Seek and you shall find?

A content analysis on the diversity of five search engines' results on political queries

Search engines are important political news sources and should thus provide users with diverse political information – an important precondition of a well-informed citizenry. The search engines' algorithmic content selection strongly influences the diversity of the content received by the users – particularly since most users highly trust search engines and often click on only the first result. A widespread concern is that users are not informed diversely by search engines, but how far this concern applies has hardly been investigated. Our study is the first to investigate content diversity provided by five search engines on ten current political issues in Germany. The findings show that sometimes even the first result is highly diverse, but in most cases, more results must be considered to be informed diversely. This unreliability presents a serious challenge when using search engines as political news sources. Our findings call for media policy measures, for example in terms of algorithmic transparency.

Seek and you shall find?

A content analysis on the diversity of five search engines' results on political queries

After British polling stations closed on June 23, 2016 – the day of the Brexit referendum – the search engine *Google* recorded an enormous rise in the query "What will happen if we leave the EU?" (O'Hare, 2016). Apart from the question of why people did not inform themselves earlier, this example illustrates how important search engines have become as a political information source (Dutton, Reisdorf, Dubois, & Blank, 2017; Newman, Fletcher, Kalogeropoulos, & Nielsen, 2019). By reducing complexity, they provide orientation in the flood of information, thus acting as powerful "gatekeepers" that select, sort, and redistribute online content (AUTHORS; Latzer, Hollnbuchner, Just, & Saurwein, 2016). Therewith, they essentially influence what users know about political issues (Granka, 2010) and the formation of public opinion (Latzer et al., 2016). However, their filtering and sorting can lead to biased information and thus entail risks for the diversity received by the users (AUTHORS). But to what extent do search engines provide comprehensive, diverse information about current political issues such as *Brexit*?

Even though this question is of high societal relevance since diverse political information is a precondition for a well-informed citizenry, it has been widely unexplored (AUTHORS). Granka (2010) postulates quantitative content analyses that measure diversity "on a per-query level" (p. 370). Starting from this desideratum, our study compares content diversity (based on the articles that are hyperlinked by the first ten search engine results) provided by five search engines – *Google*, *Ask*, *Bing, DuckDuckGo, Ixquick* – regarding ten current, controversial, political issues in Germany. We contribute to the methodological improvement of often inappropriate and overly rough indicators of content diversity (e.g., Karppinen, 2006) by developing an innovative, valid, detailed measurement of two pivotal dimensions of content diversity: *information diversity* and *diversity of speakers*. The findings show that obtaining diverse information on current political issues through search engines is possible, but by no means guaranteed – particularly if a user simply clicks on the first result, as most users do (Pan et al., 2007).

Below, we first describe how the filtering and sorting of search engines can affect content diversity, give an overview of the few existing empirical studies thereon, and derive our innovative measurement of content diversity. Afterwards, we present our findings and discuss their implications as well as our study's limitations.

Conceptual framework

How search engines influence content diversity

Diversity is considered a precondition of healthy democracies (Napoli, 1999) since it is assumed to guarantee a public debate with opposing viewpoints and a well-informed citizenry, as illustrated by the "marketplace of ideas" – an idealized metaphor of public discourse (Karppinen, 2006): citizens shall freely exchange diverse ideas and viewpoints to ensure well-informed decision-making, tolerance toward other viewpoints (Jandura & Friedrich, 2014), and the stimulation of "popular wisdom" (Donohue & Glasser, 1978, p. 592). The media should contribute to it by providing diverse content (Jandura & Friedrich, 2014), stressing the importance of diversity in media policy (Just, 2009) and in communication research.

Originally, the debate focused on human, journalistic gatekeepers, often revitalized by developments considered as potential threats to content diversity (e.g., concentration processes in the newspaper market (Donohue & Glasser, 1978), the introduction of commercial broadcasting in European countrues (Aslama, Hellman, & Sauri, 2004)). The rise of the Internet brought along the hope of unlimited content diversity online (European Commission, 2010, p. 30), but it quickly turned out that the processing capacities of the users limit diversity: in the information flood, they rely on gatekeepers to identify relevant information more than ever. Search engines were invented exactly for this purpose: they select pieces of information from a myriad of different sources based on algorithms, which entails the potential to present highly diverse content. However, in contrast to journalistic gatekeepers, search engines are not bound to any normative principles (AUTHORS; Granka, 2010). Since each act of selecting and sorting content follows specific criteria while ignoring others, both human and algorithmic gatekeeping are somewhat biased. The particular problem with search engines is the non-transparency of their selection criteria (van Hoboken, 2012) which fuels concerns of biased search results (Epstein & Robertson, 2015; Pariser, 2011) and impairments of content diversity. Two steps of search engines' gatekeeping process are pivotal (AUTHORS; Kulshrestha et al., 2019; Latzer et al., 2016):

(1) Filtering. Search engines select some pieces of information while filtering out many others. To be findable by a search engine, a website must be contained in the engine's search index, which includes every website the search engine has stored before but excludes all other websites. To become visible within a specific search engine results page (SERP), a website must be deemed relevant in terms of the specific search query. The search engine selects some items and excludes others, possibly causing bias at the *input* level (Kulshrestha et al., 2019) of the gatekeeping process.

(2) Sorting. Search engines sort and prioritize content that is relevant to the query by ranking content at the top or further down the SERP. This sorting strongly affects the content diversity received by the users (Kulshrestha et al., 2019: *ranking bias*). Although search engines allow access to an enormous content supply, most users let themselves guide by different cues, such as brands (Ieong, Mishra, Sadikov, & Zhang, 2012) or ranking (Haas & Unkel, 2017; Pan et al., 2007). Users hardly ever proceed beyond the first SERP (the first ten results in most cases; Jansen & Spink, 2006) and often only click on the first result (Pan et al., 2007). Thus, only a tiny share of content available via search engines is easily reachable by users and in fact accessible (for the difference between "available" and "accessible" see Hargittai, 2000) respectively accessed. The sorting may foster over- or underrepresentation of certain pieces of information and thereby affect content diversity.[1]

Research on content diversity provided by search engines

The potential threat of search engine bias has stimulated some research, for example on the partisanship of *Google* search snippets (Hu, Jiang, E. Robertson, & Wilson, 2019), the sources of *Google's* Top Stories (Trielli & Diakopoulos, 2019), the question of whether (socially) tabooed or controversial sub-issues are suppressed (Gerhart, 2004), and how far biased results can influence users' voting decisions (Epstein & Robertson, 2015). However, studies on content diversity provided by search engines are still extremely scarce. Most of the few existing studies (e.g., AUTHORS; Unkel & Haim, 2019; Trielli & Diakopoulos, 2019) concentrate on structural diversity by analyzing different types of sources included in the SERP. But whether structural diversity necessarily increases content diversity seems questionable. For example, according to an analysis by AUTHORS, *Google* performs worse than four other general search engines regarding source diversity due to its focus on journalistic websites and the small share of alternative website types (e.g., weblogs; see also Unkel & Haim, 2019; Metaxas & Pruksachatkun, 2017). However, particularly journalistic articles may provide more diverse informational aspects than other sources (AUTHORS). Moreover, many studies have focused on the teaser information on the SERP itself rather than including the content of the hyperlinked articles (e.g., AUTHORS; Beiler, 2013) – sparse information only allowing for a superficial measurement. Neuberger and Lobigs (2010) applied a more detailed measurement and found higher content diversity within articles accessed via *Google* than articles accessed directly via specific journalistic websites. However, their study has become outdated, and their small sample (30 results) and limitation to one single issue limit the generalizability of their findings.

Other studies have approached the field from different angles: An automated content analysis of search results on nanotechnology by Li, Anderson, Brossard and Scheufele (2014) compared the distribution of thematic foci between the first ten and the 11th to 32nd result and found a more equal distribution within the lower ranked results. This indicates that content diversity may increase when considering results ranked further down the SERP. Möller, Trilling, Helberger and van Es (2018) showed that algorithmic content selection does not necessarily limit content diversity: based on the output of a Dutch newspaper, they simulated a set of algorithm-based article recommendations (e.g., based on overall popularity or semantic filtering) and found that algorithmic selection led to comparable issue diversity as content selected by human editors. Taking into account that even the highest diversity remains ineffective if users do not make use of it, Fletcher and Nielsen (2018) conclude from survey data that using search engines to access news corresponds to the exploitation of more different media brands. However, except Neuberger and Lobigs (2010), none of these studies investigated content diversity provided by search engines in detail. To address this gap, we need a detailed, valid operationalization of content diversity.

Conceptualizing content diversity

Diversity is conceptualized very differently, depending on the object of investigation and the level of analysis (McQuail, 1992; Figure 1). Many common indicators are easy to measure and allow for comparability *across issues*, but have the disadvantage of being too rough and superficial to be linked to the normative principles of diversity (Karppinen, 2006). This applies particularly to *structural diversity* as measured by the diversity of media outlets, program types, and genres. However, the same objection can be raised for many operationalizations of *content diversity*, defined as the "heterogeneity of media content in terms of one or more specified characteristics" (van Cuilenburg, 2000, p. 52): most empirical studies use rough indicators such as different geographical locations (e.g., countries), policy fields (e.g., social policy, economic policy), and certain people and groups (e.g., political parties) (van Cuilenburg, 2007) that do not suffice to evaluate whether news coverage can contribute to a well-informed citizenry.

– Figure 1 about here –

To that purpose, we need more detailed indicators that focus on content diversity *within issues*. Two aspects are of particular importance: news coverage should (1) include a wide range of aspects and viewpoints related to each issue (Jandura & Friedrich, 2014) and (2) enable different societal groups to be heard in public (McQuail, 1992). Indeed, such a valid approach is methodologically challenging and effortful. This is probably one reason why only few studies have followed this path (for exceptions see Benson, 2009; Masini et al., 2018; Neuberger & Lobigs, 2010) – but usually at the price of being limited to one single issue. Hence, research on how far news coverage provides the preconditions of a well-informed citizenry should in general meet three requirements: it should investigate (1) *content diversity* rather than structural diversity (2) *within* rather than across *issues* (3) by means of *valid indicators*. Studies on content diversity provided by search engines should additionally (4) measure content diversity based on *hyperlinked articles* rather than on the teaser information within the SERP and (5) take user behavior into account by considering that most users only click on the *top-ranked results*.

Our study fulfills all five requirements. It analyzes two core dimensions of content diversity: (1) *Information diversity* – a newly introduced term in the international context (for the German version of the term see Neuberger and Lobigs, 2010) – refers to single informational aspects on political issues. These so-called “information units” (Geiß, 2015; Haßler, Maurer, & Oschatz, 2014) comprise for example pure facts, background information, and viewpoints. (2) *Diversity of speakers* refers to different actors (societal groups, individuals) expressing an opinion on the respective issue (including evaluations, demands). This indicator allows for a more valid measurement of *diversity of access* (McQuail, 1992) (respectively being heard in the public discourse) than *actor diversity* which measures if people/groups are only mentioned (Humprecht & Esser, 2018), paraphrased, or quoted (Masini et al., 2018), regardless of whether they express an opinion or not (for the difference between actors and speakers see Humprecht and Esser, 2018). It has to be noted that diversity of speakers is not a substitute for *diversity of opinions* (since multiple speakers can concur), but rather measures how many different actors get the opportunity to voice their opinions publicly. In contrast to previous studies (e.g., Jacobi, Kleinen-von Königslöw, & Ruigrok, 2016), we do not only measure how diverse "important" voices (particularly politicians from the main German political parties) are represented, but also consider a broad range of societal groups and ordinary people.

Our approach is based on the concept of *open diversity*[2] (in line with Humprecht and Esser, 2018; Jacobi et al., 2016), which postulates the equal visibility of people/groups and informational aspects of an issue, regardless of status and real-world distribution (Vettehen, 2005), entailing the potential to "promote change and innovation" (McQuail, 1992, p. 148) by reinforcing new ideas (van Cuilenburg, 2000). However, if realized in full, the normative ideal of open diversity can collide with the media's function to select relevant content, which is why unlimited diversity is not desirable (Vettehen, 2005). Particularly search engines are expected to "keep people from drowning in an information flood" (Saurwein, Just, & Latzer, 2015, p. 35). Here, a general problem of diversity research becomes obvious – the the lack of a threshold from which diversity can be considered sufficient. To deal with this dilemma, we made a compromise in our study: we measured content diversity based only on information units and speakers that were potentially relevant to capture an issue comprehensively (see methods section). In contrast to studies that used the empirically given value across all search engines as a baseline (e.g., Mowshowitz & Kawaguchi, 2002), we evaluated diversity measures from a normative perspective since we assume that the empirical values could show diversity deficits. Since we assume that comparisons are the best way of dealing with the lack of concrete thresholds, we compare five search engines and ten political issues.

Taking typical consumption patterns into account, we focus on the diversity of the articles hyperlinked by the first ten results of each SERP as a proxy for the maximum diversity received by users. However, most users will stop reading (far) earlier, which is why we investigate the progression of information diversity and diversity of speakers from result 1 to result 10 and address the following research questions:

RQ1: How do information diversity and diversity of speakers develop when considering more (one, two, ..., up to ten) search results (progression of diversity)?

RQ2: How does progression of information diversity and diversity of speakers differ between search engines?

RQ3: How does progression of information diversity and diversity of speakers differ between ten current political issues?

RQ4: How does the interplay between search engines and political issues influence the progression of information diversity and diversity of speakers?

Method

Sample and collection of materials

To answer the research questions, we conducted a quantitative content analysis of five search engines relevant for the German market: *Google* – by far the most important search engine (market share: more than 90 percent of all search queries in Germany; SEO, 2017) –, its "strongest" competitors (*Bing*, *Ask*) and two alternative search engines claiming to pay attention to users' privacy (*Ixquick* and *DuckDuckGo*) (SEO, 2017). Search engines with results completely based on the algorithm of *Google* (e.g., *T-Online*) or *Bing* (e.g., *Yahoo*) were excluded, even in case of a slightly higher market share compared to *Bing* or *Ask*.

Several measures aimed at increasing the ecological validity of our findings compared to previous studies. First, from all political issues controversially discussed in Germany between November 2015 and June 2016, we selected ten issues that were as different as possible, guided by variation on three criteria (Geiß, 2015; Table 1): *level of relevance* (global/supranational, national, regional), *affected societal areas* (e.g., diverse policy areas, economy, judiciary), and *occasions of news coverage* (e.g., long-running and ongoing debates, political proceedings). Second, for each selected issue, we formulated one search query (Table 1) based on the most frequently used search query according to *Google Trends*. This tool gives an overview of how likely users were to run different search queries for one issue (e.g., "TTIP" vs. "free trade agreement" vs. "transatlantic free trade agreement") across time (Trevisan, Hoskins, Oates, & Mahlouly, 2018). This procedure allowed us to identify peaks. Assuming that search engines are particularly central information sources at the time of these peaks, we gathered the data shortly thereafter. Third, to ensure comparability between search engines and issues and to minimize sources of noise – particularly personalization – in the data, the collection of materials was always conducted by the same research assistant in one place, using the same computer and browser (*Firefox*). Directly before storing the materials, privacy conditions were set wherever possible (clearing browsing history; deselecting personalization settings defaulted by *Google* and *Bing*; activating *Firefox'*s incognito mode; see also Robertson et al., 2018; Trielli & Diakopoulos, 2019). To reduce time lags due to manual storage to a minimum, the research assistant started each query on all five search engines in different browser windows (nearly) simultaneously. Then, she clicked on the first ten[3] articles hyperlinked in each SERP, considering all organic search results and news-card respectively news-triplet components (for the distinction of different types of search results see Robertson et al., 2018) as the latter are presented to the user in a similar manner as organic search results. Additional components (e.g., maps, related searches, "people also ask", snippets) were ignored. Our operational definition of "article" comprised all kinds of content we found on the hyperlinked webpages, for example journalistic articles, encyclopaedic articles (e.g., *Wikipedia*), Twitter feeds, and local authorities' websites (including embedded videos and audio recordings, but excluding further hyperlinks and previews of further articles). Each SERP and its first ten hyperlinked articles were saved. The overall sample comprises 500 articles (5 search engines×10 queries×10 articles).

– Table 1 about here –

Measurement and reliability

Formal variables – *search query, search engine, rank of search result* – were coded based on the information on the SERP. The content-related variables – *information units* and *speakers* – were coded based on the hyperlinked articles (Figure 2). For the coding, we employed detailed, issue-specific lists of information units and speakers (for an example see Appendix).

– Figure 2 about here –

Information units. We define an information unit as an aspect, fact or viewpoint on an issue (Geiß, 2015; Haßler et al., 2014). Our issue-specific lists included for example information on the background of the issues, current incidents, involved actors, and potential future developments (e.g., in the *Brexit* case: date and results of the referendum, relationship between the European Union and Great Britain and opinions thereon, political and economic (dis-)advantages).

Speakers. We define a speaker as an actor – a person, group, or institution – who explicitly pronounces an opinion (including demands or evaluations) on the respective issue (Humprecht & Esser, 2018). Actors only voicing facts were not coded, in line with the concept of *diversity of access* which considers the ability of different actors to contribute to the public discourse with their own opinion to be particularly important. Our lists include for example political actors, NGOs, experts, mass media, and ordinary citizens/the people involved in the respective issue.

To ensure that the lists only included meaningful information units, they were compiled based on (a) prior qualitative inspection of all search results on the respective issue; (b) close reading of relevant coverage in two Germany quality newspapers with different political leanings (right: Frankfurter Allgemeine Zeitung; left: Süddeutsche Zeitung) in the 10 days prior to running the respective search query; (c) complementation of logically “missing” units (e.g., when the leader of one parliamentary faction was mentioned in (a) or (b), all other faction leaders were added to the list).

The final lists contain 85 to 154 information units (mean: 115) and 105 to 207 speakers (mean: 153). For each article, all information units and speakers from the issue-specific lists that appeared at least once in it were coded. An unlimited number of information units and speakers could be coded per article, but each information unit and speaker present in the article were coded just once. According to our pretests, this procedure ensured the highest possible inter-coder reliability.

Inter-coder reliability (four student coders) was perfect (Brennan-Prediger's kappa=1.00) for all formal variables (search engine, issue, position on the SERP) and good for both information elements (.74) and speakers (.76) across all issues, based on a test of 5% of the sample. We used Brennan-Prediger's kappa since it is chance-corrected and more robust than Krippendorff's alpha regarding variables with a skewed distribution (Quarfoot & Levine, 2016).

Diversity indices. For both information units and speakers, we used the frequency distributions to calculate the standardized entropy (see also Humprecht & Esser, 2018; Jacobi et al., 2016) which is based on Shannon's H (Shannon & Weaver, 1949). It compares the Shannon entropy from the real data (H_0) with the maximum possible entropy (H_{max}) that would result if the data (n units) were perfectly equally distributed across the c bins/categories:

$$H_{std} = \frac{H_0}{H_{max}} = \frac{-\sum_{i=1}^{c} \frac{f_i}{n} \cdot \ln \frac{f_i}{n}}{-\sum_{i=1}^{c} \frac{1}{c} \cdot \ln \frac{1}{c}}$$

For instance, when 3 categories (a, b, c) are possible and n=30 units are coded, the true distribution may be f(a)=20, f(b)=8, f(c)=2; the theoretical maximum entropy would be reached at f(a)=f(b)=f(c)=10. This leads to: H_0=0.803; H_{max}=1.099; H_{std}=0.731. The standardized entropy is mathematically robust for small samples (McDonald & Dimmick,

2003) and ranges from 0 (concentration on one information element/speaker, lowest possible diversity) to 1 (completely even distribution of all information units/speakers, highest possible diversity). Theoretically, it refers to the concept of open diversity since it is sensitive to the evenness of distribution across different categories and considers if new elements really increase diversity or are rather “more of the same” and decrease diversity.[4] We did not aggregate any codes when calculating the standardized entropy so as to avoid loss of information. The entropy was not calculated on the basis of each individual article but in a cumulative manner, including the articles stepwise (result 1, results 1+2, results 1+2+3, …, results 1 to 10) in the overall entropy value, so as to measure the progress of diversity for each issue and search engine across the first ten results. The entropy values were then standardized to ensure comparability across search engines and issues. This means that they were placed in relation to the number of theoretically possible information elements or speakers within the respective issue-specific list.

Analysis. Since the distribution of averages of standardized diversity scores were unknown, we decided to base our confidence interval calculation on a bootstrapping procedure. This means that from the available diversity average scores, we drew a great number of random samples (of the same sample size the original average calculation was based on) with replacement (so-called “replicates”), in our case 10,000 replicates. This would give us an empirical sampling distribution to compare the calculated mean with. It provides an impression of how the estimated averages of diversity scores would vary according to random processes. For a two-sided 95% confidence interval, the lowest 2.5% (i.e. the lowest 250 scores) and the topmost 2.5% (the topmost 250 scores) of the sampling distribution would be cut off and the lowest and highest remaining value would be the lower and the upper bound of the confidence interval, respectively. 9,500 of 10,000 scores obtained in the bootstrap would fall into that confidence interval. This e.g. allows for nonsymmetric distributions around the average.

Findings

Progression of diversity across search engines and issues *(RQ1)*

Figure 3 displays the raw distribution of diversity scores as a "violin plot". Overall, information diversity is somewhat higher than diversity of speakers which seems reasonable since an article can be written without referring to any speakers but not without mentioning basic information elements. Besides, the lists of speakers contained a higher number of codes on average, which makes it more difficult to reach high standardized entropy values. Therefore, we focus on comparisons between search engines respectively between issues on the same dimension (information diversity respectively diversity of speakers) rather than on comparisons between both dimensions.

The progression of diversity is similar on both dimensions: considering only the first results, and then considering progressively more results (1-2, 1-3, 1-4,…) leads to a decelerating upward curve approximating a saturation point. The range of diversity values is quite large regarding the first result (observe how the "violin" plots stretch downward the y-axis), but decreases substantially if more results are taken into consideration. This finding indicates that even though the diversity provided by the first result can be quite high, there is a noticeable risk of not getting diverse (and comprehensive) information when considering only the first result. Both information diversity and diversity of speakers increase rapidly when results 2 to 5 are included, while the subsequent flattening curves show that results 6 to 10 do not add much more to diversity (decelerating saturation).

– Figure 3 about here –

Diversity progression by search engines *(RQ2)*

There are no general differences between search engines regarding their average information diversity or diversity of speakers (Figure 4). To compare the progression of diversity between the five search engines, we calculated the means of the cumulative standardized entropy across issues by search engines. The estimated averages (10,000 bootstrap replicates) in Figure 5 are similar to the overall progression. Beyond that general pattern, there are some interesting differences. On average, *Ask* provides users with the highest information diversity and diversity of speakers within the first three results. These differences level out as more search results are considered. However, the differences do not reach statistical significance even if only considering the first search result. One reason for the relatively strong performance of *Ask* is that in nine out of ten issues, *Ask* ranks entries from the online encyclopedia *Wikipedia* first (Table 2). These articles are often very comprehensive (for exceptions, see *climate change* and *refugees*) and generally very important sources for search engines (see also AUTHORS; McMahon et al., 2017; Table A1 in the Appendix).

– Figures 4 & 5 about here –

– Table 2 about here –

Progression of diversity by political issues *(RQ3)*

On average, the ten issues differ substantially in terms of both information diversity and diversity of speakers. The issues roughly formed three groups: *TTIP* had the greatest information diversity and diversity of speakers. *Böhmermann, Panama Papers, Brexit, Syria, NSU,* and *assaults in Cologne* form the middle group with moderate diversity. The *regional election in Rhineland-Palatinate* and *Refugees* form the third group with very low diversity. *Climate change* is in the moderate diversity group for information diversity and in the low diversity group regarding diversity of speakers (Figure 6).

To compare the progression of diversity between the ten political issues, we calculated the means of the cumulative standardized entropy across search engines by issues. While the overall picture of the progression of diversity is still the same on both dimensions, clear issue-specific differences become apparent (Figure 7): the overall most diversely portrayed issue, *TTIP*, is far ahead from the first to the tenth result across both dimensions while *refugees*, *climate change* (regarding diversity of speakers) and the *regional elections in Rhineland-Palatinate* lag behind across all ten results. The latter two issues, in particular, illustrate that users cannot rely on being diversely informed about political issues by search engines, even if they click on and read each of the first ten articles.

– Figures 6 & 7 about here –

The comparison between the issues illustrates the difficulties of investigating the "black boxes" of algorithms very well, as shown by two similar types of issue, *climate change* and *TTIP.* Both represent ongoing debates (occasion of coverage) on international policy issues (societal area) on a global level (level of relevance) that intensified at the time of data collection due to current events. Since both search queries used in this study were broad, the search engines could not "know" what exactly we were looking for. Despite these similarities, the results for *TTIP* were clearly more diverse. Additionally, for both *climate change* and *TTIP*, all search engines ranked a *Wikipedia* entry first (Table 2), but the different entropy indices based on this first result show that this does not always lead to a high degree of content diversity.

Progression of diversity by search engine and political issue *(RQ4)*

To this point, we reported the results based on the average values across search engines and/or issues. However, averages can mask key differences that become apparent when using a certain search engine to learn about a certain issue. In the last step of our analysis, we take such differences into account and calculate the cumulative entropy from the first result[5] to the tenth result. Figure 8 shows a homogeneous pattern concerning information diversity: despite some unsystematic outliers (e.g., the first result for *Brexit* on *DuckDuckGo*), the issue-specific patterns described above generally hold true across search engines.

When it comes to the diversity of speakers, however, the picture changes. In the case of some issues, the diversity of speakers develops similarly across all search engines (e.g., *TTIP*, *assaults Cologne*), whereas it differs more strongly between search engines for other issues (e.g., *refugees, NSU trial, regional elections*). In the case of the *NSU* trial, for example, *Ask* provides visibly higher diversity of speakers than *DuckDuckGo* across all ten results; this is certainly a descriptive result which we cannot check for statistical significance. Moreover, in the case of the *regional elections*, the diversity provided by *Ixquick* does not even increase from the first to the tenth result. This finding shows that the diversity a user gets on an issue can be influenced considerably by the search engine used, and that even using five or even ten search results does not guarantee a high diversity of speakers.

– Figure 8 about here –

Discussion

Search engines have become pivotal political information sources (Newman et al., 2019). This raises concerns since their selection criteria are non-transparent, their algorithm-based filtering and sorting might cause bias (AUTHORS), most users rely on the very first search result (Pan et al., 2007), and biased results can influence voting decisions (Epstein & Robertson, 2015). However, since empirical research on content diversity provided by search engines is widely missing, it is unclear how justified these concerns are. Our study compares the content diversity supplied by different search engines to investigate the extent to which they provide the basis for a well-informed citizenry. We analyse the progression of content diversity from the first to the tenth result provided by five search engines on ten political issues, focusing on two crucial dimensions of content diversity: *information diversity* and *diversity of speakers*. Our innovative, valid measurement thereof – using issue-specific lists of information units and speakers – provides an important conceptual and methodological improvement compared to the often inappropriate and overly rough conventional measurements of content diversity (Karppinen, 2006). The results show that users are considerably more diversely informed if they make use of more than the first results (*RQ1*), overall independent of the search engine (*RQ2*) and widely independent of the issue (*RQ3*). However, the progression of diversity differs noticeably between certain political issues *(RQ3)*. Particularly the diversity of speakers sometimes strongly depends on which specific search engine is used to inform a user about which issue (*RQ4*). Sometimes, users find comparatively diverse information even in the very first result and regardless of the search engine, but this outcome cannot be relied upon. This unreliability, combined with the non-transparency of the search algorithms' filtering and sorting and the users' preference for the first result, is a serious challenge when using search engines as political news sources.

Naturally, our study has some limitations. Our focus of ten political issues at a certain point in time allowed coding them in a level of detail that goes far beyond previous studies with much rougher indicators (Karppinen, 2006), but is still only a snapshot which brings along limited generalizability. Future studies should test how far manual and automated methods can be combined to complement the necessary level of detail with larger samples. Such studies could, for example, consider more different issues, run each query at several points in time, and compare results generated by different search queries as prior research has shown that they can substantially alter search results (Hu et al., 2019; Robertson et al., 2018). Even though the time delays in data storage were very small, automated storage (Hu et al., 2019) could completely avoid them. Moreover, we distinguished carefully between speakers from different societal areas, but our measurement of diversity of speakers neglects other important criteria of social diversity (e.g., gender, race/ethnicity, socioeconomic status). Moreover, diversity of speakers differs from diversity of viewpoints – a construct that is difficult to operationalize. Future research should intensify efforts in developing a reliable measurement of diversity of viewpoints.

Altogether, our study shows that users will not necessarily be informed diversely about politics via search engines, especially if they only rely on the first result. From this potential threat to democracy and a well-informed citizenry, we derive some recommendations for both scientific research and media policy: As search algorithms change continuously and dynamically, it is crucial to establish constant monitoring that helps to uncover potential threats. This requires long-term public funding and the development of innovative (computational) methods in order to handle a huge amount of data. Search engine companies should provide researchers with access to these data. Besides, content analyses cannot show how diversely users are *actually* informed. Therefore, we strongly recommend more research into the use of search engines (e.g. Epstein & Robertson, 2015; Haas & Unkel, 2017), particularly into the concept of exposure diversity (Napoli, 1999) that combines content analyses with investigations on the usage of search engines.

To ensure diversity, such research should go hand in hand with media policy actions. Our study illustrates how important it is for users to go beyond the first result. Thus, we see a need to strengthen users' algorithmic literacy. Media policy could require companies to provide greater transparency of search engine algorithms (AUTHORS). Of course, total transparency of search algorithms is not feasible since it would run counter to the legitimate interests of search engine companies to protect their trade secrets. Such a radical solution would also not make sense from the users' perspective as they would not be able to understand the algorithms in all their complexity. However, we think making the most influential criteria and major changes over time visible and understandable for users is a viable way (see also Schulz, cited in Fuchs, 2019). Another problem is that the typical design of SERPs nudges many users to click on the first result only. A potential way of increasing exposure diversity could be to employ a different design that focuses less on ranking and encourages users to have a closer look at different teasers and click on more results (see comparable considerations on "diversity by design": Helberger, 2011). Besides, users should have the opportunity to choose between or weight different filtering and sorting criteria.

Search engine companies will likely not implement such suggestions voluntarily since these counteract their aim to deliver the most suitable result for each user in the easiest possible way. However, recent developments in media policy calling for stronger regulation might pave the way for such measures. These discussions do not only focus on antitrust law, but increasingly also on search engines' influence on public opinion formation as a reason to demand more transparency. The German draft for the "Interstate Treaty on Media" (in

German: “Medienstaatsvertrag”) which particularly focuses on *Google* as dominating search engine can be seen as a revolutionary first step (see Schulz, cited in Fuchs, 2019).

Besides search engine regulation, media policy should continue to take measures that safeguard the diversity of journalistic media (Möller, Helberger and Makhortykh, 2019) for two reasons: First, since SERPs on political issues typically include lots of journalistic articles (AUTHORS), a high content diversity of these articles strengthens content diversity within SERPs. Second, journalistic media still account for a significant part of users’ news repertoire (Newman et al., 2019). To us, both regulating search engines and safeguarding a diverse media landscape seems a promising way of ensuring content diversity, even in a high-choice information environment.

[1] A third form of possible search engine bias – personalization – refers to different filtering and sorting between individual users ased on users’ previous search behavior and personal preferences (AUTHORS), which together with the ranking bias leads to the overall output bias (Kulshrestha et al., 2019). However, current studies (e.g., Haim et al., 2018; Robertson et al., 2018) concluded that the related concerns on “filter bubbles” (Pariser, 2011) are exaggerated; personalization is mainly conditioned by geolocation (Hannak et al., 2013), which is why we neglect this aspect in our study.

[2] Reflective diversity demands that the real world distribution of the aspects of interest (e.g., viewpoints) should be reflected in news coverage, which tends to reinforce the status quo since it neglects minority groups and positions that also contribute to diversity (McQuail, 1992).

[3] If the first SERP comprised less than ten results, she proceeded to the second SERP.

[4] As we coded each speaker and information unit only once per article, we can consider the equality of distribution only *across articles* but not *within one article*. However, an additional analysis in which we only considered the breadth of the frequency distribution (the share of considered information elements/speakers across all articles) yielded very similar results, albeit at a lower level.

[5] At this level, the entropy value in each case is based only on the frequency distribution of the first article.

References

Aslama, M., Hellman, H., & Sauri, T. (2004). Does market-entry regulation matter? Competition in television broadcasting and programme diversity in Finland, 1993-2002. *Gazette: The International Journal for Communication Studies*, *66*(2), 113–132.

Beiler, M. (2013). *Nachrichtensuche im Internet: Inhaltsanalyse zur journalistischen Qualität von Nachrichtensuchmaschinen* [Searching for news on the Internet: Content analysis on journalistic quality comparing five news search engines]. Konstanz: UVK.

Benson, R. (2009). What makes news more multiperspectival? A field analysis. *Poetics*, *37*(5-6), 402–418.

Donohue, T. R., & Glasser, T. L. (1978). Homogeneity in coverage of Connecticut Newspapers. *Journalism Quarterly*, *55*(3), 592–596.

Dutton, W. H., Reisdorf, B. C., Dubois, E., & Blank, G. (2017). *Search and politics: The uses and impacts of search in Britain, France, Germany, Italy, Poland, Spain, and the United States* (Quello Center Working Paper No. 5-1-17).

Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS* August 18, 112 (33) E4512-E4521.

European Commission (2010). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A digital agenda for Europe*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52010DC0245R(01)&from=EN

Fletcher, R., & Nielsen, R. K. (2018). Automated serendipity: The effect of using search engines on news repertoire balance and diversity. *Digital Journalism, 6*(8), 976–989.

Fuchs, C. (2019, March 11). Geplanter Medienstaatsvertrag: Mischung für Millionen [Draft of the ‚Interstate Treaty on Media‘: Mixture for millions]. *sueddeutsche.de*. https://www.sueddeutsche.de/medien/geplanter-medienstaatsvertrag-mischung-fuer-millionen-1.4364339

Geiß, S. (2015). *Die Aufmerksamkeitsspanne der Öffentlichkeit: Eine Studie zur Dauer und Intensität von Meinungsbildungsprozessen* [The public attention span: A study on the duration and intensity of opinion-forming processes]. Baden-Baden: Nomos.

Gerhart, S. (2004). Do web search engines suppress controversy? *First Monday, 9*(1).

Granka, L. (2010). The politics of search: A decade retrospective. *The Information Society, 26*(5), 364–374.

Haas, A., & Unkel, J. (2017). Ranking versus reputation: Perception and effects of search result credibility. *Behaviour & Information Technology*, *36*(12), 1285–1298.

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism*, *6*(3), 330–343.

Hannák, A., Sapieżyński, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. *Proceedings of the 22nd International Conference on World Wide Web, WWW '13* (pp. 527–537). ACM.

Hargittai, E. (2000). Open portals or closed gates? Channeling content on the world wide web. *Poetics*, *27*(4), 233–253.

Haßler, J., Maurer, M., & Oschatz, C. (2014). Media logic and political logic online and offline. *Journalism Practice*, *8*(3), 326–341.

Helberger, N. (2011). Diversity by design. *Journal of Information Policy*, *1*, 441–469.

Hu, D., Jiang, S., E., Robertson, R., & Wilson, C. (2019). Auditing the Partisanship of Google Search Snippets. *Proceedings of The World Wide Web Conference, WWW '19* (pp. 693–704). ACM.

Humprecht, E., & Esser, F. (2018). Diversity in online news. *Journalism Studies*, *19*(12), 1825–1847.

Ieong, S., Mishra, N., Sadikov, E., & Zhang, L. (2012). Domain bias in web search. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM '12* (pp. 413–422). ACM.

Jacobi, C., Kleinen-von Königslöw, K., & Ruigrok, N. (2016). Political news content in online and print newspapers: Are online editions better by electoral democratic standards? *Digital Journalism*, *4*(6), 723–742.

Jandura, O., & Friedrich, K. (2014). The quality of political media coverage. In C. Reinemann (Ed.), *Political communication* (pp. 351–373). Berlin: de Gruyter.

Jansen, B., & Spink, A. (2006). How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing and Management*, *42*(1), 248–263.

Just, N. (2009). Measuring media concentration and diversity: New approaches and instruments in Europe and the US. *Media Culture Society*, *31*, 97–117.

Karppinen, K. E. (2006). Media diversity and the politics of criteria: Diversity assessment and technocratisation of European media policy. *Nordicom Review*, *27*(2), 53–68.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2019). Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal, 22*(1), 188–227.

Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The economics of algorithmic selection on the Internet. In J. M. Bauer & M. Latzer (Eds.), *Handbook on the economics of the Internet* (pp. 395–425). Cheltenham, Northampton: Edward Elgar.

Li, N., Anderson, A. A., Brossard, D., & Scheufele, D. A. (2014). Channeling science information seekers' attention? A content analysis of top-ranked vs. lower-ranked sites in Google. *Journal of Computer-Mediated Communication*, *19*(3), 562–575.

Masini, A., van Aelst, P., Zerback, T., Reinemann, C., Mancini, P., Mazzoni, M., . . . Coen, S. (2018). Measuring and explaining the diversity of voices and viewpoints in the news. *Journalism Studies*, *19*(15), 2324–2343.

McDonald, D. G., & Dimmick, J. (2003). The conceptualization and measurement of diversity. *Communication Research*, *30*(1), 60–79.

McMahon, C., Johnson, I., & Hecht, B. (2017). The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media, ICWSM 2017* (pp. 142–151).

McQuail, D. (1992). *Media performance: Mass communication and the public interest*. London, Thousand Oaks, New Delhi: Sage.

Metaxas, P. T., & Pruksachatkun, Y. (2017). Manipulation of search engine results during the 2016 US congressional elections. *Proceedings oft he ICIW 2017*, Venice, Italy.

Möller, J., Helberger, N., & Makhortykh, M. (2019). *Filter bubbles in the Netherlands?* https://www.ivir.nl/publicaties/download/Filter-bubbles-in-the-Netherlands.pdf

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959–977.

Mowshowitz, A., & Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing & Management, 38*(1), 141–156.

Napoli, P. M. (1999). Deconstructing the diversity principle. *Journal of Communication*, *49*(4), 7–34.

Neuberger, C., & Lobigs, F. (2010). *Die Bedeutung des Internets im Rahmen der Vielfaltssicherung: Gutachten im Auftrag der Kommission zur Ermittlung der Konzentration im Medienbereich* [The role of the Internet for safeguarding diversity. Report on behalf of the commission of media concentration monitoring]. Berlin: Vistas.

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_0.pdf

O'Hare, R. (2016, June 24). Google search spike suggests many people don't know why they voted for Brexit. *Daily Mail*. http://www.dailymail.co.uk/sciencetech/article-3657997/Britain-s-Google-searches-hint-people-didn-t-know-voting-for.html

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823.

Pariser, E. (2011). *The Filter Bubble: What the internet is hiding from you*. London: Penguin Books.

Quarfoot, D., & Levine, R. A. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, *70*(4), 373–384.

Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW).

Saurwein, F., Just, N., & Latzer, M. (2015). Governance of algorithms: Options and limitations. *info*, *17*(6), 35–49.

SEO (2017). Suchmaschinen – Liste & Marktanteile: Entwicklung und Marktanteile der beliebtesten Suchmaschinen in Deutschland [Search engines and their market shares: Development and market shares of the most popular search engines in Germany]. https://seo-summary.de/suchmaschinen/

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Trevisan, F., Hoskins, A., Oates, S., & Mahlouly, D. (2018). The Google voter: Search engines and elections in the new media ecology. *Information, Communication & Society, 21*(1), 111–128.

Trielli, D., & Diakopoulos, N. (2019). Search as news curator: The role of Google in shaping attention to news information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI'19*. ACM.

Unkel, J., & Haim, M. (2019). Googling politics: Parties, sources, and issue ownerships on Google in the 2017 German Federal Election campaign. *Social Science Computer Review*. Published online ahead of print.

van Cuilenburg, J. (2000). On measuring media competition and media diversity: Concepts, theories and methods. In R. G. Picard (Ed.), *Measuring media content, quality and diversity: Approaches and issues in content research* (pp. 51–84). Turku, Finland: Turku School for Economics and Business Administration.

van Cuilenburg, J. (2007). Media diversity, competition and concentration: Concepts and theories. In E. de Bens, C. Hamelink, K. Jakkubowicz, K. Nordenstreng, J. van Cuilenburg, & R. van der Wurff (Eds.), *Media between culture and commerce* (pp. 25–54). Bristol, Chicago: intellect.

van Hoboken, J. (2012). *Search engine freedom: On the implications of the right to freedom of expression for the legal governance of web search engines*. https://pure.uva.nl/ws/files/1769429/104098_thesis.pdf

Vettehen, P. H. (2005). 'Open diversity' statistics: An illusion of 'scientific thoroughness'? *Communications*, *30*(3), 308–312.

Tables and Figures

Table 1. Classification of issues, date of data collection and search queries

Issue	Classification of issues			Date of data collection	Search queries	
	Level of relevance	*Societal area*	*Occasion of coverage*		*German*	*English translation*
Brexit	supranational	international policy	political proceeding, single event	06/27/2016	Brexit	Brexit
Böhmermann's poem[1]	binational	media policy, judiciary, culture, international political relationship	single event	04/25/2016	Böhmermann Gedicht	Böhmermann poem
Climate change	global	science	ongoing debate	12/21/2015	Klimawandel	climate change
Military intervention in Syria	global	international policy	single event	12/18/2015	Einsatz Syrien	intervention Syria
NSU judicial trial[2]	national	judiciary	series of events, judicial trials	12/18/2015	NSU Prozess	NSU judicial trial
Panama Papers[3]	global	economy, economic policy	update of latent issue	04/20/2016	Panama Papers	Panama Papers
Refugees	global	international policy	ongoing debate	11/16/2015	Flüchtlinge	refugees
Regional elections in Rhineland-Palatinate[4]	regional	regional policy	political proceedings	02/15/2016	Landtagswahl RLP	regional election RLP
(Sexual) assaults in Cologne[5]	national	society	single event	01/13/2016	Übergriffe Köln	assaults Cologne
TTIP	global	international policy, economy	ongoing debate	05/15/2016	TTIP	TTIP

[1] a poem by the German satirist Jan Böhmermann insulting the Turkish president Recep Tayyip Erdoğan, resulting in an interstate conflict

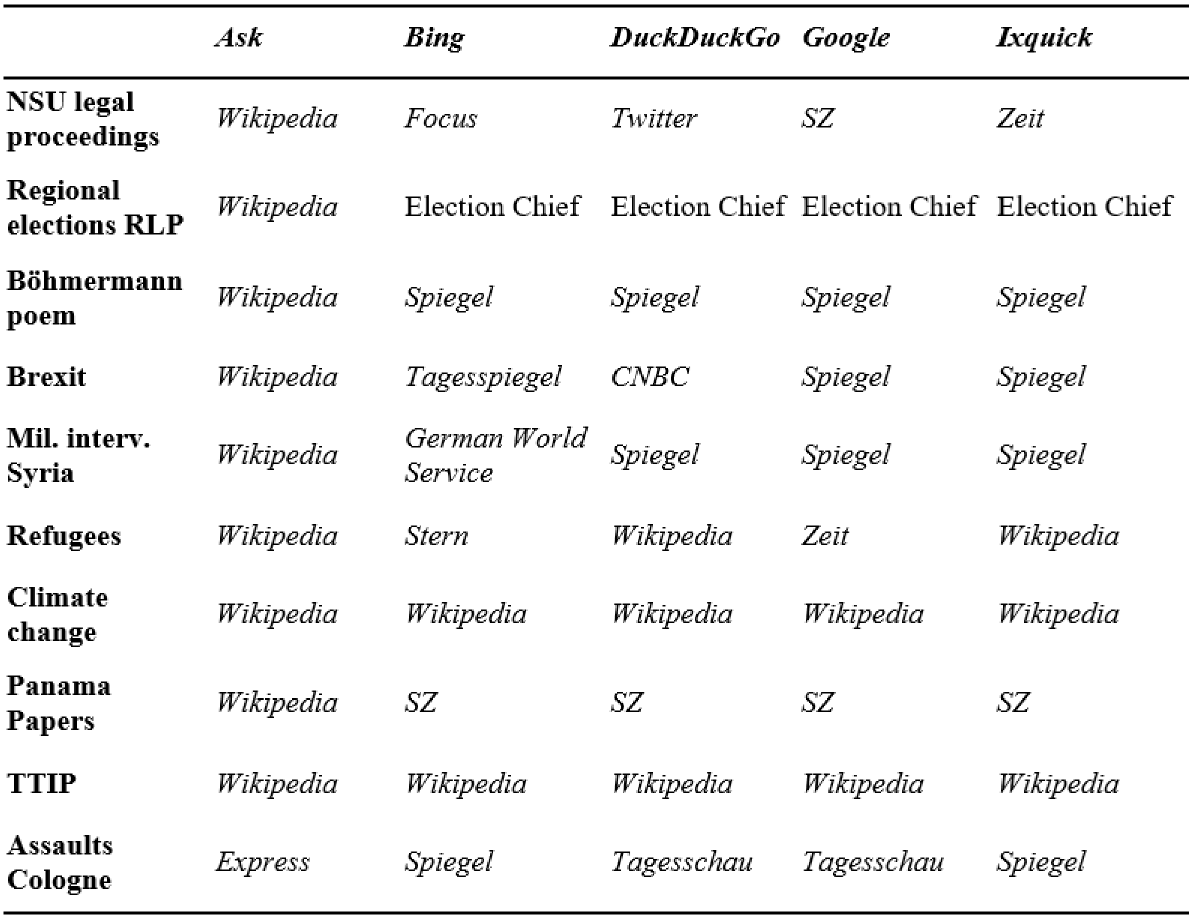
[2] judicial trial against members of the German terrorist group 'NSU' (National Socialist Underground)

[3] journalists revealing a large-scale, international financial scandal

[4] federal state ("Bundesland") in Germany

[5] a huge number of sexual assaults during a celebration of New Year's Eve primarily committed by African and Arabian people; this intensified the debate about refugees and the role of media

Table 2. First search results



	Ask	***Bing***	***DuckDuckGo***	***Google***	***Ixquick***
NSU legal proceedings	*Wikipedia*	*Focus*	*Twitter*	*SZ*	*Zeit*
Regional elections RLP	*Wikipedia*	Election Chief	Election Chief	Election Chief	Election Chief
Böhmermann poem	*Wikipedia*	*Spiegel*	*Spiegel*	*Spiegel*	*Spiegel*
Brexit	*Wikipedia*	*Tagesspiegel*	*CNBC*	*Spiegel*	*Spiegel*
Mil. interv. Syria	*Wikipedia*	*German World Service*	*Spiegel*	*Spiegel*	*Spiegel*
Refugees	*Wikipedia*	*Stern*	*Wikipedia*	*Zeit*	*Wikipedia*
Climate change	*Wikipedia*	*Wikipedia*	*Wikipedia*	*Wikipedia*	*Wikipedia*
Panama Papers	*Wikipedia*	*SZ*	*SZ*	*SZ*	*SZ*
TTIP	*Wikipedia*	*Wikipedia*	*Wikipedia*	*Wikipedia*	*Wikipedia*
Assaults Cologne	*Express*	*Spiegel*	*Tagesschau*	*Tagesschau*	*Spiegel*

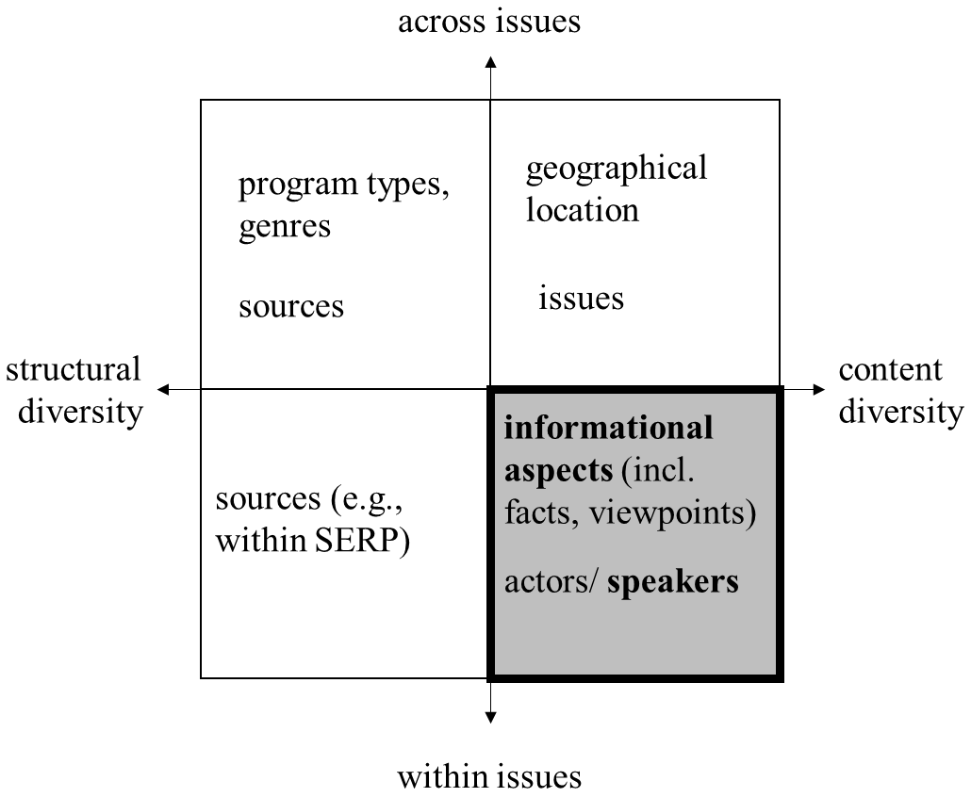


Figure 1. Conceptualization of diversity.

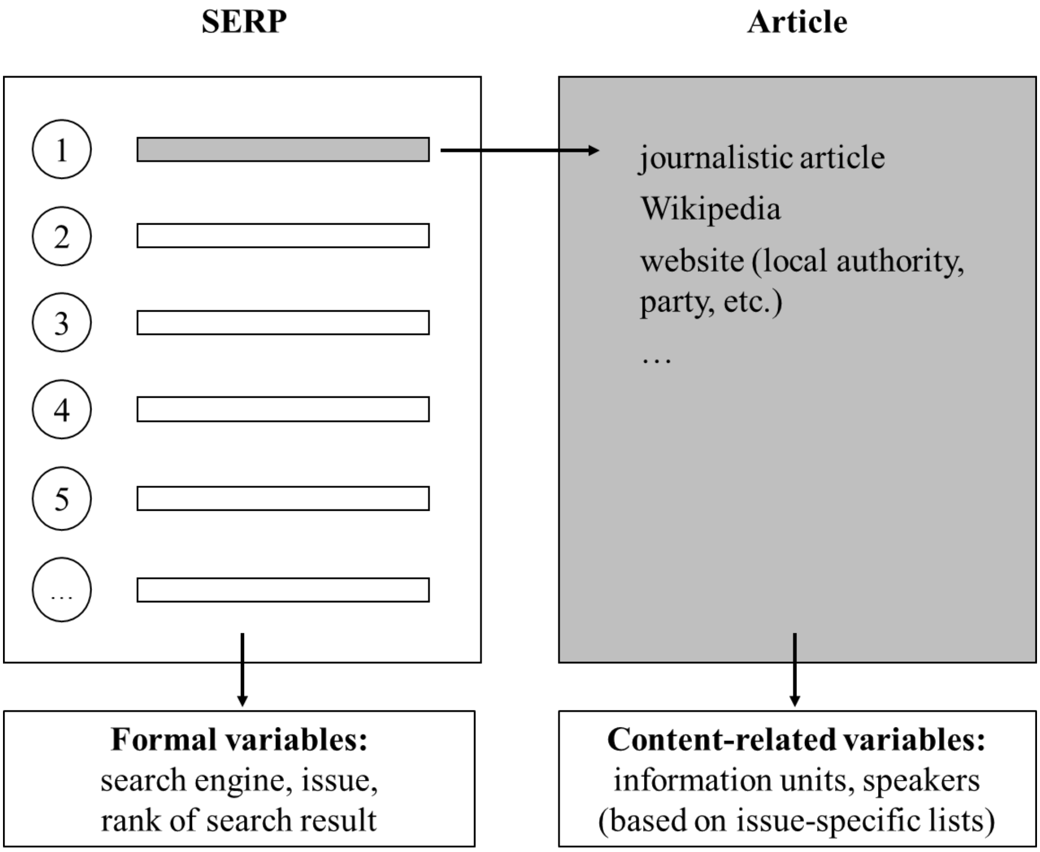

SERP
Article
1
2
3
4
5
…
journalistic article
Wikipedia
website (local authority, party, etc.)
…
Formal variables:
search engine, issue, rank of search result
Content-related variables:
information units, speakers (based on issue-specific lists)

Figure 2. Units of analysis.

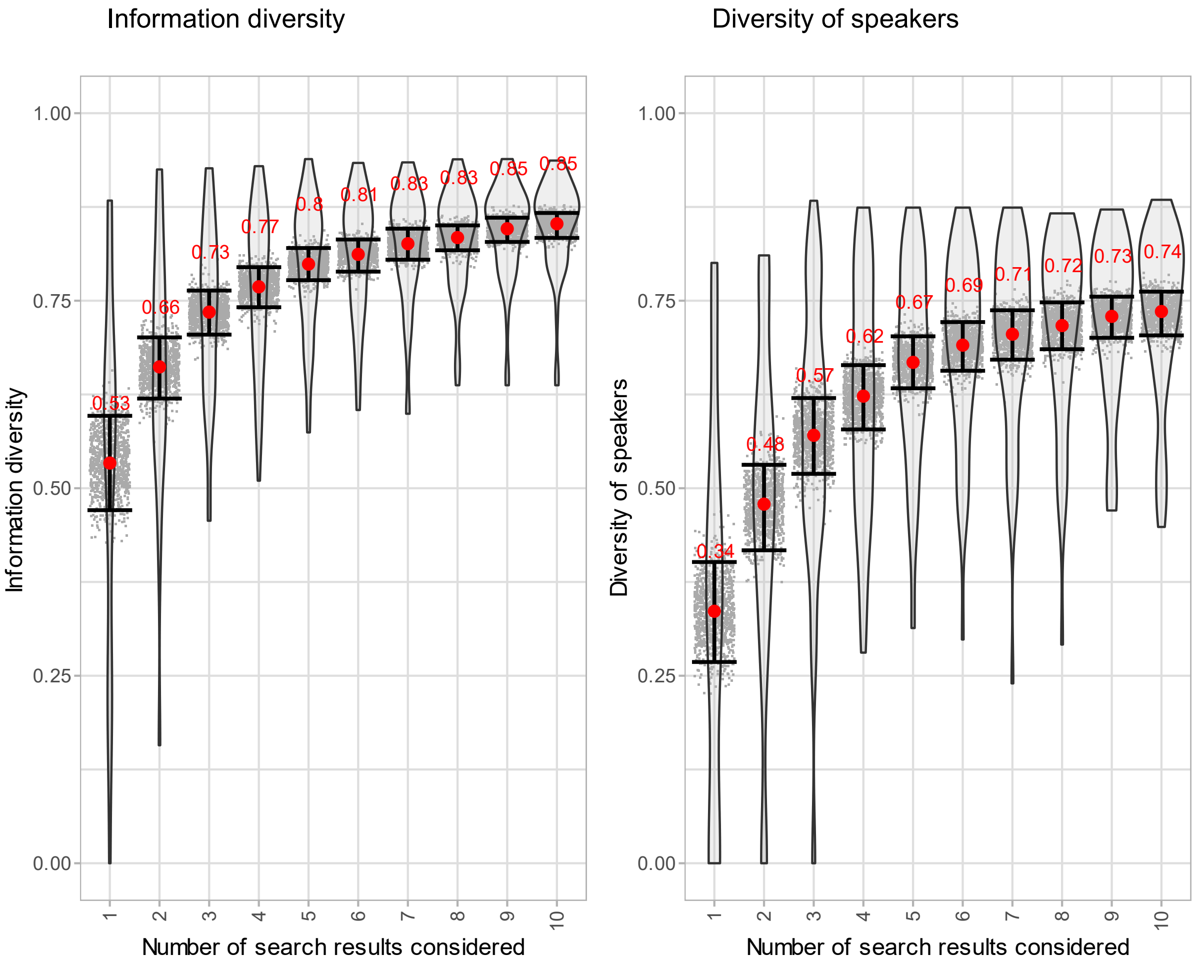


Figure 3. Progression of diversity.
95% Confidence intervals constructed from bootstrapping with 10000 replicates; 1000 randomly sampled replicates are displayed as grey dots. The raw distrubtion of diversity scores is displayed as a "violin" plot.

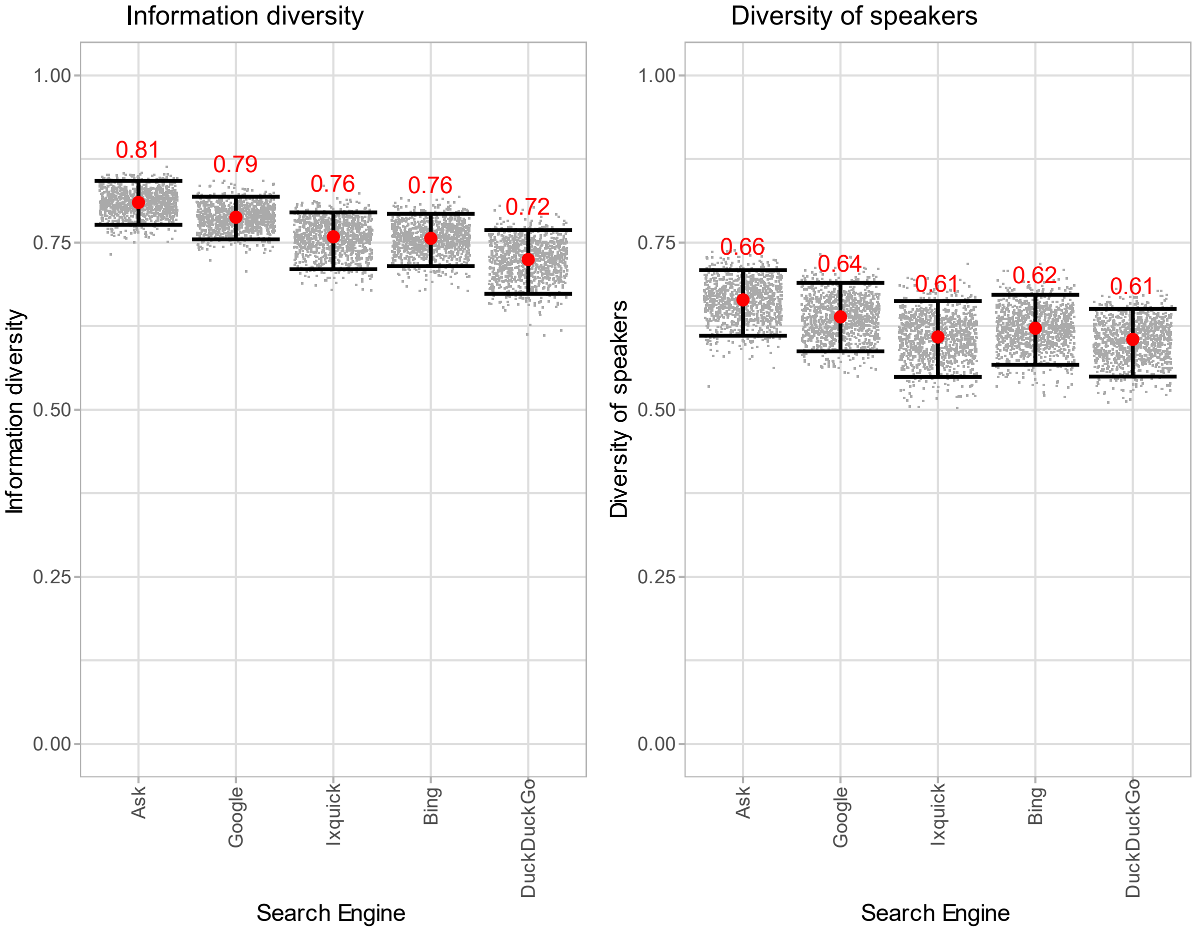


Figure 4. Diversity by search engine.
95% Confidence intervals constructed from bootstrapping with 10000 replicates; 1000 randomly sampled replicates are displayed as grey dots. The raw distrubtion of diversity scores is not displayed. Search engines are sorted descendingly regarding their information diversity.

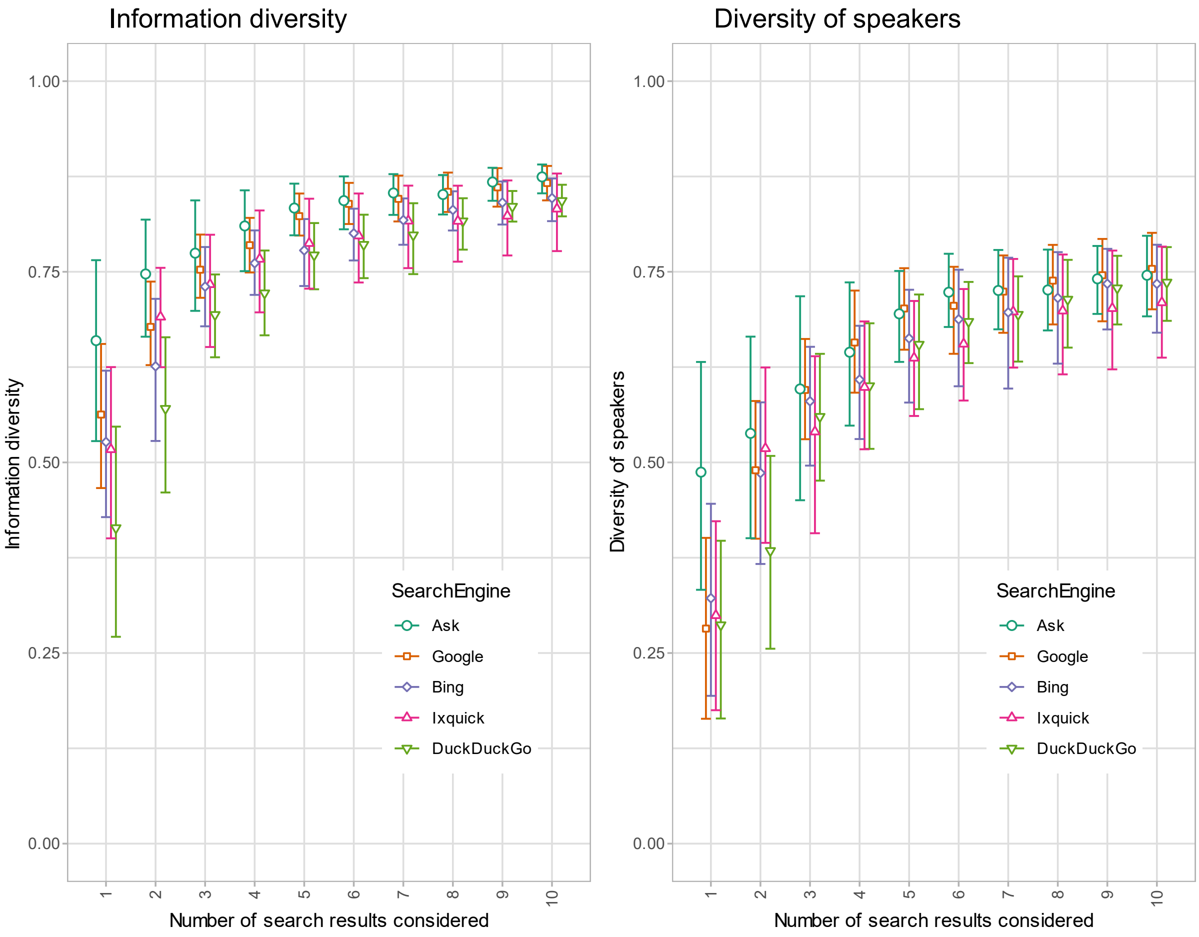


Figure 5. *Progression of diversity by search engine.*
95% Confidence intervals constructed from bootstrapping with 10000 replicates. The raw distrubtion of diversity scores is not displayed.

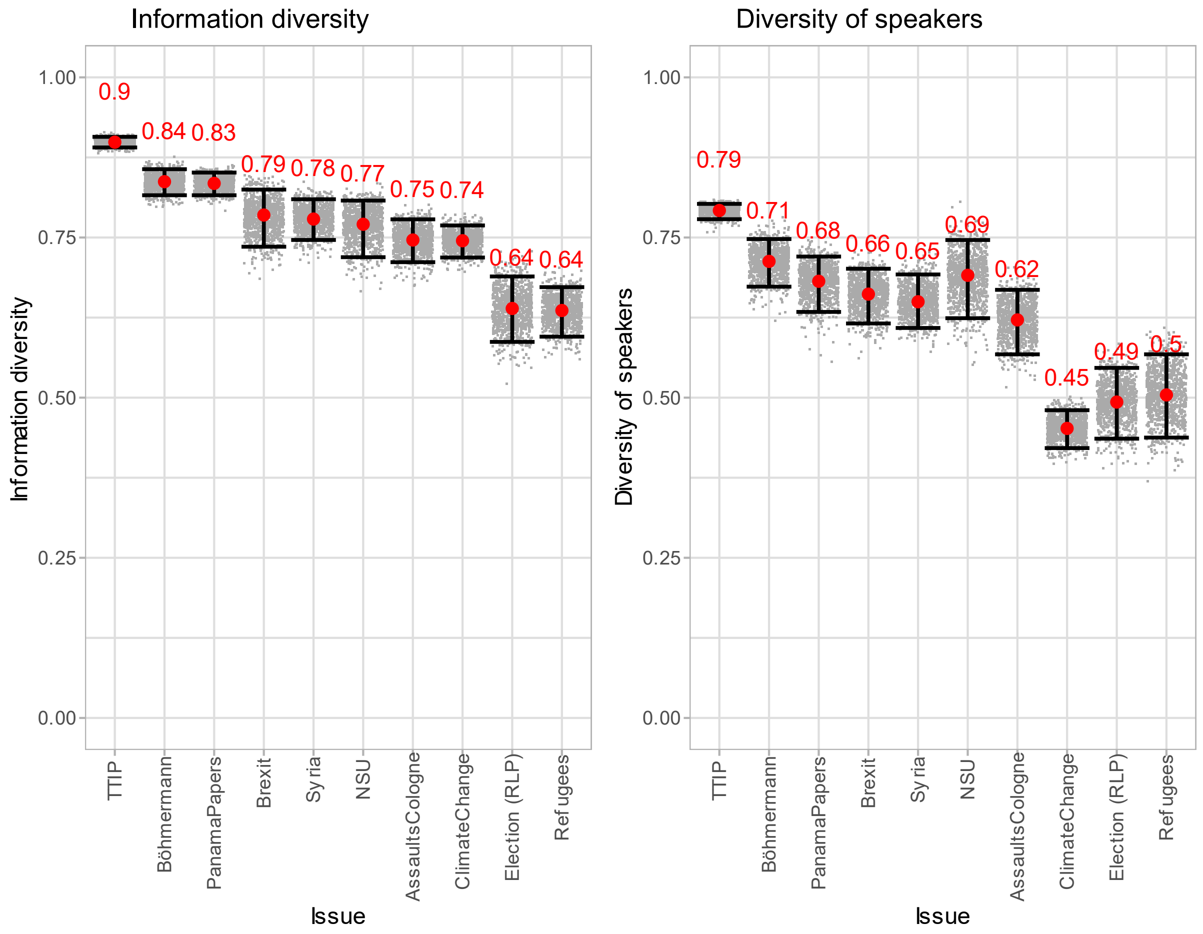


Figure 6. Diversity by issue.
95% Confidence intervals constructed from bootstrapping with 10000 replicates; 1000 randomly sampled replicates are displayed as grey dots. The raw distrubtion of diversity scores is not displayed. Issues are sorted descendingly regarding their information diversity.

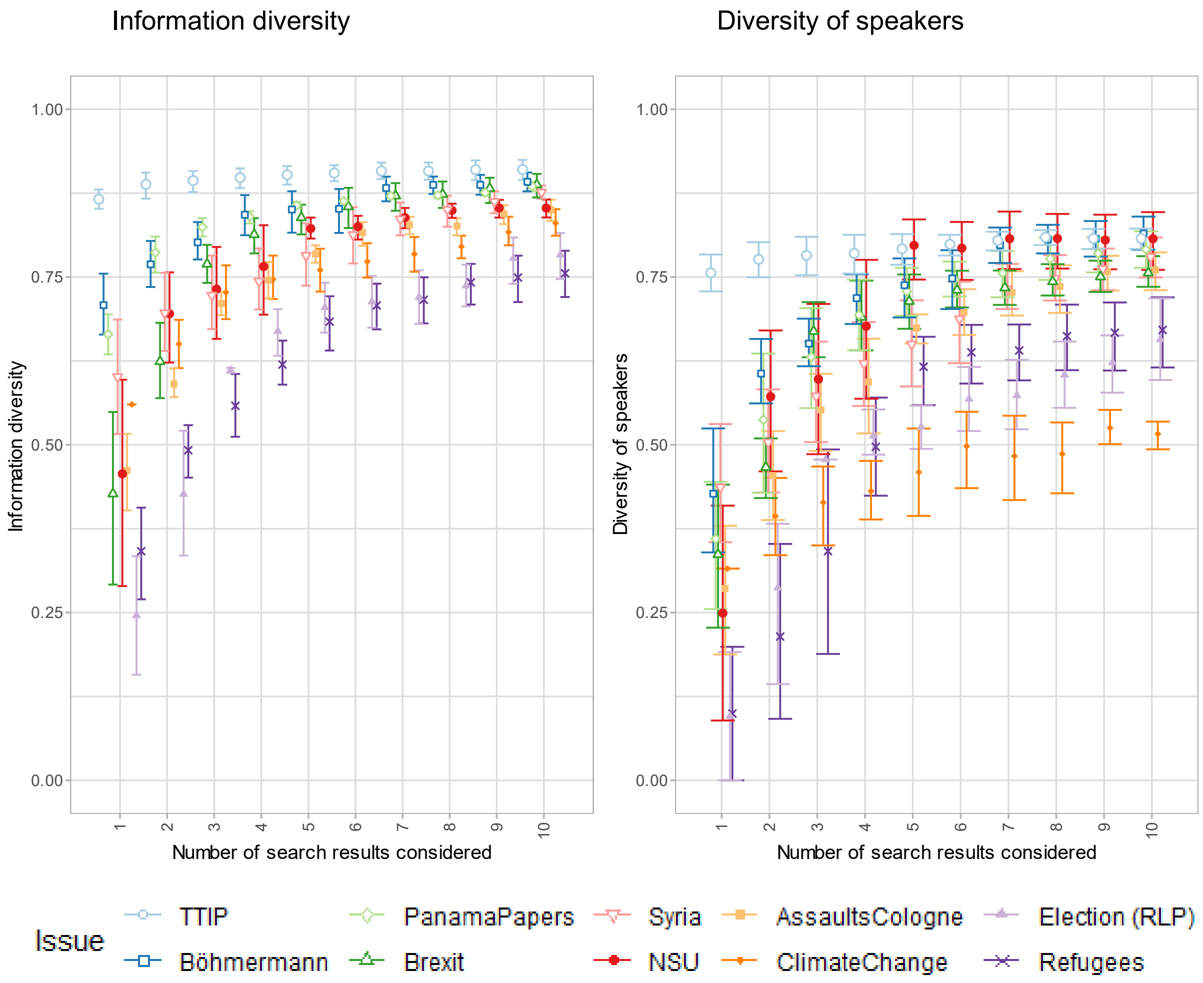


Figure 7. *Progression of diversity by issue.*
95% Confidence intervals constructed from bootstrapping with 10000 replicates. The raw distrubtion of diversity scores is not displayed.

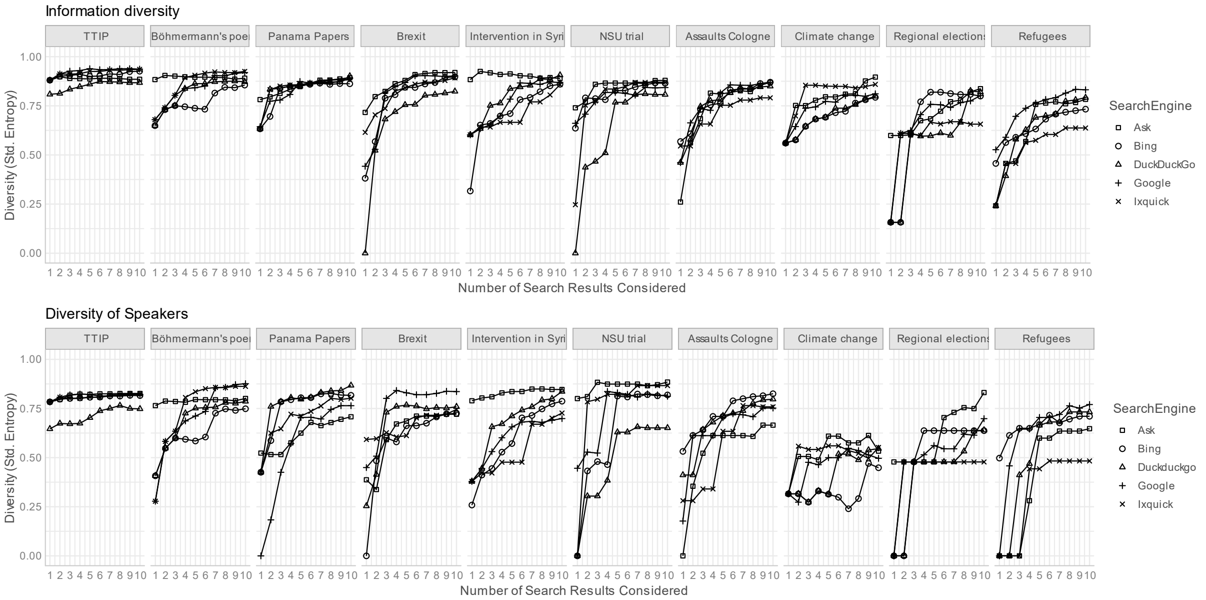


Figure 8. Progression of diversity by issue and search engine. Issues are sorted descendingly regarding their information diversity.