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:

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

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

can be performed using samples from a set of observational units (i.e., a reference data set $\mathcal{D}$) assumed to be a representative sample from the population of all possible sources. The function Δ is often referred to as a *score* function. It measures the similarity (or dissimilarity) of the two sets of observed features X and Y. Replacing X,Y with $\Delta(X,Y)$ in Equation (2) yields the score-based likelihood ratio (SLR):

$$SLR_{\Delta} = \frac{f(\Delta(X,Y)|H_s,\mathcal{D}_s)}{f(\Delta(X,Y)|H_d,\mathcal{D}_d)} \tag{3}$$

where we explicitly condition on the two data sets $\mathcal{D}_s$ (same-source) and $\mathcal{D}_d$ (different-source) used to construct the empirical densities in the numerator and denominator. The two conditioning sets are formed by restricting the reference data $\mathcal{D}$ to samples known to be from the same source and different sources, respectively. If Δ is a univariate function, then its density f is also univariate and relatively easy to estimate via any of a variety of standard parametric or non-parametric methods.

In order to compute the score-based likelihood ratio for a particular piece of evidence, we need to choose a function $\Delta(X,Y)$ that can assess the similarity of the two samples, where here we are interested in samples X and Y that are in the form of time-stamped events. In the next section we will discuss how to compute similarity functions $\Delta(X,Y)$ for such data using ideas from the marked point process literature.

Given conditioning sets $\mathcal{D}_s$ and $\mathcal{D}_d$ we can construct empirical estimates of the densities for Δ under the two competing hypotheses H_s and H_d, respectively. With these empirical distributions, and given two samples X and Y, we can compute the value of each of the density functions f evaluated at $\Delta(X,Y)$ to obtain $f(\Delta(X,Y)|H_s,\mathcal{D}_s)$ and $f(\Delta(X,Y)|H_d,\mathcal{D}_d)$. Finally we compute their ratio to get the score-based likelihood ratio SLR_{Δ} as in Equation (3). Fig. 1 provides an illustrative example of computing a SLR in this manner.

Score-based likelihood ratios can be used to quantify the strength of evidence in the following way: if $SLR_{\Delta} > 1$ we favor the hypothesis that the samples originated from the same source. Conversely, if $SLR_{\Delta} < 1$ we favor the hypothesis that the samples came from different sources. If the score-based likelihood ratio is equal to or close to 1, we say that the results are inconclusive. The further SLR_{Δ} is from 1, the more confidence we have in our conclusion.

Marked point processes

To define a score-based likelihood ratio in the context of event data we need to define a similarity score $\Delta(X,Y)$ for two observed streams of events X and Y. We do this using techniques from the modeling of marked point processes. These techniques are typically applied to problems involving statistical analysis of spatial point data—here we adapt them for temporal event data.

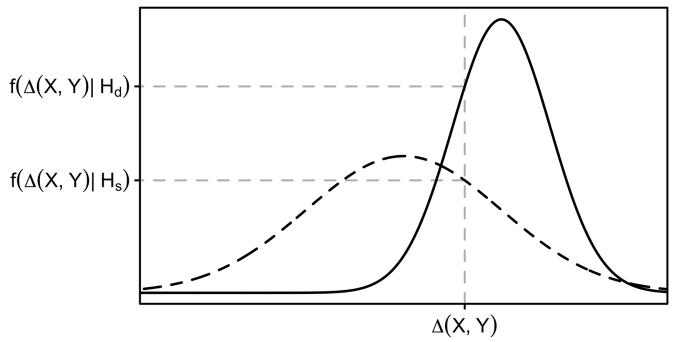


Fig. 1. Illustration of the densities of the score function Δ under the hypotheses that the samples are from the same source (H_s, dashed line) and that the samples are from different sources (H_d, solid line). The score-based likelihood ratio SLR_{Δ} is the ratio of the density functions f evaluated at $\Delta(X,Y)$.

In the discussion below we generally follow the notation of Illian et al. (2008) who define a marked point process M as a sequence of random marked points, $M = \{(t_n, m(t_n))\}$, where $m(t_n)$ is the mark of point $t_n \in \mathbb{R}^d$. The dimension d is associated with different physical interpretations (e.g., $d = 2$ for spatial data). In this paper, we restrict our focus to the one-dimensional temporal case ($d = 1$), so that the points t_n are simply timestamps, where $0 \leq t_n < t_{n+1}$ for all n. In practice we can only observe a finite window of time, and therefore we have a finite number of observations n.

In general the marks $m(t_n)$ can be either quantitative (continuous) or qualitative (categorical, discrete) and describe a particular characteristic of the objects represented by the points. The particular methodology used for analyzing marks depends on their type—different techniques are used for analyzing quantitative versus qualitative marks. Here we focus our attention on qualitative marks. If only two types of points are considered, denoted x and y, we call the process *bivariate*.

In the setting of digital forensics and user-event streams, we consider the following setup. Each user's event stream is dichotomized by the *mark*, or type of event, and represented as a bivariate marked point process in the temporal dimension. As an example, in the data we consider later in the paper, the first mark corresponds to Facebook-related browser events and the second mark corresponds to all other non-Facebook events in the browser.

Let $\{X_i, Y_i\}$ denote the i^{th} user's bivariate process where $X_i = \{t_{ij} : m(t_{ij}) = x \text{ for } j = 1,\ldots,n_i\}$ is defined as the sub-process of events of mark x. Similarly, define Y_i as the sub-process of events of mark y. Thus, the i^{th} user has n_i observed events where $t_{ij} \in \mathbb{R}^+$ and $m(t_{ij}) \in \{x, y\}$ are the time and mark of the i^{th} user's j^{th} event, respectively. Fig. 2 illustrates an example of a particular user's bivariate marked point process (this is a subset of the data that we will describe in detail later in the paper).

Illian et al. (2008) discuss a number of techniques to characterize the point distribution, while taking into account the marks, in a bivariate process. They make a distinction between first-order characteristics, indices, and higher order characteristics. The first-order characteristics consist of simple measures related to the points (i.e., the intensity λ, or the mean number of points per unit time) and the marks (i.e., the mark probabilities p_x and p_y, or the relative frequencies of each mark). Along with indices, which are neighborhood-based measurements, these characteristics describe the basic properties of the marks and points in a bivariate process. Higher order characteristics consider both the variability of the point distribution and the variability of marks and describe correlations among marks and points. These characteristics can describe longer-range dependencies in an event process, but are susceptible to certain assumptions (i.e., stationarity) that are not appropriate for the data analyzed in this paper.

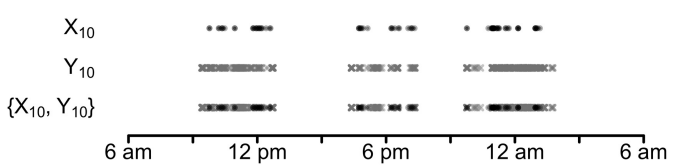


Fig. 2. Illustration of a bivariate marked point process in the temporal dimension for the first day of observation of user 10 in our data set. The lowest row (a mixture of red crosses and black points) is the bivariate process $\{X_{10}, Y_{10}\}$. The two rows above it represent the sub-processes of the event stream dichotomized by type: X_{10} (black dots of mark x, or Facebook browser events) and Y_{10} (red crosses of mark y, or non-Facebook browser events). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We focus on using two particular indices as score functions to compute likelihood ratios: the coefficient of segregation and the mingling index. Both of these indices rely on the notion of a *reference point*, which is a term related to the somewhat complicated statistical concept of the Palm distribution (Hanisch, 1984). For practical purposes, we define the reference point as an arbitrarily selected point in the process.

Coefficient of segregation

Pielou's coefficient of segregation (Pielou, 1977) is a function of (a) the ratio of the observed probability that the reference point and its nearest neighbor have different marks to (b) the same probability for independent marks,[1] defined by

$$\Delta_S(X_i, Y_i) = S(X_i, Y_i) = 1 - \frac{p_{xy} + p_{yx}}{p_x p_{\cdot y} + p_y p_{\cdot x}} \tag{4}$$

where p_{xy} (or p_{yx}) is the joint probability that the reference point has mark x and its nearest neighbor has mark y (or vice-versa), p_x and p_y are the mark probabilities, and $p_{\cdot x}$ (or $p_{\cdot y}$) is the probability that the nearest neighbor has mark x (or y) irrespective of the mark of the reference point. These probabilities are estimated based on empirical relative frequencies of the appropriate events as observed in the data. Here $\{X_i,Y_i\}$ represents the ith individual's bivariate event process as defined earlier.

The coefficient of segregation always takes values in $[-1, 1]$. If the reference point and its nearest neighbor always have the same mark, then $p_{xy} = p_{yx} = 0$ and $S(X_i,Y_i) = 1$. This corresponds to repulsion or clustering of points by their mark (i.e., points of type x always occur near each other and never near points of type y and vice-versa). If the reference point and its nearest neighbor always have different marks, then $p_{xx} = p_{yy} = 0$ which implies that $p_x = p_{yx}$ and $p_y = p_{xy}$ so $S(X_i,Y_i) < 0$ with a minimum of -1 if $p_x = p_y = 1/2$. This is the opposite of clustering, indicating that points of different marks are attracted to one another.

Mingling index

The mingling index is also based on local neighborhoods of the reference point. It compares the mark of the reference point to those of its k nearest neighbors, and is calculated by

$$\Delta_M(X_i, Y_i) = M_k(X_i, Y_i) = \frac{1}{k}\sum_{j=1}^{n_i}\sum_{\ell=1}^{k}\mathbf{I}\left[m(t_{ij}) \neq m(z_\ell(t_{ij}))\right] \tag{5}$$

where $z_\ell(t_{ij})$ denotes the ℓ^{th} nearest neighbor of the point t_{ij}. Thus, $M_k(X_i,Y_i)$ describes the mean fraction of points among the k nearest neighbors with a mark different than that of the reference point.

The mingling index can be thought of as a characterization of the mixture of marks and takes on values in $[0, 1]$. If the reference point and its k nearest neighbors tend to have the same mark, then $M_k(X_i,Y_i)$ has a small value and the process can be viewed as segregated (repulsion between points of different marks). In the opposite case, the mingling index has a large value and the process can be viewed as mixed (attraction).

In this paper we considered only the nearest neighbor so that $k = 1$. Since the process is bivariate, we can estimate $M_1(X_i,Y_i)$ via the joint probabilities p_{xy} and p_{yx} used in the estimation of the coefficient of segregation.

[1] The notion of independent marks in the denominator of Equation (4) simply refers to being able factor the joint probabilities p_{xy} and p_{yx} into the product of their marginals, so that under this assumption $p_{xy} = p_x p_{\cdot y}$ and $p_{yx} = p_y p_{\cdot x}$.

Data

The data considered in this paper comes from an *in situ* observational study conducted at a large US university in the spring of 2013 (Wang et al., 2015). In total 48 undergraduate students with Windows computers voluntarily participated in the study for a seven day period. Browser activities (such as URL requests) were automatically logged over a period of 7 days for each student. Participants were instructed to continue using their devices as normal while being logged.

The event logs from each student were dichotomized by their mark, or event type. The first mark corresponds to a Facebook event (i.e., any web browser activity occurring on facebook.com including any clicks or posts). The second mark corresponds to any non-Facebook event (i.e., any web browser activity not occurring on the aforementioned domain). We dichotomized the data in this manner to reflect the following type of situation: an individual deletes all browser-based social media activity from a device of interest in a criminal investigation in order to disassociate himself or herself from that device. The forensic examiner recovers the browser logs from that device (e.g., Oh et al., 2011) as well as the Facebook logs stored in the cloud for the individual under investigation (e.g., Roussev and McCulley, 2016). In this hypothetical situation the examiner then wants to determine the likelihood that both the cloud-based Facebook events and the device-based non-Facebook events were generated by the same individual.

The event logs from students were included in our analysis if they had at least 50 events of each type, to ensure that we would have enough data to accurately estimate the segregation and mingling score functions. Of the 48 students originally recorded, only 28 met the inclusion criteria. These students generated 66,966 log records, with 9500 (14.2%) Facebook and 57,466 (85.8%) non-Facebook browser events.

For the purposes of this study the data was de-identified by using an anonymized ID for each student. The resulting data set was in the form of < anonymous ID, timestamp, mark > triples, where mark indicated whether the browser event was related to Facebook or not. The resulting bivariate marked point processes for each student, with marks corresponding to event type, are illustrated in Fig. A.1.

Methodology

As discussed earlier, we use indices from marked point process theory to define the score function Δ for comparing event streams. This score function is used in turn to generate likelihood ratios that quantify the strength of evidence for or against the hypothesis that the events came from the same source, H_s, relative to the hypothesis that they came from different sources, H_d.

The event streams discussed in Section Data can be thought of as coming from the reference sample and the unidentified sample in the following manner:

- the reference sample X is a point process of Facebook-only browsing events, obtained from the cloud, and
- the unidentified sample Y is a point process of non-Facebook web browsing events, obtained from a device of interest.

The score-based likelihood ratio of Equation (3) is then the ratio of (i) the likelihood of observing the score function evaluated with the Facebook and non-Facebook events under the hypothesis that they were generated by the same source to (ii) the same likelihood under the hypothesis that they were generated by different sources.

For some intuition behind choosing the coefficient of segregation and mingling index as score functions, consider the first day of

observed data from two users depicted in Fig. 3. Clearly, these event streams are bursty in nature with periods of high activity followed by periods of little to no activity. It is also clear that within a particular user's event data (i.e., from the same source) that the Facebook and non-Facebook web browsing events tend to overlap in time, with Facebook events interleaved into the browsing activity. However, when comparing one user's sub-processes to those from the other user, we see little overlap and the bursts of events tend not to coincide with one another.

This suggests the coefficient of segregation between sub-processes generated by the same individual will tend to be lower than that of sub-processes from different individuals. Each user's events will tend to cluster together, so that when comparing event streams between different users the coefficient of segregation will be driven towards its maximum at one. Conversely, the mingling index between sub-processes generated by the same individual will tend to higher than that of sub-processes generated by different individuals. When comparing event streams between users, there is very little overlap so the mingling index is driven down towards its minimum at zero. These observations suggest that our score functions should have useful discriminative properties when applied to user-event data.

Evaluation of known same-source streams

To test the validity of our method in quantifying the strength of evidence, we estimated score-based likelihood ratios for both the coefficient of segregation, S, and the mingling index of the first nearest neighbor, M_1. We estimated these ratios for known same-source streams (i.e., both event streams known to be generated by the same user) via

$$SLR_{\Delta} = \frac{\widehat{f}(\Delta(X_i, Y_i)|H_s, \mathcal{D}_s)}{\widehat{f}(\Delta(X_i, Y_i)|H_d, \mathcal{D}_d)} \tag{6}$$

where Δ can be either the coefficient of segregation or mingling index measured from the i^{th} user's data via Equations (4) or (5), respectively. We denote these ratios SLR_S and SLR_{M_1}. As discussed in the Likelihood Ratio Section, score-based likelihood ratios greater (or less) than one tend to favor the hypothesis that the samples X and Y came from the same (or different) source.

Note that the probability density function f in Equation (3) has been replaced with $\widehat{f}$ in Equation (6). This represents the *estimated* empirical distribution, i.e., as estimated from data. The estimation of this density function can be done via a variety of parametric or non-parametric methods. In this paper we use kernel density estimates (KDEs) to obtain the distributions in both the numerator and denominator of Equation (6). KDEs are often used in forensic likelihood ratio analysis to avoid making parametric assumptions about the underlying distributions (e.g., Aitken and Lucy, 2004). Additional details about the KDE method used for the results in this paper are provided in Appendix B.

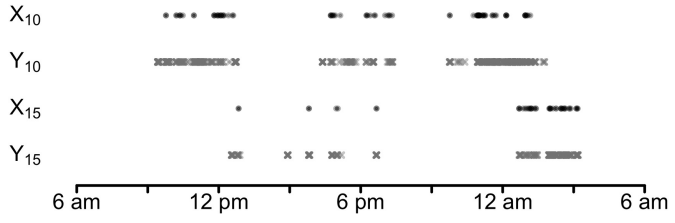


Fig. 3. Bivariate marked point processes for the first day of observation on users 10 and 15. The sub-processes X_i and Y_i denote Facebook and non-Facebook web browsing events, respectively.

The reference data set $\mathcal{D}$ is composed of multiple bivariate point processes corresponding to pairwise combinations of the $N = 28$ users' sub-processes (i.e., events from users restricted by type). It was used to create the conditioning sets $\mathcal{D}_s$ and $\mathcal{D}_d$ by imposing the restrictions that the events come from the same user or different users, respectively. To prevent overfitting to the training data, we removed all sub-processes from the i^{th} user when estimating the densities in Equation (6), in a manner that mimics leave-one-out cross validation. See Appendix C.1 for more details and notation on creating the conditioning sets.

Evaluation of known different-source streams

To ensure that our method can also correctly quantify the strength of evidence for event streams generated by different sources, we repeated the previous experiment using known different-source streams (i.e., the Facebook stream from user i and the non-Facebook stream from user j) via

$$SLR_{\Delta} = \frac{\widehat{f}(\Delta(X_i, Y_j)|H_s, \mathcal{D}_s^*)}{\widehat{f}(\Delta(X_i, Y_j)|H_d, \mathcal{D}_d^*)} \tag{7}$$

where the notation and estimation methods from the previous section hold for all variables except the conditioning sets $\mathcal{D}_s^*$ and $\mathcal{D}_d^*$, which are subsets of $\mathcal{D}_s$ and $\mathcal{D}_d$ discussed above. $\mathcal{D}_s^*$ is the set of same-source sub-processes excluding users i and j. $\mathcal{D}_d^*$ is the set of pairwise combinations of event streams from different users excluding all combinations with an event stream from either user i or j. See Appendix C.2 for more details.

Results

In this section we report on the evaluation experiments described in the Methodology.

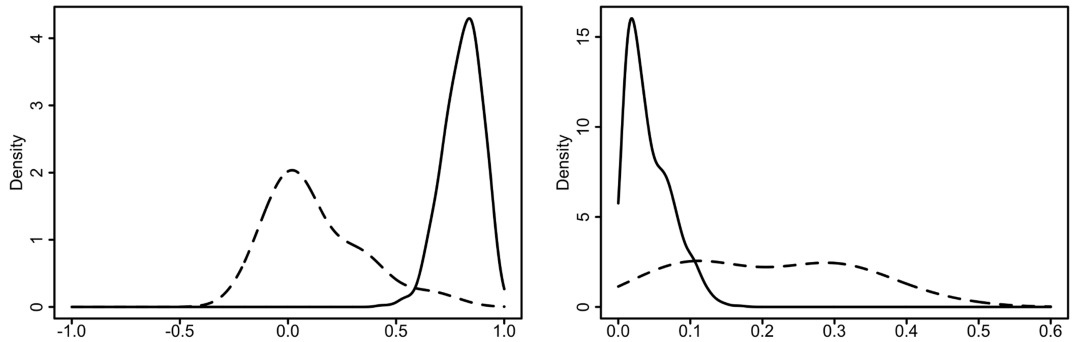
Density functions for scores

Fig. 4 shows the empirical densities for each of the coefficient of segregation and the mingling index. The densities shown in these plots were estimated with all available data, so that the same-source density used all 28 pairs of user data in $\mathcal{D}_s$ and the different-source density used all $N(N-1) = 756$ pairwise combinations of sub-processes from different users in $\mathcal{D}_d$.[2] While there is some overlap in the same- and different-source densities for both score functions, it is clear that the majority of the probability mass does not occur in the same region. This suggests that the score-based likelihood ratios will be able to accurately quantify the weight of evidence in favor of (or against) the hypothesis that the streams are from the same source.

Evaluation of known same-source streams

We first evaluated our approach for the case when the event streams are known to be from the same source, i.e., we are computing likelihood ratios for the scores $\Delta(X_i,Y_i)$. Fig. 5 shows the value of SLR_{M_1} versus SLR_S (logarithmic scale) computed with Equation (6) for the 28 pairs of same-source web browsing event streams. The threshold value of one (or zero on the log scale) is

[2] For illustrative purposes, the data from all users was used to estimate the empirical densities in Equations (6) and (7) and to generate the plots depicted in Fig. 4. The data sets $\mathcal{D}_s$ and $\mathcal{D}_d$ used to construct these plots differ from those discussed in Sections Methodology, Evaluation of known same-source streams, Evaluation of known different-source streams and Appendix C.



(a) $S(X,Y)$ (b) $M_1(X,Y)$

Fig. 4. Empirical distribution of the score functions under each hypothesis. Same-source density (H_s, dashed line) and different-source density (H_d, solid line) approximated via kernel density estimates with Gaussian kernels. Note that data from all of the users was used here for illustrative purposes—the densities depicted here differ from those used to evaluate our method (which used leave-one-out cross-validation). (a) Coefficient of segregation with bandwidths of 0.110 (H_s) and 0.025 (H_d). (b) Mingling index with $k = 1$ and bandwidths of 0.065 (H_s) and 0.009 (H_d).

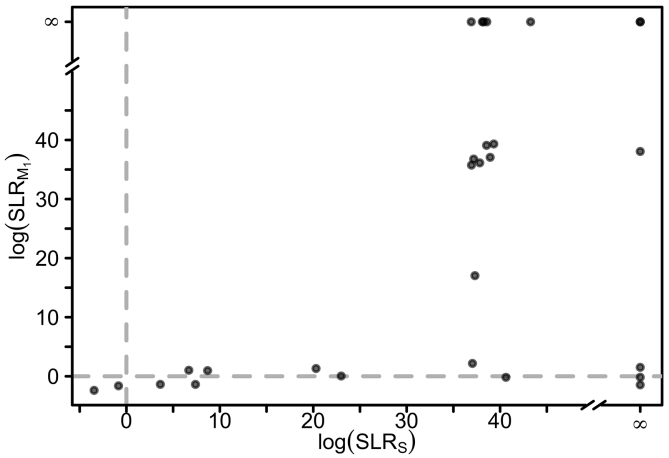


Fig. 5. Scatterplot of SLR_Δ, $\Delta \in \{S, M_1\}$, for the 28 pairs of same-source web browsing event streams calculated via Equation (6) on logarithmic scale. *Note the break in each axis to show the 11 processes with at least one infinite value for SLR_S or SLR_{M_1}.*

shown by the dashed lines. Table 1 presents the counts of the number of bivariate processes whose score-based likelihood ratios lie on either side of the threshold. It should be noted that there were 11 processes with at least one infinite value for SLR_S or SLR_{M_1}. This occurs when the likelihood of the score function evaluated at $\Delta(X_i,Y_i)$ under the hypothesis that the event streams came from different sources is numerically zero and indicates very strong support for the hypothesis that the event streams originated from the same source.

It is clear that the coefficient of segregation was more discriminative than the mingling index, with SLR_S correctly quantifying the weight of evidence in 26 of the 28 (93%) known same-source pairs.

Table 1
Counts of the measurement of the strength of evidence in known same-source pairs. Positives (+, $SLR_\Delta > 1$) and negatives (−, $SLR_\Delta < 1$) indicate that SLR_Δ favored the hypothesis that the streams are from the same source or different sources, respectively. SLR_S incorrectly quantified 2 of the 28 (7%) known matches.

		SLR_{M_1}		Total
		−	+	
SLR_S	−	2	0	2
	+	5	21	26
	Total	7	21	28

SLR_{M_1} was only able to correctly quantify 21 (75%) such pairs, all of which were also captured by SLR_S.

Of particular interest are the two bivariate processes (i.e., event streams for two users) for which both score-based likelihood ratios fail to support the same-source hypothesis, i.e., the likelihood ratios of both scores are less than one. These cases exhibit a unifying trait—they fall in the first quantile of the number of Facebook web browsing events per user (they each have less than 132 events of this type, compared to the median of 220 events per user). The sparse nature of their Facebook event data increases the coefficient of segregation and decreases the mingling index. A brief look at Fig. 4 explains why the likelihood ratios are less than the threshold values. Their score values fall in the right tail of Fig. 4a and the left tail of Fig. 4b where the same-source density is lower than the different-source density.

Evaluation of known different-source streams

We then evaluated our approach when the event streams are known to be from different sources, i.e., we are computing likelihood ratios for the scores $\Delta(X_i,Y_j)$. Fig. 6 shows the value of SLR_{M_1} versus SLR_S (logarithmic scale) computed with Equation (7) for the 756 pairwise combinations of different-source web browsing event streams. The threshold value of one (or zero on the log scale) is shown by the dashed lines. Table 2 presents the counts of the

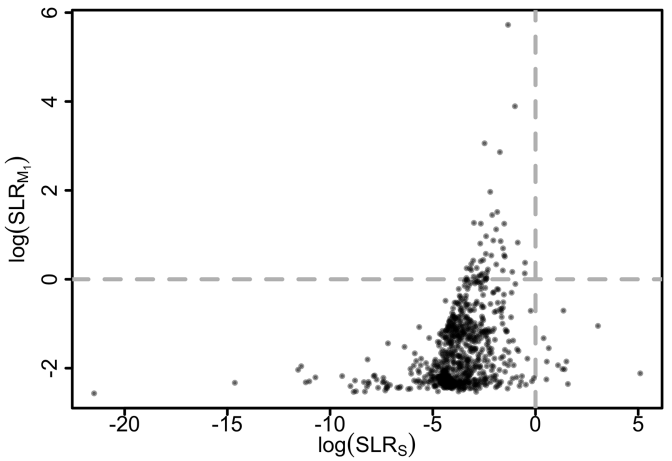


Fig. 6. Scatterplot of SLR_Δ, $\Delta \in \{S, M_1\}$, for the 756 pairwise combinations of different-source web browsing event streams calculated via Equation (7) on logarithmic scale.

Table 2
Counts of the measurement of the strength of evidence in known different-source pairs. Positives (+, $SLR_\Delta > 1$) and negatives (−, $SLR_\Delta < 1$) indicate that SLR_Δ favored the hypothesis that the streams are from the same source or different sources, respectively.

		SLR_{M_1}		Total
		−	+	
SLR_S	−	698	46	744
	+	12	0	12
	Total	710	46	756

number of bivariate processes whose score-based likelihood ratios lie on either side of the threshold.

Similar to the evaluation of known same-source event streams, the coefficient of segregation was more discriminative than the mingling index, with SLR_S correctly quantifying the weight of evidence in 744 of the 756 (98%) known different-source pairs. SLR_{M_1} was only able to correctly quantify 710 (94%) such pairs. If we use the criterion that we reject the same-source hypothesis when either SLR_S or SLR_{M_1} are below the threshold value of one, the method supports the correct hypothesis for all 756 (100%) known different-source pairs.

Discussion

Drawing on previous work in the domains of forensics and statistics, we have illustrated that score-based likelihood ratios based on marked point process indices have the potential to perform well in terms of discriminating known same- and different-source event streams from web browsing event data. The results support our earlier hypothesis, namely that the coefficient of segregation and the mingling index can be effective as discriminative score functions for user-event data.

We can combine the results of the two experiments above to obtain estimates of empirical true- and false-positive rates for our proposed method when applied to this type of data. The results indicate that using the coefficient of segregation as the score function, SLR_S, was more discriminative than using the mingling index, SLR_{M_1}. When using SLR_S, with a threshold of one, we obtain true and false positive rates of 92.9% and 1.6%, respectively (these numbers are computed by combining the results in Tables 1 and 2). If we use the less-accurate SLR_{M_1} we get true and false positive rates of 75% and 6.1%, respectively. If we restrict ourselves to only cases where both of the score-based likelihood ratios agree (i.e., both either exceed or are below one), we obtain true and false positive rates of 75% and 7.7%, respectively.

Caution must be taken when quantifying the strength of evidence using these score-based likelihood ratios. The choice of threshold is somewhat arbitrary (for example, one could choose values only greater than $1 + \omega$ and less than $1 - \gamma$ instead of the hard threshold of 1 to favor H_s and H_d, respectively). Also, the choice of which score-based likelihood ratio to use is somewhat arbitrary (for example, should we choose one, the other, or both? should we multiply them together?). Finally, note that the values reported here were obtained *only for one specific data set* and need not necessarily generalize to other data sets or different types of marked point processes. Despite these cautions, the results overall show considerable promise as a starting point for likelihood ratio analysis of user-event data.

Directions for future work

Further investigation is needed before score functions using marked point processes are well-understood enough to compute likelihood ratios that can be used in real-world forensic investigations. Such investigations should include both theoretical and empirical validation (e.g., see Meuwly et al., 2016).

An important priority for future work should be the calibration of the empirical density functions used in Equations (3), (6) and (7). The sensitivity and robustness of the method with respect to sample size (both the total number of users sampled and the number of events per user) needs further investigation. In our experimental results the score-based likelihood ratio failed to quantify the evidence in favor of the correct hypothesis (that the streams were from the same source) for two event streams with relatively little data. This suggests that the methodology will be less able to discriminate same-source and different-source hypotheses as the data becomes more sparse within streams. A potentially useful avenue for further investigation of this effect would be via simulation studies on the relationship between sparsity of user event data and the detectability of same-source event streams.

Another avenue for investigation in the context of calibration is the modeling of the empirical densities for the score functions, particularly for bandwidth selection in kernel density estimation (KDE). The value of the bandwidth in a KDE influences how smooth a density estimate is: larger values lead to smoother estimates with fatter tails, while smaller values lead to more "spiky" estimates with less probability mass in the tails. Thus, for example, smaller bandwidths will tend to lead to more extreme values for the likelihood ratios. In the results in this paper, we used a standard automated bandwidth selection technique. For our data sets this was certainly a reasonable and practical choice, but more generally further investigation is merited on the topic of the overall effect of bandwidth on the LR values when using KDEs.

Another topic worthy of further investigation relates to how the reference data set $\mathcal{D}$ and, consequently, the conditioning sets $\mathcal{D}_s$ and $\mathcal{D}_d$ are constructed for this methodology. Obtaining a representative sample of user-event streams from the population of interest could be a difficult task due to both privacy concerns and the proprietary nature of certain types of data. As an extreme case, consider a situation in which there are no background samples from which to construct these sets. Say we only have one reference sample X and one unidentified sample Y. The question then becomes how do we construct a reference data set to use in estimating the empirical densities. One potential direction to pursue would be to generate reference data sets via controlled simulations and use simulation-based techniques for likelihood-ratio construction.

Finally, while indices based on neighborhood characteristics are likely to be useful on overlapping event streams (as was the case in this paper), indices based on inter-event times could detect certain types of non-overlapping dependence patterns, e.g., when events of one mark tend to occur within a specific range of times before or after events of the other mark. This would further the applicability of the methods described here.

Conclusion

Analysis of user-generated event data is increasingly important in forensic investigation of digital evidence. However, few methodologies or tools have been developed to date that use statistical techniques, such as likelihood-ratio methods, for analysis of such data. In this paper we have taken a step towards the development of such techniques, focusing on the problem of investigating whether two event streams were generated by the same source or by different sources. We proposed an approach for generating similarity scores between event streams based on segregation and mingling indices, which are borrowed from statistical models of

marked point processes. Experimental results, based on analysis of real-world browser event streams from 28 individuals, indicate that the proposed methodology provides a useful starting point for discriminating between same-source and different-source pairs. Potential future directions that can build on these results include development of accurate calibration methods, analysis of sensitivity and robustness of the methodology, characterization of the properties of a broader range scores and indices, and additional experimental validation of the methods using both simulated and real-world data sets.

Acknowledgements

This research was partially funded through Cooperative Agreement #70NANB15H176 between the National Institute of Standards and Technology and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California, Irvine, and University of Virginia; and was also supported in part the National Science Foundation under grant number IIS-1320527 and by a Google Faculty Award (to PS).

The authors would like to acknowledge the contributions of Hal Stern for useful discussions related to the work in this paper. The authors would also like to thank the reviewers for their useful feedback on the original version of this paper.

Appendix A. Data

See Fig. A.1 for a visualization of the event streams from the 28 students.

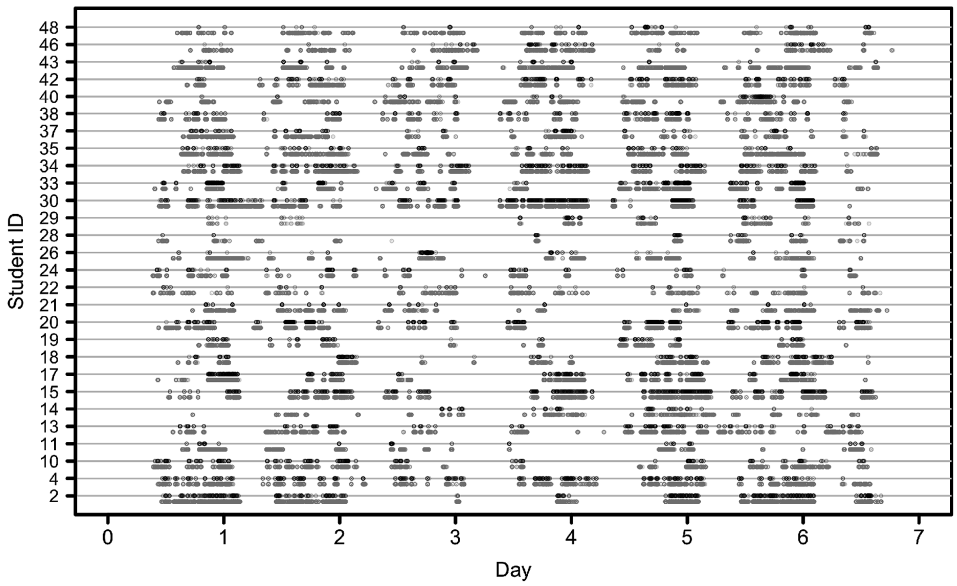


Fig. A.1. Web browsing data observed over 7 days from 28 students at a large U.S. university in the spring of 2013. Each line corresponds to data from one student with black dots on the line indicating his or her Facebook browser events and red dots immediately below the line representing non-Facebook browser events. Note that all events shown above are relative to the first day of observation for each student.

Appendix B. Kernel density estimation

Kernel density estimation (KDE) is a common choice for the non-parametric estimation of a probability density function f. The kernel function K is usually defined as any symmetric density function that satisfies the following conditions.

1. K integrates to unity: $\int K(x)dx = 1$
2. K has mean zero: $\int xK(x)dx = 0$
3. K has finite variance: $0 < \int x^2K(x)dx < \infty$

Common examples of kernel functions include the Gaussian (or Normal) and Epanechnikov kernels.

Assume that we have a collection of n points $X = \{X_1,\ldots,X_n\}$. Given a kernel function K and a bandwidth $h > 0$, a kernel density estimator is defined as

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) \tag{B.1}$$

Thus the estimated density is the average of the kernel centered at the observation X_i and scaled by h across all n observations. KDEs are essentially a local smoothing method. In the case where K is a point mass, the kernel density estimator is simply a histogram.

The choice of the kernel itself is not as important as the selection of the bandwidth h. As h decreases, the height of the peak at each observation increases resulting in undersmoothing. As h increases, the height of the peak at each observation decreases and probability mass is pushed away from the observation resulting in oversmoothing.

In this paper, we used a Gaussian kernel and automatic bandwidth selection via the "rule of thumb" from Scott (1992). All kernel density estimation was done with the density function in the R package stats (R Core Team, 2016).

Appendix C. Reference data set composition

The reference data set $\mathcal{D}$ is composed of multiple bivariate point processes corresponding to combinations of users' event streams, and is defined as $\mathcal{D} = \{\{X_i, Y_j\} : i,j \in \{1,\ldots,N\}\}$ where $N = 28$ is the total number of users.

Appendix C.1. Known same-source streams

The set $\mathcal{D}$ was used to create the conditioning sets $\mathcal{D}_s$ and $\mathcal{D}_d$ by imposing some restrictions. The set of event streams coming from the same source excluding user i is $\mathcal{D}_s = \{\{X_j, Y_j\} : j \in \{1, \ldots, N\}, j \neq i\}$, so that the probability density in the numerator of Equation (6) is estimated with the scores from the other $N-1 = 27$ users' bivariate processes. The set of event streams coming from different sources excluding user i is $\mathcal{D}_d = \{\{X_j, Y_k\} : j, k \in \{1, \ldots, N\}, j \neq k \neq i\}$, so that the denominator of Equation (6) is estimated with the scores from the other $(N-1)(N-2) = 27 \times 26$ pairwise combinations of the remaining users' sub-processes.

Appendix C.2. Known different-source streams

The conditioning sets for this experiment are subsets of those described above, so that $\mathcal{D}_s^* \subset \mathcal{D}_s$ and $\mathcal{D}_d^* \subset \mathcal{D}_d$. Here, the set of event streams coming from the same source excludes both users i and j, and is denoted $\mathcal{D}_s^* = \{\{X_k, Y_k\} : k \in \{1, \ldots, N\}, k \neq i \neq j\}$. In this manner the probability density in the numerator of Equation (7) is estimated with the scores from the other $N-2 = 26$ users' bivariate processes. The set of event streams coming from different sources excludes all bivariate processes containing a sub-process from either user i or j, and is denoted $\mathcal{D}_d^* = \{\{X_k, Y_\ell\} : k, \ell \in \{1, \ldots, N\}, k \neq \ell \neq i \neq j\}$, so that the denominator of Equation (7) is estimated with the scores from the other $(N-2)(N-3) = 26 \times 25$ pairwise combinations of the remaining users' sub-processes.

References

Aitken, C.G., Lucy, D., 2004. Evaluation of trace evidence in the form of multivariate data. J. R. Stat. Soc. Ser. C Appl. Stat. 53 (1), 109–122.

Bolck, A., Ni, H., Lopatka, M., 2015. Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. Law Probab. Risk 14 (3), 243–266.

Buchholz, F.P., Falk, C., 2005. Design and implementation of Zeitline: a forensic timeline editor. In: DFRWS.

Casey, E., 2011. Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet. Academic Press.

Eagle, N., Pentland, A.S., Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. Proc. Natl. Acad. Sci. 106 (36), 15274–15278.

Foreman, L., Champod, C., Evett, I., Lambert, J., Pope, S., 2003. Interpreting DNA evidence: a review. Int. Stat. Rev. 71 (3), 473–495.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J., 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Comput. Speech Lang. 20 (2), 331–355.

Gresty, D.W., Gan, D., Loukas, G., Ierotheou, C., 2016. Facilitating forensic examinations of multi-user computer environments through session-to-session analysis of internet history. Digit. Investig. 16, S124–S133.

Grier, J., 2011. Detecting data theft using stochastic forensics. Digit. Investig. 8, S71–S77.

Hanisch, K., 1984. Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point processes. Ser. Stat. 15 (3), 409–412.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical Analysis and Modelling of Spatial Point Patterns. John Wiley & Sons Ltd, West Sussex, England.

Ishihara, S., 2011. A forensic authorship classification in SMS messages: a likelihood ratio based approach using n-gram. In: Proc. of the Australasian Language Technology Association Workshop 2011, pp. 47–56.

Kiernan, J., Terzi, E., 2009. Constructing comprehensive summaries of large event sequences. ACM Trans. Knowl. Discov. Data 3 (4), 21.

Kirchler, M., Herrmann, D., Lindemann, J., Kloft, M., 2016. Tracked without a trace: linking sessions of users by unsupervised learning of patterns in their DNS traffic. In: Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security. ACM, pp. 23–34.

Koven, J., Bertini, E., Dubois, L., Memon, N., 2016. Invest: intelligent visual email search and triage. Digit. Investig. 18, S138–S148.

Meuwly, D., Ramos, D., Haraksim, R., 2016. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci. Int. 276, 142–153.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Bromage-Griffiths, A., 2007. Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. J. Forensic Sci. 52 (1), 54–64.

Oh, J., Lee, S., Lee, S., 2011. Advanced evidence collection and analysis of web browser activity. Digit. Investig. 8, S62–S70.

Overill, R.E., Silomon, J.A., 2010. Digital meta-forensics: quantifying the investigation. In: Proc. 4th International Conference on Cybercrime Forensics Education & Training (CFET 2010), Canterbury, UK (September 2010).

Pereira, M.T., 2009. Forensic analysis of the Firefox 3 Internet history and recovery of deleted SQLite records. Digit. Investig. 5 (3), 93–103.

Pielou, E., 1977. Mathematical Ecology. John Wiley & Sons, Inc.

R Core Team, 2016. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Roussev, V., 2016. Digital forensic science: issues, methods, and challenges. Synthesis lectures on information security. Priv. Trust 8 (5), 1–155.

Roussev, V., McCulley, S., 2016. Forensic analysis of cloud-native artifacts. Digit. Investig. 16, S104–S113.

Schlapbach, A., Bunke, H., 2007. A writer identification and verification system using HMM based recognizers. Pattern Anal. Appl. 10 (1), 33–43.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley.

Wang, Y., Niiya, M., Mark, G., Reich, S., Warschauer, M., 2015. Coming of age (digitally): an ecological view of social media use among college students. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 571–582.