INFORMATION, COMMUNICATION & SOCIETY, 2018
VOL. 21, NO. 7, 959–977
https://doi.org/10.1080/1369118X.2018.1444076

Routledge
Taylor & Francis Group

Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity

Judith Möller[a], Damian Trilling[a], Natali Helberger[b] and Bram van Es[c]

[a]Department of Communication Science, University of Amsterdam, Amsterdam, Netherlands; [b]Institute for Information Law, University of Amsterdam, Amsterdam, Netherlands; [c]eScience Center, Amsterdam, Netherlands

ABSTRACT

In the debate about filter bubbles caused by algorithmic news recommendation, the conceptualization of the two core concepts in this debate, diversity and algorithms, has received little attention in social scientific research. This paper examines the effect of multiple recommender systems on different diversity dimensions. To this end, it maps different values that diversity can serve, and a respective set of criteria that characterizes a diverse information offer in this particular conception of diversity. We make use of a data set of simulated article recommendations based on actual content of one of the major Dutch broadsheet newspapers and its users (N=21,973 articles, N=500 users). We find that all of the recommendation logics under study proved to lead to a rather diverse set of recommendations that are on par with human editors and that basing recommendations on user histories can substantially increase topic diversity within a recommendation set.

ARTICLE HISTORY
Received 30 October 2017
Accepted 20 February 2018

KEYWORDS
News; recommender systems; diversity metrics; filter bubbles; automated content classification

1. Introduction

Functioning democracies rely on the media's role to enable different groups and interests in society to participate in the democratic debate and create the opportunity for citizens to encounter diverse opinions (Owen & Smith, 2015). Therefore, diversity is one of the core public values in media law and policy. As the Council of Europe (2007) has put it, 'media pluralism and diversity of media content are essential for the functioning of a democratic society' and the right to freedom of expression 'will be fully satisfied only if each person is given the possibility to form his or her own opinion from diverse sources of information'. The close link between media diversity and democratic participation may also explain the vigor with which new technological innovations in the media landscape are being met. The impact of personalized recommendations and algorithmic filtering on users' information diet is subject to dystopian visions about filter bubbles and echo chambers. But how much

CONTACT Judith Möller j.e.moller1@uva.nl

reason do we actually have to be concerned about the diversity of algorithmic recommendations?

The aim of the present study is to better understand the effect of different recommender systems on diversity. We will do so in the form of a data-scientific experiment, in which we compare simulated recommendation sets with regard to their performance on several indicators of diversity to human editors.

Scholars of selective exposure have often argued that once given enough choice, users voluntarily lock themselves up in so-called 'echo chambers', in which they are only exposed to like-minded content (e.g. Stroud, 2011). Recently, this line of reasoning has been extended: Pariser (2011) argues that algorithmic selections contribute to a situation in which users are essentially exposed to a self-confirming feedback loop. He described the resulting spaces with like-minded information as 'filter bubbles'. That has raised concerns that societies will become increasingly polarized and fragmented. However, given that the diversity of news selected by human editors is also constrained on the individual, organizational, and structural level (Shoemaker et al., 2001) and that users play a very active role users in selecting and processing news, this line of reasoning attributes more power to algorithmic news recommendation services than they probably deserve (Zuiderveen Borgesius et al., 2016). First empirical studies indicate that, at present, algorithmic recommendation systems have not locked large segments of the audience into bubbles (Fletcher & Nielsen, 2017).

Moreover, an important lacuna in the current debate about potential filter bubbles is the lack of attention for two core concepts, *diversity* and *algorithmic selection*. So far, diversity is mainly discussed and measured in terms of political leaning of news sources and stories (e.g. Flaxman & Rao, 2016), thereby neglecting that diversity is a multi-dimensional concept that also encompasses topic plurality, genre, or the plurality in tone (Helberger et al., 2018). It is also often assumed that the appropriate level of analysis is the bubble of content created by one outlet. There has been little research so far that would translate the very relevant body of research on exposure diversity (Napoli, 2015) into measures that can be used to assess the performance of algorithmic recommender systems.

This study contributes to the current body of work on the effect of algorithmic news personalization on exposure diversity in three ways. We start out with proposing a first set of concrete indicators that can help us to assess the diversity of different recommendation algorithms. Second, we study the consequences of different algorithmic design choices in depth. In the current debate, news recommendation algorithms are often treated as a single entity and considered to be an immutable, external force. That is, however, a misrepresentation of reality. While it may be true that some or even many of the current recommender algorithms are optimized predominantly on short-term metrics such as clicks, a whole range of recommendation algorithms are available and used. In order to understand the effects of a recommendation algorithm on individual's news diet, it is important to understand that potential effects of recommendation algorithms are dependent on the exact design of the recommendation algorithm in use (see also Kitchin, 2017). The third and main contribution is the development of a new and innovative method we use to measure diversity in news recommendation sets. It goes beyond existing methods that assess diversity by classifying content as either falling in the same category (not diverse) or not falling in the same category (diverse). By taking into account that many of the categories are related to one another, we can measure the complexity of diversity

in news more accurately. For example, a story about international relations has arguably a different topic than a story about national politics, yet, compared to a story about biology they are quite similar. In order to fully assess how diverse recommendation sets are we need to include this notion of *relative* distance to our model. Finally, we use two benchmarks to evaluate the performance of the recommender systems against in terms of diversity. The first is empirical and represents the selection of news articles by human editors. The second is statistical and assumes a completely random selection over all articles.

2. Dissecting diversity

As the phenomenon of algorithmic news recommendation is still recent and constantly in development, there are few empirical studies on potentially lacking diversity of news recommendation systems in the field of social science (for an overview, see Dylko et al., 2017). The first set of studies (Dylko et al., 2017; Beam, 2014; O'Callaghan et al., 2015; Bakshy et al., 2015) have coherently provided evidence that unobtrusive and automatic selection of content can reduce exposure diversity in the context of news. However, whether that is the case or not depends on the exact settings of the algorithmic recommendation system and the data that is used to determine the recommendations as research in the field of informatics and data science has demonstrated. In fact, algorithmic design can also increase diversity (Smyth & McClave, 2001). Moreover, studies combining insights form the field of computer science with psychology have repeatedly shown that diversity in recommendation sets increases user satisfaction (Knijnenburg et al., 2012; Ziegler et al., 2005; Willemsen et al., 2016). According to this strand of literature, diverse recommendations have a distinct function in the selection decision process. They support the user in the exploratory stage of interaction with the recommender system to identify relevant items quicker (see also Bridge & Kelly, 2006).

The different results regarding diverse outcomes of news recommendation systems in social science and computer science are partially a consequence of different conceptualizations of diversity. Whereas diversity in the social scientific debate is often conceptualized as inclusion of counter-attitudinal information (Pariser, 2011), in computer science it is considered a necessary design element, often conceptualized as the inclusion of unexpected items. Diversity is mostly modeled and conceptualized at the input level in computer science, whereas social scientists analyze diversity primarily at the output level: Whereas the former study how to optimize the inclusion of surprising or unexpected items, social scientists focus on the overall output of a recommender system and assess its diversity (Sørensen & Schmidt, 2016).

2.1. Algorithmic design

The task of algorithms in the context of news recommendation systems is to structure and order the pool of available news according to predetermined principles. Yet, algorithms are created by humans and their design choices regarding the sorting and ranking of items influence the output (Bozdag, 2013). Three common approaches are:

(1) General popularity of the item among all users. This is the most simple approach, as all users get the same recommendation, leading to a Matthew effect, in which popular items become even more popular and unpopular items disappear.

(2) Semantic filtering. This approach recommends items that match the currently used item or items previously used by the same user on a number of pre-defined criteria (features). In the case of news articles, these features could be words or topics, but also the author or source of an article.
(3) Collaborative filtering or social information filtering. This process 'automates the process of 'word-of-mouth' recommendations: items are recommended to a user based upon values assigned by other people with similar taste' (Bozdag, 2013).

These methods are usually applied in hybrid forms, also including other methods like weighing items by recency or pushing content that has specific features (e.g. paid content). Of course, one can conceive of other structuring principles as well: A recommender system might inject an 'editor's pick' or even some random articles. A simple algorithm might offer the same results for each reader of a given article (item-to-item recommendation). More sophisticated algorithms do not only combine several of the logics mentioned above, but also take into account past user behavior and choices. That means that instead of finding the best match merely based on the article that is currently displayed, these algorithms select articles that resemble the spectrum of all articles that user has read before. When implemented properly, this can lead to a truly personalized news recommendation, implying that every reader of an article could be recommended a unique set of news items (Ricci et al., 2011). This personal profile based on past behavior can be additionally enriched by known user data such as age or location.

2.2. Diversity in algorithmic design

In addition to these logics, many recommendation systems also include an element of surprise: serendipity. Some authors even state that serendipity is a crucial element of all algorithmic recommendation systems (Kotkov et al., 2016). Serendipitous recommenders do not select all items they present to users according to the principles presented above. Some items are included at random. This serves two purposes: First, the chance that users lose interest because the choice set is too uniform decreases. Second, these items are needed for algorithms to learn and improve themselves. Consider the example of a user starting to use a recommendation system during a soccer world championship. Chances are high that this person will read a couple of sports articles in a row – maybe even nothing else. If an aggressively personalized recommendation system would now recommend *only* sports articles in the upcoming week, the user would not only get frustrated, but the system would have no chance of learning that the user is actually much more interested in economics than in sport. Serendipity is also needed to integrate *new* items into a recommendation system: For new articles, information about who read it is simply not available.

In computer science, there is no consensus on how serendipity should be conceptualized and operationalized, other than unexpectedness (Kotkov et al., 2016). Serendipity is not fully synonymous with diversity in this strand of literature. While both serendipity and diversity describe the relationship between user and item, diversity specifically describes the differences in content. It is usually measured as an overlap between content features (see Kunaver & Požrl, 2017). Many users appreciate diversity in the realm of news, but a key challenge is finding a balance between accuracy of the recommendation and diversity of the recommendation set (see, e.g. Ribeiro et al., 2015).

2.3. Diversity in the social sciences

In the social-scientific debate about the effects of algorithmic news recommendation on diversity, diversity is often conceptualized as ideologically cross-cutting exposure. For instance, Dylko (2016) examined the effects of user-driven versus system-driven customizable recommendation technology, and found that exposure to system-driven recommendation technology can result in political selective exposure, especially when combined with ideology-based customizability. They regard diversity as a form of selective exposure, as defined as 'proportionally high levels of consumption of information that supports one's political beliefs or attitudes, and proportionally low levels of consumption of information that is counter to one's political beliefs or attitudes' (p. 393). Beam (2014) measured diversity in terms of exposure to headlines from counter-attitudinal sources. O'Callaghan et al. (2015) showed that the YouTube recommendation system is likely to recommend extreme right-wing content if a user has just watched an extreme right-wing video. A study into Facebook's algorithmically ranked news-feed again measured diversity in the sense of 'exposure to ideologically cross cutting content' (Bakshy et al., 2015). In contrast, Haim et al. (2017) focused on source and content diversity, whereby they measured content diversity in terms of categories and source diversity in terms of percentage of articles per news outlet.

In short, most studies measure exposure diversity in the context of news personalization in terms of exposure to counter-attitudinal or counter-ideological content. This is relatively straightforward, but also rather limited. While this is a fruitful approach to study polarization in bipartisan regimes, it neglects other facets of diversity that are instrumental to especially multi-party systems (see also Benson, 2009).

In reality, news content diversity is a far richer, and far more complex phenomenon (see, e.g. Helberger et al., 2018; Sørensen & Schmidt, 2016; Masini et al., 2017; Napoli, 2015). As Van Cuilenburg (1999) puts it, diversity defines the 'heterogeneity of media content in terms of one or more specified characteristics' (p. 188). In the context of media policy, diversity is often defined and measured as diversity of sources in a media market (e.g. The Media Pluralism Monitor in Europe or the FTC's controversial Diversity Index; Napoli, 2015; Karppinen, 2013). However, for the purpose of measuring the diversity of personalized recommendations, source-centered approaches to measuring media diversity are only of little help, particularly in the European context where press self-regulation emphasizes *internal* diversity of (quality) news media. They can already been criticized on a more theoretical, normative basis. Duncan & Reid (2013) warn that they 'emphasize diversity for diversity's sake, leading to a postmodern, relativistic "anything goes" approach to public discourse, where society's most pressing concerns are drowned in a cacophony of different voices' (p. 486). More practically, most of the existing approaches look at media markets in their entirety (external diversity) (Karppinen, 2015), and less at the diversity within a particular outlet (internal diversity). Obviously, assessing the diversity of the results of personalized recommendations requires an assessment at the (internal) level of the recommendation, or respectively the output of a personalized recommendation[1].

2.4. Diversity as a function of news in democracy

But what is then a diverse recommendation set? The answer depends on what information users need in order to be able to fulfill their role in a specific model of democracy (e.g. Strömbäck, 2005; Ferree et al., 2002): Different theories of democracy (e.g. procedural, competitive, participatory, or deliberative) imply different communication needs (e.g. marketplace of ideas, creation of understanding and tolerance, or opportunity to explore news autonomously). Ideally, these different conceptualizations should be taken into account before making any claims about the impact of algorithmic filtering on the state of the democratic debate. Hence, there can be different types of diversity standing next to each other. Some maybe more suitable for personalized recommendations in the public service media, others in the commercial media, local or regional outlets, etc.

One essential characteristic of algorithmic recommendations is a user interest-driven approach: based on the different *interests* of users that an algorithmic recommender infers, it makes recommendations. This user interest-driven and topics-centered approach also reflects the move towards disintegration and decentralization of news, which is not at least a result of new, algorithmically driven media platforms, be that media services such as *Upday* or *Flipboard*, or more general purpose platforms, such as social networks and search engines. They might serve people 'more of the same' (e.g. Sunstein, 2007) and fail to represent the diversity of topics in society. This is why one dimension of the diversity that we will look at is the diversity of topics.

In a participatory model of democracy, in which the task of the media is to inform citizens in a way that allows them to actively participate in politics, exposure diversity also depends on the ratio of politically relevant content. A democratically responsible recommender from this perspective would ensure that important political topics for the public agenda reach the user. In contrast, from the perspective of a more procedural concept of exposure diversity, the ability to participate actively in politics is less prominent, whereas the importance of fundamental rights, such as freedom of expression, privacy and informational self-determination come to the fore. From this perspective, the outcome of diverse recommendation logics can be compared to what extent we can observe differences between personalized and non-personalized recommendation settings, or, in other words, how well the recommender performs in individualizing the information offer based on the signals from the user.

From a deliberative democracy point of view, not only the topics represented, but also the tone in which they are presented matters. Since the ultimate goal is (rational and fair) deliberation, this should also be reflected in the tone in which topics are represented: measured, informative, reconcillary, and non-aggressive. In contrast, under a more participative conception of exposure diversity, where diversity serves as a means to inform, enable and activate citizens to play a more active role in society, arguably, the tone can at times also veer into the direction of activist, engaging, critical and even partial and one-sided. Critical can also be the tone in a more competitive conceptualization of exposure diversity, where citizens are expected to be informed about current and future societal problems and who is the best able to solve them; but less than in a participatory model, those do not need to be sweeping, activating statements. A matter-of-fact to critical tone in describing the issues would seem more conductive to the goals under this particular conception of exposure diversity.

3. Method

We present a first attempt of an empirical assessment of recommender diversity based on the considerations outlined above. It should be noted that we are explicitly not researching viewpoint diversity. As we are using a most similar experimental design, we rely on a set of articles and recommendations from one newspaper, which limits the possibilities to study ideological diversity or diversity in style. The focus of this study is, as outlined above, topics (both in general, as the 'politicalness' of the topics). Taking these as examples, we show how to construct a multi-dimensional feature space that allows us to assess the diversity of recommendation sets. We then evaluate how different recommender systems perform on these criteria.

3.1. Research design

To study the influence of algorithmic design on diversity, we carry out a data-scientific experiment. Specifically, we compare the output of different algorithmic news recommendation for the same articles of the same news source (*de Volkskrant*, a Dutch progressive high-quality newspaper, that is oriented towards the political center, see Figure 1). Our dataset also includes recommendation sets for each of the articles that have been picked by human editors. This provides us with a unique benchmark to evaluate the performance of the recommender systems.

For N=1000 articles that were published between 19–9–2016 and 26–9–2016, we simulated which three articles out of the pool of N=21,973 articles would be recommended if one of the following logics were used:

(1) The choice of human editors
(2) The overall popularity

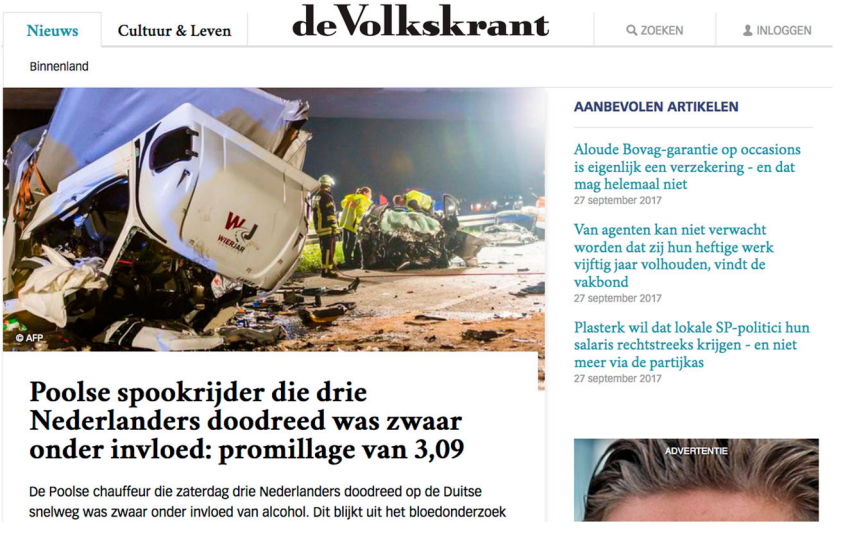

Figure 1. Recommendation box on the newspaper website. The recommendation box is displayed in the top right corner ('Aanbevolen artikelen').

(3) Collaborative filtering (item-to-item)
(4) Semantic filtering (item-to-item)
(5) Collaborative filtering (taking into account user histories)
(6) Semantic filtering (taking into account user histories)
(7) A random baseline, in which each article is equally likely to be recommended.

For the two algorithms that are based on user histories, we used the tracking history of 500 randomly selected news users.

3.2. Feature engineering

As the dataset contained rich information on each article (e.g. not only the full text, but also a category label and topic tags), we could create multiple sets of features.

3.2.1. Topic distance

We used the package Gensim (Řehůřek & Sojka, 2010) to estimate a topic model. After evaluating different models based on their perplexity and interpretability, we chose an LDA model (Blei et al., 2003) with 50 topics based on $tf \cdot idf$ representations of the documents, with additional filtering of extremely common and extremely uncommon words. We used pyLDAvis (see Sievert & Shirley, 2014) to perform multidimensional scaling on the resulting topics. As a result, each topic can be represented by its coordinates (x,y) in a two-dimensional space. As the Euclidian distance between two points is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, and as each document D is represented by a vector $\vec{w_D}$ of 50 topic weights $w_{D,1} \cdots w_{D,50}$, we can calculate the topic distance between two documents.[2]

3.2.2. Category and tag distance

Another more recent approach is the use of word embeddings to capture the aggregate inter-document difference, basically by considering the Wasserstein metric describing the minimum effort to make two documents similar. This is called the word mover distance (WMD). The great benefit of this method is that it is not sensitive to non-overlapping document vocabularies. We use a word vector model pre-trained on the NLCOW14 corpus (see Tulkens et al., 2016; Schäfer & Bildhauer, 2012), which we again apply using Gensim.

In addition, we also have human-labeled tags as proxies for the topics and we have the newspaper sections for each article as proxies for the categories. We again apply the word-distance derived from the word vector trained on the NLCOW14 corpus.

3.2.3. Ratio of politically relevant content

To estimate the topical distance from politics and related terms, we first construct a list of 20 words that are closest to 'politiek' (politics/political) according to our word vector model, plus a manual selection. This list of words is then compared with the words in the article text.

3.2.4. Tone distance

We estimated the polarity and subjectivity of the articles using the package Pattern (De Smedt & Daelemans, 2012). Pattern has been used before to estimate the sentiment of

Dutch-language newspaper coverage (Junqué de Fortuny et al., 2012). For each article, it returns a tuple with two values: the polarity, ranging from negative (−1) to positive (+1), and the subjectivity, ranging from a very objective tone (0) to a very subjective one (+1). Using these tuples as coordinates, we can situate each article in a two-dimensional sentiment space.

3.3. Diversity evaluation

Having obtained the pairwise distances, we will describe their distributions per recommender system.

Additionally, we can apply formal diversity measures such as Rao's quadratic entropy/diversity (see Ricotta & Szeidl, 2006). It was originally intended to reflect biodiversity and is given by the following equation

$$\begin{aligned} \mathcal{H}_{Rao,entropy} &= \sum\sum p_i p_j d_{i,j} \\ \mathcal{H}_{Rao,diversity} &= \sum_{j=1} p_j \sum_{i \neq j} d_{i,j} p_i \end{aligned} \tag{1}$$

where p_i,p_j represent the proportion of a species i or j respectively and $d_{i,j}$ is some distance metric. For the present paper, the species is a reflection of document categories and the distance metric is given by the similarity between the feature vectors of the document pairs.

We calculate Rao's quadratic entropy based on the main section titles, as they are readily available and unambiguous for the distance we use, which is the cosine similarity of the word vectors for these section titles. As we are interested in diversity, not from the perspective of maintainable diverse species but rather from the perspective of a diverse representation of available categories we replace the proportions p_i and p_j by $(1 - p_i)^n$ and $(1 - p_j)^n$ respectively to create slightly adapted diversity measures. This ensures that the inclusion of niche categories contribute the most to the final diversity. In the standard formulation categories with larger proportions contribute more to the diversity.

We additionally report the Shannon entropy,

$$H_{Shannon} = -\sum p_i \log_b(p_i), \tag{2}$$

in which the relative distance between the articles is not included.

It is important to note that both Rao's quadratic entropy and Shannon's entropy will give high indications of diversity even if small niche categories are severely underrepresented or even omitted.

4. Results

Given that for each document in the dataset we simulated three recommendations (see also Figure 1), the diversity of a recommender system can be conceptualized in different ways:

(1) on the level of the recommendation set, as the mean of the distances of each article towards the original article;

(2) on the level of the recommendation set, as the mean of the distances within the recommendation set;
(3) on the user level, as the diversity of the sets of all articles a user has ever been recommended;
(4) on the level of the individual recommendation.

While all of these conceptualizations can offer valuable insights, we chose to present the results of the last conceptualization, because it avoids averaging and aggregating. As we have N=1000 origin articles ×3 recommendations, all following calculations are therefore based on N=3000 'origin article'—'recommended article' pairs.

We start by looking at the descriptive statistics of the performance of the different recommender systems. As Table 1 shows, first and foremost, differences between all systems seem to be rather limited in size. Across different operationalizations and analyses, all of the recommendation logics under study proved to lead to a rather diverse set of recommendations.

We can identify some notable tendencies, though. First of all, as one would expect, the random recommender produces one of the most diverse recommendation sets. We also see that the recommendations produced by the editors are not particularly diverse. Especially if we look at the category, we see that editors tend to recommend articles from the same or very similar categories. The results with regard to topic diversity in the output of collaborative filtering are particularly interesting. If data on past user preferences are not taken into account, collaborative filtering produces the least amount of diversity. That means if the decision which articles to show is based on the selection of other readers of the same article, it is likely that articles of the same topic are presented. Yet, if users are matched on all articles they have read, we find collaborative recommendation systems are least likely to recommend a related topic. We can conclude that in terms of topic diversity, tailoring can support diverse exposure, at least in the short run.

However, the descriptive statistics in Table 1 might hide more subtle differences. For instance, it might be the case that a recommender sometimes outputs very diverse recommendations, and sometimes recommendations that are not diverse at all, which might cancel each other out. To get a better understanding of the differences, we therefore plotted the distribution of our metrics.

As Figure 2 shows, the higher average diversity for the random baseline recommender, but also the user-collaborative recommender, is mainly due to a much flatter distribution with a fatter tail: Even though the peaks of all recommenders are very close to each other,

Table 1. Means and standard deviations distances between topics, tone, category, tags, and from the word 'politiek'.

	Topic	Tone	Category	Tag	Politics
Collab. (item)	90.0 (55.1)	0.14 (0.15)	0.21 (0.23)	0.45 (0.11)	0.79 (0.02)
Semantic (item)	94.4 (56.9)	0.14 (0.14)	0.29 (0.24)	0.46 (0.10)	0.79 (0.02)
Popularity	95.6 (63.8)	0.13 (0.13)	0.00 (0.00)	0.45 (0.10)	0.80 (0.02)
Editors	91.3 (56.6)	0.14 (0.14)	0.10 (0.19)	0.43 (0.10)	0.80 (0.02)
Random	114.6 (68.1)	0.15 (0.15)	0.40 (0.18)	0.52 (0.08)	0.80 (0.02)
Collab. (user)	116.0 (67.0)	0.13 (0.12)	0.37 (0.18)	0.53 (0.07)	0.79 (0.02)
Semantic (user)	96.2 (58.7)	0.14 (0.14)	0.23 (0.24)	0.42 (0.10)	0.79 (0.02)

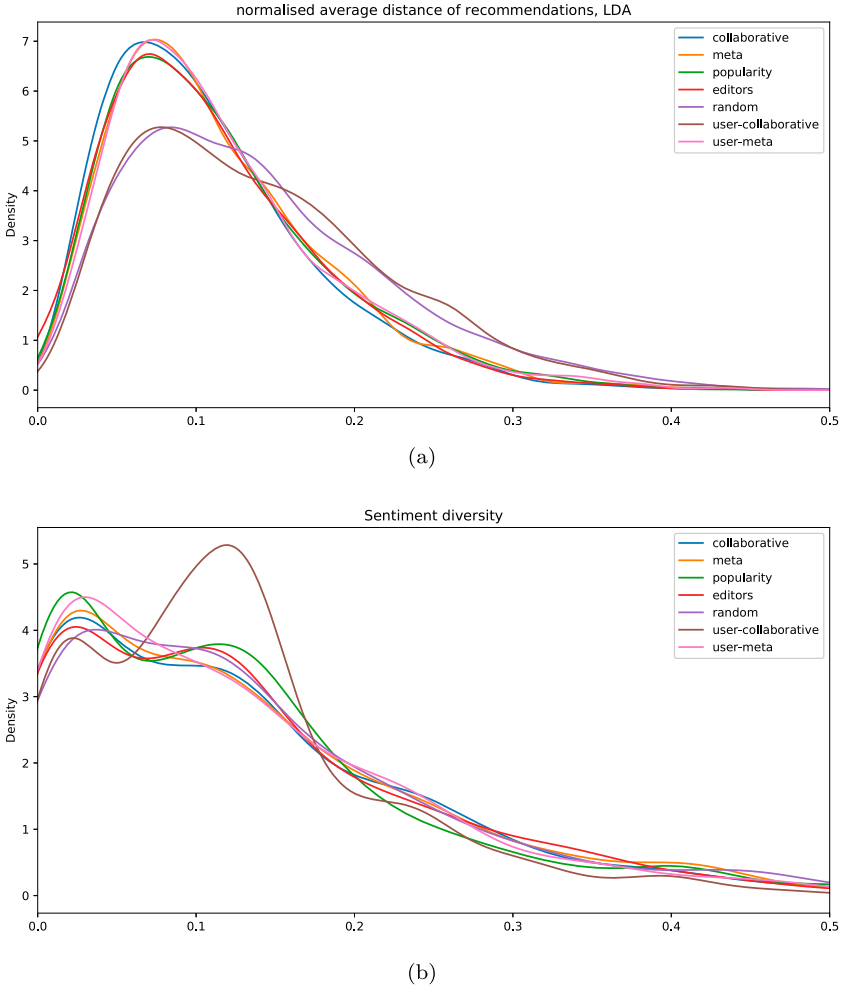

Figure 2. Topic and sentiment distance. Diversity score distribution of the recommender systems, based on the similarity of (a) LDA topic, and (b) sentiment of a recommended article compared to the currently viewed article.

the random and the user-collaborative recommender more often produce recommendations that are comparatively far off in terms of the LDA topic.

Regarding tone, in line with the descriptive statistics, we do not see many differences, except a hard to explain peak of the user-collaborative recommender.

The distribution of the categories and tags (Figure 3) is quite revealing. As the plot clearly shows human editors are more inclined to simply recommend articles with exactly the same category. Using the other extreme, a random recommender, this hardly happens. Obviously, these random recommendations can be too diverse, indicated by the pronounced tail on the right of the plots. If one aims at providing recommendations from similar, but not always the same categories, it therefore seems to be a good choice to use, for instance, any item-based recommender.

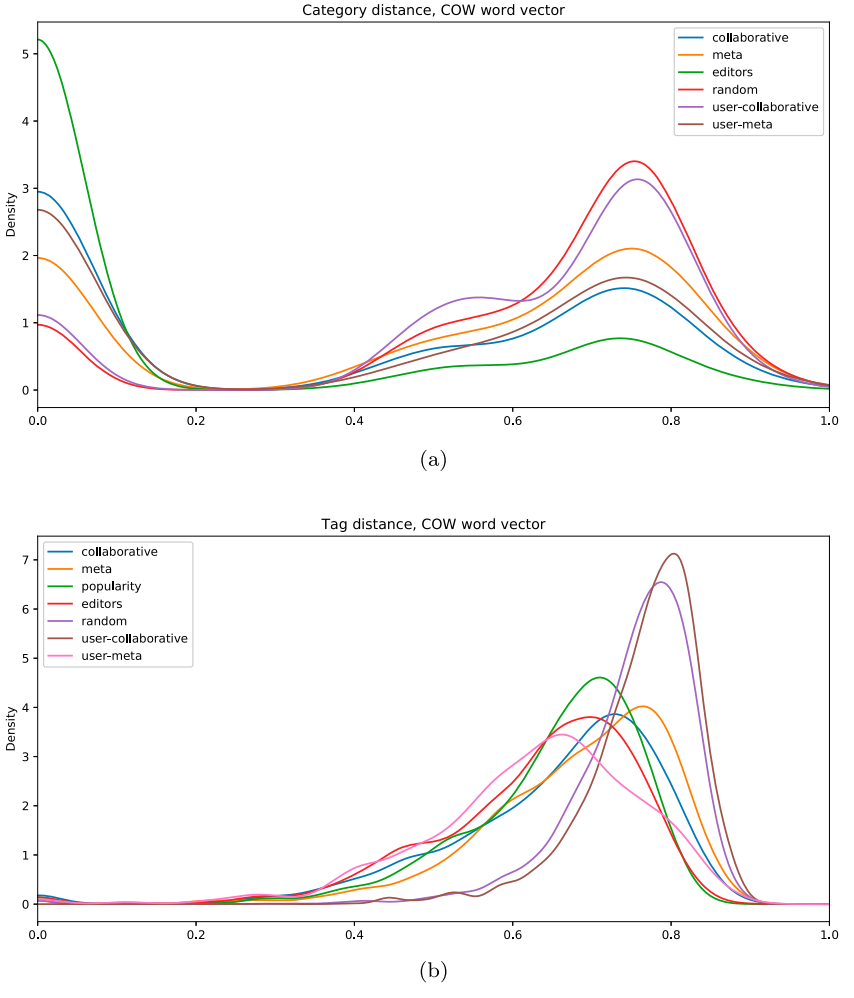

Figure 3. Category and tag distance. Diversity score distribution of the recommender systems, based on the similarity of (a) category,(b) tags, estimated using word vectors trained on the COW corpus.

Regarding the tags, which are provided by human editors to categorize topics, we see a familiar pattern: The random baseline recommender seems to frequently produce recommendations with very different tags (which is what we would expect), but so does the user-collaborative recommender. The other recommender systems and the editors do not differ much from each other.

Figure 4 indicates that the diversity as in the inclusion of political vs. non-political items does not differ between recommender systems, as we already expected from Table 1. Even though the spikes of the user-collaborative algorithm are more pronounced, they roughly correspond to the peaks of the other algorithms.

Finally, Table 2 and Figure 5 and 6 again illustrate that also if we use formal measures of entropy to assess diversity, we see comparable little differences between the systems – with, as striking exception, the collaborative user-based recommendations that are persistently more diverse than even the random recommendations.

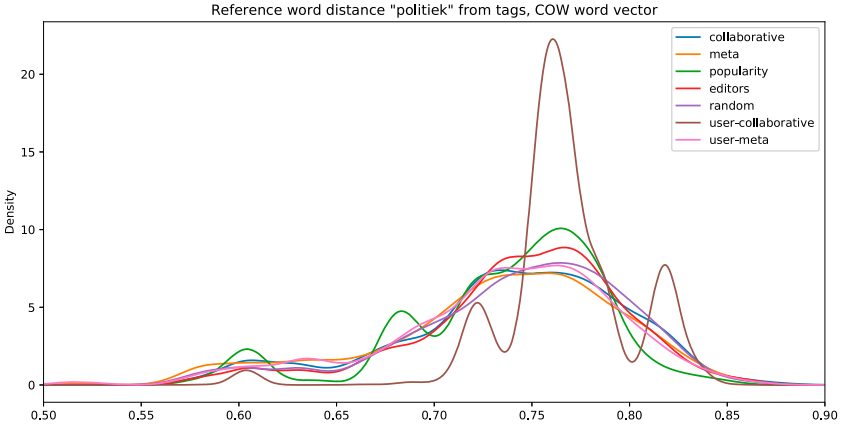

Figure 4. Distance from 'politiek'.

Table 2. Means and standard deviations of diversity metrics using topic distances and category proportions.

	Rao entropy	adapted Rao	Shannon	adapted Shannon
Collaborative (item)	0.0014 (0.0017)	0.08 (0.03)	0.88 (0.28)	0.41 (0.19)
Semantic (item)	0.0012 (0.0013)	0.08 (0.03)	0.85 (0.25)	0.39 (0.17)
Popularity	0.0019 (0.0026)	0.09 (0.03)	0.83 (0.37)	0.38 (0.24)
Editors	0.0014 (0.0019)	0.08 (0.03)	0.81 (0.32)	0.36 (0.20)
Random	0.0012 (0.0013)	0.09 (0.03)	0.83 (0.19)	0.38 (0.24)
Collaborative (user)	0.0018 (0.0016)	0.10 (0.04)	0.95 (0.15)	0.45 (0.10)
Semantic (user)	0.0013 (0.0015)	0.08 (0.03)	0.85 (0.27)	0.38 (0.18)

5. Conclusion and discussion

Our results indicate that news recommendation systems by and large match the diversity of journalistic recommendations. The personalized recommendations showed no reduction in diversity over the non-personalized recommendations; on the contrary, personalized collaborative filtering produced the highest amount of topic diversity in this controlled setting. In sum, we have established that conventional recommendation algorithms at least preserve the topic/sentiment diversity of the article supply.

Having said that, the conventional recommendation algorithms will follow the original distributions, either of the user selections of news, or of the supply of news. We saw that the long-tail was underrepresented. You can imagine that if this was a feedback system based on user selections with a long-tail that is continuously underrepresented, then, naturally, this long-tail will decrease over time. This is statistically unavoidable if no specific measures are taken to protect the long-tail. A truly diversity preserving recommendation engine will rather over-represent the 'minorities' and under-represent the 'majorities' to counter this narrowing effect of the distribution over time. This can be achieved by replacing the proportion in the diversity measures such that niche categories have a larger weight.

The general method that we suggest relies on specific methods for extracting measures like topical distance and tone. These measures can be extended with for instance; complexity, length, writing style, to not only give a diverse recommendation in the sense of meaning

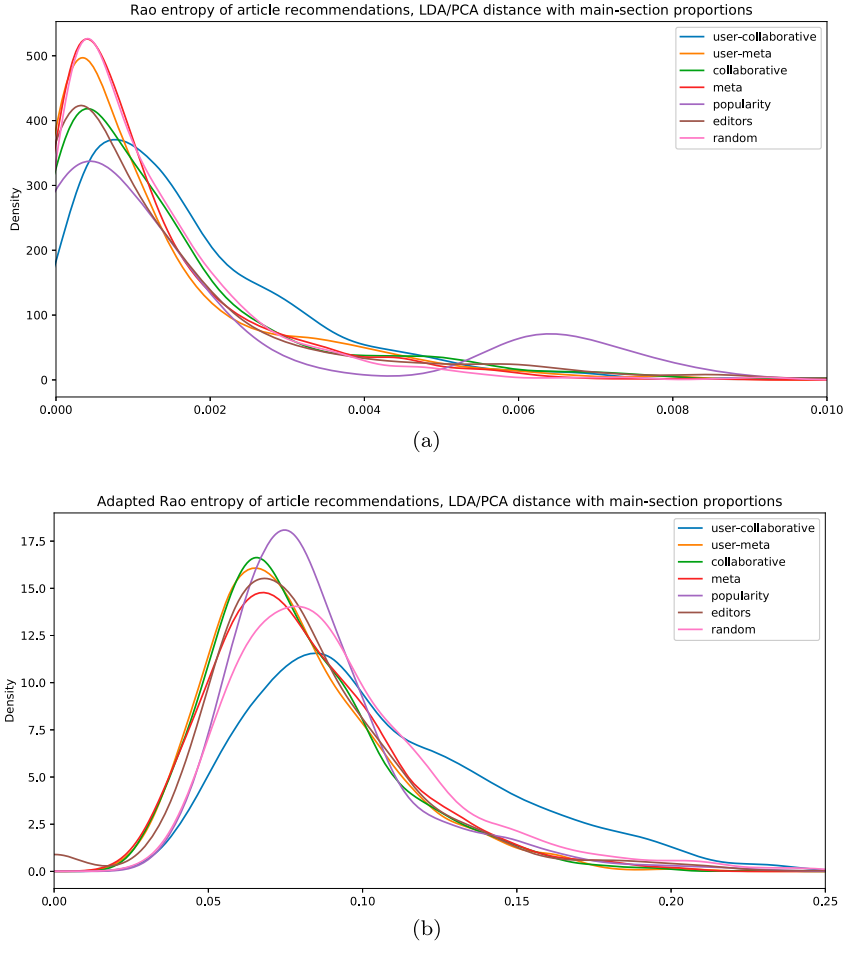

Figure 5. Diversity measure. Distributions for Rao quadratic entropy (a) and adapted Rao quadratic entropy (b).

but also in sense of the effort it takes to read the article. In that sense, we have described a paradigm that, for each measure one chooses to add, one can rely on best-in-class or state-of-the-art methods; it is method-agnostic. We relied on standard LDA for building our topic models. However, very recently, Azarbonyad et al. (2017) have proposed that hierarchical re-estimation of topic models can lead to better estimates of topic diversity. While they are interested in calculating the diversity within a document, we are more interested in differences between documents, future research should evaluate in how far their approach could improve ours. We also demonstrated the use of word embeddings to estimate the similarity of words and even documents. As the quality of these word embeddings relies heavily on the size and relevance of the corpus used to create the embeddings we encourage the generation of public word vectors dedicated to news articles.

The diversity measure we introduced was tested on simulated data from conventional recommendation engines. Next steps are to include our measures in a recommendation engine that is designed to maximize not just for instance the number of articles read or the time spent on them, but also the diversity of the different measures. This

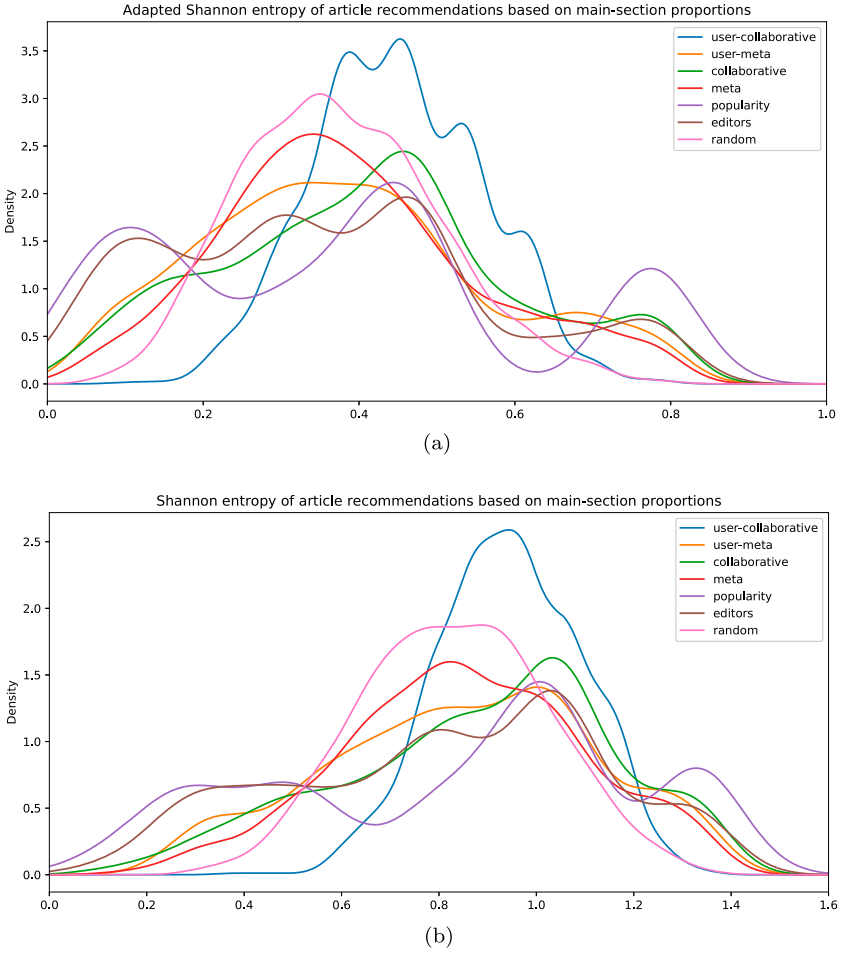

Figure 6. Diversity measure. Distributions for Shannon entropy (a) and adapted Shannon entropy (b).

recommendation engine will assume feedback between the user selections and the user recommendations, and can thus only be tested *in situ*, as such the serendipitous elements of a recommendation system can be used in a smarter way: by providing diverse news that is still relevant to the user.

Future work should explore the different approaches to maximize the diversity and elaborate the different steps that make up our general diversity framework and explore the different options, e.g. for normalizing the feature vectors and combining the features into one feature space that allows for a directional recommendation strategy. It became evident that we need better metrics of diversity to gain a better understanding of the true impact of recommender systems. To better understand the impact of these recommendation engines and whether over time there is indeed a reduction of content diversity these difference recommendation engines should be tested *in situ* over a longer period of several months. The current work should be seen in the context of the development of diversity metrics for the evaluation of recommendation engines.

It is important to note that our analyses are based on a pool of articles that were published in a high-quality broadsheet paper in one country. That means the pool of articles was quite diverse in terms of topic and tone, but only to a small extent in categories that we did not analyze (e.g. style or viewpoints). To get a better understanding of the influence of recommender systems on society at large, we need to replicate this study for a larger and more diverse pool of news. Similarly, our finding that recommender systems preserve diversity needs to be explored further. It could be that the negative effects of filter bubbles are caused not by recommenders that recommend self-enforcing information, but by repetitively showing a set of information snippets of which the most cognitive-consonant is most likely selected, which in time will lead to a narrowing of the recommended set. That means, future work on news recommendation systems needs to incorporate the notion of personalization over time, especially with regard to long-term effects of recommendations on the overall structure of information markets.

As such, this paper should be understood as a stepping stone into researching diversity as an outcome of recommendation systems as the rich and relevant concept it really is. As more and more users are receiving their news through these recommendation systems, be it on a social medium, news compilers integrated in automatic personal assistants, or on legacy news media homepages, it is crucial to understand and monitor how diverse the news menus are that are offered to the user, acknowledging that diverse can mean a lot of different things.

Notes

1. A different approach could be to compare the recommendations for different people (see also Haim et al., 2017).
2. First, we calculate the Euclidean distance for all topic pairs, resulting in a fully symmetric 50×50 matrix $\mathbf{M}$, which we call our topic dissimilarity matrix. Second, we multiply $\mathbf{M}$ with $\vec{w_{D1}}$ and $\vec{w_{D2}}$, for document 1 and 2 respectively, resulting in two matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\mathbf{A} = \mathrm{M}\sqrt{\vec{w_{D1}}\vec{w_{D1}}^{\mathrm{T}}}, \quad \mathbf{B} = \mathrm{M}\sqrt{\vec{\mathbf{w}}_{\mathbf{D2}}\vec{\mathbf{w}}_{\mathbf{D2}}^{\mathrm{T}}} \tag{3}$$

which represent the topic dissimilarity matrices weighed by the topic occurrence in the document in question. These matrices $\mathbf{A}$ and $\mathbf{B}$ contain mixed terms such as $M_{21}w_1$ which are ambiguous since it does not include the weight for one of the topics in the pair. To discard these mixed terms, we need to rewrite the above formulation to

$$\mathbf{A} = \mathrm{diag}(\mathrm{M})\mathcal{I}\sqrt{\vec{\mathbf{w}}_{\mathbf{D1}}\vec{\mathbf{w}}_{\mathbf{D1}}^{\mathrm{T}}} + (\mathrm{M} - \mathrm{diag}(\mathrm{M})\mathcal{I})\sqrt{\vec{\mathrm{w}}_{\mathrm{D1}}\vec{\mathrm{w}}_{\mathrm{D1}}^{\mathrm{T}}} - \mathrm{diag}\left(\sqrt{\vec{\mathrm{w}}_{\mathrm{D1}}\vec{\mathrm{w}}_{\mathrm{D1}}^{\mathrm{T}}}\right)\mathcal{I} \tag{4}$$

We can then calculate the Euclidian distance. We assume that each feature space has the same bounds, hence all feature spaces need to be normalized before being combined between the two matrices using the Frobenius norm: topical distance between two documents $= \|A - B\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij} - b_{ij}|^2}$.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the European Research Council, ERC-grant 638514, and the Research Priority Area of the University of Amsterdam. It was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

Notes on contributors

Judith Möller is a postdoctoral researcher at the Amsterdam School of Communication Research. In her research she focuses on the effects of political communication, in particular social media and digital media [email: J.E.Moller1@uva.nl].

Damian Trilling is an Assistant Professor at the Department of Communication Science and affiliated with the Amsterdam School of Communication Research. He is intrigued by the question how citizens nowadays are informed about current affairs and events in their society [email: d.c.trilling@uva.nl].

Natali Helberger is professor in Information Law at the Institute for Information Law. She specializes in the regulation of converging information and communications markets. Focus points of her research are the interface between technology and information law, user rights and the changing role of the user in information law and policy [email: n.helberger@uva.nl].

Bram van Es works is a Research Engineer at the eScience Centre, Amsterdam. He holds a PhD in Physics and has extensive experience in data analysis and machine learning. He has co-developed the a plugin to collect tracking data, the data pipeline and the analysis library for the personalised communication project [email: bramiozo@gmail.com].

References

Azarbonyad, H., Dehghani, M., Kenter, T., Marx, M., Kamps, J., & De Rijke, M. (2017). Hierarchical re-estimation of topic models for measuring topical diversity. In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, & J. Tait (Eds.), *Advances in information retrieval: 39th European Conference on IR research (ECIR 2017), Aberdeen, UK* (pp. 68–81). Cham: Springer.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Beam, M. A. (2014). Automating the news: How personalized news recommender system design choices impact news reception. *Communication Research*, *41*(8), 1019–1041.

Benson, R. (2009). What makes news more multiperspectival? A field analysis. *Poetics*, *37*(5), 402–418.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227.

Bridge, D., & Kelly, J. P. (2006). Ways of computing diverse collaborative recommendations. In International conference on adaptive hypermedia and adaptive web-based systems (pp. 41–50). Springer.

De Smedt, T., & Daelemans, W. (2012). Pattern for Python. *The Journal of Machine Learning Research*, *13*, 2063–2067.

Duncan, J., & Reid, J. (2013). Toward a measurement tool for the monitoring of media diversity and pluralism in South Africa: A public-centred approach. *Communication*, *39*(4), 483–500.

Dylko, I. B. (2016). How technology encourages political selective exposure. *Communication Theory*, *26*(4), 389–409.

Dylko, I. B., Dolgov, I., Hoffman, W., Eckhart, N., Molina, M., & Aaziz, O. (2017). The dark side of technology: An experimental investigation of the influence of customizability technology on online political selective exposure. *Computers in Human Behavior*, *73*, 181–190.

Ferree, M. M., Gamson, W. A., Gerhards, J., & Rucht, D. (2002). Four models of the public sphere in modern democracies. *Theory and Society*, *31*(3), 289–324.

Flaxman, S. R., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*, 298–320.

Fletcher, R., & Nielsen, R. K. (2017). Are news audiences increasingly fragmented? A cross-national comparative analysis of cross-platform news audience fragmentation and duplication. *Journal of Communication*, *67*(4), 476–498.

Haim, M., Graefe, A., & Brosius, H.-B. (2017). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism*, 1–14.

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191–207.

Junqué de Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2012). Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications*, *39*(14), 11616–11622.

Karppinen, K. (2013). *Rethinking media pluralism. Donald McGannon Research Center*. New York, NY: Fordham University Press.

Karppinen, K. (2015). The limits of empirical indicators: Media pluralism as an essentially contested concept. In P. Valcke, M. Sükösd, & R. G. Picard (Eds.), *Media pluralism and diversity* (pp. 287–296). London: Palgrave Macmillan.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4–5), 441–504.

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, *111*(Supplement C), 180–192.

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems – a survey. *Knowledge-Based Systems*, *123*(Supplement C), 154–162.

Masini, A., Aelst, P. V., Zerback, T., Reinemann, C., Mancini, P., Mazzoni, M., Damiani, M., & Coen, S. (2017). Measuring and explaining the diversity of voices and viewpoints in the news. *Journalism Studies*, *54*, 1–20.

Napoli, P. M. (2015). Assessing media diversity in the US: A comparative analysis of the FCC's diversity index and the EU's media pluralism monitor. In P. Valcke, M. Sükösd, & R. G. Picard (Eds.), *Media pluralism and diversity*. London: Palgrave. Assessing media diversity in the US: A comparative analysis of the FCC's diversity index and the EU's media pluralism monitor. In Media pluralism and diversity (pp. 141–151). Springer.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (Aug 2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, *33*(4), 459–478.

Owen, D., & Smith, G. (2015). Survey article: Deliberation, democracy, and the systemic turn. *Journal of Political Philosophy*, *23*(2), 213–234.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. New York, NY: Penguin.

Council of Europe (2007). Recommendation Rec(2007)2 of the committee of ministers to member states on media pluralism and diversity of media content.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). ELRA, Valletta, Malta. Retrieved from http://is.muni.cz/publication/884893/en.

Ribeiro, M. T., Ziviani, N., Moura, E. S. D., Hata, I., Lacerda, A., & Veloso, A. (2015). Multiobjective pareto-efficient approaches for recommender systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(4), 53.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 1–35). New York, NY: Springer.

Ricotta, C., & Szeidl, L. (2006). Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, *70*(3), 237–243.

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In Proceedings of the tenth international conference on language resources and evaluation (LREC 2012) (pp. 486–493), Istanbul, Turkey.

Shoemaker, P. J., Eichholz, M., Kim, E., & Wrigley, B. (2001). Individual and routine forces in gatekeeping. *Journalism & Mass Communication Quarterly*, *78*(2), 233–246.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). Retrieved from http://www.aclweb.org/anthology/W/W14/W14-3110.

Smyth, B., & McClave, P. (2001). Similarity vs. diversity. In International conference on case-based reasoning (pp. 347–361). Springer, Vancouver, Canada.

Sørensen, J. K., & Schmidt, J. -H. (2016). *An algorithmic diversity diet? Questioning assumptions behind a diversity recommendation system for PSM*. Working paper, Hans-Bredow-Institute for Media Research, Hamburg, Germany.

Strömbäck, J. (2005). In search of a standard: Four models of democracy and their normative implications for journalism. *Journalism Studies*, *6*(3), 331–345.

Stroud, N. J. (2011). *Niche news: The politics of news choice*. New York, NY: Oxford University Press.

Sunstein, C. R. (2007). *Republic.com 2.0*. Princeton, NJ: Princeton University Press.

Tulkens, S., Emmery, C., & Daelemans, W. (2016). Evaluating unsupervised Dutch word embeddings as a linguistic resource. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016). European Language Resources Association (ELRA)*, Paris, France.

Van Cuilenburg, J. (1999). On competition, access and diversity in media, old and new: Some remarks for communications policy in the information age. *New Media & Society*, *1*(2), 183–207.

Willemsen, M. C., Graus, M. P., & Knijnenburg, B. P. (2016). Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, *26*(4), 347–389.

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In Proceedings of the 14th international conference on World Wide Web (WWW '05). ACM Press, (PP. 22–32), Chiba, Japan.

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles?. *Internet Policy Review*, *5*(1).