Exhibit 33

Share Report Abuse Next Blog» Create Blog Sign In

The Official Google Blog

Insights from Googlers into our products, technology, and the Google culture.

Using data to help prevent fraud

3/18/2008 11:52:00 AM
Posted by Shuman Ghosemajumder, Business Product Manager for Trust & Safety

We recently began a series of posts on how we harness the power of data. Earlier we told you how data has been critical to the advancement of search technology. Then we shared how we use log data to help make Google products safer for users. This post is the newest in the series. -Ed.

Protecting our advertisers against click fraud is a lot like solving a crime: the more clues we have, the better we can determine which clicks to mark as invalid, so advertisers are not charged for them.

As we've mentioned before, our Ad Traffic Quality team built, and is constantly adding to, our three-stage system for detecting invalid clicks. The three stages are: (1) proactive real-time filters, (2) proactive offline analysis, and (3) reactive investigations.

So how do we use logs information for click fraud detection? Our logs are where we get the clues for the detective work. Logs provide us with the repository of data which are used to detect patterns, anomalous behavior, and other signals indicative of click fraud.

Millions of users click on AdWords ads every day. Every single one of those clicks -- and the even more numerous impressions associated with them -- is analyzed by our filters (stage 1), which operate in real-time. This stage certainly utilizes our logs data, but it is stages 2 and 3 which rely even more heavily on deeper analysis of the data in our logs. For example, in stage 2, our team pores over the millions of impressions and clicks -- as well as conversions -- over a longer time period. In combing through all this information, our team is looking for unusual behavior in hundreds of different data points.

IP addresses of computers clicking on ads are very useful data points. A simple use of IP addresses is determining the source location for traffic. That is, for a given publisher or advertiser, where are their clicks coming from? Are they all coming from one country or city? Is that normal for an ad of this type? Although we don't use this information to identify individuals, we look at these in aggregate and study patterns. This information is imperfect, but by analyzing a large volume of this data it is very helpful in helping to prevent fraud. For example, examining an IP address usually tells us which ISP that person is using. It is easy for people on most home Internet connections to get a new IP address by simply rebooting their DSL or cable modem. However, that new IP address will still be registered to their ISP, so additional ad clicks from that machine will still have something in common. Seeing an abnormally high number of clicks on a single publisher from the same ISP isn't necessarily proof of fraud, but it does look suspicious and raises a flag for us to investigate. Other information contained in our logs, such as the browser type and operating system of machines associated with ad clicks, are analyzed in similar ways.

These data points are just a few examples of hundreds of different factors we take into account in click fraud detection. Without this information, and enough of it to identify fraud attempted over a longer time period, it would be extremely difficult to detect invalid clicks with a high degree of confidence, and proactively create filters that help optimize advertiser ROI. Of course, we don't need this information forever; last year we started anonymizing server