RUMs with Ties: A Discrete Choice Model Allowing Multiple Winners

Flavio Chierichetti[1*], Ravi Kumar[2], Giuseppe Re[3], and Andrew Tomkins[2]

[1] Sapienza University, Rome, Italy
flavio@di.uniroma1.it
[2] Google Research, Mountain View, CA, US
ravi.k53@gmail.com, tomkins@google.com
[3] Algorand Labs, Rome, Italy
giuseppe.re@algorandlabs.com

Abstract. Discrete choice models are used to describe, explain, and predict choices made by people among a finite set of alternatives. However, standard discrete choice models come with an unrealistic assumption: that users are able to provide an *unequivocal* clear winner from any slate of alternatives. Often the user knows the winner but cannot report it, as when a UI does not allow a user to specify which of two movies they rated five stars is better. And often, among the myriad options available, the user is able to identify some good candidates, but finds it difficult to distinguish between the top contenders. In this paper, we study the problem of interacting with user choice data in which, sometimes, the user is unable to settle on a single compelling winner.

To address this issue, we introduce an extension to the well-known random utility models (RUMs), which we call *RUMs-with-Ties*, where comparisons can result in a tie. We begin with an axiomatic formulation of Luce dating to the 1950s, and provide algorithms and matching lower bounds for operating on data with ties. We also provide a comprehensive comparison of RUMs versus RUMs-with-Ties from different angles. We present theoretical results indicating that simple ways of incorporating ties into existing approaches are unlikely to perform well. We also prove in our setting that the presence of additional items, even if lower in quality, allows an algorithm to learn the highest ranked element with far fewer trials. Finally, we provide experimental evaluations of different approaches to handling indistinguishable items in choice settings and demonstrate the advantages of direct modeling of ties via our approach.

Keywords: random utility models, discrete choice, ties

1 Introduction

Users are increasingly faced with a so-called *Paradox of Choice* [46], in which more and more alternatives are available, paradoxically resulting in a decrease

* Flavio Chierichetti was partially supported by a Google Focused Research Award, by BiCi – Bertinoro international Center for informatics, and by the PRIN project 20229BCXNW.

in welfare due to the time and anxiety of understanding and weighing so many options [26]. Worse yet, as products are increasingly optimized, each generation improves by smaller amounts, causing the degree of difference between options to shrink towards zero, and exacerbating the difficulty of making a well-reasoned choice [15]. In this paper, we consider one important implication of increasingly complex decision spaces: machine learned models are frequently trained to predict a user's choice from a set of options, but users are sometimes unable to distinguish between the top options, resulting in a *tie*. A tie may come about because a user is restricted to a fixed scoring scale, such as five stars, so the user's preference may exist but the platform provider cannot observe it. It may similarly occur because the user takes actions on multiple items that do not indicate which is preferred, such as adding two items to a comparison set, or watching several short videos to completion. Or the user herself may not be able to distinguish between several top choices, such as when researching some options without coming to a final decision during a session [18].

Our models formalize the notion of a user being unable to distinguish the best option, and study approaches to accounting for these ties in modeling.

For clarity, there is a related problem that we do *not* consider: the user electing to select no item from the set of alternatives. This *non-consumption* provides relevant information [6] and helps predict future choices better [11, 12]. Non-consumption may occur for many reasons [16]. A user asked whether they drive to work or take public transit may not respond because they telecommute or commute by bicycle [53]. Or a user offered several products may find that none of them provides compelling value for money. Generally, adding more alternatives will decrease the probability of non-consumption [43]. On the other hand, our topic of study—the phenomenon of indistinguishability of top alternatives—is different in part because the probability of being unable to distinguish the best choice may either increase or decrease with more options: an obvious winning option may be added, or a previous standout best option may become conflated with a new option of similar utility.

Formalizing ties via just noticeable differences. Early efforts to model indistinguishable items in utility theory assumed transitivity: if a and b were indistinguishable, and likewise b and c, then a must be indistinguishable from c. This assumption was convenient but was recognized as unrealistic [5]. As a classic example, consider a sequence of weights ranging from 10g to 10kg, each 0.1% heavier than the previous one. Neighboring weights are indistinguishable by humans, and so under transitivity all weights in the sequence must be indistinguishable, but the lightest and heaviest are obviously different. For scalar quantities like weight, heat, or size, the psychometrics community has a long-standing notion of a *just noticeable difference* (JND), introduced by Weber and formalized by Fechner in the late 1800s [20]. Under this notion, weight w_1 might be indistinguishable from w_2, but eventually the difference between w_1 and w_i exceeds the JND and can then be sensed. Discrete choice relies on the scalar quantity of an item's overall utility, where a JND might reasonably be applied. Thus, two items (e.g., cars, movies) might have a clear winner or, if within the JND, might be "too close to call" for a user trying to distinguish them.

In 1956, R. D. Luce developed a theory extending the preferences of a user to incorporate intransitive indistinguishability using a notion of JND [35]. He developed an axiomatic framework connecting a new set-theoretic concept he called a *semiorder* to a numerical realization based on utility values of items and the JNDs between them, and showed these two formulations were equivalent. In his numeric model, each item a has a utility w_a, and there is JND value δ such that a and b are indistinguishable if $|w_a - w_b| \leq \delta$.[4] Luce formalizes an important but subtle point about JNDs, which we introduce by example. Fix δ and consider the setting where $|w_a - w_b| \leq \delta$, so a and b are indistinguishable. Now the user considers c, and realizes that, while c is indistinguishable from a, it is clearly worse than b: $w_b - w_c > \delta$. We can perform an inference by contradiction: say a is no worse than b. Then $w_a \geq w_b > w_c + \delta$, contradicting our observation that a and c are indistinguishable. While the user cannot make a direct comparison between a and b, introducing a separate element c allowed an *indirect* comparison, and the user can now *infer* that b is preferred to a. This type of indirect reasoning follows directly from the definition of indistinguishability using JNDs, and also connects to the human intuition that comparing hard-to-distinguish options to some new alternatives can provide new insights and help us to reach a conclusion. We must therefore differentiate between *observed preferences*, which occur when $w_b > w_a + \delta$ and *inferred preferences*, which occur when $|w_a - w_b| < \delta$ but $\exists c : w_c \in [\min(w_a, w_b) - \delta, \max(w_a, w_b) - \delta) \cup (\min(w_a, w_b) + \delta, \max(w_a, w_b) + \delta]$. We study below how powerful inferred preferences are compared to observed preferences.

Representing inferred preferences. Following Luce's definitions, we formalize the distinction between observed and inferred preferences and state the computational problem of deriving all possible inferred preferences. While Luce was not focused on computational questions (his work predates P vs NP!), his proof connecting semiorders with utilities and JNDs implicitly specifies an algorithm for determining inferred preferences;[5] we formalize and analyze it. We also develop an improved algorithm, which is implied by later work in the area, and we show that this second algorithm is more computationally efficient. Complementing, we also introduce a matching lower bound on the running time to conclude this is the optimal algorithm for determining all inferred preferences. Once all such preferences have been inferred, Luce showed that the final distinguishable preferences of the user may be perfectly represented as a *bucket order* in which

[4] Luce's model is slightly more general, in that each element may have a more general interval of indistinguishability around it.

[5] Luce's formulation of a semiorder starts with a binary relation $<$. If $a \not> b$ and $b \not> a$ then a and b are indistinguishable, written $a \parallel b$. To establish the equivalence with the more actionable formulation based on utilities, Luce shows two additional axioms are required. First, if $a > b$, $b \parallel c$, and $c > d$, then $a > d$. Second, if $a > b$, $b > c$, and $b \parallel d$, then either $a \not\parallel d$ or $c \not\parallel d$. These additional axioms are used in his proof via an argument that implicitly gives the information necessary to compute all inferred preferences.

the universe is partitioned into an ordered set of buckets. Items in distinct buckets are distinguishable according to the ordering of the buckets, while items within a bucket are indistinguishable. Interestingly, beginning with an intransitive distinguishability relation based on JNDs, and allowing all possible inferences, results in an (often fine-grained) *equivalence* relation in which the remaining indistinguishable elements are transitive, as the elements of a bucket must all be within δ of one another.

From user models to population models. The upshot of these observations is a model grounded in JNDs to characterize a user's preferences, including a characterization of pairs of elements that are too close to distinguish, even by inferred preferences. This bucket order may be viewed as an augmentation (incorporating JNDs) of the total ordering implied by a user's utility vector. In practical settings, we are faced not with a single user but with many users, each with potentially different utilities. In the classical theory of discrete choice, the behaviors of the entire population are captured by a distribution over utility vectors, known as a *Random Utility Model (RUM)* [8, 9, 50, 55, 56]. We define an analogous notion of *RUMs-with-Ties*, characterized by a distribution over user bucket orders, to capture the behavior of a population of users making choices with the possibility of being unable to distinguish the best item. The rest of this paper explores the properties of RUMs-with-Ties, along with a series of experiments studying their representational power and learnability. Our results regarding bucket orders, inferred preferences, and the definition of RUMs-with-Ties are covered in Sections 3 and 4.

RUMs versus RUMs-with-Ties. The first question is whether the new model is truly different from RUMs. We address this from a few perspectives. (i) We observe that any RUM may be represented as a RUM-with-Ties whose buckets are all singletons. However, the converse is not true—there are RUMs-with-Ties that cannot be approximated by any RUM on the same vocabulary plus a "no-choice" element. (ii) Alternately, we might try to use a RUM to represent just the cases in which a RUM-with-Ties provides a winner; here also, there are RUMs-with-Ties such that the winner distribution for each slate, conditioned on some element winning, cannot be approximated by any RUM. Hence, RUMs-with-Ties appear to be strictly more representationally powerful. (iii) We ask next how succinctly a RUM-with-Ties can be written down, allowing a small approximation error to provide a more robust measure. Here, we are able to extend a result of [14] for RUMs, showing that both types of models can be approximated using $\Theta(n^2 \log n)$ bits. (iv) Finally, we ask whether a deployed system receives more or less information when interacting with a RUM versus a RUM-with-Ties. On the one hand, the RUM specifies the winner of every slate, which seems like a clear improvement. On the other hand, a RUM-with-Ties might give more information as the vocabulary is larger by a single element. We study the question of the information content of responses from these two models using a statistical testing framework, asking in which model the hypotheses can be distinguished more rapidly. We show that, surprisingly, the sample complexity may either increase or decrease in the presence of ties. (See Section 5.)

Can ties be useful? We raised the tantalizing prospect that an additional item might help a user to differentiate between two very close options. In Section 6, we study the theoretical justification for this question: could adding an additional (worse) item uncover the better option with only, say, 1% as many trials as direct comparison? We answer this question in the affirmative for the most popular discrete choice model, the multinomial logit (MNL), augmented with a JND; the resulting model is known as δ-*MNL* (or δ-*logit*) [33]. We consider two close elements, a and b. Under MNL, each trial between them will add Gumbel random noise to the utilities of a and b, and with some small probability, they will become distinguishable. However, many such trials will be required to determine the winner by direct comparison between a and b. We then consider introducing a third element, c, worse than either, and allow comparisons between c and each of a and b in turn. By analyzing the information available from this indirect comparison versus the direct comparison between a and b, we show that comparisons with a well-chosen third element c can be arbitrarily more efficient by any desired multiplicative factor in the number of trails.

2 Related work

Discrete choice models have been extensively studied in disparate settings [4, 21, 31, 32]. RUMs lie at the core of choice theory [9, 16, 45, 47, 50], but they have also been used as components of larger models for voting [27] and text generation [48]. MNLs [23] are a special case of RUMs, but their simplicity makes them easier to learn from data and so are widely adopted [39, 41] and used beyond choice theory [7]. The discrete choice literature is vast [24]; we only focus on works closest to ours.

Learning choice models from preference data. There has been recent work on learning ranking and choice models from preference data [1, 47] and on learning RUMs from preferences [3, 13, 42]. Others have focused on the identifiability of RUMs [51, 56].

Learning from partial orders. Learning from incomplete preferences inducing a partial order have been studied for MNLs [34, 36, 37, 57] and RUMs [55]. These assume the underlying model is a RUM and only the observed data induces a partial order on items. However, we show that if the partial order comes from the indistinguishability of top alternatives, RUMs do not provide an accurate description of the choice distribution, while RUMs-with-Ties can.

Modeling indifference. The seminal work by [35] spearheaded different research directions on including indifference threshold in discrete choice models. There have been attempts at axiomatizing choice models resulting from the indistinguishability of top alternatives for a single "user" [10, 19]. In this regard, [40] analyze the properties of a stochastic choice function resulting from indecisiveness between close alternatives. In particular, they show that RUMs may or may not be induced by indistinguishability, validating the need for a more comprehensive model like ours. Another research direction, which we also pursue in this work, is to provide probabilistic models incorporating indistinguishability and learning algorithms for them. These include the MPD model [30] having an indifference threshold in a 2-MNL and the δ-MNL model [33] for more than two items. The latter has been used in several practical settings [11, 12, 54]. Recently, [38] provided an adaptive online learning algorithm for δ-MNL, but they only looked at pairwise preferences and not at the partial order induced by δ-MNLs. Our RUMs-with-Ties model is a generalization of δ-MNLs and accounts for more complex dynamics. Finally, some works have modeled indifference by adding a fictitious "no-choice" item to an MNL [22, 25]. We show that, both in theory and in practice (MNL++), this is quite inaccurate at predicting choices.

3 Preliminaries

Order relations. A *partial order* (or *poset*) $P = (X, \prec)$ is a binary ordering relation $\prec$ on X^2 that is transitive ($x \prec y$ and $y \prec z \implies x \prec z$) and antisymmetric ($x \prec y$ and $y \prec x \implies x = y$). If x, y are two elements of a poset such that $x \not\prec y$ and $y \not\prec x$, then we say that x and y are *incomparable* and write $x \parallel y$. An *antichain* of a poset is a set of pairwise incomparable elements. Given a poset P, we sometimes use $\prec_P$ to denote the ordering relation of P.

A *strict weak order* is a poset $P = (X, \prec)$ such that X can be partitioned into antichains $L_1, \ldots, L_k$, and for each $1 \leq i < j \leq k$, for each $x \in L_i$ and for each $y \in L_j$, it holds $x \prec y$, i.e., the antichains that make up the strict weak ordering are totally ordered. We will use *bucket order* as a synonym for strict weak order.

Let $\mathcal{S}_n$ be the set of all permutations (or *total orders*) of $[n]$. Also, let $\mathcal{B}_n$ denote the set of all the bucket orders of $[n]$; clearly, $\mathcal{B}_n \supseteq \mathcal{S}_n$. The number of bucket orders on $[n]$ is called the *Fubini number* (or *ordered Bell number*) and it satisfies $\log |\mathcal{B}_n| = \Theta(n \log n)$ [49].

Probability distributions. For a discrete distribution D, let $\operatorname{supp}(D)$ denote its support. Let $x \sim D$ denote that $x \in \operatorname{supp}(D)$ is sampled from D. For any element x, let $D(x)$ denote the probability that D assigns to x; clearly, $x \in \operatorname{supp}(D)$ if and only if $D(x) > 0$. We generalize this to any subset $S \subseteq \operatorname{supp}(D)$: $D(S) = \sum_{x \in S} D(x)$. When it is clear from the context, let $D(S)$ denote $D(S \cap \operatorname{supp}(D))$.

The *total variation distance* between distributions D and D' is

$$|D - D'|_{\mathrm{tv}} = \frac{1}{2} \sum_{\substack{x \in \operatorname{supp}(D) \\ \cup \operatorname{supp}(D')}} |D(x) - D'(x)| = \frac{1}{2} |D - D'|_1 .$$

This is also equal to the maximum, over all events S, of the absolute difference between the probabilities of S in D and in D', i.e.,

$$|D - D'|_{\mathrm{tv}} = \max_{S \subseteq \operatorname{supp}(D) \cup \operatorname{supp}(D')} |D(S) - D'(S)| .$$

Discrete choice. Let a *slate* denote any non-empty subset of $[n]$. Given a bucket order $B \in \mathcal{B}_n$ and a slate $T \subseteq [n]$, let

$$B(T) = \begin{cases} i & \text{if } i \in T \text{ satisfies } i \succ_B j, \forall j \in T \setminus \{i\}, \\ \perp & \text{otherwise.} \end{cases}$$

i.e., the unique maximum item in T according to B if it exists, and $\perp$ otherwise (i.e., in the case where the maximum item is not unique).

A *RUM* on $[n]$ is a probability distribution over $\mathcal{S}_n$.[6] As we will see later, a *RUM-with-Ties* on $[n]$ can be understood as a probability distribution D over $\mathcal{B}_n$ (although our first definition will be based on a user model). We drop the quantifier "on $[n]$" when it is obvious from the context. Given a slate $T \subseteq [n]$, we use D_T to denote the distribution of the random variable $B(T)$ for $B \sim D$, i.e., the distribution of the winner in the slate T with a random bucket order from D. Note that $\mathrm{supp}(D_T) \subseteq T \cup \{\perp\}$. A *Multinomial Logit (MNL)* is a special type of RUM where each item $i \in [n]$ is associated with *base value* w_i. A random permutation is generated by iteratively sampling without replacement, where at each step, an unsampled item is chosen w.p. proportional to the exponential of its base value. For a slate S, the probability $i \in S$ wins is $\frac{e^{w_i}}{\sum_{j \in S} e^{w_j}}$.

To define an approximation notion for RUM-with-Ties, we first define a *distance* between two RUM-with-Ties D, D', following [14]: $\mathrm{dist}(D, D') =$

$$\max_{\varnothing \subset S \subseteq [n]} |D_S - D'_S|_{\mathrm{tv}} = \max_{\substack{\varnothing \subset S \subseteq [n] \\ S' \subseteq S \cup \{\perp\}}} |D_S(S') - D'_S(S')| .$$

I.e., the distance is the maximum, over the slates S, of the total variation distance of the winner distributions of S with D and D'. Equivalently, it is the maximum over S and $S' \subseteq S \cup \{\perp\}$ of the absolute difference of the probabilities, with D and D', that the result of a random choice in S lies in S'. (Such a random choice could either return an element of S, or $\perp$ if no choice is made in S.)

We omit all proofs due to space constraints.

3.1 System 1 and System 2 models

In our paper we follow Luce's pioneering work [35] and consider the following evaluations-with-noise models. We use the terminology popularized by the Nobel laureate Daniel Kahneman in his book [28] inspired by the way in which humans think: they can use their "System 1"—an instinctive and fast mode of thought—or their "System 2"—a logical and relatively slower way of thinking.

Definition 1 (System 1 Model and G_1). *Fix $\delta \geq 0$ and let $(U_1, \ldots, U_n)$ be the user's utilities sampled from a joint value distribution $\mathcal{D}$. Given $S \subseteq [n]$ and given $\{i, j\} \in \binom{S}{2}$, the user infers $i > j$ (i beats j) if $U_i > U_j + \delta$, the user infers*

[6] RUMs are typically presented in terms of noisy item evaluations made by users. Each item is assumed to have a base value; each user samples utility for each item from a joint distribution. The user then chooses an item "rationally" as the one with the highest utility among the available ones (breaking ties, if any, u.a.r.). As the utilities are random, the family of resulting models is named "Random Utility Models," or RUMs. In an equivalent definition, the user first sorts all the items decreasingly according to their observed utilities (breaking ties, if any, u.a.r.), obtaining a permutation; given a slate, a rational user will choose the item with the highest rank in the permutation.

$j > i$ (j beats i) if $U_j > U_i + \delta$, and the user infers $i \parallel j$ (i incomparable to j) if $|U_i - U_j| \le \delta$. (Equivalently, given a slate $S \subseteq [n]$, the user sees a digraph G_1 with the elements of S as its nodes and for each $i, j \in S$, an arc from i to j if $i > j$, i.e., $U_i > U_j + \delta$.) Now, if there exists $i \in S$ such that $i > j$ for each $j \in S \setminus \{i\}$, then the user will choose i in S; otherwise, the user will choose nothing (i.e., a $\perp$ choice).

Thus, if a user is told explicitly by System 1 that an object from the slate is better (by δ) than every other object in the slate, then they will choose that object. Note that if $(U_1, \ldots, U_n)$ is such that $\Pr[\exists\{i,j\} \in \binom{[n]}{2} \mid U_i = U_j] = 0$—e.g., if the distribution is continuous and has full support—then System 1 with $\delta = 0$ is the standard RUM model. For $\delta > 0$, instead, System 1 enables the user to choose "nothing". The relation $>$ in Definition 1 is also called a *semiorder*.

Sometimes, however, users can use inductive reasoning to improve their ability to choose the best option in a slate. For instance, if the user feels that $1 > 3$ but also feels that $1 \parallel 2$ and $2 \parallel 3$, then the user can infer that 1 is the item with the largest utility. Indeed, $1 > 3 \implies U_1 > U_3 + \delta$ and $2 \parallel 3 \implies |U_2 - U_3| \le \delta \implies U_3 \ge U_2 - \delta$. Together we get $U_1 > U_3 + \delta \ge U_2$, letting the user infer 1 is better than 2. A similar argument shows that $2 > 3$. Thus, inductive reasoning allows the user to infer the full ordering $1 > 2 > 3$, whereas System 1 only allowed the user to observe that $1 > 3$. We codify this additional inference power below.

Definition 2 (System 2 Model and G_2). *Fix $\delta \ge 0$ and let $(U_1, \ldots, U_n)$ be the user's utilities sampled from a joint value distribution $\mathcal{D}$. Given a slate S and G_1 from System 1, the user produces a digraph G_2 on the same node set S containing each arc $i \to j$ for which the user can infer that $U_i > U_j$ from G_1. If $i \to j$ is in G_2, we say that $i \succ j$. Now, if there exists $i \in S$ such that $i \succ j$ for each $j \in S \setminus \{i\}$, then the user will choose i in S; otherwise, the user will choose nothing (i.e., a $\perp$ choice).*

Thus System 2 can add arcs to G_1 so to obtain a more informative rational graph G_2; the arc $i \to j$ is in G_2 only if the user can formally prove that the $U_i > U_j$. Such proofs could be direct (e.g., $i \to j$ is in G_1, thus $U_i > U_j + \delta$) or, as for the above example, might involve comparisons to other items. Note that for the example slate $S = \{1, 2, 3\}$, the user's System 1 chooses nothing from S whereas the user's System 2 is able to determine the best option in S.

Finally, we show a result on the structure of the optimal inferred G_2. This result is somewhat implicit in [35] but not in terms of G_2.

Theorem 1 ([35]). *G_2 induces a bucket order.*

3.2 RUMs-with-Ties

Finally, we recall a combinatorial model that is simpler to state yet equivalent to System 2, i.e., the choice systems representable with System 2 coincide with the choice systems representable by this combinatorial model. This model will allow us to find an optimal fit as well to succinctly represent a generic System 2 model.

Definition 3 (RUM-with-Ties). *A* RUM-with-Ties *is a probability distribution D over $\mathcal{B}_n$. Given $S \subseteq [n]$, the probability $D_S(i)$ that $i \in S$ gets chosen is equal to the probability, over bucket orders B from D, that i is the unique highest item of S in B, i.e., $D_S(i) = \Pr_{B\sim D}[B(S) = i]$. The probability $D_S(\perp)$ that no choice is made from S is defined as $D_S(\perp) = \Pr_{B\sim D}[B(S) = \perp]$.*

The following was proved in [35].

Theorem 2 ([35]). *Consider any joint value distribution $\mathcal{D}$, any $\delta \geq 0$, and any choice distribution that System 2 induces on $[n]$. Then, there is a RUM-with-Ties inducing the same choice distribution. Conversely, for any RUM-with-Ties, for each $\delta > 0$, there exists a joint value distribution $\mathcal{D}$ such that System 2 (as well as System 1) induces the same choice distribution as the RUM-with-Ties.*

3.3 System 1 vs System 2

System 1 and System 2 can be different in the ability to discern the best item in a slate. Clearly, System 2 is never worse than System 1. We will show distributions/slates for which both are identical in power, and distributions/slates such that System 2 can determine the winner while System 1 cannot.

Let $n \geq 3$ and let item $i \in [n]$ have utility value sampled independently and u.a.r. from $[i - \epsilon, i + \epsilon]$ for some $\epsilon \in (0, 1/4)$. Then, System 1 will be able to determine the winner of $S = \{1, 2, 3\}$ w.p. 1 if $\delta < 1 - 2\epsilon$ and w.p. 0 if $\delta > 1 + 2\epsilon$. On the other hand, System 2 will be able to determine the winner of S w.p. 1 if $\delta < 2 - 2\epsilon$ and w.p. 0 if $\delta > 2 + 2\epsilon$. In particular, the two systems coincide in their discriminatory power on S if $\delta < 1 - 2\epsilon$ (in which case, they will both find the winner of S), or if $\delta > 2 + 2\epsilon$ (in which case, they will both result in a no-choice). If $\delta \in (1 + 2\epsilon, 2 - 2\epsilon)$, then System 2 will return the winner of S w.p. 1, while System 1 will return the winner of S w.p. 0. Letting $\epsilon = 0$, this construction can be made deterministic.

4 Inferring preferences

In this section we present algorithms to construct G_2 from System 2 given G_1 from System 1. We obtain a quadratic-time algorithm that returns G_2 containing all and only the arcs $i \to j$ for which the user (who has only "rational" access to G_1) can prove $U_i > U_j$. We then show a quadratic lower bound on the running time of any algorithm that infers all the missing arcs by querying System 1.

4.1 Algorithm

Let $V(G)$ denote the set of nodes in the graph G and let $V = V(G_1)$. For $i \in V$, let $N_1^-(i)$ be the set of predecessors of i in G_1, i.e., nodes with a directed edge to i. Similarly, let $N_1^+(i)$ be set of successors of i in G_1, i.e., nodes with a directed edge from i. Let $\deg_1^-(i)$ (resp., $\deg_1^+(i)$) be the in-degree (resp., out-degree) of node i in the digraph G_1, i.e., $\deg_1^-(i) = |N_1^-(i)|$ and $\deg_1^+(i) = |N_1^+(i)|$.

Before introducing the algorithm, we state a relationship between the perceived utilities and the node degrees in G_1; a version of this statement also appears in [2]. Let $P_{i,j}$ be the predicate $P_{i,j}$ =“$\deg_1^-(i) < \deg_1^-(j)$ or $\deg_1^+(i) > \deg_1^+(j)$”.

Lemma 1 ([2]). *Let $\{i,j\} \in \binom{V}{2}$. Then, (i) $P_{i,j} \implies U_i > U_j$, (ii) $P_{j,i} \implies U_j > U_i$, and (iii) $\overline{P_{i,j} \vee P_{j,i}} \implies N_1^-(i) = N_1^-(j)$ and $N_1^+(i) = N_1^+(j)$. It is then impossible for $P_{i,j} \wedge P_{j,i}$ to hold.*

Let us define the *score* s_i of $i \in V$ to be $s_i = \deg_1^+(i) - \deg_1^-(i)$, i.e., the number of items that i beats minus the number of items that beat i. We point out an important consequence of Lemma 1.

Theorem 3 ([2]). *Let $\{i,j\} \in \binom{V}{2}$. Then, (i) $s_i > s_j \implies U_i > U_j$, (ii) $s_i < s_j \implies U_j > U_i$, and (iii) $s_i = s_j \implies N_1^-(i) = N_1^-(j)$ and $N_1^+(i) = N_1^+(j)$.*

We now present Algorithm 1 and prove that it produces a correct and optimal inference from G_1.

Algorithm 1 An algorithm for producing graph G_2, given G_1.

1: $V \leftarrow V(G_1)$
2: **for all** $i \in V$ **do**
3: $\quad s_i \leftarrow \deg_1^+(i) - \deg_1^-(i)$ {score}
4: $E \leftarrow \{i \to j \mid \forall i,j \in V \text{ s.t. } s_i > s_j\}$
5: **return** $G_2(V,E)$

Theorem 4. *The graph G_2 produced by Algorithm 1 contains all and only the arcs $i \to j$ for which one can prove, given access to G_1, that $U_i > U_j$. Moreover, in general, for no $\{i,j\} \in \binom{V}{2}$ it is possible to prove that $U_i = U_j$.*

We remark that the first part of Theorem 4 is implicit in [2]. Also note that Algorithm 1 can be implemented to run in quadratic time; a constructive version of the method in [35] needs cubic time.

Theorem 5. *Algorithm 1 runs in time $O(|V|^2)$.*

Note that once the scores of the nodes are computed, the bucket order corresponding to G_2 (see Theorem 1) can be constructed in time $O(n \log n)$ by sorting the nodes by scores.

4.2 Lower bound

Next, we prove a lower bound on the query complexity of any algorithm that aims to infer all the missing arcs by querying System 1. In fact, we show the same lower bound on the number of calls to System 1, even to find the best item.

Theorem 6. *An algorithm that queries System 1 on pairs of items needs $\Omega(|V|^2)$ queries in expectation to determine the best item of V, even in settings where System 2 is guaranteed to be able to determine the best item of V.*

5 The power of RUMs-with-Ties

RUMs-with-Ties are at least as powerful as classical RUMs by definition, but the precise relationship between these models is not obvious. In this section we explore different measures of the power and complexity of RUMs-with-Ties.

We begin in Section 5.1 by attempting to replicate the behavior of RUMs-with-Ties by *augmenting* regular RUMs with a "no-choice" element. In Section 5.2, we consider a related question, attempting to use RUMs to replicate the behavior of RUMs-with-Ties specifically in cases for which the RUM-with-Ties gives a winner (i.e., does not return "no-choice"). We show that RUMs-with-Ties are not approximable by RUMs in either setting. Next, in Section 5.3, we ask how many bits are required to specify almost perfectly the behavior of a RUM-with-Ties—the ability to introduce some small errors in the approximation provides a more robust measure that is not subject to corner cases of perfect fidelity. We give tight bounds on the representation complexity, essentially characterizing the compressibility of RUMs-with-Ties. These results show that RUMs-with-Ties can be written down in essentially the same space as RUMs. Finally, in Section 5.4, we ask whether the ability of RUMs-with-Ties to employ the additional "no-choice" tag is a blessing or a curse in terms of the information available to the model user.

5.1 RUMs-with-Ties vs augmented RUMs

A natural idea is to augment $[n]$ of the RUM with a special "no-choice" item $\perp$ [22]. This augmented RUM (with $n+1$ items) is only applied to slates containing $\perp$: when $\perp$ is chosen from a slate, we say that a "no-choice" happened in the original slate.

We prove that RUMs-with-Ties are more expressive than augmented RUMs. The difference is due to the non-monotonicity of the "no-choice" event in RUMs-with-Ties. Indeed, in an augmented RUM R, if $S \ni \perp$ is one of its slates, then given any $T \supset S$ it must be that $D_S(\perp) \geq D_T(\perp)$, i.e., the probability of a tie is monotonically decreasing. In a RUM-with-Ties, instead, the probability of a tie satisfies no such property, e.g., even if no tie could have happened in S, it could be that the top-most element of $T \setminus S$ is always tied with the top-most element of S. We formalize this next.

Lemma 2. *For $n = 3$, there exists a a RUM-with-Ties R on $[n]$ such that, for each augmented RUM R' on $[n] \cup \{\perp\}$, there exists at least one slate $S \subseteq [n]$, such that the ℓ_∞-distance between R's distribution on S and the distribution on $S \cap \{\perp\}$ given by R' is at least $1/2$.*

5.2 RUMs-with-Ties vs projected RUMs

A different natural reduction is to project a RUM-with-Ties T to eliminate the ties: given a set $\{S_1, \ldots, S_t\}$ of slates with their observed (no-)choice distributions $T_{S_1}, \ldots, T_{S_t}$, condition each distribution on not causing a tie; let $T'_{S_1}, \ldots, T'_{S_t}$

be the resulting distributions. Then, the hope is to fit a RUM to $T'_{S_1}, \ldots, T'_{S_t}$ obtaining a representation of all the winning events and represent the "no-choice" events separately. We prove that the error induced by projected RUMs can be large. The construction uses the fact that such a conditioning can cut off a large part of the probability space, so that the conditioning blows up the probability of some rare events.

Lemma 3. *For each $\epsilon > 0$, and for each large enough n, there exists a RUM-with-Ties T on $[n]$ whose no-ties conditioned winning distributions $T'_{S_1}, \ldots, T'_{S_{2^n}}$ have the property that, for each RUM R on $[n]$, there exists at least one slate $S_i \subseteq [n]$, such that the ℓ_∞-distance between S_i and R's distribution on S_i is at least $1 - \epsilon$.*

5.3 Succinctly representing RUMs-with-Ties

We next consider the problem of approximately representing RUMs-with-Ties with few bits. In particular, we show that RUM-with-Ties on the ground set $[n]$ can be sketched to $O(n^2 \cdot \epsilon^{-2} \log n)$ bits. This generalizes the result in [14] for classical RUMs.

Theorem 7. *Let $0 < \epsilon, \delta < 1$. There is a polynomial-time algorithm that, given any distribution D on $\mathcal{B}_n$, produces a multiset B of $O\left(\epsilon^{-2} \cdot \left(n + \ln \delta^{-1}\right)\right)$ bucket orders such that, with probability $\geq 1 - \delta$, the uniform distribution $\widetilde{D}$ on B guarantees that $\operatorname{dist}(D, \widetilde{D}) \leq \epsilon$.*

The algorithm is to sample independent bucket orders $B_1, \ldots, B_t$ from D, where $t = O(\epsilon^{-2} \cdot (n + \ln \delta^{-1}))$, and to let $\widetilde{D}$ be the uniform distribution on this multiset of samples.

An easy consequence of Theorem 7 is that any RUM-with-Ties can be approximately represented using $O(n^2 \log n)$ bits, since a bucket order can be represented using $O(n \log n)$ bits.

Corollary 1. *For each $0 < \epsilon < 1$, and for each RUM-with-Ties D, there is a data structure using $O(\epsilon^{-2} \cdot n^2 \log n)$ bits that can be used to return, for each slate $S \subseteq [n]$, a distribution $\widetilde{D}_S$ satisfying $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \epsilon$.*

Using the lower bound construction in [14], Corollary 1 can be shown to be near-optimal.

Corollary 2. *Fix a constant $0 < \alpha \leq 1/2$. A data structure for a generic RUM-with-Ties D that can be used, for each slate $S \subseteq [n]$, to return a distribution $\widetilde{D}_S$ satisfying $|D_S - \widetilde{D}_S|_{\mathrm{tv}} \leq \frac{1-\alpha}{4}$, requires at least $\frac{\alpha^3}{6} \cdot n^2 - 1$ bits.*

5.4 Comparing the information present in responses from RUMs-with-Ties vs RUMs

Our goal is to understand whether the addition of a "no-choice" response represents additional information returned by a RUM-with-Ties, or whether the loss of perfect information on the max-utility element of the slate represents a loss of information. To study this, we rely on the hypotheses testing framework and ask whether the responses from a RUM-with-Ties can more efficiently differentiate between two underlying data models than the responses from a RUM. For simplicity, instead of RUMs, we study MNLs vs MNLs-with-Ties in Systems 1 and 2. We show that, for some parameter choice, MNLs-with-Ties allow for more efficient predictions, while for some others, MNLs have more efficient predictions.

Lemma 4. *There exist settings where, using a single sample, (i) with MNLs, no algorithm can guess correctly with probability larger than* $3/4$*, while (ii) MNL-with-Ties can be used to guess the unknown hypothesis with probability* $1 - o(1)$*.*

Lemma 5. *There exist settings where, using a single sample, (i) MNLs can be used to guess the unknown hypothesis with probability at least* $3/4 - o(1)$*, while (ii) with MNLs-with-Ties, the probability of guessing correctly is at most* $1/2 + o(1)$*.*

6 Learning the best element of pair

In this section we focus on algorithms to learn the highest utility member of a pair in a δ-MNL. Let the two objects have base values x, y (wlog, $x < y$) and we will have access to outcomes of 2-slates by different users in the MNL-with-ties model. Let $\delta > 0$ be the indifference threshold and $c := \frac{y-x}{\delta}$.

Let $X \sim \text{Gumbel}(x, 1)$ and $Y \sim \text{Gumbel}(y, 1)$ be two independent r.v.'s. Then, $X - Y \sim \text{Logistic}(x - y, 1)$. We can show that if δ is large, we have a tie whp., making it necessary to sample the outcome a large number of times in order to infer if $y > x$. Suppose, instead, we can leverage a third value $Z \sim \text{Gumbel}(z, 1)$ where $x - \delta < z < y - \delta$. Then, if $\delta \gg 1$, we have that $|X - Y| \leq \delta$ (tie) whp.; conditioned on this, $|X - Z| \leq \delta$ (tie) whp.; conditioned on both, $Y - Z > \delta$ (win) whp. Thus we can infer $y > x$. E.g., if $x = \delta$, $y = 2\delta$, we can pick $z = \delta/2$ as a "lower quality" pivot value. By carefully choosing a pivot, we can significantly decrease the number of necessary slates for inference.

A simple Pairs algorithm can check N many independent 2-slates of the original two objects and then output the final winner as the object that got strictly more wins than the other.

Lemma 6. *Given* $\rho \in (0, 1)$*, the* Pairs *algorithm outputs "Y wins" with probability* $\geq 1 - \rho$ *if the number of slates is* $N_{\text{Pairs}} = 2\ln(1/\rho) \cdot \left(\frac{1}{1+e^{(1-c)\delta}} - \frac{1}{1+e^{(1+c)\delta}}\right)^{-2}$.

Suppose we have access to a pivot object Z with base value $z := (x+y)/2 + \delta = x + \delta \cdot (1 + c/2) = y + \delta \cdot (1 - c/2)$ (it would be analogous for the "symmetric" case $z = \frac{x+y}{2} - \delta$, with lower base value than x, y if $c < 1$). The Pivot algorithm checks N independent 2-slates: $N/2$ of the first object against the pivot and $N/2$ of the second object against the pivot. It tracks the pivot and outputs the final winner as the object who got the least losses against the pivot.

Lemma 7. *Given $\rho \in (0,1)$, the* Pivot *algorithm outputs "Y wins" with probability $\geq 1-\rho$ if the number of slates is* $N_{\textsc{Pivot}} = 4\ln(2/\rho)\cdot\left(\frac{1}{1+e^{-c\delta/2}} - \frac{1}{1+e^{c\delta/2}}\right)^{-2}$.

We can show that Pivot can outperform Pairs in terms of the slates used for inference by an unbounded factor.

Lemma 8. *Given $A > 1$, if $0 < c < 1$ and $\delta \geq \frac{4\ln(2)+\ln(A)}{1-c}$, then $N_{\textsc{Pairs}} > A^2 N_{\textsc{Pivot}}$.*

We can further generalize Lemma 8 as follows. Consider the following two settings. In the first (as in Pairs), we can only see outcomes of 2-slates comparing the two original objects, and we denote with $\Psi_{\textsc{Pairs}}$ the minimum number of samples needed to correctly guess the winner with probability $\geq 2/3$. In the second (as in Pivot), we are allowed to pick a pivot object to compare against any of the two original objects in a slate with size 2. We consider the experiment in which we fix a pivot and, each time, we compare it to a random object of the pair and see the outcome; again, let $\Psi_{\textsc{Pivot}}$ be the number of samples needed in this second setting for correctness with probability $\geq 2/3$.

Theorem 8. *Given an arbitrarily large $A > 1$, if $0 < c < 1$ and $\delta \geq \frac{4\ln(2)+\ln(G\cdot(A+12/c))}{1-c}$, it holds $\Psi_{\textsc{Pairs}} \geq A \cdot N_{\textsc{Pivot}} \geq A \cdot \Psi_{\textsc{Pivot}}$, where G is an absolute constant.*

Theorem 8 shows that the number of samples needed for decision with a pivot can be arbitrarily smaller than that needed with only pairs of the original objects. Moreover, Pivot algorithm can perform much better than *any* algorithm with access only to slates of the original objects.

7 Experiments with modeling ties in practice

In this section we provide a summary of an experimental analysis of RUM-with-Ties in contexts where user preferences induce bucket orders over the set of available items. We provide experimental evidence in support of the use of RUMs-with-Ties in the setting of indistinguishability of close alternatives. We have made the code publicly available on GitHub[7] for reproducibility.
Our experiments are over several real-world "user ratings" datasets: sushi [29], containing ratings on 100 sushi ingredients; travel [44], containing TripAdvisor reviews scores from 980 users on 10 activities; *young people spending habits (*ypsh*)* [52], containing a student survey on spending habits for 7 different items (e.g., branded clothing, shopping); goodbooks[8], containing ratings on 10 000 books; movielens[9], containing ratings from 270 000 users on 45 000 movies.

We interpret the ratings given by a user as utility values, with a fixed indifference threshold δ. We split each dataset randomly into training (80%) and test (20%).

[7] https://github.com/re-gius/rums-with-ties
[8] https://github.com/zygmuntz/goodbooks-10k
[9] https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

We consider the RUM-with-Ties induced by the training dataset (train-RUMwt). For comparison, we consider the following algorithms. The first, MNL++, approximates a RUM-with-Ties using an augmented MNL, i.e., an MNL augmented with the "no-choice" item [22, 25], learned from a training set using standard stochastic gradient descent. The second, δ-MNL, is an MNL with an indifference threshold δ [33]: the utilities are computed as output values of an MNL, but a winner is selected if and only if its value is at least δ more than all the other values; otherwise, the outcome is "no-choice." . The third, LP-RUMwt, solves the following LP to learn a RUM-with-Ties.[10]

$$\begin{cases} \min \frac{1}{2} \cdot \sum\limits_{\varnothing \neq S \subseteq [n]} \sum\limits_{x \in S \cup \{\perp\}} \delta_{S,x} \\ -\delta_{S,x} \leq D_S(x) - \sum\limits_{\substack{B \in \mathcal{B}_n \\ x = B(S)}} p_B \leq \delta_{S,x}, \;\; \forall S \subseteq [n], \forall x \in S \cup \{\perp\} \\ \sum\limits_{B \in \mathcal{B}_n} p_B = 1 \text{ and } p_B \geq 0 \; \forall B \in \mathcal{B}_n \end{cases} \tag{1}$$

As ground-truth, we use the RUM-with-Ties induced by the whole dataset (OPT-RUMwt). We compare the algorithms against OPT-RUMwt on the slates test set. Each of these models induces a distribution over the outcomes of each slate in the slates test set, also known as stochastic choice [17]. We compare those distributions according to the average KL-divergence between the ground-truth given by OPT-RUMwt and a models' distribution.

dataset	MNL++	δ-MNL	LP-RUMwt	train-RUMwt
SUSHI	7.9	7.3	**4.5**	**0.39**
YPSH	5.4	3.1	**2.3**	**0.026**
TRAVEL	5.5	8.9	**3.6**	**0.073**

Table 1. Average KL-divergence ($\times 10^{-3}$) achieved by the models on the slates test set. The best overall metric is in bold (always train-RUMwt). The best metric for models learned on slates (i.e., excluding train-RUMwt) is in blue.

From Table 1, we can see that train-RUMwt is by far the best model. This finding strongly supports the use of RUMs-with-Ties to model ties. However, in some cases we only have access to a fixed set of slates with their outcomes: in this setting, we cannot learn train-RUMwt, but only the other models. Thus, it is also useful to restrict the comparison to MNL++, δ-MNL, LP-RUMwt. LP-RUMwt consistently outperforms MNL++ and δ-MNL in terms of average KL-divergence, suggesting that it is better than all previous models for learning the choice distribution of a random slate. However, there is still a significant gap between LP-RUMwt and train-RUMwt. This highlights the opportunity to develop new algorithms to reduce the complexity of the previous LP so to make LP-RUMwt scalable to larger datasets but closer to train-RUMwt in terms of performance.

Finally, we use Algorithm 1 to compute the actual number of new inferences obtained in System 2. For each user u, we consider the average review scores

[10] We cannot compare directly to [38] since they provide an adaptive online learning algorithm, which assumes to have access to outcomes of any slate. Our datasets, instead, only contain a subset of all possible slates. Moreover, we focus on offline learning algorithms.

$\{r_u^i\}_{1\le i\le 10}$ as sampled utilities from its value distribution over the categories. For a fixed indifference threshold $\delta > 0$, we say that (i,j) is inferred by System 1 if $\|r_u^i - r_u^j| > \delta$; we say that (i,j) is inferred by System 2 if $i \to j$ or $j \to i$ in the G_2 graph returned by Algorithm 1. For each user u, we can therefore define the number of inferences from System 1, denoted $S_1(u,\delta)$, and from System 2, denoted $S_2(u,\delta)$. We report the average over all users of the incremental percentage of new inferences of System 2 with respect to System 1 for different values of δ: $\text{avg}_u \frac{S_2(u,\delta)-S_1(u,\delta)}{S_1(u,\delta)}\%$.

dataset	$\delta = 0.25$	$\delta = 0.5$	$\delta = 1$	$\delta = 2$
TRAVEL	2.8	10.0	25.3	-
MOVIELENS	-	6.3	42.4	-
GOODBOOKS	-	-	26.4	44.7

Table 2. Average percentage of new inferences with respect to System 1 provided by Algorithm 1 for different values of δ.

Table 2 reports the results of the experiment for the datasets. Notice that, even for seemingly low and realistic values of δ, Algorithm 1 gives significantly more inferred preferences, on average, with respect to System 1. This gives experimental evidence of the importance of System 2 new inferences and the need for an optimally efficient algorithm to compute them like our Algorithm 1.

References

1. S. Agarwal. On ranking and choice models. In *IJCAI*, pages 4050–4053, 2016.
2. F. Aleskerov, D. Bouyssou, and B. Monjardet. *Utility Maximization, Choice and Preference*. Springer, 2007.
3. M. Almanza, F. Chierichetti, R. Kumar, A. Panconesi, and A. Tomkins. RUMs from head-to-head contests. In *ICML*, pages 452–467, 2022.
4. S. P. Anderson, A. De Palma, and J.-F. Thisse. *Discrete Choice Theory of Product Differentiation*. MIT Press, 1992.
5. W. E. Armstrong. A note on the theory of consumer's behavior. *Oxford Economic Papers*, 2(1):119–122, 1950.
6. F. J. Bahamonde-Birke, I. Navarro, and J. de Dios Ortúzar. If you choose not to decide, you still have made a choice. *J. of Choice Modelling*, 22:13–23, 2017.
7. D. Bang, K. Baek, J. Kim, Y. Jeon, J.-H. Kim, J. Kim, J. Lee, and H. Shim. Logit mixing training for more reliable and accurate prediction. In *IJCAI*, pages 2812–2819, 2022.
8. A. R. Benson, R. Kumar, and A. Tomkins. On the relevance of irrelevant alternatives. In *WWW*, pages 963–973, 2016.
9. A. R. Benson, R. Kumar, and A. Tomkins. A discrete choice model for subset selection. In *WSDM*, page 37–45, 2018.
10. H. D. Block and J. Marschak. Random orderings and stochastic theories of response. In *Economic Information, Decision, and Prediction: Selected Essays: Volume I Part I Economics of Decision*, pages 172–217, 1974.
11. V. Cantillo, J. Amaya, and J. d. D. Ortúzar. Thresholds and indifference in stated choice surveys. *Trans. Res. Part B: Meth.*, 44(6):753–763, 2010.
12. V. Cantillo and J. d. D. Ortúzar. Implications of thresholds in discrete choice modelling. *Trans. Rev.*, 26(6):667–691, 2006.

13. F. Chierichetti, M. Giacchini, R. Kumar, A. Panconesi, and A. Tomkins. Approximating a rum from distributions on k-slates. In *AISTATS*, pages 4757–4767. PMLR, 2023.
14. F. Chierichetti, R. Kumar, and A. Tomkins. Light RUMs. In *ICML*, pages 1888–1897, 2021.
15. F. Chierichetti, R. Kumar, and A. Tomkins. On the number of trials needed to distinguish similar alternatives. *PNAS*, 119(31), 2022.
16. R. Corbin and A. Marley. Random utility models with equality: An apparent, but not actual, generalization of random utility models. *J. Math. Psych.*, 11(3):274–293, 1974.
17. G. Debreu. Stochastic choice and cardinal utility. *Econometrica*, pages 440–444, 1958.
18. R. Dhar. Consumer preference for a no-choice option. *J. Consum. Res.*, 24(2):215–231, 1997.
19. J.-C. Falmagne. A representation theorem for finite random scale systems. *J. Math. Psych.*, 18(1):52–72, 1978.
20. G. T. Fechner. *Vorschule der aesthetik*. Breitkopf & Härtel, Leipzig, 1876.
21. Z. Feng, Y. Zhu, and J. Cao. Modeling air travel choice behavior with mixed kernel density estimations. In *WSDM*, pages 611–620, 2017.
22. S. Gao, E. Frejinger, and M. Ben-Akiva. Cognitive cost in route choice with real-time information: An exploratory analysis. *Procedia-Social and Behav. Sci.*, 17:136–149, 2011.
23. D. H. Gensch and W. W. Recker. The multinomial, multiattribute logit choice model. *J. Mark. Res.*, 16(1):124–132, 1979.
24. M. Haghani, M. C. Bliemer, and D. A. Hensher. The landscape of econometric discrete choice modelling research. *J. Choice Modelling*, 40:100303, 2021.
25. S. Hess, M. J. Beck, and C. G. Chorus. Contrasts between utility maximisation and regret minimisation in the presence of opt out alternatives. *Transp. Res. Rec. P. A: Policy & Practice*, 66:1–12, 2014.
26. S. S. Iyengar and M. R. Lepper. When choice is demotivating: Can one desire too much of a good thing? *J. Pers. Soc. Psy.*, 79(6):995, 2000.
27. A. Jiang, L. Soriano Marcolino, A. D. Procaccia, T. Sandholm, N. Shah, and M. Tambe. Diverse randomized agents vote to win. *NIPS*, 2014.
28. D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus & Giroux, 2011.
29. T. Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *KDD*, pages 583–588, 2003.
30. K. Krishnan. Incorporating thresholds of indifference in probabilistic choice models. *Man. Sci.*, 23(11):1224–1233, 1977.
31. R. Kumar, M. Mahdian, B. Pang, A. Tomkins, and S. Vassilvitskii. Driven by food: Modeling geographic choice. In *WSDM*, pages 213–222, 2015.
32. I. Labutov, F. Schalekamp, K. Luu, H. Lipson, and C. Studer. Optimally discriminative choice sets in discrete choice models: Application to data-driven test design. In *KDD*, pages 1665–1674, 2016.
33. S. Lioukas. Thresholds and transitivity in stochastic consumer choice: a multinomial logit analysis. *Man. Sci.*, 30(1):110–122, 1984.
34. A. Liu, Z. Zhao, C. Liao, P. Lu, and L. Xia. Learning Plackett–Luce mixtures from partial preferences. In *AAAI*, pages 4328–4335, 2019.
35. R. D. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24(2):178–191, 1956.

36. J. Ma, X. Yi, W. Tang, Z. Zhao, L. Hong, E. Chi, and Q. Mei. Learning-to-rank with partitioned preference: fast estimation for the Plackett–Luce model. In *AISTATS*, pages 928–936, 2021.
37. J. Ma, X. Zhang, and Q. Mei. Fast learning of mnl model from general partial rankings with application to network formation modeling. In *WSDM*, pages 715–725, 2022.
38. Q. P. Nguyen, S. Tay, B. K. H. Low, and P. Jaillet. Top-k ranking Bayesian optimization. In *AAAI*, pages 9135–9143, 2021.
39. M.-h. Oh and G. Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *AAAI*, pages 9205–9213, 2021.
40. E. A. Ok and G. Tserenjigmid. Indifference, indecisiveness, experimentation, and stochastic choice. *Theor. Econ.*, 17(2):651–686, 2022.
41. M. Ou, N. Li, S. Zhu, and R. Jin. Multinomial logit bandit with linear utility functions. In *IJCAI*, pages 2602–2608, 2018.
42. D. C. Parkes, H. A. Soufiani, and L. Xia. Random utility theory for social choice. In *NIPS*, 2012.
43. E. A. Patall, H. Cooper, and J. C. Robinson. The effects of choice on intrinsic motivation and related outcomes: a meta-analysis of research findings. *Psych. Bull.*, 134(2):270, 2008.
44. S. Renjith, A. Sreekumar, and M. Jathavedan. Evaluation of partitioning clustering algorithms for processing social media data in tourism domain. In *RAICS*, pages 127–131, 2018.
45. A. Saha and A. Gopalan. Best-item learning in random utility models with subset choices. In *AISTATS*, pages 4281–4291, 2020.
46. B. Schwartz. *The Paradox of Choice: Why More Is Less.* Harper Perennial, 2005.
47. A. Seshadri, S. Ragain, and J. Ugander. Learning rich rankings. *NeurIPS*, 33:9435–9446, 2020.
48. Y. Sim, B. R. Routledge, and N. A. Smith. The utility of text: the case of amicus briefs and the supreme court. In *AAAI*, 2015.
49. A. Sklar. On the factorization of squarefree integers. *Proc. AMS*, 3(5):701–705, 1952.
50. K. E. Train. *Discrete Choice Methods with Simulation.* Cambridge University Press, 2003.
51. C. Turansick. Identification in the random utility model. *J. Econ. Th.*, page 105489, 2022.
52. L. Vilčeková and M. Sabo. The influence of demographic factors on attitudes toward brands and brand buying behavior of Slovak consumers. *Intl. J. Ed. Res.*, 1(11):1–10, 2013.
53. C. Weis, M. Kowald, A. Danalet, B. Schmid, M. Vrtic, K. W. Axhausen, and N. Mathys. Surveying and analysing mode and route choices in Switzerland 2010–2015. *Travel Beh. Soc.*, 22:10–21, 2021.
54. Y. Yu, K. Han, and W. Ochieng. Day-to-day dynamic traffic assignment with imperfect information, bounded rationality and information sharing. *Trans. Res. Part C: Emerging Tech.*, 114:59–83, 2020.
55. Z. Zhao, A. Liu, and L. Xia. Learning mixtures of random utility models with features from incomplete preferences. In *IJCAI*, pages 3780–3786, 7 2022.
56. Z. Zhao, T. Villamil, and L. Xia. Learning mixtures of random utility models. In *AAAI*, 2018.
57. Z. Zhao and L. Xia. Learning mixtures of Plackett–Luce models from structured partial orders. In *NeurIPS*, 2019.