# Hypothesis Testing for Generalized Thurstone Models

**Anuran Makur**$^{1,2}$**, Japneet Singh**$^{2}$ [*]
$^{1}$ Department of Computer Science,
$^{2}$ Elmore Family School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47906
amakur@purdue.edu, sing1041@purdue.edu

## Abstract

In this work, we develop a rigorous hypothesis testing method to determine whether pairwise comparison data is generated by an underlying *generalized Thurstone model* $\mathcal{T}_F$ for a given choice function $F$. Given $n$ agents, a $\mathcal{T}_F$ model assumes that each agent $i$ has a latent utility parameter $w_i$ and the probability that agent $i$ is preferred over agent $j$ in a pairwise comparison (e.g., a game) is given by $F(w_i - w_j)$. While prior work has predominantly focused on parameter estimation and uncertainty quantification for such models, our work bridges a crucial gap by developing a hypothesis testing approach for $\mathcal{T}_F$ models. We formulate this testing problem in a minimax sense by introducing a notion of separation distance between a general pairwise comparison model and the class of $\mathcal{T}_F$ models. We then derive both upper and lower bounds on the critical threshold of our minimax hypothesis testing problem, which depend on the topology of the underlying observation graph of comparisons. For example, in the setting where all possible pairwise comparisons are observed (i.e., complete observation graph), the critical threshold scales as $\Theta((nk)^{-1/2})$, where $k$ is the number of pairwise comparisons between each pair of agents. Furthermore, we propose a specific hypothesis test inspired by our separation distance for our testing problem, and assess its performance by establishing "time-uniform" upper bounds on type I and type II error probabilities using reverse martingale ideas. To complement this, we also develop a minimax risk lower bound for our testing problem using information-theoretic ideas. Additionally, we prove several auxiliary results over the course of our analysis, such as $\ell^2$-bounds on parameter estimation and "time-uniform" confidence intervals. Finally, we conduct several experiments on synthetic and real-world datasets to validate some of our theoretical results and test for $\mathcal{T}_F$ models. In the process, we also propose a data-driven approach to find the threshold of our test.

## 1 Introduction

Learning rankings from data is a fundamental problem underlying numerous applications, including recommendation systems [25], sports tournaments [4, 6], fine-tuning large language model (LLMs) [38], and social choice theory [28, 51]. The class of generalized Thurstone models (GTMs) [47, 37, 32], which fall under the broader framework of random utility models, is a widely adopted framework for ranking agents, items, or choices based on given preference data. GTMs include many other models as special cases, most notably the Bradley-Terry-Luce (BTL) model [4, 28, 33], which has been widely studied. Given $n$ agents $[n] = \{1, \ldots, n\}$, GTMs can be construed as likelihood models for pairwise comparisons between pairs of agents. In particular, a GTM $\mathcal{T}_F$ assumes that each agent $i$ is endowed with an unknown utility parameter $w_i \in \mathbb{R}$ and the probability that agent $i$ is preferred

[*] The author ordering is alphabetical.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

over agent $j$ (e.g., $i$ beats $j$ in a game) is given by $F(w_i - w_j)$, where $F$ represents a known choice function which is a cumulative distribution function (CDF).

While GTMs have been utilized in many contexts, e.g., [13, 41], they are parametric models where $n$ utility parameters characterize the model. Indeed, the assumption that pairwise comparison data is governed by a small number of parameters forms the basis of most results on GTMs [6, 50, 45, 24]. But such parametric models can sometimes be too stringent to capture the intricacies of real applications [10, 34, 49], and non-parametric models, e.g., [7, 45], have been studied as an alternative. This conversation raises an important question: *Given pairwise comparison data, can we determine whether it is comes from a specific GTM?* If it does, then we can rely on the vast GTM literature for learning, and if it does not, then we can resort to using non-parametric models.

Despite extensive research in the area, there is no systematic answer to the above question in the literature, i.e., there is no rigorously analyzed hypothesis test to determine whether given pairwise comparison data conforms to an underlying GTM model. To address this, we study the composite hypothesis testing problem of whether data obeys a GTM $\mathcal{T}_F$ for a given choice function $F$:

$$\begin{aligned} &H_0 : \mathcal{Z} \sim \mathcal{T}_F \text{ for some choice of } w \in \mathcal{W}, \\ &H_1 : \mathcal{Z} \sim \text{general pairwise comparison model that is not } \mathcal{T}_F, \end{aligned} \tag{1}$$

where $\mathcal{Z}$ denotes the pairwise comparison data, and $H_0$ and $H_1$ are the null and alternative hypotheses, respectively and $\mathcal{W}$ denotes the parameter set for weight $w$.

**Main contributions.** We analyze the composite hypothesis testing problem outlined in (1). Our main contributions include the following: **1)** We frame the hypothesis testing problem in a minimax sense (Section 2) by developing a rigorous notion of separation distance to the class of all GTMs that admits tractable analysis (Section 3, Theorem 1). **2)** We derive upper and lower bounds on the critical threshold for our test (Section 3, Theorem 2 and Proposition 2). These bounds exhibit a dependence on the graph induced by the pairwise comparison data (see Section 1) and are tight for complete graphs. **3)** We use the separation distance to propose a hypothesis test and establish various theoretical guarantees for our test. Specifically, we prove "time-uniform" type I and type II error probability upper bounds for our test (Section 3, Theorems 4 and 5), and also provide a minimax lower bound. **4)** Additionally, we obtain auxiliary results like $\ell^2$-error bounds on parameter estimation for general pairwise comparison models (Theorem 3) and "time-uniform" confidence intervals under the null hypothesis (Proposition 4). **5)** Finally, we validate our theoretical findings through synthetic and real-world experiments, proposing a data-driven approach to determine the test threshold and using the test to determine different choice functions' fit to the data (Section 4).

**Related literature.** The class of GTMs has a rich history in the analysis of preference data. Initially proposed by Thurstone [47], these models are widely used in various fields, ranging from psychology [48], economics [31], and more recent applications like aligning LLMs with human preferences [38]. Early foundational works, e.g., [47, 28, 51], explored different cumulative distribution functions $F$ for modeling choice probabilities, including Gaussian [47], logistic [4], and Laplace [11]. These models and their extensions underlie popular rating systems, such as Elo in chess [52, 13, 41] and TrueSkill in video games [17]. Several recent works have actively explored estimation techniques for Thurstone models. For instance, [50] estimated parameters of Thurstone models when the preference data is derived from general subsets of agents (not specifically pairs), and [45] focused on parameter estimation for GTMs and the effect of graph topology on the estimation accuracy.

Furthermore, a significant portion of the literature has focused on parameter estimation in the special case of the BTL model, e.g., [46, 43, 36, 8, 24], where two prominent algorithms are spectral ranking [43, 36] and maximum likelihood estimation [52, 22]. Another related line of work is on uncertainty quantification for estimated parameters [16, 20, 14, 27]. For example, [16] established the asymptotic normality of estimated parameters in the BTL model for both spectral ranking and maximum likelihood estimation, and [20] generalized the asymptotic normality results to a broader class of models such as GTMs and Mallows models.

Despite the extensive work on parameter estimation, relatively few studies have rigorously investigated hypothesis testing for such parametric models. In particular, [40] developed two-sample tests for preference data, [44] studied lower bounds for testing the independence of irrelevant alternatives (IIA) assumption (i.e., BTL and Plackett-Luce models [28, 39]), and [29] developed hypothesis tests for BTL models based on spectral methods. In contrast to these works, we develop hypothesis testing for GTMs using a maximum likelihood framework, complementing the work in [29].

Table 1: Bounds in this work on critical threshold $\varepsilon_{\mathsf{c}}$ for various induced observation graphs, where $n$ represents the number of agents and $k$ is the number of comparisons between agents per pair.

| | Complete graph | Constant-degree expander | Single cycle | Toroidal grid |
|---|---|---|---|---|
| Upper bound | $O\left(\frac{1}{\sqrt{nk}}\right)$ | $O\left(\frac{1}{\sqrt{nk}}\right)$ | $O\left(\sqrt{\frac{n}{k}}\right)$ | $O\left(\frac{1}{\sqrt{k}}\right)$ |
| Lower bound | $\Omega\left(\frac{1}{\sqrt{nk}}\right)$ | $\Omega\left(\frac{1}{\sqrt{n^2k}}\right)$ | $\Omega\left(\frac{1}{\sqrt{n^2k}}\right)$ | $\Omega\left(\frac{1}{\sqrt{n^{7/4}k}}\right)$ |

## 2 Formal model and setup

We begin by introducing a general pairwise comparison model that provides a flexible framework encompassing a broad range of established probabilistic models, including the BTL model [4, 28, 33], the Thurstone model [47], and non-parametric models [7, 45]. In this framework, we consider $n \in \mathbb{N}\backslash\{1\}$ agents (or items or choices) $[n]$ engaged in pairwise comparisons. For agents $i, j \in [n]$ with $i \neq j$, let $p_{ij} \in (0, 1)$ denote the probability that $i$ is preferred over $j$ in an "$i$ vs. $j$" pairwise comparison. This model inherently captures the asymmetric nature of pairwise comparisons, as the outcome of an "$i$ vs. $j$" comparison may differ from that of a "$j$ vs. $i$" comparison. This reflects real-world phenomena like "home advantage" that are commonly observed in sports [1, 35]. To model the fact that not all pairwise comparisons may be observed, we assume that we are given an induced observation graph $\mathcal{G} = ([n], \mathcal{E})$, where an edge $(i, j) \in \mathcal{E}$ (with $i \neq j$) exists if and only if comparisons of the form "$i$ vs. $j$" are observed. Let $E \in \{0, 1\}^{n\times n}$ be the adjacency matrix of $\mathcal{G}$, with $E_{ij} = 1$ if $(i, j) \in \mathcal{E}$ and 0 otherwise. Furthermore, we assume that the edge set $\mathcal{E}$ is symmetric (i.e., $\mathcal{G}$ is undirected), implying that if "$i$ vs. $j$" comparisons are observed, then "$j$ vs. $i$" comparisons are observed as well. Additionally, we assume that $\mathcal{G}$ is connected and is fixed a priori (see Proposition 1), independent of the outcomes of observed pairwise comparisons. Also, let $D \in \mathbb{R}^{n\times n}$ the diagonal degree matrix with $D_{ii} = \sum_{j=1}^{n} E_{ij}$ for $i \in [n]$, and $L \triangleq D - E$ be the graph Laplacian matrix. $L$ can be expressed as $L = X^{\mathrm{T}}X$, where $X \in \mathbb{R}^{(|\mathcal{E}|/2)\times n}$ is the matrix formed by collecting row vectors $x_{ij} = e_i - e_j$ for $(i, j) \in \mathcal{E}$ and $j > i$, with $e_i$ being the $i$th standard basis vector in $\mathbb{R}^n$. We now define a general pairwise comparison model.

**Definition 1** (Pairwise Comparison Model)**.** *Given an observation graph $\mathcal{G}$ over the agents $[n]$, we refer to the collection of probability parameters $\{p_{ij} : (i, j) \in \mathcal{E}\}$ as a* pairwise comparison model*.*

Furthermore, we can represent a pairwise comparison model by a *pairwise comparison matrix* $P \in [0, 1]^{n\times n}$ with

$$P_{ij} \triangleq \begin{cases} p_{ij}, & (i,j) \in \mathcal{E}, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

We remark that our ensuing analysis can be easily specialized to a symmetric setting where "$i$ vs. $j$" and "$j$ vs. $i$" comparisons are equivalent. In this case, $E$ is automatically symmetric as assumed. On the other hand, the symmetry assumption on $E$ is needed in asymmetric settings because GTMs inherently treat "$i$ vs. $j$" and "$j$ vs. $i$" comparisons as equivalent, which is not true in general models.

**GTM model.** Next, we describe a GTM for a choice function $F : \mathbb{R} \to [0, 1]$ (a special kind of CDF).

**Definition 2** (Generalized Thurstone Model)**.** *Given an observation graph $\mathcal{G}$, a pairwise comparison model is said to be a* generalized Thurstone model *(GTM) $\mathcal{T}_F$ with choice function $F : \mathbb{R} \to [0, 1]$ if there exists a weight (or utility) vector $w \in \mathcal{W}$ such that:*

$$\forall (i,j) \in \mathcal{E}, \;\; p_{ij} = F(w_i - w_j),$$

*where $\mathcal{W} \subseteq \mathbb{R}^n$ is a specified convex parameter space (usually $\mathbb{R}^n$ or a compact hypercube in $\mathbb{R}^n$).*

The GTM [37, 32] posits that every agent $i$ has a latent utility $w_i$, and uncertainty in the comparison process is modeled by independent and identically distributed (i.i.d.) noise random variables $X_1, \dots, X_n$ with absolutely continuous CDF $G : \mathbb{R} \to [0, 1]$. The discriminant variables $(w_1 + X_1, \dots, w_n + X_n)$ formed by combining utilities with the noise random variables are then compared to determine the outcomes of pairwise comparisons. Hence, the probability of preferring agent $i$ over $j$ is given by

$$\mathbb{P}(i \text{ preferred over } j) = \mathbb{P}(w_i + X_i > w_j + X_j) = \int_{-\infty}^{\infty} G(y + w_i - w_j) G'(y)\, \mathrm{d}y = F(w_i - w_j). \tag{3}$$

As noted earlier, GTMs encompass a wide range of models as special cases, e.g., Thurstone models [47], BTL models [4, 28], Dawkins models [11], etc. We can also define a pairwise probability matrix $\mathsf{F}(w) \in [0,1]^{n\times n}$ for a GTM $\mathcal{T}_F$ with weight vector $w$ via

$$(\mathsf{F}(w))_{ij} \triangleq \begin{cases} F(w_i - w_j), & (i,j) \in \mathcal{E}, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

We next describe the data generation process for GTMs and general pairwise comparison models alike. For any pair $(i,j) \in \mathcal{E}$, define the outcome of the $m$th "$i$ vs. $j$" pairwise comparison between them as the Bernoulli random variable

$$Z_{ij}^m \triangleq \begin{cases} 1, & \text{if } i \text{ preferred over } j \text{ (with probability } p_{ij}), \\ 0, & \text{if } j \text{ preferred over } i \text{ (with probability } 1 - p_{ij}), \end{cases} \tag{5}$$

for $m \in [k_{ij}]$, where $k_{ij}$ denotes the number of observed "$i$ vs. $j$" comparisons. The given pairwise comparison data is then a collection of these *independent* Bernoulli variables $\mathcal{Z} \triangleq \{Z_{ij}^m : (i,j) \in \mathcal{E},\ m \in [k_{i,j}]\}$. For convenience, we also let $Z_{ij} \triangleq \sum_{m=1}^{k_{ij}} Z_{ij}^m$ and $\hat{p}_{ij} \triangleq Z_{ij}/k_{ij}$.

**Parameter estimation for GTM.** To present our testing formulation in the sequel, we explain how the parameters of a $\mathcal{T}_F$ model are estimated given pairwise comparison data $\mathcal{Z}$ [45]. First, we define the weighted negative log-likelihood function $l : \mathcal{W} \times [0,1]^{|\mathcal{E}|} \to \mathbb{R}_+ \cup \{+\infty\}$ as

$$l(w; \{\hat{p}_{ij} : (i,j) \in \mathcal{E}\}) \triangleq - \sum_{(i,j)\in\mathcal{E}} \hat{p}_{ij} \log(F(w_i - w_j)) + (1 - \hat{p}_{ij}) \log(1 - F(w_i - w_j)). \tag{6}$$

Note that this function represents a weighted variant of the typical log-likelihood function used in parameter estimation [45, 50]. The weights of the $\mathcal{T}_F$ model are estimated by minimizing:

$$\hat{w} \triangleq \operatorname*{arg\,min}_{w\in\mathcal{W}_b} l(w; \{\hat{p}_{ij} : (i,j) \in \mathcal{E}\}), \tag{7}$$

where the constraint set $\mathcal{W}_b \triangleq \{w \in \mathcal{W} : \|w\|_\infty \le b,\ w^{\mathrm{T}}\mathbf{1} = 0\}$ for some (universal) constant $b$, $\mathbf{1} \in \mathbb{R}^n$ denotes an all-ones vector, and the constraint $w^{\mathrm{T}}\mathbf{1} = 0$ allows for identifiability of the weights.

**Assumption on comparison models.** To facilitate the analysis of the hypothesis testing problem in (1), we introduce a simplifying assumption on the class of general pairwise comparison models. We assume that the pairwise probabilities $p_{ij}$ are bounded away from 0 and 1.

**Assumption 1** (Dynamic Range)**.** *There exists a constant $\delta > 0$ such that for any pairwise comparison model under consideration, $p_{ij} \in [\delta, 1-\delta]$ for all $(i,j) \in \mathcal{E}$.*

Note that under the null hypothesis, the Assumption 1 is satisfied by all $\mathcal{T}_F$ models with weights bounded by $F^{-1}(1-\delta)/2$. Subsequently, we assume that the constant $b$ satisfies $b \ge F^{-1}(1-\delta)/2$. For any given pairwise comparison model $\{p_{ij} : (i,j) \in \mathcal{E}\}$, define $w^* \in \mathcal{W}_b$ be the weights of a $\mathcal{T}_F$ model that best approximates this pairwise comparison model in the maximum likelihood sense:

$$w^* \triangleq \operatorname*{arg\,min}_{w\in\mathcal{W}_b} l(w; \{p_{ij} : (i,j) \in \mathcal{E}\}). \tag{8}$$

Finally, we also assume in the sequel that the given choice function $F$ exhibits *strong log-concavity* and has a bounded derivative on $[-2b, 2b]$, i.e., there exists a constant $\alpha, \beta > 0$ such that:

$$\forall x \in [-2b, 2b], \quad -\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log(F(x)) \ge \alpha \quad \text{and} \quad F'(x) \le \beta. \tag{9}$$

Several popular GTMs including the BTL and Thurstone (Case V) model satisfy both the above assumptions. The following proposition highlights that $w^*$ always exists and is unique for a strictly log-concave function $F$ on $\mathcal{W}_b$.

**Proposition 1** (Existence and Uniqueness of Maximum Likelihood)**.** *Suppose the observation graph $\mathcal{G}$ is connected, the choice function $F : \mathbb{R} \to [0,1]$ is strictly log-concave on $[-2b, 2b]$, and Assumption 1 holds. Then, there exists a unique optimal solution $w^* \in \mathcal{W}_b$ satisfying (8).*

The proof is provided in **??**. It follows from Proposition 1 and Gibbs' inequality that when the pairwise comparison model is indeed a $\mathcal{T}_F$ model with weight vector $w$, then we have $w^* = w$. We discuss the potential limitations and robustness of these assumptions in **??**.

**Minimax formulation.** For any fixed graph $\mathcal{G}$, choice function $F$, and (universal) constants $\delta, \epsilon > 0$ and $b \geq F^{-1}(1-\delta)/2$, define the sets $\mathcal{M}_0$ and $\mathcal{M}_1(\epsilon)$ of $\mathcal{T}_F$ and pairwise comparison models:

$$\mathcal{M}_0 \triangleq \{P : \text{Assumption 1 holds and } \exists\, w \in \mathcal{W}_b \text{ such that } P = \mathsf{F}(w)\}, \tag{10}$$

$$\mathcal{M}_1(\epsilon) \triangleq \left\{P : \text{Assumption 1 holds and } \inf_{w \in \mathcal{W}_b} \frac{1}{n}\|P - \mathsf{F}(w)\|_{\mathrm{F}} \geq \epsilon\right\}, \tag{11}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes Frobenius norm. Now, we formalize the hypothesis testing problem in (1) as:

$$\begin{aligned} H_0 &: \mathcal{Z} \sim P \in \mathcal{M}_0, \\ H_1 &: \mathcal{Z} \sim P \in \mathcal{M}_1(\epsilon). \end{aligned} \tag{12}$$

We will discuss the separation distance $\inf_{w \in \mathcal{W}_b} \|P - \mathsf{F}(w)\|_{\mathrm{F}}$ later. For now, note that we only test on the set of observed comparisons $\mathcal{E}$ as it is not possible to determine whether the comparisons on $\mathcal{E}^c$ would conform to a $\mathcal{T}_F$ model or some other pairwise comparison model. Next, for any fixed graph $\mathcal{G}$, choice function $F$, and constants $\delta, b, \epsilon > 0$, we define the *minimax risk* as:

$$\mathcal{R}(\mathcal{G}, \epsilon) \triangleq \inf_{\phi}\Bigg\{\underbrace{\sup_{P \in \mathcal{M}_0} \mathbb{P}_{H_0}(\phi(\mathcal{Z}) = 1)}_{\mathcal{Z}\sim P \text{ under } H_0} + \underbrace{\sup_{P \in \mathcal{M}_1(\epsilon)} \mathbb{P}_{H_1}(\phi(\mathcal{Z}) = 0)}_{\mathcal{Z}\sim P \text{ under } H_1}\Bigg\}, \tag{13}$$

where the infimum is taken over all randomized decision rules $\phi(\mathcal{Z}) \in \{0, 1\}$ (with 0 corresponding to $H_0$ and 1 to $H_1$), and $\mathbb{P}_{H_0}$ and $\mathbb{P}_{H_1}$ denote the probability measures under hypotheses $H_0$ and $H_1$, respectively. Intuitively, this risk minimizes the sum of the worst-case type I and type II errors. Finally, we define the *critical threshold* of the hypothesis testing problem in (12) as the smallest value of $\epsilon$ for which the minimax risk is bounded by $\frac{1}{2}$ (cf. [40]):

$$\varepsilon_{\mathsf{c}} \triangleq \inf\left\{\epsilon > 0 : \mathcal{R}(\mathcal{G}, \epsilon) \leq \frac{1}{2}\right\}. \tag{14}$$

Note that the constant $\frac{1}{2}$ here is arbitrary and can be replaced by any constant in $(0, 1)$.

## 3 Main results

In this section, we present the main results of the paper. We first show that our notion of separation distance can be simplified for analysis, then proceed to bound the critical threshold and minimax risk, and finally, establish type I and II error bounds in the sequential setting.

Recall that to formalize (12), we defined the *separation distance* of a pairwise comparison model $P$ to the class of $\mathcal{T}_F$ models as $\inf_{w \in \mathcal{W}_b} \|P - \mathsf{F}(w)\|_{\mathrm{F}}$ (for fixed $F$). To make this separation distance more amenable to theoretical analysis, we approximate it in the next theorem with the simpler quantity $\|P - \mathsf{F}(w^*)\|_{\mathrm{F}}$, where $w^*$ is given in (8).

**Theorem 1** (Separation Distance to $\mathcal{T}_F$ Models)**.** *Let $P$ be a pairwise comparison matrix satisfying Assumption 1. Then, there exists a universal constant $c_1 > 0$ (that does not depend on $n$) such that the separation distance between $P$ and the class of $\mathcal{T}_F$ models satisfies*

$$c_1\|P - \mathsf{F}(w^*)\|_{\mathrm{F}} \leq \inf_{w \in \mathcal{W}_b} \|P - \mathsf{F}(w)\|_{\mathrm{F}} \leq \|P - \mathsf{F}(w^*)\|_{\mathrm{F}},$$

*where $w^*$ is given by (8).*

The proof is provided in **??**. The upper bound is immediate, and the lower bound utilizes the information-theoretic bounds between $f$-divergences.

**Test statistic.** We now introduce our test statistic based on the approximation derived in Theorem 1. First, we partition the observed comparison data $\mathcal{Z}$ into two (roughly) equal parts $\mathcal{Z}_1 = \{Z_{ij}^m : (i,j) \in \mathcal{E},\, m \in [\lfloor k_{ij}/2 \rfloor]\}$ and $\mathcal{Z}_2 = \mathcal{Z} \setminus \mathcal{Z}_1$. The first half of the dataset $\mathcal{Z}_1$ is used to estimate the parameters $\hat{w}$ as shown in (7). Then, we use $\mathcal{Z}_2$ to calculate the *test statistic* $T$ via

$$T \triangleq \sum_{(i,j) \in \mathcal{E}} \left(\frac{Z_{ij}(Z_{ij} - 1)}{k'_{ij}(k'_{ij} - 1)} + F(\hat{w}_i - \hat{w}_j)^2 - 2F(\hat{w}_i - \hat{w}_j)\frac{Z_{ij}}{k'_{ij}}\right)\mathbb{1}_{k'_{ij} > 1}, \tag{15}$$

where $k'_{ij} = k_{ij} - \lfloor k_{ij}/2 \rfloor$, $Z_{ij} = \sum_{m > \lfloor k_{ij}/2 \rfloor} Z_{ij}^m$ is computed as before but using only the samples in $\mathcal{Z}_2$, and $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function of $\mathcal{A}$. By construction, if $\hat{w} = w^*$, then

$\mathbb{E}[T] = \|P - \mathsf{F}(w^*)\|_{\mathrm{F}}^2$. Hence, $T$ is constructed by plugging in $\hat{w}$ in place of $w^*$ in an unbiased estimator of $\|P - \mathsf{F}(w^*)\|_{\mathrm{F}}^2$. Our proposed *hypothesis test thresholds $T$ to determine the unknown hypothesis* ($H_1$ is selected if $T$ exceeds a certain threshold). We will discuss analytical expressions for the threshold below and a data-driven manner of determining the threshold in Section 4.

**Upper bound on critical threshold.** In this section, we make the simplifying assumption that $k_{ij} = 2k$ (with $k \in \mathbb{N}$) for all $(i, j) \in \mathcal{E}$. The ensuing theorem proved in **??** establishes an upper bound on the critical threshold of the hypothesis testing problem defined in (12).

**Theorem 2** (Upper Bound on Critical Threshold)**.** *Consider the hypothesis testing problem in* (12)*, and assume that Assumption 1 holds and $k \geq 2$. Then, there exists a constant $c_2 > 0$ such that the critical threshold defined in* (14) *is upper bounded by*

$$\varepsilon_{\mathsf{c}}^2 \leq \frac{c_2 d_{\max}}{nk\lambda_2(L)},$$

*where $d_{\max} = \max_{i\in[n]} \sum_{j\in[n]\setminus i} E_{ij}$ and $\lambda_2(L)$ is the second smallest eigenvalue of Laplacian $L$.*

In analysis we select $H_1$ if $T > \gamma \frac{nd_{\max}}{k\lambda_2(L)}$ and $H_0$ otherwise, where $\gamma$ is an appropriate constant independent of $n, k$ (see **??**). The analysis relies on establishing non-trivial $\ell^2$-error bounds for parameter estimation of $\mathcal{T}_F$ models when the data is generated by a general pairwise comparison model, which is not necessarily a GTM (i.e., deriving error bounds under a potential model mismatch). The $\ell^2$-error bounds allow us to prove bounds on the mean and variance of the test statistic $T$ under both hypotheses $H_0$ and $H_1$. Then, using Chebyshev's inequality, we can bound the probabilities of error of our test under each of the hypotheses, which induces an upper bound on the critical threshold.

It is worth emphasizing the dependence of the upper bound on the topology of the observation graph $\mathcal{G}$. Notably, the connectedness of the graph ensures $\lambda_2(L) > 0$, and the value of $\lambda_2(L)$ is known for various classes of graphs, such as: **1)** Complete graphs on $n$ nodes have $d_{\max} = n - 1$ and $\lambda_2(L) = n$ yielding $\varepsilon_{\mathsf{c}} = O(1/\sqrt{nk})$; **2)** $d$-regular spectral expander graphs with constant $d$ have $d_{\max} = d$ and $\lambda_2(L) \geq d - 2\sqrt{d}$, [3] yielding $\varepsilon_{\mathsf{c}} = O(1/\sqrt{nk})$; **3)** Single cycle graphs on $n$ nodes have $d_{\max} = 2$ and $\lambda_2(L) = \Theta(1/n^2)$ yielding $\varepsilon_{\mathsf{c}} = O(\sqrt{n/k})$; **4)** Two-dimensional $\sqrt{n} \times \sqrt{n}$ toroidal grid (graph on $n$ vertices formed by cartesian product of two cycles of length $\sqrt{n}$) have $d_{\max} = 4$ and $\lambda_2(L) = 2 - 2\cos(2\pi/\sqrt{n}) = \Theta(1/n)$ yielding $\varepsilon_{\mathsf{c}} = O(1/\sqrt{k})$.

We also note that in the special case where $\mathcal{T}_F$ is a BTL model, our upper bound on $\varepsilon_{\mathsf{c}}$ recovers the bound in [29] for complete graphs. But our likelihood-based proof is quite different to the spectral ideas in [29]. Finally, we present the key $\ell^2$-error bounds for parameter estimation when data is generated by a general pairwise comparison model needed to prove Theorem 2 (as mentioned above).

**Theorem 3** ($\ell^2$-Error Bounds for Parameter Estimation)**.** *Consider any pairwise comparison model satisfying Assumption 1 with $w^*$ given by* (8) *and $\hat{w}$ constructed according to* (7) *from data generated by the model. Then, for some constant $c_3 > 0$, the following tail bound holds on the estimation error of $w^*$:*

$$\forall t \geq 1, \;\; \mathbb{P}\left(\|\hat{w} - w^*\|_2^2 \geq \frac{c_3 n\beta^2}{\alpha^2 k\lambda_2(L)F(-2b)^2}\, t\right) \leq e^{-t},$$

*where $\alpha$ is defined in* (9)*. Moreover, for any $p \geq 1$, there exists a $p$-dependent constant $c(p) > 0$ such that the expected pth moment of the error is bounded by*

$$\mathbb{E}[\|\hat{w} - w^*\|_2^p] \leq \left(\frac{c(p)n\beta^2}{\alpha^2 k\lambda_2(L)F(-2b)^2}\right)^{\frac{p}{2}}.$$

The proof is provided in **??**. In the special case where the pairwise comparison model is a GTM, our bounds recover the bounds derived in [45, Theorem 3] up to constants. However, our result is much more general because it holds for any pairwise comparison model, which requires careful formulation and development of the proof technique.

**Information-theoretic lower bounds.** We now establish information-theoretic lower bounds on the minimax risk and critical threshold for the hypothesis testing problem in (12). For simplicity and analytical tractability, assume that $k_{ij} = k \in \mathbb{N}$ for all $(i, j) \in \mathcal{E}$, and assume that the observation graph $\mathcal{G}$ is *super-Eulerian* [5], i.e., it has an Eulerian spanning sub-graph $\tilde{\mathcal{G}} = ([n], \tilde{\mathcal{E}})$ so that every vertex of $\tilde{\mathcal{G}}$ has even degree. Then, $\tilde{\mathcal{G}}$ has a *cycle decomposition* $\mathcal{C}$ by Veblen's theorem [2, 44], where $\mathcal{C}$ is a collection of simple cycles $\sigma$ that partitions the undirected edges of $\tilde{\mathcal{G}}$. The ensuing theorem proved in **??** presents our minimax risk lower bound.

**Theorem 4** (Minimax Lower Bound)**.** *Consider the hypothesis testing problem in* (12) *and assume that the observation graph $\mathcal{G}$ is super-Eulerian with spanning Eulerian sub-graph $\tilde{\mathcal{G}}$. Then, there exists a constant $c_4 > 0$ such that for any $\epsilon > 0$, the minimax risk in* (13) *is lower bounded by*

$$\mathcal{R}(\mathcal{G}, \epsilon) \geq 1 - \frac{1}{2}\sqrt{\exp\left(\frac{c_4 k^2 n^4 \epsilon^4}{|\tilde{\mathcal{E}}|^2} \sum_{\sigma \in \mathcal{C}} |\sigma|^2\right) - 1}\,,$$

*where $|\sigma|$ denotes the length of a cycle $\sigma \in \mathcal{C}$, and $\mathcal{C}$ is the cycle decomposition of $\tilde{\mathcal{G}}$.*

Our approach utilizes the Ingster-Suslina method [23], which is similar to *Le Cam's method*, but provides a lower bound by considering a cleverly chosen point and a mixture on the parameter space instead of just two points. Our specific construction is inspired by the technique introduced in [44], which establishes a lower bound for testing of IIA assumption for the BTL model and for Eulerian graph structures. We extend their approach in three significant ways. First, we generalize their method to accommodate any GTM rather than just the BTL model. Second, we use a different technique based on Theorem 1 to lower bound separation distance from the class of $\mathcal{T}_F$ models. Moreover, our work quantifies separation using Frobenius norm instead of sums of total variation distances. Third, our argument holds for a broader class of graphs, namely, super-Eulerian graphs. Note that the question of algorithmically constructing Eulerian sub-graphs of graphs has been widely studied [18].

The following proposition simplifies Theorem 4 to obtain lower bounds on the critical threshold for several classes of graphs.

**Proposition 2** (Lower Bounds on Critical Threshold)**.** *Under the assumptions of Theorem 4, the following lower bounds hold for the critical threshold defined in* (14)*:*
*1) If $\mathcal{G}$ is a complete graph with odd $n$ vertices, then $\varepsilon_{\mathsf{c}}^2 = \Omega(1/nk)$.*
*2) If $\mathcal{G}$ is a d-regular graph with constant $d \geq 2$, then $\varepsilon_{\mathsf{c}}^2 = \Omega(1/n^2k)$.*
*3) If $\mathcal{G}$ is a single cycle graph with $n$ vertices, then $\varepsilon_{\mathsf{c}}^2 = \Omega(1/n^2k)$.*
*4) If $\mathcal{G}$ is a two-dimensional $\sqrt{n} \times \sqrt{n}$ toroidal grid on $n$ vertices formed by the cartesian product of two cycles of length $\sqrt{n}$, then $\varepsilon_{\mathsf{c}}^2 = \Omega(1/n^{7/4}k)$.*

The proof of Proposition 2 involves calculating the number of simple cycles and the individual cycle lengths in the cycle decompositions $\mathcal{C}$ and is provided in **??**. The lower bounds on $\varepsilon_{\mathsf{c}}$ are then obtained from Theorem 4. We remark that our minimax upper and lower bounds on $\varepsilon_{\mathsf{c}}$ match for the complete graph case, demonstrating the minimax optimality of the threshold's scaling (up to constant factors). Moreover, they also match with respect to $k$ for other classes of graphs as well. It is worth mentioning that in the special case of BTL models with single cycle graphs, our lower bound on $\varepsilon_{\mathsf{c}}$ improves the high-level scaling behavior in [44] from $\Omega(1/\sqrt{n^3k})$ to $\Omega(1/\sqrt{n^2k})$ (when $\varepsilon_{\mathsf{c}}$ is quantified in terms of Frobenius norm). Lastly, we remark that for single cycle graphs, the gap between the upper and lower bounds in terms of $n$ intuitively holds because our lower bounds become larger when there are more cycles in $\mathcal{C}$, which is only 1 in this case. Furthermore, our test relies on the accuracy of estimating the optimal $w^*$ in (8), which worsens for graphs with smaller $\lambda_2(L)$.

**Upper bounds on type I and II error probabilities.** To complement the minimax risk lower bound in Theorem 4, we establish upper bounds on the extremal type I and II error probabilities. We will do this in the *sequential* setting, where data is observed incrementally—a common practical scenario which subsumes the standard fixed sample-size setting, cf. [30, 21]. In the sequential testing framework, at each time step, we observe a single "$i$ vs. $j$" comparison for every $(i, j) \in \mathcal{E}$. (The subsequent analysis can be extended to a general setting where we observe only one comparison for some pair $(i, j) \in \mathcal{E}$ or even a variable number of comparisons at every time step.) At time $k_1 + k$ with $k_1, k \in \mathbb{N}$, we define $T^{k_1,k}$ to be the value of the test statistic $T$ in (15), where comparisons from $k_1$ time-steps have been used to build the dataset $\mathcal{Z}_1$ to estimate parameters using $\hat{w}$, and $k$ time-steps have been used to build the dataset $\mathcal{Z}_2$ to calculate the statistic $T$. Note that $\mathcal{Z}_1$ and $\mathcal{Z}_2$ no longer need to be similar in size. Then, we can decide based on thresholding $T^{k_1,k}$ (see Theorem 5) whether to collect more data or stop and reject $H_0$ while controlling the probabilities of error. If the testing process ends without rejecting $H_0$, then we can accept $H_0$. A key observation underlying our analysis is the following *reverse martingale* property (see, e.g., [21, 30]).

**Proposition 3** (Reverse Martingale)**.** *Fix any* $k_1 \in \mathbb{N}$, *and let* $\mathcal{F}_k = \bigotimes_{(i,j)\in\mathcal{E}} \sigma(\sum_{m=k_1+1}^{k_1+k} Z_{ij}^m,$ $Z_{ij}^{k_1+k+1}, Z_{ij}^{k_1+k+2}, \dots)$ *be a non-increasing sequence of* $\sigma$*-algebras, where* $\bigotimes$ *denotes the product* $\sigma$*-algebra. Then, the sequence of test-statistics* $\{T^{k_1,k} : k \geq 2\}$ *is a reverse martingale with respect to reverse filtration* $\{\mathcal{F}_k : k \geq 2\}$*, i.e., for* $k \geq 2$, $T^{k_1,k}$ *is* $\mathcal{F}_k$*-measurable and* $\mathbb{E}\big[T^{k_1,k}|\mathcal{F}_{k+1}\big] = T^{k_1,k+1}$.

The proof is presented in **??**. This observation allows us to develop *time-uniform bounds* in terms of $k$ on type I and type II error probabilities, i.e., they hold for all $k$ larger than a constant. The next theorem, proved in **??**, presents our type I and type II error probability bounds.

**Theorem 5** (Type I and Type II Error Probability Bounds)**.** *Under the sequential setting discussed above, the following bounds hold on the extremal type I and type II error probabilities. There exist constants* $c_5, c_6, c_7, c_8, c_9$ *such that for all* $t \geq 1$, $\nu \in (0, 1/e)$, $k_1 \in \mathbb{N}$ *and* $\epsilon \geq c_5 t^{\frac{1}{2}} \tilde{\varepsilon}/n$*, we have*

$$\sup_{P\in\mathcal{M}_0} \mathbb{P}_{H_0}\left( \exists k \geq 2, T^{k_1,k} \geq c_6 t \tilde{\varepsilon}^2 + \frac{c_7 |\mathcal{E}|^{\frac{1}{2}} \ell_{k,\nu}}{k} + c_8 \tilde{\varepsilon} \sqrt{\frac{t \ell_{k,\nu}}{k}} \right) \leq \nu + e^{-t},$$

$$\sup_{P\in\mathcal{M}_1(\epsilon)} \mathbb{P}_{H_1}\left( \exists k \geq 2,\ T^{k_1,k} - (D - c_9 t^{\frac{1}{2}} \tilde{\varepsilon})^2 \leq -\frac{c_7 |\mathcal{E}|^{\frac{1}{2}} \ell_{k,\nu}}{k} - (4D + c_8 t^{\frac{1}{2}} \tilde{\varepsilon}) \sqrt{\frac{\ell_{k,\nu}}{k}} \right) \leq \nu + e^{-t},$$

*where* $D \triangleq \|P - \mathsf{F}(w^*)\|_{\mathrm{F}}$ *and* $\tilde{\varepsilon} \triangleq \sqrt{n d_{\max}/(k_1 \lambda_2(L))}$ *and* $\ell_{k,\nu} \triangleq \log(3.5 \log^2(k)/\nu)$.

We now make several remarks. Firstly, our error probability bounds encode the scalings of the thresholds to accept or reject $H_0$ (see **??**). Secondly, our bounds hold regardless of how the decision-maker assigns data collected at different time-steps to $\mathcal{Z}_1$ and $\mathcal{Z}_2$. Moreover, they provide insights on how to split the data based on the topology of the observation graph, e.g., it is advantageous to equally split the data for complete graphs. To illustrate this and help parse Theorem 5, we present corollaries of Theorem 5 for the complete and single cycle graph cases in **??**. Thirdly, our bounds clearly hold in the non-sequential fixed sample-size setting, as we can just fix a particular value of $k$. Hence, adding the two extremal probabilities of error yields upper bounds on the minimax risk. Notably, the proof of Theorem 5 requires us to develop a time-uniform version of the well-known Hanson-Wright inequality [42] specialized for our setting (see **??** in **??**). Additionally, as an intermediate step in the proof, we also obtain time-uniform *confidence intervals* under the null hypothesis $H_0$, as demonstrated in the following proposition.

**Proposition 4** (Confidence Interval for $T^{k_1,k}$)**.** *Suppose* $\hat{w}$ *is estimated as in* (7) *from the comparisons over* $k_1$ *time-steps. Then, there exists a constant* $c_6 > 0$ *such that for all* $\nu \in (0, 1/e)$ *and* $k_1 \in \mathbb{N}$,

$$\mathbb{P}_{H_0}\left( \exists k \geq 2, T^{k_1,k} \geq \|\mathsf{F}(\hat{w}) - \mathsf{F}(w^*)\|_{\mathrm{F}}^2 + c_7 \frac{\sqrt{|\mathcal{E}| \ell_{k,\nu}}}{k} + 4 \|\mathsf{F}(\hat{w}) - \mathsf{F}(w^*)\|_{\mathrm{F}} \sqrt{\frac{\ell_{k,\nu}}{k}} \right) \leq \nu.$$

Proposition 4 is established in **??**. We remark that the distribution of $\|\mathsf{F}(\hat{w}) - \mathsf{F}(w^*)\|_{\mathrm{F}}$ above can be approximated either by leveraging the asymptotic normality of $\hat{w} - w^*$ [16, 20], or by utilizing bootstrapping techniques; this gives $(1 - 2\nu)$ time-uniform confidence intervals. Additionally, the constant $c_7$ here is also the constant in our specialized version of Hanson-Wright inequality (noted above) and, for our setting, can be approximated via simulations. An empirical investigation into estimating the constant $c_6$ and the subsequent confidence intervals can be found in **??**.

**Limitations.** We outline some limitations of our results here. Firstly, our upper and lower bounds on the critical threshold exhibit gaps with respect to $n$ for graphs that are not complete. Similarly, there is a gap between the minimax risk lower bound and the upper bound induced by our type I and type II error probability bounds; such gaps are known to be quite difficult to close in theory. Furthermore, our testing approach relies on splitting the dataset into two portions, which may not be the most efficient way to use samples. Addressing any of these directions would be interesting for future work.

## 4 Experiments

In this section, we develop a data-driven approach to select the threshold for our test $T$, and conduct simulations to validate our theoretical results and its evaluations on synthetic and real-world datasets.

**Estimating the threshold.** Given a pairwise comparison dataset $\mathcal{Z} \triangleq \{Z_{ij}^m : (i,j) \in \mathcal{E},\ m \in [k_{i,j}]\}$, we employ an empirical quantile approach to determine the critical threshold for our hypothesis testing problem. We generate multiple $\mathcal{T}_F$ models with random skill scores $w \in \mathbb{R}^n$ such that $\|w\|_\infty \leq b$ for some constant $b$ and simulate $k_{ij}$ "$i$ vs. $j$" comparisons by sampling binomial random variables $\{\tilde{Z}_{ij} \sim \text{Bin}(k_{ij}, F(w_i - w_j)\}_{(i,j)\in\mathcal{E}}$. We then compute the test statistic $T$ for each simulated dataset, repeating the process a sufficient number of times to build a distribution of test statistics. Finally, we extract the 95th percentile value (or $0.95$ quantile) from this distribution as our empirical threshold.

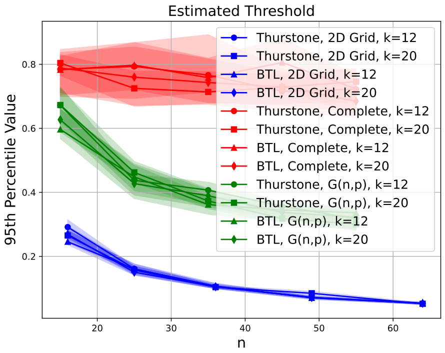

Figure 1: Estimated scaled threshold for various values of $n, k$, graph topologies, and $\mathcal{T}_F$ models.

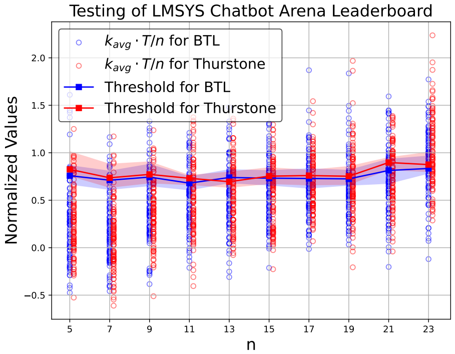

Figure 2: Scaled test statistics and the estimated thresholds evaluated on the LYMSYS dataset.

In our first experiment, we investigate the behaviour of threshold for test $T$ based on this empirical quantile approach for various values of $n$ and $k$, and for different graph topologies and $\mathcal{T}_F$ models. We considered values of $n$ ranging from $15$ to $55$ with intervals of $10$, $k \in \{12, 20\}$, and graph topologies including complete graphs, $\lceil\sqrt{n}\,\rceil \times \lceil\sqrt{n}\,\rceil$ toroidal grids and sparse graphs generated from Erdős-Rényi $\mathcal{G}(n,p)$ model with parameter $p = 2\log^2(n)/n$ and the $\mathcal{T}_F$ models included standard Thurstone (Case V) and BTL models. For each choice of parameters, we generated $400$ models by randomly sampling weights and with $b = F^{-1}(0.98)/2$ and generated synthetic comparison data. The scaled test-statistic $k\lambda_2(L)(\mathcal{G}) \cdot T/(nd_{\max})$ was computed for every parameter choice and the 95th percentile value of this scaled $T$ was identified as the threshold $\gamma$. Figure 1 plots this 95th percentile values with respect to $n$ for various parameter choices. Notably, the value $\gamma$ remains roughly constant with $n, k$ and model $F$ for complete graphs. However, for toroidal grid graphs, a significant decrease in the scaled test statistic is observed, suggesting that our error bounds may be overly conservative in estimating the deviation in $T$.

In our next experiment, we apply our test to the LMSYS chatbot leaderboard [9], a widely used benchmark for evaluating the performance of LLMs. The dataset contains a collection of pairwise comparisons between various LLMs based on their response to prompts, which are then used to obtain ELO ratings. We retain the directional nature of comparisons, where an "$i$ vs. $j$" comparison indicates model $i$ as the first response and $j$ as second during the evaluation. We rank the LLMs based on their frequency of appearance in the dataset and perform the test repeatedly on top $n$ LLMs in this ordering, with $n$ ranging from 5 to 21 with gaps of 2, for both Thurstone and BTL models. For each $n$, we plot the values in Figure 2 of (scaled) test-statistic $k_{\text{avg}} \cdot T/n$ and the obtained (scaled) thresholds using the quantile approach (with same parameters as above), where $k_{\text{avg}}$ is the average of $k_{ij}$ over all $(i,j) \in \mathcal{E}$. By randomizing over the partitioning of dataset $\mathcal{Z}$ into $\mathcal{Z}_1$ and $\mathcal{Z}_2$ and computing $T$ each time, we essentially obtain a distribution of $T$ and plot these values in Figure 2 as a scatter plot. The figure highlights that both BTL and Thurstone models perform well in modeling for smaller values of $n$ with only $10\%$ of samples above the threshold but exhibit significant deviations for larger values of $n$ (as around $60\%$ samples are above the threshold for $n = 21$). Notably, for both experiments, the error bars are $96\%$ confidence intervals (see **??** for additional details), and all results were obtained using modest computational resources within a few minutes to an hour.

**Broader impacts.** Our paper is primarily theoretical and does not have immediate societal implications. However, if practitioners apply our approach in real-world settings, it could have both positive and negative impacts. On the one hand, our work can help detecting a few kinds of biases in data such as whether the data conforms to an underlying $\mathcal{T}_F$ model. On the other hand, using any ranking mechanism based on the outcome of our test in real-world settings may unintentionally exacerbate existing inequalities, such as unequal access to resources and opportunities.

## References

[1] Home ground advantage of individual clubs in english soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44:509–521, 1995. ISSN 00390526, 14679884. doi: 10.2307/2348899.

[2] N. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory, 1736-1936*. Oxford University Press, 1986.

[3] Y. Bilu and N. Linial. Lifts, discrepancy and nearly optimal spectral gap. *Combinatorica*, 26(5): 495–519, 2006.

[4] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, December 1952.

[5] P. A. Catlin. Supereulerian graphs: a survey. *Journal of Graph theory*, 16(2):177–196, 1992.

[6] M. Cattelan, C. Varin, and D. Firth. Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150, 2013.

[7] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, February 2015.

[8] Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-$K$ ranking. *The Annals of Statistics*, 47(4):2204–2235, 2019.

[9] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

[10] D. Davidson and J. Marschak. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, 17(2), 1959.

[11] R. Dawkins. A threshold model of choice behaviour. *Animal Behaviour*, 17:120–133, 1969.

[12] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[13] A. E. Elo. *The rating of chessplayers, past and present*. Arco Pub, New York, 1978.

[14] J. Fan, J. Hou, and M. Yu. Uncertainty quantification of mle for entity ranking with covariates. *arXiv preprint arXiv:2212.09961*, 2022.

[15] T. Feder and C. Subi. Packing edge-disjoint triangles in given graphs. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 19, page 13, 2012.

[16] C. Gao, Y. Shen, and A. Y. Zhang. Uncertainty quantification in the Bradley-Terry-Luce model. *Inf. Inference*, 12(2):1073–1140, 2023. ISSN 2049-8764. doi: 10.1093/imaiai/iaac032. URL https://doi.org/10.1093/imaiai/iaac032.

[17] T. Graepel, R. Herbrich, and T. Minka. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576, 2006.

[18] N. Haghparast and D. Kiani. Spanning eulerian subgraphs of large size. *Graphs and Combinatorics*, 35(1):201–206, Jan 2019. doi: 10.1007/s00373-018-1992-7. URL https://doi.org/10.1007/s00373-018-1992-7.

[19] G. J. Hahn and W. Q. Meeker. *Statistical intervals: a guide for practitioners*, volume 92. John Wiley & Sons, 2011.

[20] R. Han, W. Tang, and Y. Xu. Statistical inference for pairwise comparison models, 2024.

[21] S. R. Howard, A. Ramdas, J. D. McAuliffe, and J. S. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 2018. URL https://api.semanticscholar.org/CorpusID:218613937.

[22] D. R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, February 2004.

[23] Y. Ingster and I. Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169. Springer Science & Business Media, 01 2003. ISBN 978-0-387-95531-5. doi: 10.1007/978-0-387-21580-8.

[24] A. Jadbabaie, A. Makur, and D. Shah. Estimation of skill distributions. arXiv:2006.08189 [stat.ML], June 2020. URL https://arxiv.org/abs/2006.08189.

[25] D. Jannach, P. Resnick, A. Tuzhilin, and M. Zanker. Recommender systems — beyond matrix completion. *Commun. ACM*, 59(11):94–102, oct 2016. ISSN 0001-0782. doi: 10.1145/2891406. URL https://doi.org/10.1145/2891406.

[26] T. P. Kirkman. On a problem in combinations. *Cambridge and Dublin Mathematical Journal*, 2:191–204, 1847.

[27] Y. Liu, E. X. Fang, and J. Lu. Lagrangian inference for ranking problems. *Oper. Res.*, 71(1): 202–223, 2023. ISSN 0030-364X.

[28] R. D. Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons Inc., New York, NY, USA, 1959.

[29] A. Makur and J. Singh. Testing for the bradley-terry-luce model. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1390–1395, 2023. doi: 10.1109/ISIT54713.2023.10206450.

[30] T. Manole and A. Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023.

[31] J. Marschak. Binary choice constraints on random utility indicator. 1960.

[32] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, New York, 2 edition, 1989.

[33] D. McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, Economic Theory and Mathematical Economics, pages 105–142, New York, NY, USA, 1973. Academic Press.

[34] D. H. McLaughlin and R. D. Luce. Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychonomic Science*, 2(1-12):89–90, 1965.

[35] B. Morley and D. Thomas. An investigation of home advantage and other factors affecting outcomes in english one-day cricket matches. *Journal of sports sciences*, 23:261–268, 3 2005. ISSN 0264-0414. doi: 10.1080/02640410410001730133.

[36] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *Proceedings of the Advances in Neural Information Processing Systems 25 (NeurIPS)*, pages 2474–2482, Lake Tahoe, NV, USA, December 3-8 2012.

[37] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238. URL http://www.jstor.org/stable/2344614.

[38] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[39] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 24(2):193–202, 1975.

[40] C. Rastogi, S. Balakrishnan, N. B. Shah, and A. Singh. Two-sample testing on ranked preference data and the role of modeling assumptions. *Journal of Machine Learning Research*, 23(225): 1–48, 2022. URL http://jmlr.org/papers/v23/20-1304.html.

[41] D. Ross. Arpad elo and the elo rating system, 2007. URL http://en.chessbase.com/post/arpad-elo-and-the-elo-rating-system.

[42] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865. URL https://doi.org/10.1214/ECP.v18-2865.

[43] Sahand Negahban and Sewoong Oh and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research, INFORMS*, 65(1):266–287, January-February 2017.

[44] A. Seshadri and J. Ugander. Fundamental limits of testing the independence of irrelevant alternatives in discrete choice. *ACM EC 2019 - Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 65–66, 1 2020. doi: 10.48550/arxiv.2001.07042.

[45] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

[46] G. Simons and Y.-C. Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, June 1999.

[47] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

[48] L. L. Thurstone. The indifference function. *The journal of social psychology*, 2(2):139–167, 1931.

[49] A. Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

[50] M. Vojnovic and S. Yun. Parameter estimation for generalized thurstone choice models. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 498–506, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/vojnovic16.html.

[51] J. I. Yellott, Jr. The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, April 1977.

[52] E. Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, December 1929.

[53] S. Zhao, E. Zhou, A. Sabharwal, and S. Ermon. Adaptive concentration inequalities for sequential decision problems. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2e65f2f2fdaf6c699b223c61b1b5ab89-Paper.pdf.