No Please, After You: Detecting Fraud in Affiliate Marketing Networks

Peter Snyder and Chris Kanich
University of Illinois at Chicago
Chicago, Illinois, USA
{psnyde2,ckanich}@uic.edu

ABSTRACT

Cookie stuffing is an activity which allows unscrupulous actors online to defraud affiliate marketing programs by causing themselves to receive credit for purchases made by web users, even if the affiliate marketer did not actively perform any marketing for the affiliate program. Using two months of HTTP request logs from a large public university, we present an empirical study of fraud in affiliate marketing programs. First, we develop an efficient, decision-tree based technique for detecting cookie-stuffing in HTTP request logs. Our technique replicates domain-informed human labeling of the same data with 93.3% accuracy. Second, we find that over one-third of publishers in affiliate marketing programs use fraudulent cookie-stuffing techniques in an attempt to claim credit from online retailers for illicit referrals. However, most realized conversions are credited to honest publishers. Finally, we present a stake holder analysis of affiliate marketing fraud and find that the costs and rewards of affiliate marketing program are spread across all parties involved in affiliate marketing programs.

Categories and Subject Descriptors

J.m [**Computer Applications**]: Miscellaneous
; K.4.4 [**Computers and Society**]: Electronic Commerce

Keywords

web security; cybercrime; economics of cybercrime

General Terms

affiliate marketing fraud, cookie-stuffing

1. INTRODUCTION

Despite the size of the online display advertising market, there are still alternative revenue opportunities for free-to-access web services. Along with models like freemium and crowdfunding, affiliate advertising is a prevalent method of creating revenue for a free website. As with any online revenue generation scheme, there are opportunities for bad actors on the Internet to defraud affiliate networks for what is potentially a substantial sum. What we don't yet understand, however, is how prevalent, successful, or damaging this fraud is. Understanding its effect on the market can inform technical solutions to the problem as well as provide motivation for how many resources should be committed to finding solutions.

Online affiliate marketing is a commercial system in which an online retailer attempts to increase traffic to their site—and hopefully their sales—by compensating third parties to promote the retailer's goods and services. Many large online businesses run affiliate marketing programs, with some of the largest run by Amazon.com[9], GoDaddy[4], eBay[7] and WalMart[22].

As with any type of commerce, dishonest parties try to subvert the initial intent of the market for person gain. **Affiliate marketing fraud** occurs when a dishonest party hacks a website, leaves a spam comment, or simply adds some code to an unrelated page which causes visitors to also visit the fraudster's affiliate link. Online retailers then pay the most recent affiliate for generating any successive sales, even if the user was completely unaware of loading the fraudster's affiliate link. This fraud has the potential not only to provide substantial revenue to the fraudster, but can also cause the affiliate program to pay a commission when no legitimate advertising was happening.

Most damaging, however would be the effect on the revenue model itself: because only the most recent affiliate gets credited for each sale, sufficiently successful attackers could reduce the revenue for legitimate affiliate advertisers so much that the entire business model no longer works, putting both those free sites out of business and further limiting the ways in which free content can be subsidized online. Understanding the technical methods that these attackers use, as well as how damaging they actually are to the business model, is key to understanding the full effects of affiliate marketing fraud.

In the course of investigating this phenomenon, this paper makes three contributions. First, we describe an automated technique for detecting affiliate marketing fraud based on analyzing HTTP request headers with approximately 93.3% accuracy. Second, we provide measurements of how frequently affiliate marketing fraud occurs relative to valid affiliate marketing activities. And third, we provide an analysis of the costs and benefits of affiliate marketing fraud, and find that the benefits and costs of affiliate marketing fraud are spread among all parties involved in affiliate marketing programs.

2. RELATED WORK

Our work exists alongside a wealth of research into the role of cybercrime and fraud in online economic activity. To name several examples, Dave et al. [6] developed a successful method for detecting click-fraud in advertising networks by looking for publishers who deviate from an expected baseline profit-per-user. Clayton and Mansfield[3] investigated the

frequency of, and motivations behind, parties providing false information to the WHOIS system when registering domains, and found that the practice was common among domains affiliated with illegal activities. Chachra et al. [2] empirically measured how, and how often, malicious parties abused domains and web services to circumvent domain-based blacklisting systems. They found that malicious parties regularly do so by piggy-backing their services on-top of popular and trusted domains, which are unfeasible to blacklist. Kanich et. al[10] measured the effectiveness of, and returns to, selling illegal pharmaceuticals promoted by through botnet produced SPAM campaigns. Moore and Edelman[13] found that affiliate marketing fraud was a common method of monetizing typosquatted domains.

A primary aim of this work is to provide a stakeholder analysis of affiliate marketing fraud. Our work builds on similar analyses that have been conducted for other types of crime and fraud. Peacock and Friedman[16] studied how the costs of payment card theft are distributed between cardholders, the banks representing card holders, banks representing merchants, the merchant, and the card network in general. They found that cardholders bare a minimum amount of risk, while merchants are most negatively affected. Kshetri[12] applied a similar analysis to click-fraud and teased apart how the behavior affected the party committing the fraud, the advertiser, and the pay-per-click network operator. Khan et al. [11] also carried out a stakeholder analysis of typosquatting, and found that despite the presence of fraud, the practice was often beneficial for all involved parties: web users, valid domain holders and typo-squatters.

Finally, a fundamental part of our methodology is inferring user intent based on HTTP header data. Schneider et. al[18] enumerated many of the difficulties of performing any analysis based on HTTP traces, including dealing with HTTP pipelining, incorrect advertised content types, and wrongly provided content length fields. Xie et. al[23] provide a method for inferring and rebuilding user browsing sessions based on collections of HTTP requests. Most relevant to our work is their focus on the *HTTP referer* field and timestamp information. Neasbitt et al. [15] built on this research in their work on reconstructing browser sessions from HTTP request logs. Their work uses an instrumented browser and replayed parts of the HTTP logs to deal with ambiguities in the methods used by Xie and our work, and in effect represents a trade off between accuracy and performance.

3. BACKGROUND

Modern website monetization programs have become increasingly complex. Here we provide an overview of the relevant actors in the space, as well as define terms which we will use for the remainder of the paper.

3.1 Affiliate Marketing Programs

Affiliate marketing programs are programs set up by online retailers to increase their sales. While these are run by both on-and-offline businesses, this paper only considers online programs. Affiliate marketing programs consist of three main parties, **online retailers** that set up the programs to increase the sales of their products, **publishers** which promote the online retailers' sites and products in exchange for compensation from the retailer, and **web users**, which are the target of the affiliate marketing programs. Online retailers hope that when web users visit websites run by publishers, the publishers will convince some web users to visit the online retailer's web site and make purchases.

Online retailers, like traditional businesses, must promote themselves and their products to generate sales, stay in business, and make a profit. However, for online retailers the need for marketing is arguably intensified by the Internet's low barriers to entry and relative lack of geographic restrictions, both of which result in greater competition and greater difficultly retaining consumers.

One technique online retailers use is affiliate marketing, where the retailer outsources some of the marketing work to third-parties on the Internet. These sites—publishers—then take on some of the responsibility for promoting the goods and products of the online retailer. In exchange, the online retailer agrees to share a portion of the profits resulting from the publisher's promotion with the publisher. Amazon, GoDaddy, eBay and Walmart run the largest affiliate marketing programs on the web, with Amazon's being by far the most popular.

Because retailers generally only pay publishers when sales occur, affiliate marketing programs are generally low risk endeavors for online retailers. The challenge for an online retailer in setting up an affiliate marketing program is instead in two other areas, 1) ensuring that the correct third party is credited for sales that publisher generated, and 2) only paying publishers for sales that they are responsible for (i.e. to not share sales that would have happened even without the publisher's promotion). In the former case, the goal is to maintain fair and ongoing relationships with promoting publishers; in the second, profit maximization by minimizing the payouts of the affiliate marketing program.

Publishers are the third-party websites that participate in affiliate marketing programs by directing visitors—and potential customers—to the online retailer's website. Honest publishers do this in many ways, such as including links on their websites or news letters, promoting and linking to products on the online retailer's website, or the online retailer's website itself. If any web users visit these links, traveling from a web property the publisher controls to the online retailer's site, and then make a purchase on the online retailer's site, the online retailer pays the publisher a portion of the sale in compensation.

Finally, **web users** are anyone on the Internet using a commodity web browser to interact with the web. These interactions might include visiting web pages controlled by publishers or making purchases from online retailers. As a group they are the indirect target of online retailers' marketing efforts; online retailers hope that affiliate marketing programs will result in more web users making purchases from the online retailers.

3.2 General Implementations

Though each affiliate marketing program studied in this paper differs in the details, each is implemented in a similar manner. Online retailers build and maintain the infrastructure for the affiliate marketing program, including means to track which publishers delivered which web users, web portals for publishers to monitor their credit, and so on. Parties wishing to monetize their web sites register with the online retailers through these web portals to become publishers. The online retailer then provides each publisher with an **affiliate identifier**, which the publisher uses to identify herself to the online retailer going forward. Whenever the

publisher creates links or directs traffic to the online retailer, she does so by combining her affiliate identifier, the online retailer's domain, and a path provided by the online retailer to create a **affiliate link**. Any web users who come to the online retailer's site using a affiliate link will have an HTTP cookie set in their browser, identifying the web user as having come from the publisher. The web user is then able to use the online retailer's site and make purchases as normal.

When the web user makes a purchase on the online retailer's site, the online retailer checks to see if the web user has an HTTP cookie identifying them as having come from a publisher. If so, the publisher receives a portion of the sale price as compensation for "delivering" the shopper to the online retailer.

3.3 Illustrative Example

To better understand how this is carried out in practice, consider the following example. *Asher* runs a website where he reviews movies. Asher decides he would like to make money from his site, so he becomes an publisher in an affiliate marketing program by Amazon, an online retailer. Amazon provides Asher with an affiliate identifier, such as "Asher123", which he uses to identify his visitors in the program.

In Amazon's affiliate marketing program, Asher is only paid when a web user visiting Asher's site clicks on a link, travels to Amazon, and makes a purchase on Amazon's site. To encourage his visitors to do this, Asher edits his site to add a link to Amazon next to each movie review. This link might read "purchase this movie on Amazon." Clicking the link would then take the web user to an affiliate link on Amazon's site, such as `http://amazon.com/example-movie?publisher=Asher123`, indicating to Amazon that the user came from Asher's site.

Mark is a web user visiting Asher's site. He sees one of the links Asher created and clicks on it. Mark is then taken to Amazon, but at a URL containing Asher's affiliate identifier, identifying to Amazon that Mark was directed by Asher. Amazon receives this request and responds by recording that Mark's browser visited Amazon due to a link from Asher's site.[1]

Mark might then browse around Amazon, or even leave his computer and continue shopping a few hours later. Mark eventually adds several items to his shopping cart and makes his purchase. When Amazon is processing the purchase, Amazon notices that Mark had recently clicked through to Amazon via a link from Asher' site. Amazon would then credit Asher a portion of Mark's purchases, using a formula based on the types and number of products being purchased[1].

3.4 Affiliate Marketing Fraud

The above example describes how an online retailer intends an affiliate marketing program to work. In practice, malicious users try to defraud the system, attempting to receive credit for sales they did not generate. This process is called **affiliate marketing fraud**, and it occurs when a fraudulent publisher tricks a web user's browser into visiting a page on the online retailer's site that the web user did not intend to visit, or in some cases, even realize their browser visited. These links cause the online retailer to record that the publisher generated the "sales lead," and gives the **fraudulent publisher** credit for any purchases the web user might make.

There are many ways this is done in practice. One common way is for a fraudulent publisher to control a website, or have broken into a website they do not control. The fraudulent publisher could then add a hidden *iframe* element to the site, and set this *iframe* to an affiliate link on the online retailer's domain. The fraudulent party would set the affiliate link to include their affiliate identifier, causing the online retailer to treat this request as though the fraudulent publisher encouraged the web user to visit the online retailer.

Thus, when the web user's browser renders the page that the fraudulent publisher has manipulated, the browser will also render the embedded *iframe*, resulting in a new request from the web user's browser to the online retailer's site. When the online retailer processes this web request, they will again set a cookie on the web user's browser, now giving the fraudulent publisher credit for delivering the web user and possibly overwriting any record that the web user was previously referred by an honest publisher.

The online retailer is generally unable to determine if the web user intended to visit the online retailer, and thus gives the fraudulent publisher credit for any purchases made by the web user. This behavior is "fraudulent" because it does not reflect the intent of the web user, and likely does not result in additional sales for the online retailer, yet results in the fraudulent publisher being paid.

Other common ways malicious parties carry out affiliate marketing fraud includes Flash objects that spawn new pages in the background, javascript redirects (either directly included or through XSS), or through malware installed on a web user's machine that opens pages on the online retailer's site carrying the fraudulent publisher's affiliate identifier. Note that regardless of the delivery mechanism, the fraudster must cause the web user's browser to visit the fraudster's affiliate link, otherwise the fraudster will not be credited for the sale.

4. DATA

To better understand affiliate marketing fraud, we looked at many affiliate marketing programs run by popular online retailers. We then looked at the incidence of affiliate marketing activity in real-world HTTP request headers. This section describes the size, structure and source of the data used in this analysis, followed by a detailed explanation of

[1] When a web browser makes a request to a URL, the web server listening to that URL respond with a *Set-Cookie* header, instructing the web browser to return a given key-value pair on all subsequent requests to the same domain. This allows web servers to keep track of users between different web requests. In the case of affiliate marketing programs, cookies are also used to keep track of which publishers delivered with web users.

Cookie stuffing is called as such because this functionality is typically implemented by setting a special cookie in the user's browser: the fraudster is then thought to be "stuffing" it into the browser, unbeknownst to the user and possibly displacing an honest affiliate.

Amazon's implementation does not change anything about the user's session cookie (a unique identifier generated for all visitors to the site), but rather stores this session to affiliate mapping server side.

Online Retailer	Domains	Cookie Setting URL	Conversion URL
Amazon	`(www\.)amazon\.com`	`^/(?:.*(dp|gp)/.*)?[&?]tag=`	`*handle-buy-box*`
GoDaddy	`^godaddy\.*`	`(?:&|\?|^|;)isc=`	`*domains/domain-configuration\.aspx*`
			`*hosting/web-hosting-config-new\.aspx*`
			`*ssl/ssl-certificates-config\.aspx*`
			`*hosting/vps-hosting-config\.aspx*`
imlive.com	`^imlive\.com$`	`(?:&|\?|^|;)wid=`	`*savebillclickout\.ashx*`
wildmatch.com	`^wildmatch\.com$`	`(?:&|\?|^|;)wid=`	`*/join/*`
eroticasians.com	`\.eroticasians\.com$`	`\?t|T=`	`*signup.html*`

Online Retailer	Affiliate Identifier	Session ID (for cookie data)
Amazon	`tag=(.*?)(?:&|$)`	`session-token=([^\;]+)`
GoDaddy	`cvosrc=(.*?)(?:&|$)`	`visitor=([^\;]+)`
imlive.com	`wid|WID=(.*?)(?:&|$)`	`spvdr=([^\;]+)`
wildmatch.com	`wid|WID=(.*?)(?:&|$)`	`vi=([^\;]+)`
eroticasians.com	`t=(.*?)(?:&|$)`	`ntc=([^\;]+)`

Table 1: Extracted regular expressions for the five most frequently observed programs (in PCRE format)

how the raw log data was processed into data structures used in the analysis described in the next section.

4.1 Affiliate Marketing Programs

Six affiliate marketing networks were analyzed in this research. These included the largest affiliate marketing programs identified in our dataset, run by Amazon and GoDaddy, as well as four affiliate marketing networks covering 164 individual affiliate marketing programs: The ClickCash network, consisting of 6 sites at the time the data was collected, the MoreNiche network, consisting of 9 sites, the PussyCache network, consisting of 8 sites, and the Sextronics network, encompassing 141 sites.

We selected these programs because they are large parties in the affiliate marketing economy, particularly Amazon and GoDaddy. A large volume of traffic to these sites was carried over unsecured HTTP connections during January and February of 2014, which allowed us to follow the related affiliate marketing activity in our dataset. Additionally, these parties use regular affiliate identifiers and predictable patterns in their affiliate links, which make it possible to extract the affiliate marketing activity from the structure of the URLs requested, without needing any information held in secret at the online retailers.

Only a subset of online affiliate marketing programs meet the above conditions. Other popular affiliate marketing programs, including Commission Junction and Google Affiliate Network, were examined but were not able to be included in this work. This was for a variety of reasons, including using non-predictable URLs in affiliate links and conversion URLs (which made detecting referrals and purchases infeasable) and carrying the majority of their traffic over HTTPS (which resulted in traffic in these networks from being omitted from our dataset).

As such, this work measures only a subset of the affiliate marketing economy. However, the substantial size of the affiliate marketing programs selected make us confident that we are capturing a large portion of affiliate marketing activity in terms of economic activity, if not in terms of parties involved. Thus, while we cannot estimate the absolute amount of affiliate marketing activity in our dataset, the findings in this paper constitute a lower bound.

Our goal was to measure, for each affiliate marketing program analyzed, how frequently web users visited affiliate marketing links, received affiliate marketing tracking cookies, and made purchases from online retailers. We also wanted to understand the ratio of honest versus fraudulent activity occurring in each network.

To do so, we registered as an publisher in each affiliate marketing program and created regular expressions to match both the affiliate links and the contained affiliate identifiers in each system. We also extracted regular expressions to detect URLs indicating a web user was making a purchase from an online retailer. We refer to these as **conversion URLs**. These patterns are included in table 1.

For online retailers that served most of their site's content over HTTP, but which directed users to HTTPS connections for the checkout process (most significantly in our dataset, Amazon and GoDaddy), we used the URL for the "add to cart" or equivalent pages as proxies for conversion URLs. As is explained in further detail below, this was done because our data set does not include any requests made over HTTPS. Conversion URLs that occurred close together in time (within an hour of each other) were coalesced into a single event. We refer to these events—visting a conversion URL for online retailers that do not force HTTPS, or visiting an "add to cart" URL for online retailers to do—as **conversion events**. While this method will result in over counting the number of purchases made from online retailers because not all conversion events result in purchases, we believe that this is the best possible approximation given the data available.

Because our dataset does not include HTTPS requests, we may also see affiliate link visits, but for users who are logged in to some online retailers (most notably Amazon), we may see no subsequent page visits within that browsing session because the entirety of the browsing, selecting, and checking out process happens within an HTTPS session. While this likely causes us to under-count conversions, we expect that loosing this subset of requests will have no effect on the relative proportions of visits to legitimate versus illegitimate affiliate links, and thus while the absolute values we present will be lower bounds, the relative values (e.g. proportion of fraudulent publishers) will not be affected.

4.2 HTTP Logs

4.2.1 Raw Data

We conducted this research on HTTP request logs taken from a large East coast university in the United States. Each

Date	Count	Size
January, 2014	895,722,435	240G
February, 2014	1,440,677,533	420G
Total	2,336,399,968	660G

Table 2: Counts of HTTP request records

record in the HTTP logs included the following fields.

- requester IP address
- `User Agent` header
- the IP of the host the request is being sent to
- the domain and the path of the resource being requested
- timestamp
- returned HTTP status code
- the HTTP referrer (if available)

We processed this raw data as follows. First, we reduced the data set by removing all logs documenting requests for any-non HTML or text asset, based on the returned MIME type, with the notable exception of HTTP redirect responses, which were also retained. The remaining logs were processed into trees, each representing part of a browsing session by a user on the network. We refer to these trees going forward as **browsing-session trees**.

4.2.2 Browsing Sessions as Trees

Conceptually, each time a web user opens her web browser, the set of visited pages can be thought of as forming a tree, with any URL typed into the browser's URL field forming the root of a tree. These "typed" URLs are roots of each tree because they are "initial" requests, not the result of visiting any "parent" page. All links clicked on from any of these root pages take users to a child page, adding a child-node to the browsing-session tree. These child-requests in turn lead to their own child-requests and so on, resulting in a tree of arbitrary depth and complexity. Because users can click the "back" button and click on a different link, or open different links from the same page in a new tab or window, each node can have multiple children. In this conception, starting a new browsing session equates to constructing a new parallel tree of requests, with the first page visited in the new window forming the root of the new tree.

Parsing the original log records into browsing-session trees eased analyzing a user's behavior in several ways. First, having a tree structure made it trivial to find the series of requests that brought a web user to a given page. Instead of needing to search through a massive flat file, answering this question with a browsing-session tree only required following the path from a request (a node) in the tree to the root of the tree. Similarly, understanding how long it took a user to move between pages also becomes trivial. It could be calculated by taking the difference between the timestamp of a request node and that of its parent node.

4.2.3 Building Browsing-Session Trees

The flat log records were converted into browsing-session trees in the following manner. First, HTTP request records were grouped by `(requester IP, user agent)` pairs. Each resulting group of records roughly corresponded to an individual web user. We found IP aliasing to not be an issue in the network in question. Out of 19,985 observed pairs of IP addresses and user agent strings, we never observed multiple user agents using the same IP address at the same time.

Second, each group of records was ordered chronologically.

Third, the following algorithm was followed for each group of records. First, allocate an empty set for each web user. This set will be used to store trees describing this web user's browser sessions. Second, consider each request, earliest to latest. If a record has an HTTP referrer, check to see if it matches any of the URLs of any record in any of the browsing-session trees in the user's tree-set (within a time window which we set to 5 minutes). If yes, add the current record to the tree as a child of the matching record. Otherwise, if the record does not have an HTTP referrer, or if the record's HTTP referrer could not be matched to an existing node, add a new browsing-session tree to the user's tree-set, and set the record as the root of the new tree. Once this process has been carried out for each record, all records will be assigned to a tree describing a browsing session for this user.

Finally, to further reduce the working set, each tree was examined to see if it contains any records associated with either an affiliate link or a conversion URL for any affiliate marketing program under consideration, using the regular expressions extracted and discussed previously. If the tree does not include any such requests, it is removed from further consideration, since these trees trivially did not include any relevant affiliate marketing activity.

4.2.4 Data Limitations

This preprocessing made a very large unsorted collection of HTTP request logs into trees representing individual browsing sessions tied together by HTTP referrer header values. While effective for our purposes, this process was imperfect: web browsers omit the HTTP referrer header in some cases (such as when the user is visiting or leaving a site requested over HTTPS, or when a web page's "content security policy"[21] is set as such), which can cause our approach to split a logical browsing session over multiple trees. However, based on manual inspection and findings from prior work on recreating browser session from HTTP requests[23] we do not believe that this limitation meaningfully impacted the accuracy of our analysis or affiliate marketing fraud detection techniques.

5. MEASUREMENTS

5.1 Fraud Detection

Measuring fraudulent activity in affiliate marketing programs requires inferring the intent of the web user based on the time and content of their HTTP requests. We developed a decision tree classifier that labels HTTP requests to affiliate links as either "fraudulent" or "honest", reflecting the intent—or lack thereof—of the web user behind those requests. Our classifier replicated trained human labeling of the same data with 93.3% accuracy, a false positive rate (i.e. an incorrect label of "fradulent") of 1.5%, and a false negative rate (i.e. an incorrect labeling of "valid") of 5.2%. The classifier requires minimal network activity (a maximum of one request per domain), and otherwise requires no data beyond what is in the HTTP request logs. We detail in the following sections how we generated and evaluated this classifier, followed by how we applied the classifier to our data to measure fraudulent affiliate marketing activity.

5.1.1 Training Data

We built our classifier based on a subset of our data (a subset of the January 2014 data), and ran our classifier on the remaining data (the remaining January 2014 records as well as data from February of the same year). 1141 browsing-session trees from the January data that contained affiliate links were manually inspected to determine if the request looked to be the result of user intent, or if the request appeared fraudulently generated.

We hand labeled each affiliate link request by examining the surrounding requests in each browsing-session tree and manually visiting each page that referred a web user to an affiliate link (the parent of the affiliate link node in the browsing-session tree). We label a request as fraudulent if any of the following apply:

- without any interaction a new window to an affiliate link was visited
- the current window was redirected to the same
- an affiliate link was requested in an *iframe*
- a flash element made a cross domain request to an affiliate link
- a request was made in any other way (such as an image or script reference) to an affiliate link without interaction

In many cases we were not able to access the content of the referring page, or there was no longer a reference to the affiliate link on the referring page. Given the age of the data at time of examination (the hand-labeling was done in October and November of 2014, the examined requests were made in January 2014) this could have occurred for a wide variety of reasons, such as site redesigns, domain expirations, or changes to content.

In cases where we were not able to find a reference to the affiliate link, and thus could not recreate the steps that caused the web user to arrive at the affiliate link from the referring page, we had to make a best effort estimate based on characteristics of the browsing-session tree. We based our decisions in these cases on whether there were request patterns in the browsing-session tree that appeared "suspicious", or out of what standard, user-initiated browsing patterns look like. Examples of such suspicious patterns include web users being redirected to the affiliate link's site almost instantly (faster than would have been possible for them to read, and maybe even fully render, the page), or referrals coming from domains with nonsensical, machine generated domains which themselves had no referrer.

To minimize the chances of false positives (e.g. false "fraudulent" labels), we rounded all decisions strongly towards "honest," thus maintaining a high burden of suspiciousness for any "fraudulent" labelings.

5.1.2 Fraud Classifier

Once we hand labeled the January data set, we built a classifier that would accurately approximate the human labels for the rest of our data, the remaining January records and the larger set of requests made in February 2014. We did so by generating a simple decision tree algorithm that was able to match the human generated labels 93.3% of the time. Our decision tree consists of three boolean questions.

1. Referrer time
We measured the amount of time that occurred between the affiliate link being visited and the parent request in the browsing-session tree. This measure represents the amount of time that the web user spent on the referring page before clicking on a link, or otherwise visiting the online retailer's page. Note that if an affiliate link did not have a detected referrer, and thus no parent in its containing browsing-session tree, it was removed from consideration and did not receive either an "honest" or "fraudulent" label. Our decision tree treats values of less than two seconds as suggesting fraud.

2. Time spent on online retailer's site
This feature captured the amount of time the web user continued browsing the online retailer's site after requesting the affiliate link. Low values here indicate that the user quickly closed the window or tab depicting the online retailer's site, or never noticed it in the first place. This measure was taken by calculating the maximum time that occurred between the affiliate link and any leaf nodes below it in the browsing-session tree. Our decision tree treats values of less than two seconds as suggesting fraud.

3. Does the referring domain offer HTTPS?
A HTTPS request was made to port 443 for each domain that referred a web user to an affiliate link. If the server responded with a valid HTTPS connection, and the certificate offered in that request had a PKI root in Mozilla's set of trusted root certificates [14], the referring domain was treated as offering HTTPS. If a referring domain gave any valid response to an HTTPS request, our decision tree treated it as indicating no fraud.

If the answer to the first two questions was less than two seconds, and if the answer to the third decision was "no", then the affiliate link request was treated as "fraudulent." In all other cases the request was treated as "honest." Put differently, if 1) a web user spent less than two seconds on the referring page before visiting the affiliate link, and 2) spent less than two seconds on the online retailer's site after visiting the affiliate link, and 3) the referring domain did not offer HTTPS, we classified the referral as "fraudulent." All other referrals were labeled "honest."

The above classifier was evaluated using standard 3-fold cross-validation, with each of the hand-labeled records being assigned randomly to one of three groups. This decision tree classifier reproduced the hand labeled values 93.3% of the time.

5.2 Measuring Affiliate Marketing Activity

Using our dataset and the above fraud detection mechanism, we were able to make several measurements which are helpful for understanding affiliate marketing programs, and the role that fraud plays in them.

5.2.1 Retailer Popularity

First, we measured how frequently each retailer appeared in our data set, and how many sessions web users created with them. We derived the first count by summing the number of requests to each online retailer in the network trace that returned an HTML document. This number is presented in the "Requests" column in table 3.

Retailer	Requests	Unique Sessions
Amazon	2,663,574	87,654
GoDaddy	7,320	364
imlive.com	731	194
wildmatch.com	3	1
eroticasians.com	3	1
Total for 166 programs	2,671,808	88,257

Table 3: Measures of how frequently each affiliate marketing programs is requested by a web user

Similarly, we captured the number of browsing sessions initiated with each online retailer by counting the number of browsing-session trees. These counts are included in the "Unique Sessions" column of table 3.

One limitation to note is that while this number is linearly related to the number of visitors to these services, it will be significantly larger than the actual number of individuals who use the service, as repeat visitors who do so in different sessions will be counted as different events.

5.2.2 Publishers

Retailer	Honest	Fraudulent	Total
Amazon	2,268	1,396	3,664
GoDaddy	5	19	24
imlive.com	4	7	11
wildmatch.com	0	1	1
eroticasians.com	1	0	1
Total for 166 programs	2,281	1,426	3,707

Table 4: Numbers of publishers in most popular affiliate marketing programs

We also measured the number of publishers appearing in our dataset, grouped by affiliate marketing program. Since each publisher is given their own affiliate identifier, and each affiliate identifier must appear in an affiliate link for the publisher to be participating in the program, we detected the number of publishers in our data set by counting the number of unique affiliate identifiers appearing in affiliate links for each program. This number is included in the "Total" column of table 4.

We then separated this measure into "honest" and "fraudulent" publishers. Fraudulent publishers are publishers that direct web users to affiliate links without the web user's intent. For our measures, "fraudulent" behavior is toxic; if a publisher is associated with both honest and fraudulent referrals, it is counted as a "fraudulent" publisher in these measures. Honest publishers are simply those that are not associated with any "fraudulent" referrals in our dataset.

A caveat to our approach is that it might over count the number of actual participants in these systems. Since individuals and companies can create multiple publisher accounts, any party can acquire multiple affiliate identifiers. While intuitively we can assume that "fraudulent" publishers are more likely than honest parties to create multiple accounts—whether to avoid detection, respond to account closures, or otherwise—the option is nevertheless present for both types of publishers. The presented numbers should therefore be treated as an upper bound on the number of observed publishers.

5.2.3 Affiliate Marketer Referrals

Retailer	Honest	Fraudulent	Total
Amazon	12,870	2,782	15,652
GoDaddy	399	98	497
imlive.com	9	13	22
wildmatch.com	0	1	1
eroticasians.com	2	0	2
Total for 166 programs	13,283	2,897	16,180

Table 5: Counts of referrals in most popular affiliate marketing programs

We also measured the number of times affiliate links were visited. This measure is an important part of understanding how significant an affiliate marketing program is to an online retailer. This measure was found by simply summing the number of affiliate links found in the data set for each affiliate marketing program. This number is included in the "Total" row of table 5.

We then distinguished fraudulent from honest referrals by using the techniques described in section 5.1. By combining the hand-generated January validity labels with the machine classified data, we get a complete labeling of all affiliate link requests as being generated by either valid user intent ("honest") or manipulated browser activity ("fraudulent").

5.2.4 Conversion Events

Finally, we determined the number of conversion events made with the tracked online retailers, and distinguished which conversions were a result of honest or fraudulent affiliate marketing. We also measured how often credit for affiliate-marketing-generated sales was stolen from an honest publisher by a fraudulent one.

First, we approximated the number of times web users were converted by an online retailer. For online retailers who have HTTPS protected checkout pages, we instead counted the number of sessions during which a web user added an item to their cart. This count, included in the "Conversion Events" column of table 6, was found by counting the number of conversion events (as defined in section 4.1) for each online retailer.

Next, we measured the number of conversion events credited to each affiliate marketing program. Conceptually this is the count of the number of conversions that were made by web user carrying affiliate marketing cookies. In practice though, this was made difficult by the fact that some online retailers (most significantly Amazon) track which publisher referred which web user server side, instead of directly in cookie values. Most examined affiliate marketing programs assign a cookie stating which publisher the web user came from (ex *affiliate=Example123*). Amazon differs by treating the referring publisher as part of the web user's larger session data. As a result, determining whether a web user making a purchase on Amazon came from an publisher required looking at Amazon requests in the web user's browsing session.

To account for these online retailers, we examined the web user's activities before each conversion event. Due to anecdotal evidence within the affiliate marketing community,

Retailer	Conversion Events	Affiliate Conversions	Honest	Fraudulent	"Stolen"
Amazon	15,624	955	781	174	0
GoDaddy	26	8	8	0	0
imlive.com	0	0	0	0	0
wildmatch.com	0	0	0	0	0
eroticasians.com	0	0	0	0	0
Total for 166 programs	15,650	963	789	174	0

Table 6: Counts of conversion events in the most commonly observed affiliate marketing programs

we set the "timeout" after which an affiliate does not receive credit for a sale at twenty four hours. Thus, if the web user visited an affiliate link within twenty four hours of visiting the conversion event, we credited the affiliate marketing associated with the most recent affiliate link for the purchase. Otherwise, if the web user did not visit any affiliate links in the hour before the conversion event, we treated the conversion as not being part of the affiliate marketing program. The count of how many conversion events were credited to an affiliate marketing is included in the "Affiliate Conversions" column of table 6.

We determined which conversions were credited to a fraudulent publisher by first extracting all conversion events from the data set. Then we considered the preceding time-window during which the online retailer would give the publisher credit for the conversion. If the web user visited no affiliate links in this time frame, the conversion was treated as if it was made without an affiliate marketing cookie. If the most recent affiliate link was labeled as "honest", then the conversion event was attributed to an honest publisher, and included in the "Honest" column of table 6. Conversely, if the most recent affiliate link was labeled as "fraudulent", it was included in the "Fraudulent" column of the same table.

Finally, we calculated the number of conversion events that were credited to a fraudulent publisher which would have otherwise been credited to an honest publisher. We call these events **affiliate marketing thefts**, since they represent instances where a fraudulent publisher party "stole" credit for a conversion from an honest publisher. We detected these events by taking the same twenty four hour window described in the previous paragraph and looking for instances where a conversion event was credited to a fraudulent publisher, but where the window also contained honest affiliate links. In the two months of data we examined, we did not find any instances of affiliate marketing theft, as is documented in the "Stolen" column of table 6.

6. ANALYSIS

Here we use the data and measurements described in the previous sections to better understand the affiliate marketing ecosystem and the different actors in it.

6.1 Market Analysis

Our measurements of affiliate marketing programs reveal several interesting aspects of these markets. First is that there are a large number of fraudulent publishers in these systems. In the largest affiliate marketing program observed, Amazon's, over one third of the observed publishers engage in "cookie-stuffing". In the case of the second largest observed affiliate marketing program, GoDaddy's, the case is even more dramatic, with the majority of observed publishers carrying out fraudulent activity. From the data we have, we conclude that a significant portion of parties participating in affiliate marketing programs are engaging in deceptive activities.

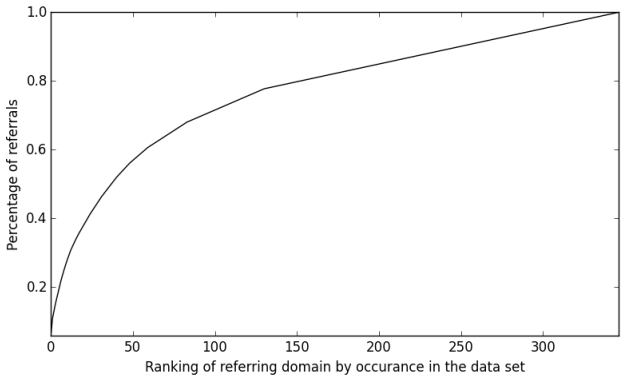

Figure 1: Number of referrals credited to each domain against each domain's frequency in dataset

However, while the number of fraudulent publishers is large, their impact is relatively small. Given their numbers, they are under represented in conversion events completed by-online retailers. Even though 38.1% of Amazon's publishers were engaged in fraud, only 18.2% of conversion events were associated with a fraudulent affiliate identifier. In the case of GoDaddy the numbers are even more lopsided, with 79.1% of observed publishers participating in "cookie-stuffing", but resulting in no purchases.

As Figure 1 shows, a relatively small number of domains contribute the majority of affiliate marketing referrals, and the majority of domains appear only once in the data. Relatively well known websites, such as *gizmodo.com*, *slickdeals.com*, *thewirecutter.com* and *dealmoon.com* refer a large number of visitors to online retailers. As these are all specifically deal focused websites which explicitly link to products throughout the site, it's very unlikely that they would have any reason to commit cookie stuffing.

Finally, the market for affiliate marketing programs is dominated by a very small number of parties. Though we examined 166 affiliate marketing programs, the largest online retailer, Amazon, captured over 98% of affiliate marketing-produced conversion events, and the top two observed online retailers combined captured every conversion event observed.

6.2 Stakeholder Analysis

The costs and benefits of affiliate marketing fraud are not distributed equally across all participants in the affiliate mar-

keting system. This section attempts to provide a breakdown of how fraud affects each relevant party.

6.2.1 Web Users

Although web users are the immediate target of fraudulent publishers, they do not bear the main costs of that fraud. Fraudulent publishers target web users—through the use of malware, hidden *iframes*, automated redirects, and other forms of browser manipulation—but beyond the possible minor inconvenience of popup windows or additional network use, web users are largely unaffected. Web users are still able to make their desired purchases from online retailers, and the price they pay is unchanged in the presence of fraud (we did not observe any affiliate marketing program where the purchase price was affected by whether the purchaser carried an affiliate marketing cookie). Additionally, the tools fraudulent publishers use to redirect web users to affiliate links are not inherently harmful, even if they may be annoying.

A possible, additional cost is the that fraud in an affiliate marketing program reaches a threshold where it is no longer profitable for the organizing online retailer, resulting in the closing of the affiliate marketing program. Were this to happen, it could pose a substantial cost to web users. One common role of affiliate marketing is to subsidize free content for web users; if publishers suddenly lost the income provided by affiliate marketing, a large amount of now-free content might move behind paywalls, or might not be produced at all.

There have been cases of online retailers closing down affiliate marketing programs, due at least in part to fraud[17, 8]. It is not possible for us to determine the degree to which fraud led to the end of these programs, but we expect fraud to have played some role in each decision. Given the limitations of our data, and the large number of affiliate marketing programs currently being operated, we can only guess that fraud has played a small-but-non-zero role in the ending of existing affiliate marketing programs, and thus that the effect of affiliate marketing fraud on web users is also small-but-non-zero.

A final possible cost that fraud imposes on web users is that gains to online retailers from affiliate marketing programs could be used by those companies to offer possible price-reductions to shoppers. Affiliate marketing fraud may thus impose costs on web users indirectly in the form of forgone price reductions. Similarly, if an affiliate marketing program causes an online retailer to lose money from investments in the program, the online retailer may chose to pass those loses on to the consumer in the form of higher prices, though market competition may also prevent the online retailer retailer from doing so. The effect of affiliate marketing fraud on the prices of goods sold by online retailers is beyond the scope of this analysis, but is mentioned here for completeness.

6.2.2 Fraudulent Publishers

Not surprisingly, fraudulent publishers benefit the most from affiliate marketing fraud. They benefit whenever a purchase is made on an online retailer's site by a web user carrying the fraudulent publisher's cookie. For Amazon, for example, this amount is between 1 and 10% of the price of each product purchased in each manipulated sale[1]. If a fraudulent publisher is able to redirect a large number of web users, or is able to target the browsers of web users who are likely to make disproportionately large purchases, the affiliate marketing fraud can provide a large return to the fraudster.

The costs a fraudulent publisher faces are minimal. While there may be some initial fixed cost in gaining control of site web users visit—purchasing a hacked site, investing time to find a site vulnerable to XSS injection, generating content users desire, etc.—the marginal cost for each additional redirected web user is nearly zero to the fraudulent party. Similarly, attacks used to redirect browsers may be automatable, further driving down the marginal cost of the attack. Fradulent publishers must also have the ability to receive funds from their chosen affiliate program, either in their own personal financial account or an account they control.

Fraudulent publishers also face possible costs of detection. If an online retailer detects that an affiliate marketing account is being used for fraudulent activity, the online retailer may choose to cancel the affiliate account, and the fraudulent publisher would then lose some unrealized revenue. There have been observed cases of online retailers closing accounts of fraudulent publishers [5, 17], though since most account closures are not likely publicized, we are unable to measure how frequently online retailers close accounts for fraud.

However, this does not shut down the traffic stream the fraudulent publisher was using, and if they still have control over the site, they can easily target a different site to extract revenue from the visitors (perhaps even using something more malicious like a drive-by download). This suggests that the costs a fraudulent publisher faces if detected are not total, since they can monetize their infrastructure in other ways.

Finally, a fraudulent publisher may also face legal costs, both private and criminal [20, 19, 8]. While the number of legal actions against fraudulent publishers seems small when compared against the large amount of fraud occurring in these markets, the magnitude of the cost of legal action to the fraudulent party may be substantial.

6.2.3 Honest Publishers

While it's unlikely that there are any benefits for honest publishers, they certainly stand to lose something in the face of affiliate marketing fraud. Honest parties face the possibility of having their referrals "stolen" by a dishonest party, thereby depriving the honest publisher of the compensation they expected. While affiliate marketing theft was never observed in our data, it is a clear possibility, and given a large enough amount of data we expect instances of affiliate marketing theft would be found. Acknowledging that though, the costs to honest publishers appears to be much more theoretical than practical; we expect the realized costs to honest publishers are minimal.

6.2.4 Online Retailers

Online retailers are the only party in the affiliate marketing system where the balance of costs and benefits of fraud is ambiguous. On one hand, online retailers benefit by having more web users directed to their site. While some methods used by fraudulent publishers may not result in more web users seeing and interacting with the online retailer's site (e.g. hidden iframes), others redirection methods will result in a new browser page being brought up, featuring the online retailer's products (e.g. pop-under windows). While it's unlikely that this irrelevant window would induce a sale, and the online retailers would need to pay the affiliate's fee, it is

possible that some sales would happen as a result of these advertisements.

Affiliate marketing fraud also imposes costs on online retailers. If sales due to fraudulent affiliate marketing referrals become a dominant part of the affiliate marketing program, then affiliate marketing theft may become a more significant issue. Honest publishers, who in our observation drive the majority of affiliate marketing based sales, may decide to leave the affiliate marketing program either in favor of a different affiliate marketing program or for alternative revenue generation schemes for their content. Were this to happen, an online retailer would be hurt by the loss of additional honestly referred sales in the future.

Online retailers are also harmed by the additional payouts they must make to fraudulent publishers. If the fraudulent publishers are claiming credit for sales that would have occurred regardless of the fraudulent marketer's actions, then the fraction of the sale the online retailer must share with the fraudulent publisher is pure cost. However, if the web users referred by the fraudulent publishers make purchases that otherwise would not have been made, then the affiliate marketer's fee is likely dwarfed by the additional revenue the sale brought the online retailer.

Of the four parties in the affiliate marketing system, online retailers are unique in facing both real costs and benefits. The net benefit to online retailers will depend on several business specific factors, and is beyond the scope of this analysis.

7. CONCLUSIONS

This paper describes a large scale investigation of the role and frequency of fraud in affiliate marketing programs. We developed an efficient, automated approach for detecting fraud with 93.3% accuracy based on HTTP request logs. Using this approach, we measured the honest and fraudulent activity of 166 affiliate marketing programs across 6 affiliate marketing networks in 2.3 billion HTTP requests. These measurements allowed us to estimate the gains and losses of the four classes of participants with a financial interest in the market: web users, honest publishers, fraudulent publishers, and online retailers who run affiliate marketing programs.

In our dataset, two affiliate marketing programs were dominant, with the programs run by Amazon and GoDaddy accounting for more than 99.9% of observed affiliate marketing activity. A significant proportion of the publisher accounts in these programs appear to be engaged in fraud (38.1%). However, the fraudulent accounts are less successful than their honest counterparts in generating referrals (17.8% of referrals appear fraudulent), and less successful in generating referrals that end up making purchases (18.1% of observed conversion events were performed by web users carrying the affiliate identifier of a fraudulent publisher).

While this data is limited in scope due to collection artifacts (ex. the lack of HTTPS data), the proportions between each type of affiliate marketing event allow us to provide a stakeholder analysis of affiliate marketing fraud and to identify how each party involved in affiliate marketing fraud is hurt or benefited by the activity. Although legitimate marketers could possibly be harmed by affiliate marketing fraud, we did not observe any such cases in this work. Only one of the four parties involved—the online retailer running the affiliate marketing program—faces substantial negative effects, and even these costs may be overwhelmed by the additional traffic and sales generated. We also identified other costs of affiliate marketing fraud, such as the possible collapse of affiliate marketing programs and the free content that affiliate marketing programs fund.

More generally, we find that stakeholder analysis is useful for putting a perspective on the harms involved in cybercrime. Performing this analysis on other attack or fraud schemes could aid the community in better understanding what types of deceptive and fraudulent behavior generate public costs, and thus where researcher effort can be focused to maximize societal benefit. Conversely, we expect that stakeholder analysis of other cybercrime activities may reveal that most of the negative externalities due to the fraudulent activity are actually quite marginal, and thus these problems are of lower concern for researcher effort.

8. REFERENCES

[1] Amazon.com, I. Amazon.com associates central - fixed advertising fee rates for specific product categories. `https://affiliate-program.amazon.com/gp/associates/help/operating/advertisingfees`, 2013. [Online; accessed 8-May-2015].

[2] Chachra, N., McCoy, D., Savage, S., and Voelker, G. Empirically characterizing domain abuse and the revenue impact of blacklisting. In *Proceedings of the Workshop on the Economics of Information Security (WEIS)(State College, PA* (2014).

[3] Clayton, R., and Mansfield, T. A study of whois privacy and proxy service abuse. In *Proceedings (online) of the 13th Workshop on Economics of Information Security, State College, PA (June 2014)* (2014).

[4] Company, G. O. Affiliate program | join the world's largest registrar - godaddy. `https://www.godaddy.com/affiliates/affiliate-program.aspx`, 2015. [Online; accessed 3-March-2015].

[5] Corbin, K. ebay affiliate pleads guilty in cookie-stuffing scam. `http://www.ecommercebytes.com/cab/abn/y13/m04/i26/s02`, 2013. [Online; accessed 8-May-2015].

[6] Dave, V., Guha, S., and Zhang, Y. Viceroi: catching click-spam in search ad networks. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (2013), ACM, pp. 765–776.

[7] eBay Inc. ebay partner network. `https://www.ebaypartnernetwork.com/files/hub/en-US/index.html`. [Online; accessed 5-March-2015].

[8] Edelman, B. Affiliate fraud litigation index. `http://www.benedelman.org/affiliate-litigation/`, 2015. [Online; accessed 6-May-2015].

[9] Inc., A. Amazon.com associates: The web's most popular and successful affiliate program. `https://affiliate-program.amazon.com/`, 2015. [Online; accessed 3-March-2015].

[10] Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., and Savage, S. Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM conference on Computer and communications security* (2008), ACM, pp. 3–14.

[11] Khan, M. T., Huo, X., and Li, Zhou amd Kanich, C. Every second counts: Quantifying the negative externalities of cybercrime via typosquatting. In

Proceedings of the 36th IEEE Symposium on Security and Privacy (2015).

[12] Kshetri, N. The economics of click fraud. *IEEE Security & Privacy 8*, 3 (2010), 45–53.

[13] Moore, T., and Edelman, B. Measuring the perpetrators and funders of typosquatting. In *Financial Cryptography and Data Security*. Springer, 2010, pp. 175–191.

[14] Mozilla. Mozilla ca certificate store. `https://www.mozilla.org/en-US/about/governance/policies/security-group/certs/`, 2015. [Online; accessed 22-February-2015].

[15] Neasbitt, C., Perdisci, R., Li, K., and Nelms, T. Clickminer: Towards forensic reconstruction of user-browser interactions from network traces. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 1244–1255.

[16] Peacock, T., and Friedman, A. Automation and disruption in stolen payment card markets. *Criminal Justice Studies 23*, 1 (2010), 33–50.

[17] Porter, W. Linkshare - lands' end versus the affiliate on typosquatting | revenewsrevenews. `http://www.revenews.com/search-engine-marketing/new-first-linkshare-lands-end-versus-the-affiliate-on-typosquatting/`, 2006. [Online; accessed 6-May-2015].

[18] Schneider, F., Ager, B., Maier, G., Feldmann, A., and Uhlig, S. Pitfalls in http traffic measurements and analysis. In *Passive and Active Measurement* (2012), Springer, pp. 242–251.

[19] U.S. District Court for Northern District of California, S. J. D. *United States vs. Brian Dunning*. 2010. [CR-10-0494].

[20] U.S. District Court for Northern District of California, S. J. D. *United States vs. Christopher Kennedy a/k/a biglevel*. 2010. [CR-10-0082].

[21] W3C. Referrer policy - w3c first public working draft, 7 august 2014. `http://www.w3.org/TR/referrer-policy/`, 2015. [Online; accessed 5-March-2015].

[22] Walmart. Home - affiliate program - walmart.com. `https://affiliates.walmart.com/`, 2014. [Online; accessed 3-March-2015].

[23] Xie, G., Iliofotou, M., Karagiannis, T., Faloutsos, M., and Jin, Y. Resurf: Reconstructing web-surfing activity from network traffic. In *IFIP Networking Conference, 2013* (2013), IEEE, pp. 1–9.