Statistical Analysis of User-Event Data in a Digital Forensics Context

Christopher Galbraith

Advised by Padhraic Smyth & Hal Stern

Departments of Statistics & Computer Science
University of California, Irvine

September 6, 2017

Overview

1. User-Generated Event Data
   - Concepts & Examples
   - Formulating Digital Forensics Questions

2. Statistical Methodology
   - Likelihood Ratio
   - Marked Point Processes

3. Case Study

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 >

Action type:

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

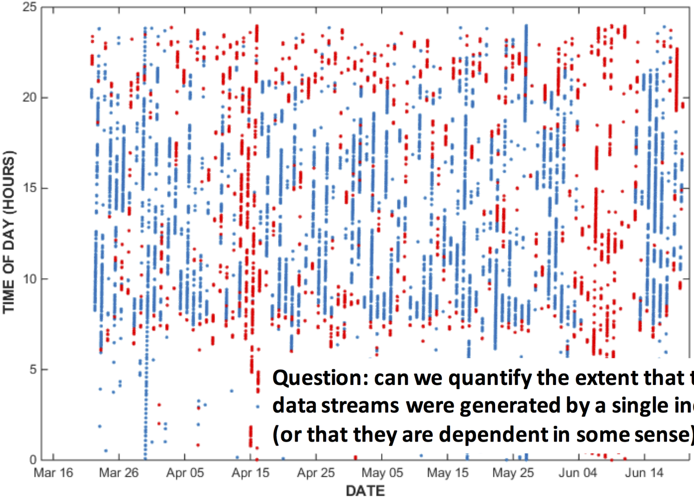
Metadata:

- Text content
- Location
- List of recipients
- .....

We focus on ID, timestamp, and type of actions

URL Visits (desktop)

URL Visits (laptop)



Question: can we quantify the extent that these data streams were generated by a single individual? (or that they are dependent in some sense)

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 in the forensics community

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

csafe
Center for Statistics and Applications in Forensic Evidence

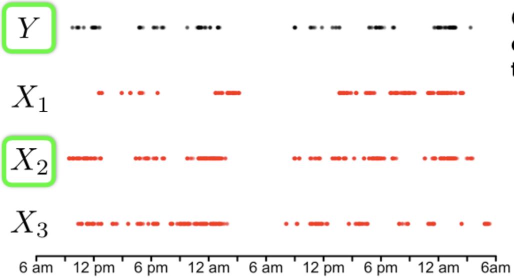
Overview

1. User-Generated Event Data
   - Concepts & Examples
   - Formulating Digital Forensics Questions
2. Statistical Methodology
   - Likelihood Ratio
   - Marked Point Processes
3. Case Study

The Likelihood Ratio

- Probabilistic framework for assessing if two samples came from the same source or not
- *LR* techniques have seen a great deal of attention in forensics as a whole
  - DNA analysis (Foreman et al., 2003)
  - Glass fragment analysis (Aitken & Lucy, 2004)
  - Fingerprint analysis (Neumann et al., 2007)
  - Handwriting analysis (Schlapbach & Bunke, 2007)
  - Analysis of illicit drugs (Bolck et al., 2015)

Likelihood Ratio



Compute the *likelihood* (or probability) of observing pairs of sequences under two assumptions:

1. **Samples are from the *same* person**

$$Pr(\{X_i, Y\}|H_s)$$

2. **Samples are from *different* people**

$$Pr(\{X_i, Y\}|H_d)$$

Likelihood Ratio

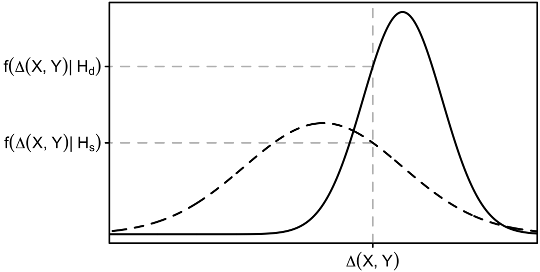
$$\frac{Pr(\{X_i, Y\}|H_s)}{Pr(\{X_i, Y\}|H_d)}$$

< 1	Samples from different sources
≈ 1	Inconclusive
> 1	Samples from same source

Score-based Likelihood Ratios

Problem: LR can be difficult to estimate.
Solution: Estimate the probability density function f of a *score function* Δ that measures the similarity of the samples X and Y, yielding the *score-based likelihood ratio*

$$SLR_{\Delta} = \frac{f(\Delta(X, Y)|H_s)}{f(\Delta(X, Y)|H_d)}$$

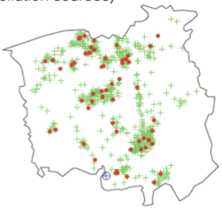


Score Functions

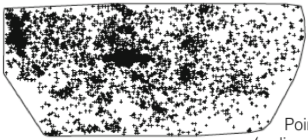
We use techniques from the analysis of *marked point processes*

- Data characterized by (a) time or location and (b) "marks"
- Marks can be continuous (e.g., height of a tree) or categorical (e.g., species of tree)
- Significant prior work in 2 dimensions (spatial data)
- Typically found in forestry, sociology, ecology, astronomy, etc.

Marked point pattern data
(disease cases and pollution sources)



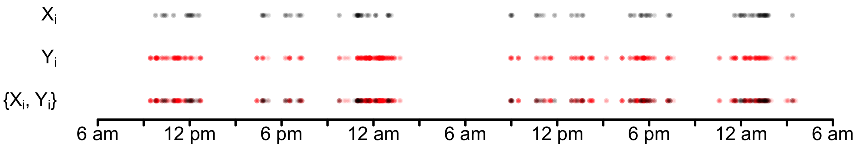
Point pattern data
(radioastronomical survey)



Figures from A. Baddeley, Spatial Point Patterns: Models and Statistics, 2012

User-Event Histories as Marked Point Processes

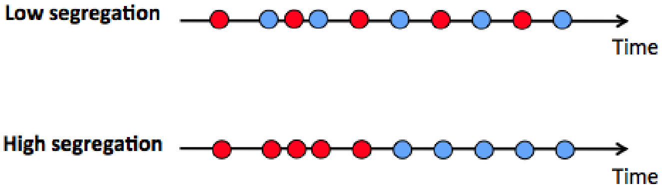
- Event streams can be viewed as marked point processes with the following properties
  - Temporal (i.e., time-stamped events)
  - Binary marks corresponding to the type of event
- Referred to as *bivariate point processes* in the literature



Coefficient of Segregation (Pielou, 1977)

Function of the ratio of observed probability that the reference point and its nearest neighbor have different marks to the same probability for independent marks

$$S(X_i, Y_i) = 1 - \frac{p_{xy} + p_{yx}}{p_x p_{\cdot y} + p_y p_{\cdot x}} \in [-1, 1]$$



Overview

1. User-Generated Event Data
   - Concepts & Examples
   - Formulating Digital Forensics Questions
2. Statistical Methodology
   - Likelihood Ratio
   - Marked Point Processes
3. Case Study

Case Study

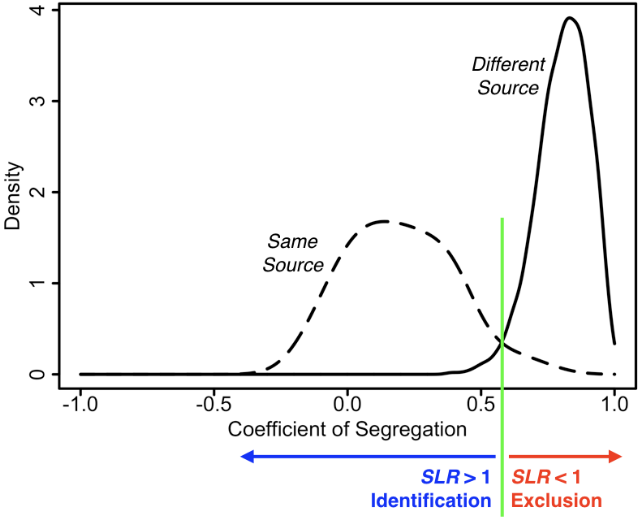
- Data from a 2013-2014 study at UCI that recorded students' browser activity for one week (Wang et al., 2015)
- Marks from dichotomized browser activity (Facebook vs. non-Facebook urls)
- Considered 55 students with at least 50 events of each type
- See our paper for more details:
  Galbraith, C., and Smyth, P. (2017). "Analyzing user-event data using score-based likelihood ratios with marked point processes." *Digital Investigation*, 22, S106-S114.


Facebook
Non-FB Browsing
Student ID
122
121
119
118
117
114
113
109
105
103
101
98
95
91
90
88
87
85
82
80
75
63
62
61
59
58
53
48
46
43
42
40
38
37
35
34
33
30
29
28
26
24
22
21
20
19
18
17
15
14
13
11
10
4
2
0
1
2
3
4
5
6
7
Day

Method

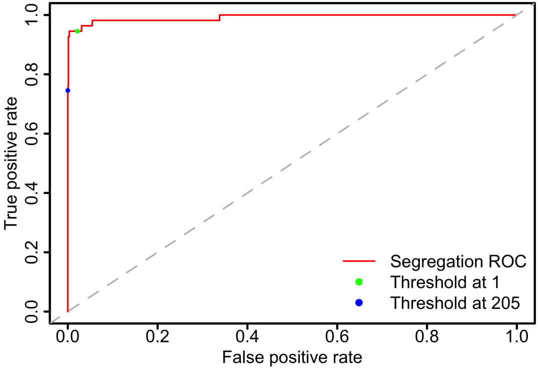
- Compute bivariate process indices for all N^2 pairwise combinations of user event streams
- For each pair $\{X_i, Y_j : i, j = 1, \ldots, N\}$ evaluate SLR_S with empirical likelihoods estimated from all *other* data
  - Leave out all event streams from users i and j
  - Estimate the probability density of the score function S under each hypothesis
  - Set SLR_S as the ratio of these estimated densities evaluated at $S(X_i, Y_j)$

Results



**Note:* Empirical density estimated using all pairwise combinations of users' data & a Gaussian KDE.

Results & Conclusions



- *SLR*s based on marked point process indices have potential to perform well in quantifying strength of evidence for user-event data
- Segregation was discriminative for web browsing event streams
- Results obtained *only for a specific data set*, may not generalize

Ongoing & Future Work

- Other score functions (inter-event times & multiple marks)
- Randomization methods
- Theoretical characterization of limits of detectability
- Obtaining more real-world data
  - Currently planning additional data collection at UC Irvine
  - Order of 100 students, months of logged data

References I

Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122.

Bolck, A., Ni, H., & Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, *14*(3), 243–266. doi: 10.1093/lpr/mgv009

Foreman, L., Champod, C., Evett, I., Lambert, J., & Pope, S. (2003). Interpreting DNA evidence: a review. *International Statistical Review*, *71*(3), 473–495.

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. West Sussex, England: John Wiley & Sons Ltd.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, *52*(1), 54–64.

Pielou, E. (1977). *Mathematical ecology*. John Wiley & Sons, Inc.

Schlapbach, A., & Bunke, H. (2007). A writer identification and verification system using HMM based recognizers. *Pattern Analysis and Applications*, *10*(1), 33–43.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley.

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).

Inter-Event Times

- Measure the time to the nearest point in X_i for each point in Y_i
- Yields a distribution of inter-event times for each $\{X_i, Y_i\}$ pair
- Can look at a variety of statistics related to these distributions:
  - Probability or cumulative density functions
  - Descriptive statistics (e.g., mean or median)
  - Statistics related to the cdf (e.g., two-sample Kolmogorov-Smirnov statistic)

$$KS = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)|$$

- In principle this contains more information than nearest neighbor indices

Randomization Methods

Problem: Given only one realization of a bivariate point process $\{X, Y\}$, how can we determine how "unusual" it is assuming that the sub-processes were generated by different individuals?

- Focus on denominator of the SLR: $f(\Delta(X, Y)|H_d)$

Solution: Simulate R realizations $\{X^r, Y^r\}$ for $r = 1, \ldots, R$, then compare the observed statistic $\Delta(X, Y)$ with a "null" distribution obtained from $\{\Delta(X^r, Y^r) : r = 1, \ldots, R\}$

- **Relabeling**: sample marks without replacement keeping times fixed
- Shifting: fix Y and shift entire sequence X or per event shifts in X
- **Simulation** of X^r from a point process (inhomogeneous, bursty) with fixed Y

Kernel Density Estimation

- Kernel function K usually defined as any symmetric density function that satisfies
  1. $\int K(x)dx = 1$
  2. $\int xK(x)dx = 0$
  3. $0 < \int x^2 K(x)dx < \infty$
- Common kernels: Gaussian, Epanechnikov, point mass (histogram)
- Let $X = \{X_1, \ldots, X_n\}$. Then given K and a bandwidth $h > 0$, a kernel density estimator is defined as

$$\hat{f}_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- Intuition: estimated density at x is the average of the kernel centered at the observation X_i and scaled by h across all n observations
- Choice of kernel really not important, but bandwidth is

The Likelihood Ratio

Following the notation of Bolck et al. (2015), define

- Evidence $E \equiv \{X, Y\}$
- X: set of observations for a reference sample from a *known source*
- Y: set of observations of the same features as X for a sample from an *unidentified source*
- H_s: same source hypothesis
- H_d: different sources hypothesis

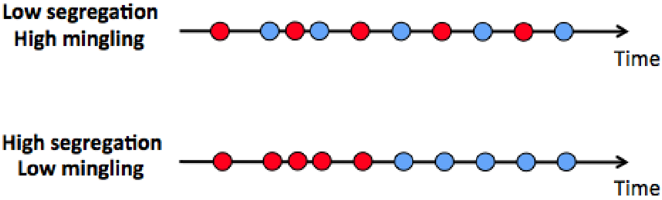
$$\underbrace{\frac{Pr(H_s|E)}{Pr(H_d|E)}}_{\textit{a posteriori}\text{ odds}} = \overbrace{\frac{Pr(E|H_s)}{Pr(E|H_d)}}^{\text{likelihood ratio}} \underbrace{\frac{Pr(H_s)}{Pr(H_d)}}_{\textit{a priori}\text{ odds}}$$

Mingling Index (Illian et al., 2008)

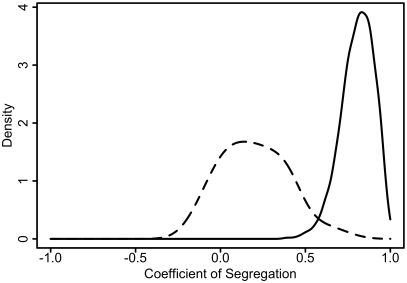
Mean fraction of points among the k nearest neighbors of the reference point that have a mark different than the reference point

$$\overline{M}_k(X_i, Y_i) = \frac{1}{k}\sum_{j=1}^{n_i}\sum_{\ell=1}^{k} \mathbb{1}\left[m(t_{ij}) \neq m(z_\ell(t_{ij}))\right] \in [0,1]$$

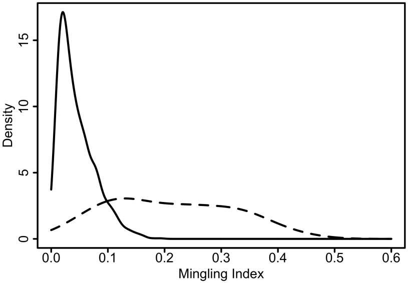
Bivariate, independent marks (stationary case) $\Rightarrow \overline{M}_k(X_i, Y_i) = 2p_xp_y$



Results – Empirical Densities



(a) Segregation



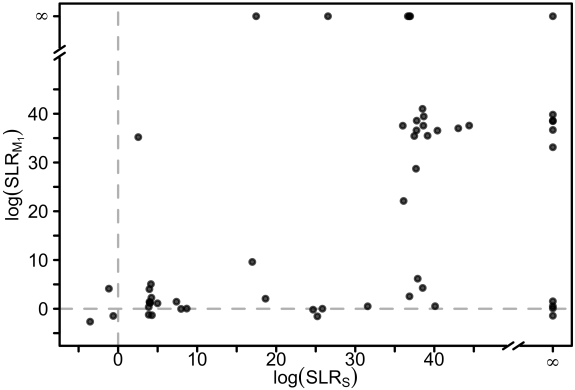
(b) Mingling

Same-source density H_s (dashed line)
Different-source density H_d (solid line)

Case Study–Reference Data Set Composition

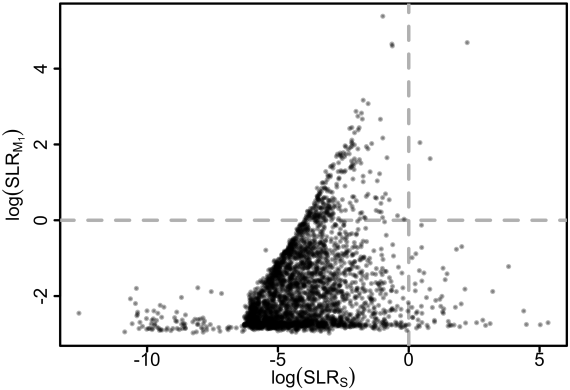
- Compute bivariate process indices $[S(X_i, Y_j)$ and $M_1(X_i, Y_j)]$ for all $N^2 = 55^2 = 3025$ pairwise combinations of user event streams
- For each pairwise combination $\{X_i, Y_j\}$ and $\Delta \in \{S, M_1\}$, compute a "leave-one-out"–like estimate of the score-based likelihood ratio
  - $\mathcal{D}_s = \big\{\{X_k, Y_k\} : k \in \{1, \ldots, N\}, k \neq i, k \neq j\big\}$
  - $\mathcal{D}_d = \big\{\{X_k, Y_\ell\} : k, \ell \in \{1, \ldots, N\}, k \neq \ell, k \neq i, k \neq j, \ell \neq i, \ell \neq j\big\}$
  - Estimate $\hat{f}(\Delta|H_s, \mathcal{D}_s)$ and $\hat{f}(\Delta|H_d, \mathcal{D}_d)$ via KDE with the "rule of thumb" bandwidth (Scott, 1992)
  - Set SLR_Δ as the ratio of these empirical densities evaluated at $\Delta(X_i, Y_j)$

Results – Evaluation of known same-source streams



		SLR_{M_1}		
		-	+	Total
SLR_S	-	2	1	3
	+	6	46	52
	Total	8	47	55

Results – Evaluation of known different-source streams



		SLR_{M_1}		
		-	+	Total
SLR_S	-	2666	240	2906
	+	61	3	64
	Total	2727	243	2970