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

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

csafe
Center for Statistics and Applications in Forensic Evidence

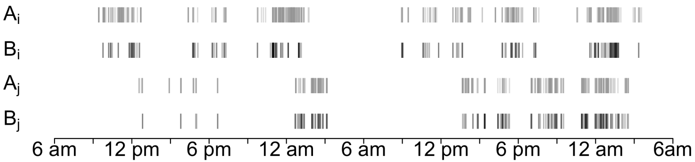
Project Goals

- 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

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

csafe
Center for Statistics and Applications in Forensic Evidence

Problem Statement



- 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 $m(t_j) \in \{\mathrm{A}, \mathrm{B}\}$ 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*

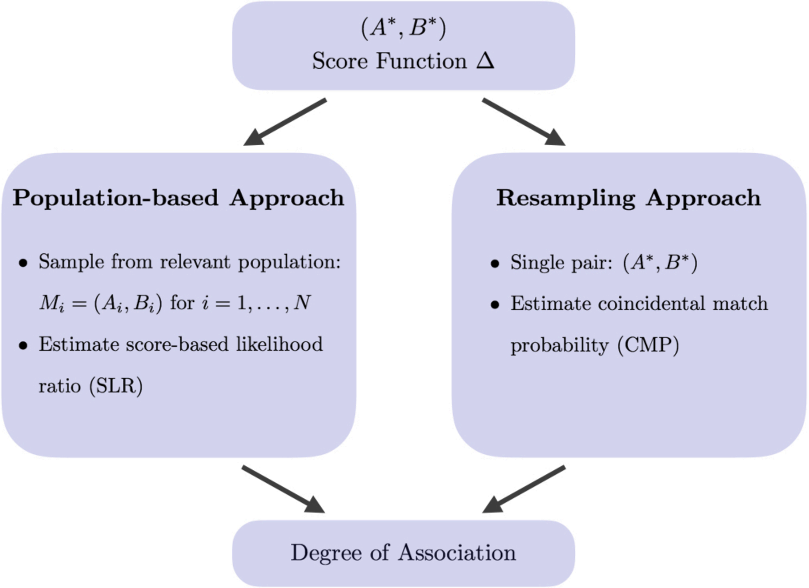
C. Galbraith, P. Smyth & H. S. Stern (2018). "Statistical Methods for Quantifying the Association Between Discrete Event Time Series." Under review by *IEEE Transactions on Information Forensics and Security.*

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*

C. Galbraith, P. Smyth & H. S. Stern (2018). "Statistical Methods for Quantifying the Association Between Discrete Event Time Series." Under review by *IEEE Transactions on Information Forensics and Security.*

Methods to Assess Degree of Association



Population-based Approach

- 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

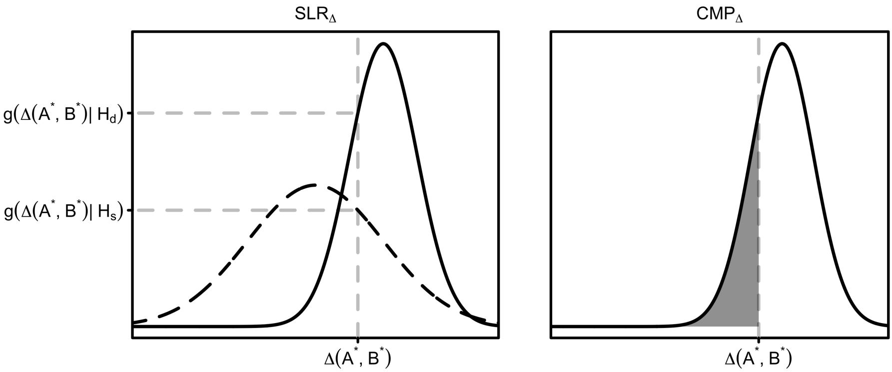
- Usually don't have sample from reference population
- Focus on the conditional likelihood given different sources
- *Coincidental match probability:* probability that a different-source pair with observed score $\Delta(A^*, B^*)$ exhibits association by chance

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

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

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

SLR vs CMP



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

Case Study Results

Table: Performance of a classifier based on SLR_Δ

Δ	TP@1	FP@1	Optimal Threshold	TP@opt	AUC
S	0.945	0.031	206	0.745	0.992
$\overline{M}_1$	0.855	0.116	218	0.473	0.946
$\overline{\mathcal{T}}_{BA}$	**0.964**	**0.029**	**49**	**0.873**	**0.996**
$med(\mathcal{T}_{BA})$	**0.964**	0.085	115	0.818	0.992

Table: Performance of a classifier based on CMP_Δ

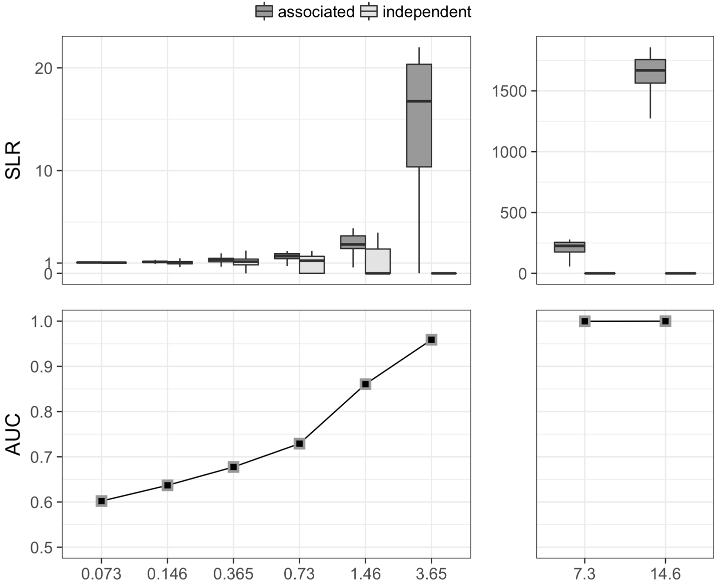
Δ	TP@5%	FP@5%	TP@0.1%	FP@0.1%	AUC
$\overline{\mathcal{T}}_{BA}$	**1.000**	**0.036**	0.982	**0.002**	**0.999**
$med(\mathcal{T}_{BA})$	**1.000**	0.176	**1.000**	0.015	0.992

Simulation

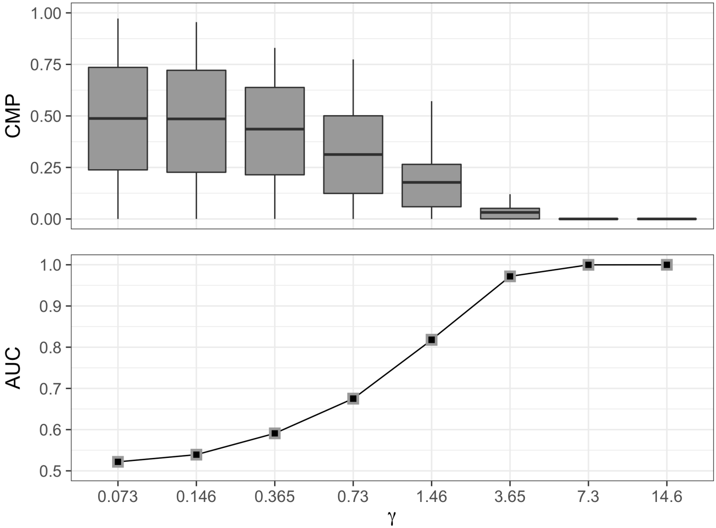
- Simulated the equivalent of one week of data for pairs of processes with varying degrees of association
  - A: Poisson process with intensity λ_A
  - B: independent Poisson process with intensity $\lambda_B = p\lambda_A$, $p \in (0, 1)$ or with probability p add Gaussian noise to event in A
- 10,000 independent & 10,000 associated pairs for each combination of parameters
- Most important factor in detecting associated pairs is the signal-to-noise ratio

$$\gamma = \frac{\overline{\mathcal{T}}_{AA}}{\sigma}$$

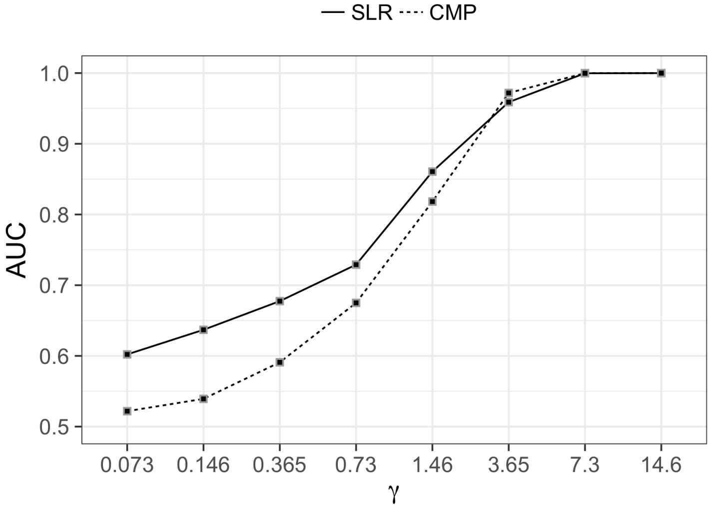
Simulation Results



Simulation Results II



Simulation Results III



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

References

Hepler, A. B., Saunders, C. P., Davis, L. J., & Buscaglia, J. (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, *219*(1), 129 - 140. doi: https://doi.org/10.1016/j.forsciint.2011.12.009

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).

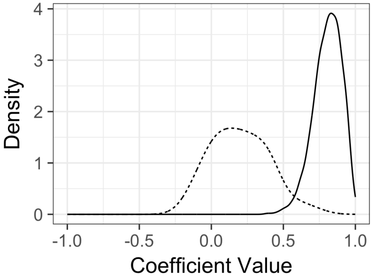


Figure: Segregation

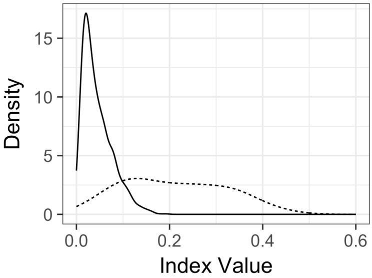


Figure: Mingling

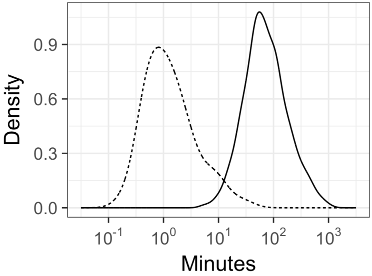


Figure: Mean IET

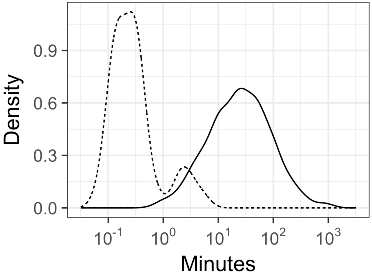


Figure: Median IET

Simulation Results IV

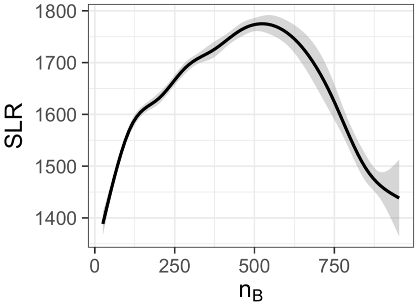


Figure: $\gamma = 14.6$

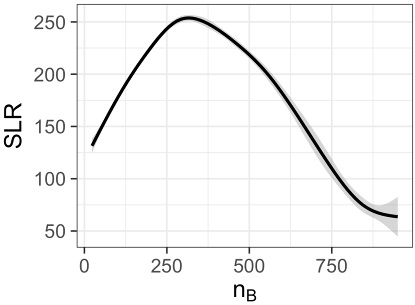


Figure: $\gamma = 7.3$

Algorithm 1 Sessionized Resampling

Input: Pair of event series (A^*, B^*)
Output: Set of resampled pairs $\mathcal{D}$

```
Fix B*
for ℓ = 1 to n_sim do
    for k = 1 to n⁻_{A*} do
        Draw t_new ~ p(t⁻)
        Set S^(ℓ)_{a,k} = S_{a,k} − t⁻_k + t_new
    end for
    Set A^(ℓ) = {S^(ℓ)_{a,k} : k = 1, ..., n⁻_{A*}}
end for
return D = {(A^(ℓ), B*) : ℓ = 1, ..., n_sim}
```

Fix B^*
for $\ell = 1$ to n_{sim} **do**
  for $k = 1$ to $n^-_{A^*}$ **do**
    Draw $t_{new} \sim p(t^-)$
    Set $S^{(\ell)}_{a,k} = S_{a,k} - t^-_k + t_{new}$
  end for
  Set $A^{(\ell)} = \{S^{(\ell)}_{a,k} : k = 1, \ldots, n^-_{A^*}\}$
end for
return $\mathcal{D} = \{(A^{(\ell)}, B^*) : \ell = 1, \ldots, n_{sim}\}$

Algorithm 2 Simulation of associated marked point processes

Input: λ_A, p, σ

Output: Simulated pair of processes (A, B)

Simulate $A = \{t_j : j = 1, \ldots, n_A\}$ from a Poisson point process with rate λ_A
Set $k = 0$
for $j = 1$ to n_A **do**
    Draw $d_j \sim \mathit{Bernoulli}(p)$
    if $d_j = 1$ **then**
        Increment $k = k + 1$
        Draw $t_k \sim \mathit{Normal}(\mu = t_j, \sigma^2)$ where $t_j \in A$
    end if
end for
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 γ is the reciprocal of the mean intensity of the simulated realizations of process A, i.e.,

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

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

$$\mathbb{E}_\tau \left(\tau_{AA}^{(i,j)} \right) = \left(\lambda_A^{(i)} \right)^{-1} \quad \forall j. \quad (2)$$

Signal-to-Noise Ratio, II

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

$$\mathbb{E}_\tau\left(\overline{\tau}_{AA}^{(\cdot,\cdot)}\right) = \mathbb{E}_\tau\left(n^{-1}\sum_{i=1}^{n}\overline{\tau}_{AA}^{(i,\cdot)}\right) \tag{3}$$

$$= n^{-1}\sum_{i=1}^{n}\mathbb{E}_\tau\left(\tau_{AA}^{(i,j)}\right) \tag{4}$$

$$= n^{-1}\sum_{i=1}^{n}\left(\lambda_A^{(i)}\right)^{-1} \tag{5}$$

$$\to \mathbb{E}_\lambda\left(\frac{1}{\lambda_A}\right) \quad \text{as } n \to \infty. \tag{6}$$

Signal-to-Noise Ratio, III

Since λ_A^{-1} is a convex function, we can apply Jensen's inequality to (6) to obtain

$$\frac{1}{\overline{\lambda}_A} \to \frac{1}{\mathbb{E}_\lambda(\lambda_A)} \leq \mathbb{E}_\lambda \left(\frac{1}{\lambda_A} \right). \tag{7}$$

Therefore, $\overline{\lambda}_A^{-1}$ 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 γ since it results in an under-estimate of the amount of noise present in the processes.