THE IMPACT OF ONLINE AFFILIATE MARKETING ON WEB SEARCH

Janek Bevendorff * Matti Wiegmann* Martin Potthast Benno Stein
Bauhaus-Universität Weimar Leipzig University
<given>.<last>@uni-weimar.de

Abstract

From small independent blogs to large commercial product review portals, online affiliate marketing has become ubiquitous on the web. According to the BVDW [1], affiliate marketing was responsible for some 14 % of all German online sales in 2019. Often unnoticed by customers, inconspicuous referral links to online retailers result in a commission being paid on conversion to the referring website, making it an easy stream of passive income. With the growing affiliate market, however, we also observe a growing conflict of interest in online content creators, particularly among those who make a living from testing and reviewing products, which manifests itself in the following development: away from providing high-quality, unbiased reviews and towards maximizing conversions. Besides obvious mass affiliate spam, we noticed a growing amount of search-engine-optimized low-quality content, some of which seem hard to detect even for the large search engines today. In our research, we therefore (1) conduct an exploratory study of the landscape of Amazon affiliate links on the web to get a grasp of the issue and (2) identify website genres that search engines need to pay particular attention to in terms of content quality. To the best of our knowledge, we are the first to formulate the quality dilemma from a search engine operator's perspective.

INTRODUCTION

Most online content is free so that it can be found with search engines—a situation that puts for-profit content creators on the web under pressure to find new methods of generating revenue from their work. Many online financing models are based on advertising, premium subscriptions, dedicated crowdfunding platforms, donations, and, increasingly, affiliate marketing. In an (offline) affiliate marketing program, the affiliate partner refers customers to a seller, and if a sale occurs as a result of it, the affiliate earns a commission. In an online scenario, this usually comes in the form of specially crafted and identifiable product links; the commission depends on clicks and click-to-sale paths [2]. Affiliate marketing relies heavily on the trust relationship between customer and affiliate [3], which makes it attractive to influencers and social media marketers [4]. Many websites use affiliate links in addition to donations and advertising.

Unfortunately, the generally impersonal nature of the web makes it easy to abuse this relationship of trust, with the result that the focus shifts from producing high-quality content to maximizing conversions. To counteract this development, search engine operators publish guidelines on what they consider to be high-quality content [5] and trustworthy affiliate referrals [6]. Unsurprisingly, these guidelines, in practice, also serve as recipes for running highly search-engine-optimized affiliate campaigns. Unlike with other financing models (particularly donations, but also ads), revenue does not directly depend on the quality of the content itself but solely on conversions from referrals. Content creators are tempted to optimize against the search engine: instead of creating user-centric content, they will produce search-engine-centric content, i.e., content that serves as a mere vehicle for a page to be indexed and listed. This strategy works as long as the content conveys sufficient trustworthiness and expertise to a casual user at first glance.

Content designed to abuse search engine rankings in order to push low-quality affiliate content is undesired but not necessarily classic "spam". Search engines need to take action against this kind of search engine optimization (SEO) by either *de-indexing* or *de-ranking* pages. However, detecting such deceptive or low-quality affiliate campaigns is difficult even for established big search businesses and requires a deep understanding of the problem.

Previous work in this area focuses primarily on search-engine-optimized web spam in general [7] and its mitigation [8–11]. While affiliate spam was found to be one of the most frequent forms of web spam, it neglects a fundamental analysis of the relationship between affiliate marketing and page quality with the advent of low-quality (pseudo) review websites. Other work on affiliate marketing abuse centers around security-related aspects, such as planting malicious cookies in users' browsers [12]. Research on fake product reviews [13], review spam [14], or review quality and helpfulness [15] deals primarily with user-contributed reviews on retail websites and not with dedicated (comparative) review websites on the web and how these may be shaped by affiliate marketing as a financing model.

The paper in hand contributes in this regard as follows: (1) We extract Amazon affiliate links and several efficiently computable content metrics inspired by Google's web quality and SEO guidelines from four Common Crawls,[1] (2) we aggregate the metrics and conduct an initial exploratory study to identify classes of affiliate websites with potentially abusive design and behavior, and (3) we introduce a categorization of affiliate websites into seven sub-genres, based on the websites' usage of affiliate links and design goals. We were able to find approximate boundaries between some genres based on only a few key metrics and identified other more problematic genres which search engines should pay particular attention to and for which more and focussed research is needed.

* equal contribution

[1] https://commoncrawl.org/

Feature	2015	2020	2021	2022
Element Descriptions				
`<img>` FWR	–	–	–	–
`<img>` TTR	↘	↗	↗	↗
`<img>` word (avg.)	–	↗	↗	↗
`<a>` FWR	–	–	↘	–
`<a>` TTR	–	↘	↘	↘
`<a>` words (avg.)	–	↗	↗	–
`<meta>` FWR	–	–	↗	↗
`<meta>` TTR	–	–	–	–
`<meta>` words (w/o 0)	–	–	–	–
`<h1>` FWR	–	–	–	–
`<h1>` TTR	–	–	–	–
`<h1>` words (w/o 0)	–	↗	↗	↗
`<title>` length	–	–	–	–
Page Structure				
`<h1>` count	–	–	–	–
`<h2>` count	–	–	–	–
`<p>+<h[1-6]>` ratio	–	↗	↗	↗
`<img>` count	↗*	–	–	–
`<a>` count	↗*	↘	↘	↘*
Data-element count	–	–	–	–
Anchor-to-content ratio	↗	↗*	↗*	↗*
Main Content				
Content word count	↗	↗	↗	↗
Content FWR	–	↘	↘*	↘*
Content TTR	↘	↘	↘	↘
Content Flesch score	–	↘*	↘*	↘*
URL Structure				
URL path depth	–	↘	↘	–
URL path length	–	↘*	↘	↘
URL number of digits	–	↘	↘	↘
URL hyphen ratio	–	↗	↗*	↗*

Table 1: Website quality features and their relationship with affiliate link counts for affiliate web pages in the four Common Crawls. The relationships are marked uncorrelated (–), increasing (↗), or decreasing (↘) for pages with 1–35 links. (*) indicates that a correlation also holds for 35–100 links.

CC Name	Total		With Affiliate Links		
	Pages	Domains	Pages	Domains	Suffixes
CC-2022-05	2.9B	35.5M	3.5M	160k	153k
CC-2021-04	3.3B	35.3M	3.8M	167k	159k
CC-2020-10	2.6B	36.1M	3.2M	152k	143k
CC-2015-11	1.7B	14.9M	4.7M	165k	121k

Table 2: Page und domain (suffix) counts of the four crawls before and after affiliate link extraction. Suffixes were calculated using the Public Suffix List.

ANALYTICAL FRAMEWORK

To explore the characteristics of affiliate websites, we derived 28 approximate characteristics of site quality that can be operationalized in an easy-to-scale manner, inspired by Google's SEO [5] and affiliate marketing guidelines [6]. We calculated these features on over 15 million affiliate pages from four Common Crawls and aggregated them as page-level macro and domain-level micro averages.

Operationalizing Page Quality

In the following, we lay out our 5-step process of operationalizing Google's guidelines:

1. Rephrasing free-text recommendations that either encourage (*"do"* or *"use"*) or discourage (*"avoid"*) an action into 61 plain, imperative statements (e.g., *"avoid keyword stuffing anchor texts"*).
2. Removal of 24 statements that cannot be measured using surface-level text features from the page HTML source code, e.g., because they require executing JavaScript (e.g., *"avoid interstitial popups"*), rendering CSS, resolving link graphs (e.g., *"avoid broken links"*), or which cannot be reproduced from a Common Crawl (e.g., *"avoid distracting advertisements"*).
3. Removal of 10 statements too expensive to calculate at scale, such as *"avoid grammar and spelling mistakes"* or *"avoid complex navigation patterns"*.
4. Removal of 10 statements that are irrelevant, such as *"use explicit image filenames (will be distorted by most CMS)"*, or included in other statements, such as *"use less than 1,000 links on a page"*.
5. Engineering of 28 numeric or boolean features for the remaining statements.

With this procedure, we compiled a feature set that measures (1) the length and lexical diversity of `<a>` anchor and `<img>` alt texts, `<meta>` descriptions, and `<h1>` headings by extracting word and character counts, type-token ratio (TTR) and function word ratios (FWR); (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, and the existence of Open Graph or JSON linked data (JSON-LD); (3) the length, diversity, and readability of the main content, and the ratio of words in affiliate link anchors to main content words as a measure of link spam; (4) length and structure of the page URLs. We further calculated the number of affiliate and non-affiliate links. Table 1 lists all features and their behaviors on the four crawls.

Since product review pages naturally use affiliate marketing, we also computed the review-to-non-review ratio (Figure 1d) depending on the affiliate link count. For this, we naively classified a website as "product review" if its headline elements contained typical review phrases such as "`Best` *n*", "`top picks`", or "`Review`".

We also calculated Bahri et al.'s GPT-2 page quality proxy measure [16] on two of the crawls but were unable to find any correlation with affiliate links or actual page quality, which suggests that the extracted page contents, even though of vastly varying quality, were not (yet) GPT-generated.

Figure 1: Page quality features averaged over the number of affiliate links on the 2020–22 Common Crawls. In relation to the number of affiliate links, we show **a)** the (log-scaled) number of individual pages and domain suffixes, **b)** the median per-page main content word count, **c)** the mean and median number of pages per domain, and **d)** the percentages of review to non-review pages, main content TTR and FWR, as well as the ratio of affiliate link anchor words to main content words.

Web Data and Extraction Pipeline

Our explorative study is based on four Common Crawls from 2022, 2021, 2020, and 2015 with 10.5 billion pages total, from which we extracted all pages containing affiliate links. For simplicity, we considered only *Amazon Associates*, as it is the largest affiliate network, and left other networks for future work. Table 2 gives an overview of the page and domain counts per crawl before and after affiliate link extraction. To get a more accurate idea of how many affiliate websites there are, we stripped the domain names of their subdomains using Mozilla's Public Suffix List (PSL).[2] Since the popular blogging platform `wordpress.com` is one of the most frequent second-level domains in the crawl but, unlike `blogspot.com`, not listed in the PSL, we added an extra rule for this domain.

In the three most recent crawls, about one in 900 pages and one in 300 domains contain Amazon affiliate links. For the 2015 crawl, these values are much higher, with one in 360 and one in 90, respectively. Overall, we detected affiliate links on 15.1 million web pages but decided to exclude the 2015 Common Crawl from further analysis due to its highly skewed distribution compared to all other crawls. This skew stems mainly from artifacts caused by obvious clusters of spam websites with many individual pages. Removing this crawl left roughly 10.4 million pages from 270,400 unique domain suffixes.

We extracted the main content of all remaining web pages using the Resiliparse library [17] and removed non-English pages and pages with fewer than 500 words. We then calculated all page-level features from the previous section on an Apache Flink cluster using the Apache Beam Python SDK. After the extraction, we also calculated the domain suffix-level mean and median as micro-average statistics for all numeric features and majority votes for boolean features. Finally, we indexed all page- and domain-level feature statistics to an Elasticsearch cluster for further analysis.

EXPLORATIVE ANALYSIS OF AFFILIATE MARKETING WEBSITES

To check our hypothesis that affiliate marketing incentivizes search-engine-centric design over page quality, we investigated page quality feature aggregates and their relationship to the number of affiliate links on a page. We found that all quality features for which we could detect a correlation do indeed indicate an overall trend towards decreasing page quality with more affiliate links on a page.

Table 1 lists the relationships of all quality features with the number of affiliate links on a page. Figure 1 shows plots of a selection of the features. We focus our analysis on pages

[2] https://publicsuffix.org/

with 35 or fewer affiliate links. Up to this threshold, pages on the order of at least 10^5 and domains on the order of at least 10^4 remain (Figure 1a), leading to stable observations across the crawls from 2020, 2021, and 2022. As mentioned previously, we excluded the 2015 Common Crawl, as it appears to contain large amounts of spam. Even though the trends of some feature statistics are reproducible there as well, many were inconclusive and distorted by spam domains with massive amounts of individual pages.

The overall strongest text quality feature across all crawls turned out to be the type-token ratio (TTR), which measures the lexical variety of a text. High amounts of repetition, as expected from repeated keywords and phrases in largely generated or heavily search-engine-optimized content, leads to an overall lower lexical variety and therefore to a lower TTR of the main content and `<a>` link anchors. Surprisingly, the TTR of (usually invisible) `<img>` alternative texts tends to increase, which may indicate keyword stuffing (i.e., the listing of many noun keywords for SEO purposes).

Another effective indicator of a low-quality website is the use of ill-formed English and frequent keyword stuffing. We measured this with the function word ratio (FWR), the Flesch reading ease score, and the word count in `<img>` alt texts, `<meta>` descriptions, `<a>` anchors, and `<h1>` headings. We found that the FWR decreased in the main content, hinting at overall lower text quality. We further observed a reduction in the Flesch reading ease score from 60 to 50, which indicates more complex texts with longer words and sentences. The average number of words in `<img>` alt texts, `<meta>` descriptions, `<a>` anchors, and `<h1>` elements also increases (excluding empty elements), another typical indicator of keyword stuffing or synthetic text.

Additionally, we found that the total number of links on a page negatively correlates with the number of affiliate links on it. This correlation means that affiliate websites tend to use, on average, fewer non-affiliate links, which would indicate an overall simpler page structure. A manual review of the most frequent websites with different numbers of affiliate links confirmed our observations. Pages with many affiliate links frequently listed excessive amounts of product titles and specifications, resulting in a low function word count, high repetitiveness, more complex words, and fewer sentence boundaries.

We found other quality features inconclusive or potentially misleading, such as the overall consistent increase in main content length. One might expect the amount of main content text to correlate with higher page quality. However, we found that websites with a higher number of affiliate links become spammier, more repetitive, and more synthetic. This observation is further supported by the increasing affiliate anchor-to-content ratio, which means that a larger portion of the main content is within affiliate link anchor texts when there are many affiliate links.

As shown in Figure 1, there appears to be a consistent relationship between several quality features and the number of affiliate links up to about 30–40. Above this range, most metrics show a sudden increase in variance and asymmetry,

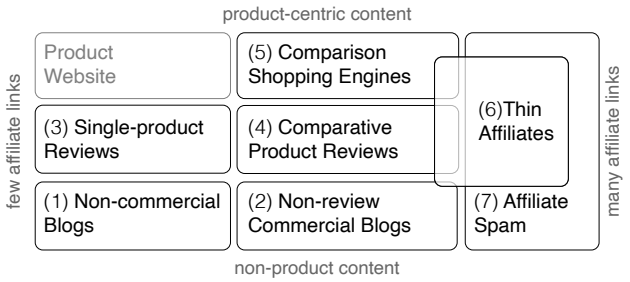

Figure 2: Schematic affiliate website genre categorization. We can characterize affiliate websites by the number of links and by the level of their integration into the main content.

influencing both the median and the arithmetic mean. This is largely explained by the exponential decrease in the number of samples with, at the same time, a sustained increase in the number of pages per domain (Figure 1c) and an overall change in genre. A manual review confirmed that websites beyond this range are almost exclusively affiliate spam or other search engine abuse. Our ad-hoc keyword classification of product review pages provides further evidence: The vast majority ($> 75\%$) of websites are marked as "product review" past the 30–40 link range.

Finally, because the 2015 Common Crawl appears to contain significantly larger spam clusters than the newer crawls, we discovered evidence of a noticeably improved crawling strategy. Yet even if much improved, the newer crawls are not perfect by any means, and the crawler still rather frequently runs into spam domains with many individual pages. This is evident since the mean and median number of pages per domain doubles between one and 35 affiliate links (Figure 1c). Beyond this threshold, there are many extreme domains (some were omitted as outliers). Limiting the number of pages crawled from sites with many affiliate links may potentially increase the quality of future crawls.

CLASSIFICATION OF AFFILIATE SITES

Our explorative analysis indicates that the affiliate count already explains much of the qualitative differences between websites. Affiliate websites with many affiliate links (40+) are often synthetic spam or thin affiliates as per our below definition, and websites with 10–40 links are increasingly low-quality, affiliate-centric review websites. However, qualitative analysis reveals that many low and medium link frequency websites use affiliate links more like ads: Websites place those links beside their content in marked sponsor sections, sidebars, footers, or on about pages. The content is not related to the linked products.

By observing these different usage patterns, we were able to classify most affiliate websites based on their quantitative (how many links) and their qualitative (depth of integration in the content) use of affiliate marketing (AM) into one of seven genres:

1. *Non-commercial blogs* that use AM to pay for expenses,

2. *commercial non-review blogs and news sites* that make at least part of their regular revenue from AM,
3. *single-product reviews* that review individual products,
4. *comparative reviews* that review product ranges,
5. *price comparison* that automatically generates listings of prices and product specifications,
6. *thin affiliate sites* that pretend to publish genuine, high-quality content, though only decorating affiliate referrals with fake or low-effort editorial text that offers no added value, and
7. *affiliate spam*, automatically generated product listings without actual content.

Figure 2 depicts the genres schematically.

Since thin affiliate sites are increasingly found above 20 links per page and spam is increasingly found above 35 links per page and almost universally above 50 links per page, our hypothesis that the quality of content decreases as the number of affiliate links increases is at least partially confirmed. The fact that this trend can already be observed with few affiliate links provides additional support, even though, at this stage, we cannot completely rule out the possibility that undetected spam sites cause part of the effect. Not all genres are, however, inherently harmful. In particular, for pages with only a few affiliate links (which makes up the vast majority of pages), we observed very little spam or other unwanted content looking at the most common sites, and overall a reflection of the diversity of the web as a whole. Pages with more than 50 affiliate links are also unproblematic, as they are almost exclusively spam and can therefore be filtered effectively based on this statistic in conjunction with other traditional web spam detection methods. So the problematic genres we need to examine more closely are the ones in between. We could already confirm empirically with a basic keyword analysis in Section 3 that a fair share of these pages can be classified as comparative review pages. We were able to validate this finding by a manual qualitative review. Since the percentage of (comparative) review pages increases as a function of the number of affiliate links with near-perfect monotonicity, we see a large gray middle area in which actual review websites mix with low-effort, low-quality reviews, and thin affiliates.

In light of our findings, telling genuine reviews apart from low-quality pseudo reviews and thin affiliates in the medium range of affiliate links seems a vital objective for search engines. Our basic page- and domain-level macro statistics are insufficient for drawing a sharp boundary between the genres. However, even highly advanced retrieval systems appear to have difficulties with these genres, as is evident from the fact that we found many of the low-quality reviews and thin affiliates that we discovered during our qualitative analysis also in Google's index and, for specific keywords, more often than not, among the top results.

CONCLUSION

In this work, we scrutinize the prevalence of affiliate marketing on the web, its impact on website quality, and its implications for search engine operations. Based on several page- and domain-level text quality features, which we calculated at scale on four different Common Crawl versions, we find indicators that page quality is (on average) negatively correlated with the number of affiliate links, even in website genres that do not count as spam. Most indicative of this finding is a reduction in the main content type-token and function word ratios and an increase in the Flesch reading ease score. These observations support our hypothesis that a conflict of interest indeed exists between the financial incentives of affiliate marketing and the creation of high-quality content, although further research is needed.

We further find that at more than 20–30 affiliate links, the vast majority of affiliate pages can be classified as "product review" pages, and beyond more than 35–40 links, pages are almost exclusively spam. From this observation, we developed a broad seven-class affiliate website genre classification based on the number of affiliate links and the amount to which a page's contents are centered around the referred-to products. While the "affiliate spam" genre is easy to detect due to its excessive use of affiliate links without any actual content, the "thin affiliate" genre and low-quality examples from the "product review" range are much harder to pin down. From the perspective of a search engine operator, these two genres, therefore, require particular attention and suitable page quality measures for ensuring high-quality search results. We also find that limiting the number of pages crawled from domains that fall into the "affiliate spam" genre might improve the quality of future Common Crawl versions.

REFERENCES

[1] BVDW, "Affiliate Marketing generiert jeden siebten Euro im E-Commerce." https://www.bvdw.org/der-bvdw/news/detail/artikel/affiliate-marketing-generiert-jeden-siebten-euro-im-e-commerce/, 2019. Last accessed: Jun 18, 2022.

[2] R. Olbrich, P. M. Bormann, and M. Hundt, "Analyzing the click path of Affiliate-Marketing campaigns," *J. Advert. Res.*, vol. 59, pp. 342–356, Sept. 2019.

[3] N. Gregori, R. Daniele, and L. Altinay, "Affiliate marketing in tourism: Determinants of consumer trust," *J. Travel Res.*, vol. 53, pp. 196–210, Mar. 2014.

[4] A. Mathur, A. Narayanan, and M. Chetty, "Endorsements on social media: An empirical study of affiliate marketing disclosures on youtube and pinterest," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, nov 2018.

[5] Google Developers, "Write high quality product reviews." https://developers.google.com/search/docs/advanced/ecommerce/write-high-quality-product-reviews, 2022. Last accessed: June 17, 2022.

[6] Google Developers, "Affiliate programs." https://developers.google.com/search/docs/advanced/guidelines/affiliate-programs, 2022. Last accessed: June 17, 2022.

[7] Z. Gyongyi and H. Garcia-Molina, "Spam: it's not just for inboxes anymore," *Computer*, vol. 38, pp. 28–34, Oct. 2005.

[8] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "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 '07, (New York, NY, USA), pp. 423–430, Association for Computing Machinery, July 2007.

[9] A. Chandra, M. Suaib, and R. Beg, "Google search algorithm updates against web spam," *Department of Computer Science & Engineering, Integral University*, vol. 3, no. 1, pp. 1–10, 2015.

[10] F. Asdaghi and A. Soleimani, "An effective feature selection method for web spam detection," *Knowledge-Based Systems*, vol. 166, pp. 198–206, Feb. 2019.

[11] J. Liu, Y. Su, S. Lv, and C. Huang, "Detecting web spam based on novel features from web page source code," *Security and Communication Networks*, vol. 2020, Dec. 2020.

[12] N. Chachra, S. Savage, and G. M. Voelker, "Affiliate crookies: Characterizing affiliate marketing abuse," in *Proceedings of the 2015 Internet Measurement Conference*, IMC '15, (New York, NY, USA), pp. 41–47, Association for Computing Machinery, Oct. 2015.

[13] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021.

[14] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Syst. Appl.*, vol. 42, pp. 3634–3642, May 2015.

[15] G. Ocampo Diaz and V. Ng, "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)*, (Melbourne, Australia), pp. 698–708, Association for Computational Linguistics, July 2018.

[16] D. Bahri, Y. Tay, C. Zheng, C. Brunk, D. Metzler, and A. Tomkins, "Generative models are unsupervised predictors of page quality: A Colossal-Scale study," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, (New York, NY, USA), pp. 301–309, Association for Computing Machinery, Mar. 2021.

[17] J. Bevendorff, M. Potthast, and B. Stein, "FastWARC: Optimizing Large-Scale Web Archive Analytics," in *3nd International Symposium on Open Search Technology (OSSYM 2021)* (A. Wagner, C. Guetl, M. Granitzer, and S. Voigt, eds.), International Open Search Symposium, Oct. 2021.