Law, Probability and Risk (2022) **21**, 91–122
Advance Access publication on December 23, 2022
https://doi.org/10.1093/lpr/mgac016

Likelihood ratios for categorical count data with applications in digital forensics

Rachel Longjohn[†], Padhraic Smyth and Hal S. Stern

University of California, Irvine, California, USA

We consider the forensic context in which the goal is to assess whether two sets of observed data came from the same source or from different sources. In particular, we focus on the situation in which the evidence consists of two sets of categorical count data: a set of event counts from an unknown source tied to a crime and a set of event counts generated by a known source. Using a same-source versus different-source hypothesis framework, we develop an approach to calculating a likelihood ratio. Under our proposed model, the likelihood ratio can be calculated in closed form, and we use this to theoretically analyse how the likelihood ratio is affected by how much data is observed, the number of event types being considered, and the prior used in the Bayesian model. Our work is motivated in particular by user-generated event data in digital forensics, a context in which relatively few statistical methodologies have yet been developed to support quantitative analysis of event data after it is extracted from a device. We evaluate our proposed method through experiments using three real-world event datasets, representing a variety of event types that may arise in digital forensics. The results of the theoretical analyses and experiments with real-world datasets demonstrate that while this model is a useful starting point for the statistical forensic analysis of user-generated event data, more work is needed before it can be applied for practical use.

Keywords: multinomial distribution; Dirichlet distribution; Bayesian inference; digital evidence; event data.

1. Introduction

Categorical count data arises across a variety of applications in forensics, e.g. counts of printed documents with toners belonging to certain resin groups (Biedermann *et al.*, 2011) or counts of licit versus various illicit drugs in a sample for chemical analysis (Mavridis and Aitken, 2009). In forensic investigations we often wish to use this data to answer source or identity questions; for example, investigators may ask whether two documents came from the same printer or two sets of geolocated events from the same individual. A widely accepted statistical approach for addressing such questions in forensics is to construct a likelihood ratio to quantify the strength of the evidence (see e.g. Champod *et al.* (2016); Stern (2017); Aitken *et al.* (2021)). The likelihood ratio approach has been widely applied in the context of DNA evidence (e.g. Evett and Weir (1998)), and there is ongoing work in a variety of other forensic evidence types, such as fingerprints (e.g. Champod and Evett

[†] Corresponding author. E-mail: rlongjoh@uci.edu

(2001)), glass evidence (e.g. Zadora and Ramos (2010)), and speaker recognition (e.g. Morrison (2009)). In this article, we outline a likelihood ratio-based framework for forensic analysis when the evidence is in the form of categorical count data.

Our work is motivated in particular by evidence in the form of user-generated event data in which investigators possess counts of different types of events generated on a digital device for persons of interest in a case. For example, an investigator may wish to determine how likely it is that two sets of mobile phone usage records were generated by the same individual or by two different individuals. With the widespread use of smartphones and other devices, statistical methods for analysing event data are becoming increasingly important in digital forensics. The majority of digital forensics research has focused on information extraction and information reconstruction from devices and the cloud (see e.g. Roussev (2016); Årnes (2017); SWGDE (2020a,b)), often producing vast amounts of extracted digital data. However, there exist comparatively few statistical forensic approaches to support quantitative investigative efforts into such data. Indeed, an Organization of Scientific Area Committees for Forensic Science (OSAC) Task Group focused on digital evidence recommended that efforts should be made to ‘strengthen [the] scientific foundations of digital/multimedia evidence by developing systematic and coherent methods for studying the principles of digital/multimedia evidence. . .as well as any associated probabilities’ (Pollitt et al., 2018).

The primary contributions of this article are two-fold: (i) the theoretical development of a likelihood ratio for categorical count data and (ii) the systematic experimental evaluation of the approach on three event datasets relevant to digital forensics. To this end, the remainder of the article is organized as follows: Section 2 introduces the mathematical notation and describes our proposed approach, and Section 3 relates this approach to other relevant research in statistics and digital forensics. Sections 4–5 analyse various theoretical properties of the likelihood ratio that results from our proposed model. We describe how our approach may be applied to digital forensics in Section 6, and present experiments and results using examples of such applications in Section 7. We conclude with a discussion of the results, future research directions, and conclusions in Sections 8–9.

2. Notation and modelling assumptions

2.1 Evidence in the form of count data

Consider evidence in the form of observed events, where each event belongs to one of K non-overlapping categories or types. For example, an event could correspond to a user of a mobile phone opening a certain software application (an ‘app’), and the category could be the particular app that was opened. We focus here on representing such data in the form of counts, e.g. the number of times a user opened each app over some time period. In what follows below we assume that we have a finite number K of predefined categories of interest (e.g. a set of commonly used apps on mobile phones) and that events belonging to other categories are not of interest.

More specifically, let the evidence be comprised of two sets of counts, where one set was generated by a known source (e.g. a person of interest or a suspect) and the other set was generated by an unknown source and can be tied to a criminal activity. Each of these sets will be a K-dimensional vector of category counts. Denote the counts from the known source as $r_1 = (r_{11}, r_{12}, \ldots, r_{1K})$ and the counts from the unknown source as $r_2 = (r_{21}, r_{22}, \ldots, r_{2K})$, such that r_{ik} is the number of events for source i in category k. Also let $N_1 = \sum_{k=1}^{K} r_{1k}$ and $N_2 = \sum_{k=1}^{K} r_{2k}$ be the total number of events observed for the known and unknown sources, respectively. Together, r_1 and r_2 constitute the evidence.

We do not impose any restrictions on the time periods over which r_1 and r_2 are observed. Depending on the application, it may make sense to only consider disjoint, overlapping, or identical time periods for r_1 and r_2 (e.g. the situation described in Section 6.1), but the proposed approach is flexible and can handle any of these scenarios.

2.2 *Source hypotheses*

We evaluate the evidence in the context of two mutually exclusive hypotheses: the same-source hypothesis H_s and the different-source hypothesis H_d. For the purposes of our model, these hypotheses can be expressed as

$$\begin{aligned} H_s &: r_2 \text{ generated by Source 1} \\ H_d &: r_2 \text{ not generated by Source 1} \end{aligned}$$

where r_2 refers to the count data from the unknown source, and Source 1 is used to refer to the known source of r_1.

The wording that we have used for H_d implies that if not generated by Source 1, r_2 was generated by another person in the general population. In practice, the starting point for generating the alternative hypothesis should be the facts of the case (Robertson *et al.*, 2016, pp. 30–33; Aitken *et al.*, 2021, pp. 615–619). Incorporating such information can be used to refine the relevant population rather than leaving it as the broad and vague 'general population'. For example, in speaker recognition, the relevant population could be refined to be people who speak a certain language with a particular regional accent (Morrison et al., 2016; Rose, 2002, pp. 64–65). One may also consider the potential relationship between Source 1 and a potential alternate source. For example, Bosma *et al.* (2020) briefly discuss if under the different-source hypothesis, the person who generated the unknown-source data is not associated with Source 1 but is from the same city or country, or if they are someone vaguely or well known to Source 1. For the purposes of our model specification, we use the broad different-source hypothesis rather than focusing on a particular relevant population, but later on we will note where in the model one could incorporate more information with a refined relevant population.

We also note that the manner by which we define these hypotheses is an example of a 'specific source' rather than a 'common source' scenario. Under the common source scenario, the same-source hypothesis would posit that the two sets of observations were generated by the same, unknown source, and the different-source hypothesis would assume that the two sets of observations originate from two different, unknown sources. In contrast, under the specific source scenario, the same-source hypothesis attributes the unknown source evidence to a specific, known source while the different-source hypothesis attributes the two sets of evidence to two different sources, one known and the other still unknown (Ommen and Saunders, 2018).

2.3 *Multinomial model*

Let $\theta_1 = (\theta_{11}, \theta_{12}, \ldots, \theta_{1K})$ represent the parameters of a categorical probability distribution for Source 1, where θ_{1k} is the probability that a single event from Source 1 belongs to category k and $\sum_{k=1}^{K} \theta_{1k} = 1$.

To model multiple events, we assume that Source 1's count data follows a multinomial distribution (the extension of a binomial distribution for $K > 2$ categories) given θ_1, i.e.

$$r_1|\theta_1 \sim \text{Multinomial}(N_1, \theta_1).$$

Note that the multinomial assumption imposes a strong 'memoryless' property on the observed events in that they are assumed to be independent (discussed further in Section 8).

The distributional assumption on r_1 holds for both the same-source and different-source hypotheses since this set of observations comes from the known source. However, the distributional assumption for r_2 depends on our source assumption. In particular, under H_s, we further assume that

$$r_2|H_s, \theta_1 \sim \text{Multinomial}(N_2, \theta_1)$$

where, because we have assumed the same source, the vector of probabilities corresponds to Source 1 and is the same as for r_1. Under H_d, however, we assume that

$$r_2|H_d, \theta_2 \sim \text{Multinomial}(N_2, \theta_2)$$

where, given that we have assumed different sources, the set of categorical probabilities $\theta_2 = (\theta_{21}, \theta_{22}, \ldots, \theta_{2K})$ can be different from that of Source 1. In what follows below we assume that r_1, r_2, N_1, and N_2 are known (the evidence) and θ_1 and θ_2 are unknown.

2.4 *Likelihood ratio*

The idea of applying the likelihood ratio to forensics has, in recent years, become widely accepted in the forensic science community as a logical means by which to assess the strength of evidence (e.g. Champod *et al.* (2016)); although there is still ongoing debate (see e.g. Lund and Iyer (2017)). The likelihood ratio arises from using Bayes' Theorem to evaluate the evidence regarding two mutually exclusive hypotheses, such as those described in Section 2.2. In the forensic context, Bayes' Theorem can be written as

$$\underbrace{\frac{Pr(H_s)}{Pr(H_d)}}_{\text{prior odds}} \cdot \underbrace{\frac{Pr(E|H_s)}{Pr(E|H_d)}}_{\text{likelihood ratio}} = \underbrace{\frac{Pr(H_s|E)}{Pr(H_d|E)}}_{\text{posterior odds}} \tag{1}$$

where E refers to the evidence. In this formulation we are implicitly conditioning on any background information, I, that may be present and that might be relevant when evaluating the evidence, such as how the evidence was collected or population information (see e.g. National Commission on Forensic Science (2015); Stern (2017)).

The likelihood ratio (*LR*) measures the relative probability of observing E under each of the two hypotheses, H_s and H_d. The philosophy behind the *LR* is that the evidence evaluator will have a set of prior odds regarding the two hypotheses before seeing the evidence. The *LR* can then be provided to the evaluator such that they can modify their prior odds to arrive at a set of posterior odds, i.e. the relative probability of the two hypotheses after observing the evidence. When $LR < 1$, the evidence is more probable under the different-source hypothesis, and when $LR > 1$, the evidence is more probable under the same-source hypothesis ($LR = 1$ is considered a neutral value). The role of the forensic investigator is often viewed as supplying the likelihood ratio, thus providing the means by which the evaluator can modify their pre-evidence beliefs (Stern, 2017).

2.5 *Bayesian computation*

To calculate the likelihood ratio, we use a Bayesian approach and treat θ_1 and θ_2 as quantities about which we are uncertain. The Bayesian approach allows us to average over our uncertainty about the unknown parameters θ_1 and θ_2, which is particularly useful with small amounts of data where there can be high uncertainty about possible parameter values. When unknown parameters are averaged over a distribution rather than estimated, $Pr(E|H_s)/Pr(E|H_d)$ can be referred to as a Bayes factor (Berger, 2013, p. 146) rather than a likelihood ratio. However, there is some history of ambiguity in the forensics literature around the use of the term 'Bayes factor' versus 'likelihood ratio' (Stern, 2017). In this article, although we opt for the Bayesian approach of marginalizing over unknown parameters, we will nonetheless use the term 'likelihood ratio' ('LR') throughout, rather than Bayes factor, to be consistent with the terminology used in other forensic science literature.

An important component of the Bayesian computation of the *LR* is the specification of the prior distribution over the probabilities θ_1 and θ_2. The prior distribution should reflect our a priori belief about the unknown parameters before we see any evidence or data. The *K*-dimensional vectors of probabilities exist in a *K*–1 dimensional simplex, where the simplex is the region defining a set of *K* probabilities that sum to 1. A well-known prior over the simplex is the Dirichlet distribution, which has the convenient property of being conjugate to the multinomial likelihood, meaning that the posterior density is also a Dirichlet distribution and can be represented in closed form. Specifically, we assume that

$$\theta_1, \theta_2 \overset{i.i.d.}{\sim} \text{Dirichlet}(\alpha)$$

where the parameters of the prior are $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ and $\alpha_k > 0$ for $k = 1, \ldots, K$. In Section 4.1 below we discuss how to conduct the Bayesian calculation of *LR*s in more detail for the multinomial-Dirichlet model of interest in this article.

To inform this later discussion of Dirichlet distributions in the context of the *LR*, we briefly review some of their general properties—later in the article we will return to a more detailed discussion of the role of the Dirichlet prior on likelihood ratios for count data. A symmetric Dirichlet distribution is one in which all of the α_k's are equal, expressing an a priori belief that each category is equally likely to occur (Figure 1(a)). The special case in which $\alpha_k = 1$ for $k = 1, \ldots, K$ is called the uniform Dirichlet distribution because it assigns equal density to all vectors θ_i that satisfy $\sum_{k=1}^{K} \theta_{ik} = 1$ (Figure 1(b)). An asymmetric Dirichlet distribution is one in which not all of the $\alpha_{k'}$s are equal (Figure 1(c)). A larger value of α_k relative to other categories indicates that we expect more events in category *k*, while a smaller value indicates that we expect fewer events in category *k*.

The mean of the Dirichlet distribution is $\frac{1}{\sum_{k=1}^{K} \alpha_k}(\alpha_1, \alpha_2, \ldots, \alpha_K)$. The concentration parameter $c = \sum_{k=1}^{K} \alpha_k$ can be thought of as the strength of the Dirichlet prior (relative to the data) and determines how concentrated the prior probability density is around the mean. For example, larger *c* values indicate the density is more tightly concentrated around its mean and that more data is required to shift the posterior density away from the prior (Figure 1(c) versus Figure 1(d)). Note that using this notation, the parameters of a symmetric Dirichlet distribution can be written as $\alpha = c \times \left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right) = \left(\frac{c}{K}, \frac{c}{K}, \ldots, \frac{c}{K}\right)$.

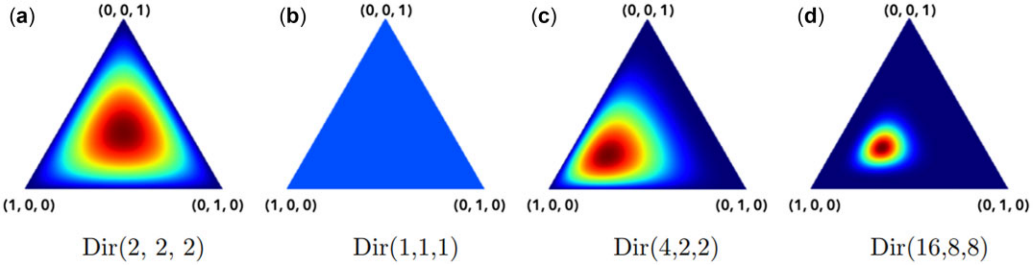


Fig. 1. Dirichlet density plots in which $K = 3$. In this case, the support of the distribution is on a 2-dimensional simplex (because of the constraint that the vector values sum to 1), which we represent using a triangle for the purposes of illustration. From left to right: (a) shows a symmetric, non-uniform Dirichlet distribution, (b) shows the uniform Dirichlet distribution. (c) shows an asymmetric Dirichlet distribution, (d) shows an asymmetric Dirichlet distribution with the same mean as (c) but with a larger concentration parameter.

3. Related work

3.1 Dirichlet-Multinomial modelling and applications

Data in the form of event counts arises in a wide variety of applications, such as population dynamics in ecology (e.g. Johnson *et al.* (2010); Richards (2008)), RNA sequencing in genetics (e.g. Lowe *et al.* (2017)), text classification in natural language processing (e.g. Blei *et al.* (2003); McCallum *et al.* (1998)) and analysis of taxa counts in microbiome studies (e.g. Chen and Li (2013); Wadsworth *et al.* (2017)). The multinomial distribution is an obvious choice for such data since it naturally extends the well-known binomial distribution. In addition, because of its conjugacy in the Bayesian framework, the Dirichlet distribution is often coupled with the multinomial in applications involving the analysis of count data (e.g. Blei *et al.* (2003); Zhang and Stern (2005); Chen and Li (2013); Wadsworth *et al.* (2017)).

Outside the context of forensic science, Puig *et al.* (2016) and Johansson and Olofsson (2007) use a Dirichlet-multinomial framework similar to the one we propose. Puig *et al.* (2016) analyse the famous Federalist papers, which are a set of essays that was published anonymously in 1787 and 1788 by Alexander Hamilton, John Jay and James Madison. The goal in Puig *et al.* (2016) is similar to ours in that they are addressing a source-based investigative question; they aim to attribute authorship of particular essays to one of the three men. However, they are comparing multiple documents' worth of source propositions rather than just two. As in our work, Johansson and Olofsson (2007) only consider two source propositions. They, however, do this only in the special case of a uniform Dirichlet prior, while our approach accommodates general Dirichlet prior distributions. The approaches in Puig *et al.* (2016) and Johansson and Olofsson (2007) also differ from ours in that their primary focus is on the posterior odds rather than on the likelihood ratio, which is the focus in forensics applications.

In forensics, Aitken *et al.* (2021, pp. 791–798) describe a likelihood ratio using a Dirichlet-multinomial model in the context of shoeprint and document analyses. Their approach treats θ as unknown but differs from ours in that θ represents the underlying proportions of the categorical evidence in the entire source population (e.g., the proportion of all printers that use each of K types of toner). In contrast, we treat θ as belonging to an individual, i.e., each individual source may have a unique probability ascribed to each category. Biedermann *et al.* (2011) also calculate a likelihood ratio using a Dirichlet-multinomial model, but unlike in our approach, they do this through the intermediate step of constructing a Bayesian network.

3.2 Statistical approaches in digital evidence

Much of the research in digital evidence so far has focused on information extraction and information reconstruction (see e.g. SWGDE (2020a,b); Årnes (2017); Roussev (2016)), and there have been relatively few statistical approaches developed specifically for quantifying the strength of digital evidence.

Galbraith and Smyth (2017) and Galbraith *et al.* (2020a) introduce a score-based likelihood ratio approach for comparing two time-stamped event streams, where the events are generated by users on a digital device. Their approach is non-parametric and requires reference population data and the specification of a score function in order to produce a likelihood ratio. This approach is extended to geolocation data in Galbraith *et al.* (2020b), and a mixture of kernel density estimates is used to compute both a likelihood ratio and a score-based likelihood ratio.

Bosma *et al.* (2020) use cell tower records for mobile phones in order to assess whether two phones were being used by the same person during a given time period. Their approach uses a combination of logistic regression and kernel density estimation to construct a likelihood ratio specifically for data that is in the form of mobile call detail records. To accomplish this, they split the training dataset into two parts, one for estimating the parameters of the logistic regression and the other for conducting the kernel density estimation.

Casey *et al.* (2020) discuss the challenges associated with evaluating location-related mobile device evidence. The examples they present are similar in nature to the applications that we discuss in Section 6.1, although they focus on specific locations of interest rather than comparing patterns of user-generated events. Note that while Galbraith *et al.* (2020b), Bosma *et al.* (2020), and Casey *et al.* (2020) are tied specifically to a device's location data, our approach is applicable to any form of user-generated event count data.

4. Likelihood ratio analysis

4.1 Solving for the likelihood ratio under the model

Below we provide an outline of the derivation of a general formula for the likelihood ratio, *LR*, using the model described in Section 2.3, with full details in Appendix A.

$$LR = \frac{P(r_1, r_2|H_s)}{P(r_1, r_2|H_d)} \tag{2}$$

$$= \frac{\int P(r_1, r_2|H_s, \theta_1)P(\theta_1|H_s)d\theta_1}{\int\int P(r_1, r_2|H_d, \theta_1, \theta_2)P(\theta_1, \theta_2|H_d)d\theta_1 d\theta_2} \tag{3}$$

$$= \frac{\int Multinom(r_1|N_1, \theta_1) \cdot Multinom(r_2|N_2, \theta_1) \cdot Dir(\theta_1|\alpha)d\theta_1}{\int Multinom(r_1|N_1, \theta_1) \cdot Dir(\theta_1|\alpha)d\theta_1 \int Multinom(r_2|N_2, \theta_2) \cdot Dir(\theta_2|\alpha)d\theta_2} \tag{4}$$

$$= \frac{B(\alpha + r_1 + r_2)B(\alpha)}{B(\alpha + r_1)B(\alpha + r_2)}, \tag{5}$$

where $B(.)$ denotes the multivariate beta function (the mathematical definition of the multivariate beta function is included in Equation (11) in the Appendix). Equation (2) expresses the general form of the likelihood ratio with the count data, r_1 and r_2, serving as the evidence. The Bayesian aspect of our approach is illustrated in Equation (3), where under H_s we marginalize over θ_1 and under H_d we marginalize over both θ_1 and θ_2. Equation (4) follows from the model's assumptions described in Section 2.3, and plugging in the appropriate probability mass and density functions gives Equation (5). Equation (5) is a closed-form expression for the ratio of integrals in Equation (4), resulting in an easily computable LR that is a function of the evidence data, r_1 and r_2, and the prior parameters, α, all represented as vectors of length K.

4.2 *Illustrative examples*

To illustrate how the formula in Equation (5) leads to behaviour expected of a likelihood ratio (Section 2.4), we consider it in the context of several examples. Suppose that there are three event categories of interest (A, B and C) and that we have no a priori information regarding these three types of events. Assume for the moment that we choose a uniform Dirichlet prior (Section 2.5) to describe our uncertainty about the θ parameters.

Suppose that we witness identical event patterns from the known and unknown sources: two A events, one B event and zero C events, i.e. $r_1 = r_2 = (2, 1, 0)$. Because the event count patterns are identical, one would expect a likelihood ratio to favour the same-source hypothesis, and in fact, applying Equation (5) in this case gives $LR = 2.14$. Interpreting as in Section 2.4, $LR = 2.14$ indicates that observing this evidence is 2.14 times more probable under H_s than under H_d.

If instead we obtained more event counts for the known source, say twenty A events, ten B events, and still zero C events (i.e., $r_1 = (20, 10, 0)$ and $r_2 = (2, 1, 0)$), then $LR = 3.88$. Possessing more data upon which to base our likelihood ratio, even for just one of the two sets of observations, enables us to make a slightly stronger statement regarding the evidence. The resulting likelihood ratios become stronger still when more data is obtained for both sources. For example, if $r_1 = r_2 = (20, 10, 0)$, then $LR = 28.02$.

Suppose now that the known source favours A events while the unknown source favours C events: $r_1 = (2, 1, 0)$ and $r_2 = (0, 1, 2)$. Under Equation (5), this leads to $LR = 0.36$ where, now that the two sets exhibit different event patterns, observing this evidence is more probable under the different-source hypothesis. Analogously to the previous examples, the resulting likelihood ratios are strengthened (in the sense that they are further from 1) by observing more data as evidence. For instance, if $r_1 = (20, 5, 5)$ and $r_2 = (0, 1, 2)$, i.e. the amount of data is mixed, then $LR = 0.19$. If $r_1 = (20, 5, 5)$ and $r_2 = (5, 5, 20)$, i.e. larger amounts of data for both sources, then $LR = 0.0008$. Table 1 summarizes the LR values for these examples.

Note that in the trivial case in which either $N_1 = 0$ or $N_2 = 0$, the likelihood ratio retains the neutral value $LR = 1$. This is straightforward to show from Equation (5) by noticing that all elements of r_1 or r_2 will be zero if $N_1 = 0$ or $N_2 = 0$. Therefore our formula for the likelihood ratio adheres to

Table 1. *LR values for the six illustrative examples. For different patterns, LR < 1 and decreases further with more data. For similar patterns, LR > 1 and increases further with more data*

	Different patterns	Similar patterns
Little data	0.3571	2.1429
Mixed	0.1925	3.8824
More data	0.0008	28.0164

the straightforward fact that if there is no available data for either the known or unknown source, looking at the likelihood ratio should be unhelpful.

4.3 *Theoretical properties of the LR under the model*

4.3.1 *Alternate formula.* We can manipulate the expression for the *LR* in Equation (5) into a form that is more amenable to interpretation (details of this derivation are in Appendix B)

$$LR = \left(\prod_{k=1, r_{2k} \geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k + s} \right) \right) \left(\prod_{s=0}^{N_2 - 1} \left(1 - \frac{N_1}{c + N_1 + s} \right) \right), \tag{6}$$

where $c = \sum_{k=1}^{K} \alpha_k$, i.e. the concentration parameter discussed in Section 2.5. Equation (6) makes explicit that the *LR* is a product of two groups of factors, where the first group of factors are ≥ 1 and the second group of factors are ≤ 1.

Note that in Equation (5), switching r_1 and r_2 does not change the value of the *LR*. This implies that there exists a second alternative formula to Equation (6) in which the roles of the source data are switched. However, since this equation does not have inferential implications beyond those of Equation (6), we only note this as a mathematical identity here and leave further details to Appendix B.

4.3.2 *Bounds on the LR.* It is straightforward to show that Equation (6) leads to the following identity

$$\prod_{s=0}^{N_2-1} \left(1 - \frac{N_1}{c + N_1 + s} \right) \leq LR < \prod_{k=1, r_{2k} \geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k + s} \right) \tag{7}$$

where the set of factors less than 1 form a lower bound for the *LR*, and the set of factors greater than 1 form an upper bound.

Let S be the set of all event categories in which both the known and unknown sources have nonzero counts, i.e. an event category k is in S if $r_{1k} > 0$ and $r_{2k} > 0$. The upper bound is only a function of α_k, r_{1k}, and r_{2k} for $k \in S$. This implies that the maximum strength the *LR* can convey in favour of the same-source hypothesis only relies on counts for event types that have been observed across both sources (Figure 2).

Also note from Equation (7) that the lower bound is only a function of N_1, N_2, and c. In practice, this means that one only needs to know the amount of data and the concentration parameter in order to limit how extreme the resulting *LR* can be in favour of the different-source hypothesis. The lower bound is only achieved when $S = \varnothing$, i.e. there are no categories for which an event was witnessed for both the known and unknown sources. This particular case is further discussed in the context of the prior in Section 5.4.

4.3.3 *Number of categories.* Choosing the event categories that are of forensic interest will have implications on the resulting likelihood ratios. Omitting relevant categories, or including irrelevant categories, masks potentially important behaviour for distinguishing between the source

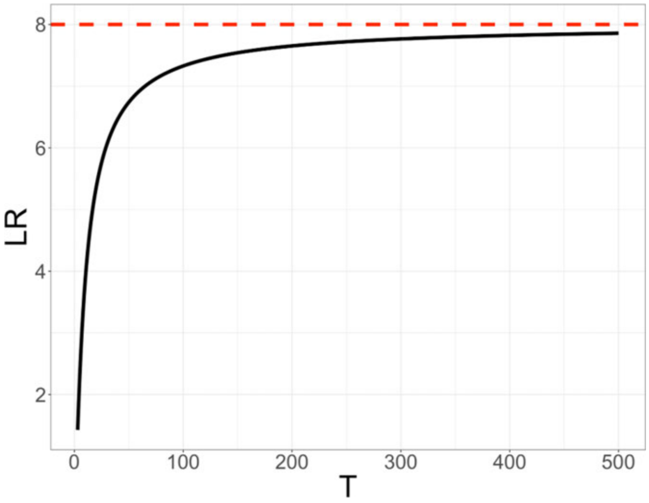


FIG. 2. Plot showing the value of the likelihood ratio using a uniform Dirichlet prior when $r_1 = r_2 = (1, 1, 1, 0_T)$, where T represents the number of 0's appended to the end of the observed count vector. The value of the *LR* is shown in black, and the upper bound is shown in red.

hypotheses. This could result in falsely quantifying the strength of the evidence in either direction—in favour of H_s or H_d. Consider, for example, the case in which investigators are examining device geolocation data for a crime and suspect in Los Angeles. If investigators decided to count the devices' visits to locations across the entire USA rather than just across southern California, the two sets of counts will appear more similar because both sets of data will likely have many shared zero counts for places not near Los Angeles. Similar to the relevant population discussion in Section 2.2, decisions regarding which events are of forensic interest should be determined by the facts of the case. For instance, in the example, it may be relevant to incorporate the entire USA in the analysis if the case involved interstate travel outside of California.

In fact, decisions about the event categories may also be informed by the relevant population. For example, consider again the case in which we have device geolocation data for a crime and a suspect in Los Angeles. Suppose also that the relevant population is determined to be people who were in California during a time period leading up to the crime. Given these circumstances, counting device visits to locations across the USA would not make sense because all potential sources will have device visits only in California and shared zero counts for locations in all other states. In fact, we prove below that in situations like this, in which case-irrelevant shared zero-count event categories are included, the resulting *LR* will be inflated.

To demonstrate the issues associated with changing the number of categories, we show what happens to the *LR* under two possible scenarios for adjusting the Dirichlet prior to accommodate a larger K. In the first scenario, the original α_k's in the prior have a fixed value as new categories are added, e.g. going from $\alpha = (1, 1, 1)$ to $\alpha = (1, 1, 1, 1)$ to $\alpha = (1, 1, 1, 1, 1)$. With this scenario, the lefthand set of factors in Equation (6) remains constant, but the concentration parameter $c = \sum_{k=1}^{K} \alpha_k$ increases, thus increasing the righthand set of factors and the likelihood ratio (Figure 2 shows an example).

$$LR = \underbrace{\left(\prod_{k=1, r_{2k} \geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k + s} \right) \right)}_{\text{constant}} \underbrace{\left(\prod_{s=0}^{N_2-1} \left(1 - \frac{N_1}{c + N_1 + s} \right) \right)}_{\text{increasing}}$$

In the second scenario, the concentration parameter c is held constant as the number of categories increases, e.g. going from $\alpha = (1,1,1)$ to $\alpha = \left(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}, \frac{3}{4}\right)$ to $\alpha = \left(\frac{3}{5}, \frac{3}{5}, \frac{3}{5}, \frac{3}{5}, \frac{3}{5}\right)$. With this strategy, the lefthand set of factors in Equation (6) increases as each α_k decreases, but c, N_1, and N_2 remain constant so that we have

$$LR = \underbrace{\left(\prod_{k=1, r_{2k} \geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k + s} \right) \right)}_{\text{increasing}} \underbrace{\left(\prod_{s=0}^{N_2-1} \left(1 - \frac{N_1}{c + N_1 + s} \right) \right)}_{\text{constant}}.$$

Thus under either of the two proposed strategies, the *LR* increases when more categories with shared zero counts are included, which, in some cases, could mean that the *LR* is artificially inflated as a result of including irrelevant categories.

It is important to note, however, that only having zero counts does not alone make a category irrelevant, nor may all irrelevant categories have zero counts. Other scenarios of including/excluding categories could be studied on a case-by-case basis.

5. The Dirichlet prior and the likelihood ratio

5.1 Revisiting the Dirichlet prior

We now revisit our discussion of the Dirichlet prior from Section 2.5 to focus in more detail on choices in selecting the prior and how these choices can affect the *LR*. Recall that in the Bayesian inference setting we have,

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\alpha) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ r_* | \theta &\sim \text{Multinomial}(N, \theta) \end{aligned}$$

where α, θ and r_* are each K-dimensional vectors. As mentioned in Section 2.5, the Dirichlet prior is a conjugate prior for the multinomial, resulting in the posterior for θ also being a Dirichlet distribution. Specifically, the posterior distribution is of the form

$$\theta | r_* \sim \text{Dirichlet}(\alpha_1 + r_{*1}, \alpha_2 + r_{*2}, \ldots, \alpha_K + r_{*K}).$$

Each of the α_k parameters in the prior are updated from α_k to $\alpha_k + r_{*k}$ having observed the data r_{*k}. The prior parameters α_k of the Dirichlet prior can be intuitively interpreted as 'pseudocounts' since they seed the parameters of the posterior distribution. The role of the prior concentration parameter $c = \sum_k \alpha_k$ can also be clearly seen (from the form of the posterior Dirichlet above) to reflect the relative strength of the prior relative to the amount of observed data, namely $N = \sum_k r_{*k}$.

In our discussion below we will focus on a number of key aspects for selecting a prior in the context of *LR* calculations for count data. In particular we discuss both non-informative priors and informative priors and discuss options for how each can be specified. We also discuss the specification of the concentration parameter c for both the non-informative and informative cases.

While ultimately, as in any Bayesian analysis, the choice of a prior should be informed by an investigator's prior knowledge (or lack of knowledge) about a particular problem, the discussion below should nonetheless provide general guidance to investigators for prior selection when evaluating our proposed approach.

We also note that arguments in favour or against a particular type of prior in the literature are often based on the prior's influence on the posterior distribution of the parameters θ. However, our focus here is on the likelihood ratio rather than on the posterior, so some of the standard Bayesian arguments about priors may not be pertinent. With this in mind, at the end of this section, after discussing both non-informative and informative priors, we analyse directly how the choice of prior can influence the likelihood ratio in Equation (6).

5.2 *Non-informative Dirichlet priors*

A necessary condition for a non-informative Dirichlet prior is that the prior is symmetric (Section 2.5), reflecting the fact that before we see the data for our problem we have no prior knowledge that any of the categories $k = 1 \dots, K$ are likely to occur more frequently than the others. The key question then becomes how to select the concentration parameter $c = \sum_{k=1}^{K} \alpha_k$, or equivalently how to select $\alpha = \left(\frac{c}{K}, \frac{c}{K}, \dots, \frac{c}{K}\right)$.

To this end, there has been considerable discussion among researchers in Bayesian methodology over the selection of an appropriate concentration parameter in order for a Dirichlet prior to be truly non-informative. A full discussion on this topic is beyond the scope for this work, but we highlight below some of the more well-known choices of non-informative Dirichlet priors from the literature.

A popular choice for a non-informative prior, given its easy interpretability, is the uniform Dirichlet distribution (also referred to as the Bayes-Laplace prior), i.e. $\alpha = (1, 1, \dots, 1)$ and $c = K$ (see e.g. Gerlach *et al.* (2009); Tuyl *et al.* (2008)). As discussed earlier, this prior assigns equal prior density to all possible K-ary vectors of event probabilities θ.

Jeffrey's prior is $\alpha = \left(\frac{1}{2}, \frac{1}{2}, \dots, \frac{1}{2}\right)$, i.e. $c = \frac{1}{2}K$, and arises from Jeffrey's invariance principle, which posits that the prior density should give an equivalent result if applied to a transformed version of the parameter, when that transformation is one-to-one. However, Jeffrey's prior can be controversial in multiparameter models, since it can yield different results than simply assuming independent non-informative prior distributions for the different components of θ. Jeffrey's prior and some of its associated issues are further discussed in Gelman *et al.* (2020) and Berger *et al.* (2015).

Berger *et al.* (2015) introduce an 'overall objective prior' which is based on the idea of marginal reference posteriors (see also Bernardo (1979) and Bernardo (2011)). Based on their work, they advocate for $c = 1$, i.e. the prior $\alpha = \left(\frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K}\right)$, as the overall objective Dirichlet prior; although Tuyl (2017) points out some issues with this prior in the case of zero counts for any of the categories in the observed data r.

Lastly, it can be argued that an appropriate parameter setting for a non-informative prior is $c = 0$, i.e. $\alpha = (0, 0, \dots, 0)$, in which we impose no 'pseudocounts'. This is an improper distribution since its density integrates to ∞ rather than to 1. In the $K = 2$ case, this is referred to as Haldane's prior (see e.g. Zellner (1996), Gelman *et al.* (2020), Zhu and Lu (2004)), but the ideas behind it can generally be extended to general K as well (see e.g. Terenin and Draper (2017)). This prior can be motivated by the fact that it is uniform in the $\log(\theta_k)$'s (i.e. the natural parameter in the exponential family representation of the multinomial, see e.g. Gelman *et al.* (2020)). In standard Bayesian inference, however, one would need to observe a count in each of the K categories in order for this prior to yield a proper posterior (a constraint that can't be guaranteed in general before we see the data). For the likelihood ratio calculations of interest in this article, this prior results in a division by zero in the formula for the LR, so Haldane's prior is not directly applicable to our framework.

5.3 Informative priors and pseudocounts

In contrast to the non-informative case, the informative prior can be used when we have relevant prior knowledge about the values of the components in θ, e.g. that certain categories are more likely to occur than others. In this situation, the α_k's are no longer equal, and there are two issues to consider. First, how the relative size of the α_k's should be determined, and second (as with the non-informative prior), how the overall concentration parameter $c = \sum_{k=1}^{K} \alpha_k$ should be specified.

In general we can write $\alpha = c \times \left(\frac{\alpha_1}{c}, \frac{\alpha_2}{c}, \ldots, \frac{\alpha_K}{c}\right)$. This form emphasizes that each term $0 < \frac{\alpha_k}{c} < 1$ can be selected as the relative prior belief in each category k, with c (the pseudocount) controlling the overall strength of the prior. The prior could be specified subjectively by an investigator, manually setting higher values for categories which are known to be more common and then selecting c by its pseudocount interpretation (e.g. how many datapoints N would be required to balance the effect of the prior in the posterior).

An alternative approach is to use a reference dataset. For example, if an investigator is analysing counts that correspond to the use of different software apps on a mobile phone, there may be relevant reference data available in the form of published data on population usage of the same software apps. In particular, consider reference data $d = (d_1, d_2, \ldots, d_K)$ in the form of counts for each of the K categories, e.g. from some reference database that is known a priori. A natural approach here would be to set the prior to be proportional to the reference counts, i.e.

$$\alpha = c \times \left(\frac{d_1}{\sum_{k=1}^{K} d_k}, \frac{d_2}{\sum_{k=1}^{K} d_k}, \ldots, \frac{d_K}{\sum_{k=1}^{K} d_k} \right)$$

where c controls the strength of the prior. A variation of this approach was discussed in a forensic investigation context by Aitken *et al.* (2021, pp. 793–798), who propose applying the equation with each of the d_k's replaced by $d_k + 1$ and choosing $c = \sum_{k=1}^{K} (d_k + 1)$. Adding 1 to each count avoids the potential issue of division by zero in the LR calculations if any of the reference values $d_k = 0$.

Any method for selecting the prior distribution should take into account the relevant population of potential sources (Section 2.2). In principle, non-informative priors should reflect a lack of prior knowledge about the relevant population's behaviour in the event categories; informative priors should reflect the prior knowledge about the relevant population's behaviour in the event categories. For setting informative priors via a reference dataset, care should be taken that this reference dataset reflects the relevant population. In practice, finding and choosing such a dataset can be quite difficult to do (see e.g. Champod *et al.* (2004)). To better understand the relationship between the choice of the prior and the resulting likelihood ratios, in the next section we will examine the impact of different priors on the LR.

5.4 *Effect of the prior on the LR*

Having discussed a number of different aspects of how the Dirichlet prior can in general be specified, we now turn our attention to examining how the choice of prior can influence the likelihood ratio. First we point out that in the case of a symmetric prior, the likelihood ratio can be written (as in Equation (6)) as

$$LR = \left(\prod_{k=1, r_{2k}\geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\frac{c}{K}+s}\right)\right) \left(\prod_{s=0}^{N_2-1} \left(1 - \frac{N_1}{c+N_1+s}\right)\right)$$

where $\alpha = \left(\frac{c}{K}, \frac{c}{K}, \ldots, \frac{c}{K}\right)$. For K, r_1, and r_2 fixed, as $c \to \infty$ each individual factor in the set of products above approaches 1; hence in the limit we have that as $c \to \infty$ the $LR \to 1$. Thus as the prior gets extremely strong, observing the data has little effect on our beliefs, and the likelihood ratio remains close to the neutral value of 1. The result also holds in the asymmetric case, i.e. $LR \to 1$, if we proportionally send all of the individual prior parameters to ∞.

More generally, we can analyse each set of factors from Equation (6) to gain additional insight:

$$LR = \underbrace{\left(\prod_{k=1, r_{2k}\geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k+s}\right)\right)}_{\text{Term (a)}} \underbrace{\left(\prod_{s=0}^{N_2-1} \left(1 - \frac{N_1}{c+N_1+s}\right)\right)}_{\text{Term (b)}}. \tag{8}$$

For each factor in Term (a), a higher α_k decreases the value of that factor, i.e. for $\alpha_x < \alpha_y$

$$1 + \frac{r_{1k}}{\alpha_x + s} > 1 + \frac{r_{1k}}{\alpha_y + s}$$

for constant r_{1k} and s. An intuitive way of thinking about this is that up-weighting a particular α_k in the prior will decrease that particular category's effect on the likelihood ratio, provided that c, N_1 and N_2 are held constant. In other words, observing a shared event in a common category does not impact the likelihood ratio as much as observing a shared event in an uncommon category.

Recall that it is exclusively Term (b) in Equation (6) that decreases the value of the LR in the calculation, and Term (b) is a function of the amount of data and the concentration parameter. The LR is exactly equal to the value of Term (b) in the case in which the known and unknown source data have no observed counts in overlapping event categories, i.e. $S = \emptyset$, to use the notation from Section 4.3.2. In this special case, it is guaranteed that the $LR \leq 1$ regardless of the setting of the prior parameters. We also consider the behaviour of Term (b) for general r_1 and r_2. For c and N_1 fixed, Term (b) decreases as N_2 increases, and similarly for c and N_2 fixed, Term (b) decreases as N_1 increases. Intuitively, this means that if more data is obtained from either source, then more overlap between the event counts, i.e. larger Term (a) values, is required to obtain a $LR > 1$ since Term (b) will be closer to 0 with more data. If instead we hold the amount of data fixed, then as c increases Term (b) also increases. Thus, for stronger values of the prior the value of Term (b) is closer to the neutral value of 1.

The value of Term (b) can become extreme when there is an imbalance between c versus N_1 and N_2. For $c \gg N_1, N_2$, Term (b) remains very close to 1 such that even in the case of no overlap the value of the LR is still close to the neutral value. For example, consider the case in which $K = 1000$,

$N_1 = N_2 = 10$, and we select a uniform Dirichlet prior such that $c = 1000$. This leads to a value of 0.91 for Term (b), which means that this is the smallest any *LR* can be with this prior and this amount of data, regardless of the amount of non-overlap observed in the counts. If, on the other hand, $c \ll N_1, N_2$, then the value of Term (b) can become extremely close to 0. For example, consider the case in which we place a uniform prior over $K = 10$ categories such that $c = 10$. If $N_1 = N_2 = 100$, then the value of Term (b) becomes 1.14×10^{-49}. This means that to obtain an $LR > 1$ one would need to observe a large amount of event count overlap between r_1 and r_2 in order to overcome the extremity of the Term (b) values.

The properties of the prior, and its effect on the *LR*, as discussed above in this section and in Section 4, will be used to motivate our prior choices in the experiments with real-world data below in Section 7.2.

6. Use cases in digital evidence

As discussed earlier, digital evidence from devices such as mobile phones and computers is increasingly common. One common form of digital evidence in this context is logfiles recording the actions that have been taken on a device by its user (see e.g. Casey (2011); Roussev (2016); Cheng *et al.* (2021)). This data is often in the form <`ID, event, timestamp`>, where the ID identifies the user/account/device, the event is one from a set of possible user actions, and the timestamp is the time at which the user initiated the event, e.g. <`John Doe's phone, Opened App X, 09/24/2021 5:15pm`>. In this section we discuss how this form of user-generated event data can be analysed by converting event records to user-event counts and then applying our likelihood-ratio approach to identity-related questions.

6.1 *Motivating scenario*

Suppose that investigators have recovered a digital device from a crime scene and have extracted from the device historical event logs of its user-generated actions. Suppose also that the device's owner is now a suspect in the investigation. A common defence in cases like this is to claim that the recovered device was stolen or otherwise not in the suspect's possession during the period of criminal activity (see e.g. Casey *et al.* (2020)). Using the device logs, we would like to be able to answer questions such as: how likely is it that these kinds of activities would be observed if the suspect's claim is false, and they actually did possess their device? Or, how likely is it that these kinds of activities would be observed if the device was not in the suspect's possession?

To answer these types of questions, we convert the event logs into count data by tallying how many times a particular action was taken on that device. We can split the counts into two time periods, one period during which the suspect is known to have had the device and another period in which the user is unknown, i.e. the time during which the suspect claims the device was stolen (Figure 3). Ideally an investigator would like to be able to systematically compare the patterns in the event data from the known and unknown sources to determine if there is support for the hypothesis that the two sets of data were actually generated by the same individual. While visual inspection of the data might provide a general intuition of how much evidence there is to support the same-source hypothesis, a quantitative evaluation of the evidence in the form a likelihood ratio (using the methods we have outlined earlier in the article) is preferable.

The count-based likelihood-ratio framework can also be used in the context of other scenarios in digital forensics. For instance, suppose it is believed that a suspect carries two phones on their

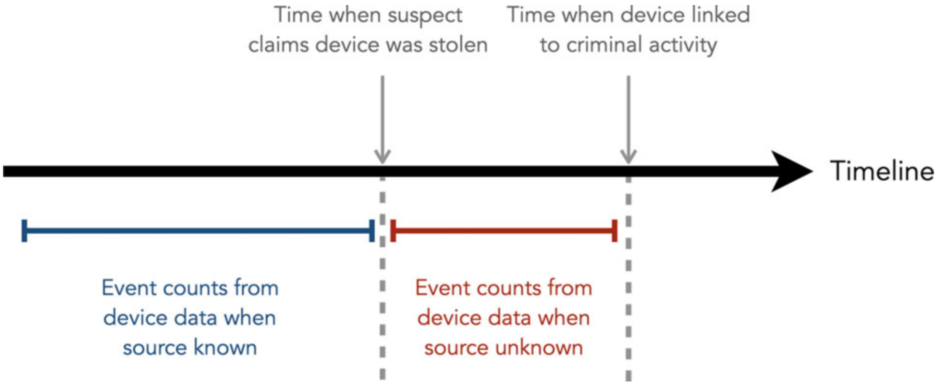

Time when suspect claims device was stolen
Time when device linked to criminal activity
Timeline
Event counts from device data when source known
Event counts from device data when source unknown

FIG. 3. Timeline in the stolen device scenario. Blue event counts correspond to the known-source counts r_1. Red event counts correspond to the unknown-source counts r_2.

person, and one of these phones was recovered from the crime scene. The proposed likelihood-ratio framework could then be used to compare patterns of activities on the recovered phone to patterns of activities on the phone still on the suspect, where the activities in general can include opening and closing of apps and files, emails or text messages sent or received, geolocations visited, and so on.

6.2 Datasets

In order to evaluate how our approach applies to digital evidence, we consider three real-world datasets that each consist of rows in the general format `<ID, event, timestamp>`. For each of these datasets, we describe below basic characteristics of the data and how the event categories were chosen[1]. Table 2 presents summary statistics.

6.2.1 *Email communications.* The first dataset consists of 285 929 emails sent between October 2003 and May 2005 (Paranjape *et al.*, 2017). We organized this data into a set of email events per sender in which each event consists of `<sender_id, recipient_id, timestamp>`. The sender and recipient IDs identify the accounts who sent and received the emails, respectively, and the timestamp is the time at which the email was sent. This resulted in 655 unique sender IDs and 946 unique recipient IDs, corresponding to 655 accounts and 946 event categories in our framework.

6.2.2 *Device geolocation records.* The second dataset comes from the Twitter Streaming API (Twitter, 2021) and consists of 74 663 geolocated tweets from Orange County, CA collected from September 2020 to March 2021. Each tweet event consists of the following information: `<user_id, latitude, longitude, timestamp>`, where the user ID identifies the Twitter account, the latitude and longitude coordinates are the location of the device from which the tweet

[1] Additional information about the datasets and how to access them is available publicly at https://ucidatalab.github.io/uci-digital-evidence/datasets/.

Table 2. *For each of the three datasets, the number of users, the number of categories (K), the mean and median number of events per user, and the number of events per category during the collection period (excluding holdout data used for setting the informative priors, as described in Section 7.2)*

			Num. events per user		Num. events per category	
Dataset	Num. users	K	Mean	Median	Mean	Median
Email	655	946	436.53	184.0	302.25	140
Geolocation	2265	1412	32.70	8.0	52.77	8
App	260	159	6957.63	1710.5	11377.26	3875

was sent, and the timestamp is the time at which the tweet occurred (Figure 4). These are public tweets for which users have opted in to sharing their geo-coordinates.

To define event categories, we mapped the latitude and longitude values to a set of census block groups (CBGs) which partition Orange County into approximately 1800 polygons covering the entire county (Figure 5). We removed CBGs in which no events were ever observed, resulting in 1412 CBGs under consideration. The data we used to define these polygons was downloaded from the Census Reporter website and comes from the American Community Survey 2019 1-year OC Total Population U.S. Census data (U.S. Census Bureau, 2019). Thus, each event category corresponds to sending a geo-located tweet from a particular CBG. Other representations of spatial information, such as geoparcels, could potentially be used as alternatives to categorize geo-located events. We used CBGs since they are well-defined, publicly accessible and provide a reasonable tradeoff between spatial resolution and data sparsity.

6.2.3 *Mobile app usage.* The third dataset consists of records of app usage on mobile phones in which each event consists of `<user_id, app_name, event_type, timestamp>` (Aliannejadi *et al.*, 2021). There are approximately 1.8 million such app usage records, generated by Android users across 86 different apps from September 2017 to May 2018. The possible event types are: Opened, Closed, User Interaction and Broken, and the app name identifies the mobile app on which the user initiated the event. For our analysis, we only consider events of the type Opened or User Interaction since almost every Opened event is subsequently followed by a Closed event, and the Broken event types may be system-generated rather than user-generated events. We define event categories by taking the combination of the app and whether the user opened or interacted with the app, e.g. Opened App X or Interacted with App Y. However, for 13 apps, no User Interaction events were observed across any of the app usage logs. For these apps, we only considered Opened events, resulting in $86 \times 2 - 13 = 159$ different event categories.

7. Computational experiments and results

7.1 Experimental set-up

For each dataset, we divide the time range over which the data was collected into two periods and count the event data for each user in each of the two time periods. For each user, there are two sets of event data: one set coming from the first half of the collection period and the other set coming from the second half of the collection period. To obtain data with which we can calculate likelihood

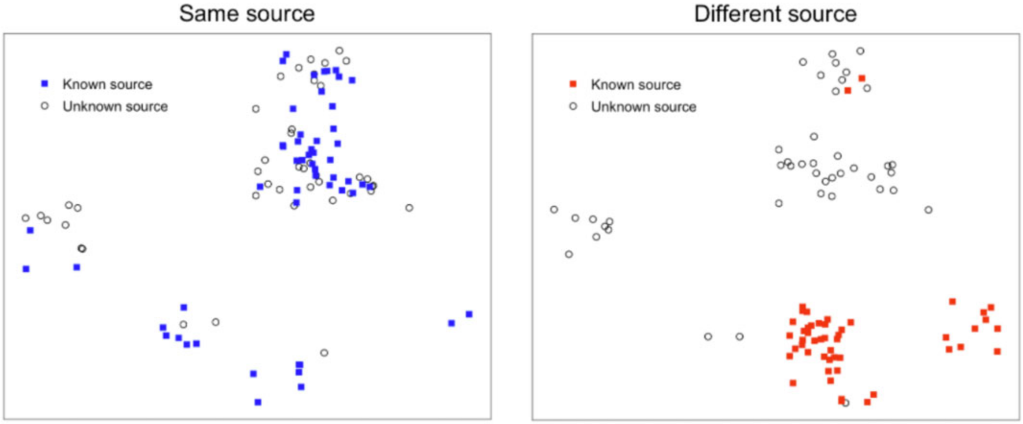


FIG. 4. Example geolocations of publicly available geo-located Tweets. The plot on the left shows Tweets generated by the same account, while the plot on the right shows Tweets generated by two different accounts.

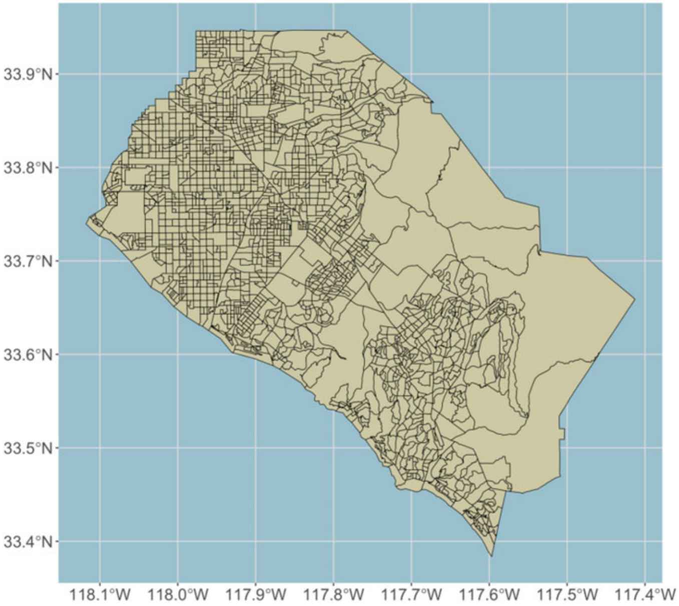


FIG. 5. Census block groups (CBGs) in Orange County, CA. Event categories for the Twitter data correspond to sending a geo-located tweet from coordinates located in a particular CBG.

ratios, we treat all of the user data coming from the first half of the collection period as having come from a known source, and the user data coming from the second half of the collection period as having come from an unknown source (though in reality we do know the user), as illustrated in Figure 6. Users with zero events during either of the two time periods were not used in the experiments.

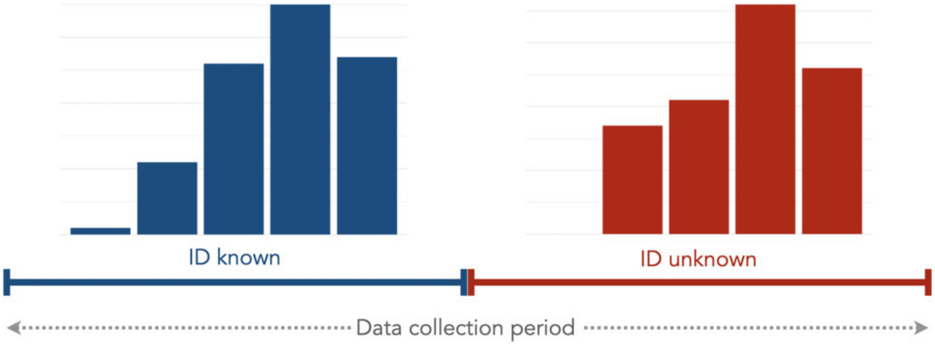


FIG. 6. Illustration depicting the general experiment setup. Count data coming from the first half of the collection period is treated as having come from a known source, and count data coming from the second half of the collection period is treated as having come from an unknown source.

For each set of event counts we treat as coming from a known source, we create a same-source pair by partnering it with that user's event data from the second half of the collection period (treated as the unknown source in *LR* calculation). We construct a different-source pair by partnering any two distinct users' event data where one set comes from the first half of the collection period and the other from the second half. For our experiments, we calculate the likelihood ratio using all of the same-source pairs and a simple random sample of the different-source pairs, with 50% same-source and 50% different-source pairs. The choice of a 50–50 split (versus some other split) was chosen for convenience and does not (in the limit as the size of the evaluation dataset increases) affect the values we obtain for evaluation metrics, nor is it intended to reflect the proportion of same- versus different-source comparisons that a forensics investigator might encounter in practice.

For the email dataset, users sending the emails are also present in the recipients but do not send emails to themselves. We circumvent this issue by removing the known sender as one of the categories within each comparison. For example, if our known source data was coming from the ID 1234, we would not count any emails to 1234 among both the known and unknown source data. Thus each likelihood ratio calculation for this dataset is computed using $946 - 1 = 945$ event categories.

For the mobile app dataset, the app usage records were not collected for the same period of time for each user, e.g. some users have data from a few days within the collection period while others have a few weeks of data. To circumvent this problem, for each set of known source data, we construct an unknown source match by pulling event data from the second half of the known user's collection period. For example, if the known user's total data collection happened from September 1 through 10, their known source data would be their event counts from September 1–5 and their unknown source event data would be their event counts from September 6–10. To find a different source to compare them to, we would only consider other users for which we have event data from September 6–10, and this is the data that would be used in the likelihood ratio calculation. This ensures that there is consistency in the data collection dates for same-source and different-source pairs in which the known source data is coming from the same user.

7.2 *Prior selection*

For each dataset, we consider two different priors in our experiments: a non-informative (symmetric) prior and an informative (asymmetric) prior. The non-informative prior is motivated by the discussion in Section 5.2. To this end, we opt to choose the uniform Dirichlet distribution as the prior, i.e. $\alpha = (1, 1, \ldots, 1)$, resulting in a concentration parameter $c = K$. We use the uniform prior for two main reasons:

- For a given dataset, choosing constant α_k values which do not depend on K ensures that we maintain a constant upper bound in the situation in which irrelevant zero-count categories have been included (Section 4.3.3).
- The uniform prior is straightforward to interpret since it assigns equal density to all probability vectors satisfying the linear constraint that $\sum_{k=1}^{K} \theta_k = 1$ (Section 2.5).

To construct the informative prior, we take a sample of data across all users to use as reference data. For the email and geolocation datasets, we hold out the data coming from the first 10% of the collection time window (corresponding to 12.4% and 13.1% of the total number of events, respectively), and for the app dataset, because the data collection period is varied across users, we hold out data coming from a random sample of 30 users (corresponding to 9% of the total number of events). We count all the events across the reference dataset, yielding marginal counts across all of the categories, denoted by $(d_1, d_2, \ldots, d_K)$. We then use a strategy similar to the one described in Section 5.3 in which we set the prior to be

$$\alpha = K \times \left(\frac{d_1 + 1}{\sum_{k=1}^{K} (d_k + 1)}, \frac{d_2 + 1}{\sum_{k=1}^{K} (d_k + 1)}, \ldots, \frac{d_K + 1}{\sum_{k=1}^{K} (d_k + 1)} \right).$$

This ensures that the likelihood ratio values from both of the priors share the factors from Term (b) in Equation (8) (also a lower bound, Section 4.3.2) since c, N_1, and N_2 will be unchanged. However, the factors in Term (a) from Equation (8) will be up-weighted or down-weighted according to their frequency in the marginal data, such that more common categories will have a smaller impact on the *LR* than with the non-informative prior, and uncommon categories will have a larger impact (Section 5.4).

With this reference data strategy, we are also effectively treating all of the users in each dataset as samples from the relevant population (Section 2.2). For the email dataset, all the users are affiliated with the same large university. For the geolocation dataset, all accounts opted in to sharing geolocation data and all have geolocated tweets in Orange County, CA during the collection period. For the app dataset, all users were recruited via Amazon Mechanical Turk. These implied relevant populations are used for convenience for our experiments, but as discussed in Section 2.2, Section 4.3.3, and Section 5.3, the selection of the relevant population is an important decision in practical applications that should be informed by the facts of the case.

7.3 Results

In Table 3, we summarize the results of the computational experiments using three metrics: true positive rate using 1 as a threshold (TPR@1), false positive rate using 1 as a threshold (FPR@1), and area under the receiver operating characteristic (AUC). We chose 1 as a threshold for the TPR and FPR because of its straightforward interpretability, as described in Section 2.4.

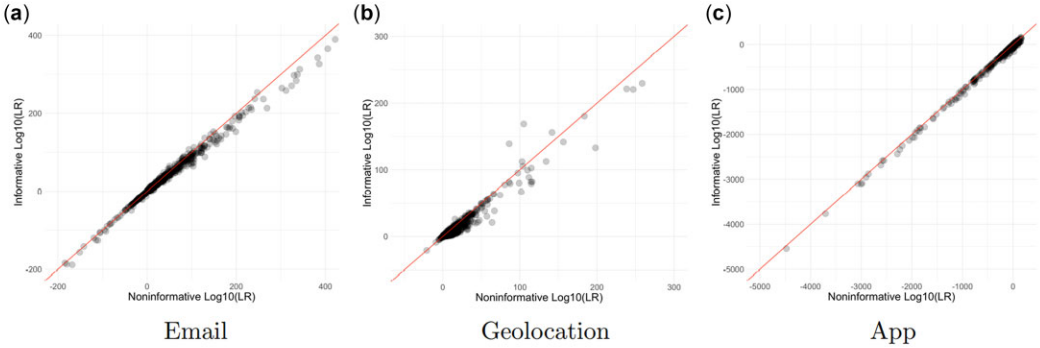
Across all three datasets, we observed AUC values generally close to 1. For the email and geolocation datasets, using the informative prior resulted in slightly higher AUC values and only slightly lower TPRs than those resulting from the non-informative prior. In contrast, for the app dataset, the informative prior yielded a lower AUC and a TPR that was substantially lower than that of the non-informative prior. We speculate that this is because much of the overlap in the same-source pairs arises from shared counts in apps that are commonly used by all users (e.g. Facebook, Google, Twitter, Instagram) and the impact of sharing these event counts is reduced using the informative prior (to be further discussed in the context of the plots below). Across all datasets, all FPRs were below 12% for both non-informative and informative priors. Using the informative prior resulted in lower FPR values than those of the non-informative prior, with the greatest relative reduction for the app dataset.

Figure 7 shows the values of the *LR* using a non-informative prior versus the *LR* using an informative prior. From these plots, it can be seen that the two different priors yield similar *LR* values, but using the informative prior often results in a slightly decreased *LR* value (below the diagonal) compared to that from the non-informative prior, and this difference can be more dramatic for larger values of the *LR*. These results align with the discussion in Section 5.4 in that the informative prior has the effect of dampening the impact of the more common categories on the *LR*. Hence a majority of the *LR* values across the experiments are slightly decreased. Less frequently, using the informative prior also takes into account the fact that a category is rare and increases this category's impact on the *LR* when it appears as a shared category in the evidence. This is most easily seen in Figure 7(b) where a few points are clearly above the identity line. In general, however, we did not find the difference in results for the non-informative and informative priors to be as impactful as the setting for the concentration parameter (discussed following Figure 8), i.e. our results were relatively insensitive to whether we used informative or non-informative priors.

Figure 8 shows boxplots of the *LR* values obtained for pairs in each dataset, as a function of the amount of data used in the *LR* calculations. Across all three datasets, we can see a separation between the *LR* values for same-source versus different-source pairs, and these values become more extreme with high amounts of data. Also for all three datasets, the median *LR* for the same-source pairs is greater than 1, while the median *LR* for the different-source pairs is less than 1. These trends

Table 3. *Likelihood ratio results are expressed as percentages for the three real-world datasets. TPR@1 and FPR@1 are the true and false positive rates, respectively, using 1 as a threshold for the LR. AUC is the area under the ROC curve (50% indicates a random classifier; 100% indicates a perfect classifier)*

	TPR@1	FPR@1	AUC
Email, $\boldsymbol{K} = 945$			
Non-informative	94.8%	10.1%	98.1%
Informative	94.7%	8.1%	98.5%
Geolocation, $\boldsymbol{K} = 1412$			
Non-informative	72.8%	6.4%	91.5%
Informative	72.7%	5.1%	92.4%
App, $\boldsymbol{K} = 159$			
Non-informative	75.0%	11.5%	86.3%
Informative	63.8%	6.2%	82.5%



Email Geolocation App

FIG. 7. Scatterplots for each of the three datasets plotting the value of the $\log_{10} LR$ for the same r_1 and r_2 when using the non-informative versus the informative prior. The $y = x$ line is plotted in red.

also hold for the experiments using the informative prior so we only show the boxplots using the non-informative prior here.

Figures 7–8 make apparent the fact that there are systematic imbalances in the values of the likelihood ratios for both the geolocation and app datasets. For the geolocation dataset, a majority of the different-source LR values are close to the neutral value of 1 (even for 'high' amounts of data), while the same-source values are more extreme. In contrast, for the app dataset, many of the same-source LR values are close to the neutral value while the different-source values become very extreme. With both of these datasets, we believe this to be a consequence of an imbalance between the concentration parameter and the amount of data (as discussed in Section 5.4). As a result of using the uniform Dirichlet distribution as the prior, the geolocation dataset has concentration parameter $c = 1412$ while the medians for N_1 and N_2 are 4 and 3 events, respectively (Table 4). Using the notation from Equation (8) in Section 4.3.2, this results in $c \gg N_1, N_2$; in fact, calculating Term (b) in Equation (8) using the median values and $c = 1412$ yields a value of 0.99 ($\log_{10} 0.99 = -0.004$, for reference in the plots). Consequently, many of the pairs of counts in the geolocation dataset remain close to the neutral value, and those that we have defined as using high amounts of data still largely suffer from such imbalance. The app dataset, on the other hand, has an imbalance in the opposite direction, with $c \ll N_1, N_2$. The uniform Dirichlet prior for the app dataset gives $c = 159$ while the medians for N_1 and N_2 are 922.5 and 597.5, respectively (Table 4). Calculating the value of Term (b) in Equation (8) using these values (medians rounded down to the nearest whole number) gives $1.83\text{e} - 306$ ($\log_{10} 1.83\text{e} - 306 = -305.74$). Thus, in order to obtain an $LR > 1$, substantial event count overlap is required. As is evident from the TPR@1 value, a majority of the same-source pairs achieve such overlap. This imbalance, however, means that many of the different-source pairs' LR values are very extreme, even in the presence of some category overlap. The problem is generally exacerbated when using the informative prior because, as mentioned before, much of the overlap in the same-source pairs arises from categories that are common across all of the users, and overlap in these categories has less of an impact when using the informative prior. The extremity of these values aside, in theory the demonstrated behaviours could be reasonable. For instance, in the presence of many potential categories and little data, one may want to have fairly neutral LR values, and any overlap between the many event categories could be a strong indicator of similarity. Similarly, in the presence of few categories but a substantial amount of data, one may deem that considerable overlap should be required to support the same-source hypothesis. However, the results of these

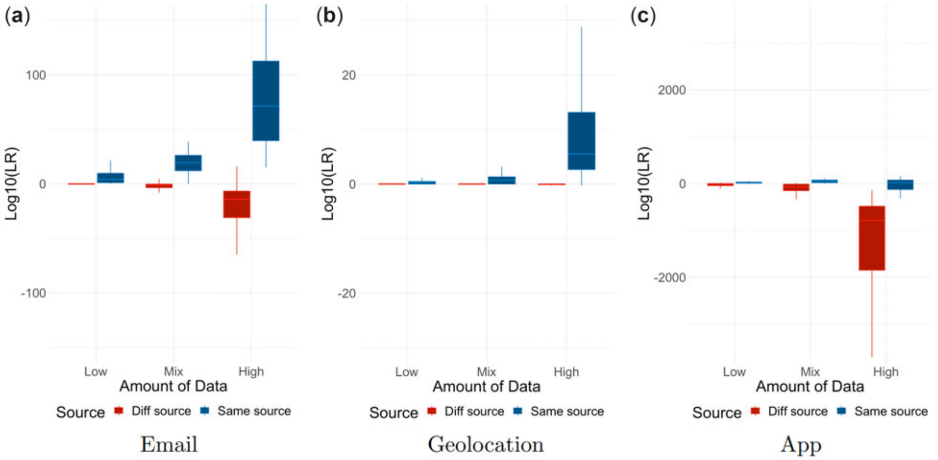


Fig. 8. Boxplots of the $\log_{10} LR$ values for same- and different-source pairs by the amount of data (using the non-informative prior). A low amount of data is the case in which both N_1 and N_2 are less than their medians, and a high amount of data is the case in which both N_1 and N_2 are greater than their medians. Cases that are neither low nor high are considered mixed. Values for these medians are in Table 4.

Table 4. *Median amount of data by dataset for the first and second halves of the collection period. Note that these medians are taken only across the same-source pairs, such that each user is only represented once in the median calculation*

Dataset	Median(N_1)	Median(N_2)
Email	122.0	44.0
Geolocation	4.0	3.0
App	922.5	597.5

experiments demonstrate that these behaviours could quickly become quite extreme. We look at the extremity of these values more closely in Table 5.

Table 5 shows that the experiments across all three of the datasets yielded likelihood ratios with extreme values, some of which were contrary to fact. The trends in these extreme values were similar between the email and geolocation datasets but generally differed for the app dataset. For the email and geolocation datasets, the percentage of extreme values increased with the amount of data, and extreme contrary-to-fact *LR* values were only observed for mixed or high amounts of data and only for the different-source pairs. The percentage of extremely small values for the app dataset increased with the amount of data but was unstable for extremely large values. This dataset also has the strongest presence of extreme contrary-to-fact *LR* values, particularly extremely small *LR* values for same-source pairs. This is consistent with the observations in Figures 7–8 and makes clearer the effects of the sensitivity to the imbalance between the concentration parameter and the amount of data discussed above. For higher amounts of data, this imbalance is only exacerbated. Overall, Table 5 suggests that the values of the *LR* are too extreme and that the model lacks suitability for the app dataset. Such behaviour could easily go unnoticed, however, if one were to only examine the performance using the metrics in Table 3 since many of the values are on the desirable side of the threshold.

To capture the extremity of these values in a performance metric, we also summarize the results of the experiments using the log-likelihood ratio cost in Table 6. The formula for the log-likelihood ratio cost is

$$C_{llr} = \frac{1}{2N_s}\sum_{i\in\mathcal{D}_s}\log_2\left(1+\frac{1}{LR_i}\right)+\frac{1}{2N_d}\sum_{i\in\mathcal{D}_d}\log_2(1+LR_i) \tag{9}$$

where N_s and N_d are the total numbers of same-source and different-source pairs respectively, $\mathcal{D}_s$ and $\mathcal{D}_d$ denote the indices of the same-source and different-source pairs respectively, and LR_i is the likelihood ratio value evaluated for pair i. This metric penalizes performance according to the magnitude of the resulting likelihood ratios rather than using a threshold (Brümmer and du Preez, 2006). Smaller values of C_{llr} indicate better performance, and always returning $LR = 1$ would give $C_{llr} = 1$. Brümmer and du Preez (2006) also showed that the C_{llr} can be decomposed into two components as follows:

$$C_{llr} = C_{llr}^{min} + C_{llr}^{cal}. \tag{10}$$

C_{llr}^{min} denotes the discrimination loss, which measures how well the method differentiates between same-source and different-source pairs. This is calculated by using the PAV algorithm to transform the resulting LR values and then recalculating the C_{llr} using these transformed values instead. The PAV-transformed LR values are optimized to achieve the minimum value of the log-likelihood ratio cost while still maintaining the rank ordering of the untransformed values. Hence, the transformed LRs have the same level of discriminatory power as the untransformed LRs, but their magnitudes will differ. C_{llr}^{cal} denotes the calibration loss, which measures the difference in the log-likelihood

Table 5. *Extreme likelihood ratio values expressed as percentages by dataset, amount of data, and same versus different source. For example, in the email dataset, 24.4% of the LR values for same-source pairs with a low amount of data (defined as in Figure 8) were extremely large. Percentages corresponding to extreme contrary-to-fact LRs are shown in bold*

	Same source		Different source	
	$LR < 10^{-10}$	$LR \geq 10^{10}$	$LR < 10^{-10}$	$LR \geq 10^{10}$
Email, $\boldsymbol{K} = 945$				
Low	0.0%	24.4%	0.0%	0.0%
Mix	0.0%	76.2%	6.0%	**1.1**%
High	0.0%	100.0%	62.5%	**3.3**%
Geolocation, $\boldsymbol{K} = 1412$				
Low	0.0%	0.0%	0.0%	0.0%
Mix	0.0%	0.0%	0.0%	0.0%
High	0.0%	32.2%	0.2%	**0.4**%
App, $\boldsymbol{K} = 159$				
Low	0.0%	67.0%	50.0%	**3.9**%
Mix	**11.1**%	75.0%	78.6%	**2.9**%
High	**42.9**%	53.6%	100.0%	0.0%

Table 6. *Log-likelihood ratio cost results. The log-likelihood ratio cost (C_{llr}) can be decomposed into the cost due to discrimination (C_{llr}^{min}) and the cost due to calibration (C_{llr}^{cal}). Smaller values of C_{llr} indicate better performance. An LR system that always returns a value of 1 for the LR would give $C_{llr} = 1$*

	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}
Email, $\boldsymbol{K} = 945$			
Non-informative	0.791	0.214	0.577
Informative	0.603	0.194	0.408
Geolocation, $\boldsymbol{K} = 1412$			
Non-informative	0.778	0.482	0.297
Informative	0.771	0.475	0.296
App, $\boldsymbol{K} = 159$			
Non-informative	113.488	0.556	112.932
Informative	140.266	0.622	139.644

ratio cost resulting from how far the magnitudes of the untransformed *LR*s are from the optimal PAV-transformed values.

From Table 6, it can be seen that both the email and geolocation datasets produced log-likelihood ratio cost values less than 1, with lower costs when using the informative prior. For all three datasets, much of the C_{llr} can be attributed to the cost due to miscalibration (C_{llr}^{cal}); the log-likelihood ratio cost results for the app dataset, in particular, are quite glaring. This may indicate that for the app data the multinomial model is especially inappropriate. For example, the model might not sufficiently account for variability in individual behaviour over time, which is supported by the large number (42.9% in Table 5) of extremely small *LR* values (contrary to fact) for same-source pairs with high amounts of data. We discuss these limitations and potential directions for handling these issues in the discussion section below.

8. Discussion

A feature of the categorical-multinomial modelling approach we investigated in this article is that the likelihood ratio is available in closed form, allowing for easy computation, supporting analysis and interpretation (such as how changes in various quantities impact the resulting likelihood ratio). For example, in a particular investigation, for a fixed set of event categories and prior settings, one could calculate multiple *LR*s under a variety of possibilities for the evidence to better understand its potential behaviour in the context of that investigation. This would be useful, for example, in analysing the potential effects of imbalances between the concentration parameter and the amount of data, as was observed with our experimental choices with the geolocation and app datasets.

In this article, we focused on particular examples of count data related to digital evidence, but the methodology is broadly applicable to count data in general, whether in the context of digital evidence or more broadly in forensic investigations in general. Relevant case-specific circumstances should be used to justify the choices for the relevant population (Section 2.2), the event categories (Section 4.3.3), and the prior distribution (Section 5). As discussed earlier, these choices can have a strong influence on the resulting likelihood ratios and should be carefully considered on a case-specific basis.

The results from the experiments with the three real-world datasets suggest that while the count-based multinomial model is able to capture useful information in terms of discrimination (Table 3), the magnitudes of the resulting likelihood ratios have weak performance in terms of calibration (Tables 5–6). One approach that has been proposed to address miscalibration is through post-hoc calibration methods that adjust the resulting likelihood ratios themselves (e.g. Morrison (2013)). However, it may also be prudent in this case to address calibration by improving upon the count-based model itself. To this end, we identify and discuss three limitations of the model below.

First, individual human behaviour can be highly variable over time, i.e. a user's behaviour can (and likely will) change over time for a number of different reasons. The same user's behaviour could appear to be completely different in two different time windows, or two different users could appear to have similar behaviour patterns. The proposed model does not take into account this time-varying aspect of users' behaviour on their devices because the probability vectors θ are assumed to be constant through time. Future work could take into account deviations from these static event probabilities, e.g. where users have daily or weekly fluctuations in their behaviour but tend to return to the same core behaviour, or where users drift in their behaviour over time.

Second, as mentioned in Section 2.3, the multinomial distribution imposes a strong memoryless assumption on the data in which the events are assumed to be independent. Because of this, dependent behaviour between events is not be captured by the multinomial model. For example, consider the situation in which one user has the sequence of events A, B, C, A, B, C, and the other user has the sequence B, A, C, B, A, C. The multinomial model treats these two as identical sets of counts, even though the sequences are quite different. In the case of geographic data, for example, if someone visited a particular census block group that person may be more likely to visit neighbouring census block groups. Further exploration to account for different kinds of event dependence in the model, such as sequential or spatial, is an avenue for future work.

Third, recall from Section 4.3.3 that an excess of shared zero counts in the event categories leads to larger likelihood ratios under the proposed model. Having a large amount of shared zero counts could be a consequence of considering irrelevant event categories, but it may also be an identifying feature of the source or arise from expected sparsity in the data rather than from meaningful similarities between the two sets of observations. The proposed model currently handles shared zero counts implicitly in the calculation; the categories' inclusion in the concentration parameter leads to an increased likelihood ratio. A future direction of this work, however, could be to explore extensions or alternatives to the proposed model which explicitly incorporate sparsity and investigate their theoretical properties in a similar manner as we have done for the multinomial model here.

These limitations that we have highlighted generally arise from behaviours that the count-based multinomial model cannot capture. Similar limitations with count-based models have been observed elsewhere in forensics, e.g. uncaptured variance (overdispersion) in authorship attribution analyses (Ishihara and Carne, 2022). Addressing these limitations that we have pointed out will help better take into account the variability in behaviour and may improve the calibration performance by making the likelihood ratios more conservative.

Beyond the model's limitations, we return to our earlier point that there is relatively little existing work on the statistical analysis of user-generated event data in forensics. As such, further study into the potential implications and consequences of using such data in forensics is recommended before statistical methods for analysing this data, including the one discussed in this article, are ready for use in forensic practice.

9. Conclusion

In this article, we develop a likelihood ratio-based technique for the statistical forensic analysis of categorical count data. We provide a closed-form solution for the likelihood ratio under our proposed model and illustrate how the resulting likelihood ratio is impacted by the amount of data in the evidence, the determination of events that are of forensic interest, and the choice of the prior used for Bayesian computation. Our approach is particularly relevant to digital forensics in which investigators wish to analyse logs of user actions generated on a digital device. Relatively few statistical methods have been developed for the forensic analysis of such data; however, with the proliferation of digital devices, the development of such methods is only likely to grow increasingly important. To this end, we demonstrate the potential efficacy of our approach through computational experiments on three datasets relevant to digital forensics. The results from these experiments suggest that the proposed methodology provides a useful starting point for the statistical forensic analysis of user-generated event data in digital forensics; however, further work is necessary before this method can be applied in practice.

Funding

This work was funded by the Centre for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 10NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College, and University of Nebraska, Lincoln.

References

Aitken C., Taroni F., and Bozza S. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, 2021.

Aliannejadi M., Zamani H., Crestani F., and Croft W. B. Context-aware target apps selection and recommendation for enhancing personal mobile assistants. *ACM Transactions on Information Systems (TOIS)*, **39**(3):1–30, 2021.

Årnes A.. *Digital Forensics*. John Wiley & Sons, 2017.

Berger J. O. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013.

Berger J. O., Bernardo J. M., and Sun D. Overall objective priors. *Bayesian Analysis*, **10**(1): 189–221, 2015.

Bernardo J. M. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**(2):113–128, 1979.

Bernardo J. M. Integrated objective Bayesian estimation and hypothesis testing. *Bayesian Statistics*, **9**:1–68, 2011.

Biedermann A., Taroni F., Bozza S., and Mazzella W. Implementing statistical learning methods through Bayesian networks (part 2): Bayesian evaluations for results of black toner analyses in forensic document examination. *Forensic Science International*, **204**(1-3):58–66, 2011.

Blei D. M., Ng A. Y., and Jordan M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**(Jan):993–1022, 2003.

Bosma W., Dalm S., van Eijk E., Harchaoui R. El, Rijgersberg E., Tops H. T., Veenstra A., and Ypma R. Establishing phone-pair co-usage by comparing mobility patterns. *Science & Justice*, **60**(2):180–190, 2020.

Brümmer N. and du Preez J. Application-independent evaluation of speaker detection. *Computer Speech & Language*, **20**(2-3):230–275, 2006.

Casey E. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. Academic Press, 2011.

Casey E., Jaquet-Chiffelle D.-O., Spichiger H., Ryser E., and Souvignet T. Structuring the evaluation of location-related mobile device evidence. *Forensic Science International: Digital Investigation*, **32**:300928, 2020.

Champod C. and Evett I. W. A probabilistic approach to fingerprint evidence. *Journal of Forensic Identification*, **51**(2):101, 2001.

Champod C., Evett I. W., and Jackson G. Establishing the most appropriate databases for addressing source level propositions. *Science & Justice: Journal of the Forensic Science Society*, **44**(3):153–164, 2004.

Champod C., Biedermann A., Vuille J., Willis S., and Kinder J. De ENFSI guideline for evaluative reporting in forensic science: A primer for legal practitioners. *Criminal Law and Justice Weekly*, **180**(10):189–193, 2016.

Chen J. and Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, **7**(1):418–442, 2013.

Cheng C. C.-C., Shi C., Gong N. Z., and Guan Y. Logextractor: Extracting digital evidence from android log messages via string and taint analysis. *Forensic Science International: Digital Investigation*, **37**:301193, 2021.

Evett I. and Weir B. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer, 1998.

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

Galbraith C., Smyth P., and Stern H. S. Quantifying the association between discrete event time series with applications to digital forensics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **183**(3):1005–1027, 2020a.

Galbraith C., Smyth P., and Stern H. S. Statistical methods for the forensic analysis of geolocated event data. *Forensic Science International: Digital Investigation*, **33**:301009, 2020b.

Gelman A., Carlin J. B., Stern H. S., and Rubin D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2020.

Gerlach R., Mengersen K., and Tuyl F. Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, **4**(1): 151–158, 2009.

Ishihara S. and Carne M. Likelihood ratio estimation for authorship text evidence: An empirical comparison of score-and feature-based methods. *Forensic Science International*, **334**:111268, 2022.

Johansson M. and Olofsson T. Bayesian model selection for Markov, hidden Markov, and multinomial models. *IEEE Signal Processing Letters*, **14**(2):129–132, 2007.

Johnson H. E., Scott Mills L., Wehausen J. D., and Stephenson T. R. Combining ground count, telemetry, and mark–resight data to infer population dynamics in an endangered species. *Journal of Applied Ecology*, **47**(5):1083–1093, 2010.

Lowe R., Shirley N., Bleackley M., Dolan S., and Shafee T. Transcriptomics technologies. *PLoS Computational Biology*, **13**(5):e1005457, 2017.

Lund S. P. and Iyer H. Likelihood ratio as weight of forensic evidence: a closer look. *Journal of Research of the National Institute of Standards and Technology*, **122**:27, 2017.

Mavridis D. and Aitken C. G. Sample size determination for categorical responses. *Journal of Forensic Sciences*, **54**(1):135–151, 2009.

McCallum A., Nigam K., et al. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, Vol. 752, No. 1, pages 41–48, 1998.

Morrison G. S. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, **125**(4): 2387–2397, 2009.

Morrison G. S. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, **45**(2):173–197, 2013.

Morrison G. S., Enzinger E., and Zhang C. Refining the relevant population in forensic voice comparison–a response to hicks et alii (2015) the importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice*, **56**(6):492–497, 2016.

National Commission on Forensic Science. Ensuring that forensic analysis is based upon task-relevant information. https://www.justice.gov/archives/ncfs/page/file/641676/download, 2015. (Accessed 27 October 2022).

Ommen D. M. and Saunders C. P. Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, **17**(2):179–197, 2018.

Paranjape A., Benson A. R., and Leskovec J. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610, 2017.

Pollitt M., Casey E., Jaquet-Chiffelle D.-O., and Gladyshev P. A framework for harmonizing forensic science practices and digital/multimedia evidence. Technical report, OSAC/NIST, 2018.

Puig X., Font M., and Ginebra J. A unified approach to authorship attribution and verification. *The American Statistician*, **70**(3):232–242, 2016.

Richards S. A. Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, **45**(1):218–227, 2008.

Robertson B., Vignaux G. A., and Berger C. E. *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, 2016.

Rose P. *Forensic Speaker Identification*. CRC Press, 2002.

Roussev V. Digital forensic science: issues, methods, and challenges. *Synthesis Lectures on Information Security, Privacy, & Trust*, **8**(5):1–155, 2016.

Stern H. S. Statistical issues in forensic science. *Annual Review of Statistics and Its Application*, **4**:225–244, 2017.

SWGDE. Best practices for mobile device forensic analysis. https://www.swgde.org/documents/published-complete-listing, 2020a. (Accessed 27 October 2022).

SWGDE. Best practices for mobile device evidence collection and preservation, handling, and acquisition. https://www.swgde.org/documents/published-complete-listing, 2020b. (Accessed 27 October 2022).

Terenin A. and Draper D. A noninformative prior on a space of distribution functions. *Entropy*, **19**(8):391, 2017.

Tuyl F. A note on priors for the multinomial model. *The American Statistician*, **71**(4):298–301, 2017.

Tuyl F., Gerlach R., and Mengersen K. A comparison of Bayes–Laplace, Jeffreys, and other priors: the case of zero events. *The American Statistician*, **62**(1):40–44, 2008.

Twitter. Twitter Streaming API. https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview, 2021. (Accessed 27 October 2022).

U.S. Census Bureau. Total Population American Community Survey 1-year estimates. https://censusreporter.org, 2019. (Accessed 27 October 2022).

Wadsworth W. D., Argiento R., Guindani M., Galloway-Pena J., Shelburne S. A., and Vannucci M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, **18**(1):1–12, 2017.

Zadora G. and Ramos D. Evaluation of glass samples for forensic purposes—an application of likelihood ratios and an information–theoretical approach. *Chemometrics and Intelligent Laboratory Systems*, **102**(2):63–83, 2010.

Zellner A. *Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons Inc., 1996.

Zhang H. and Stern H. Investigation of a generalized multinomial model for species data. *Journal of Statistical Computation and Simulation*, **75**(5):347–362, 2005.

Zhu M. and Lu A. Y. The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education*, **12**(2), 2004.

Appendix A. Derivation of likelihood ratio under the model

Below is the general Bayes theorem set-up for the likelihood ratio.

$$\frac{P(H_s|r_1,r_2)}{P(H_d|r_1,r_2)} = \underbrace{\frac{P(r_1,r_2|H_s)}{P(r_1,r_2|H_d)}}_{\text{likelihoodratio}} \frac{P(H_s)}{P(H_d)}$$

Using the assumptions described in Section 2.3, we would like to derive the formula for the likelihood ratio, *LR*.

$$\begin{aligned} LR &= \frac{\int P(r_1,r_2|H_s,\theta_1)P(\theta_1|H_s)d\theta_1}{\int\int P(r_1,r_2|H_d,\theta_1,\theta_2)P(\theta_1,\theta_2|H_d)d\theta_1 d\theta_2} \\ &= \frac{\int P(r_1,r_2|H_s,\theta_1)P(\theta_1)d\theta_1}{\int\int P(r_1|H_d,\theta_1)P(r_2|H_d,\theta_2)P(\theta_1)P(\theta_2)d\theta_1 d\theta_2} \\ &= \frac{\int P(r_1,r_2|H_s,\theta_1)P(\theta_1)d\theta_1}{\int P(r_1|H_d,\theta_1)P(\theta_1)d\theta_1 \int P(r_2|H_d,\theta_2)P(\theta_2)d\theta_2} \end{aligned}$$

Let $B(.)$ denote the multivariate beta function, which is defined as

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \qquad (11)$$

where $\Gamma(\alpha_k) = \int_0^\infty x^{\alpha_k - 1} e^{-x} dx$ is the gamma function. Focusing on the first term in the denominator first, we have

$$\begin{aligned} \int P(r_1|H_d,\theta_1)P(\theta_1)d\theta_1 &= \int \left(\binom{N_1}{r_{11}, r_{12}, \ldots, r_{1K}} \prod_{k=1}^{K} \theta_{1k}^{r_{1k}} \right) \left(\frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{1k}^{\alpha_k - 1} \right) d\theta_1 \\ &= \binom{N_1}{r_{11}, r_{12}, \ldots, r_{1K}} \frac{1}{B(\alpha)} \int \prod_{k=1}^{K} \theta_{1k}^{r_{1k}+\alpha_k-1} d\theta_1 \\ &= \binom{N_1}{r_{11}, r_{12}, \ldots, r_{1K}} \frac{B(\alpha + r_1)}{B(\alpha)} \underbrace{\int \frac{1}{B(\alpha + r_1)} \prod_{k=1}^{K} \theta_{1k}^{r_{1k}+\alpha_k-1} d\theta_1}_{\text{integral of Dirichlet}(\alpha + r_1)} \\ &= \binom{N_1}{r_{11}, r_{12}, \ldots, r_{1K}} \frac{B(\alpha + r_1)}{B(\alpha)} \end{aligned}$$

Similarly, for the other terms we have

$$\int P(r_2|H_d,\theta_2)P(\theta_2)d\theta_2 = \binom{N_2}{r_{21},r_{22},\dots,r_{2K}}\frac{B(\alpha+r_2)}{B(\alpha)}$$
$$\int P(r_1,r_2|H_s,\theta_1)P(\theta_1)d\theta_1 = \binom{N_1}{r_{11},r_{12},\dots,r_{1K}}\binom{N_2}{r_{21},r_{22},\dots,r_{2K}}\frac{B(\alpha+r_1+r_2)}{B(\alpha)}$$

Then this gives for the likelihood ratio

$$LR = \frac{B(\alpha+r_1+r_2)B(\alpha)}{B(\alpha+r_1)B(\alpha+r_2)}.$$

Appendix B. Derivation for alternative formula under equal priors assumption

In this section, we will derive the alternative formula in Equation (6).

$$LR = \frac{B(\alpha+r_1+r_2)B(\alpha)}{B(\alpha+r_1)B(\alpha+r_2)} \tag{12}$$
$$= \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{1k}+r_{2k})}{\Gamma(\sum_{k=1}^{K}(\alpha_k+r_{1k}+r_{2k}))}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)} \times \frac{\Gamma(\sum_{k=1}^{K}(\alpha_k+r_{1k}))}{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{1k})}\frac{\Gamma(\sum_{k=1}^{K}(\alpha_k+r_{2k}))}{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{2k})} \tag{13}$$
$$= \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{1k}+r_{2k})}{\Gamma(c+N_1+N_2)}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(c)} \times \frac{\Gamma(c+N_1)}{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{1k})}\frac{\Gamma(c+N_2)}{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{2k})}$$
$$= \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{1k}+r_{2k})}{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{1k})}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k+r_{2k})} \times \frac{\Gamma(c+N_2)}{\Gamma(c)}\frac{\Gamma(c+N_1)}{\Gamma(c+N_1+N_2)} \tag{14}$$

$$
\begin{aligned}
&= \left(\prod_{k=1, r_{2k}\geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \frac{\alpha_k + r_{1k} + s}{\alpha_k + s} \right) \left(\prod_{s=0}^{N_2-1} \frac{c+s}{c+N_1+s} \right) \\
&= \left(\prod_{k=1, r_{2k}\geq 1}^{K} \prod_{s=0}^{r_{2k}-1} \left(1 + \frac{r_{1k}}{\alpha_k + s} \right) \right) \left(\prod_{s=0}^{N_2-1} \left(1 - \frac{N_1}{c+N_1+s} \right) \right)
\end{aligned}
\tag{15}
$$

where $c = \sum_{k=1}^{K} \alpha_k$. The penultimate line (15) follows from the following results:

- $\Gamma(\alpha_k + r_{1k} + r_{2k}) = \Gamma(\alpha_k + r_{1k}) \prod_{s=0}^{r_{2k}-1} (\alpha_k + r_{1k} + s)$
- $\Gamma(\alpha_k + r_{2k}) = \Gamma(\alpha_k) \prod_{s=0}^{r_{2k}-1} (\alpha_k + s)$
- $\Gamma(c + N_2) = \Gamma(c) \prod_{s=0}^{N_2-1} (c + s)$
- $\Gamma(c + N_1 + N_2) = \Gamma(c) \prod_{s=0}^{N_2-1} (c + N_1 + s)$

which all arise from the fact that $\Gamma(x+1)x\Gamma(x)$ for $x \in \mathbb{R}$, which is a general property of gamma functions.

Note that the *LR* in Equation (12) is symmetric in r_1 and r_2. As pointed out in Section 4.3.1, this implies the existence of another equivalent formula. Starting with Equation (13), this formula can be derived as follows:

$$
LR = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + r_{1k} + r_{2k})}{\prod_{k=1}^{K} \Gamma(\alpha_k + r_{2k})} \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k + r_{1k})} \times \frac{\Gamma(c+N_1)}{\Gamma(c)} \frac{\Gamma(c+N_2)}{\Gamma(c+N_1+N_2)}
\tag{16}
$$

$$
\begin{aligned}
&= \left(\prod_{k=1, r_{1k}\geq 1}^{K} \prod_{s=0}^{r_{1k}-1} \frac{\alpha_k + r_{2k} + s}{\alpha_k + s} \right) \left(\prod_{s=0}^{N_1-1} \frac{c+s}{c+N_2+s} \right) \\
&= \left(\prod_{k=1, r_{1k}\geq 1}^{K} \prod_{s=0}^{r_{1k}-1} \left(1 + \frac{r_{2k}}{\alpha_k + s} \right) \right) \left(\prod_{s=0}^{N_1-1} \left(1 - \frac{N_2}{c+N_2+s} \right) \right).
\end{aligned}
\tag{17}
$$

Equations (14) and (16) are two different groupings of the factors that result from plugging in for the definition of each of the multivariate beta functions. Again applying the properties of gamma functions gives Equation (17) and the final result.