Browsing Amazon's Book Bubbles

Paul Bouchaud[1,2] [0000-0002-0310-8194]

[1] Center for Social Analysis and Mathematics (EHESS)
[2] Complex Systems Institute of Paris Île-de-France (CNRS)
paul.bouchaud@ehess.fr

Abstract. This study investigates Amazon's book recommendation system, uncovering cohesive communities of semantically similar books. The confinement within communities is extremely high, a user following Amazon's recommendations needs tens of successive clicks to navigate away. We identify a large community of recommended books endorsing climate denialism, COVID-19 conspiracy theories, and advocating conservative views on social and gender issues. Performing a collaborative filtering analysis, relying on Amazon users reviews, reveals that books reviewed by the same users tend to be co-recommended by Amazon. This study not only contributes to addressing a gap in the literature by examining Amazon's recommender systems, but also highlights that even non-personalized recommender systems may pose systemic risks by suggesting content with foreseeable negative effects on public health and civic discourse.

Keywords: Amazon · Recommender System · Filter Bubble · Collaborative Filtering

1 Introduction

"Recommendations are discovery, offering surprise and delight with what they help uncover for you. Every interaction should be a recommendation" in this study, we explore the claim made by Smith and Linden when discussing two decades of recommender systems at Amazon [21].

As early as the late 1990s, Amazon has embraced collaborative filtering [13, 7]; the evaluation of items co-purchase likelihood [17], has been pivotal for the platform, driving up to 30% of Amazon.com's page views in 2015 [18]. However, concerns have been raised about algorithmic curation potentially reinforcing exposure to like-minded content and amplifying biases [14]. While collaborative filtering algorithms alone do not inherently narrow content diversity [26], their interaction with user preferences can create "echo chambers" [23, 6]. Such effects have been observed on platforms like YouTube [8] and Twitter [2].

Despite Amazon's influence, serving over 181 million EU users [22], e-commerce platforms have received limited research attention from algorithmic auditors. Previous studies have highlighted partisan disparities in science book consumption [19] and feedback loops reinforcing user interests [6]. During the COVID-19 pandemic, Amazon faced public criticism for promoting vaccine misinformation

[4], and with studies revealing a prevalence of vaccine-hesitant books [20] and misleading search results [11].

This study employs a sock-puppet audit methodology [16] to characterize Amazon's non-personalized recommendations across a wide range of non-fiction books on the French Amazon website. We reveal tight communities of semantically similar books and identify a community co-recommending books supporting various contrarian viewpoints, from climate denialism to COVID-19 conspiracy theories. Through collaborative filtering analysis based on user reviews, we show that books reviewed by the same users tend to be recommended together. This research highlights potential systemic risks of non-personalized recommender systems on civic discourse and public health.

2 Methods

2.1 Data Collection

Amazon presents product recommendations in multi-page carousels under various labels such as 'Customers who viewed this item also viewed' and 'Frequently Bought Together'. This study focuses on non-personalized recommendations an unlogged user would encounter.

We employed a snowball sampling strategy, starting with the bestsellers from 18 non-fiction book categories on Amazon.fr, listed in [1]. Using an automated web browser that reset after each page visit, we collected recommendations and metadata for each book. We iterated the collection three times, considering the books recommended at least twice in the previous round.

In total, from the initial pool of 1 725 bestsellers, we gathered recommendations for 60 298 books between October 28 and November 4, 2023, capturing an average of 85.8% of Amazon's suggestions for each book. To assess temporal stability, we compared this dataset with a prior collection from August-September 2023, finding a 64.4% overlap in recommendations.

Our dataset comprises primarily French books (92.3%), with 81.4% as printed books, 12.3% as Kindle ebooks, and 3.6% as audiobooks. Amazon consistently recommends books in similar formats (97.9% on average). Our subsequent analysis focuses exclusively on printed books, removing redundant formats to avoid duplication.

2.2 Graph Construction

The recommendations gathered, we establish an unweighted directed graph, denoted as G, in which the vertices represent books (designated as v_i). A link between v_i and v_j is established if Amazon recommends book v_j on the page of book v_i. Filtering out non-fetched books, we end up with a graph $G = (V, E)$, with V the set of $|V| = 48\,636$ books, and E the set of $|E| = 391\,664$ edges.

2.3 Characterization

Community Structure. We employ the Leiden algorithm [25] with the Constant Potts Model [24] as the quality function to detect communities in G. This approach overcomes both the resolution limit inherent in modularity maximization [5] and Louvain's arbitrarily badly connected communities. Establishing a resolution profile we determine the resolution parameter, γ, imposing an upper limit on inter-community link density, which ensures the stability of our partitions.

Subsequently, we quantify book recommendation homophily within communities through both the fraction of a vertex's neighbors belonging to the same community [9] and, beyond their first-degree neighbors, through random walks. Specifically, we initiate 25 random walks starting from each book and compute the average length of the walks a random surfer need to transition out of the community they initiated the walk from.

Semantic Analysis. We employ TF-IDF and Non-Negative Matrix Factorization to analyze book titles and summaries. This classical approach was chosen over recent neural topic models due to its efficiency, interpretability, and simplicity, as discussed in [27]. Our analysis compares summary embeddings between books connected by recommendations in graph G and those that are not. Additionally, we measure semantic diversity within book communities relative to the entire corpus. This diversity is quantified using the geometric mean of the standard deviation of summary embeddings, a metric introduced in [12]. To ensure robustness, we conducted a sensitivity analysis, confirming that our results remain consistent across embedding dimensions ranging from 25 to 70.

2.4 Collaborative Filtering

Although the current implementation is unknown, Amazon historically used item-to-item collaborative filtering to recommend products [13]. This method suggest products by examining items frequently purchased together. To gain insight into Amazon recommender systems, and without without access to purchase or rating data, we gathered the "verified purchase" reviews for 25 151 books from main communities in G. We then performed item-to-item collaborative filtering based on users reviews, while we acknowledge reviews offer stronger signals of (dis)agreement than simple ratings or purchases.

Overall, we collected 419 460 reviews by 245 734 unique reviewers, averaging 20.5 reviews per book (median 7.0). In contrast, Amazon's can leverage an average of 748.7 ratings per book (median 52.0), in addition to purchase and navigation data, to power its recommender systems. We then compare the overlap coefficient between user sets who reviewed each pair of books and Amazon's actual book recommendations.

2.5 Search Results

To investigate how users might enter a community through Amazon search results, we conducted searches on topics with clear scientific consensus: Climate Change and COVID-19. Specifically, we performed the following queries (in French): "Climate Change CO2," "Global Warming," "IPCC" for Climate Change, and "COVID-19," "COVID pandemics," "COVID vaccine" for COVID-19.

Starting on November 1, 2023, we performed these searches daily at noon for five consecutive days. For each query, we collected the first result page, sorted by Amazon's default algorithm or by decreasing average user ratings. Other rankers such as increasing/decreasing price or publication date yielded too few books in G for meaningful analysis.

3 Results

3.1 Characterization

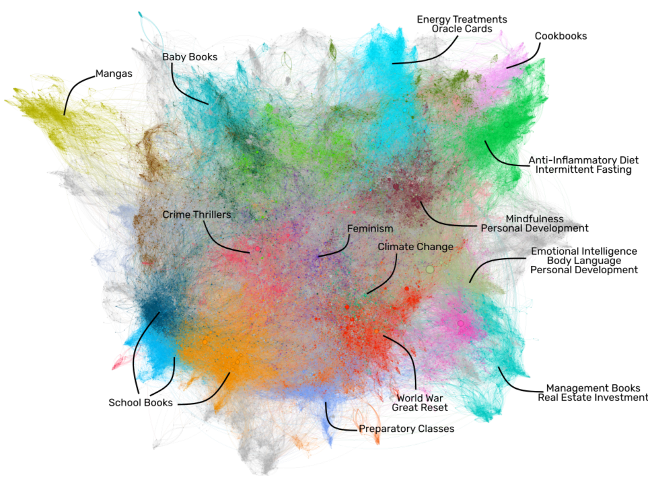


Fig. 1: Graph of Amazon book recommendations G [48 636 books, 391 664 edges], spatialized via ForceAtlas2 [10]. Vertices are color-coded by community, and their size is proportional to their in-degree.

The Leiden algorithm partitions G with high modularity ($Q = 0.86$), identifying 61 communities that encompass over 90% of the books. Figure 1 shows G with color-coded main communities. Homophily analysis reveals that 88.8% of recommended books are in the same community as the current book, rising to 94.9% for "Frequently Bought Together" items. Additionally, 53.2% of Amazon's suggested books share the same category as the current book. The likelihood that two randomly chosen books from a community belong to the same Amazon category is 5.1 times higher than for two random books. On average, 91.1% of books by authors with at least five books in G are in the same community.

Random walks based on Amazon's recommendations show strong confinement. From the 61 largest communities (covering 90% of books), 75.7% of surfers remain in their starting community after three clicks. On average, it takes 24.9 clicks (median 11) to leave a community. Confinement varies by community; for example, it takes an average of 6.9 clicks (median 4) to leave the social science community (356 books), while over 77.8 clicks (median 68) are needed to exit the coloring books community (380 books).

Figure 1 shows keywords extracted via TF-IDF for various communities in G, revealing diverse topics such as cartomancy, personal development, mangas and crime thrillers. Semantic diversity within a community is, on average, 57.9% lower than in the overall corpus. The average pairwise cosine similarity between books connected by an edge in G is 1.72 times higher than for books within the same community but not connected, and 5.41 times higher than for books from different communities without an edge. Similarly, along random walks, the similarity between a book and those recommended by Amazon after three clicks is 36.9% higher than the average similarity among books from the same community.

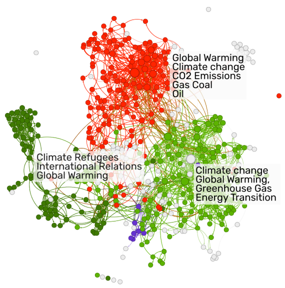
3.2 Collaborative Filtering

The user review data is sparse; among 25 100 books, users reviewed an average of 1.7 books (median 1.0). Books reviewed by the same user are 8.7 times more likely to belong to the same community in G and 6.1 times more likely to share the same Amazon book category compared to randomly selected books, considering users who reviewed at least 5 books. The cosine similarity between summary embeddings of books reviewed by the same user is 2.6 times higher than for random pairs.

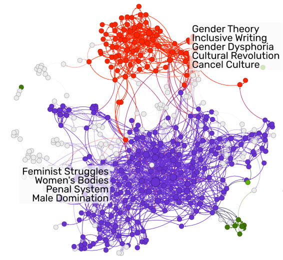
For the 10 108 books with at least 10 "verified purchase" reviews, we analyzed pairwise reviewer overlap. In 58.1% of cases, the book with the highest reviewer overlap with a seed book is from the same community in G, and in 34.5% of cases, it is recommended by Amazon (i.e., connected by an edge in G). The average reviewer overlap is 15.6 times higher for books connected by an edge in G compared to random pairs from the same community, and 9.2 times higher for books within the same community than for books from different communities.

3.3 Case studies

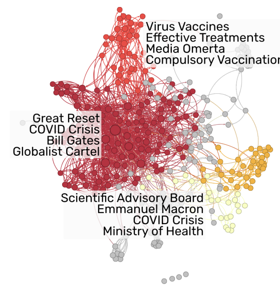
We manually curated lists of books in G discussing: Climate Change [146 books], Gender Issues (including gender identity, expression, and equality) [162 books] and COVID-19 [101 books]. These topics were chosen due to the relative abundance of available books and their social significance; aligning for instance, with the European Commission's topics of interest in their initiatives addressing misinformation [3], and the systemic risks defined in the Digital Services Act. We excluded books were these topics where not the main focus of the discussion, as well as fiction books.



(a) Graph induced by 146 books discussing Climate Change [459 books, 1 991 edges]



(b) Graph induced 117 books relating to Gender Issues [439 books, 1 862 edges]



(c) Graph induced by 101 books discussing COVID-19 [304 books, 1 620 edges]

Fig. 2: Two-hop recommendation graphs. Vertices are color-coded by G community (smaller communities are in white), their size is proportional to their in-degree.

Climate Change The 146 books addressing Climate Change primarily fall into two G communities: 42.5% in the community depicted in green on Figure 2a and 30.1% in the red community. A manual inspection reveals that books in the green community align with the scientific consensus on climate change, while 84.1% of those in the red community reject it. Apart from the two main communities, books discussing the geopolitical aspects of climate change (depicted in dark green) are accessible within two clicks from the Climate Change seed books. On average, when a user consults a climate-denialists book Amazon recommends 94.1% of books from the red community, and similarly, alongside pro-climate books, Amazon recommends 90.7% of books from the green community.

Gender Issues The graph of Amazon's book recommendations induced by two-hops from 117 books related to gender issues is depicted in Figure 2b. Again, two main communities emerge, encompassing 56.4% (in violet) and 20.5% (in red) of the books. While the books in the violet community address feminist struggles, male domination, women's rights, sexual violence, and engaged in discussions on gender identity and expression (hereafter designed as feminist/queer community), the books in the red community discuss cancel culture, inclusive writing, and "wokeism" (hereafter designed as conservative views). Interestingly, the community represented in red in Figure 2b corresponds to the same community in G that embeds climate-denialist books shown in Figure 2a. On average, when a user consults a feminist/queer book, Amazon recommends 95.7% of such books, and similarly, alongside conservative books, Amazon recommends 95.8% of conservative books.

COVID-19 The analysis of 101 books within G addressing the COVID-19 pandemic reveals that 92.1% are situated within the community, hereinafter called 'contrarian', that otherwise encompasses climate-denialist books and advocates conservative views on gender issues, previously depicted in red. In Figure 2c, the recommendation graph derived from these books is presented with vertices color-coded based on their sub-communities, detected at a higher resolution than for G. The extraction of keywords from sub-community book summaries exposes distinct thematic focuses, aligning with established taxonomies of COVID-19-related disinformation [15]. The three main sub-communities: i) endorse New World Order and Great Reset conspiracy theories; ii) challenge the established scientific consensus on vaccinations and their side effects; and iii) discuss pandemic management.

We emphasize that the assignment of a book to a particular community is not of the author's will. For instance, the book "COVID-19: The Great Reset" by Klaus Schwab and Thierry Malleret lies in the same community as those relaying conspiracy theories, due to Amazon's algorithm, which suggests five such books alongside Schwab's works. This is likely because, among the fetched books, the top 10 books with the highest reviewer overlap with Schwab's book are conspiracy-related.

Contrarian Community To gain further insight into this contrarian community, the third-largest community in G with 1 776 books, we isolated it, conducted a community detection analysis at a higher resolution than for G, and is displayed on Figure 3. Employing TF-IDF to extract keywords from book summaries within sub-communities, reveals a broad range of topics, including Freemasonry, French Politics, Foreign Policy, Cancel Culture, and Great Reset conspiracy theories.

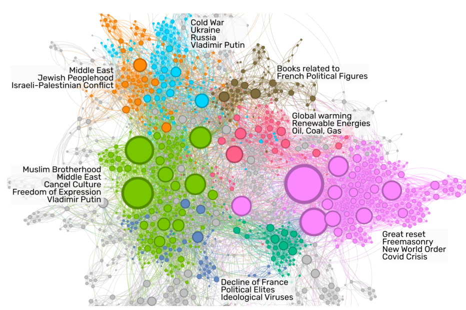


Fig. 3: Sub-graph of G induced by books of the contrarian community.

We can leverage the set of users book reviews to understand why various contrarian viewpoints coexist within the same recommendation community rather than being in distinct topic-specific communities. We observe that the average overlap among users who reviewed climate-denialist books and those reviewing books holding conservative views on gender is 6.8 times higher than the overlap between the sets of reviewers of pro-climate books and of feminist/queer books. Similarly, the reviewer overlap between COVID-19 related books and climate-denialist books (resp. conservative books) is 4.9 (resp. 4.2) times higher than the overlap between COVID-related books and pro-climate books (resp. feminist/queer books).

Beyond encompassing various disinformation narratives, this community stands out for its confinement. An average number of 15.5 (median 8) successive clicks are required for a random surfer following Amazon's recommendation to leave the contrarian communities, while 6.1 (median 4) and 6.3 (median 4) are required to leave, respectively, the feminism/queer and pro-climate communities. When the random surfers leave the contrarian community, 8.4% of them end up in the World War II and French history books community [625 books], 6.6% in community related to Personal Development and Communication [1 229 books]. Within the contrarian community, an average of 9.0 (median 5) clicks are required to leave the sub-communities, 76.6% of the random walks leaving the climate-denialist sub-community emerge in the COVID-19 sub-community, 11.2% emerge in the conservatism subcommunity.

3.4 Search Results

Performing search queries related to climate change with Amazon's default algorithmic ranking, it was observed that 51.1% of the first 10 results provided misleading information about the scientific consensus (52.5% were in the above identified contrarian community), this fraction increased to 64.1% when ranked by decreasing average user ratings. For COVID-19-related searches, when ranked according to Amazon's default algorithm, 71.7% of the top 10 results contain misinformation about COVID-19 pandemics, when ordered by decreasing average user ratings, the fraction increases to 91.1%.

4 Discussion

Our study of Amazon's book recommendation system revealed highly modular communities with strong homophily. The community of recommendation are made up of books that are semantically close, with a poorer semantic diversity than the overall book corpus; books by the same author tend to be embedded in the same community, and with users requiring many clicks to navigate away.

Examining recommendation graphs for Climate Change, Gender Issues, and COVID-19 books, we identified a community promoting climate denialism, COVID-19 conspiracy theories, and conservative views on social issues; readily accessible through search queries. Books arguing opposing viewpoints were in separate topic-specific communities. Collaborative filtering analysis showed that books reviewed by the same users tend to be recommended together, explaining why diverse contrarian viewpoints coexist within the same community.

While our data collection focused on the French Amazon market and excluded most fiction books, we believe this study provides valuable insights into a platform that significantly influences book distribution. We reveal that Amazon's non-personalized recommendation system tends to confine users within homogeneous communities, including one that spreads misinformation on climate change and COVID-19.

This study demonstrates that even non-personalized algorithms based on seemingly objective criteria can generate recommendations with potential negative effects on public health and civic discourse. Our findings contribute to discussions on algorithmic regulation, highlighting that explainability and transparency alone may not mitigate the systemic risks targeted by regulations like the Digital Services Act.

Acknowledgments. Paul Bouchaud acknowledges the Jean-Pierre Aguilar fellowship from the CFM Foundation for Research and the resources provided by the Complex Systems Institute of Paris.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Bouchaud, P.: Browsing Amazon's Book Bubbles (Nov 2023), https://hal.science/hal-04308081, working paper or preprint
2. Bouchaud, P., Chavalarias, D., Panahi, M.: Crowdsourced audit of twitter's recommender systems. Sci Rep **13**(1) (oct 2023). https://doi.org/10.1038/s41598-023-43980-4
3. Comission, E.: Second call for the european narrative observatory to fight disinformation post-covid 19
4. Dreisbach, T.: On amazon, dubious "antiviral" supplements proliferate amid pandemic (Jul 2020)
5. Fortunato, S., Barth'elemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. U.S.A. **104**(1), 36–41 (jan 2007). https://doi.org/10.1073/pnas.0605965104
6. Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., Zhang, Y.: Understanding echo chambers in e-commerce recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (jul 2020). https://doi.org/10.1145/3397271.3401431
7. Hardesty, L.: The history of amazon's recommendation algorithm (Dec 2022)
8. Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., Wojcieszak, M.: Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations (2022)
9. Interian, R., Ribeiro, C.C.: An empirical investigation of network polarization. Applied Mathematics and Computation **339**, 651–662 (dec 2018). https://doi.org/10.1016/j.amc.2018.07.066
10. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PLoS ONE **9**(6), e98679 (jun 2014). https://doi.org/10.1371/journal.pone.0098679
11. Juneja, P., Mitra, T.: Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. ACM (may 2021). https://doi.org/10.1145/3411764.3445250
12. Lai, Y.A., Zhu, X., Zhang, Y., Diab, M.: Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 1739–1746. European Language Resources Association, Marseille, France (May 2020)

13. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing **7**(1), 76–80 (2003). https://doi.org/10.1109/MIC.2003.1167344
14. Pariser, E.: The filter bubble: What the internet is hiding from you. Penguin Books (2012)
15. Posetti, J., Kalina, B.: Disinfodemic: deciphering covid-19 disinformation. https://unesdoc.unesco.org/ark:/48223/pf0000374416 (2020)
16. Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry **22**(2014), 4349–4357 (2014)
17. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web. ACM (apr 2001). https://doi.org/10.1145/371920.372071
18. Sharma, A., Hofman, J.M., Watts, D.J.: Estimating the causal impact of recommendation systems from observational data. In: Proceedings of the Sixteenth ACM Conference on Economics and Computation. ACM (jun 2015). https://doi.org/10.1145/2764468.2764488
19. Shi, F., Shi, Y., Dokshin, F.A., Evans, J.A., Macy, M.W.: Millions of online book co-purchases reveal partisan differences in the consumption of science. Nat Hum Behav **1**(4) (apr 2017). https://doi.org/10.1038/s41562-017-0079
20. Shin, J., Valente, T.: Algorithms and health misinformation: A case study of vaccine books on amazon. Journal of Health Communication **25**(5), 394–401 (may 2020). https://doi.org/10.1080/10810730.2020.1776423
21. Smith, B., Linden, G.: Two decades of recommender systems at amazon.com. IEEE Internet Comput Internet Computing **21**(3), 12–18 (may 2017). https://doi.org/10.1109/mic.2017.72
22. Team, A.A.: Amazon publishes first eu store transparency report, outlining our commitment to providing a trustworthy shopping experience (Oct 2023)
23. Thorburn, L., Stray, J., Bengani, P.: From "filter bubbles", "echo chambers", and "rabbit holes" to "feedback loops" (Apr 2023)
24. Traag, V.A., Dooren, P.V., Nesterov, Y.: Narrow scope for resolution-limit-free community detection. Phys. Rev. E **84**(1) (jul 2011). https://doi.org/10.1103/physreve.84.016114
25. Traag, V.A., Waltman, L., van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Sci Rep **9**(1) (mar 2019). https://doi.org/10.1038/s41598-019-41695-z
26. Vromman, F.V., Fouss, F.: Filter bubbles created by collaborative filtering algorithms themselves, fact or fiction? an experimental comparison. In: IEEE/WIC/ACM International Conference on Web Intelligence. ACM (dec 2021). https://doi.org/10.1145/3498851.3498945
27. Zhang, Z., Fang, M., Chen, L., Rad, M.R.N.: Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.naacl-main.285