Statistical Methods for the Forensic Analysis of Geolocated Event Data

By: Christopher Galbraith (Department of Statistics, University of California, Irvine), Padhraic Smyth (Department of Computer Science, University of California, Irvine), and Hal S. Stern (Department of Statistics, University of California, Irvine)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment.

As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development. 

https://dfrws.org
Statistical Methods for the Forensic Analysis of Geolocated Event Data

Christopher Galbraith
Padhraic Smyth
Hal S. Stern

DFRWS USA
7.22.20
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.
Outline

1  Motivation

2  Quantifying Strength of Evidence

3  Application to Geolocated Event Data

4  Future Directions and Conclusions
Motivation
DNA Samples

[Steele & Balding, 2014]
### DNA Samples

<table>
<thead>
<tr>
<th></th>
<th>Crime Scene</th>
<th>Suspect #1</th>
<th>Suspect #2</th>
<th>Suspect #3</th>
</tr>
</thead>
</table>

\[
i \Pr(\ell, I = \bullet | = ) = 1
\]

[Steele & Balding, 2014]
DNA Samples

[i] \( \Pr(\text{crime}, \text{suspect #1} | \text{suspect #2} = \text{?}) = 1 \)

[ii] \( \Pr(\text{crime}, \text{suspect #1} | \text{suspect #2} \neq \text{?}) \)

[Steele & Balding, 2014]
DNA Samples

\[ \Pr(\text{crime scene}, \text{suspect } \#1 = \text{?}) = 1 \]

\[ \Pr(\text{crime scene}, \text{suspect } \#1 \neq \text{?}) \]

[Steele & Balding, 2014]
DNA Samples

\[ \text{Pr}(\text{crime scene}, \text{suspect } \#1 | \text{suspect } \#2 = \text{suspect } \#3) = 1 \]

\[ \text{Pr}(\text{crime scene}, \text{suspect } \#1 | \text{suspect } \#2 \neq \text{suspect } \#3) \text{ CODIS Random Match Probability} \]

[Steele & Balding, 2014]
Samples from different sources

$\Pr(\text{crime scene}, \text{suspect #1} | \neq \text{suspect #2}) = 1$

Random Match Probability

CODIS

$\Pr(\text{crime scene}, \text{suspect #1} | \neq \text{suspect #2})$

Likelihood Ratio

$\frac{i}{ii}$

$< 1$ Samples from different sources

$= 1$ Inconclusive

$> 1$ Samples from same source

[Steele & Balding, 2014]
### DNA Samples

<table>
<thead>
<tr>
<th>Source</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime Scene</td>
<td></td>
</tr>
<tr>
<td>Suspect #1</td>
<td></td>
</tr>
<tr>
<td>Suspect #2</td>
<td></td>
</tr>
<tr>
<td>Suspect #3</td>
<td></td>
</tr>
</tbody>
</table>

**Likelihood Ratio**

\[
\Pr(\text{DNA}, \text{scene} | \text{?} = \text{?}) = 1
\]

- **Random Match Probability**: \(< 1\)
- **Samples from different sources**: Inconclusive
- **Samples from same source**: \(\Pr(\text{DNA}, \text{scene} | \text{?} = \text{?}) = 1\)
Extraction

[SWDGE, 2019; Roussev, 2016; Casey, 2011]

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

Extraction

SWDGE, 2019;
Roussev, 2016;
Casey, 2011

Analysis

Visualization

Buchholz and Falk, 2005;
Grier, 2011;
Koven et al., 2016;
Gresty et al., 2016;
Kirchler et al., 2016
Extraction

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

Probabilistic conclusions regarding source, e.g., Likelihood Ratio

Analysis & Visualization

[Buchholz and Falk, 2005; Grier, 2011; Koven et al., 2016; Gresty et al., 2016; Kirchler et al., 2016]
Probabilistic conclusions regarding source, e.g., Likelihood Ratio

TOPIC OF DFRWS PAPER

Analysis & Visualization

[Buchholz and Falk, 2005; Grier, 2011; Koven et al., 2016; Gresty et al., 2016; Kirchler et al., 2016]

Extraction

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

[SWDGE, 2019; Roussev, 2016; Casey, 2011]
BACKGROUND

Statistical Approaches for Evaluating Forensic Evidence
Assess the likelihood of observing

Under two competing hypotheses
Assess the likelihood of observing $A, B$ under two competing hypotheses:

- $H_s$: $(A, B)$ came from the same source
- $H_d$: $(A, B)$ came from the different sources
Wait...why aren’t we interested in the probability of the source hypothesis *given the evidence*?
Wait…why aren’t we interested in the probability of the source hypothesis given the evidence?

\[
\frac{Pr(H_s | A, B)}{Pr(H_d | A, B)} \quad \text{posterior odds}
\]
Wait...why aren’t we interested in the probability of the source hypothesis given the evidence?

\[
\frac{Pr(H_s | A, B)}{Pr(H_d | A, B)} = \frac{\text{likelihood ratio}}{\frac{Pr(A, B | H_s)}{Pr(A, B | H_d)}} \cdot \frac{Pr(H_s)}{Pr(H_d)}
\]

posterior odds

prior odds
\[
\frac{\Pr(H_s \mid A, B)}{\Pr(H_d \mid A, B)} = \frac{\Pr(A, B \mid H_s)}{\Pr(A, B \mid H_d)} \cdot \frac{\Pr(H_s)}{\Pr(H_d)}
\]

**likelihood ratio**

**posterior odds**

**prior odds**
\[
\frac{Pr(H_s | A, B)}{Pr(H_d | A, B)} = \frac{Pr(A, B | H_s)}{Pr(A, B | H_d)} \cdot \frac{Pr(H_s)}{Pr(H_d)}
\]

posterior odds

likelihood ratio

prior odds

[Stern, 2017; Thompson, 2017]
“Strength of Evidence”

\[
\frac{Pr(H_s | A, B)}{Pr(H_d | A, B)} = \frac{Pr(A, B | H_s)}{Pr(A, B | H_d)}
\]

posterior odds

“Weight of Evidence”

\[
\frac{Pr(H_s)}{Pr(H_d)}
\]

prior odds

[Stern, 2017; Thompson, 2017]
The Likelihood Ratio

Widely accepted as a “logically defensible way” to assess the strength of evidence [Willis et al., 2016]
The Likelihood Ratio

- Widely accepted as a “logically defensible way” to assess the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
The Likelihood Ratio

- Widely accepted as a “logically defensible way” to assess the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]
The Likelihood Ratio

- Widely accepted as a “logically defensible way” to assess the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]

- Studies demonstrating its understanding
The Likelihood Ratio

- Widely accepted as a “logically defensible way” to assess the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]

- Studies demonstrating its understanding
  - Misconceptions [Martire et al., 2013, Thompson and Newman, 2015, Thompson et al., 2018]
The Likelihood Ratio

- Widely accepted as a “logically defensible way” to assess the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]

- Studies demonstrating its understanding
  - Misconceptions [Martire et al., 2013, Thompson and Newman, 2015, Thompson et al., 2018]
  - Verbal Equivalents [e.g., AFSP, 2009]
Why not always use the LR?
Why not always use the LR?

- Complexity: Evidence can be high-dimensional

[Stern, 2017]
Why not always use the LR?

- **Complexity**: Evidence can be high-dimensional

- **Feature Selection**: Wide variety of features to consider

[Fine, 2016]  

[Stern, 2017]
Why not always use the LR?

- **Complexity:** Evidence can be high-dimensional

- **Feature Selection:** Wide variety of features to consider

- **Appropriate Probability Models:** Must describe variation within a given source and between different sources

\[ H_s : ? = \neq H_d : ? \]  

[Stern, 2017]
Why not always use the LR?

- **Complexity:** Evidence can be high-dimensional

- **Feature Selection:** Wide variety of features to consider

- **Appropriate Probability Models:** Must describe variation within a given source and between different sources

- **Reference Population:** Difficult to identify a *relevant* reference population to estimate model parameters & perform validation studies

[Stern, 2017]
Score-based Approaches

Measure similarity between $A$ and $B$ via a score function

$\Delta(A, B)$
Score-based Approaches

- Measure similarity between $A$ and $B$ via a score function $\Delta(A, B)$

- Score-based Likelihood Ratio: Compute a LR for the observed score

$$SLR_\Delta = \frac{g(\Delta(A, B) = \delta \mid H_s)}{g(\Delta(A, B) = \delta \mid H_d)}$$

![Same-Source vs Different-Source Likelihood Ratios](image)

$g(\Delta(A, B) = \delta \mid H_s, I)$

$g(\Delta(A, B) = \delta \mid H_d, I)$
Score-based Approaches

- Measure similarity between $A$ and $B$ via a score function
  \[ \Delta(A, B) \]

- Score-based Likelihood Ratio: Compute a LR for the observed score
  \[
  SLR_{\Delta} = \frac{g(\Delta(A, B) = \delta | H_s)}{g(\Delta(A, B) = \delta | H_d)}
  \]

Gaining popularity in a variety of forensic disciplines

- Chemical Concentrations [Bolck et al., 2015]
- Speaker Recognition [Gonzalez-Rodriguez et al., 2007]
- Fingerprints [Alberink et al., 2013; Neumann et al., 2015]
- Handwriting [Hepler et al., 2012]
Evidence Evaluation Approaches

- **Likelihood Ratio:** Models evidence directly

  \[
  LR = \frac{Pr(A, B | H_s)}{Pr(A, B | H_d)}
  \]

- **Score-based Likelihood Ratio:** Models low-dimensional summary of the evidence, \( \Delta(A, B) \)

  \[
  SLR_{\Delta} = \frac{g(\Delta(A, B) = \delta | H_s)}{g(\Delta(A, B) = \delta | H_d)}
  \]
Evidence Evaluation Approaches

- **Likelihood Ratio**: Models evidence directly
  \[ LR = \frac{Pr(A, B | H_s)}{Pr(A, B | H_d)} \]

- **Score-based Likelihood Ratio**: Models low-dimensional summary of the evidence, \( \Delta(A, B) \)
  \[ SLR_\Delta = \frac{g(\Delta(A, B) = \delta | H_s)}{g(\Delta(A, B) = \delta | H_d)} \]
CONTRIBUTION

Quantifying the Strength of Geolocated Event Evidence
Geofence Warrants
Revisiting the LR

\[ LR = \frac{Pr(A, B | H_s)}{Pr(A, B | H_d)} \]

Unknown source events

Known source events
Revisiting the LR

\[ LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} \Rightarrow \cdots \Rightarrow LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)} \]

Maths

A

Unknown source events

B

Known source events
Revisiting the LR

\[ LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} \Rightarrow \ldots \Rightarrow LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)} \]

\( LR = \)

Prob of

Known source events

Prob of

Unknown source events

\& \neq 

Random match probability

Maths

Unknown source events

Known source events
Known source events

\( B \)

Prob of
given

\( \neq \)

Random match probability
Adaptive Bandwidth
Kernel Density Estimators

[Breiman et al., 1977]
Unknown source events

Known source events

Prob of given & $\mathbb{Q}$ =
Individual Component

\[ \text{Prob of given } B \]

- **Known source events**
- **Unknown source events**

\[ & \text{?} = \]
Individual Component

Population Component

Unknown source events

Known source events

Prob of given & =

A

B

[Lichman & Smyth, 2014]
Individual Component

Population Component

Mixture
80% individual, 20% population

Prob of Known source events given

& ? =

[Lichman & Smyth, 2014]
$LR = \frac{\text{Prob of } B \text{ given known source events}}{\text{Prob of } B \text{ given random match probability}}$
LR = A

Unknown source events = B

Known source events ≠ Random match probability

Prob of

match


**Geofence Warrants Revisited**

\[ LR \approx 1137 \]

**Strong support for same-source hypothesis**

\[ LR \approx 2.8 \times 10^{-28} \]

**Exclusion**
Case Study

Collected Twitter data from May 2015 to Feb 2016

Orange County, CA
Manhattan, New York, NY

$A$ and $B$ are consecutive months from the same account

<table>
<thead>
<tr>
<th>Region</th>
<th>Accounts</th>
<th>Visits in $A$</th>
<th>Visits in $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC</td>
<td>6,714</td>
<td>44,310 (6.6)</td>
<td>38,697 (5.8)</td>
</tr>
<tr>
<td>NY</td>
<td>13,523</td>
<td>72,799 (5.4)</td>
<td>65,852 (4.9)</td>
</tr>
</tbody>
</table>
Case Study

- Collected Twitter data from May 2015 to Feb 2016
  - Orange County, CA
  - Manhattan, New York, NY

- $A$ and $B$ are consecutive months from the same account

<table>
<thead>
<tr>
<th>Region</th>
<th>Accounts</th>
<th>Visits in $A$</th>
<th>Visits in $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC</td>
<td>6,714</td>
<td>44,310 (6.6)</td>
<td>38,697 (5.8)</td>
</tr>
<tr>
<td>NY</td>
<td>13,523</td>
<td>72,799 (5.4)</td>
<td>65,852 (4.9)</td>
</tr>
</tbody>
</table>

- Results based on stratified sample based on $n_a$ and $n_b$ for different-source evidence
<table>
<thead>
<tr>
<th>Region</th>
<th>Method</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC</td>
<td>LR</td>
<td>0.380</td>
<td>0.038</td>
<td>0.845</td>
</tr>
<tr>
<td></td>
<td>SLR</td>
<td>0.614</td>
<td>0.162</td>
<td>0.783</td>
</tr>
<tr>
<td>NY</td>
<td>LR</td>
<td>0.285</td>
<td>0.089</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td>SLR</td>
<td>0.511</td>
<td>0.235</td>
<td>0.685</td>
</tr>
</tbody>
</table>

(1) LR with $\alpha(n_a)$ weights; SLR using earth mover’s distance with account weights
(2) LR & SLR threshold is 1
## Results

<table>
<thead>
<tr>
<th>Region</th>
<th>Method$^1$</th>
<th>TP Rate$^2$</th>
<th>FP Rate$^2$</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC</td>
<td>LR</td>
<td>0.380</td>
<td>0.038</td>
<td>0.845</td>
</tr>
<tr>
<td></td>
<td>SLR$_{EMD}$</td>
<td><strong>0.614</strong></td>
<td>0.162</td>
<td>0.783</td>
</tr>
<tr>
<td>NY</td>
<td>LR</td>
<td>0.285</td>
<td><strong>0.089</strong></td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td>SLR$_{EMD}$</td>
<td><strong>0.511</strong></td>
<td>0.235</td>
<td>0.685</td>
</tr>
</tbody>
</table>

(1) LR with $\alpha(n_q)$ weights; SLR using earth mover’s distance with account weights
(2) LR & SLR threshold is 1

![OC](image OC)

![NY](image NY)
Future Directions and Summary
Future Directions

- **Reference Data:** Collect & share relevant digital data amongst law enforcement & researchers, e.g., start to build CODIS-like databases.

- **Discovery:** Finding the most likely known source in a database given an unknown source sample...quickly.
Summary

- **Statistical approaches** play a key role in the **forensic analysis** of a wide variety of evidence.

- **Digital evidence** is lagging behind other forensic disciplines.

- Presented a framework for estimating **likelihood ratios** for forensic applications with **geolocated event data**.
Questions
Appendix
Score Functions

- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - Distance-based
  - Area-based

- Use distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets $A$ and $B$

- Incorporate area-based information via weights $\Omega^a$, $\Omega^b$
Earth-mover’s distance

\[ EMD(B, A \mid \Omega^b, \Omega^a) \]
Earth-mover’s distance

\[ EMD(B, A | \Omega^b, \Omega^a) \]
<table>
<thead>
<tr>
<th>Region</th>
<th>Weight</th>
<th>TP@1</th>
<th>FP@1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC</td>
<td>0.80</td>
<td>0.340</td>
<td>0.026</td>
<td>0.787</td>
</tr>
<tr>
<td></td>
<td>$\alpha(n_a)$</td>
<td><strong>0.380</strong></td>
<td>0.038</td>
<td><strong>0.845</strong></td>
</tr>
<tr>
<td></td>
<td>$\alpha(n_a</td>
<td>\gamma, \rho, \phi)$</td>
<td>0.375</td>
<td>0.037</td>
</tr>
<tr>
<td>NY</td>
<td>0.80</td>
<td>0.251</td>
<td>0.067</td>
<td>0.711</td>
</tr>
<tr>
<td></td>
<td>$\alpha(n_a)$</td>
<td><strong>0.285</strong></td>
<td>0.089</td>
<td><strong>0.768</strong></td>
</tr>
<tr>
<td></td>
<td>$\alpha(n_a</td>
<td>\gamma, \rho, \phi)$</td>
<td>0.282</td>
<td>0.088</td>
</tr>
</tbody>
</table>

### Region | $\Delta$ | Weights | TP@1 | FP@1 | AUC  |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{\text{min}}$</td>
<td>Uniform</td>
<td>0.628</td>
<td>0.202</td>
<td>0.768</td>
<td></td>
</tr>
<tr>
<td>$D_{\text{min}}$</td>
<td>Account</td>
<td>0.610</td>
<td>0.171</td>
<td>0.774</td>
<td></td>
</tr>
<tr>
<td>$D_{\text{min}}$</td>
<td>Visit</td>
<td>0.611</td>
<td>0.180</td>
<td>0.768</td>
<td></td>
</tr>
<tr>
<td>OC</td>
<td>EMD</td>
<td>Uniform</td>
<td><strong>0.654</strong></td>
<td>0.197</td>
<td><strong>0.790</strong></td>
</tr>
<tr>
<td></td>
<td>EMD</td>
<td>Account</td>
<td>0.614</td>
<td><strong>0.162</strong></td>
<td>0.783</td>
</tr>
<tr>
<td></td>
<td>EMD</td>
<td>Visit</td>
<td>0.602</td>
<td>0.169</td>
<td>0.774</td>
</tr>
<tr>
<td>$D_{\text{min}}$</td>
<td>Uniform</td>
<td>0.508</td>
<td>0.287</td>
<td>0.656</td>
<td></td>
</tr>
<tr>
<td>$D_{\text{min}}$</td>
<td>Account</td>
<td>0.494</td>
<td>0.254</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>$D_{\text{min}}$</td>
<td>Visit</td>
<td>0.493</td>
<td>0.257</td>
<td>0.663</td>
<td></td>
</tr>
<tr>
<td>NY</td>
<td>EMD</td>
<td>Uniform</td>
<td><strong>0.530</strong></td>
<td>0.253</td>
<td><strong>0.686</strong></td>
</tr>
<tr>
<td></td>
<td>EMD</td>
<td>Account</td>
<td>0.511</td>
<td>0.235</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>EMD</td>
<td>Visit</td>
<td>0.504</td>
<td><strong>0.234</strong></td>
<td>0.679</td>
</tr>
</tbody>
</table>

**LR**

**SLR**
The image contains two scatter plots, each with a title and labeled axes. The plots are described as 'Likelihood Ratio' and 'Score-based Likelihood Ratio'. The axes are labeled 'True Positive Rate' and 'False Positive Rate'. Each plot contains multiple data points, each represented by different symbols and colors, indicating various parameters or conditions.

For the 'OC' plot, the symbols and colors correspond to different parameter ranges, such as $[1.0, 2.0]$, $[2.0, 20.0]$, and $[20.0, \text{inf}]$. There are also notes for 'Mixing Weight' with values $\alpha = 0.80$ and $\alpha(n,y,\rho,\phi)$.

For the 'NY' plot, similar parameter ranges and symbols are used, with notes for 'Mixing Weight' and specific values for $\alpha$.

The plots provide a visual representation of how different parameter settings affect the likelihood ratios, allowing for the identification of trends and patterns in the data.