Understanding the Business of Online Affiliate Marketing: An Empirical Study

Haitao Xu*, Yiwen Sun*, Kaleem Ullah Qasim*, Shuai Hao†, Wenrui Ma‡§(✉)
Zhenyuan Li*, Fan Zhang*, Meng Han*, Zhao Li*¶
*Zhejiang University, Hangzhou, China
†Old Dominion University, VA, USA
‡School of Computer Science and Technology, Zhejiang Gongshang University
§Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology
¶Hangzhou Yugu Technology, Hangzhou, China
*{haitaoxu,yiwensun,lizhenyuan,fanzhang,mhan}@zju.edu.cn,†shao@odu.edu,‡mawenrui@zjgsu.edu.cn,§lzjoey@gmail.com

***Abstract*—Affiliate marketing is a revenue-sharing marketing scheme by which an affiliate, such as a blogger or YouTuber, garners commissions for promoting a merchant's goods or services, thereby aiming to foster a mutually beneficial relationship between affiliates and merchants. Despite being a multi-billion-dollar global industry, affiliate marketing remains inadequately explored, and the research community lacks a comprehensive understanding of its intricate ecosystem. In this paper, we present the first comprehensive empirical study of the affiliate marketing ecosystem. We conduct thorough measurements to assess the prevalence of affiliate marketing, estimate the market size, and elucidate the characteristics of affiliates, merchants, and intermediary affiliate networks. Over a continuous span of 13 months, we monitored four of the most prominent affiliate aggregation platforms, yielding a substantial dataset. We observed 467,219 unique offers – tasks to be undertaken by affiliates – involving 37,109 merchants and 556 affiliate networks across the four platforms. Notably, these offers would cost the merchants more than 19 million USD for the completion of all the actions pre-defined in these offers, such as signing up or making a transaction. Additionally, we compiled a large-scale dataset comprising 124,462 affiliate links, enabling us to conduct a comprehensive investigation. Finally, we propose machine learning models incorporating the characteristics of affiliate links to detect real-world affiliate marketing campaigns.**

I. Introduction

Affiliate marketing establishes a revenue-sharing marketing model, wherein a third-party affiliate marketer (hereafter referred to simply as *affiliate*)—such as a blogger or YouTuber—promotes a merchant's goods or services for a commission. It significantly contributes to online purchases, with statistics indicating that it drives 15%-30% of e-commerce sales [3] and that over 30% of new customers are acquired through affiliate links [8]. This form of marketing has evolved into a multi-billion-dollar global industry, with global spending projected to reach approximately $14.3 billion in 2023 and $15.7 billion by 2024 [1].

In addition to the aforementioned major participants (affiliates, merchants, customers) in the affiliate marketing ecosystem, *affiliate networks*, *e.g.*, Golden Goose [17], play another crucial role by serving as intermediaries between affiliates and merchants' affiliate programs. By affiliating with such networks, smaller businesses without their own affiliate programs can readily access more potential customers. Furthermore, affiliate aggregation platforms, *e.g.*, Affplus [4], have emerged as central hubs where diverse affiliate networks or merchants can post affiliate marketing tasks, *i.e.*, offers. This enables affiliate marketers to easily discover and undertake offers from various merchants.

Previous research on affiliate marketing has primarily focused on specific aspects, such as the strategies and factors influencing affiliate marketing performance [10, 14, 16, 26, 27, 28, 30, 33, 36], affiliate marketing fraud [6, 7, 9, 11, 21, 31, 32], and disclosure of affiliate marketing [13, 15, 19, 22, 34, 35, 37]. However, a comprehensive understanding of the affiliate marketing ecosystem—including its prevalence, scale, operations, usage patterns, and participant behaviors—remains understudied. To address this gap, we endeavor to explore a series of unanswered research questions on the ecosystem:

- **The prevalence of affiliate marketing.** What is the financial magnitude of the affiliate marketing ecosystem, and how many merchants and affiliate networks are actively involved?
- **The operation of affiliate marketing.** How much do merchants typically pay affiliates for specified actions, and what promotional strategies do affiliates commonly use?
- **Identification of affiliate marketing.** What distinguishes an affiliate link from a regular hyperlink, and how can these differences be leveraged to identify affiliate URLs?

Our research endeavors to conduct a large-scale empirical investigation aimed at addressing the aforementioned questions. In doing so, we select and monitor the four most prominent affiliate aggregation platforms over a span of 13 months, amassing a comprehensive dataset comprising 467,219 unique affiliate offers. We characterize those offers in terms of their categories, pricing models, payout amounts, as well as the involved merchants and affiliate networks (§IV).

To explore the affiliate marketing ecosystem in the wild, we abstain from the utilization of affiliate link identifiers proposed in prior studies [2, 22]), recognizing their susceptibility to significant false negatives due to the diversity, fluidity, and expansion of the affiliate marketing landscape. Instead, we opt to construct a set of verified affiliate link identifiers. Then, we search online using these affiliate link identifiers, amassing a dataset comprising 124,462 affiliated links embedded within webpages. By correlating this dataset with the offers collected from the four affiliate aggregation platforms, we infer the intentions of merchants, conduct behavior analysis of affiliates, and explore the distinctive characteristics of affiliate links. Our findings unveil discernible disparities between affiliate links and regular links, which could potentially be harnessed for efficient identification of affiliate marketing campaigns, thus empowering users to better discern products or services promoted through such campaigns.

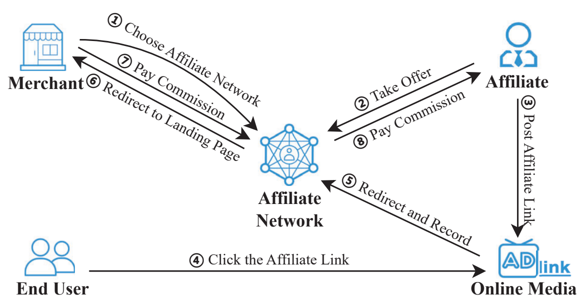


Fig. 1: A typical affiliate campaign lifecycle

II. Background

In this section, we provide an overview of the operations of affiliate marketing and prior endeavors on the identification of such campaigns.

A. Affiliate Marketing

Workflow of affiliate marketing. Figure 1 depicts the typical workflow of an affiliate marketing campaign, which commences with a merchant forging a partnership with an affiliate network to initiate a campaign (❶). Upon registration with the affiliate network, an affiliate, exemplified by a freelancer, gains access to offers tailored to their traffic sources (❷). Upon undertaking an offer and receiving approval from the affiliate network, the affiliate is furnished with a customized link, referred to as an *affiliate link*. This link typically incorporates their identifier, known as the affiliate ID, along with the corresponding merchant's identifier, termed as the merchant ID. Subsequently, the affiliate disseminates this affiliate link across various online platforms, including but not limited to blogs, videos, or social media pages (❸). When a user demonstrates interest and clicks on the affiliate link (❹), they are redirected to the merchant's landing page or product page via intermediary redirections (❺, ❻). Concurrently, the user's browser is furnished with HTTP cookies, serving to link the user's subsequent actions with the specific affiliate and facilitate accurate commission calculation. In the event that the user proceeds to fulfill a predetermined action, such as signing up or completing a transaction, the merchant is contractually obliged to remit a commission fee to the affiliate network (❼), which subsequently distributes a portion to the affiliate (❽). This standardized process is universally adhered to by ethical merchants and affiliates.

Pricing model of affiliate marketing. Affiliate marketing is a type of performance marketing, and affiliates earn money based on their performance through various pricing models such as *CPA*, *CPL*, *CPS*, *RevShare*, and *etc.* Specifically, in the CPA (Cost Per Action) model, affiliates receive compensation when users complete a predetermined action after clicking their affiliate link. In the CPL (Cost Per Lead) model, affiliates are remunerated for generating potential customer leads according to the advertiser's specifications. In the CPS (Cost Per Sale) model, affiliates earn commissions for each successful sale facilitated through their referral links. Lastly, in the RevShare model, affiliates are entitled to a percentage of the profits generated from the sales resulting from their promotional efforts.

Affiliate aggregation platforms. Affiliate aggregation platforms serve as centralized hubs where affiliate networks or merchants can post affiliate marketing offers (or tasks), while affiliates can readily find and engage with offers from a diverse range of merchants.

B. Affiliate Marketing Campaign Identification

Remarkably, there remains a lack of standardized criteria for the composition of affiliate links. According to Google Analytics [2], an affiliate link typically comprises at least two specific parameters, namely `utm_source` and `utm_medium`, which serve to identify the source sharing the link and the associated product or service. Additionally, Mathur *et al.* [22] have outlined a compilation of known affiliate link patterns in their study of endorsements on social media, as depicted in Table I. This collection encompasses a total of 108 affiliate link patterns, corresponding to 39 affiliate programs (or affiliate networks). Apart from these aforementioned studies, there appears to be scant literature addressing the definition of affiliate links. To summarize, in line with existing definitions, a URL qualifies as an affiliate link if the URL itself, the landing page URL, or any intermediate URL meets either of the following criteria: (1) it contains the two specified parameters, `utm_source` and `utm_medium`; (2) it aligns with known affiliate link patterns.

III. Methodology

The methodology utilized in this study is illustrated in Figure 2 and comprises three distinct phases. The initial phase revolves around the acquisition of a representative set of offers from affiliate aggregation platforms. The second phase involves the enhancement of affiliate link identifiers (or signatures) by leveraging established affiliate link patterns. Lastly, the third phase centers on the collection of affiliate links from the Internet via Google search employing the refined affiliate link signatures.

TABLE I: Known affiliate link patterns. Note that in the table, only one pattern is provided for each specific affiliate program for the sake of space conservation, with the values in parentheses indicating the total number of patterns available.

Admitad (2) ad.admitad.com/g/	**AffiliaXe** (2) performance.affiliaxe.com /*?aff_id=	**AliExpress** (2) s.aliexpress.com/*?af=
Amazon (10) amazon.com/*?tag=	**Amazon Prime Video** (2) primevideo.com/*?ref=	**Apple** (4) itunes.apple.com/*?at=
Audiobooks (3) affiliates.audiobooks.com /*?a_aid=*&a_bid=	**AvantLink** (2) avantlink.com/*?pw=	**Avangate** (2) secure.avangate.com /*?affiliate=
Awin (4) awin1.com/*?awinaffid=	**Banggood** (2) banggood.com/*?p=	**Book Depository** (2) bookdepository.com /*?a_aid=
Booking.com (2) booking.com/*?aid=	**Clickbank** (1) hop.clickbank.netˆ	**CJ Affiliate** (8) anrdoezrs.net/click-
DesignModo (1) designmodo.com/?u=	**Ebay** (2) rover.ebay.com/*?campid=	**Envato** (12) audiojungle.net/*?ref=
ePN† (2) buyeasy.by/cashback/	**Flipkart** (2) flipkart.com/*?affid=	**GT Omega Racing** (2) gtomegaracing.com /*?tracking=
HasOffers (2) /aff_c?*&aff_id=	**Hotellook** (2) search.hotellook.com /*?marker=	**Hotmart** (2) hotmart.net.br/*?a=
Impact Radius (2) 7eer.net/c/	**KontrolFreek** (2) kontrolfreek.com/*?a_aid=	**Ladbrokes** (2) online.ladbrokes.com /promoRedirect?key=
Makeup Geek (2) makeupgeek.com/*?acc=	**Pepperjam Network** (6) gopjn.com/t/	**Rakuten Marketing** (2) click.linksynergy.com/*?id=
Skimlinks (2) go.redirectingat.com/*?id=	**Smartex** (2) olymptrade.com /*?affiliate_id=	**RewardStyle** (1) rstyle.meˆ
ShopStyle (1) shopstyle.itˆ	**ShareASale** (3) shareasale.com/r.cfmˆ	**Studybay** (2) apessay.com/*?rid=
TataCLiQ (2) tatacliq.com/*?cid=af:	**Thermoworks** (2) thermoworks.com/*?tw=	**Zaful** (2) zaful.com/*?lkid=

†ePN: e-Commerce Partners Network

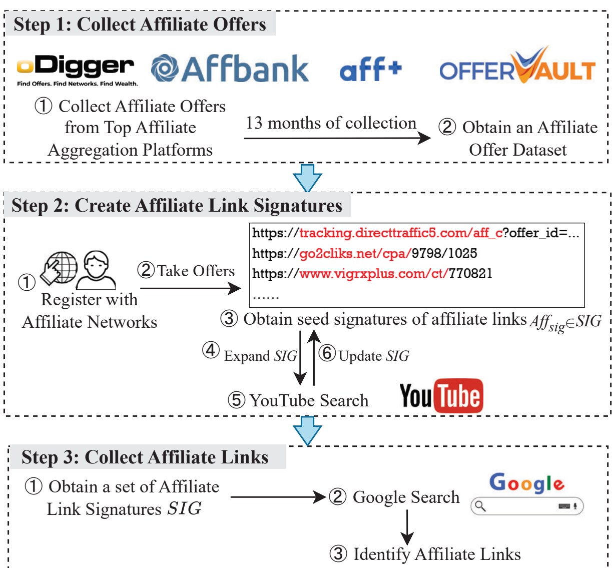


Fig. 2: Data Collection Methodology

A. Collection of Affiliate Offers

To acquire a representative dataset of offers for analysis, we have identified the four leading affiliate aggregation platforms: *Affplus* [4], *OfferVault* [25], *Odigger* [24], and *Affbank* [5], according to PropellerAds, a worldwide premium advertising platform. Employing Selenium, we conducted data collection encompassing various attributes such as title, payout, offer creation and update times, category, geolocation, affiliate network, description, and landing page URL of each offer. Data collection from these platforms spanned from April 2023 to May 2024. Ultimately, we amassed a total of 467,219 offers, with 318,739 originating from Affplus, 23,249 from oDigger, 62,196 from Affbank, and 63,035 from OfferVault.

B. Creation of Affiliate Link Signatures

While previous studies (Google Analytics [2] and Mathur et al. [22]) have offered insights into the composition of affiliate links, our manual validation indicates potential susceptibility to significant false negatives due to the diversity, fluidity, and expansion of affiliate networks. Consequently, we opt to construct a set of verified affiliate link identifiers (or signatures) based on these established affiliate link patterns.

1) Undertaking Offers and Obtaining Seed Signatures of Affiliate Links: Initially, we identified the top 20 affiliate networks boasting the highest number of offers across the four aggregation platforms. Subsequently, following a rigorous procedure, we effectively registered with and established affiliations with 14 of these networks, overcoming various challenges, notably including the submission of verifiable evidence showcasing the quality of our traffic sources.

On an affiliate network, we selectively engage with offers aligned with our traffic sources, leading to the assignment of distinct affiliate links. These links typically consist of a domain name, path, and potential parameters, with certain segments displaying unique patterns identifying the affiliate network, offer, or affiliate. Common substrings among affiliate links from various offers within the same network can serve as a signature for future link identification. We term this common URL string the affiliate link signature (Aff_{sig}).

For instance, from a collection of affiliate links such as "https://trezor.io/?offer_id=[variables]," we extract "trezor.io/?offer_id=" as the Aff_{sig}. Similarly, from Amazon affiliate links like "https://amzn.to/[variables]," we derive "https://amzn.to/" as Aff_{sig}. Through manual extraction from 200 offers we undertook across 14 affiliate networks, we identify a total of 33 unique URL patterns serving as Aff_{sig}, all of which are appended to the set SIG. It is worth noting that none of these 33 affiliate link patterns are among the known patterns discussed in §II-B. This underscores the need for enhancements based on existing patterns.

2) Expanding Seed Signatures: The incorporation of affiliate links into the *description* section below a YouTube video is widely recognized as a potent strategy for content creators to monetize their work [18, 20, 23, 29]. Notably, content creators are legally permitted to include multiple affiliate links within this section. This practice is exemplified in Figure 3, which presents a screenshot capturing the description segment of a YouTube video selected from search results generated by querying with the affiliate link signature "trezor.io/?offer_id=." It is evident that the presence of one affiliate link within the description field often indicates the likelihood of others being present. Consequently, we conduct a statistical analysis on these potential affiliate links to identify new patterns, subsequently integrating them into SIG.

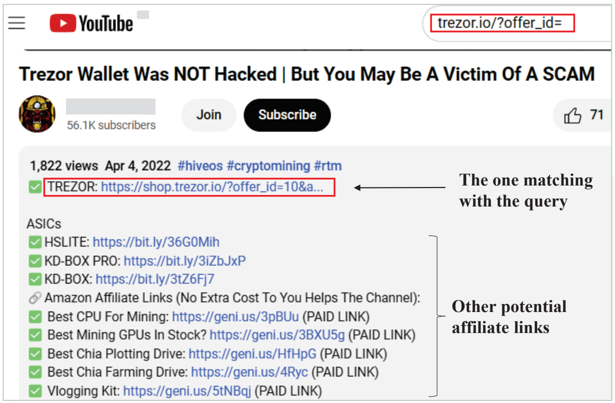


Fig. 3: A scenario wherein numerous affiliate links are situated within the YouTube description.

Applying this methodology, utilizing the 33 affiliate link signatures previously obtained in SIG as search queries, we effectively scraped a grand total of 34,223 YouTube videos featuring description sections containing at least one affiliate link. Through manual examination of these affiliate link candidates and the extraction of common strings, alongside the refinement of signatures based on known affiliate link patterns, the SIG has been expanded to encompass 177 signatures.

C. Collection of Affiliate Links via Google search

To explore the affiliate marketing landscape across the Internet, we leverage our repository of 177 distinct affiliate link signatures (Aff_{sig}) in the set SIG to perform searches on Google's search engine. This method facilitates the comprehensive collection of data from a wide array of websites.

For each signature, we retrieve the HTML source code of the resulting webpages and extract all outbound links embedded within them. Leveraging the notion that webpages containing affiliate links may host additional affiliate links, we subsequently apply the signatures in SIG again to identify affiliate links among these outbound links. Consequently, our efforts result in a compilation of 124,462 unique affiliate links and 628,776 unique non-affiliate links.

D. Ethical Concerns

We refrain from completing the offer process, terminating it immediately after obtaining the unique affiliate link generated for us. Thus, our methodology poses no additional risk to affiliate networks or merchants.

IV. Characterizing Affiliate Marketing Offers

We choose and monitor the four popular affiliate aggregation platforms, including AffBank, AffPay, Odigger, and OfferVault. Subsequently, we compile a thorough dataset comprising affiliate offers. We proceed to characterize these offers based on their categories, pricing models, payout amounts, affiliated merchants, and affiliate networks involved.

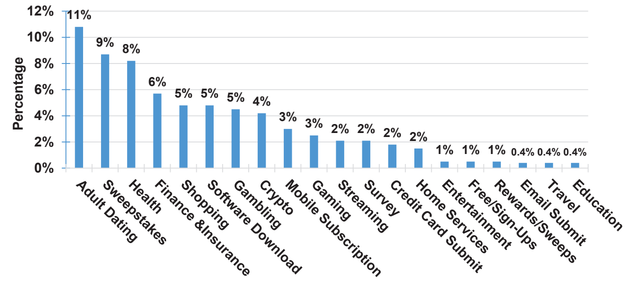


Fig. 4: Top 20 categories of offers on four platforms

TABLE II: Distribution of offers by pricing model

Pricing Model	CPA	CPL	CPS	RevShare	Others
Percentage	81.5%	4.7%	3.7%	2.5%	7.6%

A. Dataset

Over a span of 13 months, spanning from April 2023 to May 2024, we conducted daily crawls of the four affiliate offer platforms, gathering comprehensive information related to the listed offers, including their titles, posting dates, associated affiliate networks, payouts, categories, pricing models, and landing page previews. Table III presents the statistics of the dataset. In summary, across the four platforms, we identified a total of 467,219 unique offers, involving 556 affiliate networks and 37,109 merchants. The commission fees associated with these offers (presumably under a CPA model) exceeded 19 million USD.

B. Characterization of Affiliate Offers

We initially analyze the 467,219 offers by characterizing their categories, pricing models, and payouts.

Distribution of offers by category. Figure 4 depicts the distribution of the top 20 categories of offers. Notably, *Adult Dating*, *Sweepstakes*, *Health*, *Finance & Insurance*, and *Shopping* emerge as the leading five categories, collectively constituting 38.2% of all offers. Particularly noteworthy is the prominence of *Adult Dating* offers, which alone contribute 10.8%. Additionally, *Gambling* and *Crypto* (pertaining to cryptocurrency transactions) categories rank within the top 10, occupying 4.5% and 4.2% of the offers, respectively. Furthermore, *Credit Card Submit* and *Email Submit* are another two interesting categories, typically requiring affiliates to provide credit card information for online sweepstakes trials or submit valid email addresses for surveys.

Distribution of offers by pricing model. Table II presents the distribution of offers by pricing model. It shows that 81.5% offers specify the CPA model, which indicates that most affiliates receive remuneration solely upon the completion of a merchant-specified action by the end-user.

Distribution of offers by payout. Next, we aim to examine the compensation received by affiliates for engaging in offers, as well as the variances in payout across offer categories. Figure 5 illustrates the payout statistics for the top 20 offer categories. It also highlights that among all the offers collected,

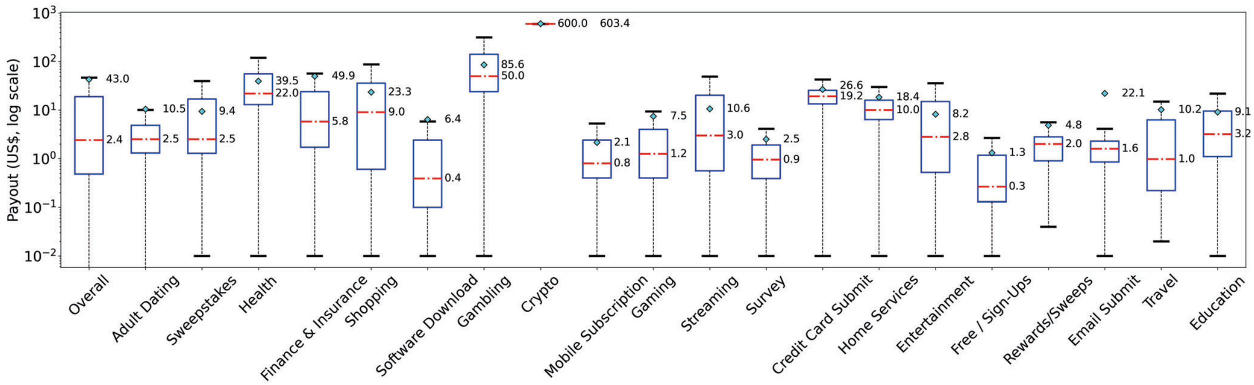


Fig. 5: Distribution of offers by payout

TABLE III: Affiliate offers observed on the four platforms

Platform	#Offers	#Affiliate Networks	Commission (US$) Total (Avg.)	# Merchants
AffBank	62,196	41	2,881,454 (51.52)	9,537
AffPay	318,739	380	12,206,391 (41.83)	32,337
Odigger	23,249	35	1,107,277 (57.46)	6,085
OfferVault	63,035	212	2,967,014 (52.17)	13,009
Total (unique)	**467,219**	556	**19,162,136** (45.21)	37,109

the median and average payouts stand at 2.4 USD and 43.0 USD, respectively (as depicted in the first box). Notably, *Crypto* offers exhibit the highest payouts, with a median of 600 USD and a mean of 603.4 USD. *Gambling* offers rank second, with an average payout of 85.6 USD per offer. Merchants in these categories typically yield substantial profits and demonstrate a greater willingness to offer high commissions to attract new clients. *Adult Dating* offers, the most prevalent on the platforms, average a payout of 10.5 USD. Conversely, *Free/Sign-Ups* offers offer the least compensation, with a median of 0.3 USD and an average of 1.3 USD, attributed to the minimal effort required from end-users.

C. Top Merchants

We also investigate the merchants sponsoring the highest number of offers and their respective categories. In Figure 6 (a), the Top 10 merchants with the most offers are showcased, collectively representing 6.8% of all observed offers. Notably, *Nutra*, the largest global retailer specializing in nutritional products, leads with 7,042 offers, followed by *Samsung* with 4,520 offers. The remaining merchants include *Amazon* (e-commerce), *Orange* (French telecom giant), *Tinder* (popular dating app), *Vodafone* (telecom giant), *McDonald's* (restaurant), *Etisalat* (UAE telecom giant), *Shein* (online retailer), and *Scanero* (Android QR code reader).

Based on the distribution of the Top 100 merchants by category depicted in Figure 6 (b), there are 17 shopping venues, 14 gambling sites, 9 software providers, 9 entertainment portals, 7 survey sites, 7 telecom giants, 7 cryptocurrency transaction platforms, and 5 adult dating apps/sites.

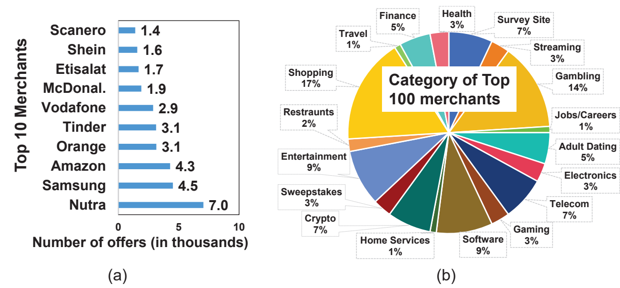


Fig. 6: (a) Top 10 merchants with the most offers; (b) Distribution of category of Top 100 merchants

D. Top Affiliate Networks

Given the pivotal role of affiliate networks, we scrutinize the top affiliate networks regarding their popularity, primary offer categories, leading target countries, pricing models, payouts, and gross commission payments for the offers.

Table IV outlines the top 10 affiliate networks with the highest number of offers across the four platforms, along with corresponding statistics. It reveals that most affiliate networks (excluding *FuzeClick* and *Appitate*) collaborate with multiple offer platforms to publish their available offers. The top three affiliate networks—*FuzeClick*, *Golden Goose*, and *MyLead*—account for over 15% of all observed offers, each publishing more than 20,000 offers over approximately one to two years. Notably, *FuzeClick* boasts the highest frequency of offer publication, at 80.8 offers per day. Additionally, many affiliate networks specialize in particular offer categories; for instance, *Algo-Affiliates* dedicates 90.7% of its offers to the *crypto* category, indicating a strong focus on partnering with cryptocurrency transaction platforms. Furthermore, certain affiliate networks such as *Golden Goose*, *MyLead*, and *TORO Advertising* prominently feature *adult dating* offers, accounting for between 26.8% and 38.0% of their respective portfolios. Meanwhile, *Zeydoo* primarily specializes in *software download* offers, constituting 56.1% of its offers, while more than a third of *AdsMain* offers focus on *sweepstakes*.

Moreover, offers typically target specific countries for traffic, with the United States (US) being highly desired. 7 out of

10 affiliate networks prioritize the US as their primary target country. The CPA pricing model emerges as the most popular among 9 out of 10 affiliate networks (excluding *MyLead*), aligning with the statistics presented in Table II. Regarding payouts, *Algo-Affiliates*, primarily specializing in *crypto* offers, offers the most generous compensation, averaging 547.5 USD per offer. *MyLead* and *AdsMain* also offer substantial payouts. Intriguingly, the top three affiliate networks offering high payouts also contribute the highest gross commission fees for their offers in our dataset, totaling 8,251,244 USD, 681,059 USD, and 206,667 USD, respectively.

V. CHARACTERIZING THE CAMPAIGNS IN THE WILD

Through Google searches for affiliate marketing campaigns in the wild using the affiliate link signatures from SIG, we compiled a substantial dataset comprising 124,462 affiliate links and 628,776 non-affiliate links for our analysis.

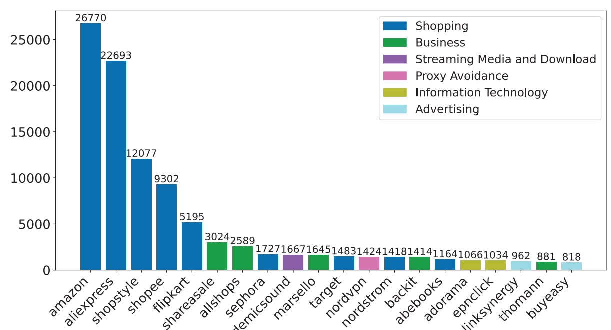


Fig. 7: Top 20 merchants and their categories

A. Merchants

We follow each affiliate link and document any redirections encountered before reaching the final landing page. By examining the landing pages of affiliate links, we obtain details about the corresponding merchants. We scrutinize the associated merchants and discern the intentions behind their product pages (*e.g.*, sign-ups or transactions).

1) Top Merchants Sponsoring Affiliate Marketing Campaigns: For each affiliate link, we extract the domain name from its corresponding landing page URL and then query the category of the domain name in *Fortiguard*, a threat intelligence platform. Subsequently, we identify the corresponding merchants of the domain names and illustrate the top 20 merchants and their categories in Figure 7.

The rationale for considering merchants instead of directly counting domain names is that a single merchant may possess multiple domain names (*e.g.*, the two domains `www.amazon.com` and `www.amazon.it` both belong to Amazon). The top 5 most frequently appearing merchants are *Amazon*, *AliExpress*, *ShopStyle*, *Shopee*, and *Flipkart*, all falling under the category of *shopping*.

Furthermore, we conduct an analysis on the ranking of domain names within each category, utilizing the Tranco top sites rankings. We scrutinize the distribution of rankings across the top 20 categories. The findings indicate that in categories such as *Adult Materials*, *Health*, *Society and Lifestyles*, *Content Servers*, and *Finance*, over 50% of domain names do not fall within the top 1 million rankings. Notably, in the *Adult Materials* category, a mere 1.41% of domain names are positioned within the top 1 million. Non-top 1 million domain names constitute a significant 98.59% of the total. These findings underscore the challenges faced by merchants utilizing affiliate marketing as a promotional strategy in garnering user attention in the absence of such marketing campaigns.

2) Merchants' Spent on Affiliate Marketing: To assess the spending of merchants per end-user's action (presumably under a CPA model) for affiliate marketing campaigns, we aim to establish correlations between identified affiliate links and offers listed on the four affiliate aggregation platforms (as discussed in §IV), for which we possess information regarding their commission fees. Specifically, for each affiliate link, we determine its corresponding merchant, then search the four affiliate aggregation platforms for offers sponsored by the same merchant. We compute the average commission fee paid by the merchant across these platforms and estimate the total payout required for each end-user's successful engagement with the affiliate link.

Out of 145,361 affiliate links, we successfully located the destination URLs and determined the corresponding merchants. For 78,938 (54.3%) of these affiliate links, we harvested matching offers on the platforms.

Among the affiliate links with matching offers, the top 5 merchants are *Amazon* (33.9% of affiliate links), *AliExpress* (28.9%), *Shopee* (11.8%), *Flipkart* (6.6%), and *Sephora* (2.2%), collectively accounting for 83.4% of the affiliate links, all falling within the *shopping* category. We compute the average payout per offer for each merchant on the four platforms (*e.g.*, \$25.4 for Amazon, \$9.7 for AliExpress, \$0.36 for Shopee, \$0.47 for Flipkart, and \$0.96 for Sephora). Subsequently, we estimate the gross commission fee paid by merchants on the 78,938 affiliate links, totaling \$1,141,068 per end-user's engagement on these affiliate links.

3) Intention of Merchants' Landing Pages: Next, we explore the purpose behind a merchant's landing page, where an end user follows an affiliate link and is finally redirected.

Owing to the dynamic nature of webpage content, as well as the complexities of design patterns and evolving user interactions, supervised methods face challenges in accurately inferring webpage intentions. Therefore, we turn to large language models (LLMs) such as GPT-3.5, which are trained on extensive portions of the Internet and proficient in capturing the nuanced structures and contexts of web content.

Due to budget constraints, we randomly select a subset of 13,000 affiliate links from the total 124,462 for this analysis of webpage intention. Specifically, we crawl the HTML source code of these affiliate links' landing pages. Subsequently, we compose prompts containing these source codes along with a list of intention labels, and feed them into GPT-3.5 Turbo, requesting the model to generate a label for each webpage's

TABLE IV: Statistics of the top 10 affiliate networks with the most offers

Affiliate Network	Offer Platforms	Observed Offers	Duration & Frequency§	Main Category (percentage)	Top Target Country	Pricing Model	Commission Fee (avg.)	Gross Commission
FuzeClick	AffPay	25,442 (5.5%)	315 days (80.8)	Software† (21.3%)	**US (33%)**	**CPA (100%)**	$3.00	$74,292
Golden Goose	AffBank, AffPay, OfferVault	23,353 (5%)	517 days (40.4)	**Adult‡ (26.8%)**	SA (5%)	**CPA (99.8%)**	$2.60	$60,611
MyLead	AffBank, AffPay Odigger, OfferVault	21,300 (4.6%)	754 days (23.6)	**Adult (30.3%)**	**DE (17.7%)**	CPS (34.8%)	**$47.9**	**$681,059**
Adgate Media	AffBank, AffPay	15,587 (3.3%)	632 days (24.7)	Survey (3.2%)	**US (30.2%)**	**CPA (100%)**	$5.73	$89,368
Algo-Affiliates	AffPay, OfferVault	15099 (3.2%)	388 days (34.3)	**Crypto (90.7%)**	**Worldwide (92.9%)**	CPA (54.2%)	**$547.5**	**$8,251,244**
Zeydoo	AffBank, AffPay Odigger, OfferVault	11,782 (2.5%)	856 days (13.8)	**Software (56.1%)**	US (4.3%)	**CPA (100%)**	$4.1	$45,799
Adscend Media	AffBank, AffPay	10,117 (2.2%)	787 days (12.9)	Free/Sign-Ups (23.1%)	**US (29.6%)**	**CPA (100%)**	$2.7	$25,590
Appitate	AffPay	8,918 (1.9%)	302 days (29.5)	—♭	US (13.1%)	**CPA (100%)**	$1.2	$10,073
TORO Advertising	AffPay, Odigger, OfferVault	8,768 (1.9%)	282 days (27.1)	**Adult (38.0%)**	US (20.5%)	**CPA (78.2%)**	**$16.3**	$127,761
AdsMain	AffPay, OfferVault	8,732 (1.9%)	498 days (12.2)	**Sweepstakes (35.5%)**	US (34.3%)	**CPA (69.3%)**	**$23.9**	$206,667

§frequency: average number of offers published per day. †Software - Software Download. ‡Adult - Adult Dating. ♭Mostly returned as N/A blank.

TABLE V: Webpage Intention Classification

Category	Confidence Score	# Webpages	Percentage
Financial Transactions	95%	7,475	57.50%
Information Retrieval	85%	1,181	9.08%
User Interactions	98%	835	6.42%
Acquisitions	90%	946	7.28%
System Functions & Operat.	87%	1,799	13.84%
User Verification & Regist.	95%	700	5.38%
Total	92%	13,000	100%

source code. The classification results of webpage intentions are presented in Table V.

The results overwhelmingly support our assumption that a significant proportion of the analyzed webpages exhibit transactional intentions, particularly in the realm of *e-commerce*. Specifically, 57.50% of the analyzed webpages fall under the *Transactional* category, with the model demonstrating an average confidence score of 95% for these classifications.

To validate the accuracy of GPT's classification results, we manually examined the webpage content of 100 webpages and assigned labels for their intentions. We found that our manually assigned labels align with those produced by GPT, thereby demonstrating the effectiveness of leveraging LLMs for such webpage intention inference tasks.

B. Affiliates

Next, we aim to characterize affiliates, a pivotal role within the ecosystem responsible for disseminating merchants' product information to end users, by analyzing where and how frequently affiliates post affiliate links.

To conduct this behavior analysis of affiliates, the initial step is to identify the unique identifier associated with each affiliate, referred to as the *affiliate ID*. Subsequently, we gather the affiliate links posted online by each affiliate, utilizing their respective affiliate ID as a means of identification.

Step 1: affiliate ID extraction. An affiliate ID is commonly embedded within the query parameters or the path of an affiliate link, serving the purpose of tracking the traffic and sales generated by the affiliate. Notably, the location of an affiliate ID within an affiliate link typically remains consistent for links generated by the same affiliate network. With this observation, we discern specific patterns from affiliate links.

For instance, all affiliate links generated by *SellHealth*, a health-related affiliate network, adhere to the format "*https://[merchant_name]/ct/[affiliate_id]/xxx*." By crafting regular expressions tailored to extract the designated segment (*e.g.*, */ct/[affiliate_id]*) as an affiliate ID, we effectively identify affiliate IDs. This process yields 324 affiliate IDs.

Step 2: affiliate link collection. We conduct individual queries for each of the 324 affiliate IDs using Google Search and utilize the Serpapi API to collect the search results. This procedure results in a total of 2,944 affiliate links, along with corresponding website information and the date time.

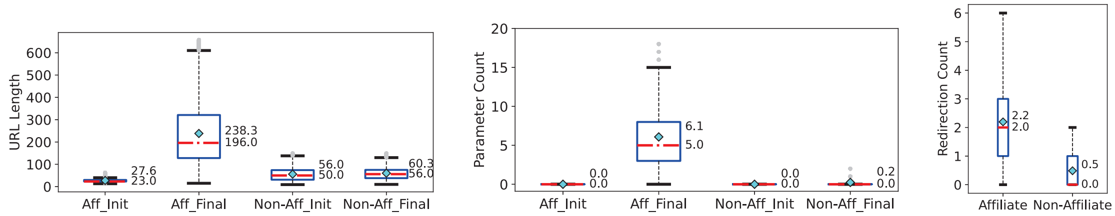
Online platforms where affiliate links are shared. The top 5 platforms favored by affiliates for publishing affiliate links are *YouTube*, *Facebook*, *Pinterest*, *Quora*, and *Twitter*. Among all 324 affiliates analyzed, 55.86% utilize YouTube as their primary platform, followed by 19.44% who choose Facebook, 10.94% who prefer Pinterest, 7.72% who select Quora, and 5.56% who opt for Twitter. Additionally, on average, each affiliate utilizes 3 different platforms for posting affiliate links.

Work routine of affiliates. We analyze the work habits of the top 10 affiliates who have published the most affiliate links in our dataset. Collectively, these affiliates have published between 52 and 254 affiliate links, averaging 105 links each.

By examining the time information associated with the online platforms where affiliate links are embedded, we observe significant variations in the length of time these affiliates have been active. The most experienced affiliate has been consistently engaged in marketing activities since 2011, spanning an impressive 4,293 days. Conversely, the least experienced affiliate began activities in December 2023. On average, the top 10 affiliates have been active for approximately **1,522 days**, roughly four years.

In terms of product promotion, each of the top 10 affiliates promotes an average of 9 distinct products. They typically post 2-3 affiliate links per active day. It is worth noting that Google search results may not capture all behaviors of an affiliate and therefore these statistics data represents the lower bound only.

Notably, one affiliate stands out by consistently posting an average of 35 affiliate links daily on *Rumble*, an online video-



(a) Difference of URL length (b) Difference of parameters (c) Redirection count

Fig. 8: Difference between affiliate links and non-affiliate links in URL length, parameters and redirection count

sharing platform, during late August 2023. Remarkably, all of these affiliate links are dedicated to promoting a product intricately associated with *matched betting*. Such behavior deviates from typical affiliate marketing practices, where affiliates often diversify platforms, promote various products or services, and distribute affiliate links evenly to broaden their audience and maximize revenue potential.

C. *Affiliate Links*

With our dataset comprising affiliated and non-affiliated links established, next we delve into the distinct characteristics of affiliate links compared to non-affiliate links. To achieve this, we follow each link and document any redirections encountered before reaching the final landing page. For affiliate links specifically, we also capture the HTML source code of the landing page. This redirection data enables us to quantify the number of redirections between a link and its landing page URL, as well as track any changes in URL length and the number of embedded parameters. Through this analysis, we aim to uncover differences between affiliate and non-affiliate links both before and after redirections.

1) Difference between Affiliate Links and Non-Affiliate Links: We conduct a comparative analysis between affiliate links and non-affiliate links across various dimensions, such as URL length, parameters before and after redirection, and the number of redirections.

Difference in URL Length. Figure 8 (a) depicts the difference between affiliate links and non-affiliate links in terms of URL length. It is evident that affiliate links tend to have a shorter length, averaging approximately 27 characters. This is often attributed to the utilization of URL shortening services prior to publication. Upon redirection to the final destination page, the average length of affiliate links increases to 238 characters, reflecting the addition of parameters. Conversely, non-affiliate links exhibit a longer initial length, averaging around 56 characters. These links typically undergo minimal changes when redirected, with an average final length of approximately 60 characters, as non-affiliate links generally do not undergo redirection.

Difference in URL Parameter. Figure 8 (b) illustrates the disparity between affiliate and non-affiliate links concerning the number of parameters within a URL. Initially, affiliate links typically lack parameters. However, as they undergo multiple redirections, parameters are appended to these links to denote referral sources, affiliate IDs, product categories, and other pertinent information. On average, final affiliate links encompass 6 parameters. In contrast, both the original and final non-affiliate links generally remain devoid of any parameters.

Difference in URL Redirection. Figure 8 (c) highlights the difference between affiliate and non-affiliate links in terms of the number of redirections. It is notable that affiliate links undergo an average of 2 redirections before reaching the landing page, whereas most non-affiliate links exhibit zero redirections.

The occurrence of multiple redirections for affiliate links can be attributed to two primary motives. Firstly, it facilitates the tracking of user behavior by affiliate networks. Through a sequence of redirections and intermediate links, affiliates can accurately monitor the specific affiliate link through which users entered, as well as track their subsequent actions on the final destination page, such as purchases, registrations, and other interactions. This tracking mechanism is crucial for precisely calculating commissions or rewards.

Secondly, multiple redirections are utilized to obscure the actual affiliate links. By employing a chain of redirections, the true destination of the affiliate link remains concealed, thereby preventing users from bypassing the affiliate system and directly accessing the target page. This ensures that affiliates receive their rightful commissions.

Furthermore, due to the inclusion of multiple parameters in affiliate links, many affiliates utilize URL shortening services to convert them into more concise and visually appealing links. This not only enhances user experience but also improves click-through rates.

2) Methods of Affiliate Links Embedded: To analyze the embedding methods of affiliate links within webpages, we parse the HTML source code and extract tags containing these links. This encompasses instances where affiliate links appear within the text area of a tag (*e.g.*, *<title>[affiliate_link]</title>*) and within attribute values of a tag (*e.g.*, *link*).

Overall, we extracted a total of 172,564 tags, corresponding to 32,899 unique affiliate links. Table VI presents the top 10 most frequently occurring tags and the specific text or attributes where affiliate links are embedded. Remarkably, the top 10 tags collectively represent 99.02% of the total.

The most prevalent tag is `<script>`, occurring 70,419 times and accounting for 40.81% of all tags. Notably, all but one of these affiliate links appear within the text of `<script>` (*i.e.*, part of JavaScript code), while one appears in

TABLE VI: Methods of Affiliate Links Embedded in HTML

Tag	in_attr[1]		in_text[2]	Total (%)
	Attribute	Total		
script	**#data-json** (1)	**1**	**70,418**	70,419 (40.81%)
meta	**#content** (42,560)	42,560	0	42,560 (24.66%)
a	#href (30,657) #aria-label (1,072) #title (1,006)	32,868	2,300	35,168 (20.38%)
span	#title (37) #data-sheets-value (1) #name (1)	39	5,772	5,811 (3.37%)
title	-	-	4,655	4,655 (2.7%)
img	#alt (3,784)	3,834	6	3,840 (2.23%)
h2	-	-	2,888	2,888 (1.67%)
div	#alt (377) #aria-label (525) #title (172)	1,260	1,472	2,732 (1.58%)
p	-	-	2,306	2,306 (1.34%)
link	#href (294) #title (125)	419	72	491 (0.28%)

[1] Affiliate links appear within the attribute value of an HTML tag, *e.g.*, `<a href="https://affiliate.link">link</a>`. Top 3 attribute value names of each tag are listed.

[2] Affiliate links appear within the text of an HTML tag, *e.g.*, `<title>https://affiliate.link</title>`.

the attribute named `data-json`. The second most common tag is `<meta>`, constituting 24.66% of the total, with affiliate links found exclusively within the ``content'' attribute values. The third most frequent tag is `<a>`, making up 20.38%. Together, these three tags encompass roughly 85% of the total.

D. Affiliate Link Identification

TABLE VII: Performance of Affiliate Link Classifiers

Classifiers	Accuracy	Precision	Recall	F1 score
Random Forest	0.967	0.949	0.848	0.888
Logistic Regression	0.913	0.819	0.612	0.696
KNN	0.864	0.571	0.853	0.675
Decision Tree	0.958	0.925	0.817	0.857
Gaussian Naive Bayes	0.899	0.738	0.633	0.674

To systematically detect affiliate marketing campaigns in the wild, we constructed a dataset consisting of 100,000 affiliate links and 100,000 non-affiliate links. We then trained a meticulously curated ensemble of machine learning (ML) classifiers on this corpus for the binary classification task of determining whether a URL is an affiliate or non-affiliate link.

For each URL in the dataset, we engineered a 12-dimensional feature vector encapsulating attributes ranging from affiliate-specific lexemes in URL parameters to quantitative metrics such as URL length and parameter count, as shown in §. A normalization procedure was executed to ensure a homogenized feature space conducive to ML algorithms.

The dataset was partitioned into training and testing subsets, adhering to an 80-20 ratio. To evaluate the models' generalizability and predictive capabilities, a 5-fold cross-validation scheme was implemented during the training phase. The classification results are presented in Table VII, indicating that the Random Forest model achieved the best performance, with an accuracy of 0.967, a precision of 0.949, a recall of 0.848, and an F1 score of 0.888. Thus, machine learning models incorporating the characteristics of links can be effectively leveraged to detect real-world affiliate marketing campaigns.

VI. Related Work

There have been research efforts dedicated to examining various facets of affiliate marketing. Here, we present an overview of prior studies closely aligned with our objectives.

Factors influencing affiliate marketing performance. Patrick *et al.* [28] utilize various technology adoption models to discern the determining factors influencing the intention to use affiliate marketing among SMEs (small-to-medium-sized enterprises). Suryanarayana *et al.* [33] propose a game-theoretic model to analyze agent behavior within affiliate networks. Monika *et al.* [10] specifically concentrate on Amazon's program to identify profitable affiliate marketing niches. Furthermore, Olbrich *et al.* [26] examine the click path of affiliate marketing and analyze the mutual influence between the design parameters of affiliates and search-engine advertising by merchants. Additionally, Ghosal *et al.* [16] evaluate the impact of affiliate marketing strategies on the online shopping attitudes of millennials. Syrdal *et al.* [36] apply the Elaboration Likelihood model to explore the linguistic features of affiliate marketing content and assess the effects of these features on engagement behavior.

Affiliate fraud. Although affiliate marketing holds promise for merchants, it also carries inherent risks, with affiliate fraud being a primary concern. This fraudulent activity is typically aimed at generating false actions, such as auto-clicking or cookie-stuffing, to earn illegitimate commissions [31]. Bede *et al.* [7] shed light on four fraudulent information-seeking behaviors—stuffing, sniffing, squatting, and stalking—and explore strategies to mitigate potential profit loss for Small and Medium-sized Businesses (SMBs). Chachra *et al.* [11] developed a Chrome extension, AffTracker, to detect affiliate cookies and gather data from domains susceptible to fraudulent affiliate activities. They identify the merchant categories most targeted by such scams and the third-party affiliate networks most frequently implicated. However, their focus is limited to six major affiliate programs. Additionally, Mangio *et al.* [21] propose a classification method to differentiate between influencer and non-influencer falsities in affiliate marketing. Furthermore, Chu *et al.* [12] identify six distinct illicit methods employed by creators and Multi-Channel Networks (MCNs) on YouTube, shedding light on their execution tactics.

Disclosure of affiliate marketing. In parallel, research has delved into the disclosure practices within affiliate marketing, examining their impact on sellers, users, and affiliate networks [13, 15, 19, 22, 34, 37]. Mathur *et al.* [22] provide a comprehensive list of known affiliate link patterns, which our study significantly builds upon. Swart *et al.* [35] design and implement AdIntuition, an online tool that automatically identifies YouTube videos containing affiliate marketing content, and conduct a user study to assess its performance.

VII. Conclusion

In this paper, we present a comprehensive investigation into the affiliate marketing ecosystem by compiling an extensive dataset of affiliate offers and links. We systematically analyze the usage patterns and operational behaviors of various participants in affiliate marketing. We develop an enhanced approach for identifying affiliate links by assembling a comprehensive set of affiliate link signatures and propose machine learning algorithms for systematically and efficiently detecting affiliate marketing campaigns in the wild. Our study addresses a significant gap in understanding of affiliate marketing.

Acknowledgment

This work was supported in part by the National Key R&D Program of China (No. 2023YFB3106800), the National Natural Science Foundation of China (No. 62272410), and the Key R&D Program of Zhejiang Province (No. 2023C01039). Wenrui Ma is the corresponding author.

References

[1] 2023. The State of Affiliate Marketing: Benchmark Report. https://shorturl.at/pJUS9.

[2] 2024. Google Analytics: Collect campaign data with custom URLs. https://shorturl.at/8JwFX.

[3] 99firms. 2023. https://shorturl.at/6JizI.

[4] aff+. 2023. https://www.affplus.com/.

[5] AffBank. 2023. https://affbank.com/.

[6] B. Amarasekara and A. Mathrani. 2016. Controlling risks and fraud in affiliate marketing: A simulation and testing environment. In *PST 2016*.

[7] B. Amarasekara, A. Mathrani, and C. Scogings. 2020. Stuffing, Sniffing, Squatting, and Stalking: Sham Activities in Affiliate Marketing. *Library Trends* (2020).

[8] Awin. 2021. https://www.awin.com/gb/power-100.

[9] A. Baharuddin et al. 2023. Affiliate Marketer Advertising Fraud: A Normative Legal and Prophetic Hadith Perspective. *Journal for Integrative Islamic Studies* (2023).

[10] M. Bangari, D. Panwar, and D. Kumar. 2019. Determinants of Profitable Niche of Affiliate Marketing in Amazon. *Pacific Business Review International* (2019).

[11] N. Chachra et al. 2015. Affiliate Crookies: Characterizing Affiliate Marketing Abuse. In *IMC*.

[12] A. Chu, A. Arunasalam, M. Ozmen, and Z. Celik. 2022. Behind the Tube: Exploitative Monetization of Content on YouTube. In *USENIX Security*.

[13] D. DeAndrea and M. Vendemia. 2016. How Affiliation Disclosure and Control Over User-Generated Comments Affects Consumer Health Knowledge and Behavior. *J Med Internet Res* (2016).

[14] Y. Dwivedi, N. Rana, and M. Alryalat. 2017. Affiliate marketing: An overview and analysis of emerging literature. *The Marketing Review* (2017).

[15] Benjamin E. and Wesley B. 2015. Risk, Information, and Incentives in Online Affiliate Marketing. *Journal of Marketing Research* (2015).

[16] D. Ghosal, B. Prasad, and M. Behera. 2020. Impact of Affiliate marketing on e-buying behavior of Millennial–A TAM based Approach with Text Analysis. (2020).

[17] Golden Goose. 2023. https://gg.agency/.

[18] John H. 2022. How to Improve Your Affiliate Marketing with YouTube. https://tinyurl.com/4cjh5ccx.

[19] M. Hai, Z. Fang, and Z. Li. 2022. Does Business Group's Conscious of Social Responsibility Enhance its Investment Efficiency? *Sustainability* (2022).

[20] Salma M. 2023. YouTube Affiliate Marketing: 8 Tips. https://tinyurl.com/y4b53mtz.

[21] F. Mangio and G. Di Domenico. 2022. All that glitters is not real affiliation: How to handle affiliate marketing programs in the era of falsity. *Business Horizons* (2022).

[22] A. Mathur et al. 2018. Endorsements on Social Media: An Empirical Study of Affiliate Marketing Disclosures on YouTube and Pinterest. In *ACM HCI*.

[23] Tayyaba N. 2021. How to promote affiliate links with YouTube. https://shorturl.at/goPtr.

[24] Odigger. 2023. https://shorturl.at/T08Ak.

[25] OfferVault. 2023. https://www.offervault.com/.

[26] R. Olbrich, P. Bormann, and M. Hundt. 2019. Analyzing the Click Path Of Affiliate-Marketing Campaigns: Interacting Effects of Affiliates' Design Parameters With Merchants' Search-Engine Advertising. *Journal of Advertising Research* (2019).

[27] R. Olbrich et al. 2019. The effect of social media and advertising activities on affiliate marketing. *International Journal of Internet Marketing and Advertising* (2019).

[28] Z. Patrick and O. Hee. 2021. Affiliate Marketing in SMEs: The Moderating Effect of Developmental Culture.. In *Journal of Social Sciences & Humanities*.

[29] Niall R. 2023. Youtube Affiliate Marketing in 2023. https://www.authorityhacker.com/youtube-affiliate/.

[30] L. Rolim, J. Simões, and D. Figueiredo. 2020. Network and revenue of the clube hurb affiliate marketing program: a story of two tales. In *IEEE/ACM ASONAM*.

[31] Peter Snyder and Chris Kanich. 2016. Characterizing fraud and its ramifications in affiliate marketing networks. *Journal of Cybersecurity* (2016).

[32] P. Snyder et al. 2015. No Please, After You: Detecting Fraud in Affiliate Marketing Networks. In *WEIS*.

[33] S. Suryanarayana, D. Sarne, and S. Kraus. 2021. Information design in affiliate marketing. *Autonomous Agents and Multi-Agent Systems* (2021).

[34] S. Suryanarayana et al. 2019. Information Disclosure and Partner Management in Affiliate Marketing. In *DAI*.

[35] M. Swart et al. 2020. Is This An Ad?: Automatically Disclosing Online Endorsements On YouTube With AdIntuition. In *ACM CHI*.

[36] H. Syrdal et al. 2023. Influencer marketing and the growth of affiliates: The effects of language features on engagement behavior. In *Journal of Business Research*.

[37] J. Zhang, S. Lee, and Y. Tan. 2023. The Effects of the FTC Policy and Affiliation Disclosures on Review Video Engagement: Evidence from YouTube. (2023).