Statistical Analysis of User-Event Data for Digital Forensics

Chris Galbraith[1] & Padhraic Smyth[2]
[1]Department of Statistics
[2]Department of Computer Science

BACKGROUND

Event histories recording user activities are routinely logged on devices such as computers and mobile phones. For a particular user these logs typically consist of a list of events where each event consists of a timestamp and some metadata associated with the event. As digital devices become more prevalent, these types of user event histories are encountered with increasing regularity during forensic investigations. As an example, an investigator might be trying to determine if two event histories, corresponding to different usernames, were in fact generated by the same individual.

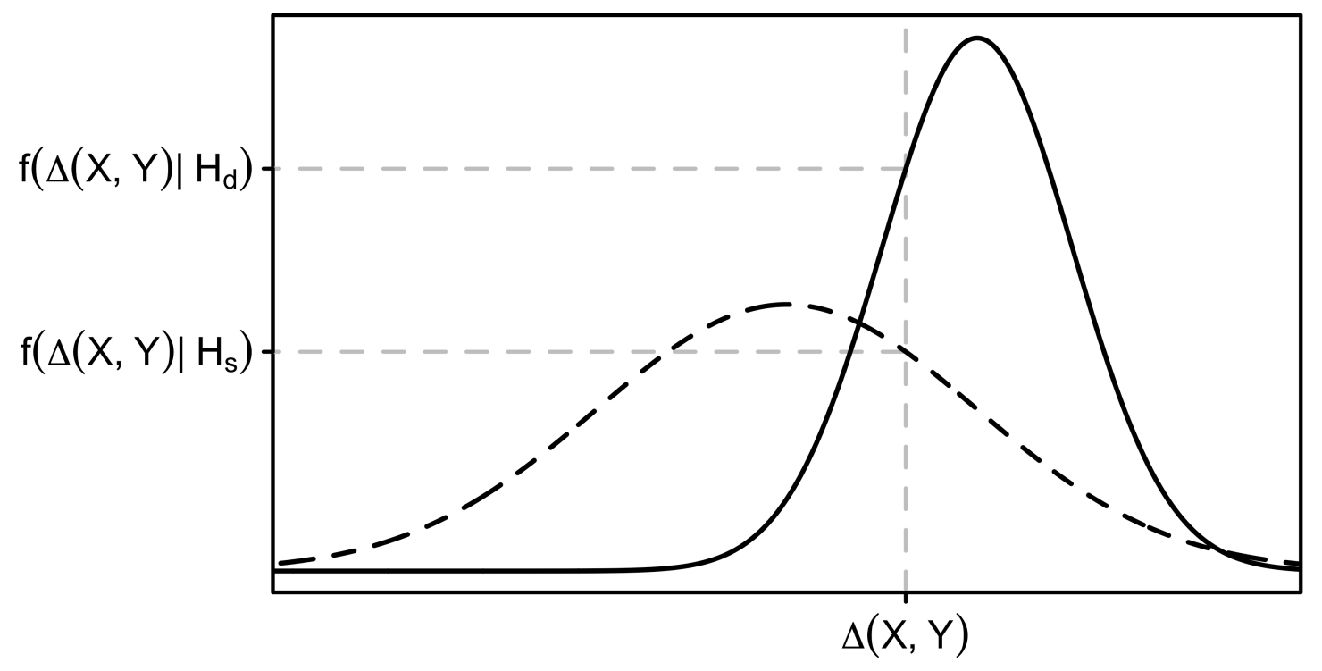
SCORE-BASED LIKELIHOOD RATIO

Let *E* = {*X*, *Y*} where *X* is a set of observations for a reference sample from a known source, and Y is a set of observations of the same features as X for a sample from an unidentified source [1]. The *likelihood ratio* is the ratio of the probability of observing the evidence *E* under two competing hypotheses (that the samples come from the same source, H_s, or different sources, H_d). The LR arises in the application of Bayes' theorem to this situation:

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

Instead of modeling the probability of observing the evidence *E* directly, one can model the probability density of a *score function* Δ that measures the similarity between the samples *X* and *Y*. This yields the *score-based likelihood ratio:*

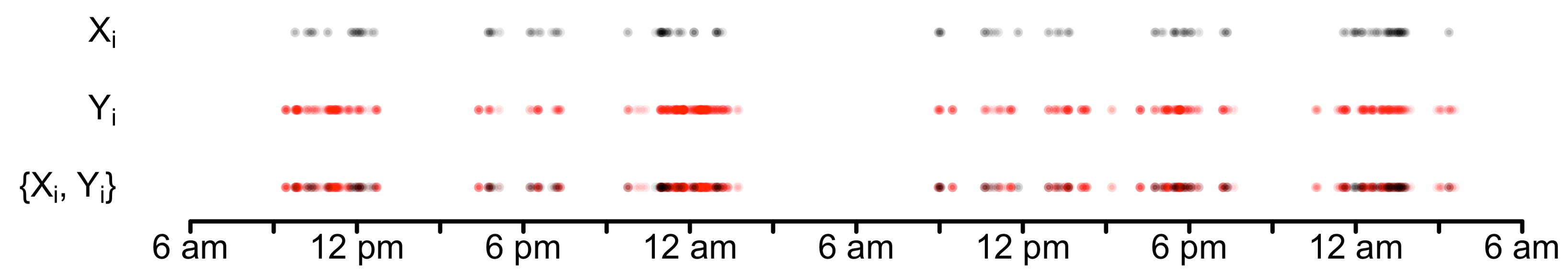
$$SLR_\Delta = \frac{f(\Delta(X,Y)|H_s)}{f(\Delta(X,Y)|H_d)}$$



BIVARIATE POINT PROCESSES

The event stream for the i^{th} user can be viewed as one-dimensional bivariate point processes $M_i = \{X_i, Y_i\}$ where $X_i = \{t_{ij} : m(t_{ij}) = x \text{ for } j = 1, \ldots, n_i\}$ is the sub-process of events of type *x* (similar definition for Y_i). We use indices (coefficient of segregation, *S*, and mingling index, M_1) as score functions.

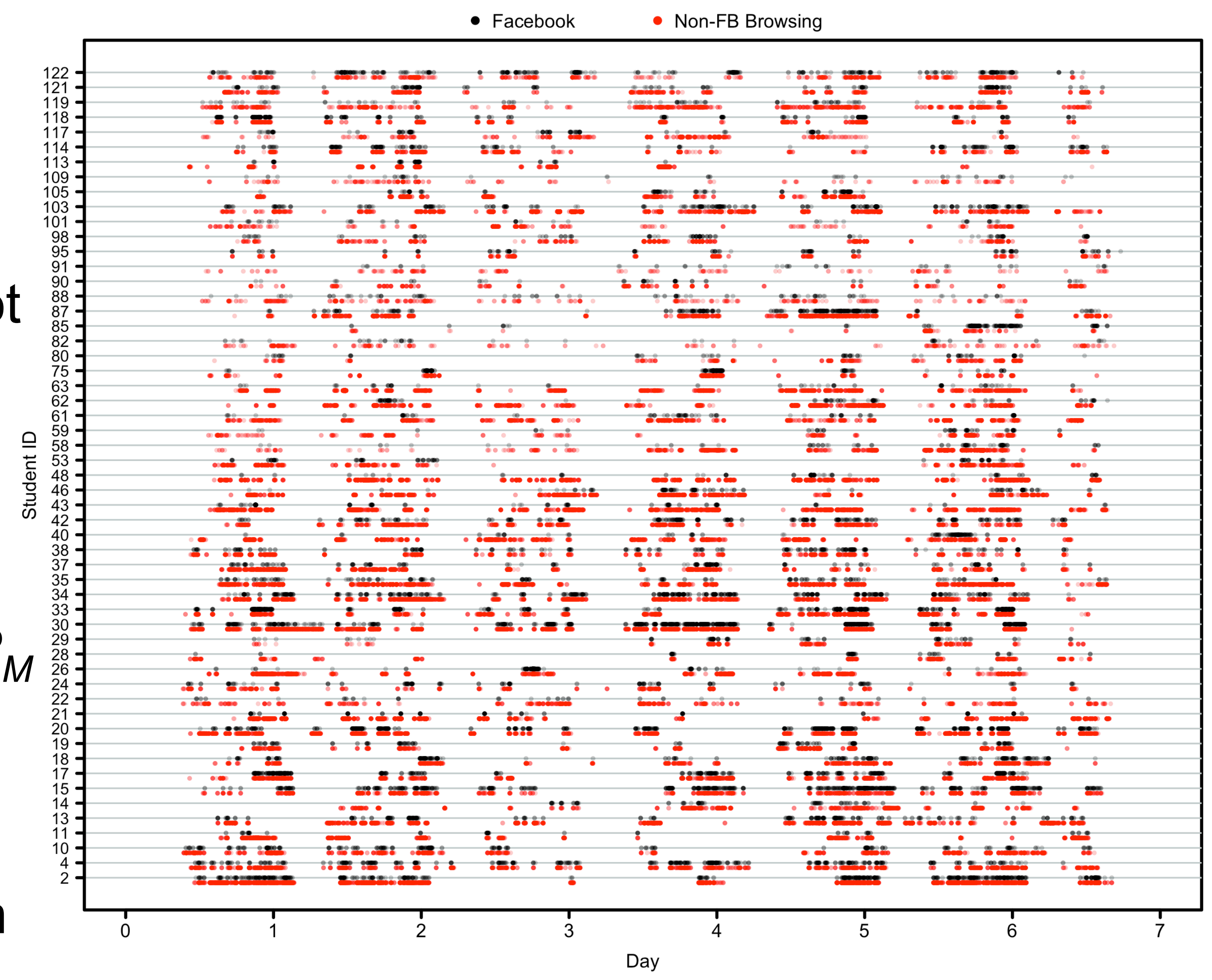
$$S(X_i, Y_i) = 1 - \frac{p_{xy} + p_{yx}}{p_x p_{\cdot y} + p_y p_{\cdot x}} \in [-1, 1] \qquad \overline{M}_k(X_i, Y_i) = \frac{1}{k} \sum_{j=1}^{n_i} \sum_{\ell=1}^{k} \mathbf{I}\left[m(t_{ij}) \neq m(z_\ell(t_{ij}))\right] \in [0, 1]$$



APPLICATION

DATA

Event streams from 55 students observed for one week in a study conducted at UCI that used logging software to record all browser activity on their computers [3]. Bivariate point processes created for each student by dichotomizing browsing events to Facebook (any url whose root domain is facebook.com) versus non-Facebook.

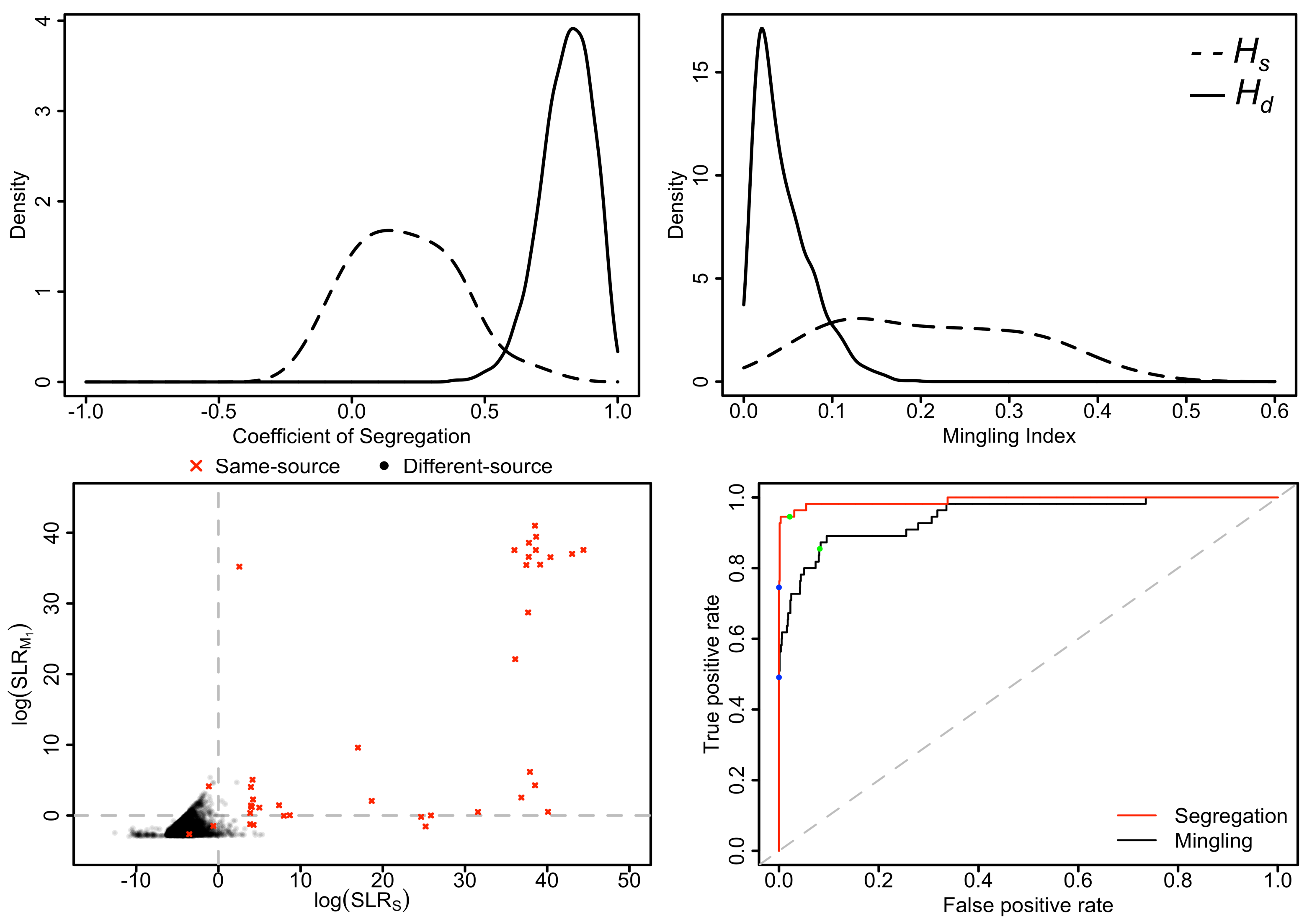


METHOD

1. Compute indices for all 55^2=3025 pairwise combinations of event streams.
2. For each pair $\{X_i, Y_j\}$, evaluate SLR_S and SLR_M with densities estimated from all *other* data.
   - Leave out all event streams from users *i* and *j* in reference data sets.
   - Empirical densities estimated via KDE with Gaussian kernel and rule of thumb bandwidth [2].

RESULTS

Score-based likelihood ratios based on marked point process indices have potential to perform well in quantifying strength of evidence for user-event data. In particular, segregation and mingling were discriminative score functions for web browsing event streams. Note that results obtained *only for specific data set and may not generalize to others.*



FUTURE WORK

- Randomization methods (given only one set of event streams)
- Inter-event times
- Multiple types (>2) of events
- Change detection in this context

REFERENCES

1.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.
2.Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. Wiley.
3.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*, 571–582.

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.