11-2013

Mining fraudulent patterns in online advertising

Richard J. OENTARYO
Singapore Management University, roentaryo@smu.edu.sg

Ee-peng LIM
Singapore Management University, eplim@smu.edu.sg

Mining Fraudulent Patterns in Online Advertising

Richard J. Oentaryo and Ee-Peng Lim
Living Analytics Research Centre
Singapore Management University
80 Stamford Road, Singapore 178902

1 Introduction

Advances in web technologies have rendered online advertising as an effective means for small and large businesses to target different market segments on the fly. Online advertising is a huge industry. According to Gartner Inc., worldwide online advertising revenue is projected to hit $11.4 billion in 2013, up from $9.6 billion in 2012 [1]. Global revenue will also reach $24.5 billion in 2016, with online advertising creating opportunities for app developers, advertising networks, and service providers in various regions. An online advertising ecosystem is typically coordinated by an advertising *commissioner*, acting as a broker between advertisers and content publishers. An *advertiser* plans a budget, provides some commissioner(s) with ads, and agrees on commission for customer actions (e.g., clicking ad, bidding in an auction, etc.). Meanwhile, *content publishers* contract with each commissioner to post ads on their websites, and receive commission based on the traffic they drive to the advertisers.

In such a business ecosystem, building and maintaining trusts amongst the advertisers, commissioners, and publishers has been a critical issue. For example, it is important that an advertiser pays commission to its publishers (via commissioners) accordingly, based on the agreed scheme. Conversely, it is imperative for a publisher to provide traffic reports based on genuine interest of their site visitors in the ad contents. Commissioners also need to exercise a fair accounting practice, ensuring that the advertisers' dollars are well spent and that (honest) publishers get their commission appropriately. Unfortunately, the real-world situations are not ideal. In pay-per-click advertising, for instance, an automated script or program may imitate a legitimate visitor when clicking an ad, so as to earn unjustified revenue. This issue, also known as *click fraud*, has over the years ailed many advertising businesses. Click fraud degrades the reliability of online advertising, and leads to substantial monetary loss in the long run [2, 3].

Detecting click fraud is a challenging task, not only because fraudulent traffic often mimics that of legitimate site visitors but also due to the evolving fraud techniques and strategies. There

is a need for a fraud detection system that can identify malicious publishers from large volume of click traffic in an efficient and robust manner. Our own endeavor consists of studying click traffic data provided by our Singapore-based industry partner, an advertising commissioner that operates a global mobile advertising network with millions of consumers accessing internet contents on mobile phones/devices. In Q1 2012, the network had delivered more than 45 billion ad banners to over 10,000 publisher sites, reaching 300 million unique users each month. Of particular interest is how we can determine whether a publisher is fraudulent, based on click traffics generated by the visitors of its sites. Note that we aim at detecting fraud at *publisher level* (rather than visitor level), and we assume that all advertisers are trustworthy.

Currently, our industry partner uses an in-house developed mechanism to identify fraudulent publishers semi-automatically. However, such mechanism is limited and may not be easily adapted to new fraudulent patterns. We conjecture that *data mining* approaches [3, 4] can be used to improve the existing mechanism, while reducing the efforts for manual interventions. Data mining offers a means to automatically identify fraudulent publishers and discover the latent signatures that differentiate them from legitimate (normal) publishers. It is worth noting that the click traffic data supplied by our partner are unusual and challenging, involving heterogeneous numerical and categorical variables, noisy or missing patterns, and highly skewed distribution of the (fraudulent vs. legitimate) publishers. In this short paper, we demonstrate how data mining approaches can be effectively employed in real-world, practical setting. Several key insights on fraudulent patterns are also discussed.

2 Related Works

Broadly, click fraud detection approaches fall into three categories: *manual*, *cryptographic*, and *data analytic* approaches [2]. In the first approach, a set of tools is utilized to identify fraud depending on the behavior of advertisements in the sites, e.g., how click through rate (CTR)—or the ratio of clicks to impressions—of an ad deviates from the system-wide norm. (This is similar to the approach currently adopted by our partner). However, this method can only detect simple types of fraud (e.g., traffics with low-quality CTR), and it is fairly easy for fraudsters to intrude, sample, or mimic the system-wide metrics. In the cryptographic approach, the site visitors were required to cooperate by providing their information to the publishers. Unfortunately, this often demands changing the advertising business model, making it unscalable and non-transparent to the visitors. It also leads to privacy issues, as the commissioner needs to identify every visitor.

Finally, the data analytic approaches generally utilize data mining and machine learning methods, aiming at finding key insights on fraud that are statistically reliable, unknown before, and actionable from data [3, 4]. These methods are non-intrusive to the advertising business model, and can complement the classic and cryptographic methods. Moreover, they can operate on anonymized data without intruding the visitors' privacy, yet can reveal specific patterns that characterize fraudulent traffics. Data analysis can be performed at individual publisher level (assuming publishers are independent from one another), or at group level (whereby publishers may form a coalition to attack) [2]. The latter is a more sophisticated form of fraud, allowing fraudsters to not only gain more with less resource, but also reduce the risk of getting detected. Studies on our partner's data, however, reveal little evidence of such coalition attack (see Section 3). Our work thus focuses on identifying fraud at the individual publisher level.

3 Problem

3.1 Datasets

The datasets supplied by our partner comprise a *publisher database* and a *click database*, which record the publishers' account profiles and their click activities, respectively. The fields of the two databases are listed in Table 1 and Table 2. Each publisher has three possible statuses:

- *OK*: Publishers who are deemed as having healthy traffic
- *Observation*: Publishers who may have just started their traffics, or their traffic statistics deviates from system wide average. There is no conclusive stand with these publishers yet.
- *Fraud*: Publishers who are deemed as fraudulent with clear proof. Their accounts are suspended and their earnings will not be paid.

Accordingly, given the profile and click activities of a publisher, our research objective is to conduct exploratory data analysis and perform data mining in order to learn the inherent traits that distinguish malicious (i.e., *Fraud*) publishers from normal (i.e., *OK* or *Observation*) ones.

We note that most of the fields in the publisher and click databases have been anonymized for privacy protection. The count statistics of the publishers and clicks are summarized in Table 3. The publishers and clicks data supplied by our partner consist of three sets taken from different time periods. Each click dataset captures the click traffics in a 3-day period, while each publisher dataset records the publishers who received at least one click during that period. Also, the *Fraud* and *Observation* publishers form very small portion of the population in comparison to the OK publishers, implying a highly-imbalanced class distribution that poses a major challenge.

Table 1. Fields in the publisher database

Field	Description
publisherid	Identifier of a publisher (anonymized)
bankaccount	Bank account of a publisher (field may be empty)
address	Mailing address of a publisher (encrypted; field may be empty)
status	Label of a publisher, i.e., *OK*, *Observation*, or *Fraud*

Table 2. Fields in the click database

Field	Description
clickid	Identifier of a click (anonymized)
numericip	Public IP address of a visitor (anonymized)
deviceua	Phone model used by a visitor (anonymized)
publisherid	Identifier of a publisher (anonymized)
campaignid	Identifier of an ad campaign (anonymized)
country	Country where the site visitor comes from
clicktime	Click timestamp (in *yyyy-mm-dd* format)
referredurl	URL where ad banners are clicked (anonymized; field may be empty)
channel	Publisher's channel type (e.g., adult sites, entertainment, etc.)

Table 3. Statistics of the publishers and clicks data at different time periods

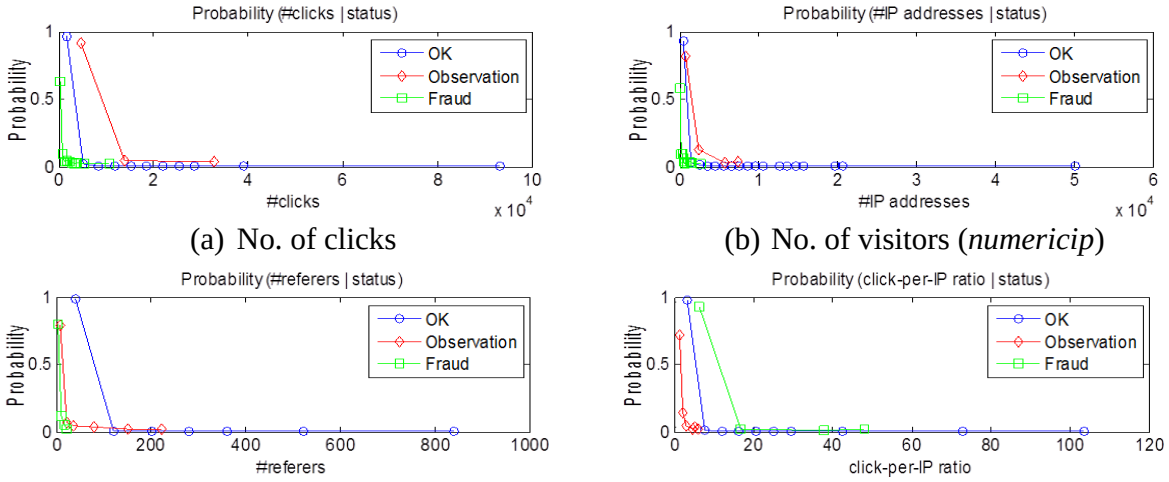
Period	No. of clicks	No. of publishers		
		Fraud	Observation	OK
09-11 Feb 2012	3,173,834	72 (2.34%)	80 (2.60%)	2,929 (95.06%)
23-24 Feb 2012	2,689,005	85 (2.77%)	84 (2.74%)	2,895 (94.49%)
08-10 Mar 2012	2,598,815	82 (2.73%)	71 (2.37%)	2,847 (94.90%)

3.2 Exploratory Analysis

We first analyzed the basic statistics of the publishers, derived by grouping the entries in the click database by publisher. For each publisher, we computed the probability distribution (i.e., normalized frequency) of the number of clicks, number of visitors (*numericip*), number of referrers (*referredurl*), and the ratio of clicks over visitors. Figure 1 shows the four distributions in the 09-11 Feb 2012 data, grouped by the publisher status (i.e., *OK*, *Observation*, and *Fraud*). Interestingly, Figure 1(a) suggests that the *Fraud* publishers have lower click probability than the *OK* publishers (which may be attributed to the fact that the traffic of a publisher will be blocked as soon as it is deemed as fraudulent). Figure 1(b) shows a similar observation for the visitors.

Meanwhile, Figure 1(c) indicates that, for *referrerurl* distribution, the *Fraud* publishers are rather different from the *OK* ones, and have low probability similar to the *Observation* publishers.

Further investigation revealed that many *Fraud* publishers have empty *referredurl* fields. Thus, the signatures—hereafter called *features*—derived from *referredurl* can be good indicators for fraud. Lastly, the distribution plot in Figure 1(d) shows that *Fraud* publishers have higher click per visitor ratio than the other groups, suggesting that the former focus on more efficient use of resources to inflate the clicks. This motivates us to explore other ratio-based features (e.g., click per referrer ratio, click per country ratio, etc.). Further details shall be given in Section 4.



(a) No. of clicks (b) No. of visitors (*numericip*)

(c) No. of referrers (*referrredurl*) (d) Click per visitor ratio

Figure 1. Distribution of the click data (09-11 February 2012)

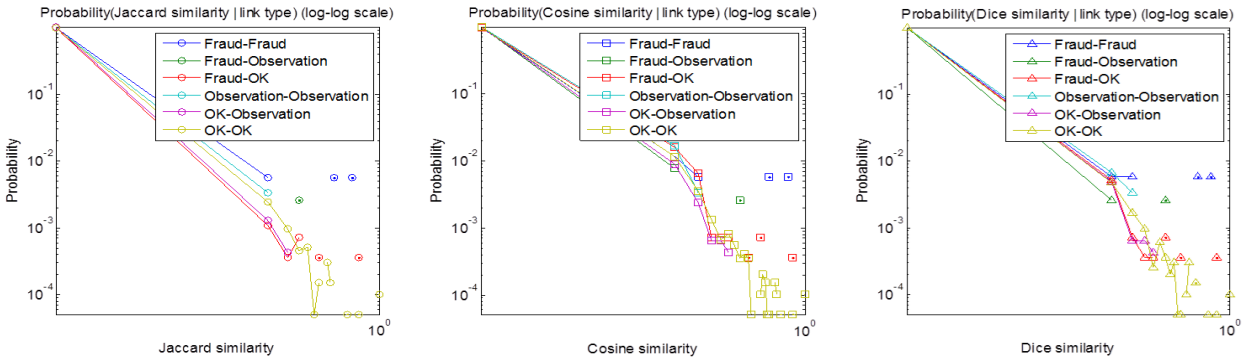


Figure 2. Distribution of similarity among different publisher groups (09-11 February 2012)

To investigate whether fraudsters form a coalition to perform attacks, we also constructed a *similarity graph*, whereby each node represents a publisher and an edge is created when two publishers have at least one common visitor. The similarity score between any two publishers are then computed using the Jaccard, cosine, and Dice similarity indices [5]. Figure 2 presents the distribution of the similarity scores between different publisher types (in logarithmic scale). As

seen, the similarity scores are generally low, suggesting that publishers unlikely share resources or know each other. Although here the similarity between a pair of *Fraud* publishers tends to be higher than other pairs, we found that the differences are (statistically) insignificant. We can thus conclude that the likelihood of fraudsters forming coalition is small.

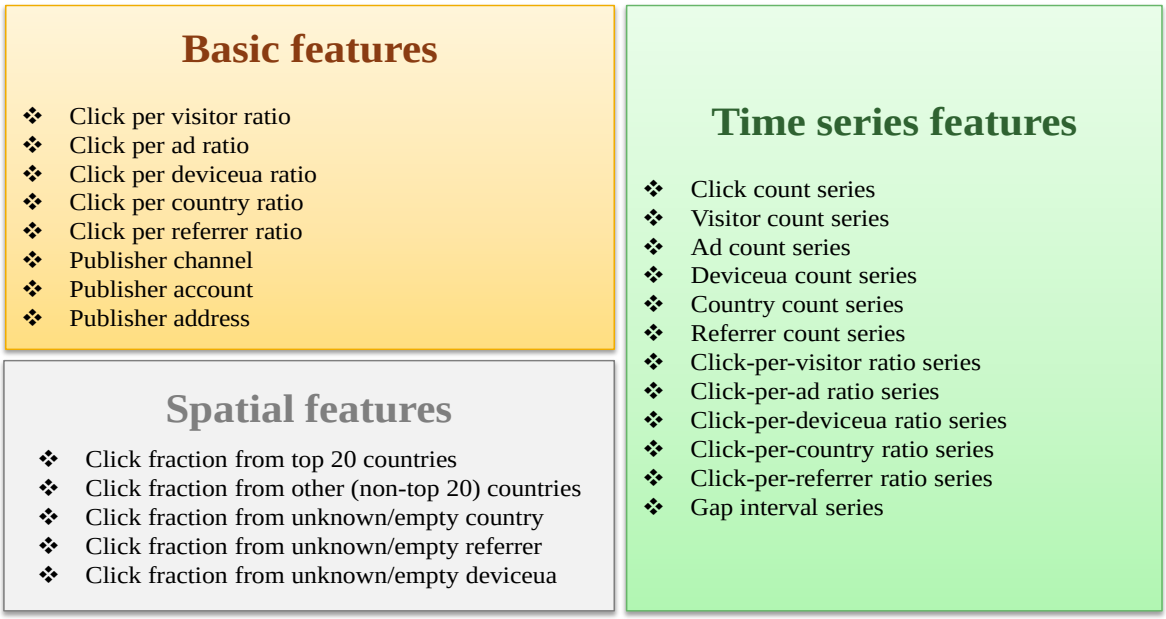


Figure 3. List of features extracted from the publisher and click databases

4 Method

4.1 Feature Extraction

Based on our analysis in Section 3.2, we extracted several basic ratio features out of the click and publisher databases, as listed in Figure 3. These features do not account for detailed temporal and spatial aspects of the traffics. With these basic features alone, however, the detection results were found inadequate. To improve the results, we performed fine-grained analysis on the spatiotemporal aspects of the click traffics, leading to two new types of features: *spatial* and *time series*. Figure 3 also lists the spatial and time series features considered in our work.

For the spatial features, we computed for each publisher the fraction of clicks from different *country*, *referredurl*, and *deviceua*. For *country*, we considered top 20 countries in terms of their total number of clicks. We further computed the click fraction from unknown (empty) country, and collated the click fraction from the remaining (non-top 20) countries. Finally, we calculated the click fractions from unknown/empty *referredurl* and *deviceua*, which we believe correlate strongly with the probability of a publisher being fraudulent.

For the time series features, we broke down the 3-day span of each dataset into windows of *1 minute* long, and tracked several values of interest (e.g., number of clicks per minute, number of visitors per minute, etc.). This resulted in a time series vector of length 4,320 (= 3 days in minutes) for each value type. We also experimented with longer time interval (e.g., 1 hour and 1 day), but the results were significantly worse. This suggests that fraudulent clicks tend to be *bursty*, taking place repeatedly at a very short time period.

Next, we computed several statistical features aggregating the time series values over the 3-day period: *count*, *mean*, *maximum*, *sum*, *sum of square*, and *standard deviation*. To capture trending patterns in the time series, we also derived the same set of statistical features for the *positive* and *negative* gradients of the series. The gradient at time window *t*, *g(t)*, is computed as a finite difference between two consecutive series values *v(t)* and *v(t-1)*, i.e., $g(t) = v(t) - v(t-1)$. The gradient is positive if $g(t) > 0$, and negative if $g(t) < 0$. Further, we considered the *gap interval* time series, which refers to the span between the timestamps of two consecutive clicks. Adding all the basic, spatial, and time series features, we have a total of 257 features. Not all the features were useful though, and feature ranking can be carried out to simplify the model.

4.2 Classification Algorithms

To identify *Fraud* publishers, we employed two types of algorithms: *single classifier*, and *ensemble classifier*. The latter is about combining multiple hypotheses from many weak (single) classifiers in attempt to form a *strong classifier* that should give better and more accurate results. In highly-imbalanced classification tasks, the ensemble models are also expected to yield better performances. For the single classifiers, we explored various linear models, and found that logistic regression [6] gave good performances. Logistic regression is an extension of linear regression [4], which models the relationship between one or more prediction variables and a binary outcome variable. As our ensemble classifier, we experimented with decision tree-based models and obtained the best results from the extremely randomized trees (ERT) [7]. ERT creates a collection of decision trees, each modeling a subset of data space. It also includes an extra randomization step when choosing the node-splitting threshold [7].

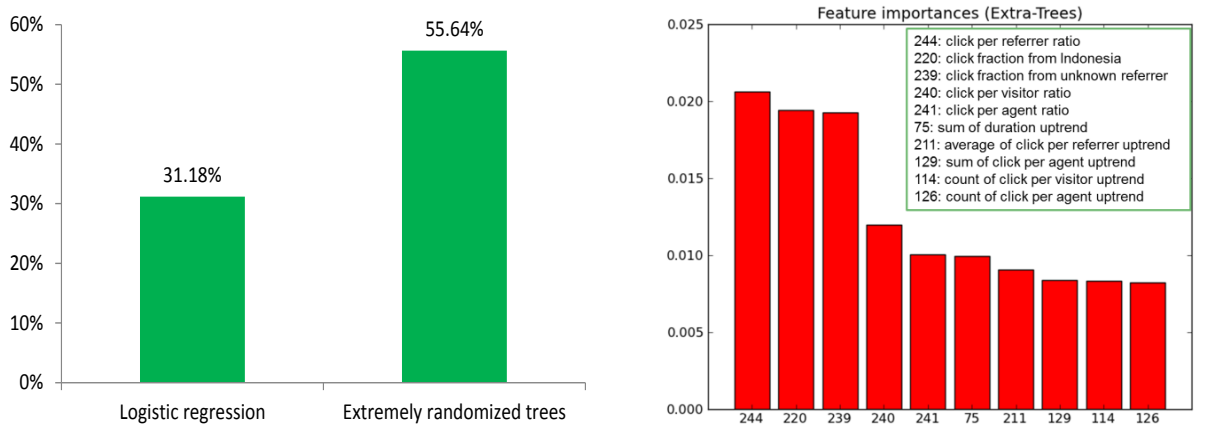
5 Results and Insights

5.1 Detection Performances

To evaluate the performance of our classifiers, we considered the *average precision* (AP) metric [8], which favors methods that are capable of ranking few useful items ahead of the rest.

Such criterion is suitable for detecting rare instances such as fraudulent publishers [8]. Figure 4(a) shows the prediction results of the classifiers. It can be observed that the ERT produced much higher AP than the logistic regression, which exemplifies the benefits of ensemble classifiers over single classifiers. It is worth noting that attaining over 50% AP score is considered very satisfactory, given the highly-skewed distribution of the publisher status.

A plausible explanation for the superiority of ERT is that ensemble model exploits the local different behavior of the weak learners to enhance the overall system accuracy. Moreover, using mixture of classifiers (instead of choosing just one) can reduce the risk of selecting a poorly performing classifier, thus reducing the output variance. Also key to a successful ensemble model is the diversity of its base learners. In the case of ERT, the additional randomization step helps promote diversity, yielding faster ensemble construction as well as reduced output variance.



(a) Classifier performances (b) Top 10 features produced by ERT

Figure 4. Classifier performances on fraud detection task

5.2 Key Indicators

In Figure 4(b), we show the top 10 features discovered and ranked by our best classifier (i.e., the ERT), whereby feature importance was estimated from the expected fraction of the samples the tree components contribute to [7]. These features reveal several interesting observations. First, the top and third-ranked features suggest that *Fraud* publishers tend to have empty (unknown) referrer, which results in high click per referrer ratio and high click fraction of empty *referredurl* field. Secondly, high click traffic from high-risk countries, such as Indonesia, is good indicator for fraudulent behaviors. We can also see that the ratio features and their related time series features, in particular the sum, count, and positive gradient features, may be indicative of

fraudulent behaviors. This suggests that fraudulent traffic tends to exhibit an uptrend pattern in a short time period. All in all, we can conclude that fine-grained analysis of the spatial and temporal aspects of the click traffics are important for accurate fraud detection.

6 Conclusion

Click fraud is one of the most daunting problems in online advertising, and is a central issue of trust between advertisers, advertising commissioners, and content publishers. In this paper, we analyze a real-world online advertising dataset, and present data mining approaches to extract fraudulent publisher patterns. In our experiments, we demonstrate how granular analysis of time series is crucial for deriving informative features for fraudulent publishers. We also show that spatial features such as click fraction from high-risk countries may be used as indicators for fraud. Finally, the combination of ensemble classifiers and feature ranking has shown good results in the face of highly-skewed class distribution, noisy/missing patterns, and mixed variable types, thus exemplifying its potentials in complex fraud identification tasks.

Acknowledgment

This work is supported by National Research Foundation under its International Research Centre @ Singapore Funding Initiative, and administered by the IDM Programme Office.

References

[1] Gartner Inc., "Gartner says worldwide mobile advertising revenue to reach $11.4 billion in 2013," Available at: http://www.gartner.com/newsroom/id/23062152013.

[2] A. Metwally, *et al.*, "DETECTIVES: DETEcting Coalition hiT Inflation attacks in adVertising nEtworks Streams," in *Proceedings of the ACM International World Wide Web Conference*, Banff, Alberta, Canada, 2007, pp. 241-250.

[3] C. Phua, *et al.*, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010.

[4] J.-W. Han, *et al.*, *Data mining: Concepts and techniques*: Morgan Kaufmann, 2006.

[5] B. Markines, *et al.*, "Evaluating similarity measures for emergent semantics of social tagging," in *Proceedings of the ACM International World Wide Web Conference*, Madrid, Spain, 2009, pp. 641-650.

[6] R.-E. Fan, *et al.*, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871-1874, 2008.

[7] P. Geurts, *et al.*, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3-42, 2006.

[8] M. Zhu, "Recall, precision, and average precision," Working Paper 2004-09, University of Waterloo, 2004.