesirable ones? In this paper, we first introduce a joint representation of rankings and user dynamics using Markov decision processes. Then, we show that this representation greatly simplifies the construction of *consequential ranking models* that trade off the immediate utility and the long-term welfare. In particular, we can obtain optimal consequential rankings just by applying weighted sampling on the rankings provided by models that maximize measures of immediate utility. However, in practice, such a strategy may be inefficient and impractical, specially in high dimensional scenarios. To overcome this, we introduce an efficient gradient-based algorithm to learn parameterized consequential ranking models that effectively approximate optimal ones. We showcase our methodology using synthetic and real data gathered from Reddit and show that ranking models derived using our methodology provide ranks that may mitigate the spread of misinformation and improve the civility of online discussions.

1 Introduction

Rankings are ubiquitous across a large variety of online services, from search engines, online shops and recommender systems to social media and online dating. Consequently, it has become much easier to find information, products, jobs, opinions or even potential romantic partners—rankings have undoubtedly increased the utility users obtain from online services. However, ranking have also been blamed to play a major role in an increasing number of missteps, particularly in the context of social and information systems, from fueling the spread of misinformation [41], increasing polarization [13] and degrading social discourse [44] to undermining democracy [40]. As the decisions taken by ranking models become more consequential to individuals and society, one must ask: what went wrong in these cases?

Current ranking models are typically designed to optimize immediate measures of utility, which often reward instant gratification. For example, one of the guiding technical principles behind the optimization of ranking models in the information retrieval literature, the Probability Ranking Principle (PRP) [30], states that the optimal ranking should order items in terms of probability of

relevance to the user. However, such measures of immediate utility do not account for long-term consequences. As a result, ranking models often have an unexpected cost to the long-term welfare. In this work, our goal is to design consequential ranking models which understand the long-term consequences of their proposed rankings.

More specifically, we focus on a problem setting that fits a variety of real-world applications, including those mentioned previously: at every time step, a ranking model receives a set of items and ranks these items on the basis of a measure of immediate utility[1] and a set of features. Items may appear over time and be present at several time steps. Moreover, their corresponding features may change over time and these changes may be due to the influence of previous rankings. For example, the number of likes, votes, or comments—the features—that a post—the item—published by a user receives in social media depends largely on its ranking position [12, 14, 17, 21]. Moreover, for every sequence of rankings, there is an associated long-term (cost to the) welfare, whose specific definition is application dependent. For example, in information integrity, the welfare may be defined as the average number of posts including misinformation at the top of the rankings over time. Then, our goal is to construct consequential ranking models that optimally trade off fidelity to the original ranking model maximizing immediate utility and long-term welfare.

Our contributions. In this paper, we first introduce a joint representation of ranking models and user dynamics using Markov decisions processes (MDPs), which is particularly well-fitted to faithfully characterize the above problem setting[2]. Then, we show that this representation greatly simplifies the construction of consequential ranking models that trade off fidelity to the rankings provided by a model maximizing immediate utility and the long-term welfare. More specifically, we apply Bellman's principle of optimality and show that it is possible to derive an analytical expression for the optimal consequential ranking model in terms of the original ranking model and the cost to the welfare. This means that we can obtain optimal consequential rankings just by applying weighted sampling on the rankings provided by the original ranking model using the (exponentiated) cost to welfare. However, in practice, such a naive sampling will be inefficient, specially in the presence of high-dimensional features. Therefore, we design a practical and efficient gradient-based algorithm to learn parameterized consequential ranking models that effectively approximate optimal ones[3].

Finally, we showcase our methodology using synthetic and real data gathered from Reddit. Our results show that consequential ranking models derived using our methodology provide ranks that may mitigate the spread of misinformation and improve the civility of online discussions without significant deviations from the rankings provided by models maximizing immediate utility measures.

Further related work. The work most closely related to ours is devoted to construct either fair rankings [1, 3, 5, 32–34, 45, 46] or diverse rankings [4, 7, 28, 29]. However, this line of research defines fairness and diversity in terms of exposure allocation on an individual ranking. In contrast, we consider sequences of rankings, we characterize the consequences of these rankings on the user dynamics, and focus on improving the welfare in the long-term. Other related work includes recent approaches to address the learning-to-rank problem from the perspective of reinforcement learning [34, 11, 42]. However, these approaches consider the construction of a single optimal ranking as an MDP in which the state defined at the level of an item/position. In contrast, our MDP considers a sequence of rankings.

Finally, there is a paucity of work on the delayed impact of machine learning algorithms [15, 23, 25] and recommender systems [35, 31]. However, the former has focused on classification tasks and considered simple one-step feedback models and the latter on the tradeoff between exploration and exploitation.

2   Rankings and User Dynamics

In this section, we first introduce our joint representation of rankings and user dynamics, starting from the problem setting it is designed for. Then, we formally define consequential rankings as the solution to a particular reinforcement learning problem.

---

[1]Our methodology does not need to observe the immediate utility the ranking model based their rankings on.

[2]In this work, for ease of exposition, we assume all users get exposed to the same rankings, as in, *e.g.*, Reddit. However, our methodology can be readily extended to the scenario in which each user get exposed to a different ranking, as in, *e.g.*, Twitter.

[3]We will release an open-source implementation of our algorithm with the final version of the paper.

Problem setting. Let p_θ be a particular ranking model[4] (or, equivalently, ranking algorithm). At each time step $t \in \{1, \ldots, T\}$, the ranking model receives a set of n items and these items are characterized by a feature matrix $\boldsymbol{X}(t) \in \mathbb{R}^{n \times p}$, where the i-th row $\boldsymbol{X}_i(t)$ contains the feature values for item $i \in [n]$ and p is the number of features per item. Here, we assume that items may appear over time and be present at several time steps. Moreover, their corresponding feature values may change over time. For example, think of the number of likes, votes or comments that a post receives in social media—they are often used as features to decide the ranking of the post and they change over time.

Then, the ranking model provides a ranking $\boldsymbol{y}(t)$ of the items on the basis of their set of features and a (hidden) measure of immediate utility. A ranking $\boldsymbol{y}(t) = (y_1(t), \ldots, y_n(t))$ is defined as a permutation of the n rank indices, *i.e.*, the model ranks item i in position $y_i(t)$, where highest rank is position 1. In addition, we also define the ordering $\boldsymbol{\omega}(t) = (\omega_1(t), \ldots, \omega_n(t))$ of a ranking as a permutation of the n item indices, *i.e.*, the model ranks item $\omega_i(t)$ in position i. The ranking and orderings are related by $w_{y_i(t)}(t) = i$ and $y_{w_i(t)}(t) = i$. Here, we assume that the provided ranking at time step t may influence the feature matrix at time step $t + 1$. This is in agreement with recent empirical studies [12, 14, 17, 21], which have shown that the posts (the items) that are ranked highly receive a higher number of likes, comments or shares (the features).

Finally, given a trajectory of feature matrices and rankings $\tau = \{(\boldsymbol{X}(t), \boldsymbol{y}(t))\}_{t=0}^{T}$ there is an additive cost to the welfare, $c(\tau) = \sum_{t=0}^{T} c(\boldsymbol{X}(t), \boldsymbol{y}(t))$, where $c(\boldsymbol{X}(t), \boldsymbol{y}(t))$ is an arbitrary immediate cost whose specific definition is application dependent. For example, in information integrity, the welfare may be defined as the average number of posts including misinformation at the top of the rankings over time. In the remainder, we will say that a trajectory τ is *induced* by a ranking model p_θ.

Joint representation of rankings and user dynamics. The above problem setting naturally fits the following joint representation of rankings and user dynamics using Markov decision processes (MDPs) [36], which also has an intuitive causal interpretation:

$$p_\theta(\tau \mid \boldsymbol{X}(t_0), \boldsymbol{y}(t_0)) = \prod_{t=1}^{T} \underbrace{p_\theta(\boldsymbol{y}(t) \mid \boldsymbol{X}(t))}_{\text{ranking model}} \underbrace{p(\boldsymbol{X}(t) \mid \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))}_{\text{user dynamics}}, \tag{1}$$

where the first term represents the particular choice of ranking model[5], the second term represents the distribution for the user dynamics, which determines the feature matrix at any given time step, and the initial feature matrix $\boldsymbol{X}(t_0)$ and ranking $\boldsymbol{y}(t_0)$ are given. Moreover, the above representation makes two major assumptions, which are also illustrated in Figure 3 in Appendix A:

(i) To provide a ranking for a set of items at time step t, the ranking model only uses the feature matrix corresponding to that set of items. More formally, given the feature matrix $\boldsymbol{X}(t)$, the ranking $\boldsymbol{y}(t)$ provided by the ranking model is conditionally independent of previous feature matrices $\boldsymbol{X}(t')$, $t' < t$. In most practical scenarios, ranking models optimizing for immediate utility satisfy this assumption.

(ii) The dynamics of the feature matrices, which characterize the user dynamics, are Markovian. That means, given the feature matrix $\boldsymbol{X}(t-1)$ and ranking $\boldsymbol{y}(t-1)$, the feature matrix $\boldsymbol{X}(t)$ is conditionally independent of previous feature matrices $\boldsymbol{X}(t')$ and rankings $\boldsymbol{y}(t')$, $t' < t$. This is a natural assumption taken in the state-of-the-art models (*e.g.*, Hawkes processes with exponential kernels [8, 10, 37]).

— *Ranking model:* Our approach is agnostic to the particular choice of ranking model—it provides a methodology to derive consequential rankings that are optimal under a ranking model. In our experiments, we showcase our methodology for one well-known ranking model, the Plackett-Luce (P-L) ranking model [24, 27], which is best described in terms of the orderings of the rankings. More specifically, under the P-L model, at each time step t, the ranking $\boldsymbol{y}(t)$, with ordering $\boldsymbol{\omega}(t)$, is sampled from a distribution

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}(t) \mid \boldsymbol{X}(t)) = \prod_{k=1}^{n} f_k(\boldsymbol{X}(t)) = \prod_{k=1}^{n} \frac{\exp\left(\boldsymbol{\theta}^T \boldsymbol{X}_{\omega_k}(t)\right)}{\sum_{k'=k}^{N} \exp\left(\boldsymbol{\theta}^T \boldsymbol{X}_{\omega_{k'}}(t)\right)} \tag{2}$$

---

[4]Unless stated otherwise, the notation p_θ does not imply that θ is a parameter within a class of ranking models, but it just serves as a placeholder to identify a specific ranking model.

[5]In our work, we consider probabilistic ranking models, which assign a probability to each ranking. It would be interesting to extend our methodology to deterministic ranking models.

where $\boldsymbol{\theta}$ is a given parameter. In the above, we can think of $\boldsymbol{\theta}^T \boldsymbol{X}_{\omega_k}(t)$ as a *quality score* associated to the item ω_k, which controls the probability that this item is ranked at the top—the higher the quality score, the higher the probability that the item is ranked first. In practice, the quality score of the above P-L ranking model may be computed using a complex nonlinear function [39], *e.g.*, a neural network.

— *User dynamics:* Our approach only requires to be able to sample $\boldsymbol{X}(t)$ from any arbitrary model for the transition probability $p(\boldsymbol{X}(t) \,|\, \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))$, which may be estimated using historical ranking and user data. Here, in contrast with the ranking model, the user dynamics are not something that one can decide upon—they are given.

Consequential rankings. Let p_{θ_0} be an existing ranking model that optimizes some hidden immediate utility and $c(\cdot)$ a cost to the welfare. Then, we construct a consequential ranking model p_θ, which optimally trades off the fidelity to the original ranking model and the cost to the long-term welfare, by solving the following optimization problem:

$$\underset{p_\theta}{\text{minimize}} \quad \mathbb{E}_{\tau \sim p_\theta} \left[S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) \right], \tag{3}$$

with

$$S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) = c(\tau) + \lambda \log \frac{p_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}{p_{\theta_0}(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}, \tag{4}$$

where the expectation is taken over all the trajectories $\tau = \{(\boldsymbol{X}(t), \boldsymbol{y}(t))\}_{t=0}^T$ of feature matrices and rankings of length T under the ranking model p_θ and $\lambda \geq 0$ is a given parameter which controls the trade off between the fidelity to the original ranking model and the long-term cost to the welfare. In Eq. 4, the first term penalizes trajectories that achieve a large cost to the welfare and the second term penalizes ranking models whose induced trajectories differ more from those that the original model would induce, where the terms associated to the user dynamics $p(\boldsymbol{X}(t) \,|\, \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))$ cancel. Moreover, the choice of trajectory length T will depend on the definition of long-term—accounting for longer-term consequences to the welfare will require larger trajectory lengths T.

Finally, note that our measure of fidelity has a natural interpretation in terms of the Kullback-Leibler (KL) divergence [20], which has been extensively used as a *distance* measure between probability distributions, leading to the formulation of reinforcement learning as probabilistic inference [22, 18, 48]. More specifically, we can write the expectation of the second term as the KL divergence between the original and the consequential ranking model, *i.e.*, $KL[p_\theta(\cdot \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) \,||\, p_{\theta_0}(\cdot \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))]$.

3   Building Consequential Rankings

In this section, we tackle the optimization problem defined by Eq. 3 from the perspective of reinforcement learning and show that the optimal consequential ranking model can be expressed in terms of the original ranking model.

We can first break the above problem into small recursive subproblems using Bellman's principle of optimality [2]. This readily follows from the fact that, under the representation introduced in Section 2, the ranking model and the user dynamics are a Markov decision process (MDP). More specifically, Bellman's principle tells us that the optimal ranking model should satisfy the following recursive equation, which is called the Bellman optimality equation:

$$V_t(\boldsymbol{X}, \boldsymbol{y}) = \min_{p_\theta} \ell(\boldsymbol{X}, \boldsymbol{y}) + \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p_\theta(\boldsymbol{X}', \boldsymbol{y}' \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[V_{t+1}(\boldsymbol{X}', \boldsymbol{y}') \right] \tag{5}$$

with $V_T(\boldsymbol{X}, \boldsymbol{y}) = \ell(\boldsymbol{X}, \boldsymbol{y})$. The function $V_t(\boldsymbol{X}, \boldsymbol{y})$ is called the value function and the function $\ell(\boldsymbol{X}, \boldsymbol{y})$ is called immediate loss. Moreover, in our problem, it can be readily shown that the immediate loss adopts the following form:

$$\begin{aligned} \ell(\boldsymbol{X}, \boldsymbol{y}) &= c(\boldsymbol{X}, \boldsymbol{y}) + \lambda \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p_\theta(\boldsymbol{X}', \boldsymbol{y}' \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p_\theta(\boldsymbol{X}', \boldsymbol{y}' \,|\, \boldsymbol{X}, \boldsymbol{y})}{p_{\theta_0}(\boldsymbol{X}', \boldsymbol{y}' \,|\, \boldsymbol{X}, \boldsymbol{y})} \right] \\ &= c(\boldsymbol{X}, \boldsymbol{y}) + \lambda KL(p_\theta(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y}) \,||\, p_{\theta_0}(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})). \end{aligned}$$

Within the loss function, the first term penalizes the immediate cost to the welfare and the second term penalizes consequential ranking models whose induced transition probability differs from that induced by the original ranking model.

In general, Bellman optimality equations are difficult to solve. However, the structure of our problem will help us find an analytical solution. Inspired by Todorov [38], we proceed as follows. Let $Z_t(\boldsymbol{X}, \boldsymbol{y}) = \exp(-V_t(\boldsymbol{X}, \boldsymbol{y}))$. Then, we can rewrite the minimization in the RHS of Eq. 5 as

$$\min_{p_\theta} \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p_\theta(\boldsymbol{X}', \boldsymbol{y}' \mid \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p_\theta(\boldsymbol{X}', \boldsymbol{y}' \mid \boldsymbol{X}, \boldsymbol{y})}{p_{\theta_0}(\boldsymbol{X}', \boldsymbol{y}' \mid \boldsymbol{X}, \boldsymbol{y}) Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')} \right],$$

where we have dropped λ and $c(\boldsymbol{X}, \boldsymbol{y})$ because they do not depend on p_θ. Then, we can use Eq. 1 to factorize both transition probabilities in the numerator and the denominator within the logarithm and, as a result, the terms $p(\boldsymbol{X}' \mid \boldsymbol{X}, \boldsymbol{y})$ cancel and we obtain:

$$\min_{p_\theta} \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p_\theta(\boldsymbol{X}', \boldsymbol{y}' \mid \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p_\theta(\boldsymbol{y}' \mid \boldsymbol{X}')}{p_{\theta_0}(\boldsymbol{y}' \mid \boldsymbol{X}') Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')} \right].$$

The above equation resembles a KL divergence, however, note that the fraction within the logarithm does not depend on $(\boldsymbol{X}, \boldsymbol{y})$ and the denominator $p_{\theta_0}(\boldsymbol{y}' \mid \boldsymbol{X}') Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')$ is not normalized to one. If we multiply and divide the fraction by the following normalization term:

$$G[Z_{t+1}](\boldsymbol{X}') = \mathbb{E}_{\boldsymbol{y}' \sim p_{\theta_0}(\boldsymbol{y}' \mid \boldsymbol{X}')}[Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')], \tag{6}$$

we obtain:

$$\min_{p_\theta} - \mathbb{E}_{\boldsymbol{X}' \sim p(\boldsymbol{X}' \mid \boldsymbol{X}, \boldsymbol{y})} \left[\log G[Z_{t+1}](\boldsymbol{X}') \right] + \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p_\theta(\boldsymbol{X}', \boldsymbol{y}' \mid \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p_\theta(\boldsymbol{y}' \mid \boldsymbol{X}') G[Z_{t+1}](\boldsymbol{X}')}{p_{\theta_0}(\boldsymbol{y}' \mid \boldsymbol{X}') Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')} \right].$$

In the above equation, note that the first term does not depend on p_θ and the second term achieves its global minimum of zero if the numerator and the denominator are equal. Thus, the optimal consequential ranking model is just given by:

$$p_\theta^*(\boldsymbol{y} \mid \boldsymbol{X}) = \frac{p_{\theta_0}(\boldsymbol{y} \mid \boldsymbol{X}) Z_{t+1}(\boldsymbol{X}, \boldsymbol{y})}{G[Z_{t+1}](\boldsymbol{X})}. \tag{7}$$

Finally, if we substitute back p_θ^* into the Bellman equation, given by Eq. 5, and write it in terms of Z_t, we can also find the function Z_t using the following recursive expression:

$$Z_t(\boldsymbol{X}, \boldsymbol{y}) = \exp\left(-c(\boldsymbol{X}, \boldsymbol{y}) + \lambda \mathbb{E}_{\boldsymbol{X}' \sim p(\boldsymbol{X}' \mid \boldsymbol{X}, \boldsymbol{y})} \left[\log G[Z_{t+1}](\boldsymbol{X}') \right] \right),$$

with $Z_T(\boldsymbol{X}, \boldsymbol{y}) = -\log c(\boldsymbol{X}, \boldsymbol{y})$. The above result has an important implication. It means that we can use sampling methods to obtain (unbiased) samples from the optimal consequential ranking, *e.g.*, stratified sampling [9], as shown in Appendix B. However, in practice, these sampling methods may be inefficient and have high variance if the original ranking model p_{θ_0} produces rankings that have very low probability under the optimal consequential ranking model. This will be specially problematic in the presence of high-dimensional feature vectors due to the curse of dimensionality. In the next section, we will present a practical method for approximating $p_\theta^*(\boldsymbol{y} \mid \boldsymbol{X})$, which iteratively adapts a parameterized consequential ranking model using a stochastic gradient-based algorithm.

4 A Stochastic Gradient-Based Algorithm

In this section, our goal is to find a consequential ranking model p_θ within a class of parameterized ranking models $\mathcal{P}(\Theta)$ that approximates well the optimal consequential ranking model p_θ^*, given by Eq. 7, that minimizes the objective function in Eq. 3, *i.e.*, $\mathbb{E}_{\tau \sim p_\theta}\left[S_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) \right]$.

To this aim, we introduce a general gradient-based algorithm, which only requires the class of parameterized ranking models $\mathcal{P}(\Theta)$ to be differentiable. In particular, we resort to stochastic gradient descent (SGD) [19], *i.e.*, $\theta^{(j+1)} = \theta^{(j)} + \gamma_j \nabla_\theta \mathbb{E}_{\tau \sim p_\theta}\left[S_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) \right]\big|_{\theta=\theta^{(j)}}$, where $\gamma_j > 0$ is the learning rate at step $j \in \mathbb{N}$. Here, it may seem challenging to compute a finite sample estimate of the gradient of the objective function $\mathbb{E}_{\tau \sim p_\theta}\left[S_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) \right]$ since the derivative is taken with respect to the parameters of the ranking model p_θ, which we are trying to learn. However, we can overcome this challenge using the log-derivative trick as in [43], which allows us to write the gradient as:

$$\nabla_\theta \mathbb{E}_{\tau \sim p_\theta}\left[S_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) \right] = \mathbb{E}_{\tau \sim p_\theta}\left[\left(S_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) + \lambda \right) \nabla_\theta \log p_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) \right], \tag{8}$$

Algorithm 1 It trains a parameterized consequential ranking model.

Require: Cost to welfare $c(\cdot)$, parameter λ, original ranking model p_{θ_0}, $(\boldsymbol{X}(0), \boldsymbol{y}(0))$, # of iterations M, mini batch size B, and learning rate γ.

1: $\theta^{(0)} \leftarrow$ INITIALIZERANKINGMODEL()
2: **for** $j = 1, \ldots, M$ **do** $\triangleright$ iterations
3:     $\mathcal{D} \leftarrow$ MINIBATCH(p_θ, B) $\triangleright$ sample mini batch
4:     $\nabla \leftarrow 0$
5:     **for** $\tau^{(i)} \in \mathcal{D}$ **do**
6:         $S \leftarrow c(\tau^{(i)}) + \lambda \log \frac{p_{\theta(j)}(\tau^{(i)} \mid \boldsymbol{X}(0), \boldsymbol{y}(0))}{p_{\theta_0}(\tau^{(i)} \mid \boldsymbol{X}(0), \boldsymbol{y}(0))}$
7:         $\nabla \leftarrow \nabla + (S + \lambda) \nabla_\theta \log p_{\theta(j)}(\tau^{(i)} \mid \boldsymbol{X}(0), \boldsymbol{y}(0))$
8:     $\theta^{(j+1)} \leftarrow \theta^{(j)} + \gamma \frac{\nabla}{B}$
9: **return** $\theta^{(M)}$

---

where $\nabla_\theta \log p_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0))$ is often referred as the score function [16]. The overall procedure is summarized in Algorithm 1, where MINIBATCH(p_θ, B) samples a minibatch of size B from $p_\theta(\tau)$ and INITIALIZERANKINGMODEL() initializes the parameters of the ranking model.

Remarks. Note that, to compute an empirical estimate of the gradient in Eq. 8, we only need to be able to sample from the user dynamics $p(\boldsymbol{X}(t) \mid \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))$, since the explicit dependence cancels out within $S_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0))$, as pointed out in Section 2. Moreover, depending on the choice of parameterized family of ranking models, one may be able to compute the score functions analytically. In our experiments, the class of Plackett-Luce (P-L) ranking models allows for that, *i.e.*,

$$\nabla_\theta \log p_\theta(\tau \mid \boldsymbol{X}(0), \boldsymbol{y}(0)) = \sum_{t=1}^{T} \sum_{k=1}^{n} \left(\theta^T - \nabla_\theta \log \sum_{k'=k}^{n} \exp(\theta^T \boldsymbol{X} \omega_{k'}(t)) \right),$$

where the second term within the logarithm in the last equation is the derivative of the log-sum-exp function, whose analytical expression can be found elsewhere. Finally, if we think of the parameterized ranking model p_θ as a policy, our algorithm resembles policy gradient algorithms used in the reinforcement learning literature [36]. This connection opens up the possibility of using variance reduction techniques used in policy gradient to improve the empirical estimation of the gradient [47].

5 Experiments on synthetic data

In this section, our goal is compare the performance of ranking models that maximize an immediate measure of utility against consequential rankings in a problem setting with known user dynamics satisfying the Markov property.

Experimental setup. Each trajectory has length $T = 30$ and, at each time step $t \in \{0, \ldots, T\}$, the ranking model receives a set of $n = 10$ posts $\mathcal{I}(t)$ and ranks them. Given a set of items $\mathcal{I}(t)$ and a ranking $\boldsymbol{y}(t)$, we assume that the set of items $\mathcal{I}(t+1)$ is just a copy of the set of items $\mathcal{I}(t)$ where the $d \sim \text{Poisson}(1)$ posts at the bottom of the ranking $\boldsymbol{y}(t)$ are replaced by new posts. Each post i has two features $\boldsymbol{X}_i(t) = [p_i, a_i(t)]$, where p_i is the (static) probability that the post is misinformation and $a_i(t)$ is the (dynamic) rate of shares at time t, initialized with $a_i(0) = 0$. There are high risk posts ($p_i = 0.6$) and low risk posts ($p_i = 0.1$) and a post is either high risk or low risk uniformly at random. Thus, whether the actual post is misinformation or not is a latent variable $m_i \sim \text{Bernouilli}(p_i)$, which is unobserved by the ranking model. The instantaneous rate of shares for each item i is given by $a_i(t+1) = \exp(-2(t - s_i)) (a_i(t) + \alpha_i + 0.1 y_i(t))$, where s_i is the time when the post was first ranked by the ranking model, α_i is the virality, and a post is either viral ($\alpha_i = 10$) or non viral ($\alpha_i = 0.1$) uniformly at random. Here, note that rate of shares of an item increases if the item is ranked at the top, as observed in previous empirical studies.

The original ranking model p_{θ_0} aims to promote viral posts on the top $K = 3$ positions of the ranking at each time t, *i.e.*, $u(t) = \sum_{k=1}^{K=3} a_{\omega_k(t)}(t)$, where $\boldsymbol{\omega}(t)$ is the ordering of the ranking $\boldsymbol{y}(t)$. To this aim, it uses a Plackett-Luce (P-L) model, given by Eq. 2, with $\theta \in [0, 1]$. The consequential ranking models p_θ aim to trade off fidelity to the original model and the long-term presence of misinformation on the top $K = 3$ positions of the rankings, *i.e.*, $c(\tau) = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K=3} p_{\omega_k(t)}$. Here, we consider two consequential rankings models: (i) an optimal consequential ranking model p_θ^*, which provides

6

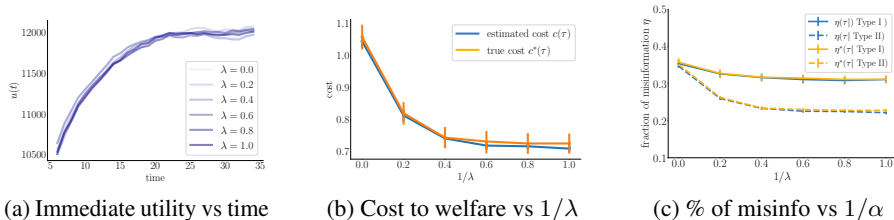| (a) Immediate utility vs time | (b) Cost to welfare vs $1/\lambda$ | (c) % of misinfo vs $1/\alpha$ |

Figure 1: Performance of the original ranking model ($1/\lambda \to 0$) and the optimal consequential ranking model ($1/\lambda > 0$) in synthetic data. Panels (a) and (b) show that, as $1/\lambda$ increases, the consequential ranking model is able to reduce the (true) cost to welfare without significantly decreasing the immediate utility. Panel (c) shows that, as $1/\lambda$ increases, the fraction of misinformation for viral posts on the top 3 positions is lower than the fraction of misinformation for non viral posts.

rankings by applying weighted sampling on the rankings provided by the original ranking model p_{θ_0}; and, (ii) a Plackett-Luce (P-L) consequential ranking model p_θ, which is learned using Algorithm 1 with $M = 400$ iterations and $B = 50$ as batch size. Moreover, we experiment with different values of the parameter λ, which controls the trade off between fidelity to the original model and cost to the welfare and, for each experiment, we perform 200 repetitions.

Quality of the rankings. We first compare the original ranking model p_{θ_0} and the optimal consequential ranking model p_θ^* in terms of three quality metrics: (i) the immediate utility $u(t)$; (ii) the cost to welfare $c(\tau)$; and, (iii) the true cost to welfare $c^*(\tau)$, defined as $c^*(\tau) = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K=3} m_{\omega_k(t)}$. Figure 1(a-b) summarizes the results, where note that the original ranking model p_{θ_0} is just the optimal consequential ranking model with $1/\lambda \to 0$. The results show that: (i) the consequential ranking model achieves lower (true) cost to welfare than the original ranking model; and, (ii) as $1/\lambda$ increases, the consequential ranking model is able to reduce the (true) cost to welfare without significantly decreasing the immediate utility. Next, we investigate whether the optimal consequential ranking model treats viral and non-viral posts differently. Intuitively, the ranking model should be more willing to change the rank of high risk viral posts than that of high risk non viral posts. To confirm this intuition, we compute the fraction of estimated and true misinformation, $\eta(\tau)$ and $\eta^*(\tau)$, in the top $K = 3$ positions of the rankings over time for both viral ($\alpha = 10$) and non viral ($\alpha = 0.1$) posts. Figure 1(c) summarizes the results, which show that, as we increase $1/\lambda$, the fraction of misinformation for viral posts on the top 3 positions is lower than the fraction of misinformation for non viral posts. Appendix C provides an in-depth comparison between optimal and P-L consequential ranking models.

6 Experiments on real data

In this section, we compare the performance of ranking models that maximize an immediate measure of utility and parameterized consequential rankings models using data from Reddit, a popular social news aggregation platform[6]. Before we proceed further, we would like to acknowledge that:

(i) Since we do not have access to the ranking algorithm used by Reddit (or any other social media platform), our experiments are a proof of concept, which demonstrate the practical potential of our methodology on real data using a simple P-L ranking model. Evaluating the efficacy of our methodology across a wide range of deployed ranking algorithms is left as future work.

(ii) Our experiments use observational data, *i.e.*, they are open loop. As a result, the rankings only influence the immediate utility and the cost of welfare but not the user dynamics. However, our evaluation is likely to be conservative—consequential rankings may achieve a greater reduction of the cost to welfare in an interventional experiment. For example, in the context of uncivil behavior, there is empirical evidence that users are more likely to post uncivil comments if they are exposed to uncivil comments before [6, 26]. Therefore, penalizing the rank of uncivil comments over time may prevent other users from engaging into uncivil behavior.

Dataset description and experimental setup. We used a publicly available Reddit dataset[7], which contains (nearly) all publicly available comments to link submissions posted by Reddit users from

---

[6]Due to the size of the dataset, we were unable to run the weighted sampling procedure needed to implement optimal consequential rankings models.

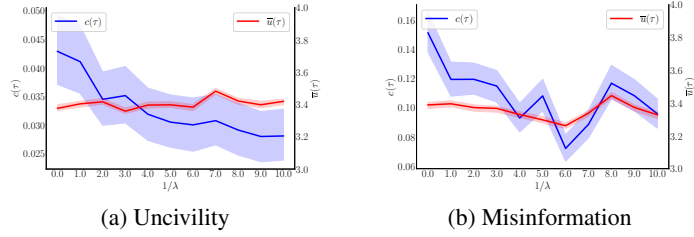[7]https://archive.org/details/2015_reddit_comments_corpus.

(a) Uncivility	(b) Misinformation

Figure 2: Performance of the original ranking model ($1/\lambda \to 0$) and the optimal consequential ranking model ($1/\lambda > 0$) in real Reddit data in terms of average immediate utility $\overline{u}(\tau)$ and cost to the welfare $c(\tau)$. Lines are averages and shaded areas are 95% confidence intervals over all submissions in the test set.

October 2007 to May 2015. In our experiments, we focused on the links submissions to the subreddit Politics and selected the set of submissions with more than 10 and less than 60 comments. After these preprocessing steps, our dataset comprised 3,173 submissions and 68,016 comments.

In a first set of experiments, we focus on the civility of the comments in each submission, as measured by an uncivility score ϕ. In a second set of experiments, we focus on the misinformation spread by the comments of each submission, as measured by an unreliability score γ. Appendix D contains more details on the definition and estimation of both scores. In both sets of experiments, we use 42,267 comments from 1,973 submissions as training set for learning the parameterized consequential ranking models and 25,749 comments from 1,200 submissions as test set for evaluation.

Each submission corresponds to one trajectory whose length T is just the number of comments in the submission, *i.e.*, each time step corresponds to the time at which a new comment was created. Then, at each time step $t \in \{0, \ldots, T\}$, the ranking model ranks the latest set of $n = 10$ comments $\mathcal{I}(t)$. Moreover, each comment i has three features $\boldsymbol{X}_i(t) = [\tau_i, \phi_i, \gamma_i]$, where τ_i is the time (in seconds) elapsed since the first comment was posted, ϕ_i is the uncivility score and γ_i is the unreliability score At each time t, the original ranking model p_{θ_0} aims to promote the most recent comment to the top of the ranking, *i.e.*, its immediate utility $u(t)$ is defined as $u(t) = \tau_{\omega_1(t)}$, where $\omega_1(t)$ is the item at the top of the rank $\boldsymbol{y}(t)$. To this aim, it uses a Plackett-Luce (P-L) model, given by Eq. 2, with $\theta = [1.25 \cdot 10^{-4}, 0, 0]$. For the first set of experiments, the consequential ranking models p_θ aim to trade off fidelity to the original ranking model and the civility of the comments on the top $K = 6$ positions of the rankings, *i.e.*, $c(\tau) = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K=6} \phi_{\omega_k(t)}$. In a second set of experiments, the consequential ranking models p_θ aim to trade off fidelity to the original ranking model and the misinformation in the comments on the top $K = 6$ positions of the rankings, *i.e.*, $c(\tau) = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K=6} \gamma_{\omega_k(t)}$. In both cases, the consequential ranking models are Plackett-Luce (P-L) models, which we learned using Algorithm 1 with $M = 20$ iterations and $B = 100$ as batch size, and experiment with $\lambda = [0.0, 1.0, 2.0, ..., 10.0]$.

Quality of the rankings. We first compare the original ranking model p_{θ_0} and the consequential P-L ranking models p_θ using the average immediate utility and the cost to welfare. Here, note that, in the first set of experiments, the cost to welfare measures the degree of uncivility of the top ranking positions while, in the second set of experiments, it measures the amount of misinformation. Figure 2 summarizes the results, where note that the original ranking model p_{θ_0} is just the optimal consequential ranking model with $1/\lambda \to 0$. The results show that the consequential P-L ranking models are able to reduce the cost to the welfare up to 30% at a minimum cost in terms of immediate utility—they are able to reduce the degree of uncivility and the amount of misinformation at the top ranking positions without significant changes to the original reverse chronological ranking.

7   Conclusions

We have initiated the design of (parameterized) consequential ranking models that optimally trade off between the fidelity to ranking models optimizing for immediate utility and the long-term welfare. Our work opens up many interesting avenues for future work. For example, we have considered probabilistic ranking models and a fidelity measure based on KL divergence. A natural next step is to augment our methodology to allow for deterministic ranking models and consider other fidelity measures between rankings. It would be very interesting to apply our framework to more sophisticated ranking models. Moreover, we have assumed that the models that optimize for immediate utility

are optimal. However, they may be suboptimal in terms of the sum of immediate utility over time since it is unclear that current ranking models are designed to account for the consequences that their proposed rankings have on the feature matrices. It would be very interesting to account for this in future work. It would also be interesting to experiment with settings in which the cost to welfare cannot be factorized into items, *e.g.*, information diversity. Finally, we have evaluated our algorithm using observational real data, however, it would be very revealing to perform an evaluation based on interventional experiments.

References

[1] A. Asudehy, H. Jagadishy, J. Stoyanovichz, and G. Das. Designing fair ranking schemes. *arXiv preprint arXiv:1712.09752*, 2017.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000.

[3] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. *arXiv preprint arXiv:1805.01788*, 2018.

[4] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[5] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

[6] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW)*, 2017.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[8] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. G. Rodriguez. Learning and forecasting opinion dynamics in social networks. In *Advances in Neural Information Processing Systems*, pages 397–405, 2016.

[9] R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE, 2005.

[10] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564, 2016.

[11] Y. Feng, J. Xu, Y. Lan, J. Guo, W. Zeng, and X. Cheng. From greedy selection to exploratory decision-making: Diverse ranking with policy-value networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 125–134, 2018.

[12] M. Gomez-Rodriguez, K. P. Gummadi, and B. Schoelkopf. Quantifying information overload in social media and its impact on social contagions. In *ICWSM*, 2014.

[13] Herrman, John. Inside facebook's political-media machine. *New York Times*, 2016.

[14] N. Hodas and K. Lerman. How visibility and divided attention constrain social contagion. In *SocialCom*, 2012.

[15] L. Hu and Y. Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the Web Conference*, 2018.

[16] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

[17] J. Kang and K. Lerman. Vip: Incorporating human cognitive biases in a probabilistic model of retweeting. In *ICSC*, 2015.

[18] H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

[19] J. Kiefer, J. Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[20] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[21] K. Lerman and T. Hogg. Leveraging position bias to improve peer recommendation. *PloS one*, 9(6):e98914, 2014.

[22] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

[23] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[24] R. D. Luce. The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3):215–233, 1977.

[25] H. Mouzannar, M. I. Ohannessian, and N. Srebro. From fair decision making to social equality. In *FAT**, 2019.

[26] A. Muddiman and N. J. Stroud. News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, 67(4):586–609, 2017.

[27] R. L. Plackett. The analysis of permutations. *Applied Statistics*, pages 193–202, 1975.

[28] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum*, volume 43, pages 46–52. ACM, 2009.

[29] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM, 2008.

[30] S. E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

[31] T. Schnabel, P. N. Bennett, S. T. Dumais, and T. Joachims. Short-term satisfaction and long-term coverage: Understanding how users tolerate algorithmic exploration. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 513–521, 2018.

[32] A. Singh and T. Joachims. Equality of opportunity in rankings. In *Workshop on Prioritizing Online Content, Neural Information Processing Systems*, 2017.

[33] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2018.

[34] A. Singh and T. Joachims. Policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056*, 2019.

[35] A. Sinha, D. F. Gleich, and K. Ramani. Deconvolving feedback loops in recommender systems. In *Advances in Neural Information Processing Systems*, pages 3243–3251, 2016.

[36] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[37] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schoelkopf, and M. Gomez-Rodriguez. Optimizing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 2019.

[38] E. Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.

[39] T. Tran, D. Phung, and S. Venkatesh. Choice by elimination via deep neural networks. *arXiv preprint arXiv:1602.05285*, 2016.

[40] S. Vaidhyanathan. Facebook wins, democracy loses. *New York Times*, 2017.

[41] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 2018.

[42] Z. Wei, J. Xu, Y. Lan, J. Guo, and X. Cheng. Reinforcement learning to rank with markov decision process. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 945–948, 2017.

[43] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[44] J. C. Wong. Former facebook executive: social media is ripping society apart. *The Guardian*, 2017.

[45] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 2017.

[46] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *CIKM*, 2017.

[47] T. Zhao, H. Hachiya, G. Niu, and M. Sugiyama. Analysis and improvement of policy gradient estimation. In *Advances in Neural Information Processing Systems*, pages 262–270, 2011.

[48] B. D. Ziebart, J. A. Bagnell, and A. K. Dey. Modeling interaction via the principle of maximum causal entropy. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

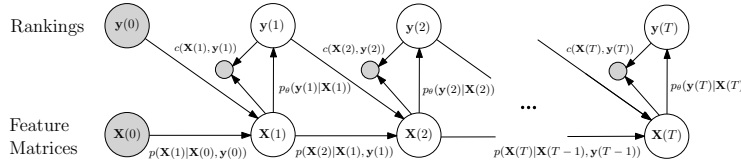## A   Graphical representation of rankings and user dynamics

Figure 3: Our joint representation of rankings and user dynamics using Markov decision processes (MDPs). The ranking model $p_\theta(\boldsymbol{y}(t) \mid \boldsymbol{X}(t))$ provides a ranking $\boldsymbol{y}(t)$ for a set of items on the basis of the feature matrix $\boldsymbol{X}(t)$ of the items and both the feature matrix and the provided ranking result in a cost to welfare $c(\boldsymbol{X}(t), \boldsymbol{y}(t))$. The distribution of user dynamics $p(\boldsymbol{X}(t+1) \mid \boldsymbol{X}(t), \boldsymbol{y}(t))$ determines the feature matrix $\boldsymbol{X}(t+1)$ on the basis of the previous feature matrix $\boldsymbol{X}(t)$ and ranking $\boldsymbol{y}(t)$.

B   Algorithm for sampling from optimal consequential ranking

Algorithm 2 shows the stratified sampling [9] method for obtaining samples from optimal consequential ranking. Within the algorithm, $\mathrm{SAMPLE}(p_{\theta_0}, \kappa)$ samples κ trajectories from $p_{\theta_0}(\tau)$ and $\mathrm{STRATIFIEDSAMPLER}(\mathcal{D}', W)$ generates $|\mathcal{D}'|$ samples weighted by W using stratified sampling.

Algorithm 2 It samples from an optimal consequential ranking model given p_{θ_0}.

Require: Cost to welfare $c(\cdot)$, parameter λ, original ranking model p_{θ_0}, $(\boldsymbol{X}(0), \boldsymbol{y}(0))$, # of samples B, # of samples κ to compute $G[Z_T]$.

1:   $\mathcal{D} \leftarrow \mathrm{SAMPLE}(p_{\theta_0}, \kappa)$ ▷ samples for estimating $G[Z_T]$.
2:   $G[Z_T] \leftarrow 0$
3:   **for** $c(\tau_i) \in \mathcal{D}$ **do**
4:       $G[Z_T] \leftarrow G[Z_T] + \exp\left(-\lambda^{-1}c(\tau_i)\right)/\kappa$
5:   $\mathcal{D}' \leftarrow \mathrm{SAMPLE}(p_{\theta_0}, B)$ ▷ unweighted samples.
6:   $W \leftarrow []$ ▷ array of weights.
7:   **for** $c(\tau_i) \in \mathcal{D}'$ **do**
8:       $W[i] \leftarrow \exp\left(-\lambda^{-1}c(\tau_i)\right)/\kappa/G[Z_T]$
9:   $W \leftarrow W/\mathrm{SUM}(W)$
10: **return** $\mathrm{STRATIFIEDSAMPLER}(\mathcal{D}', W)$

C   Additional experiments on synthetic data

First, we compare the performance of the optimal consequential ranking model computed via weighted sampling and the P-L ranking model learned using Algorithm 1 using the same quality metrics as in the previous section. Figure 4(a-b) summarizes the results. We observe that both the optimal and P-L consequential ranking models achieve similar values of immediate utility over time. However, the optimal model has a competitive advantage in terms of cost to the welfare, which becomes more pointed as $1/\lambda$ grows. These findings suggest that, the larger the value of $1/\lambda$, the more difficult is to learn a P-L model that approximates effectively the optimal model.

Next, we compare the scalability of both models in terms of the number of samples needed per ranking. Figure 4(c) summarizes the results, which shows that, as $1/\lambda$ grows, it becomes computationally prohibitive to generate optimal consequential rankings using weighted sampling due to the growing difference between p_θ^* and p_{θ_0}. This questions the practicality of weighted sampling to generate optimal consequential rankings.

D   Uncivility and unreliability scores

To estimate the uncivility score ϕ for each comment, we apply sentiment analysis on the text of the comments using the software package *Pattern*[8] and, for each comment, obtain two quantities:

---

[8]https://www.clips.uantwerpen.be/pages/pattern-en

(a) Immediate utility vs time   (b) Cost to welfare vs $1/\lambda$ (c) Average # samples per ranking
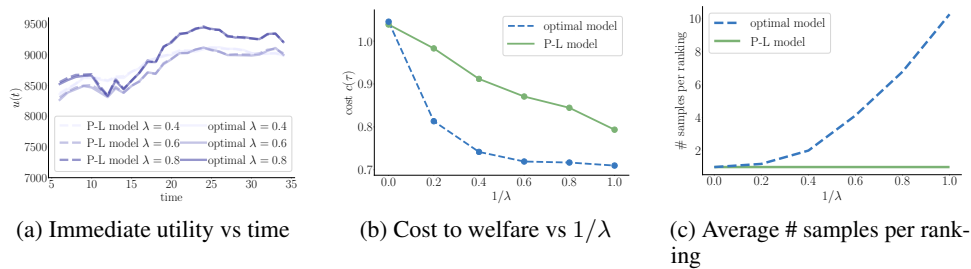
Figure 4: Performance and scalability of the optimal consequential ranking model and the P-L consequential ranking model learned using Algorithm 1. Panels (a-b) show that the optimal consequential ranking model achieves a better trade off between immediate utility and cost to the welfare than the P-L consequential ranking model. However, Panel (c) shows that the weighted sampling the optimal model uses to generate rankings quickly become computationally prohibitive in terms of # samples needed per ranking as $1/\lambda$ increases and the difference between the original ranking model and the optimal consequential ranking model increases.

mood and polarity. The mood of a comment can take one of the following four values: indicative, imperative, conditional and subjunctive. The polarity of a comment is a real number in $[-1, 1]$, where lower (higher) values indicate more negative (positive) words in the text. Then, we define the uncivility score ϕ of a comment as the absolute value of the polarity of the comment if the polarity is negative and the mood of the comment is indicative or imperative and zero otherwise

We estimate the unreliability score γ for each comment by estimating the average unreliability score of the domains that appeared in each of them, as estimated by aggregating publicly available data from Politifact and Snopes[9]. More specifically, our combined dataset contains fact checking information for $17,804$ unique urls from $4,540$ unique domains. For each url, it assigns a label that indicates the reliability of its content. We used these labels to assign a numerical unreliability score for each url. More specifically, if the url is labeled as "false", "pants-fire", "mfalse" or "legend", we set the unreliability score to 1. If the url is labeled as "true", "mtrue" or "mostly-true", then we set the unreliability score to -1. And, if the url is labeled using some other label value, we set the unreliability score to 0. We computed an unreliability score for each domain, which measures its level of (un)trustworthiness, by taking the average of the unreliability scores of individual urls from the domain. Then, we define the unreliability score γ of a comment as the average unreliability score of the domain(s) of the link(s) used in the comment if the average is negative and zero otherwise. Here, also note that, if a comment does not contain any links or the domain(s) of the link(s) does not appear in our dataset, we set the unreliability score for that comment to 0.

Tables 1 and 2 provide a few examples of comments with a high uncivility score and domains with a high unreliability score.

Comment	Uncivility (ϕ)
If you once tell a lie, the truth is ever after your enemy.	0.0
I dream of a world where your bigoted stupid ideas don't have the popular shield of faith.	0.1
Shut the f**k up and die already you POS warmongering profiteer.	0.4
Crap? Or pap. Take your pick.	0.8
i blame the evil KOCH BROTHERS!	1.0

Table 1: Examples of sentences with different levels of uncivility, as estimated by the feature ϕ. Comments with higher levels of uncivility typically correspond to those that use foul language.

---

[9]https://www.kaggle.com/arminehn/rumor-citation/version/3

Url	Misinformation (γ)
aids.gov	0.0
pbs.org	0.26
breitbart.com	0.56
lifeisajoke.com	1.0

Table 2: Examples of domains that spread different amounts of misinformation, as estimated by the feature γ.

E Additional experiments on real data

In the experiments on real data in the main paper, we have assumed that the original ranking model p_{θ_0} maximizes immediate utility. Moreover, we have taken for granted that there is a negative correlation between the immediate utility achieved by the consequential ranking model and the Kullback-Leibler (KL) divergence between its induced probability $p_\theta(\tau)$ and the probability $p_{\theta_0}(\tau)$ induced by the original ranking model. However, in practice, it may happen that the original ranking model is not optimal in terms of immediate utility maximization. This may be specially the case if the original ranking model greedily maximizes immediate utility and does not take into account how their proposed rankings change the user dynamics. Figure 5 demonstrates that, in our first set of experiments, there is indeed a negative correlation between the immediate utility and the KL divergence, however, in our second set of experiments, there is a positive correlation. This suggests that the original ranking model in the second set of experiments is suboptimal.
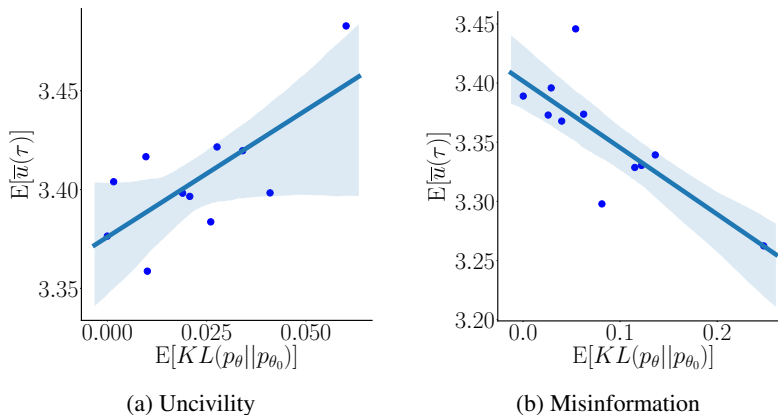

(a) Uncivility

(b) Misinformation

Figure 5: Average immediate utility $\mathbb{E}[\overline{u}(\tau)]$ achieved by the consequential ranking model vs the average Kullback-Leibler (KL) divergence $\mathbb{E}[KL(p_\theta \,||\, p_{\theta_0})]$ between its induced probability $p_\theta(\tau)$ and the probability $p_{\theta_0}(\tau)$ induced by the original ranking model.