Quantifying the Association Between Discrete Event Time Series

Chris Galbraith†, Padhraic Smyth‡ & Hal S. Stern†

†Department of Statistics, ‡Department of Computer Science

Problem Statement

Consider a pair of user-generated event series

$$M = (A, B) = \{(t_j, m(t_j)) : j = 1, \ldots, n\}$$

where $t_j \in \mathbb{R}^+$ is the time and $m(t_j) \in \{\text{A}, \text{B}\}$ is the type of the j^{th} event. We want to quantify the likelihood that the pair was generated by the same source.

Measures of Association

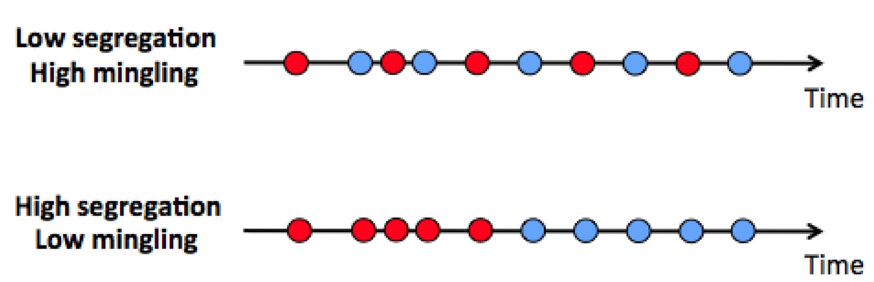
Score Functions using Nearest Neighbors

- *Coefficient of Segregation* [3]: function of the ratio of the probability that a reference point (i.e., a randomly selected event in (A, B)) and its nearest neighbor have different marks to the same probability for independent marks.

$$S(A, B) = 1 - \frac{p_{AB} + p_{BA}}{p_A p_{\cdot B} + p_B p_{\cdot A}} \in [-1, 1]$$

- *Mingling Index* [4]: mean fraction of points among the k nearest neighbors whose type is different than that of the reference point

$$\overline{M}_k(A, B) = \frac{1}{nk} \sum_{j=1}^{n} \sum_{\ell=1}^{k} \mathbb{I}\left[m(t_j) \neq m(z_\ell(t_j))\right] \in [0, 1]$$



Score Functions using Inter-Event Times

Assume that $n_B < n_A$ and fix series B. We then measure the time from each event in B to the closest event in series A in either direction

$$\mathcal{T}_{BA} \equiv \left\{\tau_{BA,j} : j = 1, \ldots, n_B\right\}$$

$$\text{where } \tau_{BA,j} = \min_{k \in \{1,\ldots,n_A\}} |t_{b,j} - t_{a,k}|$$

- Mean inter-event time from B to A

$$\overline{\mathcal{T}}_{BA} = \frac{1}{n_B} \sum_{j=1}^{n_B} \tau_{BA,j} \in (0, \infty)$$

- Median inter-event time from B to A

$$med(\mathcal{T}_{BA}) \in (0, \infty)$$

Population-based Approach

Given

- Pair of interest: (A^*, B^*)
- Score function: Δ
- Sample of N pairs of event time series with known sources: $M_i = (A_i, B_i)$ for $i = 1, \ldots, N$

Method

- 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 [1]

Resampling Approach

Given

- Pair of interest: (A^*, B^*)
- Score function: Δ

Method

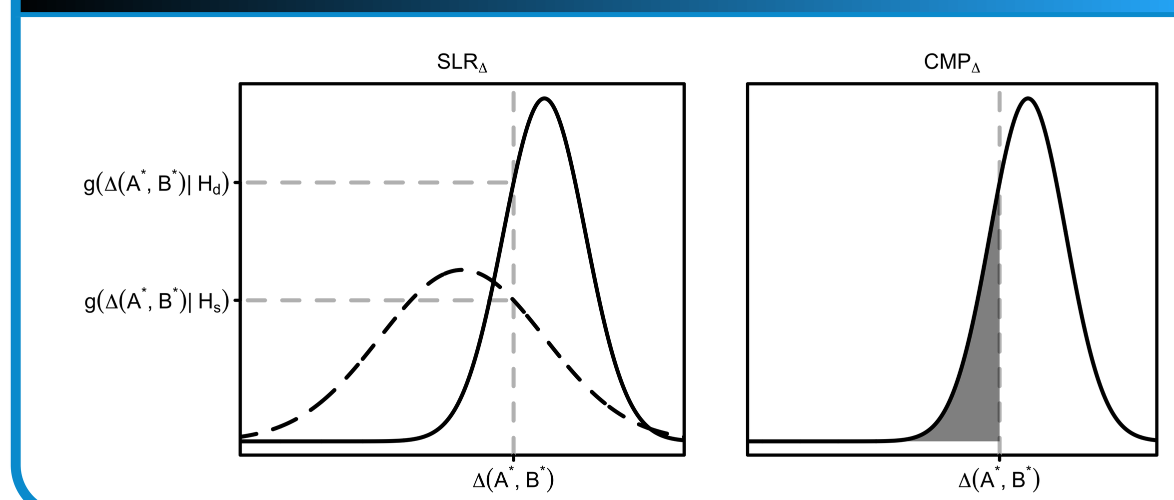
- Focus on the denominator of SLR_Δ
- *Coincidental match probability:* probability that a different-source pair with observed score $\Delta(A^*, B^*)$ exhibits association by chance

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

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

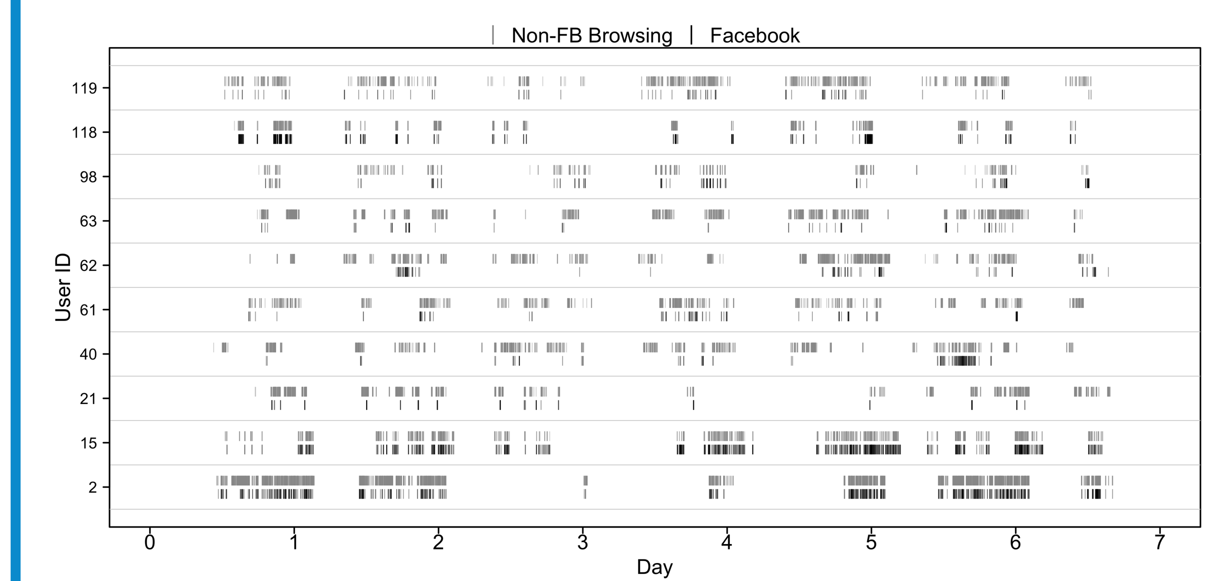
$$\widehat{CMP}_\Delta = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}[\Delta(A^{(i)}, B^{(i)}) < \Delta(A^*, B^*)]$$

Comparison of Approaches



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 [2]
- Event series created by dichotomizing browsing events to Facebook versus non-Facebook urls
- Only considered 55 students with at least 50 web browsing events of each type



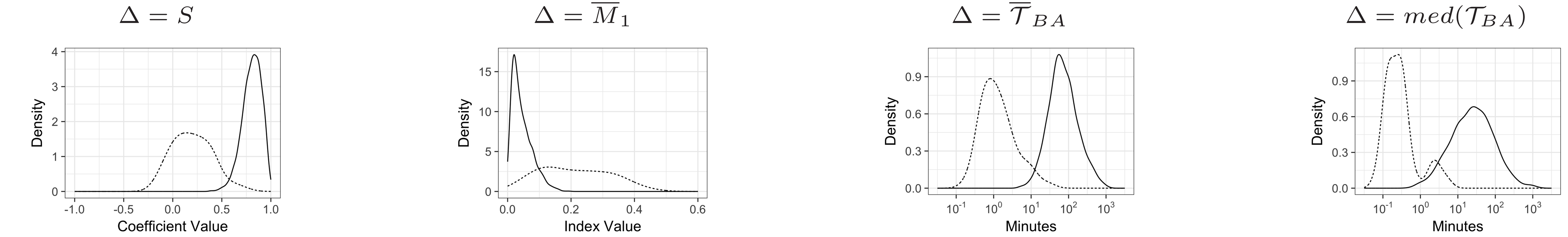
Performance of a classifier based on SLR_Δ

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

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

$\Delta = S$ $\quad$ $\Delta = \overline{M}_1$ $\quad$ $\Delta = \overline{\mathcal{T}}_{BA}$ $\quad$ $\Delta = med(\mathcal{T}_{BA})$

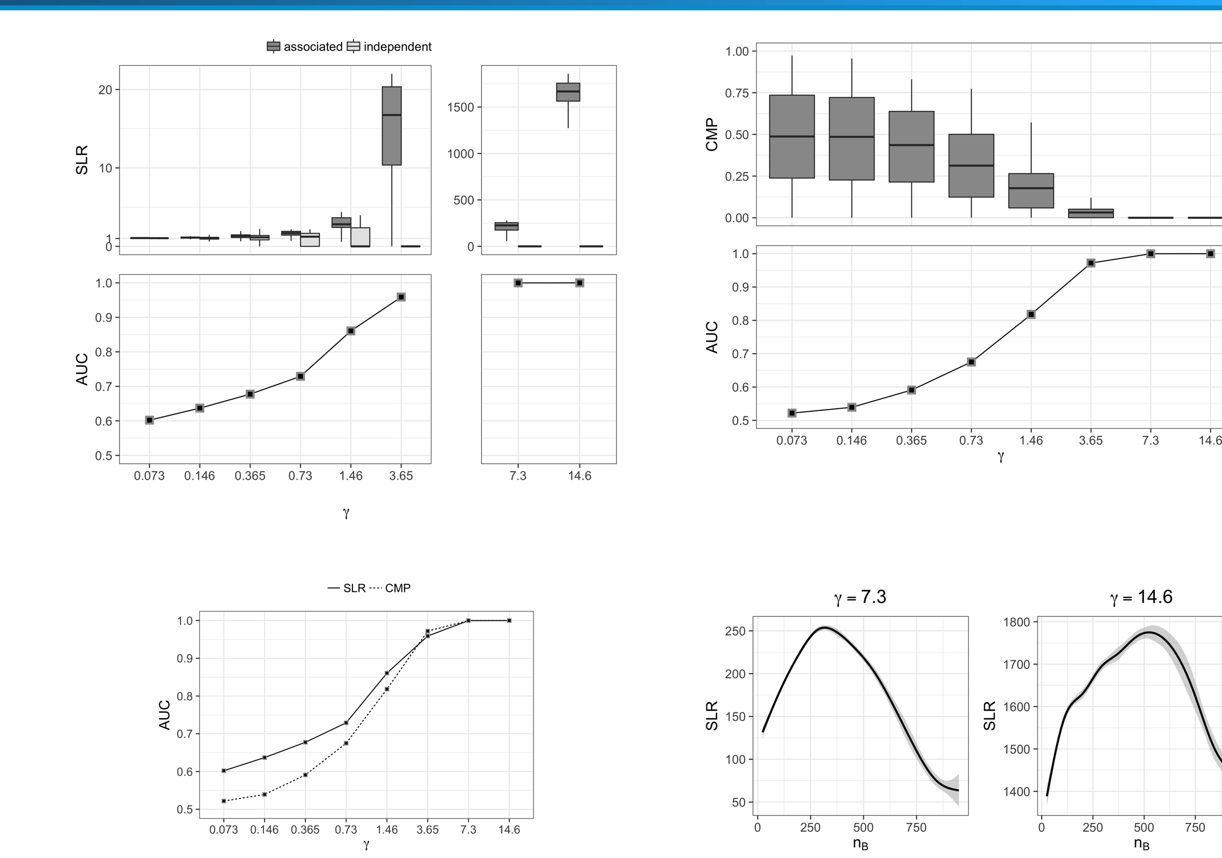


Simulation

- Simulated 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

$$\lambda_B = p\lambda_A \text{ with } p \in (0, 1)$$

  or w.p. p add Gaussian noise to event in A
- 10,000 independent & 10,000 associated pairs for each combination of $<p, \sigma, \text{distribution of } \lambda_A>$
- Most important factors in detecting associated pairs:
  - Number of events in series B: n_B
  - Signal-to-noise ratio: $\gamma = \overline{\mathcal{T}}_{AA}/\sigma$



Conclusions

- Resampling approach shows promise in situations where no reference data is available
- Population-based SLR is preferred, given
  - Better performance for weakly associated pairs
  - Similar performance for strongly associated pairs
  - Well-established in forensic investigation

References

[1] A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia, "Score-based likelihood ratios for handwriting evidence," *Forensic Science International*, vol. 219, no. 1, pp. 129 – 140, 2012.

[2] Y. Wang, M. Niiya, G. Mark, S. Reich, and M. Warschauer, "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, 2015.

[3] E. C. Pielou, *Mathematical Ecology*. John Wiley & Sons, Inc., 1977.

[4] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan, *Statistical Analysis and Modelling of Spatial Point Patterns*. England: John Wiley & Sons Ltd, 2008.

Acknowledgement

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.

Contact Information

Web www.ics.uci.edu/~galbraic
Email galbraic@uci.edu