Analyzing Statistical Tests of Search Engine Bias

Patrick Maillé, Nils Peyrouset, Bruno Tuffin

Analyzing Statistical Tests of Search Engine Bias

Patrick Maillé
IMT Atlantique, IRISA, UMR CNRS 6074
F-35700 Rennes, France
patrick.maille@imt.fr

Nils Peyrouset
ENSAE Paris
F-91120 Palaiseau, France
nils.peyrouset@ensae.fr

Bruno Tuffin
Inria, Univ. Rennes, IRISA
F-35700 Rennes, France
bruno.tuffin@inria.fr

December 21, 2023

Abstract

Search engines play a significant role in shaping the Internet, but there are concerns about the motivation behind their rankings: some content and service providers have complained about being ranked abusively low to favor engines' commercially-close content, initiating the so-called search neutrality debate. There are similar worries about possible orientations towards some ideologies or political points of view. Whatever the opinion of the reader on this sensitive issue of the definition and need for search engine neutrality, it is important to have at our disposal tools to monitor search engines behavior and understand deviations from "expected" results. The goal of this paper is to review existing *statistical* tests of potential bias by search engines, compare and combine them both formally and numerically. We end up with a battery of ANOVA and Dixon Q tests for which we characterize the bias detection probability and false positive probability, the latter requiring to be minimal if tests have to be used officially and publicly. We also run the tests on a campaign of searches on real-life search engines and discuss the outcome.

1 Introduction

The Internet is now omnipresent in our lives. Its success is mostly due, if we forget about the technical aspects, to its free and open characteristic: anything can be easily reached from everywhere. But in many cases, reaching a "relevant" content is done through so-called *search engines* which provide an ordered list of links in response to a dialed request term. Search engines are therefore key actors in the digital world landscape. It is for example estimated that Google, the most used search engines in the world, processes in 2023 approximately 99,000 search queries every second, translating to 8.5 billion searches per day and approximately 2 trillion global searches per year[1]. Search engines therefore shape the way the Internet is used and what is easily reached. As a consequence, there are natural questions about their ranking algorithms and how "neutral" or selfless they are with respect to the displayed content. It is the so-called *search neutrality debate* [11], echoing the *network neutrality* debate on the behavior of Internet Service Providers (ISPs) which were accused of prioritizing some traffic or block others and therefore of providing an unfair competitive advantage to preferred content or applications [15]. Search engines are another type of important gatekeepers between users and content, and some consider that a neutrality principle should similarly apply to them. The expression " search neutrality" is traced to have been first used by Andrew Odlyzko in [17] in 2009, predicting that it will become the next Internet issue when network neutrality is settled.

[1] https://blog.hubspot.com/marketing/google-search-statistics

The potential occurrence of bias in the rankings was raised in 2009 by Adam Raff, a co-founder of Foundem–a "vertical search" service whose goal is to help consumers by offering a comparison of several online offers. Raff indeed complained about being ranked abusively low by Google with respect to Google Shopping, accusing the company of stifling competition. Google's chairman Eric Schmidt had to appear before a Senate committee [18] and the Federal Trade Commission (FTC), the national regulator [2]. The FTC closed the case, stating that Google did not violate U.S. antitrust laws, since the algorithm purpose was to improve users' experience and not to hurt competitors, even if it does. The European commission did not fully agree with the FTC statement and acted differently: in 2017 it fined Google €2.42 billion for breaching European Union (EU) antitrust rules. Google was declared to have abused its market dominance as a search engine by giving an illegal advantage to another Google product, its own comparison shopping service [9]. Note that the EU is still very active on the topic in a broad sense, releasing in 2020 the Digital Markets Act (DMA) [4] which "aims to guarantee a competitive and fair digital sector, allowing innovative digital businesses to grow and ensuring the safety of users online". The largest search engines Bing and Google enter the framework, and must satisfy requirements about transparency and accountability. The regulation deals not only with economic-oriented potential bias but also with politics or ideology-oriented bias, another aspect on which search engines can have an impact on society by orienting beliefs[2]. Empirical evidence of the potential impact of such bias has been shown in [8] in a lab. In 2018 though, a study over Google on candidates running for the U.S. elections showed no evidence of political bias[3]. It is in any case a topic worthy of supervision, and there are techniques to compare the results produced by two search engines on conservative/liberal perspective towards a given controversial topic [10].

Whatever the belief of the reader on SEs needing or not to be neutral, it is better in any case to have tools at our disposal to monitor their activities and understand their behavior. It is indeed of interest for society/institutions to understand their impact on people's daily life, and for end users to make an informed choice in a (potentially) competitive environment and go with the SE that better corresponds to what they wish.

A natural or expected behavior of a SE is something hard to characterize: what does it mean to be neutral? We typically expect the SE to provide a ranked list of *relevant* links for the request, but the meaning of relevance can be seen as something subjective. There are numerous factors in the notion of relevance, such as search intent, location and personalization, among others. Establishing whether a search result is irrelevant or not is therefore a difficult, if not dubious, task. Any chosen definition will probably be controversial, and a consensus will doubtfully be reached, so we choose to follow some papers of the literature which instead look at the differences between rankings of existing SEs [16, 14, 20]. The purpose is therefore to detect outliers in SE results. *Outlier detection* is the principle of detecting abnormal or unusual results in a bunch of existing ones [1, 12], a deviation from others being potentially (but not necessarily) due to a controversial motivation.

That is our purpose in this paper: detecting different results from a search engine which could later be singled out for a deeper investigation. But while many outlier detection techniques are based on heuristics, we focus here on *non-heuristic* statistical tests, describe and compare the existing ones. The list of tests we use is based on two papers, [16] and [14], the ones we have noted to have applied statistical tests. In [16] the difference of the results produced by an SE with respect to the *aggregated* SEs is computed thanks to the *cosine similarity*, a standard measure in information retrieval [19]. The significance of differences is measured thanks to analysis of variance (ANOVA) tests [3]. On the other hand, [14] computes a weighted score of web pages (or links) from their ranking (visibility) over all SEs and uses those page scores to define a score for each SE for all their rankings. Several versions of Dixon's Q test [5, 7], designed to point out outliers on small-sized samples, are then performed. The two papers [16] and [14] make use of different but according to us valuable notions of bias/relevance and two different types of statistical tests. Our contributions in the present paper are therefore several:

- While the notion of bias in [16] does not take the results *position* into account but only

[2] `https://www.forbes.com/sites/kevinanderton/2020/12/13/middle-schooler-proves-google-search-results-influence-political-opinions-infographic/`

[3] `https://engineering.stanford.edu/magazine/article/are-search-results-biased-along-partisan-lines`

whether they appear or not among the results, the score in [14] assigns a weight to each position, corresponding to its *click-through-rate* (CTR) that is its probability to be clicked, representing the importance or visibility of each position; we extend the notion of bias to account for this importance.

- We then point out that each notion of bias/relevance can be combined with the two types of test (ANOVA and Dixon Q) in the literature. We discuss how these combinations can be applied and make a formal comparison.
- We compare all combinations on synthetic data. The goal is to artificially build a biased SE and compare the probability of detection when the real relevances of web pages are random; the higher the probability, the better. We similarly study the probability of false positive detection when actually no SE is voluntarily biased. Limiting this probability is important if one wants to point at an SE and avoid being a wrong accuser, which could lead to defamation charges.
- We also perform the tests and compare them on a set of real searches.

The remaining of the paper is organized as follows. Section 2 reviews the two developed tests (and the used bias/relevance notions) in the literature, namely in [16] and [14]. Section 3 extends the notion of bias in [16] by adding a weight to the position corresponding to its importance and Section 4 describes the various combinations of notions with the different tests and makes a formal comparison. Section 5 tests the efficiency of SE bias detection methods for all the *(metric, test)* combinations on synthetic data. Section 6 performs the same combination of tests on a campaign of real searches. Section 7 concludes and gives some directions for future research.

2 Outlier tests in the literature

The purpose of this section is to review the two notable existing tests of bias in the literature. Each work is made of two main components (on which we will later test all the different combinations): the considered notion of bias and the statistical test which is performed. We are going to describe both.

We start by introducing the main notations in order to uniformize the presentation of the tests. Let $\mathcal{I} = \{1, \dots, I\}$ be the set of SEs, of cardinality I, $\mathcal{L}$ be the set of pages/links that can be found on the Internet. Let $\mathcal{K}$ be a set of n queries corresponding to a given domain/theme for which we want to investigate if some SEs' output can be suspected to be biased.

2.1 Bias based on cosine similarity and ANOVA

We formally describe here the test developed in [16], described here to fit the notations of [14].

2.1.1 Notion of bias

The notion of bias is defined for a subset $\mathcal{T} \subset \mathcal{K}$ of t queries, even if in practice later we will see that often $\mathcal{T}$ will be of cardinality $t = 1$.

Let $v_{i,\ell,k} = 1$ if Page $\ell \in \mathcal{L}$ is displayed by SE $i \in \mathcal{I}$ for query $k \in \mathcal{T}$, and 0 otherwise. Let also

$$v_{i,\ell} := \sum_{k \in \mathcal{T}} v_{i,\ell,k} \tag{1}$$

be the number of times Page $\ell \in \mathcal{L}$ is displayed by SE $i \in \mathcal{I}$ for the set $\mathcal{T}$, and

$$v_{\ell} := \sum_{i \in \mathcal{I}} v_{i,\ell}$$

be the total number of times Page ℓ is displayed by the full list of SEs over all searches in $\mathcal{T}$.

The notion of bias in [16] of SE i is mathematically defined by

$$b_i(\mathcal{T}, \mathcal{I}) := 1 - \frac{\sum_{\ell \in \mathcal{L}} v_{i,\ell} v_{\ell}}{\left(\sum_{\ell \in \mathcal{L}} v_{i,\ell}^2 \sum_{\ell \in \mathcal{L}} v_{\ell}^2 \right)^{1/2}} \tag{2}$$

The term $\frac{\sum_{\ell\in\mathcal{L}} v_{i,\ell}v_\ell}{\left(\sum_{\ell\in\mathcal{L}} v_{i,\ell}^2 \sum_{\ell\in\mathcal{L}} v_\ell^2\right)^{1/2}}$ is the cosine similarity, often used in information retrieval [13]. A similarity measure is a function which computes the degree of similarity between a pair of vectors (or documents). The vectors being compared are here (indexed over pages) the number of times each page is displayed by a singled out SE (SE i), and by all SEs, i.e., respectively $(v_{i,\ell})_{\ell\in\mathcal{L}}$ and $(v_\ell)_{\ell\in\mathcal{L}}$. The idea is to measure the differences, in proportion, of occurrence of each page for SE i with respect to the full set of SEs. Cosine similarity between two vectors is between 0 and 1, and equals 1 for collinear vectors, corresponding here to the same proportion of display of each page in both vectors. The $1-\cdot$ operation is for the bias to be 0 if results are similar and 1 if they are totally orthogonal.

This notion of bias is used in the next subsection to statistically measure if an SE is significantly different from the others.

2.1.2 ANOVA tests

The tests used in [16] are ANOVA tests [3]. Since they are long to rigorously describe, we prefer to refer the reader to Appendix A for a full and formal presentation, and limit ourselves in this subsection to the main principles; it avoids to divert from the main message.

There are so-called one-way and two-way ANOVA tests [3]. Remark though that both types of tests assume the data to be normally-distributed with the same (although unknown) variance σ^2, something not verified in practice by our data made of bias values.

The *one-way ANOVA* test in [16] considers domains separately. In each domain, it computes the bias (for a given SE) for query $k \in \mathcal{K}$ (hence considering $t = 1$ in (1) and n queries yielding n bias values). It then tests for the given domain:

H_0: "the mean bias is the same for each of the n queries" versus H_1 : "not all biases are equal".

Under the normal assumption, we can get critical values for a statistics, above which assumption H_0 is rejected at the wished significance level.

The *two-way ANOVA* test studies the effects of several factors instead of one. (Table 2 in [16] considers in our context as effects the SE and the query, and shows test results on each domain.) It builds a statistics to test, again for a wished significance level

H_0: "SEs have no impact on bias values" versus H_1 : "The are differences over SEs",

and a similar one to test

H_0: "Queries have no impact on bias values" versus H_1 : "The are differences over queries".

Finally, another ANOVA test can be applied, the *two-way ANOVA with interaction*. The previous two-way ANOVA considers no interaction between factors, assumed additive. This is corrected by considering a sample of biases for each domain and each SE. Data are samples of n biases computed per SE and per domain for n queries, each bias being again computed with $t = 1$. Three null hypotheses H_0 are tested:

- H_0^r: "no SE effect"
- H_0^c: "no domain effect"
- H_0^{int}: "no SE and domain interaction".

Again, each time statistics are built and rejection tests available at a pre-specified confidence level.

2.1.3 Tests application

In [16], the tests have been applied in the following ways (identifying here the data over which tests are run, that are generically denoted by X_{rj} in the test descriptions in the Appendix):

- 16 SEs (at the time) were identified and 25 queries organized in 5 queries of 5 domains, from which 25 biases were computed for each SE ($t = 1$ then). An interesting one-way ANOVA test is performed for each domain: investigating whether all SEs provide similar biases: $X_{i,k}$ is the bias for SE i and keyword k.
- For each domain, based on the 16×5 biases for the 16 SEs and 5 queries, Table 2 of [16] tests assumption H_0 that the biases are the same for the different SEs as well as H_0: the keywords have similar biases.

- Two-way ANOVA tests are applied in Table 3 of [16] since domains (or SEs) may experience different biases. $X_{r,j}$ is here the bias for SE r on domain j obtained from the 5 queries. H_0 is therefore either "SEs have no impact on bias" (they yield similar values of bias) and/or "domains have no impact on bias values".

2.2 SNIDE

Paper [14] on the other hand tries to detect an abnormal output by an SE with respect to the others, based on Dixon's Q test.

2.2.1 Notions of score

A notion of score is introduced in [14] and defined for a page/link as its average visibility among SEs, where the visibility of a position in the ranking is assumed proportional to its average observed *click-through-rate* (CTR). Note that we consider the visibility of a given position is the same for all SEs, but this can be generalized. Let q_π be the visibility value for position π, with $q_1 \geq q_2 \geq \cdots$. In this paper, as in [14] we take for those visibility values the CTR values obtained through measurements in [6] and displayed in Table 1.

Table 1: CTR values used in the paper, taken from [6]

q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}
0.364	0.125	0.095	0.079	0.061	0.041	0.038	0.035	0.03	0.022

The score $R_{\ell,k}$ of a page $\ell \in \mathcal{L}$ for a search term $k \in \mathcal{K}$ is defined as the average visibility of that page over the rankings from all considered SEs for that search term:

$$R_{\ell,k} = \frac{1}{I} \sum_{i=1}^{I} q_{\pi_i(\ell,k)} \tag{3}$$

where $\pi_i(\ell, k)$ denotes the position of page ℓ on SE i for query k.

From the definition of the score for pages, we can define a score $S_{i,k}$ for SE i, for a given search term k:

$$S_{i,k} = \sum_{\ell \in \mathcal{L}} q_{\pi_i(\ell,k)} R_{\ell,k}. \tag{4}$$

It is the total "page score visibility" of its results for that search term, quantifying how "consensual" SE i is with respect to the other SEs. That metric can be used to check if an SE has a score (corresponding to visibilities of presented pages) abnormally low with respect to others.

All the tests built from the scores of SEs or pages are based on Dixon's Q test.

2.2.2 Dixon's Q test

Dixon's Q test [5, 7] aims at detecting an abnormal value on a *small sample*. Assume we have I values, here too assumed normally distributed and sorted in increasing order, $x_1 \leq x_2 \leq \cdots \leq x_I$.

The statistics

$$Q = \begin{cases} Q_{10} = \frac{x_I - x_{I-1}}{x_I - x_1} & \text{if } 3 \leq I \leq 7 \\ Q_{11} = \frac{x_I - x_{I-1}}{x_I - x_2} & \text{if } 8 \leq I \leq 10 \\ Q_{21} = \frac{x_I - x_{I-2}}{x_I - x_2} & \text{if } 11 \leq I \leq 13 \\ Q_{22} = \frac{x_I - x_{I-2}}{x_I - x_3} & \text{if } 14 \leq I \leq 25. \end{cases} \tag{5}$$

is used to investigate whether the largest value is an outlier.

Under H_0*: all values are from the same normal distribution*, H_0 is rejected at confidence level $1 - \alpha$ if $Q > q_{\alpha,I}$, where values $q_{\alpha,I}$ are tabulated (see [7]).

Similarly, if we rather want to investigate whether the smallest value x_1 is an outlier, we will consider the statistics

$$Q = \begin{cases} Q_{10} = \frac{x_2 - x_1}{x_I - x_1} & \text{if } 3 \leq I \leq 7 \\ Q_{11} = \frac{x_2 - x_1}{x_{I-1} - x_1} & \text{if } 8 \leq I \leq 10 \\ Q_{21} = \frac{x_3 - x_1}{x_{I-1} - x_1} & \text{if } 11 \leq I \leq 13 \\ Q_{22} = \frac{x_3 - x_1}{x_{I-2} - x_1} & \text{if } 14 \leq I \leq 25. \end{cases} \tag{6}$$

2.2.3 Test applications

Paper [14] considers $I = 15$ SEs and applies Dixon's Q test for several purposes, each time for a given search term $k \in \mathcal{K}$:

1. **Abnormal SE score**. As evoked previously, the values used for the test are therefore $x_i = S_{i,k}$, the SE scores, sorted in increasing order. We consider the statistics Q defined in (6), and will say that the SE with the smallest score is an outlier and potentially biased at risk α if $Q > q_{\alpha,I}$.
2. **Investigating SEs disregarding the most visible link**. If ℓ^* is the page in $\mathcal{L}$ with the highest score for the considered search term k, the values x_i for $1 \leq i \leq I$ are the ordered scores/visibilities among the I SEs, that is, $x_i = q_{\pi_i(\ell^*,k)}, 1 \leq i \leq I$. We again consider the statistics Q defined in (6), and will say that the SE with the smallest score is an outlier and abnormally "hides" the most relevant page (as per the consensus ranking) at risk α if $Q > q_{\alpha,I}$.
3. **Investigating if the top-ranked page of each SE is also visible at other SEs**. Consider for each SE i the top-ranked page $\ell'(i)$, and then take $(x_i)_{1 \leq i \leq I}$ as the ordered visibilities $(q_{\pi_i(\ell'(j),k)})_{1 \leq j \leq I}$ of that page among SEs. It gives I tests for $1 \leq i \leq I$ suggesting that at risk α the visibility of the top-ranked page at SE j (necessarily the largest visibility q_1) is abusively high with respect to the visibility of that page at other SEs if the statistics Q defined in (5) verifies $Q > q_{\alpha,I}$.
4. **Investigating if an SE ranks first a page not considered relevant by others**. The values used for the test are the ordered I scores of each SE's top-ranked page, i.e., $x_i = R_{\tilde{\pi}_i(1,k),k}, 1 \leq i \leq I$ with $\tilde{\pi}_i(\cdot, k)$ the inverse permutation of $\pi_i(\cdot, k)$: $\tilde{\pi}_i(s, k)$ is the index of the page displayed in position s by SE i for the request k. We consider the statistics Q defined in (6), and will say that the SE whose top-ranked page has the smallest score is an outlier and potentially biased at risk α if $Q > q_{\alpha,I}$.

In [14], a campaign was analyzed over 767 most searched keywords at the time. Average deviations by SEs from the consensus are computed and compared, as well as similarities of results between SEs and percentages of failed tests.

3 Bias taking into account the Click Through Rate

In the definition of bias in (2), the rank of a page is indifferent, we just care about whether the page is "shown" or not. Though, being shown first, second or at any other position does not have the same impact. We therefore propose to take into account a value associated with the position in the rankings and representing the visibility of the position: the CTR of each position, similarly to what is done in Section 2.2.1 when defining the score of a page and then of an SE. We believe that it will then be much more representative of the bias *induced* for the user.

To formally integrate the CTR, instead of writing $v_{i,\ell,k} = 1$ if Page $\ell \in \mathcal{L}$ is displayed by SE $i \in \mathcal{I}$ for query $k \in \mathcal{K}$ in (2), we count $v_{i,\ell,k} = q_{\pi_i(\ell,k)}$, that is the CTR of the position $\pi_i(\ell, k)$ of Page ℓ on SE i for search term k.

From this modified notion of bias taking into account the impact of the position, the ANOVA tests as proposed in [16] can be performed exactly as described in Section 2.1.2. This new version will be used in our numerical experiments in the next sections.

4 Formal comparison

Having described the two different notions of bias/similarity and score, plus the used statistical tests, we can make the following remarks:

- To our knowledge the only two tools for SE bias detection based on statistical grounds are those described in Sections 2.1 and 2.2, even if based on the (strong) normality assumption of the data.
- The ANOVA tests are based on several (t) queries, even if in practice for the one-way tests we have $t = 1$. On the other hand the page score (3) is defined for a single search, but we can redefine it for the set $\mathcal{T} \subset \mathcal{K}$ of t queries

$$R_{\ell,\mathcal{T}} = \sum_{k\in\mathcal{T}} R_{\ell,k} = \sum_{k\in\mathcal{T}} \frac{1}{I} \sum_{i=1}^{I} q_{\pi_i(\ell,k)}, \tag{7}$$

  and make similar tests to those done in [14] (a top-ranked page is then defined for example as the one with the largest aggregated score in (7)). When considering different domains with a set of queries, we will compute the score over a domain $\mathcal{K}$, such that $t = n$. Applying such an "averaging" makes also the normality assumption more reasonable.
- The proposition in Section 3 to take into account the bias will make the work in [16] more "impact-friendly" and make the comparison with the use of the score described in Section 2.2.1 more relevant. With that extended definition of page scores, SE scores can be accordingly computed as

$$S_{i,\mathcal{T}} = \sum_{\ell\in\mathcal{L}} \left(\sum_{k\in\mathcal{T}} q_{\pi_i(\ell,k)} \right) R_{\ell,\mathcal{T}}, \tag{8}$$

  quantifying each SE's propensity to show high-score pages.
- Different things are actually tested in [16] and [14]. It is interesting to wonder whether something tested by an ANOVA test can also be by Dixon's Q test and reciprocally. We will then see how complementary and/or competitive the tests are. It is what we propose to do in the next section.
  - Testing if there is a suspected bias among SEs is typically addressed in both papers, by the one-way ANOVA in [16] and testing a suspicious SE score in [14]. Their respective accuracy can therefore be compared.
  - Dixon's Q test *identifies* one doubtful value with respect to the others while one-way ANOVA "just" say that there are significant differences.
  - The other tests in [14] are typically not managed by the notion of bias and ANOVA tests.
  - The two-way ANOVA test requires two characteristics and a sample in each characteristic, something not addressed by Dixon's Q text, focusing on the distribution of a (one-dimensional) sample.
- Since many tested hypotheses are different, the two papers can be seen as complementary.
- As both papers can test whether an SE behaves differently from the others, which is a relevant issue when seeking a suspect behavior from an SE, we can make a comparison of the outputs. In both cases H_0 is "we cannot differentiate the output from an SE with respect to the others". The test is done by comparing the SE biases through ANOVA, or looking at the different scores through Dixon's Q test (both can be done for any value of t).

It therefore makes sense to compare the power of both tests to investigate if one or the other detects more often voluntary biases from SEs. Robustness to false positives is also of interest, in particular from the point of view of a regulator for whom false positives could lead to legal issues.

Exchanging the metrics and the tests would also be of interest and be part of the comparison, potentially leading to an experimentally more robust proposition:

- We can first apply Dixon's Q test to bias values of SEs and investigate he assumption H_0: "All bias values are statistically not differentiable" against H_1: "Biases show statistical differences".
- We could also apply ANOVA tests to SE scores (one or two ways) if decomposing scores on (individual) queries in subdomains similarly to [16]. An SE bias is then just replaced by an SE score.

5 Tests with synthetic data

We aim in this section to test the efficiency of SE bias detection methods, i.e., of all *(metric, test)* combinations, where the metric is the SE score or the SE bias (with CTR or not), and the test is Dixon's Q or ANOVA. We also aim at studying false positives.

To control the setting, we use synthetic data here: we generate (randomly) "real relevance" values for pages related to requests, that can be imperfectly estimated by SEs. When no SE is biased, each one ranks pages according to its relevance estimations, resulting in possibly different rankings. To model the presence of bias, we assume one SE will always show a specific page (the one it has economic interests in) at the top position.

In Subsection 5.1 we present the specific mathematical model considered and some performance results from simulations when the bias (when present) affects all requests. By contrast, in Subsection 5.2 we enrich the model by adding the notion of "domain" related to a request as is done in [16], and refine the behavior of the biased SE, by assuming it pushes its page of interest to the top of the ranking only when the request belongs to a specific domain. Such a domain-specific bias might be harder to detect among a large number of requests spanning several domains.

5.1 Dixon Q test and one-way ANOVA

In this subsection, we do not use the notion of domain (remark again that it was not considered in [14] either). This can be interpreted in two different ways: either the bias (when present) is the same over all domains, or all the requests that are considered–that are chosen by the entity investigating bias, like a regulator–belong to the same domain, i.e., the one for which bias is suspected.

5.1.1 Generation of page relevance (and SE estimation)

To simulate the behavior of unbiased and biased SEs, we assume that the detection test is performed over a number n of request terms. We consider a given set of w pages, so that $w = |\mathcal{L}|$ according to our notations at the beginning of Section 2. For each query term k, each page $\ell \in \mathcal{L}$ has a "real relevance" value $V_{\ell,k}$, that we generate according to a uniform distribution over the interval $[0, 1]$. But the estimation of that (objective) relevance value can be difficult for SE, which we model through an additive estimation noise following a centered normal distribution with standard deviation σ_1, independent of everything else. Summarizing, for a given request k, the estimated relevance of page ℓ for SE i is generated as

$$\hat{V}_{\ell,k}(i) := \underbrace{V_{\ell,k}}_{\sim\mathcal{U}(0,1)} + \underbrace{\nu_{\ell,i,k}}_{\sim\mathcal{N}(0,\sigma_1^2)} . \tag{9}$$

5.1.2 Simulating the behavior of a biased search engine

We then consider two bias situations, on which to test our bias detection methods. For each query k:

- in the **unbiased setting**, each SE i simply sorts pages by decreasing order of their estimated relevances $\hat{V}_{\ell,k}(i)$,
- the only change in the **biased setting** is that one SE (say, SE $i = 1$) is biased, and will always push to the first position the same page (say, page 1) whatever its relevance. All other rankings are based on estimated relevances.

Note this setting is the one considered in [14], where only Dixon's Q test was performed to detect outliers among SE scores. Here however, when we vary the number n of requests, we apply the extended definition of page scores introduced in (7).

5.1.3 Statistical tests and results

For one-way ANOVA tests, the bias (with or without CTR) as described in (2) is computed with $t = 1$ for each of the n search terms, as in [16] and as described in Section 2.1.2. We therefore have a sample of size n for each SE, allowing to perform variance-based tests.

Figures 1 and 2 show the detection rates for the six bias detection tests (stemming from each combination of a bias metric–SE score, bias without CTR, bias with CTR–and a statistical test–Dixon's Q outlier detection, one-way ANOVA) when respectively varying the number n of requests and the standard deviation of the relevance estimation "error" $\nu_{\ell,i,k}$. Each point is the empirical average over 10^6 experiments, and 95% confidence intervals are so thin that they visually reduce to a single point.

Finally, each figure contains two graphs, one for the unbiased setting–hence, showing the false positive rates–and the other one for the biased setting, suggesting a "true positive": note however that the tests do not identify which SE is potentially biased (this could be done for Dixon's test by checking whether the outlier SE is indeed the biased one, but we kept the test as is for comparison purposes). Figure 1 *(left)* however suggests most one-way ANOVA positives are actually true positives, as the rate of false positives with ANOVA is very low (below 1% for all tested situations). In contrast, Dixon's Q test is less conservative, and the rate of false positives is significantly larger: when performed over a single request it is around 35%, and decreases when the number of requests increases. Note that all one-way ANOVA tests need at least two requests to work: at least two experiments are required to estimate variances, see (11) in the appendix when expressing the ANOVA statistics.

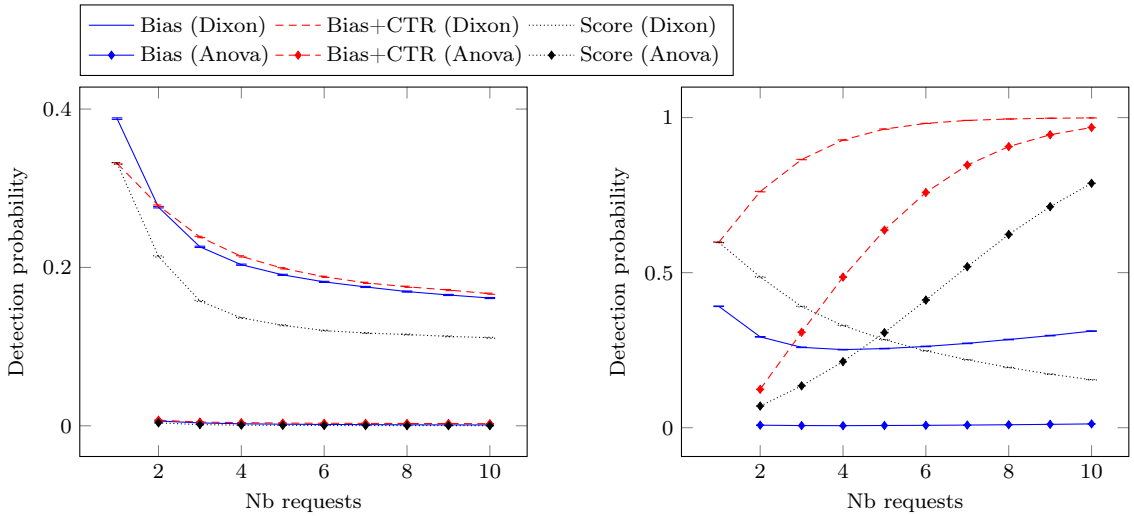

Figure 1: False positive *(left)* and true positive *(right)* rates when varying the number of requests, for $\sigma_1 = 0.1$.

Most tests tend to improve (less false positives, more true positives) when the number of requests increases, which could be expected as the test is based on more data, however:

- For the "bias without CTR" metric and Dixon's test, the rate of true positives first decreases with n, before increasing. We interpret that as follows: the rate of false positives being quite high for small values of n, many instances that led to a positive test would also have been positive even without SE 1 favoring any page. Then as more data is used, the accuracy of the test improves and most positive tests are really "true positives", i.e., are a consequence of SE 1 biased behavior.

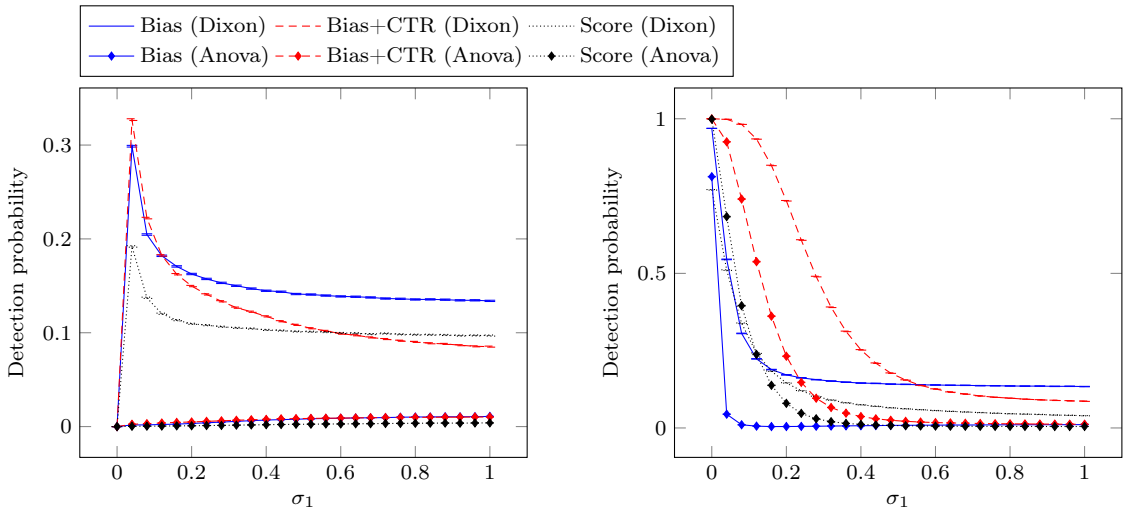

Figure 2: False positive *(left)* and true positive *(right)* rates (over 5 requests) when varying the standard deviation σ_1 of the relevance estimation error.

- The rate of true positives consistently decreases with n for the score metric and Dixon's test, which could be surprising but can be understood from the definition of page scores, given in (7): as SE 1 is systematically giving the maximum visibility to Page 1, that page accumulates a higher score over the n requests from that SE alone, with respect to the other pages whose scores average out over the multiple requests. When n increases, even if being the only biased SE, SE 1 manages to artificially make Page 1 the highest-score page, and finally, as SE 1 systematically gives that page the top position, it obtains a high score (see (8)) and is not detected by the test.

Figure 1 suggests that for each test method, the three metrics give comparable false-positive rates, but yield different true-positive rates, the higher ones being for the "bias with CTR" metric. Depending on the number of requests the best-performing test to apply with that metric can change: the choice between ANOVA and Dixon's Q tests depend on the willingness to avoid false positives. If avoiding them is preferred, ANOVA test has to be chosen, even more as n gets larger since the gap between the two (red) curves of true probability detection then shrinks.

Figure 2 shows the impact of the estimation noise amplitude, namely the standard deviation σ_1 of the estimation error in (9). Again we notice that Dixon-based tests yield a much higher false positive rate than ANOVA-based tests:

- For $\sigma_1 = 0$ there is no randomness so that in the absence of bias all SEs behave exactly the same and no outlier is found;
- But even with very small positive values of σ_1, the estimation noise is sufficient to have at least one SE show a slightly different ranking, that is singled out by Dixon's tests as an outlier;
- Interestingly, as σ_1 increases the rate of false positives decreases, since this increases the variety of rankings among (still unbiased) SEs, so that an outlier is less often identified.

Regarding true positives, as could be expected the rate decreases with σ_1 since it becomes harder for the tests to detect abnormalities when the variety of rankings gets larger. When σ_1 is small, Bias+CTR is (again) the metric to be preferred, with ANOVA if a limited false positive (type-II error) is important, and Dixon's Q test if it is not the case. When σ_1 is large, the combination Bias+Dixon has the largest detection probability, but at the expense of a large type-II error.

5.2 Performance of bias detection methods for domain-specific bias

We now investigate a subtler bias behavior that an SE may implement, which consists in favoring a webpage only when the request is associated to a given domain. For all other types of requests, the SE ranks pages neutrally, i.e., according to their estimated relevance.

5.2.1 Domain-specific bias model

To encompass a domain-dependent behavior, we assume that each request can be associated to a domain d in a set $\mathcal{D}$ of $D = |\mathcal{D}|$ domains, and we modify our model as follows.

- For each *domain* $d \in \mathcal{D}$, each page ℓ has an average relevance, denoted by $\bar{V}_{\ell,d}$, among the possible requests related to that domain. In our experiments we will take independent uniformly distributed random variables on $[0, 1]$ for each page ℓ and domain d.
- Then the impact of a specific request k identified as related to a domain $d(k)$, is modeled by a slight change from $\bar{V}_{\ell,d}$, to yield $V_{\ell,k}$, the (true) relevance of page ℓ for request k. In this paper we use an additional centered Gaussian random variable with standard deviation σ_2.
- Each SE i then imperfectly estimates that relevance $V_{\ell,k}$ as was already assumed in Subsection 5.1 and expressed mathematically in (9).

In the end, for a request k, the estimated relevance of page ℓ for SE i is taken as

$$\hat{V}_{\ell,k}(i) = \overbrace{\underbrace{\bar{V}_{\ell,d(k)}}_{\sim\mathcal{U}(0,1)} + \underbrace{\nu_{\ell,k}}_{\sim\mathcal{N}(0,\sigma_2^2)}}^{\text{True relevance } V_{\ell,k}} + \underbrace{\nu_{\ell,j,k}}_{\sim\mathcal{N}(0,\sigma_1^2)} , \tag{10}$$

where $d(k)$ is the domain to which the request k is associated.

5.2.2 Statistical tests

We keep using the same three metrics as in the previous section (SE score, bias with and without CTRs), but now the collected results depend on the domain: for a given set $\mathcal{D}$ of domain, we simulate a number n of requests performed for each domain $d \in \mathcal{D}$, so that the tests we use are based on nD requests, all performed for each SE.

The statistical tests we apply focus on those two dependent variables for each request (the SE and the domain), in order to detect outliers. Namely, we use the ANOVA two-way tests with interaction, aiming at signaling

- a general difference among SEs (row effect, if we imagine data organized with SEs as rows and domains as columns–with the standard notations of ANOVA tests given in the appendix–with n data points in each cell);
- a non-additive row-column interaction, i.e., a difference in how the domain affects SEs (this is done with the two-way interaction test).

Note that a column effect can also be tested (differences among domains) but is not of real interest for our purposes.

5.2.3 Experimental results

Our experimental results (false-positive and true-positive rates) are shown in Figures 3 to 6, when varying one parameter in the model. The default parameter values we used are given in Table 2.

All figures show low false-negative rates (below 1%), and true-positive rates evolve as could be expected with some model parameters:

- bias on only one domain is harder to detect when there are more domains (as shows the decrease of true positives in terms of the number of domains on the right panel of Figure 3, while there is no major trend for the false positives on the left panel),

Variable	Notation	Value
Number of domains	D	5
Number of requests per domain	n	5
Standard deviation of page relevance within each domain	σ_2	0.05
Standard deviation of relevance estimation error	σ_1	0.1

Table 2: Default values for the simulations of domain-specific bias detection

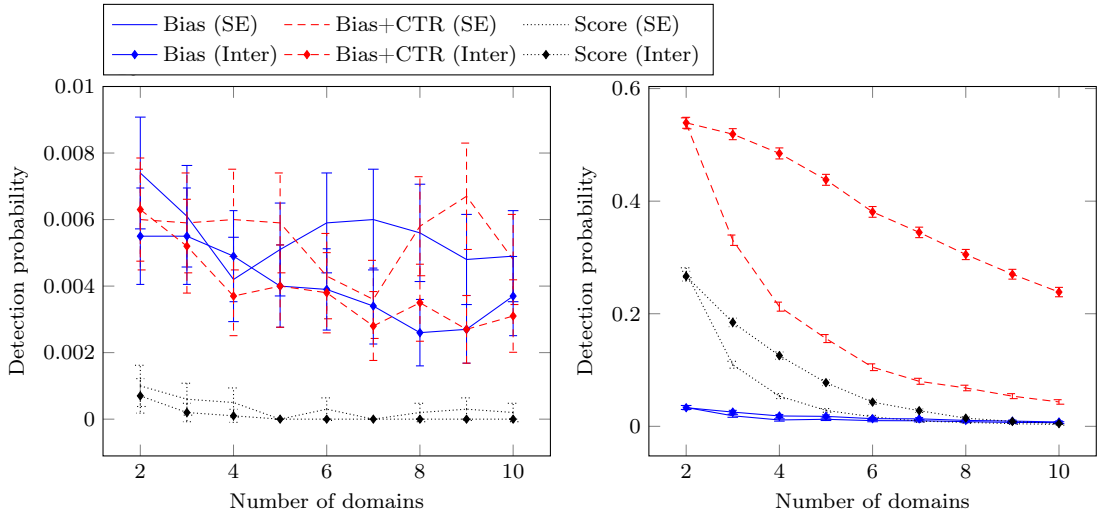

Figure 3: False positive *(left)* and true positive *(right)* rate when the number of domains varies, for two-way ANOVA tests aimed at identifying an SE effect and a SE-domain interaction.

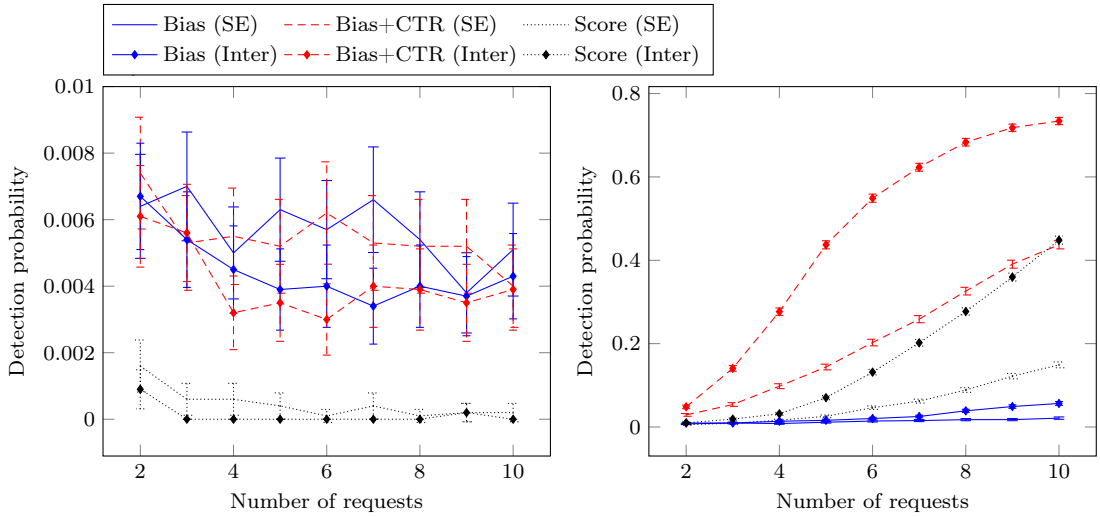

Figure 4: False positive *(left)* and true positive *(right)* rates when the number of requests varies, for two-way ANOVA tests aimed at identifying an SE effect and a SE-domain interaction.

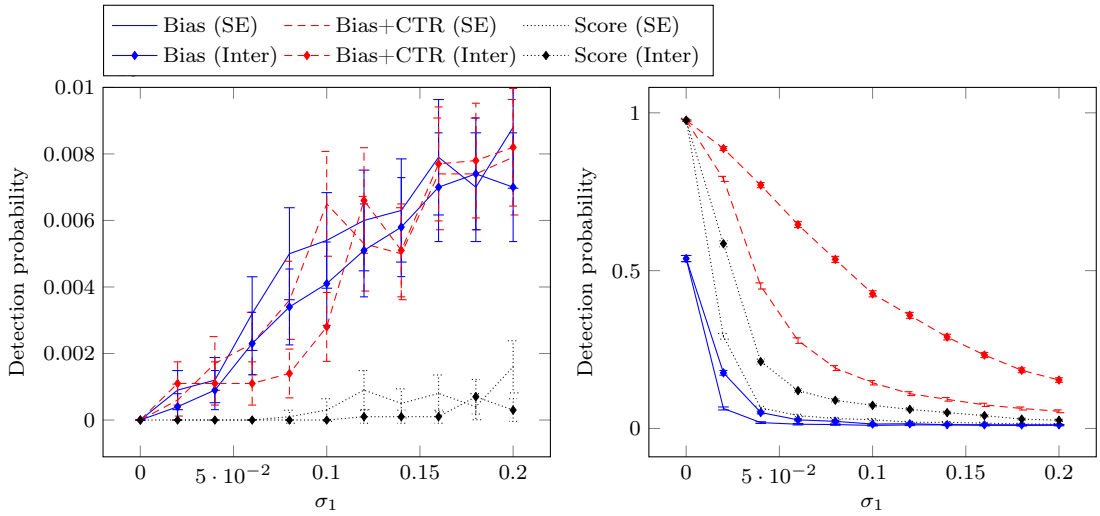

Figure 5: False positive *(left)* and true positive *(right)* rates when σ_1 varies, for two-way ANOVA tests aimed at identifying an SE effect and a SE-domain interaction.

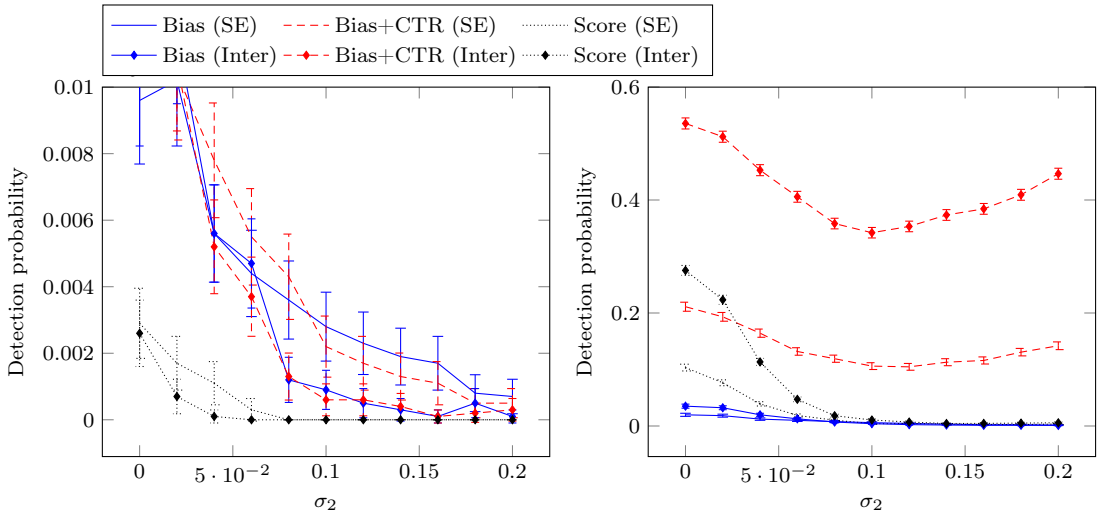

Figure 6: False positive *(left)* and true positive *(right)* when σ_2 varies, for two-way ANOVA tests aimed at identifying an SE effect and a SE-domain interaction.

- bias is easier to detect if the test is based on more data, e.g., when more requests are performed for each domain (right panel of Figure 4),
- when SEs have very noisy estimations of real relevances (σ_1 large), their rankings differ so much that bias can go undetected (as illustrated by the decrease in terms of σ_1 on the right panel of Figure 5). A larger σ_1 also induces as expected more false positives, but the values remain below 1%.

Finally, the impact of the standard deviation σ_2 of page relevance within each domain, shown in Figure 6, is not completely clear. One might expect a higher σ_2 to make page relevances less domain-dependent, so that detecting bias on a specific domain gets harder, and it's actually what we observe for two of our metrics, namely SE score and bias without CTR. For the bias-with-CTR metric, that gives the largest true-positive rate, the detection rate first decreases with σ_2 until it equals σ_1, and then increases when σ_2 exceeds σ_1. A possible interpretation can be that the higher variance over true relevances renders the test less sensitive to noise estimation errors, making it easier to detect bias.

The tests based on the score provide less false positives than the two others. For those based on bias without including the CTR, the probability of correct detection is the lowest while having a high false detection probability: the two other combinations, which are among the propositions of the present paper, are therefore rather advised.

6 Tests on a campaign

We also ran experiments over real searches in 2023. We considered 4 different domains/topics, some of them potentially subject to bias, and $n = 5$ keywords in each domain. The full list is given in Table 3.

Domains	Sport	Travel	Cooking	Grammar
keywords	Soccer world cup	Visit Paris	Home-made sushi	Fish plural
	Olympic games	Event Las Vegas	Pizza recipe	Irregular verbs
	NBA	Plane ticket	Mojito	Subjonctive
	NCAA final four	Stay in Rome	Cooking advice	Grammar corrector
	Golf US Open	Best travel agency	Scrambled eggs	When to use the

Table 3: List of considered domains and keywords

The results were collected thanks to the SNIDE tool `https://snide.inria.fr/`[4], gathering the results from the following 14 search engines at the time:

- AllTheInternet
- AOL
- Ask
- Bing
- DirectHit
- Duckduckgo
- Ecosia
- Google
- Lilo
- Lycos
- Qwant
- Startpage
- Yahoo
- Yandex.

All the search results are available at

`https://www.dropbox.com/scl/fo/yjvt7c0c1bcufuuxnu1o0/h?rlkey=ndbkd1s3jtbfwvk8dtfezqbmy&dl=0`

as JSON files.

We ran the statistical tests described in the previous sections at confidence level 99%. More precisely, we performed the following tests.

1. For each individual query, a Dixon Q test on the bias of each SE evaluated as in (2) on the number of times the pages are displayed as in [16] or using the CTR as proposed in Section 3.
2. Still for each individual query, we applied the Dixon Q tests described in Section 2.2.3: testing if there is an abnormally low SE score (symptomatic of an SE bias); investigating SEs disregarding the most visible link; investigating if the top-ranked page of each SE is also visible at other SEs; and investigating if an SE ranks first a page not considered relevant by others.

[4]Note that the tool has to be permanently updated; there is no guarantee it will be fully maintained. To keep the paper self-contained, we provide links to the obtained output.

Metric	AllTheInternet	AOL	Ask	Bing	DirectHit	Duckduckgo	Ecosia	Google	Lilo	Lycos	Qwant	Startpage	Yahoo	Yandex
Bias	77	**0.4**	51	5	**0.2**	16	20	26	49	**0.6**	37	70	**0.4**	**0.0**
Bias+CTR	6	1	34	3	12	70	43	32	35	1	31	54	**0.8**	**0.5**
Score	11	**0.4**	5	3	39	39	11	7	36	**0.9**	6	12	**0.3**	**0.9**

Table 4: Results from one-way ANOVA tests targeting a domain effect for each SE: the values are p-values multiplied by 100, and positive tests (p-values below 1%) are highlighted in bold.

Factor	Bias	Bias+CTR	Score
Domain effect	1	0	1
SE effect	0	0	0
Interaction effect	0	0	0

Table 5: Two-way ANOVA test results

3. We also applied a Dixon Q test to biases of each SE per domain, when the biases are computed as per (2) from the $n = 5$ requests in the domain.
4. For each domain, we performed the one-way ANOVA test to the hypothesis that there is no difference of values between SEs or between requests when the considered metric is i) the bias in [16], ii) the bias with the CTR, and iii) the score in Section 3.
5. The two-way ANOVA tests with interaction were also used, to test if there are differences between SEs or between domains, again with the three possible metrics.

For Dixon Q tests of items 1 to 3 above, we observe no rejection at all on the "Sports", "Cooking", "Grammar" or "Travel" domains, so we do not display the corresponding tables. For the sensitive domain "Travel" for which some SEs may propose services and often considered subject to bias, not being able to observe a clear deviation is notable.

We also performed one-way ANOVA tests to possibly detect, within each domain, significant differences among SEs (SE effect); again at a confidence level 99% no statistically significant discrepancy among SEs was detected with any of our three metrics.

Some one-way ANOVA tests came out positive, namely, those aiming at identifying a domain effect for each SE. The results (p-values) are given in Table 4 and suggest that the metric values differ a lot among domains for some SEs. However, we should recall that the study only spans over 4 different domains (and 5 requests per domain) and that the data does not satisfy the assumptions upon which the test is build, namely some domains may experience different variances over requests due to more obvious responses or a larger set of relevant values. More extensive campaigns are probably needed to confirm or not that effect.

The fact that Dixon's Q test never rejects the null hypothesis but that it happens for ANOVA tests may be seen in contradiction with the results from the synthetic tests, but note that those synthetic results were obtained in average, meaning that for a single realization a different result can be obtained, and for a very specific case of voluntary bias.

Two-way ANOVA test results are presented in Table 5. For two-ways ANOVA, interestingly there is no detected effect of SEs on bias, but there is a detected difference due to domains. It may come from the different diversity, that is, variance, of the relevant pages to be referenced depending on the treated topic. However the tests do not highlight an interaction (SE×Domain) effect.

7 Conclusions

This paper builds on and extends research aimed at detecting possible bias that search engines may implement, based on the comparison of the results provided by several search engines. More precisely, we consider three types of metrics that can be used to measure bias (two from the literature, and a new one that we introduce) and two families of statistical tests to detect significant discrepancies, namely Dixon Q tests for outlier detection and variance analysis tests. This offers a wide range of *metric+test* combination possibilities that can be used in practice. We compare their performance through extensive simulations with synthetically-generated data to analyze their sensitivity in a controlled setting, when (known) bias is present or not. In addition, we apply those combinations to real data collected through a publicly-available tool, in order to show what types of results they can provide on real data.

While the test results should always be taken with care, especially with small-scale campaigns, the methodology developed in this paper offers a series of options that can pinpoint possible unfair behavior by a search engine, and trigger further investigation on that engine (even possibly, on a specific domain). The various tests of the literature we have described are often complementary, but when substitutable we have illustrated that they do not have the same sensitivity to so-called type I (false negative) or type II (false positive) errors, giving room to the user to select what is more important to them. From a legal perspective, if the goal is to prosecute for a voluntary bias, minimizing false positives seems the right option. Moreover, our new tests including the CTR to the bias computation and the score seem worthwhile additions with respect to just computing the bias from the number of appearances in rankings.

Future work can include larger-scale bias measurement campaigns to strengthen or maybe contradict the first observations we found here (mostly showing no suspicious behavior). But it could also focus on simulation settings, by building and studying more complex scenarios than the one we consider in this paper, for example by implementing a subtler biased behavior, where a search engine wanting to favor a webpage would introduce randomness in its ranking to render the detection more difficult. Such settings may show different best-performing combinations for the metric and test choices.

References

[1] C. C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2nd edition, 2016.

[2] J. Brill. Statement of the Commission regarding Google's search practices, 2013. `http://www.ftc.gov/public-statements/2013/01/statement-commission-regarding-googles-search-practices` . Last accessed Nov 2021.

[3] G. Casella. *Statistical Design*. Springer Texts in Statistics. Springer New York, 2008.

[4] European Commission, Content Directorate-General for Communications Networks, Technology, J Sunderland, F Herrera, S Esteves, I Godlovitch, L Wiewiorra, S Taş, P Kroon, M Stronzik, D Baischew, L Nett, S Tenbrock, S Strube Martins, A Streel, J Kalliala, J Huerta Bravo, W Maxwell, and A Renda. *Digital Markets Act – Impact assessment support study – Annexes*. Publications Office, 2020. `https://op.europa.eu/en/publication-detail/-/publication/2a69fd2a-3e8a-11eb-b27b-01aa75ed71a1/language-en`.

[5] R. B. Dean and W. J. Dixon. Simplified statistics for small numbers of observations. *Analytical Chemistry*, 23(4):636–638, 1951.

[6] R. Dejarnette. Click-through rate of top 10 search results in Google, 2012. `http://www.internetmarketingninjas.com/blog/search-engine-optimization/click-through-rate`, last accessed June 28, 2017.

[7] W. J. Dixon. Processing data for outliers. *Biometrics*, 9(1):74–89, 1953.

[8] R. Epstein and R.E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Psychological and Cognitive Sciences*, 112(33):4512–4521, 2015.

[9] European Commission. Antitrust: Commission fines google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service. Press release, available at `https://ec.europa.eu/commission/presscorner/detail/en/IP_17_1784`, June 2017.

[10] G. Gezici. Biased or not?: The story of two search engines. In *ACM FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[11] J. Grimmelmann. Some skepticism about search neutrality. *The Next Digital Decade: Essays on the Future of the Internet*, page 435, January 2011.

[12] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, October 2004.

[13] B. Li and L. Han. Distance weighted cosine similarity measure for text classification. In Hujun Yin, Ke Tang, Yang Gao, Frank Klawonn, Minho Lee, Thomas Weise, Bin Li, and Xin Yao, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2013*, pages 611–618, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[14] P. Maillé, G. Maudet, M. Simon, and B. Tuffin. Are Search Engines Biased? Detecting and Reducing Bias using Meta Search Engines. *Electronic Commerce Research and Applications*, 2022.

[15] P. Maillé and B. Tuffin. *From Net Neutrality to ICT neutrality.* Springer, 2022.

[16] A. Mowshowitz and A. Kawaguchi. Measuring search engine bias. *Information Processing & Management*, 41(5):1193 – 1205, 2005.

[17] A. Odlyzko. Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets. *Review of Network Economics*, 8(1):40–60, 2009.

[18] D. Rushe. Eric Schmidt Google senate hearing – as it happened, 2012. `http://www.guardian.co.uk/technology/blog/2011/sep/21/eric-schmidt-google-senate-hearing` . Last accessed Nov 2021.

[19] A. Singhal. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43, 2001.

[20] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38, 2010.

A ANOVA tests

We describe the tests using the notations of the paper so that their application can be easily retrieved.

One-way ANOVA The first (rejection) test implemented in [16] is a one-way ANOVA test. Recall first how this test works to clarify [16]. It assumes normally-distributed with the same (although unknown) variance σ^2. It also assumes that we have I independent samples, each of size m. The members of ith sample ($1 \leq i \leq I$) are $X_{i1}, \ldots X_{in}$ normal with mean μ_i and variance σ^2. The goal of the one-way ANOVA is to investigate hypothesis H_0, here

$$H_0 : \mu_1 = \cdots = \mu_I (= \mu) \text{ versus } H_1 : \text{not all the means are equal.}$$

The μ_i are unknown but estimated by

$$\mu_i \approx X_{i\cdot} = \frac{1}{n} \sum_{k=1}^{n} X_{ik}$$

while μ, also not known under H_0, is approximated by

$$\mu \approx X_{\cdot\cdot} = \frac{1}{I} \sum_{i=1}^{I} X_{i\cdot} = \frac{1}{In} \sum_{i=1}^{m} \sum_{k=1}^{n} X_{ik}.$$

Let

$$SS_W = \sum_{i=1}^{I} \sum_{k=1}^{n} (X_{ik} - X_{i\cdot})^2$$

be the *within samples sum of squares* and and

$$SS_b = n\sum_{i=1}^{I}(X_{i\cdot} - X_{\cdot\cdot})^2$$

be the *between samples sum of squares.* Define the statistic

$$TS = \frac{SS_b/(I-1)}{SS_W/(nI-I)}. \tag{11}$$

Under H_0, TS has an F-distribution with $I-1$ numerator and $nI-I$ denominator degrees of freedom. Define $f_{I-1,nI-I,\alpha}$ is the $100(1-\alpha)$ percentile of this distribution ($\mathbb{P}(F_{I-1,nI-I} > f_{I-1,nI-I,\alpha}) = \alpha$) where $F_{i,j}$ represent an F-random variable with i numerator and j denominator degrees of freedom. The one-way ANOVA test at significance level α for H_0

$$\text{rejects } H_0 \text{ if } TS = \tfrac{SS_b/(I-1)}{SS_W/(nI-I)} > F_{I-1,nI-I,\alpha}$$

$$\text{accept } H_0 \quad \text{otherwise.}$$

Values $f_{i,j,\alpha}$ are tabulated.

Two-way ANOVA The two-way ANOVA studies the effects of several factors instead of one. Let X_{ik} be the value obtained when the first factor is i and the second is k. Again the X_{ik} are assumed independent normal random variables with a common variance σ^2, but the mean is affected by the two factors (instead of just one). One-way ANOVA was assuming $\mathbb{E}[X_{ik}] = \mu_i = \mu + \alpha_i$ and testing $\alpha_i = 0\ \forall i$. Now, $\mathbb{E}[X_{ik}] = \mu_{ik} = a_i + c_k$ as an *additive* model. With a "$\cdot$" notation, let

$$\mu_{i\cdot} = \sum_{k=1}^{n}\mu_{ik}/n \quad \mu_{\cdot k} = \sum_{i=1}^{I}\mu_{ik}/I \quad \mu_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{k=1}^{n}\mu_{ik}/(nI)$$

and

$$a_{\cdot} = \sum_{i=1}^{I} a_i/I \quad b_{\cdot} = \sum_{k=1}^{n} c_k/n.$$

Define

$$\begin{aligned}
\mu &= \mu_{\cdot\cdot} = a_{\cdot} + c_{\cdot} \\
\alpha_i &= \mu_{i\cdot} - \mu = a_i - a_{\cdot} \\
\beta_k &= \mu_{\cdot k} - \mu = c_k - c_{\cdot}
\end{aligned}$$

μ is called the grand mean, α_i the deviation from the grand mean due to row i, β_k the deviation from the grand mean due to column k. Define again

$$\begin{aligned}
X_{i\cdot} &= \sum_{k=1}^{n} X_{ik}/n \text{ average of the values in row } i \\
X_{\cdot k} &= \sum_{i=1}^{I} X_{ik}/I \text{ average of the values in column } k \\
X_{\cdot\cdot} &= \sum_{i=1}^{I}\sum_{k=1}^{n} X_{ik}/(nI) \text{ average of all data values}
\end{aligned}$$

such that

$$\begin{aligned}
\mathbb{E}[X_{i\cdot}] &= \mu + \alpha_i \\
\mathbb{E}[X_{\cdot k}] &= \mu + \beta_k \\
\mathbb{E}[X_{\cdot\cdot}] &= \mu.
\end{aligned}$$

The estimators of $\mu = \mathbb{E}[X_{\cdot\cdot}]$, $\alpha_i = \mathbb{E}[X_{i\cdot} - X_{\cdot\cdot}]$ and $\beta_k = \mathbb{E}[X_{\cdot k} - X_{\cdot\cdot}]$ are respectively $\hat{\mu} = X_{\cdot\cdot}$, $\hat{\alpha}_i = X_{i\cdot} - X_{\cdot\cdot}$ and $\hat{\beta}_k = X_{\cdot k} - X_{\cdot\cdot}$.

We want to test

$$H_0 : \alpha_i = 0 \ \ \forall i$$

versus

$$H_1 : \text{ not all } \alpha_i = 0.$$

(We could also be interested in testing whether there is a column effect:

$$H_0 : \beta_k = 0 \ \ \forall k \qquad \text{versus} \qquad H_1 : \text{ not all } \beta_k = 0.)$$

Let

$$SS_e = \sum_{i=1}^{I} \sum_{k=1}^{n} (X_{ik} - X_{i\cdot} - X_{\cdot k} + X_{\cdot\cdot})^2,$$

called *the error sum of squares*, and

$$SS_r = n \sum_{i=1}^{I} (X_{i\cdot} - X_{\cdot\cdot})^2$$

called *the row sum of squares.* We use the statistic

$$TS = \frac{SS_r/(I-1)}{SS_e/((I-1)(n-1))}.$$

Under H_0, at significance α,

$$\text{Reject } H_0 \quad \text{if} \quad TS = \tfrac{SS_r/(I-1)}{SS_e/((I-1)(n-1))} > F_{I-1,(I-1)(n-1),\alpha}$$

$$\text{Accept } H_0 \quad \text{otherwise}$$

Two-way ANOVA with interaction There also exists a two-way analysis of variance with interaction. The weakness of previous model is that row and column effects are additive, they have no interaction. We change this by introducing the general form $\mu_{id} = \mathbb{E}[X_{id}]$. We (still for some variables) note

$$\begin{aligned}
\mu &= \mu_{\cdot\cdot} \quad \text{grand mean, average of all mean values} \\
\alpha_i &= \mu_{i\cdot} - \mu_{\cdot\cdot} \quad \text{amount by which the average of row } i \text{ exceeds } \mu \\
\beta_d &= \mu_{\cdot j} - \mu_{\cdot\cdot} \quad \text{amount by which the average of column } d \text{ exceeds } \mu \\
\gamma_{id} &= \mu_{id} - \mu_{i\cdot} - \mu_{\cdot d} + \mu_{\cdot\cdot} \quad \text{amount by which } \mu_{id} \text{ exceeds the sum of } \mu \\
&\qquad\qquad \text{and increments due to row } i \text{ and column } d
\end{aligned}$$

$$\text{so that } \mu_{id} = \mu + \alpha_i + \beta_d + \gamma_{id}.$$

To test $\gamma_{id} = 0$, we will need more than one observation for each pair of factors, namely X_{idk} ($k \in \mathcal{K}$) observations for each pair (i, d) of factors, normally distributed with mean μ_{id} and variance σ^2. We generalize the notions of estimators under the "$\cdot$" notation. Let

$$\begin{aligned}
SS_e &= \sum_{k\in\mathcal{K}} \sum_{i=1}^{I} \sum_{d=1}^{D} (X_{idk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_d - \hat{\gamma}_{id})^2 \\
&= \sum_{k\in\mathcal{K}} \sum_{i=1}^{I} \sum_{d=1}^{D} (X_{idk} - X_{id\cdot})^2
\end{aligned}$$

$$\begin{aligned}
SS_{int} &= \sum_{r=i}^{I} \sum_{d=1}^{D} n(X_{id\cdot} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_d)^2 \\
&= \sum_{i=1}^{I} \sum_{d=1}^{D} n(X_{id\cdot} - X_{i\cdot\cdot} - X_{\cdot d\cdot} + X_{\cdots})^2
\end{aligned}$$

The null hypotheses H_0 we are interested in are:

- H_0^r: $\alpha_i = 0 \ \forall i$ (no row effect)

- H_0^c: $\beta_d = 0\ \forall d$ (no column effect)
- H_0^{int}: $\gamma_{id} = 0\ \forall i, d$ (no row and column interaction).

With again the "$\cdot$" notation, We test H_0^{int}: $\gamma_{id} = 0\ \forall i, d$ at significance α:

$$\text{Reject } H_0^{\text{int}} \quad \text{if} \quad TS = \frac{SS_{int}/((D-1)(I-1))}{SS_e/(DI(n-1))} > F_{(D-1)(I-1),DI(n-1),\alpha}$$

$$\text{Accept } H_0^{\text{int}} \quad \text{otherwise}$$

We test H_0^r: $\alpha_i = 0\ \forall i$ at significance α:

$$\text{Reject } H_0^r \quad \text{if} \quad TS = \frac{SS_r/(I-1)}{SS_e/(DI(n-1))} > F_{(I-1),DI(n-1),\alpha}$$

$$\text{Accept } H_0^{\text{int}} \quad \text{otherwise}$$