Browsing Amazon's Book Bubbles

Paul Bouchaud

Browsing Amazon's Book Bubbles

Paul Bouchaud
paul.bouchaud@iscpif.fr
Center for Social Analysis and Mathematics (EHESS)
Complex Systems Institute of Paris Île-de-France (CNRS)
Paris, France

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, "New World Order" narratives, 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.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative filtering**; *Empirical studies in HCI*.

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" [37], in this study, we explore the claim made by Smith and Linden when discussing two decades of recommender systems at Amazon. Early on, platforms sought to personalize online experiences, rearranging the shelves of digital stores to align with users' inferred interests. Still prevailing nowadays, recommender systems allocate the attention of millions of users, notably driving up to 80% of hours of content streamed on Netflix in 2015 [13]. To achieve insightful recommendations while accommodating a vast user and item base, [31] suggested evaluating similarity between pairs of items as the likelihood of individuals who purchase one item also buying another. Item-to-item collaborative filtering approach, proven to be highly effective, has been embraced by Amazon for product recommendations since the late 1990s [14, 23]. This strategy played a pivotal role, recommendations driving 30% of Amazon.com's page views in 2015 [34]. However, the widespread use of algorithmic curation has raised concerns about the potential risk of reinforcing exposure to like-minded content and amplifying existing biases [25]. While collaborative filtering algorithms, in isolation, do not inherently narrow the diversity of recommended content [42], their coupling with individual users' own preferences and biases can lead users into narrow content spaces, commonly referred to as "echo chambers", restricting access to a diverse range of content [12, 39]. These algorithmic distortions have been observed in online platforms such as YouTube [15, 30], Twitter [5, 17], Meta [10] or Google Search [29].

Even though e-commerce platforms wield substantial influence, with Amazon alone serving over 181 million users in the European Union [38], they have garnered limited research attention. In 2017, [35] highlighted partisan disparities in science consumption by analyzing book co-purchases on Amazon and Barnes & Noble. [12] identified feedback loops between users and Alibaba Taobao's recommender systems, reinforcing users' existing interests, leading to echo chambers. Interest in algorithmic curation on e-commerce platforms surged during the COVID-19 pandemic, spurred by media criticism of Amazon for promoting vaccine misinformation [4, 9, 28]. [36] discovered a two-to-one ratio of vaccine-hesitant books over supportive ones, and [20] revealed Amazon's tendency to rank misleading search results higher than debunking ones and found a filter bubble effect. Misinformation in books can wield significant influence as they are perceived as authoritative expressions of expertise, commonly consulted in researching specific topics [26]. Furthermore, their cost directly contributes to the monetization of disinformation, a challenge underscored in the EU Code of Practice on Disinformation [1].

Instead of focusing on COVID-19 and vaccine-related misinformation, this study examines Amazon's non-personalized recommendations for a wide range of non-fiction books. Our analysis reveals a recommendation landscape characterized by tight communities of semantically similar books, from which tens of successive clicks are required for a user, following Amazon recommendations, to navigate away. We discover a community within which Amazon co-recommends books supporting climate denialism, COVID-19 conspiracy theories, "New World Order" narratives, and conservative views on social and gender issues. In an effort to partially open Amazon's algorithmic black box, we performed a collaborative filtering analysis based on Amazon user reviews, revealing that books reviewed by the same users tend to be recommended together by Amazon. In addition to addressing a gap in the literature by investigating Amazon's recommender systems, this study emphasizes that even non-personalized recommender systems can pose system risks and have foreseeable negative effects on civic discourse and public health.

2 METHODS

2.1 Data Collection

When shopping on Amazon, numerous product recommendations are prominently featured. These recommendations are typically presented within multipage carousels under various denominations such as 'Customers who viewed this item also viewed', 'Customers who bought this item also bought', 'More articles to discover' or 'Inspired by your browsing history'. Amazon also selects up to two additional items, presented as "Frequently Bought Together" allowing the user to add them all to their cart in a single click.

In this study, we focus on non-personalized recommendations, gathering the suggestions an unlogged Amazon user, without any browsing history or cookies, would encounter upon visiting the website. To gather these recommendations, we leveraged an automated web browser, resetting itself after each page visit. In addition to the recommendations, we collected the metadata associated to each book, such as its title, description, author(s), publisher or Amazon category.

In order to create a comprehensive overview of the Amazon book recommendations landscape, we employed a snowballing methodology initiated with the selection of the top 150 bestsellers from within 18 non-fiction book categories, such as "News, Politics and Society", "Earth and Environmental Sciences", or "Business and Stock Market". From this initial pool of 1 725 books, we systematically collected recommendations offered by Amazon.fr. Subsequently, we treated these newly acquired books as seeds for the next iteration; this iterative procedure was repeated three times. To ensure the scalability of our data collection efforts, in between each iteration, books recommended only once were pruned. In total, we collected the Amazon recommendations associated with 60 298 books. Our coverage is such that, on average, for each book we retrieved, we also captured 85.8% of the books suggested by Amazon. For further details about the data collection, please refer to the Appendix.

The data collection spanned from October 28, 2023, to November 4, 2023. To assess the temporal stability of the recommendations, we compared this recent dataset with a prior collection of 31k books gathered between August 23, 2023, and September 5, 2023. Results showed a 64.4% overlap in recommendations between these periods. On average, 89.0% of previously suggested books were located within the same book community as the book in the recent dataset; communities detected later on.

The books within our dataset are primarily written in French 92.3%, a smaller portion in English 7.0%, and are distributed across various formats: 81.4% as printed books (paperback, pocket, or hardcover), 12.3% as Kindle ebooks, and 3.6% as audiobooks. Amazon consistently recommends books in a similar format, with an average of 97.9% of recommended books of the same format as the current book. In the subsequent analysis, our focus is exclusively on printed books; to avoid duplicate entries, redundant formats were removed.

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 $|E| = 429\,363$ edges.

2.3 Characterization

2.3.1 Community detection. To gain insight into the structure of G, we perform a community detection. The Leiden algorithm [41] was preferred over the conventional Louvain algorithm [24] as the latter can lead to arbitrarily badly connected communities. To overcome the resolution-limit inherent in modularity maximization [11], we adopt the Leiden algorithm while incorporating the Constant Potts Model [40] as the quality function. Establishing a resolution profile, we determine the appropriate resolution parameter, γ, which ensures the stability of our partitions. Within the Constant Potts Model, γ imposes an upper limit on inter-community link density.

2.3.2 Confinement. After identifying the communities, we quantify book recommendation homophily with respect to their respective communities. Specifically, we determine the fraction of a vertex's neighbors belonging to the same community as the vertex itself [18]. To assess the recommendations' confinement beyond their first-degree neighbors, we perform random walks on the graph. Starting from a given book, a surfer randomly clicks on a book suggested by Amazon, i.e. randomly selects a neighbor within G. This process is repeated until the random surfer transitions out of the community they initiated the walk from. Initiating 25 walks from every book $v \in V$, we compute the average length of the walks needed to leave each community.

2.3.3 Semantic Analysis. To explore the content of books within different communities, we analyse their title and summary. Despite recent advancements in neural natural language processing, we opted for a classical approach that combines TF-IDF (Term Frequency-Inverse Document Frequency) with Non-Negative Matrix Factorization (NMF). This choice was motivated by the simplicity, efficiency, and robustness of TF-IDF/NMF, as highlighted in [43]. NMF decomposes the term-document matrix generated through TF-IDF into two non-negative matrices: one representing terms and topics, and the other representing topics and documents. Such decomposition facilitates straightforward interpretation [22]. We compare the summary embedding of books either tied or not by a recommendation in G. Additionally, we will compare the semantic diversity of books within a given community to the overall corpus. We compute the semantic diversity as the geometric mean of the standard deviation of the summary embeddings [21]. We verified that the results remain consistent across a broad range of embedding dimensions.

2.4 Collaborative filtering

Historically, Amazon recommendations were formulated through an item-to-item collaborative filtering approach. Due to a lack of transparency, we do not know how the recommendations are formulated nowadays. Yet, to gain further insights into Amazon recommendations, we followed the work of Linden, Smith, and York [23] on collaborative filtering. This method aims to identify the most similar matches for a given item by inspecting items that customers frequently purchase together. However, we lacked access to purchase or ratings records, and relied solely on customer

reviews as our data source. We acknowledge that reviewing an article provides a stronger signal of (dis)agreement compared to simple ratings or purchases. This consideration should be kept in mind when analyzing the item-to-item similarities.

Then, in addition to book recommendations, we collected "verified purchase" user reviews for 25 151 books, sampled from the main communities of G. This dataset comprises 419 460 reviews contributed by 245 734 unique reviewers, resulting in an average of 20.5 reviews per book (median 7.0). For comparison, to train its recommendation engine, Amazon has access to an average of 748.7 ratings per book (median 52.0) in addition to purchase and navigation records (unavailable to the public).

To assess the relationship between pairs of books, we calculate the overlap coefficient between the two sets of users who have reviewed them. Finally, we compare the recommendations made by Amazon with the books we found to be most similar through review-based collaborative filtering.

2.5 Search Results

Following the assessment of community confinement in Amazon book recommendations, we explore how a user may enter a community through search results. To this end, we perform a handful of search queries related to Climate Change and COVID-19, topics underpinned by unequivocal scientific consensus. The search queries were chosen to maintain a broad and impartial scope, based on keywords extracted from books summary through TF-IDF. Yet, we do not claim to be extensive or, due to lack of publicly available information, representative of real users search queries on Amazon. Specifically, for Climate Change we performed the following queries (in French): "Climate Change CO2", "Global Warming", "IPCC", and for COVID-19 related queries: "COVID-19", "COVID pandemics" and "COVID vaccine". Starting on November 1st, 2023, we performed the search queries daily at noon for five consecutive days. For each query, we collected the first result page, which was sorted by either Amazon's default algorithm or decreasing average user ratings. Additionally, Amazon offers the possibility to rank results according to increasing/decreasing price or by date of publication; nevertheless, the fraction of books returned by those rankers, being in G is too low for any meaningful analysis.

3 RESULTS

3.1 Characterization

3.1.1 Community Detection & Confinement. The Leiden algorithm build a partition of G with a high modularity of $Q = 0.86$, identifying 61 communities encompassing more than 90% of the books. We display G on Figure 1, with the main communities colored-coded. Computing the homophily reveals that 88.8% of the recommended books belong to the same community as the current book. For items presented by Amazon as "Frequently Bought Together" this fraction increased to 94.9%. To make sense of these communities of recommended books, one can first observed that 53.2% of the book suggested by Amazon belong to the same book category as the currently visited book, also the probability that two randomly chosen books from a community belong to the same Amazon book category is 5.1 times higher than for two randomly selected books. On

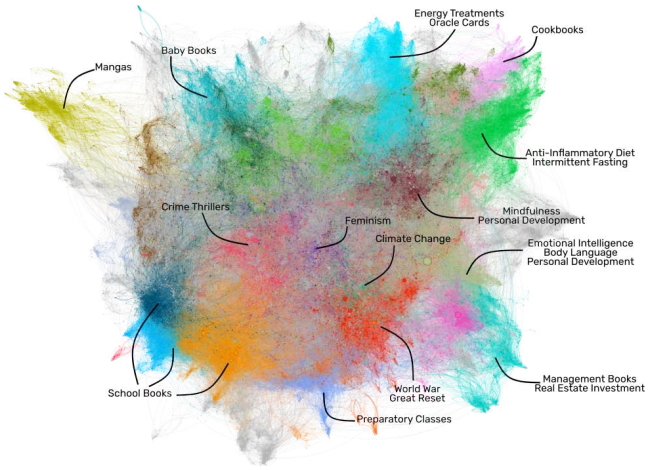

Figure 1: Graph of Amazon book recommendations G [48 636 books, 391 664 edges], spatialized via ForceAtlas2 [19]. Vertices are color-coded by community, and their size is proportional to their in-degree. Additionally, keywords associated with books within the main communities are displayed. These keywords were extracted from book summary through TF-IDF.

average, 91.1% of books authored by individuals who have written at least 5 books in G are found within the same book community.

Random walks based on Amazon's book recommendations exhibit strong confinement; walks initiated from the 61 largest communities, accounting for 90% of the books. After three successive click, 75.7% of random surfers are still in the same community as the one they started from; on average 24.9 (median 11) successive clicks are needed to leave a community. Notably, the confinement differs between communities. While 6.9 clicks are required on average (median 4) to leave the social science books community [356 books in G], one needs more than 77.8 successive clicks on average (median 68) to leave the colouring books community [380 books in G].

3.1.2 Semantic Analysis. We display on Figure 1, the keywords, extracted through TF-IDF, for some communities of G. One observes a wide array of topics such as cartomancy, personal development, mangas, crime thrillers or school books. The semantic diversity within a community is, on average, 57.9% poorer compared to the overall corpus. The average pairwise cosine similarity between summary embeddings of books connected by an edge in G is 1.72 times higher than the similarity between pairs of books from the same community but not connected by an edge, and 5.41 times higher than for pairs of books without an edge and from different communities. Computing the semantic similarity of books visited along random walks reveals a semantic confinement. The similarity between a book and those recommended by Amazon after three successive clicks is 36.9% higher than the average cosine similarity between pairs of books from the same community.

3.2 Collaborative filtering

The collected reviews are extremely sparse, among the 25.1k fetched books, each user reviewed an average of 1.7 books (median 1.0). Two books reviewed by the same user are 8.7 times more likely to belong to the same community in G than two randomly picked books, and 6.1 times more likely to belong to the same Amazon book category — computation restricted to users having reviewed at least 5 books in our pool. Similarly, the cosine similarity between summary embeddings of books reviewed by the same user is 2.6 times higher than for random pairs of books.

Considering the 10 108 books with at least 10 "verified purchase" reviews, we assessed the pairwise reviewer overlap. In 58.1% of cases, the book with the highest reviewer overlap with a seed book is affiliated with the same community in G, and, in 34.5% of cases, it is recommended by Amazon, i.e. linked by edge in G. The average overlap between reviewers is 15.6 times higher for pairs of books linked by an edge in G compared to random pairs of books from the same community. Likewise, the average overlap between reviewers is 9.2 times higher for random pairs of books from the same community in G compared to pairs of 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 [8], 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.

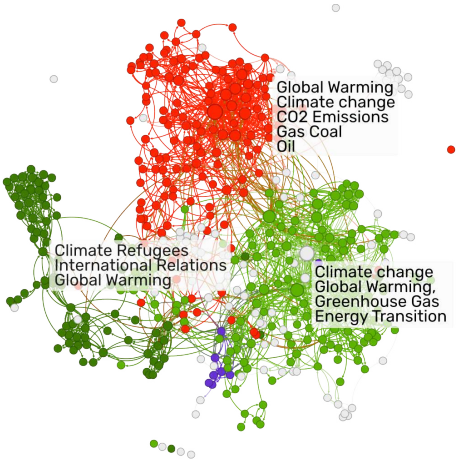

Figure 2: Two-hop recommendation graph induced by 146 books discussing Climate Change [459 books, 1 991 edges]. Vertices are color-coded by G community (smaller G communities are in white), and their size is proportional to their in-degree.

3.3.1 Climate Change. The 146 books addressing Climate Change primarily fall into two G communities: 42.5% in the community depicted in green on Figure 2 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.

Based on the taxonomy of climate-change contrarian claims by Coan et al. [7], a manual annotation reveals that climate-denialist books on French Amazon predominantly emphasize two narratives:

- "Climate science is unreliable" (43.2% of the book in the red community): asserting a lack of consensus, challenging the credibility of models, suggesting bias among scientists, labeling them as alarmists, or accusing them of participating in conspiracies. Some compare climate advocacy with a form of religion.
- "Climate-solution won't work" (20.5% of the book in the red community): arguing that climate policies are either harmful or ineffective, express doubts about the efficacy of clean energy (in particular wind turbines), assert that fossil fuels are abundant and affordable, or advocate for nuclear energy.

Other contrarian claims identified by [7] were found, including assertions that CO2 is beneficial for the planet, challenging the expected rate of ice melting, or attributing climate change to natural cycles. 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.

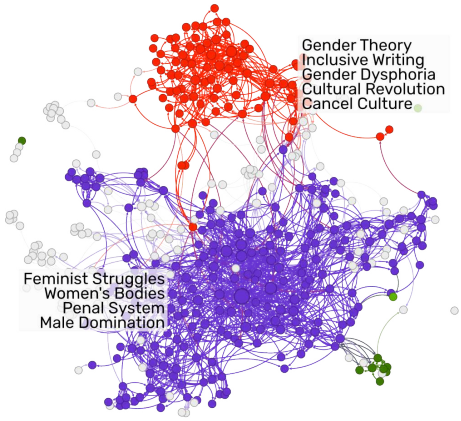

Figure 3: Two-hop Recommendation graph induced by 117 books relating to Gender Issues [439 books, 1 862 edges]. Vertices are color-coded by G community (smaller G communities in white), and their size is proportional to their in-degree.

3.3.2 Gender Issues. The graph of Amazon's book recommendations induced by two-hops from 117 books related to gender issues

is depicted in Figure 3. Again, two distinct 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 3 corresponds to the same community in G that embeds climate-denialist books shown in Figure 2.

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.

3.3.3 COVID-19. The analysis of 101 books within G addressing the COVID-19 pandemic reveals that 92.1% are situated within the community that otherwise encompasses climate-denialist books and advocates conservative views on gender issues, previously depicted in red. In Figure 4, 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.

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" [33] by Klaus Schwab —the founder of the World Economic Forum— and Thierry Malleret lies in the same community as those relaying conspiracy theories. This categorization is a result of Amazon's recommendation algorithm, which displays, in addition to two other Schwab's writings, five conspiracy theory books. Inspecting the set of reviews clarifies these recommendations, as, within the set of fetched books, the top 10 books with the highest overlap of reviewers with Klaus Schwab's book are relaying conspiracy theories.

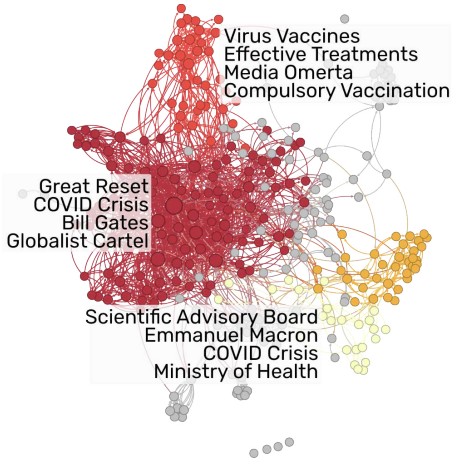

Figure 4: Two-hop Recommendation graph induced by 101 books discussing COVID-19 [304 books, 1 620 edges]. Vertices are color-coded by G community (smaller G communities in white), and their size is proportional to their in-degree.

The extraction of keywords from sub-community book summaries exposes distinct thematic focuses, aligning with established taxonomies of COVID-19-related disinformation [27]. 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.

3.3.4 Contrarian Community. To gain a deeper understanding of why various contrarian viewpoints coexist within the same recommendation community rather than being in distinct topic-specific communities, we leverage the set of users book reviews. Our analysis revealed 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 and 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).

To gain further insight into this contrarian community, the third-largest community in G with 1776 books, we isolated it, conducted a community detection analysis at a higher resolution than for G, and is displayed on Figure 5. 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.

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 [1229 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

In this study, we delved into Amazon's book recommendation system. Our analysis brought to light the existence of highly modular communities, characterized by substantial homophily —88.8% of recommended books belonged to the same community as the currently

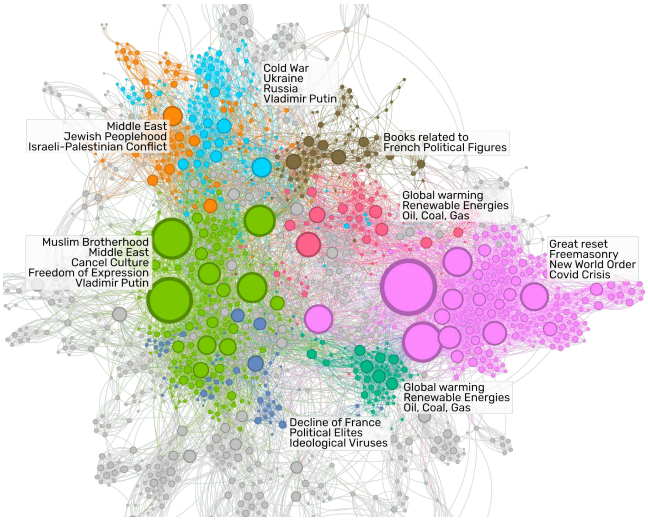

Figure 5: Sub-graph of G induced by books of the contrarian community [1776 books, 12 364 edges]. Vertices in the graph are color-coded by sub-communities, and their size is proportional to their in-degree. Additionally, keywords extracted through TF-IDF from books summary are displayed.

displayed book. 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. Users following Amazon recommendations find themselves deeply entrenched within a book community, often requiring tens of successive clicks to navigate away.

Exploring recommendation graphs induced by Climate Change, Gender Issues, and COVID-19 related books, we identified a community housing books promoting climate denialism, conspiracy theories on COVID-19 and on the "New World Order", conservative views on social and gender issues. In contrast, books arguing the opposite side of the discussions are in topic-specific communities. Once in this contrarian community, a user, randomly following Amazon's recommendations, needs 15.5 (median 8) successive clicks to navigate away. By performing broad scope search queries on Climate Change and COVID-19, we reveal that users can easily enter this community. Consistent with [16], a substantial portion of the results misguide the public on the underlying scientific consensus, regardless of ranking by Amazon's algorithmic ranker or average customer review.

In an attempt to partially unveil the inner workings of the Amazon algorithmic black box, we conducted a collaborative filtering analysis [23, 37] utilizing Amazon user reviews —by lack of access to ratings, purchase records, or page impression records. This analysis underscores that books reviewed by the same users tend to be recommended together by Amazon, shedding light on why diverse contrarian viewpoints across various topics coexist within the same recommendation community —because the reviewers do tend to overlap.

Despite its high coverage, we do not claim our data collection to be extensive or to provide a comprehensive overview of the entire Amazon book landscape. By performing the data collection through snowballing from socially-related bestsellers, we for example, ignore most fiction books. Additionally, the data collection was limited to the French Amazon, the third-largest market in the European Union with 34.6 million monthly active users [38]. However, one should note that widespread misinformation has also been observed in the United States [16] and Belgium [6]. Nonetheless, we argue that this article makes a valuable contribution to the existing literature by providing insights into a platform that, despite its substantial influence in the distribution of millions of books each year, has received relatively less attention compared to social media giants. Revealing that Amazon's non-personalized book recommendation systems tend to confine users within homogeneous communities. While refraining from taking a normative stance on cross-exposure [3, 32], we underscore the existence of a community harboring diverse contrarian viewpoints, spreading misinformation and conspiracy theories, misleading the public on the scientific consensus regarding Climate Change or COVID-19.

Leveraging on publicly available information, within the constraints of Amazon's opacity, the study highlights that even non-personalized algorithms, potentially relying on "objective" criteria like co-purchases or user reviews, can generate content recommendations with foreseeable negative effects on public health and civic discourse. These findings contribute to the broader discussion on algorithmic regulation, emphasizing that explainability and transparency, while crucial for accountability, do not inherently mitigate the systemic risks targeted by the regulations such as the Digital Services Act [2].

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.

REFERENCES

[1] 2022. The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
[2] 27/10/2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *OJ* L 277 (27/10/2022), 1–102.
[3] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* 115, 37 (aug 2018), 9216–9221. https://doi.org/10.1073/pnas.1804840115
[4] Julia Belluz. 2016. Amazon is a giant purveyor of medical quackery. https://www.vox.com/2016/9/6/12815250/amazon-health-products-bogus
[5] Paul Bouchaud, David Chavalarias, and Maziyar Panahi. 2023. Crowdsourced audit of Twitter's recommender systems. *Sci Rep* 13, 1 (oct 2023). https://doi.org/10.1038/s41598-023-43980-4
[6] AIFORENSICS & CheckFirst. 2023. Study of Amazon's Recommendation System.
[7] Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Sci Rep* 11, 1 (nov 2021). https://doi.org/10.1038/s41598-021-01714-4
[8] European Comission. [n. d.]. https://digital-strategy.ec.europa.eu/en/funding/second-call-european-narrative-observatory-fight-disinformation-post-covid-19
[9] Tom Dreisbach. 2020. On Amazon, dubious "antiviral" supplements proliferate amid pandemic. https://www.npr.org/2020/07/27/894825441/on-amazon-dubious-antiviral-supplements-proliferate-amid-pandemic
[10] Brendan Nyhan et al. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620, 7972 (jul 2023), 137–144. https://doi.org/10.1038/s41586-023-06297-w

[11] Santo Fortunato and Marc Barthélemy. 2007. Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1 (jan 2007), 36–41. https://doi.org/10.1073/pnas.0605965104

[12] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. 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. https://doi.org/10.1145/3397271.3401431

[13] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The Netflix Recommender System. *ACM Trans. Manage. Inf. Syst. Transactions on Management Information Systems* 6, 4 (dec 2015), 1–19. https://doi.org/10.1145/2843948

[14] Larry Hardesty. 2022. The history of Amazon's recommendation algorithm. https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

[15] Muhammad Haroon, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq, and Magdalena Wojcieszak. 2022. YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations. arXiv:2203.10666 [cs.CY]

[16] Eslam Hussein and Hoda Eldardiry. 2020. Investigating Misinformation in Online Marketplaces: An Audit Study on Amazon. arXiv:2009.12468 [cs.IR]

[17] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2021. Algorithmic amplification of politics on Twitter. *Proc. Natl. Acad. Sci. U.S.A.* 119, 1 (dec 2021). https://doi.org/10.1073/pnas.2025334119

[18] Ruben Interian and Celso C. Ribeiro. 2018. An empirical investigation of network polarization. *Appl. Math. Comput.* 339 (dec 2018), 651–662. https://doi.org/10.1016/j.amc.2018.07.066

[19] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 9, 6 (jun 2014), e98679. https://doi.org/10.1371/journal.pone.0098679

[20] Prerna Juneja and Tanushree Mitra. 2021. Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3411764.3445250

[21] Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 1739–1746. https://aclanthology.org/2020.lrec-1.215

[22] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (oct 1999), 788–791. https://doi.org/10.1038/44565

[23] G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7, 1 (2003), 76–80. https://doi.org/10.1109/MIC.2003.1167344

[24] M. E. J. Newman and M. Girvan. 2004. Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 2 (feb 2004). https://doi.org/10.1103/physreve.69.026113

[25] Eli Pariser. 2012. *The filter bubble: What the internet is hiding from you.* Penguin Books.

[26] Andrew Perrin. 2016. Book reading 2016. https://www.pewresearch.org/internet/2016/09/01/book-reading-2016/

[27] J Posetti and B Kalina. 2020. Disinfodemic: deciphering COVID-19 disinformation. https://unesdoc.unesco.org/ark:/48223/pf0000374416.

[28] Matt Reynolds. 2019. Amazon sells "autism cure" books that suggest children drink toxic, bleach-like substances. https://www.wired.co.uk/article/amazon-autism-fake-cure-books

[29] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact. on Human-Computer Interaction* 2, CSCW (nov 2018), 1–22. https://doi.org/10.1145/3274417

[30] Leonardo Sanna, Salvatore Romano, Giulia Corona, and Claudio Agosti. 2021. YTTREX: Crowdsourced Analysis of YouTube's Recommender System During COVID-19 Pandemic. In *Information Management and Big Data*. Springer International Publishing, 107–121. https://doi.org/10.1007/978-3-030-76228-5_8

[31] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM. https://doi.org/10.1145/371920.372071

[32] Frank M. Schneider and Carina Weinmann. 2021. In Need of the Devil's Advocate? The Impact of Cross-Cutting Exposure on Political Discussion. *Polit Behav* 45, 1 (may 2021), 373–394. https://doi.org/10.1007/s11109-021-09706-w

[33] Klaus Schwab and Thierry Malleret. 2020. *Covid-19: The great reset.* World Economic Forum.

[34] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. ACM. https://doi.org/10.1145/2764468.2764488

[35] Feng Shi, Yongren Shi, Fedor A. Dokshin, James A. Evans, and Michael W. Macy. 2017. 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

[36] Jieun Shin and Thomas Valente. 2020. Algorithms and Health Misinformation: A Case Study of Vaccine Books on Amazon. *Journal of Health Communication* 25, 5 (may 2020), 394–401. https://doi.org/10.1080/10810730.2020.1776423

[37] Brent Smith and Greg Linden. 2017. Two Decades of Recommender Systems at Amazon.com. *IEEE Internet Comput Internet Computing* 21, 3 (may 2017), 12–18. https://doi.org/10.1109/mic.2017.72

[38] About Amazon Team. 2023. Amazon publishes first EU Store Transparency Report, outlining our commitment to providing a trustworthy shopping experience. https://www.aboutamazon.eu/news/policy/amazon-publishes-first-eu-store-transparency-report-outlining-our-commitment-to-providing-a-trustworthy-shopping-experience

[39] Luke Thorburn, Jonathan Stray, and Priyanjana Bengani. 2023. From "filter bubbles", "Echo chambers", and "rabbit holes" to "feedback loops". https://techpolicy.press/from-filter-bubbles-echo-chambers-and-rabbit-holes-to-feedback-loops/

[40] V. A. Traag, P. Van Dooren, and Y. Nesterov. 2011. Narrow scope for resolution-limit-free community detection. *Phys. Rev. E* 84, 1 (jul 2011). https://doi.org/10.1103/physreve.84.016114

[41] V. A. Traag, L. Waltman, and N. J. van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9, 1 (mar 2019). https://doi.org/10.1038/s41598-019-41695-z

[42] Flore Vancompernolle Vromman and François Fouss. 2021. Filter bubbles created by collaborative filtering algorithms themselves, fact or fiction? An experimental comparison. In *IEEE/WIC/ACM International Conference on Web Intelligence*. ACM. https://doi.org/10.1145/3498851.3498945

[43] Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. 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. https://doi.org/10.18653/v1/2022.naacl-main.285

A DATA COLLECTION

The data collection was performed through snowballing starting from 1 725 unique seeds collected from the top hundred bestsellers of 18 Amazon book categories: News, Politics and Society; Social issues in society; Politics; Current affairs; Economics; Major conflicts and geopolitics; History and current affairs; Media and communication; Business and the Stock Market; Humanities; Science; Science for all; Earth, water and environmental sciences; Environment; Ecology; Life sciences, biology and genetics; Health, Fitness and Diet; Religions and Spiritualities.

From the 1 725 initial seeds, the first iteration identified 14 847 unique books, 7 485 of them recommended at least twice. The second iteration identified 41 463 books, 23 663 of them recommended at least twice and the third iteration identified 94 139 books, 57 781 of them recommended at least twice. Removing duplicates, we finally collected 60 298 books.

Amazon introduces product recommendations under multiple designations. For the examination of non-personalized recommendations, we exclusively retained the following: 'Customers who viewed this item also viewed', 'Related to items you viewed', 'What other items are customers buying after viewing this item?', 'Customers who bought this item also bought', 'More articles to discover', 'Customers who read this book also read', 'People who viewed this content also viewed', 'Popular products based on this article', 'Products related to this article'

Received 26 November 2023