Digital Investigation 22 (2017) S106–S114

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2017 USA — Proceedings of the Seventeenth Annual DFRWS USA

Analyzing user-event data using score-based likelihood ratios with marked point processes

Christopher Galbraith [a, *], Padhraic Smyth [b]

[a] Department of Statistics, University of California, Irvine, Bren Hall 2019, Irvine, CA 92697, USA
[b] Department of Computer Science, University of California, Irvine, Bren Hall 3019, Irvine, CA 92697, USA

Keywords:
Digital forensics
Likelihood ratio
Marked point process
Event data
Density estimation
Time series

A B S T R A C T

In this paper we investigate the application of score-based likelihood ratio techniques to the problem of detecting whether two time-stamped event streams were generated by the same source or by two different sources. We develop score functions for event data streams by building on ideas from the statistical modeling of marked point processes, focusing in particular on the coefficient of segregation and mingling index. The methodology is applied to a data set consisting of logs of computer activity over a 7-day period from 28 different individuals. Experimental results on known same-source and known different-source data sets indicate that the proposed scores have significant discriminative power in this context. The paper concludes with a discussion of the potential benefits and challenges that may arise from the application of statistical analysis to user-event data in digital forensics.

Introduction

Event histories recording user activities are routinely logged on devices such as computers and mobile phones. For a particular user these logs typically consist of a list of events where each event consists of a timestamp and some metadata associated with the event. For example, with popular Web browsers (such as Chrome, Internet Explorer, and Firefox) a variety of events related to user actions are logged on the local device. Examples of such actions include content downloads, URL requests, search history, and so on. Log files of user activity are also often accessible via cloud storage, for example for user events related to email activity, social media activity (such as Facebook and Twitter), and remote file storage and editing.

As digital devices become more prevalent, these types of user event histories are encountered with increasing regularity during forensic investigations. As an example, an investigator might be trying to determine if two event histories, corresponding to different usernames, were in fact generated by the same individual.

The primary contribution of this paper is the development of quantitative likelihood ratio techniques for forensic analysis of user-generated time-series in the form of event data. In particular we investigate score-based likelihood ratio methods in the context of determining whether two event histories are related, e.g., whether or not they were generated by the same individual. We focus in this paper on events that correspond to URL requests generated in a browser—however, the methodology we propose is broadly applicable to event data in general.

We begin by discussing related work, both in digital forensics as well as in score-based likelihood ratio methodologies and applications. We then discuss the theoretical foundations of the likelihood ratio and motivate the score-based likelihood ratio in the context of digital forensics. We then introduce relevant ideas from *marked point processes*, a statistical framework that has been widely used to analyze spatial point data, which we apply here to sequential event data streams. In particular we focus on the use of segregation and mingling indices as the basis for our score functions, and we describe how these techniques can be applied to evaluating the likelihood that two event streams were generated by the same source (or individual). We apply this methodology to a data set of event histories for 28 individuals, focusing on user activity related to social media. The results indicate that score functions based on marked point processes can have significant discriminative power for event-based data sets. In the final section of the paper we discuss both the promise and challenges involved in developing statistical analysis methods for event histories in the context of forensic investigations.

* Corresponding author.
E-mail addresses: galbraic@uci.edu (C. Galbraith), smyth@ics.uci.edu (P. Smyth).

http://dx.doi.org/10.1016/j.diin.2017.06.009

Related work

We will discuss two general threads of related work in this section: (i) methods for exploring and analyzing user event histories in the context of digital forensics and (ii) likelihood-ratio techniques for evaluating whether two samples originated from the same source. There has been relatively little overlap of these two topics in prior work, with a few exceptions (e.g., Ishihara, 2011; Overill and Silomon, 2010).

Analysis of user-event logs

In digital forensics there is significant interest in the development of tools that can assist in the investigation of user-generated event logs from computers and mobile devices (Casey, 2011; Roussev, 2016). These event logs may be stored locally on the device (Oh et al., 2011; Pereira, 2009) or (increasingly) in cloud storage (Roussev and McCulley, 2016). To help investigators better understand and explore these large data sets there has been a variety of work in recent years on techniques for visualization and analysis of such logs. Examples include interactive timeline analysis (e.g., Buchholz and Falk, 2005), exploring data theft using file copying timestamps (Grier, 2011), visualization of email histories (Koven et al., 2016), analyzing session to session similarities of Internet usage (Gresty et al., 2016), and linking user sessions via network traffic information (Kirchler et al., 2016). Beyond the field of digital forensics, in areas such as machine learning and data mining, a variety of general purpose event mining and analysis algorithms and tools have also been developed for exploration of event data, using techniques such as automated summarization (e.g., Kiernan and Terzi, 2009) and social network analysis (e.g., Eagle et al., 2009). In general, however, much of this prior work on event data is oriented towards data exploration, rather than on the development of statistical methodologies to answer specific questions in a digital forensics setting.

Score-based likelihood ratios in forensics

Although not commonly employed in digital forensics, likelihood ratio techniques have seen a great deal of attention in forensics as a whole. In forensic analysis a common question is whether two (or more) samples of interest come from the same source or not. Likelihood ratio (LR) methods provide a probabilistic framework for assessing the relative likelihood of the two competing hypotheses (same-source or different-source) given observed evidence. LR methods have been widely accepted in the practice of forensic science, particularly in DNA analysis (Foreman et al., 2003). In other areas such as glass fragment analysis (Aitken and Lucy, 2004), speaker recognition (Gonzalez-Rodriguez et al., 2006), fingerprint analysis (Neumann et al., 2007), handwriting analysis (Schlapbach and Bunke, 2007), and analysis of illicit drugs (Bolck et al., 2015), the use and application of LR techniques is still an area of ongoing research and investigation.

In the *direct LR approach* the probabilities (or likelihood) of the observed measurements (under some appropriate distributional model) are computed under both hypotheses being considered. *Score-based LR methods* differ to the direct approach in that they focus on distributions of similarities (or dissimilarities) between samples. These similarities are often one-dimensional, which can be easier to work with compared to modeling the often high-dimensional observations in the direct LR approach. The two approaches, score-based LR and direct LR, provide different tradeoffs in terms of flexibility and robustness (e.g., see Bolck et al. (2015) for a discussion of this tradeoff in the context of forensic analysis of chemical profiles of drugs). In this paper we focus on the score-based LR approach. This is motivated by the fact that the type of data we are analyzing, namely event time series, can be difficult to model directly (in terms of making appropriate distributional assumptions), making the score-based approach appealing and more directly applicable in this context.

The likelihood ratio

In the discussion below on likelihood ratios we will generally follow the notation of Bolck et al. (2015). The LR is the ratio of two conditional probabilities, where each probability corresponds to the strength of evidence under a particular hypothesis. The evidence, E, corresponds to observed data and can take different forms such as measurements related to DNA, fingerprints, or user-event streams. Let $E = \{X,Y\}$ where X is a set of observations (measured "features") for a reference sample from a known source (i.e., a sample from a suspect), and Y is a set of observations of the same features as X for a sample from an unidentified source (i.e., a sample recovered from the crime scene).

The likelihood ratio is the ratio of the probability of observing the evidence E under two competing hypotheses. The first hypothesis is that the samples come from the same source, H_s. The second hypothesis is that the samples come from different sources, H_d. The LR arises in the application of Bayes' theorem to this situation:

$$\underbrace{\frac{Pr(H_s|E)}{Pr(H_d|E)}}_{a\ posteriori\ \text{odds}} = \overbrace{\frac{Pr(E|H_s)}{Pr(E|H_d)}}^{\text{likelihood ratio}} \quad \underbrace{\frac{Pr(H_s)}{Pr(H_d)}}_{a\ priori\ \text{odds}} \tag{1}$$

The likelihood ratio serves the purpose of updating the *a priori* odds to form the *a posteriori* odds (i.e., the ratio of the probability of the hypothesis H_s to the probability of the hypothesis H_d after observing the evidence E) by comparing the probability of observing the evidence if the samples are from the same source versus different sources. In practice a forensic examiner may present a likelihood ratio involving a specific type of evidence to either the judge or jury, who then update their personal prior odds. This process is repeated for multiple forms of evidence until the decision maker can formulate their posterior odds to arrive at a final judgment. In this paper we focus specifically on the likelihood ratio in Equation (1) above, and in particular on statistical models and estimation techniques related to $Pr(E|H_s)$ and $Pr(E|H_d)$.

In practice we are often working with evidence E in the form of continuous measurements, requiring the use of probability density functions f (rather than probabilities Pr) to define the likelihood ratio:

$$LR = \frac{f(E|H_s)}{f(E|H_d)} = \frac{f(X,Y|H_s)}{f(X,Y|H_d)} \tag{2}$$

The likelihood ratio in Equation (2) is sometimes referred to as a feature-based likelihood ratio, where f is the joint density of the multivariate feature vectors X and Y. As mentioned earlier, estimating high-dimensional joint densities tends to be unreliable when the dimensionality of the data (the number of features in X and Y) is large. In particular, the number of observations required to reliably estimate a joint density to a required degree of accuracy tends to increase exponentially as a function of dimensionality (e.g., Scott, 1992).

One technique to sidestep this issue is to compute a function Δ of the observed samples X and Y and estimate the probability density function of $\Delta(X,Y)$, where $\Delta(X,Y)$ is typically a one-dimensional scalar-valued function of X and Y. This estimation