Quantifying the Association Between Discrete Event Time Series

Christopher Galbraith
Advised by Padhraic Smyth & Hal S. Stern

University of California, Irvine

May 25, 2018
Project Goals

- 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 by forensics community
Project Goals

- 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 by forensics community
Consider a pair of user-generated event series \( M = (A, B) \) such that

\[
M = \{(t_j, m(t_j)) : j = 1, \ldots, n\}
\]

where \( t_j \in \mathbb{R}^+ \) is the time and \( m(t_j) \in \{A, B\} \) is the type of the \( j^{th} \) event.

We want to quantify the likelihood that the pair was generated by the same source.
Approach

1. Determine suitable measures to quantify association between two event series $A$ and $B$.
2. Quantify the likelihood that a pair $(A, B)$ was generated by the same source or by different sources, given a measure of association.
   - *Assessing the strength or degree of association*

Approach

1. Determine suitable measures to quantify association between two event series $A$ and $B$.

2. Quantify the likelihood that a pair $(A, B)$ was generated by the same source or by different sources, given a measure of association.
   - *Assessing the strength or degree of association*

Methods to Assess Degree of Association

\[ (A^*, B^*) \]
Score Function \( \Delta \)

**Population-based Approach**
- Sample from relevant population:
  \[ M_i = (A_i, B_i) \text{ for } i = 1, \ldots, N \]
- Estimate score-based likelihood ratio (SLR)

**Resampling Approach**
- Single pair: \((A^*, B^*)\)
- Estimate coincidental match probability (CMP)

Degree of Association
Two competing hypotheses:

\[ H_s : (A^*, B^*) \text{ came from the same source} \]
\[ H_d : (A^*, B^*) \text{ came from different sources} \]

Use sample \( M_i = (A_i, B_i) \) for \( i = 1, \ldots, N \) to estimate the score-based likelihood ratio

\[ SLR_\Delta = \frac{g(\Delta(A^*, B^*)|H_s)}{g(\Delta(A^*, B^*)|H_d)} \]

Different interpretations of the denominator (Hepler et al., 2012)
Resampling Approach

- Usually don’t have sample from reference population
- Focus on the conditional likelihood given different sources
- **Coincidental match probability**: probability that a different-source pair with observed score $\Delta(A^*, B^*)$ exhibits association by chance

\[
CMP_\Delta = Pr(\Delta(A, B) < \Delta(A^*, B^*)|H_d)
\]

- Use resampling in time to simulate different-source pairs $(A^{(i)}, B^{(i)})$ and estimate

\[
\hat{CMP}_\Delta = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}[\Delta(A^{(i)}, B^{(i)}) < \Delta(A^*, B^*)]
\]
SLR vs CMP

\[ g(\Delta(A^*, B^*)| H_d) \]

\[ g(\Delta(A^*, B^*)| H_s) \]

\[ \Delta(A^*, B^*) \]
Case Study

- Data from a 2013-2014 study at UCI that placed logging software on 124 students’ computers that recorded all browser activity for one week (Wang et al., 2015)
- Event series created by dichotomizing browsing events to Facebook versus non-Facebook related urls
- Only considered 55 students with at least 50 web browsing events of each type
### Table: Performance of a classifier based on $SLR_\Delta$

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>TP@1</th>
<th>FP@1</th>
<th>Optimal Threshold</th>
<th>TP@opt</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>0.945</td>
<td>0.031</td>
<td>206</td>
<td>0.745</td>
<td>0.992</td>
</tr>
<tr>
<td>$\overline{M}_1$</td>
<td>0.855</td>
<td>0.116</td>
<td>218</td>
<td>0.473</td>
<td>0.946</td>
</tr>
<tr>
<td>$\overline{T}_{BA}$</td>
<td><strong>0.964</strong></td>
<td><strong>0.029</strong></td>
<td><strong>49</strong></td>
<td><strong>0.873</strong></td>
<td><strong>0.996</strong></td>
</tr>
<tr>
<td>$med(T_{BA})$</td>
<td><strong>0.964</strong></td>
<td>0.085</td>
<td>115</td>
<td>0.818</td>
<td>0.992</td>
</tr>
</tbody>
</table>

### Table: Performance of a classifier based on $CMP_\Delta$  

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>TP@5%</th>
<th>FP@5%</th>
<th>TP@0.1%</th>
<th>FP@0.1%</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overline{T}_{BA}$</td>
<td>1.000</td>
<td><strong>0.036</strong></td>
<td>0.982</td>
<td><strong>0.002</strong></td>
<td>0.999</td>
</tr>
<tr>
<td>$med(T_{BA})$</td>
<td><strong>1.000</strong></td>
<td>0.176</td>
<td><strong>1.000</strong></td>
<td>0.015</td>
<td>0.992</td>
</tr>
</tbody>
</table>
Simulated the equivalent of one week of data for pairs of processes with varying degrees of association

- $A$: Poisson process with intensity $\lambda_A$
- $B$: independent Poisson process with intensity $\lambda_B = p\lambda_A$, $p \in (0, 1)$
  or with probability $p$ add Gaussian noise to event in $A$

- 10,000 independent & 10,000 associated pairs for each combination of parameters

- Most important factor in detecting associated pairs is the signal-to-noise ratio

$$\gamma = \frac{\overline{T}_{AA}}{\sigma}$$
Simulation Results II

Quantifying Association

May 25, 2018 13 / 17
Simulation Results III

C. Galbraith (UCI)

Quantifying Association

May 25, 2018 14 / 17
Conclusions

- The resampling approach shows promise in situations where no reference data is available.
- The population-based SLR is still the preferred method, given:
  - Better performance for pairs exhibiting weak association.
  - Similar performance to the CMP for strongly associated pairs.
  - Well-established approach in forensic investigation.
Future Directions

- Preparing R package `assocr` for release
- Potential collaboration with Los Alamos National Laboratory
- Extend methodology (spatial data, exclusion patterns, etc)
- Develop theory of detectability
- Develop methods for identification

Simulation Results IV

Figure: $\gamma = 14.6$

Figure: $\gamma = 7.3$
Algorithm 1 Sessionized Resampling

Input: Pair of event series \((A^*, B^*)\)

Output: Set of resampled pairs \(\mathcal{D}\)

1: Fix \(B^*\)
2: \textbf{for} \(\ell = 1\) \textbf{to} \(n_{\text{sim}}\) \textbf{do}
3: \hspace{1em} \textbf{for} \(k = 1\) \textbf{to} \(n_{A^*}\) \textbf{do}
4: \hspace{2em} \text{Draw} \(t_{\text{new}} \sim p(t^-)\)
5: \hspace{2em} \text{Set} \(S_{a,k}^{(\ell)} = S_{a,k} - t_{k}^- + t_{\text{new}}\)
6: \hspace{1em} \textbf{end for}
7: \hspace{1em} \text{Set} \(A^{(\ell)} = \{S_{a,k}^{(\ell)} : k=1,\ldots,n_{A^*}\}\)
8: \hspace{1em} \textbf{end for}
9: \textbf{return} \(\mathcal{D} = \{(A^{(\ell)}, B^*) : \ell = 1,\ldots,n_{\text{sim}}\}\)
Algorithm 2 Simulation of associated marked point processes

Input: $\lambda_A, p, \sigma$

Output: Simulated pair of processes $(A, B)$

1: Simulate $A = \{t_j : j = 1, \ldots, n_A\}$ from a Poisson point process with rate $\lambda_A$
2: Set $k = 0$
3: for $j = 1$ to $n_A$ do
4: Draw $d_j \sim \text{Bernoulli}(p)$
5: if $d_j = 1$ then
6: Increment $k = k + 1$
7: Draw $t_k \sim \text{Normal}(\mu = t_j, \sigma^2)$ where $t_j \in A$
8: end if
9: end for
10: return $B = \{t_k : k = 1, \ldots, n_B = \sum_{j=1}^{n_A} d_j\}$
Signal-to-Noise Ratio, I

Recall that the numerator of the signal-to-noise ratio $\gamma$ is the reciprocal of the mean intensity of the simulated realizations of process $A$, i.e.,

$$
\overline{\lambda}_A^{-1} = \left[ n^{-1} \sum_{i=1}^{n} \lambda_A^{(i)} \right]^{-1}.
$$

(1)

where $n$ is the number of simulated processes and $\lambda_A^{(i)}$ is the intensity of the $i^{th}$ realization of process $A$. Since each realization of $A$ is a Poisson process, the inter-event times $\tau_{AA}^{(i,j)}$ for $j = 1, \ldots, n_A^{(i)}$ are distributed $i.i.d.$ Exponential($\lambda_A^{(i)}$), and their expectation is

$$
\mathbb{E}_{\tau} \left( \tau_{AA}^{(i,j)} \right) = \left( \lambda_A^{(i)} \right)^{-1} \quad \forall j.
$$

(2)
Note that each realization of $A$ is independent of the other $n - 1$ realizations. Thus the expected inter-event time across the realizations of $A$ is

$$\mathbb{E}_{\tau} \left( \overline{\tau}_{AA}^{(i,j)} \right) = \mathbb{E}_{\tau} \left( n^{-1} \sum_{i=1}^{n} \overline{\tau}_{AA}^{(i,j)} \right)$$

$$= n^{-1} \sum_{i=1}^{n} \mathbb{E}_{\tau} \left( \tau_{AA}^{(i,j)} \right)$$

$$= n^{-1} \sum_{i=1}^{n} \left( \lambda_A^{(i)} \right)^{-1}$$

$$\rightarrow \mathbb{E}_{\lambda} \left( \frac{1}{\lambda_A} \right) \quad \text{as } n \rightarrow \infty.$$
Since \( \lambda_A^{-1} \) is a convex function, we can apply Jensen’s inequality to (6) to obtain

\[
\frac{1}{\lambda_A} \rightarrow \frac{1}{\mathbb{E}_\lambda(\lambda_A)} \leq \mathbb{E}_\lambda \left( \frac{1}{\lambda_A} \right). \tag{7}
\]

Therefore, \( \lambda_A^{-1} \) is a lower bound on the expected inter-event time across the simulated realizations of process \( A \). It is more conservative to use than (5) for calculating \( \gamma \) since it results in an under-estimate of the amount of noise present in the processes.