Information Systems 36 (2011) 1105–1123

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

CATCH: A detecting algorithm for coalition attacks of hit inflation in internet advertising

Chulyun Kim [b], Hui Miao [a], Kyuseok Shim [a,*]

[a] School of Electrical Engineering and Computer Science, Seoul National University, Kwanak, P.O. Box 34, Seoul 151-742, Republic of Korea

[b] Department of Software Design and Management, Kyungwon University, Bokjeong-Dong, Sujeong-Gu, Seongnam, Gyeonggi-Do 461-701, Republic of Korea

ARTICLE INFO

Available online 6 May 2011

Keywords:
Click fraud
Hit inflation
Coalition attack
Internet advertising
Graph mining
Data mining

ABSTRACT

As the Internet flourishes, online advertising becomes essential for marketing campaigns for business applications. To perform a marketing campaign, advertisers provide their advertisements to Internet publishers and commissions are paid to the publishers of the advertisements based on the clicks made for the posted advertisements or the purchases of the products of which advertisements posted. Since the payment given to a publisher is proportional to the amount of clicks received for the advertisements posted by the publisher, dishonest publishers are motivated to inflate the number of clicks on the advertisements hosted on their web sites. Since the click frauds are critical for online advertising to be reliable, the online advertisers make the efforts to prevent them effectively. However, the methods used for click frauds are also becoming more complex and sophisticated.

In this paper, we study the problem of detecting coalition attacks of click frauds. The coalition attacks of click fraud is one of the latest sophisticated techniques utilized for click frauds because the fraudsters can obtain not only more gain but also less probability of being detected by joining a coalition. We introduce new definitions for the coalition and propose the novel algorithm called CATCH to find such coalitions. Extensive experiments with synthetic and real-life data sets confirm that our notion of coalition allows us to detect coalitions much more effectively than that of previous work.

1. Introduction

The World Wide Web becomes the most useful source of information for our daily life and thus people spend a fair amount of their time to find useful information in the Internet. Online advertising becomes essential and utilizes the Internet for developing a marketing campaign to attract the Internet users. In the online advertising, potential customers can be easily directed to a particular advertisement which provides detailed information and ordering methods of the advertised products. This indicates that the Internet advertising is one of the most important business models of the Internet industry.

To perform a marketing campaign, advertisers provide their advertisements with a budget to advertising commissioners, such as Google AdSense, which are brokers between advertisers and Internet publishers. The Internet publishers make contracts with commissioners for displaying the advertisements on their web sites and then commissions are paid to publishers for visitors' clicking their advertisements or purchasing the advertised products via publishers' web sites. For Internet advertising campaigns, there are several payment policies such as pay-per-impression, pay-per-click and pay-per-sale [16,7,6,15,17]. Among these options, the pay-per-click is

[*] Corresponding author.
E-mail addresses: cykim@kyungwon.ac.kr (C. Kim), huim@kdd.snu.ac.kr (H. Miao), shim@ee.snu.ac.kr (K. Shim).

0306-4379/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2011.04.001

the most popular policy and we will assume the pay-per-click policy. It is easy to extend our technique to other policies.

As the payment given to a publisher is proportional to the number of clicks for the advertisements posted by the publisher, a dishonest publisher (i.e., fraudster) may cheat by inflating the number of clicks for the advertisements hosted on his web sites and driving up the bills sent to advertisers. This method is known as *click inflation* or *click fraud* [6]. Click fraud will hinder the reliability of online advertising system. In a short-term, invalid response to the advertisements inflated by click fraud may increase the revenue of online advertising. However, it diminishes the effectiveness of advertising in the Internet and the market for online advertising will eventually contract in a long-term. Furthermore, it may result in expensive litigations from unsatisfied advertisers as the recent news articles on the click fraud lawsuits indicate. For instance, in March 2006, Google agreed to the $90 Million settlement in the class action lawsuit over click fraud filed by Lane's Gifts and Collectibles [2,3]. In July 2005, Yahoo also settled a class action lawsuit against click fraud [4], where Checkmate Strategic Group alleged that Yahoo did not work enough to prevent click fraud and Yahoo agreed to pay the plaintiffs the amount of $4.95 million. Recently, in July 2009, Facebook was sued for click fraud by RootZoo in the lawsuit [1]. Therefore, it is very important for the commissioners to actively work on preventing click fraud to convince their advertisers the fairness of their accounting practices.

An elementary attack is for a publisher to repeatedly click advertisements hosted by the publisher himself. However, the attack can be simply blocked by removing duplicate clicks within a short period of time from the same visitor [6,17,20]. Another method is to use a script to make every visitor automatically click advertisements [6]. A fraudster can use a pair of pages to sophisticatedly enforce the visitor to click the posted advertisements. Only one of the pair is informed to the commissioner and the other is not reported to the commissioner. When a visitor browses the informed page to the commissioner, the web page does nothing to the visitor so that commissioner cannot notice the automatic fraud clicking when the commissioner monitors the informed page. However, if a visitor surfs the unknown page to the commissioner, the visitor is redirected to the informed page of the pair and clicking the advertisements in the redirected page is performed automatically [6,20]. In another attacking method, a publisher makes fake visitors and inflates the click counts by forging their identifications [23].

The attack methods mentioned previously are performed by a single fraudster. If a fraudster excessively reuses his own resources to generate more attacks, the possibility to be detected increases quickly. In order not to be detected easily by using these attacks, the fraudster needs to utilize more physical resources. However, it is not easy to increase the amount of physical resources used by a fraudster blindly due to high expenses. To overcome the limitation, fraudsters frequently form a group and launch a coalition attack by sharing their resources together (i.e., machines or IPs) [21]. By joining a coalition, a fraudster expects not only more gain but also less probability to be detected. Thus, coalition attacks are getting popular to inflate the click counts of advertisements without the high cost of increasing the resources.

Example 1. Let us consider the case in Fig. 1(a) where advertisements published by Alice are clicked by a visitor v_1 who is actually herself. Note that a visitor can give a publisher at most a certain number of clicks in a period without being detected and we assume that two clicks are the maximum here. However, when Alice joins a coalition such as in Fig. 1(b), she can get more clicks in each period without installing more resources. The coalition consists of three fraudsters who share their machines and produce the clicks for each other. Thus the revenue for each fraudster grows without increasing the resources of each fraudster.

In [21], Metwally et al. pioneered the work on fraud click detection methods and proposed the detection method named DETECTIVES for the coalition attacks. The proposed method is based on the similarity between visitors of publishers. The similarity between two publishers is defined by the Jaccard coefficient [14] between the sets of their visitors. The similarities are used to build a graph where each publisher becomes a vertex and an edge is introduced between a pair of publishers if the similarity between the two publishers is greater than a threshold. Then, each complete subgraph (i.e., clique) in the graph is defined as a coalition.

Example 2. In Fig. 1(b), Alice, Bruce and Charles get clicks from the same visitors, v_1, v_2 and v_3. Thus the Jaccard coefficient between every pair of publishers is exactly 1. Given a minimum threshold 0.8, the transformed graph is a triangle with three vertices, Alice, Bruce and Charles. Since it is a complete graph, they are considered as a coalition by DETECTIVES.

Although the definition of a coalition introduced in [21] is interesting, we observed several limitations and problems in this definition as follows:

- Visitors may click multiple advertisements published by the same publisher. Thus, clicking once and clicking

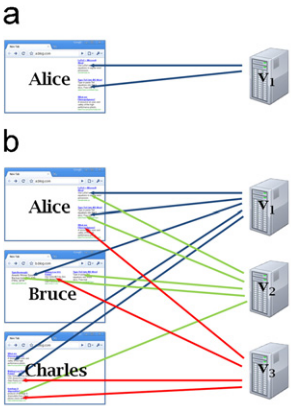

Fig. 1. Example of coalition attack. (a) Attack by a single publisher. (b) Attack by a caolition.