Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines

Janek Bevendorff[1(✉)], Matti Wiegmann[2], Martin Potthast[2,3], and Benno Stein[2]

[1] Leipzig University, Leipzig, Germany
janek.bevendorff@uni-weimar.de
[2] Bauhaus-Universität Weimar, Weimar, Germany
[3] ScaDS.AI, Leipzig, Germany

Abstract. Many users of web search engines have been complaining in recent years about the supposedly decreasing quality of search results. This is often attributed to an increasing amount of search-engine-optimized but low-quality content. Evidence for this has always been anecdotal, yet it's not unreasonable to think that popular online marketing strategies such as affiliate marketing incentivize the mass production of such content to maximize clicks. Since neither this complaint nor affiliate marketing as such have received much attention from the IR community, we hereby lay the groundwork by conducting an in-depth exploratory study of how affiliate content affects today's search engines. We monitored Google, Bing and DuckDuckGo for a year on 7,392 product review queries. Our findings suggest that all search engines have significant problems with highly optimized (affiliate) content—more than is representative for the entire web according to a baseline retrieval system on the ClueWeb22. Focussing on the product review genre, we find that only a small portion of product reviews on the web uses affiliate marketing, but the majority of all search results do. Of all affiliate networks, Amazon Associates is by far the most popular. We further observe an inverse relationship between affiliate marketing use and content complexity, and that all search engines fall victim to large-scale affiliate link spam campaigns. However, we also notice that the line between benign content and spam in the form of content and link farms becomes increasingly blurry—a situation that will surely worsen in the wake of generative AI. We conclude that dynamic adversarial spam in the form of low-quality, mass-produced commercial content deserves more attention. (Code and data: https://github.com/webis-de/ECIR-24).

Keywords: Web Search Quality · Search Engine Optimization · Web Spam

J. Bevendorff and M. Wiegmann—Equal contribution.

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14610, pp. 56–71, 2024.
https://doi.org/10.1007/978-3-031-56063-7_4

1 Introduction

Web search engines are possibly the most important information access technologies today. It may therefore be a troubling sign that a noticeable number of social media users are sharing their observation that search engines are becoming less and less capable of finding genuine and useful content satisfying their information needs. Reportedly, a torrent of low-quality content, especially for product search, keeps drowning any kind of useful information in search results.

Previous research has shown that most pages returned by web search engines have some degree of search engine optimization (SEO) [22], with conflicting effects on users' perception of page quality [33]. The dynamics of search engine optimization and the web in general have always been a problem for search providers. It's not far-fetched to assume a connection between SEO and a perceived degradation of quality and to ask whether search providers are losing this battle. Unfortunately, search providers offer little insight into their efforts to curb SEO and SEO's dynamic nature is difficult to capture in a static and standardized test collection, which may explain why SEO has received relatively little attention from the research community in terms of retrieval effectiveness studies. Zobel [37] argues in this context that retrieval research is therefore susceptible to Goodhart's Law, owing to the difficulty of quantifying the qualitative goal of user satisfaction. Hence, measuring relevance may yield only initially convincing but ultimately impractical results.

In this paper, we systematically investigate for the first time whether and to which degree "Google is getting worse." We focus on comparative product reviews that offer tests and purchase recommendations as a key indicator of search quality. Such reviews often contain affiliate product links, which refer customers to a seller. The referring entity (the "affiliate") then receives a commission for clicks or purchases resulting from the referral. Affiliate marketing is essentially built on the trust of customers in the affiliate [15]. However, since users often trust their search engines already [20,31], the affiliate inherits this trust as a byproduct of a high ranking. This creates a conflict of interest between affiliates, search providers, and users. With "relevance" being an imperfect metric, affiliates then turn to optimizing rankings instead of investing in high-quality reviews.

Our first contribution is an investigation of the SEO properties of comparative review pages found on the result pages of Google (by proxy of Startpage), Bing, and DuckDuckGo for 7,392 product review queries (Sect. 3). We compare these findings with the results of the BM25 baseline search engine ChatNoir [5] and the raw ClueWeb22 [30] (Sect. 4). We find that the majority of high-ranking product reviews in the result pages of commercial search engines (SERPs) use affiliate marketing, and significant amounts are outright SEO product review spam. The baseline system retrieves both at much lower rates, more consistent with the overall low base rate of affiliate marketing in the ClueWeb22 as a whole. We also find strong correlations between search engine rankings and affiliate marketing, as well as a trend toward simplified, repetitive, and potentially AI-generated content. Our second contribution is a longitudinal analysis of the ongoing competition between SEO and the major search engines over the period of one year (Sect. 5). We find that search engines do intervene and that ranking

updates, especially from Google, have a temporary positive effect, though search engines seem to lose the cat-and-mouse game that is SEO spam.

2 Related Work

SEO is an integral part of today's web. Lewandowski et al. [22] estimate that at least 80% of all web pages use it in some form. To show this, they employ 41 measures to assess the degree of optimization, among which are checks for SEO plugins, lists of URLs, page-level HTML features, or load speed. A recent study by Schultheiß et al. [33] investigates the compatibility between SEO and content quality on medical websites with a user study. The study finds an inverse relationship between a page's optimization level and its perceived expertise, indicating that SEO may hurt at least subjective page quality. Other studies and expert interviews by the same authors conclude that despite its prevalence, (German) lay users are largely oblivious of SEO and its effects [21,34], and that less knowledgeable users tend to trust Google in particular more than others [33].

SEO is a double-edged sword. On the one hand, it makes high-quality pages easier to find, but is on the other hand also a sharp tool for pushing up low-quality results in the search rankings. This necessarily begs the question whether topical relevance is a good proxy for utility and user satisfaction. User-based effectiveness measures are typically modeled after the notion of *gain*, where a user interacts with the result list, accumulating utility from encountered documents until they decide to stop [6,25,26]. Costs incurred by interacting with complex search result pages (SERPs) can also be included to improve stopping rank prediction [3]. However, the framework is mostly descriptive and still relies on a good definition of utility, which in practice comes down to user click data and topical relevance judgments by third-party annotators. This leaves a gap [37] between the quantitative detached proxy measures of relevance and the qualitative goal of utility that can be widened by adversarial optimization.

Epstein and Robertson [12] demonstrate the power of SEO to influence the outcome of elections, but to our knowledge, little research has been published on how to combat it. Recent retrieval systems consider ranking fairness [28,32,36] to avoid biasing the results towards individual providers. Although motivated differently, this can potentially avoid over-ranking individual highly optimized pages, but De Jonge and Hiemstra [10] already demonstrate that fairness measures are insufficient to prevent SEO in general. Kurland and Tennenholtz [19] also find that besides generic spam detection, little research on adversarial content optimization exists. They further seem to agree with us in that this may be due to the difficulty of modeling competitive processes with static test collections and thus call for a rigorous game-theoretical search modeling framework.

In our study, we also investigate the role of affiliate marketing in product reviews and its relationship with SEO. Previous work on affiliate marketing often focuses on fraud rather than SEO abuse. Most affiliate fraud falls into one of four categories [1,11]: (1) conversion hijacking via adware or loyalty software (i.e., adding affiliate tags to links on the fly through a malicious client-side software), (2) cookie-stuffing [8,35] (i.e., planting malicious cookies in users' browsers),

(3) typo squatting (i.e., buying domain names with typos and redirecting to the target domain with an added affiliate tag), and (4) user tracking with affiliate cookies. Affiliate link spam is often a part of "long-tail" SEO spam [16,23], where low-frequency queries are targeted with spam gateway pages to dominate niche queries. Unfortunately, spam mitigation research [2,7,9,24] rarely considers affiliate marketing at all.

Since our work focuses on product reviews, it is also related to research on fake reviews [27], review spam in general [17], as well as review quality and helpfulness assessment [29]. However, these studies focus more on user-contributed reviews on retail websites and less on editorial content in blogs or dedicated product test and review portals.

3 Data and Feature Extraction

To analyze prevalence and impact of SEO spam in product reviews on search engines, we created a large collection of top 20 SERPs for 7,392 product review queries. The SERPs were scraped repeatedly over the course of a year from Startpage (a privacy frontend to Google), Bing, and DuckDuckGo. The linked pages were archived as Web ARChive (WARC[1]) files. From those, we extracted a set of page-level features inspired by Google's SEO [14] and affiliate marketing guidelines [13] that indicate text complexity and quality, HTML page structure, the use of affiliate marketing, and whether a page looks like a product review.

Product Review Queries. To find review pages for a wide range of products, we curated a large list of search queries of the form "best *product category*," where *product category* is a placeholder for a category taken from one of two publicly available product category taxonomies: (1) the GS1 Global Product Classification (GPC, November 2021) and (2) the Google Product Taxonomy (GPT, Version 2021-09-21). We combined the leaf nodes of both taxonomies and cleaned them up by excluding categories we expected would produce atypical results or that don't lend themselves for actual product reviews, such as live animals, crops, fresh produce, and large vehicles like airplanes or boats. We manually reviewed the resulting queries and discarded near-duplicates and queries with artifacts or poor wording, resulting in a final list of 7,392 unique search queries. The lists contains several typical product search queries like *best headphones*, but also many long-tail queries like *best anvils* or *best alphabet toys*.

Commercial Search Engines. We retrieved results from Startpage, Bing, and DuckDuckGo as representative commercial search engines. Although DuckDuckGo claims to utilize many different data sources, we found the results to be extremely similar to Bing. For all queries, we retrieved the top 20 English-language (organic, non-ad) SERPs, resulting in about 148,000 hits and 128,000 unique URLs per scrape. The first collection was done after Google's July product reviews update on August 24th, 2022. The scrape was repeated around every two weeks starting October 26th, 2022, until September 19th, 2023. The period

[1] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/.

spans eleven substantial updates to Google's ranker (among which are three helpful content updates and three product review updates). From the retrieved result lists, we crawled the corresponding web pages, and archived them into WARC files. For the Startpage results, we also included page assets and rendered screenshots for later use.

Baseline Retrieval. We employed the research search engine ChatNoir [5] as a baseline, which offers access to the ClueWeb22B [30], a recent collection of the 200 million most popular web pages, via a web-based API. ChatNoir at its core uses an Okapi BM25 retrieval model and hence serves as a basic and purely document-based whitebox retrieval baseline. The querying and archiving process is analogous to the commercial search engines.

SEO Page Features. For our analysis of the page SEO properties, we compiled a set of page features inspired by Google's SEO [14] and affiliate marketing guidelines [13]. We adapted the guidelines that could be operationalized most easily at page level without rendering the page or executing dynamic content.

The features resulting from this process measure (1) the length and lexical diversity of the main content (extracted using the Resiliparse library [4]), `<a>` anchor and `<img>` alt texts, `<meta>` descriptions, and `<h1>` headings by extracting word and character counts, type-token ratios (TTR; the ratio of unique words to total word count), function word ratios (FWR; the ratio of function words to total word count), and Flesch reading ease scores [18]; (2) the structuredness of a page by counting `<h1>`, `<h2>`, `<img>`, and `<a>` tags, the ratio of `<p>` and `<h[1-6]>` elements to main content words, the use of Open Graph or JSON linked data (JSON-LD), and the existence of breadcrumb navigations; (3) the length and depth of the page URL, i.e., how many files and directories follow after the domain; (4) the number and ratio of affiliate links, the ratio between affiliate links and main content, the number and ratio of site-internal links, as well as the use of `nofollow` link relations; (5) the reuse of topic keywords from headings in the remaining content to measure keyword stuffing.

The features were calculated on a total of 6.6 million results from Startpage (Google), Bing, and DuckDuckGo, and 122,000 results from ChatNoir. We used English as the search language, though we also received a few German results due to unavoidable geo-personalization. Pages without detectable main content were discarded. We extracted the same features also from another 79 million English-language pages from the raw ClueWeb22B dataset as a representation of the long-tail web behind the retrieval frontends.

Review Classification. To test whether a page actually is a review, we performed a simple regular-expression-based keyword classification of the `<h[1-6]>` page headings. Phrases such as "best ...," "top picks," "our favorite ...," "how to use ...," "...we've tested," "what is the best ...," "...review," or various combinations thereof with numbers or other keywords are indicative of review content. We evaluated this approach by drawing a balanced random sample of 100 positively and 100 negatively classified pages from all Startpage and Bing

scrapes. These were then annotated manually as either *review* or *non-review* by the two main authors of this paper with almost perfect agreement (Cohen's $\kappa = .96$). Based on this ground truth, the review classification accuracy was 79% and the precision 85%, which we find decent for such a simple classifier.

Affiliate Link Analysis. We counted the number of affiliate links placed on a page by comparing all anchor URLs to a list of typical patterns that we compiled for the nine largest and most influential online affiliate networks. These are in alphabetical order: *Ali Express*, *Amazon Associates*, *Awin*, *CJ*, *ClickBank*, *eBay*, *FlexOffers*, *Refersion*, and *ShareASale*. The list is based on publicly available web information about affiliate network market shares and participating seller counts. We consider a web page to use affiliate marketing if at least one anchor URL in the HTML source matches one of the patterns. To increase the recall of this method, we resolved short links from `bit.ly`, `amzn.to`, `ebay.us`, and `fxo.co` with a single HEAD request prior to matching them against the list of patterns.

Website Content Categorization. For a qualitative analysis of the contents of the SERPs, we manually annotated the top-30 domains of each scrape (cf. Table 1) with the following seven classes: (1) *Authentic Review Sites* serving high-quality comparative reviews and real product tests (e.g., `nytimes.com/wirecutter`, `consumerreports.org`); (2) *Magazines*, news papers, or other editorial pages which also discuss and review products as a (less serious) side hustle—often as a separate division or on a dedicated subdomain (e.g., `nymag.com`); (3) *Review Content Farms* producing low-effort product listicles, pseudo-reviews, and buyer's guides in large quantities, but with (superficial) editorial content on the side—these are sometimes also found on separate subdomains of otherwise more reputable sites (e.g., `reviewed.com`); (4) *Review Spam* consisting of seemingly generated product listicles without any genuine editorial content (e.g., `blinkx.tv`); (5) *Web Shops* like `amazon.com`; (6) *Social Media* or other community sites with user-generated content. (7) *Other* sites like product manufacturer websites or anything else that doesn't fall into the other classes.

The annotation was done independently by the two main authors with substantial agreement (Cohen's $\kappa = .70$, Accuracy $= .76$). The raters disagreed in 7 cases between review farms and magazines and in 5 cases between review farms and spam. This only speaks to the overall low quality standards for affiliate web content, making it increasingly difficult to distinguish benign content from low-grade or spam content. We resolved these disagreements in favor of the site.

4 SEO Spam in Product Reviews

Our first concern in this work is the general prevalence of low-quality SEO content and spam in search engine results and its driving motivation. We therefore first analyze which of the content features are predictive of rank and thus indicate potential SEO engineering on a page. We then analyze quantitatively and qualitatively the use of affiliate marketing as a measure of monetization level.

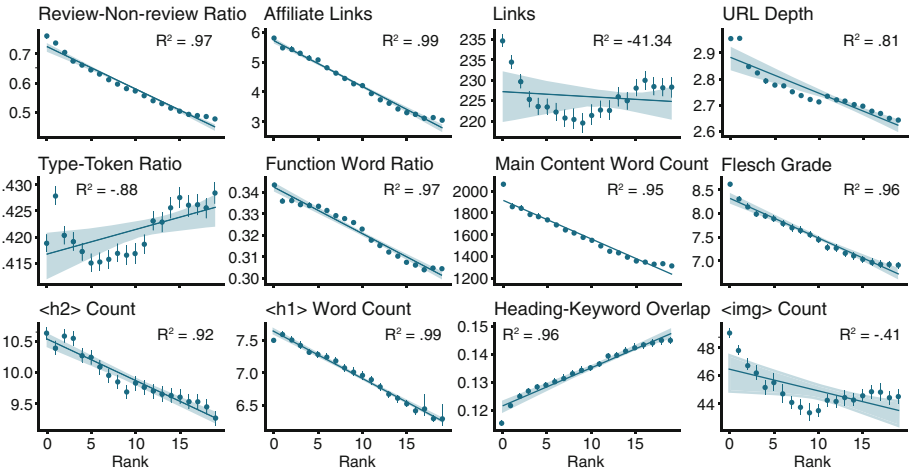

Fig. 1. Selected correlations between rank (independent variable) and the average number of affiliate links, review status, etc. across all Startpage and Bing scrapes. Most averages (not all shown here) correlate either perfectly with rank or have at least a non-linear, non-monotonic relationship. Error bars indicate the bootstrapped 95% confidence intervals of the rank bins. Shaded areas indicate the 95% confidence intervals of the regression line. Note: This plot shows global trends and allows no conclusions about individual pages, since each point is only the mean over all pages at that rank.

Measuring SEO in Review Pages. Across the SERPs of all Startpage and Bing scrapes, we find that rank is indeed a very good predictor of most of our page features. The inverse is not necessarily true, meaning that our page features are not effective SEO exploits, but are nonetheless able to measure SEO at a global population level with highly-correlated sample means. Figure 1 shows a selection of SEO features and their correlation with rank, which we discuss in the following.[2] Cleaned of extreme pages with more than 40 affiliate links per page, these features point to SEO engineering in that pages with better (i.e., numerically lower) ranks are more repetitive (FWR: $r = -.99$; TTR: $r = .59$, $p = .006$) but also more readable (Flesch: $r = .94$). They are also indicators of lower-quality, possibly mass-produced, or even AI-generated content. The FWR relationship is stronger than the TTR relationship, which breaks down for the top-5 ranks. Highly ranked pages also have shallower URLs ($r = -.92$) and longer main content text ($r = -.98$), and are more structured, i.e., they have a lower ratio of text paragraphs to headings ($r = -.99$). Contrary to our initial expectations, pages with better ranking also have a lower overlap between heading keywords and body text ($r = .98$), which indicates that headings become—although more keyword-heavy–also more generic and less specific to the content on the page. Based on manual inspection, a possible explanation could be that "Review" pages become increasingly "thinner" with more affiliate links, i.e., they become more list-like with only a bit of filler text for every featured product that

[2] r is Pearson's correlation coefficient with $p \ll .001$, unless stated otherwise.

provides little value for a user. Some features are more weakly correlated, but still show remarkable non-monotonic and non-linear or piecewise linear relationships (such as number of links or images on a page, which are negatively and only approximately linearly correlated until rank 10 and then reverse direction).

The inverse correlations where the page features themselves are used as independent variables (binned and cleaned of outliers beyond the 95th percentile) to predict the rank also hold, although with smaller effect sizes. Given the nature of global averages, the effect sizes are large enough to detect a trend, but too small to predict the actual ranking. Thus, we can say that longer pages are on average ranked higher ($r = -.91$) and have more affiliate links ($r = -.81$), but the absolute regression coefficient is quite shallow and close to a horizontal line around an average rank 10 (of 20) with medium to high determination ($R^2 = .81$). This means our features can measure certain global effects of SEO, but are not sufficient ranking factors in and of themselves (which would have been quite surprising anyway).

Affiliate Marketing on the Web. Our analysis of the relationship between SEO and the use of affiliate marketing in product reviews in particular reveals a strong positive relationship. First, pages with affiliate links are much more common on Startpage (29%), Bing (42%), and DuckDuckGo (41%) SERPs than on ChatNoir SERPs (18%) and vastly more common than in the ClueWeb22 overall (2.35%). The largest affiliate network across all result pages is *Amazon Associates* by an order of magnitude, followed by *Awin*, *ShareASale*, *CJ*, and *eBay*. There are major differences in the overall numbers of pages with affiliate links returned by the different search engines. Startpage retrieves on average ca. 12,000 pages with 1–10 affiliate links for all product queries. Bing and DuckDuckGo return almost 20,000 in the earlier scrapes and ca. 16,000 in later scrapes (more on this in Sect. 5). ChatNoir retrieves the fewest affiliate pages with only 9,400. For the range of 10–20 links, the search engines return 9,000 (Startpage), 13,000–18,000 (Bing/DuckDuckGo), and 8,200 pages (ChatNoir).

Second, higher-ranked pages have clearly more affiliate links ($r = -.99$, see Fig. 1). Comparing the mean of all pages with the median and 95th percentile of affiliate pages, this is best explained by a mix of both individual pages with high affiliate counts and more affiliate pages in general among the top ranks. We find no conclusive relationship between rank and normal (non-affiliate) links on a page across the whole top-20 range ($r = -.15$, $p = .523$). This confirms that highly ranked pages have indeed more affiliate links and not only more links in general, though non-affiliate links are indeed correlated for first 10 ranks ($r = -.92$, $p = .0002$). The inverse relationship is not nearly as strong ($r = -.62$, $p = .001$) with weak linear determination ($R^2 = -.59$), so, thankfully and unsurprisingly, affiliate links alone cannot predict the rank.

Third, our qualitative site content classification (see Table 1) shows that several spam domains are frequently among the top ranks, some with hundreds of links per page (see also Fig. 3). All inspected pages with more than 100 affiliate links were from spam sites and pages with more than 20 links were at least increasingly likely to be from spam or low-quality affiliate review farm sites

Table 1. Number of websites per review content category for all search engine scrapes (top 20 websites for Startpage, Bing, DuckDuckGo, top 30 for ChatNoir).

Class	Startpage		DDG		Bing		ChatNoir	
Authentic Review Site	1	(2%)	3	(5%)	2	(3%)	0	(0%)
Magazine	14	(31%)	17	(27%)	14	(23%)	2	(7%)
Review Farm	7	(16%)	9	(15%)	9	(15%)	3	(10%)
Spam	4	(9%)	19	(31%)	14	(23%)	1	(3%)
Web Shop	14	(31%)	10	(16%)	9	(15%)	15	(50%)
Social Media	5	(11%)	4	(6%)	3	(5%)	5	(17%)
Other	0	(0%)	0	(0%)	11	(18%)	4	(13%)

designed primarily to harvest clicks. Of all search engines, ChatNoir returns the fewest pages with excessive amounts of affiliate links and the fewest sites in the spam and review farm categories. Bing and DuckDuckGo are especially vulnerable and frequently return up to 2–5 times as many spam pages as Startpage or ChatNoir. As a result, we will base most further analyses only on pages with fewer than 40 affiliate links, which corresponds to the 95th percentile of Startpage results (90th for Bing/DuckDuckGo, 96th for ChatNoir, 99.9th for the ClueWeb22). We find this amount of blatant yet well-ranked affiliate spam peculiar and concerning and it goes to show how important thorough spam filtering is. It is unclear whether on- or off-page SEO (such as link networks) helped in making these spam domains visible. We found through `archive.org` that some of the identified spam sites were quite likely sold or hijacked (such as `socialmoms.com` or `distrotest.net`), while others (like `pulptastic.com`) intentionally mix spam reviews with (possibly generated or scraped) editorial content.

Reviews Vs. Non-reviews. Of the retrieved pages, more than half are identified as review pages by our keyword classification (Startpage: 54%, Bing/DuckDuckGo: 59%). Again, we see that the global mean likelihood of being a review ($r = -.98$) is almost perfectly predicted by a page's rank. We take this as strong evidence that the search engines fundamentally understood our information need. ChatNoir retrieves fewer review pages in total: only 39% of the result pages are reviews (ClueWeb22 base rate: 7%). There are several valid explanations for this behavior: (1) the retrieval algorithm is worse and does not retrieve review pages well, (2) the ClueWeb is a smaller dataset containing fewer review pages in total; (3) SEO on the review pages targets live search engines and ChatNoir simplistic BM25 model ignores many features that identify those pages as relevant to more advanced retrieval systems, such as Google's; (4) Google and Bing as the two major search engines and thus primary SEO targets are particularly vulnerable, which is also a consequence of (3).

Reviews being mostly monetized with affiliate marketing—which relies on trust and visibility—makes at least some amount of SEO quite likely, which

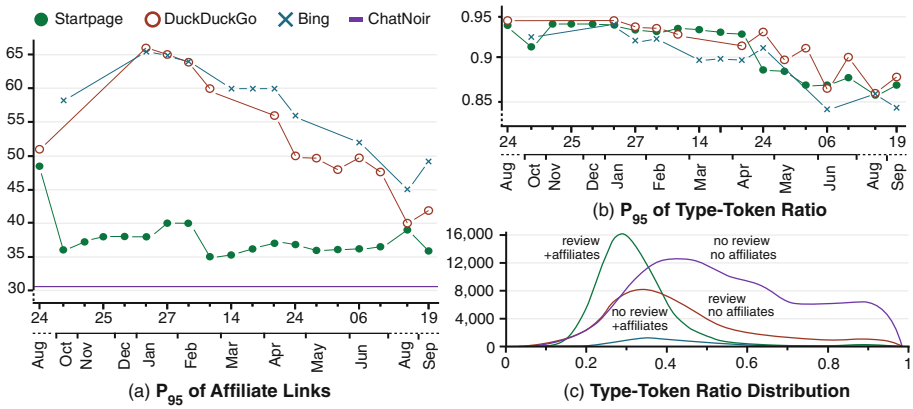

Fig. 2. (a) 95th percentile of per-page affiliate links in the SERPs for all search engines over time (including pages with more than 40 affiliate links). (b) 95th percentile of TTR over time. (c) Count and TTR distribution split by review vs. non-review and with vs. without affiliate links.

conflicts with users' needs for accurate and unbiased information. This is a compelling argument, because a page's likelihood of both being a review but also spam share the same predictor: affiliate links. Figure 2c shows the distribution of type-token ratio (TTR) values over all pages. Review pages that use affiliate marketing have the overall lowest TTR. Review pages without affiliate links and non-review pages with affiliate links have a slightly higher TTR and non-review pages without affiliate links have the highest. This shows that highly commercialized pages are on average simpler and use more repetitive vocabulary, which is a strong indicator of lower-quality content.

5 Temporal Analysis of Product Reviews

Our second major contribution in this work is the temporal analysis of the search quality in terms of (1) the prevalence of pages from the "Review Content Farm" or "Spam" categories (Sect. 3) and (2) SEO-indicating content features.

A temporal analysis of the product review search results should reveal one of three trends: (1) Search engines are truly getting worse, i.e., they are losing the battle against SEO content. In this case we should see a long-term increase in spam and a decline in overall quality. (2) Search engines are winning the upper hand and we see the inverse of this. (3) SEO is a constant battle and we see repeated breathing patterns of review spam entering and leaving the results as search engines and SEO engineers take turns adjusting their parameters. Our evidence suggests that all search engines have some success to show for. Particularly Bing and DuckDuckGo have substantially improved their results, albeit on an overall lower level than Startpage (Google). Yet despite these gains, it seems like (3) is the most likely scenario.

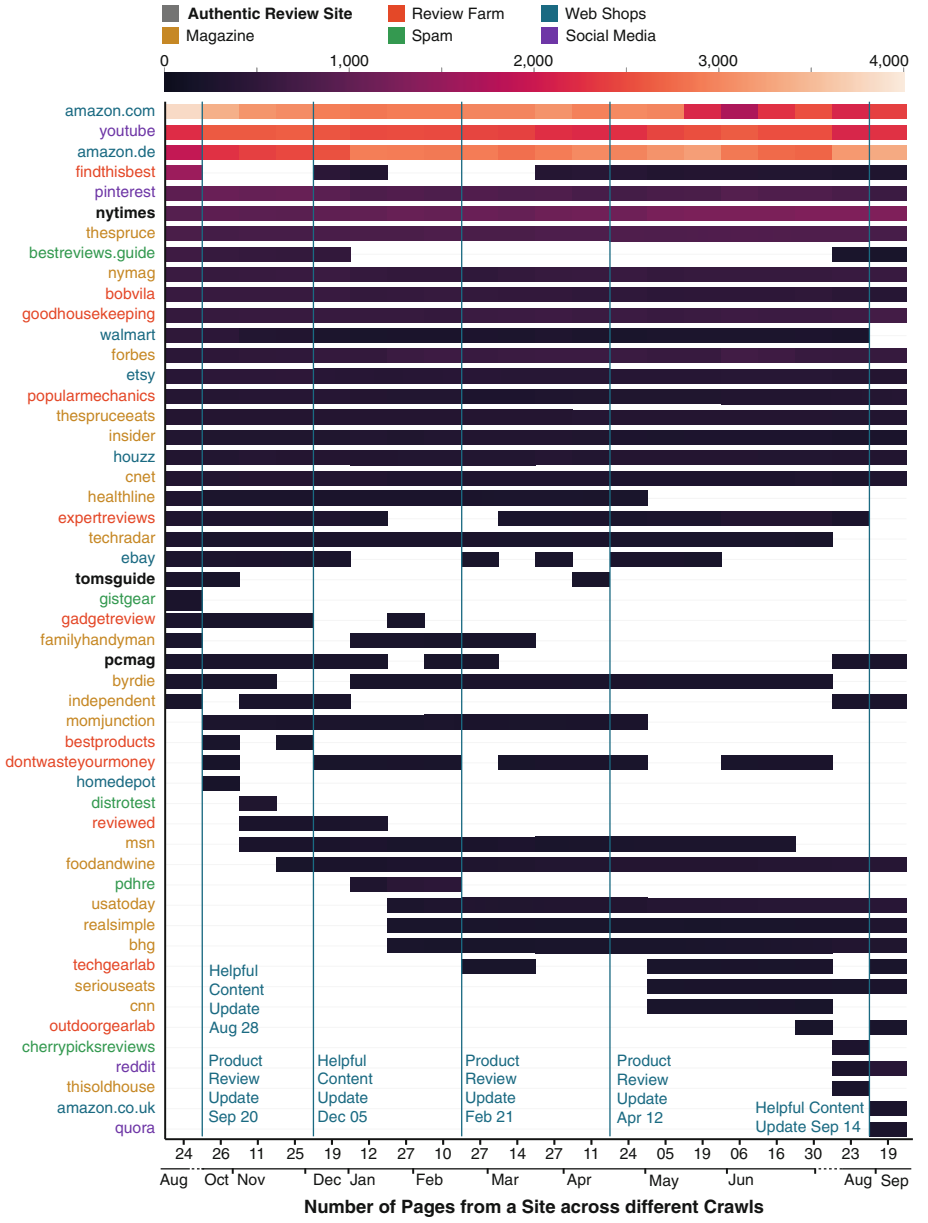

Fig. 3. The 30 most common sites from the Startpage SERPs and their frequencies over time (as page counts per scrape). Blue lines indicate Google's ranker updates. Sites were manually categorized based on the content they served at their first retrieval time. (Color figure online)

Frequent Websites Over Time. Figure 3 shows the union page counts of the 30 most common domains from each of our Startpage scrapes on a time axis, as well as notable ranker updates that happened between the scrapes. Each domain was manually classified into one of the categories from Sect. 3.

We see that *review spam* sites are usually short-lived and de-indexed or penalized quickly, especially after ranker updates. *Review farm* sites and some more frequent *spam* sites such as `findthisbest.com` and `bestreviews.guide` are more persistent and remain throughout multiple consecutive scrapes. However, they often vanish (at least for a while) after Google's ranker updates, either immediately or up to two weeks later, which is the usual update rollout time. The "August 2022 Helpful Content Update" and the "September 2022 Product Reviews Update" combined had the most profound effect, shaving off ranks from all major sites and causing several spam sites to disappear entirely.

Reputable sites, (*web shops*, *social media*, *authentic review sites*) often appear in many consecutive scrapes at similar frequencies. *Magazines* and pure *product review farms* are somewhat, but not as persistent. The most common category of new pages that enter the top 30 for the first time seems to be *magazines*, which hints that having a separate low-quality review section to support a site's primary content is a successful and lucrative business model.

Table 1 shows that *magazines* are also the dominant category besides *web shops*. Bing and DuckDuckGo are also notably less robust against *review spam* than Startpage. Some of the *review spam* sites we identified still existed at the time of the last scrape. Most of them, however, were already defunct at that point. A common pattern we observed is that these start off as seemingly legitimate review pages, but at some point flip to delivering mass spam in the form of scraped or directly embedded Amazon search results. It is unclear what exactly triggers the sudden disappearance of such a spam page, but a significant update to Google's ranking seems like a likely explanation.

SEO Content Over Time. Our SEO content features analyzed over time reveal positive effects of the ranker updates, particularly in the average number of affiliate links per page. Figure 2a shows the 95th percentile of average per-page affiliate link counts by search engine. After September 2022, a large drop can be observed in the Startpage results, while Bing and DuckDuckGo continue to climb until around January. The February update produces another dent in Startpage's curve and the April update yet another albeit smaller one. On the other hand, we also see affiliate pages slowly regaining their lost momentum between updates, indicating a constant struggle. Yet, even after all updates, the commercial search engines still have significantly higher percentile scores than the ChatNoir baseline and although the baseline does return spam domains, it usually does so with much lower frequency.

Interestingly, Google started downranking at least some affiliate pages in the last two of our scrapes, starting end of August, 2023 ($r = -.92$; previously: $r \approx -.99$), resulting in a significantly flatter regression coefficient ($\beta = -.07$, $R^2 = .81$; previously: $\beta \approx -.20$, $R^2 \approx .98$). Whether this is a short-lived change or a lasting trend remains to be seen.

The most profound change during our measurement period, however, is in Bing's (and thus DuckDuckGo's) affiliate link count per page. Between February and August 2023, Bing reduced their 95th percentile by almost 70%, yet still hovered at a rather high level at the end of the measurement period. With Bing being less transparent about their ranker updates, we cannot tell if this is due to a massive crackdown on affiliate spam on their behalf or just windfall gains from Google's updates causing certain pages to disappear. The effects are softened, though still visible if we exclude pages with more than 40 affiliate links.

We see further in Fig. 2b that the average page type-token ratio has decreased consistently over time across all search engines, which shows that while mass affiliate spam may have been contained to some degree, the overall content quality may not have improved.

6 Conclusion

In this paper, we investigate the common observation that "Google is getting worse" by examining its search results for its susceptibility to SEO-driven low-quality content along with those of other major search engines, and in comparison to baselines. We focus on product review search, which we consider particularly vulnerable to affiliate marketing due to its inherent conflict of interest between users, search providers, and content providers.

We conduct two main analyses. First, we investigate what kind of content is retrieved by product review queries and how much SEO influences rankings in this web genre. We correlate page-level quality attributes with search engine rank and find strong relationships between them. Although we cannot predict the rank of individual pages, at the population level, we can conclude that higher-ranked pages are on average more optimized, more monetized with affiliate marketing, and they show signs of lower text quality.

Second, we examine how search results change over time and whether the changes made by search engine operators improve the overall quality of the results. We find that search engines measurably target SEO and affiliate spam with their ranker updates. Google's updates in particular are having a noticeable, yet mostly short-lived, effect. In fact, the Google results seem to have improved to some extent since the start of our experiment in terms of the amount of affiliate spam. Yet, we can still find several spam domains and also see an overall downwards trend in text quality in all three search engines, so there is still quite a lot of room for improvement.

The constant struggle of billion-dollar search engine companies with targeted SEO affiliate spam should serve as an example that web search is a dynamic game with many players, some with bad intentions. Addressing this kind of dynamic, fast-changing, and monetization-driven adversarial SEO content is difficult to do with static evaluation. Going forward, we plan to evaluate how we can better build and evaluate truly robust web IR systems in competitive environments.

Acknowledgments. This publication has received funding from the European Commission under grant agreement № 101070014 (OpenWebSearch.eu).

References

1. Amarasekara, B., Mathrani, A., Scogings, C.: Stuffing, sniffing, squatting, and stalking: sham activities in affiliate marketing. Libr. Trends **68**(4), 659–678 (2020)
2. Asdaghi, F., Soleimani, A.: An effective feature selection method for web spam detection. Knowl.-Based Syst. **166**, 198–206 (2019)
3. Azzopardi, L., Thomas, P., Craswell, N.: Measuring the utility of search engine result pages: an information foraging based measure. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, pp. 605–614. Association for Computing Machinery, New York, NY, USA, 27 June 2018. https://doi.org/10.1145/3209978.3210027
4. Bevendorff, J., Potthast, M., Stein, B.: FastWARC: optimizing large-scale web archive analytics. In: Wagner, A., Guetl, C., Granitzer, M., Voigt, S. (eds.) 3rd International Symposium on Open Search Technology (OSSYM 2021). International Open Search Symposium, October 2021
5. Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Elastic ChatNoir: search engine for the ClueWeb and the common crawl. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 820–824. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_83
6. Carterette, B.: System effectiveness, user models, and user utility: a conceptual framework for investigation. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 903–912. Association for Computing Machinery, New York, NY, USA, 24 July 2011. https://doi.org/10.1145/2009916.2010037
7. Castillo, C., Donato, D., Gionis, A., Murdock, V., Silvestri, F.: Know your neighbors: web spam detection using the web topology. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 423–430. Association for Computing Machinery, New York, NY, USA, July 2007
8. Chachra, N., Savage, S., Voelker, G.M.: Affiliate crookies: characterizing affiliate marketing abuse. In: Proceedings of the 2015 Internet Measurement Conference, IMC 2015, pp. 41–47. Association for Computing Machinery, New York, NY, USA, October 2015. https://doi.org/10.1145/2815675.2815720
9. Chandra, A., Suaib, M., Beg, R.: Google search algorithm updates against web spam. Inform. Eng. Int. J. **3**(1), 1–10 (2015)
10. De Jonge, T., Hiemstra, D.: UNFair: search engine manipulation, undetectable by amortized inequity. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, pp. 830–839. Association for Computing Machinery, New York, NY, USA, 12 June 2023. https://doi.org/10.1145/3593013.3594046
11. Edelman, B., Brandi, W.: Information and incentives in online affiliate marketing. Citeseer (2013)
12. Epstein, R., Robertson, R.E.: The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proc. Nat. Acad. Sci. U.S.A. **112**(33), E4512–21 (2015). https://doi.org/10.1073/pnas.1419828112
13. Google Search Central: Affiliate programs (2022). https://developers.google.com/search/docs/advanced/guidelines/affiliate-programs. Accessed 17 June 2022
14. Google Search Central: Write high quality product reviews (2022). https://developers.google.com/search/docs/advanced/ecommerce/write-high-quality-product-reviews. Accessed 17 June 2022

15. Gregori, N., Daniele, R., Altinay, L.: Affiliate marketing in tourism: determinants of consumer trust. J. Travel Res. **53**(2), 196–210 (2014). https://doi.org/10.1177/0047287513491333
16. Gyongyi, Z., Garcia-Molina, H.: Spam: it's not just for inboxes anymore. Computer **38**(10), 28–34 (2005)
17. Heydari, A., Tavakoli, M.A., Salim, N., Heydari, Z.: Detection of review spam: a survey. Expert Syst. Appl. **42**(7), 3634–3642 (2015)
18. Kincaid, J.P., Fishburne, R.P. Jr., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel (1975)
19. Kurland, O., Tennenholtz, M.: Competitive search. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, pp. 2838–2849. Association for Computing Machinery, New York, NY, USA, 7 July 2022. https://doi.org/10.1145/3477495.3532771
20. Lewandowski, D., Kerkmann, F., Rümmele, S., Sünkler, S.: An empirical investigation on search engine ad disclosure. J. Am. Soc. Inf. Sci. **69**(3), 420–437 (2018)
21. Lewandowski, D., Schultheiß, S.: Public awareness and attitudes towards search engine optimization. Behav. Inf. Technol. **42**(8), 1025–1044 (2023). https://doi.org/10.1080/0144929X.2022.2056507
22. Lewandowski, D., Sünkler, S., Yagci, N.: The influence of search engine optimization on Google's results: a multi-dimensional approach for detecting SEO. In: WebSci, pp. 12–20. ACM (2021)
23. Liao, X., Liu, C., McCoy, D., Shi, E., Hao, S., Beyah, R.A.: Characterizing long-tail SEO spam on cloud web hosting services. In: Bourdeau, J., Hendler, J., Nkambou, R., Horrocks, I., Zhao, B.Y. (eds.) Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, 11–15 April 2016, pp. 321–332. ACM (2016). https://doi.org/10.1145/2872427.2883008
24. Liu, J., Su, Y., Lv, S., Huang, C.: Detecting web spam based on novel features from web page source code. Secur. Commun. Netw. **2020** (2020)
25. Moffat, A., Thomas, P., Scholer, F.: Users versus models: what observation tells us about effectiveness metrics. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM 2013, pp. 659–668. Association for Computing Machinery, New York, NY, USA, 27 October 2013. https://doi.org/10.1145/2505515.2507665
26. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. Secur. **27**(1), 1–27 (2008). https://doi.org/10.1145/1416950.1416952
27. Mohawesh, R., et al.: Fake reviews detection: a survey. IEEE Access **9**, 65771–65802 (2021)
28. Morik, M., Singh, A., Hong, J., Joachims, T.: Controlling fairness and bias in dynamic learning-to-rank. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, pp. 429–438. Association for Computing Machinery, New York, NY, USA, 25 July 2020. https://doi.org/10.1145/3397271.3401100
29. Ocampo Diaz, G., Ng, V.: Modeling and prediction of online product review helpfulness: a survey. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 698–708. Association for Computational Linguistics, Melbourne, Australia, July 2018
30. Overwijk, A., Xiong, C., Liu, X., VandenBerg, C., Callan, J.: ClueWeb 22: 10 billion web documents with visual and semantic information. arXiv (2022). https://doi.org/10.48550/ARXIV.2211.15848. https://arxiv.org/abs/2211.15848

31. Purcell, K., Rainie, L., Brenner, J.: Search engine use 2012 (2012)
32. Raj, A., Ekstrand, M.D.: Measuring fairness in ranked results: an analytical and empirical comparison. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, pp. 726–736. Association for Computing Machinery, New York, NY, USA, 7 July 2022. https://doi.org/10.1145/3477495.3532018
33. Schultheiß, S., Häußler, H., Lewandowski, D.: Does search engine optimization come along with high-quality content?: A comparison between optimized and non-optimized health-related web pages. In: CHIIR, pp. 123–134. ACM (2022)
34. Schultheiß, S., Lewandowski, D.: "Outside the industry, nobody knows what we do" SEO as seen by search engine optimizers and content providers. J. Doc. **77**(2), 542–557 (2020). https://doi.org/10.1108/JD-07-2020-0127
35. Snyder, P., Kanich, C.: Characterizing fraud and its ramifications in affiliate marketing networks. J. Cybersecur. **2**(1), 71–81 (2016)
36. Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking: a survey, 25 March 2021
37. Zobel, J.: When measurement misleads: the limits of batch assessment of retrieval systems. SIGIR Forum **56**(1), 1–20 (2023). https://doi.org/10.1145/3582524.3582540