Quantifying the Association Between Discrete Event Time Series

Christopher Galbraith[†]

Padhraic Smyth[‡] & Hal S. Stern[†]

[†]Department of Statistics

[‡]Department of Computer Science

UCIRVINE

UNIVERSITY of CALIFORNIA · IRVINE

July 31, 2018

Logs of User-Generated Event Data

Browser requests
Web searches
Email activity
Phone/SMS
Social media activity
GPS locations
File access
Network activity
Exercise/movement
.....

User Event Data

< ID, timestamp, action type, metadata >

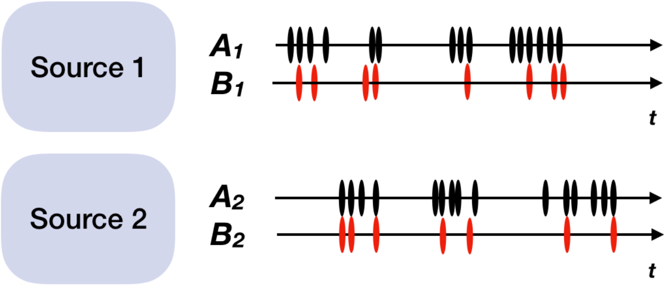
Text content
Location
List of recipients
.....

Web clicks
Web searches
Emails sent
Social media posts
Files edited
.....

We focus on ID, timestamp, and type of actions

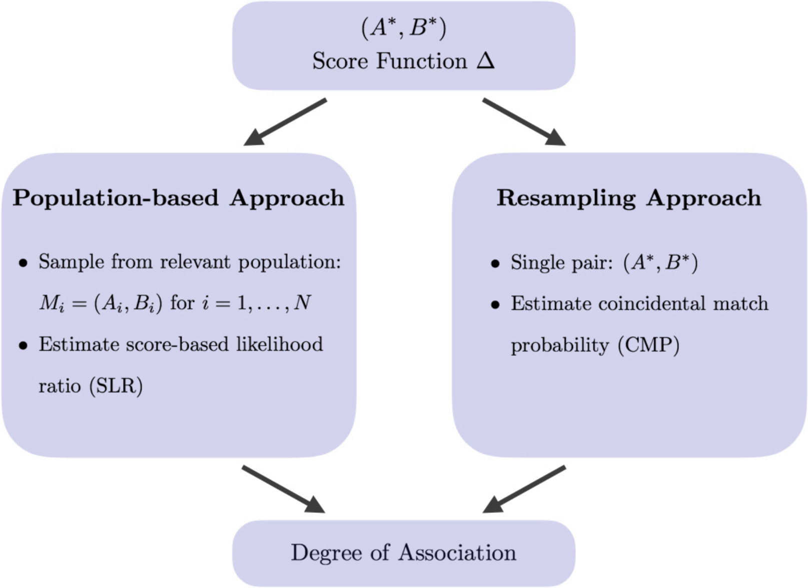
Problem Statement

- Consider a pair of user-generated event series $M = (A, B)$
  - Each series fully characterized by event times
  - Event types differ between series
- Quantify the likelihood that the pair was generated by the same source



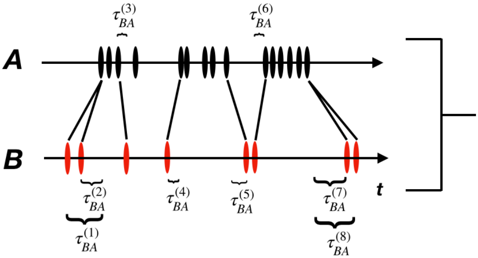
WLOG assume that $n_B < n_A$.

Methodology



Score Functions

- Need to determine suitable measures to quantify association between two event series A and B.
  - Nearest-neighbor indices (from marked point process literature)
  - **Distribution of inter-event times**



$$\Delta(A,B) = \overline{\tau}_{BA} = \frac{1}{n_B}\sum_{i=1}^{n_B} \tau_{BA}^{(i)}$$

Population-based Approach

- Two competing propositions:

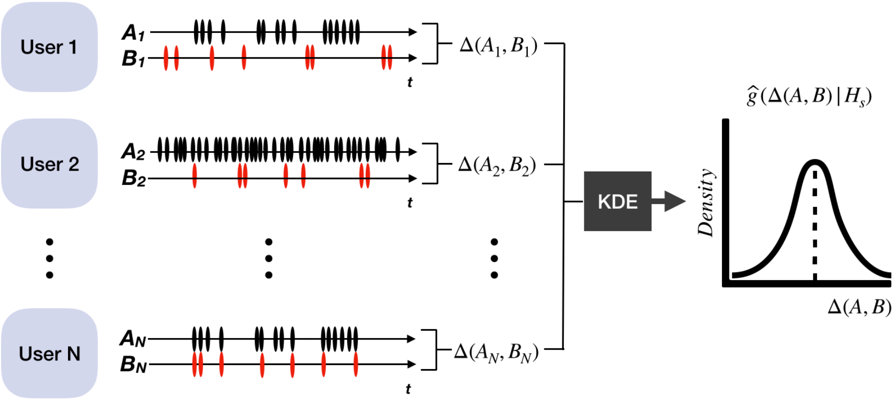
$$H_s : (A^*, B^*) \text{ came from the same source}$$
$$H_d : (A^*, B^*) \text{ came from different sources}$$

- Use sample $M_i = (A_i, B_i)$ for $i = 1, \ldots, N$ to estimate the *score-based likelihood ratio* for the observed score $\Delta(A^*, B^*)$

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

- Different interpretations of denominator lead to different *SLR*s (Hepler et al., 2012)

Estimation of g

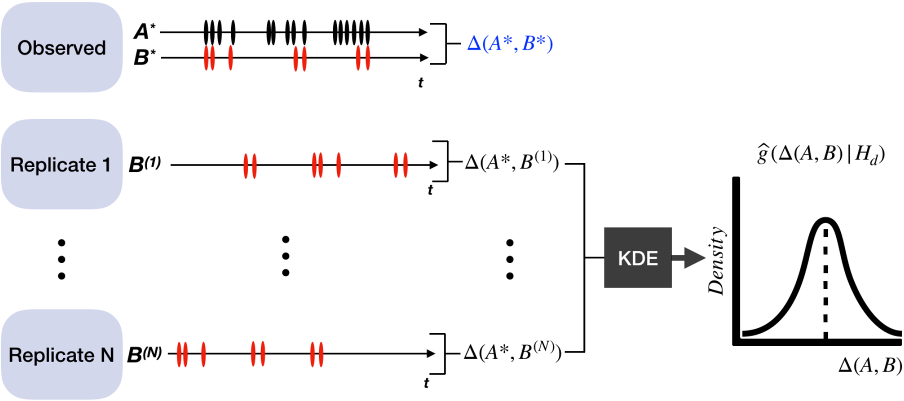


To estimate $g(\Delta(A,B)|H_d)$, repeat this process using all pairwise combinations of event series $(A_i, B_j) \ni i \neq j$.

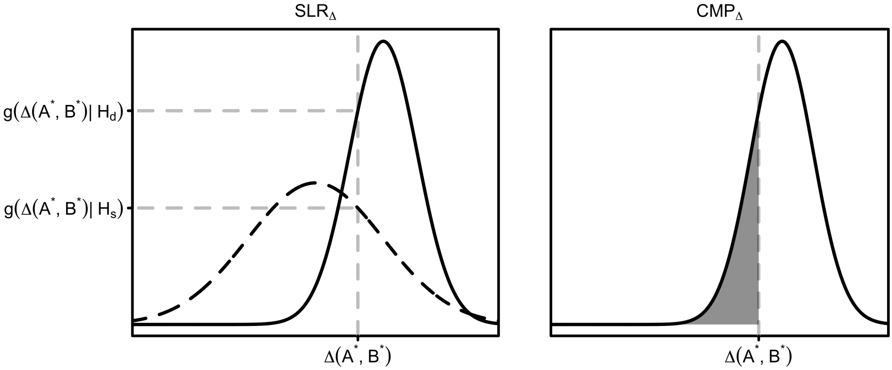
Resampling Approach

- *Coincidental match probability:* probability that a different-source pair with observed score $\Delta(A^*, B^*)$ exhibits association by chance

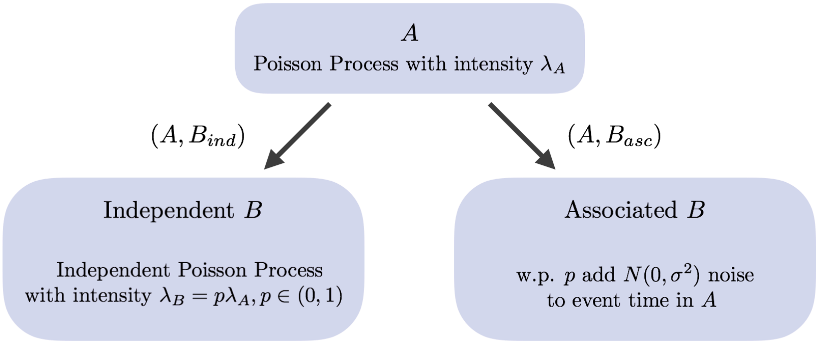
$$CMP_\Delta = Pr(\Delta(A, B) < \Delta(A^*, B^*)|H_d)$$



Comparison of Approaches



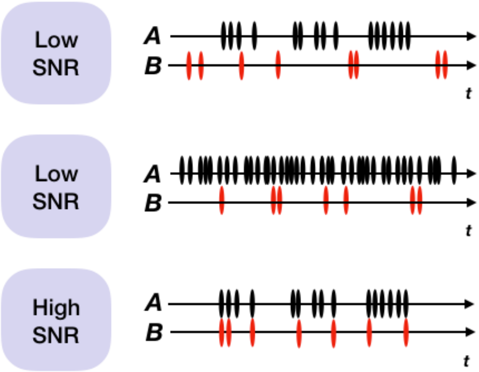
Simulation Study



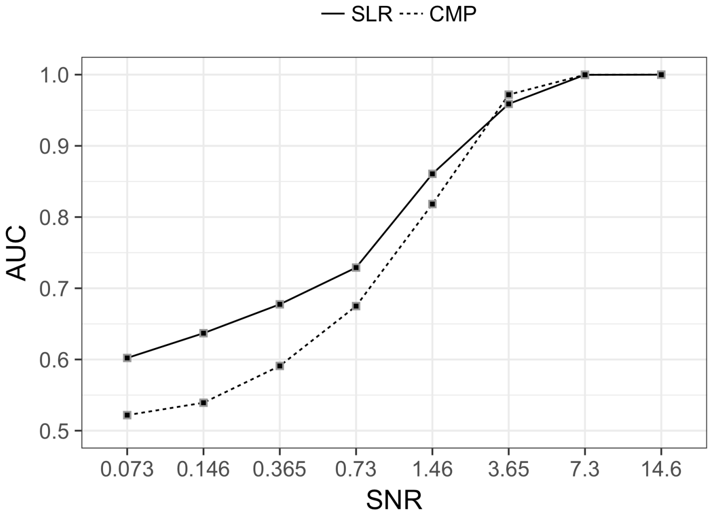
- Simulated the equivalent of one week of data for 20k pairs of processes (10k independent & 10k associated)
- Repeated for various combinations of (λ_A, p, σ)

Signal-to-Noise Ratio

$$\text{SNR} = \frac{\overline{\tau}_{AA}}{\overline{\tau}_{BA}} = \frac{\text{mean IET for process } A}{\text{mean IET from } B \text{ events to nearest } A \text{ event}}$$



Simulation Results



$^{*}p = 0.20$

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
- Considered 55 students with at least 50 web browsing events of each type

Case Study Results

Method	Score Function Δ	TP Rate*	FP Rate*	AUC
Population-based	Near-neighbor (mingling)	85.5	11.6	94.6
Population-based	Near-neighbor (segregation)	94.5	3.1	99.2
Population-based	Inter-event Time (mean)	96.4	2.9	99.6
Resampling	**Inter-event Time (mean)**	**98.2**	**0.2**	**99.9**

**Population-based methods use SLR with a threshold of 1*

**Sampling-based method uses CMP with threshold of 0.1%*

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
- R implementation available on Github: `assocr`

Future Directions

- Extend methodology
  - Spatial data
  - Other types of association (e.g., exclusion and ‘causal’ patterns)
  - Incorporate more (> 2) types of events
- Develop methods for identification
- Develop theory of detectability

Acknowledgements

UCI DataLab

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

csafe
Center for Statistics and Applications in Forensic Evidence

The material presented here is based upon work supported by the National Institute of Science and Technology under Award No. 70NANB15H176. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Institute of Science and Technology, nor of the Center for Statistics and Applications in Forensic Evidence.

References

Galbraith, C., & Smyth, P. (2017). Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digital Investigation*, *22*(Supplement), S106 - S114. doi: https://doi.org/10.1016/j.diin.2017.06.009

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).

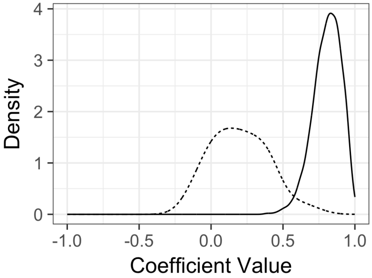


Figure: Segregation

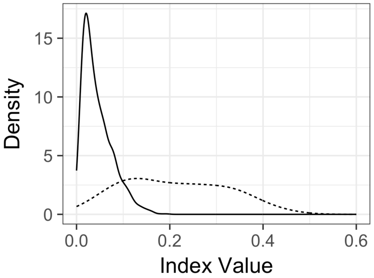


Figure: Mingling

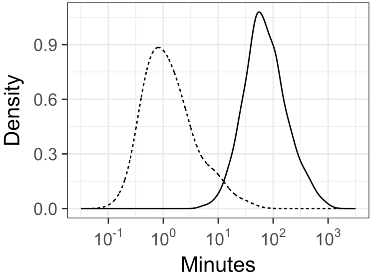


Figure: Mean IET

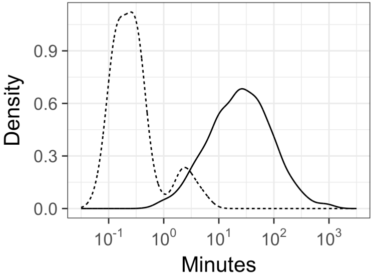
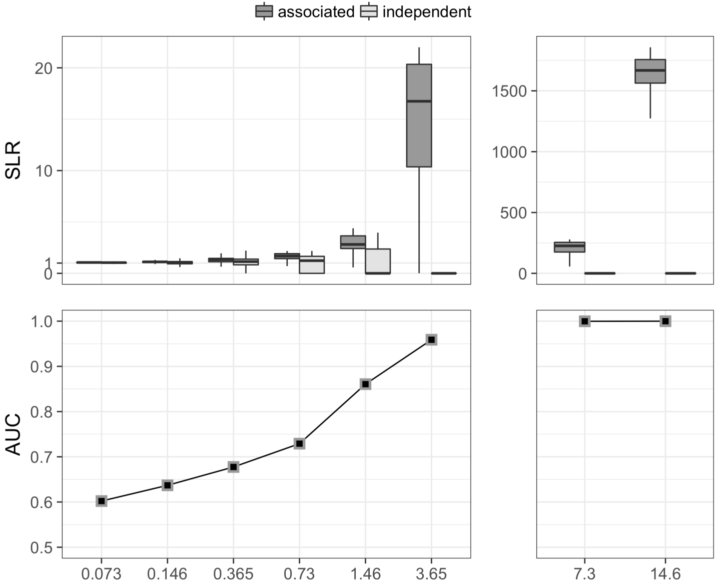
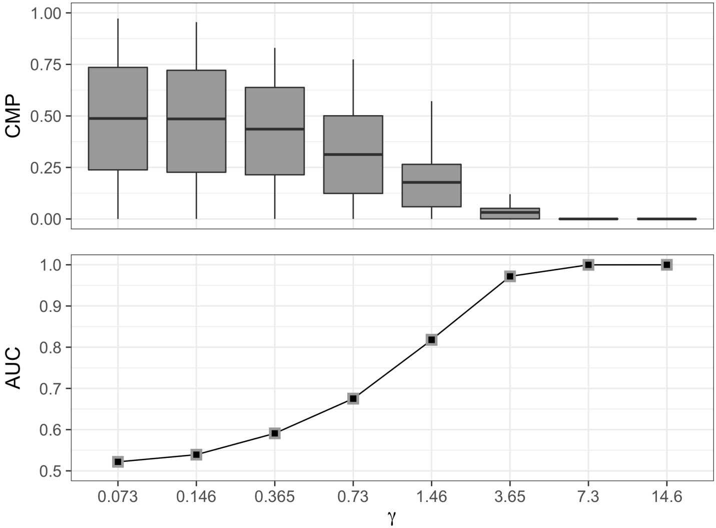


Figure: Median IET

Simulation Results



Simulation Results



Simulation Results

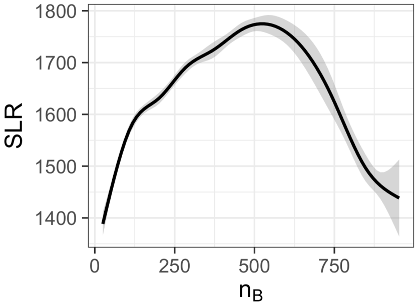


Figure: $\gamma = 14.6$

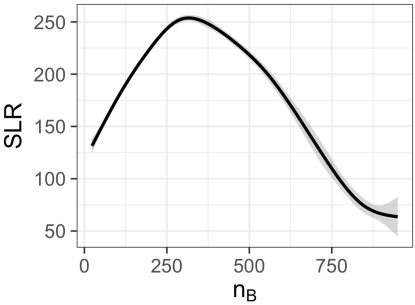


Figure: $\gamma = 7.3$