On the Design of Consequential Ranking Algorithms

Behzad Tabibian
Reasonal

Vicenç Gomez
Universitat Pompeu Fabra

Abir De
MPI-SWS

Bernhard Schölkopf
MPI-IS

Manuel Gomez Rodriguez
MPI-SWS

Abstract

Ranking models are typically designed to 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 anticipate the consequences of their proposed rankings and 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 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 illustrate our methodology using synthetic and real data gathered from Reddit and show that our consequential rankings 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. They have undoubtedly increased the utility users obtain from online services. However, rankings have also been blamed for negative developments, particularly in the context of social and information systems, from fueling the spread of misinformation (Vosoughi et al., 2018), increasing polarization (Herrman, J., 2016) and degrading social discourse (Wong, 2017), to undermining democracy (Vaidhyanathan, 2017). 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* (Robertson, 1977), 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 anticipate 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, an existing ranking model receives a set of items and ranks these items on the basis of a measure of immediate (possibly unknown) utility[1] and a set of features. Items may appear over time and be present at several time steps. Moreover, their corresponding features may also 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—

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

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

the item—published by a user receives in social media depends largely on its ranking position (Hodas & Lerman, 2012; Gomez-Rodriguez et al., 2014; Lerman & Hogg, 2014; Kang & Lerman, 2015). 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 based on the number of posts including misinformation at the top of the rankings, averaged over time. Our goal is then to construct consequential ranking models that optimally trade off fidelity to the original ranking model maximizing immediate utility and long-term welfare[2].

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[3]. Then, we show that this representation greatly simplifies the construction of consequential ranking models that trade off fidelity to the rankings provided by a ranking 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 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, especially 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[4].

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

Related work. Our work relates to several lines of research: (i) ranking algorithms; (ii) delayed impact of machine learning algorithms; (iii) optimal control and reinforcement learning; and, (iv) reducing the spread of misinformation and polarization.

— *Ranking algorithms:* the work most closely related to ours is devoted to construct either fair rankings (Singh & Joachims, 2017, 2018, 2019; Zehlike et al., 2017) or diverse rankings (Carbonell & Goldstein, 1998; Clarke et al., 2008). However, this line of research defines fairness and diversity in terms of exposure allocation on an individual ranking rather than in a sequence of rankings. 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.

— *Delayed impact of ML algorithms:* the delayed impact of machine learning algorithms has not been studied until very recently (Hu & Chen, 2018; Liu et al., 2018; Mouzannar et al., 2019; Chen et al., 2019). However, most of these recent approaches have focused on classification tasks and have considered simple one-step feedback models. In contrast, in this work, we focus on rankings and consider a multiple step feedback model based on Markov decision processes (MDPs).

— *Optimal control and reinforcement learning:* the work most closely related to ours within the extensive literature on optimal control and reinforcement learning is devoted to improving the functioning of social and information systems (Wang et al., 2017; Zarezade et al., 2018). However, this line of work has mainly focused on representations based on temporal point processes and have not considered rankings. Recently, a framework based on survival process has been proposed to optimize click through rate using reinforcement learning (Zheng et al., 2018).

— *Reducing the spread of misinformation and polarization:* the literature on algorithms for reducing the spread of misinformation (Balmau et al., 2018; Kim et al., 2018; Tschiatschek et al., 2018) and reducing polarization (Garimella et al., 2017b,a) is expanding very rapidly (see Kumar & Shah (2018) for an excellent review of recent work). However, to the best of our knowledge, previous work has not approached the problem from the perspective of ranking algorithms.

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.

Problem setting. Let p be a particular ranking model (or, equivalently, ranking algorithm). At each time step $t \in \{1, \ldots, T\}$, the ranking model receives a set of

[2] In practice, one can only measure a welfare *proxy*, however, for brevity, we will refer to welfare proxy as welfare. Moreover, the effectiveness of our methodology will depend on the quality of the welfare proxies at our disposal.

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

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

n items and these items are characterized by a feature matrix $\boldsymbol{X}(t) \in \mathbb{R}^{n \times m}$, where the i-th row $\boldsymbol{X}_i(t)$ contains the feature values for item $i \in [n]$ and m 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 also 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 $\boldsymbol{\omega}_i(t)$ 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 (Gomez-Rodriguez et al., 2014; Hodas & Lerman, 2012; Kang & Lerman, 2015; Lerman & Hogg, 2014), 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 cost 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.

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) (Sutton & Barto, 2018), which also has an intuitive causal interpretation:

$$\begin{aligned} & p(\tau \,|\, \boldsymbol{X}(t_0), \boldsymbol{y}(t_0)) \\ & = \prod_{t=1}^{T} p(\boldsymbol{X}(t), \boldsymbol{y}(t) \,|\, \boldsymbol{X}(t-1), \boldsymbol{y}(t-1)) \\ & = \prod_{t=1}^{T} \underbrace{p(\boldsymbol{y}(t) \,|\, \boldsymbol{X}(t))}_{\text{ranking model}} \; \underbrace{p(\boldsymbol{X}(t) \,|\, \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))}_{\text{user dynamics}}, \quad (1) \end{aligned}$$

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 $\boldsymbol{X}(t')$, $t' < t - 1$.

(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 $\boldsymbol{y}(t')$, $t' < t - 1$.

We would like to highlight that, in most practical scenarios, ranking models optimizing for immediate utility satisfy the first assumption. However, depending on the choice of features, the second assumption may be violated and thus the representation of the user dynamics becomes an approximation. It would be very interesting, albeit challenging, to lift the second assumption in future work.

Next, we elaborate further on the specifics of the ranking model and the distribution of the user dynamics.

— *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 model, Plackett-Luce (P-L) ranking (Luce, 1977; Plackett, 1975), which is best described in terms of the orderings of the rankings. 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

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}(t) \,|\, \boldsymbol{X}(t)) = \prod_{k=1}^{n} f_k(\boldsymbol{X}(t)), \qquad (2)$$

with

$$f_k(\boldsymbol{X}(t)) = \frac{\exp\left(\boldsymbol{\theta}^T \boldsymbol{X}_{\omega_k}(t)\right)}{\sum_{k'=k}^{N} \exp\left(\boldsymbol{\theta}^T \boldsymbol{X}_{\omega_{k'}}(t)\right)}, \qquad (3)$$

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

[5] In our work, we consider probabilistic ranking models, which assign a probability to each ranking. Extending the methodology to deterministic ranking models is left for future work.

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 (Tran et al., 2016), *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_0 be an existing ranking model[6] that optimizes some hidden immediate utility and $c(\cdot)$ be a given cost to the welfare. Then, we construct a consequential ranking model p^*, 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:

$$\underset{p}{\text{minimize}} \quad \mathbb{E}_{\tau \sim p} \left[S(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) \right], \qquad (4)$$

with

$$S(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) = c(\tau) + \lambda \log \frac{p(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}{p_0(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}, \qquad (5)$$

where the expectation is taken over all the trajectories τ of feature matrices and rankings of length T induced by the ranking model p_0. 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. In Eq. 5, the parameter $\lambda \geq 0$ controls the trade off between the fidelity to the original ranking model and the long-term cost to the welfare. Note that, for $\lambda \to \infty$, the optimal ranking p^* coincides with the original ranking p_0. Moreover, 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 induced by the original model, since the terms associated to the user dynamics $p(\boldsymbol{X}(t) \,|\, \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))$ cancel.

Finally, note that, from the perspective of reinforcement learning, we are solving a *forward problem*, where the cost is given, rather than an *inverse problem*, where the cost is inferred. Moreover, our measure of fidelity has a natural interpretation in terms of the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), which has been extensively used as a distance measure between distributions, leading to a formulation of reinforcement learning as probabilistic inference (Levine, 2018; Kappen et al., 2012; Ziebart et al., 2010). 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.*,

$$\begin{aligned} KL[p(\cdot \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) \,||\, p_0(\cdot \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))] \\ = \mathbb{E}_{\tau \sim p} \left[\log \frac{p(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}{p_0(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))} \right]. \end{aligned}$$

In the next section, we will exploit this interpretation to greatly simplify the construction of consequential rankings.

3 BUILDING CONSEQUENTIAL RANKINGS

In this section, we tackle the optimization problem defined by Eq. 4 from the perspective of reinforcement learning and show that the optimal consequential ranking model p^* 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 (Bertsekas, 2000). 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:

$$\begin{aligned} V_t(\boldsymbol{X}, \boldsymbol{y}) = \min_{p} \; & \ell(\boldsymbol{X}, \boldsymbol{y}) \\ & + \lambda \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[V_{t+1}(\boldsymbol{X}', \boldsymbol{y}') \right] \qquad (6) \end{aligned}$$

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:

$$\begin{aligned} \ell(\boldsymbol{X}, \boldsymbol{y}) &= c(\boldsymbol{X}, \boldsymbol{y}) \\ &\quad + \lambda \mathbb{E}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p(\boldsymbol{X}', \boldsymbol{y}' \,|\, \boldsymbol{X}, \boldsymbol{y})}{p_0(\boldsymbol{X}', \boldsymbol{y}' \,|\, \boldsymbol{X}, \boldsymbol{y})} \right] \\ &= c(\boldsymbol{X}, \boldsymbol{y}) + \lambda KL(p(\cdot\,, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y}) \,||\, p_0(\cdot\,, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})). \end{aligned}$$

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

[6] In our experiments, we will approximate the existing ranking model using a P-L ranking model. We will fit the parameters of this P-L ranking model from historical rankings provided the original ranking model via regularized maximum likelihood estimation (MLE) (Hunter, 2004).

help us find an analytical solution. Inspired by Todorov (2009), 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 right hand side of Eq. 6 as

$$\min_{p} \mathop{\mathbb{E}}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p(\boldsymbol{X}', \boldsymbol{y}' \,|\, \cdot)}{p_0(\boldsymbol{X}', \boldsymbol{y}' \,|\, \cdot) Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')} \right],$$

where we have dropped λ and $c(\boldsymbol{X}, \boldsymbol{y})$ because they do not depend on p and, for brevity, we have replaced the conditionals (X, Y) inside the logarithm with $\cdot$. 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 $p(\boldsymbol{X}' \,|\, \boldsymbol{X}, \boldsymbol{y})$ cancel and we obtain:

$$\min_{p} \mathop{\mathbb{E}}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p(\boldsymbol{y}' \,|\, \boldsymbol{X}')}{p_0(\boldsymbol{y}' \,|\, \boldsymbol{X}') Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')} \right].$$

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 $p_0(\boldsymbol{y}' \,|\, \boldsymbol{X}') Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')$ is not normalized to one. If we multiply and divide the fraction by the following normalization term:

$$G[Z_{t+1}](\boldsymbol{X}') = \mathbb{E}_{\boldsymbol{y}' \sim p_0(\boldsymbol{y}' | \boldsymbol{X}')}[Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')], \qquad (7)$$

we obtain:

$$\begin{aligned} \min_{p} - & \mathop{\mathbb{E}}_{\boldsymbol{X}' \sim p(\cdot \,|\, \boldsymbol{X}, \boldsymbol{y})} [\log G[Z_{t+1}](\boldsymbol{X}')] \\ & + \mathop{\mathbb{E}}_{(\boldsymbol{X}', \boldsymbol{y}') \sim p(\cdot, \cdot \,|\, \boldsymbol{X}, \boldsymbol{y})} \left[\log \frac{p(\boldsymbol{y}' \,|\, \boldsymbol{X}') G[Z_{t+1}](\boldsymbol{X}')}{p_0(\boldsymbol{y}' \,|\, \boldsymbol{X}') Z_{t+1}(\boldsymbol{X}', \boldsymbol{y}')} \right]. \end{aligned}$$

Here, note that the first term does not depend on p 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:

$$p^*(\boldsymbol{y} \,|\, \boldsymbol{X}) = \frac{p_0(\boldsymbol{y} \,|\, \boldsymbol{X}) Z_{t+1}(\boldsymbol{X}, \boldsymbol{y})}{G[Z_{t+1}](\boldsymbol{X})}. \qquad (8)$$

The above equation reveals that the optimal consequential ranking model $p^*(\boldsymbol{y} \,|\, \boldsymbol{X})$ does implicitly depend on time due to Z_{t+1}. Finally, if we substitute back the above expression into the Bellman equation, given by Eq. 6, we can also find the function Z_t using the following recursive expression:

$$\begin{aligned} Z_t(\boldsymbol{X}, \boldsymbol{y}) = & \exp\big(-c(\boldsymbol{X}, \boldsymbol{y}) \\ & + \lambda \mathbb{E}_{\boldsymbol{X}' \sim p(\boldsymbol{X}' \,|\, \boldsymbol{X}, \boldsymbol{y})} [\log G[Z_{t+1}](\boldsymbol{X}')]\big), \end{aligned}$$

with $Z_T(\boldsymbol{X}, \boldsymbol{y}) = -\log c(\boldsymbol{X}, \boldsymbol{y})$. This 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 (Douc & Cappé, 2005), as shown in Algorithm 1, where $\textsc{Sample}(p_0, \kappa)$ samples κ trajectories from $p_0(\tau)$ and $\textsc{StratifiedSampler}(\mathcal{D}', W)$ generates $|\mathcal{D}'|$ samples weighted by W using stratified sampling.

Algorithm 1 It samples from an optimal consequential ranking model given p_0.

```
Require: Cost to welfare $c(\cdot)$, parameter $\lambda$, original ranking model $p_0$, $(\boldsymbol{X}(0), \boldsymbol{y}(0))$, # of samples $B$, # of samples $\kappa$ to compute $G[Z_T]$.

$\mathcal{D} \leftarrow \textsc{Sample}(p_0, \kappa)$        ▷ samples for estimating $G[Z_T]$.
$\Lambda[Z_T] \leftarrow 0$
for $c(\tau_i) \in \mathcal{D}$ do
    $\Lambda[Z_T] \leftarrow \Lambda[Z_T] + \exp\left(-\lambda^{-1} c(\tau_i)\right) / \kappa$
$\mathcal{D}' \leftarrow \textsc{Sample}(p_0, B)$        ▷ unweighted samples.
$W \leftarrow []$        ▷ array of weights.
for $c(\tau_i) \in \mathcal{D}'$ do
    $W[i] \leftarrow \exp\left(-\lambda^{-1} c(\tau_i)\right) / \kappa / G[Z_T]$
$W \leftarrow W / \textsc{Sum}(W)$
return $\textsc{StratifiedSampler}(\mathcal{D}', W)$
```

Unfortunately, in practice, these sampling methods may be inefficient and have high variance if the original ranking model p_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 $p^*(\boldsymbol{y} \,|\, \boldsymbol{X})$, which iteratively adapts a parameterized consequential ranking model $p^*_\theta(\boldsymbol{y} \,|\, \boldsymbol{X})$ using a stochastic gradient-based algorithm.

4 A GRADIENT-BASED ALGORITHM

In this section, our goal is to find a parameterized consequential ranking model p^*_θ within a class of parameterized ranking models $\mathcal{P}(\Theta)$ (*e.g.* PL models in Eq. 2) that approximates well the optimal consequential ranking model p^*, given by Eq. 8, *i.e.* $p^*_\theta \approx p^*$. To this aim, we minimize the parameterized version of the objective function in Eq. 4, *i.e.*,

$$\mathbb{E}_{\tau \sim p_\theta} \left[S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) \right]. \qquad (9)$$

where,

$$S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) = c(\tau) + \lambda \log \frac{p_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}{p_0(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))}$$

More specifically, 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

Algorithm 2 Training a parameterized consequential ranking model.

```
Require: Cost to welfare $c(\cdot)$, parameter $\lambda$, original ranking model $p_0$, $(\boldsymbol{X}(0), \boldsymbol{y}(0))$, # of iterations $M$, mini batch size $B$, and learning rate $\gamma$.
$\theta^{(0)} \leftarrow$ InitializeRankingModel()
for $j = 1, \ldots, M$ do                                   ▷ iterations
    $\mathcal{D} \leftarrow$ Minibatch($p_\theta, B$)                    ▷ sample mini batch
    $\nabla \leftarrow 0$
    for $\tau^{(i)} \in \mathcal{D}$ do
        $S \leftarrow c(\tau^{(i)}) + \lambda \log \frac{p_{\theta^{(j)}}(\tau^{(i)} \mid \boldsymbol{X}(0), \boldsymbol{y}(0))}{p_0(\tau^{(i)} \mid \boldsymbol{X}(0), \boldsymbol{y}(0))}$
        $\widetilde{\nabla} \leftarrow \nabla_\theta \log p_{\theta^{(j)}}(\tau^{(i)} \mid \boldsymbol{X}(0), \boldsymbol{y}(0))$
        $\nabla \leftarrow \nabla \; + \; (S + \; \lambda) \, \widetilde{\nabla}$
    $\theta^{(j+1)} \leftarrow \theta^{(j)} + \gamma \, \frac{\nabla}{B}$
return $\theta^{(M)}$
```

(SGD) (Kiefer & Wolfowitz, 1952), *i.e.*,

$$\theta^{(j+1)} = \theta^{(j)} + \gamma_j \nabla_\theta \, \mathbb{E}_{\tau \sim p_\theta} \left[S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) \right] |_{\theta = \theta^{(j)}} \,,$$

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 $\mathbb{E}_{\tau \sim p_\theta} [S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))]$ 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 Williams (1992), *i.e.*,

$$\begin{aligned} \nabla_\theta \mathbb{E}_{\tau \sim p_\theta} & [S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))] \\ &= \mathbb{E}_{\tau \sim p_\theta} [(S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) + \lambda) \times \\ & \qquad \nabla_\theta \log p_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))] \,, \end{aligned} \tag{10}$$

where $\nabla_\theta \log p_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))$ is often referred as the score function (Hyvärinen, 2005). This yields the following unbiased finite sample Monte-carlo estimator for the gradient:

$$\begin{aligned} \nabla_\theta \mathbb{E}_{\tau \sim p_\theta} & [\; S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))] \approx \\ & \sum_{i=1}^{B} \Big(S_\theta(\tau^{(i)} \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) + \lambda \Big) \times \\ & \qquad \nabla_\theta \log p_\theta(\tau^{(i)} \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)), \end{aligned} \tag{11}$$

where B is the number of sampled trajectories from the joint distribution $p_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))$ induced by the ranking model p_θ. The overall procedure is summarized in Algorithm 2, 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. 10, we only need to be able to sample from the user dynamics $p(\boldsymbol{X}(t) \,|\, \boldsymbol{X}(t-1), \boldsymbol{y}(t-1))$, since the explicit dependence cancels out within $S_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0))$, 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. More specifically, it readily follows from Eq. 2 that

$$\begin{aligned} \nabla_\theta \log p_\theta(\tau \,|\, \boldsymbol{X}(0), \boldsymbol{y}(0)) &= \nabla_\theta \sum_{t=1}^{T} \sum_{k=1}^{n} \log f_k(\boldsymbol{X}(t)) \\ = \nabla_\theta \sum_{t=1}^{T} \sum_{k=1}^{n} & \left(\theta^T \boldsymbol{X}_{\omega_k}(t) - \log \sum_{k'=k}^{n} \exp(\theta^T \boldsymbol{X} \omega_{k'}(t)) \right) \\ = \sum_{t=1}^{T} \sum_{k=1}^{n} & \left(\theta^T - \nabla_\theta \log \sum_{k'=k}^{n} \exp(\theta^T \boldsymbol{X} \omega_{k'}(t)) \right), \end{aligned}$$

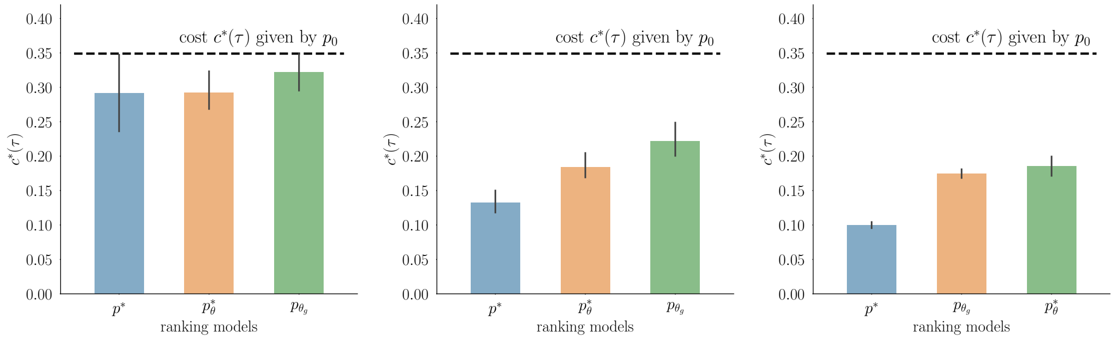
where the second term within the logarithm in the last equation is the derivative of the log-sum-exp function, whose analytical expression can be found elsewhere. Finally, if we think of the parameterized ranking model p_θ as a policy, our algorithm resembles policy gradient algorithms used in the reinforcement learning literature (Sutton & Barto, 2018). This connection opens up the possibility of using variance reduction techniques used in policy gradient to improve the empirical estimation of the gradient (Zhao et al., 2011).

5 EXPERIMENTS ON SYNTHETIC DATA

In this section, we compare the performance achieved by the original ranking models, which maximize an immediate measure of utility, the optimal consequential rankings models, implemented using Algorithm 1, the P-L consequential ranking model learned using Algorithm 2, and a non-trivial greedy baseline, which downranks items with high values of cost to welfare in an heuristic manner, using synthetic data.

Experimental setup. Each trajectory has length $T = 20$ and, at each time step $t \in \{1, \ldots, T\}$, the ranking model receives a set $\mathcal{I}(t)$ of $n = 4$ posts 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 $\mathcal{I}(t)$ where the $d \sim$ Poisson(1) posts at the bottom of the ranking $\boldsymbol{y}(t)$ are replaced by new posts.

Each post i has two features $\boldsymbol{X}_i(t) = [r_i, a_i(t)]$, where r_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 ($r_i = 0.6$) and low risk posts ($r_i = 0.1$) 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



(a) $\bar{u} \in (0.852, 0.862)$

(b) $\bar{u} \in (0.862, 0.872)$

(c) $\bar{u} \in (0.872, 0.882)$

Figure 1: Performance of the original ranking model p_0, the optimal consequential ranking model p^*, implemented using Algorithm 1, the P-L consequential ranking model p^*_θ, learned using Algorithm 2, and the greedy baseline p_{θ_g} on synthetic data. It shows the true cost to welfare $c^*(\tau)$ for three different ranges of average utility $\bar{u} = \sum_{t=1}^{T} u(t)/T$ for all models. Here, we tuned over the parameters λ (for p^* and p^*_θ) and d (for p_{θ_g}) to obtain the corresponding range for the average utility. The results show that the consequential ranking models p^* and p^*_θ outperform the greedy baseline p_{θ_g} in terms of the cost to welfare $c^*(\tau)$ and that the optimal consequential ranking model p^* performs best.

$m_i \sim \text{Bernoulli}(r_i)$, which is unobserved by the ranking model. The instantaneous rate of shares for each item i is given by:

$$a_i(t+1) = \exp(-2(t - s_i)) \times \tag{12}$$

$$(a_i(t) + \alpha_i + 0.02(5.0 - y_i(t))) \, , \tag{13}$$

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_0 aims to rank posts according to the number of shares $a(t)$ at each time t, *i.e.*, its immediate utility $u(t)$ is defined as

$$u(t) = \zeta(t) \tag{14}$$

where $\zeta(t)$ is the Kendall-Tau correlation between the ordering induced by the ranking $\boldsymbol{y}(t)$ and the ordering induced by the sorted items according to $a(t)$. To this aim, it uses a Plackett-Luce (P-L) model, given by Eq. 2, with $\theta = [0, 20]$.

The cost to welfare measures the long-term presence of misinformation on the top position of the rankings. More specifically, it is defined as

$$c(\tau) = \frac{1}{T} \sum_{t=1}^{T} r_{\omega_k(t)}. \tag{15}$$

Moreover, we compare the original ranking model with three ranking models, which aim to trade off fidelity to the original model and the cost to welfare:

(i) An optimal consequential ranking model p^*, which is implemented using Algorithm 1.

(ii) A Plackett-Luce (P-L) consequential ranking model p^*_θ, which is learned using Algorithm 2 with $M = 100$ iterations and $B = 50$ as batch size.

(iii) A greedy baseline p_{θ_g}, which is a P-L ranking model with parameters $\theta = [-d, \; 20]$, which downranks items i with nonzero misinformation probability, *i.e.*, $r_i > 0$. Here, d is a given parameter that controls how much we downrank such items.

For the P-L consequential ranking model and the greedy baseline, we experiment with different values of the parameters λ and d, respectively. Finally, for each experiment, we perform 8,000 repetitions.

Quality of the rankings. We compare the original ranking model p_0, the optimal consequential ranking model p^*, the P-L consequential ranking model p^*_θ and the greedy baseline p_{θ_g} in terms of two quality metrics: (i) the immediate utility $u(t)$, given by Eq. 14; and (ii) the true cost to welfare $c^*(\tau)$, defined as

$$c^*(\tau) = \frac{1}{T} \sum_{t=1}^{T} m_{\omega_1(t)}. \tag{16}$$

Figure 1 summarizes the results, which show that: (i) the (optimal and P-L) consequential ranking models outperform the greedy baseline in terms of the cost to welfare, for three different ranges of utility; (ii) the optimal consequential ranking achieves a significantly better tradeoff between the fidelity to the original ranking model and the cost to welfare, than the P-L ranking model, as one may have expected; and, (iii) the optimal consequential ranking model reduces the (true) cost to welfare without decreasing its fidelity to the original ranking model.

In Appendix B we provide more experiments on synthetic data. More specifically, we compare the running time of Algorithm 1 and Algorithm 2 and we show that the optimal consequential ranking model ranks viral and non-viral high-risk posts differently.

6 EXPERIMENTS ON REAL DATA

In this section, we compare the performance achieved by the original ranking models, which maximize an immediate measure of utility, the P-L consequential ranking model learned using Algorithm 2, and the same greedy baseline introduced in Section 5 using Reddit data[7]. 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) We consider a batch reinforcement learning setting. 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.

Dataset description. We used a publicly available Reddit dataset[8], which contains (nearly) all publicly available comments to link submissions posted by Reddit users from 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. The average length of a comment thread in our dataset is 21, with median of 17 and maximum length of 60. In a first set of experiments, we focus on the civility of the comments in each submission, as measured by an incivility 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 C contains more details on the definition and estimation of both scores. In both sets of experiments, we use 1,973 submissions as training set for learning the parameterized consequential ranking models and the remaining 1,200 submissions as test set for evaluation, and we repeat our experiments for three different random sets of training and test sets.

Experimental setup. 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 $n = 5$ comments $\mathcal{I}(t)$[9].

Each comment i has three features $\boldsymbol{X}_i(t) = [l_i, \phi_i, \gamma_i]$, where l_i is the number of comments posted until time step i, ϕ_i is the incivility score and γ_i is the unreliability score. At each time t, the original ranking model p_0 aims to promote the most recent comment to the top of the ranking, *i.e.*, its immediate utility $u(t)$ is defined as

$$u(t) = \zeta(t) \tag{17}$$

where $\zeta(t)$ is the Kendall Tau correlation between the ordering induced by the ranking $\boldsymbol{y}(t)$ and the inverse chronological ordering. To this aim, it uses a Plackett-Luce (P-L) model, fitted by maximizing the likelihood function over traces with reverse chronological order.

In the first set of experiments, the cost to welfare measures the long-term presence of uncivil comments on the top position of the rankings. More specifically, it is defined as

$$c(\tau) = \frac{1}{T} \sum_{t=1}^{T} \phi_{\omega_1(t)}. \tag{18}$$

In the second set of experiments, the cost to welfare measures the long-term presence of unreliable comments on the top position of the rankings. More specifically, it is defined as

$$c(\tau) = \frac{1}{T} \sum_{t=1}^{T} \gamma_{\omega_1(t)}. \tag{19}$$

Similarly as in Section 5, we compare the original ranking model with two ranking models, which aim to trade off fidelity to the original model and the cost to welfare:

(i) A Plackett-Luce (P-L) consequential ranking model p^*_θ, which is learned using Algorithm 2 with $M = 20$ iterations and $B = 100$ as batch size; and,

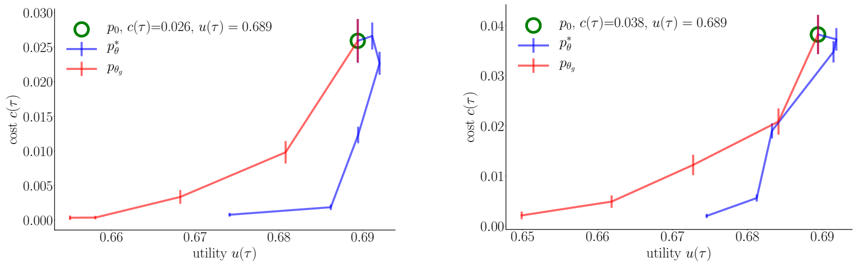
(ii) A greedy baseline p_{θ_g} with parameters $\theta = [1,\ -d,\ 0]$ for the first set of experiments and $\theta = [1,\ 0,\ -d]$ for the second set of experiments. Here, the greedy baseline downranks items i with nonzero incivility (or unreliability) score *i.e.*, $\phi_i > 0$ (or $\gamma_i > 0$).

For both ranking models (i-ii), we experiment with different values of the parameters λ and d, respectively. Finally, for each experiment, we perform 8,000 repetitions.

[7] Due to the size of the dataset, we were unable to run Algorithm 1 and thus we could not experiment with the optimal consequential rankings model.

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

[9] Experiments with $n > 5$ give qualitatively the same results because comments with high score of incivility/unreliability are rare in the dataset.



(a) Uncivility

(b) Misinformation

Figure 2: Cost to welfare $c(\tau)$ vs. utility $u(\tau)$ achieved by the original ranking model p_0, the consequential PL-ranking model p^*_θ and the greedy baseline p_{θ_g} on Reddit data. The consequential PL-ranking model p^*_θ achieves a better trade off between the fidelity to the original ranking model ranking models optimizing immediate utility and the long-term welfare than the greedy baseline p_{θ_g}.

Results. We first compare the original ranking model p_0, the consequential P-L ranking models p^*_θ and the greedy baseline p_{θ_g} in terms of the tradeoff between cost to welfare $c(\tau)$ and the immediate utility given by Eq. 17. Here, note that, in the first set of experiments, the cost to welfare measures the degree of incivility of the top ranking positions (Eq. 18) while, in the second set of experiments, it measures the amount of misinformation (Eq. 19). Figure 2 summarizes the results, which shows that (i) our consequential PL-ranking model p^*_θ can trade off between the fidelity to ranking models optimizing immediate utility and the long-term welfare more effectively than the greedy baseline p_{θ_g}, and (ii) the PL-ranking model p^*_θ is able to reduce the degree of incivility 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 (1) the fidelity to ranking models optimizing for immediate utility and (2) long-term welfare. More specifically, we have first introduced a joint representation of rankings and user dynamics using Markov decisions processes. Exploiting this representation, we have shown that we can obtain optimal consequential rankings just by applying weighted sampling on the rankings provided by the model optimizing for immediate utility. However, in practice, such a strategy may be inefficient and impractical, specially in high dimensional scenarios. To overcome this, we introduced an efficient gradient-based algorithm to learn parameterized consequential ranking models that effectively approximate the optimal ones. Finally, we have experimented on synthetic and real data to show the efficacy of our parameterized consequential ranking models.

Our work opens up several venues 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. Finally, we have evaluated our algorithm using observational real data, however, it would be very valuable to perform interventional experiments.

Acknowledgments. V. Gomez has received funding from “la Caixa” Foundation (ID 100010434), under the agreement LCF/PR/PR16/51110009 and by the Volkswagen Foundation, Courage project (95.566).

References

Balmau, O., Guerraoui, R., Kermarrec, A., Maurer, A., Pavlovic, M., & Zwaenepoel, W. 2018. Limiting the Spread of Fake News on Social Media Platforms by Evaluating Users’ Trustworthiness. *arXiv:1808.09922*.

Bertsekas, D. 2000. *Dynamic Programming and Optimal Control*. 2nd edn. Athena Scientific.

Carbonell, J., & Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. *In: SIGIR*.

Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., & Chi, E. 2019. Top-k off-policy correction for a REINFORCE recommender system. *In: WSDM*.

Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., & MacKinnon, I. 2008. Novelty and diversity in information retrieval evaluation. *In: SIGIR*.

Douc, Randal, & Cappé, Olivier. 2005. Comparison of resampling schemes for particle filtering. *In: ISPA*.

Garimella, K., Gionis, A., Parotsidis, N., & Tatti, N. 2017a. Balancing information exposure in social networks. *In: NeurIPS*.

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. 2017b. Reducing controversy by connecting opposing views. *In: WSDM*.

Gomez-Rodriguez, M., Gummadi, K. P., & Schoelkopf, B. 2014. Quantifying Information Overload in Social Media and Its Impact on Social Contagions. *In: ICWSM*.

Herrman, J. 2016. Inside Facebook's Political-Media Machine. *New York Times*.

Hodas, N., & Lerman, K. 2012. How visibility and divided attention constrain social contagion. *In: SocialCom*.

Hu, L., & Chen, Y. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. *In: WWW*.

Hunter, D. 2004. MM algorithms for generalized Bradley-Terry models. *Ann. Stat.*, **32**(1).

Hyvärinen, A. 2005. Estimation of Non-Normalized Statistical Models by Score Matching. *JMLR*, **6**, 695–709.

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

Kappen, H., Gómez, V., & Opper, M. 2012. Optimal control as a graphical model inference problem. *Mach. Learn.*, **87**(2).

Kiefer, J., & Wolfowitz, J. 1952. Stochastic estimation of the maximum of a regression function. *Ann. of Math. Stat.*, **23**(3).

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., & Gomez-Rodriguez, M. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. *In: WSDM*.

Kullback, S., & Leibler, R. 1951. On information and sufficiency. *Ann. of Math. Stat.*, **22**(1).

Kumar, S., & Shah, N. 2018. False information on web and social media: A survey. *arXiv:1804.08559*.

Lerman, K., & Hogg, T. 2014. Leveraging position bias to improve peer recommendation. *PloS one*, **9**(6).

Levine, S. 2018. Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review. *arXiv:1805.00909*.

Liu, L., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. 2018. Delayed Impact of Fair Machine Learning. *In: ICML*.

Luce, R. 1977. The choice axiom after twenty years. *Journal of mathematical psychology*, **15**(3).

Mouzannar, H., Ohannessian, M., & Srebro, N. 2019. From Fair Decision Making to Social Equality. *In: FAT**.

Plackett, R. 1975. The analysis of permutations. *Applied Statistics*.

Robertson, S. 1977. The probability ranking principle in IR. *J. Doc.*, **33**(4).

Singh, A., & Joachims, T. 2017. Equality of Opportunity in Rankings. *In: NeurIPS workshop*.

Singh, A., & Joachims, T. 2018. Fairness of Exposure in Rankings. *In: SIGKDD*.

Singh, A., & Joachims, T. 2019. Policy Learning for Fairness in Ranking. *In: NeurIPS*.

Sutton, Richard S, & Barto, Andrew G. 2018. *Reinforcement learning: An introduction*. MIT press.

Todorov, E. 2009. Efficient computation of optimal actions. *PNAS*, **106**(28).

Tran, T., Phung, D., & Venkatesh, S. 2016. Choice by elimination via deep neural networks. *arXiv:1602.05285*.

Tschiatschek, S., Singla, A., Gomez-Rodriguez, M., Merchant, A., & Krause, A. 2018. Fake News Detection in Social Networks via Crowd Signals. *In: WWW*.

Vaidhyanathan, Siva. 2017. Facebook wins, democracy loses. *New York Times*.

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

Wang, Y., Williams, G., Theodorou, E., & Song, L. 2017. Variational policy for guiding point processes. *In: ICML*.

Williams, Ronald J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, **8**(3-4).

Wong, J. 2017. Former Facebook executive: social media is ripping society apart. *The Guardian*.

Zarezade, A., De, A., Upadhyay, U., Rabiee, H., & Gomez-Rodriguez, M. 2018. Steering Social Activity: A Stochastic Optimal Control Point of View. *JMLR*.

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

Zhao, T., Hachiya, H., Niu, G., & Sugiyama, M. 2011. Analysis and improvement of policy gradient estimation. *In: NeurIPS*.

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N., Xie, X., & Li, Z. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. *In: WWW*.

Ziebart, B., Bagnell, J., & Dey, A. 2010. Modeling Interaction via the Principle of Maximum Causal Entropy. *In: ICML*.