Analyzing User-Event Data Using Score-based Likelihood Ratios with Marked Point Processes

Chris Galbraith

University of California, Irvine

August 9, 2017
Motivation

- Event histories of user activities (aka event streams or user-event data) are routinely logged on devices including computers and mobile phones.
- Typically consist of <event, timestamp, metadata>.
- As digital devices become more prevalent, these user event histories are encountered with increasing regularity.
- Investigators want to determine the likelihood that two event histories were generated by the same individual.
Motivation
Motivation
Related Work

- **Visualization tools**
  - Tools to assist the investigation of user-generated event logs from computers and mobile devices (Casey, 2011; Roussev, 2016)
  - Interactive timeline analysis (Buchholz & Falk, 2005)
  - Visualization of email histories (Koven et al., 2016)

- **Automated summarization & session similarity**
  - Analyzing session to session similarities of Internet usage (Gresty et al., 2016)
  - Linking user sessions via network traffic information (Kirchler et al., 2016)
  - Automated summarization of event data (Kiernan & Terzi, 2009)

- **Model based approaches**
  - Social network analysis (Eagle et al., 2009)
My Work

- Advised by Padhraic Smyth (and Hal Stern) under CSAFE
- Develop statistical methodologies to address questions of interest
  - Are two event streams from the same individual or not?
  - Are there unusual and significant changes in behavior?
- Develop testbed data sets to evaluate these methodologies
- Develop open-source software for use in the forensics community
Overview

1. The Likelihood Ratio
   - Feature-based
   - Score-based

2. Marked Point Processes
   - Bivariate Point Processes
   - Summary Statistics

3. Case Study
The Likelihood Ratio

- Probabilistic framework for assessing if two samples came from the same source or not
- $LR$ techniques have seen a great deal of attention in forensics as a whole
  - DNA analysis (Foreman et al., 2003)
  - Glass fragment analysis (Aitken & Lucy, 2004)
  - Speaker recognition (Gonzalez-Rodriguez et al., 2006)
  - Fingerprint analysis (Neumann et al., 2007)
  - Handwriting analysis (Schlapbach & Bunke, 2007)
  - Analysis of illicit drugs (Bolck et al., 2015)
DNA – *LR* Gold Standard

**Focus on alleles known to vary across the population.**

Compute the *likelihood* (or probability) of observing pairs of sequences under *two assumptions*.

1. Samples are from the *same* person
2. Samples are from *different* people
Feature-based Likelihood Ratio

1. Samples are from the *same* person

\[ Pr(\{X_i, Y\} | H_s) \]

2. Samples are from *different* people

\[ Pr(\{X_i, Y\} | H_d) \]

**Likelihood Ratio**

\[
\frac{Pr(\{X_i, Y\} | H_s)}{Pr(\{X_i, Y\} | H_d)}
\]

- **< 1** Samples from different sources
- **= 1** Inconclusive
- **> 1** Samples from same source
Score-based Likelihood Ratios

**Problem:** LR can be difficult to estimate.

**Solution:** Estimate the probability density function $f$ of a *score function* $\Delta$ that measures the similarity of the samples $X$ and $Y$, yielding the *score-based likelihood ratio*

$$SLR_\Delta = \frac{f(\Delta(X, Y) | H_s)}{f(\Delta(X, Y) | H_d)}$$
Overview

1. The Likelihood Ratio
   - Feature-based
   - Score-based

2. Marked Point Processes
   - Bivariate Point Processes
   - Summary Statistics

3. Case Study
Marked Point Processes

I follow the notation of Illian et al. (2008), who define a marked point process $M$ as a sequence of random marked points

$$M = \{(t_n, m(t_n)) : n = 1, 2, \ldots\}$$

where $m(t_n)$ is the mark of the point $t_n \in \mathbb{R}^d$

- Marks can be continuous or categorical (or both if multiple marks)
- Typically found in forestry, sociology, ecology, astronomy, etc.
Event streams can be viewed as marked point processes with the following properties:

- Temporal (i.e., time-stamped events)
- Binary marks corresponding to the type of event

We refer to these as *bivariate point processes*.
Coefficient of segregation, $S$ (Pielou, 1977): function of the ratio of observed probability that the reference point and its nearest neighbor have different marks to the same probability for independent marks

$$S(X_i, Y_i) = 1 - \frac{p_{xy} + p_{yx}}{p_x p_y + p_y p_x} \in [-1, 1]$$
Mingling index, $\overline{M}_k$ (Illian et al., 2008): mean fraction of points among the $k$ nearest neighbors of the reference point that have a mark different than the reference point

$$\overline{M}_k(X_i, Y_i) = \frac{1}{k} \sum_{j=1}^{n_i} \sum_{\ell=1}^{k} 1 \left[ m(t_{ij}) \neq m(z_{\ell}(t_{ij})) \right] \in [0, 1]$$
Overview

1. The Likelihood Ratio
   - Feature-based
   - Score-based

2. Marked Point Processes
   - Bivariate Point Processes
   - Summary Statistics

3. Case Study
Case Study

- Data from a 2013-2014 study at UCI that recorded students’ browser activity for one week (Wang et al., 2015)
- Dichotomize browser activity
  - Reference sample of Facebook-only events
  - Unidentified sample of non-Facebook events
- Considered 28 students with at least 50 events of each type
Compute bivariate process indices for all $N^2$ pairwise combinations of user event streams

For each pair $\{X_i, Y_j : i, j = 1, \ldots, N\}$ evaluate $SLR_S$ and $SLR_{M_1}$ with empirical likelihoods estimated from all other data

- Leave out all event streams from users $i$ and $j$
- Estimate the probability density of the score function $\Delta$ under each hypothesis
- Set $SLR_\Delta$ as the ratio of these estimated densities evaluated at $\Delta(X_i, Y_j)$
Results – Empirical Densities

(a) Segregation

(b) Mingling

Same-source density $H_s$ (dashed line)
Different-source density $H_d$ (solid line)
### Results – Classification Accuracy

<table>
<thead>
<tr>
<th>$SLR_{M_1}$</th>
<th>&lt; 1</th>
<th>&gt; 1</th>
<th>Total</th>
<th>$SLR_S$</th>
<th>&lt; 1</th>
<th>&gt; 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1</td>
<td></td>
<td></td>
<td>2</td>
<td>&gt; 1</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>&gt; 1</td>
<td>5</td>
<td>21</td>
<td>26 (93%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>21</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table**: Known same-source pairs

<table>
<thead>
<tr>
<th>$SLR_{M_1}$</th>
<th>&lt; 1</th>
<th>&gt; 1</th>
<th>Total</th>
<th>$SLR_S$</th>
<th>&lt; 1</th>
<th>&gt; 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1</td>
<td></td>
<td>698</td>
<td>46</td>
<td>&gt; 1</td>
<td></td>
<td></td>
<td>744 (98%)</td>
</tr>
<tr>
<td>&gt; 1</td>
<td>12</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>710 (94%)</td>
<td></td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table**: Known different-source pairs
Results – ROC Curve
Conclusions

- SLRs based on marked point process indices have potential to perform well in quantifying strength of evidence for user-event data.
- Segregation and mingling were discriminative score functions for web browsing event streams.
- Results obtained only for specific data set and may not generalize to others.
Future Work

- Other score functions (inter-event times & multiple marks)
- Theoretical characterization of limits of detectability
- Randomization methods
- Obtaining more real-world data
  - Currently planning additional data collection at UC Irvine
  - Order of 100 students, months of logged data


Feature-based Likelihood Ratio

Following the notation of Bolck et al. (2015), define

- Evidence $E \equiv \{X, Y\}$
- $X$: set of observations for a reference sample from a *known source*
- $Y$: set of observations of the same features as $X$ for a sample from an *unidentified source*
- $H_s$: same source hypothesis
- $H_d$: different sources hypothesis

$$\frac{Pr(H_s | E)}{Pr(H_d | E)} = \frac{Pr(E | H_s)}{Pr(E | H_d)} \frac{Pr(H_s)}{Pr(H_d)}$$

*a posteriori odds*  
*likelihood ratio*  
*a priori odds*
Kernel Density Estimation

- Kernel function $K$ usually defined as any symmetric density function that satisfies
  1. $\int K(x)dx = 1$
  2. $\int xK(x)dx = 0$
  3. $0 < \int x^2K(x)dx < \infty$

- Common kernels: Gaussian, Epanechnikov, point mass (histogram)

- Let $X = \{X_1, \ldots, X_n\}$. Then given $K$ and a bandwidth $h > 0$, a kernel density estimator is defined as

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

- Intuition: estimated density at $x$ is the average of the kernel centered at the observation $X_i$ and scaled by $h$ across all $n$ observations

- Choice of kernel really not important, but bandwidth is
Compute bivariate process indices \([S(X_i, Y_j) \text{ and } M_1(X_i, Y_j)]\) for all \(N^2 = 55^2 = 3025\) pairwise combinations of user event streams.

For each pairwise combination \(\{X_i, Y_j\}\) and \(\Delta \in \{S, M_1\}\), compute a “leave-one-out”–like estimate of the score-based likelihood ratio:

- \(D_s = \{\{X_k, Y_k\} : k \in \{1, \ldots, N\}, k \neq i, k \neq j\}\)
- \(D_d = \{\{X_k, Y_\ell\} : k, \ell \in \{1, \ldots, N\}, k \neq \ell, k \neq i, k \neq j, \ell \neq i, \ell \neq j\}\)
- Estimate \(\hat{f}(\Delta|H_s, D_s)\) and \(\hat{f}(\Delta|H_d, D_d)\) via KDE with the “rule of thumb” bandwidth (Scott, 1992)

Set \(SLR_\Delta\) as the ratio of these empirical densities evaluated at \(\Delta(X_i, Y_j)\).
Results – Evaluation of known same-source streams

\[
\begin{array}{c|cc|c}
 SLR_M_1 & - & + & \text{Total} \\
 SLR_S & 2 & 0 & 0 \\
 & 5 & 21 & 26 \\
 \text{Total} & 7 & 21 & 28 \\
\end{array}
\]
Results – Evaluation of known different-source streams

<table>
<thead>
<tr>
<th></th>
<th>$SLR_{M_1}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SLR_S$</td>
<td>- 698 46</td>
<td>744</td>
</tr>
<tr>
<td></td>
<td>+ 12 0</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>710 46</td>
<td>756</td>
</tr>
</tbody>
</table>