Altruistic Collective Action in Recommender Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

Users of online platforms based on recommendation systems (RecSys) (e.g., TikTok, X, YouTube) *strategically* interact with content to influence future recommendations. On some platforms, users have been documented to form large-scale grassroots collectives encouraging others to purposefully interact with algorithmically suppressed content in order to "boost" its recommendation; we term this behavior *user altruism*. We study a game between users and a RecSys, where users provide (potentially manipulated) ratings of platform content, and the RecSys—limited by preference learning ability—provides each user her approximately most-preferred item. We compare users' social welfare under truthful preference reporting and under a class of collective strategies capturing user altruism. In our theoretical analysis, we provide sufficient conditions to ensure *strict* increases in user social welfare under user altruism and provide an algorithm to find an effective collective strategy. Interestingly, for commonly assumed recommender utility functions, strategies also improve the welfare of the RecSys! Our theoretical analysis is complemented by simulations of collective strategies on the GoodReads dataset, and an online survey of real users' altruistic behaviors. Our findings serve as a proof-of-concept of the reasons why RecSys may incentivize users to collectivize and interact with content altruistically. Indeed, the class of actions we present improve a minority group's welfare while not decreasing the welfare of any other user. Thus, as long as there exist even minimally altruistic agents, the RecSys implicitly incentivizes agents to perform algorithmic collective action when possible.

1 Introduction

Social networking and media platforms such as TikTok, Instagram, and YouTube learn about users by tracking their interactions with content via likes, comments, views, etc. Using these interactions, they can infer user preferences and provide new content recommendations. While the technical algorithm is a black box, the basic idea of how it functions is well-known to users [21, 24]. Using heuristics, some users strategically interact with content to purposefully tailor the recommendations they receive [14, 45, 65]. Consider, for example, a user who enjoys a type of niche content. Though this content is not generally popular, the user would like it to feature prominently in their feed. Knowing that engaging with content will likely cause it to be recommended again in the future, the user purposefully likes, comments, and watches this type of content **more often** than they personally would like in the moment. There are several game theoretic models of self-interested strategic behavior [34, 15].

However, existing theoretical models do not acknowledge that recommendations are informed by both a user's own interactions and the interactions of **other** users. Collaborative filtering, a widely-adopted recommendation methodology [35, 63], allows the recommender to infer the preferences of one user from another by embedding users into a shared latent space. When one user engages with content, the recommender may predict how much engagement others would have with the item based on the similarity of user embeddings. Real users also have a heuristic understanding of how their engagements impact others' recommendations, and use this information to interact strategically to impact other users' platform experiences [45, 24]. This behavior has been particularly relevant for users who want to combat algorithmic suppression and injustice. Indeed, Karizat et al. [45] document that users believe certain types of content are suppressed by the algorithm and thus *altruistic*[1] users *intentionally* engage with such content to "boost" it via algorithmic recommendation to others.

On an individual level, it is unlikely that *one* altruistic user attempting to boost the popularity of suppressed content would impact algorithmic recommendations on a large scale. However, as many marginalized user and creator groups report recommendation inequity or suppression [53, 32, 10], users have organized *large-scale grassroots movements* encouraging other users to *purposefully like, comment, interact and follow the content of those who are suppressed by the algorithm* [18, 52].

1.1 Our contributions

Model. In Section 2, we present a game-theoretic model of users' correlated strategies in recommendation systems (RecSys) and we provide a proof-of-concept analysis of the reasons why *collaborative filtering or matrix factorization-based RecSys may incentivize users to collectivize and behave altruistically* as documented by HCI research and real-world events. To the best of our knowledge, ours is the *first* theoretical model capturing *strategic* behavior of users in RecSys that are *not self-interested*.

Altruistic behavior. We focus on a large class of settings where the preferences of the population over different items satisfy a *majority-minority* relationship, i.e., users are mostly clustered by preferences with groups being more/less mainstream. In Section 3, we compare the recommendations and social welfare that users receive when they interact with content according to their *personal* preferences to those they receive under a class of simple *correlated* interaction strategies; we call these strategies "altruistic". Under reasonable conditions, our class of correlated strategies improves social welfare and recommendations beyond the truthful interaction baseline. We construct an algorithm, robust to misspecification in the information shared by users, to find such a strategy. These results provide the first theoretical groundwork for the documented collectivist behavior.

RecSys utility. Interestingly, considering two commonly used utility functions for the RecSys (engagement-based and user-welfare-based), users' collective action is also good for the recommender (Section 3.4)! Intuitively, this is because the strategies improve recommendation by increasing the users' total engagement with platform content, which in turn enables a platform to sell more ads.

Empirical results. We conduct two empirical studies to supplement our theoretical results. The first (Section 4.1) is an experiment Goodreads book reviews. Similarly to movements in the prominent "BookTok" community where TikTok users organize to interact with content from marginalized groups to battle unfair algorithmic promotion [52], we simulate collectives of readers of the most popular genre increasing engagement with less popular book genres. This **improves minority group welfare** by as much as **15 times**! The second (Section 4.2) is a survey given to 100 Prolific participants. We find that the proportion of users who intentionally attempt to impact other people's recommended feeds is relatively large and provide (textual) descriptions of users' underlying altruistic reasoning.

Finally, Section 5 includes a discussion on our model assumptions and future research directions.

1.2 Related work

Details on connections to Human Computer Interaction (HCI), Algorithmic Collective Action, Theoretical RecSys Modeling, Strategic Classification, and Matrix Completion are in Appendix B.1.

2 Model and preliminaries

Notation. Matrices are capital, bolded (i.e. $\mathbf{X} \in \mathbb{R}^{m\times n}$), vectors are lower-case, bolded (i.e. $\mathbf{z} \in \mathbb{R}^d$), and one-dimensional variables are lower-case (i.e. $y \in \mathbb{R}$). Of a matrix, $\mathbf{X}$, the ith *column* is $\mathbf{X}_i$ (an exception to lowercase vectors), the jth *row* is $\mathbf{x}_j$, and the jth row, ith column element is $x_{j,i}$. Sets are capital calligraphic letters (i.e. $\mathcal{U}$). The complement is $\mathcal{U}^C$. A table of notation is in Appendix A.

Model summary. We model a setting where a RecSys (aka *"learner"*) wishes to recommend an item from a set of n to each user (aka *"agent"*) from a set of m. $\mathbf{R}^\star \in \mathbb{R}^{m\times n}$ is the *ground truth* personal preference matrix for the m users over the n items. $\widetilde{\mathbf{R}} \in \mathbb{R}^{m\times n}$ is the *revealed preference matrix*. Each element, $\tilde{r}_{u,i}$, is a numerical representation of a user's interactions (such as likes, watches, etc) with an item, i, and is called a "rating". Users may interact with items differently from their true

[1]"Altruism" indicates that a user's utility function depends positively on the welfare of others as in Becker [5]

Protocol 1 Learner's protocol

LEARNING PHASE:

Learner gets $\widehat{\mathbf{R}}$, the $k^\star$-truncated SVD of $\widetilde{\mathbf{R}}$ s.t. $k^\star < \text{rank}(\widetilde{\mathbf{R}})$. `// "learned" preferences`

RECOMMENDATION PHASE:

Learner shows agents $u \in [m]$ their top item $\texttt{top}(u) \in [n]$ according to $\widehat{\mathbf{R}}$.

personal preferences (i.e. $\widetilde{\mathbf{R}} \neq \mathbf{R}^\star$). Using $\widehat{\mathbf{R}}$, a transformation of $\widetilde{\mathbf{R}}$, the learner gives each agent an estimated top item. The process $\mathbf{R}^\star \to \widetilde{\mathbf{R}} \to \widehat{\mathbf{R}}$ abstracts matrix completion (MC) preference learning and recommendation. MC is discussed briefly in Section 2.1 and at length in Appendix C.1.

2.1 Learner's protocol

The learner (he) wants to recommend each agent (she, when referred to individually) her top item. However, access to a complete and perfect preference matrix is unrealistic in practice. Prominent recommendation methodologies (e.g., matrix completion) query user ratings until a stopping point and then approximate unknown preferences via rank minimization (see Appendix C.1). We adopt the tractable abstraction of *low-rank approximation* on $\widetilde{\mathbf{R}}$. The learner sees $\widehat{\mathbf{R}}$, a representation of what he may have learned from agents' realized item interactions. The formalization is given in Protocol 1.

Three remarks are in order. First, we leave discussion of the learner's *utility* to Section 3.4. Second, if the RecSys must learn preferences through matrix completion, he queries ratings until his low rank approximation represents users' preferences well. Expected exploration length should depend on information retention of different rank reductions of the unknown complete matrix. We capture this through the optimization of Definition 2.1. Third, each user u's recommended item, $\texttt{top}(u)$, comes from $\widehat{\mathbf{R}}$. In other words, $\texttt{top}(u)$ may be different than the user's truthfully most preferred item.

The *Social Welfare* of the system is a measure of how good recommendations are in sum:

$$\text{SW}(\widetilde{\mathbf{R}}, \alpha) = \textstyle\sum_{u \in [m]} r^\star_{u,\texttt{top}(u)},$$

where α is a learner parameter related to exploration length, formally defined below. SW depends on the true preferences and the recommendation $\texttt{top}(u)$, which depends on revealed preferences and α.

2.1.1 Learning phase

The recommender gets $\widehat{\mathbf{R}}$, a reduced information version of the fully realized ratings of $\widetilde{\mathbf{R}}$. $\widehat{\mathbf{R}}$ is a $k^\star$-truncated SVD, i.e., $\widehat{\mathbf{R}} = \sum_{j \in [k^\star]} \sigma_j \mathbf{u_j} \mathbf{v_j}^\top$. Where $\widetilde{\mathbf{R}} = \sum_{j \in [\text{rank}(\widetilde{\mathbf{R}})]} \sigma_j \mathbf{u_j} \mathbf{v_j}^\top$.

What is $k^\star$? In the matrix completion analogy, $k^\star$ is the rank of the estimated complete preference matrix after some ratings are queried. $k^\star$ should be such that the estimated matrix represents the variation of preferences well while not requiring too many queries. How matrix completion algorithms deal with this in practice varies. We will model this process as a learner who has a marginal "budget" of "variation" he can allow himself to lose. See Appendix C.2 for a formal discussion of variation.

Definition 2.1 **(α-loss tolerant learner)** *An α variance loss tolerant learner gets $\widehat{\mathbf{R}}$ where $k^\star$ is:*

$$\min_{k \in [\text{rank}(\widetilde{\mathbf{R}})]} k \quad \textit{s.t.,} \quad \sigma_{k+1}(\widetilde{\mathbf{R}}) \leq \alpha$$

Recall from principal component (PC) analysis, that the kth singular value captures the relative variation retained by the kth PC. An α-loss tolerant learner has the lowest rank $\widehat{\mathbf{R}}$, such that increasing rank by 1 will not improve the proportion of information retained from $\widetilde{\mathbf{R}}$ according to budget α.

2.1.2 Recommendation phase

The learner estimates top item(s) for each user, $\mathcal{I}_{\text{top}}(u) := \arg\max_{i \in [n]} \hat{r}_{u,i}$. He breaks ties in favor of the *most popular* of the top-rated items $\mathcal{I}^{\text{pop}}_{\text{top}}(u)$, i.e.,

$$\texttt{top}(u) \sim \text{Unif}(\mathcal{I}^{\text{pop}}_{\text{top}}(u)), \quad \text{where} \quad \mathcal{I}^{\text{pop}}_{\text{top}}(u) := \max_{i \in \mathcal{I}_{\text{top}}(u)} \quad \|\widehat{\mathbf{R}}_i\|_1$$

2.2 Agent preferences

For the main body, we focus on a simple class of ground truth preference matrices.

Definition 2.2 (Majority-minority matrix) $\mathbf{R} \in \mathbb{R}_{\geq 0}^{m \times n}$ *is a* majority-minority matrix *if there exists a partition of users* $\mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{\text{MIN}} = [m]$ *(where* $\mathcal{U}_{\text{MAJ}}, \mathcal{U}_{\text{MIN}} \subseteq [m]$*) and a partition of items* $\mathcal{I}_{\text{MAJ}} \cup \mathcal{I}_{\text{MIN}} = [n]$ *(where* $\mathcal{I}_{\text{MAJ}}, \mathcal{I}_{\text{MIN}} \subseteq [m]$*) such that if* $u \in \mathcal{U}_{\text{MAJ}}$ *and* $i \in \mathcal{I}_{\text{MIN}}$ *or* $u' \in \mathcal{U}_{\text{MIN}}$ *and* $i' \in \mathcal{I}_{\text{MAJ}}$*, then* $r_{u,i}, r_{u'i'} = 0$*. Further, no user has ratings of all 0s (i.e.,* $\forall u \in [m]$*,* $\sum_{i \in [n]} r_{u,i} > 0$*).*

These are matrices where each user group has an exclusively preferred set of items. In Appendix G, we present analogous results under a more complex class of non-exclusive preference matrices.

Remark 2.1 *The welfare of each user is invariant under any re-ordering of users and items.*

The proof of this remark is in Appendix C.3. We will order the rows and columns of the true preference matrices such that for some $\bar{m} \in [m]$ and $\bar{n} \in [n]$, $\mathcal{U}_{\text{MAJ}} = [\bar{m}]$ and $\mathcal{I}_{\text{MAJ}} = [\bar{n}]$. Therefore, $\mathbf{R}^\star$ is a block-diagonal matrix where the blocks are: $\mathbf{R}^\star_{\text{MAJ}} \in \mathbb{R}_{\geq 0}^{\bar{m} \times \bar{n}}$ and $\mathbf{R}^\star_{\text{MIN}} \in \mathbb{R}_{\geq 0}^{(m-\bar{m}) \times (n-\bar{n})}$. We name these matrix blocks "majority" and "minority" because they represent user groups that are more/less dominant in the system, respectively. To represent dominance mathematically, consider:

Assumption 2.1 (Singular Value Gap) *Let* $\mathbf{R}$ *be a majority-minority matrix,* $k_{\text{MAJ}} = \operatorname{rank}(\mathbf{R}_{\text{MAJ}})$*, and* $\mathcal{G}(\mathbf{R}) := (\sigma_1(\mathbf{R}_{\text{MIN}}), \sigma_{k_{\text{MAJ}}}(\mathbf{R}_{\text{MAJ}}))$*. If* $\mathcal{G}(\mathbf{R}) \neq \emptyset$*, then* $\mathbf{R}$ *has a* singular value gap.

The preference matrices of the main body will be majority-majority matrices with a singular value gap. See Appendix C.4 for an example. Recall that for block-diagonal matrices, SVD is the sum the blocks' SVDs. We can use this to compute $k^\star$ when the learner is α-loss tolerant such that $\alpha \in \mathcal{G}(\widetilde{\mathbf{R}})$:

Proposition 2.1 *For any revealed majority-minority preference matrices* $\widetilde{\mathbf{R}}$ *with a singular value gap and any* α*-loss tolerant learner where* $\alpha \in \mathcal{G}(\widetilde{\mathbf{R}})$ *it must be the case that* $k^\star = k_{\text{MAJ}}$*.*

Rating. We consider two agent behavioral models: *truthful* and *altruistic*. *Truthful* agents interact with content according to their true, personal preferences: $\tilde{r}_{u,i} = r^\star_{u,i}$. *Altruistic* agents gain utility from *other users* having good recommendations. Does the system incentivize truthfulness to personal preference for altruistic agents? Are there computationally tractable dominating rating strategies?

3 Welfare Analysis

We will see that when there exist (1) a majority class of users who like popular items (2) a minority class who like niche items (3) limited RecSys exploration, too little information may be learned about minority preferences. This yields good majority recommendations, but popular item recommendations to the minority. Fortunately, if altruistic majority users purposefully interact with niche content, they may force the learner to be sensitive to minority preferences and strictly improve recommendations!

3.1 Social welfare of majority-minority preference matrices under truthfulness

In a majority-minority matrix satisfying Assumption 2.1, the majority is more "important" in the low-rank approximation. If $\widetilde{\mathbf{R}} = \mathbf{R}^\star$, some learners will only recommend accurately to the majority.

Theorem 3.1 (Truthfulness is good for majority, bad for minority) *Let* $\mathbf{R}^\star$ *be a majority-minority matrix satisfying Assumption 2.1 and* $\mathbf{R}^\star = \widetilde{\mathbf{R}}$*. If* $\alpha \in \mathcal{G}(\mathbf{R}^\star)$ *(the singular value gap), then majority users get their top item, while minority users get popular items they do not like. Formally:*

$$r^\star_{u,\texttt{top}(u)} = \begin{cases} \max_{i \in [n]} r^\star_{u,i} & u \in \mathcal{U}_{\text{MAJ}} \\ 0 & u \in \mathcal{U}_{\text{MIN}} \end{cases}, \quad \text{SW}(\mathbf{R}^\star, \alpha) = \textstyle\sum_{u \in \mathcal{U}_{\text{MAJ}}} \arg\max_{i \in [n]} r^\star_{u,i}$$

3.2 Improving social welfare via simple collective rating strategies

Can collaborative rating distortion force the learner to retain minority information? We focus on aiding "picky" minority users, who are "hard to learn" because in MC, all their preferences may be estimated as 0 if exploration is insufficient (Assumption C.1 and Thm C.2). In a worst-case sense, they are particularly in need. Appendix results use a weaker version of pickiness (Assumption G.5).

Definition 3.1 (Picky Users) *We say that item* $i^\star$ *is a picky item with picky user group* $\mathcal{U}_{i^\star}$ *if*

$$r_{u,i^\star} > 0 \iff u \in \mathcal{U}_{i^\star}, \forall u \in [m] \quad \textit{and} \quad r_{u,i} = 0 \; \forall u \in \mathcal{U}_{i^\star}, \; \forall i \neq i^\star$$

For the following results, we model collectives of majority users *uprating* a picky minority item by a collaboratively selected amount, η. Generalized misreporting strategies are analyzed in Appendix G.

$(\eta, \mathcal{U}_\mathrm{A})$-Altruistic uprating. Let $\mathbf{R}^\star$ be a majority-minority matrix, and let $\mathcal{U}_\mathrm{A} \subseteq \mathcal{U}_\mathrm{MAJ}$ be an altruistic subset of majority users. All $u \in \mathcal{U}_\mathrm{A}$ rate $\eta \in \mathbb{R}_{>0}$ for picky item $i^\star > \bar{n}$, other ratings are truthful. Thus, $\widetilde{\mathbf{R}}$ is the same as $\mathbf{R}^\star$ except for elements indexed $(u, i^\star), \forall u \in \mathcal{U}_\mathrm{A}$.

Next, we derive sufficient conditions on $\mathcal{U}_\mathrm{A}$ and η such that if $\widetilde{\mathbf{R}}$ were reported, the picky users *and* all majority users have maximized welfare. To do so, we define another useful singular value gap.

Definition 3.2 **($(\eta, \mathcal{U}_\mathrm{A})$-Sufficient Singular Value Gap)** *For a given $\eta \in \mathbb{R}_{>0}$, $\mathcal{U}_A \subseteq \mathcal{U}_\mathrm{MAJ}$, and majority-minority preference matrix, $\mathbf{R}$, define the following space, $\mathcal{G}(\mathbf{R}, \mathcal{U}_A, \eta)$:*

$$\mathcal{G}(\mathbf{R}, \mathcal{U}_A, \eta) := \left(\sigma_1(\mathbf{R}_\mathrm{MIN}), \sqrt{\min\{\sigma_{k_\mathrm{MAJ}}(\mathbf{R}_\mathrm{MAJ})^2, \eta^2|\mathcal{U}_A| + \mathrm{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\mathrm{AV}_{i^A_{\bar{n}}}}\right)$$

where $\mathrm{ASV}_{i^\star} = \|\mathbf{R}_{i^\star}\|_2^2$ is the aggregate square value of item $i^\star$ and $\mathrm{AV}_{i^A_{\bar{n}}} = \max_{i \in [\bar{n}]} \sum_{u \in \mathcal{U}_A} r_{u,i}$ is the largest aggregate value of a popular item for altruists.

Now we are ready to formally state when altruistic strategies are effective:

Theorem 3.2 **(Social Welfare as a function of $\mathcal{U}_\mathrm{A}$ and η)** *Let $\mathbf{R}^\star$ be a majority-minority matrix with a picky item $i^\star > \bar{n}$ and some $(\eta, \mathcal{U}_\mathrm{A})$-altruistic uprating such that $\mathbf{R}^\star$ has $(\eta, \mathcal{U}_\mathrm{A})$-sufficient singular value gap. If $\eta < \min_{u \in \mathcal{U}_\mathrm{MAJ}} \max_{i \in [n]} r^\star_{u,i}$ and $\alpha \in \mathcal{G}(\mathbf{R}^\star, \mathcal{U}_\mathrm{A}, \eta)$, then we have that*

$$\mathtt{top}(u) \in \begin{cases} \mathit{arg}\max_{i \in [n]} r^\star_{u,i} & u \in \mathcal{U}_\mathrm{MAJ} \cup \mathcal{U}_{i^\star} \\ [\bar{n}+1] & u \in \mathcal{U}_\mathrm{MIN} \setminus \mathcal{U}_{i^\star} \end{cases}, \quad \mathrm{SW}(\widetilde{\mathbf{R}}, \alpha) = \textstyle\sum_{u \in (\mathcal{U}_\mathrm{MAJ} \cup \mathcal{U}_{i^\star})} \max_{i \in [n]} r^\star_{u,i}$$

This is a *strict* improvement (Corollary D.1)! This yields *sufficient conditions* on η (given $\mathcal{U}_\mathrm{A}$) for strict SW increase (Corollary D.2), which enables the construction of an effective η finder.

3.3 Algorithms to find effective altruism (EA)

Algorithm 1 returns an effective η using only arithmetic operations/comparisons, which suggests, given sufficient info-sharing, it is computationally reasonable that users find effective strategies.

Algorithm 1 Find an effective η

Require: $\sigma_{k_\mathrm{MAJ}}(\mathbf{R}^\star_\mathrm{MAJ})$, α, $\bar{n}$, $\mathrm{ASV}_{i^\star}$, $\mathrm{AV}_{i^A_{\bar{n}}}$, κ, $|\mathcal{U}_\mathrm{A}|$

$N_\mathrm{up} \leftarrow \min\left\{(\sigma_{k_\mathrm{MAJ}}(\mathbf{R}^\star_\mathrm{MAJ})^2 - \alpha^2)/(\sqrt{\bar{n}}\mathrm{AV}_{i^A_{\bar{n}}}), \kappa\right\}$ `// upper bound on feasible` η

$d \leftarrow \bar{n}\mathrm{AV}_{i^A_{\bar{n}}}{}^2 + 4|\mathcal{U}_\mathrm{A}|(\alpha^2 - \mathrm{ASV}_{i^\star})$ `// find discriminant`

if $d < 0$ **then**

  $N_\mathrm{lo} \leftarrow N_\mathrm{up}/2$ `// no real` η `lower bound exists`

else if $d \geq 0$ **then**

  $N_\mathrm{lo} \leftarrow (\sqrt{\bar{n}}\mathrm{AV}_{i^A_{\bar{n}}} + \sqrt{d})/(2|\mathcal{U}_\mathrm{A}|)$ `// lower bound on feasible` η

if $N_\mathrm{lo} < N_\mathrm{up}$ **then return** $(N_\mathrm{lo} + N_\mathrm{up})/2$ `// return if exists feasible` η

$N_\mathrm{up} \leftarrow \min\left\{(\sqrt{\bar{n}}\mathrm{AV}_{i^A_{\bar{n}}} - \sqrt{d})/(2|\mathcal{U}_\mathrm{A}|), N_\mathrm{up}\right\}$ `// new upper bound on feasible` η

if $N_\mathrm{up} > 0$ **then return** $N_\mathrm{up}/2$ `// return an` η `if upper bound` > 0

return 0 `// sufficient conditions can't be satisfied`

Theorem 3.3 **(Algorithm 1 returns an effective η)** *Let $\mathbf{R}^\star$ be a majority-minority matrix satisfying Assumption 2.1 with a picky item at index $i^\star > \bar{n}$ and $\alpha \in \mathcal{G}(\mathbf{R}^\star)$, then Algorithm 1, using $(\sigma_{k_\mathrm{MAJ}}(\mathbf{R}^\star_\mathrm{MAJ})$, α, $\bar{n}$, $\mathrm{ASV}_{i^\star}$, $\mathrm{AV}_{i^A_{\bar{n}}}$, κ, $|\mathcal{U}_\mathrm{A}|)$ as parameters, returns either:*

- *$\eta \in \mathbb{R}_{>0}$ such that social welfare is improved if all $u \in \mathcal{U}_A$ uprate $i^\star$ by η.*
- *0 if and only if there is no η correlated strategy that satisfies our feasible conditions.*

It may not be reasonable to assume that the shared information is perfect. In Appendix D.3, we prove that even with incorrect information, the $\hat{\eta}$ produced by Algorithm 1 may satisfy true conditions.

3.4 Learner's welfare

Minority group's welfare improves while not decreasing the welfare of any other user. Thus, given minimally altruistic agents and the right conditions, algorithmic collective action is *incentivized*. Is this good design? We analyze the learner's welfare under two utility functions similar to prior work. Surprisingly, altruistic strategies improve his welfare! Proofs are in Appendix D.4.

Benevolent learner. A benevolent learner is one whose welfare is *user* social welfare, $U_{\text{BEN}} := \text{SW}(\mathbf{R}^\star, \alpha)$. We directly get Corollary D.4: a strict increase in benevolent learner utility!

Engagement-based learner. To RecSys, users' ratings represent watchtimes, views, likes, comments, etc., on which he can sell ad space. Thus, he gains utility directly from ratings (even those which are "fake"). Suppose $U_{\text{EN}} := \sum_{i \in [n]} \sum_{u \in [m]} |\tilde{r}_{u,i}|$. EA strategies are uprating schemes: a set of users submit η instead of 0. Thus, because engagement-based utility is simply the sum of ratings, clearly an EA strategy yields higher utility than truthfulness. The formal proposition is left to Appendix D.4.2

4 Empirical results

We present two empirical contributions. The first is a simulation of altruistic users on a real dataset. The second is a survey of 100 users asked about their interactions with recommendation algorithms.

4.1 Experiment

We run a simulation similar to movements in online book communities, where due to inequities [53], users intentionally interact with marginalized authors' content to correct for algorithms' lack of promotion [52]. We construct a matrix of Goodreads users' interactions with different genres. The number of books reviewed of a genre is that user's rating of it. Whether a genre is a majority / minority item comes from the reviews it has. User groups are defined by a user's most reviewed genre. We simulate a subset of romance (the most reviewed genre) readers reviewing additional books of a less-popular genre. Even though assumptions of the main body are not satisfied, for certain α's, there is social welfare improvement! Methodology and additional results are in Appendix E.1.

4.1.1 Results

We present social welfare change when $1/3$ or $1/2$ of romance users uprate. For each minority genre uprated, the percent increase in social welfare of minority groups is large, as much as a 15 times (Figure 3)! In all cases, the total welfare improvement is between 8% and 10% (Table 5). Like in the theoretical results, α must be in a particular range for each result. These ranges are presented in Table 4 in Appendix E.2).

4.2 Survey

Both literature and real-world evidence prove *in certain contexts* users engage with content to affect others' recommendations in altruistic ways. Is it realistic to expect that correlated rating strategies are actually *widespread* in a recommendation system? We present examples of strategies and preliminary results on the scale of altruism through a survey of random users. All details are in Appendix E.2.

4.2.1 Results

Of 100 responders, all used algorithm-based platforms. The majority of participants (92) believed their interactions affect their recommended feed; a smaller majority (57) believed their interactions affected others (Table 6). A surprisingly high amount of users indicate correlated rating strategies (Table 7). 32 users had intentionally interacted or avoided to influence other feeds. Those who have intentionally interacted to affect others were fairly consistent in reasoning; 16/20 discussed promoting content from specific sources they liked or morally supported and 6/20 mentioned some form of charity. In Appendix E.2.2 we provide textual examples and reasoning provided by users.

5 Conclusion

We model a RecSys in which users' preferences influence each other's recommendations. Constructing a class of simple correlated rating strategies, we find that users are able to strictly improve social welfare beyond truthful preference reporting. These strategies represent altruistic manipulations: users in the majority are able to improve the minority group's recommendations. We provide a robust algorithm to find an effective strategy and prove that the learner' utility is also improved under altruism. We supplement our theoretical results with empirics: (1) a simulation of altruism on the Goodreads dataset and (2) an online survey of real users. We are the first to lay the groundwork for the theoretical analysis of recommendations as multi-agent models in which users are *not exclusively self-interested*. There are several interesting lines of future work, and we elaborate on them in Appendix F.

References

[1] Dheeraj Baby and Soumyabrata Pal. Online Matrix Completion: A Collaborative Approach with Hott Items, August 2024. URL http://arxiv.org/abs/2408.05843. arXiv:2408.05843 [cs].

[2] Ian Ball. Scoring Strategic Agents, May 2024. URL http://arxiv.org/abs/1909.01888. arXiv:1909.01888 [econ].

[3] Joachim Baumann and Celestine Mendler-Dünner. Algorithmic collective action in recommender systems: promoting songs by reordering playlists. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

[4] Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information Discrepancy in Strategic Learning. In *Proceedings of the 39th International Conference on Machine Learning*, June 2022. URL https://proceedings.mlr.press/v162/bechavod22a.html.

[5] Gary S. Becker. A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093, 1974. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/1830662.

[6] Omri Ben-Dov, Jake Fawkes, Samira Samadi, and Amartya Sanyal. The role of learning algorithms in collective action. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 3443–3461. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ben-dov24a.html.

[7] Omer Ben-Porat and Moshe Tennenholtz. A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a9a1d5317a33ae8cef33961c34144f84-Paper.pdf.

[8] Omer Ben-Porat, Itay Rosenberg, and Moshe Tennenholtz. Content Provider Dynamics and Coordination in Recommendation Ecosystems. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18931–18941. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/dabd8d2ce74e782c65a973ef76fd540b-Paper.pdf.

[9] Omer Ben-Porat, Lee Cohen, Liu Leqi, Zachary C. Lipton, and Yishay Mansour. Modeling Attrition in Recommender Systems with Departing Bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6072–6079, June 2022. doi: 10.1609/aaai.v36i6.20554. URL https://ojs.aaai.org/index.php/AAAI/article/view/20554.

[10] Sam Biddle, Paulo Victor Ribeiro, and Tatiana Dias. TikTok Told Moderators: Suppress Posts by the Ugly and Poor. *The Intercept*, March 2020. URL https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/.

[11] Emmanuel J. Candes and Yaniv Plan. Matrix Completion With Noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010. ISSN 1558-2256. doi: 10.1109/JPROC.2009.2035722. URL https://ieeexplore.ieee.org/abstract/document/5454406.

[12] Emmanuel J. Candes and Benjamin Recht. Exact Matrix Completion via Convex Optimization, May 2008. URL http://arxiv.org/abs/0805.4471. arXiv:0805.4471 [cs].

[13] Emmanuel J. Candes and Terence Tao. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010. ISSN 1557-9654. doi: 10.1109/TIT.2010.2044061. URL https://ieeexplore.ieee.org/document/5452187/.

[14] Sarah H. Cen, Andrew Ilyas, Jennifer Allen, Hannah Li, and Aleksander Madry. Measuring strategization in recommendation: Users adapt their behavior to shape future content, 2024. URL https://arxiv.org/abs/2405.05596.

[15] Sarah H. Cen, Andrew Ilyas, and Aleksander Madry. User Strategization and Trustworthy Algorithms. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 202, New York, NY, USA, December 2024. Association for Computing Machinery. ISBN 979-8-4007-0704-9. doi: 10.1145/3670865.3673545. URL https://doi.org/10.1145/3670865.3673545.

[16] Yiling Chen, Yang Liu, and Chara Podimata. Learning Strategy-Aware Linear Classifiers. In *Advances in Neural Information Processing Systems*, volume 33, pages 15265–15276. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/ae87a54e183c075c494c4d397d126a66-Abstract.html.

[17] Stephane Chretien and Sebastien Darses. Perturbation bounds on the extremal singular values of a matrix after appending a column, December 2014. URL http://arxiv.org/abs/1406.5441. arXiv:1406.5441 [math].

[18] CNN. TikTokers stand in solidarity with black creators to protest censorship | CNN, 2020. URL https://www.cnn.com/2020/05/19/us/tiktok-black-lives-matter-trnd/index.html.

[19] Lee Cohen, Saeed Sharifi-Malvajerdi, Kevin Stangl, Ali Vakilian, and Juba Ziani. Bayesian Strategic Classification. *arXiv.org*, February 2024. URL https://arxiv.org/abs/2402.08758v1.

[20] Sarah Dean, Sarah Rich, and Benjamin Recht. Recommendations and user agency: the reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 436–445, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372866. URL https://doi.org/10.1145/3351095.3372866.

[21] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173694. URL https://doi.org/10.1145/3173574.3173694.

[22] Michael Ann DeVito. How transfeminine tiktok creators navigate the algorithmic trap of visibility via folk theorization. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), November 2022. doi: 10.1145/3555105. URL https://doi.org/10.1145/3555105.

[23] Michael Ann DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173694. URL https://doi.org/10.1145/3173574.3173694.

[24] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "i always assumed that i wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 153–162, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702556. URL https://doi.org/10.1145/2702123.2702556.

[25] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Group-Fair Classification with Strategic Agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 389–399, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 979-8-4007-0192-4. doi: 10.1145/3593013.3594006. URL https://dl.acm.org/doi/10.1145/3593013.3594006.

[26] Bailey Flanigan, Ariel D Procaccia, and Sven Wang. Distortion under public-spirited voting. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, page 700, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597722. URL https://doi.org/10.1145/3580507.3597722.

[27] Susan A. Gelman and Cristine H. Legare. Concepts and folk theories. *Annual review of anthropology*, 40:379–398, October 2011. ISSN 0084-6570. doi: 10.1146/annurev-anthro-081309-145822. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3579644/.

[28] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic Classification in the Dark. In *Proceedings of the 38th International Conference on Machine Learning*, July 2021. URL https://proceedings.mlr.press/v139/ghalme21a.html.

[29] Wenshuo Guo, Karl Krauth, Michael Jordan, and Nikhil Garg. The Stereotyping Problem in Collaboratively Filtered Recommender Systems. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pages 1–10, New York, NY, USA, November 2021. Association for Computing Machinery. ISBN 978-1-4503-8553-4. doi: 10.1145/3465416.3483298. URL https://dl.acm.org/doi/10.1145/3465416.3483298.

[30] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 111–122, New York, NY, USA, January 2016. Association for Computing Machinery. ISBN 978-1-4503-4057-1. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

[31] Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[32] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. "honestly, i think tiktok has a vendetta against black creators": Understanding black content creator experiences on tiktok. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), October 2023. doi: 10.1145/3610169. URL https://doi.org/10.1145/3610169.

[33] Keegan Harris, Chara Podimata, and Zhiwei Steven Wu. Strategic Apple Tasting, October 2023. URL http://arxiv.org/abs/2306.06250. arXiv:2306.06250 [cs].

[34] Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. Recommending to Strategic Users, February 2023. URL http://arxiv.org/abs/2302.06559. arXiv:2302.06559 [cs].

[35] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004. ISSN 1046-8188. doi: 10.1145/963770.963772. URL https://doi.org/10.1145/963770.963772.

[36] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[37] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The Disparate Effects of Strategic Manipulation. *arXiv.org*, August 2018. doi: 10.1145/3287560.3287597. URL https://arxiv.org/abs/1808.08646v4.

[38] Xinyan Hu, Meena Jagadeesan, Michael I. Jordan, and Jacob Steinhardt. Incentivizing high-quality content in online recommender systems, 2024. URL https://arxiv.org/abs/2306.07479.

[39] Benjamin Hébert and Weijie Zhong. Engagement maximization, 2025. URL https://arxiv.org/abs/2207.00685.

[40] Nicole Immorlica, Meena Jagadeesan, and Brendan Lucier. Clickbait vs. quality: How engagement-based optimization shapes the content landscape in online platforms. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 36–45, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645353. URL https://doi.org/10.1145/3589334.3645353.

[41] Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[42] Prateek Jain and Soumyabrata Pal. Online Low Rank Matrix Completion, March 2023. URL http://arxiv.org/abs/2209.03997. arXiv:2209.03997 [cs].

[43] I.T. Joliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 978-0-387-95442-4. doi: 10.1007/b98835. URL http://link.springer.com/10.1007/b98835.

[44] Aditya Karan, Nicholas Vincent, Karrie Karahalios, and Hari Sundaram. Algorithmic collective action with two collectives. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1468–1483, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732098. URL https://doi.org/10.1145/3715275.3732098.

[45] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021. doi: 10.1145/3476046. URL https://doi.org/10.1145/3476046.

[46] Gur Keinan and Omer Ben-Porat. Modeling churn in recommender systems with aggregated preferences, 2025. URL https://arxiv.org/abs/2502.18483.

[47] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion from Noisy Entries. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/hash/aa942ab2bfa6ebda4840e7360ce6e7ef-Abstract.html.

[48] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, August 2009. ISSN 1558-0814. doi: 10.1109/MC.2009.263. URL https://ieeexplore.ieee.org/document/5197422/.

[49] Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in Collaborative Filtering. In *Recommender Systems Handbook*. Springer, Boston, MA, November 2021. doi: 10.1007/978-1-0716-2197-4_3. URL https://link.springer.com/chapter/10.1007/978-1-0716-2197-4_3.

[50] Sagi Levanon and Nir Rosenfeld. Strategic Classification Made Practical. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6243–6253. PMLR, July 2021. URL https://proceedings.mlr.press/v139/levanon21a.html. ISSN: 2640-3498.

[51] David Liu, Jackie Baek, and Tina Eliassi-Rad. When collaborative filtering is not collaborative: Unfairness of pca for recommendations, 2023. URL https://arxiv.org/abs/2310.09687.

[52] Tyler McCall. BookTok's Racial Bias. *The Cut*, November 2022. URL https://www.thecut.com/2022/11/booktok-racial-bias-tiktok-algorithm.html.

[53] Alysia De Melo. The influence of booktok on literary criticisms and diversity. *Social Media + Society*, 10(4):20563051241286700, 2024. doi: 10.1177/20563051241286700. URL https://doi.org/10.1177/20563051241286700.

[54] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 230–239, 2019.

[55] Ashlee Milton, Leah Ajmani, Michael Ann DeVito, and Stevie Chancellor. "i see me here": Mental health content, community, and algorithmic curation on tiktok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581489. URL https://doi.org/10.1145/3544548.3581489.

[56] Stephen J. Payne. CHAPTER 6 - Users' Mental Models: The Very Ideas. In John M. Carroll, editor, *HCI Models, Theories, and Frameworks*, Interactive Technologies, pages 135–156. Morgan Kaufmann, San Francisco, 2003. ISBN 978-1-55860-808-5. doi: https://doi.org/10.1016/B978-155860808-5/50006-X. URL https://www.sciencedirect.com/science/article/pii/B978155860808550006X.

[57] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[58] Kenny Peng, Manish Raghavan, Emma Pierson, Jon Kleinberg, and Nikhil Garg. Reconciling the accuracy-diversity trade-off in recommendations. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1318–1329, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645625. URL https://doi.org/10.1145/3589334.3645625.

[59] Ariel D. Procaccia and Jeffrey S. Rosenschein. The Distortion of Cardinal Preferences in Voting. In Matthias Klusch, Michael Rovatsos, and Terry R. Payne, editors, *Cooperative Information Agents X*, pages 317–331, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-38570-7.

[60] Benjamin Recht. A Simpler Approach to Matrix Completion. *J. Mach. Learn. Res.*, 12(null): 3413–3430, December 2011. ISSN 1532-4435.

[61] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, January 2010. ISSN 0036-1445, 1095-7200. doi: 10.1137/070697835. URL http://epubs.siam.org/doi/10.1137/070697835.

[62] Princess Sampson, Ro Encarnacion, and Danaë Metaxa. Representation, self-determination, and refusal: Queer people's experiences with targeted advertising. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1711–1722, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594110. URL https://doi.org/10.1145/3593013.3594110.

[63] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative Filtering Recommender Systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, pages 291–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-72079-9. doi: 10.1007/978-3-540-72079-9_9. URL https://doi.org/10.1007/978-3-540-72079-9_9.

[64] Dorothee Sigg, Moritz Hardt, and Celestine Mendler-Dünner. Decline now: A combinatorial model for algorithmic collective action. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713966. URL https://doi.org/10.1145/3706598.3713966.

[65] Ellen Simpson and Bryan Semaan. For you, or for"you"? everyday lgbtq+ encounters with tiktok. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3), 01 2021. doi: 10.1145/3432951. URL https://doi.org/10.1145/3432951.

[66] Rushabh Solanki, Meghana Bhange, Ulrich Aïvodji, and Elliot Creager. Crowding out the noise: Algorithmic collective action under differential privacy, 2025. URL https://arxiv.org/abs/2505.05707.

[67] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2, January 2009. ISSN 1687-7470. doi: 10.1155/2009/421425. URL https://dl.acm.org/doi/10.1155/2009/421425.

[68] Mengting Wan and Julian J. McAuley. Item recommendation on monotonic behavior chains. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada,*

October 2-7, 2018, pages 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

[69] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1248. URL https://doi.org/10.18653/v1/p19-1248.

[70] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. Agent, gatekeeper, drug dealer: How content creators craft algorithmic personas. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359321. URL https://doi.org/10.1145/3359321.

[71] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians, 2014. URL https://arxiv.org/abs/1405.0680.

Supplementary Material

A Table of notation

To assist the reader, we include a table summarizing our paper notation below.

Table 1: Paper notation in the main body

Notation	Explanation
$\mathbf{R}$	generic preference matrix (elements $r_{u,i}$)
$\mathbf{R}^\star$	ground truth preference matrix
$\widetilde{\mathbf{R}}$	revealed preference matrix
$\widehat{\mathbf{R}}$	low-rank approximation of preference matrix
$\mathbf{R}_{\text{MAJ}}, \mathbf{R}_{\text{MIN}}$	submatrices for majority and minority groups
α	learner's loss tolerance parameter
$k^\star$	learner's chosen rank
k_{MAJ}	rank of the majority submatrix
$\text{TVR}(k^\star, \mathbf{R})$	total variation retained after $k^\star$-truncated svd on $\mathbf{R}$
$\texttt{top}(u)$	top item recommended to user u
$\mathcal{I}_{\text{top}}(u)$	set of user u's top items under $\widehat{\mathbf{R}}$
$\mathcal{I}_{\text{top}}^{\text{pop}}(u)$	most popular top item(s) for user u
$\mathcal{U}$	set of users
$\mathcal{U}_{\text{MAJ}}, \mathcal{U}_{\text{MIN}}$	majority and minority user sets
$\mathcal{I}_{\text{MAJ}}, \mathcal{I}_{\text{MIN}}$	majority and minority item sets
$\mathcal{G}(\mathbf{R})$	singular value gap range for matrix $\mathbf{R}$
$i^\star$	picky item index preferred by minority users
$\mathcal{U}_{i^\star}$	group of users who like only item $i^\star$
$\mathcal{U}_{\text{A}}$	subset of majority users who are altruists
η	positive uprating amount used by altruistic users
$\hat{\eta}$	correlated strategy returned by Algorithm 1
$\text{ASV}_{i^\star}$	squared norm of column $i^\star$ in $\mathbf{R}^\star$
$\text{AV}_{i^{\text{A}}_{\tilde{n}}}$	max total (true) preference for a popular item among altruists
$\mathcal{G}(\mathbf{R}, \mathcal{U}_{\text{A}}, \eta)$	uprating-aware singular value gap
κ	smallest top rating among majority users
$\text{SW}(\mathbf{R}, \alpha)$	social welfare under matrix $\mathbf{R}$ and parameter α
$\hat{\mathbf{z}}, \mathbf{z}^\star$	estimated and true parameter vectors for Algorithm 1

B Supplementary material for Section 1

B.1 Extended related works

Our work is related to four major streams of literature: Human Computer Interaction, RecSys modeling, strategic classification, and collaborative filtering & matrix completion. We provide extensive details on each of these below.

Human computer interaction and algorithmic collective action. There is a breadth of human computer interaction literature that serves as the inspiration for our theoretical modeling of user strategic behavior. It has long been clear that even "normal" users are aware of social platforms' recommendation algorithms [24]. HCI researchers study the complex mental models people develop to understand algorithms as folk theories [56, 27]. How a user forms their algorithmic folk theories is heavily impacted by the way in which they present themselves to the platform (posts, statuses, etc) and interact with content (likes, comments, etc.) [23]. Thus, there are many works that conduct interview studies with users of different intersections about their experiences/theories related to curation/recommendation algorithms. These intersections include YouTube creators [70], LGBTQ+ TikTok users [65], transfeminine TikTok creators [22], Queer people targeted by ads [62], participants of online mental health communities[55], and black content creators [32]. Across the intersections of these interview studies and in a recent large scale experiment [14], results indicate that people use their folk theories to strategically interact (e.g. liking, commenting, watching, etc) with platform content in order to tailor their personal recommendation feeds when possible. However, a user's folk theories (including those documented in the interview studies above) are not limited to how the algorithm impacts them personally. Many users of the online book discussion community known as "BookTok", for example, theorize that the recommendation algorithm generally suppresses content from marginalized creators [53]. Karizat et al. [45] document the relationship between such folk theories and altruistic actions users take in order to ameliorate bad/harmful recommendation for other groups. While there are not yet large scale experiments on altruistically strategic interactions, grassroots movements among communities engaging in this to support BLM [18] and marginalized authors [52] have been reported. Our theoretical model serves as a proof-of-concept mathematically for altruistic interaction and our survey indicates that altruistic interaction may be fairly commonplace. This is relevant to a theoretical line of work called algorithmic collective action (ACA) first presented by Hardt et al. [31] who study a setting in which agents participate in *coordinated* strategies to manipulate a learning algorithm. Further works have considered modifications primarily to the agent-side such as: multiple and varied collectives [44] and combinatorial actions [64]. And others complicate the learner's problem such as: ACA under a differentially-private model [66] or distributionally robust optimizers [6]. Our model focuses on ACA in recommendation systems, which are not covered under the existing theoretical works. Empirically, Baumann and Mendler-Dünner [3] examine a very similar ACA phenomenon by simulating collective playlist reordering to improve song promotion on platforms like Spotify. Considering playlist order as a form of rating, our work can be interpreted as a theoretical foundation to their results.

RecSys modeling. There is a large body of work on the modeling of recommender systems through a game-theoretic or economic lens. Two are most relevant to this work as they model users themselves behaving strategically. First, Haupt et al. [34] model strategic users and a recommender system as a two-phase Stackelberg game in which the recommender commits to a policy that maps a user's interactions to a content recommendation and then users strategically interact with content during a learning phase. In our model, we abstract away from the recommender learning preferences over rounds and allow the interactions of one user to impact the recommendations of another. Second, Cen et al. [15] also model the interaction as a repeated two-player (Stackelberg) game. At each round, the learner provides recommendations to the user based on an estimate of user strategies, and the user responds with a behavior generated by an [evolving] user strategy. Authors conclude that trustworthy algorithms are those that do not incentivize a user to be strategic and guarantee κ payoff to the naive agent. Our model differs from this one because we consider a multi-agent game, though their conception of trustworthiness could likely be made applicable to our model if we consider payoff guarantees across all users.

Other relevant game-theoretic models consider some sort of user or item social welfare (though not strategic users). Dean et al. [20] find that under certain conditions in top-N collaborative filtering methodologies, some items, despite being in the system, would never be recommended to any type of user. Guo et al. [29] expand upon this problem and find that it can be ameliorated when users are represented by multiple feature vectors. We take a more user-centered approach in our analysis of social welfare. Peng et al. [58] analyze the recommender's accuracy/diversity trade-off when providing content and conclude that users' consumption cost constraints should imply more diversity is utility-maximizing for users. Because our results focus on top-1 recommendation according to a fixed preference matrix, this is not an issue for the learner we study. Some game-theoretic recommender system models focus primarily on welfare or mechanisms for the learner. Hébert and Zhong [39] consider the recommender's problem as an optimal information design over a sequence of rounds to keep a user engaged with the platform for as long as possible, Ben-Porat et al. [9] consider the learner as a multi-armed bandit wanting to avoid user attrition, and Keinan and Ben-Porat [46] create a model that maximizes engagement to avoid user churn. While not the primary focus of our work, we consider a version of engagement maximization for the learner's utility function in section 3.4. Additionally, there are a handful of modeling works that consider content creators as strategic. Some of these directions include characterizing/learning equilibria of the content creation marketplace [41, 8], modeling the incentives creators have to invest in creating quality or clickbait content [40], designing learning algorithms for the recommender that will incentivize quality content generation [38], designing fair and stable recommenders under content creator strategization[7], and . While content creators are a type of user, creation of new content is not the type of rating interaction we consider with our preference matrices (see section 5 for discussion of ratings).

Finally, while not explicitly a model of content recommendation, there is also a connection between our modeling of altruistic agents and ideas of public spirit. Flanigan et al. [26] model a voting setting in which a public-spirited agent, who may otherwise suffer distortion in social welfare due to the nature of voting with cardinal preferences [59], accounts for other agents' utility functions in the submission of her cardinal preferences to the voting mechanism. Our conception of agents intentionally manipulating ratings to help others is similar, but because in our model every agent receives her own recommendation (rather than the election in which one candidate is chosen) and the strategy space is continuous the public-spirited voting regime is significantly different.

Strategic classification. Given that we focus on strategic individuals in recommender systems, there is a breadth of relevant work on strategic agents in classification problems. Presented first by Hardt et al. [30], in this setting there is a learner who publishes a classifier to agents whose utility depends on their classification. Top-1 recommendation resembles strategic multi-class classification where the learner takes as input user features and outputs (a type of) content each user would like. There exists a large thread of learner-centric strategic classification literature, such as algorithms that are in some way robust to agent strategization such as incentive-awareness [30, 50, 16], truthfulness [33] or strategyproofness [2]. Although we address learner welfare in Section 3.4, more relevant to our work is strategic classification literature focusing on user-centric perspectives. Particularly relevant are works on fairness across user groups. Hu et al. [37] address unequal outcomes in terms of misclassification and Milli et al. [54] study unequal outcomes in terms of effort to manipulate one's features. Both papers, including [25] examine interventions that the learner could take and how that impacts outcomes for different user groups. By contrast, we present a *user* intervention, show that altruistic users are incentivized to act in this manner, and that no user is made worse-off. We consider the setting where altruistic agents strategize collectively via a coordinating mechanism which may not know everything about the learner and participating agents. Hence a related thread of literature analyzes when agents have incomplete information about the classifier, e.g., because it is purposefully withheld or too complicated for everyday people to understand. Bechavod et al. [4] study the setting where information is inferred and shared within sub-populations, others assume agents have knowledge of a distributional prior [19], and look at when transparent or opaque policies give rise to more accurate classifications [28]. In contrast to these papers, which study incomplete knowledge of the agents on the classifier, we consider incomplete information of a coordinating mechanism regarding both the participating strategic agents and the learner.

Collaborative filtering and matrix completion. In this methodology, since users rarely rate or view all available items estimates of user's preferences for unseen items depend on their past ratings and the ratings of other users [63, 49, 67]. Formally, this can be viewed as a matrix completion problem where the matrix represents user-item ratings (e.g., stars or likes). Without any knowledge of the matrix properties, this problem is impossible: the remaining entries could be anything. Consequently, ratings matrices are commonly assumed to be low-rank. A series of papers provide bounds on the number of random matrix entries required to perfectly recovery low-rank matrices that also satisfy some coherence conditions (which loosely tell us how informative ratings are about one another) [13, 12, 60]. Other papers study extensions of this problem: when there is noise [11, 47] and when recovery is online and sampling is active [1, 42]. Our learner protocol is a tractable abstraction of matrix completion-based recommendation by directly considering the problem of low-rank approximations of completed preferences matrices.

Matrix completion in recommendation systems is particularly relevant to our use of rank reduction for the learner. Matrix factorization [48] approximates ratings as dot products of low-dimensional user and item embeddings. When the objective is to minimize squared estimation error, it is equivalent to PCA. Nuclear norm minimization [12], is a convex relaxation of the rank function and in some cases provably recovers the rank-minimizing solution subject to agreement with known entries [61].

Many of the common techniques such as matrix factorization [48] and nuclear norm minimization [12] rely on the assumption that the matrix is low-rank. implicitly assuming that preference data may be represented in a low-dimensional latent space. If he does not receive ratings certain key user/item pairs (because of extremely disparate user preferences or bad luck, for example) that prove a user/item (i.e. row or column) is linearly independent from others, he will likely assume users/items are more similar than they really are. We point the reader to appendix section C.1 for some technical results. Studying this problem through an explicitly matrix factorization lens, however, is difficult in generality. Thus our model represents this by first having a fully completed matrix and then requiring some level of rank reduction.

C Supplementary Material for Section 2

C.1 Theoretical connections to matrix completion

Realistically, the recommender learns (potentially non-deterministic) user preferences over rounds in which users are served items and provide ratings (likes, comments, watchtimes, etc). Over these rounds, he may fill in the preference matrix with estimates how much each user likes the shown item. Because there are often a large number of missing entries in the full ratings matrix, assuming a simpler structure for the ratings matrix is necessary to create estimates of the remaining entries. A common method in collaborative filtering is matrix completion. While there are many methods and optimization problems used in practice such as matrix factorization [48], here we discuss the connection of low-rank approximation to a specific method that has been extensively in the literature: nuclear norm minimization subject to agreement with observed entries [12, 60, 13].

In particular, we assume that $\mathbf{R}$ is some unknown low-rank matrix which we want to recover. Define Ω to be the set of (u, i) pairs that have been observed so that $(u, i) \in \Omega$ if user u has seen item i. Formally,

$$\Omega := \{(u, i) \in [m] \times [n] : r_{u,i} \text{ has been observed}\}.$$

If we know that $\mathbf{R}$ is the only rank-d matrix that agrees with observed entries then the following rank-minimization problem will return $\mathbf{R}$:

$$\begin{aligned} \text{minimize}_{\mathbf{X} \in \mathbb{R}^{\mathbf{m} \times \mathbf{n}}} \quad & \text{rank}(X) \\ \text{subject to} \quad & x_{u,i} = r_{u,i} \quad \forall (u, i) \in \Omega. \end{aligned} \tag{1}$$

However, this problem is NP-hard. Consequently, the aforementioned papers consider the nuclear norm, $\|\cdot\|_*$ which is the sum of the singular values and solve the following problem instead:

$$\begin{aligned} \text{minimize}_{\mathbf{X} \in \mathbb{R}^{\mathbf{m} \times \mathbf{n}}} \quad & \|X\|_* \\ \text{subject to} \quad & x_{u,i} = r_{u,i} \quad \forall (u, i) \in \Omega. \end{aligned} \tag{2}$$

The nuclear norm is a proxy for rank in the same way that the L_1 norm is a proxy for the L_0 norm of a vector and can be minimized via semi-definite programming. And, when singular values are at most 1, it is the best convex lower approximation of the rank function [61].

In our paper, the α loss tolerant learner picks a rank such that the sum of the remaining singular values is small (each is at most α). The nuclear norm minimization problem makes this sum as small as possible. Therefore, it tends to "smooth" out potential discrepancies in observed (or unobserved) entries. To give an example of this, we will now discuss a similar result to Theorem 3.1 on the recovery of minority preferences.

The following assumption specifies a specific type of user preferences.

Assumption C.1 (Distinct preferences) *Users are in one of two disjoint groups, G_A and G_B, where the ratings of user u satisfy the following:*

$$\mathbf{R}_{u\cdot} = \begin{cases} \mathbf{a} & u \in G_A \\ \mathbf{b} & u \in G_B \end{cases},$$

where $\mathbf{a}$ and $\mathbf{b}$ are linearly independent and for $i_B^* = \text{argmax}_{i \in [n]} b_i$, $a_{i_B^*} = 0$.

Theorem C.2 (Estimated Minority Item Ratings are Zero) *Assume that $\mathbf{R}$ satisfies Assumption C.1, and for all $u \in G_B$, $(u, i_B^*) \notin \Omega$. Then the solution $\hat{\mathbf{R}}$ to Problem 1 will satisfy $\hat{\mathbf{R}}_{ui_B^*} = 0$ for all $u \in [m]$.*

The proof is very simple and relies on the following lemma:

Lemma C.1 (Lemma 2.3. of [61]) *Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of the same dimensions. If $\mathbf{A}\mathbf{B}^\top = 0$ and $\mathbf{A}^\top \mathbf{B} = 0$ then $\|\mathbf{A} + \mathbf{B}\|_* = \|\mathbf{A}\|_* + \|\mathbf{B}\|_*$.*

Proof. WLOG let $i_B^* = n$. Assume for the sake of contradiction that $\hat{\mathbf{R}}_{\cdot,n} \neq 0$. Then we can write $\hat{\mathbf{R}}$ as $\mathbf{X} + \mathbf{Y}$ where $\mathbf{X}$ is equal to $\hat{\mathbf{R}}$ on all columns but the last column where it is zero, and $\mathbf{Y}$ is zero except for the last column where it is equal to $\hat{\mathbf{R}}$.

By construction, $\mathbf{X}\mathbf{Y}^\top = 0$ and $\mathbf{X}^\top\mathbf{Y} = 0$. Therefore, by Lemma C.1

$$|\hat{\mathbf{R}}\|_* = \|\mathbf{X}\|_* + \|\mathbf{Y}\|_* > \|\mathbf{X}\|_*.$$

The constraint $x_{u,i} = r_{u,i}$ for all $(u,i) \in \Omega$ is satisfied for $\mathbf{X}$ since all users who saw item n (if any) gave it a rating of zero. Therefore, $\hat{\mathbf{R}}$ cannot be the optimal solution. □

C.2 Further discussion of “variation" retention

We can consider the following formal definition of “variation retention" that a particular rank reduction would have:

Definition C.1 *The total variation retained when doing a $k^\star$-truncated SVD to approximate $\widetilde{\mathbf{R}}$ is:*

$$\mathrm{TVR}(k^\star) := \frac{\sum_{j\in[k^\star]} \sigma_j(\widetilde{\mathbf{R}})}{\sum_{j\in[\mathrm{rank}(\mathbf{R})]} \sigma_j(\widetilde{\mathbf{R}})}$$

Our TVR is version of what is commonly referred to as an “explained variance ratio”[57] or Cummulative Percentage of Total Variation [43] for PCA these are the same function though in terms of the *eigenvalues* of $\widetilde{\mathbf{R}}$.

This means that an equivalent definition of the α-loss tolerant learner would be:

Definition C.2 (equivalent α-loss tolerant learner) *An α variance loss tolerant learner gets $\widehat{\mathbf{R}}$ where $k^\star$ is the minimum such that* $\mathrm{TVR}(k^\star+1) - \mathrm{TVR}(k^\star) \leq \alpha \cdot (\sum_{j\in[\mathrm{rank}(\widetilde{\mathbf{R}})]} \sigma_j(\widetilde{\mathbf{R}}))^{-1}$.

That is, the α-loss tolerant learner just has a budget of $\frac{\alpha}{\sum_{j\in[\mathrm{rank}(\widetilde{\mathbf{R}})]} \sigma_j(\widetilde{\mathbf{R}})}$ on the increase in TVR as k is increased. This type of learner wants the minimum rank possible such that increasing rank doesn’t improve the total variance retained very much.

C.3 Proof of remark 2.1

Proof of Remark 2.1. Let π_R be a permutation of users (rows) and π_C be a permutation of items (columns) and $\mathbf{P}_R$, $\mathbf{P}_C$ the corresponding permutation matrices. Then the permuted ratings matrix is given by

$$\mathbf{R}' = \mathbf{P}_R\mathbf{R}\mathbf{P}_C.$$

Claim 1: Let $\hat{\mathbf{R}}$ be the rank-k PCA of $\mathbf{R}$. Then $\hat{\mathbf{R}}' = \mathbf{P}_R\mathbf{R}\mathbf{P}_C$ is the rank-k PCA of $\mathbf{R}'$.

Recall that $\hat{\mathbf{R}}$ is a rank-k matrix minimizing the sum of squared errors. For any rank-k matrix $\mathbf{X}$:

$$\begin{aligned}
\|\mathbf{R} - \mathbf{X}\|_F^2 &= \sum_{u\in[m]}\sum_{i\in[n]} (r_{ui} - x_{ui})^2 \\
&= \sum_{u\in[m]}\sum_{i\in[n]} (r_{\pi_R(u),\pi_C(i)} - x_{\pi_R(u),\pi_C(i)})^2 \\
&= \|\mathbf{P}_R\mathbf{R}\mathbf{P}_C - \mathbf{P}_R\mathbf{X}\mathbf{P}_C\|_F^2 \\
&= \|\mathbf{R}' - \mathbf{P}_R\mathbf{X}\mathbf{P}_C\|_F^2
\end{aligned}$$

Thus:

$$\hat{\mathbf{R}} \in \operatorname*{arg\,min}_{\mathbf{X}:\mathrm{rank}(\mathbf{X})=k} \|\mathbf{R} - \mathbf{X}\|_F \iff \hat{\mathbf{R}}' \in \operatorname*{arg\,min}_{\mathbf{X}:\mathrm{rank}(\mathbf{X})=k} \|\mathbf{R}' - \mathbf{X}\|_F$$

Claim 2: Let μ and μ' be the recommendations based on $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}'$, respectively. Then for all $u \in [m]$:

$$r_{u,\mu_u} = r'_{\pi_R(u),\mu'_{\pi_R(u)}}.$$

By construction, $\hat{r}'_{\pi_R(u),\pi_C(i)} = \hat{r}_{u,i}$ for all $(u,i) \in [m] \times [n]$. Thus,

$$\begin{aligned} &\mu_u \in \arg\max_{i\in[n]} \hat{r}_{u,i} \\ \iff\ &\mu_u \in \arg\max_{i\in[n]} \hat{r}'_{\pi_R(u),\pi_C(i)} \\ \iff\ &\pi_C(\mu_u) \in \arg\max_{i\in[n]} \hat{r}'_{\pi_R(u),i} \end{aligned}$$

Further, $\|\mathbf{R}_{:,i}\|_1 = \|\mathbf{R}_{:,\pi_C(i)}\|_1$. Therefore, the recommendation will be the same regardless of ordering. $\square$

C.4 Example of a majority-minority matrix with a singular value gap

Example C.1 *A very simple majority-minority matrix is a binary matrix where every user likes just one item and there are 4 items: 2 popular items liked by m_{MAJ} users and 2 less-popular items liked by $m_{\text{MIN}} < m_{\text{MAJ}}$ users. Ordering users by which item they like, and listing the popular items first the we can write $\mathbf{R}$ as*

$$\mathbf{R} = \begin{pmatrix} \mathbf{1}_{m_{\text{MAJ}}} & 0 & 0 & 0 \\ 0 & \mathbf{1}_{m_{\text{MAJ}}} & 0 & 0 \\ 0 & 0 & \mathbf{1}_{m_{\text{MIN}}} & 0 \\ 0 & 0 & 0 & \mathbf{1}_{m_{\text{MIN}}} \end{pmatrix}$$

where $\mathbf{1}_m \in \mathbb{R}^m$ is a vectors of all 1s, one for each of the users who like that item. Likewise:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{\text{MAJ}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\text{MIN}} \end{pmatrix}$$

where $\mathbf{R}_{\text{MAJ}} \in \mathbb{R}^{2m_{\text{MAJ}}\times 2} = \begin{pmatrix} \mathbf{1}_{m_{\text{MAJ}}} & 0 \\ 0 & \mathbf{1}_{m_{\text{MAJ}}} \end{pmatrix}$ has the ratings of all users who like the popular items and $\mathbf{R}_{\text{MIN}} \in \mathbb{R}^{2m_{\text{MIN}}\times 2} = \begin{pmatrix} \mathbf{1}_{m_{\text{MIN}}} & 0 \\ 0 & \mathbf{1}_{m_{\text{MIN}}} \end{pmatrix}$ has the ratings of all users who like the less-popular items. The Singular Value Gap space is $\mathcal{G}(\mathbf{R}) = (\sqrt{m_{\text{MIN}}}, \sqrt{m_{\text{MAJ}}})$

C.5 Proof of Proposition 2.1

Proof. Since $\widetilde{\mathbf{R}}$ is a block matrix, we have that $\sigma_{k_{\text{MAJ}}+1}(\widetilde{\mathbf{R}}) = \sigma_1(\widetilde{\mathbf{R}}_{\text{MIN}}) \leq \alpha$, which implies that $k^\star \leq k_{\text{MAJ}}$. Likewise: $\sigma_{k_{\text{MAJ}}}(\widetilde{\mathbf{R}}) = \sigma_{k_{\text{MAJ}}}(\widetilde{\mathbf{R}}_{\text{MAJ}}) > \alpha$, which implies $k^\star \geq k_{\text{MAJ}}$. Thus $k^\star = k_{\text{MAJ}}$. $\square$

D Supplementary Material for Section 3

D.1 Supplementary material for truthful social welfare results

D.1.1 Proof of Theorem 3.1

Proof. By Proposition 2.1, the learner reduces $\mathbf{R}^\star$ to rank k_{MAJ}, meaning $\widehat{\mathbf{R}} = \sum_{i\in[k_{\text{MAJ}}]} \sigma_i \mathbf{u_i}\mathbf{v_i}^\top$, where this is a sum over the k_{MAJ} largest singular values. Let $U_{\text{MAJ}}\Sigma_{\text{MAJ}}V_{\text{MAJ}}^\top$, $U_{\text{MIN}}\Sigma_{\text{MIN}}V_{\text{MIN}}^\top$ be a SVD for $\mathbf{R}^\star_{\text{MAJ}}$ and $\mathbf{R}^\star_{\text{MIN}}$, respectively. Then the following is a SVD for $\mathbf{R}^\star$:

$$\begin{pmatrix} U_{\text{MAJ}} & \mathbf{0} \\ \mathbf{0} & U_{\text{MIN}} \end{pmatrix} \begin{pmatrix} \Sigma_{\text{MAJ}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\text{MIN}} \end{pmatrix} \begin{pmatrix} V_{\text{MAJ}}^\top & \mathbf{0} \\ \mathbf{0} & V_{\text{MIN}}^\top \end{pmatrix} = \begin{pmatrix} U_{\text{MAJ}}\Sigma_{\text{MAJ}}V_{\text{MAJ}}^\top & \mathbf{0} \\ \mathbf{0} & U_{\text{MIN}}\Sigma_{\text{MIN}}V_{\text{MIN}}^\top \end{pmatrix}.$$

Note that

$$\mathbf{\Sigma} := \begin{pmatrix} \Sigma_{\text{MAJ}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\text{MIN}} \end{pmatrix}$$

is not necessarily a perfectly ordered diagonal of singular values because even if Σ_{MAJ} and Σ_{MIN} are ordered, we do not assume full rank of $\mathbf{R}^\star_{\text{MAJ}}$, meaning that some diag elements of Σ_{MAJ} may be 0.

However by Assumption 2.1, the first k_{MAJ} singular values belong to Σ_{MAJ}. Also using the definition of k_{MAJ} as the rank of $\mathbf{R}^\star_{\text{MAJ}}$,

$$\widehat{\mathbf{R}} = \begin{pmatrix} U_{\text{MAJ}}\Sigma_{\text{MAJ}}V_{\text{MAJ}}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^\star_{\text{MAJ}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Firstly, we will show that for all $u \in \mathcal{U}_{\text{MAJ}}, r^\star_{ui^\star_u} = \max_{i\in[n]} r^\star_{u,i}$. Because all ratings for majority users are preserved, for all $u \in \mathcal{U}_{\text{MAJ}}$:

$$\arg\max_{i\in[n]} \hat{r}_{u,i} = \arg\max_{i\in[n]} r^\star_{u,i}.$$

Hence, $i^\star_u \in \arg\max_{i\in[n]} r^\star_{u,i}$ and $r^\star_{ui^\star_u} = \max_{i\in[n]} r^\star_{u,i}$.

Secondly, we will show that for all $u \in \mathcal{U}_{\text{MIN}}, r^\star_{u,i^\star_u} = 0$. For all minority users:

$$\arg\max_{i\in[n]} \hat{r}_{u,i} = [n]$$

since their ratings are represented by a vector of 0s. Therefore, they will be recommended an item from the arg max of

$$\text{maximize}_{i\in[n]} \|\widehat{\mathbf{R}}_i\|_1$$

But clearly (because $\widehat{\mathbf{R}}$ is simply the zero-padded majority matrix) this can be rewritten as

$$\text{maximize}_{i\in[\bar{n}]} \|\mathbf{R}^\star_{\text{MAJ}_i}\|_1$$

By assumption, $\sum_{u\in\mathcal{U}_{\text{MIN}}} r^\star_{u,i} = 0$ for all $i \in [\bar{n}]$. Thus for all $u \in \mathcal{U}_{\text{MIN}}$, $i^\star_u \in [\bar{n}]$ and $r^\star_{u,i^\star_u} = 0$. $\square$

D.2 Supplementary material for EA social welfare results

D.2.1 Proof of Theorem 3.2

We will first refer to known lower-bounds on matrix singular values when appending a column:

Lemma D.1 (Corollary 3.5 of Chretien and Darses [17]) *Let d be a positive integer and let $\mathbf{M} \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix with rank $k \leq d$, whose eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_d$. For some $\mathbf{a} \in \mathbb{R}^d$, and $c \in \mathbb{R}$ let $\mathbf{A}$ be given by*

$$\mathbf{A} = \begin{pmatrix} c & \mathbf{a}^\top \\ \mathbf{a} & \mathbf{M} \end{pmatrix}$$

Then

$$\lambda_{k+1}(\mathbf{A}) \geq \min(c, \lambda_k) - \|\mathbf{a}\|_2.$$

And bounds on matrix singular values when removing columns:

Lemma D.2 (Corollary 7.3.6 of Horn and Johnson [36]) *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a hermitian matrix and let $\hat{\mathbf{A}} \in \mathbb{C}^{m\times(n-1)}$ or $\in \mathbb{C}^{(m-1)\times n}$ be a hermitian matrix obtained by deleting any column or row from $\mathbf{A}$. Define $r := \text{rank}(\mathbf{A})$ and $\sigma_r(\hat{\mathbf{A}}) = 0$ if $m \geq n$ and a column is deleted or if $m \leq n$ and a row is deleted. Then:*

$$\sigma_1(\mathbf{A}) \geq \sigma_1(\hat{\mathbf{A}}) \geq \sigma_2(\mathbf{A}) \geq \sigma_2(\hat{\mathbf{A}}) \geq \cdots \geq \sigma_r(\mathbf{A}) \geq \sigma_r(\hat{\mathbf{A}})$$

Proof of Theorem 3.2.

WLOG we shall assume that $i^\star = \bar{n} + 1$ and $\mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{i^\star} = [m_1]$ (see Remark 2.1). In order to prove that social welfare is the sum of majority AND picky item users' top ratings, we shall go first prove the following claims:

Claim D.1 *Given $\widetilde{\mathbf{R}}$, the learner will pick $k^\star = k_{\text{MAJ}} + 1$.*

Claim D.2 *Let $\widehat{\mathbf{R}}$ be the rank $k^\star$ SVD truncation of $\widetilde{\mathbf{R}}$. We have that*

$$\hat{r}_{u,i} = \begin{cases} r^\star_{u,i} & u \in \mathcal{U}_{\text{MAJ}}, i \in [\bar{n}] \\ \tilde{r}_{u,i} & u \in (\mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{i^\star}), i = i^\star \\ 0 & ow \end{cases}.$$

Proof of Claim D.1.

Recall that the learner picks $k^\star$ by solving

$$\begin{aligned} &\min_{k \le \min\{m,n\}} \quad k \\ &\text{s.t. } \sigma_{k+1}(\widetilde{\mathbf{R}}) \le \alpha \end{aligned}$$

First, we will show that $\sigma_{k_{\text{MAJ}}+1}(\widetilde{\mathbf{R}}) > \alpha$ which implies that $k^\star \ge k_{\text{MAJ}} + 1$.

We shall show this by invoking Lemma D.1 to bound the $(k_{\text{MAJ}} + 1)$th singular value of a matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times (\bar{n}+1)}$ made up of the first $\bar{n} + 1$ columns of $\widetilde{\mathbf{R}}$. We will then show that this is weakly smaller than $\sigma_{k_{\text{MAJ}}+1}(\widetilde{\mathbf{R}})$ using lemma D.2.

Let matrix $\mathbf{X} \in \mathbb{R}^{m \times \bar{n}}$ be a matrix made up of the first $\bar{n}$ columns of $\mathbf{R}^\star$ (or equivalently, $\widetilde{\mathbf{R}}$).

Construct $\mathbf{A} \in \mathbb{R}^{(\bar{n}+1)\times(\bar{n}+1)}$ according to Lemma D.1: with $\mathbf{X}^\top \mathbf{X} \in \mathbf{R}^{\bar{n} \times \bar{n}}$, $c = \widetilde{\mathbf{R}}_{i^\star}^\top \widetilde{\mathbf{R}}_{i^\star} \in \mathbb{R}$, $\mathbf{a} = \mathbf{X}^\top \widetilde{\mathbf{R}}_{i^\star} \in \mathbb{R}^m$. This satisfies the conditions of Lemma D.1 when $k = k_{\text{MAJ}}$ and $d = \bar{n}$. Evaluating for each value in the bound of Lemma D.1:

$$\begin{aligned} c = \widetilde{\mathbf{R}}_{i^\star}^\top \widetilde{\mathbf{R}}_{i^\star} &= \sum_{u \in [m]} \tilde{r}_{u,i^\star}^2 \\ &= \sum_{u \in \mathcal{U}_{\text{A}}} \eta^2 + \sum_{u \in \mathcal{U}_{i^\star}} (r_{u,i^\star}^\star)^2 && \text{Definition 3.1} \\ &= \eta^2 |\mathcal{U}_{\text{A}}| + \|\mathbf{R}_{i^\star}^\star\|_2^2 && \text{By construction of EA strategy} \end{aligned}$$

Additionally:

$$\begin{aligned} \|\mathbf{a}\|_2^2 = \sum_{i \in [\bar{n}]} \left(\mathbf{X}^\top \widetilde{\mathbf{R}}_{i^\star}\right)_i^2 &= \sum_{i \in [\bar{n}]} \left(\sum_{u \in [m]} r_{u,i}^\star \tilde{r}_{u,i^\star} \right)^2 \\ &= \sum_{i \in [\bar{n}]} \left(\sum_{u \in \mathcal{U}_{\text{MAJ}}} r_{u,i}^\star \tilde{r}_{u,i^\star} \right)^2 && \text{exclusivity of } \mathcal{I}_{\text{MAJ}} \text{ and } \mathcal{I}_{\text{MIN}} \\ &= \sum_{i \in [\bar{n}]} (\eta \sum_{u \in \mathcal{U}_{\text{A}}} r_{u,i}^\star)^2 = \eta^2 \sum_{i \in [\bar{n}]} (\sum_{u \in \mathcal{U}_{\text{A}}} r_{u,i}^\star)^2 && \text{By construction of EA strategy} \\ &\le \eta^2 \bar{n} \max_{i \in [\bar{n}]} (\sum_{u \in \mathcal{U}_{\text{A}}} r_{u,i}^\star)^2 \\ &= \bar{n} (\eta \text{AV}_{i_{\bar{n}}^{\text{A}}})^2 \end{aligned}$$

Thus $\|\mathbf{a}\|_2 \le \eta \sqrt{\bar{n}} \text{AV}_{i_{\bar{n}}^{\text{A}}}$.

To get the bound of Lemma D.1, we also need singular values of $\mathbf{X}$ (equivalently, eigenvalues of $\mathbf{M} := \mathbf{X}^\top \mathbf{X}$). Clearly, the non-zero singular values of $\mathbf{X}$ and $\mathbf{R}_{\text{MAJ}}^\star$ are the same because $\mathbf{X}$ is simply $\mathbf{R}_{\text{MAJ}}^\star$ but padded with zeroes, thus:

$$\lambda_{k_{\text{MAJ}}} = \sigma_{k_{\text{MAJ}}}(\mathbf{X})^2 = \sigma_{k_{\text{MAJ}}}(\mathbf{R}_{\text{MAJ}}^\star)^2$$

We have then that

$$\begin{aligned} \sigma_{k_{\text{MAJ}}+1}(\widetilde{\mathbf{R}})^2 &\ge \sigma_{k_{\text{MAJ}}+1}(\tilde{\mathbf{X}})^2 && \text{Lemma D.2} \\ &= \lambda_{k_{\text{MAJ}}+1}(\mathbf{A}) && \text{By construction of } \mathbf{A} \\ &\ge \min(c, \lambda_{k_{\text{MAJ}}}) - \|\mathbf{a}\|_2 && \text{Lemma D.1} \\ &\ge \min(\eta^2 |\mathcal{U}_{\text{A}}| + \|\mathbf{R}_{i^\star}^\star\|_2^2, \sigma_{k_{\text{MAJ}}}(\mathbf{R}_{\text{MAJ}}^\star)^2) - \eta \sqrt{\bar{n}} \text{AV}_{i_{\bar{n}}^{\text{A}}} \\ &> \alpha^2 && \text{Theorem 3.2 assumption} \end{aligned}$$

This implies that $\sigma_{k_{\text{MAJ}}+1}(\widetilde{\mathbf{R}}) > \alpha$ and that $k^\star \geq k_{\text{MAJ}} + 1$.

Now we will show that $k^\star \leq k_{\text{MAJ}} + 1$. Note that for all $i > \bar{n} + 1$ and $u \in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{i^\star}$, $\tilde{r}_{u,i} = 0$. Likewise, for all $i \leq \bar{n} + 1$ and $u \notin \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{i^\star}$, $\tilde{r}_{u,i} = 0$. WLOG and because $i^\star$ is picky, we can represent $\widetilde{\mathbf{R}}$ as a block diagonal matrix:

$$\widetilde{\mathbf{R}} = \begin{pmatrix} \widetilde{\mathbf{R}}_{\text{MAJ}'} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{R}}_{\text{MIN}'} \end{pmatrix}$$

where $\widetilde{\mathbf{R}}_{\text{MAJ}'} \in \mathbb{R}^{m_1 \times \bar{n}+1}$ has reported ratings for users $u \in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{i^\star}$ and items $i \leq \bar{n} + 1$ and $\widetilde{\mathbf{R}}_{\text{MIN}'} \in \mathbb{R}^{(m-m_1)\times n-(\bar{n}+1)}$ has reported ratings for users $u \in (\mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{i^\star})$ and items $i > \bar{n} + 1$.

Recall that the singular values of a block diagonal matrix are simply a concatenation of the singular values of the two blocks. Since $\operatorname{rank}(\widetilde{\mathbf{R}}_{\text{MAJ}'}) \leq \operatorname{rank}(\mathbf{R}^\star_{\text{MAJ}}) + 1 = k_{\text{MAJ}} + 1$, it has no more than $k_{\text{MAJ}} + 1$ nonzero singular values. It follows that at least one of the $k_{\text{MAJ}} + 2$ largest singular values of $\widetilde{\mathbf{R}}$ are a singular value of $\tilde{\mathbf{R}}_2$. Therefore:

$$\begin{aligned}
\sigma_{k_{\text{MAJ}}+2}(\widetilde{\mathbf{R}}) &\leq \sigma_1(\widetilde{\mathbf{R}}_{\text{MIN}'}) \\
&\leq \sigma_1(\mathbf{R}^\star_{\text{MIN}}) && \text{Lemma D.2} \\
&< \alpha && \text{Theorem 3.2 assumption}
\end{aligned}$$

This implies that $k^\star \leq k_{\text{MAJ}} + 1$. So $k^\star = k_{\text{MAJ}} + 1$ as desired. □

Proof of Claim D.2.

Recall that the $k^\star$-truncated SVD is $\sum_{j \in [k^\star]} \sigma_j \mathbf{u_j} \mathbf{v_j}^\top$ where $[k^\star]$ are the $k^\star$ largest singular values. In the above claims we showed that $\sigma_{k_{\text{MAJ}}+1}(\widetilde{\mathbf{R}}) \geq \min\{\sigma_{k_{\text{MAJ}}}(\mathbf{R}^\star_{\text{MAJ}})^2, \eta^2 |\mathcal{U}_{\text{A}}| + \text{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}}$ and so by assumption, the $k_{\text{MAJ}} + 1$ largest singular values are all strictly greater than $\sigma_1(\widetilde{\mathbf{R}}_{\text{MIN}'})$.

In the proof of Theorem 3.1, we showed that the SVD of a block diagonal matrix can be decomposed into a sum of terms for each block. Therefore, because the $k_{\text{MAJ}} + 1$ largest singular values are all strictly greater than $\sigma_1(\widetilde{\mathbf{R}}_{\text{MIN}'})$, it must be case that for $\widetilde{\mathbf{R}}$, $\sum_{j \in [k_{\text{MAJ}}+1]} \sigma_j \mathbf{u_j} \mathbf{v_j}^\top$, where $[k_{\text{MAJ}} + 1]$ are the $k_{\text{MAJ}} + 1$ largest singular values, form the following matrix:

$$\begin{pmatrix} \widetilde{\mathbf{R}}_{\text{MAJ}'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Of course, from Claim D.1, we have that $k^\star = k_{\text{MAJ}} + 1$, thus this completes Claim D.2. □

Now, we shall use Claims D.1 and D.2 to prove our Theorem result.

For all $u \in \mathcal{U}_{\text{MAJ}}$ and all $i \neq i^\star$, $r^\star_{u,i} = \hat{r}_{u,i}$. To show that user u will be recommended a top item it therefore suffices to show that picky item, $i^\star$, will not become a top item. This is true by construction:

$$\hat{r}_{u,i^\star} \leq \eta < \min_{u' \in \mathcal{U}_{\text{MAJ}}} \max_{i \in [n]} r^\star_{u'i} \leq \max_{i \in [n]} r^\star_{u,i} \quad \forall u \in \mathcal{U}_{\text{MAJ}}$$

For all $u \in \mathcal{U}_{i^\star}$, $r^\star_{u,i} = \hat{r}_{u,i}$ for all $i \in [n]$. Thus $\arg\max_{i \in [n]} \hat{r}_{u,i} = \arg\max_{i \in [n]} r^\star_{u,i} = i^\star$.

Lastly, for all $u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{i^\star}$:

$$\arg\max_{i \in [n]} \hat{r}_{u,i} = [n]$$

Therefore, we will have

$$\texttt{top}(u) \in \arg\max_{i \in [n]} \|\widehat{\mathbf{R}}_i\|_1 = \arg\max_{i \in [\bar{n}+1]} \|\widehat{\mathbf{R}}_i\|_1 \subseteq [\bar{n} + 1]$$

and recall that $r^\star_{u,i} = 0 \; \forall i \in [\bar{n} + 1], u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{i^\star}$
As such:

$$\begin{aligned}
\text{SW}(\widetilde{\mathbf{R}}, \alpha) &= \sum_{u \in \mathcal{U}_{\text{MAJ}}} \max_{i \in [n]} r^\star_{u,i} + \sum_{u \in \mathcal{U}_{i^\star}} \max_{i \in [n]} r^\star_{u,i} + \sum_{u \in \mathcal{U}_{\text{MIN}}} 0 \\
&= \sum_{u \in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{i^\star}} \max_{i \in [n]} r^\star_{u,i}.
\end{aligned}$$

□

D.2.2 Formal statement of strict welfare increases

Corollary D.1 (Altruism Improves SW) *When the assumptions of Theorem 3.2 hold, we have that*

$$\rho := \frac{\text{SW}(\widetilde{\mathbf{R}},\alpha)}{\text{SW}(\mathbf{R}^\star,\alpha)} = 1 + \frac{\sum_{u\in\mathcal{U}_{i^\star}} r^\star_{u,i^\star}}{\sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i}} > 1$$

Proof of D.1. Recall that

$$\mathcal{G}(\mathbf{R},\mathcal{U}_{\text{A}},\eta) = \left(\sigma_1(\mathbf{R}_{\text{MIN}}), \sqrt{\min\{\sigma_{k_{\text{MAJ}}}(\mathbf{R}_{\text{MAJ}})^2, \eta^2|\mathcal{U}_{\text{A}}| + \text{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}}}\right)$$

and

$$\mathcal{G}(\mathbf{R}) = (\sigma_1(\mathbf{R}_{\text{MIN}}), \sigma_{k_{\text{MAJ}}}(\mathbf{R}_{\text{MAJ}}))\,.$$

Since $\eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}}$ is strictly positive, we have that the upper-bound of $\mathcal{G}(\mathbf{R},\mathcal{U}_{\text{A}},\eta)$ is smaller while they are both lower-bounded by $\sigma_1(\mathbf{R}_{\text{MIN}})$. Consequently,

$$\alpha \in \mathcal{G}(\mathbf{R},\mathcal{U}_{\text{A}},\eta) \implies \alpha \in \mathcal{G}(\mathbf{R}).$$

Thus, by Theorem 3.1, $\text{SW}(\mathbf{R}^\star,\alpha) = \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i}$. Additionally, by Theorem 3.2, $\text{SW}(\widetilde{\mathbf{R}},\alpha) = \sum_{u\in\mathcal{U}_{\text{MAJ}}\cup\mathcal{U}_{i^\star}} \max_{i\in[n]} r^\star_{u,i}$. Taking the ratio:

$$\begin{aligned}\frac{\text{SW}(\widetilde{\mathbf{R}},\alpha)}{\text{SW}(\mathbf{R}^\star,\alpha)} &= \frac{\sum_{u\in\mathcal{U}_{\text{MAJ}}\cup\mathcal{U}_{i^\star}} \max_{i\in[n]} r^\star_{u,i}}{\sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i}}\\ &= 1 + \frac{\sum_{u\in\mathcal{U}_{i^\star}} \max_{i\in[n]} r^\star_{u,i}}{\sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i}} > 1.\end{aligned}$$

Where the last inequality follows by our assumption that $\mathcal{U}_{i^\star}$ is non-empty, and individuals provide positive ratings for item $i^\star$. Additionally, note that the denominator is well-defined as $\mathbf{R}_{\text{MAJ}}$ has at least 1 positive entry by the fact that it has a minimum singular of at least α. □

D.2.3 Formal statement of sufficient conditions

Corollary D.2 (Sufficient conditions on altruistic uprating to improve social welfare) *Let* $\mathbf{R}^\star$ *be a majority-minority matrix with a picky item* $i^\star > \bar{n}$ *and* $\alpha > \sigma_1(\mathbf{R}^\star_{\text{MIN}})$. *Then, an* $(\eta,\mathcal{U}_A)$*-altruistic uprating improves social welfare if the following hold:*

$$0 < \eta < \kappa, \quad \alpha < \sqrt{\min\{\sigma_{k_{\text{MAJ}}}(\mathbf{R}^\star_{\text{MAJ}})^2, \eta^2|\mathcal{U}_{\text{A}}| + \text{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}}}$$

where $\kappa := \min_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i}$.

Proof of Corollary D.2. Suppose we have an altruistic rating $(\eta,\mathcal{U}_{\text{A}})$ for the picky item such that the conditions above hold. Then we must have that

$$\alpha \in (\sigma_1(\mathbf{R}^\star_{\text{MIN}}), \sqrt{\min\{\sigma_{k_{\text{MAJ}}}(\mathbf{R}^\star_{\text{MAJ}})^2, \eta^2|\mathcal{U}_{\text{A}}| + \text{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}}})$$

The space in this interval is $\mathcal{G}(\mathbf{R}^\star,\mathcal{U}_{\text{A}},\eta)$ by definition. Equivalently, $\alpha \in \mathcal{G}(\mathbf{R}^\star,\mathcal{U}_{\text{A}},\eta)$. Since α clearly exists, $\mathcal{G}(\mathbf{R}^\star,\mathcal{U}_{\text{A}},\eta) \neq \emptyset$, which means $\mathbf{R}^\star$ has $(\eta,\mathcal{U}_{\text{A}})$-sufficient singular value gap. Thus by Theorem 3.2 and Corollary D.1, social welfare is improved by the manipulated matrix. □

D.2.4 Proof of Theorem 3.3

Proof of thm 3.3. Let η be the returned output of Algorithm 1. Note that the index of the picky item is $\bar{n}+1$ without loss of generality to any $i^\star > \bar{n}$, see remark 2.1. Thus we will return to $i^\star$ as if it were $\bar{n}+1$ for the sake of this proof. There are two parts to Theorem 3.3 that we present as claims and prove sequentially for the cases when η returned by Algorithm 1 is positive or zero.

Claim D.3 *If* $\eta > 0$ *is returned by Algorithm 1, then playing* η *will improve social welfare.*

Proof of Claim D.3. By Corollary D.2, it is sufficient to show that η (when $\eta \neq 0$) satisfies the following:

1. $\alpha^2 < \min\{\sigma_{k_{\text{MAJ}}}(\mathbf{R}^\star_{\text{MAJ}})^2, \eta^2|\mathcal{U}_{\text{A}}| + \text{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}}$
2. $\eta > 0$

3. $\eta < \kappa$

We first focus on the inequality:

$$\alpha^2 < \min\{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2, \eta^2|\mathcal{U}_{\mathrm{A}}| + \mathrm{ASV}_{i^\star}\} - \eta\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} \tag{3}$$

Where $\eta > 0$, the inequality above is equivalent to both of the following statements holding:

$$\alpha^2 < \sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \eta\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} \iff \eta < \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}} \tag{4}$$

and

$$\alpha^2 < \eta^2|\mathcal{U}_{\mathrm{A}}| + \mathrm{ASV}_{i^\star} - \eta\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} \iff \eta^2|\mathcal{U}_{\mathrm{A}}| - \eta\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \mathrm{ASV}_{i^\star} - \alpha^2 > 0 \tag{5}$$

Clearly equation 4 is an upper bound on η. We shall analyze equation 5 to get explicit bounds on η

Let $f(\eta)$ be the quadratic of equation 5 in terms of η with discriminant $d := \bar{n}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}{}^2 + 4|\mathcal{U}_{\mathrm{A}}|(\alpha^2 - \mathrm{ASV}_{i^\star})$. Now we need to understand for which set of $\eta \in \mathbb{R}$, $f(\eta) > 0$. Notice that, by standard properties of quadratic functions, if $d \geq 0$, $f(\eta) > 0$ where $\eta \in \left[\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}\right]^C$ and if $d < 0$, $f(\eta) > 0 \quad \forall \eta \in \mathbb{R}$. Consequently, the set of feasible η for equation 3 to hold break into the following cases:

1. Case 1: $d < 0$, therefore equation 5 does not constrain η and only equation 4 and positivity is important:

$$\eta \in \left(0, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}}\right)$$

2. Case 2: $d \geq 0$, η must be feasible for both equation 5 and 4 and positive.

$$\eta \in \left[\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}\right]^C \cap \left(0, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}}\right)$$

Note that $\frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}} > 0$ because $\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}}) > \alpha$ by setting assumptions.

We can further rewrite Case 2. Notice that by setting $\nabla f = 0$, the minimum of $f(\eta)$ is at $\eta = \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}}{2|\mathcal{U}_{\mathrm{A}}|}$ which is greater than 0 by setting assumptions. Thus, it must be that $\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|} > 0$ because $\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}$ is the right hand root $f(\eta)$.

1. Case 1: $d < 0$

$$\eta \in \left(0, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}}\right)$$

2. Case 2: $d \geq 0$,

$$\eta \in \left(0, \min(\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right) \cup \left(\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}}\right)$$

Now we finally add the $\eta < \kappa$ to the sufficient conditions. this becomes a part of both case's upper bounds:

1. Case 1: $d < 0$,

$$\eta \in \left(0, \min(\kappa, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^\star_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right)$$

2. Case 2: $d \geq 0$,

$$\eta \in \left(0, \min(\kappa, \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right) \cup \left(\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \min(\kappa, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right)$$

Note that we have shown that these cases are equivalent to the sufficient conditions we must prove are met.

It is easy to see that in either case, when the relevant space is non-empty, Algorithm 1 returns an η in the space because the algorithm first checks the discriminant and then constructs the relevant range(s) (if nonempty). □

Claim D.4 *Algorithm 1 returns* 0 *if and only if there is no* η *correlated strategy that satisfies our feasible conditions.*

Proof of Claim D.4. In our proof of Claim D.3, we showed that an equivalent way to characterize an η that satisfies our sufficient conditions for SW improvement is the following:

1. Case 1: $d < 0$,

$$\eta \in (0, \min(u, \kappa))$$

2. Case 2: $d \geq 0$,

$$\eta \in (0, \min(\kappa, r_1, u)) \cup (r_2, \min(\kappa, u))$$

Where

$$d := \bar{n}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}{}^2 + 4|\mathcal{U}_{\mathrm{A}}|(\alpha^2 - \mathrm{ASV}_{i^\star})$$

$$u := \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})$$

$$r_1 := \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}$$

$$r_2 := \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}$$

and $\bar{n}, |\mathcal{U}_E A|, \sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}}), \kappa, \mathrm{ASV}_{i^\star}, \mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}$ are parameters of the algorithm

Suppose there does not exist an η that is feasible according to our cases. It must be the case that $d \geq 0$, because otherwise there is clearly always a feasible η as $\kappa, u > 0$ by setting assumptions. Since there is no feasible η, it must be the case that

$$\eta \in \left(0, \min(\kappa, \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right) \cup \left(\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \min(\kappa, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right)$$

is an empty space. Algorithm 1 first checks the LHS set. If it is empty, it checks the RHS, and if that is empty, it returns 0. Therefore if η is infeasible, 0 is returned.

Suppose 0 is returned by the algorithm. It clearly could not have been the case that $d < 0$ because given $\kappa, u > 0$, $d < 0$ would never result in a returned 0. Thus we consider $d \geq 0$. In this case, evaluating the If statements, 0 is clearly returned only if

$$\eta \in \left(0, \min(\kappa, \frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} - \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right) \cup \left(\frac{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}} + \sqrt{d}}{2|\mathcal{U}_{\mathrm{A}}|}, \min(\kappa, \frac{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}^{\star}_{\mathrm{MAJ}})^2 - \alpha^2}{\sqrt{\bar{n}}\mathrm{AV}_{i^{\mathrm{A}}_{\bar{n}}}})\right)$$

is empty. Thus we have that if 0 is returned, there must be no feasible η (for our sufficient conditions). □

Having proven both Claims, we have shown both parts of Theorem 3.3 hold, so we may conclude the full proof. □

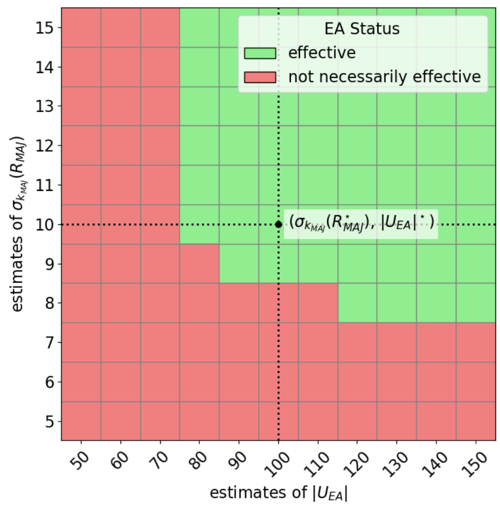


Figure 1: Robustness of Algorithm 1 to imperfect knowledge of $|\mathcal{U}_{\mathrm{A}}|$ and $\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}})$. Squares are green if for estimates $(|\mathcal{U}_{\mathrm{A}}|, \sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}}))$ taken as input into Algorithm 1, the outputted value of η satisfies the sufficient conditions for effective altruism given in Theorem 3.2.

D.3 Supplementary material for robustness

D.3.1 Empirical robustness example

In the following example we illustrate that even slight mis-estimates of the required parameters will still yield an η returned by the algorithm that satisfies the conditions for effective altruism.

Example D.1 (Simple Robustness Example) *Let $\mathbf{R}^{\star}$ be the very simple majority-minority matrix from example C.1 with $m_{\mathrm{MIN}} = 10$ and $m_{\mathrm{MAJ}} = 100$. Let the learner's loss-tolerance parameter be $\alpha = \sqrt{15}$ and assume half of users who like item 1 and half of users who like item 2 are altruistic. With perfect information, the output of Algorithm 1 is $\eta^{\star} \approx .886$.[2] For $\eta < 1$, we have that $\eta < \min_{u \in \mathcal{U}_{\mathrm{MAJ}}} \max_{i \in [n]} r^{*}_{u,i} = 1$, additionally, $\alpha > \sigma_1(\mathbf{R}_{\mathrm{MIN}})$. Therefore, to satisfy the sufficient conditions of Theorem 3.2, it suffices to check that*

$$f(\eta) = \min\{\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}})^2, \eta^2|\mathcal{U}_A| + \mathrm{ASV}_{i^{\star}}\} - \eta\sqrt{\bar{n}}\mathrm{AV}_{i^A_{\bar{n}}} - \alpha^2 > 0.$$

Solving this analytically (see proof of Theorem 3.3), we get that all $\eta \in [0.78, 1)$ satisfy $f(\eta) > 0$, that is, there is a significant range of η that result provably in effective altruism.

To test how robust Algorithm 1 is to imperfect information, we created a sets of 11 estimates for $\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}})$ and $|\mathcal{U}_A|$ centered at true values of 10 and 100, respectively. Estimates of $\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}})$ range from 5 to 15 with a granularity of 1. Estimates of $|\mathcal{U}_A|$ range from 50 to 150 with a granularity of 10. For the mesh of 121 parameter pairs, we test if Algorithm 1, using these estimates as input, returns an η such that $f(\eta) > 0$. We find that all tested value pairs such that $\sigma_{k_{\mathrm{MAJ}}}(\mathbf{R}_{\mathrm{MAJ}}) \geq 10$ and $|\mathcal{U}_A| > 80$ result provably in effective altruism (see Figure 1).

D.3.2 Main robustness result

Theorem D.1 (Robustness of Algorithm 1 to misspecifications) *Let $\mathbf{R}^{\star}$ be a majority-minority matrix. Under the assumptions of Theorem 3.2, further assume κ, n and $\|\mathbf{R}^{\star}\|_1, \|\mathbf{R}^{\star}\|_2$ are public knowledge. Thus, let $\mathbf{z} := (\sigma_{k_{maj}}(\mathbf{R}_{\mathrm{MAJ}}), \alpha, \bar{n}, \mathrm{ASV}_{i^{\star}}, \mathrm{AV}_{i^A_{\bar{n}}}, |\mathcal{U}_A|)$ be the vector of unknown parameters and call $\mathbf{z}^{\star}$ be the vector of true parameters, $\hat{\mathbf{z}}$ the vector of estimates, and $\hat{\eta}$ the value returned by the algorithm. Define*

- *A function, f, of the unknown parameters and parameterized by a feasible η,*
$$f(\mathbf{z}; \eta) := \min(\sigma^2, \eta^2|\mathcal{U}_A| + \mathrm{ASV}_{i^{\star}}) - \eta\sqrt{\bar{n}}AV - \alpha^2$$
- *A function, L, of preference matrix, $\mathbf{R}$ and parameterized by a feasible η,*
$$L(\mathbf{R}; \eta) := \sqrt{4\|\mathbf{R}\|_2^2 + (\eta\|\mathbf{R}\|_1^2)/4 + \eta^2 n + \max\{4\|\mathbf{R}\|_2^2, 1 + \eta^4\}}$$

[2] Code for this example is submitted as supplementary material.

When $\hat{\eta} > 0$, if $\|\hat{\mathbf{z}} - \mathbf{z}^\star\|_2 < {}^{f(\hat{\mathbf{z}};\hat{\eta})}/_{L(\mathbf{R}^\star;\hat{\eta})}$ then the conclusions of Theorem 3.2 still hold when $\hat{\eta}$ is played by altruistic users.

In order to prove this, we will invoke Lipschitz bounds on a function that is a minimum of two Lipschitz functions, so the following lemma will be helpful:

Lemma D.3 *Let $f(\mathbf{z}) = \min\{f_1(\mathbf{z}), f_2(\mathbf{z})\}$ where f_1 and f_2 are Lipschitz on a convex region $\mathcal{D}$ with constants L_1 and L_2, respectively. Then f is Lipschitz on $\mathcal{D}$ with constant $L = \max\{L_1, L_2\}$.*

Proof of Theorem D.1. Consider two arbitrary points $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{D}$. Assume without loss of generality that $f(\mathbf{z}_1) \geq f(\mathbf{z}_2)$. If $f_1(\mathbf{z}_1) \geq f_2(\mathbf{z}_1)$:

$$\begin{aligned}
|f(\mathbf{z}_1) - f(\mathbf{z}_2)| &= |f_1(\mathbf{z}_1) - \min\{f_1(\mathbf{z}_2), f_2(\mathbf{z}_2)\}| \\
&= |f_1(\mathbf{z}_1) + \max\{-f_1(\mathbf{z}_2), -f_2(\mathbf{z}_2)\}| \\
&= |\max\{f_1(\mathbf{z}_1) - f_1(\mathbf{z}_2), f_1(\mathbf{z}_1) - f_2(\mathbf{z}_2)\}| \\
&\leq |\max\{f_1(\mathbf{z}_1) - f_1(\mathbf{z}_2), f_2(\mathbf{z}_1) - f_2(\mathbf{z}_2)\}| \\
&\leq \max\{|f_1(\mathbf{z}_1) - f_1(\mathbf{z}_2)|, |f_2(\mathbf{z}_1) - f_2(\mathbf{z}_2)|\} \\
&\leq \max\{L_1\|\mathbf{z}_1 - \mathbf{z}_2\|, L_2\|\mathbf{z}_1 - \mathbf{z}_2\|_2\} \\
&= \max\{L_1, L_2\}\|\mathbf{z}_1 - \mathbf{z}_2\|_2
\end{aligned}$$

By making a symmetric argument for $f_1(\mathbf{z}_1) < f_2(\mathbf{z}_1)$ we get the same bound. Thus, f is Lipschitz on $\mathcal{D}$ with constant $L = \max\{L_1, L_2\}$ as desired. □

With this lemma, we will now proceed with the full proof.

Proof of Theorem D.1. For simplicity of notation, we will denote $\sigma_{k_{maj}}(\mathbf{R}^\star_{\text{MAJ}})$ as σ.

Recall from Corollary D.2, it suffices to show that

1. $\hat{\eta} > 0$
2. $\hat{\eta} < \kappa$
3. $0 < \min\{(\sigma^\star)^2, \hat{\eta}^2|\mathcal{U}^\star_{\text{A}}| + \text{ASV}_{i^\star}{}^\star\} - \hat{\eta}\sqrt{\bar{n}^\star}\text{AV}_{i^{\text{A}}_{\bar{n}}}{}^\star - (\alpha^\star)^2$

for the true parameters, $\mathbf{z}^\star$ and the returned $\hat{\eta}$. Because $\hat{\eta} > 0$, by assumption, clearly the first condition is satisfied. Also, the 2nd condition must be satisfied because $\hat{\eta}$ was returned by Algorithm 1, and Theorem 3.3 asserts that any nonzero η returned by the algorithm satisfies $\hat{\eta} < \kappa$.

Thus all we must prove is that $f(\mathbf{z}^\star; \hat{\eta}) > 0$. Given that $f(\hat{\mathbf{z}}; \hat{\eta}) > 0$ by Theorem 3.3

Note that we can equivalently write $f(\mathbf{z}; \eta) = \min\{f_1(\mathbf{z}; \eta), f_2(\mathbf{z}; \eta)\}$ where

$$f_1(\mathbf{z}; \eta) = \sigma^2 - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}} - \alpha^2$$

$$f_1(\mathbf{z}; \eta) = \eta^2|\mathcal{U}_{\text{A}}| + \text{ASV}_{i^\star} - \eta\sqrt{\bar{n}}\text{AV}_{i^{\text{A}}_{\bar{n}}} - \alpha^2$$

Because $\eta, \text{AV}_{i^{\text{A}}_{\bar{n}}}, \bar{n}, \alpha > 0$ by setting assumptions.

We note that for the remainder of this proof, we are exclusively interested in $f(\cdot; \hat{\eta})$ (i.e. f with the the returned $\hat{\eta}$ as the parameter), so for notational simplicity, we often drop parameter, $\hat{\eta}$.

Assume that f is L-Lipschitz on some region $\mathcal{D}$ (we will prove later that this is true for a suitably defined $\mathcal{D}$). Then, given $\|\mathbf{z} - \hat{\mathbf{z}}\|_2 < \frac{\Delta(\hat{\mathbf{z}})}{L}$ and $\hat{\mathbf{z}}, \mathbf{z} \in \mathcal{D}$:

$$\begin{aligned}
f(\mathbf{z}) &= f(\hat{\mathbf{z}}) + (f(\mathbf{z}) - f(\hat{\mathbf{z}})) \\
&\geq f(\hat{\mathbf{z}}) - |f(\mathbf{z}) - f(\hat{\mathbf{z}})| \\
&\geq f(\hat{\mathbf{z}}) - L\|\mathbf{z} - \hat{\mathbf{z}}\|_2 \quad f \text{ is } L\text{-Lipschitz} \\
&> f(\hat{\mathbf{z}}) - f(\hat{\mathbf{z}}) \quad \text{by assumption} \\
&= 0
\end{aligned}$$

Therefore, to show the desired $f(\hat{\mathbf{z}}; \hat{\eta}) > 0$, it suffices to show that f is L-Lipschitz on $\mathcal{D}$. We will do this by finding a bound on the maximum gradient norm of f_1 and f_2. First we will compute the gradient of both functions:

$$\begin{aligned}\|\nabla f_1(\mathbf{z})\|_2^2 &= \|(\frac{\partial f_1}{\partial \sigma}(\mathbf{z}), \frac{\partial f_1}{\partial \alpha}(\mathbf{z}), \frac{\partial f_1}{\partial \bar{n}}(\mathbf{z}), \frac{\partial f_1}{\partial \mathrm{ASV}_{i^\star}}(\mathbf{z}), \frac{\partial f_1}{\partial \mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}}(\mathbf{z}), \frac{\partial f_1}{\partial |\mathcal{U}_{EA}|}(\mathbf{z}))^\top\| \\ &= \|(2\sigma, -2\alpha, \frac{\hat{\eta}\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}}{2\sqrt{\bar{n}}}, 0, -\hat{\eta}\sqrt{\bar{n}}, 0)^\top\|_2^2 = 4\sigma^2 + 4\alpha^2 + \frac{\hat{\eta}\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}{}^2}{4\sqrt{\bar{n}}} + \hat{\eta}^2\bar{n} \qquad \text{Definition of } f_1(\mathbf{z})\end{aligned}$$

$$\begin{aligned}\|\nabla f_2(\mathbf{z})\|_2^2 &= \|(\frac{\partial f_2}{\partial \sigma}(\mathbf{z}), \frac{\partial f_2}{\partial \alpha}(\mathbf{z}), \frac{\partial f_2}{\partial \bar{n}}(\mathbf{z}), \frac{\partial f_2}{\partial \mathrm{ASV}_{i^\star}}(\mathbf{z}), \frac{\partial f_2}{\partial \mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}}(\mathbf{z}), \frac{\partial f_2}{\partial |\mathcal{U}_{EA}|}(\mathbf{z}))^\top\| \\ &= \|(0, -2\alpha, \frac{\hat{\eta}\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}}{2\sqrt{\bar{n}}}, 1, -\hat{\eta}\sqrt{\bar{n}}, \hat{\eta}^2)^\top\|_2^2 = 4\alpha^2 + \frac{\hat{\eta}\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}{}^2}{4\sqrt{\bar{n}}} + 1 + \hat{\eta}^2\bar{n} + \hat{\eta}^4 \qquad \text{Definition of } f_2(\mathbf{z})\end{aligned}$$

To bound the gradient norms we will use bounds on the parameters in terms of the primitives of the matrix. Recall that $\hat{\eta}$ is fixed parameter (it is what Algorithm 1 recommends). Thus, the gradient norm depend on $\alpha, \bar{n}, \mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}, \frac{1}{\sqrt{\bar{n}}}$. Since all variables are positive, it suffices to find an upper-bound in terms of the known values of n, $\|\mathbf{R}^\star\|_1$ and $\|\mathbf{R}^\star\|_2$. As $1 \leq \bar{n} \leq n$ we have that $\bar{n} \leq n$ and $\frac{1}{\sqrt{n}} \leq 1$. By assumption, we have that $\alpha < \sigma_{k_{maj}}(\mathbf{R}^\star_{\mathrm{MAJ}}) = \sigma$. Additionally, that $\sigma_{k_{maj}}(\mathbf{R}^\star_{\mathrm{MAJ}}) \leq \sigma_1(\mathbf{R}^\star) = \|\mathbf{R}^\star\|_2$. Hence $\alpha < \sigma < \|\mathbf{R}^\star\|_2$. Lastly,

$$\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}} = \max_{i \in [\bar{n}]} \sum_{u \in \mathcal{U}_{\mathtt{A}}} r_{u,i} \leq \max_{i \in [n]} \sum_{u \in [m]} r_{u,i} = \|\mathbf{R}^\star\|_1$$

Hence we can define $\mathcal{D}$ to be the space where each of these bounds hold:

$$\mathcal{D} = \{\mathbf{z} : 0 < \alpha < \sigma \leq \|\mathbf{R}^\star\|_2, 1 \leq \bar{n} \leq n, \mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}} \leq \|\mathbf{R}^\star\|_1\}$$

Since it is defined by a set of linear inequalities, $\mathcal{D}$ is convex.

Further we can bound the norms of the gradients for all $\mathbf{z} \in \mathcal{D}$:

$$\|\nabla f_1(\mathbf{z})\|_2^2 = 4\sigma^2 + 4\alpha^2 + \frac{\hat{\eta}\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}{}^2}{4\sqrt{\bar{n}}} + \hat{\eta}^2\bar{n} \leq 8\|\mathbf{R}^\star\|_2^2 + \frac{\hat{\eta}\|\mathbf{R}^\star\|_1^2}{4} + \hat{\eta}^2 n$$

$$\|\nabla f_1(\mathbf{z})\|_2^2 = 4\alpha^2 + \frac{\hat{\eta}\mathrm{AV}_{i^{\mathtt{A}}_{\bar{n}}}{}^2}{4\sqrt{\bar{n}}} + 1 + \hat{\eta}^2\bar{n} + \hat{\eta}^4 \leq 4\|\mathbf{R}^\star\|_2^2 + \frac{\hat{\eta}\|\mathbf{R}^\star\|_1^2}{4} + 1 + \hat{\eta}^2 n + \hat{\eta}^4$$

Consequently, for $L_1 = \sqrt{8\|\mathbf{R}^\star\|_2^2 + \frac{\hat{\eta}\|\mathbf{R}^\star\|_1^2}{4} + \hat{\eta}^2 n}$ and $L_2 = \sqrt{4\|\mathbf{R}^\star\|_2^2 + \frac{\hat{\eta}\|\mathbf{R}^\star\|_1^2}{4} + 1 + \hat{\eta}^2 n + \hat{\eta}^4}$ we have that, by the Mean Value Theorem, for any $\mathbf{z}, \mathbf{z}' \in \mathcal{D}$ and $i \in \{1, 2\}$:

$$|f_i(\mathbf{z}) - f_i(\mathbf{z}^{'})| \leq \sup_{\mathbf{z} \in \mathcal{D}} \|\nabla f_i(\mathbf{z})\|_2 \|\mathbf{z} - \mathbf{z}'\|_2 \leq L_i \|\mathbf{z} - \mathbf{z}'\|_2$$

That is, f_i is L_i-Lipschitz on $\mathcal{D}$.

Applying Lemma D.3, we can conclude that $f = \min\{f_1, f_2\}$ is Lipschitz with $L = \max\{L_1, L_2\}$.
□

D.3.3 Additional Corollary to bound EA group misestimation

We now have a bound for how perturbed a vector of parameters for the correlating mechanism may be such that the resulting $\hat{\eta}$ still works. There are certain parameters that we imagine are more/less likely to be incorrect. In particular, it is likely the case that estimating who is altruistic, i.e. cares about the minorities and intends to participate in a grassroots uprating movement, will be difficult. Thus we would be particularly interested in robustness to estimations on the EA user group. From Theorem D.1, we directly can derive a corollary on just incorrect estimations of how many and which agents participate in the movement.

Corollary D.3 (Robustness of Algorithm 1 to estimation of EA group) *Under the assumptions and definitions of Theorem D.1, let it be the case that the algorithm must estimate which majority users care about minorities, but all other required parameters about true preferences and the learner are correct. Then if*

$$(\widehat{|\mathcal{U}_A|} - |\mathcal{U}_A|^\star)^2 + (\widehat{\mathrm{AV}_{i^A_{\bar{n}}}} - \mathrm{AV}_{i^A_{\bar{n}}}{}^\star)^2 < \frac{f(\hat{\mathbf{z}};\hat{\eta})}{L(\mathbf{R}^\star;\hat{\eta})}$$

and $\hat{\eta} > 0$, then the conclusions of Theorem 3.2 will still hold even when the true group of altruists uprate by $\hat{\eta}$.

Proof. Recall that

$$\mathbf{z} := (\sigma_{k_{maj}}(\mathbf{R}_{\text{MAJ}})\,\alpha, \bar{n}, \text{ASV}_{i^\star}, \text{AV}_{i^{\text{A}}_{\bar{n}}}, |\mathcal{U}_{\text{A}}|)$$

The first 4 elements of $\mathbf{z}$ do not depend on the group of EA users, therefore those are estimated correctly. Thus $(\hat{\mathbf{z}} - \mathbf{z}^\star)_j = 0\ \forall j \in [4]$. The inequality then follows directly from Theorem D.1. □

D.4 Supplementary material for Section 3.4

D.4.1 Supplementary material for benevolent learner

Corollary D.4 (EA increases utility for the benevolent learner) *Under the assumptions for Theorem 3.2, an α-loss tolerant learner with a benevolent utility function would achieve*

$$U^{\text{TRUE}}_{\text{BEN}} = \textstyle\sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i}, \quad U^{A}_{\text{BEN}} = \textstyle\sum_{u\in(\mathcal{U}_{\text{MAJ}}\cup\mathcal{U}_{i^\star})} \max_{i\in[n]} r^\star_{u,i}$$

when agents are truthful or follow an effective altruist strategy, respectively, and $U^{A}_{\text{BEN}} > U^{\text{TRUE}}_{\text{BEN}}$.

Proof of Corollary D.4. This follows directly from Theorems 3.1 and 3.2 because by definition of benevolence, $\mathtt{U}^{\mathtt{TRUE}}_{BEN} := \text{SW}(\mathbf{R}^\star, \alpha)$ and $\mathtt{U}^{\mathtt{EA}}_{BEN} := \text{SW}(\widetilde{\mathbf{R}}, \alpha)$. □

D.4.2 Supplementary material for engagement-based learner

Proposition D.1 (EA increases utility for the engagement-based learner) *Under assumptions for Theorem 3.2, and α-loss tolerant learner with engagement-based utility would achieve*

$$U^{\text{TRUE}}_{\text{EN}} = \textstyle\sum_{i\in[n]}\sum_{u\in[m]} |r^\star_{u,i}|, \quad U^{A}_{\text{EN}} = \textstyle\sum_{i\in[n]}\sum_{u\in[m]} |r^\star_{u,i}| + \eta|\mathcal{U}_A|$$

when agents are truthful or follow an effective altruist strategy, respectively, and $U^{A}_{\text{EN}} > U^{\text{TRUE}}_{\text{EN}}$.

Proof of Proposition D.1. When agents report truthfully, $r^\star_{u,i} = \tilde{r}_{u,i}$.

Under the η EA correlated strategy, recall that the reported preference matrix is:

$$\widetilde{\mathbf{R}} = \begin{pmatrix} r^{\text{MAJ}^\star}_{11} & \dots & r^{\text{MAJ}^\star}_{1,\bar{n}} & \eta & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r^{\text{MAJ}^\star}_{u,1} & \dots & r^{\text{MAJ}^\star}_{u,\bar{n}} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r^{\text{MAJ}^\star}_{\bar{m},1} & \dots & r^{\text{MAJ}^\star}_{\bar{m},\bar{n}} & \eta & \dots & 0 \\ \mathbf{0} & \dots & \mathbf{0} & & \mathbf{R}^\star_{\text{MIN}} & \end{pmatrix}$$

Which is just the true matrix, but with 0s swapped to be η for the EA users.

A feasible η is such that $\eta > 0$ and $|\mathcal{U}_{\text{A}}| > 0$ by assumptions of Theorem 3.2 hence the strict inequality. □

E Supplementary material for Section 4

E.1 Experiment

Code is available in the Github codebase.

Table 2: Popularity of genres. For each genre, this table depicts the number of users in that genre group (the set of users who review that genre more than any other genre) and the total number of book reviews for that genre.

Genre	Users in genre group	Total reviews (in thousands)
romance	7627	1044
young-adult	5061	869
fantasy, paranormal	4427	823
mystery, thriller, crime	1087	650
history, historical fiction, biography	668	320
children	3	114
comics, graphic	12	71
non-fiction	7	47
poetry	0	14

E.1.1 Data information

Dataset. We use Goodreads Datasets collected in 2017 by a UCSD lab [69, 68] to gather information about how much different readers (users) engage with books from different genres (items). We get user engagement data over genres by merging "English review subset for spoiler detection"* and "Extracted fuzzy book genres"*. The review dataset contains 1.3 million total book reviews about 25,000 books from 19,000 users. The genre dataset labels books as "romance", "young-adult", "fantasy", "paranormal", "mystery, thriller, crime", "historical, historical fiction, biography", "comics, graphic", "children", "non-fiction", "poetry". Because of genre overlap (i.e. a novel may be romance and fiction) the extracted genres are not perfectly exclusive and we choose to exclude the fiction genre which is a super-set of most of the genres.

Ratings matrix. For all the reviews, we create an indicator for each of the genres. We get the number of reviews a user submits for each genre.[3] This enabled us to construct a ratings matrix containing data on $n \sim 19k$ users and their ratings of $m = 9$ genres. A user's *rating* for a genre describes their level of engagement with books of that genre, as measured by the number of reviews they provide. We also define each user's favorite genre as that which she has reviewed the most, and genre groups as a partition of users according to their favorite genre. So the set of users in the Romance group, for instance, are those who review books labeled as Romance more than any other genre.

Table 2 presents values of genres, ordered by popularity (i.e. total number of reviews), and the number of users in the corresponding genre group.

Majority and minority users. To create groups of majority and minority users and majority and minority items, we order the genres by popularity. Majority genres are the $\bar{n} = 5$ most popular ('romance', 'young-adult', 'fantasy, paranormal', 'mystery, thriller, crime', 'history, historical fiction, biography') and minority genres are the 4 least popular ('comics, graphic', 'children', 'non-fiction', 'poetry'). A reader is a majority (minority) user if their favorite genre is a majority (minority) genre. In total, we have $\bar{m} = 18,870$ majority users which constitute 99.88% of all users, and 22 minority users which constitute 0.12% of all users (implications of this very large gap are discussed in section 4.1.1).

Singular values. The singular values of the ratings matrix are presented in Figure 2. Notice that the first singular value is significantly larger than all others. Thus an α learner within the large range of $(\sigma_2(\mathbf{R}), \sigma_1(\mathbf{R})]$ (which is about $4,950$ to $22,450$) would conclude that the he optimally approximates the full ratings matrix with only the first principle component. For all baseline social welfare calculations (i.e. before altruistic behavior) we will assume that the learner's α is within this range. Equivalently, under the true ratings matrix, we assume that the learner picks an α such that $k^\star = 1$.

[3] Because some books are labeled under multiple genres, this means if user, u reviews a romance poetry book she has done a review of the romance genre and a review of the poetry genre.

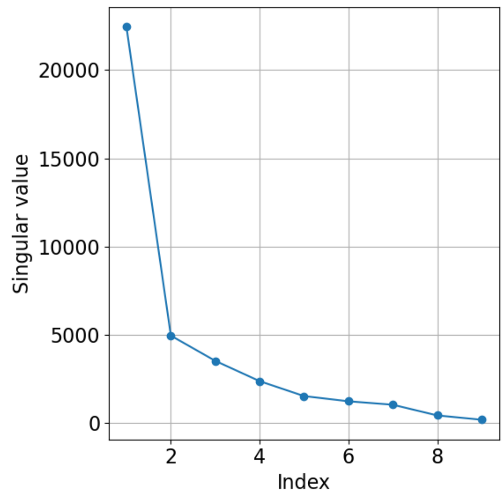


Figure 2: Singular values for Goodreads ratings matrix. The sharp decline from the first to second singular values supports a rank-1 approximation strategy for the learner.

E.1.2 Methodology

EA behavior. Altruistic behavior is modeled as a subset of majority users uprating (i.e writing a review) one of the minority genres. We example different fractions of users whose favorite genre is romance (the most popular), uprating one of three minority genres with a value equal to $\frac{2}{3}$ of their rating for romance. While this is a slight departure from collective uprating of a fixed value η, it is more realistic in this setting given that users have very different levels of engagement on Goodreads and therefore are likely to be altruistic in different capacities. We test $\frac{1}{2}$ or $\frac{1}{3}$ of romance readers for three minority genres–children, comics-graphic, or non-fiction. We exclude uprating of poetry since it is not the favorite genre for any user.

Learner's protocol. Following our theoretical model, the learner is an α-loss tolerant learner. The learner picks a $k^{\star}$ based on the size of the singular values of the reported ratings matrix. He then does a rank $k^{\star}$ approximation of the reported ratings matrix, and top-1 genre recommendation based on the low-rank approximation. For each altruistic behavior tested, we report the α range necessary to induce the increases in social welfare.

Social welfare. The utility of a user is their true rating for the genre that they are recommended. The social welfare is the sum of utilities for all users. We can use this to calculate social welfare improvement: ρ, the social welfare with EA behavior divided by the social welfare without EA behavior.

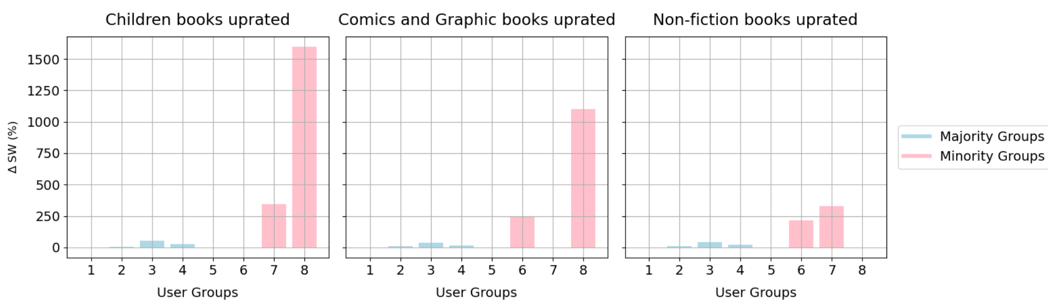
Table 3: Genres, index, and majority-minority classification. This table shows the genres, it's corresponding index, and whether it is classified as a "majority" or "minority" genre. Poetry is excluded since no users have poetry as their favorite genre.

Genre index	Genre	Majority-minority classification
1	romance	majority
2	young-adult	majority
3	fantasy, paranormal	majority
4	mystery, thriller, crime	majority
5	history, historical fiction, biography	majority
6	children	minority
7	comics, graphic	minority
8	non-fiction	minority

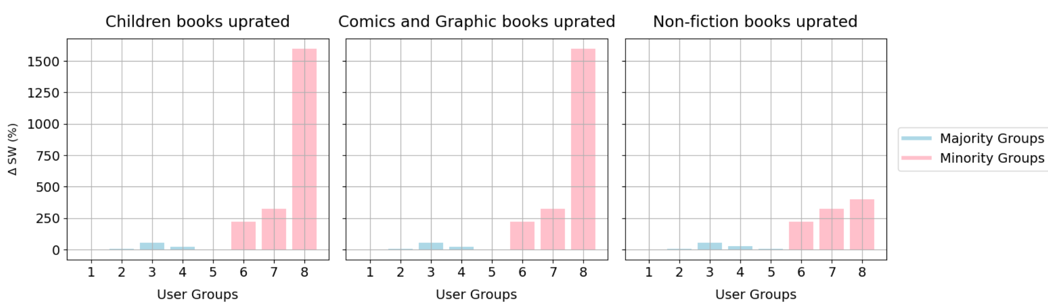
E.1.3 Additional results

For completeness, we will present the full results for both $1/2$ and $1/3$ users uprating. For both sizes of EA user groups, we see significant welfare improvements, particularly for the minority user groups.

Interestingly, it is not always the uprated group that is helped the most. When $1/2$ of romance users uprate minority genres, there is little change in the welfare for that genre, however, EA behavior still positively impacts the other minority groups. Conversely, when $1/3$ of romance users uprate the minority items, we see significant welfare improvements across all minority genre groups (including the one uprated).



(a) $1/2$ Romance users uprating



(b) $1/3$ Romance users uprating

Figure 3: Welfare improvements by user group. See Table 3 for the favorite genre groups indices correspond to.

In Table 4, we detail the range that α must in be for each EA social welfare improvement result we get on the dataset. The *row* denotes the fraction of romance users who uprate a minority genre and the *column* denotes which minority genre is uprated.

Table 4: $(\eta, \mathcal{U}_\mathrm{A})$-Singular value gaps. These are the ranges that α must be in for each EA strategy.

Fraction of Romance Users	Children	Comics & Graphic	Non-fiction
$1/2$	$(4948, 5447)$	$(4948, 5609)$	$(4948, 5797)$
$1/3$	$(4948, 4976)$	$(4948, 5039)$	$(4948, 5185)$

Table 4 represents the ranges of learner's information loss tolerances that would induce the SW results we present. Recall that information loss tolerance is a constraint on singular values (Definition 2.1). 4948 is $\sigma_2(\mathbf{R}^\star)$ (Data information in Appendix E.1.1 and Figure 2) and the second value of each range is $\sigma_2(\tilde{\mathbf{R}})$. Thus, were the learner in each range: (1) under truthful reporting he would have provided recommendations using a rank-1 approximation matrix, (2) under altruism there's enough variation in the 2nd principal component such that he instead is "convinced" to use a rank-2 matrix.

Table 5: Percent total social welfare improvement for different EA strategies. This table shows the percent increase in welfare for *all* users under different fractions of Romance users uprating (rows) and different genres being uprated (columns). In all cases, the total welfare improvement is between 8% and 10%, corresponding to $\rho > 1$.

Fraction of romance users	Children	Comics & Graphic	Non-fiction
1/2	9.4%	7.9%	9.3%
1/3	9.5%	9.5%	9.8%

E.1.4 Limitations

While we are able to see welfare improvements, there are a some limitations to this empirical study. Firstly, there are very few minority users. It would be more interesting and meaningful if minority users constituted a greater fraction of total readers, as this would translate to greater social welfare improvements and hence greater incentives to participate in EA. A counter to this is that the benefits are essentially for free (no groups are meaningfully hurt by EA behavior), and that helping even a small number of users is important. Secondly, the singular value gap is rather small given the size of the singular values. Hence, it may be unrealistic to assume that the learner follows this truncation protocol. This is potentially a limitation of the particular dataset that we have chosen. Thirdly, while this altruistic behavior is similar to the real-world movement we model, the groupings we have defined are not the same. On BookTok, groups of users have been documented to intentionally increase engagement with books from with marginalized author groups such as black or indigenous authors[52] rather than genres. We did not have the data to test this behavior perfectly.

E.2 Survey

E.2.1 Additional results

Survey time took participants an average of 8 minutes and 33 seconds.

Table 6: Number of participants (out of 100) and response to algorithmic impact questions.

Do your interactions affect...	Yes	No	Unsure
your own future recommendations?	92	6	2
other people's recommendations?	57	15	28

Table 7: Number of participants (out of 100) and response to strategic interaction questions.

Have you intentionally...	Yes	No	Unsure
interacted to affect **your** recommendations?	68	27	5
avoided interacting to affect **your** recommendations?	62	30	8
interacted to affect **others'** recommendations?	20	79	1
avoided interacting to affect **others'** recommendations?	20	75	5

E.2.2 Additional example quotes

Intentionally interacted w/ boosting purpose. *"I just do this to support creators and help them grow"*

Intentionally interacted w/ charity purpose.

1. *"I remember seeing a woman on TikTok who was raising money for a personal cause through GoFundMe, and she asked for support in getting her video on more people's For You pages. I intentionally interacted with her post by liking, sharing, and leaving supportive comments. I also saved the video to my collection to boost its visibility."*

2. *"I saw a woman who was fighting for custody of her daughter so I purposely liked it, commented, and shared the video to hopefully get the Tiktok algorithm to push that video out."*

Avoided interaction w/ political purpose. *"Anything political, I 100% REFUSE to interact with anything political other than to ad it to my filters or block because engaging in politics is just too dangerous."*

Avoided interaction w/ misinformation/harm purpose. *"I once avoided liking or commenting on a sensational news post on Facebook because I didn't want to boost its visibility or contribute to spreading misinformation. I knew that interacting with it would make it more likely to appear in others' feeds. By ignoring it, I hoped the platform's algorithm would deprioritize it for others as well."*

Avoided interaction w/ privacy purpose. *"I do not want people to see what I am interested in in my mental health feeds"*

E.2.3 Methodology

Our survey was IRB-exempt as an online survey to adults in the US.

We ran our survey to 100 US-based Prolific users on May 7th, 2025. Prior to the finalized version, we ran two pilot studies each of 5 users (all of whom excluded from the final study) to ensure questions and format was understandable. Each participant was compensated \$2.70. Participants were pre-screened to ensure residence in the United States, a Prolific approval rate ≥ 95, and a Prolific join date no later than Sept. 1st 2024.

Survey questions were divided into 5 sections: 1. Demographics, 2. Recommender System Use, 3. Self-interested Strategization, 4. Altruistic Strategization, and 5. Fairness Beliefs about Recommender Systems. The order in which participants received sections 3 and 4 were randomized. A full list of questions can be found in the appendix (E.2.5). To understand users' knowledge and theories about concepts such as collaborative filtering, we asked participants whether they believed their interactions with content affect their own and others' recommendations and how much. To understand whether users use this knowledge in order to interact strategically, we asked participants whether they ever *intentionally* interact(avoid) content with the purpose of increasing(decreasing) its recommendation to themselves/others. To understand whether any strategic behavior may be driven by altruistic beliefs, we ask participants whether they believe accuracy of recommendations and promotion of content is fair across different user groups. We manually add theme/topic codes to textual responses. Authors manually added $\{0, 1\}$ codes to indicate textual responses that mention boosting/promoting specific creators, charity, politics, harmful/misinformative content, and privacy.

E.2.4 Prolific details

Study label: Survey
Study name: Users' Interactions with Content Recommendation Algorithms
Study Description: This MIT research survey is a part of an experiment to understand people's interactions with algorithms on social media, streaming, and music platforms. You will be asked about your behavior and underlying reasoning when engaging with content recommended to you by these platforms' algorithms.

E.2.5 Online survey questions

The survey consists of five blocks of questions:

1. The first block elicited demographic characteristics: age, employment status, education level, race, state of residence, ethnicity, and gender.
2. The second block asks questions about recommendation system use. We broke down recommender systems into three types: social media (e.g., Twitter, Instagram), streaming (e.g., Netflix, Hulu), and music platforms (e.g., Spotify, Sound Cloud). For each type, we asked (1) what specific platforms respondents used in the last week, and (2) how often they currently use those platforms.
3. The third block pertains to self-interested strategization on recommender systems. First, we elicit information about *awareness*: do respondents believe that their interactions with platform contents impact their future recommendations both in general and for specific types

of interactions (e.g., likes, comments, or subscriptions). Second, we ask about *strategic behavior*: have respondents have ever intentionally interacted (or avoided interacting) with content in order to impact their future recommendations. If they answer yes, we ask for frequency of this behavior and for them to provide an example.

4. The fourth block mirrors the previous block but asks about *altruistic* strategization. We first ask about awareness: do respondents believe that their interactions with platform contents impact *other people's* future recommendations both in general and for specific types of interactions. Second, we ask about strategic behavior: have respondents have ever intentionally interacted (or avoided interacting) with content in order to impact future recommendations for *other people*. If they answer yes, we ask for frequency of this behavior and for them to provide an example.

5. The last block asks about people's beliefs on the accuracy and fairness of recommender systems. We ask if they think (a) the accuracy of content recommendations and (b) the amount of content promotion is fair or not fair for different types of users, and if they think that companies should undertake efforts to increase fairness across users.

Below are the exact questions on the survey including survey logic.

Recommendation Survey

Start of Block: Intro

Q32 This MIT research survey is a part of an experiment to understand people's interactions with algorithms on social media, streaming, and music platforms. You will be asked about your behavior and underlying reasoning when interacting with the content recommended to you by these platforms' algorithms.   You understand that no personally identifiable information provided by you during the research will be disclosed to others without your written permission, except if necessary to protect your rights or welfare, or if required by law.   You understand that your answers should be honest and original descriptions of your experience with using online platforms.

○ yes (1)

○ no (2)

End of Block: Intro

Start of Block: Demographics

Q36 What is your Prolific ID?

____

*

Q34 What is your age?

____

Q35 Which of the following best describes your employment status?

- student (1)
- employed (part-time) (2)
- employed (full-time) (3)
- retired (4)
- unemployed (5)
- other (6) ____

Q36 What is the highest level of education you have completed?

- less then high-school (1)
- high-school (2)
- some college (3)
- college (4)
- graduate degree (5)

Q42 Choose one or more races that you consider yourself to be

- ▢ White or Caucasian (1)
- ▢ Black or African American (2)
- ▢ American Indian/Native American or Alaska Native (3)
- ▢ Asian (4)
- ▢ Native Hawaiian or Other Pacific Islander (5)
- ▢ Other (6)
- ▢ Prefer not to say (7)

Q43 In which state do you currently reside?

▼ Alabama (1) ... I do not reside in the United States (53)

Q47 What is your ethnicity? (select all that apply)

- ▢ German (1)
- ▢ British (2)
- ▢ French (3)
- ▢ Mexican (4)
- ▢ Puerto Rican (5)
- ▢ Cuban (6)
- ▢ African American (7)
- ▢ Haitian (8)
- ▢ Nigerian (9)
- ▢ Chinese (10)
- ▢ Indian (11)
- ▢ Filipino (12)
- ▢ Lebanese (13)
- ▢ Iranian (14)
- ▢ Egyptian (15)
- ▢ Native Hawaiian (16)

- [ ] Samoan (17)
- [ ] Fijian (18)
- [ ] Something Else (please specify) (19) ____
- [ ] Prefer not to say (20)

Q31 How do you describe yourself?

- ○ Male (1)
- ○ Female (2)
- ○ Non-binary / third gender (3)
- ○ Prefer to self-describe (4) ____
- ○ Prefer not to say (5)

End of Block: Demographics

Start of Block: Data-gathering on recommendation system use

Q28 Which of the following social media platforms did you use in the last week? (select all that apply)

- ☐ Facebook (1)
- ☐ Instagram (2)
- ☐ TikTok (3)
- ☐ X/Twitter (4)
- ☐ Tumblr (5)
- ☐ Threads (7)
- ☐ Truth Social (8)
- ☐ other (6) ____
- ☐ None of the above (9)

Q26 How often do you currently use social media platforms?

- ○ never (1)
- ○ once a week (2)
- ○ many times a week (3)
- ○ once a day (4)
- ○ many times a day (5)

Q1 Which of the following streaming platforms did you use in the last week? (select all that apply)

- ▢ Youtube (1)
- ▢ Disney+ (2)
- ▢ Twitch (3)
- ▢ Netflix (4)
- ▢ Hulu (5)
- ▢ Paramount+ (6)
- ▢ Amazon Prime Video (7)
- ▢ other (8) ____
- ▢ None of the above (9)

Q30 How often do you currently use streaming platforms?

- ○ never (1)
- ○ once a week (2)
- ○ many times a week (3)
- ○ once a day (4)
- ○ many times a day (5)

Q29 Which of the following music platforms did you use in the last week? (select all that apply)

- ☐ Spotify (1)
- ☐ Apple Music (2)
- ☐ Sound Cloud (3)
- ☐ other (4) ____
- ☐ None of the above (5)

Q31 How often do you currently use music platforms?

- ○ never (1)
- ○ once a week (2)
- ○ many times a week (3)
- ○ once a day (4)
- ○ many times a day (5)

End of Block: Data-gathering on recommendation system use

Start of Block: Intro to interactions

Q39 The following questions are about your interactions with content (posts, videos, songs, etc) on social media, streaming, and music platforms. An interaction with an item is a general term to include *any* action. Interactions include, but are not limited to, watching/listening, liking, subscribing, commenting, favoriting, sharing, reviewing, etc.

End of Block: Intro to interactions

Start of Block: Self-Interested Strategization

1180

Q38 Recommendations made to **you** In this section, we will ask about the recommendations **you** get from the platform and your interactions with content.

Page Break

Q2 Do you think that your interactions with a platform's content (i.e. likes, comments, subscriptions, watch times) affect which items are recommended to **you** in the future?

- yes (1)
- no (2)
- unsure (3)

Display this question:

If Do you think that your interactions with a platform's content (i.e. likes, comments, subscription... = yes

Or Do you think that your interactions with a platform's content (i.e. likes, comments, subscription... = unsure

Q5 When you interact with platform content in the following ways, how much do you think it impacts what is recommended to **you** in the future?

	unsure (1)	not at all (2)	somewhat (4)	significantly (5)
You like an item (1)	○	○	○	○
You comment on an item (2)	○	○	○	○
You share an item (5)	○	○	○	○
You watch a video or listen to a song all the way through (3)	○	○	○	○
You watch a video or listen to a song multiple times (4)	○	○	○	○
You subscribe to an item or creator (6)	○	○	○	○

Page Break

Q40 We will now ask about your intentions when interacting with platform content. When you intentionally interact with a piece of content this means that you have interacted in this way with a specific reason or purpose in mind.

Page Break

Q33 Have you ever intentionally interacted with (i.e. liked, commented on, subscribed to, watched, etc) an item with the purpose to make it show up in **your** future recommendations?

- yes (1)
- no (2)
- unsure (3)

Display this question:

If Have you ever intentionally interacted with (i.e. liked, commented on, subscribed to, watched, et... = yes

Q35 How often do you intentionally interact with an item so that it will be recommended to **you**?

- less than once a week (1)
- once a week (2)
- many times a week (3)
- once a day (4)
- many times a day (5)

Display this question:

If Have you ever intentionally interacted with (i.e. liked, commented on, subscribed to, watched, et... = yes

Q33 Please briefly describe a time when you intentionally interacted with an item so that it would be recommended to **you**.

____

____

____

____

____

Q37 Have you ever intentionally *avoided* interacting with (i.e. liking, commenting on, subscribing to, watching, etc) an item with the purpose to make the platform recommend it *less* frequently to **you** in the future?

- yes (1)
- no (2)
- unsure (3)

Display this question:

If Have you ever intentionally avoided interacting with (i.e. liking, commenting on, subscribing to,... = yes

Q36 How often do you intentionally *avoid* interacting with an item so that it will be recommended *less* frequently to **you**?

- less than once a week (1)
- once a week (2)
- many times a week (3)
- once a day (4)
- many times a day (5)

Display this question:

If Have you ever intentionally avoided interacting with (i.e. liking, commenting on, subscribing to,... = yes

Q34 Please briefly describe a time when you intentionally *avoided* interacting with an item so that that it would be recommended *less* frequently to **you**.

____

____

____

____

____

Q46 This question is an attention check to ensure data quality. Please select "once a week" and thank you for your attention.

- ○ less than once a week (1)
- ○ once a week (2)
- ○ many times a week (3)
- ○ once a day (4)
- ○ many times a day (5)

End of Block: Self-Interested Strategization

Start of Block: EA Strategization

Q37 Recommendations made to **other people** In this section, we will ask about the recommendations **other people** get from the platform and your interactions with content.

Page Break

Q17 Do you think that your interactions with a platform's content (i.e. likes, comments, subscriptions, watch times, etc.) affect which items are recommended to **other people** in the future?

- yes (1)
- no (2)
- unsure (3)

Display this question:

If Do you think that your interactions with a platform's content (i.e. likes, comments, subscription... = yes

Or Do you think that your interactions with a platform's content (i.e. likes, comments, subscription... = unsure

Q33 When you interact with platform content in the following ways, how much do you think it impacts what is recommended to **other people** in the future?

	unsure (1)	not at all (2)	somewhat (3)	significantly (4)
You like an item (1)	○	○	○	○
You comment on an item (2)	○	○	○	○
You share an item (5)	○	○	○	○
You watch a video or listen to a song all the way through (3)	○	○	○	○
You watch a video or listen to a song multiple times (4)	○	○	○	○
You subscribe to an item or creator (6)	○	○	○	○

1188

Page Break

Q41 We will now ask about your intentions when interacting with platform content. When you intentionally interact with a piece of content this means that you have interacted in this way with a specific reason or purpose in mind.

Page Break

Q34 Have you ever intentionally interacted with (i.e. liked, commented on, subscribed to, watched, etc) an item with the purpose to make it show up in **other people's** future recommendations?

- yes (1)
- no (2)
- unsure (3)

Display this question:

If Have you ever intentionally interacted with (i.e. liked, commented on, subscribed to, watched, et... = yes

Q39 How often do you intentionally interact with an item so that it will be recommended **other people**?

- less than once a week (1)
- once a week (2)
- many times a week (3)
- once a day (4)
- many times a day (5)

Display this question:

If Have you ever intentionally interacted with (i.e. liked, commented on, subscribed to, watched, et... = yes

Q35 Please briefly describe a time when you intentionally interacted with an item so that it would be recommended to **other people**.

____

____

____

____

____

Q38 Have you ever intentionally *avoided* interacting with (i.e. liking, commenting on, subscribing to, watching, etc) an item with the purpose to make the platform recommend it *less* frequently to **other people** in the future?

- yes (1)
- no (2)
- unsure (3)

Display this question:

If Have you ever intentionally avoided interacting with (i.e. liking, commenting on, subscribing to,... = yes

Q40 How often do you intentionally *avoid* interacting with an item so that it will be recommended *less* frequently to **other people**?

- less than once a week (1)
- once a week (2)
- many times a week (3)
- once a day (4)
- many times a day (5)

Display this question:

If Have you ever intentionally avoided interacting with (i.e. liking, commenting on, subscribing to,... = yes

Q36 Please briefly describe a time when you intentionally *avoided* interacting with an item so that it would be recommended *less* frequently to **other people**.

____

____

____

____

____

Q47 This question is an attention check to ensure data quality. Please select "many times a day" and thank you for your attention.

- less than once a week (1)
- once a week (2)
- many times a week (3)
- once a day (4)
- many times a day (5)

End of Block: EA Strategization

Start of Block: Care for other users

Q23 *The following question asks about how **accurate** platforms' content recommendations are. When a platform recommends an item to someone, the accuracy is how well the item matches what the person actually wants to see.* Do you believe the **accuracy** of content recommendations for different types of users (gender, ethnicity, age, etc.) is equal or unequal?

- very unequal (1)
- somewhat unequal (2)
- neutral (3)
- somewhat equal (4)
- very equal (5)

Q41 *The following question asks about content **promotion.** The amount of promotion an item gets is how often a platform recommends it to others.* Do you believe the amount platforms **promote** different types of users' (gender, ethnicity, age, etc.) content is fair or not fair?

- very unfair (1)
- unfair (2)
- neutral (3)
- fair (4)
- very fair (5)

Q25 Do you support or oppose the idea that social media and streaming companies should undertake efforts to increase the fairness of their platforms?

- support (1)
- oppose (2)
- unsure (3)

End of Block: Care for other users

F Supplementary material for Section 5

There are several interesting lines of future work that may loosen assumptions we have made or expand upon our study.

"Evil" Altruism Following the classic economics lens[5], our altruistic agents care about the welfare of other agent(s). However, we intentionally consider only altruistic strategies that also do not decrease the welfare any agent would have received otherwise. This is to emphasize that as long as users are even minimally altruistic, the recommender system *naturally incentivizes* them to do this algorithmic collective action. There are many other directions in which future work could study collective altruistic rating strategies. In particular, altruism does not have to improve social welfare. Altruistic users could desire to improve the recommendation of a small group and not care about *all* others. One could model collectives purposefully promoting content that appeals to the small group of users, but is offensive (and would cause negative utility if recommended) to others.

Cost to Altruistic Actions Our survey and other literature [18, 52, 45] indicate that real-world participation in the collective action we study theoretically often consists of intentional likes, shares, or supportive comments. Supportive comments are often simple or generic, such as "Oh no, it looks like I've accidentally commented for the algorithm."[45], not including a direct reference to the information in the content itself. Unlike traditional protests where physical presence is required, this action requires less overt effort. As such, we model uprating with the implicit assumption that it comes at no effort cost. That said, future work may reconsider this modeling decision. Sophisticated recommender systems may be trained to ignore overly simple strategies such as bot-like comments, thus requiring higher effort from altruistic users. Additionally, uprating via comment is a public display of support, so there may be interesting reputational costs to account for.

Alternative Learner Utilities While we also show user EA strategies yield the learner higher welfare when he follows the protocol and evaluates his welfare according to engagement or benevolent utility, two remarks are in order: (1) the learner welfare achieved is not necessarily *optimal*, and (2) there are many other possible learner utility functions to consider. For the first point, further work could analyze this setting as a multi-agent Stackelberg game in which some agents have altruistic utilities and the principal explicitly maximizes benevolent, engagement, or an alternative utility. For the second point, a natural alternative in which altruism hurts the learner is one that explicitly penalizes dimensionality. Recall that altruism causes the optimal rank reduction for our α-loss tolerant learner to increase from $\bar{n}$ to $\bar{n}+1$. On a large scale, it would thus be bad for a dimensionality-sensitive learner to incentivize altruism for many different groups and items.

G Generalized classes of matrices and rating strategies

In this subsection, we will derive analogous social welfare results for a more complicated class of preference matrices and altruistic strategies than those which are in the main body of this manuscript. This preference matrix class will not be limited to majority minority groups with exclusive preferences and strategic altruism is not limited to uprating. Because of these complexities, the proof techniques necessary will be significantly different than in the main body, but the majority of results will have analogies. Additionally, we have not been able to provide a simple algorithm for the computation of an altruistic strategy as we do in the simpler case.

G.1 $\mathcal{M}$: A class of popularity gap matrices

Consider a tuple: $(\mathbf{R}, \bar{n})$ where: $\mathbf{R} \in [0,1]^{m\times n}$ is a [normalized] preference matrix and $\bar{n}$ is some integer value $0 < \bar{n} < n$. Call the first $\bar{n}$ items (columns) of $\mathbf{R}$ the *popular* items and the remaining, the *unpopular* items. Define $\kappa_{(\mathbf{R},\bar{n})} \in \mathbb{R}_{\geq 0}$ to be $\max_{i' \in \{(\bar{n}+1),\ldots,n\}} \|\mathbf{R}_{i'}\|_1$, the greatest L-1 norm for any unpopular item. The $\mathbf{R}$ matrix but with the preferences for unpopular items zeroed out will be important for the remainder of our analysis, so we define this as follows.

Definition G.1 (Popular Preferences Matrix, $\mathbf{R}'(\bar{n})$) *Let $\mathbf{R}'(\bar{n}) \in [0,1]^{m\times n}$ be a matrix s.t.:*

$$r'_{u,i} = \begin{cases} r_{u,i}, & \textit{if } i \leq \bar{n} \\ 0, & \textit{otherwise} \end{cases}$$

Thus $\mathbf{R}'(\bar{n})$ is a block matrix where $\mathbf{A}(\bar{n}) \in [0,1]^{m\times\bar{n}}$ is the popular item block of $\mathbf{R}$.:

$$\mathbf{R}'(\bar{n}) = \begin{pmatrix} \mathbf{A}(\bar{n}) & \mathbf{0}^{m\times(n-\bar{n})} \end{pmatrix} \tag{6}$$

Because the learner is interested in recovering a best item for each user, we shall define a set of top item(s) for a user u with respect to the matrix, $\mathbf{R}$:

Definition G.2 (User u's Top Item(s)) *Define $\mathcal{I}_{top}(\mathbf{R}, u)$ to be a set of top items for a user u according to preference matrix $\mathbf{R}$:*

$$\mathcal{I}_{top}(\mathbf{R}, u) := \arg\max_{i \in [n]} r_{u,i}$$

Note that $|\mathcal{I}_{top}(\mathbf{R}, u)| \geq 1$.

We shall define user groups based on whether a user's top item(s) is(are) popular or unpopular:

Definition G.3 (Majority User) *A majority user for a particular $\mathbf{R}$ preference matrix and $\bar{n}$ is one who has top rated item $i \in \mathcal{I}_{top}(\mathbf{R}, u)$, such that $i \leq \bar{n}$, meaning i is one of the popular items. Formally, we define the set of majority users for a particular $\mathbf{R}$ preference matrix:*

$$\mathcal{U}_{\text{MAJ}} := \{u : \exists i \in \mathcal{I}_{top}(\mathbf{R}, u) \quad s.t. \quad i \in [\bar{n}]\}$$

Definition G.4 (Minority User) *A minority user for a particular $\mathbf{R}$ preference matrix and $\bar{n}$ is one who has item $i \in \mathcal{I}_{top}(\mathbf{R}, u)$, such that $i > \bar{n}$, meaning i is one of the unpopular items. Formally, we define the set of minority users for a particular $\mathbf{R}$ preference matrix:*

$$\mathcal{U}_{\text{MIN}} := \{u : \exists i \in \mathcal{I}_{top}(\mathbf{R}, u) \quad s.t. \quad i \in \{(\bar{n}+1), \ldots, n\}\}$$

For some results, it will be useful to make assumptions that a tuple, $(\mathbf{R}, \bar{n})$ is such that there exist nonempty user majority/minority groups and they are exclusive.

Assumption G.1 (Minority/Majority User assumptions) *Preference matrix $\mathbf{R}$ and $\bar{n}$ is such that the following is true of majority and minority user groups:*

1. *Majority and Minority user sets are exclusive:*

$$\mathcal{U}_{\text{MIN}} \cap \mathcal{U}_{\text{MAJ}} = \emptyset$$

2. *There is at least one minority user:*

$$\forall m : |\mathcal{U}_{\text{MIN}}| > 0$$

Remark G.1 *Note that the majority/minority exclusivity of assumption G.1 is a weaker assumption than the exclusivity assumption in the main body of the paper as assumption G.1 exclusivity does not imply that majority and minority users' preferences are entirely exclusive, only that their* ***top items*** *are exclusive. Formally:*

$$\forall u \in \mathcal{U}_{\text{MAJ}}, \forall i \in \{\bar{n}+1, \ldots, n\}, \quad i \notin \mathcal{I}_{top}(\mathbf{R}, u)$$
$$\forall u' \in \mathcal{U}_{\text{MIN}}, \forall i' \in [\bar{n}], \quad i' \notin \mathcal{I}_{top}(\mathbf{R}, u')$$

We will now construct a class of preference matrix, $\mathbf{R}$ and popular item index $\bar{n}$ tuples. In order to do this, we define the assumptions that tuples belonging to this class must satisfy. These assumptions will be about the difference between particular ratings of the preference matrix, so before proceeding we define $\Delta(\mathbf{R}, \bar{n})$. This will have a similar function to the singular value gaps of the class of matrices used in the main body in that we will use Δ to impose a gap in popularity between items.

Definition G.5 (Sufficient Ratings Gap, $\Delta(\mathbf{R}, \bar{n})$) *The sufficient ratings gap is a function of the ratings of $\mathbf{R}$ and which items are popular, $\bar{n}$*

$$\Delta(\mathbf{R}, \bar{n}) := \frac{2^{\frac{5}{2}} \kappa_{(\mathbf{R}, \bar{n})} n^{\frac{3}{2}}}{[\sigma_{\bar{n}}(\mathbf{R}'(\bar{n}))]^2} \tag{7}$$

Assumption G.2 (Majority users' top item(s) are sufficiently highly rated) *Majority users don't care about all items equally: $[n] \setminus \mathcal{I}_{top}(\mathbf{R}, u) \neq \emptyset$ and there is a sufficient gap between a majority user's top rating (which may appear on multiple items) and her other ratings:*

$$\max_{i \in [n] \setminus \mathcal{I}_{top}(\mathbf{R}, u)} r_{u,i} < \max_{i \in [n]} r_{u,i} - \Delta(\mathbf{R}, \bar{n}) \quad \forall u \in \mathcal{U}_{\text{MAJ}} \tag{8}$$

Assumption G.3 (Minority users have sufficient preference for a popular item) *Each minority user likes at least one popular item by a sufficient amount:*

$$\exists i \in \bar{n} \quad s.t. \quad r_{u,i} > \Delta(\mathbf{R}, \bar{n}) \quad \forall u \in \mathcal{U}_{\text{MIN}} \tag{9}$$

Note that because assumption G.3 states that each minority user likes at least one popular item by some small amount, the majority-minority matrices discussed in the main body cannot satisfy this assumption as they are block matrices that impose complete exclusivity in preference. Thus, while the class we will construct here does not technically include matrices analyzed in the main body, the appendix class can be viewed as "more general" because it handles the settings in which preferences matrices do not have the $\mathbf{R}_{\text{MIN}}, \mathbf{R}_{\text{MAJ}}$ block structure that creates exclusivity between all items a majority user likes and all items a minority user likes.

One may consider assumption G.3 as reminiscent of a non-zero support assumption for minority users' preference over popular items while the main body imposes zero-support over the same space.

We can now define the key class of preference matrix and popular item index tuples that we will use for the remainder of the results in this section.

Definition G.6 (Popularity Gap Class, $\mathcal{M}$) *The following is an important class of tuples where the popular items (whose indices lies in $[\bar{n}]$) are sufficiently more highly rated by a variety of users than the unpopular items:*

$$\mathcal{M} := \{(\mathbf{R}, \bar{n}) : \textit{Assumption G.2, G.3 hold.}\} \tag{10}$$

Remark G.2 (The Meaning of Popularity) *We note that for any valid $\mathbf{R} \in [0,1]^{m \times n}$ and $\bar{n}$ where assumptions G.2 and G.3 are true, it must be the case that the following is true:*

$$2^{\frac{5}{4}} n^{\frac{3}{4}} \sqrt{\kappa_{(\mathbf{R},\bar{n})}} < \sigma_{\bar{n}}(\mathbf{R}'(\bar{n})) \tag{11}$$

Intuitively, equation 11 means that popular items (those in $[\bar{n}]$) are sufficiently well-liked by enough users such that their associated singular values are big relative to the magnitude of minority items' ratings (whose l1-norms are upper bounded by $\kappa_{(\mathbf{R},\bar{n})}$).

We call $\mathcal{M}$ the "Popularity Gap Class" following the intuition detailed in remark G.2. That is, in order for the assumptions to be potentially satisfied, it must be the case that the singular value of the associated Popular Preferences Matrix, $\mathbf{R}'(\bar{n})$, dominates over the a function of the unpopular ratings. Much like the singular value gap assumption of the main body, ensuring that equation 11 holds essentially insures that the items labeled as popular by $\bar{n}$ are actually mathematically popular.

G.2 Learner's selection of optimal truncation rank

In the section above, we present a class of $(\mathbf{R}, \bar{n})$. However, the α-loss tolerant learner must approximate the received preference matrix, $\widetilde{\mathbf{R}}$, to an optimal rank less than or equal to $\operatorname{rank}(\widetilde{\mathbf{R}})$. Given that the tuple $(\widetilde{\mathbf{R}}, \bar{n}) \in \mathcal{M}$, in this section we will show that $k^* = \bar{n}$ for learners whose α total variance loss budget is within a particular range.

G.3 Preliminary: useful singular value bounds on $\mathbf{R}$ if $(\mathbf{R}, \bar{n}) \in \mathcal{M}$

Recall that α-loss tolerant learners are defined in terms of how large the next singular value after truncation would be. Thus, it will be useful to have bounds on the singular values of $\mathbf{R}$ and $\mathbf{R}'(\bar{n})$.

Corollary G.1 (Corollary of Lemma D.2) *Let matrix $\tilde{\mathbf{A}} \in \mathbb{C}^{m \times (n-j)}$ where $j < n$. Define $\sigma_r(\hat{\mathbf{A}}) = 0$ for singular values lost to column deletion. Then the following relation of singular values holds:*

$$\sigma_i(\mathbf{A}) \geq \sigma_i(\tilde{\mathbf{A}}) \geq \sigma_{i+j}(\mathbf{A})$$

Proof. This is easily seen by induction on j using the Horn and Johnson lemma as a $j = 1$ base case. □

Proposition G.1 ($\bar{n}$ and $\bar{n}+1$ singular value bounds) *If a tuple $(\mathbf{R}, \bar{n}) \in \mathcal{M}$ then the singular values of $\mathbf{R}$ satisfy the following relations:*

$$\sigma_{\bar{n}}(\mathbf{R}) \geq 2^{\frac{5}{4}} n^{\frac{3}{4}} \sqrt{\kappa_{(\mathbf{R},\bar{n})}}$$

$$\sigma_{\bar{n}+1}(\mathbf{R}) \leq \sqrt{(n-\bar{n})\kappa_{(\mathbf{R},\bar{n})}}$$

Proof. The first inequality comes from the fact that assumptions G.2 and G.3 hold. Clearly $\frac{2^{\frac{5}{2}}\kappa_{(\mathbf{R},\bar{n})}n^{\frac{3}{2}}}{[\sigma_{\bar{n}}(\mathbf{R}'(\bar{n}))]^2} \leq 1$ if assumption G.2 is true because otherwise the minimum difference between a top items and next ratings is greater than what $\mathbf{R} \in [0,1]^{m \times n}$ allows. This yields $\sigma_{\bar{n}}(\mathbf{R}'(\bar{n})) \geq 2^{\frac{5}{4}} n^{\frac{3}{4}} \sqrt{\kappa_{(\mathbf{R},\bar{n})}}$ Because $\mathbf{R}'(\bar{n})$ is the same as $\mathbf{R}$ with the unpopular columns removed (and replaced with zeros, which does not affect singular values) we can invoke Corollary G.1 to get the desired inequality in terms of $\mathbf{R}$.

Now we will show the second inequality. Define matrix $\mathbf{B} \in [0,1]^{m \times (n-\bar{n})}$ to be matrix $\mathbf{R}$ but where popular item columns have been removed. We have the following:

$$\begin{aligned} \sigma_{1+\bar{n}}(\mathbf{R}) &\leq \sigma_1(\mathbf{B}) \quad \text{Corollary G.1} \\ &= ||\mathbf{B}||_2 \quad \text{Def of spectral norm} \\ &\leq ||\mathbf{B}||_F \quad \text{Matrix Norm Equivalences} \\ &\leq \sqrt{(n-\bar{n})\kappa_{(\mathbf{R},\bar{n})}} \end{aligned}$$

To get the last inequality, recall that $||\mathbf{X}||_F = \sqrt{\mathrm{tr}(\mathbf{X}^\top \mathbf{X})}$ and note that

$$(\mathbf{B}^\top \mathbf{B})_{i,j} \leq \max_{i \in [n-\bar{n}]} \mathbf{B_i}^\top \mathbf{1} = \max_{i \in [n-\bar{n}]} ||\mathbf{B_i}||_1 = \kappa_{(\mathbf{R},\bar{n})}$$

where $\mathbf{B_i}$ is the ith column vector of $\mathbf{B}$. $\mathrm{tr}(\mathbf{X}^\top \mathbf{X})$ must thus be upper bounded by $(n-\bar{n})\kappa_{(\mathbf{R},\bar{n})}$ because there are $n-\bar{n}$ diagonal elements of $\mathbf{B}^\top\mathbf{B}$ each upper bounded by $\kappa_{(\mathbf{R},\bar{n})}$. ☐

G.3.1 A learner whose $k^* = \bar{n}$

Intuitively, we have shown in the above preliminaries that if $(\mathbf{R},\bar{n}) \in \mathcal{M}$, then the $\bar{n}$-th singular value of $\mathbf{R}$ must be relatively big while the next singular values must be quite small. This should mean that retaining singular values 1 through $\bar{n}$ is "important" while the remainder of the singular values do not contribute very much.

Definition G.7 **($\bar{n}$, R-Singular Value Gap)** *For any $\mathbf{R} \in [0,1]^{m \times n}$ and $\bar{n} \in [n]$ s.t. $(\mathbf{R},\bar{n}) \in \mathcal{M}$, define the space*

$$\mathcal{G}(\bar{n},\mathbf{R}) := \{y \in \mathbb{R} : y \in \left(\sqrt{(n-\bar{n})\kappa_{(\mathbf{R},\bar{n})}}, 2^{\frac{5}{4}} n^{\frac{3}{4}} \sqrt{\kappa_{(\mathbf{R},\bar{n})}}\right)\}$$

Much like in the main body of the paper, a learner whose α parameter falls into this gap will select to truncate exactly to dimension $\bar{n}$. Formally:

Proposition G.2 **($k^\star = \bar{n}$ for the α-loss tolerant learner)** *For all $\widetilde{\mathbf{R}} \in [0,1]^{m \times n}$ such that $(\widetilde{\mathbf{R}},\bar{n}) \in \mathcal{M}$, If the α loss tolerant learner is such that $\alpha \in \mathcal{G}(\bar{n},\widetilde{\mathbf{R}})$, then it must be the case that $k^* = \bar{n}$.*

Proof. By proposition G.1, $\sigma_{\bar{n}} > \alpha$ while $\sigma_{\bar{n}+1} < \alpha$. By properties of singular values, $\sigma_j \geq \sigma_{\bar{n}} \forall j \leq \bar{n}$, thus $\bar{n}$ is the minimum k such that $\sigma_{k+1} < \alpha$ ☐

Remark G.3 **($\mathcal{G}(\bar{n},\mathbf{R})$ is nonempty)** *Note that for any $\bar{n} \geq 1$, $2^{\frac{5}{4}} n^{\frac{3}{4}} > \sqrt{n-\bar{n}}$ therefore the space $\mathcal{G}(\bar{n},\widetilde{\mathbf{R}})$ is not empty for any reasonable tuple.*

Obviously, this range will limit the type of learners we discuss, however it is important to note that it is always nonempty (see remark G.3) and for large n, this range for α is also very big, therefore this is a non negligible space of general α-loss tolerant learners.

G.4 Recommendations and top-1 social welfare when $\widetilde{\mathbf{R}}$ s.t. $(\widetilde{\mathbf{R}},\bar{n}) \in \mathcal{M}$

We will derive the recommendations made and resulting social welfare when the received preference matrix is such that $(\widetilde{\mathbf{R}},\bar{n}) \in \mathcal{M}$ and the α-loss tolerant learner is parametrized such that $k^\star = \bar{n}$.

G.4.1 Preliminary: SVD truncation error bounds

First we will remind the reader of an equivalent representation of truncated SVD and derive a useful lemma to upper bound the approximation error.

Recall that for a preference matrix, $\mathbf{R} \in [0,1]^{m\times n}$, a $k^\star$-truncated SVD approximation is equivalent to solving the following optimization problem:

$$\begin{aligned} \text{minimize}_{\mathbf{\Pi}\in\mathbb{R}^{n\times n}} \quad & \|\mathbf{R} - \mathbf{R}\mathbf{\Pi}\|_F^2 \\ \text{subject to} \quad & \mathbf{\Pi} = \mathbf{U}\mathbf{U}^\top \\ & \mathbf{U} \in \mathbb{R}^{n\times k^\star} \\ & \mathbf{U}^\top\mathbf{U} = \mathbf{I}_{k^\star} \end{aligned} \tag{12}$$

Where $\mathbf{I}_{k^\star}$ is a $k^\star$-dimensional identity and clearly $\mathbf{\Pi}$ is a projection matrix.

We can define $\widehat{\mathbf{R}} = \mathbf{R}\mathbf{\Pi}^\star$ where $\mathbf{\Pi}^\star$ is the minimizer and this $\widehat{\mathbf{R}}$ is equivalent to the $k^\star$-truncated SVD. We can derive a bound on how close the optimal projection matrix, $\mathbf{\Pi}^\star$, is to $\mathbf{I}_{n,\bar{n}}$, a "partial" identity matrix where only the first $\bar{n}$ diagonal elements are 1s. Functionally, because $\widehat{\mathbf{R}} = \mathbf{R}\mathbf{\Pi}^\star$, this will be an upper bound on how close $\widehat{\mathbf{R}}$ is to just being the first $\bar{n}$ columns of $\mathbf{R}$ with the remaining columns zeroed out. We note that this bound (and its proof) is a version of Theorem 1 from [51].

Proposition G.3 *Let $\mathbf{\Pi}^*_{\bar{n}} \in \mathbb{R}^{n\times n}$ be the optimal projection operator of $\mathbf{R}$ to its $\bar{n}$-truncated SVD. Assume that $\sigma_{\bar{n}}(\mathbf{R}'(\bar{n})) > 0$. We have the following:*

$$||\mathbf{\Pi}^*_{\bar{n}} - \mathbf{I}_{n,\bar{n}}||_F \leq \frac{\Delta(\mathbf{R},\bar{n})}{2\sqrt{n}}, \tag{13}$$

where $\mathbf{I}_{n,\bar{n}}$ is a $n\times n$ matrix where the first $\bar{n}$ diagonal entries are 1 and all other entries are 0.

In order to show this, we will invoke a well-known matrix theory result that we define as a lemma and prove for completeness below. As a preliminary, recall from matrix analysis that between two matrices, we may compare their subspaces using *principal angles*. In particular, between two matrices, $\mathbf{U}, \mathbf{U}' \in \mathbb{R}^{m\times\bar{n}}$ made up of orthonormal columns, the vector of principal angles is $\mathbf{d} := (\cos^{-1}\sigma_1, \ldots \cos^{-1}\sigma_{\bar{n}})$ where σ_i is the ith singular value of $\mathbf{U}^\top\mathbf{U}'$. We will denote $\sin\Theta(\mathbf{U},\mathbf{U}') := \mathrm{diag}(\mathbf{d})$.

Lemma G.1 *Let $\mathbf{U}, \mathbf{U}' \in \mathbb{R}^{m\times\bar{n}}$ be matrices with orthonormal columns.*

$$\|\sin\Theta(\mathbf{U},\mathbf{U}')\|_F = \frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_F.$$

Proof. Let $\mathbf{\Pi} := \mathbf{U}\mathbf{U}^\top$ and $\mathbf{\Pi}' := \mathbf{U}'\mathbf{U}'^\top$ notice that these are projection matrices.

$$\begin{aligned}
\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_F^2 &= \|\mathbf{\Pi} - \mathbf{\Pi}'\|_F^2 \\
&= \mathrm{Tr}\left((\mathbf{\Pi} - \mathbf{\Pi}')^\top(\mathbf{\Pi} - \mathbf{\Pi}')\right) && \text{def of Frobenius norm} \\
&= \mathrm{Tr}\left(\mathbf{\Pi}^\top\mathbf{\Pi} + \mathbf{\Pi}'^\top\mathbf{\Pi}' - \mathbf{\Pi}^\top\mathbf{\Pi}' - \mathbf{\Pi}'^\top\mathbf{\Pi}\right) \\
&= \mathrm{Tr}(\mathbf{\Pi}^2) + \mathrm{Tr}(\mathbf{\Pi}'^2) - \mathrm{Tr}(\mathbf{\Pi}\mathbf{\Pi}') - \mathrm{Tr}(\mathbf{\Pi}'\mathbf{\Pi}) && \text{projection symmetric, trace linear} \\
&= \mathrm{Tr}(\mathbf{\Pi}^2) + \mathrm{Tr}(\mathbf{\Pi}'^2) - 2\mathrm{Tr}(\mathbf{\Pi}\mathbf{\Pi}') && \text{trace cyclic} \\
&= \mathrm{Tr}(\mathbf{\Pi}) + \mathrm{Tr}(\mathbf{\Pi}') - 2\mathrm{Tr}(\mathbf{\Pi}\mathbf{\Pi}') && \text{projection idempotent} \\
&= 2\bar{n} - 2\mathrm{Tr}(\mathbf{\Pi}\mathbf{\Pi}') && \text{trace of projection = rank} \\
&= 2\bar{n} - 2\mathrm{Tr}\left((\mathbf{U}^\top\mathbf{U}')^\top(\mathbf{U}^\top\mathbf{U}')\right) \\
&= 2\bar{n} - 2\sum_{i\in[\bar{n}]}\cos^2(d_i) && \sigma_i(\mathbf{U}^\top\mathbf{U}') = \cos(d_i) \\
&= 2\left(\sum_{i\in[\bar{n}]} 1 - \cos^2(d_i)\right) \\
&= 2\left(\sum_{i\in[\bar{n}]} \sin^2(d_i)\right) && \text{trig identity} \\
&= 2\|\sin\Theta(\mathbf{U},\mathbf{U}')\|_F^2
\end{aligned}$$

Taking square root and dividing by $\sqrt{2}$ on both sides gives the desired identity. □

Proof of Proposition G.3. Let $\mathbf{C} = \mathbf{R}^\top \mathbf{R}$ and $\mathbf{C}' = \mathbf{R}'(\bar{n})^\top \mathbf{R}'(\bar{n})$, thus $\mathbf{C}, \mathbf{C}' \in \mathbb{R}^{n \times n}$. Let $\mathbf{U}, \mathbf{U}' \in \mathbb{R}^{n \times \bar{n}}$ be matrices whose columns correspond to the $\bar{n}$ normalized eigenvectors of the $\bar{n}$ largest eigenvalues of $\mathbf{C}, \mathbf{C}'$.

We will complete this proof by going through the following claims:

1. $\mathbf{U}'\mathbf{U}'^\top = \mathbf{I}_{n,\bar{n}}$
2. $\frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}}\|_F = \|\sin \Theta(\mathbf{U}, \mathbf{U}')\|_F$
3. $\|\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}}\|_F \leq \frac{2\sqrt{2}n\kappa}{[\sigma_{\bar{n}}(\mathbf{R}'(\bar{n}))]^2}$

For some notational cleanliness, for this proof, we will refer to $\kappa_{(\mathbf{R},\bar{n})}$ as simply κ.

Claim G.1

$$\mathbf{U}'\mathbf{U}'^\top = \mathbf{I}_{n,\bar{n}}$$

where $\mathbf{I}_{n,\bar{n}}$ is a $n \times n$ matrix where the first $\bar{n}$ diagonal entries are 1 and all other entries are 0.

Proof of Claim G.1.

Notice that because

$$\mathbf{R}'(\bar{n}) = \begin{pmatrix} \mathbf{A}(\bar{n}) & \mathbf{0}^{m \times (n-\bar{n})} \end{pmatrix} \tag{14}$$

We have that:

$$\mathbf{C}' = \begin{pmatrix} \tilde{\mathbf{C}} & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \tag{15}$$

Where $\tilde{\mathbf{C}} \in \mathbb{R}^{\bar{n} \times \bar{n}}$.

Notice that $\tilde{\mathbf{C}}$ is a symmetric $\bar{n} \times \bar{n}$ matrix, thus it has an orthonormal eigenbasis. Consider orthonormal matrix $\mathbf{V}' \in \mathbb{R}^{\bar{n} \times \bar{n}}$ s.t. that it's columns are eigenvectors of $\tilde{\mathbf{C}}$. By orthonormality, $\mathbf{V}'\mathbf{V}'^\top = \mathbf{I}_{\bar{n}}$. Define $\tilde{\mathbf{V}}' \in \mathbb{R}^{n \times \bar{n}}$ such that it is $\mathbf{V}'$ padded with zeros. Clearly, $\tilde{\mathbf{V}}'$ is a normalized matrix of the $\bar{n}$ eigenvectors corresponding to the $\bar{n}$ largest eigenvectors of $\mathbf{C}'$, therefore $\mathbf{U}'$ is simply $\tilde{\mathbf{V}}'$. Thus we can see that $\mathbf{U}'\mathbf{U}'^\top = \mathbf{I}_{n,\bar{n}}$. □

Claim G.2

$$\frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}}\|_F = \|\sin \Theta(\mathbf{U}, \mathbf{U}')\|_F$$

Proof of Claim G.2. By Claim G.1,

$$\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}} = \mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top.$$

Further, since $\mathbf{U}, \mathbf{U}'$ are composed of normalized orthogonal columns, by Lemma G.1, we have that

$$\frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_F = \|\sin \Theta(\mathbf{U}, \mathbf{U}')\|_F.$$

Put together, we have the desired equality:

$$\frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}}\|_F = \frac{1}{\sqrt{2}}\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_F = \|\sin \Theta(\mathbf{U}, \mathbf{U}')\|_F.$$

□

Claim G.3

$$\|\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}}\|_F \leq \frac{2\sqrt{2}n\kappa}{(\sigma_{\bar{n}}(\mathbf{R}'(\bar{n})))^2}$$

Proof of Claim G.3.

From Yu et al [71] Theorem 2 using $r = 1$ and $s = \bar{n}$, we have:

$$||\sin\Theta(\mathbf{U}, \mathbf{U}')||_F \leq \frac{2\min(\sqrt{\bar{n}}||\mathbf{C} - \mathbf{C}'||_{op}, ||\mathbf{C} - \mathbf{C}'||_F)}{\min(\lambda_0 - \lambda_1(\mathbf{C}'), \lambda_{\bar{n}}(\mathbf{C}') - \lambda_{\bar{n}+1})}$$

where $\lambda_0 = \infty$ and $\lambda_{\bar{n}+1} = 0$ by construction. Note that by assumption, $\sigma_{\bar{n}}(\mathbf{R}'(\bar{n})) > 0$, so the denominator is well-defined.

Thus we have:

$$\begin{aligned}
\|\mathbf{U}\mathbf{U}^\top - \mathbf{I}_{n,\bar{n}}\|_F &\leq \frac{2\sqrt{2}\min(\sqrt{\bar{n}}||\mathbf{C} - \mathbf{C}'||_{op}, ||\mathbf{C} - \mathbf{C}'||_F)}{\min(\lambda_0 - \lambda_1(\mathbf{C}'), \lambda_{\bar{n}}(\mathbf{C}') - \lambda_{\bar{n}+1})} && \text{Claim 2}\\
&= \frac{2\sqrt{2}\min(\sqrt{\bar{n}}||\mathbf{C} - \mathbf{C}'||_{op}, ||\mathbf{C} - \mathbf{C}'||_F)}{\min(\infty - \lambda_1(\mathbf{C}'), \lambda_{\bar{n}}(\mathbf{C}'))} && \text{by construction}\\
&\leq \frac{2\sqrt{2}\|\mathbf{C} - \mathbf{C}'\|_F}{\min(\infty - \lambda_1(\mathbf{C}'), \lambda_{\bar{n}}(\mathbf{C}'))}\\
&= \frac{2\sqrt{2}\|\mathbf{C} - \mathbf{C}'\|_F}{\lambda_{\bar{n}}(\mathbf{C}')}\\
&= \frac{2\sqrt{2}\|\mathbf{C} - \mathbf{C}'\|_F}{(\sigma_{\bar{n}}(\mathbf{R}'(\bar{n})))^2} && \text{Def of singular value}\\
&\leq \frac{2\sqrt{2}n\kappa}{(\sigma_{\bar{n}}(\mathbf{R}'(\bar{n})))^2}
\end{aligned}$$

Where the last inequality is as follows. Notice that

$$\begin{aligned}
\|\mathbf{C} - \mathbf{C}'\|_F &= \|\mathbf{R}^\top\mathbf{R} - \mathbf{R}'(\bar{n})^\top\mathbf{R}'(\bar{n})\|_F\\
&= \left\| \begin{pmatrix} \mathbf{A}(\bar{n})^\top\mathbf{A}(\bar{n}) & \mathbf{A}(\bar{n})^\top\mathbf{B}(\bar{n}) \\ \mathbf{B}(\bar{n})^\top\mathbf{A}(\bar{n}) & \mathbf{B}(\bar{n})^\top\mathbf{B}(\bar{n}) \end{pmatrix} - \begin{pmatrix} \mathbf{A}(\bar{n})^\top\mathbf{A}(\bar{n}) & 0 \\ 0 & 0 \end{pmatrix} \right\|_F\\
&= \left\| \begin{pmatrix} 0 & \mathbf{A}(\bar{n})^\top\mathbf{B}(\bar{n}) \\ \mathbf{B}(\bar{n})^\top\mathbf{A}(\bar{n}) & \mathbf{B}(\bar{n})^\top\mathbf{B}(\bar{n}) \end{pmatrix} \right\|_F
\end{aligned}$$

Where $\mathbf{A}(\bar{n}) \in [0,1]^{m\times\bar{n}}$ are the popular columns of the preference matrix and $\mathbf{B}(\bar{n}) \in [0,1]^{m\times(n-\bar{n})}$ are the unpopular columns of the matrix. Thus we can upper bound every element of $\mathbf{C} - \mathbf{C}'$:

$$(\mathbf{C} - \mathbf{C}')_{ij} \leq \max_{i>\bar{n}} \mathbf{R_i}^\top\mathbf{1} = \max_{i>\bar{n}} \|\mathbf{R_i}\|_1 = \kappa.$$

There are n^2 elements in $(\mathbf{C} - \mathbf{C}')$ Therefore

$$\|\mathbf{C} - \mathbf{C}'\|_F \leq \sqrt{n^2\kappa^2} \leq n\kappa.$$

$\square$

To reach the final statement of the theorem, notice that $\mathbf{U}\mathbf{U}^\top$ creates the projection matrix, $\mathbf{\Pi}^\star$ that minimizes the optimization problem 12 and recall that $\Delta(\mathbf{R}, \bar{n}) := \frac{4\sqrt{2}\kappa_{(\mathbf{R},\bar{n})}n\sqrt{n}}{[\sigma_{\bar{n}}(\mathbf{R}'(\bar{n}))]^2}$ $\square$

G.4.2 Recommendations and social welfare bounds

Now that we have some idea of what $\widehat{\mathbf{R}}$ will be from proposition G.3, we can make statements about recommendations and resulting social welfare particular α learners will give when $(\widetilde{\mathbf{R}}, \bar{n}) \in \mathcal{M}$.

Theorem G.4 (Recommendations are good for majority, bad for minority) *Let $\widetilde{\mathbf{R}}$ be a reported preference matrix such that, for some $\bar{n} \in [n]$, $(\widetilde{\mathbf{R}}, \bar{n}) \in \mathcal{M}$. Let there also be an α loss tolerant learner s.t. $k^\star = \bar{n}$, or sufficiently, $\alpha \in \mathcal{G}(\bar{n}, \widetilde{\mathbf{R}})$.*

After learner protocol, all majority users are accurately given one of their top popular items, while minority users are given a popular item. Formally, top-1 item recommendation on $\widehat{\mathbf{R}}$ satisfies the following two properties:

$$\arg\max_{i\in[n]} \hat{r}_{u,i} \subseteq \mathcal{I}_{top}(\widetilde{\mathbf{R}}, u) \cap [\bar{n}] \quad \forall u \in \mathcal{U}_{\text{MAJ}} \tag{16}$$

and

$$\arg\max_{i\in[n]} \hat{r}_{u,i} \subseteq [\bar{n}] \quad \forall u \in \mathcal{U}_{\text{MIN}} \tag{17}$$

Proof. For notational cleanliness in this proof, we will write κ to refer to $\kappa_{(\widetilde{\mathbf{R}},\bar{n})}$ and $\widetilde{\mathbf{R}}'$ to refer to $\widetilde{\mathbf{R}}'(\bar{n})$.

Note that if we consider $\alpha \in \mathcal{G}(\bar{n}, \widetilde{\mathbf{R}})$, by proposition G.2, $k^\star = \bar{n}$ thus, $\widehat{\mathbf{R}} := \widetilde{\mathbf{R}}\mathbf{\Pi}^*_{\bar{n}}$. By Proposition G.3, if $\frac{2\sqrt{2}\kappa n}{\sigma_{\bar{n}}(\widetilde{\mathbf{R}}')} = 0$ then $\mathbf{\Pi}^*_{\bar{n}} = \mathbf{I}_{n,\bar{n}}$. When this is the case, $\widehat{\mathbf{R}} = \widetilde{\mathbf{R}}'$ and the optimal solution for our top-1 item selection problem is such that properties 17 and 16 hold. We want to show that when the Frobenius norm difference of Proposition G.3 is small, under assumptions G.2 and G.3, the top-1 item selection problem's solution is as if the Frobenius norm difference were 0, so properties 17 and 16 still hold. Let the rows of $\widetilde{\mathbf{R}}$ be $\tilde{\mathbf{r}}_u^\top \in [0,1]^n$ for $u \in [m]$ and the columns of $\mathbf{\Pi}^*_{\bar{n}}$ be $\mathbf{v}_i \in \mathbb{R}^n$ for $i \in [n]$. We shall the consider the problem as follows:

$$\begin{pmatrix} \tilde{r}_{1,1}, & \dots & \tilde{r}_{1,n} \\ \vdots & \dots & \vdots \\ \tilde{r}_{m,1}, & \dots & \tilde{r}_{m,n} \end{pmatrix} \begin{pmatrix} 1+\varepsilon_{1,1} & 0+\varepsilon_{1,2} & \dots & 0+\varepsilon_{1,\bar{n}} & \dots & 0+\varepsilon_{1,n} \\ 0+\varepsilon_{2,1} & 1+\varepsilon_{2,2} & \dots & 0+\varepsilon_{2,\bar{n}} & \dots & 0+\varepsilon_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0+\varepsilon_{i,1} & 0+\varepsilon_{i,2} & \dots & 1+\varepsilon_{i,\bar{n}} & \dots & 0+\varepsilon_{i,n} \\ 0+\varepsilon_{i+1,1} & 0+\varepsilon_{i+1,2} & \dots & 0+\varepsilon_{i+1,\bar{n}} & \dots & 0+\varepsilon_{i+1,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0+\varepsilon_{n,1} & 0+\varepsilon_{n,2} & \dots & 0+\varepsilon_{n,\bar{n}} & \dots & 0+\varepsilon_{n,n} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{r}}_1^\top \mathbf{v}_1 & \dots & \tilde{\mathbf{r}}_1^\top \mathbf{v}_n \\ \vdots & \dots & \vdots \\ \tilde{\mathbf{r}}_m^\top \mathbf{v}_1 & \dots & \tilde{\mathbf{r}}_m^\top \mathbf{v}_n \end{pmatrix} \tag{18}$$

Where $\mathbf{\Pi}^*_{\bar{n}}$ is some perturbed $\mathbf{I}_{n,\bar{n}}$ matrix such that $\sqrt{\sum_{i\in[n]}\sum_{i'\in[n]} |\varepsilon_{i,i'}|^2} \le \frac{2\sqrt{2}n\kappa}{(\sigma_{\bar{n}}(\widetilde{\mathbf{R}}'))^2}$. To show the properties 17 and 16 hold for an $\widetilde{\mathbf{R}}$ that satisfies our $\mathcal{M}$ assumptions, it is useful to define some bounds on how far off $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i$ may be from $\tilde{r}_{u,i}$ with respect to a bound on the norm of the perturbation.

Claim G.4 *Fix any $u \in [m]$ and any $i \in [\bar{n}]$ and define an upper bound $x \ge ||\varepsilon_i||_2$ where $\varepsilon_i \in \mathbb{R}^n$ is the ith column of perturbations. The estimate of $\tilde{r}_{u,i}$, $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i$, is lower bounded: $\tilde{r}_{u,i} - \sqrt{n}x \le \tilde{\mathbf{r}}_u^\top \mathbf{v}_i$.*

Proof of Claim G.4. We can rewrite $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i = \tilde{r}_{u,i} + \tilde{r}_{u,i}\varepsilon_{i,i} + \sum_{i'\in[n]\setminus i} \tilde{r}_{u,i'}\varepsilon_{i',i}$. Construct a vector $\mathbf{a} \in \mathbb{R}^n$ such that when $\varepsilon_{i',i} \le 0$, $a_{i'} = -\varepsilon_{i',i}$ otherwise $a_{i'} = 0$. Because $\widetilde{\mathbf{R}} \in [0,1]^{m\times n}$ we have that $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \ge \tilde{r}_{u,i} - \tilde{r}_{u,i}a_{i,i} - \sum_{i'\in[n]\setminus i} \tilde{r}_{u,i'}a_{i',i}$. Invoking the upper bound of 1 on $\tilde{r}_{u',i'}$: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \ge \tilde{r}_{u,i} - \sum_{i'\in[n]} a_{i',i}$. By construction, $a_{i'} \ge 0\ \forall i' \in [n]$. Thus equivalently: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \ge \tilde{r}_{u,i} - ||\mathbf{a}||_1$. Because $a_{i'}$ are equal to $-\varepsilon_{i',i}$ or 0, $||\mathbf{a}||_1 \le ||\varepsilon_i||_1$ where $\varepsilon_i \in \mathbb{R}^n$ is the ith column of perturbations and we can replace $\mathbf{a}$: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \ge \tilde{r}_{u,i} - ||\varepsilon_i||_1$. From the l1-l2 norm inequality and the l2 norm bound on the perturbation column: $||\varepsilon_i||_1 \le \sqrt{n}||\varepsilon_i||_2 \le \sqrt{n}x$. We have: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \ge \tilde{r}_{u,i} - \sqrt{n}x$. ☐

Additionally, we use an analogous proof to show the following claim as well:

Claim G.5 *Fix any $u \in [m]$ and any $i \in [\bar{n}]$ and define an upper bound $x \ge ||\varepsilon_i||_2$ where $\varepsilon_i \in \mathbb{R}^n$ is the ith column of perturbations. The estimate of $\tilde{r}_{u,i}$, $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i$, is upper bounded: $\tilde{r}_{u,i} + \sqrt{n}x \ge \tilde{\mathbf{r}}_u^\top \mathbf{v}_i$.*

Claim G.6 *Fix any $u \in [m]$ and any $i \in \{\bar{n}+1, \dots, n\}$ and define an upper bound $x \ge ||\varepsilon_i||_2$ where $\varepsilon_i \in \mathbb{R}^n$ is the ith column of perturbations. The estimate of $\tilde{r}_{u,i}$, $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i$, is upper bounded: $\sqrt{n}x \ge \tilde{\mathbf{r}}_u^\top \mathbf{v}_i$.*

Proof of Claim G.6. We can rewrite $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i = \sum_{i'\in[n]} \tilde{r}_{u,i'}\varepsilon_{i',i}$.Construct a vector $\mathbf{a} \in \mathbb{R}^n$ such that when $\varepsilon_{i',i} \ge 0$, $a_{i'} = \varepsilon_{i',i}$ otherwise $a_{i'} = 0$. Because $\widetilde{\mathbf{R}} \in [0,1]^{m\times n}$ we have that $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \le \sum_{i'\in[n]} \tilde{r}_{u,i'}a_{i',i}$. Invoking the upper bound of 1 on $\tilde{r}_{u',i'}$: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \le \sum_{i'\in[n]} a_{i',i}$. By construction, $a_{i'} \ge 0\ \forall i' \in [n]$. Thus equivalently: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \le ||\mathbf{a}||_1$. Because $a_{i'}$ are equal to $\varepsilon_{i',i}$ or 0, $||\mathbf{a}||_1 \le ||\varepsilon_i||_1$ where $\varepsilon_i \in \mathbb{R}^n$ is the ith column of perturbations and we can replace $\mathbf{a}$: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \le ||\varepsilon_i||_1$. From the l1-l2 norm inequality and the l2 norm bound on the perturbation column: $||\varepsilon_i||_1 \le \sqrt{n}||\varepsilon_i||_2 \le \sqrt{n}x$. We have: $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \le \sqrt{n}x$. ☐

Now using the above claims we shall show that given any reported preference matrix $\widetilde{\mathbf{R}}$ that satisfies assumption G.2 and G.3, if learner does $\bar{n}$-truncated SVD, properties 16 and 17 hold.
We shall first show that property 16 holds. By property G.3, the Frobenius norm for the whole perturbation is upper bounded. Thus, the L_2 norm for any individual perturbation vector is also upper bounded by the same value. Thus we invoke Claims G.4, G.5, and G.6 with the Frobenius norm bound: $x := \frac{\Delta(\widetilde{\mathbf{R}},\bar{n})}{2\sqrt{n}}$. For a given majority user $u \in \mathcal{U}_{\text{MAJ}}$, we want lower bounds on estimates for the popular top items and upper bounds on estimates for other items:

1. $\forall i \in (\mathcal{I}_{\text{top}}(\widetilde{\mathbf{R}}, u) \cap [\bar{n}])$, $\tilde{r}_{u,i} - \frac{\Delta(\widetilde{\mathbf{R}},\bar{n})}{2} \leq \tilde{\mathbf{r}}_u^\top \mathbf{v}_i$
2. $\forall i' \in (\mathcal{I}_{\text{top}}(\widetilde{\mathbf{R}}, u)^C \cap [\bar{n}])$, $\tilde{r}_{u,i'} + \frac{\Delta(\widetilde{\mathbf{R}},\bar{n})}{2} \geq \tilde{\mathbf{r}}_u^\top \mathbf{v}_{i'}$
3. $\forall i'' \in \{\bar{n}+1, \dots n\}$, $\quad \frac{\Delta(\widetilde{\mathbf{R}},\bar{n})}{2} \geq \tilde{\mathbf{r}}_u^\top \mathbf{v}_{i''}$

By definition of a majority user, $(i^{(u)}_{(top)} \cap [\bar{n}])$ is not empty. But by assumption G.2, $\nexists i, i', i''$ such that $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \leq \tilde{\mathbf{r}}_u^\top \mathbf{v}_{i'}$ or $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i \leq \tilde{\mathbf{r}}_u^\top \mathbf{v}_{i''}$ for any $u \in \mathcal{U}_{\text{MAJ}}$. It must be the case that $\arg\max_{i\in[n]} \hat{r}_{u,i} \subseteq (i^{(u)}_{(top)} \cap [\bar{n}]) \quad \forall u \in \mathcal{U}_{\text{MAJ}}$ and property 16 holds.
Now we shall show that property 17 holds. Invoking Claims G.4 and G.6 using property G.3 for the L_2 bound, for a given $u \in \mathcal{U}_{\text{MIN}}$ we want the lower bound on estimates of popular items to compare to the upper bound on estimates for unpopular items:

1. $\forall i \in [\bar{n}], \quad \tilde{r}_{u,i} - \frac{\Delta(\widetilde{\mathbf{R}},\bar{n})}{2} \leq \tilde{\mathbf{r}}_u^\top \mathbf{v}_i$
2. $\forall i' \in \{\bar{n}+1, \dots n\}, \quad \frac{\Delta(\widetilde{\mathbf{R}},\bar{n})}{2} \geq \tilde{\mathbf{r}}_u^\top \mathbf{v}_{i'}$

But by Assumption G.3, there exists at least one i such that $\tilde{\mathbf{r}}_u^\top \mathbf{v}_i > \tilde{\mathbf{r}}_u^\top \mathbf{v}_{i'} \quad \forall i'$. Therefore $\arg\max_{i\in[n]} \hat{r}_{u,i} \subseteq [\bar{n}] \quad \forall u \in \mathcal{U}_{\text{MIN}}$ and property 17 holds. □

Corollary G.2 (Upper Bound on Social Welfare with Truthful Users) *Additionally, if assumption G.1 holds, and users are truthful (i.e. $\mathbf{R}^\star = \widetilde{\mathbf{R}}$), we have:*

$$\text{SW}(\mathbf{R}^\star, \alpha) \leq |\mathcal{U}_{\text{MIN}}|\underline{R} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^\star_{u,i} < \textit{Max SW Possible} \tag{19}$$

Where $\underline{R} := \max_{u\in\mathcal{U}_{\text{MIN}}} \max_{i'\in[\bar{n}]} r^\star_{u,i'}$

Proof. Let $\texttt{top}^\star(u) \in \mathcal{I}_{\text{top}}(\mathbf{R}^\star, u)$ be some (truthfully) top item for a user u. Recall from the notation in the main body of our paper that $\texttt{top}(u)$ represents the recommended item to user u

$$\begin{aligned}
\text{SW}(\mathbf{R}^\star, \alpha) &= \sum_{u\in[m]} r^\star_{u,\texttt{top}(u)} \\
&= \sum_{u\in\mathcal{U}_{\text{MIN}}} r^\star_{u,\texttt{top}(u)} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} r^\star_{u,\texttt{top}^\star(u)} && \text{(property 16)} \\
&= \sum_{u\in\mathcal{U}_{\text{MIN}}} r^\star_{u,\arg\max_{i\in[\bar{n}]} \hat{r}_{u,i}} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} r^\star_{u,\texttt{top}^\star(u)} && \text{(property 17)} \\
&\leq |\mathcal{U}_{\text{MIN}}|\underline{R} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} r^\star_{u,\texttt{top}^\star(u)}
\end{aligned}$$

By Assumption G.1, the strict inequality holds as well. □

Intuitively, Theorem G.4 and Corollary G.2 highlight something concerning: when users are truthful to this type of α learner, majority users get their best recommendations, while minority users do not, instead they get recommended some popular item which does not reflect their greatest preferences.

G.5 Improving top-1 social welfare via altruism

We are interested in whether and how agents who are in the majority defined by $(\mathbf{R}^\star, \bar{n})$ could improve top-1 social welfare given the learner is α-loss tolerant such that $\alpha \in \mathcal{G}(\mathbf{R}^\star, \bar{n})$ by falsifying ratings on just one minority item.

We consider altruistic misreporting on a truthful preference matrix, $\mathbf{R}^\star$. Much like altruism in the main body of the paper, this altruism transforms $\mathbf{R}^\star$ into $\widetilde{\mathbf{R}}$, which is the same matrix except the $\bar{n}+1$th vector has been changed. However, to generalize the "uprating" in that section, we now no longer limit altruistic agents to only increasing their rating. Rather, $\mathbf{r}^\star$ is changed to any $\tilde{\mathbf{r}} \in [0,1]^m$, though still with the constraint that minority users' ratings remain constant to model **altruism** specifically.

Remark G.4 (Minority item reordering) *Like the main body of the paper, we focus, WLOG, on item $\bar{n}+1$ when it comes to altruism. Minority items can be reordered with no consequence.*

Definition G.8 (General Altruistic Rating) *Consider a ground truth preference matrix $\mathbf{R}^\star$ such that for some $\bar{n}$, $(\mathbf{R}^\star, \bar{n}) \in \mathcal{M}$. WLOG, we consider a [general] altruistic rating strategy to be one in which the $\bar{n}+1$th column vector, $\mathbf{R}^\star_{\bar{n}+1}$, is replaced with $\tilde{\mathbf{r}}$ under the constraints:*

$$\tilde{r}_u = r^*_{u,\bar{n}+1} \quad \forall u \in \mathcal{U}_{\text{MIN}}, \quad \tilde{\mathbf{r}} \in [0,1]^m$$

Such uprating will result in the learner receiving a strategically manipulated preference matrix, $\widetilde{\mathbf{R}} \in [0,1]^{m \times n}$ instead of the true matrix $\mathbf{R}^\star$.

Naturally, because the goal of manipulating ratings of item $\bar{n}+1$ is to help minority users who like it, it will be important to establish that there exists enough [true] preference for item $\bar{n}+1$ such that it is "worthwhile" manipulating.

Assumption G.5 (Manipulated item is sufficiently liked) *For a given $(\widetilde{\mathbf{R}}^\star, \bar{n}) \in \mathcal{M}$, define $\mathcal{U}_{\text{SWITCH}} := \{u : u \in \mathcal{U}_{\text{MIN}}, (\bar{n}+1) \in \mathcal{I}_{top}(\mathbf{R}^\star, u)\} \subseteq \mathcal{U}_{\text{MIN}}$ to be the set of minority users who have a top item which is item $\bar{n}+1$ and assume $|\mathcal{U}_{\text{SWITCH}}| \neq 0$. Assume the following is true of ground truth preferences:*

$\exists \delta \in \mathbb{R}_{>0}$ *such that:*

1. *For minority users whose top item is not $\bar{n}+1$, the variation of popular item ratings is not too large:*
$$\sum_{u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{\text{SWITCH}}} \max_{i \in [\bar{n}+1]} r^\star_{u,i} \leq \sum_{u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{\text{SWITCH}}} \min_{i \in [\bar{n}+1]} r^\star_{u,i} + \delta \tag{20}$$

2. *The switch users like item $\bar{n}+1$ sufficiently more than popular items:*
$$\sum_{u \in \mathcal{U}_{\text{SWITCH}}} r^\star_{u,(\bar{n}+1)} > \sum_{u \in \mathcal{U}_{\text{SWITCH}}} \max_{i \in [\bar{n}]} r^\star_{u,i} + \delta \tag{21}$$

Just like the main body of the paper, we shall now derive sufficient conditions on $\tilde{\mathbf{r}}$ in order to improve social welfare beyond the truthful baseline. Intuitively, these sufficient conditions will represent the following: altruistic rating increases the $\bar{n}+1$ singular value and thus ensures that $(\widetilde{\mathbf{R}}, \bar{n}+1) \in \mathcal{M}$ now allowing some minority users to become a part of the majority, giving these "switch" users all the benefits of being majority from Theorem G.4.

Because we want conditions for $(\widetilde{\mathbf{R}}, \bar{n}+1) \in \mathcal{M}$ while assuming that $(\mathbf{R}^\star, \bar{n}) \in \mathcal{M}$ we will need $\Delta(\widetilde{\mathbf{R}}, \bar{n}+1)$ based on the given $\mathbf{R}^\star, \bar{n}$. By extension, we need $\sigma_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1))$. Rather than use this singular value directly, we use a lower bound, which may be calculated without taking the SVD of $\widetilde{\mathbf{R}}'(\bar{n}+1)$.

Definition G.9 ($\hat{\sigma}_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1))$) *For a given $(\mathbf{R}^\star, \bar{n}) \in \mathcal{M}$, we can estimate the altruistic matrix's $\bar{n}+1$th singular value:*

$$\hat{\sigma}_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1)) := \sqrt{\min(\tilde{\mathbf{r}}^\top \tilde{\mathbf{r}}, [\sigma_{\bar{n}}(\mathbf{R}^{\star\prime}(\bar{n}))]^2) - ||\tilde{\mathbf{r}}^\top \mathbf{A}(\bar{n})||_2}$$

Where $\mathbf{A}(\bar{n}) \in [0,1]^{m \times \bar{n}}$ are the first $\bar{n}$ columns of $\mathbf{R}^\star$

With this estimate in hand, we can proceed with our estimate of $\Delta(\widetilde{\mathbf{R}}, \bar{n}+1)$.

Definition G.10 (Altruistic Sufficient Ratings Gap, $\Delta(\tilde{\mathbf{r}}; \mathbf{R}^\star, \bar{n})$) *For a given $(\mathbf{R}^\star, \bar{n}) \in \mathcal{M}$, we define the sufficient ratings gap needed for a particular altruistic strategy, $\tilde{\mathbf{r}}$, to be:*

$$\Delta(\tilde{\mathbf{r}}; \mathbf{R}^\star, \bar{n}) := \frac{2^{\frac{5}{2}} n^{\frac{3}{2}} \kappa_{(\mathbf{R}^\star, \bar{n}+1)}}{\left[\hat{\sigma}_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1))\right]^2}$$

Where $\kappa := \max_{i' \in \{(\bar{n}+2), \ldots, n\}} \|\mathbf{R}^\star_{i'}\|_1$

Our sufficient conditions for $\tilde{\mathbf{r}}$ will ensure that the altruistic strategy is such that $(\widetilde{\mathbf{R}}, \bar{n}+1) \in \mathcal{M}$ and that learner will select $k^\star = \bar{n}+1$. The sufficient conditions can be evaluated without actually calculating any resulting singular values of $\widetilde{\mathbf{R}}$, which may be expensive to do over the entire space of feasible $\tilde{\mathbf{r}}$

Proposition G.4 (Sufficient Conditions for Effective Altruism) *Let there be some ground truth preference matrix, $\mathbf{R}^\star$, such that for some $\bar{n} \in [n]$, $(\mathbf{R}^\star, \bar{n}) \in \mathcal{M}$ and assumptions G.1 and G.5 hold. Also let there be an α-loss tolerant learner such that $\alpha \in \mathcal{G}(\mathbf{R}^\star, \bar{n})$.*

The following are sufficient conditions on $\tilde{\mathbf{r}}$ (definition G.8) to ensure $\mathrm{SW}(\widetilde{\mathbf{R}}, \alpha) > \mathrm{SW}(\mathbf{R}^\star, \alpha)$:

1. $\alpha < \hat{\sigma}_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1))$
2. $\forall u \in \mathcal{U}_{\text{MAJ}}: \quad \tilde{r}_u < \max_{i \in [n]} r^\star_{u,i} - \Delta(\tilde{\mathbf{r}}; \mathbf{R}^\star, \bar{n})$
3. $\forall u \in \mathcal{U}_{\text{MAJ}}: \quad \max_{i \in [n] \setminus \mathcal{I}_{top}(\mathbf{R}^\star, u)} r^\star_{u,i} < \max_{i \in [n]} r^\star_{u,i} - \Delta(\tilde{\mathbf{r}}; \mathbf{R}^\star, \bar{n})$
4. $\forall u \in \mathcal{U}_{\text{SWITCH}}: \quad \max_{i \in [n] \setminus \mathcal{I}_{top}(\mathbf{R}^\star, u)} r^\star_{u,i} < r^\star_{u,\bar{n}+1} - \Delta(\tilde{\mathbf{r}}; \mathbf{R}^\star, \bar{n})$
5. $\forall u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{\text{SWITCH}}: \quad 0 < \max_{i \in [\bar{n}+1]} r^\star_{u,i} - \Delta(\tilde{\mathbf{r}}; \mathbf{R}^\star, \bar{n})$

Proof. We shall break this proof into the following claims:

Claim G.7 *$\mathrm{SW}(\mathbf{R}^\star, \alpha)$ is upper bounded by:*

$$\sum_{u \in \mathcal{U}_{\text{MIN}}} \max_{i \in [\bar{n}]} r^\star_{u,i} + \sum_{u \in \mathcal{U}_{\text{MAJ}}} \max_{i \in [n]} r^\star_{u,i}$$

Claim G.8 *The altruistically transformed matrix and $\bar{n}+1$ index falls into the popularity gap class, $\mathcal{M}$. Formally:*

$$(\tilde{\mathbf{R}}, \bar{n}+1) \in \mathcal{M}$$

Claim G.9 *$\mathrm{SW}(\widetilde{\mathbf{R}}, \alpha)$ is bounded from below by:*

$$\sum_{u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{\text{SWITCH}}} \min_{i \in [\bar{n}+1]} r^*_{u,i} + \sum_{u \in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{\text{SWITCH}}} \max_{i \in [n]} r^*_{u,i}$$

and thus yields desired (strict) inequality.

Proof of Claim G.7.

By assumption, $(\mathbf{R}^\star, \bar{n}) \in \mathcal{M}$ and $\alpha \in \mathcal{G}(\mathbf{R}^\star, \bar{n})$ thus by proposition G.2, if the users were to submit preferences truthfully, the learner will reduce to rank $\bar{n}$. Thus, by proposition G.4:

$$\arg\max_{i \in [n]} \hat{r}_{u,i} \subseteq \mathcal{I}_{\text{top}}(\mathbf{R}^\star, u) \cap [\bar{n}] \quad \forall u \in \mathcal{U}_{\text{MAJ}}$$

$$\arg\max_{i \in [n]} \hat{r}_{u,i} \subseteq [\bar{n}] \quad \forall u \in \mathcal{U}_{\text{MIN}}$$

Which directly yields the desired social welfare upper bound because majority users get their actual maximum value, while minority users cannot do any better than their maximum value amongst the popular items, which is strictly less than the actual maximum value over all items by assumption G.1. □

Proof of Claim G.8. Recall that a preference matrix $\mathbf{R}$ such that $(\mathbf{R}, \bar{n}) \in \mathcal{M}$ looks like this:

$$\mathbf{R} = \begin{pmatrix} \mathbf{P} & \mathbf{U} \end{pmatrix} \tag{22}$$

Where $\mathbf{P} \in [0,1]^{m\times\bar{n}}$ and $\mathbf{U} \in [0,1]^{m\times(n-\bar{n})}$ are the matrices of popular and unpopular item ratings respectively. Construct the following:

$$\mathbf{X} = \begin{pmatrix}\tilde{\mathbf{r}} & \mathbf{P}\end{pmatrix} \tag{23}$$

Where $\tilde{\mathbf{r}} \in [0,1]^{m\times 1}$ is the $\bar{n}+1$ modified column vector of $\mathbf{R}^\star$ (ie. the 1st column of $\mathbf{U}$) to represent majority users' altruism. Thus $\mathbf{X} \in [0,1]^{m\times(\bar{n}+1)}$. Let $\mathbf{A} := \mathbf{X}^\top\mathbf{X}$. Thus we clearly have

$$\mathbf{A} = \begin{pmatrix} c & \mathbf{a}^\top \\ \mathbf{a} & \mathbf{M}\end{pmatrix}$$

Where:

1. $\mathbf{M} = \mathbf{P}^\top\mathbf{P} \in \mathbb{R}^{\bar{n}\times\bar{n}}$
2. $\mathbf{a}^\top := \tilde{\mathbf{r}}^\top\mathbf{P} \in [0,1]^{1\times\bar{n}}$
3. $c := \tilde{\mathbf{r}}^\top\tilde{\mathbf{r}} \in \mathbb{R}$

Note that the eigenvalues of $\mathbf{A}$ would be the same as the squared nonzero singular values of $\widetilde{\mathbf{R}}'(\bar{n}+1)$. Thus we can use the lower bound given by Lemma D.1 to get a lower bound on $\bar{n}+1$th singular value of $\tilde{\mathbf{R}}'(\bar{n}+1)$. From Lemma D.1:

$$\left[\sigma_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1)\right]^2 \geq \min(\tilde{\mathbf{r}}^\top\tilde{\mathbf{r}}, \left[\sigma_{\bar{n}}(\mathbf{R}^{\star'}(\bar{n}))\right]^2) - ||\tilde{\mathbf{r}}^\top\mathbf{P}||_2 = \left[\hat{\sigma}_{\bar{n}+1}(\widetilde{\mathbf{R}}'(\bar{n}+1))\right]^2$$

Note that this means that our estimate of delta:

$$\Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n}) \geq \Delta(\widetilde{\mathbf{R}},\bar{n}+1)$$

So now that we've established that our estimate is an upper bound on the true Δ our sufficient conditions clearly ensure that assumptions G.2, G.3 would hold on $(\tilde{\mathbf{R}},\bar{n}+1)$ using the real $\Delta(\widetilde{\mathbf{R}},\bar{n}+1)$. We write this out explicitly below:
Assumption G.2: This holds because the users who will be the new majority under $(\tilde{\mathbf{R}},\bar{n}+1)$ are now $u \in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{\text{SWITCH}}$.

1. $\forall u \in \mathcal{U}_{\text{MAJ}}: \quad \tilde{r}_u < \max_{i\in[n]} r^\star_{u,i} - \Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n}) \leq \max_{i\in[n]} r^\star_{u,i} - \Delta(\widetilde{\mathbf{R}},\bar{n}+1)$
2. $\forall u \in \mathcal{U}_{\text{MAJ}}: \quad \max_{i\in[n]\setminus\mathcal{I}_{\text{top}}(\mathbf{R}^\star,u)} r^\star_{u,i} < \max_{i\in[n]} r^\star_{u,i} - \Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n}) \leq \max_{i\in[n]} r^\star_{u,i} - \Delta(\widetilde{\mathbf{R}},\bar{n}+1)$
3. $\forall u \in \mathcal{U}_{\text{SWITCH}}: \quad \max_{i\in[n]\setminus\mathcal{I}_{\text{top}}(\mathbf{R}^\star,u)} r^\star_{u,i} < r^\star_{u,\bar{n}+1} - \Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n}) \leq r^\star_{u,\bar{n}+1} - \Delta(\widetilde{\mathbf{R}},\bar{n}+1)$

Assumption G.3: Minority users is slightly more subtle because on $(\tilde{\mathbf{R}},\bar{n}+1)$ the new minority group is $\subseteq \mathcal{U}_{\text{MIN}} \cup \mathcal{U}_{\text{SWITCH}}$ (recall that majority and minority groups are not necessarily exclusive unless stated). However, once again by construction, the properties $\tilde{\mathbf{r}}$ satisfies ensure that assumption G.3 is satisfied on all $u \in \mathcal{U}_{\text{MIN}} \cup \mathcal{U}_{\text{SWITCH}}$. Because $\forall u \in \mathcal{U}_{\text{SWITCH}}$:

$$\max_{i\in[n]\setminus\mathcal{I}_{\text{top}}(\mathbf{R}^\star,u)} r^\star_{u,i} < r^\star_{u,\bar{n}+1} - \Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n}) \leq r^\star_{u,\bar{n}+1} - \Delta(\widetilde{\mathbf{R}},\bar{n}+1)$$

Which automatically implies

$$0 < r^\star_{u,\bar{n}+1} - \Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n}) \leq r^\star_{u,\bar{n}+1} - \Delta(\widetilde{\mathbf{R}},\bar{n}+1)$$

And then satisfaction of assumption G.3 for $u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{\text{SWITCH}}$ follows directly from the the $\tilde{\mathbf{r}}$ properties again because $\Delta(\widetilde{\mathbf{R}},\bar{n}+1)$ is upper bounded by $\Delta(\tilde{\mathbf{r}};\mathbf{R}^\star,\bar{n})$

Of course it is also the case that $\widetilde{\mathbf{R}} \in [0,1]^{m\times n}$ Thus we have $(\widetilde{\mathbf{R}},\bar{n}+1) \in \mathcal{M}$ as desired. $\square$

Proof of Claim G.9. We shall now leverage the fact that $(\tilde{\mathbf{R}},\bar{n}+1) \in \mathcal{M}$ to get a lower bound on social welfare. First, we prove that if this α learner sees $\widetilde{\mathbf{R}}$, he will reduce to rank $\bar{n}+1$. We have from assumptions that $\alpha \in (\sqrt{(n-\bar{n})\kappa_{(\mathbf{R}^\star,\bar{n})}}, \sqrt{\min(\tilde{\mathbf{r}}^\top\tilde{\mathbf{r}}, [\sigma_{\bar{n}}(\mathbf{R}^{\star\prime}(\bar{n}))]^2) - ||\tilde{\mathbf{r}}^\top\mathbf{P}||_2})$. We need to prove that this guarantees we also have:

$$\alpha \in (\sigma_{\bar{n}+2}(\tilde{\mathbf{R}}), \sigma_{\bar{n}+1}(\tilde{\mathbf{R}}))$$

We shall start with the LHS:

$$\sqrt{\min(\tilde{\mathbf{r}}^\top\tilde{\mathbf{r}}, [\sigma_{\bar{n}}(\mathbf{R}^{\star\prime}(\bar{n}))]^2) - ||\tilde{\mathbf{r}}^\top\mathbf{P}||_2} \leq \sigma_{\bar{n}+1}(\tilde{\mathbf{R}}'(\bar{n}+1)) \qquad \text{(Claim G.8)}$$
$$\leq \sigma_{\bar{n}+1}(\tilde{\mathbf{R}}) \qquad \text{(Corollary G.1)}$$

Now the RHS:

$$\sqrt{(n-\bar{n})\kappa_{(\mathbf{R}^\star,\bar{n})}} \geq \sqrt{(n-\bar{n})\kappa_{(\tilde{\mathbf{R}},\bar{n}+1)}} \qquad \text{(Definition of } \kappa\text{)}$$
$$\geq \sigma_{\bar{n}+2}(\tilde{\mathbf{R}}) \qquad \text{(Proposition G.1)}$$

Thus the learner will reduce to rank $\bar{n}+1$. Because $(\tilde{\mathbf{R}}, \bar{n}+1) \in \mathcal{M}$ and the learner will rank reduce to $\bar{n}$, we can now invoke proposition G.4:

$$\arg\max_{i\in[n]} \hat{r}_{u,i} \subseteq \mathcal{I}_{\text{top}}(\widetilde{\mathbf{R}}, u) \cap [\bar{n}+1] \quad \forall u \in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{\text{SWITCH}}$$
$$\arg\max_{i\in[n]} \hat{r}_{u,i} \subseteq [\bar{n}+1] \quad \forall u \in \mathcal{U}_{\text{MIN}}$$

From this we get the lower bound we want because users $\in \mathcal{U}_{\text{MAJ}} \cup \mathcal{U}_{\text{SWITCH}}$ will receive their top item since we ensure $\mathcal{U}_{\text{MAJ}}$ top items are unchanged by the sufficient conditions that guarantee $\forall u \in \mathcal{U}_{\text{MAJ}}: \quad \tilde{r}_u < \max_{i\in[n]} r^\star_{u,i}$ and ratings for users $\in \mathcal{U}_{\text{SWITCH}}$ are unchanged.

Users $u \in \mathcal{U}_{\text{MIN}} \setminus \mathcal{U}_{\text{SWITCH}}$ might receive something as bad as their worst $i \in [\bar{n}+1]$ item:

$$\text{SW}(\widetilde{\mathbf{R}}, \alpha) \geq \sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \min_{i\in[\bar{n}+1]} r^\star_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}\cup\mathcal{U}_{\text{SWITCH}}} \max_{i\in[n]} r^\star_{u,i}$$

Thus we have the following ρ bound by invoking the assumption that the switch users sufficiently like their top item (Assumption G.5) (colored for clarity):

$$\begin{aligned}
\rho &\geq \frac{LB(\text{SW}(\widetilde{\mathbf{R}}, \alpha))}{UB(\text{SW}(\mathbf{R}^\star, \alpha)}\\
&= \frac{\sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \min_{i\in[\bar{n}+1]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}\cup\mathcal{U}_{\text{SWITCH}}} \max_{i\in[n]} r^*_{u,i}}{\sum_{u\in\mathcal{U}_{\text{MIN}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}}\\
&= \frac{\sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \min_{i\in[\bar{n}+1]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{SWITCH}}} \max_{i\in[n]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}}{\sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}}\\
&> \frac{\sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \min_{i\in[\bar{n}+1]} r^*_{u,i} + \delta + \sum_{u\in\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}}{\sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}} \qquad \text{equation 21}\\
&\geq \frac{\sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \min_{i\in[\bar{n}+1]} r^*_{u,i} + \delta + \sum_{u\in\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}}{\delta + \sum_{u\in\mathcal{U}_{\text{MIN}}\setminus\mathcal{U}_{\text{SWITCH}}} \min_{i\in[\bar{n}+1]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{SWITCH}}} \max_{i\in[\bar{n}]} r^*_{u,i} + \sum_{u\in\mathcal{U}_{\text{MAJ}}} \max_{i\in[n]} r^*_{u,i}} \qquad \text{equation 20}\\
&= 1
\end{aligned}$$

Which yields the desired $\rho > 1$ □

□

G.6 Extra results on $\mathcal{M}$

While we have explained how the $\mathcal{M}$ class of tuples represents matrices with a popularity gap, it is not intuitively clear exactly what combinations of $\mathbf{R}$ and $\bar{n}$ may work. Can one $\mathbf{R}$ have multiple values of $\bar{n}$ such that $(\mathbf{R}, \bar{n}), (\mathbf{R}, \bar{n}') \in \mathcal{M}$? We can show conditions that, given $(\mathbf{R}, \bar{n}) \in \mathcal{M}$, for $\bar{n}' > \bar{n}$, $(\mathbf{R}, \bar{n}') \notin \mathcal{M}$. There are perhaps other interesting propositions about this class we leave to future work.

Proposition G.5 (Greater $\bar{n}$ does not satisfy assumptions for $\mathcal{M}$) *Define*

$$\underline{\kappa}_{(\mathbf{R},\bar{n})} := \min_{i'\in\{(\bar{n}+1),\ldots,n\}} ||\mathbf{R}_{\mathbf{i}'}||_1$$

If a tuple $(\mathbf{R}, \bar{n}) \in \mathcal{M}$ and $\underline{\kappa}_{(\mathbf{R},\bar{n})} > \frac{(n-\bar{n})\kappa_{(\mathbf{R},\bar{n})}}{4\sqrt{2}n\sqrt{n}}$ then $\nexists \bar{n}' \in \{\bar{n}+1, \ldots, n\}$ s.t. $(\mathbf{R}, \bar{n}')$ satisfies the assumptions of the previous subsection.

Proof. Define $\bar{n}' \in \{\bar{n}+1, \ldots, n\}$, $\mathbf{R}'(\bar{n}') \in [0,1]^{m \times \bar{n}'}$ to be the matrix $\mathbf{R}$, but with all columns $i > \bar{n}'$ set to be 0 vectors, and $\kappa_{(\mathbf{R},\bar{n}')} := \max_{i' \in \{(\bar{n}'+1),\ldots,n\}} ||\mathbf{R}_{\mathbf{i}'}||_1$. We want to show that $\sigma_{\bar{n}'}(\mathbf{R}'(\bar{n}')) < \sqrt{4\sqrt{2}n\sqrt{n}\kappa_{(\mathbf{R},\bar{n}')}}$. If this is the case, it definitely cannot be true that $(\mathbf{R}, \bar{n}')$ satisfies the assumptions of the previous subsection because it would require the difference between top and next rating to be greater than 1.

By corollary G.1 and proposition G.1, $\sigma_{\bar{n}+1}(\mathbf{R}'_{\bar{n}'}) \leq \sqrt{\kappa_{(\mathbf{R},\bar{n})}(n-\bar{n})}$. Note that this is in terms of $\kappa_{(\mathbf{R},\bar{n})}$ and not in terms of $\kappa_{(\mathbf{R},\bar{n}')}$. Using the asssumption, we have that: $\frac{\underline{\kappa}_{(\mathbf{R},\bar{n})}(4\sqrt{2}n\sqrt{n})}{n-\bar{n}} > \kappa_{(\mathbf{R},\bar{n})}$. Thus we can write:

$$\sigma_{\bar{n}+1}(\mathbf{R}'_{\bar{n}'}) \leq \sqrt{\kappa_{(\mathbf{R},\bar{n})}(n-\bar{n})} < \sqrt{4\sqrt{2}n\sqrt{n}\underline{\kappa}_{(\mathbf{R},\bar{n})}} \leq \sqrt{4\sqrt{2}n\sqrt{n}\kappa_{(\mathbf{R},\bar{n}')}}$$

Where the last inequality follows because $\underline{\kappa}_{(\mathbf{R},\bar{n})}$ is minimum. $\square$