КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 519.6
IRSTI 50.05, 50.41
https://doi.org/10.53364/24138614_2026_40_1_24

S.A. Aibaruly
[1]Kazakh-British Technical University,
Almaty, Kazakhstan

E-mail: akbarsaimzhan@gmail.com

PERFORMANCE OF RANKING METHODS ON CYCLIC PREFERENCE DATA: A COMPARATIVE STUDY

Abstract. *Ranking methods are fundamental tools in data analysis, applied across recommendation systems, social choice, and decision-making. While methods such as HodgeRank leverage topological properties of preference data, their empirical performance on real-world cyclic datasets remains understudied. This paper presents a comprehensive comparison of four ranking methods—HodgeRank, Borda Count, Bradley-Terry, and Spectral Ranking—across three datasets with varying cyclicity levels (55% to 96%). Performance is evaluated using pairwise accuracy (PA); stability is validated via 5-fold cross-validation; statistical significance of differences is assessed via McNemar's test with Bonferroni correction; and top-k agreement is quantified using Jaccard similarity J and Kendall's τ.*

Key findings reveal a non-monotonic relationship between cyclicity and HodgeRank performance: PA peaks at moderate cyclicity (0.851 at $\beta_1 = 82\%$) but degrades at both low (0.574 at 55%) and high (0.791 at 96%) extremes. Bradley-Terry achieves the highest average accuracy (0.767) with strong cross-validation stability. On highly cyclic data (SUSHI3), three methods—Bradley-Terry, Borda, and Spectral — reach perfect top-10 consensus ($J = 1.000$), while HodgeRank diverges from all three ($J = 0.250$, $\tau = -0.283$), a difference statistically confirmed (McNemar $\chi^2 = 119.4$, $p < 0.001$) despite aggregate PA differing by only 5.8 percentage points. These results demonstrate that aggregate accuracy metrics alone are insufficient for deployment decisions in ranking systems.

Keywords: *ranking methods, HodgeRank, cyclicity, preference aggregation, pairwise comparison, Bradley-Terry model, Borda Count, combinatorial Hodge theory.*

Introduction.

Ranking is a fundamental problem in data analysis that arises across numerous domains including recommendation systems, sports analytics, search engines, and social choice theory. Given pairwise comparisons between items, the goal is to produce a total ordering that best reflects aggregate preferences. However, real-world preference data often exhibit intransitivity—cyclic patterns where item A is preferred to B, B to C, yet C to A [1]—making the ranking problem non-trivial. This phenomenon, documented extensively in behavioral economics and cognitive psychology, reflects fundamental aspects of human decision-making rather than mere measurement error.

Recent work by Singh and Davidov [2] provides a principled framework for detecting and modeling cyclicality in paired comparison data. Their approach demonstrates that identifying cyclic patterns reduces to a model selection problem, with guarantees on large sample properties for distinguishing between gradient-based and cyclic preference structures.

Various ranking methods have been proposed to handle such data. Classical approaches include Borda Count [3], which performs simple aggregation of pairwise values, and the Bradley-Terry model [4], which employs a probabilistic framework with item strength parameters estimated via maximum likelihood. These methods date to the 18th and 20th centuries respectively but remain widely used due to their simplicity and robustness.

More recently, HodgeRank [5] applies Hodge theory from algebraic topology [6], decomposing pairwise preferences into a gradient component (consistent preferences), a curl component (locally cyclic patterns), and a harmonic component (globally cyclic patterns). Together, the curl and harmonic components constitute the non-gradient variation captured by β_1. This topological perspective provides an elegant mathematical framework for reasoning about intransitivity through the cyclicity ratio β_1, which quantifies the proportion of preference variation not explained by the gradient component—encompassing both curl and harmonic contributions. The method has inspired subsequent work on higher-order structures [7] and applications to temporal networks.

Despite HodgeRank's theoretical elegance, its empirical performance on real-world data with varying cyclicity levels remains understudied. Prior evaluations focus primarily on synthetic data [5] or specific application domains, leaving open questions about performance across naturally occurring cyclicity regimes. While intuition suggests methods should degrade uniformly as cyclicity increases, the actual relationship may be more complex. Moreover, aggregate accuracy metrics may mask important differences in which specific items methods rank highly—a critical consideration for applications like recommendation systems [8] where users primarily interact with top-k results.

This paper addresses these gaps through experiments on three real-world datasets spanning cyclicity from 55% to 96%. Our contributions are fourfold: (1) Characterizing HodgeRank's non-monotonic relationship with cyclicity, with peak performance at moderate levels; (2) Demonstrating that HodgeRank selects fundamentally different top items on highly cyclic data despite reasonable aggregate accuracy; (3) Showing Bradley-Terry's consistent robustness across all cyclicity levels; (4) Providing evidence-based selection guidelines for practitioners.

Materials and research methods.

We evaluate four ranking methods on three real-world datasets with varying cyclicity levels. This section formalizes the ranking problem, describes each method in detail, defines our evaluation metric, and explains the experimental protocol.

Let $V = \{1, \dots, n\}$ denote n items to be ranked. We observe pairwise preferences as a skew-symmetric matrix $Y \in R^{n\,x\,n}$ where $Y_{ij} > 0$ indicates that item i is preferred to item j across observed comparisons, $Y[j,i] = -Y[i,j]$ by skew-symmetry, and $Y[i,i] = 0$ by convention. The magnitude $|Y[i,j]|$ reflects the strength of preference, either through frequency of comparison outcomes or average rating differences depending on data source. The weight matrix $W \in \mathbb{R}^{\mathrm{nxn}}$ encodes comparison frequencies, with $W[i,j]$ denoting the number of times items i and j were directly compared. W is non-negative ($W[i,j] \geq 0$ for all i,j), with $W[i,j] > 0$ if and only if $(i,j) \in E$. We assume W is symmetric ($W[i,j] = W[j,i]$) and define the set of compared pairs as $E = \{(i,j) : W[i,j] > 0, i < j\}$.

Our goal is to produce a ranking $\pi: V \to \{1, \dots, n\}$ that maximizes agreement with observed preferences Y. We evaluate ranking quality using pairwise accuracy (PA), defined as the fraction of compared pairs for which the ranking correctly predicts preference direction. Beyond aggregate accuracy, we also analyze top-k agreement to assess whether methods agree on which specific items deserve high ranks, a critical consideration for practical applications.

HodgeRank [5] models preferences as edge flows on a weighted undirected graph $G = (V, E, W)$, where nodes represent items and edges represent comparisons. We assume G is connected; if not, each connected component is processed independently. Under connectivity, the graph Laplacian $L = D - W$ is symmetric positive semidefinite with exactly one zero

eigenvalue, whose null space is span {1}—the space of constant vectors. Each edge $(i,j) \in E$ carries a flow $Y[i,j]$, which may be inconsistent across cycles in the graph. The method seeks a score vector $s \in \mathbb{R}^n$ that minimizes the weighted L^2 discrepancy between observed preferences and the gradient field induced by scores [7]:

$$min_s \|Y - grad(s)\|_W^2 \quad (1)$$

where $grad(s)[i,j] = s[i] - s[j]$ represents the potential difference between items i and j under score vector s; Y is the observed preference matrix (skew-symmetric); W is the weight matrix encoding comparison frequencies; $\|X\|_W^2 W = \Sigma_{ij} W[i,j] \cdot X[i,j]^2$ denotes the weighted Frobenius norm; and s is the score vector to be estimated.

The weighted Frobenius norm prioritizes well-observed edges over sparse comparisons, naturally handling incomplete and unbalanced data. Intuitively, $grad(s)[i,j]$ represents the "potential difference" that would be induced by score vector s, analogous to voltage differences in electrical circuits. The optimization seeks scores that make these induced differences match observed preferences as closely as possible in the weighted L^2 sense.

The solution to Equation (1) is obtained via the graph Laplacian $L = D - W$, where $D = diag(W \cdot 1)$ is the weighted degree matrix with $D[i,i] = \Sigma_j W[i,j]$. Under the connectivity and non-negativity assumptions, L is symmetric positive semidefinite ($L \geq 0$), with all eigenvalues $\lambda_i \geq 0$ and exactly one zero eigenvalue $\lambda_1 = 0$. The first-order optimality conditions yield the linear system $Ls = b$, where $b = Y \cdot 1 \in \mathbb{R}^n$ is the divergence representing the net preference sum for each item. Since L has a one-dimensional null space spanned by constant vectors (adding a constant to all scores does not change gradients), we use the Moore-Penrose pseudoinverse $L^\dagger$ to obtain the minimum-norm solution: $s = L^\dagger \cdot b$. This solution is unique up to additive constants; we center scores by setting $\Sigma_i s[i] = 0$.

The cyclicity ratio β_1 quantifies the proportion of preference variation that cannot be explained by any gradient field. The notation β_1 is inspired by the first Betti number of a graph—the integer $|E| - |V| + 1$ counting independent cycles—which reflects the graph's capacity to support cyclic preference patterns; a larger cycle space generally leads to a higher ratio β_1:

$$\beta_1 = \frac{\|Y - grad(s)\|_W^2}{\|Y\|_W^2} \quad (2)$$

where $s* = L^\dagger \cdot b$ is the optimal score vector from Equation (1); $\|Y - grad(s*)\|_W^2$ measures the non-gradient component (curl and harmonic) representing intransitive preferences; $\|Y\|_W^2$ measures total preference variation; and $\beta_1 \in [0,1]$ is the cyclicity ratio.

By the Hodge decomposition theorem [6], any edge flow on a graph can be uniquely decomposed into a gradient (divergence-free in the dual sense, image of d_0), a curl (image of d_1*), and a harmonic component (kernel of the graph Hodge Laplacian Δ_1). The ratio β_1 measures what fraction of total variation lies in the non-gradient (curl and harmonic) components. When $\beta_1 = 0$, preferences are perfectly gradient (fully transitive), meaning there exists a score vector s such that $Y[i,j] = s[i] - s[j]$ for all edges. When $\beta_1 = 1$, preferences are entirely non-gradient (maximally cyclic), with zero projection onto any gradient space. In practice, real-world datasets typically exhibit $\beta_1 \in [0.5, 1.0]$, reflecting mixtures of consistent global trends and local intransitivities.

Borda Count [3] computes scores as $s\_BC[i] = \Sigma_j Y[i,j]$, the sum of pairwise preference values for each item. Since Y is skew-symmetric, this equals the number of pairwise wins minus losses when $Y[i,j] \in \{-1, 0, +1\}$, or weighted wins minus losses for continuous preference

values. Items are ranked by sorting scores in descending order: $\pi(i) < \pi(j)$ if $s\_BC[i] > s\_BC[j]$.

The Bradley-Terry model [4] assumes pairwise outcomes arise from underlying item strengths $\pi \in \mathbb{R}^{n}_{+}$ according to $P(i\ beats\ j) = \pi_i / (\pi_i + \pi_j)$. Given observed comparison data, we estimate π via maximum likelihood using the MM (Minorization-Maximization) algorithm [9]. The iterative updates converge to local maxima of the log-likelihood function, typically within 50 iterations on our datasets.

The MM (Minorization-Maximization) algorithm [9] provides efficient and numerically stable parameter estimation. Starting from uniform initialization $\pi^{(0)} = 1$, the algorithm iteratively updates each strength parameter via $\pi^{(t+1)}[i] = (\Sigma_j w[i,j]) / (\Sigma_j n[i,j]/(\pi^{(t)}[i] + \pi^{(t)}[j]))$, where $w[i,j]$ is the number of times i beat j and $n[i,j] = w[i,j] + w[j,i]$ is the total number of comparisons. This update is guaranteed to increase the log-likelihood at each iteration, converging to a local maximum. We iterate until convergence ($||\pi^{(t+1)} - \pi^{(t)}||\infty < 10^{-6}$) or maximum 100 iterations. The final ranking is produced by sorting log-strengths: $s = log(\pi)$.

Spectral Ranking [10] applies singular value decomposition to the weighted preference matrix $Y_W = Y \odot \sqrt{W}$. The ranking is derived from the first left singular vector scaled by the largest singular value: $s = U[:,0] \cdot \Sigma[0,0]$. This can be interpreted as finding the rank-1 approximation minimizing $\left||Y_W - s \cdot v^{T}|\right|_F$.

We evaluate rankings using Pairwise Accuracy (PA): $PA = \left(\frac{1}{|E|}\right) \Sigma_{(i,j)\in E} I\left[sign(Y_{i,j}) = sign\left(\pi^{(-1)}(i) - \pi^{(-1)}(j)\right)\right]$, is item i's rank position (lower values indicate higher ranks) and $\mathbb{1}[\cdot]$ is the indicator function. $PA \in [0, 1]$; higher values indicate better agreement with observed preferences.

To assess performance stability, we employ 5-fold cross-validation on edges: edges are randomly split into 5 folds; each fold trains on 80% of edges and evaluates PA on the held-out 20%, yielding mean ± std across folds. For MovieLens (1.29M edges), we subsample 20,000 edges before splitting and build compact matrices from the 3,104 active items within the subsample, keeping CV tractable while preserving a valid stability estimate.

Statistical significance of PA differences between method pairs is assessed via McNemar's test on the paired binary per-edge accuracy vectors (correct/incorrect per edge). With 6 pairwise comparisons among 4 methods, we apply Bonferroni correction: $\alpha = 0.05/6 \approx 0.0083$.

For top-k agreement we report two metrics: (a) Jaccard similarity $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, measuring set overlap of the top-k items; (b) Kendall's τ restricted to the union of top-k sets from both methods, measuring rank correlation on the most relevant items. $J = 1$ indicates identical item sets; $\tau = -1$ indicates completely reversed ordering.

We evaluate on three real-world datasets representing different domains and cyclicity characteristics. SUSHI3 [11] contains 5,000 users ranking 100 sushi types, yielding 4,809 edges with extremely high cyclicity (95.6%). This dataset reflects subjective taste preferences in Japanese cuisine where individual variation creates pervasive intransitivity.

MovieLens [12] contains 610 users rating 9,724 movies on a 1–5-star scale. We extract pairwise preferences by comparing ratings within each user, retaining only edges with at least 5 comparisons for reliability. The resulting graph has 8,452 edges with moderate cyclicity (82.2%), reflecting a mixture of consensus on blockbusters and individual taste variation.

Amazon Appliances [13] contains 47 users reviewing 48 kitchen appliances with ratings 1-5. This small dataset has only 227 edges with low cyclicity (55.2%). The limited user base makes this our most challenging dataset from a statistical estimation perspective. Table 1 summarizes dataset characteristics.

Table 1 – Dataset characteristics

Dataset	Users	Items	Edges	Avg Degree	Cyclicity β_1
SUSHI3	5,000	100	4,809	96.2	95.6%
MovieLens	610	9,724	8,452	1.7	82.2%
Amazon	47	48	227	9.5	55.2%

Note: Cyclicity computed using Equation (2). Higher values indicate more intransitive preferences.

All methods are implemented in Python 3.8 using NumPy 1.21 and SciPy 1.7. HodgeRank uses scipy. linalg. pinv with tolerance $rcond = max(M,N) \cdot \varepsilon \cdot \sigma_{max}$, where ε is machine epsilon and σ_{max} is the largest singular value of L. Near-zero eigenvalues below this threshold are treated as zero, which may amplify noise for sparse graphs with low algebraic connectivity λ_2. Bradley-Terry employs the MM algorithm [9] with convergence tolerance 10^{-6} and maximum 100 iterations.

Results and their discussion.

Tables 2-4 and Figures 1-5 present the main results for all methods across datasets. Bradley-Terry achieves highest average accuracy (0.767), followed by Borda (0.748), HodgeRank (0.739), and Spectral (0.678). However, performance varies substantially by dataset, with no universal dominance across all cyclicity levels.

Table 2 – Pairwise accuracy comparison across methods

Method	Amazon	MovieLens	SUSHI3	Average
BradleyTerry	0.579 [0.560±0.053]	0.873 [0.832±0.003]	0.849 [0.840±0.007]	0.767
Borda	0.579 [0.551±0.041]	0.818 [0.797±0.004]	0.848 [0.842±0.002]	0.748
HodgeRank	0.574 [0.569±0.032]	0.851 [0.825±0.003]	0.791 [0.785±0.006]	0.739
Spectral	0.509 [0.528±0.041]	0.799 [0.658±0.126]	0.727 [0.784±0.049]	0.678

Note: Full-data PA shown first; CV mean±std (5-fold) in brackets. Spectral shows high variance on MovieLens (±0.126), indicating instability on sparse moderate-cyclicity data.

Table 3 – McNemar's test (Bonferroni $\alpha = 0.0083$)

Method pair	Amazon	MovieLens	SUSHI3
BT vs Borda	$ns\ (p = 1.000)$	$\chi^2 = 33359, p < 0.0001$	$ns\ (\chi^2 = 0.2, p = 0.698)$
BT vs HodgeRank	$ns\ (p = 1.000)$	$\chi^2 = 9438, p < 0.0001$	$\chi^2 = 119.4, p < 0.0001$
BT vs Spectral	$\chi^2 = 13.1, p = 0.0003$	$\chi^2 = 54837, p < 0.0001$	$\chi^2 = 343.9, p < 0.0001$
Borda vs HodgeRank	$ns\ (p = 1.000)$	$\chi^2 = 8329, p < 0.0001$	$\chi^2 = 133.1, p < 0.0001$
Borda vs Spectral	$\chi^2 = 13.1, p = 0.0003$	$\chi^2 = 3051, p < 0.0001$	$\chi^2 = 317.3, p < 0.0001$
HodgeRank vs Spectral	$\chi^2 = 10.6, p = 0.0012$	$\chi^2 = 19513, p < 0.0001$	$\chi^2 = 69.0, p < 0.0001$

Note: Bonferroni-corrected α=0.0083; ns = not significant. Amazon ns reflects ~84% tied edges. BT = Bradley-Terry.

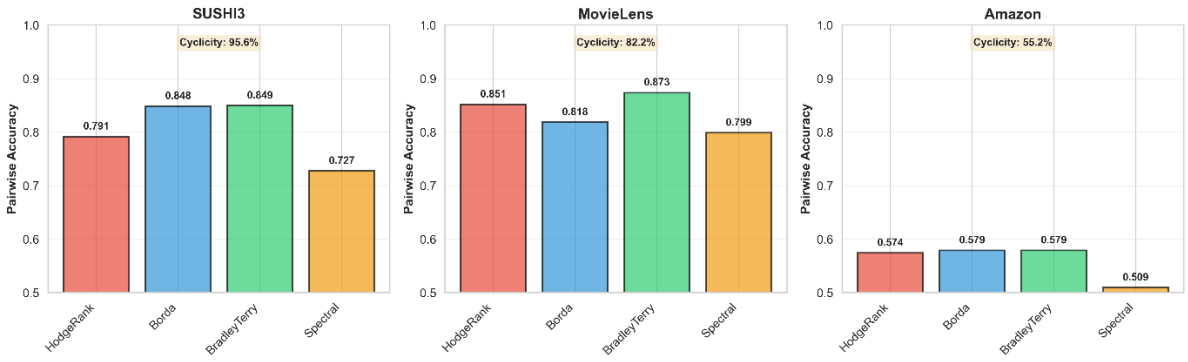

Figure 2 – Method Comparison by Dataset

On Amazon, all methods except Spectral achieve similar low accuracy ($PA \approx 0.58$); Spectral underperforms at 0.509, and McNemar's test finds no significant differences between Bradley-Terry, Borda, and HodgeRank (all $p > 0.1$). This reflects two compounding factors: (1) only 227 edges with 47 users, and (2) approximately 84% of edges are ties after averaging ratings across users, leaving fewer than 40 non-tied edges for meaningful discrimination. On MovieLens, Bradley-Terry achieves highest accuracy (0.873), with HodgeRank second (0.851). This dataset represents HodgeRank's optimal regime where moderate cyclicity provides sufficient gradient signal without excessive noise.

On SUSHI3, both Borda (0.848) and Bradley-Terry (0.849) significantly outperform HodgeRank (0.791) by 5.7–5.8 percentage points (McNemar $\chi^2 = 133.1$ and $\chi^2 = 119.4$ respectively, $p < 0.001$ after Bonferroni correction). Notably, Borda and Bradley-Terry are statistically equivalent ($\chi^2 = 0.2, p = 0.698$), indicating that on high-cyclicity data simple aggregation performs as well as the probabilistic model. For a graph with 4,809 edges, this translates to approximately 279 more pairwise errors for HodgeRank compared to Bradley-Terry, or 274 errors compared to Borda. Spectral Ranking consistently underperforms across all datasets, suggesting its rank-1 approximation approach is less suitable for ranking with incomplete comparisons.

Figure 1 reveals a non-monotonic relationship between cyclicity and HodgeRank performance, challenging conventional expectations. Contrary to uniform degradation, HodgeRank achieves highest accuracy at moderate cyclicity (PA = 0.851 at β_1 = 82.2%) while performing worse at both low (PA = 0.574 at β_1 = 55.2%) and high (PA = 0.791 at β_1 = 95.6%) extremes. This inverted-U pattern suggests HodgeRank requires both sufficient structural signal and adequate data density to function effectively.

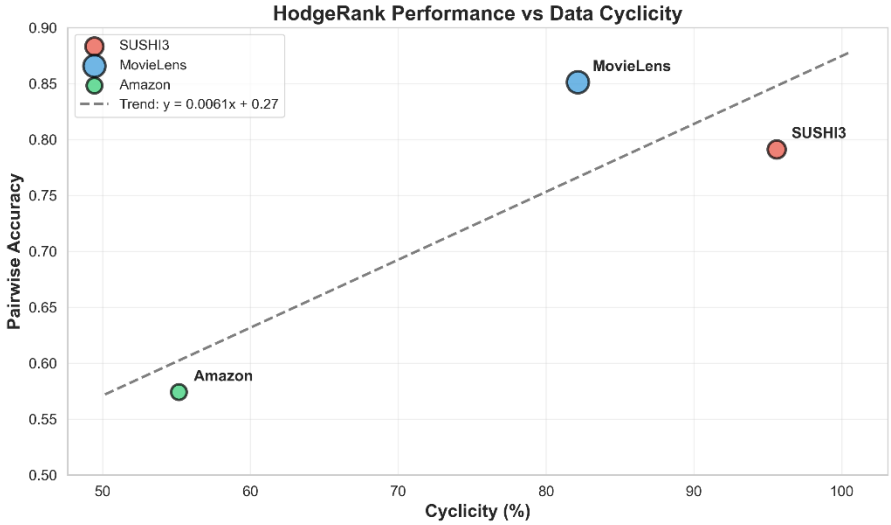

Figure 1 – HodgeRank Performance vs Data Cyclicity

The Amazon result is particularly noteworthy. Despite having the lowest cyclicity, HodgeRank achieves worst performance, suggesting data sparsity defeats gradient-based optimization through Laplacian conditioning issues. With mean degree 9.5, L has small algebraic connectivity λ_2 (the second-smallest eigenvalue, also called the Fiedler value), leading to a large condition number $\kappa(L) = \frac{\lambda_{max}}{\lambda_2}$. The pseudoinverse $L^{\dagger}$ effectively divides by these small eigenvalues, amplifying noise in the preference signal—a phenomenon analogous to numerical instability in ill-conditioned linear systems. The MovieLens result (β_1 = 82.2%) represents HodgeRank's peak where 18% gradient variance provides sufficient discriminative power. On SUSHI3, β_1 = 95.6% means gradient explains only 4.4% of variation—projecting onto this low-variance subspace loses most information.

Figures 3-5 examine top-10 rankings qualitatively, revealing differences masked by aggregate metrics. While pairwise accuracy provides useful summary statistics, users of ranking systems [8] primarily interact with top-k results, making qualitative agreement on highly-ranked items critical for practical deployment.

Table 4 – Top-10 agreement (Jaccard J / Kendall τ)

Method pair	Amazon J / τ	MovieLens J / τ	SUSHI3 J / τ
BT vs Borda	0.429 / −0.077	0.000 / −0.516	1.000 / 0.733
BT vs HodgeRank	0.333 / −0.010	0.538 / 0.051	0.250 / −0.283
BT vs Spectral	0.176 / −0.324	0.000 / −0.484	1.000 / 0.778
Borda vs HodgeRank	0.538 / 0.256	0.000 / −0.516	0.250 / −0.283
Borda vs Spectral	0.429 / 0.187	0.667 / 0.545	1.000 / 0.689
HodgeRank vs Spectral	0.176 / −0.132	0.000 / −0.632	0.250 / −0.317

Note: J = Jaccard similarity of top-10 item sets (1 = identical, 0 = no overlap); τ = Kendall's τ on union of top-10 sets (1 = same order, −1 = reversed).

Table 4 reveals key patterns. On SUSHI3, Bradley-Terry, Borda, and Spectral achieve three-way perfect top-10 consensus ($J = 1.000$ for all three pairs, τ ranging from 0.689 to 0.778), all selecting premium tuna varieties. HodgeRank shares only 25% of items with either method ($J = 0.250$) and ranks the overlap in near-opposite order ($\tau = -0.283$), selecting structurally distinct items such as kurumaebi and ishigakidai. On MovieLens, Bradley-Terry and Borda have zero top-10 overlap ($J = 0.000, \tau = -0.516$) despite similar PA: Borda selects broadly popular films (Pulp

Fiction, The Godfather), while Bradley-Terry selects films that consistently win direct comparisons but are less frequently rated.

Figure 3 (Amazon, β_1=55.2%) shows moderate consensus despite low cyclicity. Product_13 appears in multiple methods' top-5 (green), but substantial method-specific choices (red) indicate sparse data causes different signal extraction.

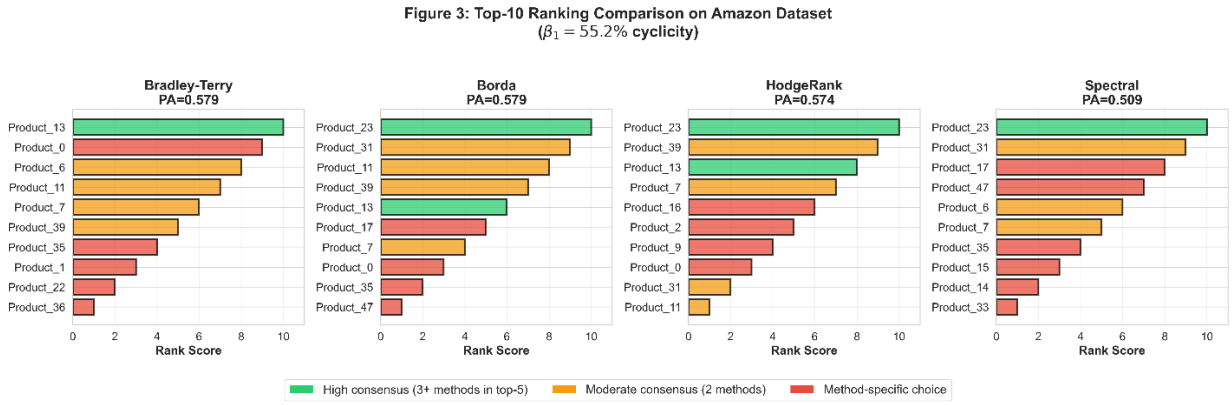

Figure 3 – Amazon Top-10 Rankings

Figure 4 (MovieLens, $\beta_1 = 82.2\%$) reveals a clear split between method families. Borda and Spectral converge on broadly popular blockbusters — both include Pulp Fiction, The Godfather, and Star Wars in their top-10 ($J = 0.667, \tau = 0.545$), reflecting aggregate popularity signals. Bradley-Terry and HodgeRank instead select niche high-quality films that consistently win direct pairwise comparisons but are less frequently rated: Bradley-Terry and HodgeRank partially converge on niche high-quality films ($J = 0.538$), both selecting Captain Fantastic, Paradise Lost, and Trial. Bradley-Terry additionally ranks Man for All Seasons and Man Bites Dog, while HodgeRank additionally selects Tea with Mussolini and 7th Voyage of Sinbad. This family split explains the zero overlap between Borda and Bradley-Terry ($J = 0.000, \tau = -0.516$) despite their similar aggregate PA of 0.818 vs 0.873.

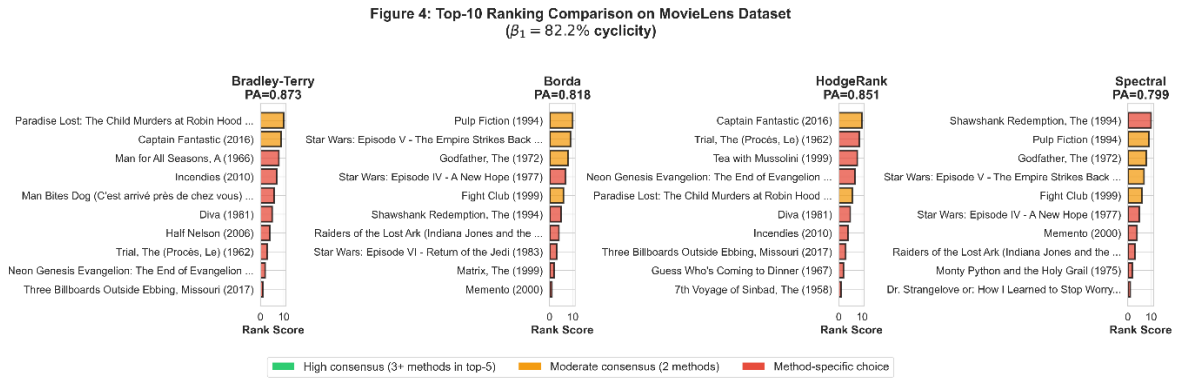

Figure 4 – MovieLens Top-10 Rankings

Figure 5 (SUSHI3, β_1=95.6%) reveals the most dramatic divergence. Bradley-Terry, Borda, and Spectral achieve strong consensus (green), all three ranking the same premium tuna varieties in their top-5: chu_toro, toro, negi_toro, maguro, negi_toro_maki. These selections align with known Japanese cuisine preferences. HodgeRank diverges substantially, selecting kurumaebi (prawn), ishigakidai (parrotfish), zuke, hiramasa—items with minimal consensus with classical methods (Jaccard=0.250, Kendall's $\tau = -0.283$ vs both Bradley-Terry and Borda).

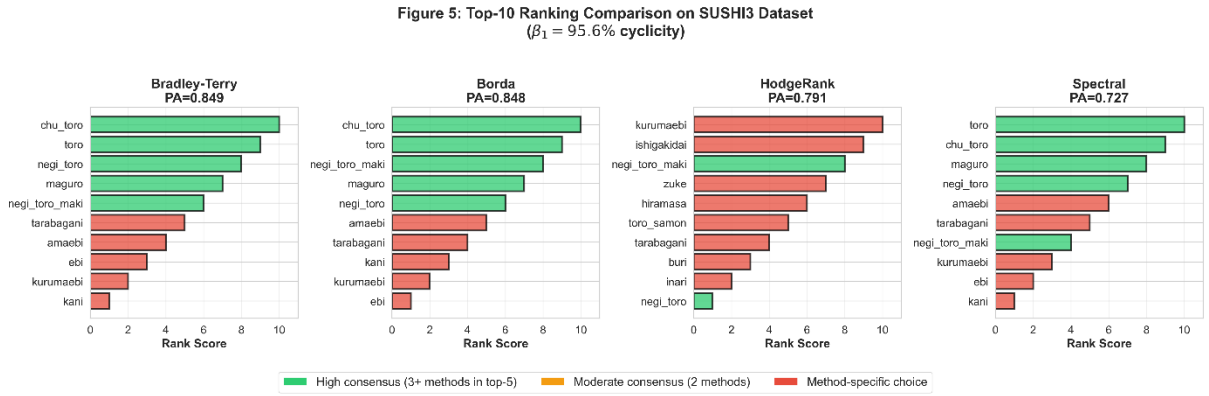

Figure 5 – SUSHI3 Top-10 Rankings

This divergence is severe. While HodgeRank's PA of 0.791 trails Bradley-Terry's 0.849 by only 5.8 percentage points, the methods recommend fundamentally different items. A recommendation system [8] deploying HodgeRank would present users with entirely different sushi than one using Bradley-Terry or Borda. This demonstrates that aggregate metrics alone provide insufficient information for deployment decisions—similar overall accuracy can mask dramatic differences in specific recommendations.

Our results reveal two key findings about HodgeRank's behavior on cyclic data. First, the relationship between cyclicity and performance is non-monotonic rather than uniformly degrading. HodgeRank achieves peak performance at moderate cyclicity (MovieLens, β_1=82.2%) while underperforming at both extremes. The Amazon result demonstrates that data sparsity can defeat gradient-based optimization [6] through conditioning issues independent of cyclicity levels. On SUSHI3, projection onto the 4.4%-variance gradient space discards most information contained in the dominant cyclic component.

Second, at high cyclicity, HodgeRank selects fundamentally different top items than classical methods despite similar aggregate accuracy. This reflects how gradient optimization emphasizes structural patterns in the low-variance subspace over aggregate popularity captured by simple aggregation methods [3]. Whether this difference is beneficial depends on application requirements—surfacing consensus preferences favors classical methods, while discovering structural patterns invisible to aggregation might favor HodgeRank.

Bradley-Terry's consistent performance (average PA=0.767) across all cyclicity levels makes it the most reliable choice when data characteristics are unknown. Its probabilistic framework [4] naturally absorbs inconsistencies as stochastic noise in strength parameters rather than treating them as structural contradictions. Borda Count performs similarly (average PA=0.748) through direct aggregation [3] without enforcing global consistency constraints.

For practitioners deploying ranking systems [8], we recommend: (1) Measure cyclicity using Equation (2) before selecting methods—this diagnostic provides crucial structural information; (2) Use Bradley-Terry as default when characteristics are unknown—its consistent performance minimizes risk; (3) Consider HodgeRank only when $\beta_1 < 0.85$ AND data has sufficient density (suggested minimum: mean degree > 10); (4) On highly cyclic data ($\beta_1 > 0.90$), prefer simple methods that impose no structural constraints; (5) Always examine top-k rankings qualitatively, not just aggregate metrics—similar PA values can mask dramatic recommendation differences.

Limitations include: (1) Only three datasets, though spanning wide cyclicity range—the non-monotonic pattern of HodgeRank requires more datasets for confirmation; (2) Default hyperparameters for all methods; (3) CV for MovieLens uses a 20,000-edge subsample rather than the full 1.29M edges due to computational constraints; (4) Amazon results are difficult to interpret due to ~84% tied edges after preference averaging; (5) No examination of temporal or dynamic preferences. Future work should pursue theoretical analysis of cyclicity thresholds, develop adaptive methods blending approaches based on local structure, and extend to temporal

preferences. Applications of higher-order Hodge theory [7] to simplicial complexes may overcome limitations of gradient-based approaches on cyclic pairwise data.

Conclusion.

This paper presents a comprehensive comparison of ranking methods across varying cyclicity levels, combining aggregate metrics with quantitative and qualitative top-k analysis. Through experiments on three real-world datasets (Amazon, MovieLens, SUSHI3) representing different domains and cyclicity regimes (55%-96%), we characterize how data structure affects algorithmic performance.

Our key findings are: (1) Bradley-Terry [4] achieves highest average accuracy (0.767) and most consistent performance across cyclicity levels, making it the recommended general-purpose method; (2) HodgeRank [5] exhibits non-monotonic behavior, performing best at moderate cyclicity (PA=0.851 at β_1=82.2%) but degrading at both low (PA=0.574 at β_1=55.2%) and high (PA=0.791 at β_1=95.6%) extremes; (3) At high cyclicity (SUSHI3), HodgeRank stands alone against a three-method consensus (Jaccard=0.250 vs all others, $\tau = -0.283$), despite aggregate accuracy differing by only 5.8 percentage points, a difference confirmed statistically significant (McNemar $\chi^2 = 119.4, p < 0.001$); (4) Data sparsity (Amazon) defeats HodgeRank through Laplacian conditioning issues independent of cyclicity effects.

These findings challenge assumptions about topologically sophisticated methods universally outperforming classical baselines [1]. HodgeRank's gradient-based approach [5,6] proves sensitive to both data sparsity and excessive cyclicity. For practitioners deploying ranking systems [8], method selection requires understanding both data characteristics (cyclicity, density) and application requirements (consensus versus structural discovery). By providing both negative results (HodgeRank's limitations) and positive guidelines (when each method excels), we enable principled, context-aware algorithmic selection.

References

1. Spearing, J., Tawn, J., Irons D. and Paulden, T. (2023). "Modeling Intransitivity in Pairwise Comparisons with Application to Baseball Data," J. Comput. Graph. Statist., vol. 32, no. 4, pp. 1383–1392, https://doi.org/10.1080/10618600.2023.2177299

2. Singh, R. & Davidov, O. (2025). "The Analysis of Paired Comparison Data in the Presence of Cyclicality and Intransitivity," https://doi.org/10.48550/arXiv.2406.11584

3. Fox, N. B. & Bruyns, B. (2025). "An Evaluation of Borda Count Variations Using Ranked Choice Voting Data," Soc. Choice Welf., https://doi.org/10.1007/s00355-025-01638-2

4. Wu, W., Niezink, N. & Junker, B. (2022). "A Diagnostic Framework for the Bradley–Terry Model," J. R. Stat. Soc. Ser. A, vol. 185, no. S2, pp. S461–S484, https://doi.org/10.1111/rssa.12959

5. Oliveira, L. R. N., Lunardi, J. T., Calçada, M., Pereira, A.L., Jesuz, D. A. F. & Costa, C. (2024). "HodgeRank as a New Tool to Explore the Structure of a Social Representation," Front. Phys., vol. 12, art. 1333727, 2024. https://doi.org/10.3389/fphy.2024.1333727

6. Lim, L.H. (2020). "Hodge Laplacians on graphs," SIAM Rev., vol. 62, no. 3, pp. 685-715, https://doi.org/10.1137/18M1223101

7. Schaub, M. T., Benson, A. R., Horn, P., Lippner, G. & Jadbabaie, A. (2020). "Random walks on simplicial complexes and the normalized Hodge 1-Laplacian," SIAM Rev., vol. 62, no. 2, pp. 353-391, https://doi.org/10.1137/18M1201019

8. Li, C., Ishak, I., Ibrahim, H., Zolkepli, M., Sidi, F. & Li, C. (2023). "Deep learning-based recommendation system: systematic review and classification," IEEE Access, vol. 11, pp. 113790-113835, https://doi.org/10.1109/ACCESS.2023.3323353

9. J. Newman, M. E. (2023). " Efficient computation of rankings from pairwise comparisons," The Journal of Machine Learning Research, vol. 24, pp. 11259-11283, https://dl.acm.org/doi/abs/10.5555/3648699.3648937

10. Ruijian Han, Wenlu Tang & Yiming Xu. (2025). "Statistical inference for pairwise comparison models", https://doi.org/10.48550/arXiv.2401.08463

11. Yanagi, T, Ikeda, S, Sukegawa, N. & Takano, Y. (2025). " Privacy-preserving recommender system. using the data collaboration analysis for distributed datasets" PLOS ONE, vol. 20, pp. 1-14, https://doi.org/10.1371/journal.pone.0319954

12. Fan, Yu-Chen, Ji, Yitong, Zhang, Jie, Sun, Aixin. (2024). "Our Model Achieves Excellent Performance on MovieLens: What Does It Mean?" ACM Trans. Inf. Syst., vol. 42, pp. 1-25, https://doi.org/10.1145/3675163

13. Yupeng Hou, Jiacheng Li, Zhankui He, A Yan, Xiusi Chen & Julian McAuley. (2024). "Bridging Language and Items for Retrieval and Recommendation.", https://doi.org/10.48550/arXiv.2403.03952

ЦИКЛДІК ҚАЛАУЛАР ДЕРЕКТЕРІ БОЙЫНША РЕЙТИНГТЕУ ӘДІСТЕРІНІҢ ӨНІМДІЛІГІ: САЛЫСТЫРМАЛЫ ЗЕРТТЕУ

Аңдатпа. *Рейтингтеу әдістері ұсыныс жүйелері, әлеуметтік таңдау және шешім қабылдау салаларында кеңінен қолданылатын деректерді талдаудың негізгі құралдары болып табылады. HodgeRank сияқты әдістер қалаулар деректерінің топологиялық қасиеттерін пайдаланғанымен, олардың нақты циклдік деректер жиынтықтарындағы эмпирикалық өнімділігі жеткілікті зерттелмеген. Бұл мақалада төрт рейтингтеу әдісі — HodgeRank, Borda Count, Bradley-Terry және Spectral Ranking — циклдық деңгейлері әртүрлі (55%-дан 96%-ға дейін) үш деректер жиынтығында кешенді салыстырылады. Өнімділік жұптық дәлдік (PA) арқылы бағаланады; тұрақтылық 5 қатпарлы қиылыспалы тексеру арқылы расталады; айырмашылықтардың статистикалық маңыздылығы Бонферрони түзетімімен Макнемар сынағы арқылы тексеріледі; top-k келісімі Жаккар ұқсастығы J және Кендалл τ коэффициенті арқылы сандық бағаланады.*

Негізгі нәтижелер HodgeRank өнімділігі мен циклдық арасындағы сызықтық емес байланысты ашады: PA орташа циклдықта шыңына жетеді (0.851, β_1 = 82%-да), бірақ төмен (0.574, 55%-да) және жоғары (0.791, 96%-да) шекті мәндерде нашарлайды. Bradley-Terry ең жоғары орташа дәлдік (0.767) пен жоғары қиылыспалы тексеру тұрақтылығын көрсетеді. Жоғары циклдік SUSHI3 деректерінде үш әдіс — Bradley-Terry, Borda және Spectral — top-10 элементтерінде толық консенсусқа жетеді (J = 1.000), ал HodgeRank барлық үш әдістен түбегейлі ауытқиды (J = 0.250, τ = −0.283) — бұл ауытқу Макнемар сынағымен статистикалық тұрғыдан расталды (χ^2 = 119.4, p < 0.001), жиынтық дәлдіктің айырмашылығы тек 5.8 пайыздық тармақ болса да. Нәтижелер рейтингтеу жүйелерін орналастыру туралы шешімдерді қабылдауда жиынтық дәлдік метрикаларының жеткіліксіздігін көрсетеді.

Түйін сөздер: *рейтингтеу әдістері, HodgeRank, циклдық, қалаулар агрегациясы, жұптық салыстыру, Bradley-Terry моделі, Borda Count, комбинаторлық Ходж теориясы.*

ПРОИЗВОДИТЕЛЬНОСТЬ МЕТОДОВ РАНЖИРОВАНИЯ НА ЦИКЛИЧЕСКИХ ДАННЫХ О ПРЕДПОЧТЕНИЯХ: СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ

Аннотация. *Методы ранжирования являются фундаментальными инструментами анализа данных, применяемыми в рекомендательных системах, теории социального выбора и принятии решений. Хотя методы, подобные HodgeRank, используют топологические свойства данных о предпочтениях, их эмпирическая производительность на реальных циклических наборах данных остаётся недостаточно изученной. В данной статье представлено комплексное сравнение четырёх методов ранжирования — HodgeRank, Borda Count, Bradley-Terry и Spectral Ranking — на трёх наборах данных с различными уровнями цикличности (от 55% до 96%). Производительность оценивается по попарной точности (PA); устойчивость подтверждается 5-кратной перекрёстной проверкой; статистическая значимость различий проверяется критерием Макнемара с поправкой Бонферрони; согласованность top-k результатов измеряется индексом Жаккара J и коэффициентом Кендалла τ.*

Ключевые результаты выявляют нелинейную зависимость между цикличностью и производительностью HodgeRank: PA достигает пика при умеренной цикличности (0.851 при $\beta_1 = 82\%$), но снижается как при низкой (0.574 при 55%), так и при высокой (0.791 при 96%) цикличности. Bradley-Terry демонстрирует наивысшую среднюю точность (0.767) и высокую стабильность по результатам перекрёстной проверки. На высокоцикличных данных SUSHI3 три метода — Bradley-Terry, Borda и Spectral — достигают полного консенсуса в top-10 ($J = 1.000$), тогда как HodgeRank принципиально расходится со всеми тремя ($J = 0.250$, $\tau = -0.283$) — это расхождение статистически подтверждено критерием Макнемара ($\chi^2 = 119.4$, $p < 0.001$) при разнице в попарной точности всего 5.8 п.п. Результаты показывают, что агрегированных метрик точности недостаточно для принятия обоснованных решений о развёртывании систем ранжирования.

Ключевые слова: *методы ранжирования, HodgeRank, цикличность, агрегация предпочтений, попарное сравнение, модель Bradley-Terry, метод Борда, комбинаторная теория Ходжа.*

Авторлар туралы мәлімет

Саимжан Әлиақбар Айбарұлы	магистрант, Қазақстан-Британ техникалық университеті,Алматы қ., Қазақстан E-mail: akbarsaimzhan@gmail.com

Сведение об авторах

Саимжан Алиакбар Айбарулы	магистрант, Казахстанско-Британский технический университет, г. Алматы, Казахстан, E-mail: akbarsaimzhan@gmail.com

Information about the authors

Saimzhan Aliakbar Aibaruly	Master's student, Kazakh-British Technical University, Almaty, Kazakhstan, E-mail: akbarsaimzhan@gmail.com