Improving Click Fraud Detection by Real Time Data Fusion

Mehmed Kantardzic[1], Chamila Walgampaya[1], Brent Wenerstrom[1], Oleksandr Lozitskiy[1],
Sean Higgins[2], Darren King[2]
[1] CECS Department, Speed School of Engineering, University of Louisville, Louisville, KY 40292, USA
{mmkant01, ckwalg01, bkwene01, o0lozi01 } @louisville.edu
[2] Hosting.com, Louisville, KY 40202, USA
sean@srhiggins.com, dking@Hosting.com

Abstract

Click fraud is a type of internet crime that occurs in pay per click online advertising when a person, automated script, or computer program imitates a legitimate user of a web browser clicking on an ad, for the purpose of generating a charge per click without having actual interest in the target of the ad's link. Most of the available commercial solutions are just click fraud reporting systems, not real-time click fraud detection and prevention systems. A new solution is proposed in this paper that will analyze the detailed user click activities based on data collected form different sources. More information about each click enables better evaluation of the quality of click traffic. We utilize the multi source data fusion to merge client side and server side activities. Proposed solution is integrated in our CCFDP V1.0 system for a real-time detection and prevention of click fraud. We have tested the system with real world data from an actual ad campaign where the results show that additional real-time information about clicks improve the quality of click fraud analysis.

1. Introduction

Web search is a fundamental technology for navigating the Internet and it provides information access to millions of users per day. Internet search engine companies, such as Google, Yahoo and MSN, have revolutionized not only the use of the Internet by individuals but also the way businesses advertise to consumers [9, 10, 16]. Typical search engine queries are short and reveal a great deal of information about user preferences. This gives search engine companies a unique opportunity to display highly targeted ads to the user. These search services are expensive to maintain and depend upon advertisement revenue to remain free [10]. Many services such as Google, Yahoo and MSN generate advertisement revenue by selling clicks. This business model is known as pay-per-click system.

In the pay-per-click model internet providers are paid by the traffic they drive to a company's web ads. There is an incentive for dishonest service providers to inflate the number of clicks their sites generate. In addition, dishonest advertisers tend to simulate clicks on the advertisements of their competitor to deplete their advertising budgets [12]. This fraudulent behavior results in bad reputations and often in extra costs. Generation of such invalid clicks either by humans or software with the intension to make money or deplete competitor's budget is known as *click fraud.*

There is no globally accepted mechanism to detect click fraud. Therefore, most of the search marketing industries optimize traffic using its own binary paradigm into valid clicks and invalid clicks utilizing their own metrics[2]. But these solutions are still not mature. The bottom line is, with the exception of a small percentage of obviously genuine or robot clicks, the vast majority of clicks simply cannot be classified to be valid or invalid. Rather, they are of a certain quality. Depending on various attribute-values clicks are of higher or lower quality.

Neither commercial solutions for click fraud detection (usually based only on client side data) nor the search engine solutions (mainly using server data) offer comprehensive methodology to detect suspicious activities in the click traffic including click fraud[6, 7]. We initiated this research with an assumption that more data about each click collected from different sources will result in better estimation of the click quality. The research problem was data collection and data fusion from asynchronous sources.

Multi source data fusion refers to the acquisition, processing and synergistic combination of information gathered by various knowledge sources to provide a better understanding of a phenomenon [15]. In addition to the statistical advantages, the use of multiple data sources may increase the accuracy with which an event can be observed and characterized; in our case an event is a user click on the web site. Data fusion has been applied most prominently to military applications such as battlefield surveillance and tactical situation assessment. Data fusion has also emerged in commercial applications such as robotics, manufacturing, medical diagnosis, and remote sensing [8].

A new approach is proposed in this paper that will analyze the detailed user activities using and integrating data from both server and client side collaboratively to better evaluate the quality of clicks. Also, we are summarizing some historical log data and real-time data about click context for better description of each click event. We utilize appropriate data fusion techniques to merge all recorded activities about click traffic. The main goal is an improvement of a click fraud analysis using extended data set which is integrated in real-time. We have tested our approach with data from an actual ad campaign and it shows that about 27% of click traffic in 2007 and almost 50% of click traffic in 2008 appears to be fraudulent which represents a significant increase in fraud rates compared to traditional commercial solutions.

978-1-4244-3555-5/08/$25.00 ©2008 IEEE

The paper organized as follows. In section 2 an introduction to the pay-per-click model and click fraud categories is given. Section 3 discusses available solutions to detect click fraud in advertising. Section 4 describes the data collection process, while in section 5 a detailed discussion about the fusion of data is given followed by our click fraud detection algorithm in section 6. Section 7 discusses data analysis and results, and conclusions and future work are given in section 8.

2. Pay-Per-Click model

A typical Internet traffic model is depicted in Figure 1[3]. An advertiser can be a company or an individual who would like to display their advertisements on other websites. Publishers are the websites that accept contracts through advertisers to display advertisements. The commissioner is an independent entity that has agreements with both the advertiser and the publisher. It can be a search engine or other advertisement agent. The commissioner keeps track of the advertisers' budgets so that they are not over-spent. Often, the advertiser's site does not administer the advertisement pay-per-click model itself, but rather employs a third party ad network, here referred to as the click fraud detection service, to administer the pay-per-click or click-through program on its behalf[2]. The source of clicks (potentially a source of click fraud) can be either human or software based.

First, a Web user requests a web page on the publisher site. The requested page is loaded along with the advertisement on the Web users' browser. If the Web user clicks on an advertisement hypertext link (for example a banner ad or logo) on that page, the publisher redirects the Web user request to the commissioner's server. The commissioner logs the click for accounting purposes. The commissioner then redirects the Web user's browser to the advertiser site. The publishers are paid based on the click traffics they drive to the advertiser's web site. The commissioner earns a percentage of this revenue. Sometimes these payments are based on number of sales generated in the advertiser's website, rather than the volume of traffic drives by the publisher[3].

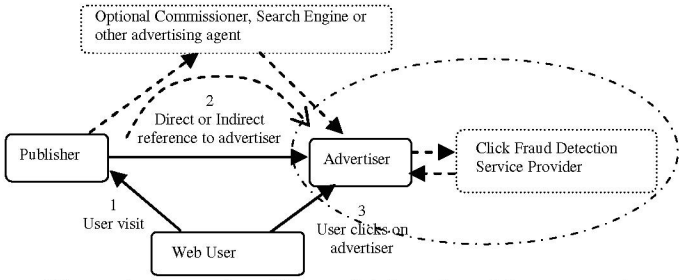

Figure 1. General traffic model for advertising network

2.1 Click fraud in advertising networks

In general click fraud occurs in pay-per-click online advertising when a person, automated script or computer program imitates a legitimate user of a web browser clicking on an ad, for the purpose of generating an improper charge per click. There are two main kinds of fraud in advertising networks, hit shaving, and hit inflation. Hit shaving is the fraud performed by an advertiser, who does not pay commission on some of the traffic received from publishers, by claiming that it received less traffic than the true numbers[11]. Hit inflation is the reverse. Publishers are overpaid by the traffic they drive to advertisers. There is an incentive for them to falsely increase the numbers of clicks their sites generate. We can categorize hit inflation into three main groups for the sake of detection and prevention[3]:

- *Active Human Click Fraud:*

In this category clicks are generated by human activity and the entire web request process is completed, which means that not only the web page is loaded, but also the images, Flash, JavaScipt code, etc. A human may have activity on the web page such as mouse clicks, key strokes and scroll bar activities and the page view time is at least a couple of seconds. These fraudulent clicks are generated mostly by people in developing countries who are hired for a nominal wage.

- *Active Software Click Fraud:*

Specific software initiates the click fraud. Usually, the web page request is not completed and only the initial text web page is requested without the images, JavaScript code or videos. No mouse or keyboard activities are performed, and the page view time is less than one second. A web page request to a server usually contains multiple follow up requests. It first loads the texts in the web page followed by images, multimedia, etc. In click agent software these requests are less likely to happen and such clues can be utilized to identify software clicks. Although some click agent software can generate these follow up requests, they are still different from the browser requests generated by real users, such as missing mouse clicks or having a short page view time.

- *Background Click Fraud:*

This fraudulent activity corresponds to adware and spyware run in the background of the client computer without being known by the user. It hijacks browser sessions and sends out web requests to multiple ad servers. This spyware or adware installed on a client computer without the consent of its user. Such software might pop-up an advertisement window, sometimes pop-under, or might not pop up a window at all. We may include popular clickbots in this category. A clickbot is a software robot that clicks on ads and issues HTTP requests for advertiser web pages. We can classify clickbots into two main categories based on how they are built. First type is "for-sale" clickbots. These clickbots can be purchased for a nominal fee, which runs on individual machines. Second is the "malware" type clickbots. These robots spread from one computer, server or network to another. Usually, clickbots are part of larger botnets, which are networks of software robots that are designed to follow commands given from a particular agent known as the botmaster[3].

3. Commercial solutions and research efforts in click fraud detection

Several commercial solutions, e.g. Authenticlick[2], Clicklab LLC[4], Web Traffic Intelligence Inc[17] etc., are

available for click fraud detection. They all use similar technology by adding a sampler or collecting JavaScript on a page to track. Most of them have draw backs. First, they cannot detect traffic generated by robots if robotic software does not run JavaScript. Second, they don't have a way to detect and prevent click fraud dynamically. Most of them are click fraud reporting solutions compared to preferable click fraud prevention systems. Third, some of them do not use real-time user activity information which is important to detect automatically suspicious click traffic. In terms of our research, these systems are usually based only on one type of click data without comprehensive and integrated information about a click event.

Research activities in the domain of click fraud are concentrated primarily on solutions for specific problems and algorithms, not on integrated click fraud detection systems. Anupam et al presented one of the first and highly referenced papers, which does not describe a solution for the click fraud problem but presents a specific type of hit inflation attack[1]. Reiter et al. brought the problem of hit shaving to the attention of the technical community[13]. They explore simple and immediately useful approaches to enable referrers to monitor the number of click throughs for which they should pay. Metwally's group proposed a solution based on Bloom Filters to detect duplicates in data streams which can be used to detect click fraud[11]. Tuzhilin presents an independent investigation on Google's effort to detect click fraud. He pointed out that "one inherent weakness of Google's data collection effort that is important for detecting invalid clicks, is their inability to get full access to all the clicking activities of the visitors of the advertised websites"[14]. Gandhi et al. proposed a new type of camouflaged click fraud attack on the advertising infrastructure so called "badvertisement"[5]. This stealthy attack can be thought of as a threatening mutation of spam and phishing attacks. The target of this attack is the unwitting advertiser and it could be very serious with significant revenue potential for its perpetrators. The attack was experimentally verified by corrupting the JavaScript file that is required to be downloaded and executed by a client's web browser to publish sponsored advertisements. Mahdian et al. from Microsoft research proposed click based algorithms which are resistant to click fraud. They showed their algorithms satisfying additional technical assumptions are fraud resistant in the sense that a devious user can not change the expected payment of the advertiser per impression[10].

4. Data Collection

Our click fraud detection and prevention system CCFDP V1.0 consists of three main modules: a) click data collection and fusion, b) online scoring of click traffic, and c) extended online traffic monitoring and reporting [3]. In this paper we elaborate characteristics and advantages of our data collection module.

Data are collected asynchronously on both server and client side, and stored in the Global Fraudulent Database (GFD) through the server side log (S) and client side log (C). Also, we are collecting and storing in the GFD background information of the user activities as the clicktracking log (T). The data collection setup is a five step process shown in Figure 2:

(1) Web user, which could be also a fraudulent program, sends a web request for a website (by clicking on an ad).
(2) The web server (publisher) sends server side data and query to GFD.
(3) GFD logs the sever side data and calculate a temporary fraud score.
(4) If the score is below the threshold value, web server sends back response with a tracking code to client computer. If the received fraud score is higher than a threshold, the web server will block the web page request and sends a warning page instead.
(5) The JavaScript tracking code executes on client computer and keeps sending client side log with static and dynamic parameters back to GFD.

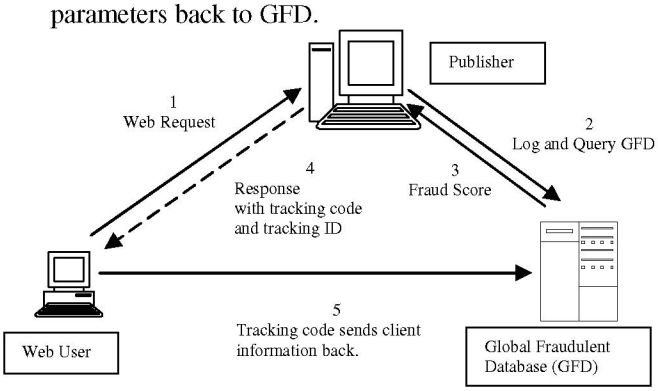

Figure 2. Data collection setup

We are collecting through this process more than 20 attributes of each click. A brief description of key attributes is given in Table 1.

Attribute	Log Sources	Description
IP	S,C	IP address of the user access point
Referrer	S,C	From which server the request is directed
JavaScript details	S, C	What is the user agent? (for example Explorer or Mozilla)
Allowcookie	S, C	Does the web browser accept cookies or not?
Tracking ID	S, C, T	Unique tracking ID (permanent or temporary)
Timezone	S, C, T	Time zone of the user computer

Table 1. Key attributes in extended click record

Hosting.com website was used for the web advertising campaign to collect and analyze click data. Our service providers are Google, Yahoo, MSN and Miva. The campaign is started on January 7th, 2007 and still in process of collecting data. As of June 30th, 2008 we have collected around 1,400,000 natural and 28,000 paid click data.

5. Data fusion

Our data sources for click fraud collection and data fusion can be classified into three categories:

a. *Direct sources* which include: server computer, client computer with clicktracking information.
b. *Indirect sources*: real-time buffer (data records are collected in real-time) and fraudulent database GFD (data columns are generated in extended real-time)
c. *History data*: baselines and Blocking Database (BD).

Data belonging to direct sources are collected in real-time and stored as a record of the click without any preprocessing. Indirect sources define some derived click attributes which require a significant amount of preprocessing. These transformations take place in a real-time for buffer data (recent server data about clicks), and require extended real-time for detection and recording outliers in the current time window compared with various baselines. Data fusion from these sources includes the following activities:

- Server and client side data are fused based on identical tracking ID. If IDs are not the same, the equivalence of IPs is used as an alternative criterion for fusion of partial records.
- Fusion of clicktracking data with client data is much more complex because of their 1:n relation: for each record on the client side there may be more than several user activities recorded in the clicktracking data table. Because we are performing data fusion in real-time we do not have the possibility to "wait" for all click tracking data, and we make "temporary" integration based on a pre-specified window of clicktracking records.
- Characteristics of the clicks are not only "static" based on server and client information. There are also "dynamic" elements about each click describing the context which depends upon the clicks before and after the current one. For example, a single click for a given IP may not be suspicious, but hundred of consecutive clicks from the same IP will make this IP highly suspicious. Similar analysis can be performed with other parameters registered on the server side such as referrer or country. We are using the real-time buffer (recent server click records) to detect these outliers online and to transform them into additional context based characteristics of clicks.
- Collection of data about clicks is extended in real-time giving information about clicks when compared with standard baselines for key parameters.
- Each IP, referrer, etc. (key parameters) may have some history on the given site. These characteristics are included in the record about current click. For example, if the current click is based on a referrer which has a history of suspicious activity, this fact will be included in the record and computation of a click score.
- The blocking database lists highly suspicious IPs, referrers, and countries. Our implementation of the blocking database allows online changes and therefore a more efficient blocking process compared with traditional commercial solutions.

Previous steps are only illustrative examples of a data fusion process, while the details are given in the CCFDP documentation [3]. The integrated structure of a click record which includes all context information is shown in Figure 3.

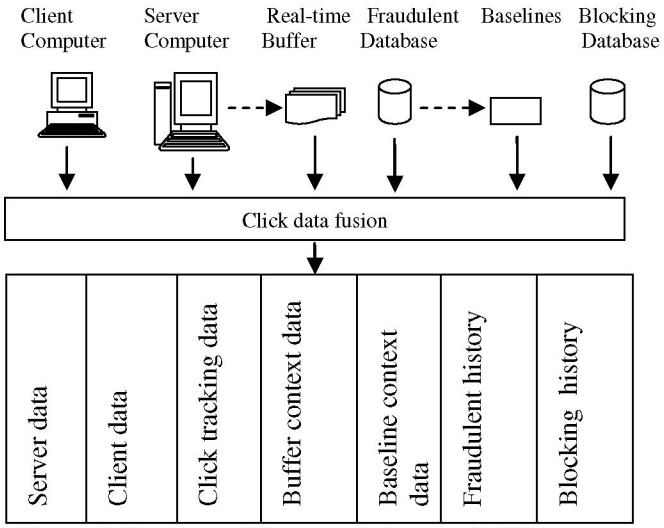

Figure 3: Structure of the record about each click

6. Fraudulent traffic scoring in CCFDP system

Scoring the level of fraudulent activities in the given click requires: a) partial scoring of some characteristics of the fused click record, and b) combining these partial scores into final integrated score S. Score values are normalized over the interval [0, 1], where S=0 represents valid clicks, S=1 is a fraudulent click, and 0<S<1 values are interpreted as suspicious clicks. The sum of all scores for the clicks in the given campaign, normalized by the total number of clicks, represents a percentage of invalid clicks. When the new clicks are coming, the scoring process includes:

1. Collecting the values for partial scores r_i from the GFD database for IP, useragent and referrer (at the beginning they are not in database, so r_i values are 0).
2. Computing score values r_i for all other parameters (using corresponding heuristic "rules").
3. Combining all score values for the final score *S* for the given click. If partial scores are r_1, r_2, …,r_n and experimentally chosen corresponding weights are w_1, w_2, …,w_n for a given click[3], then the final integrated score is expressed as

$$Score\ S = \frac{\prod_{i=1}^{n} wiri}{\prod_{i=1}^{n} wiri + \prod_{i=1}^{n} (1 - wiri)}$$

4. If click score in (3) is *S*=0, do not include any new records in GFD database (for IPs, useragents or referrers).
5. If click score in (3) is *S*>0, and there is no previous records for key click attributes (IP, useragent or referrer), create these new records with corresponding *S* values (and increase the count of the number of clicks for a given attribute).
6. If some attributes of the click with S>0 already exist in the GFD database, then adjust their historic scores by weighted averaging S for each attribute separately (IP, useragent, referrer, etc.)

7. Based on values in fraudulent database GFD for each key attribute, dynamically are modified "blocked lists" in the blocking database BD. Specific predictor is going in the "blocked list" if its score value in GFD database is above the given threshold T (not necessarily T=1, the threshold may be T=0.9).

7. Experimental results

We performed a large number of experimental analyses applying our CCFDP system to an available real world data set. In this paper we are explaining only the subset of these experiments which are related to a data fusion process. We confirmed that an extended click record, which includes context data, is a sound basis for better estimation of click traffic quality.

Our results in Figure 4 show significant increase in fraudulent activities in 2008 (61%) comparing click traffic in 2007 (44%). A large number of clicks in 2008are fraud (50%) with 3% of them would have been prevented using buffered context data and the dynamic blocking database.

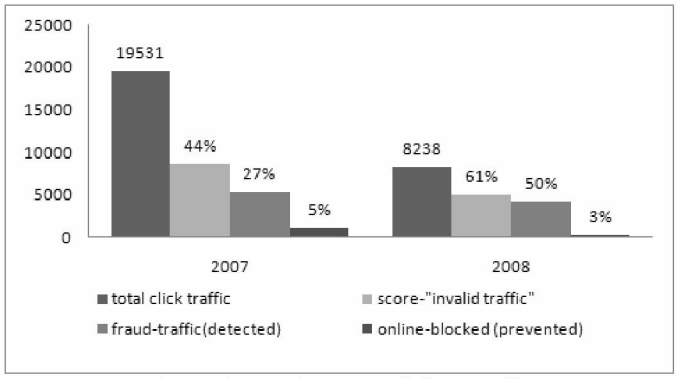

Figure 4: Total vs. Fraudulent traffic

The referrer is one of the key attributes in the detection of click fraud. The referrer parameter alone delivers little information about the traffic, but once fused with client side data, it provides useful insight. For example, if traffic from a particular referrer always refuses JavaScript, it may become highly suspicious and our scoring algorithm automatically increases fraud scores for this referrer. Table 2 represents details of traffic from the top referrers. Of course Google and its partner network are the top referring sites in our ad campaign. Our results show that about 40% of Google only traffic and 23% of Google partner network traffic are fraudulent. There is considerable portion of the traffic that do not disclose the referrer. These situations are identified and fraud scores are updated online identifying significant level of fraud traffic, especially the large number of fraudulent clicks that would have been prevented (about 17%).

The CCFDP system detects the country details in real-time and attached it to the sever side data. Even though hosting.com ad campaign is focused on US traffic we have seen only 28% traffic is from the US in 2007 even though it is 88% in the first quarter of 2008. The fusion process enables us to detect software clicks made mainly by robots whenever

Referrer	Total traffic	Score	Online blocked (prevented)	Fraud traffic (detected)	Fraud traffic %
Google	4869	2771	10	1966	40%
Google Syndication	2240	896	0	518	23%
Searchportal	2342	874	18	269	12%
Sedoparking	593	231	0	79	13%
Mywebsearch	621	265	0	93	15%
Yahoo	101	63	12	53	53%
MSN	53	25	0	17	32%
No referrer	4660	1895	786	1270	27%

Table 2: Referrer analysis

there is a missing client side entry for a server side entry in the fused data. In 2007, 16% traffic from the US is reported by software clicks and in 2008 this is at 34%. Clicks from some countries are becoming more suspicious when we are analyzing only software click traffic. In our case, these are the countries such as TR, GH, and VN and their parameters in the GFD are recorded, and they influenced all future clicks from these countries. Summary of these data is given in Table 3.

2007				2008			
Total traffic		SW clicks		Total traffic		SW clicks	
Country	% total traffic	Country	% software clicks	Country	% total traffic	Country	% software clicks
US	28%	US	16%	US	88%	US	34%
IN	9%	TR	2%	CA	5%	CA	3%
MA	5%	IN	1%	GB	3%	GB	2%
PH	3%	GH	1%	IL	1%	IL	1%
MY	2%	VN	1%	DE	0.9%	AU	0.4%
MX	2%	IL	1%	AU	0.6%	DE	0.4%
PK	2%	MA	1%	IN	0.1%	NL	0.1%

Table 3: Total traffic vs. Software traffic

Figure 5 shows the distribution of scores of clicks in 2007. Area 1 represents most of the valid clicks. In the fused record these correspond to the records with attributes which do not have presence in the fraudulent database and all key attributes satisfies the requirements defined in the algorithm to be a legitimate click. Area II shows the suspected clicks. These are records with the attributes present in the fraudulent database or attributes that exceeds certain threshold values. Area III includes online blocked traffic using blocking database BD, and double clicks click traffic using buffered context data. Online blocked traffic is identified as software

clicks based on the fused records with highly suspicious scores. For example, repetition of the same IP or referrer multiple times will block all future clicks with the given attributes. These scenarios are easy to detect in the fused data compared to analyzing either server side or client side data alone.

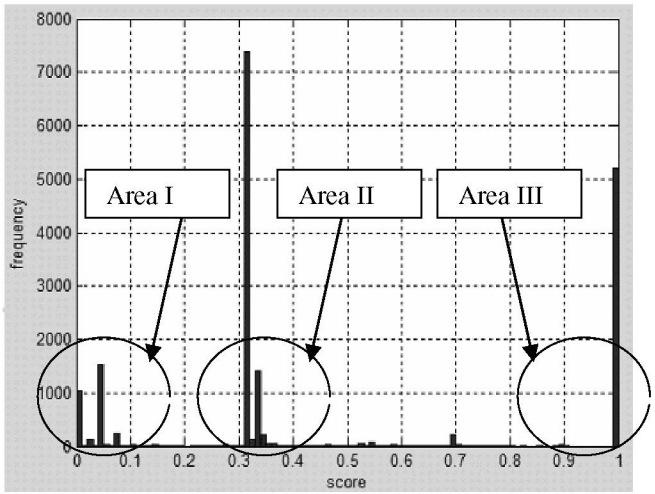

Figure 5: Distribution of scores for clicks in 2007

8. Conclusion

Most of the search marketing industries optimize traffic using its own binary paradigm into valid clicks and invalid clicks utilizing their own metrics. But with the exception of a small percentage of obvious genuine and robot traffic, the vast majority of clicks are cannot be determined to be valid or invalid. Rather, they are of a certain quality which depends on user intention.

In this paper we proposed a real-time detection and prevention system for click fraud CCFDP V1.0 using data fusion to enhance the description of each click, and to obtain better estimation of a click traffic quality. The system provides solutions to some unsolved problems in the available commercial solutions such as detecting software click fraud and preventing click fraud dynamically. The CCFDP system analyzes the detailed user activities on both the server side and client side collaboratively to better evaluate the quality of clicks. Extended click record includes also context data available in fraudulent and blocking databases. Our system analyze the extended click record and assigns a score S for each click based on the quality of the click. We have tested the system with data from an actual ad campaign in 2007 and 2008. Results show that a higher percentage of click fraud detection and prevention is present even with most popular search engines such as Google. Intensive experimental evaluation of the CCFDP system should be performed in the next period. Besides a fine tuning of some parameters, we have to address security and scalability aspects.

9. Acknowledgments

This research has been partially funded by National Science Foundation (NSF) under grant #0637563 and Kentucky Science and Technology Corp. (KSTC) under grant #KSTC-144-401-07-018.

10. References

1. Anupam V. et al., On the Security of Pay-per-Click and Other Web Advertising Schemes, Computer Networks, Vol. 31, Issues 11--16, 1999.
2. Auntheticlick Inc., www.authenticlick.net, Issaquah, Washington, USA.
3. CCFDP V 1.0, Final report for STTR-NSF project, #0637563, January 2008.
4. Clicklab, LLC, http://www.ClickLab.com, McLean, Virginia.
5. Gandhi M., et al., Badvertisements: Stealthy Click-Fraud with Unwitting Accessories, Anti-Phishing and Online Fraud, Part I Journal of Digital Forensic Practice, Volume 1, Special Issue 2, November, 2006.
6. Ge L., Kantardzic M., King D., CCFDP: Collaborative Click Fraud Detection and Prevention System, 18th International Conference on Computer Application in Industry and Engineering – CAINE 2005, Honolulu, November, 2005.
7. Ge L., Kantardzic M., King D., Collaborative Click Fraud Detection and Prevention System (CCFDP) Discovers Software-Based Click Fraud, IADIS e-Commerce 2005 Conference, Porto, Portugal, December 2005.
8. Hall D.L. et al., An Introduction to Multisensor Data Fusion, Proceedings of The IEEE, Vol. 85, No. 1, 1997.
9. Mahdian M., Theoretical challenges in the design of advertisement auctions, The Capital Area Theory Symposia, University of Maryland, Spring 2006.
10. Mahdian M. et al., Click fraud resistant methods for learning click-through rates, Proceedings of the First International Workshop on Internet and Network Economics, Lecture Notes in Computer Science 3828, 2005, pp. 34-45.
11. Metwally A., Agrawal, D., El Abbadi, A., Detectives: detecting coalition hit inflation attacks in advertising networks streams, Proceedings of the 16th international Conference on WWW, Alberta, Canada, 2007.
12. Metwally A. et al., Duplicate detection in click streams, Proceedings of the 14th international conference on World Wide Web, Japan, May 2005.
13. Reiter M. et al., Detecting Hit-Shaving in Click-Through Payment Schemes, Proceedings of the 3rd USENIX Workshop on Electronic Commerce, Massachusetts, USA 1998.
14. Tuzhilin A., The Lane's Gifts v. Google Report, 2006.
15. Varshney P. K., Multisensor data fusion, Proceedings of the 13th international Conference on industrial and Engineering Applications of Artificial intelligence and Expert Systems, New Orleans, Louisiana, USA, 2000.
16. Wang, H., Lee, M. K., and Wang, C. Consumer privacy concerns about Internet marketing. Communications of the ACM Vol. 41, Issue 3, 1998
17. Web Traffic Intelligence, Inc., www.clickalyzer.com, Thermopolis, WY., USA.