Measuring Commonality in Recommendation of Cultural Content to Strengthen Cultural Citizenship

ANDRES FERRARO and GUSTAVO FERREIRA, McGill University, Montréal, Canada
FERNANDO DIAZ, Carnegie Mellon University, Pittsburgh, United States of America
GEORGINA BORN, University College London, London, United Kingdom

Recommender systems have become the dominant means of curating cultural content, significantly influencing the nature of individual cultural experience. While the majority of academic and industrial research on recommender systems optimizes for personalized user experience, this paradigm does not capture the ways that recommender systems impact cultural experience in the aggregate, across populations of users. Although existing novelty, diversity, and fairness studies probe how recommender systems relate to the broader social role of cultural content, they do not adequately center culture as a core concept and challenge. In this work, we introduce commonality as a new measure of recommender systems that reflects the degree to which recommendations familiarize a given user population with specified categories of cultural content. Our proposed commonality metric responds to a set of arguments developed through an interdisciplinary dialogue between researchers in computer science and the social sciences and humanities. With reference to principles underpinning public service media (PSM) systems in democratic societies, we identify universality of address and content diversity in the service of strengthening cultural citizenship as particularly relevant goals for recommender systems delivering cultural content. We develop commonality as a measure of recommender system alignment with the promotion of a shared cultural experience of, and exposure to, diverse cultural content across a population of users. Moreover, we advocate for the involvement of human editors accountable to a larger value community as a fundamental part of defining categories in the service of cultural citizenship. We empirically compare the performance of recommendation algorithms using commonality with existing utility, diversity, novelty, and fairness metrics using three different domains. Our results demonstrate that commonality captures a property of system behavior complementary to existing metrics and suggests the need for alternative, non-personalized interventions in recommender systems oriented to strengthening cultural citizenship across populations of users. Moreover, commonality demonstrates both consistent results under different editorial policies and robustness to missing labels and users. Alongside existing fairness and diversity metrics, commonality contributes to a growing body of scholarship developing "public good" rationales for digital media and machine learning systems.

A. Ferraro currently affiliated to Pandora-SiriusXM.
F. Diaz work done while at Google.
This research was funded in part by the European Research Council Advanced Grant "Music and Artificial Intelligence: Building Critical Interdisciplinary Studies" or MusAI (2021–26), PI Prof. Georgina Born, grant agreement no. 101019164 based at University College London, UK; and in part by the Canada CIFAR AI Chairs program.
Authors' addresses: A. Ferraro and G. Ferreira, McGill University, Montréal, Canada; e-mails: andresferraro@acm.org, gustavo.ferreira@mila.quebec; F. Diaz, Carnegie Mellon University, Pittsburgh, USA; e-mail: diazf@acm.org; G. Born, University College London, London, UK; e-mail: g.born@ucl.ac.uk.

ACM 2770-6699/2024/03-ART10
https://doi.org/10.1145/3643138

CCS Concepts: • **Information systems;** • **Computing methodologies** → *Artificial intelligence*;

Additional Key Words and Phrases: Recommender systems, evaluation, cultural content, movies, music, literature, cultural citizenship, diversity

ACM Reference Format:
Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2024. Measuring Commonality in Recommendation of Cultural Content to Strengthen Cultural Citizenship. *ACM Trans. Recomm. Syst.* 2, 1, Article 10 (March 2024), 32 pages. https://doi.org/10.1145/3643138

1 INTRODUCTION AND BACKGROUND

Online platforms that host cultural content such as music, movies, and literature use recommender systems to suggest and distribute items from their catalogs employing the principle of personalization. Generally, we measure the degree to which a recommender system succeeds in personalization by adopting various offline metrics (e.g., precision, NDCG) and online metrics (e.g., clickthrough rate, consumption) [27]. Evaluation using these metrics is appealing in commercial settings, because they are aligned with revenue-generating metrics such as retention and subscriptions. As a result, personalization remains a central principle of academic and industrial research on recommender systems.

However, increasing evidence suggests that, while the degree of personalization is one desirable property of a recommender system, it does not capture the wider effects of recommender systems in aggregate, nor does it measure the effects of recommender systems across a population of users. This is important, because personalized recommendations are likely to have cumulative effects, shaping the wider cultures and societies within which they are being used [2].

We advocate that the design of recommender systems delivering cultural content broaden its foundation to include not just personalization and associated commercial interests but also appropriate normative principles oriented to furthering the democratic well-being and the cultural and social development of contemporary societies. By "normative," we refer to principles considered to provide models of morally, ethically, and/or politically right or just action or behavior at the level of societies and communities as well as individuals. And, just as domains such as criminal justice or lending have associated normative values related to justice and fairness, the distribution of cultural content does as well. For guidance on normative principles appropriate to the provision of cultural content, we turn to the principles underpinning **public service media (PSM)** systems [3]. Public service media refers to the longstanding existence of various channels of content distribution and related media organizations that are designed to be accountable to the public and may also be publicly funded [50]. Examples include the British Broadcasting Corporation, the Canadian Broadcasting Corporation, and the Australian Broadcasting Corporation.[1]

From the set of normative principles guiding PSM, in this article, we identify universality of address and content diversity in the service of strengthening cultural citizenship as particularly relevant normative principles for recommender systems delivering cultural content. If personalization attempts to maximize individual user satisfaction with a platform, then the promotion of cultural citizenship entails disseminating a diversity or plurality of cultural content to stimulate intercultural and intracultural dialogue and common exposure to cultural diversity. As a result,

[1]PSM organizations are found throughout Europe and in many member states of the Commonwealth. Several PSM organizations may be found within a single country and thereby constitute a PSM ecology or system, as is the case, for example, in the UK, Germany, and Australia. Transnational PSM institutions also exist, such as the European Broadcasting Union, an alliance composed of PSM organizations from countries that lie within the European Broadcasting Area or are members of the Council of Europe.

the distribution of cultural content can enhance both social integration and pluralistic cultural experience across communities. In making these arguments, we contribute to a growing body of scholarship developing public good rationales for digital media and machine learning systems [3, 15, 16, 76, 78, 106, 110]. Later, we expand upon and further justify these arguments.

Recognizing the role played by evaluation metrics in embodying values such as personalization or fairness, we derive a new evaluation metric based on the principle of commonality. In the current work, our metric measures the degree to which a recommender system familiarizes a given population of users with specific under-represented categories of cultural content in a certain medium. These categories, identified by human editors answerable to a knowledgeable community, work in concert with the metric to promote cultural citizenship. More concretely, our commonality metric provides editors with an instrument to counteract undesirable biases associated with racism, sexism, and the neglect of non-Western content in the cultural content being recommended, and to deliver this more diverse experience commonly across a population of users. However, in principle, the commonality metric could also be applied to achieve other normatively desired kinds of content exposure—for example, to impartial news.

To better understand our metric, we include a series of quantitative analyses of its behavior. Using data from three media—movies, music, and literature—we compare commonality with existing utility, diversity, and fairness metrics. Our results demonstrate that our new metric is not correlated with existing metrics (i.e., it captures the properties our conceptual propositions require while other metrics do not) while maintaining comparable robustness.

To date, criticisms of recommender systems and machine learning systems for their capacity to reproduce forms of bias and discrimination have been based on the evaluation of such biases at the level of individual users. Relatively little attention has been paid to identifying means of both counteracting biases and enhancing diversity in recommended cultural content by evaluating their performance in the promotion of common experiences across a population of users. As we show in this work, recommender systems can be developed explicitly to promote a value such as diversity by counteracting racist and sexist biases and the neglect of non-Western content—and they can advance these progressive changes as common experiences, thus enhancing cultural citizenship. In this way recommender design, and evaluation in particular, can support the wider cultural changes called for by those critical of the lack of diversity and biases evident in recommender systems , as well as by those sympathetic to these criticisms from the recommender system community [7, 36, 37, 71, 79, 115].

Our research process consisted of sustained interdisciplinary dialogues between two computer science researchers designing music recommender systems (Diaz, Ferraro) and two media and communication scholars with expertise in PSM from the humanities and social sciences (Born, Ferreira). Over the course of a year, we instructed each other in appropriate background research, sharing ideas and deepening our mutual engagement in both directions. In this way, we translated terms from one "side" to the other, while also responding to critical questioning about the relevance of key concepts, and subsequently adapting the latter. Such translation across disciplinary domains is difficult and may be incomplete. Nonetheless, our experience is that it can produce hybrid thinking that can in turn generate powerful new concepts and tools. Indeed, systematic interdisciplinary practices of this kind can move beyond the tendency for one domain to provide merely a service to the other [5] and instead makes possible reflexive critical thought on both "sides" that builds towards new, higher-level syntheses.

From a methodological perspective, evaluation based on normative principles differs from evaluation based on individual-focused metrics that model user satisfaction (e.g., NDCG, AP). While the traditional utility metrics can be validated with, for example, user studies, surveys, or downstream metrics (e.g., retention), metrics based on normative principles such as ours need to be

validated using more rigorous conceptual grounding and theoretical translation, since there is no quantitative ground truth to validate against.

For this reason, this article has to combine conceptual perspectives with formal and empirical research on evaluating recommender systems for cultural content. Our general question is how to measure the cultural effects of recommender systems across a whole user population. We begin with a rigorous conceptual grounding, in Section 2 framing four propositions for recommender systems, and in Sections 3 and 4, advancing the relevant principles for translation. Subsequently, in Section 5, we define commonality and report on the results of our experiments related to our three research questions. Finally, in Section 6, we reflect on our results, discussing the potential of commonality to address cultural citizenship, and conclude with questions for future work on commonality and other normative principles.

2 FOUR PROPOSITIONS FOR RECOMMENDER SYSTEMS

The interdisciplinary research dialogue described above resulted in four related propositions at the core of our research, as follows:

First, we propose that it is timely for the design of recommender systems delivering cultural content to move beyond a commercial orientation focused primarily on individualized interests.[2] We suggest that recommender design should, in addition, pursue complementary design paradigms guided by normative principles intended to promote the democratic development of contemporary cultures and societies, as this enhances human flourishing. In this way, we link our work to "a computational politics wedded to emancipation and human flourishing" [100].

Second, we propose that as well as a focus on personalization, recommender system design should acknowledge the aggregate and cumulative influences of recommender systems we have described, which have the potential to mediate wider cultural and social changes, and in this light develop ways of analyzing and modifying these influences in progressive ways that seek to achieve the goals described in the previous paragraph. In this respect, our work participates in the "values in design" debate [42, 43, 62], which addresses the challenges of reflexively "incorporating human values adequately into formal models" [8]. "Values in design" recognizes that the development of formal models for machine learning systems tend to fall back on "internalist" tendencies, in that "only considerations that are legible within the language of algorithms," for example, accuracy and efficiency, "are recognized as important design and evaluation considerations" [6]. The result is that design responses "to questions concerning human values such as fairness [become] problematic," because "problems with quantification [affect] everything downstream" [8, 70]. It is to suggest a new approach to identifying values for recommender design that we turn below to studies of the principles guiding public service media systems. As Benjamin Fish and Luke Stark [8] comment, "expanding formal models to include social values," what Ben Green and Salomé Viljoen [6] call "formalist incorporation," may be "situationally and strategically useful," even if it is imperative to be aware of how "such solutions are insufficient as full remedies to the inherent limitations of formal modeling" [8].

Third, as a concrete means of addressing these two propositions, we turn to evaluation metrics as a place to incorporate alternative "values in design" into the development of recommender systems. Specifically, we have developed a metric named "commonality" that measures the degree to which recommendations familiarize a given user population with specified categories of content chosen

[2] Although concerns about personalization often center on filter bubbles that keep individuals in taste echo chambers, our proposition in this article goes further. We are concerned not only with the social effects of individualization but with the need for greater diversity in the curation and distribution of cultural content and the benefits of promoting a shared diversity of cultural experience.

to promote certain "values in design" [39]. In addition to translating normative principles from the social sciences into mathematical representations, we conducted a series of experiments to assess the novelty and usefulness of commonality, in the context of other, related evaluation metrics.

Fourth, we draw on research that identifies the normative principles underlying **public service media (PSM)** systems, principles that then can be translated into a quantitative metric. In turning to previous research on principles embodied in PSM systems, we note both the powerful insights that can be derived from this research and its limitations with respect to the challenge of translating earlier normative concepts into contemporary digital platforms. Hence, although there is a literature concerned with how PSM organizations and older private media organizations are adapting to platformization and personalization (e.g., References [10, 23, 51, 102, 109, 110]), as well as papers by PSM-based researchers on these topics and specifically on recommender design (e.g., References [9, 41]), attempts to adapt PSM's earlier normative principles to the platform present are less advanced. We stress that, in this study, we are not concerned primarily or exclusively with PSM organizations in themselves nor with their approaches to recommender systems (see, e.g., Reference [58]). Rather, we take writings on the normative foundations of PSM as a source of potentially relevant concepts and then attempt to translate these relevant principles into the design of recommender systems. Our work does not aim to be an intervention in PSM but to have general implications for recommender design.

3 THE CASE FOR PRINCIPLES TO INFORM THE DESIGN OF RECOMMENDER SYSTEMS

Recommender systems have become the dominant means of curating cultural content in the digital era. Curation—or the selection, organization, and promotion of content to be made available to consumers—has, however, a deep history. For centuries, consumers have encountered cultural content through intermediaries—publishers, gallerists, patrons, impresarios—who collected works of culture, music, and art, organized and categorized those works, and made them available to audiences and consumers. From the 2000s, the term curation began to be used to refer to the collection and organization of content on the internet. Indeed, the present has been depicted as an era of "curationism"—an "acceleration of the curatorial impulse to become a dominant way of thinking and being... [in] an attempt to make affiliations with, and to court, various audiences and consumers" [4]. In general, "the introduction of new technologies has both introduced new methods of curation and expanded the breadth of individuals deemed fit to be curators" [1]. Yet, curation is not just an individual activity: It forges interconnections between curators, artists, audiences, and the industries, institutions, and online platforms supporting cultural production [81]. Today, the normative question of how curation by online platforms should be organized, and the forms it should take, is a pressing one.

When the curation enacted by online platforms' recommender systems is multiplied across the billions of recommendations presented to users, it significantly influences the nature of individual cultural experiences [105]. Yet, in marked contrast with earlier eras of curation, this influence is multiplied and magnified cumulatively not only across time but across populations, cultures, and regions. In the short term, recommender systems clearly influence individual cultural consumption and taste. In the medium and long term, by employing data on consumer behavior and repeatedly influencing consumer choices, recommender systems can shape cultural literacies as well as population-wide trends in cultural consumption and cultural taste [17]. They also participate in the commodification of the data generated by consumers as they engage with online platforms [44]. Moreover, the collection and mining of consumer data implemented by recommender systems, a type of "monitoring-based marketing" [3], takes place at a much larger scale and is more rapid, recursive, and intensive in comparison with earlier, non-computational methods

of applying market research to identify and shape what consumers might want. In some ways, this may be a productive kind of power and control exercised over consumers; yet, as a critic notes, "users have little choice over whether this data is generated and little say in how it is used," and "in this sense we might describe the generation and use of this data as... [an] alienated or estranged dimension of [users'] activity" [3].

Recommender systems therefore implement a higher degree of automatized intervention than previous forms of curation in the way not only individuals but societies and communities encounter cultural content. And despite their intended personalized address, recommender systems have cumulative effects in shaping the wider cultures and societies within which they are being employed.

Yet, perhaps because academic and industrial research on recommender systems has converged on personalization as a paradigm, these cumulative effects have been relatively unexplored by research in the recommender systems community. Some metrics linked to personalization and "user relevance" (e.g., NDCG, precision, clickthrough rate) align with user retention and other longer-term individual-level metrics that, when aggregated, can correlate with business metrics like revenue [103]. Even metrics like diversity, often motivated by broadening the range of categories to which users are exposed, regularly need to be justified by individual-level metrics like retention [2]. And when efforts have been made to design recommender systems for the "social good," the focus has been on "understanding the unique personal preferences" of users [61]. Thus, despite being concerned with and sensitive to the broader social role of cultural content, these previous metrics focus on individual exposure of content providers or effectiveness for users and do not capture the wider, aggregate shaping effects of recommender systems on patterns of cultural consumption, taste, and literacy as described above.

3.1 Applying Normative Principles from Public Service Media to Recommendation: Universality, Diversity, and Cultural Citizenship

In employing the principles underpinning public service media systems in democratic societies, we take the view that "a public service rationale is as pertinent as ever in the digital era" [3]. The normative ideas underpinning public service media developed over the last century in the context of governments seeking to strengthen their societies' democratic and representative channels of communication as well as means of reliable public information provision and cultural exchange [11, 50, 93, 96]. Although these normative principles originated in national democratic polities, they have also been applied transnationally, for example, in the European Union through the auspices of the European Broadcasting Union. They are therefore not limited in their purview to national media systems [32, 33, 54, 55]. A common misconception is to equate PSM systems with state-controlled media; however, as media institutions and ecologies created for the purposes mentioned, and bolstered by regulatory frameworks, they are designed to be independent of the state, to exhibit a certain autonomy and political impartiality, and to be publicly accountable [11]—although in reality this status may be fragile or imperfectly achieved.

A substantial body of research in media and political theory has identified the normative principles informing PSM systems and how these systems function as a communicative infrastructure for democratic societies. Central among those principles are universality (or commonality), diversity, and citizenship [12, 15, 16, 18, 85, 92]. We consider this triad particularly relevant for recommender systems delivering cultural content, since together they answer calls in democratic media theory and political theory for digital media systems to enhance cultural citizenship [12]. The concept of cultural citizenship has become foundational for democratic political theories in the past two decades; indeed, "one of the striking developments in recent political discourse has been the increasing confluence of culture and citizenship" [29]. Cultural citizenship has been defined

as a fourth stage of citizenship that responds to recognition of the social transformations and challenges posed by globalization, increased migration, the growing heterogeneity of the populations of nation states, and the intensification of identity politics among subaltern and marginalized groups [75, 88].[3] Given these profound changes, cultural citizenship draws attention to a "new domain of cultural rights [involving] the right to symbolic presence, dignifying representation" and "the maintenance and propagation of distinct cultural identities" [83]. Hence, for theorists of cultural citizenship, "cultural pluralism is viewed as something which enriches rather than threatens the fabric of society" [29]. In this light, it becomes clear that, to promote cultural citizenship, PSM organizations and other democratic channels of cultural production and distribution have a responsibility to curate and disseminate a plurality of cultural content with the intention of stimulating both intercultural and intracultural dialogue, as well as the acceptance of, and respect for, cultural diversity [12, 14]. In this way, PSM and other democratic media can act both as a force "for social cohesion and integration" and as a forum for pluralistic cultural experience and exchange among those many groups and communities that coexist and interact in democratic societies [54].

Music, movies, and literature, as expressive media, add further dimensions to these ideas. Some political theorists argue that the dialogical mechanisms required by democratic pluralism should not be confined to the classic concerns of public sphere theory—information, reason, and cognition—but should also engage matters of identity and affective experience. Hence, the political philosopher Martha Nussbaum draws attention to emotion as a basic component of ethical reasoning, arguing that a compassionate citizenry depends on access to pluralistic cultural repertoires that engage audiences' emotions and thereby enhance their capacity for mutual recognition, empathy, and toleration. For Nussbaum, such processes are essential for the well-being and the development of democratic societies [80]. Arguably, then, cultural citizenship is the principal form for the exercise of citizenship in the multicultural societies characterizing the contemporary world. If we take seriously the role of mediated cultural content, such as that curated by recommender systems, in influencing users' tastes and thereby conditioning the wider public culture, then, by analogy with the concern in democratic theory with the formation of an educated and informed citizenry, we might add a concern with the formation of a *culturally* mature and aware, culturally pluralistic citizenry [18, 84]. In this sense, digital platforms distributing cultural content—such as music, movies, and literature—can be understood as primary "theaters" for contemporary pluralism and consequently bear an obligation to provide a diversity of cultural experience. As Stuart Hall, the leading critical race theorist, noted, "The quality of life for black or ethnic minorities depends on the whole society knowing more about the 'black experience"' [46], an experience that can be grasped most compellingly through access to the diverse riches of black cultural production—whether music, movies, or literature. Platforms curating cultural content therefore have the capacity, and arguably the responsibility, to play a vital role in fostering cultural citizenship—itself a precondition for the processes of ethical, social, and cultural development that underlie the general condition of citizenship [12].

Both universality or commonality—that is, the provision of common cultural experiences—and diversity of cultural experience, or exposure to diverse cultural content, are therefore essential to the strengthening of cultural citizenship. As Georgina Born argues, both "mutual cultural recognition and the expansion of cultural referents... are dynamics essential to the well-being of pluralist societies. But this does not obviate the need also for integration—for the provision of common [cultural] experience and the fostering of common identities" [12]. Scholarship on these matters emphasizes, further, that implementing principles such as universality (commonality), diversity, and

[3]In Thomas Marshall's classic sociological account of the historical emergence of citizenship [72], he divides it into three stages or "elements"—civil, political, and social [38]. Theorists of cultural citizenship conceive of it as a fourth stage.

citizenship requires "alternative success metrics... focused on [media systems'] impact on democracy" and which address users "as citizens and not just... as consumers" [106]. Such metrics will enable democratically oriented media and platforms to better fulfill the present need to advance "cultural citizenship and the needs of the digital society" [54]. Recommender system design intended to strengthen cultural citizenship therefore requires us to implement universality—via a commonality metric—and to deliver diversity of cultural experience, a challenge to which we turn now.

3.2 Human Editing, Value Communities, and Recommender Systems as Sociotechnical Assemblages

Given the importance of pluralistic cultural experience in strengthening cultural citizenship, a core challenge for this research is the need to boost the diversity of cultural content to which a population of users is exposed by recommendation. Unlike existing ideas of diversity employed in the recommender system literature, we consider that diversity for cultural items such as movies, music, and literature can be conceptualized in a range of ways. They include, first, diversity of content in terms of artistic and cultural expression, which can be equated with the need to ensure that a range of genres are present as well as intra-generic differences, generic margins, and niches; and, second, diversity of the source or producer of the content, according to region, territory, industry source, or culture of origin as well as under-represented demographics among producers of the content (musicians, filmmakers, writers). The two—diversity of content and of source—are potentially related, in that greater diversity of source or producer is likely to favor, although it does not guarantee, greater diversity of content. However, judgments about what kinds of content and source diversity are desirable are intrinsically context-dependent and culturally dependent. In this sense, they necessitate human editorial processes that draw on knowledgeable and communally validated categorizations—both of the subtleties of demographic categories and of the complex contours of cultural genres.

A key assumption in our work is therefore that human editors must be involved in these judgments, and that their role is to reflect on the diversity of a recommender with respect to a given category or categories by drawing on insights generated by a larger "value community" knowledgeable about relevant cultural expressions and their social conditions [13]. By value community, we refer to the existence of communities sharing cultural interests and tastes, among them genre communities, who broadly embody an evolving consensus about the cultural interests or genres they enjoy and their relationship to categories of social identity and about which members have varying degrees of expertise. The consensual judgments of value emerging from a value community are, then, relational, and, as Pierre Bourdieu suggests, they will inevitably encompass a lively and shifting dissensus within the consensus [19]. The human editors we envisage therefore act as conduits for these larger communities of interest and judgment, and their judgments are legitimized and validated by this relationship.[4]

The aim is to achieve a diverse mix of content and sources that appeals beyond personalization and that avoids the risks of employing reified models of both identities and genres. Editors' judgments, moreover, will necessarily evolve over time and will be repeatedly replenished by evaluating (via a commonality metric) the performance of the recommendation of the diverse categories selected across a user population. It is the resulting universal promotion of a plurality of cultural experiences, relative to a given social context and cultural situation, that is likely to cumulatively enhance cultural citizenship; over time, it may also foster progressive cultural and social change.

[4]Our framework for involving value communities in the judgments of value and category validation that inform the work of human editors has resonances with the "participatory turn in AI design" [31]. However, our aim is not only that of "empowering stakeholders," although that is certainly one aim, but of honing and legitimizing the editorial processes underlying the curation of diversity in relation to the cultural content being recommended.

A related conceptual step is necessitated by the key role we are proposing for editors responsive to wider value communities. It is to expand how we think about "recommender systems" to include human editors, the value communities validating their knowledge, and user populations (or audiences). In this light, we propose that recommender systems can be productively conceived as sociotechnical assemblages that include the social knowledge and social labor that go into the processes described. As Nick Seaver puts it, "algorithms are not autonomous technical objects, but complex sociotechnical systems," and "while discourses about algorithms sometimes describe them as "unsupervised," working without a human in the loop, in practice there are no unsupervised algorithms. If you cannot see a human in the loop, you just need to look for a bigger loop" [97]. Designating recommender (or algorithmic) systems as sociotechnical assemblages implies, then, that these "technologies are embedded in the social context that produces them" [90].

4 CONCEPTUALIZING DIVERSITY IN RELATION TO CULTURAL CITIZENSHIP AND IMPLEMENTING IT IN RECOMMENDER SYSTEMS

If diversity of common cultural experience is a precondition for enhancing cultural citizenship, then the question is how diversity should be conceptualized in relation to recommender systems to achieve this end. As discussed above, diversity of content and diversity of source are perhaps the most obvious vectors of diversity. But it is certainly possible to imagine additional forms of diversity in relation to recommender systems focused, for example, on diversity of consumption experience and of user controls [68, 114]. This might include the potential to design diversity into the navigational architecture of recommendation by avoiding "similarity" and promoting difference; or by offering controls to users that endow the algorithm with greater legibility and increase users' agency to pursue diverse pathways through a given recommendation space.

Although these and other approaches to diversity might be productive for design, the recommender system community has mainly addressed diversity in terms of promoting diversity in some abstract space (e.g., a vector space) or through fairness measures, approaches that equate in some ways to what what we have called diversity of content and/or of source. Yet, this existing work, even if it is concerned with and sensitive to the broader social role of cultural content, does not adequately support the rich set of goals system designers might have and the values they might want to implement through design. Typically, diversity metrics are limited to the goal of capturing the variety of content *within* a recommendation list; they may consider categorizations of the content, distances in a latent space, or simply how many different items are recommended [2, 47, 57, 95]. Aligned with the goals of personalization, the formulation of these diversity metrics sometimes optionally consider the relevance of the content for users, assuming that what an individual user consumed in the past indicates what they will still be interested in, and that recommendations should be limited to such categories. And while related novelty metrics measure the newness of items or categories of recommendation, they are still individualized, and they are also agnostic about *what type* of content is new to the user.

In a similar way, to evaluate fairness means to be concerned with increasing diversity by seeking to redress the problematic under-representation of certain categories of source and content. But these approaches tend to adhere to fixed and pre-given definitions of genres and identities, in this way risking the reification of those categories and untethering them from processes of community validation of the kind we advocate in Section 3.2. Existing work on fairness addresses specific topics around increasing biases as well as the under-representation of particular groups (for a complete survey, see References [34] and [30]). Provider fairness metrics typically consider how many different groups of content providers appear in recommendations and assume a given distribution that it is desired to match. Consumer fairness metrics consider disparate treatments of the system in relation to different groups of consumers. Recent research has proposed more

general multi-stakeholder fairness metrics, acknowledging the impact recommender systems have on different groups of individuals [20, 74, 99].

A standard argument is that by reflecting biases embedded in the datasets, recommender systems create a feedback loop reinforcing such biases [28, 71]. The loop can be identified in popularity bias, which may reflect a mainstream bias in cultural domains. In music recommendations, this tendency reinforces "popular artists, at the expense of discarding less-known music" [82]. Against this propensity, diversity can be theorized as a key criterion for user satisfaction, providing music discovery for users who "do not want to listen to the highest rated song" within a system "over and over again" [69]. Both in provider and consumer fairness, recommender systems have therefore been shown to reproduce or exacerbate wider conditions of cultural and social discrimination against certain social groups.

On the consumer fairness side, movie recommender systems can reinforce biases against minority groups of users, as they reproduce user choices "through different iterations of users interaction" with the system [71]. In particular, there can be stronger bias amplification in recommendations for female users. In the music domain, gender bias, rooted "in cultural practices historically related with socio-political power differentials," can be "propagated by CF-based recommendations" based on user ratings [98]. Similar research demonstrates the propagation of users' gender biases on movie "recommendation algorithms," which "generally distort preference biases present in the input data and do so in sometimes unpredictable ways" [67]. There is also evidence of how popularity and demographic biases in both music and movie recommendation tend to affect user utility grouped by age and gender; models tend to perform better for male users and vary significantly across age groups [35]. Other studies show that popularity bias may lead to unfair treatment of users with little interest in popular items in the context of music recommendations [64], recommendations' accuracy can vary for different groups of users depending on their openness to listening to music beyond the mainstream [63], and finally, multiple recommender systems with different levels of popularity bias may affect users of different genders differently [66].

On the provider fairness side, music recommenders have been shown to under-represent female and non-binary artists, affecting users' listening behaviors as "higher proportions of female artists in recommended streaming is predictive of higher proportions of female artists in organic streaming" [37]. Still, in the music domain, certain systems reveal to have an imbalance of exposure between female and male artists and tend to confirm the feedback loop that moves male artists to the top [40]. With respect to books, "there are efforts in many segments of the publishing industry to improve representation of women, ethnic minorities, and other historically underrepresented groups" [36]. Yet, recommender systems tend to propagate disparities present in user profiles [36, 91].

Overall, it is striking that the various forms of bias shown by the under-representation of cultural content with respect to gender, race, class, and region (i.e., diversity of source or provider) correspond to wider core-periphery dynamics and geographical inequalities in the cultural industries [21, 104, 112]. It seems that recommender systems often mirror these inequalities, promoting Western-centric popular cultural content, in the English language, released by major producers [113]. Increasing diversity of source and producer, both as an issue of equity in itself and as it bears on diversity of content, is therefore a huge hurdle in achieving recommender systems oriented to enhancing cultural citizenship.

Aligned with our work, some recent research [86, 89] addresses normative issues of diversity in relation to recommender systems in the news domain. This attests to the importance and relevance of aligning recommender systems with human values as a key research direction in the community. Sanne Vrijenhoek et al. [89], building on earlier work [49, 101], propose a metric sensitive to four different models of democracy in relation to news recommendation. In contrast, our

research addresses a more general issue highlighted by the principles of PSM: the normative democratic importance of common exposure to certain kinds of content—in this study, content diversity modelled also on normative grounds—which we translate into and embed in our metric.

5 MEASURING COMMONALITY

In light of our discussion in previous sections about the importance of addressing the cumulative cultural and social effects of recommender systems, we contend that it is also crucial to identify means of developing systems aligned with the principles of commonality and cultural citizenship. Because of the importance of quantitative metrics in research and engineering of recommender systems, we now turn to translating our conceptual work into measurable properties. This requires developing an evaluation metric to measure the common experiences of diversity at the aggregate level. Assuming a democratic media environment, we seek to evaluate whether a recommender system contributes to the strengthening of cultural citizenship by systematically promoting diversity of source and content within a given type of cultural content (in our experiments, movies, music, and literature). In this way, evaluation has the potential to assist in counteracting sexist and racist biases and the neglect of non-Western and non-mainstream content across a user population. This also provides a means of evaluating the extent to which a given recommender system is contributing to the kinds of wider cultural changes called for by anti-racist and feminist critics as well as by those sympathetic to criticisms of existing recommender systems.

5.1 Metric Definition

We are interested in measuring the extent to which users, in response to algorithmic recommendations, gain a shared familiarity with a diverse set of content. This requires us to consider four concerns when evaluating a system according to commonality. First, given the importance of the plurality of cultural content (Section 3.1) and the role of human editors in supporting it (Section 3.2), we need to define how to represent the categories and classes of items selected by editors (Section 5.1.1). Second, given that we are interested in promoting cultural citizenship (Section 3.1), we need to quantify the extent to which an individual user is familiarized with a category selected by editors (Section 5.1.2). Third, because the promotion of cultural citizenship requires a *common* cultural experience (Section 3.1), we need to quantify the *joint* familiarity across users with a category selected by editors (Section 5.1.3). Finally, since editors will usually select multiple categories, we need to aggregate per-category commonality into a composite metric (Section 5.1.4).

5.1.1 Selecting Categories. The promoted categories, we suggest, will be identified and curated by editors in a relevant field seeking to promote a plurality of cultural content in the service of strengthening cultural citizenship (Section 3.2). We can contrast human curation with statistical methods for selecting under-represented categories, such as those used in some algorithmic fairness work (e.g., Reference [73]). Purely statistical methods, because they are agnostic to the cultural and social dimension of items, can result in under-represented content misaligned with the goals of enhancing diversity of common cultural experience (e.g., lower quality or toxic content). As described in Section 3.2, editors make curatorial decisions drawing on insights generated by their knowledge about cultural expressions and social conditions with the goal of achieving a diverse mix of content and sources that avoids the risks of employing reified models of both identities and genres. These editors may opt to promote, for example, movies by female directors or those produced for non-Western markets.

Given the large body of criticism of bias and unfairness in the recommender system literature, and for the purpose of testing the commonality metric, in what follows, we chose to work experimentally with widely recognized under-represented categories of source or producer in the

three chosen media (movies, music, and literature). The under-represented categories come in three broad clusters, which are: female and non-binary gendered producers, artists or authors; independent production; and non-Western sources. At the same time, boosting diversity by promoting these under-represented source and producer categories bears directly on—and is very likely to increase—the diversity of content in each case. However, it is important to point out that the approach and the principles set out in this article can be applied in alternative ways, employing different categories and boosting different vectors of diversity. The distinctive facet of our work is not so much the attempt to redress specific kinds of under-representation—although we are certainly concerned with this challenge both in itself and as a key component of the goal of enhancing cultural citizenship. Rather, it is our ambition to find means of binding such an intervention to larger normative ambitions (strengthening cultural citizenship) and to find means of evaluating the effects of this intervention (that is, increasing the diversity of cultural experience) not just on individuals but universally, across populations of users.

5.1.2 Measuring Familiarity. To measure familiarity, we make the following assumptions:

- **tabula rasa users:** a user begins their recommendation session with no background in the relevant categories (Section 5.1.1);
- **single turn sessions:** a user engages with exactly one ranked list of recommendations;
- **exposure improves familiarity:** the familiarity of a user with a category improves if they are exposed to an item of that category, even if it is not consumed.

The *tabula rasa* users assumption is consistent with existing search and recommendation evaluation that assumes users are completely unfamiliar with any unrated items (in the case of traditional search) or any unseen subtopics (in the case of the intent-aware search [116]). As a result, our measure of familiarity will be a lower bound on a system's true effectiveness. The single turn sessions assumption is also consistent with current batch evaluation practice in recommender systems. Finally, we assume that users are presented with recommendations through an interface where navigating recommendations results in improved familiarity. In "radio-style" interfaces where recommended tracks are streamed without an affordance to skip, users necessarily must consume content to serially browse the recommendations. Even in situations where recommendations are provided as a ranked list with summaries (e.g., thumbnails, snippets), the exposure to the presence of an item category can improve the probability of familiarity; for example, Kay et al. [59] found that shifting the categorical distribution of search results can impact a user's assumptions about the distribution of content. While these assumptions clarify our exposition, we believe extensions to users with heterogeneous backgrounds, multi-turn recommendations, and more refined notions of consumption and familiarity are valuable topics for future work.

We can measure a user's familiarity with a category after having interacted with a ranked list of recommendations by connecting it to the notion of *recall* found in existing search and recommendation evaluation. Recall refers to the fraction of all relevant content that a user has come across in a ranking. It is often used to measure the coverage of relevant content—in our case, items in an editorially selected category—that a user will encounter in a session.

More formally, let π_u be a ranking of n items from the catalog $\mathcal{D}$ for user $u \in \mathcal{U}$. As with most recommender system evaluation, we assume that a user scans linearly from the top-ranked item downward. We can measure familiarity as the recall of items in category g at position k,

$$\mathrm{R}(\pi_u, k, g) = \frac{|\pi_{u,:k} \cap \mathcal{D}_g|}{|\mathcal{D}_g|}, \tag{1}$$

where $\mathcal{D}_g$ is the set of items labeled with category g.

Although we could use a fixed cutoff k, this may not capture users that terminate their scan of the list before or after the kth item. We can explicitly model the probability that a user stops at a particular rank position as a multinomial over positions, $\Pr(k)$. Carterette [24] demonstrates that this model underlies most popular ranked list evaluation metrics. Specifically, in our experiments, we adopt the browsing model used in rank-biased precision [25, 77],

$$\Pr(k) = (1-\gamma)\gamma^{k-1}, \tag{2}$$

where the patience parameter $\gamma \in (0,1)$ controls how deep into the ranked list the user is likely to progress, regardless of relevance. If $\gamma \leq 0.5$, then we are modeling a user who stops scanning early in the ranking, while higher values of γ models a user who scans more deeply.[5]

Combining Equations (1) and (2), we can compute, given a ranking π_u, the familiarity of u with category g as the expected recall,

$$\Pr(F_{u,g}|\pi_u) = \sum_{i=1}^{n} \Pr(i)\,\mathrm{R}(\pi_u, i, g), \tag{3}$$

where $F_{u,g}$ is a binary random variable indicating that the user is familiar with the category.

5.1.3 Commonality. As discussed in Section 3.1, the notion of commonality stresses the importance of a *shared* cultural experience for cultural citizenship. Translating the concept of a shared experience can be addressed in multiple ways. However, we might average the familiarity across users. In this case, we would be computing the expected familiarity a user will have in a category. However, the arithmetic mean can be dominated by outliers. Consider the following two distributions, A and B, of familiarity values for 10 users,

	1	2	3	4	5	6	7	8	9	10
A	1	1	1	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ
B	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

In distribution A, the expected familiarity is higher than that for distribution B. However, in distribution A, only three of the users really have a shared cultural experience, since those users with familiarity of ϵ[6] effectively are unfamiliar with the category. In distribution B, every user has at least partial familiarity, suggesting that it is better aligned with our objective of a shared experience.

As an alternative to the arithmetic mean, we can measure the probability that every user simultaneously gains familiarity with the editorially selected categories. This models familiarity as a binary random variable estimated by recall. The joint distribution models the probability that every user is familiar with a category simultaneously. This approach is better aligned with our objective of commonality, since it explicitly models a collectively shared experience. Formally, given a set of editorially selected categories $\mathcal{G}$, we can compute the joint distribution of familiarity with respect to a single category $g \in \mathcal{G}$ as,

$$\begin{aligned} \mathrm{C}_g(\pi) &= \Pr(F_{1,g}, \ldots, F_{m,g}|\pi) \\ &= \prod_{u \in \mathcal{U}} \Pr(F_{u,g}|\pi_u). \end{aligned} \tag{4}$$

[5]In all our experiments, we adopt $\gamma = 0.5$ as a method common in user-based evaluation metrics for retrieval and recommendation [25].

[6]Note that ϵ here is used to indicate that there is a marginal value of familiarity for those users.

In our example distribution, the joint probability for distribution B is larger than that of A because familiarity is higher in general. Put another way, the lowest familiarity is higher for B than for A. In practice, although our browsing model is strictly positive (i.e., greater than zero), to avoid numerical precision issues, we use the logarithm of commonality, which is rank equivalent with commonality.

5.1.4 Aggregation. While some system designers may be comfortable analyzing commonality disaggregated by category, many will want a summary of commonality across groups. This may be due to convenience (e.g., a leaderboard) or out of a desire to measure robust performance across a given set of categories, in the interest of promoting a shared plurality of cultural content.

As such, we developed an aggregated commonality metric, summarizing performance across categories. Although we could aggregate per-category commonality using an arithmetic mean, this might be unstable due to different category sizes and, as a result, uncalibrated metric values (Equation (4)). Instead, we adopt Borda's rank aggregation method. Assume that we have a set $\mathcal{S}$ of systems, each associated with a set of per-user rankings $\{\pi^s\}_{s\in\mathcal{S}}$. We begin, for each category $g \in \mathcal{G}$, by generating a system ranking according to $\mathrm{C}_g(\pi^s)$. We then assign the top-ranked system with a value of 1, the second-ranked system with a value of 2, down to $|\mathcal{S}|$ for the last-ranked system. Aggregating these "votes" across categories results in a final system score, where lower values are better.

5.2 Mathematical Analysis

In Section 4, we discussed prior approaches to measuring the diversity and fairness of recommender systems in the context of cultural citizenship. Given that our commonality metric is developed with cultural citizenship in mind, we turn to contrasting the specific mathematical differences between commonality and prior metrics.

5.2.1 Aggregation. The fundamental conceptual shift from individualized to collective metrics is reflected in our adoption of the joint probability of user events. As a simple contrast, consider **normalized discounted cumulative gain (NDCG)** [56]. When computing the aggregate metric, we look at the sample mean over users,

$$\mathbb{E}_{\mathcal{U}}[\mathrm{NDCG}(\pi)] = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathrm{NDCG}(\pi_u),$$

where $\mathcal{R}_u$ is the set of items relevant to user u. In this case, we can see that NDCG sums the metric value across users, while commonality (Equation (4)) multiplies familiarity across users. This results in a metric that is much more sensitive to supporting collective familiarity and shared cultural experiences.

As a general case, diversity and fairness metrics operate similarly. We adopt an individualized metric, compute its value for a user (e.g., answering "how fair/diverse is this ranking for this user?"), and then compute the sample mean. As a result, we can imagine situations where the lack of diversity or fairness for some users is "compensated for" by users whose recommendations are more diverse or fair.

One exception to this is fairness metrics based on measuring the Kullback-Leibler divergence between the distribution of categories in recommendations from a uniform distribution over categories. Let θ be the distribution of exposure over recommended groups and $\theta^* = \frac{1}{|\mathcal{G}|}$ a uniform distribution over groups [60]. We can reduce the sample mean of this metric to a rank equivalent

quantity based on the sum of group joint probabilities,

$$\begin{aligned}\mathbb{E}_{\mathcal{U}}[\Delta_{\mathrm{KL}}(\pi)] &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \mathrm{D}_{\mathrm{KL}}(\theta^*\|\theta_u)\\ &= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\sum_{g\in\mathcal{G}} \frac{1}{|\mathcal{G}|}\log\left(\frac{\frac{1}{|\mathcal{G}|}}{\theta_{u,g}}\right)\\ &\overset{\text{rank}}{=} -\sum_{u\in\mathcal{U}}\sum_{g\in\mathcal{G}}\log\left(\theta_{u,g}\right)\\ &= -\sum_{g\in\mathcal{G}}\log\left(\prod_{u\in\mathcal{U}}\theta_{u,g}\right).\end{aligned}$$

In this case, we can see that, like commonality, Δ_{KL} includes a product of per-user metrics. However, there are two slight differences. First, the metric being multiplied is the relative exposure of a category in the user's ranking as opposed to the familiarity. While these may sometimes be correlated, there are certainly situations where we might observe high $\theta_{u,g}$ and a low $F_{u,g}$, meaning that Δ_{KL} would be inappropriate for measuring the shared familiarity. Second, the aggregation of joint metrics in Δ_{KL} uses a simple sum aggregation, which is possible, in part, because $\theta_{u,g}$ is calibrated across groups while $F_{u,g}$ may not be (i.e., differences in sizes of categories may lead to different ranges of empirical values). Note that this observation may be unique to using the Kullback-Leibler divergence, which includes a logarithmic term.

Another way to interpret category commonality is as the geometric mean of the recall of a category, connecting it to geometric mean average precision [87]. In the context of utility metrics, Valcarce et al. [108] experiment with geometric mean performance, finding that it is more robust than the arithmetic mean when dealing with samples of users. We will return to this observation in Section 5.3.5.

5.2.2 Category Metric. A second difference between commonality and prior fairness, diversity, and novelty metrics is in the category-level metric.

Fairness metrics tend to emphasize divergence from some reference distribution of categories [60]. In the previous section, we saw the example of Δ_{KL}, where, when using the uniform distribution over categories as a reference, the aggregated metric reduces to the magnitude of categories in the recommendations. These per-user quantities capture the *presence* of the category as opposed to the *comprehensiveness* of the category (i.e., recall). Even in the case of non-uniform reference distributions, exposure distributions are normalized in such a way that any recall information is removed, implying that, while similar in form, fairness metrics are mathematically measuring a different phenomenon.

While fairness metrics capture the divergence of category exposure from some reference distribution, diversity metrics measure the support of the exposure distribution in some space. Consider the **expected intra-list distance (EILD)**. Assume that we consider all recommended items,

$$\mathbb{E}_{\mathcal{U}}[\mathrm{EILD}(\pi)] = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\Pr(i)\Pr(j-i)\delta(\pi_{u,i},\pi_{u,j}),$$

Table 1. Research Questions

	Section	Research Question
RQ1	Correlation with existing metrics (Section 5.3.3)	Is commonality correlated with previous metrics when ranking systems under different conditions (e.g., when editors create different categories)?
RQ2	Robustness to missing category labels (Section 5.3.4)	How robust is commonality—and other metrics—to changes in the items labeled on the different categories?
RQ3	Generalization from sampled users (Section 5.3.5)	How consistent is the commonality metric when it addresses a subset of users as compared to the full population of users?

where δ is a linear distance function between items. If each item belongs to only one category, then we can set $\delta(i,j) = 1 - \sum_{g \in \mathcal{G}} \mathrm{I}(i \in \mathcal{D}_g)\,\mathrm{I}(j \in \mathcal{D}_g)$ and derive a rank-equivalent metric,

$$\mathbb{E}_{\mathcal{U}}[\mathrm{EILD}(\pi)] \overset{\text{rank}}{=} -\sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \underbrace{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \Pr(i) \Pr(j-i)\,\mathrm{I}(\pi_{u,i} \in \mathcal{D}_g)\,\mathrm{I}(\pi_{u,j} \in \mathcal{D}_g)}_{\text{exposure of items in } \mathcal{D}_g}.$$

From this, we can see that, like commonality, diversity computes the exposure of items in a category. Like fairness metrics, diversity metrics measure presence as opposed to comprehensiveness.

In terms of novelty, take the **expected profile distance (EPD)** metric [111]. Assume that we consider all recommended items and relevant items in the users profile,

$$\mathbb{E}_{\mathcal{U}}[\mathrm{EPD}(\pi)] = \frac{C}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i=1}^{n} \Pr(i) \sum_{r \in \mathcal{R}_u} \delta(\pi_{u,i}, r),$$

where C is a constant and δ is a linear distance function between items. If each item belongs to only one category, then we can set $\delta(i,j) = 1 - \sum_{g \in \mathcal{G}} \mathrm{I}(i \in \mathcal{D}_g)\,\mathrm{I}(j \in \mathcal{D}_g)$ and derive a rank-equivalent metric,

$$\mathbb{E}_{\mathcal{U}}[\mathrm{EPD}(\pi)] \overset{\text{rank}}{=} -\sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \underbrace{\sum_{r \in \mathcal{R}_u} \mathrm{I}(r \in \mathcal{D}_g) \sum_{i=1}^{n} \Pr(i)\,\mathrm{I}(\pi_{u,i} \in \mathcal{D}_g)}_{\textit{new} \text{ exposure of items in } \mathcal{D}_g}.$$

From this perspective—and with this distance function—we can see that, for each category, the metric measures magnitude of exposure of items in that category *for users who have already engaged with an item in that category*. For example, in the movie domain, taking the category of West African film, if a user has already positively rated at least one movie in that category, then this metric would measure how many more West African films are recommended.

5.3 Empirical Analysis

Although we synthesized concepts from the PSM literature and the evaluation literature to develop our commonality metric, we are interested in understanding its empirical behavior as an evaluation metric. Therefore, in this section, we conduct several experiments[7] to study and compare different properties of the commonality metric. For clarity, in Table 1, we summarize the three research

[7]The code to reproduce the experiments is hosted here: https://github.com/andrebola/commonality-recsys-tors

Table 2. Summary of the Resulting Datasets Used in the Experiments

Domain	Dataset	# Users	# Items	# Ratings/Interactions	Density %
Movies	MovieLens-1m [48]	3,706	6,040	1,000,209	4.47
Music	LFM-2b [94]	18,711	28,341	1,758,838	0.33
Literature	LibraryThing	7,279	37,232	749,401	0.28

questions that we address by each experiment in the following subsections. In Section 5.3.3, we first look at how different the information captured by commonality is from that captured by existing metrics. More specifically, we want to know if commonality is correlated with previous metrics when ranking systems under different conditions (RQ1). In Section 5.3.4, we look at how robust commonality is to changes in the items labeled on the different categories (RQ2). Finally, in Section 5.3.5, we investigate how sensitive commonality is to a reduction in the size of the user population used to compute the metric (RQ3).

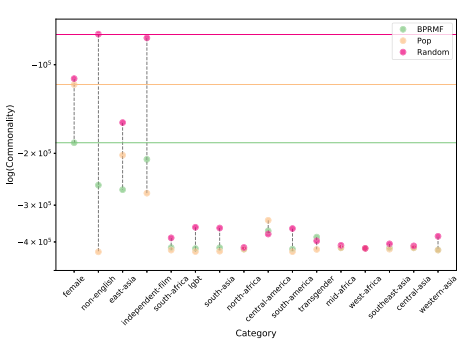
It is important to highlight that for all these research questions, we are focusing on the consumption of *diverse categories* of items that are generally under-represented by existing recommender systems.

These analyses use a fixed experimental setup consisting of multiple, publicly available datasets, which we use to compare commonality with existing metrics. Our analyses focus on the behavior of the commonality metric under different possible editorial policies. So, while a production system would employ editors who act as conduits for value communities, we select categories such that they are representative in terms of both size and coverage of what we might expect from a human editor.

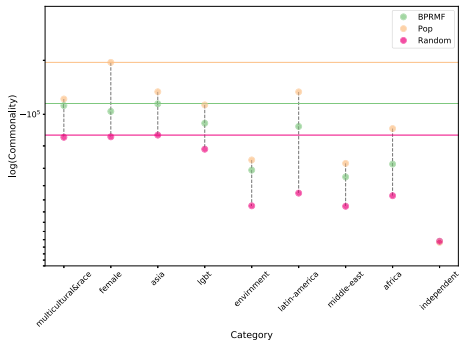
5.3.1 Data. We consider three recommender system domains dealing with cultural content: movies, music, and literature. For each dataset, in addition to publicly available data, we selected categories (i.e., $\mathcal{G}$) based on their historic under-representation to assess the behavior of our metric. In Table 2, we summarize the dataset used in the experiments. Figures 1 describe the "raw" values of commonality for each dataset.

Movies. We use the MovieLens-1m dataset [48], which contains 1,000,209 ratings of approximately 3,900 movies from 6,040 users from the MovieLens platform. Using a separate dataset,[8] we augmented the MovieLens movies with metadata including country of production, gender of the director, original language, and keywords collected from the movie's description. For this dataset, we used rankings from multiple recommendation systems prepared by Valcarce et al. [107]. Following the method described by the authors, we converted to binary relevance labels considering ratings of 4 and 5 as relevant. We selected categories of movies that are typically under-represented by movie recommender systems. Specifically, we consider female directors (under-representation by gender); independent film (under-representation by industry sector); and several sources of non-Western film (under-representation due to geographical and linguistic inequality). We use categorical gender data, acknowledging the limitations of this framing [53]. For geographic categories, we use the country of production for the following regions of the world: South America, Central America, North Africa, South Africa, West Africa, Mid Africa, Southeast Asia, South Asia, Western Asia, Central Asia, and East Asia. We consider, broadly, non-English language movies as a separate category. And, finally, we use keywords to create categories with selected movies whose categories contain "independent films," "LGBT," and "transgender." We manually checked whether these keywords can be trusted to represent specific identities.

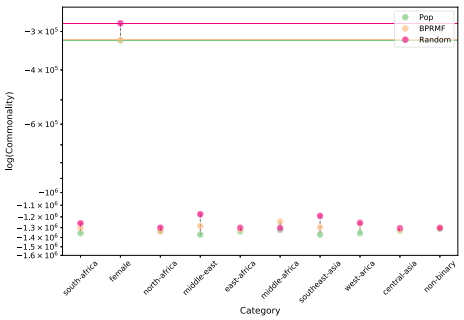
[8] https://www.themoviedb.org/



(a) Commonality for MovieLens-1m



(b) Commonality for libraryThing



(c) Commonality for LFM-2b

Fig. 1. Per-category commonality values for recommendations based on random, popularity, and BPRMF models for three datasets. Horizontal lines indicate the mean commonality across categories.

Music. For music, we use LFM-2b dataset [94], which is the largest dataset containing users' listening events for 120,000 Last.fm users and over two billion listening events. We enriched this dataset with additional information about the artists collected from MusicBrainz.org, a large collaborative database of music information. From this dataset, we only considered interaction between the years 2013 and 2020. After removing items that were listened to by fewer than 15 users and users that listened to less than 15 items, the resulting dataset had 18,711 users and 28,341 items. From these items, a total of 2,712 belong to at least one of the categories we selected to enhance diversity. We selected eight categories related to non-Western regions of the world where artists are primarily based (North Africa, East Africa, Middle Africa, South Africa, West Africa, Middle East, Central Asia, and Southeast Asia). We additionally selected two categories based on artists' gender information collected from Musicbrainz: The two categories are female artists and non-binary artists. We acknowledge that the category of female artists represents a group much larger than the other categories but, as we have shown, they still suffer under-representation due to both industrial and recommender biases when compared to male artists. We trained 12 recommender systems for the last.fm data using the Elliot library. We use models based on MF2020, NeuMF, RP3beta, BPRMF and iALS trained both with binarized and original input; plus two baselines: Popularity and Random.

Literature. For literature, we use the LibraryThing dataset,[9] containing user book ratings. We use a subset of the dataset containing 7,279 users and 37,232 items. We collected information for these

[9] https://cseweb.ucsd.edu/~jmcauley/datasets.html#social_data

books from librarything.com. From the information, we selected the following categories: Africa, Asia, Latin-America, Middle-East, Environment,[10] Female, LGBT, Independent, and Multicultural & Race. A total of 8,171 books correspond to at least one category. For this dataset, we also use rankings from multiple recommendation systems prepared by Valcarce et al. [107]. Following the method described by the authors, we first convert the original ratings to a scale of 1–5 and then convert to binary relevance labels considering ratings of 4 and 5 as relevant.

5.3.2 Baseline Metrics. For all the experiments, we compare commonality against three classes of metrics: utility metrics, diversity metrics [111], and fairness metrics [22]. We measure utility-focused properties using precision (P), recall (R), and NDCG. We measure fairness across categories $\mathcal{G}$ using disparate exposure (U) [45] and the divergence family of metrics (Δ_{abs}, Δ_{sq}, Δ_{KL}) using the probability of exposure to categories [60]. We measure diversity of categories $\mathcal{G}$ using α – NDCG and IA – ERR and novelty using **Expected Intra-List Distance** (EILD) and **Expected Profile Distance** (EPD) with distances based on genre representations of items.

5.3.3 Correlation with Existing Metrics. In our first analysis, we were interested in understanding the correlation between commonality and existing metrics (Section 5.3.2). Observing consistently high correlations between commonality and existing metrics across domains would suggest redundancy, reducing the need for a new metric. We measure this correlation across three different editorial regimes for selecting items within each category (i.e., which subset of $\mathcal{D}_g$). In the first condition, we assume that editors select *all* items in $\mathcal{D}_g$. For example, if items authored by women were a broad category of interest, then an editor in this condition would be interested in promoting a comprehensive familiarity with all items authored by women. In this condition, larger categories, while sometimes more likely to be recommended naturally (if the category is already popular) will be more difficult to achieve high familiarity with, even when explicitly programmed; smaller categories, however, will, by chance, have a lower probability of exposure but, with explicit programming, can reach high familiarity. In part to address this, we consider a second regime wherein an editor may downsample items from the largest category. Our final regime considers downsampling items from all categories until they have equal size. In this section, our analyses help us understand inter-metric correlation as a function of different editorial conditions.

To compare metrics, we compute the Kendall's τ correlation between system rankings according to commonality and existing metrics.

Full Category Selection. In our first analysis, we compare the correlation when using an editorial policy that selects all items in a category for promotion. Our results (Table 3) indicate that none of our utility and fairness metrics show a strong consistent correlation with commonality. While some domains show stronger correlations for some of these metrics, there is no evidence that commonality is redundant with these metrics. In terms of diversity and novelty, both EILD and EPD show stronger correlation with commonality consistently across domains.

Downsampling the Dominant Category. To understand the relation between commonality and previous metrics under different editorial policies, we now look at how selecting a subset of items in the larger category would affect the rank correlations. In this analysis, we progressively remove category labels for items from the dominant category, simulating a policy that de-emphasizes familiarity with the comprehensive set of items in a category. We randomly downsample the dominant category to percentages of the original size, including 10%, 30%, 50%, 70%, and 90%.

[10]This category includes literature on topics such as global warming and climate change. We identified this as a relevant topic that editors could potentially choose to raise awareness in the audience. Note that while this is a progressive category, it does not strictly follow our selection of categories to boost the diversity of user exposure to cultural content.

Table 3. Full Category Selection

	movies	music	literature
Utility			
P	−0.119	**0.512**	0.053
R	−0.138	**0.574**	0.043
NDCG	−0.205	**0.450**	0.072
Fairness			
U	**0.721**	0.326	**0.763**
Δ_{KL}	**0.616**	**0.698**	0.062
Δ_{sq}	**0.348**	**0.636**	**−0.647**
Δ_{abs}	**0.339**	**0.605**	**−0.647**
Diversity and Novelty			
α − NDCG	−0.062	**0.512**	0.254
IA − ERR	−0.100	**0.512**	0.254
EILD	**0.730**	**0.853**	**0.782**
EPD	**0.702**	**0.822**	**0.763**

Correlation between commonality utility, fairness, and diversity metrics. Kendall's τ between rankings of runs with Bonferroni correction to correct for multiple comparisons (bold: $p < 0.05$).

We repeat the process five times, averaging correlations across runs. We present our results in Figure 2. In general, we observe similar correlations to our first analysis, namely, that EILD and EPD show stronger correlation with commonality consistently across domains. This suggests that, even though we downsample labels, commonality is degraded similarly across systems, leaving their rank order unaffected. This stability is unsurprising, since our Borda aggregation method disregards absolute scores and instead uses relative rank-based system weights.

Downsampling All Categories. In our final analysis, we measure how metrics correlate with commonality when editors downsample items from categories to have similar size. We scale the size of the categories such that 0% keeps all the original items and 100% reduces all the categories to the size of the smallest category. Similar to our last analysis, we sample items from each category, removing the annotations of the items that are not selected and repeat the process five times, averaging the results. Note that in the our previous analysis, we only reduce the largest category, while in this case, we reduce the most dominant categories until they are all uniform; we downsample items at 10%, 30%, 50%, 70%, 90%, and 100% of the original category size, where, at 100%, all categories have the same size and 0% means that they all have the original size (as in our first analysis). Our results, presented in Figure 3, show relatively stable correlations across sampling rates except in the case of the divergence fairness metrics (Δ_{sq}, Δ_{abs}), where the correlation improves as category sizes become more similar. This follows from the fact that uniform exposure across categories is precisely the objective of the divergence class of fairness metrics. Although degrading slightly when categories have more uniform sizes, both EILD and EPD still have higher consistent correlation with commonality.

Summary. Our analysis indicates that, across the editorial policies we considered, diversity metrics (EILD and EPD) maintain high correlation with commonality. Although these specific fairness and diversity metrics show stronger correlation than other metrics, their absolute correlation varies substantially across domains and remains relatively far from perfect correlation. Returning to our earlier analysis, much of this correlation is due to the fact that these measures, unlike utility and fairness metrics, capture the exposure of promoted content *on average* while commonality captures the exposure *simultaneously*.

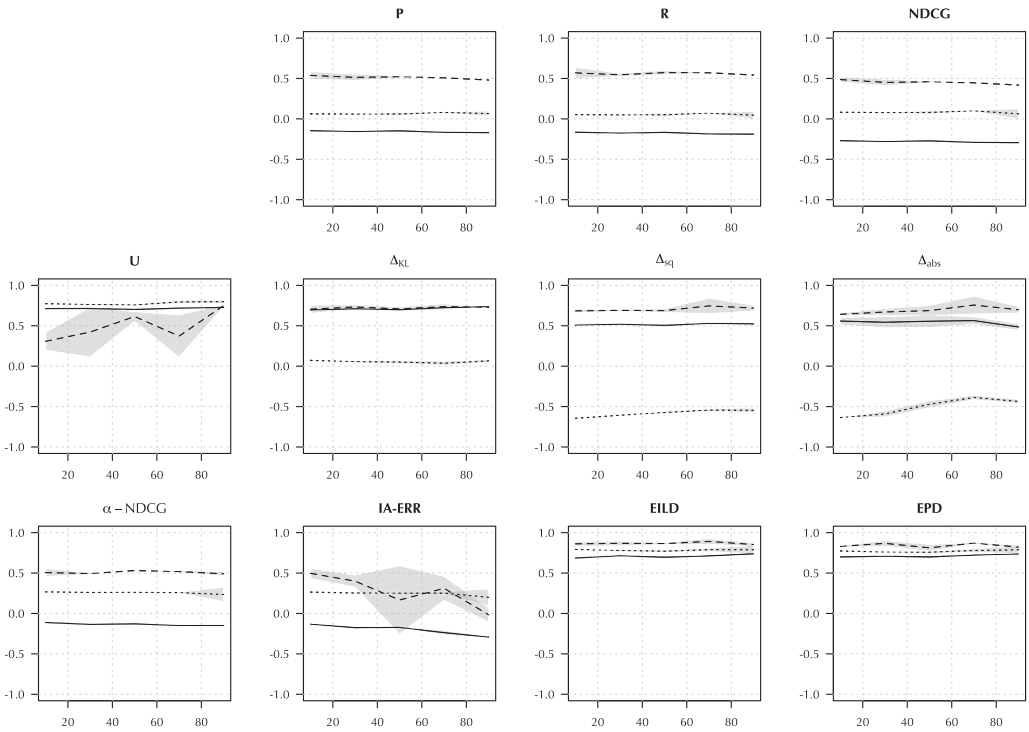


Fig. 2. Correlation with commonality when downsampling the dominant category. Horizontal axis (all plots): percentage downsampled. Vertical axis (all plots): τ correlation with commonality. Top row: Utility metrics; middle row: fairness metrics; bottom row: diversity metrics. Solid line: movies; dashed line: music; dotted line: literature. Lines show mean across five trials. Shaded regions indicate one standard deviation around the mean.

5.3.4 Robustness to Missing Category Labels. In our second analysis, we evaluate the robustness of commonality to missing category labels. To do this, we remove category labels for items in each category and measure the correlation between the metric computed with incomplete category labels and the metric with complete category labels. This is different from our earlier analysis, because we are simulating errors induced when editors have incomplete information about the complete set of items they *would like* to select. We present the results of this analysis in Figure 4. Commonality degrades with increasing label noise due to impact on recall estimates. That said, since all systems are uniformly subject to incomplete data, the degradation in correlation is slight. As expected, utility metrics that do not use category labels show strong correlation with complete label information, regardless of missing labels. Fairness, diversity, and novelty metrics—with the exception of Δ_{KL} and IA – ERR—show more dramatic degradation compared to commonality.

5.3.5 Generalization from Sampled Users. Since offline evaluation approximates performance for a full population of users within a sample, we were interested in understanding the stability of commonality under smaller samples. In this analysis, we evaluate commonality on a random subset of users and measure whether the ranking of systems changes significantly. To test how well a metric generalizes to a larger population of users, we randomly sample a subset of users in the range 10%–90% and measure the correlation between the metric computed over the different

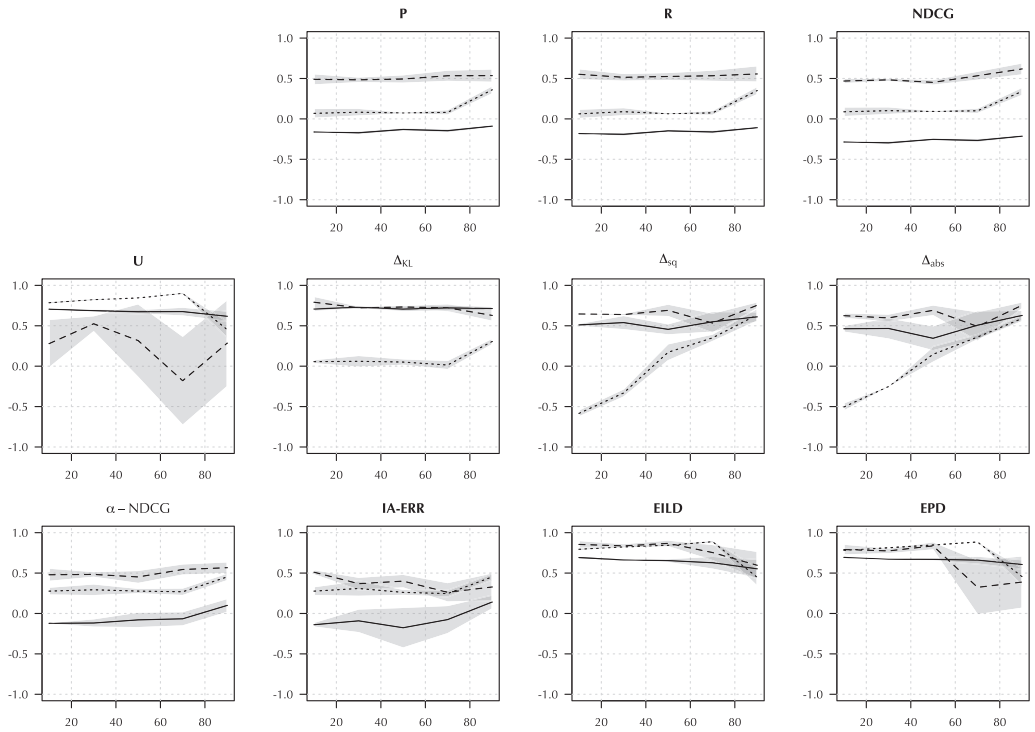


Fig. 3. Correlation with commonality when downsampling all categories. Horizontal axis (all plots): percentage downsampled. Vertical axis (all plots): τ correlation with commonality. Top row: Utility metrics; middle row: fairness metrics; bottom row: diversity metrics. Solid line: movies; dashed line: music; dotted line: literature. Lines show mean across five trials. Shaded regions indicate one standard deviation around the mean.

subsets using Kendall's τ, preferring metrics where the ranking of the systems is consistent across the different user samples. We show the results for this analysis in Figure 5. Commonality performance degrades slightly with smaller samples, although not as catastrophically as divergence-based fairness metrics or IA − ERR, consistent with prior literature [107]. Utility metrics tended to be robust, consistent with prior literature [107].

6 DISCUSSION

Since commonality when linked to other progressive cultural principles (here, diversity) is a normative property we seek to promote in recommender systems, we have emphasized clear connections between it and the formal properties of our metric (e.g., diverse curation, familiarity). This exercise involved substantial translational work between disciplines—between ideas from the social sciences and humanities, and perspectives from recommender engineering. Specifically, we derived normative principles from the literature on public service media and translated them into guiding principles for the design of quantitative evaluation. In contrast with other evaluation metrics—including many based on personalization—we do not have a latent or delayed quantity to validate the metric. As such, conceptual analysis and theoretical development play a necessary and an exceptionally important role in the overall research we are presenting here, and it has been imperative to combine these conceptual perspectives with mathematical and empirical research, as we have shown.

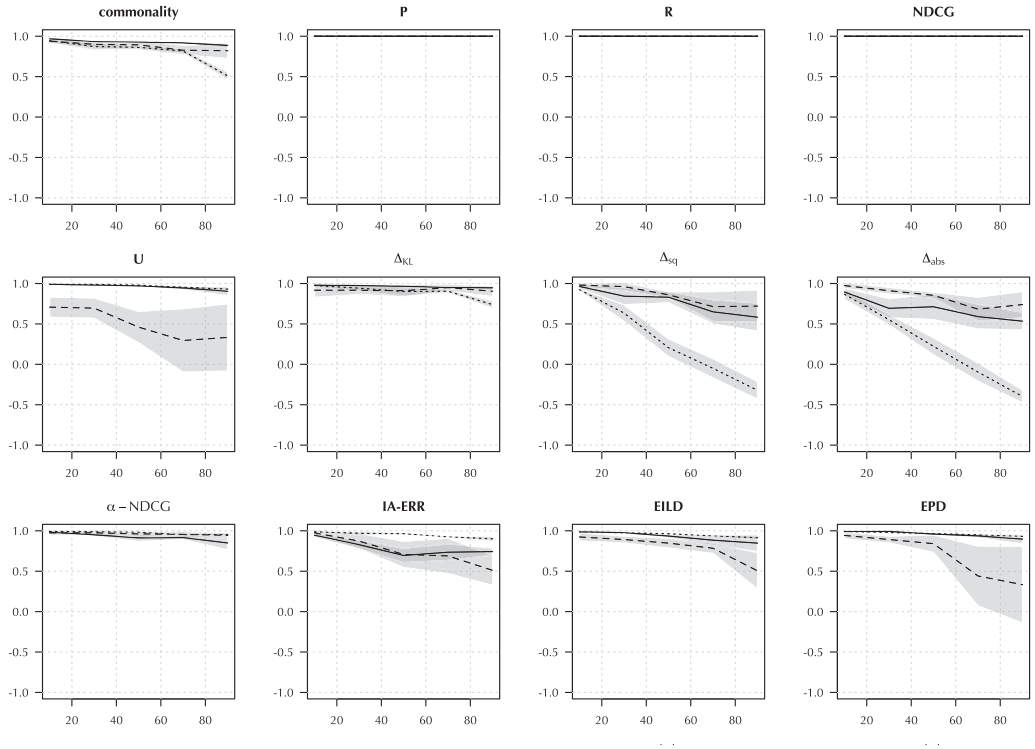


Fig. 4. Robustness to missing category labels. Category labels were progressively removed from items and then the correlation between the system ranking with partial labels and the system ranking with complete labels was measured. Horizontal axis (all plots): percentage downsampled. Vertical axis (all plots): τ correlation with complete labels. Top row: Commonality and utility metrics; middle row: fairness metrics; bottom row: diversity metrics. Solid line: movies; dashed line: music; dotted line: literature. Lines show mean across five trials. Shaded regions indicate one standard deviation around the mean.

We were, in part, motivated by the proposition that existing evaluation metrics fail to capture broader principles associated with the promotion both of universality (commonality) and of cultural citizenship. The mathematical comparison of commonality with existing metrics (Section 5.2) demonstrated that formal properties of commonality were absent in existing metrics. Our empirical results, in response to RQ1 (Section 5.3.3), further support this proposition based on the inconsistent correlation between commonality and existing metrics. We posed RQ2 (Section 5.3.4) to address various types of noise in the labelled data in a real application. We showed that commonality degrades less than other fairness, diversity, and novelty metrics, as it loses complete data. However, RQ3 referred to changes in the user population, delving into a highly relevant issue for this study: variations in the audiences or users addressed by each recommender system. In this case, our commonality metric degrades slightly with smaller samples, although not as catastrophically compared to other fairness metrics or some diversity metrics.

Human editors—and the value communities they channel and from whom they derive validated categories and judgments—play an important role in the assemblage supporting our evaluation metric. Even though we used examples of categories justified by existing literature, by envisaging editors answerable to knowledgeable communities that would guide category definition and assignment, we were able to investigate this metric performance while attending in the broader

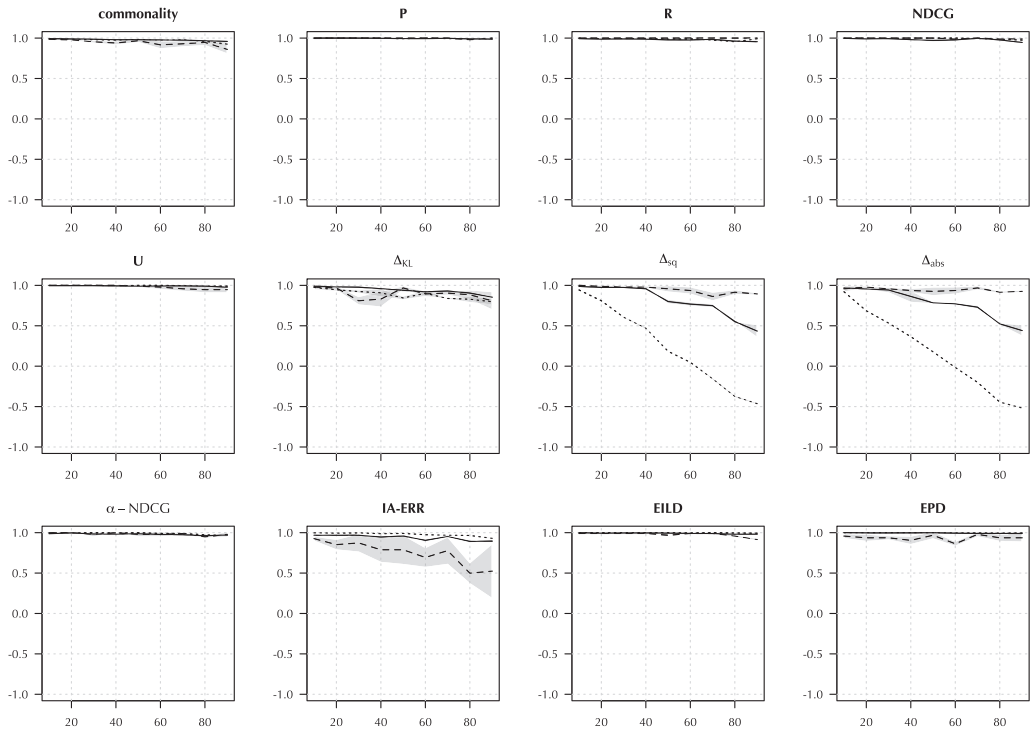


Fig. 5. Generalization from sampled users. Correlation between rankings of systems using metrics with samples of users and rankings of systems using metrics with full samples of users. Horizontal axis (all plots): percentage downsampled. Vertical axis (all plots): τ correlation with complete labels. Top row: Commonality and utility metrics; middle row: fairness metrics; bottom row: diversity metrics. Solid line: movies; dashed line: music; dotted line: literature. Lines show mean across five trials. Shaded regions indicate one standard deviation around the mean.

design of the assemblage to SSH concerns about the risks of identity essentialism.[11] While the categories we selected were limited by labels in our datasets, the general behavior of metrics we observed are representative of the diversity in size and prominence that we expect in practice. Given that, in this series of experiments, we ourselves substituted for the editors we envisage, we are interested in future work in exploring the extent to which, and how, categories and items selected by those editors would affect the results. Moreover, in some cases, editors may desire more fine-grained control over category importance. In this situation, we can easily adapt our Borda count method to incorporate weights for categories [52].

In previous work [39], we aggregated commonality values using mean commonality across groups instead of Borda count (Section 5.1). While mean aggregation is appropriate when aggregated values are calibrated across groups, it can degrade in the presence of outliers, which can occur due to differences in category sizes [39, Figure 1(b)] . Borda aggregation, however, preserves only the rank position of each system during aggregation, discarding the magnitude of differences

[11]The risks of essentialism alluded to here are denounced in decolonial data feminist writing, which argues that "predatory data's algorithmically driven platforms and 'predictive' architectures have massified reductive classification schemes" [26]. The alternative envisaged is to promote "explicitly pluralistic, coalitional knowledge" practices, a version of which we are modelling here through the editors and their relationship to evolving value communities.

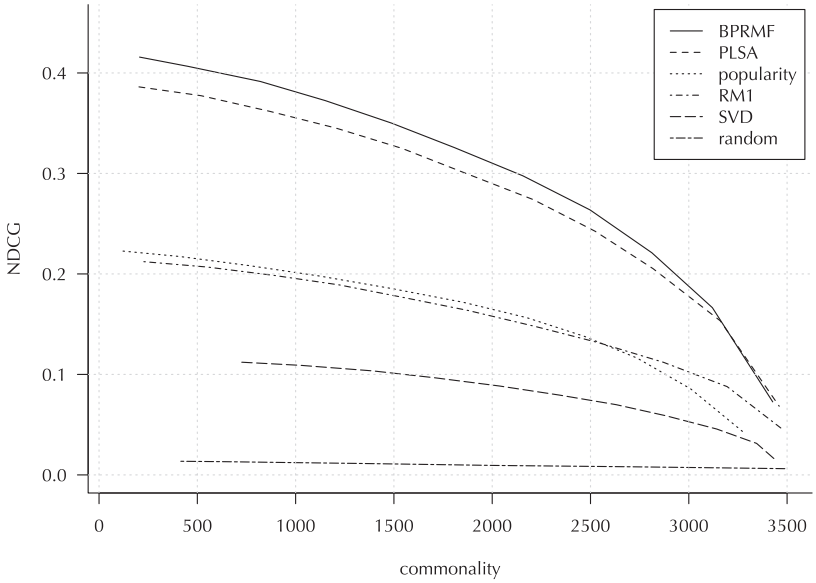


Fig. 6. Utility-commonality tradeoff using interleaved promotion. Commonality and NDCG of personalization-focused algorithms post-processed by interleaved promotion. Results for the movies domain.

in commonality between runs. In general, we find that Borda aggregation is necessary to compute a stable aggregation.

So far, we have discussed commonality with respect to personalization-focused models. Yet, we need to broadly consider the effects that increasing commonality could have on accuracy metrics. As an example, in this section, we apply a simple mitigation strategy implemented as a post-processing of the output of these personalization algorithms so we can focus on understanding commonality rather than on the Pareto-optimal algorithm.

While measurement provides us with a means to assess and compare the commonality of different algorithms, we need to also understand potential tradeoffs when we intervene to mitigate commonality in an existing system. Toward this end, we can apply a simple mitigation strategy implemented as a post-processing of the output of these personalization algorithms. Specifically, we developed a simple interleaved promotion algorithm that boosts in-category items within existing personalization-focused recommendations. For each user, we order items in a category according to their positions in the personalized ranking and construct a promoted-content ranking by selecting items from each category round robin. We select the top-ranked item in the final interleaved list by sampling an item from either the top of the original personalized ranking with probability p or the top of the combined promoted content ranking with probability $1 - p$. We remove the selected item from its source list. We do the same for the second-ranked item and continue this procedure until we have completed the ranking. A high value of our hyperparameter p will recover the original ranking; a low value of p will return the combined promoted content ranking; values in between will be a combination of the two. We present a detailed description of the algorithm in Appendix A and results in Figure 6. We observe that interleaving allows us to increase the commonality while smoothly degrading utility. In most cases, tradeoff Pareto curves dominate each other, indicating that relative utility performance can be largely maintained across commonality targets. That said, some runs with lower baseline NDCG performance reverse order under interleaving. This indicates possible systematic under-exposure of content mitigated by interleaving.

7 CONCLUSION

In this work, motivated by defining metrics for recommendation of cultural content, we developed a method to measure alignment with principles of cultural citizenship that we adapted from the PSM literature. Our proposed commonality metric emphasizes shared familiarity, the simultaneous exposure of users to content from selected categories. This definition, captured by the joint probability of familiarity events, is worth exploring, we have suggested, for theoretical, normative, and pragmatic reasons.

In addition to commonality, we introduced a relatively simple model of familiarity based on recall. We believe there is opportunity to develop alternative models of familiarity that consider a user's previous experience with the category or other contextual information. However, the design of a familiarity model should be aligned with the concept of shared experience, meaning that, even if a user has engaged with content from a category in the past, *re-exposing* them may promote commonality at the risk of over-satiating users with niche interests, a topic of recent research [65].

Our results demonstrate that existing high-utility recommendation algorithms under-perform in terms of commonality. We believe that exploring the space of commonality-informed recommendation can produce algorithms that perform substantially better in terms of commonality while maintaining high utility.

By introducing earlier the idea of a recommender system as a sociotechnical assemblage, we point to how future research could attend more to other components of this assemblage, beyond the algorithm, that also bear on both commonality and shared exposure to cultural diversity, or their lack. Regarding cultural diversity, such components might include the catalogues of content on which the system draws, and the larger institutional configuration within which recommender systems are designed and operationalized. Our focus in this article on the importance of the commonality metric, then, should not be mistaken for the view that developing a new metric is in itself sufficient to advance and achieve the goals we have articulated: recommender systems in the public interest that can enhance cultural citizenship.

Moreover, given that—as affirmed by the 'values in design' debate—recommender system design can consciously start out from normative principles such as those set out in the first part of this article, with the ambition of modelling public interest interventions, then such a design will not conform to the usual reductions and abstractions that have until now structured mathematical formalization in recommender system design. Yet, one of the key challenges recognized by 'values in design' is how the translation of normative principles into formalizations will always, inevitably, entail some kind of reduction and abstraction from those principles. The question then becomes: Which reductions and abstractions can be tolerated while retaining and upholding core dimensions of normativity? This article attempts to engage with this key challenge rigorously, imaginatively, and innovatively. The second, empirical and experimental part could not possibly model all dimensions of the normative perspectives we set out in the first part; that would take several papers. We have attempted to model two of the most important principles: universality (commonality) and diversity (of source and content), in the service of progress towards a third—cultural citizenship. We suggest that this is in itself a satisfactory achievement, and we propose to follow up other aspects of the normative perspectives laid out in the first part in future papers.

Two aims for future work seem particularly fruitful. First, we are interested in how the commonality metric attuned to increasing diversity of shared cultural experience might enable us to track these processes over time as, potentially, they cumulatively affect a given population of users. This builds on our founding assumption that existing personalized recommender systems are having cumulative effects—effects that have not yet been identified and studied by the recommender

system community. In the same way, we assume that the commonality metric could also be tracked over time to identify the cumulative effects of the interventions we describe, making explicit the potential for cumulative changes in cultural exposure among the user population—and potentially bringing to light certain kinds of progressive cultural change.

Second, and more generally, our conviction is that, depending on the wider social and cultural goals of the research guiding recommender system design, our commonality metric could be extended to address other normative principles—to take an example of great significance to democratic societies (and as mentioned earlier), balance or impartiality in the recommendation of news and information.

APPENDIX

A INTERLEAVED PROMOTION ALGORITHM

ALGORITHM 1: Algorithm to Incorporate Promoted Content Into a Personalization-based Ranking

```
Function interleave User u, Float p
    allRecs = getRecommendation (u);
    List idealRec = new List(items);
    while idealRec.size() < 100 do
        List currentCategories = new List(categories);
        forall Category c in categories do
            if c not in currentCategories then
                itemC = allRecs.getNextItemByCategory(c);
                idealRec.push(itemC);
                catsItemC = getAllCategories(itemC);
                currentCategories.addAll(catsItemC);
            end
        end
    end
    forall Item i in idealRec do
        if random number > p: then
            nextItem = idealRec.pop();
        end
        else
            nexItem = allRecs.getNextItemNotAdded();
        end
        newRec.push(nextItem);
    end
    return newRec;
end
```

REFERENCES

[1] Alation. 2022. *Data Curation: Enabling Self-Service Analytics for Data-Driven Organizations (White Paper).* https://www.alation.com/resourcecenter/whitepapers/data-curation-enabling-self-service-analytics-for-data-driven-organizations/. Accessed 22-02-2024.

[2] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of the Web Conference*. Association for Computing Machinery, New York, NY, 2155–2165. DOI: https://doi.org/10.1145/3366423.3380281

[3] Mark Andrejevic. 2013. Public service media utilities: Rethinking search engines and social networking as public goods. *Media Int. Austral.* 146, 1 (2013), 123–132.

[4] David Balzer. 2014. *Curationism: How Curating Took Over the Art World and Everything Else*. Coach House Books.

[5] Andrew Barry and Georgina Born. 2013. Interdisciplinarity: Reconfigurations of the social and natural sciences. In *Interdisciplinarity*. Routledge, 1–56.

[6] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 19–31.

[7] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Soc. Forces* 98, 4 (12 2019), 1–3. DOI: https://doi.org/10.1093/sf/soz162

[8] Benjamin Fish and Luke Stark. 2021. Reflexive design for fairness and other human values in formal models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*. Association for Computing Machinery, New York, NY, 89–99. DOI: https://doi.org/10.1145/3461702.3462518

[9] Christina Boididou, Di Sheng, Felix J. Mercer Moss, and Alessandro Piscopo. 2021. Building public service recommenders: Logbook of a journey. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 538–540.

[10] Tiziano Bonini. 2017. The participatory turn in public service media. In *Public Service Media Renewal. Adaptation to Digital Network Challenges*, M. Głowacki and A. Jaskiernia (Eds.). Peter Lang, New York, NY, 101–115.

[11] Georgina Born. 2005. *Uncertain Vision: Birt, Dyke and the Reinvention of the BBC*. Vintage. Retrieved from https://books.google.es/books?id=LZNfNmeSqoYC

[12] Georgina Born. 2006. Digitising democracy. In *What Can Be Done?: Making the Media and Politics Better*, John Lloyd and Jean Seaton (Eds.). Blackwell, Oxford, 102–23.

[13] Georgina Born. 2010. The social and the aesthetic: For a post-Bourdieuian theory of cultural production. *Cultur. Sociol.* 4, 2 (2010), 171–208.

[14] Georgina Born. 2012. Mediating the public sphere: Digitisation, pluralism, and communicative democracy. *Beyond Habermas: Democ., Knowl. Pub. Sphere* (2012), 119–146.

[15] Georgina Born. 2018. Principles of public service for the 21st century. In *A Future for Public Service Television*, Des Freedman and Vana Goblot (Eds.). The MIT Press. DOI: https://doi.org/10.7551/mitpress/9781906897710.003.0015

[16] Georgina Born. 2018. Taking the principles of public service media into the digital ecology. In *A Future for Public Service Television*, Des Freedman and Vana Goblot (Eds.). The MIT Press, 181–190. DOI: https://doi.org/10.7551/mitpress/9781906897710.003.0025

[17] Georgina Born, Jeremy Morris, Fernando Diaz, and Ashton Anderson. 2021. Artificial intelligence, music recommendation, and the curation of culture. *Schwartz Reisman Institute for Technology and Society White Paper*. https://tspace.library.utoronto.ca/bitstream/1807/129105/1/Born-Morris-etal-AI_Music_Recommendation_Culture.pdf. Accessed 22-02-2024.

[18] Georgina Born and Tony Prosser. 2001. Culture and consumerism: Citizenship, public service broadcasting and the BBC's fair trading obligations. *Mod. Law Rev.* 64, 5 (2001), 657–687. Retrieved from http://www.jstor.org/stable/1097275

[19] Pierre Bourdieu. 1971. Systems of education and systems of thought. In *Knowledge and Control*, Michael F. D. Young (Ed.). Collier Macmillan, New York, NY, 189–207.

[20] Robin D. Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. 2016. Towards multi-stakeholder utility evaluation of recommender systems. *UMAP (Extend. Proc.)* 750 (2016).

[21] Minerva Campos-Rabadán. 2020. Inequalities within the international film arena. A framework for studying Latin American film festivals. *Comunicación y Medios* 29, 42 (2020), 70–82. DOI: https://doi.org/10.5354/0719-1529.2020.57224

[22] Rocío Cañamares, Pablo Castells, and Alistair Moffat. 2020. Offline evaluation options for recommender systems. *Inf. Retr. J.* 23 (2020), 387–410.

[23] Roberto Suárez Candel. 2012. *Adapting Public Service to the Multiplatform Scenario: Challenges, Opportunities and Risks; Final Report of the Project "Redefining and Repositioning Public Service Broadcasting in the Digital and Multiplatform Scenarion; Agents and Strategies."* Hans-Bredow-Inst. für Medienforschung an der Univ. Hamburg, Verlag.

[24] Ben Carterette. 2011. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*. ACM, New York, NY, 903–912.

[25] Pablo Castells and Alistair Moffat. 2022. Offline recommender system evaluation: Challenges and new directions. *AI Mag.* 43, 2 (2022), 225–238.

[26] Anita Say Chan. 2022. *Predatory Data: Eugenics in Big Tech and Our Fight for an Independent Future*. University of California Press, Oakland, CA.

[27] Praveen Chandar, Fernando Diaz, and Brian St. Thomas. 2020. Beyond accuracy: Grounding evaluation metrics for human-machine learning systems. *Adv. Neural Inf. Process. Syst.* https://nips.cc/virtual/2020/tutorial/16652

[28] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18)*. Association for Computing Machinery, New York, NY, 224–232.

[29] Gerard Delanty. 2002. Two conceptions of cultural citizenship: A review of recent literature on culture and citizenship. *Global Rev. Ethnopolit.* 1, 3 (2002), 60–66.
[30] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in recommender systems: Research landscape and future directions. *User Model. User-adapt. Interact.*, Springer, 1–50.
[31] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in AI design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO'23)*. Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3617694.3623261
[32] Boguslawa Dobek-Ostrowska, Michał Głowacki, Karol Jakubowicz, and Miklos Sükösd. 2010. *Comparative Media Systems: European and Global Perspectives*. Central European University Press.
[33] Karen Donders. 2011. *Public Service Media and Policy in Europe*. Springer.
[34] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in recommender systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, 679–707. DOI: https://doi.org/10.1007/978-1-0716-2197-4_18
[35] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 172–186.
[36] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18)*. Association for Computing Machinery, New York, NY, 242–250. DOI: https://doi.org/10.1145/3240323.3240373
[37] Avriel Epps-Darling, Henriette Cramer, and Romain Takeo Bouyer. 2020. Artist gender representation in music streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR'20)*. 248–254.
[38] Adalbert Evers and Anne-Marie Guillemard. 2012. *Social Policy and Citizenship: The Changing Landscape*. Oxford University Press.
[39] Andres Ferraro, Gustavo Ferreira, Diaz Fernando, and Georgina Born. 2022. Measuring commonality in recommendation of cultural content: Recommender systems to enhance cultural citizenship. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys'22)*. Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3523227.3551476
[40] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the Conference on Human Information Interaction and Retrieval*. 249–254.
[41] Benjamin Fields, Rhia Jones, and Tim Cowlishaw. 2012. The case for public service recommender algorithms. In *Proceedings of the FATREC Workshop*.
[42] Mary Flanagan, Daniel C. Howe, and Helen Nissenbaum. 2008. *Embodying Values in Technology: Theory and Practice*. Cambridge University Press, 322–353. DOI: https://doi.org/10.1017/CBO9780511498725.017
[43] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.
[44] Christian Fuchs. 2012. Class and exploitation on the Internet. In *Digital Labor*, Routledge, 211–224.
[45] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. The winner takes it all: Geographic imbalance and provider (Un)fairness in educational recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'21)*. Association for Computing Machinery, New York, NY, 1808–1812. DOI: https://doi.org/10.1145/3404835.3463235
[46] Stuart Hall. 1993. *Which Public, Whose Service?* British Film Institute.
[47] Jungkyu Han and Hayato Yamana. 2017. A survey on recommendation methods beyond accuracy. *IEICE Trans. Inf. Syst.* 100-D (2017), 2931–2944.
[48] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (2015), 1–19.
[49] Natali Helberger. 2019. On the democratic role of news recommenders. *Digit. Journal.* 7, 8 (2019), 993–1012.
[50] David Hendy. 2013. *Public Service Broadcasting*. Bloomsbury Publishing.
[51] Alfred Hermida and Amanda Ash. 2010. Wikifying the CBC: Reimagining the remit of public service media. In *International Symposium on Online Journalism*, University of Texas, Austin. Citeseer.
[52] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. 1994. Decision combination in multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 16, 1 (1994), 66–75.
[53] Anna Lauren Hoffmann. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Inf., Commun. Societ.* 22, 7 (2019), 900–915.

[54] Karol Jakubowicz. 2007. Public service broadcasting in the 21st century. In *From Public Service Broadcasting to Public Service Media*. RIPE@ 2007, Nordicom, 29–49.

[55] Karol Jakubowicz. 2010. PSB 3.0: Reinventing European PSB. In *Reinventing Public Service Communication*. Springer, 9–22.

[56] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[57] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of individual traits on diversity-aware music recommender user interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP'18)*. Association for Computing Machinery, New York, NY, 291–299. DOI: https://doi.org/10.1145/3209219.3209225

[58] Elliot Jones. 2022. *Inform, Educate, Entertain...and Recommend?* Technical Report. Ada Lovelace Institute.

[59] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. Association for Computing Machinery, New York, NY, 3819–3828.

[60] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael D. Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the Web Conference*. 1065–1075.

[61] Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. 2016. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 11–14.

[62] Cory Knobel and Geoffrey C. Bowker. 2011. Values in design. *Commun. ACM* 54, 7 (2011), 26–28.

[63] Dominik Kowald, Peter Muellner, Eva Zangerle, Christine Bauer, Markus Schedl, and Elisabeth Lex. 2021. Support the underground: Characteristics of beyond-mainstream music listeners. *EPJ Data Sci.* 10, 1 (2021), 14.

[64] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Proceedings of the European Conference on Information Retrieval*. Springer, 35–42.

[65] Liu Leqi, Fatma Kilinc-Karzan, Zachary C. Lipton, and Alan L. Montgomery. 2021. Rebounding bandits for modeling satiation effects. *Advances in Neural Information Processing Systems* 34 (2021), 4003–4014.

[66] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing item popularity bias of music recommender systems: Are different genders equally affected? In *Proceedings of the 15th ACM Conference on Recommender Systems*. 601–606.

[67] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. 2019. Crank up the volume: Preference bias amplification in collaborative recommendation. In *Proceedings of the 1st Workshop on Recommendation in Multi-stakeholder Environments*.

[68] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J. Vera Liao, James Choi, Kaiyue Fan, Sean A. Munson, and Alexis Hiniker. 2021. How the design of YouTube influences user sense of agency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'21)*. Association for Computing Machinery, New York, NY.

[69] Malcolm Slaney and William White. 2006. Measuring playlist diversity for recommendation systems. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia*. 77–82.

[70] Momin M. Malik. 2020. A hierarchy of limitations in machine learning. *arXiv preprint arXiv:2002.05193* (2020).

[71] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2145–2148.

[72] Thomas H. Marshall. 1950. *Citizenship and Social Class*. Cambridge.

[73] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. In *Proceedings of the 27th ACM Conference on Information and Knowledge Management*.

[74] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2021. Ethical aspects of multi-stakeholder recommendation systems. *Inf. Societ.* 37, 1 (2021), 35–45.

[75] Toby Miller. 2001. Introducing... cultural citizenship. *Soc. Text* 19, 4 (2001), 1–5.

[76] Hallvard Moe. 2008. Dissemination and dialogue in the public sphere: A case for public service media online. *Media, Cult. Societ.* 30, 3 (2008), 319–336.

[77] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (Dec. 2008), 2:1–2:27.

[78] Graham Murdock. 2005. Building the digital commons. In *Cultural Dilemmas in Public Service Broadcasting*. Nordicom Goteborg, Sweden, 213–231.

[79] Safiya Umoja Noble. 2018. *Algorithms of Oppression*. New York University Press, New York. DOI: https://doi.org/doi:10.18574/nyu/9781479833641.001.0001

[80] Martha C. Nussbaum. 2003. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press.

[81] Hans Ulrich Obrist, Lionel Bovier, and Birte Theiler. 2008. *A Brief History of Curating*. JRP/Ringier Zurique.

[82] Òscar Celma and Pedro Cano. 2008. From hits to niches? Or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-scale Recommender Systems and the Netflix Prize Competition*. 1–8.
[83] Jan Pakulski. 1997. Cultural citizenship. *Citizen. Stud.* 1, 1 (1997), 73–86.
[84] Bhikhu Parekh et al. 2000. *The Parekh Report: The Future of Multi-Ethnic Britain*. Birminham Education Authority Publication, Birmingham.
[85] Marc Raboy. 1996. *Public Broadcasting for the 21st Century*. Vol. 17. Indiana University Press.
[86] Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021. No NLP task should be an island: Multi-disciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics, 45–55.
[87] Stephen Robertson. 2006. On GMAP: And other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. Association for Computing Machinery, New York, NY, 78–83.
[88] Renato Rosaldo. 1994. Cultural citizenship in San Jose, California. *PoLAR* 17 (1994), 57.
[89] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio—Rank-aware divergence metrics to measure normative diversity in news recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys'22)*. Association for Computing Machinery, New York, NY, 208–219. DOI: https://doi.org/10.1145/3523227.3546780
[90] Maria Sapignoli. 2021. Anthropology and the AI-turn in global governance. *Am. J. Internat. Law* 115 (2021), 294–298.
[91] Shrikant Saxena and Shweta Jain. 2021. *Exploring and Mitigating Gender Bias in Recommender Systems with Explicit Feedback*. CoRR abs/2112.02530 (2021). arXiv:2112.02530. https://arxiv.org/abs/2112.02530
[92] Paddy Scannell. 1989. Public service broadcasting and modern public life. *Media, Cult. Societ.* 11, 2 (1989), 135–166.
[93] Paddy Scannell and David Cardiff. 1991. *A Social History of British Broadcasting: 1922–1939: Serving the Nation*. Basil Blackwell.
[94] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis. In *Proceedings of the Conference on Human Information Interaction and Retrieval (CHIIR'22)*. Association for Computing Machinery, New York, NY, 337–341. DOI: https://doi.org/10.1145/3498366.3505791
[95] Markus Schedl and David Hauger. 2015. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. Association for Computing Machinery, New York, NY, 947–950. DOI: https://doi.org/10.1145/2766462.2767763
[96] Jean Seaton. 2021. The BBC: Guardian of public understanding. In *Guardians of Public Value*. Palgrave Macmillan, Cham, 87–110.
[97] Nick Seaver. 2018. What should an anthropology of algorithms do? *Cultur. Anthropol.* 33, 3 (2018), 375–385.
[98] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2020. Exploring artist gender bias in music recommendation. In *Proceedings of the 2nd Workshop on the Impact of Recommender Systems*.
[99] Nasim Sonboli, Robin Burke, Michael D. Ekstrand, and Rishabh Mehrotra. 2022. The multisided complexity of fairness in recommender systems. *AI Mag.* 43, 2 (2022), 164–176. DOI: https://doi.org/10.1002/aaai.12054
[100] Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Soc. Stud. Sci.* 48, 2 (2018), 204–231.
[101] Jesper Strömbäck. 2005. In search of a standard: Four models of democracy and their normative implications for journalism. *Journal. Stud.* 6, 3 (2005), 331–345.
[102] Trine Syvertsen, Karen Donders, Gunn Enli, and Tim Raats. 2019. Media disruption and the public interest. *Nordic J. Media Stud.* 1, 1 (2019), 11–28.
[103] Georgios Theocharous, Philip S. Thomas, and Mohammad Ghavamzadeh. 2015. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 1806–1812.
[104] Tamas Tofalvy and Júlia Koltai. 2023. "Splendid Isolation": The reproduction of music industry inequalities in Spotify's recommendation system. *New Media & Society* 25, 7 (2023), 1580–1604.
[105] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An audit of misinformation filter bubbles on YouTube: Bubble bursting and recent behavior changes. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 1–11.
[106] Klaus Unterberger and Christian Fuchs. 2021. *The Public Service Media and Public Service Internet Manifesto*. University of Westminster Press.
[107] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18)*. Association for Computing Machinery, New York, NY, 260–268. DOI: https://doi.org/10.1145/3240323.3240347

[108] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2020. Assessing ranking metrics in top-N recommendation. *Inf. Retriev. J.* 23, 4 (2020), 411–448.

[109] Hilde Van den Bulck. 2008. Can PSB stake its claim in a media world of digital convergence? The case of the Flemish PSB management contract renewal from an international perspective. *Convergence* 14, 3 (2008), 335–349.

[110] Hilde Van den Bulck and Hallvard Moe. 2018. Public service media, universality and personalisation through algorithms: Mapping strategies and exploring dilemmas. *Media, Cult. Societ.* 40, 6 (2018), 875–892.

[111] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11)*. Association for Computing Machinery, New York, NY, 109–116. DOI: https://doi.org/10.1145/2043932.2043955

[112] Deb Verhoeven, Bronwyn Coate, and Vejune Zemaityte. 2019. Re-distributing gender in the global film industry: Beyond #MeToo and #MeThree. *Media Industr. J.* 6, 1 (2019).

[113] Michael Matthias Voit and Heiko Paulheim. 2021. Bias in knowledge graphs—An empirical study with movie recommendation and different language editions of DBpedia. In *Proceedings of the Language, Data, and Knowledge Conference*.

[114] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22)*. Association for Computing Machinery, New York, NY, 1251–1261.

[115] Sarah Myers West. 2020. Redistribution and rekognition: A feminist critique of algorithmic fairness. *Catalyst: Femin., Theor., Technosci.* 6, 2 (2020).

[116] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. 2003. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR'03)*. ACM Press, New York, NY, 10–17.

Received 10 December 2022; revised 18 December 2023; accepted 3 January 2024