# Spatial DNA: Measuring Similarity of Geolocation Data Sets

Christopher Galbraith & Padhraic Smyth; University of California, Irvine

## Problem Statement

Consider a pair of user-generated event point patterns

$$M = (A, B) = \{(x_i, m(x_i)) : i = 1, \dots, n\}$$

where $x_i \in \mathbb{R}^d$ is the location and $m(x_i) \in \{A, B\}$ is the type of the $i^{th}$ event. We want to quantify the likelihood that the pair was generated by the same source.

## Measures of Association

### Temporal Point Patterns

- Score functions using nearest neighbors: Compute summary statistics of marks for neighborhoods around randomly selected points in the pattern. We utilize the *coefficient of segregation* $S$.
- Score functions using inter-event times: Measure the time from each event in $B$ to the closest event in $A$ in either direction

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

where $\tau_{BA,j} = \min_{j \in \{1,\dots,n_A\}} |x_{b,k} - x_{a,j}|$ and $x \in \mathbb{R}^+$

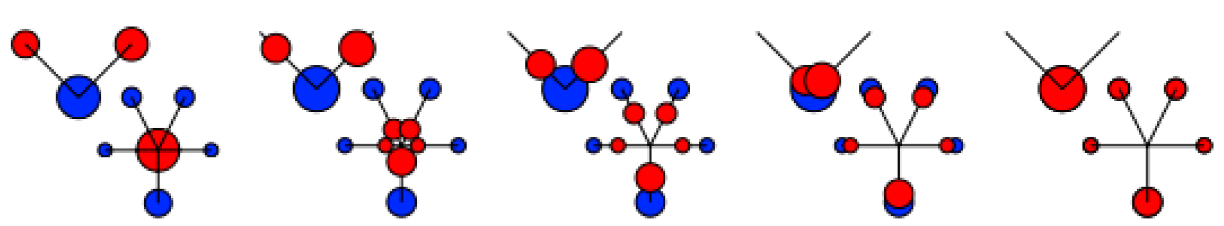
We compute either the mean $\overline{\mathcal{T}}_{BA}$ or median $med(\mathcal{T}_{BA})$.

### Spatial Point Patterns

We utilize the *earth mover's distance* [1] applied to the empirical distribution of point patterns $A$ and $B$. Given the locations and weights of the points, EMD is the solution to an optimal transport problem for transforming points in one set ($A$) to those of another ($B$)

$$EMD(A, B) = \min_{F \in \mathbb{F}(A,B)} \sum_{j=1}^{n_A} \sum_{k=1}^{n_B} f_{jk} d_{jk}$$

where $F$ is a member of the set of feasible flows from $A$ to $B$, $\mathbb{F}(A, B)$, thus the optimization is constrained.



## Population-based Approach [2]

- 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 $i = 1, \dots, N$ to estimate the *score-based likelihood ratio*

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

## Resampling Approach [3]

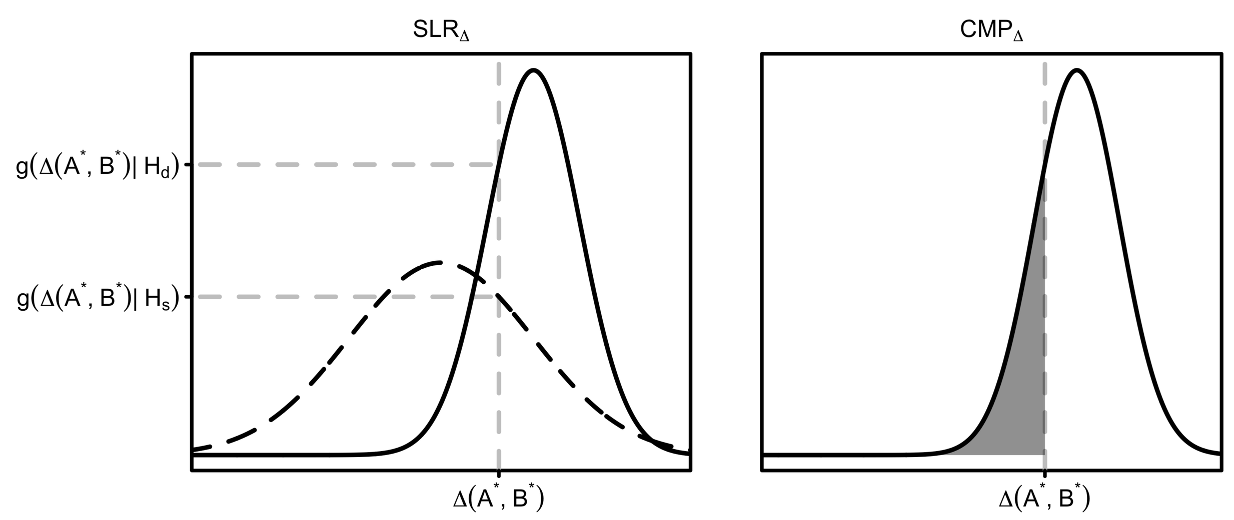
- Focus on the denominator of $SLR_\Delta$
- *Coincidental match probability:* probability that a different-source pair with observed score $\Delta(A^*, B^*)$ exhibits association by chance

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

- Estimator: simulate different-source pairs to compute

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

## Comparison of Approaches




## References

[1] Scott Cohen. *Finding color and shape patterns in images.* Number 1620. Stanford University, Department of Computer Science, 1999.

[2] Christopher Galbraith and Padhraic Smyth. Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digital Investigation*, 22:S106–S114, 2017.

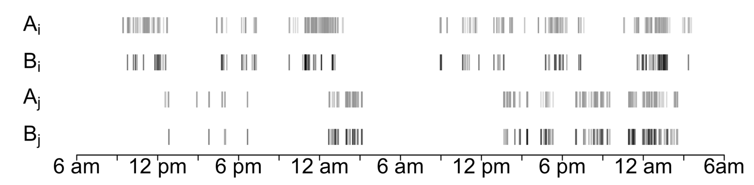
[3] Christopher Galbraith, Padhraic Smyth, and Hal S. Stern. Quantifying the association between discrete event time series with applications to digital forensics. *Submitted to J. R. Stat. Soc. A*, 2019.

[4] Moshe Lichman. *Context-Based Smoothing for Personlized Prediction Models.* UC Irvine, 2017.

## Temporal Applications

### UCI Student Data

Marks correspond to user-generated events on different domains collected by browser logging software.




| $\Delta$ | TP@1 | FP@1 | AUC |
|---|---|---|---|
| Segregation (S) | 0.945 | 0.031 | 0.992 |
| Mean IET | **0.964** | **0.029** | **0.996** |
| Median IET | **0.964** | 0.085 | 0.992 |

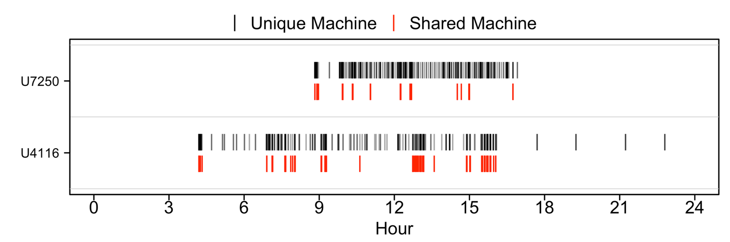
Performance of a classifier based on $\widehat{SLR}_\Delta$

| $\Delta$ | TP@.05 | FP@.05 | TP@.001 | FP@.001 | AUC |
|---|---|---|---|---|---|
| Mean IET | **1.000** | **0.036** | 0.982 | **0.002** | **0.999** |
| Median IET | **1.000** | 0.176 | **1.000** | 0.015 | 0.992 |

Performance of a classifier based on $\widehat{CMP}_\Delta$

### LANL Authentication Data

Marks correspond to logins on different computers in the Los Alamos National Laboratory.

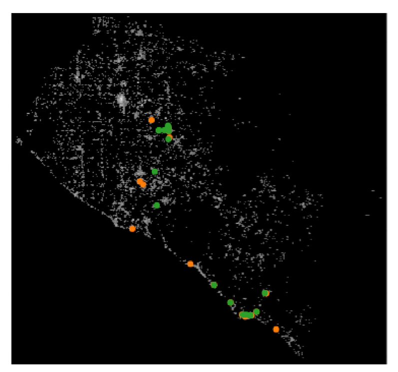



| Match | Unique | Shared | S | Mean IET | Median IET |
|---|---|---|---|---|---|
| Y | U4116 | U4116 | 0.000 | 0.000 | 0.000 |
| | U7250 | U7250 | 0.000 | 0.000 | 0.000 |
| N | U4116 | U7250 | 0.952 | 0.197 | 0.637 |
| | U7250 | U4116 | 0.333 | 0.545 | 0.266 |

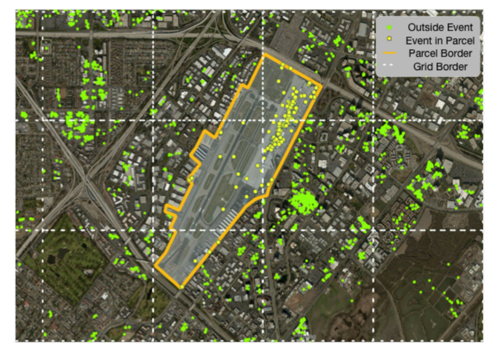
$\widehat{CMP}_\Delta$

## Spatial Application

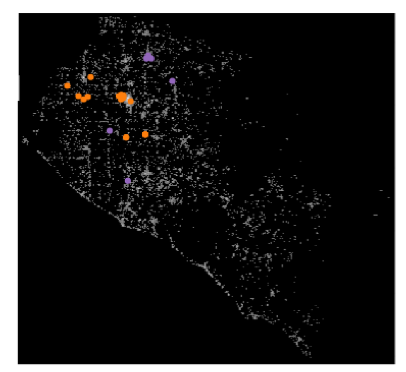
- Geolocated event data collected from Twitter users in southern California in July and August 2013
- Filtered to "visits" (grouped nearby tweets within an hour window as one effective event).
  - Population: 546k visits, 103k users
  - Point pattern: 28k visits; 223 users with at least 20 visits in successive months
- Geoparcel data collected from the Southern California Association of Government



Same-source pair



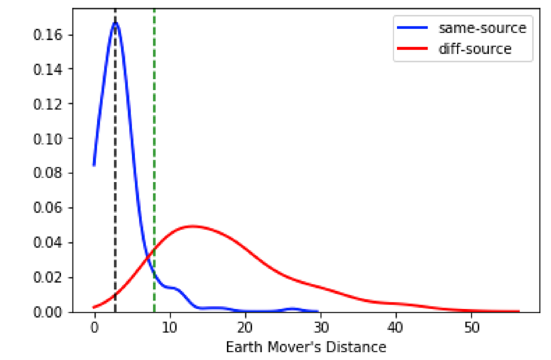
John Wayne Airport parcel [4]



Different-source pair

| $\Delta$ | TP@.05 | FP@.05 | TP@.001 | FP@.001 | AUC |
|---|---|---|---|---|---|
| Mean IED | 0.441 | 0.008 | 0.257 | 0.004 | 0.931 |
| Median IED | 0.230 | **0.004** | 0.153 | 0.004 | 0.822 |
| EMD | **0.734** | 0.028 | **0.423** | 0.004 | **0.948** |

Performance of a classifier based on $\widehat{CMP}_\Delta$ using all same-source pairs and a random sample of 250 different-source pairs.




This work was partially funded by CSAFE through Cooperative Agreement # 70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia