

User-centric clickprints on the Web

Balaji Padmanabhan

The Wharton School

University of Pennsylvania

balaji@wharton.upenn.edu

<http://opim.wharton.upenn.edu/~balaji>

Yinghui Yang

Graduate School of Management

University of California, Davis

yyang@ucdavis.edu

<http://faculty.gsm.ucdavis.edu/~yyang>

As part of this research we address the question of whether humans have unique signatures - or clickprints - when they browse the Web. The importance of being able to answer this can be significant given applications to electronic commerce in general and in particular online fraud detection, a major problem in electronic commerce costing the economy billions of dollars annually. Solutions to this question can help address the following two problems – detection and verification. Detection is the ability to determine who an online user is when the user has not yet identified himself or herself in any way to a merchant. When a user has identified himself or herself, verification is the ability to check that the revealed identity is correct.

We distinguish between site-centric and user-centric clickprints in the following manner. If a specific site can determine unique clickprints based on user-activity at that site then site-centric clickprints exist. On the other hand, if user sessions are viewed at the level of the client – i.e. individual browsing patterns across sites – then clickprints in these data are user-centric clickprints. In our research we are examining both site-centric and user-centric clickprints. The

framework to address these is the same, which we describe below. However in this work we will empirically examine whether user-centric clickprints exist.

The unique clickprint problem can be formulated as the following problem.

(A1) Given a feature construction mechanism F that constructs features from user sessions, a classifier C , and a Web browsing data set D , are there unique clickprints implicit in F and C that exist in D ?

Solving problem (A1) may appear trivial, since all that is needed is to construct features, build a classifier and check its accuracy. However this seems trivial only because it ignores perhaps the most interesting aspect of the problem – the unit of analysis at which the search is done. Specifically, it is common in the literature to treat a single session at a Web site as the unit of analysis and build models that predict purchase in the session, or as is the case here, models that predict the user corresponding to the session. This approach implicitly assumes that an online signature is something that manifests itself in the chosen unit of analysis, here a single Web session. In practice this may often not be the case. Our experiments show that the accuracies of predicting a user based on treating a single session as the unit of analysis is often not very high.

However over time – across many sessions - more information is revealed. In other words, we do not assume that *every* Web session has information to uniquely identify individuals – in fact the examples offered above suggest otherwise. However there may be some level of aggregation, agg , such that every agg sessions may have enough information to uniquely distinguish individuals. Solving the unique clickprint determination problem is then the same as determining this level of aggregation agg , which leads to the following problem statement:

(A2) Given a feature construction mechanism F , classifier C , and a Web browsing dataset D , find the smallest level of aggregation agg at which unique clickprints, implicit in F and C exist in D .

The importance of aggregation also has to do with how features are constructed (the procedure F) for groups of sessions, as opposed to single sessions. A natural approach is to determine a set of variables that can be constructed for a single session, and to then use groups of

sessions to learn the distributions of these variables. For example, After, say, ten sessions, we may determine that the distribution is $N(4, 1.2)$, and this may be sufficiently different from all other distributions. While higher levels of aggregation may also provide unique clickprints, we seek the smallest level of aggregation at which we can do so. Finally note that it is certainly possible that given some set of variables, the individual and joint distributions of these, estimated at any level of aggregation, may still be inadequate to uniquely distinguish individuals – perhaps because there are a very large number of individuals or perhaps because unique clickprints just do not exist. In such cases we will require that our solution procedure indicate that no level of aggregation is adequate to uniquely distinguish the users.

We empirically estimate the minimum level of *agg* for several leading online sites based on Web browsing data of 50,000 users over a period of one year. These results show evidence of site-centric clickprints. We find that even with a very basic set of features unique clickprints may exist when the number of users is low. When the number of users mixed was larger the aggregation required is significantly higher, suggesting that it is more difficult to distinguish users as the size of the crowd increases. However the numbers in these experiments may well only represent an upper bound since in practice online merchants can construct a far richer set of features about users. Such a set may result in learning unique clickprints with even fewer sessions than what is estimated here.

We are currently working on user-centric data with a richer set of behavioral features (user-centric Web browsing data captures a lot more features of a user's browsing activities such as the most visited Web sites, the starting site, the ending site etc.). These features can correspond to the more detailed site-centric browsing behavior if a Web site captures information about the pages a user visits at its site. We expect that lower number of sessions will be needed to identify user-centric clickprints and that this may also scale to a larger number of users. In this paper we will report detailed empirical

results that evaluate this hypothesis and attempt to answer the question of whether or when user-centric clickprints exist in Web browsing data.