

Analyzing User-Event Data Using Score-based Likelihood Ratios with Marked Point Processes

Chris Galbraith

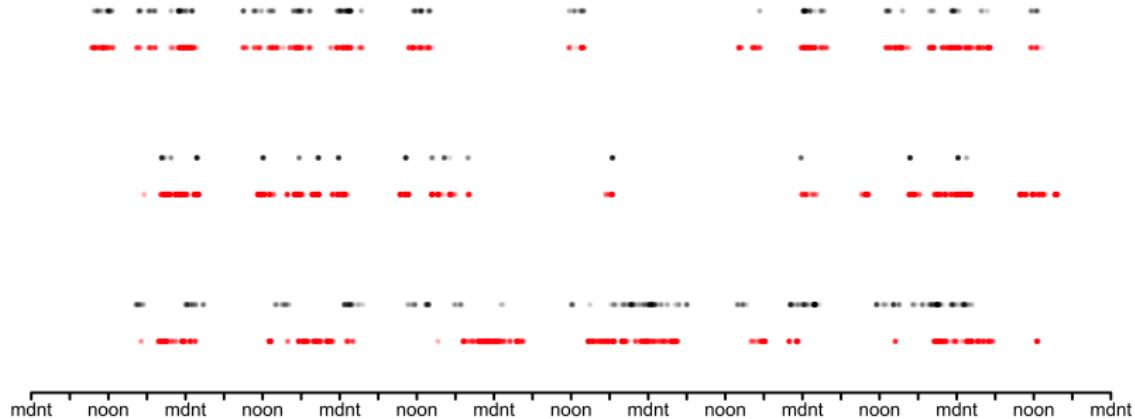
University of California, Irvine

August 9, 2017

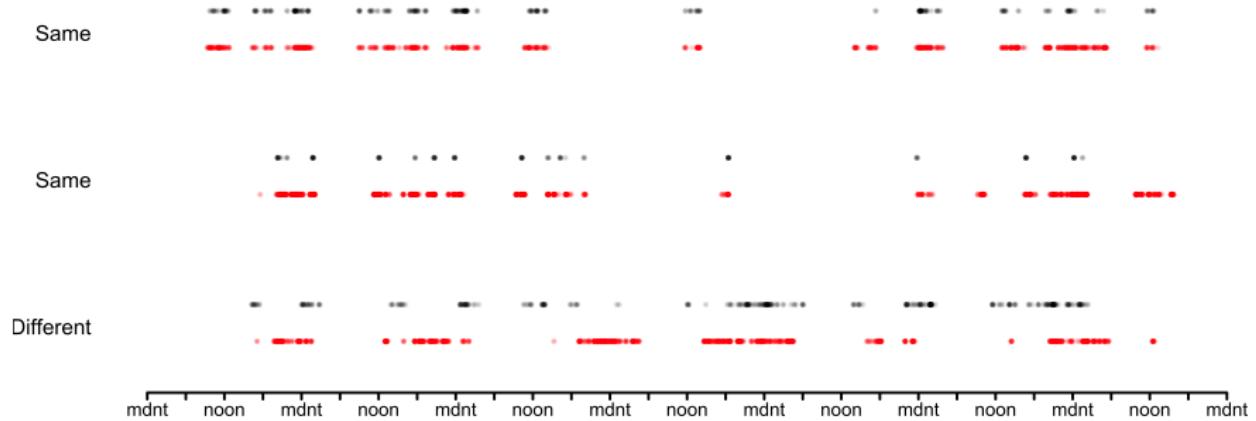
Motivation

- Event histories of user activities (aka *event streams* or *user-event data*) are routinely logged on devices including computers and mobile phones
- Typically consist of <event, timestamp, metadata>
- As digital devices become more prevalent, these user event histories are encountered with increasing regularity
- Investigators want to determine the likelihood that two event histories were generated by the same individual

Motivation



Motivation



Related Work

- Visualization tools
 - Tools to assist the investigation of user-generated event logs from computers and mobile devices (Casey, 2011; Roussev, 2016)
 - Interactive timeline analysis (Buchholz & Falk, 2005)
 - Visualization of email histories (Koven et al., 2016)
- Automated summarization & session similarity
 - Analyzing session to session similarities of Internet usage (Gresty et al., 2016)
 - Linking user sessions via network traffic information (Kirchler et al., 2016)
 - Automated summarization of event data (Kiernan & Terzi, 2009)
- Model based approaches
 - Social network analysis (Eagle et al., 2009)

My Work

- Advised by Padhraic Smyth (and Hal Stern) under CSAFE
- 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



Overview

1 The Likelihood Ratio

- Feature-based
- Score-based

2 Marked Point Processes

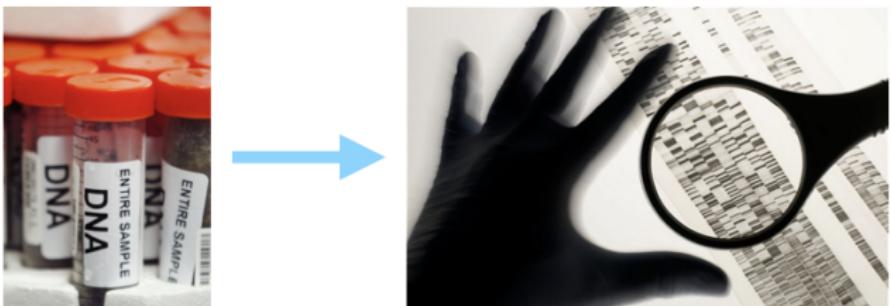
- Bivariate Point Processes
- Summary Statistics

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)
 - Speaker recognition (Gonzalez-Rodriguez et al., 2006)
 - Fingerprint analysis (Neumann et al., 2007)
 - Handwriting analysis (Schlapbach & Bunke, 2007)
 - Analysis of illicit drugs (Bolck et al., 2015)

DNA – LR Gold Standard



Focus on alleles known to vary across the population.

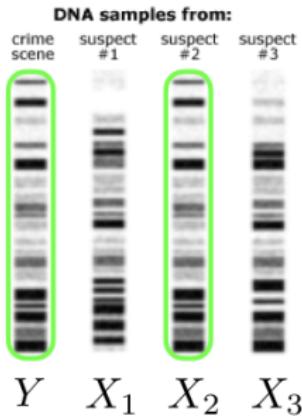
DNA samples from:

crime scene	suspect #1	suspect #2	suspect #3
[Gel Electrophoresis Pattern]	[Gel Electrophoresis Pattern]	[Gel Electrophoresis Pattern]	[Gel Electrophoresis Pattern]

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

1. Samples are from the *same* person
2. Samples are from *different* people

Feature-based Likelihood Ratio



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

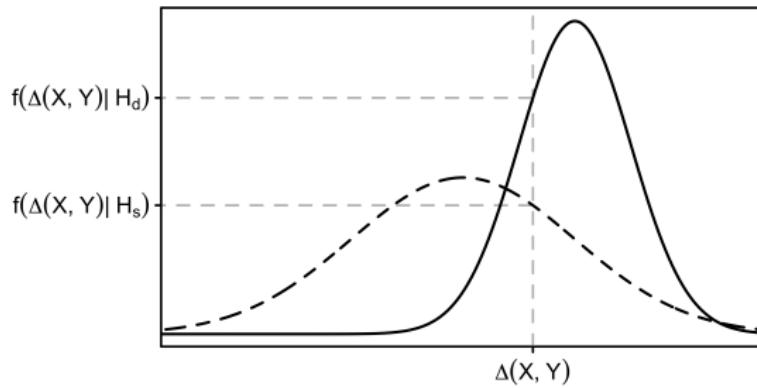
> 1 Samples from same source

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*

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



Overview

1 The Likelihood Ratio

- Feature-based
- Score-based

2 Marked Point Processes

- Bivariate Point Processes
- Summary Statistics

3 Case Study

Marked Point Processes

- I follow the notation of Illian et al. (2008), who define a *marked point process* M as a sequence of random marked points

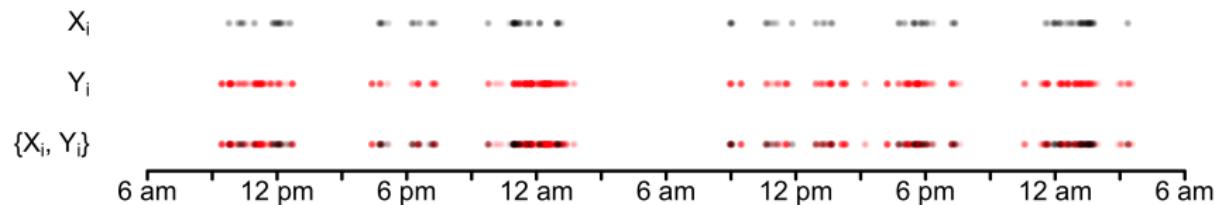
$$M = \{(t_n, m(t_n)) : n = 1, 2, \dots\}$$

where $m(t_n)$ is the mark of the point $t_n \in \mathbb{R}^d$

- Marks can be continuous or categorical (or both if multiple marks)
- Typically found in forestry, sociology, ecology, astronomy, etc.

User-Event Histories as Bivariate Point Processes

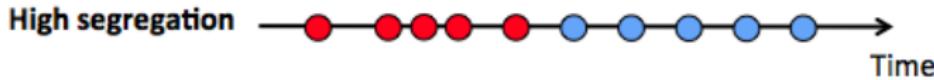
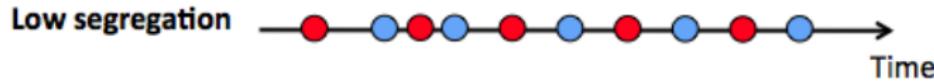
- Event streams can be viewed as marked point processes with the following properties
 - Temporal (i.e., time-stamped events)
 - Binary marks corresponding to the type of event
- We refer to these as *bivariate point processes*



Bivariate Process Indices

Coefficient of segregation, S (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

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



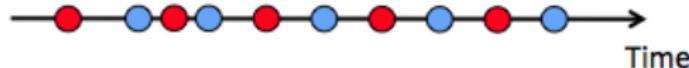
Bivariate Process Indices

Mingling index, \overline{M}_k (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

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

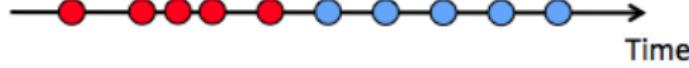
Low segregation

High mingling



High segregation

Low mingling



Overview

1 The Likelihood Ratio

- Feature-based
- Score-based

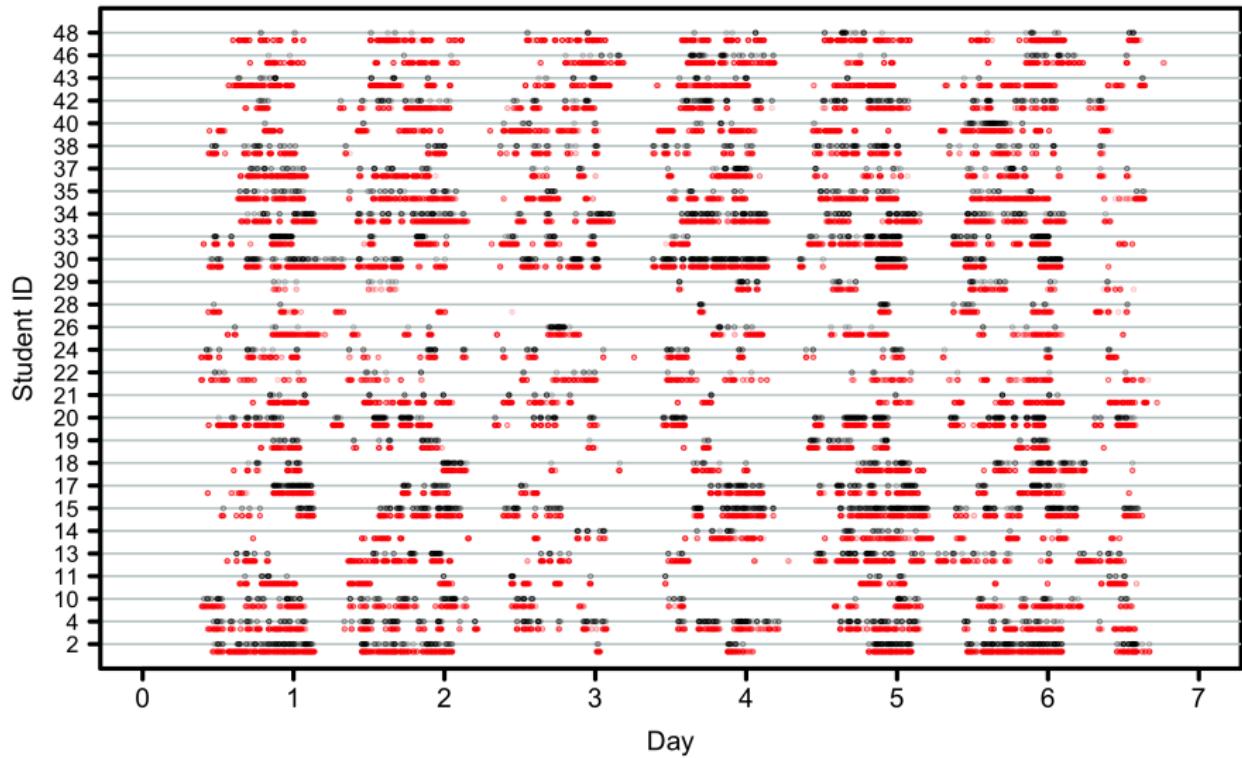
2 Marked Point Processes

- Bivariate Point Processes
- Summary Statistics

3 Case Study

Case Study

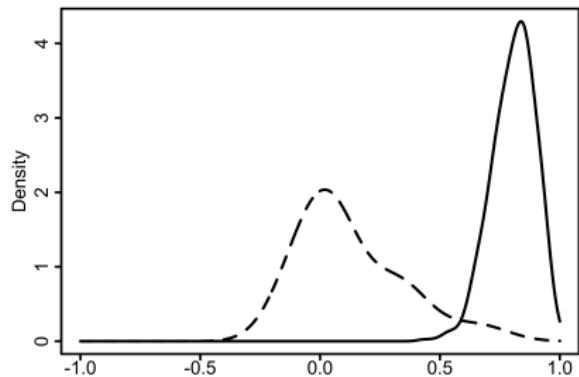
- Data from a 2013-2014 study at UCI that recorded students' browser activity for one week (Wang et al., 2015)
- Dichotomize browser activity
 - Reference sample of Facebook-only events
 - Unidentified sample of non-Facebook events
- Considered 28 students with at least 50 events of each type



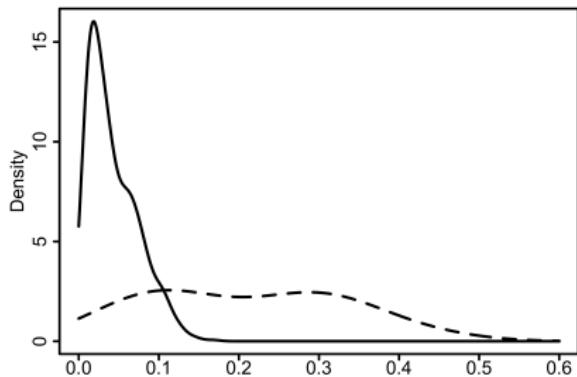
Method

- Compute bivariate process indices for all N^2 pairwise combinations of user event streams
- For each pair $\{X_i, Y_j : i, j = 1, \dots, N\}$ evaluate SLR_S and SLR_{M_1} 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 Δ under each hypothesis
 - Set SLR_Δ as the ratio of these estimated densities evaluated at $\Delta(X_i, Y_j)$

Results – Empirical Densities



(a) Segregation



(b) Mingling

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

Results – Classification Accuracy

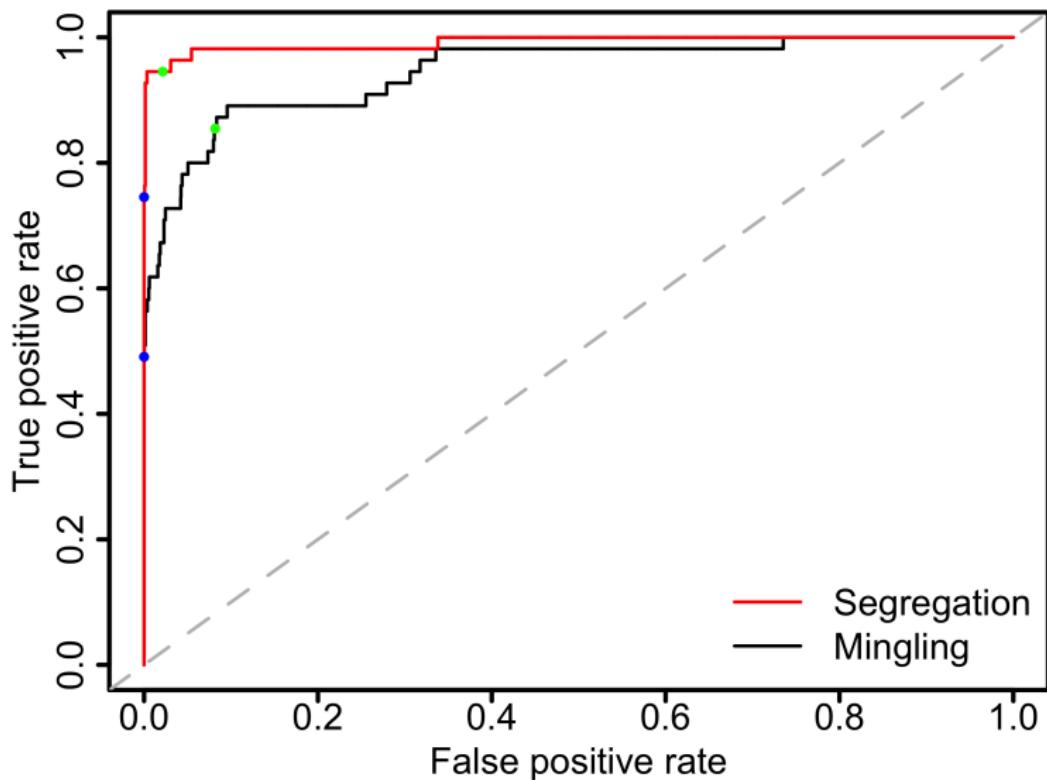
		SLR_{M_1}		Total
		< 1	> 1	
SLR_S	< 1	2	0	2
	> 1	5	21	26 (93%)
	Total	7	21 (75%)	28

Table: Known same-source pairs

		SLR_{M_1}		Total
		< 1	> 1	
SLR_S	< 1	698	46	744 (98%)
	> 1	12	0	12
	Total	710 (94%)	46	756

Table: Known different-source pairs

Results – ROC Curve



Conclusions

- *SLRs based on marked point process indices have potential to perform well in quantifying strength of evidence for user-event data*
- Segregation and mingling were discriminative score functions for web browsing event streams
- Results obtained *only for specific data set* and may not generalize to others

Future Work

- Other score functions (inter-event times & multiple marks)
- Theoretical characterization of limits of detectability
- Randomization methods
- 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
- Buchholz, F. P., & Falk, C. (2005). Design and implementation of Zeitline: a forensic timeline editor. In *Dfrws*.
- Casey, E. (2011). *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic Press.
- Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36), 15274–15278.
- Foreman, L., Champod, C., Evett, I., Lambert, J., & Pope, S. (2003). Interpreting DNA evidence: a review. *International Statistical Review*, 71(3), 473–495.
- Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, 20(2), 331–355.
- Gresty, D. W., Gan, D., Loukas, G., & Ierotheou, C. (2016). Facilitating forensic examinations of multi-user computer environments through session-to-session analysis of internet history. *Digital Investigation*, 16, S124–S133.
- Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. West Sussex, England: John Wiley & Sons Ltd.

References II

- Kiernan, J., & Terzi, E. (2009). Constructing comprehensive summaries of large event sequences. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(4), 21.
- Kirchler, M., Herrmann, D., Lindemann, J., & Kloft, M. (2016). Tracked without a trace: Linking sessions of users by unsupervised learning of patterns in their DNS traffic. In *Proceedings of the 2016 ACM workshop on artificial intelligence and security* (pp. 23–34).
- Koven, J., Bertini, E., Dubois, L., & Memon, N. (2016). Invest: Intelligent visual email search and triage. *Digital Investigation*, 18, S138–S148.
- 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.
- Roussev, V. (2016). Digital forensic science: Issues, methods, and challenges. *Synthesis Lectures on Information Security, Privacy, & Trust*, 8(5), 1–155.
- Schlappbach, 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).

Feature-based 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

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

Kernel Density Estimation

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

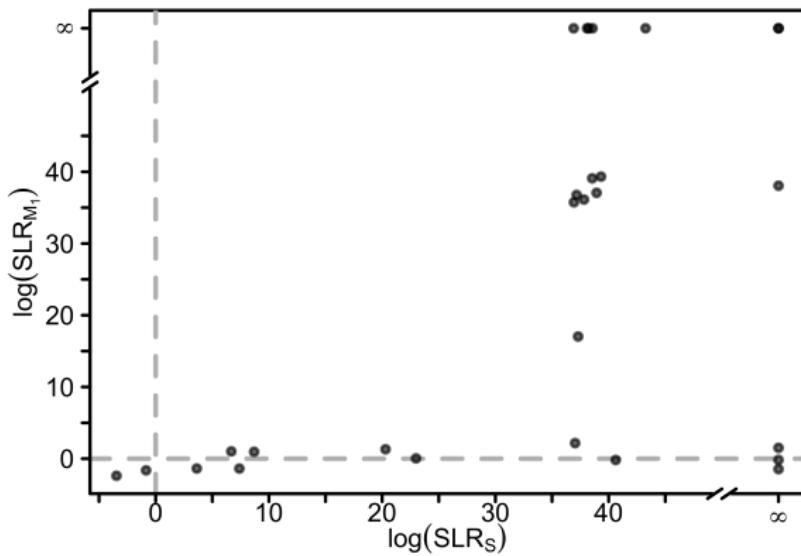
$$\hat{f}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- 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

Case Study–Reference Data Set Composition

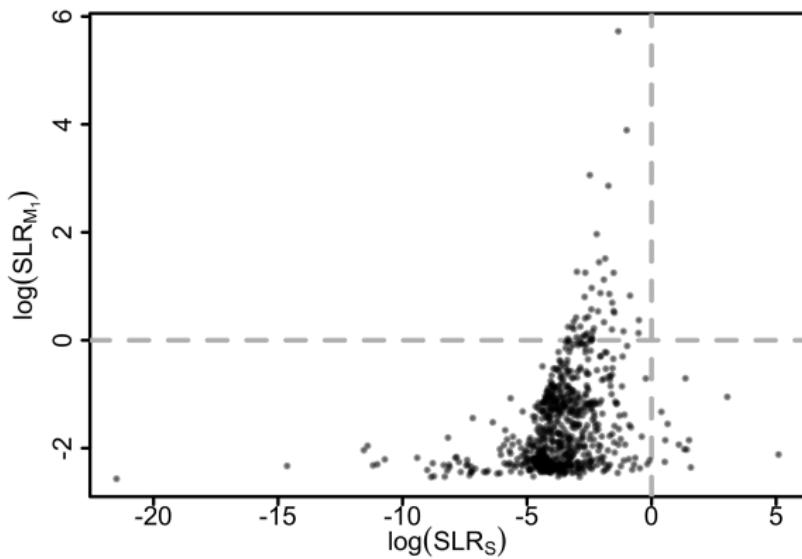
- Compute bivariate process indices $[S(X_i, Y_j)$ and $M_1(X_i, Y_j)]$ for all $N^2 = 55^2 = 3025$ pairwise combinations of user event streams
- For each pairwise combination $\{X_i, Y_j\}$ and $\Delta \in \{S, M_1\}$, compute a “leave-one-out”–like estimate of the score-based likelihood ratio
 - $\mathcal{D}_s = \{\{X_k, Y_k\} : k \in \{1, \dots, N\}, k \neq i, k \neq j\}$
 - $\mathcal{D}_d = \{\{X_\ell, Y_\ell\} : \ell \in \{1, \dots, N\}, \ell \neq i, k \neq j, k \neq \ell, \ell \neq i, \ell \neq j\}$
 - 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	0	0
	+	5	21	26
Total	7	21	28	

Results – Evaluation of known different-source streams



		SLR_{M_1}		Total
		-	+	
SLR_S	-	698	46	744
	+	12	0	12
Total	710	46	756	