Accuracy, Exposure, and Serendipity in Movie Recommendation

Tulio Castro[1], Gisele Tessari Santos[1]

Centro Universitário Ibmec BH, Brazil
tulio.castro14@gmail.com gisele.tessari@ibmec.edu.br

Abstract. Accuracy-oriented recommenders concentrate exposure on a small fraction of the catalog. On a large Letterboxd–IMDb dataset we factor the pipeline into representation and decision policy and vary one factor at a time. With the policy held fixed at inner-product ranking, none of five representations exposes more than 14% of the catalog, and KGCN attains the best Recall@50. Holding the representation fixed at KGCN and changing only the policy takes Coverage@50 from 0.14 to 0.65, and the same ablation replicates on a second representation. Exposure is therefore governed mainly by the decision stage, and the mechanism is measurable: item embedding norms correlate with popularity (ρ up to 0.97), so inner-product ranking carries a popularity prior that cosine removes and ε-greedy bypasses. We recommend KGCN with inner-product ranking and tail exploration: it trades 9% of its recall for 4.4× the coverage, Pareto-dominates every other representation, and preserves substantially more recall than the cosine alternatives. Decomposing Seren@K into hit rate and surprise-per-hit shows that rewarding surprise raises surprise-per-hit but lowers the hit rate, so measured serendipity falls. That paradox is partly an artifact of offline evaluation.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Computing methodologies** → **Neural networks**.

Keywords: catalog exposure, GNN, heterogeneous graph, Letterboxd, ranking policy, recommendation, serendipity

1. INTRODUÇÃO

Recomendadores otimizados para acurácia tendem a repetir itens populares: o recall sobe, mas a exposição se concentra e se estreita ao longo do tempo [Klimashevskaia et al. 2024]. No conjunto Letterboxd–IMDb usado aqui, quatro baselines clássicos expõem menos de 9% dos 51 mil filmes elegíveis. O custo é a perda de serendipidade, isto é, de descoberta simultaneamente inesperada e relevante [Kotkov et al. 2016].

Duas famílias de solução atuam em estágios distintos. A primeira enriquece a *representação* com embeddings sobre grafos bipartidos, heterogêneos ou de conhecimento [He et al. 2020; Schlichtkrull et al. 2018; Hu et al. 2020; Wang et al. 2019]. A segunda altera a *política de decisão* com exploração explícita [Li et al. 2010]. Como são usualmente avaliadas em protocolos diferentes, permanece aberta a questão de qual estágio controla a exposição do catálogo.

Este trabalho mede separadamente a contribuição de cada estágio para acurácia, cobertura e serendipidade sob um único protocolo offline. A hipótese é que a exposição é governada principalmente pela política de ranqueamento. O desenho fatorial varia um fator por vez: cinco representações sob produto interno e quatro políticas sobre uma representação fixa, com a ablação replicada em duas representações.

A contribuição central é uma atribuição empírica sustentada por quatro frentes. Primeiro, constrói-se um conjunto Letterboxd–IMDb com cerca de 20 milhões de interações e limpeza auditada. Segundo,

Copyright©2026 Permission to copy without fee all or part of the material printed in KDMiLe is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

comparam-se baselines, GNNs, KGCN e políticas sob o mesmo protocolo, com intervalos bootstrap e testes pareados. Terceiro, uma ablação com representação fixa, replicada em dois modelos, atribui a variação de cobertura ao ranqueamento e identifica seu mecanismo geométrico. Quarto, a decomposição da serendipidade em taxa de acerto e surpresa por acerto explica por que recompensar surpresa amplia a cobertura, mas reduz a serendipidade observada.

2. TRABALHOS RELACIONADOS

Representação e decisão. LightGCN simplifica a propagação colaborativa [He et al. 2020], enquanto KGCN incorpora semântica externa com atenção dependente do usuário [Wang et al. 2019]; ambos mantêm o ranqueamento por produto interno e privilegiam acurácia. Em outra linha, políticas exploratórias alteram a decisão por bônus de incerteza ou substituição aleatória [Li et al. 2010], mas sua avaliação offline é enviesada sem propensões ou logs aleatorizados [Li et al. 2011; Gilotte et al. 2018].

Exposição e lacuna. Métricas além da acurácia incluem surpresa, novidade e cobertura [Ge et al. 2010; Kotkov et al. 2016; Duricic et al. 2023], e estudos longitudinais mostram concentração crescente do catálogo [Klimashevskaia et al. 2024]. Entretanto, representação e política raramente são isoladas no mesmo protocolo. Este trabalho preenche essa lacuna variando cada estágio separadamente sobre a mesma base e as mesmas métricas.

3. DADOS

A coleta expandiu uma conta-semente pelas conexões públicas do Letterboxd [Letterboxd Limited 2026][1]. O *snapshot* de abril de 2026 contém 13.977 usuários, 422.932 filmes, 20.168.063 interações e 4.265.332 arestas sociais. O IMDb [Internet Movie Database 2025][2] forneceu ano, gêneros, elenco e direção; somente perfis e conteúdos públicos foram coletados.

A linkagem determinística combinou título normalizado e ano, com tolerância de ± 1, resultando em 291.522 filmes canônicos e 94,1% das interações cobertas. A limpeza removeu registros sem evidência de consumo e inconsistências temporais, descartando menos de 4% das linhas. A matriz permanece esparsa (0,34% dos pares observados) e concentrada (Gini 0,913), embora a atividade mediana seja de 895 interações por usuário.

Após filtragem 5-core, a divisão temporal produz treino em 1990–2023 (3.190.938 interações; 10.265 usuários; 51.285 filmes), validação em 2024 (642.256; 8.866) e teste em 2025–abr. 2026 (620.317; 8.538). Outros 573 mil registros de usuários frios foram preservados. A versão pública no Kaggle [Silva and Santos 2026][3] pseudonimiza usuários e omite dados pessoais; interações e arestas sociais seguem CC BY-NC 4.0, e os metadados, a licença não comercial do IMDb.

4. METODOLOGIA

Todo recomendador avaliado aqui é a composição de dois estágios executados em sequência, e não um modelo único (Figura 1). O primeiro estágio, a *representação*, lê o grafo e produz um vetor por usuário e um vetor por filme. O segundo estágio, a *política de decisão*, recebe esses vetores, atribui um escore a cada filme elegível e devolve as 50 primeiras posições. A lista final passa uma vez por cada estágio, nunca por dois recomendadores em cadeia. Uma configuração é, portanto, sempre um par: KGCN com produto interno, ou HetLGCN com cosseno, e é esse par que a Tabela I reporta em cada linha. Dizer que a representação está fixa significa que o primeiro estágio é o mesmo em todas as linhas comparadas e apenas o segundo muda.

[1] `https://letterboxd.com`. Acesso: abr. 2026.
[2] `https://developer.imdb.com/non-commercial-datasets/`. Acesso: abr. 2026.
[3] `https://www.kaggle.com/datasets/tliocastro/letterboxd-sample-apr-2026`. Acesso: abr. 2026.

Notação e embeddings. U é o conjunto de usuários e I o de filmes elegíveis; $u \in U$ denota usuário, $i \in I$ filme, e hist(u) o histórico de treino de u. Modelos paramétricos produzem embeddings e_u, e_i, com d=64 ou d=32 conforme a arquitetura.

4.1 Grafo Heterogêneo

O grafo $G = (V, E)$ é definido por V, conjunto de nós, e E, conjunto de arestas. Ele tem 4 tipos de nó (usuário, filme, gênero, pessoa), 7 tipos de aresta, 110.310 nós e cerca de 9,9M arcos bidirecionais. O painel (a) da Figura 1 mostra um trecho com todos os tipos: `rated` e `liked` separam nota alta de curtida, `follows` é a única relação entre nós do mesmo tipo e a única mantida direcionada, e as quatro relações de conteúdo ligam o filme a gêneros e pessoas. Caminhos usuário–filme–pessoa–filme propagam semântica. Cada modelo consome um subgrafo distinto: o subgrafo social (`rated`, `liked`, `follows`), o grafo completo, ou apenas o lado de conteúdo do item.

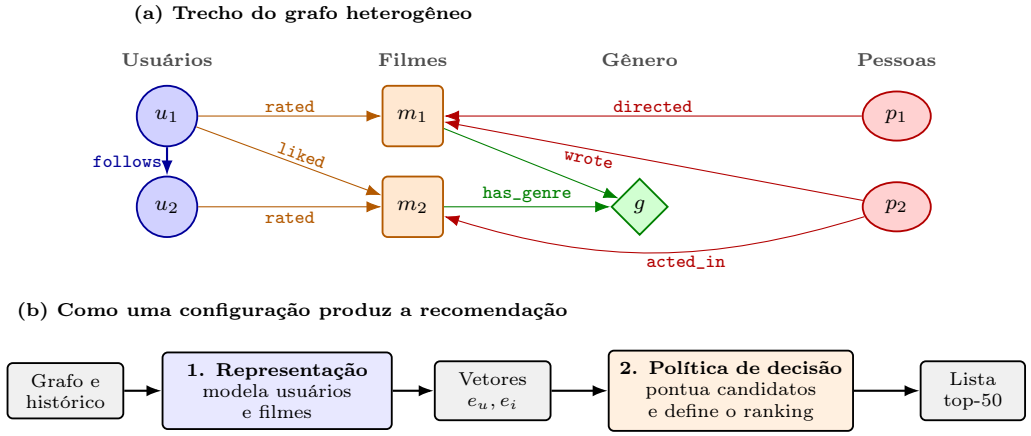

Fig. 1. (a) Exemplo do grafo heterogêneo. (b) Pipeline de avaliação: a representação produz e_u, e_i, e a política de decisão os converte em uma lista top-50.

Opções avaliadas. No estágio de representação, consideram-se Popularity, Item-KNN, BPR-MF, LightGCN-3, KGCN e HetLGCN; Popularity e Item-KNN pontuam diretamente, sem produzir embeddings. No estágio de decisão, consideram-se produto interno, cosseno e cosseno com surpresa, cada um com ou sem ε-greedy.

Comparações experimentais. A comparação de representações mantém a política fixa em produto interno. A ablação sobre KGCN fixa essa representação e varia a política; a replicação sobre HetLGCN repete o mesmo procedimento com uma segunda representação.

4.2 Arquiteturas Avaliadas

Avaliam-se seis representações: cinco na comparação com produto interno e o HetLGCN, usado na replicação da ablação de política. *Popularity* ranqueia por popularidade global e *Item-KNN* soma as similaridades cosseno item–item com o histórico do usuário; nenhuma aprende embeddings. *BPR-MF* [Rendle et al. 2009] aprende embeddings com a perda BPR (*Bayesian Personalized Ranking*), que favorece o item positivo sobre um negativo amostrado, e *LightGCN-3* [He et al. 2020] os propaga no grafo bipartido usuário–filme por médias normalizadas de vizinhos, em 3 camadas.

HetLGCN estende esse mesmo LightGCN ao grafo heterogêneo. Na Eq. 1, $\mathbf{h}_v^{(l)}$ é a representação do nó v na camada l, t o tipo de v, $\mathcal{R}(t)$ as relações incidentes, $\mathcal{N}_r(v)$ os vizinhos pela relação r, $\widehat{A}_r$

a adjacência normalizada e L o total de camadas:

$$\mathbf{h}_v^{(l+1)} = \frac{1}{|\mathcal{R}(t)|} \sum_{r \in \mathcal{R}(t)} \sum_{u \in \mathcal{N}_r(v)} \widehat{A}_r[u,v] \cdot \mathbf{h}_u^{(l)}, \quad \mathbf{h}_v = \frac{1}{L+1} \sum_{l=0}^{L} \mathbf{h}_v^{(l)}. \tag{1}$$

Na configuração reportada, o HetLGCN consome o subgrafo social (`rated`, `liked`, `follows`); é essa a representação mantida fixa na ablação de política.

A abordagem proposta emprega o *KGCN* [Wang et al. 2019], que propaga no grafo de conhecimento do item (`movie`→`genre`/`person`). Para usuário u, relação r e vizinho v, o escore de atenção $g(u,r) = \langle e_u, e_r \rangle$ é normalizado por softmax nos vizinhos amostrados $\mathcal{N}_S(i)$,

$$\alpha_{uiv} = \frac{\exp(g(u, r_{iv}))}{\sum_{v' \in \mathcal{N}_S(i)} \exp(g(u, r_{iv'}))}, \qquad \tilde{e}_i = \sum_{v \in \mathcal{N}_S(i)} \alpha_{uiv} e_v, \tag{2}$$

e $\tilde{e}_i$ combina-se com e_i pelo agregador soma, $v_i^u = \tanh\big(W(e_i + \tilde{e}_i) + b\big)$, cujo produto interno com e_u alimenta a perda BPR. Como a atenção depende de u, o mesmo vizinho de conteúdo pesa de forma diferente para usuários diferentes, fato retomado na Seção 5.2. Os hiperparâmetros estão na legenda da Tabela I.

4.3 Políticas de Decisão

Avaliam-se três regras para converter os vetores em ranking, cada uma diferindo da anterior em um único aspecto. A primeira é o *produto interno* $\langle e_u, e_i \rangle$, padrão dos modelos treinados com BPR. A segunda é o *cosseno*, que difere apenas por dividir pelas normas, descartando a magnitude dos vetores. Sobre ela define-se uma pontuação com surpresa explícita: na Eq. 3, α e β pesam os dois objetivos e c_u é o centróide histórico,

$$\text{score}(u,i) = \underbrace{\alpha \cos(e_u, e_i)}_{\text{relevância}} + \underbrace{\beta\,(1 - \cos(e_i, c_u))}_{\text{surpresa}}, \tag{3}$$

com $c_u = \text{L2}\Big(\frac{1}{|\text{hist}(u)|} \sum_{j \in \text{hist}(u)} e_j\Big)$, a média L2-normalizada dos filmes no histórico. O peso β controla a surpresa: $\beta = 0$ recupera o cosseno puro e $\beta > 0$ favorece filmes distantes do gosto habitual. A varredura de β da Seção 5.4 usa essa forma.

A terceira regra acrescenta *exploração ε-greedy* a qualquer uma das anteriores: cada posição da lista é, com probabilidade $\varepsilon = 0{,}10$, preenchida por um item não visto amostrado uniformemente, em vez do melhor item segundo o escore. A substituição ocorre nas posições finais da lista de $K_{\text{rec}} = 50$, de modo que o efeito é analisado em @50. Nenhuma das três regras tem parâmetros aprendidos além dos da representação, o que mantém a ablação restrita ao estágio de decisão.

4.4 Métricas

A avaliação segue protocolos offline de ranqueamento top-K [Herlocker et al. 2004; Shani and Gunawardana 2011]. Para cada usuário, todos os filmes elegíveis são pontuados e os itens de treino removidos; não se amostram negativos, pois métricas amostradas distorcem comparações [Krichene and Rendle 2020]. Recall@K e NDCG@K medem acurácia, a segunda ponderando a posição do acerto [Järvelin and Kekäläinen 2002]; Coverage@K é a fração do catálogo que aparece nas listas; e Nov@K é a média de $-\log_2$ da popularidade relativa dos itens recomendados, de modo que valores maiores indicam itens menos populares.

Seja rec_u^K a lista top-K de u, rel_u os itens relevantes no teste, $H_u(K) = \text{rec}_u^K \cap \text{rel}_u$ os acertos, e_i o embedding do item e c_u o centróide histórico de u. A serendipidade por usuário é a surpresa média

entre os acertos, e zero quando não há acerto [Ge et al. 2010; Kotkov et al. 2016]:

$$\text{Seren@K}(u) = \begin{cases} \frac{1}{|H_u(K)|} \sum_{i \in H_u(K)} \left(1 - \cos(e_i, c_u)\right), & |H_u(K)| > 0, \\ 0, & |H_u(K)| = 0. \end{cases} \tag{4}$$

Reporta-se a média macro sobre os $|U_{\text{test}}|$ usuários de teste. Como os usuários sem acerto contribuem com zero, essa média fatora exatamente em dois termos:

$$\underbrace{\text{Seren@K}}_{\text{serendipidade medida}} = \underbrace{\text{HitUser@K}}_{\text{fração com acerto}} \times \underbrace{\text{CSeren@K}}_{\text{surpresa por acerto}}, \tag{5}$$

em que HitUser@K é a fração de usuários com ao menos um acerto no top-K e CSeren@K é a média de Seren@K(u) restrita a esses usuários. A identidade é exata e separa duas causas que a métrica agregada confunde: recomendar itens menos óbvios e simplesmente acertar menos. Seren@K é calculada no espaço de embeddings do HetLGCN para todos os modelos, garantindo uma representação comum de comparação.

5. RESULTADOS

A Tabela I organiza onze configurações nas três comparações do desenho fatorial da Figura 1 (8.538 usuários, 51.285 itens, $\approx$ 620k interações relevantes): representações com produto interno, políticas sobre KGCN e políticas sobre HetLGCN. A primeira varia a representação; as duas seguintes fixam uma representação e variam a política. Comparar suas amplitudes permite atribuir cada métrica a um estágio.

Nov@50 maior indica itens menos populares; CSeren@50 corresponde a Seren@50/HitUser@50 (Eq. 5), e o negrito marca o melhor valor por coluna. Os hiperparâmetros foram selecionados na validação por Recall@50, com *early stopping* de paciência 5: BPR-MF e LightGCN-3 usam d=64, lr 0,01 e 4 negativos; LightGCN-3 e HetLGCN usam 3 camadas; KGCN usa d=32, single-hop e S=8; e a exploração usa ε=0,10.

Table I. Resultados das representações e políticas no conjunto de teste.

	R@50	NDCG@50	C@50	Nov@50	HitUser@50	Seren@50	CSeren@50
Representações com produto interno							
Popularity	0.0224	0.0283	0.0082	10.50	0.4399	0.0263	0.0599
Item-KNN	0.0253	0.0338	0.0867	10.97	0.4708	0.0376	0.0799
BPR-MF	0.0252	0.0313	0.0394	10.59	0.4792	0.0301	0.0629
LightGCN-3	0.0230	0.0284	0.0215	10.51	0.4416	0.0259	0.0586
KGCN	**0.0308**	**0.0387**	0.1407	11.50	**0.5162**	0.0555	0.1075
Políticas sobre KGCN							
produto interno + ε-greedy	0.0279	0.0363	0.6246	12.07	0.4966	0.0536	0.1080
cosseno	0.0237	0.0304	0.2025	12.67	0.4687	**0.0833**	0.1778
cosseno + ε-greedy	0.0215	0.0284	0.6533	13.13	0.4507	0.0814	**0.1806**
Políticas sobre HetLGCN							
cosseno	0.0106	0.0163	0.6903	12.98	0.3390	0.0312	0.0921
cosseno + ε-greedy	0.0096	0.0152	**0.8563**	**13.42**	0.3216	0.0312	0.0971

5.1 Atribuição do Efeito ao Estágio

Sob produto interno, variar a representação move C@50 de 0,008 a 0,141, amplitude de 0,133. Sobre o KGCN fixo, variar a política leva a cobertura de 0,141 a 0,653, amplitude de 0,512; no HetLGCN, a replicação a leva de 0,690 a 0,856. A variação produzida pela decisão é, portanto, maior e se repete em duas representações.

A acurácia apresenta o padrão complementar: o R@50 do KGCN na comparação de representações (0,0308) é 2,9 vezes o melhor valor das políticas sobre HetLGCN (0,0106), diferença que a troca

de política não recupera. Assim, a representação estabelece o teto de acurácia, enquanto a política controla principalmente a exposição; a Figura 2 resume esse compromisso.

Contrastes pareados. Os contrastes usam bootstrap pareado sobre os mesmos 8.538 usuários, 10.000 reamostras, em Recall@50. Cada passo da escada de políticas tem custo mensurável e nenhum IC cruza zero: sobre o KGCN, trocar produto interno por cosseno custa $\Delta = -0{,}0071$ (IC95% $[-0{,}0087; -0{,}0055]$) e acrescentar ε-greedy custa $\Delta = -0{,}0029$ (IC95% $[-0{,}0034; -0{,}0024]$); sobre o HetLGCN, o ε-greedy custa $\Delta = -0{,}0010$ (IC95% $[-0{,}0012; -0{,}0008]$), todos com $p < 0{,}001$. O custo é real, mas pequeno diante do ganho de exposição: a configuração recomendada na Seção 5.3 ainda *supera* o Item-KNN, o melhor baseline em recall, por $\Delta = +0{,}0026$ (IC95% $[+0{,}0011; +0{,}0042]$; $p = 0{,}0008$), expondo sete vezes mais catálogo.

5.2 Mecanismo da Exposição

Modelos treinados com BPR ranqueiam por $\langle e_u, e_i \rangle = \|e_u\| \, \|e_i\| \cos(e_u, e_i)$. Como a norma dos itens se correlaciona com a popularidade de treino (Spearman $\rho = 0{,}37$ no BPR-MF e $\rho = 0{,}97$ no LightGCN-3), o produto interno incorpora um *prior* de popularidade. Isso explica por que o LightGCN-3, apesar da propagação no grafo, expõe menos catálogo que o BPR-MF (0,022 contra 0,039).

O cosseno remove a norma do escore, enquanto o ε-greedy insere itens não vistos independentemente dela. Com essas políticas, a popularidade típica dos itens recomendados cai de cerca de 2.000 interações no Popularity para 440 no cosseno sobre KGCN e 320 com exploração. O KGCN também reduz parcialmente o *prior*: sua atenção dependente do usuário (Eq. 2) altera a representação do item conforme o perfil, deslocando o ranking global e elevando cobertura e serendipidade mesmo com produto interno.

5.3 Configuração Recomendada

A configuração recomendada é KGCN com produto interno e ε-greedy: R@50 = 0,0279 e C@50 = 0,625. Em relação ao KGCN puro, ela cede 9,4% do recall por 4,4 vezes mais cobertura e é a única que supera todas as representações com produto interno simultaneamente nos dois eixos. Também supera o Item-KNN em recall por $\Delta = +0{,}0026$ (IC95% $[+0{,}0011; +0{,}0042]$; $p = 0{,}0008$), expondo sete vezes mais catálogo.

Adicionar o cosseno eleva a cobertura apenas de 0,625 para 0,653, mas reduz o recall de 0,0279 para 0,0215. Por isso, o pequeno ganho adicional de exposição não compensa a perda de acurácia. Configurações sobre HetLGCN alcançam cobertura ainda maior, porém com recall substancialmente menor. A recomendação prioriza, portanto, o melhor compromisso entre acurácia e exposição no protocolo offline, sem afirmar superioridade de utilidade percebida (Seção 6).

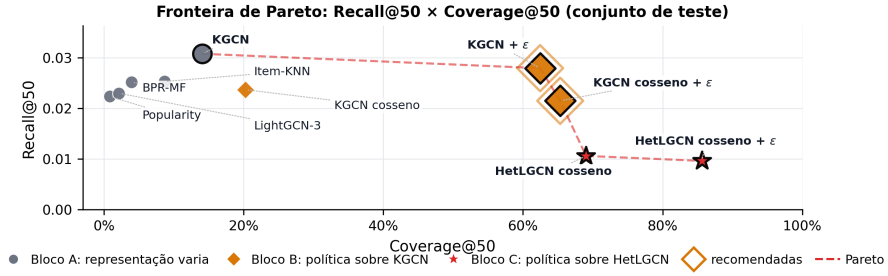

Fig. 2. Fronteira de Pareto entre Recall@50 e Coverage@50 no teste. O losango destacado indica a configuração recomendada; a linha tracejada conecta os pontos não dominados.

5.4 O Paradoxo da Serendipidade

A varredura de β testa se recompensar surpresa aumenta a serendipidade medida. A Tabela II mostra que não: a cobertura cresce de 0,705 ($\beta = 0$) para 0,995 ($\beta = 0{,}5$), mas Seren@50 cai de 0,0327 para 0,0150 em $\beta = 1{,}0$. A decomposição da Eq. 5 mostra que os dois fatores se movem em direções opostas. A surpresa por acerto sobe de forma monótona e acentuada, de 0,088 para 1,007; nesse extremo, os poucos acertos são praticamente ortogonais ao centróide do usuário, ou seja, o termo de surpresa faz exatamente o que lhe foi pedido. O que colapsa é o outro fator: a fração de usuários com algum acerto cai de 0,373 para 0,015. Como Seren@50 é o produto dos dois, o colapso do acerto domina.

Table II. Efeito de β ($\alpha = 1$; varredura na validação). CSeren@50 sobe enquanto HitUser@50 colapsa; o produto (Seren@50) cai.

β	R@50	C@50	HitUser@50	CSeren@50	Seren@50
0,0	0.0117	0.705	**0.3728**	0.0877	**0.0327**
0,3	0.0060	0.973	0.2551	0.1235	0.0315
0,5	0.0028	**0.995**	0.1621	0.1826	0.0296
0,7	0.0008	0.994	0.0693	0.3694	0.0256
1,0	0.0001	0.612	0.0149	**1.0067**	0.0150

O paradoxo, portanto, não é que a surpresa deixe de funcionar, e sim que uma métrica condicionada a acertos observados não consegue creditá-la. Separam-se aí duas afirmações frequentemente reunidas na literatura: tornar a lista mais surpreendente e produzir acertos serendipitosos. A primeira cresce, em CSeren@50 e Nov@50, enquanto Seren@50 diminui. A distinção também delimita o alcance do resultado: CSeren@50, medido só sobre acertos observados, é o componente menos sensível a itens relevantes porém ausentes do log, e HitUser@50 é exatamente o componente que esse viés penaliza (Seção 6).

6. DISCUSSÃO E LIMITAÇÕES

O desenho fatorial indica que a representação estabelece o teto de acurácia e a política controla principalmente a exposição. Essa conclusão, porém, vem de um protocolo observacional: apenas itens presentes no teste são considerados relevantes, embora um filme de cauda longa possa estar ausente por nunca ter sido exposto. Esse viés penaliza políticas exploratórias, sobretudo em HitUser@K. O crescimento de CSeren@50 enquanto Seren@50 cai sugere, portanto, que parte do paradoxo é um artefato de medição. Uma avaliação não enviesada exigiria propensões registradas, logs aleatorizados ou teste online [Li et al. 2011; Gilotte et al. 2018], recursos ausentes no *snapshot* utilizado.

Bootstrap e testes pareados mostram diferenças consistentes no log, mas não eliminam o viés causal. O conjunto também reflete uma rede social filtrada, notas majoritariamente positivas e a exclusão de usuários frios da avaliação. Assim, os resultados identificam *onde* a exposição é decidida, mas não demonstram maior utilidade ou melhor experiência online.

7. CONCLUSÃO

Este trabalho separou representação em grafo e política de decisão e as variou isoladamente sobre um conjunto Letterboxd–IMDb de larga escala. Com a política fixa, nenhuma das cinco representações expõe mais de 15% do catálogo; fixando a representação no KGCN, a política leva a cobertura de 0,14 a 0,65, e a ablação se replica em uma segunda representação. A exposição é decidida no estágio de ranqueamento, e o mecanismo é mensurável: a norma do embedding correlaciona-se com popularidade (ρ de 0,37 a 0,97), e o produto interno a explora enquanto o cosseno a descarta. A acurácia, em contrapartida, segue limitada pela representação, e o KGCN detém o melhor R@50. Recomenda-se

KGCN com produto interno e ε-greedy: a configuração cede 9,4% do recall por 4,4 vezes mais catálogo, domina em Pareto todas as outras representações e preserva mais acurácia que as alternativas com cosseno.

Segue disso que serendipidade não é apenas distância no espaço de embeddings. Decomposta em taxa de acerto e surpresa por acerto, a surpresa cresce monótona com β, até a ortogonalidade em $\beta = 1$, enquanto a taxa de acerto colapsa: recompensar surpresa funciona, mas o protocolo offline não consegue creditá-la. Trabalhos futuros devem enfrentar cold-start, Thompson Sampling e, sobretudo, avaliação contrafactual com propensões ou feedback explícito.

REFERENCES

Duricic, T., Kowald, D., Lacic, E., and Lex, E. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. *Frontiers in Big Data* vol. 6, pp. 1251072, 2023.

Ge, M., Delgado-Battenfeld, C., and Jannach, D. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. pp. 257–260, 2010.

Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A., and Dollé, S. Offline A/B testing for recommender systems. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM)*. pp. 198–206, 2018.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 639–648, 2020.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22 (1): 5–53, 2004.

Hu, Z., Dong, Y., Wang, K., and Sun, Y. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020 (WWW)*. pp. 2704–2710, 2020.

Internet Movie Database. IMDb non-commercial datasets. `https://developer.imdb.com/non-commercial-datasets/`, 2025.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20 (4): 422–446, 2002.

Klimashevskaia, A., Jannach, D., Elahi, M., and Trattner, C. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 2024. arXiv:2308.01118.

Kotkov, D., Wang, S., and Veijalainen, J. A survey of serendipity in recommender systems. *Knowledge-Based Systems* vol. 111, pp. 180–192, 2016.

Krichene, W. and Rendle, S. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. pp. 1748–1757, 2020.

Letterboxd Limited. Letterboxd — Social film discovery. `https://letterboxd.com`, 2026.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*. pp. 661–670, 2010.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*. pp. 297–306, 2011.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*. pp. 452–461, 2009.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., and Welling, M. Modeling relational data with graph convolutional networks. In *Proceedings of the 15th Extended Semantic Web Conference (ESWC)*. pp. 593–607, 2018.

Shani, G. and Gunawardana, A. Evaluating recommendation systems. In *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds.). Springer, Boston, MA, pp. 257–297, 2011.

Silva, T. C. and Santos, G. T. Letterboxd sample Apr 2026. `https://www.kaggle.com/datasets/tliocastro/letterboxd-sample-apr-2026`, 2026. Kaggle dataset, CC BY-NC 4.0.

Wang, H., Zhao, M., Xie, X., Li, W., and Guo, M. Knowledge graph convolutional networks for recommender systems. In *Proceedings of The World Wide Web Conference 2019 (WWW)*. pp. 3307–3313, 2019.