Affiliate Advertising Fraud and an Investigatory Framework for Fraud

Mathew Miehling, Prof William J Buchanan, Alistair Lawson and Alan Batey

Centre for Distributed Computing and Security,
Edinburgh Napier University
{m.miehling,w.buchanan,al.lawson}@napier.ac.uk
http://cdcs.napier.ac.uk

Abstract. This paper outlines the details of a real-life scam, which involves creating fraudulent Web sites which pretend to sell goods, but are actually used to perform click-through crime or use identity fraud to claim commission on the sale of goods. It involves an investigation of real-life investigatory data, which outlines the methodology used to implement an investigatory framework. This novel framework allows an investigator to use anonymised data, which still has the context of the investigation.

1 Introduction

At present, an increasing number of businesses are turning toward an online presence for their advertising campaigns, and many are moving toward Affiliate Advertising programmes. These programmes offer web page publishers revenue for driving business toward a merchant's site and, according to a report released by an international digital marketing firm, Econsultancy, generated £4.62 billion worth of sales in the UK alone last year [6]. Unfortunately, as this market matures, individuals have begun to discover malicious ways of manipulating their revenue figures in order to earn money fraudulently through a variety of scams. The aim of the investigatory framework defined in this paper is to assist advertising networks to detect malicious activity. One of the key aspects of this is the ability to analyse the risks within websites and present them to the investigator.

In the UK alone, it is estimated that £38.4 billion is identified on fraud in general [4] with £27 billion allocated directly to cybercrime [5]. Affiliate Advertising Fraud falls into the Online Scams subsection which is estimated to have an annual total cost of £1.4 billion according to the Office of Cyber Security and Information Assurance [5]. In fraud related investigations there is sometimes a need to investigate a crime without imparting any personal bias on the evidence, thus the aim of this work which is carried out in collaboration with the Financial Services Authority (FSA) in London, is to produce a novel investigation infrastructure in which scripting can be used to define the complete investigation process, and where each step of this process can be entirely matched to the requirements of the investigation. The key objectives of the project are:

- Produce a proof-of-concept investigatory infrastructure.
- Investigate and implement criminal fraud classifications (online advertising fraud, credit card fraud, ID theft, and so on).
- Evaluate investigatory framework to validate our hypothesis.

2 Online Advertising Programmes

A typical online advertising scheme consists of three entities:

- Advertisers, synonymous with publisher and affiliate, are the people who create the content that is responsible for driving traffic to a merchant's site.
- Advertising Network, also called the affiliate network, which acts as a middle-man between the publishers and merchants that are a part of their network.
- Merchants are the entities that are actually selling the product or services.

There are two main types of advertising programmes: Pay Per Click (PPC) and Pay Per Action (PPA) [8]. In a PPC advertising programme, if a publisher's advertising link is clicked, the publisher receives a set amount of money. This programme often has a low payout per click as it is highly prone to abuse. Many PPA programmes work through the use of tracking cookies, so that when a user visits a publisher's page, and clicks on an advertisement link, a cookie is placed on that user's computer. The merchant can then use the tracking cookie to credit the proper merchant with the commission.

Figure 1 provides an overview of a typical affiliate programme scam. Valid publishers (AdvertiserB and AdvertiserC) create web pages with content related to the products in order to drive business to the merchant's site. AdvertiserA, though, has set up a site whose sole purpose is to facilitate advertising fraud either by using fake or stolen identity information to generate sales commission, or to generate pure click-though commission.

Many of the original scams include either using a script or manually clicking repeatedly on an advertising link in order to either inflate their revenue, or to deplete the daily allotted advertising budget of a competitor [17]. In order to see where the clicks originate an advertiser may examine their click logs that are often provided to both merchants and publishers by several of the larger advertising networks [21]. They are often from the same IP address, and are sometimes even from a remote region in which the services offered on the advertiser's site are not even offered [14].

As the advertising networks began to crack down on the basic form of click fraud, the malicious users have been driven to come up with methods that were more difficult to detect. From these efforts has come the practice of forcing a click, where users visiting a site are forced to click a link that they would not normally not click on [8]. The most common example is to display the advertising link in a pop-up window as the user browses to a publisher's site. Once the link is loaded, the advertiser's tracking cookie has been put onto the user's PC and the merchant will credit that advertiser even if their page is not the reason that the user makes a purchase [26]. Another example of forced click can be found on

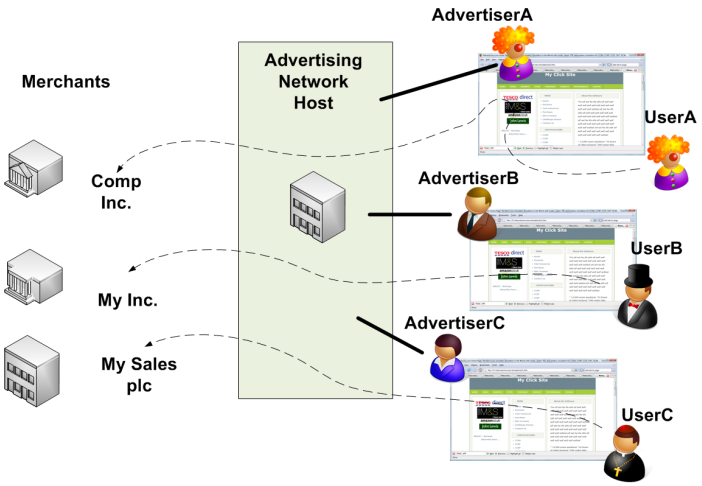

Fig. 1. Affiliate Fraud

many online voucher sites. These sites exist so that users can find codes to be entered when shopping online to receive discounts. If a user searches on these voucher sites looking for deals for ordering pizza, they may find a page full of great deals that would then require that they click a link to reveal the code needed to take part in the deal. When this link is clicked, the tracking cookie for that publisher is then put onto that user's PC [29] and the pizza merchant will credit that publisher with commission for the sale, even though they already knew which merchant they wanted to use.

3 Literature Review

This section outlines some of the key literature related to the research.

3.1 Privacy Preservation

The world of research is powered by data, and without access to research data, it would be near impossible for the research community to validate any possible breakthroughs [30]. However, when this data contains sensitive information about people, there is often moral and legal reasons to preserve the privacy of any individuals in order to remain compliant with laws, such as for the Data Protection Act (DPA) [16]. The idea is to enable the release of research data, but to ensure that no individual can be identified using any other information, including information outside the supplied data set [23].

A table in a data set could be thought of as a group of rows with each row corresponding to a specific person or organisation. For simplicity's sake, we refer to these people and organisations as respondents throughout the rest of this

paper. Rows contain two different types of fields: *public* and *private*. The public fields contain information that most people would not mind sharing and that, by itself, cannot be used to specifically identify a respondent. These fields may include information such as age, date of birth, and zip/post code. The private fields contain information that a person would not want shared or that could be used to positively identify a respondent. This information includes fields such as surname, social security number, national insurance number, salary, medical condition, and other sensitive information.

There are several types of variables used throughout data sets, where a categorical variable is a variable with multiple possible categories, but without order to those categories [3]. For example, hair colour has many possible categories such as brown, blonde, black or ginger. However, there is no inherent way to order the categories by value or importance so hair colour is said to be categorical. The main goal of privacy preservation schemes is to enable the sharing of important data without allowing a non-trusted person to infer the values of any private fields related to any respondents.

Generalisation and Suppression Samarati and Sweeney introduced the concept of k-anonymity in order to achieve a goal of privacy preservation. In order to be considered k-anonymous, any record in the table must exactly match at least *k-1* other records. This ensures that no records can be matched directly back to an individual. In order to achieve k-anonymity, the data provider must be capable of distinguishing *quasi-identifiers (QID)* present in a table. *Quasi-identifiers* are defined as public attributes that may also be available in external data sources [12] and may be used to re-identify records [11]. Some examples of common QIDs include age, zip/post code, gender and ethnicity.

Almost all k-anonymity schemes implement some form of generalisation and/or suppression, in order to meet the requirements. It is possible to generalise the values of quasi-identifiers, such that k-anonymity is achieved [27]. In generalisation techniques the goal is to replace quasi-identifiers with a generic value that maps to multiple specific values. For example, instead of displaying the town a person is from, the data provider could generalise the value to the state, region or even country. The trade-off with generalisation is that as a value becomes more general, context is lost [22].

Having been considered the de-facto anonymisation technique by many experts in the field, k-anonymity has had several improvements made to it along the way such as l-diversity and t-closeness. Table 3.1 shows a truncated example of a patient database from a fictional hospital. Based upon the data shown in the table, this data set is two-anonymous because the QIDs (gender and age in this case) resolve to at least two records. For example's sake, let's say that a non-trusted person, who is a a person with authorisation to view the records, but one that may be acting outside of the remit of their authorisation, is curious about the diagnosis of Jim. The snooper looks up Jim's age on his Facebook account in order to further narrow down the results. The non-trusted person is then left with the records for Bob, Jim and Ben (but he cannot see the names,

because that field is suppressed). The attacker can now infer with 80% certainty that Jim's diagnosis is positive because two out of the three possible respondents are positive. l-diversity was developed to ensure that the values of private fields contain enough diversity to prevent the problem showcased in Tbl. 3.1 [19]. In response to l-diversity, Li et al. developed the concept of t-closeness saying that not only did the values of sensitive fields need to be diverse, but that the difference in the distribution of these values in the equivalence class, and in the overall table, must be within the threshold t.

Name (Suppressed)	Gender	Age	Diagnosis (Sensitive)
Bob	Male	50-60	Positive
Jim	Male	50-60	Positive
Ben	Male	50-60	Negative
George	Male	50-60	Positive
Larry	Male	50-60	Positive
Joy	Female	20-30	Negative
Claire	Female	20-30	Positive
Greg	Male	20-30	Positive
Moe	Male	20-30	Negative

Aside from modifications to the original theory of k-anonymity, new implementations have been developed in order to simultaneously solve a second problem along with preserving privacy. For example, many early generalisation techniques were designed to support only one sensitive field per respondent. In practice, this is highly unlikely, and He and Naughton have addressed this issue with their k-anonymous implementation that they have called top-down, local generalization [15]. Often, though, generalisation is not enough to achieve k-anonymity, so using a voter registration list purchased for twenty dollars along with an insurance database thought to be anonymised by the data provider, Sweeney was able to track down the medical records of the governor of Massachusetts using linking techniques [28]. In order to prevent this attack, a technique called suppression can be used [23] along with generalisation.

Suppression is the process of not disclosing particular fields in the database that could be used in such a linking attack [27]. If the Zip code field of the voting records or an insurance database in [28] had been suppressed, the task of re-identifying the governor would have been marginally more difficult. Unfortunately, suppression vastly decreases the quality of the data, and may even alter statistics making the data useless for many purposes including our own.

Perturbation Rather than simply omitting data, one could choose to modify certain fields in order to reduce a non-trusted person's ability to re-identify any specific responses using a method known as perturbation [25]. One of the original perturbation methods, defined in [13], is the Post Randomisation Method (PRAM). Gouweleeuw et al. were able to perturb a file in such a way that answers

to specific questions were unable to be traced back to a particular respondent, but analysts were still able to make valid inferences about the original data. An analyst attempting to make inferences from the file must make corrections to account for randomly introduced errors. Because the complete distribution of errors is known by the analyst, Gouweleeuw et al. argue, and are supported by [24], that the process is only slightly more cumbersome than normal for categorical variables. Some, such as Aggarwal argue that because perturbation is designed for randomisation, and there is no guarantee of the privacy that k-anonymity offers [2], which is a definite drawback.

Substitution As laid out in our previous work [20], the anonymisation portion of this framework is crucial because it is to be used within an investigatory setting, where it is a difficult task to anonymise a table of data and maintain all of the relationships between entities in the table. Unfortunately, the methods defined in the previous sections tend to transform the data into a format that no longer preserves the original context, and thus a substitution [9] method is used to bypass this restriction.

The method used in this paper is thus to use method called Table Internal Synchronization [10]. This data substitution method is more like the blanket substitution mentioned in [9], but it takes the process a step further and ensures that if the name "Mat Miehling" is changed to "Fred Smith", in the first instance it will also be changed to "Fred Smith" in every subsequent instance. This is essential to meet the context preserving goal of our framework and to maintain the relationship between respondents.

4 Details of the scam

The following defines the context of the real-life scam, without revealing the details of it. According to the incident report given to us by the police, the affiliate network, AffiliateNow, received a complaint from Merchant2 about suspected fraudulent behaviour. The user involved, Stan Smith of 416 Cherry Street in Gotham City, had been sending fraudulent leads to Merchant2 and earning commission from them. Upon internal investigation by Merchant2's fraud team, Merchant2 had decided not to honour the commission earned by Stan Smith. Merchant2 then raised an incident report with AffiliateNow to warn other merchants of his fraudulent behaviour and to have him removed from the network.

The AffiliateNow employee investigating Stan Smith's case found that the traffic being sent to merchants from Stan Smith's affiliate account was coming from the same referring site and many of the IP addresses were repeated. The repeated IP addresses were all from foreign countries and visiting sites that only offered services in the UK. AffiliateNow suspended Stan Smith's account and issued a warning to all affected merchants about Stan Smith's account.

That is as far as the incident report we have received seems to go. However, upon further investigation, we have found several links from Stan Smith to other accounts in the affiliate database. The originally reported account is linked to 5

other affiliate accounts in the database, four of which are listed in Stan Smith's name with the 5th having his name in the cheque payable field and the name Edward Smith as the account holder. These different accounts have four unique physical addresses, two of which have been listed by other affiliates as their address.

Three of the accounts we examined have different names in the cheque payable and account holder fields. Of the five affiliates with bank account information on file, three also listed a different name on the bank account than that of the account holder. The greatest anomaly that we discovered involved the telephone numbers listed for the affiliates. Surprisingly, only six out of the 28 affiliates examined listed mobile phones when asked for a phone number. This is helpful to our analysis because landline telephone numbers can be traced back to a general area. Only one of the remaining 22 numbers, however, had a dialling code consistent with the address information provided by the affiliate. This should be a significant clue that something is amiss with the accounts of these affiliates.

In looking at the customer database, several inconsistencies are present. A system designed to seek out these anomalies may enable affiliate networks to flag accounts for a closer inspection by an employee. For example, if a detection system had been run in our case study it may have picked up that Stan Smith was registered to multiple physical addresses. It may have also picked up that multiple users were registered to these addresses as well. Linking these accounts together may enable the affiliate network to remove large chunks of fraudulent accounts with a single investigation rather than a new investigation for each account.

We believe that the most suspicious detail in the customer records is the fact that none of the telephone numbers originate from the area listed in the address details of the affiliate. A person listed as living in Gotham City may have a phone number with a dialling code for Shelbyville, for example. A comparison between the dialling code and postcode of the listed address could easily mark such an account at a high risk for being fraudulent if a landline phone number is provided during the affiliate registration process.

Another tell-tale sign of fraudulent behaviour is an in-depth look at the affiliate's site. If the site consists mainly of banners and ads, or is in some other way inappropriate for the products being advertised, the page may belong to a malicious affiliate. If the affiliate does not use proper grammar and complete sentences it may be a sign that the site was hastily made. If the images appear broken or are taken from another site, something suspicious may also be going on with the affiliate. These are a couple of the more obvious signs that the site may have been created simply to host the ads and scripts necessary to generate a fraudulent income.

Less obvious signs might be found in the code of the affiliate's site. Websites that contain scripts used by known fraudsters such as the code example shown in Figure 2 should probably be looked at more closely. If an affiliate is producing dozens of sites for their operation, they are likely to all have a similar layout and

similar mistakes in their code. Running the website through a HTML and CSS validation checker on suspected pages may produce similar results, which could be an indication of multiple accounts involved in dodgy behaviour.

Weighing each of these categories and keeping track of an affiliate's score while running these tests could give an indication of whether or not the affiliate is genuine, fraudulent or undetermined. In the case of fraudulent and undetermined, the case could be moved to the fraud team of the affiliate network for further investigation.

Apart from the affiliate database and sites, a good indication that an account is involved with fraudulent behaviour has duplicate IP addresses appearing from the same affiliate on multiple merchants. An occasional duplicate IP address is not necessarily fraudulent, but the same duplicate IP address(es) multiple times in a small time period is pretty suspicious. Another method of combating the rising number of malicious affiliates is to prevent them from joining a programme in the first place. Edelman posits that it may be possible to prevent fraudsters from joining an affiliate programme all together ([7]). He found that if a merchant pays their affiliates in arrears with compensation to offset the extra time before payment is received, there exists a certain point at which it is no longer profitable for fraudsters, or bad-type agents as he calls them, to participate in the programme. Unfortunately, according to a recent survey ([1]) of over 450 affiliates, 57% of good-type affiliates decide whether to join an affiliate programme based upon how often a programme pays out. With the majority of affiliates basing their preference of programme on how soon they start earning, extending that wait may decrease the number of good affiliates a merchant or affiliate network can attract.

5 Investigatory Framework for Fraud

The current research aims to provide an anonymisation framework for investigations that also preserves context, but which is still useful for investigators (see Fig. 3). The framework is designed in a modular fashion to allow for customisation in the level of assistance and methods of visualising the data available to the user. The anonymisation portion of the framework takes in affiliate data from an advertising network's customer database and substitutes fake values into every field of the table. Once a value has been assigned a substitution that value will always have the same substitute in this data set. This allows the user to maintain relationships between entities which is essential to our context-preserving element of the framework. The original data is then securely stored unperturbed and out of reach of the investigator. In order to allow for quicker anonymisation, any field that is not needed by the user can be marked for masking or exclusion from the resulting anonymised table. Figure 3 outlines the current implementation, where the investigatory engine assesses the risks related to the crime, such as for, in the case of affiliate fraud, that the remote sites have a large number of URLs within each page, or that there are a large number of broken links within the site. Agents are then used to gather this data from the sites under investiga-

Anon_ID	Tel Area	Anon_First	Anon_Last	Snailmail	Snailmail2	Logins	Earned
100026	Gotham City (A)	Stan	Smith	1 Spooner Street	Gotham City (G)	13	£0.00
100027	Gotham City (A)	Edward	Smith	1 Brookside Close	Gotham City (J)	41	£0.00
100006	Mobile	Stan	Smith	10 Evergreen Terrace	Gotham City (D)	206	£801.00
100003	Gotham City (B)	Robert	Johnson	10 Evergreen Terrace	Gotham City (D)	183	£1,180.00
100004	Liberty City	Robert	Johnson	10 Evergreen Terrace	Gotham City (D)	416	£8,103.10
100007	Sunnydale	Frank	Smith	10 Evergreen Terrace	Gotham City (D)	232	£2,404.25
100008	Shelbyville	Richard	Miller	10 Evergreen Terrace	Gotham City (D)	15	£535.00
100002	Springfield	James	Williams	10 Evergreen Terrace	Gotham City (D)	10	£150.00
100010	Gotham City (B)	Joseph	Garcia	10 Evergreen Terrace	Gotham City (D)	11	£285.00
100013	Gotham City (C)	Daniel	Taylor	10 Evergreen Terrace	Gotham City (D)	147	£3,781.96
100014	Mobile	Paul	Martin	10 Evergreen Terrace	Gotham City (D)	3	£0.00
100011	Petoria	Thomas	Anderson	10 Evergreen Terrace	Gotham City (D)	166	£1,052.00
100009	Mobile	Charles	Davis	10 Evergreen Terrace	Gotham City (D)	94	£28.00
100012	Gotham City (I)	Christopher	Anderson	10 Evergreen Terrace	Gotham City (D)	220	£50.00
100015	Gotham City (D)	Mark	White	10 Evergreen Terrace	Gotham City (D)	13	£0.00
100001	Mobile	John	Doe	10 Evergreen Terrace	Gotham City (D)	7	£0.00
100005	Ogdenville	Craig	Smith	10 Evergreen Terrace	Gotham City (D)	56	£69.00
100000	Mobile	Mike	Rogers	10 Evergreen Terrace	Gotham City (D)	17	£0.00
100022	Gotham City (E)	Christopher	Anderson	416 Cherry Street	Gotham City (C)	316	£1,602.70
100025	Gotham City (F)	Aaron	Robinson	416 Cherry Street	Gotham City (C)	45	£45.00
100023	Gotham City (G)	Caleb	Harris	416 Cherry Street	Gotham City (C)	18	£480.00
100016	North Haverbrook	Matthew	Brown	416 Cherry Street	Gotham City (C)	14	£220.00
100021	Gotham City (B)	Nicholas	Wilson	416 Cherry Street	Gotham City (C)	13	£300.00
100018	Mobile	Stan	Smith	416 Cherry Street	Gotham City (C)	267	£587.00
100017		Tony	Jones	416 Cherry Street	Gotham City (C)	74	£132.24
100024	Gotham City (F)	Stan	Smith	416 Cherry Street	Gotham City (C)	66	£318.10
100019	Gotham City (H)	Craig	Smith	416 Cherry Street	Gotham City (C)	16	£128.00

Fig. 2. Example data

tion, and presented to the investigator in a way that prioritizes the risks involved. The investigator can then choose to select the key investigatory parameters, in order to more quickly achieve the required results. Once the investigator has investigated the case, the remapping process can then be used to determine the mapping back from the substitution to the real data, which can then reveal the actual details of the criminal investigation.

6 Conclusions and Future Work

This paper has outlined a basic methodology for preserving the context of an investigation, based on a real-life scam. Unfortunately there is often very literature published which relates to the actual detail of fraudulent activity, often because it can be difficult to publish the details of an investigation. This is unfortunate as fraud is a growing problem in the UK and world-wide, thus methods must be put in-place to be able to investigate these activities, and for businesses to use risk-based models to assess whether their partners have malicious activities. The methods used in this paper have been used in a real-life case, and thus have proven success in preserving the context of a problem, while preserving the anonymity of those involved, until some form of crime can be implied. Current

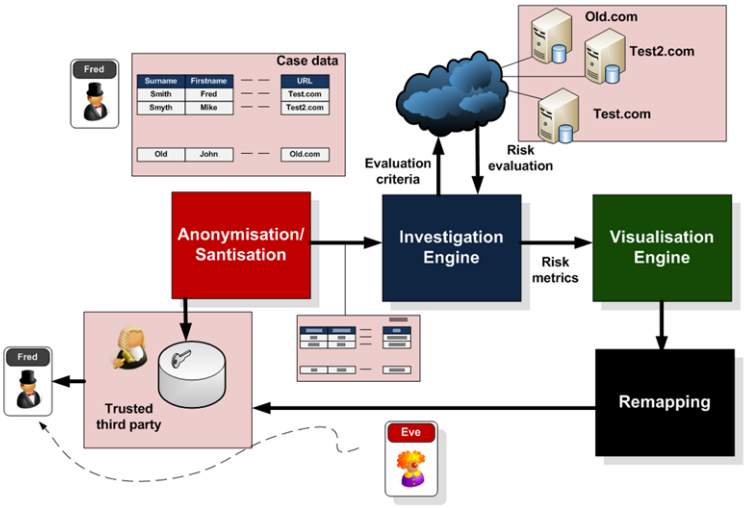

Fig. 3. Framework Overview

work involves defines the risks involved in affiliate crime, and will be used to better inform future risk models for businesses involved in affiliate marketing.

Bibliography

[1] AffStat. 2009 affiliate summit affstat report. Technical report, 2009. URL http://affstat.com/.

[2] Charu C. Aggarwal. On unifying privacy and uncertain data models. *Data Engineering, International Conference on*, 0:386–395, 2008. doi: http://doi.ieeecomputersociety.org/10.1109/ICDE.2008.4497447.

[3] Alan Agresti. *Analysis of Ordinal Categorical Data*. Wiley, second edition, 2010.

[4] National Fraud Authority. Annual fraud indicator. Technical report, National Fraud Authority, January 2011. URL http://www.attorneygeneral.gov.uk/nfa/WhatAreWeSaying/Documents/AFI%202011.pdf.

[5] Detica. The cost of cyber crime. Technical report, Office of Cyber Security and Information Assurance in the Cabinet Office, February 2011. URL http://www.cabinetoffice.gov.uk/sites/default/files/resources/the-cost-of-cyber-crime-full-report.pdf.

[6] Econsultancy. Affiliate marketing buyer's guide 2010. Technical report, Econsultancy Digital Marketers United, August 2010. URL http://econsultancy.com/uk/reports/affiliate-marketing-buyers-guide.

[7] Edelman. Cpa advertising fraud: Forced clicks and invisible windows. Technical report, 2008. URL http://www.benedelman.org/news/100708-1.html.

[8] Ben Edelman. Cpa advertising fraud: Forced clicks and invisible windows, October 2008. URL http://www.benedelman.org/news/100708-1.html.

[9] Dale Edgar. Data sanitization techniques. White paper, Net 2000 Ltd, 2004.

[10] Dale Edgar. Data masking: What you need to know. White paper, Net 2000 Ltd, 2010.

[11] Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627 – 637, 2008. ISSN 1067-5027. doi: DOI:10.1197/jamia.M2716. URL http://www.sciencedirect.com/science/article/B7CPS-4T9DCS2-B/2/dd9631a3af70d2e800e45564408655e6.

[12] Arik Friedman, Ran Wolff, and Assaf Schuster. Providing k-anonymity in data mining. *The VLDB Journal*, 17:789–804, July 2008. ISSN 1066-8888. doi: http://dx.doi.org/10.1007/s00778-006-0039-5. URL http://dx.doi.org/10.1007/s00778-006-0039-5.

[13] Jose Gouweleeuw, Peter Kooiman, Leon Willenborg, and Paul De Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14(4):463–478, December 1998.

[14] Brian Grow, Ben Elgin, and Moira Herbst. Click fraud: The dark side of online advertising, October 2006. URL http://www.businessweek.com/magazine/content/06_40/b4003001.htm.

[15] Yeye He and Jeffrey F. Naughton. Anonymization of set-valued data via top-down, local generalization. *Proc. VLDB Endow.*, 2:934–945, August 2009. ISSN 2150-8097. URL `http://portal.acm.org/citation.cfm?id=1687627.1687733`.

[16] United Kingdom. Data protection act, 1998.

[17] N. Kshetri. The economics of click fraud. *Security Privacy, IEEE*, 8(3):45 –53, 2010. ISSN 1540-7993. doi: 10.1109/MSP.2010.88.

[18] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106 –115, 2007. doi: 10.1109/ICDE.2007.367856.

[19] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1, March 2007. ISSN 1556-4681. doi: http://doi.acm.org/10.1145/1217299.1217302. URL `http://doi.acm.org/10.1145/1217299.1217302`.

[20] Mathew Miehling, William J Buchanan, John L Old, Alan Batey, and Arshad Rahman. Analysis of malicious affiliate network ac-tivity as a test case for an investigatory framework. In *Proceedings of 9th European Conference on Information Warfare and Security*, July 2010.

[21] Yanlin Peng, Linfeng Zhang, J. Chang, and Yong Guan. An effective method for combating malicious scripts clickbots. In Michael Backes and Peng Ning, editors, *Computer Security ESORICS 2009*, volume 5789 of *Lecture Notes in Computer Science*, pages 523–538. Springer Berlin / Heidelberg, 2009. URL `http://dx.doi.org/10.1007/978-3-642-04444-1_32`.

[22] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13:1010–1027, 2001. ISSN 1041-4347. doi: http://doi.ieeecomputersociety.org/10.1109/69.971193.

[23] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.37.5829`.

[24] Natalie Shlomo. Releasing microdata: Disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality*, 2 (1), 2008. URL `http://repository.cmu.edu/jpc/vol2/iss1/7`.

[25] Natalie Shlomo and Ton De Waal. Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics*, 24 (2):229–253, June 2008.

[26] Rajendran Sriramachandramurthy, Siva K. Balasubramanian, and Monica Alexandra Hodis. Spyware and adware: How do internet users defend themselves? *American Journal of Business*, 24(2), Fall 2009.

[27] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. 2002. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.7384`.

[28] Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, October 2002. ISSN 0218-4885. doi: 10.1142/S0218488502001648. URL http://portal.acm.org/citation.cfm?id=774544.774552.

[29] A. Tuzhilin. The lane's gifts v. google report. Technical report, Stern School of Business at New York University, 2006. URL http://www.reference.com/go/http://googleblog.blogspot.com/pdf/Tuzhilin_Report.pdf.

[30] Andrew Vickers. Whose data set is it anyway? sharing raw data from randomized trials. *Trials*, 7:1–6, 2006. URL http://dx.doi.org/10.1186/1745-6215-7-15. 10.1186/1745-6215-7-15.