Technical Report
1247

Bayes Factors for Detecting Social Group Changes

M.B. Hurley
J.J. Liu
D.C. Shah

17 December 2019

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

This material is based upon work supported by the Department of Homeland Security under Air Force Contract No. FA8702-15-D-0001.

This report is the result of studies performed at Lincoln Laboratory, a federally funded research and development center operated by Massachusetts Institute of Technology. This material is based upon work supported by the Department of Homeland Security under Air Force Contract No. FA8702-15-D-0001. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

Massachusetts Institute of Technology
Lincoln Laboratory

Bayes Factors for Detecting Social Group Changes

M.B. Hurley

J.J. Liu

Group 104

D.C. Shah

Group 52

Technical Report 1247

17 December 2019

Lexington Massachusetts

This page intentionally left blank.

ABSTRACT

The AI Software Architectures & Algorithms Group at MIT Lincoln Laboratory has used Bayes factors (or Bayes t-tests) to measure the similarity between pairs of datasets that can be modelled as draws from Poisson, binomial, or multinomial distributions. More recently, similar Bayes factors or t-tests have been obtained to determine whether a group has merged with or split from another group. It is again assumed that the available data samples can be modeled with Poisson, binomial, or multinomial distributions. A motivation for this report is that it has often been difficult to find desired Bayes factors in the academic literature. This report is intended to provide a repository of the analytical derivations that lead to the Bayes factors for similarity, merger, or splits for Poisson, binomial, and multinomial distributions. Simulation results are presented to clarify the strengths and weaknesses of the derived Bayes factors. Analysis results from the Reddit social networking site are used to demonstrate the utility of the Poisson similarity and merger Bayes factors for real-world applications.

ACKNOWLEDGMENTS

The authors would like to thank our sponsors for supporting our research interests and providing the needed resources so that this work could be completed. We especially would like to thank them for their patience while we worked through many difficult derivations and challenges of writing this technical report. The authors would like to thank the group leaders for their guidance and encouragement to explore new methods for evaluating the similarity between data samples and the more complicated process of detecting groups that have merged or split using Bayesian methods.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Page

LIST OF FIGURES

LIST OF FIGURES
(Continued)

LIST OF FIGURES
(Continued)

LIST OF FIGURES
(Continued)

Figure No. **Page**

LIST OF FIGURES

(Continued)

LIST OF FIGURES

(Continued)

Figure No. **Page**

LIST OF TABLES

This page intentionally left blank.

1. INTRODUCTION

This technical report is part of a larger study to detect changes in online social group dynamics. The primary question associated with the content of this report is "If a group disbands, what is the likelihood that group members have reconstituted as a new group or have merged with another group?" This report describes the derivation and application of likelihood ratios that can be used to determine that a group has disbanded and merged with a second group by evaluating the activity levels of the members in the groups before and after a potential merger. Other studies in the larger project examined the communication content of the members of the groups to detect changes. The Reddit social networking site was used as the principal real-world data source for this analysis.

The intent of this report is to document the derivation of the likelihood ratios (Bayes factors or Bayesian t-tests) for detecting changes in online social groups based upon communication rates (posting and commenting). This report derives Bayes factors or Bayesian t-tests for three common probability distribution functions: the Poisson, binomial, and multinomial functions. The Poisson function is the class of function most applicable to the interests of the overall study: the detection of merging or splitting of groups based upon the activity levels of the members of the group. Equations based upon the binomial and multinomial distribution functions were derived in case future applications are identified for these functions. The equations are generic in that Bayes factors can be determined for any process that can be modeled with these probabilistic functions and where the process may undergo mergers or splits.

In addition to the Bayes factors for mergers and splits, Bayes factors have been used to mesure the similarity between data samples. Derivations are provided for the Poisson, binomial and multinomial probability density functions. The Bayes factor equations for the similarity measures can be found in the literature with great difficulty, but the derivations are extremely challenging to find. This document provides details of the derivations for these Bayes factors to serve as a guide for researchers that are new to Bayes factors. It should be noted that other equations that may be found in the literature can differ slightly from the derivations within this report. These differences are due to different assumptions adopted for likelihood functions and prior probability distributions. Researchers should evaluate the derivations that they require for their specific applications. The derivations in this report should simplify the process of deriving new Bayes factors for their specific assumptions.

The document is structured to provide the equations for different Bayes factors, followed by a discussion of the implications of the factors. The appendix contains detailed derivations for each factor. The structure of the appendices is to hierarchically describe the derivation of different component integrals is a way intended to clarify the overall derivations as much as possible. In many cases, the integrals are convoluted and care is required to follow the mathematics. The majority of the derivations are been placed in appendices to spare the casual reader from the grueling mathematics when their only interest is in the results and not the process.

This page intentionally left blank.

2. STATE OF THE ART

The history of Bayesian t-tests or Bayes factors is relatively sparse. The traditional frequentist statistical approach to t-tests as well as other named tests dominates the literature. A challenge with the literature is that although the test functions are often described, the detailed derivations are extremely difficult to find. There seems to be an assumption within the research community that any researcher of reasonable skill should be able to easily reproduce the derivation. This report provides the details so that other researchers will not have to repeatedly waste their time deriving Bayes factors and, if unfamiliar with integration, should be able to extend the provided derivations to new applications with some reasonable amount of effort.

Przyborowski and Wilenski are credited with deriving the first frequentist test for whether two samples are drawn from the same Poisson distribution [1]. Their approach followed a traditional frequentist approach to hypothesis testing, whereas in this paper, a Bayesian approach has been adopted. Their approach involved estimating the probability that the two measurements were obtained from a Poisson function for the case where an unknown and equivalent mean falls below a threshold value. In the Bayesian approach, the second hypothesis is that the two samples are drawn from two different Poisson distributions and that a ratio of the two probabilities can be used to estimate the likelihood that the samples come from the same distribution.

Harold Jeffreys is generally credited with first defining a Bayesian test statistic for one sample that was eventually given the name “Bayes factor” [2]. I. J. Good looked at a Bayesian significance test for multinomial distributions. The goal was to identify if a multinomial sample was better described by a non-uniform multinomial distribution or by the null hypothesis of a uniform distribution across the elements of the multinomial distribution [3]. The task of interest in this technical report is to determine whether two samples are from the same or different probability distributions where the distributions are members of the same probability family. It is assumed that the distribution parameters are unknown and are not of interest, subject to the constraints that prior probability distributions place on those parameters.

Extensions from a one-sample to a two-sample Bayesian t-test are credited to Gönen et al. [4] and Rouder et al. [5]. The Bayes factor generally quantifies the relative likelihood of two hypotheses: H_0, which is called the null hypothesis and, for two-sample comparisons, is associated with the likelihood that both samples are drawn from the same distribution, and H_1, which is associated with the likelihood that the two samples are drawn from different distributions. Bayes factors and other test statistics are used to determine whether a treatment applied to a given population results in a change in the characteristics of the treated population. Examples include drug trials, sociology experiments, and disease studies. The ratio is selected to be $p\left(H_1\right)/p\left(H_0\right)$ because the interest is in the change. Gronau, et al. call for more adoption of Bayes factors or Bayesian two-sample t-tests for data analysis [6]. In this report, the interest is in determining whether two samples are from the same distribution or two different distributions, so the inverse ratio $p\left(H_0\right)/p\left(H_1\right)$ is adopted. The multiplicative inverse of Bayes factors may be required to compare with the equations in other papers.

A number of different papers on either Bayes factors or two-sample Bayesian t-tests can be found in the literature. Gronau et al. provide a good overview of the two-sample Bayesian t-test papers [6]. Gönen et al. [4] primarily focus on t-tests for normally distributed samples. Sides et al. provide the Bayes factor for two samples drawn from Poisson distributions [7]. Zhao and Tang provide the Bayes factor equations for two samples drawn from binomial distributions [8]. This report does not provide any details on the Bayes factors for normally distributed data, but does provide details on the extension from binomial to multinomial distributions. Other papers related to Bayes factors and two-sample Bayesian t-tests include Jeffreys [9], Etz and Wagenmakers [10], and Rouder et al. [5], among others. Zhao et al. focus on determining required sample sizes to obtain a selected level of confidence with binomial experiments [8]. Sides [11] and Sides et al. [7] similarly focus on determining required sample sizes for both Poisson and binomial experiments.

A point to note is that some of the prior literature describes equations that rely on derived statistics, like various t statistics that are obtained from data samples, along with the number of counts N, and the number of degrees of freedom ν, to estimate Bayes factors or t-test values. The methods that are derived here directly use the counts to estimate Bayes factors instead of secondary t statistics.

A second point to note is that different choices for prior probability distributions for the statistical parameters in the likelihood density functions will result in different Bayes factors. This analysis selects conjugate priors for the likelihood functions. There are discussions in the literature as to which prior probabilities are appropriate for different situations and possible problems with various choices for prior probability. These discussions are not considered in this report. It is possible that in the future, these arguments may be considered as part of a follow-on evaluation of the validity of different prior probability distribution functions for estimating Bayes factors.

Section 5 compares some of the derived Bayes factors with other similarity and distance measures. There is a vast literature dedicated to different similarity and distance measures that cannot be covered here. Sung-Hyuk Cha provides a comprehensive survey of distance/similarity measures between probability density functions [12]. Cha defines a hierarchy of distance/similarity functions grouped into nine families. The first eight families are the L_p Minkowski family, the L_1 family, the intersection family, the inner product family, the fidelity or squared-chord family, the squared L_2 or χ^2 family, the Shannon entropy family, and the combinations family. The combinations family are measures that are combinations from the first seven families. Cha proposes a ninth family, the vicissitude family, with a list of six additional measures that do not previously appear in the literature and are extensions of measures from the other families based upon syntactic relationships between measures. Overall, Cha lists 62 different measures distributed across the nine families. The analysis of the relationships between the different measures is somewhat limited. The probabilistic log likelihood scores proposed in this report do not fit neatly into any of the families that Cha defines. An additional tenth family may be required for these probabilistic measures.

The similarity/distance measures adopted for the analysis in this report include the dot product (inner product family), Euclidean distance (Minkowski family), Manhattan or city block distance (Minkowski family), Canberra distance (L_1 family), Jaccard similarity (inner product family), and Bray-Curtis dissimilarity which is directly related to the Sorensen similarity measure

(L_1 family), and cosine similarity (inner product family). A comprehensive comparison between the probabilistic similarity scores and the 48 measures would be interesting, but beyond the resources that are available for this project.

This page intentionally left blank.

3. BAYES FACTORS FOR GROUP SIMILARITY

The equations for Bayes factors or two-sample Bayesian t-tests for the similarity of pairs of samples has been determined by others for the Poisson, binomial, and multinomial distributions. These results are repeated here because they can be difficult to find in the statistics literature. This is usually because they are often of secondary interest to the main research theme of these papers. If found, the results of the derivations are usually provided without a detailed description of how the equations were obtained.

3.1 THE BAYES FACTOR FOR A POISSON SIMILARITY TEST

An overview of the derivation of the Bayes factor for two sample counts, n_A and n_B, that are either drawn from the same Poisson distributions or two different Poisson distributions is provided here. The Poisson distribution function is

$$p\left(n|\lambda\right)=\frac{\lambda^{n}e^{-\lambda}}{\Gamma\left(n+1\right)}, \tag{1}$$

where n is the sample count and λ is the Poisson function parameter that is both the mean and standard deviation. The conjugate prior distribution for the Poisson distribution is the gamma distribution function,

$$p\left(\lambda|\alpha,\beta\right)=\frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma\left(\alpha\right)}, \tag{2}$$

which is different from the gamma function, notated as $\Gamma\left(x\right)$. The parameters for the conjugate prior distribution are α, which is the shape parameter, and β, which is the rate parameter.

The posterior probability for two samples of counts, n_A and n_B, being generated by the same Poisson generator is

$$p\left(H_0|n_A,n_B\right)=\frac{1}{Z}\int_0^{\infty}p\left(n_A|\lambda\right)p\left(n_B|\lambda\right)p\left(\lambda|\alpha,\beta\right)d\lambda. \tag{3}$$

It is assumed here that the time periods for collecting the counts are the same. The posterior probability that the two samples are generated by different Poisson generators is

$$p\left(H_1|n_A,n_B\right)=\frac{1}{Z}\int_0^{\infty}p\left(n_A|\lambda_A\right)p\left(\lambda_A|\alpha,\beta\right)d\lambda_A\int_0^{\infty}p\left(n_B|\lambda_B\right)p\left(\lambda_B|\alpha,\beta\right)d\lambda_B. \tag{4}$$

Appendix A.1 provides the detailed derivation of the Poisson likelihood ratio function, which leads to

$$R=\frac{p\left(H_0|n_A,n_B\right)}{p\left(H_1|n_A,n_B\right)}=\frac{\rho_S\left(1+\beta\right)^{\left(n_A+n_B+2\alpha\right)}\Gamma\left(n_A+n_B+\alpha\right)\Gamma\left(\alpha\right)}{\rho_{\bar{S}}\beta^{\alpha}\left(2+\beta\right)^{\left(n_A+n_B+\alpha\right)}\Gamma\left(n_A+\alpha\right)\Gamma\left(n_B+\alpha\right)}. \tag{5}$$

The numerator is the hypothesis that the two samples are drawn from the same Poisson distribution, and the denominator is the hypothesis that the two distributions are drawn from two different Poisson distributions. The prior probabilities for similar or different generators are ρ_S and $\rho_{\bar{S}}$. For this derivation, a common conjugate prior distribution has been assumed for all Poisson generators. It is possible that the hypothesis that different generators produce the two samples could also assume different prior knowledge about the likely values for the Poisson rates, λ. This extension has not been derived.

If the two samples are collected for different durations, then corrections are required to the ratio to account for the relative collection times. The ratio is now given as

$$R = \frac{\rho_S \left(r_A + \beta\right)^{(n_A+\alpha)} \left(r_B + \beta\right)^{(n_B+\alpha)} \Gamma\left(n_A + n_B + \alpha\right) \Gamma\left(\alpha\right)}{\rho_{\bar{S}} \beta^{\alpha} \left(r_A + r_B + \beta\right)^{(n_A+n_B+\alpha)} \Gamma\left(n_A + \alpha\right) \Gamma\left(n_B + \alpha\right)}. \tag{6}$$

This function assumes that the conjugate prior rate variable, λ, is valid for a standard collection period. The collection period for sample n_A is scaled by r_A with respect to the standard time period. A similar scale factor, r_B, is adopted for sample n_B. Appendix A.1.1 provides the detailed derivation for the Poisson likelihood ratio function, scaled for different sample durations. The rate variable, λ, is a variable of integration and does not appear in the likelihood ratio test, but the relative scales between the sample periods and the default prior period still appear in the Bayes factor.

3.2 BAYES FACTORS FOR MULTINOMIAL AND BINOMIAL SIMILARITY

The Bayes factor for two vectors of counts for a canonical set, $\boldsymbol{d}_A$ and $\boldsymbol{d}_B$, can be obtained in a similar manner to the Bayes factors for Poisson distributions. The binomial distribution can be treated as a multinomial distribution with two dimensions. The multinomial distribution is

$$p\left(\boldsymbol{d}|\boldsymbol{q}\right) = \frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_i\right) + 1\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_i + 1\right)} \prod_{i=1}^{k} \boldsymbol{q}_i^{\boldsymbol{d}_i}, \tag{7}$$

where $\boldsymbol{d}_i$ are the sample counts and $\boldsymbol{q}_i$ the probabilities for the i categories.

The conjugate prior probability distribution for the multinomial distribution is the Dirichlet distribution,

$$p\left(\boldsymbol{q}|\boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_i\right)} \prod_{i=1}^{k} \boldsymbol{q}_i^{\boldsymbol{\alpha}_i - 1}, \tag{8}$$

where $\boldsymbol{\alpha}_i$ are the prior parameters for the i categories.

Appendix A.2 provides the detailed derivation for both the multinomial and binomial similarity likelihood ratio function. The derivation results in

$$R = \frac{p\left(H_0|\boldsymbol{d}_A, \boldsymbol{d}_B\right)}{p\left(H_1|\boldsymbol{d}_A, \boldsymbol{d}_B\right)} = \frac{\rho_S \mathrm{B}'\left(\boldsymbol{d}_A + \boldsymbol{d}_B + \boldsymbol{\alpha}\right) \mathrm{B}'\left(\boldsymbol{\alpha}\right)}{\rho_{\bar{S}} \mathrm{B}'\left(\boldsymbol{d}_A + \boldsymbol{\alpha}\right) \mathrm{B}'\left(\boldsymbol{d}_B + \boldsymbol{\alpha}\right)}, \tag{9}$$

where a generalized beta function is defined as

$$\mathrm{B}'\left(\boldsymbol{\mu}\right) = \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\mu}_i\right)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\mu}_i\right)}. \tag{10}$$

Unlike the ratio test for the Poisson distributions, the duration of sample collection does not have to be the same because the ratio test is not sensitive to differences in total counts, but only to how the counts are distributed across the different elements of the categorical set. It is still the case that increases in the total counts improve the sensitivity of the likelihood ratio test.

The binomial Bayes factor can be directly obtained from Equation 9 by using two-dimensional vectors for the counts $\boldsymbol{d}_A$ and $\boldsymbol{d}_B$.

This page intentionally left blank.

4. BAYES FACTORS FOR GROUP MERGERS AND SPLITS

Although Bayes factors for similarity tests can be found in the statistics literature (with some difficulty), no similar functions have been found for determining if two groups have merged or an initial group has split into two independent subgroups. It is possible that this document is the first description of functions for detecting group mergers or splits through Bayes factors. The motivation for this work is that social media sites, like Reddit, contain groups that may occasionally be removed or disperse for a variety of reasons. Statistical measures would be useful to determine whether a group has dispersed or has simply combined with another group to continue the same activities. If a removed group simply forms a new group, the previously described Bayes factor similarity functions can be used to detect this case.

Because the merge-split Bayes factors are independent of time, the same factors apply to both mergers and splits. Three samples are required for the merge-split Bayes factors presented in this paper: two samples from the independent groups and one sample that is possibly from a combined group or one of the two independent groups. As was the case with the Bayes factors used to measure two-sample similarity, the merge-split Bayes factors samples are assumed to consist of measures of activity levels of the members in the groups.

This paper describes two models that have been used to derive Bayes factors. Both models assume that there are two time periods. For mergers, two independent group are assumed to exist at the first time period. One of the groups no longer exists at the end of this time period and the identity of the terminated group is known. The members of the terminated group are assumed to 1) have merged en masse with the other group or 2) dispersed across other groups. The derived Bayes factors only evaluate whether the terminated group has or has not combined with the potential destination group.

The first and simpler model assumes that there are generators that produce activity statistics for the two groups at the first time period. This model assumed that the generator parameters of the two initial groups do not change between the two time periods. The model for the merger hypothesis assumes that the generator for the group activity statistics at the second time period is a combination of the generators for the two individual groups prior to the merger. The hypothesis of no merger assumes that the generator for the potential destination group at the second time period is unchanged from the first time period. It is assumed that the terminated group and the potential destination group are distinguishable. In this second hypothesis, there is no generator for the terminated group during the second time period.

The second and more complex model assumes that there is a probability that the generator parameters for the users in the potential destination group may have changed between the two time periods. This second model has four hypotheses, whereas the first model only has two: hypothesis 1) the terminated group has merged with the destination group, which has not changed behavior; hypothesis 2) the terminated group has merged with the destination group, which has changed behavior; hypothesis 3) the terminated group has disbanded and the destination group has not changed behavior; and hypothesis 4) the terminated group has disappeared and the destination group has changed behavior. All hypotheses assume that the generator parameters for the members of the terminated group have not changed; otherwise mergers and splits would not be detectable. This Bayes factor requires prior probabilities for the four hypotheses so that a merger test function can be derived. These prior probabilities are set based upon the likelihood that the potential destination group may have changed behavior. The Bayes factor for this model uses a numerator constructed from the sum of hypotheses 1) and 2) and a denominator constructed from the sum of hypotheses 3) and 4). The Bayes factor for the first model can be produced from the second model by only using hypotheses 1) and 3) for the merger test. This second model was created because the analysis of Reddit data showed that subreddits change activity levels over time and that explicitly accounting for possible activity changes could lead to better detection of mergers. Results will show that although the more complex model appears to operate correctly, the results as applied to Reddit data were not fully satisfactory.

One may note that an assumption associated with the derivations is that all time periods are of identical duration for both models. No work has been performed to handle the more general case where the before and after time periods are different, or even the extreme case where the sampling periods are all different. There may be some applications of the Poisson process models where this could be relevant. The binomial and multinomial process models are insensitive to different time periods, as stated previously. This extension for the Poisson process model could be produced if a need ever arises.

4.1 THE BAYES FACTOR FOR POISSON MERGERS

The first merger scenario examined is one where the generators for the number of events are Poisson distributions. For the scenarios of interest, the assumption is that the counts of member activities follow a Poisson distribution. This assumption makes it possible to obtain analytical solutions for the likelihood ratio functions.

With regard to community activity levels, the derivations presented here are for the case where the activity level samples are from a single individual. It is further assumed that the Bayes factor can be calculated for each individual and the individual factors combined to get an overall Bayes factor for group merger. This requires an assumption that the activity of each individual is statistically independent of all the other users, even though this assumption would not be expected to hold for social activities, especially given that users' social activities trigger additional social activity from other users, which then spur other users to participate. Because this is new research, the independence assumption will be adopted, with the anticipation that the overall Bayes factors for merger can still give a reasonable indication of the likelihood of a merger or split.

It is common practice to estimate the log of Bayes factors. This provides better numerical precision in comparison to straight probability ratios. Many mathematics packages, like MATLAB, contain specific log probability functions, like gammaln() in MATLAB. The product of Bayes factors then becomes a sum of log Bayes factors. The derivations in this report will neglect the use of logarithms, although it is anticipated that actual implementations will take advantage of the numerical stability that this provides.

4.1.1 Poisson Mergers with Consistent Generators

The Bayes factor for the model where the potential destination group remains unchanged before and after a potential merger is of the form

$$R_{P2} = \frac{p\left(d_B, d_S, d_R | G_B, G_S, G_B \text{ \& } G_S\right) p\left(G_B, G_S, G_B \text{ \& } G_S\right)}{p\left(d_B, d_S, d_R | G_B, G_S, G_S\right) p\left(G_B, G_S, G_S\right)}, \tag{11}$$

where d_B, d_S, and d_R, are the counts for a user in the terminated, potential destination, and the resultant groups. The variables G_B and G_S represent the Poisson generators for the terminated group and the potential destination group. At this stage, it is assumed that the destination group members do not change behaviors after the merger. The hypothesis for the numerator is for merger and the hypothesis for the denominator is against merger. The numerator indicates that the resultant group is a combination of the generators of the original terminated group and the potential destination group. The combination of Poisson distributions is simple in that the new rate λ is a sum of the constituent rates.

The notation can be compressed with the following replacements:

$$\begin{aligned}
\rho_{MS} =& p\left(G_B, G_S, G_B \text{ \& } G_S\right), \\
\rho_{\bar{M}S} =& p\left(G_B, G_S, G_S\right), \\
p_{MS} =& p\left(d_B, d_S, d_R | G_B, G_S, G_B \text{ \& } G_S\right), \\
p_{\bar{M}S} =& p\left(d_B, d_S, d_R | G_B, G_S, G_S\right).
\end{aligned} \tag{12}$$

Then

$$R_{P2} = \frac{\rho_{MS}\ p_{MS}}{\rho_{\bar{M}S}\ p_{\bar{M}S}}. \tag{13}$$

Appendix B.1 contains the detailed derivation that eventually produced the likelihood ratio equation for Poisson distributions with consistent generators:

$$\begin{aligned}
R_{P2} =& \frac{\rho_{MS}}{\rho_{\bar{M}S}} \frac{\Gamma\left(d_S + \alpha_S\right) \Gamma\left(d_B + d_S + d_R + \alpha_B + \alpha_S\right)}{\Gamma\left(d_B + d_S + \alpha_B + \alpha_S\right) \Gamma\left(d_S + d_R + \alpha_S\right)} \left(\frac{1 + \beta_B}{2 + \beta_B}\right)^{d_B + \alpha_B} \times \\
& {}_2F_1\left(d_B + \alpha_B, -d_R; d_B + d_S + \alpha_B + \alpha_S; \left(\frac{\beta_B - \beta_S}{2 + \beta_B}\right)\right).
\end{aligned} \tag{14}$$

The parameters α_B and α_S are the prior distribution shape parameters for the terminated (banned) group and the potential destination group both before and after the potential merger. The parameters β_B and β_S are the prior distribution rate parameters for the same groups. The function ${}_2F_1$ is the Gaussian (or ordinary) hypergeometric function.

If the prior rate parameters β_B and β_S are equal, the hypergeometric function ${}_2F_1$ drops out of the equation because ${}_2F_1\left(a, b; c; 0\right) = 1$,

$$R_{P2,\beta} = \frac{\rho_{MS}}{\rho_{\bar{M}S}} \frac{\Gamma\left(d_S + \alpha_S\right) \Gamma\left(d_B + d_S + d_R + \alpha_B + \alpha_S\right)}{\Gamma\left(d_B + d_S + \alpha_B + \alpha_S\right) \Gamma\left(d_S + d_R + \alpha_S\right)} \left(\frac{1 + \beta}{2 + \beta}\right)^{d_B + \alpha_B}. \tag{15}$$

4.1.2 Poisson Mergers with the Potential for Changes in Generators

While analyzing the behavior of Equation 14, there was a desire to be able to handle the more general case where the users in the destination group may have changed their behavior between time periods, regardless of the influence of new members from a terminated group. Although additional prior information is required in terms of the likelihood that groups will change behavior over time, the Bayes factor for this second model could be useful in certain environments. For the case where the potential destination group may change behavior, two additional hypotheses are required. The two additional prior probabilities are $\rho_{M\bar{S}}$ and $\rho_{\bar{M}\bar{S}}$; the prior probability that there has been a merger and the potential destination group has changed behavior and the prior probability of no merger and a change in the potential destination group's behavior.

Additional notations can be defined for the cases where the generator changes for the potential destination group after the potential merger point,

$$\begin{aligned} p_{M\bar{S}} =& p\left(d_B, d_S, d_R | G_B, G_S, G_B \;\&\; G'_S\right), \\ p_{\bar{M}\bar{S}} =& p\left(d_B, d_S, d_R | G_B, G_S, G'_S\right), \end{aligned} \tag{16}$$

where the prime superscript indicates the changed generator. The generator for the banned user is assumed to remain the same. The ratio for merged versus not merged with the four different hypotheses is

$$R_{P4} = \frac{\rho_{MS}\; p_{MS} + \rho_{M\bar{S}}\; p_{M\bar{S}}}{\rho_{\bar{M}S}\; p_{\bar{M}S} + \rho_{\bar{M}\bar{S}}\; p_{\bar{M}\bar{S}}}. \tag{17}$$

Because the generator may change between the two time periods, two additional prior parameters are added for the changed generator. The scale parameter for the post-possible merger generator is α_R and the rate parameter is β_R.

The detailed derivations are found in Appendix B.1, and especially Appendix B.1.2. The likelihood ratio for the model where the potential destination group may have changed behavior after a potential merger is much more complicated. It is easier to list the equations for the contributions to R_{P4} instead of the full equation. The first term for merger with a consistent destination generator is given by

$$\begin{aligned} p_{MS} =& \left(2+\beta_B\right)^{-(d_B+\alpha_B)} \left(2+\beta_S\right)^{-(d_S+d_R+\alpha_S)} \times \\ & \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)} \times \\ & {}_2F_1\left(d_B+\alpha_B, -d_R; d_B+d_S+\alpha_B+\alpha_S; \left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right). \end{aligned} \tag{18}$$

The second term for merger with a changed destination generator is

$$\begin{aligned} p_{M\bar{S}} =& \left(2+\beta_B\right)^{-(d_B+\alpha_B)} \left(1+\beta_S\right)^{-(d_S+\alpha_S)} \beta_R^{\alpha_R} \left(1+\beta_R\right)^{-(d_R+\alpha_R)} \times \\ & \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+d_R+\alpha_B+\alpha_R\right)}{\Gamma\left(d_B+\alpha_B+\alpha_R\right)} \times \\ & {}_2F_1\left(d_B+\alpha_B, -d_R; d_B+\alpha_B+\alpha_R; \left(\frac{\beta_B-\beta_R+1}{2+\beta_B}\right)\right). \end{aligned} \tag{19}$$

The third term for no merger with a consistent destination generator is

$$p_{\bar{M}S} = (1+\beta_B)^{-(d_B+\alpha_B)} (2+\beta_S)^{-(d_S+d_R+\alpha_S)} \Gamma(d_S+d_R+\alpha_S) . \tag{20}$$

The fourth term for no merger with a changed destination generator is

$$\begin{aligned} p_{\bar{M}\bar{S}} = &(1+\beta_B)^{-(d_B+\alpha_B)} (1+\beta_S)^{-(d_S+\alpha_S)} \beta_R^{\alpha_R} (1+\beta_R)^{-(d_R+\alpha_R)} \times \\ &\frac{\Gamma(d_S+\alpha_S)\Gamma(d_R+\alpha_R)}{\Gamma(\alpha_R)} . \end{aligned} \tag{21}$$

Terms that are common to all four equations have been removed from the equivalent equations in the appendix. It may be noted that although the consistent model had the ${}_2F_1$ term in p_{MS} equal to 1 when $\beta_B = \beta_S$, the same simplification does not occur for the probability of merger with a different generator, $p_{M\bar{S}}$.

4.2 BAYES FACTOR FOR BINOMIAL MERGERS

Although the Poisson distributions were of the most interest for the research team's applications, the Bayes factors for binomial and multinomial distributions of categorical count data were also derived. Examples of where this might be useful would be when user activity volumes can be placed into different categories, such as different discussion topics. An example from Reddit might a subreddit that connects buyers and sellers. The items under sale can be placed into different categories and the volume of sales measured across the categories on a day-by-day or week-by-week basis. For group mergers and splits, this could indicate that the same types of sales are being conducted before and after a merger, irrespective of overall volume of sales.

Although the derivation for the similarity between sample counts for binomial and multinomial distributions was previously combined into one derivation, the derivations in this section are significantly more challenging and complex. The derivations are kept separate in an attempt to improve clarity of exposition.

The binomial and multinomial functions estimate the probability of obtaining vectors of counts over a set of elements, given a probability vector that describes their distribution over a set of categories. For these equations, the variables $\boldsymbol{d}_B$, $\boldsymbol{d}_S$, and $\boldsymbol{d}_R$ are vectors of counts across a set of elements. The total number of counts are n_B, n_S, and n_R. The subscript B is for the terminated (or banned) group, S is for the potential destination group before the possible merger, and R for the potential destination group after the possible merger.

4.2.1 Binomial Merger Bayes Factor with Consistent Generators

For the binomial functions, the $d_\bullet$ symbols represent the integer counts for first element of the two-element vectors, $\boldsymbol{d}_\bullet$. The n variables contain the sums over the two-element sets. The second terms of the two-element sets appear as $n - d_\bullet$ in the following equations.

Appendix B.3 provides the detailed derivation of the binomial merger Bayes factor for both models. The likelihood ratio function for binomial mergers with a consistent generator for the potential destination group both before and after the potential merger is of the general form of Equation 13. The two likelihood terms are

$$
\begin{aligned}
p'_{MS} =& \sum_{t=0}^{d_R} \frac{1}{\Gamma\left(d_R - t + 1\right)} \times \\
& \sum_{u=0}^{n_R - d_R} \left(-1\right)^u B\left(n_S - d_S + \beta_S, d_S + \alpha_S + t + u\right) \times \\
& \sum_{v=0}^{n_R - d_R - u} \left(-1\right)^v B\left(n_B - d_B + \beta_B, d_B + \alpha_B + d_R - t + v\right) \times \\
& \frac{B\left(d_R - t + v + \alpha_\gamma, t + u + \beta_\gamma\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R - d_R - u - v + 1\right)B\left(\alpha_\gamma, \beta_\gamma\right)}, \\
p'_{\bar{M}S} =& \frac{B\left(d_S + d_R + \alpha_S, n_S - d_S + n_R - d_R + \beta_S\right)B\left(d_B + \alpha_B, n_B - d_B + \beta_B\right)}{\Gamma\left(d_R + 1\right)\Gamma\left(n_R - d_R + 1\right)}.
\end{aligned}
\tag{22}
$$

The two terms are primarily composed of gamma functions and binomial functions. The likelihood of merger p'_{MS} is a triply nested sum, with summation indices t, u, and v. The sums involve various combinations of the counts for the resulting destination group: d_R and n_R. When these numbers are large, there will be many terms for p'_{MS} and numerical precision may become difficult to maintain. The α_γ and β_γ variables are the prior probability mixing parameters for the merged distribution. For the case where these prior parameter values are 1, the prior distribution on the mixing between the combined distributions is equivalent to a uniform distribution.

The Bayes factor is constructed from these terms with an equation in the form of Equation 13.

4.2.2 Binomial Merger Bayes Factor with the Potential Changes in Generators

For the cases where the generator for the potential destination group may change, the Bayes factor is of the general form of Equation 17. The additional likelihood terms for the cases where the destination generator may change are

$$
\begin{aligned}
p'_{M\bar{S}} =& \frac{B\left(d_S + \alpha_S, n_S - d_S + \beta_S\right)}{B\left(\alpha_R, \beta_R\right)} \sum_{t=0}^{d_R} \frac{1}{\Gamma\left(d_R - t + 1\right)} \\
& \sum_{u=0}^{n_R - d_R} \left(-1\right)^u B\left(\beta_R, \alpha_R + t + u\right) \times \\
& \sum_{v=0}^{n_R - d_R - u} \left(-1\right)^v B\left(n_B - d_B + \beta_B, d_B + \alpha_B + d_R - t + v\right) \times \\
& \frac{B\left(d_R - t + v + \alpha_\gamma, t + u + \beta_\gamma\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R - d_R - u - v + 1\right)B\left(\alpha_\gamma, \beta_\gamma\right)}, \\
p'_{\bar{M}\bar{S}} =& \frac{B\left(d_S + \alpha_S, n_S - d_S + \beta_S\right)B\left(d_B + \alpha_B, n_B - d_B + \beta_B\right)}{B\left(\alpha_R, \beta_R\right)\Gamma\left(d_R + 1\right)\Gamma\left(n_R - d_R + 1\right)} \times \\
& B\left(d_R + \alpha_R, n_R - d_R + \beta_R\right).
\end{aligned}
\tag{23}
$$

Again, the merger term contains a triply nested sum involving the counts for the resultant group, d_R and n_R. A common term, $\Gamma(n_R+1)$, appears in all four terms for the binomial likelihood ratio and cancels. It has been removed from these four equations.

4.3 MULTINOMIAL MERGER BAYES FACTOR

The multinomial merger Bayes factor is even more complex than the binomial merger Bayes factor. As was described previously, the counts are now over a set of k categorical variables. Although the binomial Bayes factor contains terms that are triply nested sums over the count of occurrences for one of two categories, the multinomial ratio function will have a set of $k-1$ triply nested sums, for a total of $3k-3$ nested sums. These sums are over the resultant vector of counts, with the counts for the last category mingled across the other nested sums.

As for the Poisson and binomial Bayes factors, two models have been derived: one with two terms and consistent generators, and one with four terms and the possibility that the destination group sample was generated with a different generator after the potential merger. Equations 13 and 17 are the general equations for the two models.

Instead of presenting pairs of terms, as was done for the Poisson and binomial derivations, all four terms will be presented for the multinomial derivation. The appropriate terms can be selected to produce the Bayes factor for the desired model.

Appendix B.4 provides the detailed derivations of the multinomial merger Bayes factor terms. The first term is

$$
\begin{aligned}
p_{MS} = &\sum_{t_{k-1}=0}^{\boldsymbol{d}_{R_{k-1}}} \sum_{u_{k-1}=0}^{\boldsymbol{d}_{R_k}} \sum_{v_{k-1}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}} \Bigg(\sum_{t_{k-2}=0}^{\boldsymbol{d}_{R_{k-2}}} \sum_{u_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}} \sum_{v_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}-u_{k-2}} \Bigg(\times \\
&\cdots \Bigg(\sum_{t_1=0}^{\boldsymbol{d}_{R_1}} \sum_{u_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)} \sum_{v_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)-u_1} \times \\
&\Bigg(\left[(-1)^{\sum_{r=1}^{k-1} u_r+v_r}\right] \frac{B\left(\sum_{r=1}^{k-1} \boldsymbol{d}_{R_r}-t_r+v_r+\alpha_{\gamma_1}, \sum_{r=1}^{k-1} t_r+u_r+\alpha_{\gamma_2}\right)}{B\left(\alpha_{\gamma_1},\alpha_{\gamma_2}\right)} \times \\
&\frac{\Gamma\left(\boldsymbol{d}_{B_k}+\boldsymbol{\alpha}_{B_k}\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{d}_{B_q}+\boldsymbol{\alpha}_{B_q}+\boldsymbol{d}_{R_q}-t_q+v_q\right)}{\Gamma\left(\boldsymbol{d}_{B_k}+\boldsymbol{\alpha}_{B_k}+\left(\sum_{q=1}^{k-1} \boldsymbol{d}_{B_q}+\boldsymbol{\alpha}_{B_q}+\boldsymbol{d}_{R_q}-t_q+v_q\right)\right)} \times \\
&\frac{\Gamma\left(\boldsymbol{d}_{S_k}+\boldsymbol{\alpha}_{S_k}\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{d}_{S_r}+\boldsymbol{\alpha}_{S_r}+t_r+u_r\right)}{\Gamma\left(\boldsymbol{d}_{S_k}+\boldsymbol{\alpha}_{S_k}+\sum_{r=1}^{k-1} \boldsymbol{d}_{S_r}+\boldsymbol{\alpha}_{S_r}+t_r+u_r\right)} \times \\
&\left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_{R_r}+1\right)}{\Gamma\left(\boldsymbol{d}_{R_r}-t_r+1\right)}\right] \frac{\Gamma\left(\boldsymbol{d}_{R_k}+1\right)}{\Gamma\left(\boldsymbol{d}_{R_k}-\left(\sum_{q=1}^{k-1} u_q+v_q\right)+1\right) \prod_{r=1}^{k-1} \Gamma\left(v_r+1\right)} \Bigg)\Bigg) \cdots \Bigg)\Bigg).
\end{aligned}
\tag{24}
$$

The second term is

$$p_{\overline{M}S} = \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Si} + \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Si}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Si} + \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Si}\right)}. \tag{25}$$

The third term is

$$\begin{aligned}
p_{M\overline{S}} = {} & \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{Ri}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_{Ri}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{S_i} + \boldsymbol{\alpha}_{S_i}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{S_j} + \boldsymbol{\alpha}_{S_j}\right)} \\
& \sum_{t_{k-1}=0}^{\boldsymbol{d}_{R_{k-1}}} \sum_{u_{k-1}=0}^{\boldsymbol{d}_{R_k}} \sum_{v_{k-1}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}} \Bigg(\sum_{t_{k-2}=0}^{\boldsymbol{d}_{R_{k-2}}} \sum_{u_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}} \sum_{v_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}-u_{k-2}} \Bigg(\times \\
& \cdots \Bigg(\sum_{t_1=0}^{\boldsymbol{d}_{R_1}} \sum_{u_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r + v_r\right)} \sum_{v_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r + v_r\right)-u_1} \times \\
& \Bigg(\left[(-1)^{\sum_{r=1}^{k-1} u_r + v_r}\right] \frac{B\left(\sum_{r=1}^{k-1} \boldsymbol{d}_{R_r} - m_r + v_r + \alpha_{\gamma_1}, \sum_{r=1}^{k-1} t_r + u_r + \alpha_{\gamma_2}\right)}{B\left(\alpha_{\gamma_1}, \alpha_{\gamma_2}\right)} \times \\
& \frac{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k}\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - m_q + v_q\right)}{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k} + \left(\sum_{q=1}^{k-1} \boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)\right)} \frac{\Gamma\left(\boldsymbol{\alpha}_{R_k}\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{\alpha}_{R_r} + t_r + u_r\right)}{\Gamma\left(\boldsymbol{\alpha}_{R_k} + \sum_{r=1}^{k-1} \boldsymbol{\alpha}_{R_r} + t_r + u_r\right)} \times \\
& \left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_{R_r} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_r} - t_r + 1\right)}\right] \frac{\Gamma\left(\boldsymbol{d}_{R_k} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_k} - \left(\sum_{q=1}^{k-1} u_q + v_q\right) + 1\right) \prod_{r=1}^{k-1} \Gamma\left(v_r + 1\right)} \Bigg)\Bigg) \cdots \Bigg)\Bigg).
\end{aligned} \tag{26}$$

The fourth term is

$$p_{\overline{MS}} = \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{Ri}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_{Ri}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Si} + \boldsymbol{\alpha}_{Si}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Si} + \boldsymbol{\alpha}_{Si}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Ri}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Ri}\right)}. \tag{27}$$

The nested summation can become quite deep as the number of categories k becomes large. Note the "$\cdots$" ellipses in Equations 24 and 26 indicate the general depth of nesting. These functions are likely to prove extremely difficult to numerically calculate, especially because the (-1) terms cause a summation where additional terms partially cancel during evaluation. There may be specific cases where these functions can be calculated with reasonable accuracy. These could be useful for evaluating numerical integration results for those applications that try to avoid analytical integration.

The Bayes factor would be constructed from these terms with an equation in the form of Equation 17.

5. RESULTS FROM SIMULATIONS

5.1 SIMULATIONS TO TEST SIMILARITY METRICS

5.1.1 Simulations of Poisson Similarity Tests

Figure 1 shows the simulation results where two samples are drawn from the same Poisson generator function for a range of mean values, as indicated on the x-axis. A set of $10,000$ pairs of samples was drawn and tested with the Poisson log likelihood ratio test of Equation 6. The probability of a correct decision is $\approx 99\%$ by the point that the mean is 10 counts.

Figure 2 shows simulation results where samples are drawn from different Poisson generators over a range of mean values. The 3D plot shows the two mean values with the x- and y-axes, and the approximate probability that two samples will be declared to be from the same generator with the z-axis. The x- and y-axis data are plotted with log units. As before, $10,000$ pairs of samples were drawn for each pair of means to estimate the probability of declaring that two samples are from the same generator. It can be seen that there is a ridge of high probability for the cases where the two means are close to each other. For example, for means of ≈ 100 and ≈ 500, the probability of declaring the two samples as being from different generators is ≈ 0.001. In the absence of additional information, there are ranges in the pairs of sample counts where the Poisson log likelihood ratio test will have a high probability for selecting the decision that the samples are from the same generator. Although simulations provide useful information on what models are doing, care has to be exercised in interpreting simulations in those cases where the analyst has more information than the models—such as the true values for the means.

5.1.2 Comparison of Poisson Similarity to Euclidean Distances

To illustrate the accuracy of the Poisson similarity measure of Equation 6, a comparison was made to Euclidean distance measures. For the simulated data, Manhattan and Chebyshev distances are equivalent, because the data are of a single dimension. Other set-based distance measures are not appropriate for this analysis.

The simulated data were generated by selecting mean distribution values for Poisson distribution draws from a gamma distribution with a selected set of shape parameters, α, and rate parameters, β. MATLAB gamma probability distributions use a scale parameter instead of a rate parameter, and the two are inverses of each other. If using MATLAB, the inverse must be accounted for. For the simulation, two sets of data were generated with 10,000 sample pairs drawn for each set. The first set was generated with pairs drawn from the same Poisson distribution, a different mean was selected for each pair and was drawn from a gamma distribution prior. This composed the set where the correct decision is that the samples are from the same distribution. A second set of 10,000 pairs of samples was also generated. This set was composed of pairs where each element was drawn from a different Poisson distribution function; a different Poisson mean was used for each measurement with each mean drawn from the same gamma distribution prior. This composed the set where the correct decision was that the pair were not from the same Poisson distribution. The two datasets can be used to plot Receiver Operating Characteristics (ROC) curves of the probability of correct detection versus the probability of false alarm for different decision thresholds.

ROC curves were drawn for the Poisson similarity measure and the Euclidean distance between measurement sample counts.

The Poisson similarity measure function was implemented in MATLAB and takes, as input, the two sample counts, an optional prior probability for similarity, defaulting to 0.5, two optional gamma prior values for the scale and rate, defaulting to 0.1 for both, and two time duration scale factors for each sample, measured with respect to the time duration assumed for the prior distribution. These default to 1.0.

Figure 3 shows the ROC curve where the actual parameters for the gamma distribution function used to generate Poisson mean values are shape = 1.0 and rate = 10.0, and the presumed shape and rate parameters are the same for the Poisson similarity measure. The ROC curve for the Poisson similarity measure dominates the Euclidean distance curve for all possible choices for probability of false alarm P_{FA}.

Figure 4 shows the case where the Poisson similarity measure uses the prior gamma distribution parameters, shape = 0.1 and rate = 0.1. In this case, the Poisson similarity measure slightly underperforms the Euclidean distance in the mid-range of the x-axis while performing slightly better in the range of low probability of false alarm.

Figure 5 shows the case where the Poisson similarity measure uses the prior gamma distribution parameters, shape = 1.0 and rate = 1.0. This prior distribution indicates greater confidence on the likely values for the Poisson means than that plotted in Figure 4. In this case, the Poisson similarity measure underperforms the Euclidean distance to a much greater degree in the mid-range of the x-axis, although performing slightly better in the range of low probability of false alarm. This demonstrates that the selection of parameters for the prior distribution can have a significant impact on the performance of the Poisson similarity measure. If prior information on likely mean values are available, then taking advantage of that information will improve performance. If the priors are overconfident, it will hurt performance.

5.1.3 Simulations of Binomial Similarity Tests

Similar simulations as done for the Poisson similarity measure can be done for the multinomial similarity measure. Only the binomial case is considered so that the plots are simpler to produce and analyze. Figure 6 shows the empirical probability of correctly classifying two samples drawn from the same binomial distribution as a function of total sample size. For this evaluation, $10,000$ draws were made to estimate the probability that pairs of samples were from the same distribution. The binomial probability distribution vector was set to $(0.5, 0.5)$. The prior probability for the two samples being from the same distribution was set to 0.5. The Dirichlet prior parameter vector for the multinomial measure function was set to $\alpha = (1, 1)$. The general trend is that the probability of correctly determining that two samples are from the same binomial distribution increases with increasing sample counts, as expected. It can also be seen that there is a saw-tooth structure to the curve, caused by the quantization of possible counts, which must be integer values. For a sample size of 1, there is an even chance that the correct decision will be made. More disturbing is that for samples with a total of two counts, the probability is about 0.37 for making the correct decision. It is apparently the case that random guessing would produce a better chance of guessing correctly. This interpretation is incorrect because the similarity measure does not have access to the information on the correct binomial probability distribution. This a flaw of this form of analysis, which will be shortly addressed.

A more appropriate evaluation is to draw the binomial probability distribution from a Dirichlet prior distribution and evaluate the probability of making the correct decision that the two samples of binomial counts are drawn from the same generator. Again, $10,000$ draws of sample pairs were made, but each pair was drawn from a binomial distribution where the probability distribution was drawn from a Dirichlet distribution. For this specific analysis, the draw was from a uniform distribution, which was $\alpha = (0.5, 0.5)$. In this case, the probability of making the correct decision (when the total sample count was 1) was better than random chance. More encouraging was that the probability of correctly determining that the two-count samples were from the same binomial distribution was 0.53, which is better than randomly guessing. The saw-tooth pattern due to quantization of the counts to integer values is still visible. Clearly, better decisions can be made with larger sample counts. It may be noted that the total counts can be different between the two sample vectors. The multinomial similarity measure naturally accounts for this. This additional analysis does not appear in this paper.

A surface plot of the probability for classifying two samples as being drawn from the same binomial distribution is shown in Figure 8. Again, $10,000$ pairs of samples were drawn to estimate the empirical probability that two samples were from the same generator. The prior probability for the decision that the samples were the same was 0.5 and the Dirichlet prior parameter was $\alpha = (0.5, 0.5)$. The plot shows the probabilities for the first two dimensions of the binomial probability vectors associated with two different generators, $p_A(1)$ and $p_B(1)$. The values for $p_A(2)$ and $p_B(2)$ are, of course, one minus the plotted values for the corresponding element. As with the Poisson surface plot in Figure 2, there is a ridge of high probability when the two distributions are similar.

Figure 9 shows ROC performance curves for the Euclidean distance, cosine distance, and the multinomial similarity measure. The simulated data were an equal split of sample count vectors generated from either identical or different binomial distribution functions with $10,000$ samples in each set. Each sample was generated from a unique binomial distribution drawn from a Dirichlet prior distribution with the parameter $\alpha = (1, 1)$. The matching set had count vectors drawn from the same generator. The Manhattan, Chebyshev, and Bray-Curtis distances were also examined. Although the distance values differed between these three measures, there is a linear relation between these measures such that these three measures and the Euclidean distance measure produced the same ROC curves. Only the Euclidean distance ROC curve has been plotted. Note that this equivalence is due to how the simulated data were generated. Other datasets may not have the same equivalence between these measures. The figure shows that the binomial similarity measure and the Euclidean distance measure slightly outperform the cosine distance measure. The binomial similarity measure slightly outperforms the Euclidean distance measure for high P_D.

Figure 10 shows similar ROC performance curves to Figure 9, but with the presumed prior Dirichlet distribution parameter set to $\alpha = (0.1, 0.1)$. There is some performance loss for the binomial similarity measure, with slightly worse performance than the Euclidean and cosine distance curves, although not severe. Figure 11 shows the case where the presumed prior Dirichlet distribution parameter was set to $\alpha = (10, 10)$. The performance loss is much more significant for this case than for the other two cases. This is a case where the estimate on the prior distribution is more confident about the binomial probability distribution than is warranted, demonstrating that it is better to err on the side of being overly uncertain as opposed to overly confident.

5.2 SIMULATIONS TO TEST MERGER BAYES FACTORS FOR POISSON DISTRIBUTIONS

A selected number of simulations have been executed to examine how the merger measures perform under a few different cases. Simulations have not been executed for binomial and multinomial merger measures. This is primarily because the main focus of the research was on mergers for datasets that were best modelled with Poisson distributions instead of binomial and multinomial distributions. Whereas the binomial merger measures are not overly complicated to convert to computer software, the multinomial measure, with nested triple summations, would require a major effort to convert to computer software and achieve reasonable numerical precision for the range of possible parameters. This effort will be pursued when a need arises.

A curve for the Poisson merger Bayes factor is shown in Figure 12. This curve was calculated for the case where the data generator was the same before and after the possible merger. The prior probability for merger was set at 0.5. The parameters for the gamma distribution prior were $\alpha = 0.01$ and $\beta = 0.01$. The counts for the two groups before the potential merger was fixed at 10. The counts for the group after the potential merger ranged from 0 to 100 with a step size of 1. For low values, the probability is that the two groups did not merge, as given by negative values for the log Bayes factor on the y-axis. The transition occurs at 15 counts and the log Bayes factor continues to become more positive as the counts for the resultant group continue to increase.

The continual increase in the probability for merger with the consistent generator model can be problematic for real situations. It is more likely that the destination group has changed behavior rather than merged with the other group if the activity counts are dramatically larger than the sum of the counts for the two original groups. The merger measures of Section B.1 were derived to account for possible changes in the generator before and after the merger event. Figure 13 shows a curve for a simulation run where there is some probability that the destination group activity changed after the possible merger event. The prior probabilities for the four cases are 10% for merger with no generator change, 20% for merger with a generator change, 30% for no merger with no generator change, and 40% for no merger with a generator change. The parameters for the gamma distribution prior were $\alpha = 0.01$ and $\beta = 0.01$. The counts for the two groups before the potential merger was fixed at 10. The counts for the group after potential merger ranged from 0 to 100 with a step size of 1. The log Bayes factor is seen to dramatically increase between the step from 0 to 1 count, indicating that the probability of merger significantly increases for a single step, although still very unlikely. This dramatic rise looks to be accurate for the derived equations, although possibly not representative of real systems. The preferred decision is for no merger until the destination group has 15 counts, when the probability for merger becomes more likely. The probability for merger is most likely for a count of 29, where it begins to trend to a lower probability for merger. The preferred decision switches toward no merger at a count of 46. The shape of the curve is influenced by the counts, the prior probability distribution and the parameters of the prior gamma distribution function. This may be more desirable performance for a scoring function to detect mergers between groups. However, it requires that the prior parameters be determined from a meta-analysis or selected to produce the desired behaviors.

The Poisson merger Bayes factor in Section 4.1 contains gamma functions and hypergeometric functions. The MATLAB implementation of these simulations was plagued by numerical precision errors for large values. These were quite severe for counts greater than an order of 100 for the hypergeometric functions. The simulation software was written to use logarithms of the equations in Section 4.1 to increase numerical precision for the gamma functions. The applicable function in MATLAB is the *gammaln* function. The python scipy package appeared to be less susceptible to precision errors, but would require more study to evaluate its sensitivity. Production software for these merge-split measures will have to be carefully written and validated to avoid these potential numerical precision problems.

Figure 14 shows ROC curves for a simulation of the possible merger of two groups containing randomly selected groups of members from a pool of 100 members. The members were drawn from the pool using a Dirichlet distribution with the parameter vector $\alpha = (6, 50)$ to randomly determine if a member belonged to a group. It was possible for a member to belong to both groups. Activity levels for each member of a group were drawn from a Poisson distribution where the mean parameter λ for the member was drawn from a gamma distribution function with $\alpha = 0.1$ and $\beta = 0.01$. These parameters result in a mean of 10 counts and a standard deviation of 32 counts for the Poisson distribution. The simulation generated 1000 runs each for merger and non-merger datasets. For a merger, the counts for the merged group were generated from the two λ values used to generate counts for the two initial groups. If a member belonged to both groups, both λ values were added to generate that member's activity level. If a member belonged to a single group, that member's λ value as used to generate the activity level for the merged group. For the case of no merger, the λ values of the potential destination group members were used to generate the counts. The parameters for the Poisson merger measure were set to the following values for the simulation: the prior probability for merger was set to 0.5, and the prior parameters for all of the prior gamma distributions for the three groups were set to $\alpha = 0.001$ and $\beta = 0.001$. This was a more conservative selection than the actual parameters that were selected to generate the count data, which were $\alpha = 0.1$ and $\beta = 0.01$.

A number of different distance measures were selected to compare with the Bayes factors. Most of these functions could be invoked with the MATLAB *pdist* function. These included the Euclidean, Manhattan, Chebyshev, cosine, and Spearman distances. These distances were calculated by adding the two prior potential merger groups' counts and comparing to the post potential merger group's counts. The *pdist* function also provides distance measures for the Jaccard and Hamming distances. In this case, the counts were converted to binary vectors that indicated which members were active or inactive. MATLAB functions were also written to calculate the Bray-Curtis dissimilarity measure and the Canberra distance. The cosine and Spearman distances were problematic to calculate when all the counts were zero for a group because the measure is undefined for this case. These cases were eliminated from the estimation of the ROC curves.

Although many of these measures were undefined for zero counts for a member across all groups, the Poisson merger Bayes factor provides a slight preference toward indicating merger for those members of the pool with all zero counts. The Poisson merger measures only included those members assigned to one or both of the two groups, instead of including all 100 members in the calculation.

The ROC curves in Figure 14 show that the Poisson merger measure significantly dominates all the other distance measures examined. The Jaccard and Spearman measures are the next best, but do not come close to the performance of the Poisson merger measure. Although this is only one simulation with a given structure for group membership and activity levels, the Poisson equations are able to apply significantly more available information to detect mergers than the other measures.

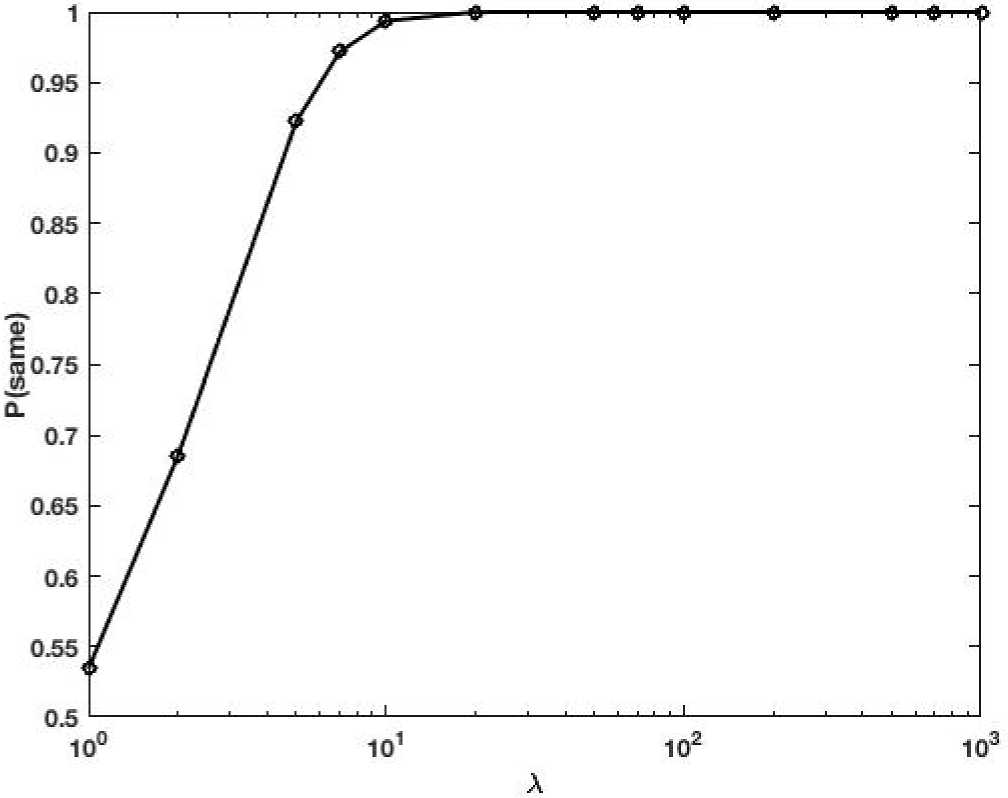


Figure 1. Empirical probability estimates for correctly declaring that two samples are drawn from the same distribution over a range of Poisson means.

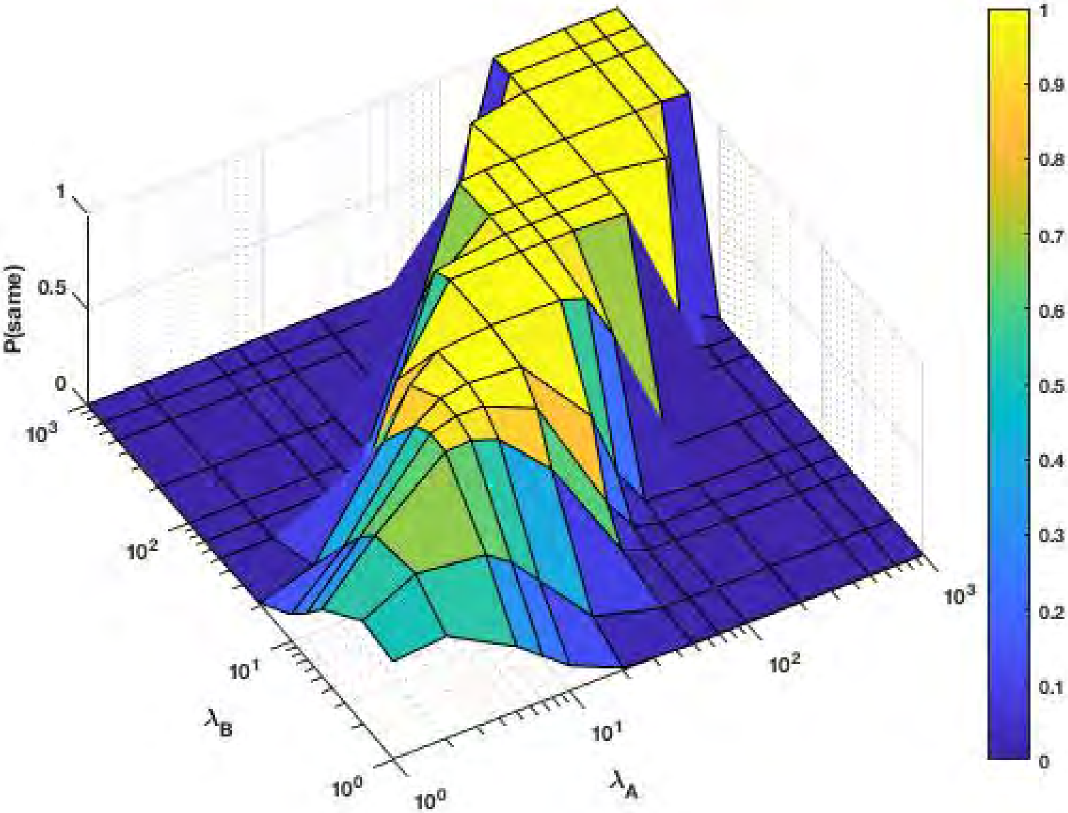


Figure 2. Empirical probability estimates of similarity for two samples drawn from different Poisson generators over a range of means.

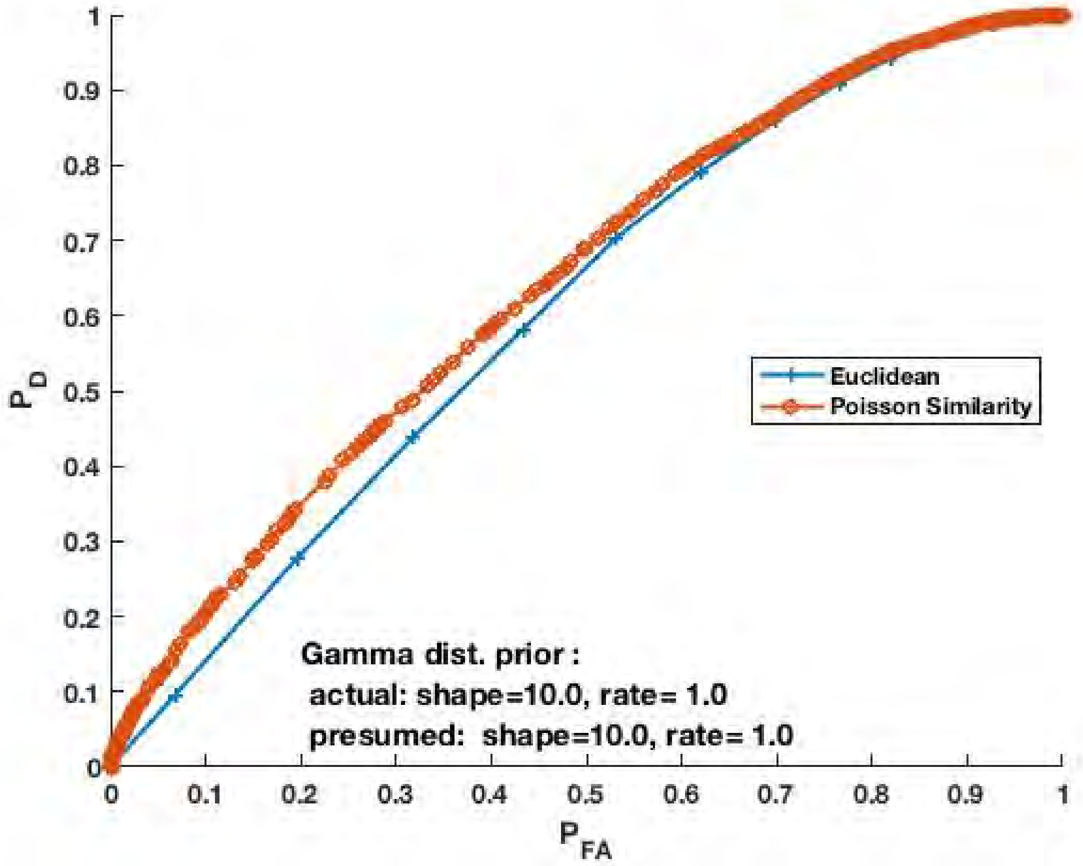


Figure 3. The ROC curves compare the performance of the Poisson similarity Bayes factor to the Euclidean distance measure where Poisson means are drawn from a gamma distribution with shape = 10.0 *and rate* = 1.0*. The Poisson similarity measure uses a prior gamma distribution shape* = 10.0 *and a rate* = 1.0*.*

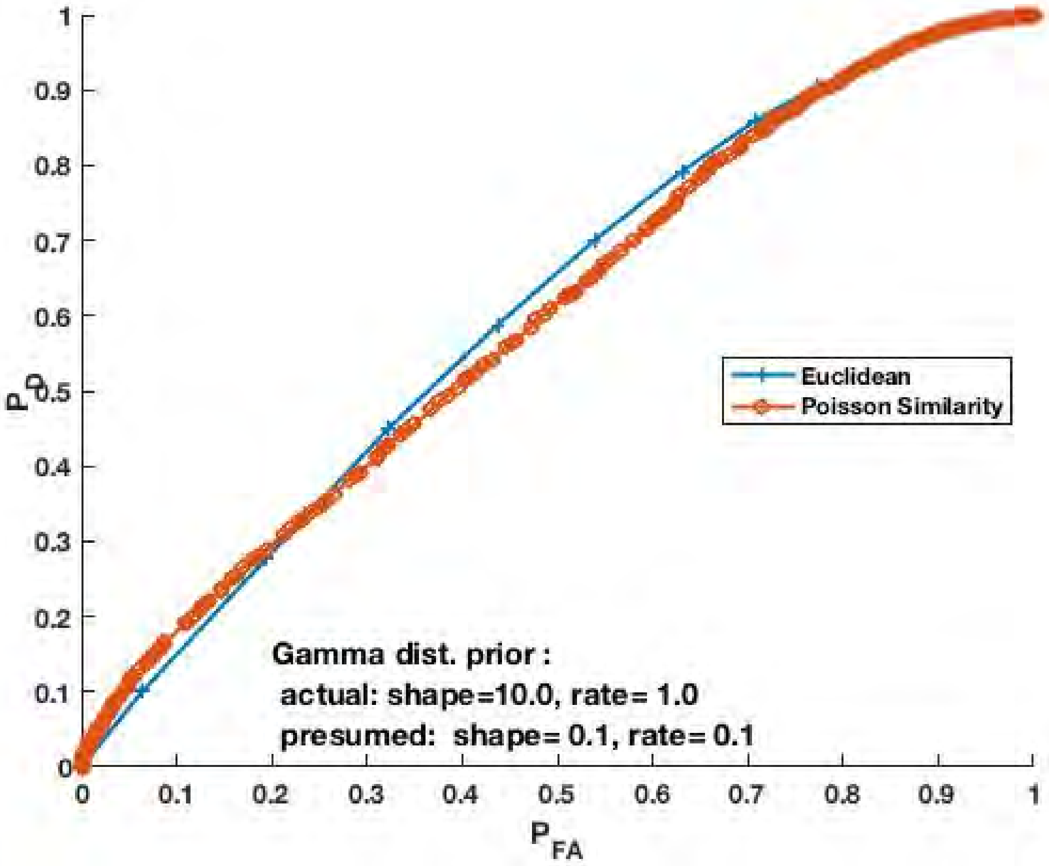


Figure 4. The ROC curves compare the performance of the Poisson similarity Bayes factor to the Euclidean distance measure where Poisson means are drawn from a gamma distribution with shape $= 10.0$ *and rate* $= 1.0$*. The Poisson similarity measure uses a prior gamma distribution shape* $= 0.1$ *and a rate* $= 0.1$*.*

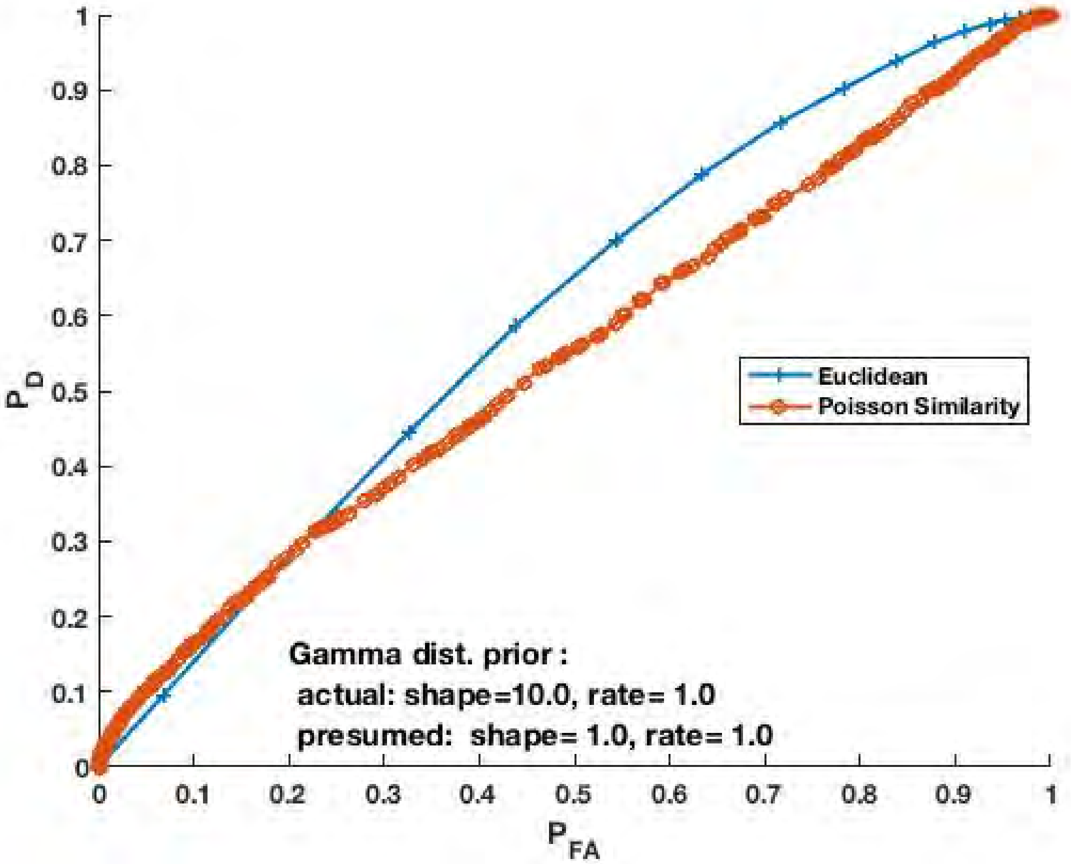


Figure 5. The ROC curves compare the performance of the Poisson similarity Bayes factor to the Euclidean distance measure where Poisson means are drawn from a gamma distribution with shape = 10.0 *and rate* = 1.0*. The Poisson similarity measure uses a prior gamma distribution shape* = 1.0 *and a rate* = 0.1*.*

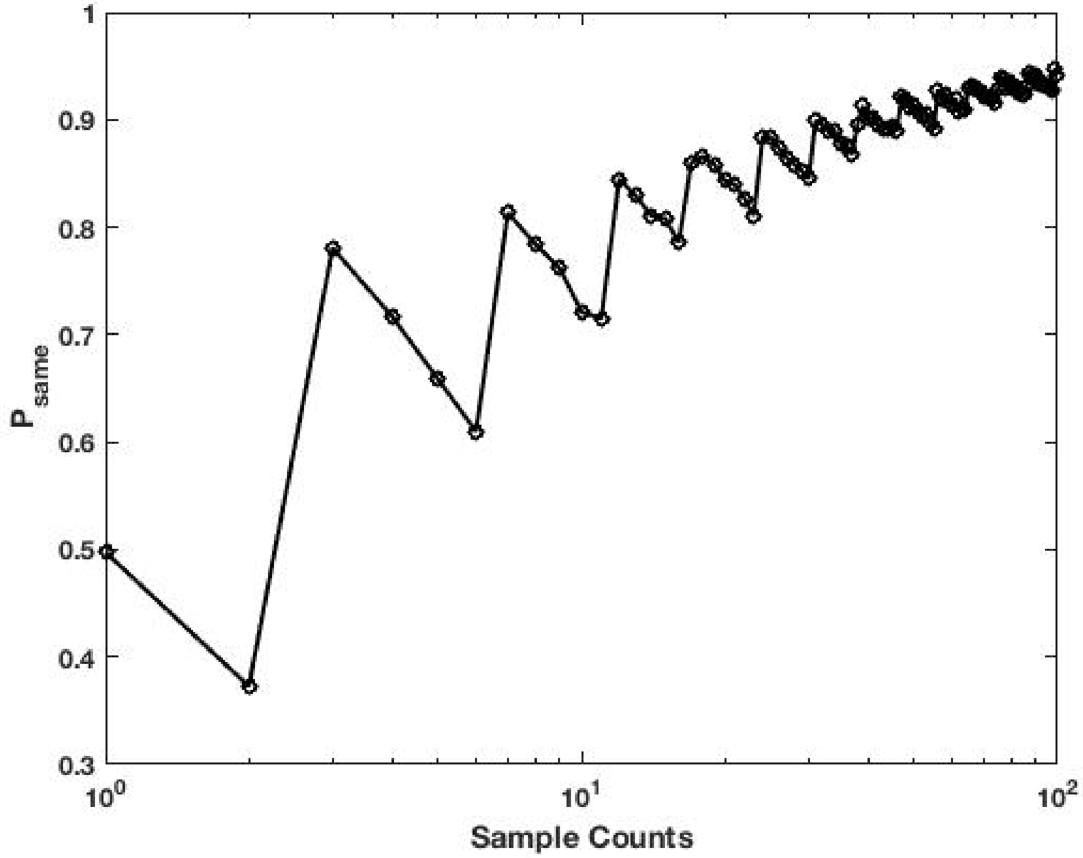


Figure 6. The empirical probability of successfully detecting that two sample vectors are drawn from the same binomial distribution, which was fixed at $(0.5, 0.5)$. *The total number of counts for each vector is plotted on the x-axis.*

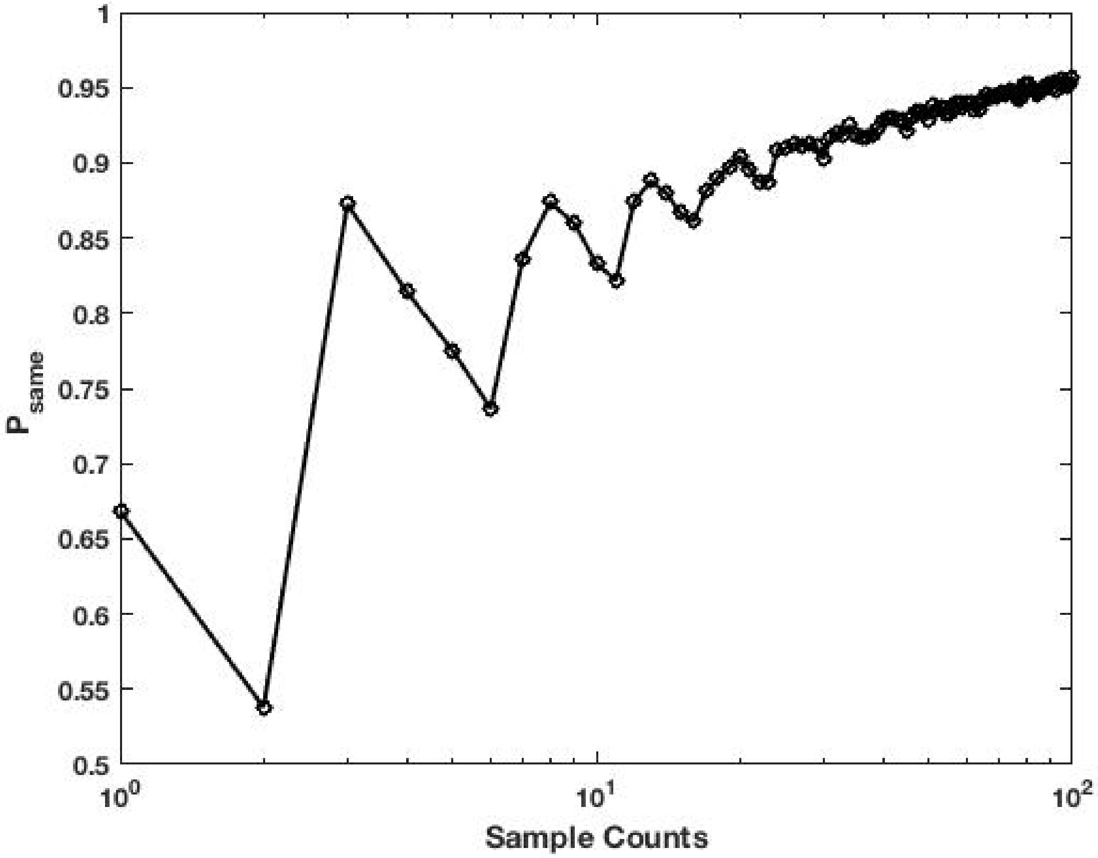


Figure 7. The empirical probability of successfully detecting that two samples are drawn from the same binomial distribution. The categorical probability distribution for each sample pair was drawn from a Dirichlet prior with the parameter vector $(1, 1)$*. The total number of counts for each vector is plotted on the x-axis.*

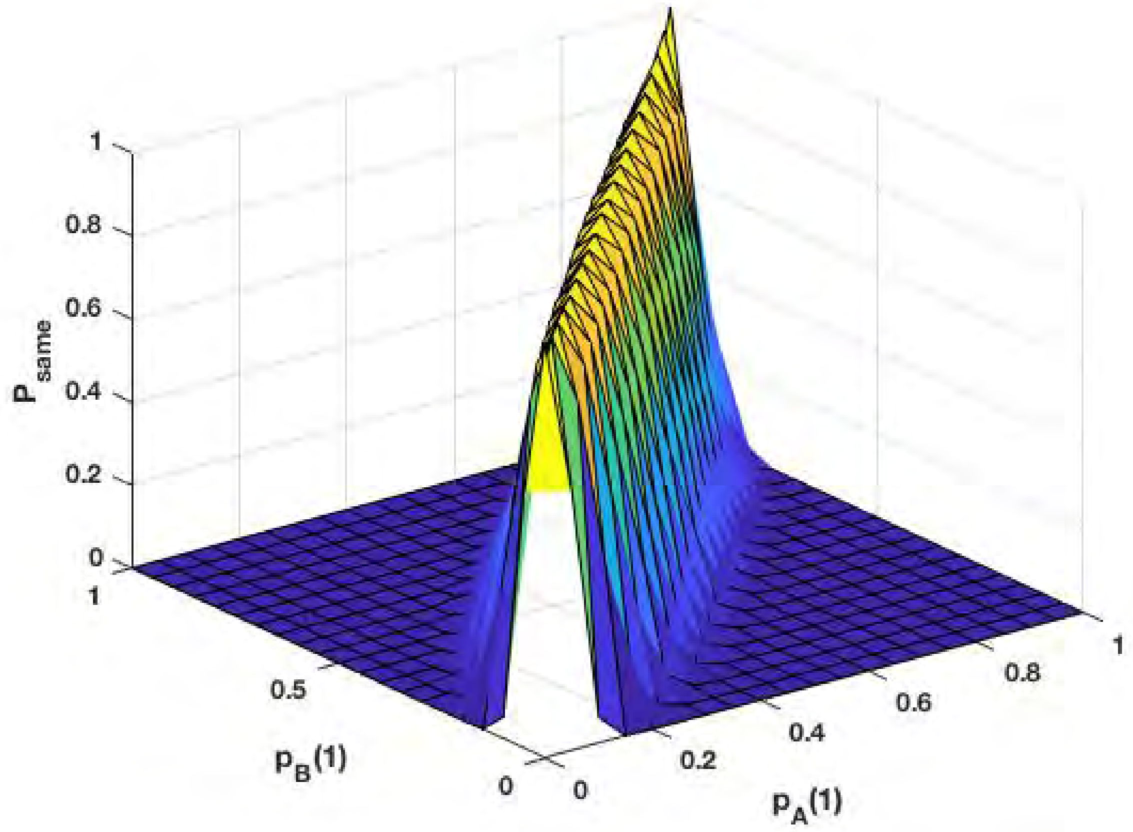


Figure 8. The empirical probability estimates for two samples being declared to be drawn from the same distribution over a range of means for different binomial generators.

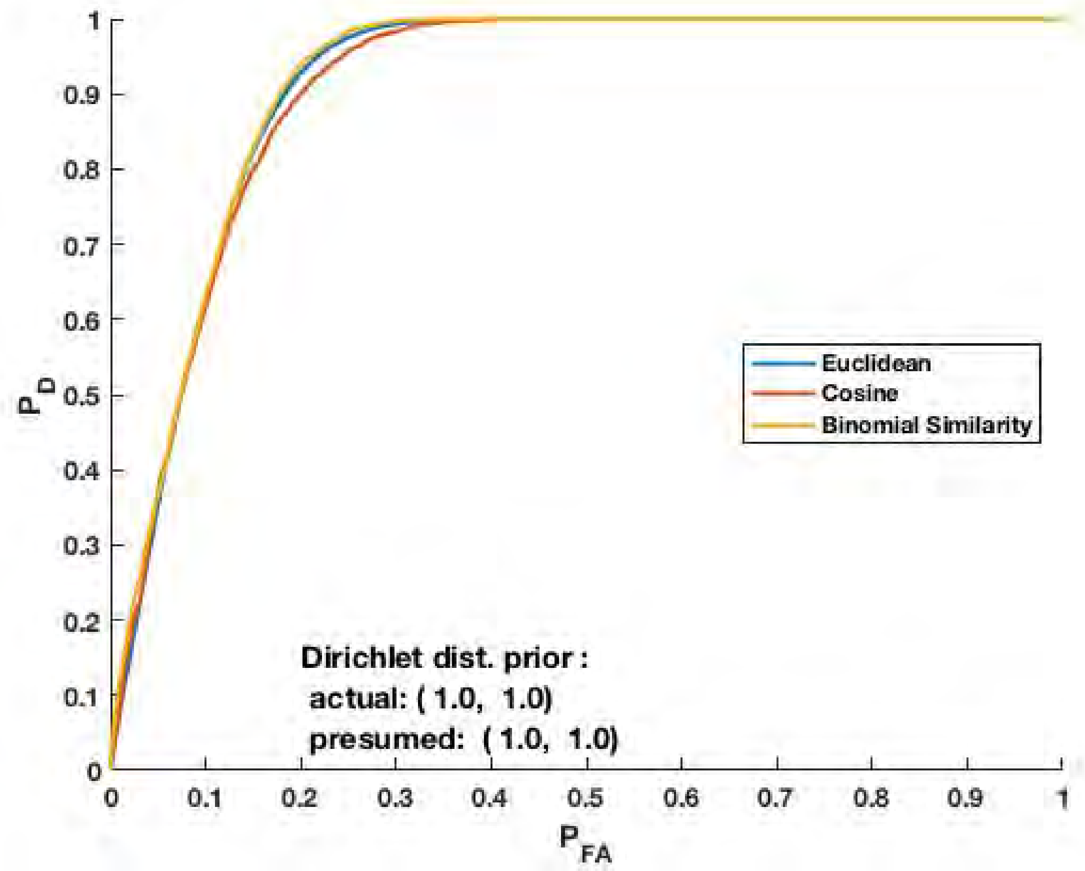


Figure 9. *The ROC curves compare the Euclidean, cosine distance, and binomial similarity Bayes factor for samples drawn equally from the same and different generators. The draws for the binomial probability distributions were drawn from a Dirichlet distribution with a parameter vector of* $(1, 1)$*. The prior distribution vector for the binomial similarity measure was also* $(1, 1)$*.*

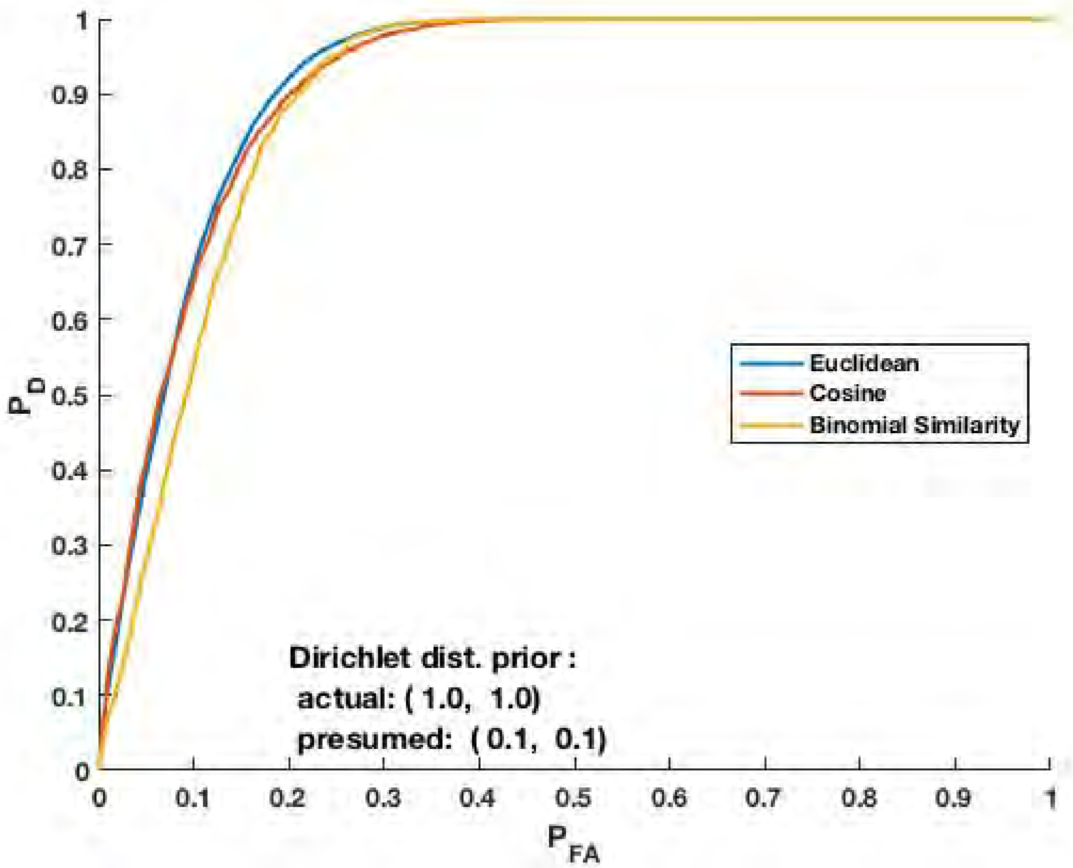


Figure 10. The ROC curves compare the Euclidean, cosine distance, and binomial similarity Bayes factor for samples drawn equally from the same and different generators. The draws for the binomial probability distributions were drawn from a Dirichlet distribution with a parameter vector of $(1, 1)$*. The presumed prior Dirichlet distribution vector was* $(0.1, 0.1)$.

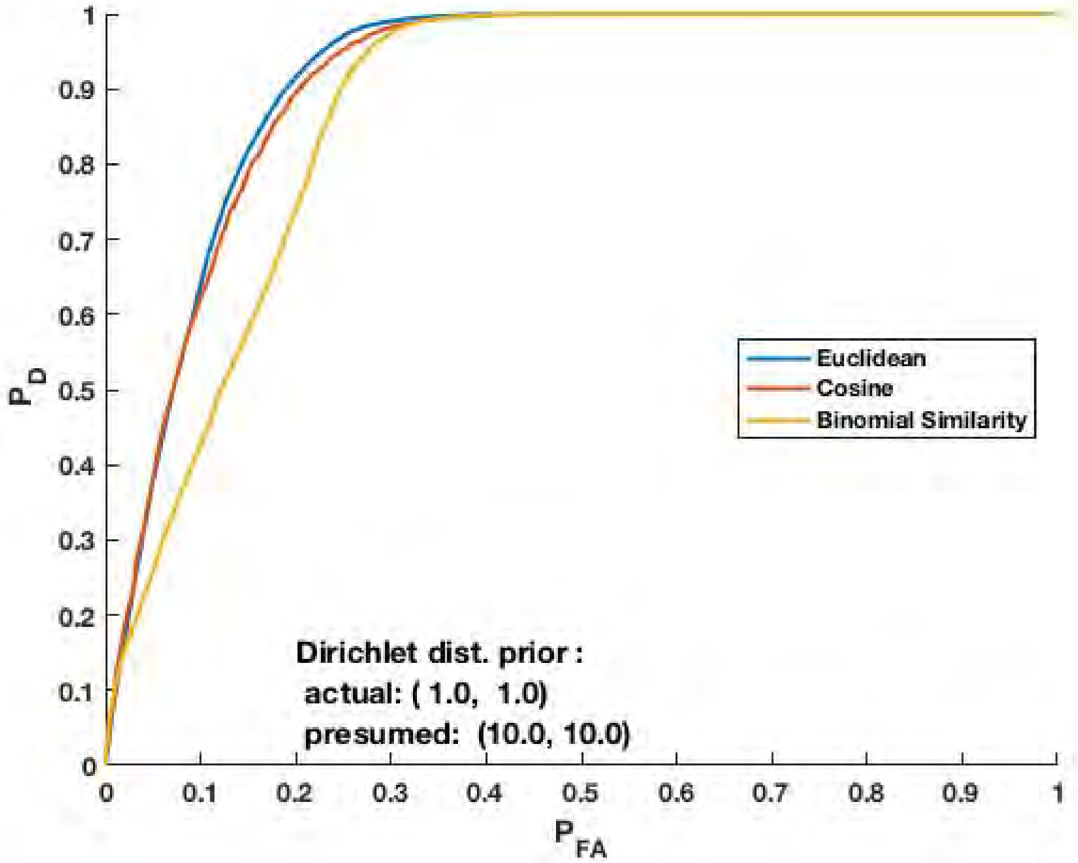


Figure 11. The ROC curves compare the Euclidean, cosine distance, and binomial similarity Bayes factor for samples drawn equally from the same and different generators. The draws for the binomial probability distributions were drawn from a Dirichlet distribution with a parameter vector of (1, 1)*. The presumed prior Dirichlet distribution vector was* (10, 10)*.*

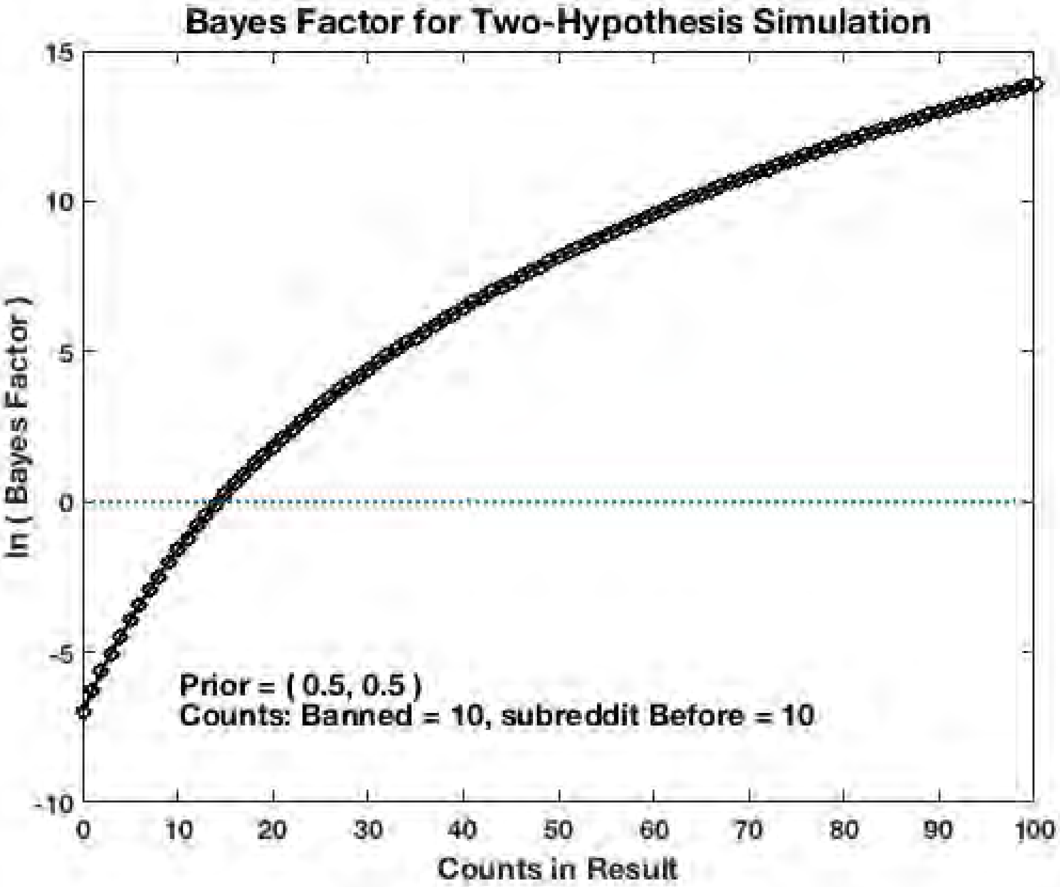


Figure 12. The log Bayes factor curve for Poisson distributions plotted as a function of the counts for the potential destination group of one member's activity. The generator function for the potential destination group is presumed to be the same before and after the possible merger event.

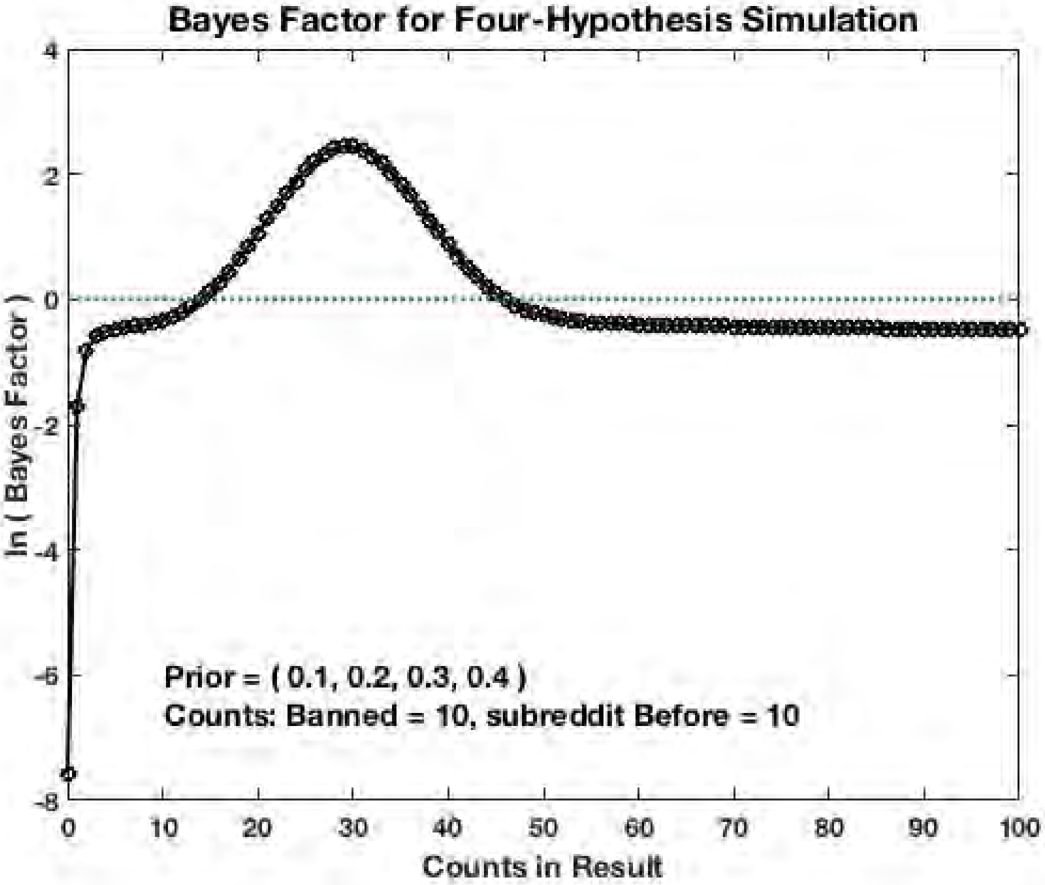


Figure 13. The log Bayes factor curve for Poisson distributions plotted as a function of the counts for the potential destination group of one member's activity. The generator function for the potential destination group is presumed to potentially change after the possible merger event.

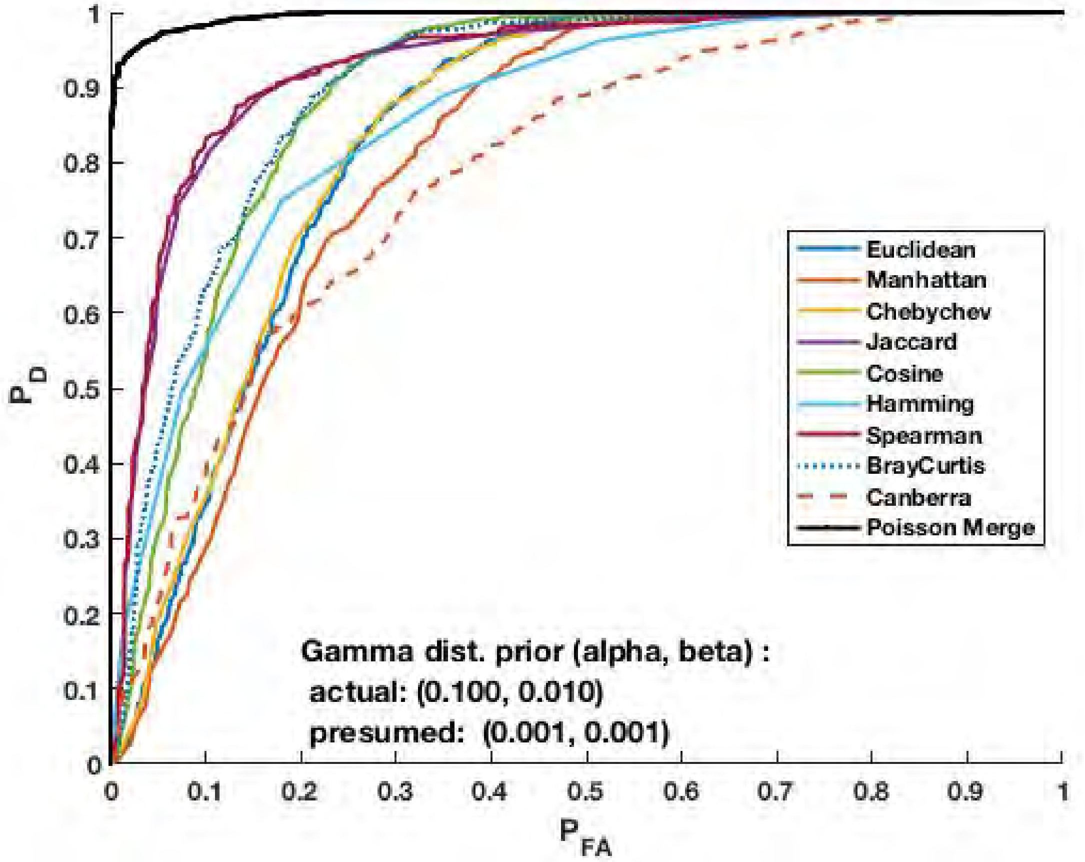


Figure 14. The ROC curves compare the performance of the Poisson merger Bayes factor and a number of other commonly used distance measures.

6. ANALYSIS OF REDDIT DATA

Although the simulation results in Section 5.2 are highly encouraging, application of the Poisson merger measure to real datasets can provide a much stronger indication of the utility of this measure. The Reddit social media platform can be used to extract real-world data to search for changes in posting and commenting behaviors and for mergers between groups (or *subreddits*). Although the Reddit administrators are very liberal in regard to what activities are permitted in subreddits, there are certain activities that can cause subreddits to be banned, such as engaging in or promoting illegal activities, involuntary pornography, sexual or suggestive content involving minors, encouraging or inciting violence, and harassment.

The data used in this report include more than 118 million posts created by more than 8 million unique users on Reddit between March 2016 and September 2017, collected from the archive of Reddit made publicly available on Google BigQuery, omitting known bots and default subreddits. Bots are computer programs that automatically generate posts and comments; 988 bots (190 active) were identified by manually inspecting abnormally active users ($>$ 3000 comments per month) and by scraping /r/botWatcher for recently mentioned bots. Default subreddits (those to which all users are automatically subscribed upon account creation) are linked at the top of every Reddit page and are typically among the most active subreddits; 49 default subreddits (as of March 2016) were omitted from this analysis.

6.1 CHANGE DETECTION IN SUBREDDITS

Equation 5 or 6 in Section 3.2 can be used to measure the week-to-week similarity in Reddit post and comment rates for subreddit members. Equation 5 is appropriate when the time durations are the same, whereas Equation 6 is required when the time durations are different. These equations are appropriate for cases where the activity levels can be modelled with Poisson distributions.

Figure 15 shows the results of an analysis of user comment volumes. The subreddits *CFB* (focused on college football), *nfl* (focused on the National Football League), and *Patriots* (focused on the NFL Patriots team) were selected for the analysis. The top plot is the overall log Bayes factor for changes in user comment activity levels from week to week. The log Bayes factors for each user's change in posting volumes are calculated individually and then summed to obtain the overall log Bayes factors. The values for the *Patriots* log Bayes factors have been scaled by a factor of five so that the peaks are more noticable. The magnitude of the *Patriots* log Bayes factors are smaller because the number of active users is significantly less than that of the other two subreddits.

The second plot is the Euclidean distance between vectors where the elements are user comment volume. Users not actively commenting in a given week are coded as a zero count. The Euclidean distance between pairs of vectors from week to week is plotted. The third plot uses the same user activity vectors to plot the cosine distance, $1 - \cos(\theta)$, between vector pairs. Larger values indicate a greater difference between the weeks. The fourth plot is the count of the number of active dataset commented during any given week. The fifth plot is the overall number of comments for a given week.

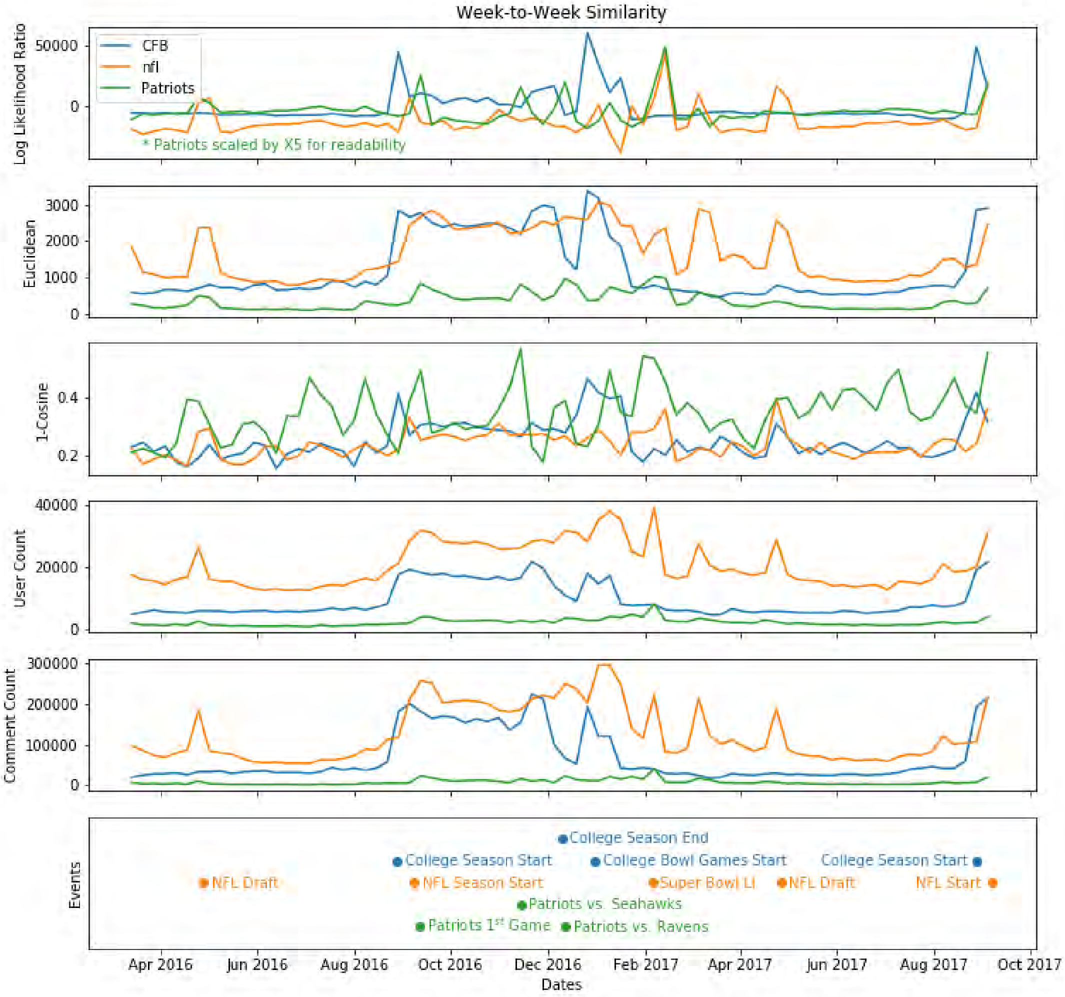


Figure 15. The figure shows the results from an analysis of football associated subreddits: CFB, nfl, and Patriots. Week-to-week differences are plotted for the overall log Bayes factor, the Euclidean distance, and the cosine distance as determined from user comment activity levels. Also plotted are the number of active users and the number of comments posted for each week. The bottom box displays significant events related to the three subreddits that can be observed in the results.

The last and lowest plot shows the dates of significant events realted to the three highlighted subreddits that may have influenced activity levels. Most of the significant events can be seen in the three metrics. The NFL Draft dates are easy to detect for the *nfl* and *Patriots* subreddits; the

NFL draft is irrelevant to college football and is not evident in the *CFB* subreddit. The seasons are easy to detect in all measures in the respective subreddits associated with either college football or NFL football. Significant games can also be seen in the plots, such as the Patriots–Seahawks and Patriots–Ravens games. Activity changes around the time of Super Bowl LI are seen in both *nfl* and *Patriots*. The Patriots played in the Super Bowl in 2016 and thus caused the significant activity change in the *Patriots* subreddit during this event.

The peaks in the cosine distance plots are difficult to discern due to a sensitivity to random fluctuations. The Euclidean distance scales are different for the three subreddits, making it challenging to determine what distance would be appropriate for a change detection threshold. The log Bayes factor plot is much easier to select a threshold for substantial change detection, and it would be quite possible to select a consistent threshold for the three subreddits, even with the factor of five scaling for the *Patriots* Bayes factors.

Any timescale could have been chosen for similarity analysis, from periods of minutes and hours to weeks, months, and years. As expected, shorter time periods have fewer counts and contain greater variation due to increased sensitivity to small sample size fluctuations. Longer durations provide greater sample support and higher accuracy, but will be less sensitive to high frequency, short duration changes. An advantage that the log Bayes factor provides is that values are closer to zero when sample support is low. Figure 16 shows day-to-day results over the 2016 college and NFL football season, similar to the week-to-week analysis in Figure 15. With the higher resolution, the increased activity during game days can be seen in all three subreddits. It may be noted that the bye weeks in the Patriots schedule can be seen, both during the regular season and during the playoffs, although the playoff byes are not noted. It can also be seen that the Super Bowl generates a significantly greater change in activity levels than the other games. The log Bayes factors for the *Patriots* subreddit are again scaled by a factor of five for observability.

The football focused subreddits are fairly well behaved with the log Bayes factors in that a threshold of zero is reasonable for detecting activity changes. This is not generally true for all subreddits. There are a population of subreddits involved in trading and selling items. Their user activity patterns change dramatically from week to week, such that it is rare for two consecutive weeks to be similar. For example, Figure 17 shows the week-to-week log Bayes factors for activity changes for the *pokemontrades* subreddit. The ratio is consistently high across the entire time period, indicating that user activities are always changing in an extreme fashion. This is primarily because a user's interest in trading Pokémon paraphernalia changes dramatically from week to week. It can be seen that there are two spikes in dissimilarity for the weeks of 1 August 2016 and 21 November 2016. Pokémon Go was released in the summer of 2016, and Pokémon Sun and Moon were released in November. There was also a special Pokémon Go Thanksgiving event on 23 November 2016 that may have contributed to the dramatic change in activity. Further analysis has not been conducted.

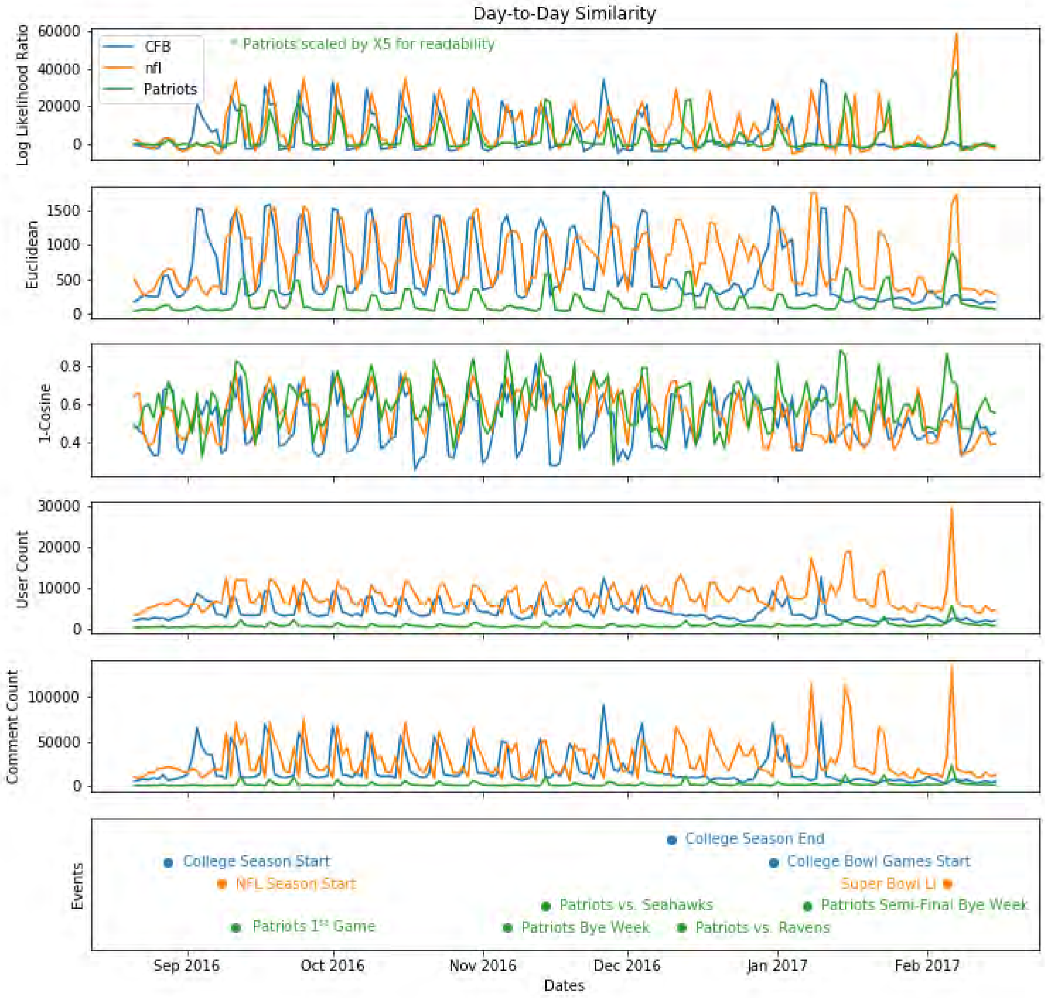


Figure 16. Shown are similar results to Figure 15, but for daily differences over the 2016 college and NFL football season for the subreddits CFB, nfl, and Patriots. Day-to-day differences are plotted for the overall log Bayes factor, the Euclidean distance, and the cosine distance based upon user comment activity volumes. Also plotted are the number of active users and number of comments posted for each week. The bottom box displays significant events related to the three subreddits that can be observed in the results.

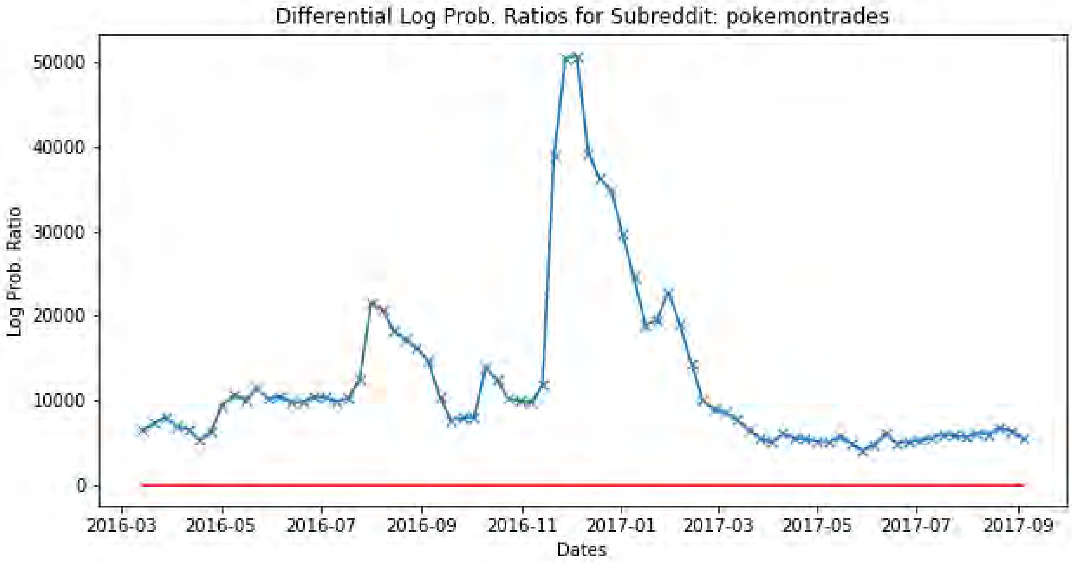


Figure 17. The week-to-week log Bayes factor measures for the subreddit pokemontrades. The dissimilarity is relatively high because the subreddit functions as a marketplace for trading and selling Pokémon-related items. This causes a dramatic change in user activities as their interests wax and wane for various Pokémon-related items.

6.2 DETECTION OF BANNED SUBREDDITS RECONSTITUTING IN OTHER SUBREDDITS

The merger Bayes factors of Section 4.1 for Poisson distribution functions can be used to determine whether members of a banned subreddit have either migrated to another subreddit or created a new subreddit in which to continue their banned activities. For this exercise, it was assumed that users did not re-register with new accounts in order to obscure their identify. The dataset consisted of 13 subreddits that were banned between 1 October 2016 and 31 January 2017. Table 1 lists the banned subreddits and the dates of their banning. Preliminary processing was used to eliminate all potential destination subreddits for the banned members by requiring that at least one member from the banned account had commented in the potential destination subreddit in the week after the banning. The numbers of potential destination subreddits are shown in the 'Possible Subreddits' column. The rightmost two columns show the number of subreddits with log Bayes factor scores above a zero threshold for two cases: the first is where the activities of all members of the potential destination subreddit are used in the calculations and the second is where only the activities of members of the banned subreddit are used in the calculations.

The Google BigQuery database was used to extract the daily user activity levels across the relevant subreddits. Both the post rates and comment rates were studied, but this report only presents the results for the comment rates, which were more informative than the results from post rate data. Using the banning date for a subreddit in Table 1, the last recorded comment date for the banned subreddit was determined with a database query. The total number of comments were extracted for each user for the week before the last comment date. The last comment date was often a few weeks before the banning date because the BigQuery repository only scrapes Reddit posts and comments every few weeks. This causes the database to often miss the final weeks of activity for banned or deleted subreddits. This characteristic of the repository made the analysis somewhat problematic.

After extracting the total comment counts for individual users over the last week of available data, the viable destination subreddits were processed to extract similar one-week user comment counts. Two sets were extracted for each of these subreddits: the activity levels for the same week where comments for the banned subreddit data were available, and the activity levels of the members of the potential destination subreddit for one week after the banning. These three vectors were used to estimate the likelihood of merger as well as to calculate differences with other distance or similarity measures.

The merger log Bayes factors were calculated for all pairs of banned subreddits and potential destination subreddits using the equations of Section 4.1. The Bayes factors were calculated individually for each user, based upon their activity counts in the three datasets. If a user was not active in a subreddit for a given time period, the activity count was set to zero. The overall merger Bayes factor score was the sum of the merger Bayes factor scores for all users. This assumed that user activities were not correlated. Clearly, this is unlikely to be true for social networks, but the hope is that any biases on the overall scores would be small enough so that the most likely subreddits involved in possible mergers with banned subreddits could still be identified.

The prior probability values for the log Bayes factor functions were set to $\alpha = 0.001$ and $\beta = 0.001$ for the gamma distribution function. A uniform distribution was chosen for the prior probability for merger. A threshold of zero was selected as the threshold for consideration as a possible destination for the banned members.

Two different analyses were run. The first analysis constructed activity vectors from all the users who were active in a potential destination subreddit. The Bayes factor tests were found to be sensitive to the inclusion of large numbers of users in very active subreddits with high fluctuations in activity levels. A second analysis was performed that constructed the activity vectors using only the members of the banned subreddit in the banned subbreddit and in a potential destination subreddit.

Table 1 shows the number of subreddits with merger Bayes factors above the zero threshold for both types of user activity vectors. When all members of a potential destination subreddit were used, a larger fraction of the merger scores were above the zero threshold. It is possible to raise the threshold to reduce the number of subreddits that were selected as possible merger targets and that an analyst would be required to review, if desired. The second analysis, where only banned users are included, resulted in a significantly higher rejection of subreddits when using a zero threshold. In most cases, no matches were found for a threshold of zero on the merger Bayes factor. Only *top10gifts* and *whathappensonsnapchat* had scores above threshold for this reduced dataset. It is very likely that shifts in the activity patterns of the members in the potential destination subreddit caused the higher acceptance fraction in the first analysis.

TABLE 1

A Selection of Banned Subreddits from October 2016 Through January 2017 with Counts of Possible Destination Subreddits Based Upon Two Different Threshold Limits

Subreddit Name	Date Banned	Possible Subreddits	Above Threshold (All)	Above Threshold (Banned)
candid	15 Dec 2016	128	113	0
hotwife_cuckold	01 Jan 2017	215	123	0
houstonr4r	07 Oct 2016	49	48	0
idg0d	30 Sep 2016	25	19	0
idgod	01 Oct 2016	50	32	0
leftwithsharpedge	15 Dec 2016	431	302	0
nsfwshoops	07 Dec 2016	37	32	0
pedofriends	20 Dec 2016	929	201	0
pharmacyreviews	11 Jan 2017	217	139	0
publichealthwatch	06 Nov 2016	490	236	0
rawcelebs	01 Jan 2016	377	253	0
top10gifts	21 Jan 2017	80	76	35
whathappensonsnapchat	15 Jan 2017	101	79	1

The *top10gifts* subreddit had a sizable number of potential subreddits above threshold. Table 2 shows the comment activity of the users from the banned subreddit *top10gifts* and a list of the top five possible subreddit destinations. Only Abdul_Marx and DianeBcurious were active in these five subreddits, and the log Bayes factors correlate with the number of comments. None of the users from the banned subreddits were active at all in the top five possible subreddits during the last week of available data for the banned subreddit. Given this more in-depth look, as well a more detailed look at *whathappensonshapchat*, it is highly unlikely that any of the 13 banned subreddits reconstituted in another subreddit.

6.2.1 Comparison of the Poisson Merger Bayes Factor to Other Measures

A comparison can be made between the Poisson Bayes factor merger score results and a number of other distance measures to get a better sense of how well the Bayes factor scores perform with respect to other common similarity and distance measures. The user activity levels were represented as vectors, as in the previous subsection. This means that other similarity and distance measures such as the dot product, Euclidean distance, Manhattan distance, Canberra distance, and cosine distance can also be used to detect mergers. Two different methods were developed for detecting mergers with these measures. For generalization purposes, it can be assumed that the function $S\left(\overrightarrow{x}, \overrightarrow{y}\right)$ is a general representation function for similarity (distance) measures between

TABLE 2

The Number of Comments and Overall Log Bayes Factors for the Five Most Likely Destination Subreddits for *top10gifts* After Banning

User	top10gifts	facepalm	magicTCG	Aquariums	Christianity	therewasanattempt
Abdul_Marx	1	7	0	0	3	2
DianeBcurious	1	0	5	4	0	0
Tommybundy6745	1	0	0	0	0	0
cmg232	1	0	0	0	0	0
Scores	-	6.08	5.75	5.52	5.24	4.83

vectors. If the similarity (distance) between a banned subreddit vector and the 'after' subreddit vector is greater (less) than the similarity (distance) between a banned subreddit vector and the 'before' subreddit vector, then there is a possibility of a merger. The notation used in the following descriptions are that b represents the banned subreddit, B represents the potential destination subreddit before banning, and A represents the potential destination subreddit after banning.

The second comparison of similarity (distance) vectors is performed with a combined vector that is composed of the banned subreddit vector and the 'before' subreddit vector. The notation to represent this combined vector is 'C'. This second comparison looks to see if the similarity (distance) between the C vector and the A vector is greater (less) than the similarity (distance) between the B vector and the A vector. If so, then it is possible that a merger may have occurred.

The first merger detection metric is defined to be

$$S_{M_a}\left(\overrightarrow{b},\overrightarrow{B},\overrightarrow{A}\right) = S\left(\overrightarrow{b},\overrightarrow{A}\right) - S\left(\overrightarrow{b},\overrightarrow{B}\right), \tag{28}$$

where S_{M_a} is a similarity (distance) measure. The second merger detection metric is similar to the first. The equation is

$$S_{M_b}\left(\overrightarrow{b},\overrightarrow{B},\overrightarrow{A}\right) = S\left(\overrightarrow{b}+\overrightarrow{B},\overrightarrow{A}\right) - S\left(\overrightarrow{B},\overrightarrow{A}\right). \tag{29}$$

A second class of similarity measures consists of set measures. The activity vectors can be converted to sets of active users. These measures include Jaccard similarity and Bray-Curtis

TABLE 3

Selected Measures of Distance and Similarity

Vector Distances	
Euclidean Distance	$d_E\left(\overrightarrow{a}, \overrightarrow{b}\right) = \left(\sum_i (a_i - b_i)^2\right)^{1/2}$
Manhattan Distance	$d_M\left(\overrightarrow{a}, \overrightarrow{b}\right) = \sum_i (a_i - b_i)$
Canberra Distance	$d_C\left(\overrightarrow{a}, \overrightarrow{b}\right) = \sum_i \frac{\left\lvert \overrightarrow{a}_i - \overrightarrow{b}_i \right\rvert}{\lvert \overrightarrow{a}_i \rvert + \left\lvert \overrightarrow{b}_i \right\rvert}$
Dot Product Similarity	$s_D\left(\overrightarrow{a}, \overrightarrow{b}\right) = \sum_i a_i b_i$
Cosine Similarity	$s_C\left(\overrightarrow{a}, \overrightarrow{b}\right) = \frac{s_D\left(\overrightarrow{a}, \overrightarrow{b}\right)}{d_E\left(\overrightarrow{a}, \overrightarrow{a}\right) d_E\left(\overrightarrow{b}, \overrightarrow{b}\right)}$
Set Similarities	
Jaccard Similarity	$s_J(\{a\}, \{b\}) = \frac{\lvert\{a\}\rvert \cap \lvert\{b\}\rvert}{\lvert\{a\}\rvert \cup \lvert\{b\}\rvert}$
Bray-Curtis Dissimilarity	$d_{BC}(\{a\}, \{b\}) = 1 - \frac{\lvert\{a\}\rvert \cap \lvert\{b\}\rvert}{\lvert\{a\}\rvert + \lvert\{b\}\rvert}$

dissimilarity. The activity levels do not factor into the calculation of these measures and only take into account membership.

Table 3 provides a general outline of the formulas used to estimate the selected measures for the comparison to the merger Bayes factor scores. Because differences are used to detect mergers, a zero threshold can be used with the difference calculations, as with Bayes factors.

Figures 18 through 32 show the histograms of differences between pairs of similarities or distances for the *top10gifts* banned subreddit. Figures 18 through 24 show the results for the case that all users' activity levels are used to construct the activity vectors. Figures 26 through 32 show the results where the activity vectors are constructed from only the activities of the users who are active in the banned subreddit during the last week of stored data within the database. The histograms on the left side of the figures show the differences between the banned vector and the 'after' vector versus the banned vector and the 'before' vector. The histograms on the right side of the figures show the differences between the combined vector and the 'after' vector versus the 'before' and 'after' vector.

Figures 18 through 24 have been included for completeness. Although the dot product distance measures are similar between the use of the full activity vectors versus the vectors including only users from the banned subreddits, the other common distance and similarity measures show greater differences and generally better results with the vectors that only use the banned subreddit members' activity levels. The merger Bayes factor score histograms are also different between the two analyses, which are plotted in Figures 25 and 33. Most of the following discussion focuses on the figures where the activity vectors only include the banned subreddit members.

Figures 26 and 27 show the histograms for the Euclidean and Manhattan distance measures. The distribution shows that there are a few subreddits with high difference values, and these are presumed to be the most likely subreddits out of the 80 subreddits to have merged with *top10gifts*.

Figure 28 has peaks in the histograms at the ends of the range for the banned vector comparison. The Canberra distance is relatively insensitive to activity level differences between many of the subreddits. The expectation is that the Canberra distance decreased if there was a merger. The comparison involving the combined vector shows that the Canberra distance always increased, so no mergers would be indicated for this case.

Figure 29 shows the histograms for the dot product similarity measure. There is a nice range of distributions, with a few having large positive similarity differences. It is also insensitive to the inclusion of other user activities in the vectors for the potential destination subreddit.

Figure 30 shows the histograms for the cosine similarity measure. There are only a few peaks in the histogram. This is because the number of members in the banned subreddit are so low that there is little to distinguish the similarities between potential destination subreddits. In many cases, the magnitude of some vectors is zero, which precludes the calculation of a cosine. The algorithm was written to return a value of -1 in these instances. These values have not been excluded from the histograms and are the cause of large negative values. This measure performs poorly when the number of users is low. The same problem can be seen with the Jaccard similarity difference histograms in Figure 31. The Bray-Curtis dissimilarity difference histograms in Figure 32 has the same problem, although there are more small peaks across the histogram, but the end peaks are still dominant. Although results look somewhat better for the same measures with the activity vectors for all users, the cosine, Jaccard, and Bray-Curtis measures still show some of the same problems (see Figures 22, 23, and 24).

Table 4 summarizes the counts of potential destination subreddits with similarity differences above a zero threshold for *top10gifts*. Although this report does not contain figures that show results for the similarity analysis where all subreddit user activities were included in the activity vectors, Table 4 contains a summary of this analysis in the second and third columns. The fourth and fifth columns summarize the results of a similarity analysis restricted to the activity of the members of the banned subreddit. Columns two and four show the counts for the comparison of measures between the banned and the ‘after’ vectors versus the banned and the ‘before’ vectors. Columns three and five show the counts for the comparison of measures between the combined and the ‘after’ vector versus the ‘before’ and the ‘after’ vector. The acceptance rates for the merger log Bayes factors are shown at the bottom of the table. The merger Bayes factor appears to be sensitive to the inclusion of all subreddit user activity, most likely due to fluctuations in the potential group members’ activity levels. With the restriction to consider only banned user activity, its performance is similar to many of the other measures.

Tables 5 and 6 show the top five possible destination subreddits for the different distance and similarity measures for both versions of calculating scores for merger, as given by Equations 28

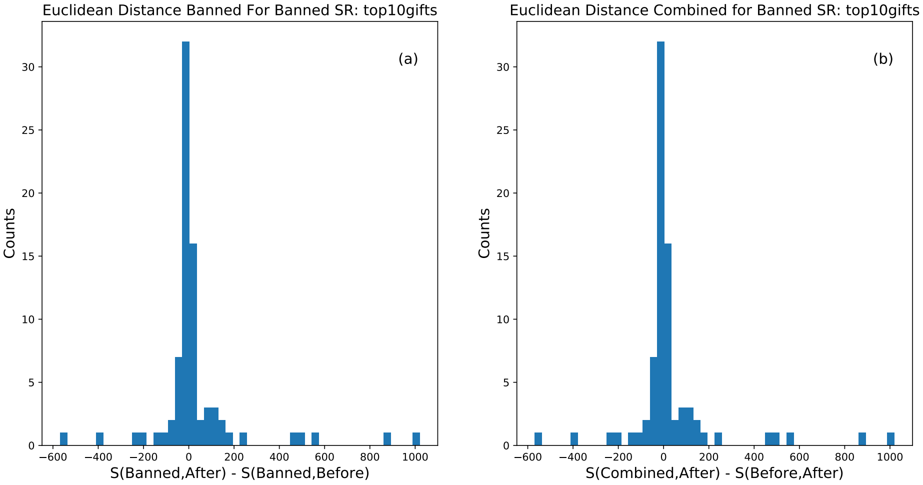


Figure 18. Histograms of the Euclidean distance differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

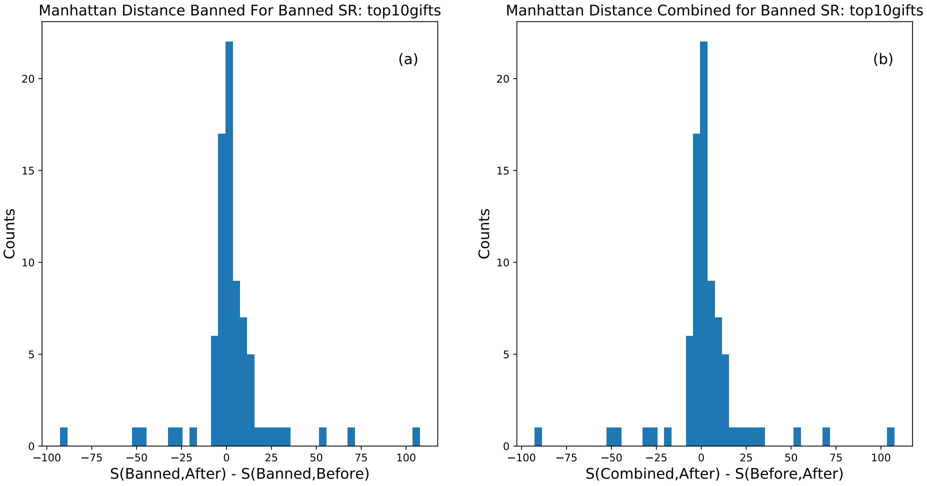


Figure 19. Histograms of the Manhattan distance differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

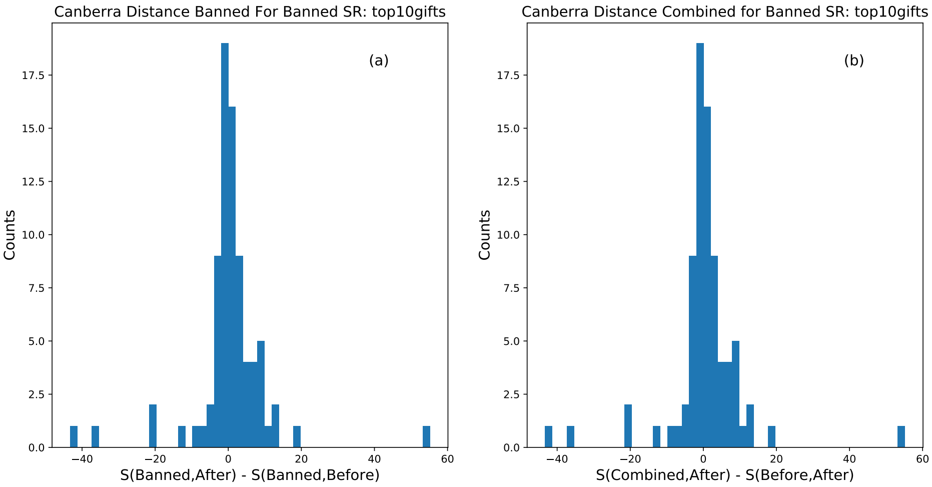


Figure 20. Histograms of the Canberra distance differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

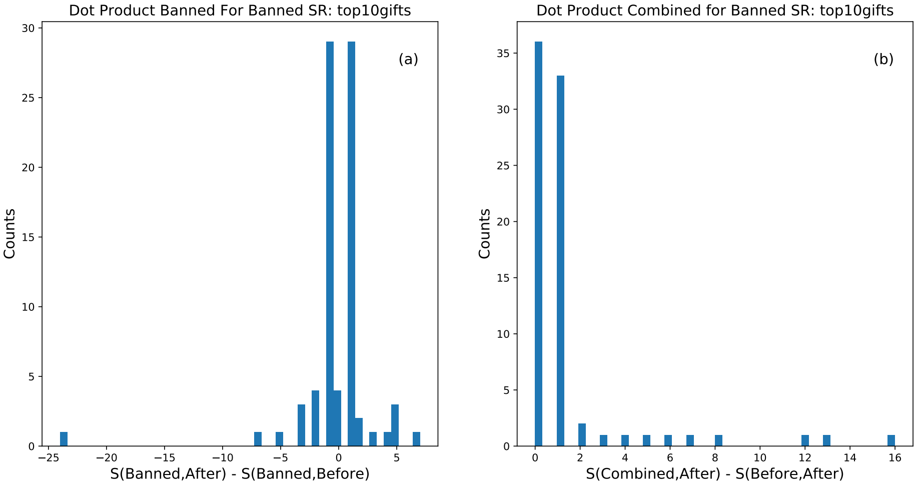


Figure 21. Histograms of the dot product similarity differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

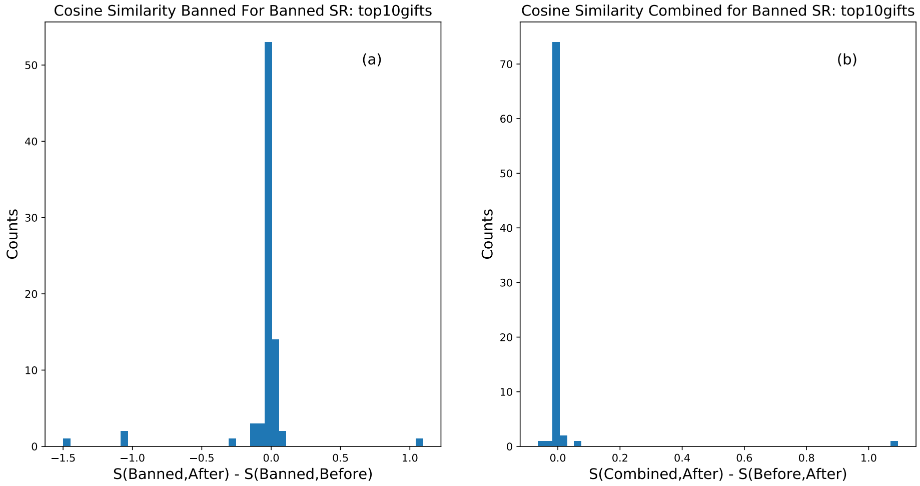


Figure 22. Histograms of the cosine similarity differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

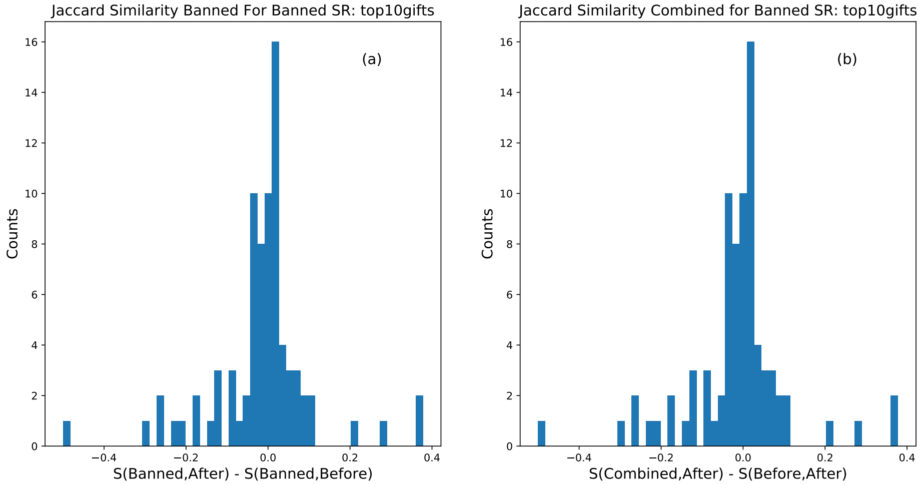


Figure 23. Histograms of the Jaccard similarity differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

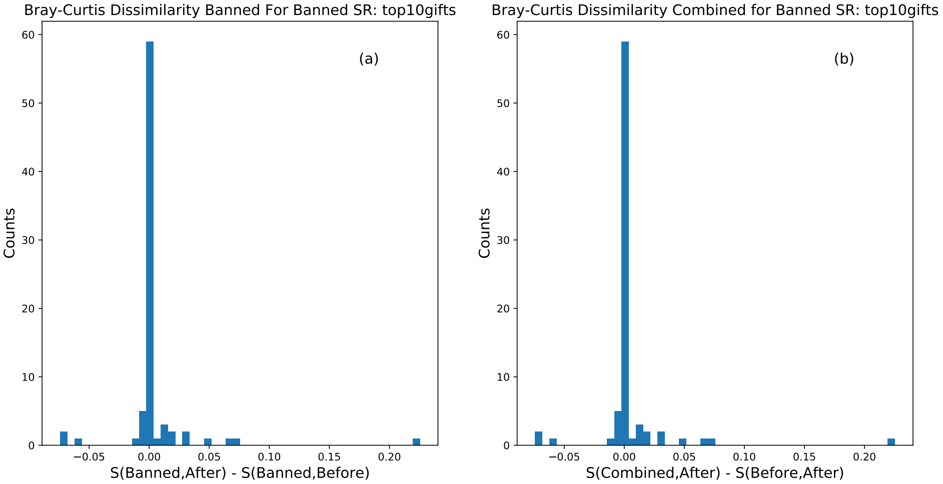


Figure 24. Histograms of the Bray-Curtis dissimilarity differences for the subreddits potentially merged with top10gifts. The activity vectors include all subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

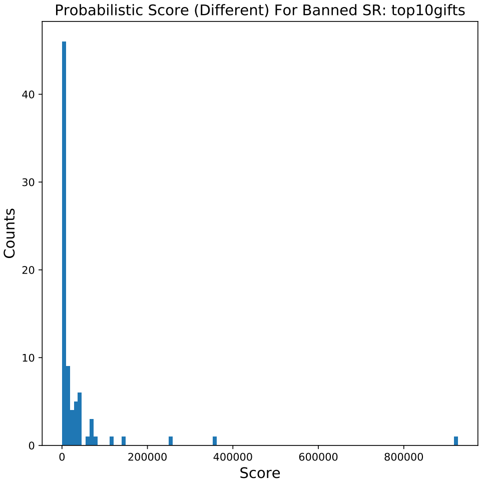


Figure 25. A histogram of the merger log Bayes factors for the subreddits that potentially merged with top10gifts. The activity vectors include all subreddit members.

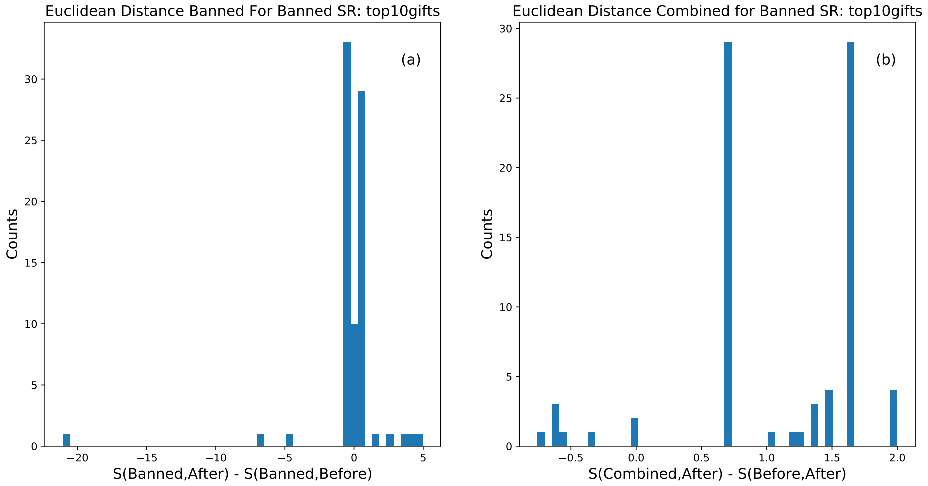


Figure 26. Histograms of the Euclidean distance differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

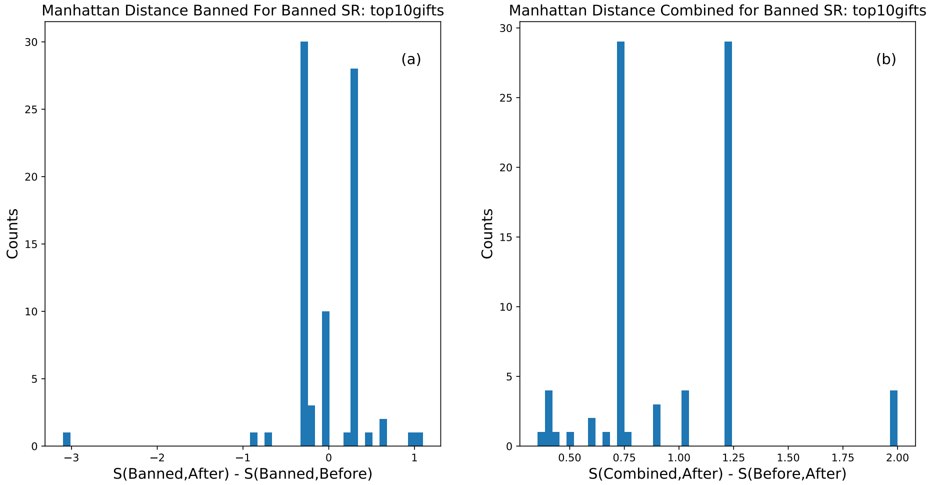


Figure 27. Histograms of the Manhattan distance differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

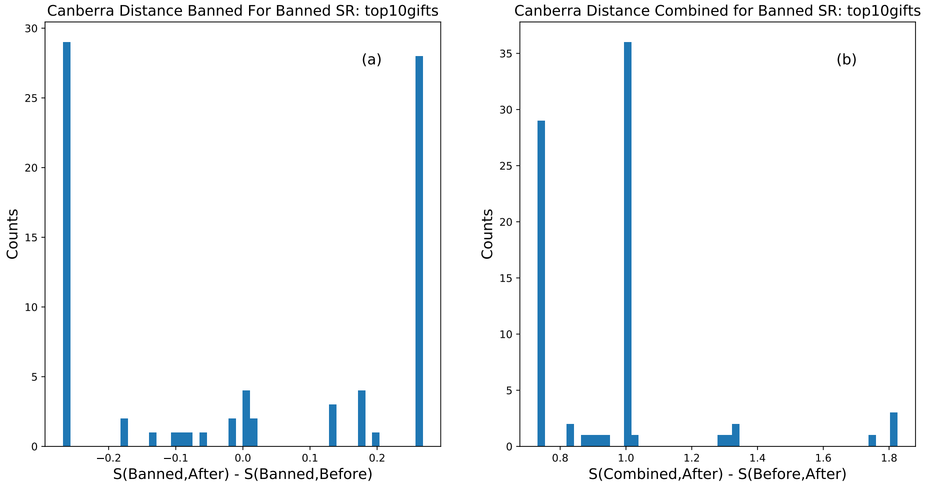


Figure 28. Histograms of the Canberra distance differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

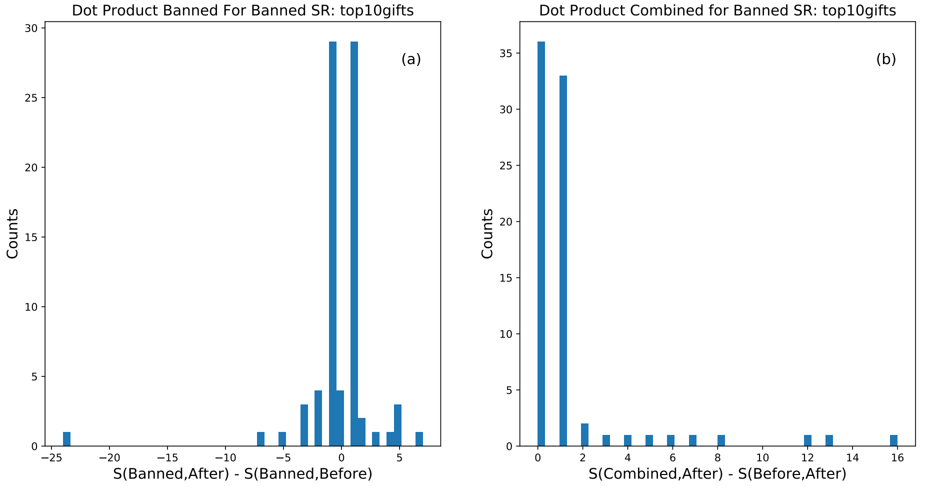


Figure 29. Histograms of the dot product similarity differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

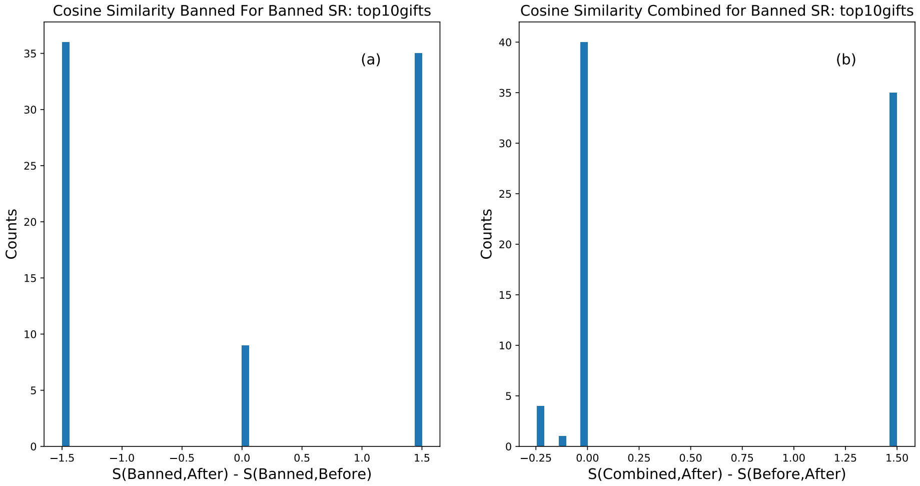


Figure 30. Histograms of the cosine similarity differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

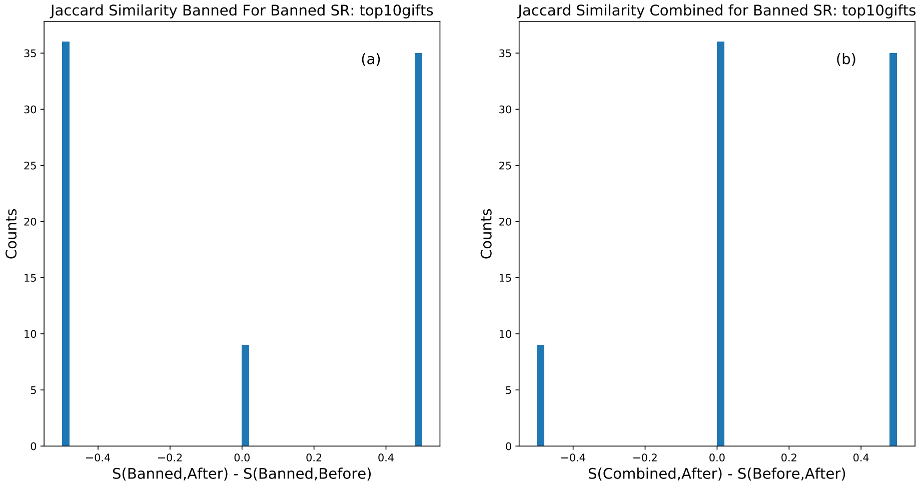


Figure 31. Histograms of the Jaccard similarity differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

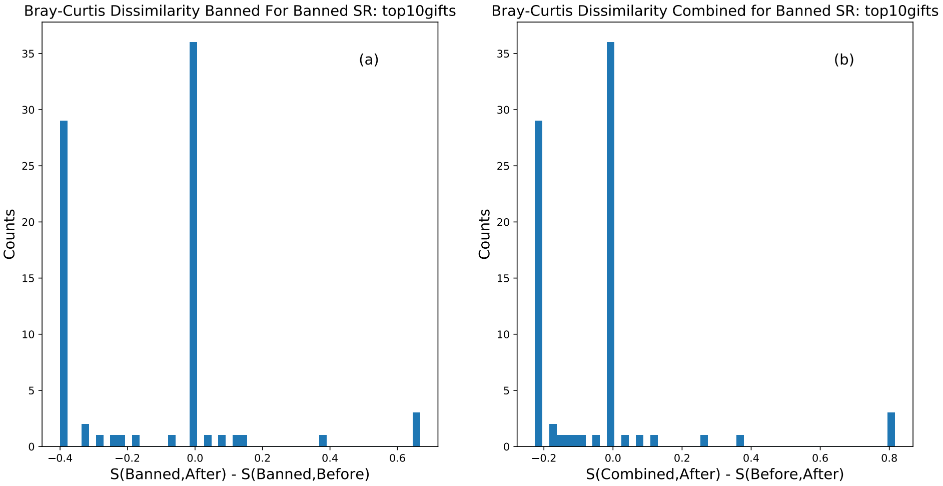


Figure 32. Histograms of the Bray-Curtis dissimilarity differences for the subreddits potentially merged with top10gifts. The activity vectors only include banned subreddit members. Plot (a) is the difference between the 1) distance between the 'banned' and 'before' vectors and 2) the 'banned' and 'after' vectors. Plot (b) is the difference between 3) the distance between the combined activity vector and the 'after' vector and 4) the distance between the 'before' and 'after' vector.

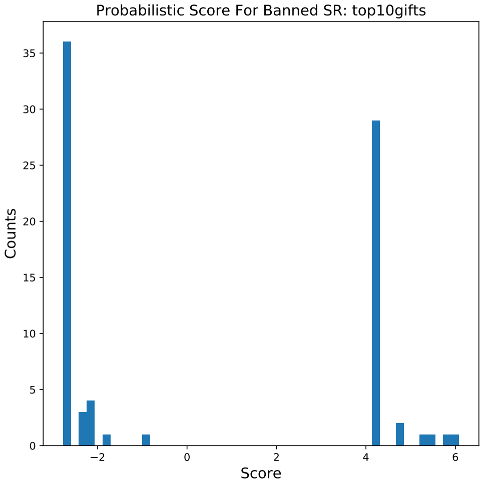


Figure 33. A histogram of the merger log Bayes factors for the subreddits that potentially merged with top10gifts. The activity vectors only include banned subreddit members.

TABLE 4

The Number of Subreddits with Similarity (Distance) Metric Differences Above (Below) a Threshold of Zero for *top10gifts*

Measure	**Counts (out of 80)**			
	All SR users		**Banned SR users**	
	BA - bB	**cA - BA**	**BA - bB**	**cA - BA**
Euclidean Distance	40	6	36	6
Manhattan Distance	30	0	36	0
Canberra Distance	36	0	38	0
Dot Product	37	44	37	44
Cosine Similarity	38	33	35	35
Jaccard Similarity	37	35	35	35
Bray-Curtis Dissimilarity	36	36	38	36
	bBA		**bBA**	
Merger Bayes Factor Score	76		35	

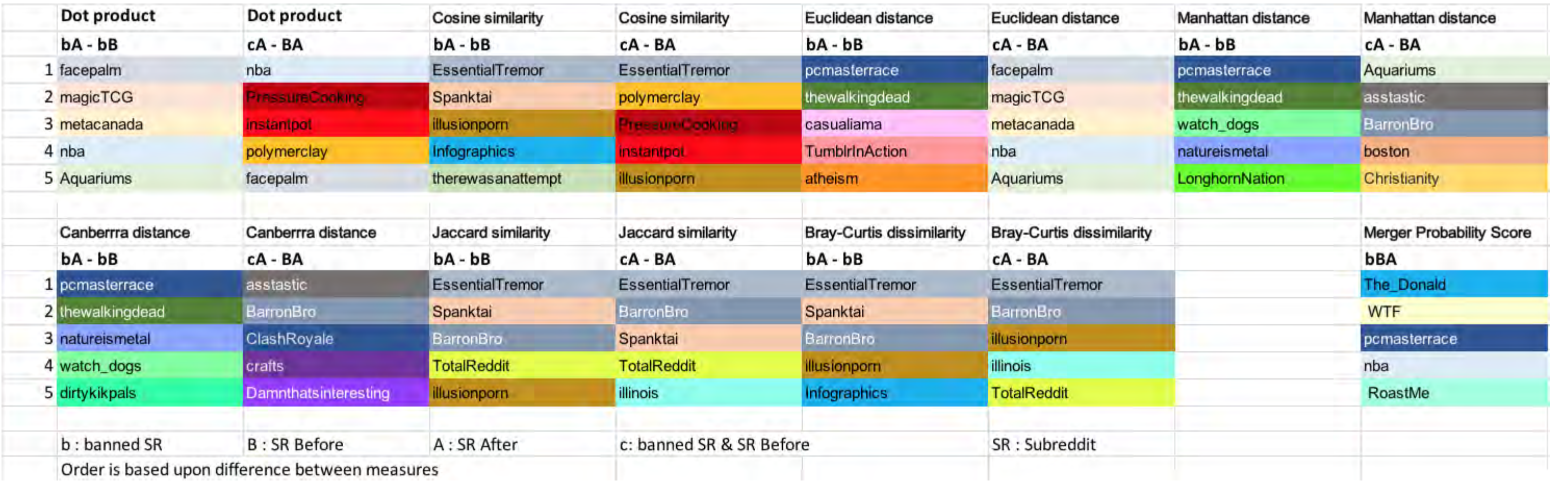
and 29. Table 5 shows the list for the case where all user activity data were included in the potential destination subreddit vectors, and Table 6 shows the results where only the activity data for the banned user list are extracted for the potential destination subreddit vectors.

A comparison between the two tables shows that both dot product calculations produced the same top-five list, although the lists from the banned to before and after evaluation and the combined evaluation are different. These measures are relatively immune to changes in activity levels for other members. The other measures are much more sensitive to the destination subreddit activity for other users. There is also some sensitivity to subreddits with large volumes of user activity, like *nba*, *nfl*, and *The Donald*.

If it is assumed that the top-five lists in Table 6 (where only the users from the banned subreddit are considered) are more meaningful than the lists in Table 5, then examination of Table 6 shows that the Canberra distance lists, the Jaccard similarity lists, and the Bray-Curtis dissimilarity lists are very similar. The subreddits *asstastic*, *BarronBro*, and *boston* rank in similar order for all six lists. The Canberra distance lists and the Bray-Curtis lists are identical, even though one measure relies on activity levels and the other only considers memberships. Only the Canberra distance, the Jaccard distance, and the Bray-Curtis dissimilarity top-five lists are the same between the similarity measure for the banned activity to the before and after activities and the similarity measure for the combined activity and the before activity to the after activity. The other measures have different top-five lists between the two calculations. It should be noted that differences can be found in rank ordering further down the lists beyond what is shown in Table 6, meaning that the tables do not prove that different measures are equivalent.

TABLE 5

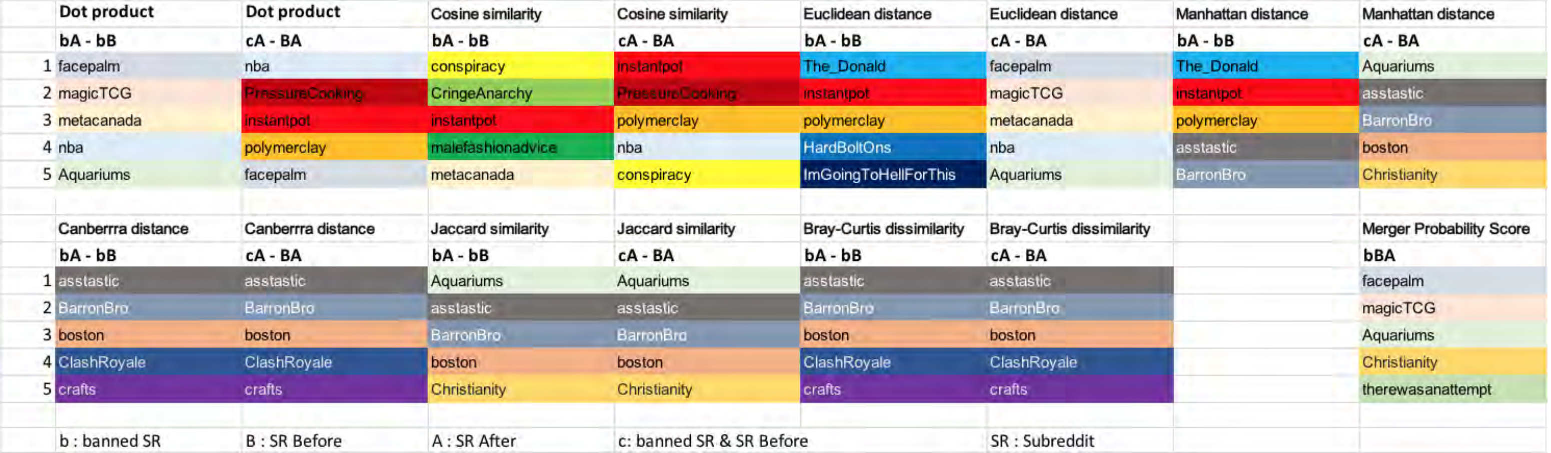
The Top Five Most Likely Destination Subreddits for *top10gifts* from Different Ranking Methods Using All Members



	Dot product	Dot product	Cosine similarity	Cosine similarity	Euclidean distance	Euclidean distance	Manhattan distance	Manhattan distance
	bA - bB	**cA - BA**	**bA - bB**	**cA - BA**	**bA - bB**	**cA - BA**	**bA - bB**	**cA - BA**
1	facepalm	nba	EssentialTremor	EssentialTremor	pcmasterrace	facepalm	pcmasterrace	Aquariums
2	magicTCG	PressureCooking	Spanktai	polymerclay	thewalkingdead	magicTCG	thewalkingdead	asstastic
3	metacanada	instantpot	illusionporn	PressureCooking	casualiama	metacanada	watch_dogs	BarronBro
4	nba	polymerclay	Infographics	instantpot	TumblrInAction	nba	natureismetal	boston
5	Aquariums	facepalm	therewasanattempt	illusionporn	atheism	Aquariums	LonghornNation	Christianity
	Canberrra distance	Canberrra distance	Jaccard similarity	Jaccard similarity	Bray-Curtis dissimilarity	Bray-Curtis dissimilarity		Merger Probability Score
	bA - bB	**cA - BA**	**bA - bB**	**cA - BA**	**bA - bB**	**cA - BA**		**bBA**
1	pcmasterrace	asstastic	EssentialTremor	EssentialTremor	EssentialTremor	EssentialTremor		The_Donald
2	thewalkingdead	BarronBro	Spanktai	BarronBro	Spanktai	BarronBro		WTF
3	natureismetal	ClashRoyale	BarronBro	Spanktai	BarronBro	illusionporn		pcmasterrace
4	watch_dogs	crafts	TotalReddit	TotalReddit	illusionporn	illinois		nba
5	dirtykikpals	Damnthatsinteresting	illusionporn	illinois	Infographics	TotalReddit		RoastMe
	b : banned SR	B : SR Before	A : SR After	c: banned SR & SR Before		SR : Subreddit		
	Order is based upon difference between measures							

TABLE 6

The Top Five Most Likely Destination Subreddits for *top10gifts* from Different Ranking Methods Using Banned Members Only



	Dot product	Dot product	Cosine similarity	Cosine similarity	Euclidean distance	Euclidean distance	Manhattan distance	Manhattan distance
	bA - bB	**cA - BA**	**bA - bB**	**cA - BA**	**bA - bB**	**cA - BA**	**bA - bB**	**cA - BA**
1	facepalm	nba	conspiracy	instantpot	The_Donald	facepalm	The_Donald	Aquariums
2	magicTCG	PressureCooking	CringeAnarchy	PressureCooking	instantpot	magicTCG	instantpot	asstastic
3	metacanada	instantpot	instantpot	polymerclay	polymerclay	metacanada	polymerclay	BarronBro
4	nba	polymerclay	malefashionadvice	nba	HardBoltOns	nba	asstastic	boston
5	Aquariums	facepalm	metacanada	conspiracy	ImGoingToHellForThis	Aquariums	BarronBro	Christianity
	Canberrra distance	Canberrra distance	Jaccard similarity	Jaccard similarity	Bray-Curtis dissimilarity	Bray-Curtis dissimilarity		Merger Probability Score
	bA - bB	**cA - BA**	**bA - bB**	**cA - BA**	**bA - bB**	**cA - BA**		**bBA**
1	asstastic	asstastic	Aquariums	Aquariums	asstastic	asstastic		facepalm
2	BarronBro	BarronBro	asstastic	asstastic	BarronBro	BarronBro		magicTCG
3	boston	boston	BarronBro	BarronBro	boston	boston		Aquariums
4	ClashRoyale	ClashRoyale	boston	boston	ClashRoyale	ClashRoyale		Christianity
5	crafts	crafts	Christianity	Christianity	crafts	crafts		therewasanattempt
	b : banned SR	B : SR Before	A : SR After	c: banned SR & SR Before		SR : Subreddit		

The top-five lists for the cosine and Jaccard similarity measures that are determined from the activity vectors that include only the banned subbreddit members are not very meaningful. As can be seen in Figures 30 and 23, there are large peaks for the maximum values. The top-five list has many ties, and the top five are primarily determined by alphabetical order.

An examination of the top-five lists for the merger Bayes factor score show that the measure is sensitive to the activity levels of users who are not from the banned subreddit. Better results might have been obtained if data were available for the banned subreddits up to the date of the banning. It might also be possible to obtain better results by drawing the 'before' activity data from potential destination subreddits for the week before the banning, with the hope that the subreddit will not have evolved as dramatically in its activity levels. One benefit from the merger Bayes factor score is that it is less sensitive to fluctuations in counts between low user activity levels and high user activity levels, unlike the dot product, cosine similarity, Euclidean distance, and Manhattan distances. In examining Table 6, *facepalm*, *magicTCG*, and *Aquariums* should be examined to see whether users from the banned subreddit have moved their activities to these subreddits.

A more in-depth evaluation of the Reddit merger data has been conducted, but is not presented in this report. The results are that the subreddits that were banned during the period from 1 October 2016 to 31 January 2017 do not appear to have moved their activities to other subreddits. For the most part, it looks like those members that remained active after the banning independently moved to other subreddits to conduct different activities. These more detailed analyses are not presented here because the main purpose of the analysis was to determine whether Bayes factors could be used to detect subreddit mergers, and the length of the report is close to excessive. The results are not conclusive because of the missing weeks of activity data for the banned subreddits. Results might be more encouraging if these data were available. However, it does look like the Bayes factors can find and rank subreddits that may have mergers, as can be seen with Table 2. A consideration for the use of Bayes factors should be whether the system to be evaluated matches the generative merger models in this paper, which include Poisson, binomial, and multinomial distributions. It may be noted that only the Poisson model has been fully developed. Further work would be required to complete software development and validate computational stability for the binomial and multinomial Bayes factors.

7. SUMMARY

In summary, this report documents the analytical derivation of Bayes factors for a number of different models associated with detecting similarity, mergers, and splits between groups. Although some of the equations for probabilistic similarity measures are available in the literature, the detailed derivations that describe those equations are often extremely difficult to find. This technical report attempts to document these derivations so that in the future researchers will not have to take the equations on faith or spend valuable time figuring out how to derive the equations for themselves. In addition, this technical report derives Bayes factors for detecting mergers and splits between datasets. No descriptions of these factors have been found in the open literature and the derivations are believed to be original. The detailed derivations are again provided so that researchers do not have to repeat the work. In the unfortunate case where errors may have occurred in the derivations or typesetting, researchers will have an easier time identifying and correcting those errors.

This page intentionally left blank.

APPENDIX A: DERIVATION OF BAYES FACTORS TO TEST THE SIMILARITY OF DISTRIBUTIONS

This section contains the detailed derivations of Bayes factors that test if two samples are drawn from the same or different probability distribution functions.

A.1 POISSON SIMILARITIES

The posterior probability for two samples of counts, n_A and n_B, being generated by the same Poisson generator is

$$p\left(H_0|n_A,n_B\right)=\frac{1}{Z}\int_0^\infty p\left(n_A|\lambda\right)p\left(n_B|\lambda\right)p\left(\lambda|\alpha,\beta\right)d\lambda. \tag{A.1}$$

It is assumed that the time periods for collecting the counts are the same. The posterior probability for the case where two samples are generated by different Poisson generators is

$$p\left(H_1|n_A,n_B\right)=\frac{1}{Z}\int_0^\infty p\left(n_A|\lambda_A\right)p\left(\lambda_A|\alpha,\beta\right)d\lambda_A\int_0^\infty p\left(n_B|\lambda_B\right)p\left(\lambda_B|\alpha,\beta\right)d\lambda_B. \tag{A.2}$$

The parameter Z is a normalization term that can be neglected for the Bayes factor because it will cancel. The Bayes factor is

$$R=\frac{\int_0^\infty p\left(n_A|\lambda\right)p\left(n_B|\lambda\right)p\left(\lambda|\alpha,\beta\right)d\lambda}{\int_0^\infty p\left(n_A|\lambda_A\right)p\left(\lambda_A|\alpha,\beta\right)d\lambda_A\int_0^\infty p\left(n_B|\lambda_B\right)p\left(\lambda_B|\alpha,\beta\right)d\lambda_B}. \tag{A.3}$$

The integral in the numerator expands to

$$p\left(H_0|n_A,n_B\right)=\frac{1}{Z}\int_0^\infty \frac{\lambda^{n_A}e^{-\lambda}}{\Gamma\left(n_A+1\right)}\frac{\lambda^{n_B}e^{-\lambda}}{\Gamma\left(n_B+1\right)}\frac{\beta^\alpha\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma\left(\alpha\right)}d\lambda, \tag{A.4}$$

where the Poisson distribution functions and the gamma distribution function have been inserted.

Terms are collected together to give

$$p\left(H_0|n_A,n_B\right)=\frac{\beta^\alpha}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\Gamma\left(\alpha\right)}\int_0^\infty \lambda^{n_A+n_B+\alpha-1}e^{-(2+\beta)\lambda}d\lambda. \tag{A.5}$$

Use Gradshteyn and Ryzhik integral 3.381.4 [13]

$$\int_0^\infty x^{\nu-1}e^{-\mu x}dx=\mu^{-\nu}\Gamma\left(\nu\right), \tag{A.6}$$

to get

$$p\left(H_0|n_A,n_B\right)=\frac{\beta^\alpha\left(2+\beta\right)^{-(n_A+n_B+\alpha)}\Gamma\left(n_A+n_B+\alpha\right)}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\Gamma\left(\alpha\right)}. \tag{A.7}$$

The probability for two samples being generated from two different Poisson distributions is

$$p\left(H_1|n_A,n_B\right)=\frac{1}{Z}\int_0^\infty p\left(n_A|\lambda_A\right)p\left(\lambda_A|\alpha,\beta\right)d\lambda_A\int_0^\infty p\left(n_B|\lambda_B\right)p\left(\lambda_B|\alpha,\beta\right)d\lambda_B. \tag{A.8}$$

It is possible to keep the prior probability distribution parameters unique, but most applications of the equation will adopt the assumption that the prior knowledge about the two distributions is the same. Inserting the probability density functions results in

$$p\left(H_1|n_A,n_B\right)=\frac{1}{Z}\int_0^\infty\frac{\lambda_A^{n_A}e^{-\lambda_A}}{\Gamma\left(n_A+1\right)}\frac{\beta^\alpha\lambda_A^{\alpha-1}e^{-\beta\lambda_A}}{\Gamma\left(\alpha\right)}d\lambda_A\int_0^\infty\frac{\lambda_B^{n_B}e^{-\lambda_B}}{\Gamma\left(n_B+1\right)}\frac{\beta^\alpha\lambda_B^{\alpha-1}e^{-\beta\lambda_B}}{\Gamma\left(\alpha\right)}d\lambda_B. \quad \text{(A.9)}$$

Collecting terms leads to

$$\begin{aligned}p\left(H_1|n_A,n_B\right)=&\frac{\beta^{2\alpha}}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\left(\Gamma\left(\alpha\right)\right)^2}\int_0^\infty\lambda_A^{n_A+\alpha-1}e^{-(1+\beta)\lambda_A}d\lambda_A\times\\&\int_0^\infty\lambda_B^{n_B+\alpha-1}e^{-(1+\beta)\lambda_B}d\lambda_B.\end{aligned} \quad \text{(A.10)}$$

Again, use Gradshteyn and Ryzhik integral 3.381.4 [13] to get

$$p\left(H_1|n_A,n_B\right)=\frac{\beta^{2\alpha}\left(1+\beta\right)^{-\left(n_A+\alpha\right)}\Gamma\left(n_A+\alpha\right)\left(1+\beta\right)^{-\left(n_B+\alpha\right)}\Gamma\left(n_B+\alpha\right)}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\left(\Gamma\left(\alpha\right)\right)^2}. \quad \text{(A.11)}$$

Collect terms to get

$$p\left(H_1|n_A,n_B\right)=\frac{\beta^{2\alpha}\left(1+\beta\right)^{-\left(n_A+n_B+2\alpha\right)}\Gamma\left(n_A+\alpha\right)\Gamma\left(n_B+\alpha\right)}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\left(\Gamma\left(\alpha\right)\right)^2}. \quad \text{(A.12)}$$

The ratio function of Equation A.3 is then

$$R=\frac{\frac{\beta^\alpha(2+\beta)^{-\left(n_A+n_B+\alpha\right)}\Gamma(n_A+n_B+\alpha)}{Z\Gamma(n_A+1)\Gamma(n_B+1)\Gamma(\alpha)}}{\frac{\beta^{2\alpha}(1+\beta)^{-\left(n_A+n_B+2\alpha\right)}\Gamma(n_A+\alpha)\Gamma(n_B+\alpha)}{Z\Gamma(n_A+1)\Gamma(n_B+1)(\Gamma(\alpha))^2}}. \quad \text{(A.13)}$$

Cancel common terms in numerators and denominators to get

$$R=\frac{\left(2+\beta\right)^{-\left(n_A+n_B+\alpha\right)}\Gamma\left(n_A+n_B+\alpha\right)\Gamma\left(\alpha\right)}{\beta^\alpha\left(1+\beta\right)^{-\left(n_A+n_B+2\alpha\right)}\Gamma\left(n_A+\alpha\right)\Gamma\left(n_B+\alpha\right)}. \quad \text{(A.14)}$$

Invert negative exponents to get the final equation,

$$R=\frac{\left(1+\beta\right)^{\left(n_A+n_B+2\alpha\right)}\Gamma\left(n_A+n_B+\alpha\right)\Gamma\left(\alpha\right)}{\beta^\alpha\left(2+\beta\right)^{\left(n_A+n_B+\alpha\right)}\Gamma\left(n_A+\alpha\right)\Gamma\left(n_B+\alpha\right)}. \quad \text{(A.15)}$$

A.1.1 Poisson Likelihood Ratios for Different Time Periods

If the time periods are different for collecting the samples, the mathematics becomes just slightly more complicated. The rate λ can be defined as the expected number of counts in a given time period. The nature of the Poisson distribution means that the variable λ can be scaled by the ratio between the time span of the collected sample and a reference time span.

$$p\left(n|r\lambda\right) = \frac{\left(r\lambda\right)^{n} e^{-r\lambda}}{\Gamma\left(n+1\right)}. \tag{A.16}$$

The conjugate prior parameters α and β are defined with respect to the reference time period. The integrals in Section A.1 can be repeated with a scale factor r for the time periods. The integral from Equation A.1 is revised to

$$p\left(H_0|n_A, n_B\right) = \frac{\beta^{\alpha}\left(r_A + r_B + \beta\right)^{-(n_A+n_B+\alpha)}\Gamma\left(n_A+n_B+\alpha\right)}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\Gamma\left(\alpha\right)} \tag{A.17}$$

and the integral in Equation A.12 to

$$p\left(H_1|n_A, n_B\right) = \frac{\beta^{2\alpha}\left(r_A+\beta\right)^{-(n_A+\alpha)}\left(r_B+\beta\right)^{-(n_B+\alpha)}\Gamma\left(n_A+\alpha\right)\Gamma\left(n_B+\alpha\right)}{Z\Gamma\left(n_A+1\right)\Gamma\left(n_B+1\right)\left(\Gamma\left(\alpha\right)\right)^2}. \tag{A.18}$$

The ratio in Equation A.15 is revised to

$$R = \frac{\left(r_A+\beta\right)^{(n_A+\alpha)}\left(r_B+\beta\right)^{(n_B+\alpha)}\Gamma\left(n_A+n_B+\alpha\right)\Gamma\left(\alpha\right)}{\beta^{\alpha}\left(r_A+r_B+\beta\right)^{(n_A+n_B+\alpha)}\Gamma\left(n_A+\alpha\right)\Gamma\left(n_B+\alpha\right)}. \tag{A.19}$$

A.2 BINOMIAL AND MULTINOMIAL SIMILARITIES

The derivation for both the binomial and multinomial simiilarity measures are related, so that derivation for the multinomial distribution naturally produces an equation for the binomial distribution. As before, the two hypotheses are that the two data samples are from the same generator or from two different generators. The probability for a single generator producing two samples is given by the integral

$$p\left(H_0|\boldsymbol{d}_A, \boldsymbol{d}_B\right) = \frac{1}{Z}\int p\left(\boldsymbol{d}_A|\boldsymbol{q}\right)p\left(\boldsymbol{d}_B|\boldsymbol{q}\right)p\left(\boldsymbol{q}|\boldsymbol{\alpha}\right)d\Theta, \tag{A.20}$$

where $\boldsymbol{d}$ are the vectors of counts for the two samples, Z is a normalization term for the integrals for the two hypotheses, and Θ is the probability simplex for the probability vector $\boldsymbol{q}$. The vector $\boldsymbol{\alpha}$ is the parameter vector for the prior probability function.

The integral for the second hypothesis with two generators is

$$p\left(H_1|\boldsymbol{d}_A, \boldsymbol{d}_B\right) = \frac{1}{Z}\int p\left(\boldsymbol{d}_A|\boldsymbol{q}_A\right)p\left(\boldsymbol{q}_A|\boldsymbol{\alpha}\right)d\Theta_A\int p\left(\boldsymbol{d}_B|\boldsymbol{q}_B\right)p\left(\boldsymbol{q}_B|\boldsymbol{\alpha}\right)d\Theta_B. \tag{A.21}$$

The normalization term can be neglected because we will be interested in a ratio test and the term will cancel. The Bayes factor can then be written as

$$R = \frac{\int p\left(\boldsymbol{d}_A|\boldsymbol{q}\right)p\left(\boldsymbol{d}_B|\boldsymbol{q}\right)p\left(\boldsymbol{q}|\boldsymbol{\alpha}\right)d\Theta}{\int p\left(\boldsymbol{d}_A|\boldsymbol{q}_A\right)p\left(\boldsymbol{q}_A|\boldsymbol{\alpha}\right)d\Theta_A\int p\left(\boldsymbol{d}_B|\boldsymbol{q}_B\right)p\left(\boldsymbol{q}_B|\boldsymbol{\alpha}\right)d\Theta_B}. \tag{A.22}$$

As noted previously, the equation can be enhanced to include unique prior probabilities for the two hypotheses, but this extension will not be derived.

The integral for the numerator will be derived in detail. The solutions to the two integrals in the denominator follow directly from the first integral.

$$\begin{gathered}I_0 = \int \left[\frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}\right)+1\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ai}+1\right)} \prod_{i=1}^{k} \boldsymbol{q}_i^{\boldsymbol{d}_{Ai}} \right] \left[\frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi}\right)+1\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi}+1\right)} \prod_{i=1}^{k} \boldsymbol{q}_i^{\boldsymbol{d}_{Bi}} \right] \times \\ \left[\frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_i\right)} \prod_{i=1}^{k} \boldsymbol{q}_i^{\boldsymbol{\alpha}_i-1} \right] d\Theta.\end{gathered} \tag{A.23}$$

The Γ functions can be moved outside the integral and the exponents collected. Details on the simplex integral limits Θ can be expanded now.

$$\begin{gathered}I_0 = \frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}\right)+1\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ai}+1\right)} \frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi}\right)+1\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi}+1\right)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k-1} \Gamma\left(\boldsymbol{\alpha}_i\right)} \times \\ \int_0^1 \int_0^{1-\boldsymbol{q}_1} \cdots \int_0^{1-\sum_{i=1}^{k-2} \boldsymbol{q}_i} \prod_{i=1}^{k-1} \boldsymbol{q}_i^{\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i-1} \boldsymbol{q}_k^{\boldsymbol{d}_{Ak}+\boldsymbol{d}_{Bk}+\boldsymbol{\alpha}_k-1} d\boldsymbol{q}_1 d\boldsymbol{q}_2 \cdots d\boldsymbol{q}_{k-1}.\end{gathered} \tag{A.24}$$

The sum of $\boldsymbol{q}_i$ must equal 1, so the final dimension $\boldsymbol{q}_k$ has a single fixed value equal to $1 - \sum_{i=1}^{k-1} \boldsymbol{q}_i$ and is effectively a delta function at this value. The Γ functions will be ignored for now as we work through the integrals. Integration over the probability simplex space Θ is then a series of upper limits on each term that is one minus the values of the previously listed simplex dimensions.

$$\begin{gathered}I_{0.1} = \int_0^1 \int_0^{1-\boldsymbol{q}_1} \cdots \int_0^{1-\sum_{i=1}^{k-3} \boldsymbol{q}_i} \prod_{i=1}^{k-2} \boldsymbol{q}_i^{\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i-1} \times \\ \int_0^{1-\sum_{i=1}^{k-2} \boldsymbol{q}_i} \boldsymbol{q}_{k-1}^{\boldsymbol{d}_{A,k-1}+\boldsymbol{d}_{B,k-1}+\boldsymbol{\alpha}_{k-1}-1} \left(1 - \sum_{i=1}^{k-2} \boldsymbol{q}_i - \boldsymbol{q}_{k-1}\right)^{\boldsymbol{d}_{A,k}+\boldsymbol{d}_{B,k}+\boldsymbol{\alpha}_k-1} d\boldsymbol{q}_1 d\boldsymbol{q}_2 \cdots d\boldsymbol{q}_{k-2} d\boldsymbol{q}_{k-1}.\end{gathered} \tag{A.25}$$

The analytic solution to the innermost integral is 3.191.1 in Gradshteyn and Ryzhik [13]:

$$\int_0^u x^{\nu-1} \left(u-x\right)^{\mu-1} dx = u^{\mu+\nu-1} \mathrm{B}\left(\mu,\nu\right); \quad \left[\mathrm{Re}\mu > 0, \mathrm{Re}\nu > 0\right], \tag{A.26}$$

where B is the beta function,

$$\mathrm{B}\left(\mu,\nu\right) = \frac{\Gamma\left(\mu\right)\Gamma\left(\nu\right)}{\Gamma\left(\mu+\nu\right)}. \tag{A.27}$$

The solution to the innermost integral results in

$$\begin{gathered}I_{0.1} = \int_0^1 \int_0^{1-\boldsymbol{q}_1} \cdots \int_0^{1-\sum_{i=1}^{k-3} \boldsymbol{q}_i} \prod_{i=1}^{k-2} \boldsymbol{q}_i^{\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i-1} \left(1 - \sum_{i=1}^{k-2} \boldsymbol{q}_i\right)^{\left(\sum_{i=k-1}^{k} \boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i\right)-1} \times \\ \mathrm{B}\left(\boldsymbol{d}_{A,k}+\boldsymbol{d}_{B,k}+\boldsymbol{\alpha}_k, \boldsymbol{d}_{A,k-1}+\boldsymbol{d}_{B,k-1}+\boldsymbol{\alpha}_{k-1}\right) d\boldsymbol{q}_1 d\boldsymbol{q}_2 \cdots d\boldsymbol{q}_{k-2}.\end{gathered} \tag{A.28}$$

The $\boldsymbol{q}_{k-2}$ terms can be explicitly extracted from the summations to perform the next integral, which is of a similar form to the integral in A.25,

$$\begin{aligned} I_{0.1} =& \mathrm{B}\left(\boldsymbol{d}_{A,k}+\boldsymbol{d}_{B,k}+\boldsymbol{\alpha}_k, \boldsymbol{d}_{A,k-1}+\boldsymbol{d}_{B,k-1}+\boldsymbol{\alpha}_{k-1}\right) \int_0^1 \int_0^{1-\boldsymbol{q}_1} \cdots \int_0^{1-\sum_{i=1}^{k-3}\boldsymbol{q}_i} \prod_{i=1}^{k-3} \boldsymbol{q}_i^{\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i-1} \times \\ & \boldsymbol{q}_{k-2}^{\boldsymbol{d}_{A,k-2}+\boldsymbol{d}_{B,k-2}+\boldsymbol{\alpha}_{k-2}-1} \left(1-\sum_{i=1}^{k-3}\boldsymbol{q}_i-\boldsymbol{q}_{k-2}\right)^{\left(\sum_{i=k-1}^{k}\boldsymbol{d}_{A,i}+\boldsymbol{d}_{B,i}+\boldsymbol{\alpha}_i\right)-1} \\ & d\boldsymbol{q}_1 d\boldsymbol{q}_2 \cdots d\boldsymbol{q}_{k-2}. \end{aligned} \tag{A.29}$$

This integral produces another beta function with the first term being a summation over the counts and prior parameters from the probability dimensions that have already been integrated and the second term being the sum of the count and prior parameter for the most recently integrated dimension.

$$\begin{aligned} I_{0.1} =& \mathrm{B}\left(\boldsymbol{d}_{A,k}+\boldsymbol{d}_{B,k}+\boldsymbol{\alpha}_k, \boldsymbol{d}_{A,k-1}+\boldsymbol{d}_{B,k-1}+\boldsymbol{\alpha}_{k-1}\right) \times \\ & \mathrm{B}\left(\left(\sum_{i=k-1}^{k}\boldsymbol{d}_{A,i}+\boldsymbol{d}_{B,i}+\boldsymbol{\alpha}_i\right), \boldsymbol{d}_{A,k-2}+\boldsymbol{d}_{B,k-2}+\boldsymbol{\alpha}_{k-2}\right) \\ & \int_0^1 \int_0^{1-\boldsymbol{q}_1} \cdots \int_0^{1-\sum_{i=1}^{k-4}\boldsymbol{q}_i} \prod_{i=1}^{k-4} \boldsymbol{q}_i^{\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i-1} \times \boldsymbol{q}_{k-3}^{\boldsymbol{d}_{A,k-3}+\boldsymbol{d}_{B,k-3}+\boldsymbol{\alpha}_{k-3}-1}. \\ & \left(1-\sum_{i=1}^{k-4}\boldsymbol{q}_i-\boldsymbol{q}_{k-3}\right)^{\left(\sum_{i=k-2}^{k}\boldsymbol{d}_{A,i}+\boldsymbol{d}_{B,i}+\boldsymbol{\alpha}_i\right)-1} d\boldsymbol{q}_1 d\boldsymbol{q}_2 \cdots d\boldsymbol{q}_{k-3}. \end{aligned} \tag{A.30}$$

The remaining integrals produce additional beta functions with the same general form so that full integral results in

$$I_{0.1} = \prod_{i=1}^{k-1} \mathrm{B}\left(\sum_{j=i+1}^{k}\left(d_{A,j}+\boldsymbol{d}_{B,j}+\boldsymbol{\alpha}_j\right), \boldsymbol{d}_{A,i}+\boldsymbol{d}_{B,i}+\boldsymbol{\alpha}_i\right). \tag{A.31}$$

Various pairs of Γ functions cancel when the product of beta functions are expanded into Γ functions, resulting in

$$I_{0.1} = \frac{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{d}_{A,i}+\boldsymbol{d}_{B,i}+\boldsymbol{\alpha}_i\right)}{\Gamma\left(\sum_{i=1}^{k}\boldsymbol{d}_{A,i}+\boldsymbol{d}_{B,i}+\boldsymbol{\alpha}_i\right)}. \tag{A.32}$$

The integrals for the other hypothesis of separate generators produce similar equations, except that it is a product of two independent terms, one for each dataset. It is possible to have different prior parameters $\boldsymbol{\alpha}$ for each integral, but it has been assumed here that the same prior parameters are used for each integral. The ratio then becomes

$$R = \frac{\frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}\right)+1\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Ai}+1)} \frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi}\right)+1\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Bi}+1)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)} \frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i\right)}}{\frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}\right)+1\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{A,i}+1\right)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)} \frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Ai}+\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{A,i}+\boldsymbol{\alpha}_i\right)} \times \frac{\Gamma\left(\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi}\right)+1\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Bi}+1)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)} \frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i\right)}}. \tag{A.33}$$

Cancel common terms in numerators and denominators to get

$$R = \frac{\frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}+\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i\right)}}{\frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Ai}+\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai}+\boldsymbol{\alpha}_i\right)} \frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_i\right)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha}_i)}}. \tag{A.34}$$

Reduce to a single division to get

$$R = \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ai} + \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_i\right) \Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai} + \boldsymbol{\alpha}_i\right) \Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_i\right) \prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_i\right)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{d}_{Ai} + \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_i\right) \prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ai} + \boldsymbol{\alpha}_i\right) \prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_i\right) \Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_i\right)}. \tag{A.35}$$

If the beta function can be generalized to the function

$$\mathrm{B}'\left(\boldsymbol{\mu}\right) = \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\mu}_i\right)}{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\mu}_i\right)}, \tag{A.36}$$

then the Bayes factor can be written as

$$R = \frac{\mathrm{B}'\left(\boldsymbol{d}_A + \boldsymbol{d}_B + \boldsymbol{\alpha}\right) \mathrm{B}'\left(\boldsymbol{\alpha}\right)}{\mathrm{B}'\left(\boldsymbol{d}_A + \boldsymbol{\alpha}\right) \mathrm{B}'\left(\boldsymbol{d}_B + \boldsymbol{\alpha}\right)}, \tag{A.37}$$

where the variables within the beta functions are assumed to be similarly sized vectors. The Bayes factor for the binomial distribution is the specific case where the number of categories k is 2.

APPENDIX B: SPLIT/MERGER BAYES FACTORS DERIVATIONS

This section contains detailed derivations for the Bayes factors that can be used to determine whether samples may be drawn from merged or split distribution functions.

The specific problem that is examined is the case that there are two groups with activity levels that can be interpreted as having been generated by combined or individual generative models. The scenario is that at a moment in time, one of the groups ceases to exist, as indicated by the subscript B for 'banned.' The second group that the banned group may have potentially merged with is indicated by the subscript S. After the potential merger, the second group is represented by the subscript R. The samples for the post-ban group are either drawn from a generator for the original group S, or are drawn from a combination of the generators for B and S. The specific form of the resultant generator depends on the statistical functions selected to represent the distribution of measurements.

The main question is whether the activity of the first group has moved to another group. The measurements for the banned group, potential source group, and the resultant group are d_B, d_S, and d_R. The generators for the banned and potential source groups are represented by G_B and G_S, where the actual parameters depend upon the nature of the generative model. The resultant generator is represented as G_B & G_S if the groups have merged, and by G_S if the banned group no longer exists. Note that the & sign does not indicate a specific mathematical operation, but any general operation that combines the two generators into one. The nature of this operation depends on the statistical functions selected to model the generators.

The simplest model for detecting a merger is the following case: If the groups have merged, then the generative model should be a combination of the two generators before the two mergers. If not, the generator should be unchanged from the previous sample period. As a statistical test, we're looking for the Bayes factor,

$$R = \frac{p\left(d_B, d_S, d_R | G_B, G_S, G_B \text{ \& } G_S\right) p\left(G_B, G_S, G_B \text{ \& } G_S\right)}{p\left(d_B, d_S, d_R | G_B, G_S, G_S\right) p\left(G_B, G_S, G_S\right)}, \tag{B.1}$$

where prior probabilities for the two hypotheses are included in the factor. A merger is more likely if $R > 1.0$.

The use of Equation B.1 for the cases where there are no measurements can result in what may be interpreted as undesirable decisions, although the prior probabilities can result in a tie decision if no data are available. A more complicated model was developed to provide a way to obtain more reasonable results for the limit of no collected measurements. This model is one where the generators can either remain the same or change over time. There are then four probabilities to calculate: two that contribute to the numerator and two that contribute to the denominator of the Bayes factor. These two models are described in more detail in Section 4. For both models, it is assumed that the generator of the banned group does not change. Otherwise this group cannot be detected based on its activity patterns. The probability that the other group's activity pattern has changed must not be equal to 1 because the merger or split would then be impossible to detect.

The more complex model contains four possible hypotheses: 1) there is a merger and the destination generator has not changed, 2) there is a merger and the destination generator has changed, 3) there is no merger and the destination generator is unchanged, and 4) there is no merger and the destination generator has changed. It is easiest to define the numerator and denominator terms for the Bayes factor as

$$\begin{aligned} R_N &= p\left(d_B, d_S, d_R | G_B, G_S, G_B \text{ \& } G_S\right) p\left(G_B, G_S, G_B \text{ \& } G_S\right) \\ &\quad + p\left(G_B, G_S, G_B \text{ \& } G'_S | d_B, d_S, d_R\right) p\left(G_B, G_S, G_B \text{ \& } G'_S\right) \qquad \text{(B.2)} \\ R_D &= p\left(d_B, d_S, d_R | G_B, G_S, G_S\right) p\left(G_B, G_S, G_S\right) \\ &\quad + p\left(d_B, d_S, d_R | G_B, G_S, G'_S\right) p\left(G_B, G_S, G'_S\right). \qquad \text{(B.3)} \end{aligned}$$

The prime on the generator G'_S indicates that the generator has changed after the potential merger. The Bayes factor for merged or not merged with the four different hypotheses is

$$R = \frac{R_N}{R_D}. \qquad \text{(B.4)}$$

The specific nature of the probability terms is dependent upon the probability density functions chosen for the generators. The ratio is also sensitive to the prior probabilities for the four hypotheses. The following sections derive Bayes factors for the Poisson, binomial, and multinomial distributions. The binomial distribution can be considered a special case of the multinomial distribution. In some cases, we provide independent derivations for the two and in other cases, we derive the multinomial case and then obtain the binomial Bayes factor from the multinomial Bayes factor.

B.1 POISSON MERGERS AND SPLITS

The Poisson distribution function is given by Equation 1. The conjugate prior distribution for the Poisson function is the Dirichlet distribution function (Equation 2). The Poisson distribution function has the useful property that a merger of two Poisson functions is a Poisson function with a parameter that is the sum of the parameters from the individual functions,

$$\lambda_M = \lambda_B + \lambda_S. \qquad \text{(B.5)}$$

The probability for a merger when the generators are constant through time is given by the formula

$$p_{MS} = \frac{1}{Z} \int_0^\infty P\left(d_B | \lambda_B\right) P\left(d_S | \lambda_S\right) P\left(d_R | \lambda_B + \lambda_S\right) P\left(\lambda_B | \alpha_B, \beta_B\right) P\left(\lambda_S | \alpha_S, \beta_S\right) d\lambda_B d\lambda_S. \qquad \text{(B.6)}$$

The probability of no merger is

$$p_{\bar{M}S} = \frac{1}{Z} \int_0^\infty P\left(d_B | \lambda_B\right) P\left(d_S | \lambda_S\right) P\left(d_R | \lambda_S\right) P\left(\lambda_B | \alpha_B, \beta_B\right) P\left(\lambda_S | \alpha_S, \beta_S\right) d\lambda_B d\lambda_S. \qquad \text{(B.7)}$$

For the case with the possibility of the destination generator changing, the probability of merger with a changed generator is

$$\begin{aligned} p_{M\bar{S}} = \frac{1}{Z} \int_0^\infty & P(d_B | \lambda_B) P\left(d_S | \lambda_S\right) P\left(d_R | \lambda_B + \lambda_R\right) \times \\ & P\left(\lambda_B | \alpha_B, \beta_B\right) P\left(\lambda_S | \alpha_S, \beta_S\right) P\left(\lambda_R | \alpha_R, \beta_R\right) d\lambda_B d\lambda_S d\lambda_R. \end{aligned} \qquad \text{(B.8)}$$

The probability of no merger with a changed generator is

$$\begin{aligned} p_{\bar{M}\bar{S}} = \frac{1}{Z}\int_0^\infty & P\left(d_B|\lambda_B\right) P\left(d_S|\lambda_S\right) P\left(d_R|\lambda_R\right) \times \\ & P\left(\lambda_B|\alpha_B,\beta_B\right) P\left(\lambda_S|\alpha_S,\beta_S\right) P\left(\lambda_R|\alpha_R,\beta_R\right) d\lambda_B d\lambda_S d\lambda_R. \end{aligned} \tag{B.9}$$

B.1.1 The Probability of Merger with Consistent Generators

The probability for merger with consistent generators is given by Equation B.6. The insertion of the Poisson probabilities and conjugate prior functions produces

$$\begin{aligned} p_{MS} = \frac{1}{Z}\int_0^\infty & \frac{\lambda_B^{d_B} e^{-\lambda_B}}{\Gamma\left(d_B+1\right)} \frac{\lambda_S^{d_S} e^{-\lambda_S}}{\Gamma\left(d_S+1\right)} \frac{\left(\lambda_B+\lambda_S\right)^{d_R} e^{-\left(\lambda_B+\lambda_S\right)}}{\Gamma\left(d_R+1\right)} \times \\ & \frac{\beta_B^{\alpha_B}\lambda_B^{\alpha_B-1} e^{-\beta_B\lambda_B}}{\Gamma\left(\alpha_B\right)} \frac{\beta_S^{\alpha_S}\lambda_S^{\alpha_S-1} e^{-\beta_S\lambda_S}}{\Gamma\left(\alpha_S\right)} d\lambda_B d\lambda_S. \end{aligned} \tag{B.10}$$

Collecting terms and rearranging the equation results in

$$\begin{aligned} p_{MS} = & \frac{1}{Z}\frac{\beta_B^{\alpha_B}\beta_S^{\alpha_S}}{\Gamma\left(d_B+1\right)\Gamma\left(d_S+1\right)\Gamma\left(d_R+1\right)\Gamma\left(\alpha_B\right)\Gamma\left(\alpha_S\right)} \times \\ & \int_0^\infty \lambda_B^{d_B+\alpha_B-1}\lambda_S^{d_S+\alpha_S-1}\left(\lambda_B+\lambda_S\right)^{d_R} e^{-(2+\beta_B)\lambda_B} e^{-(2+\beta_S)\lambda_S} d\lambda_B d\lambda_S. \end{aligned} \tag{B.11}$$

It may be noted that this equation is symmetric under the interchange of variables in B and S. The order of integration should not affect the result, although the two approaches may provide results that do not look symmetric under this interchange. As an aside, this is a way to discover equalities between different functional forms of the same equation.

For ease of notation, define a constant K_1,

$$K_1 = \frac{1}{Z}\frac{\beta_B^{\alpha_B}\beta_S^{\alpha_S}}{\Gamma\left(d_B+1\right)\Gamma\left(d_S+1\right)\Gamma\left(d_R+1\right)\Gamma\left(\alpha_B\right)\Gamma\left(\alpha_S\right)}. \tag{B.12}$$

Break the integral into an inner and outer integral,

$$p_{MS} = K_1 \int_0^\infty \lambda_S^{d_S+\alpha_S-1} e^{-(2+\beta_S)\lambda_S} \int_0^\infty \lambda_B^{d_B+\alpha_B-1}\left(\lambda_B+\lambda_S\right)^{d_R} e^{-(2+\beta_B)\lambda_B} d\lambda_B d\lambda_S. \tag{B.13}$$

Use Gradshteyn and Ryzhik 3.383.4 [13] to get

$$\begin{aligned} \int_u^\infty x^{\nu-1}\left(x-u\right)^{\mu-1} e^{-\gamma x} dx = & \gamma^{-\frac{\mu+\nu}{2}} u^{\frac{\mu+\nu-2}{2}} \Gamma\left(\mu\right) \exp\left(-\frac{\gamma u}{2}\right) \times \\ & W_{\frac{\nu-\mu}{2},\frac{1-\mu-\nu}{2}}\left(\gamma u\right), \left[\operatorname{Re}\mu > 0, \operatorname{Re}\gamma u > 0\right]. \end{aligned} \tag{B.14}$$

The function $W_{(a,b)}(x)$ is a Whittaker function, a special function. Assign the match between terms in Equation B.13 and Equation B.14 to be

$$\begin{aligned} x &= \lambda_B + \lambda_S, && \text{(B.15)}\\ u &= \lambda_S, && \text{(B.16)}\\ \nu &= (d_R + 1), && \text{(B.17)}\\ \mu &= (d_B + \alpha_B), && \text{(B.18)}\\ \gamma &= (2 + \beta_B). && \text{(B.19)} \end{aligned}$$

Because we're only working on integrating the inner integral at this point,

$$dx = d\lambda_B. \tag{B.20}$$

Convert terms to match the Gradshteyn and Ryzhik integral,

$$p_{MS} = K_1 \int_0^\infty \lambda_S^{d_S+\alpha_S-1} e^{-(2+\beta_S)\lambda_S} \int_u^\infty (x-u)^{\mu-1} (x)^{\nu-1} e^{-\gamma(x-u)} dx du. \tag{B.21}$$

We end up with a factor of $\exp(\gamma u)$ inside the inner integral because of how we've had to reformat the integral to match Gradshteyn and Ryzhik's formula. The expansion results in $\exp((2+\beta_B)\lambda_S)$. This term can be combined with the $\exp(-(2+\beta_S)\lambda_S)$ in the outer integral to give the term $\exp(\beta_B - \beta_S)\lambda_S$.

Insert the solution and convert from Gradshteyn and Ryzhik terms to the previous terms to get

$$\begin{aligned} p_{MS} =& K_1 (2+\beta_B)^{-\frac{(d_B+\alpha_B)+(d_R+1)}{2}} \times \\ & \int_0^\infty \lambda_S^{d_S+\alpha_S-1} \lambda_S^{\frac{(d_B+\alpha_B)+(d_R+1)-2}{2}} e^{(\beta_B-\beta_S)\lambda_S} \Gamma(d_B+\alpha_B) e^{-\frac{(2+\beta_B)\lambda_S}{2}} \times \\ & W_{\frac{(d_R+1)-(d_B+\alpha_B)}{2}, \frac{1-(d_B+\alpha_B)-(d_R+1)}{2}} ((2+\beta_S)\lambda_S) d\lambda_S. \end{aligned} \tag{B.22}$$

Move constant terms outside the integral and cancel where possible to get

$$\begin{aligned} p_{MS} =& K_1 (2+\beta_B)^{-\frac{(d_B+\alpha_B)+(d_R+1)}{2}} \Gamma(d_B+\alpha_B) \int_0^\infty \lambda_S^{\frac{d_B+2d_S+d_R+\alpha_B+2\alpha_S-3}{2}} e^{-\frac{(2\beta_S-\beta_B+2)}{2}\lambda_S} \times \\ & W_{\frac{(d_R+1)-(d_B+\alpha_B)}{2}, \frac{-(d_B+d_R+\alpha_B)}{2}} ((2+\beta_B)\lambda_S) d\lambda_S. \end{aligned} \tag{B.23}$$

A suitable integral formula can be found in a more recent edition of Gradshteyn and Ryzhik with formula 7.621.3 [14] to get

$$\begin{aligned} \int_0^\infty t^{\alpha_w} e^{-s_w t} W_{\lambda_w,\mu_w}(qt) dt = & \frac{\Gamma\left(\alpha_w + \mu_w + \frac{3}{2}\right) \Gamma\left(\alpha_w - \mu_w + \frac{3}{2}\right)}{\Gamma(\alpha_w - \lambda_w + 2)} q^{\mu_w + \frac{1}{2}} \left(s_w + \frac{q}{2}\right)^{-\left(\alpha_w + \mu_w + \frac{3}{2}\right)} \times \\ & F\left(\alpha_w + \mu_w + \frac{3}{2}, \mu_w - \lambda_w + \frac{1}{2}; \alpha_w - \lambda_w + 2; \frac{2s_w - q}{2s_w + q}\right), \\ & \left[\text{Re } \left(\alpha_w \pm \mu_w + \frac{3}{2}\right) > 0, \text{Re } s_w > -\frac{q}{2}, q > 0\right], \end{aligned} \tag{B.24}$$

where F is a hypergeometric function, in this case, specifically ${}_2F_1$.

The Whittaker function $W_{\lambda,\mu}(z)$ is symmetric under replacement of μ with $-\mu$, which is why the first limit condition allows for either a plus or minus sign on μ. Another symmetry in the double integral is the interchange of λ_B with λ_S simultaneously with the interchange of d_B and d_S. These symmetries can be used to obtain four different functions for the double integral. There will be an opportunity for fortuitous cancellation of terms in the eventual Bayes factor calculation with the right selection of the integration order.

The first two of four assignments of terms between Equation B.23 and B.24 is

$$\alpha_w = \frac{d_B + 2d_S + d_R + \alpha_B + 2\alpha_S - 3}{2}, \tag{B.25}$$

$$s_w = \frac{(2\beta_S - \beta_B + 2)}{2}, \tag{B.26}$$

$$\lambda_w = \frac{d_R + 1 - (d_B + \alpha_B)}{2}, \tag{B.27}$$

$$\mu_w = \pm\frac{d_B + d_R + \alpha_B}{2}, \tag{B.28}$$

$$q = 2 + \beta_B. \tag{B.29}$$

Calculations to check the constraint in Equation B.24 result in

$$\alpha_w + \mu_w + \frac{3}{2} = d_B + d_S + d_R + \alpha_B + \alpha_S, \tag{B.30}$$

$$\alpha_w - \mu_w + \frac{3}{2} = d_S + \alpha_S, \tag{B.31}$$

$$s_w + \frac{q}{2} = 2 + \beta_S. \tag{B.32}$$

Either sign on μ_w will work if the prior $\alpha_s > 0$. The needed terms in Equation B.24 where μ is positive are

$$\alpha_w + \mu_w + \frac{3}{2} = d_B + d_S + d_R + \alpha_B + \alpha_S, \tag{B.33}$$

$$\alpha_w - \mu_w + \frac{3}{2} = d_S + \alpha_S, \tag{B.34}$$

$$s_w + \frac{q}{2} = 2 + \beta_S, \tag{B.35}$$

$$s_w - \frac{q}{2} = \beta_S - \beta_B, \tag{B.36}$$

$$\mu_w - \lambda_w + \frac{1}{2} = d_B + \alpha_B, \tag{B.37}$$

$$\alpha_w - \lambda_w + 2 = d_B + d_S + \alpha_B + \alpha_S. \tag{B.38}$$

The integral is then

$$\begin{aligned}p_{MS} =&K_1 \left(2+\beta_B\right)^{-\frac{d_B+d_R+\alpha_B+1}{2}} \frac{\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)\Gamma\left(d_S+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_S+\alpha_B\right)}\times\\ &\left(2+\beta_B\right)^{\frac{d_B+d_R+\alpha_B+1}{2}}\left(2+\beta_S\right)^{-\left(d_S+d_B+d_R+\alpha_S+\alpha_B\right)}\times\\ &F\left(d_B+d_S+d_R+\alpha_B+\alpha_S, d_B+\alpha_B; d_B+d_S+\alpha_B+\alpha_S; \left(\frac{\beta_S-\beta_B}{2+\beta_S}\right)\right).\end{aligned}\tag{B.39}$$

Cancel some terms to get

$$\begin{aligned}p_{MS,1} =&K_1 \frac{\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_S+\alpha_B\right)}\times\\ &\left(2+\beta_S\right)^{-\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}\times\\ &F\left(d_B+d_S+d_R+\alpha_B+\alpha_S, d_B+\alpha_B; d_B+d_S+\alpha_B+\alpha_S; \left(\frac{\beta_S-\beta_B}{2+\beta_S}\right)\right).\end{aligned}\tag{B.40}$$

If the negative value of μ is used,

$$\alpha_w+\mu_w+\frac{3}{2}=d_S+\alpha_S,\tag{B.41}$$

$$\alpha_w-\mu_w+\frac{3}{2}=d_B+d_S+d_R+\alpha_B+\alpha_S,\tag{B.42}$$

$$s_w+\frac{q}{2}=2+\beta_S,\tag{B.43}$$

$$s_w-\frac{q}{2}=\beta_S-\beta_B,\tag{B.44}$$

$$\mu_w-\lambda_w+\frac{1}{2}=-d_R,\tag{B.45}$$

$$\alpha_w-\lambda_w+2=d_B+d_S+\alpha_B+\alpha_S.\tag{B.46}$$

The integral is then

$$\begin{aligned}p_{MS,2} =&K_1 \frac{\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)}\times\\ &\left(2+\beta_B\right)^{-\left(d_B+d_R+\alpha_B\right)}\left(2+\beta_S\right)^{-\left(d_S+\alpha_S\right)}\times\\ &F\left(d_S+\alpha_S, -d_R; d_B+d_S+\alpha_S+\alpha_B; \left(\frac{\beta_S-\beta_B}{2+\beta_S}\right)\right).\end{aligned}\tag{B.47}$$

If the order of integration is reversed, the result is the same as the exchange of subscript labels, S and B, in Equations B.40 and B.47:

$$\begin{aligned}p_{MS,3} =&K_1 \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)}\times\\ &\left(2+\beta_B\right)^{-\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}\times\\ &F\left(d_B+d_S+d_R+\alpha_B+\alpha_S, d_S+\alpha_S; d_B+d_S+\alpha_B+\alpha_S; \left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right),\end{aligned}\tag{B.48}$$

and

$$\begin{aligned} p_{MS,4} =& K_1 \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)} \times \\ & \left(2+\beta_S\right)^{-\left(d_S+d_R+\alpha_S\right)}\left(2+\beta_B\right)^{-\left(d_B+\alpha_B\right)} \times \\ & F\left(d_B+\alpha_B,-d_R;d_B+d_S+\alpha_B+\alpha_S;\left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right). \end{aligned} \tag{B.49}$$

Six identities for hypergeometric functions can be obtained, displayed here as a chain of equalities:

$$\begin{aligned} & \left(2+\beta_S\right)^{-\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)} \times \\ & F\left(d_B+d_S+d_R+\alpha_B+\alpha_S,d_B+\alpha_B;d_B+d_S+\alpha_B+\alpha_S;\left(\frac{\beta_S-\beta_B}{2+\beta_S}\right)\right) \\ =& \left(2+\beta_B\right)^{-\left(d_B+d_R+\alpha_B\right)}\left(2+\beta_S\right)^{-\left(d_S+\alpha_S\right)} \times \\ & F\left(d_S+\alpha_S,-d_R;d_B+d_S+\alpha_B+\alpha_S;\left(\frac{\beta_S-\beta_B}{2+\beta_S}\right)\right) \\ =& \left(2+\beta_B\right)^{-\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)} \times \\ & F\left(d_B+d_S+d_R+\alpha_B+\alpha_S,d_S+\alpha_S;d_B+d_S+\alpha_B+\alpha_S;\left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right) \\ =& \left(2+\beta_S\right)^{-\left(d_S+d_R+\alpha_S\right)}\left(2+\beta_B\right)^{-\left(d_B+\alpha_B\right)} \times \\ & F\left(d_B+\alpha_B,-d_R;d_B+d_S+\alpha_B+\alpha_S;\left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right). \end{aligned} \tag{B.50}$$

B.1.2 The Probability of No Merger with Consistent Generators

The probability of no merger with consistent generators through time is given by Equation B.7. The insertion of the Poisson probabilities and conjugate prior functions produces

$$p_{\bar{M}S} = \frac{1}{Z}\int_0^\infty \frac{\lambda_B^{d_B}e^{-\lambda_B}}{\Gamma\left(d_B+1\right)}\frac{\lambda_S^{d_S}e^{-\lambda_S}}{\Gamma\left(d_S+1\right)}\frac{\lambda_S^{d_R}e^{-\lambda_S}}{\Gamma\left(d_R+1\right)}\frac{\beta_B^{\alpha_B}\lambda_B^{\alpha_B-1}e^{-\beta_B\lambda_B}}{\Gamma\left(\alpha_B\right)}\frac{\beta_S^{\alpha_S}\lambda_S^{\alpha_S-1}e^{-\beta_S\lambda_S}}{\Gamma\left(\alpha_S\right)}d\lambda_B d\lambda_S. \tag{B.51}$$

Collecting and rearranging terms results in

$$\begin{aligned} p_{\bar{M}S} =& \frac{1}{Z}\frac{\beta_B^{\alpha_B}\beta_S^{\alpha_S}}{\Gamma\left(d_B+1\right)\Gamma\left(d_S+1\right)\Gamma\left(d_R+1\right)\Gamma\left(\alpha_B\right)\Gamma\left(\alpha_S\right)} \times \\ & \int_0^\infty \lambda_B^{d_B+\alpha_B-1}e^{-\left(1+\beta_B\right)\lambda_B}d\lambda_B \int_0^\infty \lambda_S^{d_S+d_R+\alpha_S-1}e^{-\left(2+\beta_S\right)\lambda_S}d\lambda_S. \end{aligned} \tag{B.52}$$

Using the constant K_1, defined in Equation B.12, gives

$$p_{\bar{M}S} = K_1 \int_0^\infty \lambda_B^{d_B+\alpha_B-1}e^{-\left(1+\beta_B\right)\lambda_B}d\lambda_B \int_0^\infty \lambda_S^{d_S+d_R+\alpha_S-1}e^{-\left(2+\beta_S\right)\lambda_S}d\lambda_S. \tag{B.53}$$

The two equations are solved with Gradshteyn and Ryzhik 3.381.4 [13] (see Equation A.6) to produce

$$p_{\bar{M}S} = K_1 \left(1+\beta_B\right)^{-(d_B+\alpha_B)} \Gamma\left(d_B+\alpha_B\right) \left(2+\beta_S\right)^{-(d_S+d_R+\alpha_S)} \Gamma\left(d_S+d_R+\alpha_S\right). \tag{B.54}$$

Collecting like terms reduces the equation to

$$p_{\bar{M}S} = K_1 \Gamma\left(d_B+\alpha_B\right) \Gamma\left(d_S+d_R+\alpha_S\right) \left(1+\beta_B\right)^{-(d_B+\alpha_B)} \left(2+\beta_S\right)^{-(d_S+d_R+\alpha_S)}. \tag{B.55}$$

B.1.3 The Bayes Factor for Two Hypotheses with Consistent Generators

If the prior probabilities for the two hypotheses, merged and not merged, are represented by ρ_{MS} and $\rho_{\bar{M}S}$ then the Bayes factor is

$$R_2 = \frac{\rho_{MS}\; p_{MS}}{\rho_{\bar{M}S}\; p_{\bar{M}S}}. \tag{B.56}$$

The probability P_{MS} has four different forms that can be selected from the ratio function (Equations B.40, B.47, B.48, and B.49). A comparison with the final results for $p_{\bar{M}S}$ with the four variants shows that $p_{MS,4}$ (Equation B.49) will provide for the most cancellation of terms;

$$\begin{aligned} R_2 =& K_1 \frac{\rho_{MS}}{\rho_{\bar{M}S}} \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{K_1\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)\Gamma\left(d_B+\alpha_B\right)\Gamma\left(d_S+d_R+\alpha_S\right)} \times \\ & \frac{\left(2+\beta_S\right)^{-(d_S+d_R+\alpha_S)}\left(2+\beta_B\right)^{-(d_B+\alpha_B)}}{\left(1+\beta_B\right)^{-(d_B+\alpha_B)}\left(2+\beta_S\right)^{-(d_S+d_R+\alpha_S)}} \times \\ & F\left(d_B+\alpha_B, -d_R; d_B+d_S+\alpha_B+\alpha_S; \left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right). \end{aligned} \tag{B.57}$$

The next step is to cancel a few more terms and invert the form of the β term to get

$$\begin{aligned} R_2 =& \frac{\rho_{MS}}{\rho_{\bar{M}S}} \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)\Gamma\left(d_S+d_R+\alpha_S\right)} \left(\frac{1+\beta_B}{2+\beta_B}\right)^{d_B+\alpha_B} \times \\ & F\left(d_B+\alpha_B, -d_R; d_B+d_S+\alpha_B+\alpha_S; \left(\frac{\beta_B-\beta_S}{2+\beta_B}\right)\right). \end{aligned} \tag{B.58}$$

If the prior parameters β_B and β_S are equal, the hypergeometric function F drops out of the equation because $F\left(a, b; c; 0\right) = 1$.

$$R_{2,\beta} = \frac{\rho_{MS}}{\rho_{\bar{M}S}} \frac{\Gamma\left(d_S+\alpha_S\right)\Gamma\left(d_B+d_S+d_R+\alpha_B+\alpha_S\right)}{\Gamma\left(d_B+d_S+\alpha_B+\alpha_S\right)\Gamma\left(d_S+d_R+\alpha_S\right)} \left(\frac{1+\beta}{2+\beta}\right)^{d_B+\alpha_B}. \tag{B.59}$$

B.2 POISSON BAYES FACTORS INCLUDING CHANGES IN GENERATORS OVER TIME

The Bayes factors that result from the assumption that the generative models do not vary through time may not be valid for many real situations. It is no surprising to see activity patterns change over time in social networks, which is part of the attraction for members. This section extends the Bayes factor calculations to assume that there is some probability that parameters in the generative model for the potential host may change with time. For this derivation, it is assumed that the members of the banned subreddit do not change their behavior because otherwise they would be undetectable.

B.2.1 Merger with Generator Changes

The probability for merger where the host activity pattern changes is given in Equation B.8, and expands to

$$\begin{aligned} p_{M\bar{S}} = \frac{1}{Z} \int_0^\infty & \frac{\lambda_B^{d_B} e^{-\lambda_B}}{\Gamma(d_B+1)} \frac{\lambda_S^{d_S} e^{-\lambda_S}}{\Gamma(d_S+1)} \frac{(\lambda_R+\lambda_B)^{d_R} e^{-(\lambda_R+\lambda_B)}}{\Gamma(d_R+1)} \\ & \frac{\beta_B^{\alpha_B} \lambda_B^{\alpha_B-1} e^{-\beta_B\lambda_B}}{\Gamma(\alpha_B)} \frac{\beta_S^{\alpha_S} \lambda_S^{\alpha_S-1} e^{-\beta_S\lambda_S}}{\Gamma(\alpha_S)} \frac{\beta_R^{\alpha_R} \lambda_R^{\alpha_R-1} e^{-\beta_R\lambda_R}}{\Gamma(\alpha_R)} d\lambda_B d\lambda_S d\lambda_R. \end{aligned} \tag{B.60}$$

Extract constant terms from the integral and collect terms. Use the constant term K_1 defined in Equation B.12 to get

$$\begin{aligned} p_{M\bar{S}} = K_1 & \frac{\beta_R^{\alpha_R}}{\Gamma(\alpha_R)} \times \\ & \int_0^\infty \lambda_B^{d_B+\alpha_B-1} e^{-(2+\beta_B)\lambda_B} \lambda_S^{d_S+\alpha_S-1} e^{-(1+\beta_S)\lambda_S} (\lambda_R+\lambda_B)^{d_R} e^{-(1+\beta_R)\lambda_R} \lambda_R^{\alpha_R-1} d\lambda_B d\lambda_S d\lambda_R. \end{aligned} \tag{B.61}$$

The λ_S terms separate into an independent integral, while the λ_B and λ_R terms are coupled. This results in

$$\begin{aligned} p_{M\bar{S}} = K_1 & \frac{\beta_R^{\alpha_R}}{\Gamma(\alpha_R)} \int_0^\infty \lambda_S^{d_S+\alpha_S-1} e^{-(1+\beta_S)\lambda_S} d\lambda_S \\ & \int_0^\infty \lambda_B^{d_B+\alpha_B-1} e^{-(2+\beta_B)\lambda_B} (\lambda_R+\lambda_B)^{d_R} e^{-(1+\beta_R)\lambda_R} \lambda_R^{\alpha_R-1} d\lambda_B d\lambda_R. \end{aligned} \tag{B.62}$$

The first integral is again solved with Gradshteyn and Ryzhik 3.381.4 [13] (see Equation A.6) to get

$$\begin{aligned} p_{M\bar{S}} = K_1 & \frac{\beta_R^{\alpha_R}}{\Gamma(\alpha_R)} (1+\beta_S)^{-(d_S+\alpha_S)} \Gamma(d_S+\alpha_S) \\ & \int_0^\infty \lambda_B^{d_B+\alpha_B-1} e^{-(2+\beta_B)\lambda_B} (\lambda_R+\lambda_B)^{d_R} e^{-(1+\beta_R)\lambda_R} \lambda_R^{\alpha_R-1} d\lambda_B d\lambda_R. \end{aligned} \tag{B.63}$$

The double integral is of a similar form to Equation B.11, with λ_R replacing λ_S, $d_S = 0$, α_R replacing α_S, and $\beta_R - 1$ replacing β_S. Equation B.49 is selected as the version to use because of its selection for the previous Bayes factor calculation. After replacement, the result is

$$\begin{aligned}p_{M\bar{S}} =& K_1 \frac{\beta_R^{\alpha_R}}{\Gamma\left(\alpha_R\right)} \left(1+\beta_S\right)^{-\left(d_S+\alpha_S\right)} \Gamma\left(d_S+\alpha_S\right) \\ & \frac{\Gamma\left(\alpha_R\right) \Gamma\left(d_B+\alpha_B\right) \Gamma\left(d_B+d_R+\alpha_B+\alpha_R\right)}{\Gamma\left(d_B+\alpha_B+\alpha_R\right)} \times \\ & \left(1+\beta_R\right)^{-\left(d_R+\alpha_R\right)} \left(2+\beta_B\right)^{-\left(d_B+\alpha_B\right)} \times \\ & F\left(d_B+\alpha_B, -d_R; d_B+\alpha_B+\alpha_R; \left(\frac{\beta_B-\beta_R+1}{2+\beta_B}\right)\right).\end{aligned} \tag{B.64}$$

A few terms cancel and others can be collected to give

$$\begin{aligned}p_{M\bar{S}} =& K_1 \beta_R^{\alpha_R} \left(2+\beta_B\right)^{-\left(d_B+\alpha_B\right)} \left(1+\beta_S\right)^{-\left(d_S+\alpha_S\right)} \left(1+\beta_R\right)^{-\left(d_R+\alpha_R\right)} \\ & \frac{\Gamma\left(d_B+\alpha_B\right) \Gamma\left(d_S+\alpha_S\right) \Gamma\left(d_B+d_R+\alpha_B+\alpha_R\right)}{\Gamma\left(d_B+\alpha_B+\alpha_R\right)} \times \\ & F\left(d_B+\alpha_B, -d_R; d_B+\alpha_B+\alpha_R; \left(\frac{\beta_B-\beta_R+1}{2+\beta_B}\right)\right).\end{aligned} \tag{B.65}$$

It is not possible to neglect the hypergeometric function in this case because the last term in the parentheses is not zero when β_B and β_R are equal.

B.2.2 No Merger with Poisson Probability Generator Changes

The probability for no merger for the case where the generative model parameters may change is given by Equation B.9. All integral terms are separable into three different integrals, all solved with Gradshteyn and Ryzhik 3.381.4 [13] (see Equation A.6). Terms in Equation B.62 can be carried over to give

$$\begin{aligned}p_{\bar{M}\bar{S}} = & K_1 \frac{\beta_R^{\alpha_R}}{\Gamma\left(\alpha_R\right)} \left(1+\beta_B\right)^{-\left(d_B+\alpha_B\right)} \Gamma\left(d_B+\alpha_B\right) \left(1+\beta_S\right)^{-\left(d_S+\alpha_S\right)} \Gamma\left(d_S+\alpha_S\right) \\ & \left(1+\beta_R\right)^{-\left(d_R+\alpha_R\right)} \Gamma\left(d_R+\alpha_R\right).\end{aligned} \tag{B.66}$$

Collecting common terms gives

$$\begin{aligned}p_{\bar{M}\bar{S}} = & K_1 \frac{\beta_R^{\alpha_R} \Gamma\left(d_B+\alpha_B\right) \Gamma\left(d_S+\alpha_S\right) \Gamma\left(d_R+\alpha_R\right)}{\Gamma\left(\alpha_R\right)} \\ & \left(1+\beta_B\right)^{-\left(d_B+\alpha_B\right)} \left(1+\beta_S\right)^{-\left(d_S+\alpha_S\right)} \left(1+\beta_R\right)^{-\left(d_R+\alpha_R\right)}.\end{aligned} \tag{B.67}$$

B.2.3 Likelihood Ratio for Poisson Probability Generator Changes

There are now four terms, so the prior probabilities are represented with ρ_{MS}, $\rho_{\bar{M}S}$, $\rho_{M\bar{S}}$, and $\rho_{\bar{M}\bar{S}}$. The Bayes factor between the two sets of merged and not-merged hypotheses is

$$R_4 = \frac{\rho_{MS}\ p_{MS} + \rho_{M\bar{S}}\ p_{M\bar{S}}}{\rho_{\bar{M}S}\ p_{\bar{M}S} + \rho_{\bar{M}\bar{S}}\ p_{\bar{M}\bar{S}}}. \tag{B.68}$$

The K_1 constant and the $\Gamma\left(d_B+\alpha_B\right)$ is common to all four terms and cancel in the Bayes factor. No other terms are common across all four probabilities.

B.3 BINOMIAL MERGERS AND SPLITS

Although the Bayes factors for the similarity between samples from both the binomial and multinomial probability functions could be derived from a single description, the derivations for mergers is significantly more complicated for these distribution functions. This complexity arises from the model of how binomial and multinomial functions merge. The Poisson function is easily modeled because the distribution parameters simply add. The merger of parameters for binomial and multinomial functions do not obey such a simple combination rule. A merged distribution is dependent on the relative strengths of the contributions of two distributions to the merged distribution.

The binomial distribution is

$$P\left(d;p,n\right)=\binom{n}{d}p^{d}\left(1-p\right)^{n-d}, \tag{B.69}$$

where n is the total number of samples, and d is the number of observed events in the first of two categories in the n samples. The variable p is the probability for observing an event of category one. The conjugate prior for the binomial distribution is a beta distribution,

$$P\left(p\right)=\frac{p^{\alpha-1}\left(1-p\right)^{\beta-1}}{B\left(\alpha,\beta\right)}, \tag{B.70}$$

where α and β are prior shape parameters. The merged probability p_R is chosen to be a mixture of the two initial probabilities, p_B and p_S.

$$p_R=\gamma p_B+\left(1-\gamma\right)p_S;\quad 0\leq\gamma\leq 1, \tag{B.71}$$

where γ is the mixing parameter for the two different probabilities. This parameter defines the relative proportion of the two combined distributions in the merged distribution. If information were available on sample durations, then this information could also be incorporated into the value of γ. For this derivation, a beta distribution will eventually be assumed for the prior probability $p\left(\gamma\right)$. A uniform prior can be obtained from a beta distribution by defining the α and β terms associated with the function to be equal to 1.

B.3.1 Binomial Probability of Merger with the Same Generator

The posterior probability for merger with consistent distribution function parameters for the potential host group, before and after merger, is

$$\begin{aligned}p_{MS}=\frac{1}{Z}\int_0^1\int_0^1\int_0^1 &\binom{n_B}{d_B}p_B^{d_B}\left(1-p_B\right)^{n_B-d_B}\binom{n_S}{d_S}p_S^{d_S}\left(1-p_S\right)^{n_S-d_S}\\ &\binom{n_R}{d_R}\left(\gamma p_B+\left(1-\gamma\right)p_S\right)^{d_R}\left(1-\left(\gamma p_B+\left(1-\gamma\right)p_S\right)\right)^{n_R-d_R}\frac{p_B^{\alpha_B-1}\left(1-p_B\right)^{\beta_B-1}}{B\left(\alpha_B,\beta_B\right)}\\ &\frac{p_S^{\alpha_S-1}\left(1-p_S\right)^{\beta_S-1}}{B\left(\alpha_S,\beta_S\right)}p\left(\gamma\right)dp_Bdp_Sd\gamma.\end{aligned} \tag{B.72}$$

Rearrange terms to get

$$\begin{gathered}p_{MS} = \frac{1}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R} \\ \int_0^1\int_0^1 p_B^{d_B+\alpha_B-1}\left(1-p_B\right)^{n_B-d_B+\beta_B-1}\int_0^1 p_S^{d_S+\alpha_S-1}\left(1-p_S\right)^{n_S-d_S+\beta_S-1} \\ \left(\gamma p_B+\left(1-\gamma\right)p_S\right)^{d_R}\left(1-\gamma p_B-\left(1-\gamma\right)p_S\right)^{n_R-d_R}p\left(\gamma\right)dp_Bdp_Sd\gamma.\end{gathered} \tag{B.73}$$

The general double integral that is needed is of the form of

$$\begin{gathered}g_{II}\left(a,b,c,d,e,f,u_x,u_y\right) = \int_0^{u_x} x^{d-1}\left(u_x-x\right)^{a-1}\times \\ \int_0^{u_y} y^{e-1}\left(u_y-y\right)^{b-1}\left(\gamma x+\left(1-\gamma\right)y\right)^f\left(\gamma\left(u_x-x\right)+\left(1-\gamma\right)\left(u_y-y\right)\right)^c dydx,\end{gathered} \tag{B.74}$$

which is derived in Appendix B.5.3. The substitutions are

$$\begin{aligned}x &= p_B, \\ y &= p_S, \\ a &= n_B-d_B+\beta_B, \\ b &= n_S-d_S+\beta_S, \\ c &= n_R-d_R, \\ d &= d_B+\alpha_B, \\ e &= d_S+\alpha_S, \\ f &= d_R, \\ u_x &= 1, \\ u_y &= 1,\end{aligned} \tag{B.75}$$

which gives

$$\begin{gathered}p_{MS} = \frac{1}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R}\sum_{t=0}^{d_R}\frac{\Gamma\left(d_R+1\right)}{\Gamma\left(d_R-t+1\right)}\times \\ \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(n_S-d_S+\beta_S,d_S+\alpha_S+t+u\right)\times \\ \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^p B\left(n_B-d_B+\beta_B,d_B+\alpha_B+d_R-t+v\right)\times \\ \frac{\Gamma\left(n_R-d_R+1\right)}{\Gamma\left(p+1\right)\Gamma\left(n_R-d_R-u-v+1\right)}\int_0^1\left(1-\gamma\right)^{t+u}\gamma^{d_R-t+v}p\left(\gamma\right)d\gamma.\end{gathered} \tag{B.76}$$

The relationship between the binomial coefficient and gamma functions,

$$\binom{n}{k} = \frac{\Gamma\left(u+1\right)}{\Gamma\left(k+1\right)\Gamma\left(u-k+1\right)}, \tag{B.77}$$

can be applied to get

$$\begin{gathered}p_{MS} = \frac{\Gamma\left(n_R+1\right)}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\sum_{t=0}^{d_R}\frac{1}{\Gamma\left(d_R-t+1\right)}\times\\ \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(n_S-d_S+\beta_S,d_S+\alpha_S+t+u\right)\times\\ \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B,d_B+\alpha_B+d_R-t+v\right)\times\\ \frac{1}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)}\int_0^1\left(1-\gamma\right)^{t+u}\gamma^{d_R-t+v}p\left(\gamma\right)d\gamma.\end{gathered}\tag{B.78}$$

The integral over γ remains to be solved. Inserting the Dirichlet prior for $p\left(\gamma\right)$ gives the integral

$$I_\gamma = \int_0^1\left(1-\gamma\right)^{t+u}\gamma^{d_R-t+v}p\left(\gamma\right)d\gamma = \int_0^1\left(1-\gamma\right)^{t+u}\gamma^{d_R-t+v}\frac{\gamma^{\alpha_\gamma-1}\left(1-\gamma\right)^{\beta_\gamma-1}}{B\left(\alpha_\gamma,\beta_\gamma\right)}d\gamma.\tag{B.79}$$

Extracting and combining terms leads to

$$I_\gamma = \frac{1}{B\left(\alpha_\gamma,\beta_\gamma\right)}\int_0^1\left(1-\gamma\right)^{t+u+\beta_\gamma-1}\gamma^{d_R-t+v+\alpha_\gamma-1}d\gamma.\tag{B.80}$$

The solution is given by Equation B.120.

$$I_\gamma = \frac{B\left(d_R-t+v+\alpha_\gamma,t+u+\beta_\gamma\right)}{B\left(\alpha_\gamma,\beta_\gamma\right)}.\tag{B.81}$$

The integral for mixing parameter γ can be inserted into the full integral.

$$\begin{gathered}p_{MS} = \frac{\Gamma\left(n_R+1\right)}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\sum_{t=0}^{d_R}\frac{1}{\Gamma\left(d_R-t+1\right)}\times\\ \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(n_S-d_S+\beta_S,d_S+\alpha_S+t+u\right)\times\\ \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B,d_B+\alpha_B+d_R-t+v\right)\times\\ \frac{B\left(d_R-t+v+\alpha_\gamma,t+u+\beta_\gamma\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)B\left(\alpha_\gamma,\beta_\gamma\right)}.\end{gathered}\tag{B.82}$$

B.3.2 Binomial Probability of No Merger with the Same Generator

The probability of no merger, for the case where the generative model parameters for the potential host do not change, is

$$\begin{aligned}p_{\bar{M}S} = \frac{1}{Z}\int_0^1\int_0^1 \binom{n_B}{d_B} p_B^{d_B}\left(1-p_B\right)^{n_B-d_B}\binom{n_S}{d_S} p_S^{d_S}\left(1-p_S\right)^{n_S-d_S} \\ \binom{n_R}{d_R} p_S^{d_R}\left(1-p_S\right)^{n_R-d_R}\frac{p_B^{\alpha_B-1}\left(1-p_B\right)^{\beta_B-1}}{B\left(\alpha_B,\beta_B\right)}\frac{p_S^{\alpha_S-1}\left(1-p_S\right)^{\beta_S-1}}{B\left(\alpha_S,\beta_S\right)}dp_B dp_S.\end{aligned} \tag{B.83}$$

Extract constant terms and combine common terms to get

$$\begin{aligned}p_{\bar{M}S} = \binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R}\frac{1}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)}\times \\ \int_0^1 p_B^{d_B+\alpha_B-1}\left(1-p_B\right)^{n_B-d_B+\beta_B-1}dp_B\times \\ \int_0^1 p_S^{d_S+d_R+\alpha_S-1}\left(1-p_S\right)^{n_S-d_S+n_R-d_R+\beta_S-1}dp_S.\end{aligned} \tag{B.84}$$

The probability terms separate into independent integrals that are solved with Gradshteyn and Ryzhik 3.191.1 [13].

$$\begin{aligned}p_{\bar{M}S} = \binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R}\frac{1}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)}\times \\ B\left(d_B+\alpha_B, n_B-d_B+\beta_B\right)B\left(d_S+d_R+\alpha_S, n_S-d_S+n_R-d_R+\beta_S\right).\end{aligned} \tag{B.85}$$

B.3.3 Binomial Probability of Merger with a Different Generator

The probability of a merger where the generative model parameters for the potential destination group have changed is

$$\begin{aligned}p_{M\bar{S}} = \frac{1}{Z}\int_0^1\int_0^1\int_0^1\int_0^1 \binom{n_B}{d_B} p_B^{d_B}\left(1-p_B\right)^{n_B-d_B}\binom{n_S}{d_S} p_S^{d_S}\left(1-p_S\right)^{n_S-d_S}\times \\ \binom{n_R}{d_R}\left(\gamma p_B+\left(1-\gamma\right)p_R\right)^{d_R}\left(1-\left(\gamma p_B+\left(1-\gamma\right)p_R\right)\right)^{n_R-d_R}\times \\ \frac{p_B^{\alpha_B-1}\left(1-p_B\right)^{\beta_B-1}}{B\left(\alpha_B,\beta_B\right)}\frac{p_S^{\alpha_S-1}\left(1-p_S\right)^{\beta_S-1}}{B\left(\alpha_S,\beta_S\right)}\frac{p_R^{\alpha_R-1}\left(1-p_R\right)^{\beta_R-1}}{B\left(\alpha_R,\beta_R\right)}p\left(\gamma\right)dp_B dp_S dp_R d\gamma.\end{aligned} \tag{B.86}$$

Extract constant terms, combine common terms, and extract an independent integral in p_S to get

$$\begin{aligned}p_{M\bar{S}} = \frac{1}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)B\left(\alpha_R,\beta_R\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R}\times \\ \int_0^1 p_S^{d_S+\alpha_S-1}\left(1-p_S\right)^{n_S-d_S+\beta_S-1}dp_S\int_0^1\int_0^1\int_0^1 p_B^{d_B+\alpha_B-1}\left(1-p_B\right)^{n_B-d_B+\beta_B-1}\times \\ \left(\gamma p_B+\left(1-\gamma\right)p_R\right)^{d_R}\left(1-\left(\gamma p_B+\left(1-\gamma\right)p_R\right)\right)^{n_R-d_R}\times \\ p_R^{\alpha_R-1}\left(1-p_R\right)^{\beta_R-1}p\left(\gamma\right)dp_B dp_R d\gamma.\end{aligned} \tag{B.87}$$

The independent integral is again solved with Gradshteyn and Ryzhik 3.191.1 [13] to get

$$\begin{aligned}p_{M\bar{S}} = {} & \frac{1}{ZB\left(\alpha_B, \beta_B\right)B\left(\alpha_S, \beta_S\right)B\left(\alpha_R, \beta_R\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R}\times \\ & B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right)\int_0^1\int_0^1 p_B^{d_B+\alpha_B-1}\left(1-p_B\right)^{n_B-d_B+\beta_B-1}\times \\ & \int_0^1\left(\gamma p_B+\left(1-\gamma\right)p_R\right)^{d_R}\left(1-\left(\gamma p_B+\left(1-\gamma\right)p_R\right)\right)^{n_R-d_R}\times \\ & p_R^{\alpha_R-1}\left(1-p_R\right)^{\beta_R-1}p\left(\gamma\right)dp_R dp_B d\gamma.\end{aligned} \tag{B.88}$$

The next integral can be solved with g_{II} in Appendix B.5.3. The substitutions are

$$\begin{aligned} x &= p_B, \\ y &= p_R, \\ a &= n_B-d_B+\beta_B, \\ b &= \beta_R, \\ c &= n_R-d_R, \\ d &= d_B+\alpha_B, \\ e &= \alpha_R, \\ f &= d_R, \\ u_x &= 1, \\ u_y &= 1.\end{aligned} \tag{B.89}$$

This results in

$$\begin{aligned}p_{M\bar{S}} = {} & \frac{1}{ZB\left(\alpha_B, \beta_B\right)B\left(\alpha_S, \beta_S\right)B\left(\alpha_R, \beta_R\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\binom{n_R}{d_R}\times \\ & B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right)\int_0^1\sum_{t=0}^{d_R}\frac{\Gamma\left(d_R+1\right)}{\Gamma\left(d_R-t+1\right)}\times \\ & \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(\beta_R, \alpha_R+t+u\right)\left(1-\gamma\right)^{t+u}\times \\ & \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B, d_B+\alpha_B+d_R-t+v\right)\times \\ & \frac{\Gamma\left(n_R-d_R+1\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)}\gamma^{d_R-t+v}p\left(\gamma\right)d\gamma.\end{aligned} \tag{B.90}$$

Apply the relationship between binomial coefficients and gamma functions in Equation B.77 and collect terms in the mixing parameter, γ, to get

$$\begin{aligned}
p_{M\bar{S}} = & \frac{\Gamma\left(n_R+1\right)}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)B\left(\alpha_R,\beta_R\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}\times \\
& B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right)\sum_{t=0}^{d_R}\frac{1}{\Gamma\left(d_R-t+1\right)}\times \\
& \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(\beta_R, \alpha_R+t+u\right)\times \\
& \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B, d_B+\alpha_B+d_R-t+v\right)\times \\
& \frac{1}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)}\int_0^1 \gamma^{d_R-t+v}\left(1-\gamma\right)^{t+u}p\left(\gamma\right)d\gamma.
\end{aligned} \tag{B.91}$$

The final step is to insert the prior beta distribution for γ and integrate over the mixing parameter γ, which produces a beta function term, as in Equations B.80 and B.80. This results in

$$\begin{aligned}
p_{M\bar{S}} = & \frac{\Gamma\left(n_R+1\right)}{ZB\left(\alpha_B,\beta_B\right)B\left(\alpha_S,\beta_S\right)B\left(\alpha_R,\beta_R\right)}\binom{n_B}{d_B}\binom{n_S}{d_S}B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right)\times \\
& \sum_{t=0}^{d_R}\frac{1}{\Gamma\left(d_R-t+1\right)}\sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(\beta_R, \alpha_R+t+u\right)\times \\
& \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B, d_B+\alpha_B+d_R-t+v\right)\times \\
& \frac{B\left(d_R-t+v+\alpha_\gamma, t+u+\beta_\gamma\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)B\left(\alpha_\gamma,\beta_\gamma\right)}.
\end{aligned} \tag{B.92}$$

B.3.4 Binomial Probability of No Merger with a Different Generator

The probability of no merger where the destination group has changed generators is given by

$$\begin{aligned}
p_{\bar{M}\bar{S}} = & \frac{1}{Z}\int_0^1\int_0^1\int_0^1 \binom{n_B}{d_B}p_B^{d_B}\left(1-p_B\right)^{n_B-d_B}\binom{n_S}{d_S}p_S^{d_S}\left(1-p_S\right)^{n_S-d_S}\binom{n_R}{d_R}p_R^{d_R}\left(1-p_R\right)^{n_R-d_R} \\
& \frac{p_B^{\alpha_B-1}\left(1-p_B\right)^{\beta_B-1}}{B\left(\alpha_B,\beta_B\right)}\frac{p_S^{\alpha_S-1}\left(1-p_S\right)^{\beta_S-1}}{B\left(\alpha_S,\beta_S\right)}\frac{p_R^{\alpha_R-1}\left(1-p_R\right)^{\beta_R-1}}{B\left(\alpha_R,\beta_R\right)}dp_Bdp_Sdp_R.
\end{aligned} \tag{B.93}$$

Extract constant terms and combine common terms. Separate out the independent integrals to get

$$\begin{aligned} p_{\bar{M}\bar{S}} = {} & \frac{1}{ZB\left(\alpha_B, \beta_B\right) B\left(\alpha_S, \beta_S\right) B\left(\alpha_R, \beta_R\right)} \binom{n_B}{d_B} \binom{n_S}{d_S} \binom{n_R}{d_R} \times \\ & \int_0^1 p_B^{d_B+\alpha_B-1} \left(1-p_B\right)^{n_B-d_B+\beta_B-1} dp_B \int_0^1 p_S^{d_S+\alpha_S-1} \left(1-p_S\right)^{n_S-d_S+\beta_S-1} dp_S \\ & \int_0^1 p_R^{d_R+\alpha_R-1} \left(1-p_R\right)^{n_R-d_R+\beta_R-1} dp_R. \end{aligned} \tag{B.94}$$

All three integrals are solved with Gradshteyn and Ryzhik 3.191.1 [13], resulting in

$$\begin{aligned} p_{\bar{M}\bar{S}} = {} & \frac{1}{ZB\left(\alpha_B, \beta_B\right) B\left(\alpha_S, \beta_S\right) B\left(\alpha_R, \beta_R\right)} \binom{n_B}{d_B} \binom{n_S}{d_S} \binom{n_R}{d_R} \times \\ & B\left(d_B+\alpha_B, n_B-d_B+\beta_B\right) B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right) B\left(d_R+\alpha_R, n_R-d_R+\beta_R\right). \end{aligned} \tag{B.95}$$

B.3.5 Common Terms in the Binomial Probabilities

The common terms in the four probabilities are

$$C_p = \frac{1}{ZB\left(\alpha_B, \beta_B\right) B\left(\alpha_S, \beta_S\right)} \binom{n_B}{d_B} \binom{n_S}{d_S}. \tag{B.96}$$

These cancel in the Bayes factor. The four terms can be then be defined as

$$
\begin{aligned}
p'_{MS} = {} & \Gamma\left(n_R+1\right)\sum_{t=0}^{d_R}\frac{1}{\Gamma\left(d_R-t+1\right)}\times \\
& \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(n_S-d_S+\beta_S, d_S+\alpha_S+t+u\right)\times \\
& \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B, d_B+\alpha_B+d_R-t+v\right)\times \\
& \frac{B\left(d_R-t+v+1, t+u+1\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)}, \\
p'_{\bar{M}S} = {} & \binom{n_R}{d_R} B\left(d_S+d_R+\alpha_S, n_S-d_S+n_R-d_R+\beta_S\right) B\left(d_B+\alpha_B, n_B-d_B+\beta_B\right), \\
p'_{M\bar{S}} = {} & \frac{\Gamma\left(n_R+1\right)}{B\left(\alpha_R,\beta_R\right)} B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right)\sum_{t=0}^{d_R}\frac{1}{\Gamma\left(d_R-t+1\right)} \\
& \sum_{u=0}^{n_R-d_R}\left(-1\right)^u B\left(\beta_R, \alpha_R+t+u\right)\times \\
& \sum_{v=0}^{n_R-d_R-u}\left(-1\right)^v B\left(n_B-d_B+\beta_B, d_B+\alpha_B+d_R-t+v\right) \\
& \frac{B\left(d_R-t+v+1, t+u+1\right)}{\Gamma\left(v+1\right)\Gamma\left(n_R-d_R-u-v+1\right)}, \\
p'_{\bar{M}\bar{S}} = {} & \binom{n_R}{d_R}\frac{1}{B\left(\alpha_R,\beta_R\right)} B\left(d_S+\alpha_S, n_S-d_S+\beta_S\right) B\left(d_B+\alpha_B, n_B-d_B+\beta_B\right) \\
& B\left(d_R+\alpha_R, n_R-d_R+\beta_R\right).
\end{aligned}
\tag{B.97}
$$

B.4 MULTINOMIAL MERGERS AND SPLITS

Although the similarity measures between samples from both the binomial and multinomial probability functions could be derived from a single description, the derivations for mergers is significantly more complicated. This complexity arises from modeling how binomial and multinomial functions merge. The Poisson function is easily modeled because the distribution parameters are simply added together for the merged parameter. Because the binomial and multinomial functions model distributions of occurrences across two or more categories, the merged distributions are dependent upon the relative contribution of two distributions to the merged distribution. In general, this would be unknown for combinations of multinomial samples. In specific cases, the means used to select the samples can provide information to restrict the possible combinations of the merged samples. This derivation will assume that the mixture of generative distributions is unknown.

The multinomial probability density function is

$$P\left(\boldsymbol{d};\boldsymbol{p}\right)=\frac{\Gamma\left(\left(\sum_{i=1}^{k}\boldsymbol{d}_i\right)+1\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{d}_i+1\right)}\prod_{i=1}^{k}\boldsymbol{p}_i^{\boldsymbol{d}_i};\quad \sum_{i=1}^{k}\boldsymbol{p}_i=1. \tag{B.98}$$

The conjugate prior for the multinomial density function is a Dirichet distribution

$$P\left(\boldsymbol{\alpha};\boldsymbol{p}\right)=\frac{\Gamma\left(\sum_{i=1}^{k}\boldsymbol{\alpha}_i\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{\alpha}_i\right)}\prod_{i=1}^{k}\boldsymbol{p}_i^{\boldsymbol{\alpha}_i-1};\quad \sum_{i=1}^{k}\boldsymbol{p}_i=1. \tag{B.99}$$

The merger operation selected for this derivation is a mixing between the multinomial probability vectors for the two separate samples,

$$\boldsymbol{p}_R=\gamma\boldsymbol{p}_B+\left(1-\gamma\right)\boldsymbol{p}_S;\quad 0\leq\gamma\leq 1. \tag{B.100}$$

The mixing parameter γ is used to combine the two generators. This is equivalent to the relative likelihood of choosing one generator over another for the creation of the next sample.

B.4.1 Multinomial Probability of Merger with the Same Generator

The posterior probability for a merger where the generator of the destination group does not change is

$$p_{MS}=\frac{1}{Z_{MS}}\int_{\Theta_B}\int_{\Theta_S}\prod_{i=1}^{k}\boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1}\boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{\alpha}_{Si}-1}\left(\gamma\boldsymbol{p}_{Bi}+\left(1-\gamma\right)\boldsymbol{p}_{Si}\right)^{\boldsymbol{d}_{Ri}}d\Theta_B d\Theta_S, \tag{B.101}$$

where Z_{MS} is a normalization term,

$$\frac{1}{Z_{MS}}=\frac{\Gamma\left(1+\sum_{i=1}^{k}\boldsymbol{d}_{Bi}\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{d}_{Bi}+1\right)}\frac{\Gamma\left(1+\sum_{i=1}^{k}\boldsymbol{d}_{Si}\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{d}_{Si}+1\right)}\frac{\Gamma\left(1+\sum_{i=1}^{k}\boldsymbol{d}_{Ri}\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{d}_{Ri}+1\right)}\frac{\Gamma\left(\sum_{i=1}^{k}\boldsymbol{\alpha}_{Bi}\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{\alpha}_{Bi}\right)}\frac{\Gamma\left(\sum_{i=1}^{k}\boldsymbol{\alpha}_{Si}\right)}{\prod_{i=1}^{k}\Gamma\left(\boldsymbol{\alpha}_{Si}\right)}. \tag{B.102}$$

The derivation of the integral over the probability simplex is given in Appendices B.5.3 and B.5.4 for Equation B.101. The result of the derivation is given in Equation B.228 and with appropriate replacement of variables is

$$
\begin{aligned}
p_{MS} = \frac{1}{Z_{MS}} & \sum_{t_{k-1}=0}^{\boldsymbol{d}_{R_{k-1}}} \sum_{u_{k-1}=0}^{\boldsymbol{d}_{R_k}} \sum_{v_{k-1}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}} \Bigg(\sum_{t_{k-2}=0}^{\boldsymbol{d}_{R_{k-2}}} \sum_{u_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}} \sum_{v_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}-u_{k-2}} \Bigg(\times \\
& \cdots \Bigg(\sum_{t_1=0}^{\boldsymbol{d}_{R_1}} \sum_{u_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)} \sum_{v_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)-u_1} \times \\
& \Bigg(\left[(-1)^{\sum_{r=1}^{k-1} u_r+v_r} \right] \gamma^{\sum_{r=1}^{k-1} \boldsymbol{d}_{R_r}-t_r+v_r} (1-\gamma)^{\sum_{r=1}^{k-1} t_r+u_r} \times \\
& \frac{\Gamma\left(\boldsymbol{d}_{B_k}+\boldsymbol{\alpha}_{B_k}\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{d}_{B_q}+\boldsymbol{\alpha}_{B_q}+\boldsymbol{d}_{R_q}-t_q+v_q\right)}{\Gamma\left(\boldsymbol{d}_{B_k}+\boldsymbol{\alpha}_{B_k}+\left(\sum_{q=1}^{k-1} \boldsymbol{d}_{B_q}+\boldsymbol{\alpha}_{B_q}+\boldsymbol{d}_{R_q}-t_q+v_q\right)\right)} \times \qquad\qquad \text{(B.103)} \\
& \frac{\Gamma\left(\boldsymbol{d}_{S_k}+\boldsymbol{\alpha}_{S_k}\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{d}_{S_r}+\boldsymbol{\alpha}_{S_r}+t_r+u_r\right)}{\Gamma\left(\boldsymbol{d}_{S_k}+\boldsymbol{\alpha}_{S_k}+\sum_{r=1}^{k-1} \boldsymbol{d}_{S_r}+\boldsymbol{\alpha}_{S_r}+t_r+u_r\right)} \times \\
& \left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_{R_r}+1\right)}{\Gamma\left(\boldsymbol{d}_{R_r}-t_r+1\right)} \right] \frac{\Gamma\left(\boldsymbol{d}_{R_k}+1\right)}{\Gamma\left(\boldsymbol{d}_{R_k}-\left(\sum_{q=1}^{k-1} u_q+v_q\right)+1\right) \prod_{r=1}^{k-1} \Gamma\left(v_r+1\right)} \Bigg) \Bigg) \cdots \Bigg) \Bigg).
\end{aligned}
$$

There are $k-1$ sets of triple sums that involve the values in the merged matrix of counts, $\boldsymbol{d}_R$. The result is a bit intimidating and numerical solutions for the equation are likely to be fraught with problems from round-off errors. This is especially true when $\boldsymbol{d}_R$ has very large counts within it. The dominant variable in the triple sums is $\boldsymbol{d}_{R_k}$ and it may be possible to select the smallest value from the matrix $\boldsymbol{d}_R$. This would significantly reduce the number of sums.

Integration to eliminate γ can now be performed. Assuming that the prior for γ is a two-term Dirichlet distribution, the prior parameters for γ are α_{γ_1} and α_{γ_2} and using the integral for γ in

Equation B.233, the result is

$$\begin{aligned}
p_{MS} = \frac{1}{Z_{MS}} &\sum_{t_{k-1}=0}^{\boldsymbol{d}_{R_{k-1}}} \sum_{u_{k-1}=0}^{\boldsymbol{d}_{R_k}} \sum_{v_{k-1}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}} \Bigg(\sum_{t_{k-2}=0}^{\boldsymbol{d}_{R_{k-2}}} \sum_{u_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}} \sum_{v_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}-u_{k-2}} \Bigg(\times \\
&\cdots \Bigg(\sum_{t_1=0}^{\boldsymbol{d}_{R_1}} \sum_{u_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)} \sum_{v_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)-u_1} \times \\
&\Bigg(\left[(-1)^{\sum_{r=1}^{k-1} u_r+v_r}\right] \frac{B\left(\sum_{r=1}^{k-1} \boldsymbol{d}_{R_r} - t_r + v_r + \alpha_{\gamma_1}, \sum_{r=1}^{k-1} t_r + u_r + \alpha_{\gamma_2}\right)}{B\left(\alpha_{\gamma_1}, \alpha_{\gamma_2}\right)} \times \\
&\frac{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k}\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)}{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k} + \left(\sum_{q=1}^{k-1} \boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)\right)} \times \\
&\frac{\Gamma\left(\boldsymbol{d}_{S_k} + \boldsymbol{\alpha}_{S_k}\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{d}_{S_r} + \boldsymbol{\alpha}_{S_r} + t_r + u_r\right)}{\Gamma\left(\boldsymbol{d}_{S_k} + \boldsymbol{\alpha}_{S_k} + \sum_{r=1}^{k-1} \boldsymbol{d}_{S_r} + \boldsymbol{\alpha}_{S_r} + t_r + u_r\right)} \times \\
&\left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_{R_r} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_r} - t_r + 1\right)}\right] \frac{\Gamma\left(\boldsymbol{d}_{R_k} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_k} - \left(\sum_{q=1}^{k-1} u_q + v_q\right) + 1\right) \prod_{r=1}^{k-1} \Gamma\left(v_r + 1\right)} \Bigg) \Bigg) \cdots \Bigg) \Bigg).
\end{aligned} \tag{B.104}$$

B.4.2 Multinomial Probability of No Merger with the Same Generator

This integral is simpler:

$$p_{\bar{M}S} = \frac{1}{Z_{\bar{M}S}} \int_{\Theta_B} \int_{\Theta_S} \prod_{i=1}^{k} \boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{\alpha}_{Si}-1} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Ri}} d\Theta_B d\Theta_S, \tag{B.105}$$

where

$$Z_{\bar{M}S} = Z_{MS}. \tag{B.106}$$

This can be rewritten as

$$p_{\bar{M}S} = \frac{1}{Z_{\bar{M}S}} \int_{\Theta_B} \int_{\Theta_S} \prod_{i=1}^{k} \boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{d}_{Ri}+\boldsymbol{\alpha}_{Si}-1} d\Theta_B d\Theta_S. \tag{B.107}$$

This is separable into two independent integrals,

$$p_{\bar{M}S} = \frac{1}{Z_{\bar{M}S}} \int_{\Theta_B} \prod_{i=1}^{k} \boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1} d\Theta_B \int_{\Theta_S} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{d}_{Ri}+\boldsymbol{\alpha}_{Si}-1} d\Theta_S. \tag{B.108}$$

These are both solved with the general integral of the probability simplex with Equation B.133,

$$p_{\bar{M}S} = \frac{1}{Z_{\bar{M}S}} G_I\left(\boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right) G_I\left(\boldsymbol{d}_{Si} + \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Si}\right). \tag{B.109}$$

The probability of no merger where the generative model for the potential destination group remains the same across the two time periods is

$$p_{\bar{M}S} = \frac{1}{Z_{\bar{M}S}} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Si} + \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Si}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Si} + \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Si}\right)}. \tag{B.110}$$

B.4.3 Probability of Merger with a Different Generator

It is assumed for this integral that the destination dataset is drawn from a different multinomial generative model for the merged pair. This results in

$$p_{M\bar{S}} = \frac{1}{Z_{M\bar{S}}} \int_{\Theta_B} \int_{\Theta_S} \int_{\Theta_R} \prod_{i=1}^{k} \boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{\alpha}_{Si}-1} \left(\gamma \boldsymbol{p}_{Bi} + (1-\gamma)\, \boldsymbol{p}_{Ri}\right)^{\boldsymbol{d}_{Ri}} \boldsymbol{p}_{Ri}^{\boldsymbol{\alpha}_{Ri}-1} d\Theta_B d\Theta_S d\Theta_R, \tag{B.111}$$

where $Z_{M\bar{S}}$ is a normalization term,

$$\begin{aligned} \frac{1}{Z_{M\bar{S}}} = & \frac{\Gamma\left(1 + \sum_{i=1}^{k} \boldsymbol{d}_{Bi}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi} + 1\right)} \frac{\Gamma\left(1 + \sum_{i=1}^{k} \boldsymbol{d}_{Si}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Si} + 1\right)} \frac{\Gamma\left(1 + \sum_{i=1}^{k} \boldsymbol{d}_{Ri}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ri} + 1\right)} \times \\ & \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{Bi}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_{Bi}\right)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{Si}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_{Si}\right)} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{Ri}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_{Ri}\right)}. \end{aligned} \tag{B.112}$$

This normalization term $Z_{M\bar{S}}$ is related to Z_{MS} by

$$\frac{1}{Z_{M\bar{S}}} = \frac{1}{Z_{MS}} \frac{\Gamma\left(\sum_{i=1}^{k} \boldsymbol{\alpha}_{Ri}\right)}{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{\alpha}_{Ri}\right)}. \tag{B.113}$$

The triple integral separates into a product of two independent integrals that are a single integral and a double integral. Both integrals are in the forms of previously derived integrals and only require the insertion of the correct terms:

$$\begin{aligned} p_{M\bar{S}} = & \frac{1}{Z_{M\bar{S}}} \int_{\Theta_S} \prod_{i=1}^{k} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{\alpha}_{Si}-1} d\Theta_S \times \\ & \int_{\Theta_B} \int_{\Theta_R} \prod_{i=1}^{k} \boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1} \left(\gamma \boldsymbol{p}_{Bi} + (1-\gamma)\, \boldsymbol{p}_{Ri}\right)^{\boldsymbol{d}_{Ri}} \boldsymbol{p}_{Ri}^{\boldsymbol{\alpha}_{Ri}-1} d\Theta_B d\Theta_R. \end{aligned} \tag{B.114}$$

The first integral is solved with G_I, Equation B.133. The second integral is solved with G_{II}, Equation B.228:

$$p_{M\bar{S}} = \frac{1}{Z_{M\bar{S}}} G_I\left(\boldsymbol{d}_S + \boldsymbol{\alpha}_S\right) G_{II}\left(\boldsymbol{d}_B + \boldsymbol{\alpha}_B, \boldsymbol{\alpha}_R, \boldsymbol{d}_R\right). \tag{B.115}$$

This expands to

$$\begin{aligned}
p_{M\bar{S}} = {} & \frac{1}{Z_{M\bar{S}}} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{S_i} + \boldsymbol{\alpha}_{S_i}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{S_j} + \boldsymbol{\alpha}_{S_j}\right)} \\
& \sum_{t_{k-1}=0}^{\boldsymbol{d}_{R_{k-1}}} \sum_{u_{k-1}=0}^{\boldsymbol{d}_{R_k}} \sum_{v_{k-1}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}} \Bigg(\sum_{t_{k-2}=0}^{\boldsymbol{d}_{R_{k-2}}} \sum_{u_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}} \sum_{v_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}-u_{k-2}} \Bigg(\times \\
& \cdots \Bigg(\sum_{t_1=0}^{\boldsymbol{d}_{R_1}} \sum_{u_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)} \sum_{v_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)-u_1} \times \\
& \Bigg(\left[(-1)^{\sum_{r=1}^{k-1} u_r+v_r}\right] \gamma^{\sum_{r=1}^{k-1} \boldsymbol{d}_{R_r}-t_r+v_r} \left(1-\gamma\right)^{\sum_{r=1}^{k-1} t_r+u_r} \times \\
& \frac{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k}\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)}{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k} + \left(\sum_{q=1}^{k-1} \boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)\right)} \frac{\Gamma\left(\boldsymbol{\alpha}_{R_k}\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{\alpha}_{R_r} + t_r + u_r\right)}{\Gamma\left(\boldsymbol{\alpha}_{R_k} + \sum_{r=1}^{k-1} \boldsymbol{\alpha}_{R_r} + t_r + u_r\right)} \times \\
& \left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_{R_r} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_r} - t_r + 1\right)}\right] \frac{\Gamma\left(\boldsymbol{d}_{R_k} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_k} - \left(\sum_{q=1}^{k-1} u_q + v_q\right) + 1\right) \prod_{r=1}^{k-1} \Gamma\left(v_r + 1\right)} \Bigg) \Bigg) \cdots \Bigg) \Bigg).
\end{aligned} \tag{B.116}$$

It is assumed that the prior distribution function for the γ term is a two-parameter Dirichlet distribution, the prior parameters for γ are represented with α_{γ_1} and α_{γ_2}. Inserting the prior distribution function and using the integral for γ in Equations B.233, the result is

$$\begin{aligned}
p_{M\bar{S}} = {} & \frac{1}{Z_{M\bar{S}}} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{S_i} + \boldsymbol{\alpha}_{S_i}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{S_j} + \boldsymbol{\alpha}_{S_j}\right)} \\
& \sum_{t_{k-1}=0}^{\boldsymbol{d}_{R_{k-1}}} \sum_{u_{k-1}=0}^{\boldsymbol{d}_{R_k}} \sum_{v_{k-1}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}} \Bigg(\sum_{t_{k-2}=0}^{\boldsymbol{d}_{R_{k-2}}} \sum_{u_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}} \sum_{v_{k-2}=0}^{\boldsymbol{d}_{R_k}-u_{k-1}-v_{k-1}-u_{k-2}} \Bigg(\times \\
& \cdots \Bigg(\sum_{t_1=0}^{\boldsymbol{d}_{R_1}} \sum_{u_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)} \sum_{v_1=0}^{\boldsymbol{d}_{R_k}-\left(\sum_{r=2}^{k-1} u_r+v_r\right)-u_1} \times \\
& \Bigg(\left[(-1)^{\sum_{r=1}^{k-1} u_r+v_r}\right] \frac{B\left(\sum_{r=1}^{k-1} \boldsymbol{d}_{R_r} - t_r + v_r + \alpha_{\gamma_1}, \sum_{r=1}^{k-1} t_r + u_r + \alpha_{\gamma_2}\right)}{B\left(\alpha_{\gamma_1}, \alpha_{\gamma_2}\right)} \times \\
& \frac{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k}\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)}{\Gamma\left(\boldsymbol{d}_{B_k} + \boldsymbol{\alpha}_{B_k} + \left(\sum_{q=1}^{k-1} \boldsymbol{d}_{B_q} + \boldsymbol{\alpha}_{B_q} + \boldsymbol{d}_{R_q} - t_q + v_q\right)\right)} \frac{\Gamma\left(\boldsymbol{\alpha}_{R_k}\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{\alpha}_{R_r} + t_r + u_r\right)}{\Gamma\left(\boldsymbol{\alpha}_{R_k} + \sum_{r=1}^{k-1} \boldsymbol{\alpha}_{R_r} + t_r + u_r\right)} \times \\
& \left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_{R_r} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_r} - t_r + 1\right)}\right] \frac{\Gamma\left(\boldsymbol{d}_{R_k} + 1\right)}{\Gamma\left(\boldsymbol{d}_{R_k} - \left(\sum_{q=1}^{k-1} u_q + v_q\right) + 1\right) \prod_{r=1}^{k-1} \Gamma\left(v_r + 1\right)} \Bigg) \Bigg) \cdots \Bigg) \Bigg).
\end{aligned} \tag{B.117}$$

B.4.4 Probability of No Merger with a Different Generator

This integral is straightforward because all terms separate into three independent integrals:

$$p_{M\bar{S}} = \frac{1}{Z_{\bar{M}\bar{S}}} \int_{\Theta_B} \int_{\Theta_S} \int_{\Theta_R} \prod_{i=1}^{k} \boldsymbol{p}_{Bi}^{\boldsymbol{d}_{Bi}+\boldsymbol{\alpha}_{Bi}-1} \boldsymbol{p}_{Si}^{\boldsymbol{d}_{Si}+\boldsymbol{\alpha}_{Si}-1} \boldsymbol{p}_{Ri}^{\boldsymbol{d}_{Ri}+\boldsymbol{\alpha}_{Ri}-1} d\Theta_B d\Theta_S d\Theta_R, \tag{B.118}$$

where $Z_{\bar{M}\bar{S}}$ is equal to $Z_{M\bar{S}}$.

The probability of no merger when the generative model for the destination distribution changes across the two time periods is

$$p_{\bar{M}S} = \frac{1}{Z_{\bar{M}\bar{S}}} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Bi} + \boldsymbol{\alpha}_{Bi}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Si} + \boldsymbol{\alpha}_{Si}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Si} + \boldsymbol{\alpha}_{Si}\right)} \frac{\prod_{i=1}^{k} \Gamma\left(\boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Ri}\right)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_{Ri} + \boldsymbol{\alpha}_{Ri}\right)}. \tag{B.119}$$

The four terms are combined to form the Bayes factor using the form of Equation 17, which are summarized in Section 4.3.

B.5 GENERAL FORMULAS FOR DOUBLE INTEGRALS

This appendix section includes general integrals required for the merge/split integrals.

B.5.1 General Integral I

The first general integral that is required is Gradshteyn and Ryzhik 3.191.1 [13]:

$$g_I(\mu, \nu) = \int_0^u x^{\nu-1} (u-x)^{\mu-1} \, dx = u^{\mu+\nu-1} B(\mu, \nu). \tag{B.120}$$

B.5.2 General Integral I Expanded Over a Probability Simplex

The integrals of interest in the main body of the paper are of the form of Equation B.120, but performed over a probability simplex Θ. The typical equation to solve is

$$G_I(\boldsymbol{d}) = \int_{\Theta} \prod_{i=1}^{k} \boldsymbol{x}_i^{\boldsymbol{d}_i - 1} d\Theta. \tag{B.121}$$

Here, d is a k-dimensional vector of counts or counts with prior terms (reals). The variable $\boldsymbol{x}$ is also a k-dimensional vector and is used here to represent a categorical probability distribution that must satisfy the equation

$$\sum_{i=1}^{k} \boldsymbol{x}_i = 1. \tag{B.122}$$

The probability simplex then requires that the integrals over individual terms satisfy this condition, and the integral expands out to

$$G_I = \int_0^1 \int_0^{1-\boldsymbol{x}_1} \int_0^{1-\sum_{i=1}^{2} \boldsymbol{x}_i} \cdots \int_0^{1-\sum_{i=1}^{k-1} \boldsymbol{x}_i} \prod_{i=1}^{k-1} \boldsymbol{x}_i^{\boldsymbol{d}_i - 1} \left(1 - \sum_{j=1}^{k-1} \boldsymbol{x}_i\right)^{\boldsymbol{d}_k - 1} d\boldsymbol{x}_1 d\boldsymbol{x}_2 \cdots d\boldsymbol{x}_{k-1} \tag{B.123}$$

over the probability simplex. For simplification, define

$$\boldsymbol{u}_j = 1 - \sum_{i=1}^{j-1} \boldsymbol{x}_i, \tag{B.124}$$

which has the relationships

$$\begin{aligned} \boldsymbol{u}_1 &= 1; [j = 1], && \text{(B.125)} \\ \boldsymbol{u}_j &= \boldsymbol{u}_{j-1} - \boldsymbol{x}_j; [1 < j < k]. && \text{(B.126)} \end{aligned}$$

Substitution of these definitions into the integral results in

$$G_I = \int_0^{\boldsymbol{u}_1} \int_0^{\boldsymbol{u}_2} \int_0^{\boldsymbol{u}_3} \cdots \left[\prod_{i=1}^{k-2} \boldsymbol{x}_i^{\boldsymbol{d}_i - 1}\right] \int_0^{\boldsymbol{u}_{k-1}} \boldsymbol{x}_{k-1}^{\boldsymbol{d}_{k-1}-1} \left(\boldsymbol{u}_{k-1} - \boldsymbol{x}_{k-1}\right)^{\boldsymbol{d}_k - 1} d\boldsymbol{x}_1 d\boldsymbol{x}_2 \cdots d\boldsymbol{x}_{k-1}. \tag{B.127}$$

Note that the k-th integral is fully constrained because the value $\boldsymbol{x}_k = 1 - \sum_{i=1}^{k-1} \boldsymbol{x}_i$ is equivalent to integration of a delta function. The integral for $\boldsymbol{x}_{k-1}$ is solved with Gradshteyn and Ryzhik 3.191.1 [13] to get

$$G_I = \int_0^{\boldsymbol{u}_1} \int_0^{\boldsymbol{u}_2} \int_0^{\boldsymbol{u}_3} \cdots \int_0^{\boldsymbol{u}_{k-2}} \left[\prod_{i=1}^{k-2} \boldsymbol{x}_i^{\boldsymbol{d}_i - 1}\right] \boldsymbol{u}_{k-1}^{\boldsymbol{d}_k + \boldsymbol{d}_{k-1} - 1} B\left(\boldsymbol{d}_{k-1}, \boldsymbol{d}_k\right) d\boldsymbol{x}_1 d\boldsymbol{x}_2 \cdots d\boldsymbol{x}_{k-2}. \tag{B.128}$$

Expansion of $\boldsymbol{u}_{k-1}$ results in the next integral,

$$\begin{aligned} G_I = B\left(\boldsymbol{d}_{k-1}, \boldsymbol{d}_k\right) \int_0^{\boldsymbol{u}_1} \int_0^{\boldsymbol{u}_2} \int_0^{\boldsymbol{u}_3} \cdots \left[\prod_{i=1}^{k-3} \boldsymbol{x}_i^{\boldsymbol{d}_i - 1}\right] \\ \int_0^{\boldsymbol{u}_{k-2}} \boldsymbol{x}_{k-2}^{\boldsymbol{d}_{k-2}-1} \left(\boldsymbol{u}_{k-2} - \boldsymbol{x}_{k-2}\right)^{\boldsymbol{d}_k + \boldsymbol{d}_{k-1} - 1} d\boldsymbol{x}_1 d\boldsymbol{x}_2 \cdots d\boldsymbol{x}_{k-2}, \end{aligned} \tag{B.129}$$

which is solved with the same formula to get

$$\begin{aligned} G_I = B\left(\boldsymbol{d}_{k-1}, \boldsymbol{d}_k\right) \int_0^{\boldsymbol{u}_1} \int_0^{\boldsymbol{u}_2} \int_0^{\boldsymbol{u}_3} \cdots \left[\prod_{i=1}^{k-3} \boldsymbol{x}_i^{\boldsymbol{d}_i - 1}\right] \\ \boldsymbol{u}_{k-2}^{\boldsymbol{d}_k + \boldsymbol{d}_{k-1} + \boldsymbol{d}_{k-2} - 1} B\left(\boldsymbol{d}_{k-2}, \boldsymbol{d}_k + \boldsymbol{d}_{k-1}\right) d\boldsymbol{x}_1 d\boldsymbol{x}_2 \cdots d\boldsymbol{x}_{k-3}. \end{aligned} \tag{B.130}$$

Continuation of the process to complete the remaining integrals results in

$$G_I = \prod_{i=1}^{k-1} B\left(\boldsymbol{d}_i, \sum_{j=i+1}^{k} \boldsymbol{d}_j\right). \tag{B.131}$$

The beta function can be expanded into terms of Γ functions,

$$G_I = \prod_{i=1}^{k-1} \frac{\Gamma\left(\boldsymbol{d}_i\right) \Gamma\left(\sum_{j=i+1}^{k} \boldsymbol{d}_j\right)}{\Gamma\left(\sum_{j=i}^{k} \boldsymbol{d}_j\right)}. \tag{B.132}$$

Pairs of the Γ functions that contain sums cancel across the product of fractions so that the result is

$$G_I(\boldsymbol{d}) = \frac{\prod_{i=1}^{k} \Gamma(\boldsymbol{d}_i)}{\Gamma\left(\sum_{j=1}^{k} \boldsymbol{d}_j\right)}. \tag{B.133}$$

B.5.3 General Double Integral II

The following general double integral is useful in the derivation of merging and splitting probability estimates:

$$\begin{aligned} g_{II}(a,b,c,d,e,f,u_x,u_y) = \int_0^{u_x} x^{d-1}(u_x - x)^{a-1} \times \\ \int_0^{u_y} y^{e-1}(u_y - y)^{b-1}(\gamma x + (1-\gamma)y)^f(\gamma(u_x - x) + (1-\gamma)(u_y - y))^c \, dydx. \end{aligned} \tag{B.134}$$

Terms f and c are defined to be integers.

The y integral solution is available using Gradshteyn and Ryzhik, 3.211 [13],

$$\begin{aligned} \int_0^1 x^{\lambda-1}(1-x)^{\mu-1}(1-ux)^{-\rho}(1-vx)^{-\sigma}\,dx = B(\mu,\lambda)\,F_1(\lambda,\rho,\sigma,\lambda+\mu;u,v)\,; \\ [\mathrm{Re}\ \lambda > 0, \mathrm{Re}\ \mu > 0]\,, \end{aligned} \tag{B.135}$$

where $B(\mu,\lambda))$ is the beta function, and $F_1(\lambda,\rho,\sigma,\lambda+\mu;u,v)$ is a hypergeometric function in two parameters. This function is defined in Gradshteyn and Ryzhik 9.180.1 [13] to be

$$F_1(\alpha,\beta,\beta',\gamma;x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{(\alpha)_{m+n}(\beta)_m(\beta')_n}{(\gamma)_{m+n}} x^m y^n. \tag{B.136}$$

Pochhammer notation is used in the formula, where

$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}. \tag{B.137}$$

Define

$$g_{II,y} = \int_0^{u_y} y^{e-1}(u_y - y)^{b-1}(\gamma x + (1-\gamma)y)^f(\gamma(u_x - x) + (1-\gamma)(u_y - y))^c\,dy \tag{B.138}$$

and

$$y = u_y z. \tag{B.139}$$

Then

$$g_{II,y} = (u_y)^{e+b-1}\int_0^1 z^{e-1}(1-z)^{b-1}(\gamma x + (1-\gamma)u_y z)^f(\gamma(u_x - x) + (1-\gamma)u_y(1-z))^c\,dz. \tag{B.140}$$

Continuing the transformation leads to

$$g_{II,y} = \left(u_y\right)^{e+b-1} \left(\gamma x\right)^f \int_0^1 z^{e-1} \left(1-z\right)^{b-1} \left(1 + \frac{\left(1-\gamma\right) u_y}{\gamma x} z\right)^f \times \\ \left(\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y - \left(1-\gamma\right) u_y z\right)^c dz \tag{B.141}$$

and

$$g_{II,y} = \left(u_y\right)^{e+b-1} \left(\gamma x\right)^f \left(\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y\right)^c \\ \int_0^1 z^{e-1} \left(1-z\right)^{b-1} \left(1 + \frac{\left(1-\gamma\right) u_y}{\gamma x} z\right)^f \left(1 - \frac{\left(1-\gamma\right) u_y}{\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y} z\right)^c dz. \tag{B.142}$$

Comparing this equation with Equation B.135, the following correspondences can be made:

$$\lambda = e, \tag{B.143}$$

$$\mu = b, \tag{B.144}$$

$$\rho = -f, \tag{B.145}$$

$$\sigma = -c, \tag{B.146}$$

$$u = -\frac{\left(1-\gamma\right) u_y}{\gamma x}, \tag{B.147}$$

$$v = \frac{\left(1-\gamma\right) u_y}{\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y}, \tag{B.148}$$

which means that

$$g_{II,y} = \left(u_y\right)^{e+b-1} \left(\gamma x\right)^f \left(\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y\right)^c \\ B\left(b,e\right) F_1 \left(e, -f, -c, e+b; -\frac{\left(1-\gamma\right) u_y}{\gamma x}, \frac{\left(1-\gamma\right) u_y}{\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y}\right). \tag{B.149}$$

This result can be inserted into the equation for g_{II} to get

$$g_{II} = \left(u_y\right)^{e+b-1} B\left(b,e\right) \left(\gamma\right)^f \int_0^{u_x} x^{d-1} \left(u_x - x\right)^{a-1} \left(x\right)^f \left(\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y\right)^c \\ F_1 \left(e, -f, -c, e+b; -\frac{\left(1-\gamma\right) u_y}{\gamma x}, \frac{\left(1-\gamma\right) u_y}{\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y}\right) dx. \tag{B.150}$$

After a little more clean up,

$$g_{II} = u_y^{e+b-1} B\left(b,e\right) \gamma^f \int_0^{u_x} x^{d+f-1} \left(u_x - x\right)^{a-1} \left(\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y\right)^c \\ F_1 \left(e, -f, -c, e+b; -\frac{\left(1-\gamma\right) u_y}{\gamma x}, \frac{\left(1-\gamma\right) u_y}{\gamma \left(u_x - x\right) + \left(1-\gamma\right) u_y}\right) dx. \tag{B.151}$$

It is the case that when f and c are negative integers, the F_1 function is a finite sum of polynomials with Pochhammer symbols. From Wolfram [15],

$$\left(a\right)_n = \frac{\left(-1\right)^n \left(-a\right)!}{\left(-a-n\right)!}; \quad a \leq 0, \; n \leq -a, \tag{B.152}$$

or with replacement of the negative a with a positive s,

$$(-s)_n = \frac{(-1)^n\,(s)!}{(s-n)!} = \frac{(-1)^n\,\Gamma(s+1)}{\Gamma(s-n+1)};\quad s \geq 0,\ n \leq s. \tag{B.153}$$

When $n > s$, the result is zero, which terminates the summation. This leads to F_1 being a finite sum of polynomials. For the specific case here,

$$F_1(e,-f,-c,e+b;-x,y) = \sum_{m=0}^{f}\sum_{n=0}^{c} \frac{(e)_{m+n}\,\Gamma(f+1)\,\Gamma(c+1)}{(e+b)_{m+n}\,\Gamma(f-m+1)\,\Gamma(c-n+1)} x^m\,(-y)^n. \tag{B.154}$$

The integral g_{II} is now

$$\begin{aligned} g_{II} = (u_y)^{e+b-1}\,B(b,e)\,\gamma^f \int_0^{u_x} x^{d+f-1}\,(u_x-x)^{a-1}\,(\gamma(u_x-x)+(1-\gamma)\,u_y)^c \times \\ \sum_{m=0}^{f}\sum_{n=0}^{c} \frac{(e)_{m+n}\,\Gamma(f+1)\,\Gamma(c+1)}{(e+b)_{m+n}\,\Gamma(f-m+1)\,\Gamma(c-n+1)} \times \\ \left(\frac{(1-\gamma)\,u_y}{\gamma x}\right)^m \left(-\frac{(1-\gamma)\,u_y}{\gamma(u_x-x)+(1-\gamma)\,u_y}\right)^n dx. \end{aligned} \tag{B.155}$$

The summations can be moved outside the integral and terms rearranged to move a number of them outside the integral,

$$\begin{aligned} g_{II} = (u_y)^{e+b-1}\,B(b,e)\,\gamma^f \sum_{m=0}^{f}\sum_{n=0}^{c} \frac{(e)_{m+n}\,\Gamma(f+1)\,\Gamma(c+1)}{(e+b)_{m+n}\,\Gamma(f-m+1)\,\Gamma(c-n+1)} \times \\ (-1)^n\,((1-\gamma)\,u_y)^{m+n}\,\gamma^{-m} \int_0^{u_x} x^{d+f-m-1}\,(u_x-x)^{a-1}\,(\gamma(u_x-x)+(1-\gamma)\,u_y)^{c-n}\,dx. \end{aligned} \tag{B.156}$$

The remaining integral can be represented with

$$g_{II,x} = (-\gamma)^{c-n} \int_0^{u_x} x^{d+f-m-1}\,(u_x-x)^{a-1} \left(x - \left(u_x + \frac{1-\gamma}{\gamma}u_y\right)\right)^{c-n} dx. \tag{B.157}$$

This integral can be reformulated using the replacement

$$x = u_x z. \tag{B.158}$$

The integral is then

$$g_{II,z} = (-\gamma)^{c-n} \int_0^1 u_x^{d+f-m-1} z^{d+f-m-1} u_x^{a-1}\,(1-z)^{a-1} \left(u_x z - \left(u_x + \frac{1-\gamma}{\gamma}u_y\right)\right)^{c-n} u_x dz. \tag{B.159}$$

Extract the u_x terms to get

$$g_{II,z} = (-\gamma)^{c-n}\,u_x^{d+f+a+c-m-n-1} \int_0^1 z^{d+f-m-1}\,(1-z)^{a-1} \left(z - \left(1 + \frac{1-\gamma}{\gamma u_x}u_y\right)\right)^{c-n} dz. \tag{B.160}$$

Reformat and pull out a multiplicative constant from the last term to get

$$\begin{aligned} g_{II,z} = (-\gamma)^{c-n}\, u_x^{d+f+a+c-m-n-1} \left(-\left(\frac{\gamma u_x + (1-\gamma)\,u_y}{\gamma u_x}\right)\right)^{c-n} \times \\ \int_0^1 z^{d+f-m-1}\,(1-z)^{a-1}\left(1-\left(\frac{\gamma u_x}{\gamma u_x + (1-\gamma)\,u_y}\right)z\right)^{c-n} dz. \end{aligned} \tag{B.161}$$

Cancel some terms to get

$$\begin{aligned} g_{II,z} = u_x^{d+f+a-m-1}\left(\gamma u_x + (1-\gamma)\,u_y\right)^{c-n} \times \\ \int_0^1 z^{d+f-m-1}\,(1-z)^{a-1}\left(1-\left(\frac{\gamma u_x}{\gamma u_x + (1-\gamma)\,u_y}\right)z\right)^{c-n} dz. \end{aligned} \tag{B.162}$$

The integral can now be solved with Gradshteyn and Ryzhik 3.197.3 [13] to get

$$\begin{aligned} \int_0^1 x^{\lambda-1}\,(1-x)^{\mu-1}\left(1-\beta' x\right)^{-\nu} dx = B(\lambda,\mu)\ {}_2F_1\left(\nu,\lambda;\lambda+\mu;\beta'\right) \\ \left[\operatorname{Re}\lambda > 0, \operatorname{Re}\mu > 0, \left|\beta'\right| < 1\right]. \end{aligned} \tag{B.163}$$

The mapping is

$$\lambda = d+f-m, \tag{B.164}$$

$$\mu = a, \tag{B.165}$$

$$\nu = -(c-n), \tag{B.166}$$

$$\beta' = \left(\frac{\gamma u_x}{\gamma u_x + (1-\gamma)\,u_y}\right), \tag{B.167}$$

so

$$\begin{aligned} g_{II,z} = u_x^{d+f+a-m-1}\left(\gamma u_x + (1-\gamma)\,u_y\right)^{c-n} \times \\ B(d+f-m,a)\ {}_2F_1\left(-(c-n), d+f-m; d+f-m+a; \left(\frac{\gamma u_x}{\gamma u_x + (1-\gamma)\,u_y}\right)\right). \end{aligned} \tag{B.168}$$

The result of the integration can be inserted back into the larger equation to give

$$\begin{aligned} g_{II} = u_y^{e+b-1} B(b,e)\,\gamma^f \sum_{m=0}^{f}\sum_{n=0}^{c} \frac{(e)_{m+n}\,\Gamma(f+1)\,\Gamma(c+1)}{(e+b)_{m+n}\,\Gamma(f-m+1)\,\Gamma(c-n+1)} \times \\ (-1)^n \left((1-\gamma)\,u_y\right)^{m+n} \gamma^{-m} \left(\gamma u_x + (1-\gamma)\,u_y\right)^{c-n} u_x^{a+d+f-m-1} B(a,d+f-m) \times \\ {}_2F_1\left(-(c-n), d+f-m; a+d+f-m; \frac{\gamma u_x}{\gamma u_x + (1-\gamma)\,u_y}\right). \end{aligned} \tag{B.169}$$

The series expansion of ${}_2F_1$ is

$${}_2F_1(a,b;c;z) = \sum_{p=0}^{\infty} \frac{(a)_p\,(b)_p}{(c)_p}\frac{x^p}{p!}. \tag{B.170}$$

When a is a negative integer,

$$ {}_2F_1\left(a,b;c;z\right) = \sum_{p=0}^{|a|} (-1)^p \binom{|a|}{p} \frac{(b)_p}{(c)_p} x^p. \tag{B.171}$$

Insert this function into Equation B.169 to get

$$\begin{aligned} g_{II} = u_y^{e+b-1} B\left(b,e\right) \gamma^f \sum_{m=0}^{f} \sum_{n=0}^{c} \frac{(e)_{m+n} \Gamma\left(f+1\right) \Gamma\left(c+1\right)}{(e+b)_{m+n} \Gamma\left(f-m+1\right) \Gamma\left(c-n+1\right)} \times \\ (-1)^n \left(\left(1-\gamma\right) u_y\right)^{m+n} \gamma^{-m} \left(\gamma u_x + \left(1-\gamma\right) u_y\right)^{c-n} u_x^{a+d+f-m-1} B\left(a,d+f-m\right) \times \\ \sum_{p=0}^{c-n} (-1)^p \binom{c-n}{p} \frac{(d+f-m)_p}{(a+d+f-m)_p} \left(\frac{\gamma u_x}{\gamma u_x + \left(1-\gamma\right) u_y}\right)^p . \end{aligned} \tag{B.172}$$

Collect terms and reorganize to get

$$\begin{aligned} g_{II}\left(a,b,c,d,e,f,\gamma,u_x,u_y\right) = B\left(b,e\right) \sum_{m=0}^{f} B\left(a,d+f-m\right) \frac{\Gamma\left(f+1\right)}{\Gamma\left(f-m+1\right)} \times \\ \sum_{n=0}^{c} (-1)^n \frac{(e)_{m+n} \Gamma\left(c+1\right)}{(e+b)_{m+n} \Gamma\left(c-n+1\right)} \left(1-\gamma\right)^{m+n} \times \\ \sum_{p=0}^{c-n} (-1)^p \binom{c-n}{p} \frac{(d+f-m)_p}{(a+d+f-m)_p} \gamma^{f-m+p} \times \\ u_x^{a+d+f-m+p-1} u_y^{b+e+m+n-1} \left(\gamma u_x + \left(1-\gamma\right) u_y\right)^{c-n-p} . \end{aligned} \tag{B.173}$$

In this arrangement, there are three terms in u_x and u_y, similar to the starting equation (B.134) for the general integral for g_{II}, but with different exponents.

Expand everything into gamma functions:

$$\begin{aligned} g_{II}\left(a,b,c,d,e,f,\gamma,u_x,u_y\right) = \frac{\Gamma\left(b\right) \Gamma\left(e\right)}{\Gamma\left(b+e\right)} \sum_{m=0}^{f} \frac{\Gamma\left(a\right) \Gamma\left(d+f-m\right)}{\Gamma\left(a+d+f-m\right)} \frac{\Gamma\left(f+1\right)}{\Gamma\left(f-m+1\right)} \times \\ \sum_{n=0}^{c} (-1)^n \frac{\Gamma\left(e+m+n\right) \Gamma\left(b+e\right) \Gamma\left(c+1\right)}{\Gamma\left(e\right) \Gamma\left(b+e+m+n\right) \Gamma\left(c-n+1\right)} \left(1-\gamma\right)^{m+n} \times \\ \sum_{p=0}^{c-n} (-1)^p \frac{\Gamma\left(c-n+1\right)}{\Gamma\left(p+1\right) \Gamma\left(c-n-p+1\right)} \times \\ \frac{\Gamma\left(d+f-m+p\right) \Gamma\left(a+d+f-m\right)}{\Gamma\left(d+f-m\right) \Gamma\left(a+d+f-m+p\right)} \times \\ \gamma^{f-m+p} u_x^{a+d+f-m+p-1} u_y^{b+e+m+n-1} \left(\gamma u_x + \left(1-\gamma\right) u_y\right)^{c-n-p} . \end{aligned} \tag{B.174}$$

Delete common terms to get

$$\begin{aligned} g_{II}\left(a,b,c,d,e,f,\gamma,u_x,u_y\right) = \sum_{m=0}^{f} \frac{\Gamma\left(f+1\right)}{\Gamma\left(f-m+1\right)} \times \\ \sum_{n=0}^{c} \left(-1\right)^n \frac{\Gamma\left(b\right)\Gamma\left(e+m+n\right)\Gamma\left(c+1\right)}{\Gamma\left(b+e+m+n\right)} \left(1-\gamma\right)^{m+n} \times \\ \sum_{p=0}^{c-n} \left(-1\right)^p \frac{1}{\Gamma\left(p+1\right)\Gamma\left(c-n-p+1\right)} \frac{\Gamma\left(a\right)\Gamma\left(d+f-m+p\right)}{\Gamma\left(a+d+f-m+p\right)} \gamma^{f-m+p} \times \\ u_x^{a+d+f-m+p-1} u_y^{b+e+m+n-1} \left(\gamma u_x + \left(1-\gamma\right)u_y\right)^{c-n-p}. \end{aligned} \tag{B.175}$$

Replace some terms with beta functions:

$$\begin{aligned} g_{II}\left(a,b,c,d,e,f,\gamma,u_x,u_y\right) = \sum_{m=0}^{f} \frac{\Gamma\left(f+1\right)}{\Gamma\left(f-m+1\right)} \times \\ \sum_{n=0}^{c} \left(-1\right)^n B\left(b,e+m+n\right) \left(1-\gamma\right)^{m+n} \times \\ \sum_{p=0}^{c-n} \left(-1\right)^p B\left(a,d+f-m+p\right) \frac{\Gamma\left(c+1\right)}{\Gamma\left(p+1\right)\Gamma\left(c-n-p+1\right)} \gamma^{f-m+p} \times \\ u_x^{a+d+f-m+p-1} u_y^{b+e+m+n-1} \left(\gamma u_x + \left(1-\gamma\right)u_y\right)^{c-n-p}. \end{aligned} \tag{B.176}$$

B.5.4 Reversed Order of Integration

The order of integration for g_{II} could be reversed, with integration over x performed before integration over y. The result can be determined from Equation B.174 with the following substitutions:

$$a \leftrightarrow b, \tag{B.177}$$

$$d \leftrightarrow e, \tag{B.178}$$

$$u_x \leftrightarrow u_y, \tag{B.179}$$

$$\gamma \leftrightarrow 1-\gamma, \tag{B.180}$$

which means that g_{II} can also be written as

$$\begin{aligned} g_{II}\left(a,b,c,d,e,f,\gamma,u_x,u_y\right) = \sum_{m=0}^{f} \frac{\Gamma\left(f+1\right)}{\Gamma\left(f-m+1\right)} \times \\ \sum_{n=0}^{c} \left(-1\right)^n B\left(a,d+m+n\right) \left(\gamma\right)^{m+n} \times \\ \sum_{p=0}^{c-n} \left(-1\right)^p B\left(b,e+f-m+p\right) \frac{\Gamma\left(c+1\right)}{\Gamma\left(p+1\right)\Gamma\left(c-n-p+1\right)} \left(1-\gamma\right)^{f-m+p} \times \\ u_y^{b+e+f-m+p-1} u_x^{a+d+m+n-1} \left(\gamma u_x + \left(1-\gamma\right)u_y\right)^{c-n-p}. \end{aligned} \tag{B.181}$$

B.5.5 General Integral II Expanded Over a Probability Simplex

As with general integral G_I, general integral G_{II} can be expanded to an integration over a probability simplex. The general integral is of the form of

$$G_{II}\left(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\right)=\int_{\Theta_x}\int_{\Theta_y}\prod_{i=1}^{k}\boldsymbol{x}_i^{\boldsymbol{a}_i-1}\boldsymbol{y}_i^{\boldsymbol{b}_i-1}\left(\gamma\boldsymbol{x}_i+\left(1-\gamma\right)\boldsymbol{y}_i\right)^{\boldsymbol{c}_i}d\Theta_x d\Theta_y, \tag{B.182}$$

with the same conditions as for general integral I in Equations B.122, which means that both

$$\boldsymbol{x}_k \;=\; 1-\sum_{i=1}^{k-1}\boldsymbol{x}_i, \tag{B.183}$$

$$\boldsymbol{y}_k \;=\; 1-\sum_{i=1}^{k-1}\boldsymbol{y}_i. \tag{B.184}$$

As in Equations B.124 through B.126, the following variables can be defined:

$$\boldsymbol{u}_{\boldsymbol{x}_j} \;=\; 1-\sum_{i=1}^{j-1}\boldsymbol{x}_j;\quad j\in\{1,2,3,\cdots,k\}, \tag{B.185}$$

$$\boldsymbol{u}_{\boldsymbol{y}_j} \;=\; 1-\sum_{i=1}^{j-1}\boldsymbol{y}_j;\quad j\in\{1,2,3,\cdots,k\}, \tag{B.186}$$

which both have the relationships

$$\boldsymbol{u}_{\boldsymbol{x}_1} \;=\; 1;\quad \left[j=1\right], \tag{B.187}$$

$$\boldsymbol{u}_{\boldsymbol{x}_j} \;=\; \boldsymbol{u}_{\boldsymbol{x}_{j-1}}-\boldsymbol{x}_j;\quad \left[j\in\{2,3,4,\cdots,k\}\right], \tag{B.188}$$

$$\boldsymbol{u}_{\boldsymbol{y}_1} \;=\; 1;\quad \left[j=1\right], \tag{B.189}$$

$$\boldsymbol{u}_{\boldsymbol{y}_j} \;=\; \boldsymbol{u}_{\boldsymbol{y}_{j-1}}-\boldsymbol{y}_j;\quad \left[j\in\{2,3,4,\cdots,k\}\right]. \tag{B.190}$$

The integral can now be written as

$$\begin{aligned}G_{II}=&\int_{\Theta_{\boldsymbol{x}_{k-2}}}\int_{\Theta_{\boldsymbol{y}_{k-2}}}\prod_{i=1}^{k-2}\boldsymbol{x}_i^{\boldsymbol{a}_i-1}\boldsymbol{y}_i^{\boldsymbol{b}_i-1}\left(\gamma\boldsymbol{x}_i+\left(1-\gamma\right)\boldsymbol{y}_i\right)^{\boldsymbol{c}_i}\times\\&\int_0^{\boldsymbol{u}_{\boldsymbol{x}_{k-1}}}\int_0^{\boldsymbol{u}_{\boldsymbol{y}_{k-1}}}\boldsymbol{x}_{k-1}^{\boldsymbol{a}_{k-1}-1}\boldsymbol{y}_{k-1}^{\boldsymbol{b}_{k-1}-1}\left(\gamma\boldsymbol{x}_{k-1}+\left(1-\gamma\right)\boldsymbol{y}_{k-1}\right)^{\boldsymbol{c}_{k-1}}\times\\&\left(\boldsymbol{u}_{\boldsymbol{x}_{k-1}}-\boldsymbol{x}_{k-1}\right)^{\boldsymbol{a}_k-1}\left(\boldsymbol{u}_{\boldsymbol{y}_{k-1}}-\boldsymbol{y}_{k-1}\right)^{\boldsymbol{b}_k-1}\times\\&\left(\gamma\left(\boldsymbol{u}_{\boldsymbol{x}_{k-1}}-\boldsymbol{x}_{k-1}\right)+\left(1-\gamma\right)\left(\boldsymbol{u}_{\boldsymbol{y}_{k-1}}-\boldsymbol{y}_{k-1}\right)\right)^{\boldsymbol{c}_k}d\boldsymbol{x}_{k-1}d\boldsymbol{y}_{k-1}d\Theta_{\boldsymbol{x}_{k-2}}d\Theta_{\boldsymbol{y}_{k-2}},\end{aligned} \tag{B.191}$$

where the integrals over $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are integrals of delta functions with values that are represented as $\boldsymbol{u}_{\boldsymbol{x}_{k-1}}-\boldsymbol{x}_{k-1}$ and $\boldsymbol{u}_{\boldsymbol{y}_{k-1}}-\boldsymbol{y}_{k-1}$, which follows from Equations B.185 through B.190. This means that

$$\begin{aligned} G_{II} = \int_{\Theta_{\boldsymbol{x}_{k-2}}} \int_{\Theta_{\boldsymbol{y}_{k-2}}} \prod_{i=1}^{k-2} \boldsymbol{x}_i^{\boldsymbol{a}_i-1} p_{\boldsymbol{y}_i}^{\boldsymbol{b}_i-1} \left(\gamma \boldsymbol{x}_i + (1-\gamma)\, \boldsymbol{y}_i\right)^{\boldsymbol{c}_i} \times \\ g_{II} \left(\boldsymbol{a}_k, \boldsymbol{b}_k, \boldsymbol{c}_k, \boldsymbol{a}_{k-1}, \boldsymbol{b}_{k-1}, \boldsymbol{c}_{k-1}, \boldsymbol{u}_{\boldsymbol{x}_{k-1}}, \boldsymbol{u}_{\boldsymbol{y}_{k-1}}\right) d\Theta_{\boldsymbol{x}_{k-2}} d\Theta_{\boldsymbol{y}_{k-2}}, \end{aligned} \tag{B.192}$$

where g_{II} is defined in Equation B.174, provided in Appendix B.5.3.

Equation B.174 can be used to expand out the g_{II} term

$$\begin{aligned} G_{II} = & \sum_{m_{k-1}=0}^{\boldsymbol{c}_{k-1}} \frac{\Gamma\left(\boldsymbol{c}_{k-1}+1\right)}{\Gamma\left(\boldsymbol{c}_{k-1}-m_{k-1}+1\right)} \times \\ & \sum_{n_{k-1}=0}^{\boldsymbol{c}_k} (-1)^{n_{k-1}}\, B\left(\boldsymbol{b}_k, \boldsymbol{b}_{k-1}+m_{k-1}+n_{k-1}\right) (1-\gamma)^{m_{k-1}+n_{k-1}} \times \\ & \sum_{p_{k-1}=0}^{\boldsymbol{c}_k-n_{k-1}} (-1)^{p_{k-1}} B\left(\boldsymbol{a}_k, \boldsymbol{a}_{k-1}+\boldsymbol{a}_{k-1}-m_{k-1}+p_{k-1}\right) \times \\ & \frac{\Gamma\left(\boldsymbol{c}_k+1\right)}{\Gamma\left(p_{k-1}+1\right)\Gamma\left(\boldsymbol{c}_k-n_{k-1}-p_{k-1}+1\right)} \gamma^{\boldsymbol{c}_{k-1}-m_{k-1}+p_{k-1}} \times \\ & \int_{\Theta_{\boldsymbol{x}_{k-2}}} \int_{\Theta_{\boldsymbol{y}_{k-2}}} \prod_{i=1}^{k-2} \boldsymbol{x}_i^{\boldsymbol{a}_i} \boldsymbol{y}_i^{\boldsymbol{b}_i} \left(\gamma \boldsymbol{x}_i + (1-\gamma)\, \boldsymbol{y}_i\right)^{\boldsymbol{c}_i} \times \\ & \boldsymbol{u}_{\boldsymbol{x}_{k-1}}^{\boldsymbol{a}_k+\boldsymbol{a}_{k-1}+\boldsymbol{c}_{k-1}-m_{k-1}+p_{k-1}-1} \boldsymbol{u}_{\boldsymbol{y}_{k-1}}^{\boldsymbol{b}_k+\boldsymbol{b}_{k-1}+m_{k-1}+n_{k-1}-1} \times \\ & \left(\gamma \boldsymbol{u}_{\boldsymbol{x}_{k-1}} + (1-\gamma)\, \boldsymbol{u}_{\boldsymbol{y}_{k-1}}\right)^{\boldsymbol{c}_k-n_{k-1}-p_{k-1}} d\Theta_{\boldsymbol{x}_{k-2}} d\Theta_{\boldsymbol{y}_{k-2}}. \end{aligned} \tag{B.193}$$

A double integral in $\boldsymbol{x}_{k-2}$ and $\boldsymbol{y}_{k-2}$ can be completed with the expansions in Equations B.188 and B.190. The general form of the result is similar to that of g_{II}, but with different values for a', b', c', d', e', and f', where the prime symbol indicates parameters in g_{II}. The next double integral has become more complex with the accumulation of terms from k:

$$a'_{k-2} = \boldsymbol{a}_k + \boldsymbol{a}_{k-1} + \boldsymbol{c}_{k-1} - m_{k-1} + p_{k-1}, \tag{B.194}$$

$$b'_{k-2} = \boldsymbol{b}_k + \boldsymbol{b}_{k-1} + m_{k-1} + n_{k-1}, \tag{B.195}$$

$$c'_{k-2} = \boldsymbol{c}_k - n_{k-1} - p_{k-1}, \tag{B.196}$$

$$d'_{k-2} = \boldsymbol{a}_{k-2}, \tag{B.197}$$

$$e'_{k-2} = \boldsymbol{b}_{k-2}, \tag{B.198}$$

$$f'_{k-2} = \boldsymbol{c}_{k-2}. \tag{B.199}$$

For later compactness of notation, it is convenient to define the following variables,

$$m_k = 0, \tag{B.200}$$

$$n_k = 0, \tag{B.201}$$

$$p_k = 0. \tag{B.202}$$

As the series of nested double integral progresses, the terms for each level of $g_{II_{k-l}}$ are, in general, for $1 \leq l \leq k-1$:

$$\begin{aligned}
a'_{k-l} &= \left(\sum_{q=k-l+1}^{k} \boldsymbol{a}_q\right) + \left(\sum_{q=k-l+1}^{k-1} \boldsymbol{c}_q - m_q + p_q\right), && \text{(B.203)}\\
b'_{k-l} &= \sum_{q=k-l+1}^{k} \boldsymbol{b}_q + m_q + n_q, && \text{(B.204)}\\
c'_{k-l} &= \boldsymbol{c}_k - \left(\sum_{q=k-l+1}^{k} n_q + p_q\right), && \text{(B.205)}\\
d'_{k-l} &= \boldsymbol{a}_{k-l}, && \text{(B.206)}\\
e'_{k-l} &= \boldsymbol{b}_{k-l}, && \text{(B.207)}\\
f'_{k-l} &= \boldsymbol{c}_{k-l}. && \text{(B.208)}
\end{aligned}$$

Another definition for the recursive components for the condition $2 \leq l \leq k-1$ is

$$\begin{aligned}
a'_{k-l} &= a'_{k-l+1} + d'_{k-l+1} + f'_{k-l+1} - m_{k-l+1} + p_{k-l+1}, && \text{(B.209)}\\
b'_{k-l} &= b'_{k-l+1} + e'_{k-l+1} + m_{k-l+1} + n_{k-l+1}, && \text{(B.210)}\\
c'_{k-l} &= c'_{k-l+1} - n_{k-l+1} - p_{k-l+1}. && \text{(B.211)}
\end{aligned}$$

These are used later to simplify the result of the integration.

When $l = k-1$, $\boldsymbol{u}_{\boldsymbol{x}_1} = 1$ and $\boldsymbol{u}_{\boldsymbol{y}_1} = 1$, the full integral can be cast into a series of nested triple sums containing the product of terms collected from the nested double integrals,

$$\begin{aligned}
G_{II} = &\sum_{m_{k-1}=0}^{f'_{k-1}} \sum_{n_{k-1}=0}^{c'_{k-1}} \sum_{p_{k-1}=0}^{c'_{k-1}-n_{k-1}} \cdots \sum_{m_1=0}^{f'_1} \sum_{n_1=0}^{c'_1} \sum_{p_1=0}^{c'_1-n_1} \times\\
&\prod_{r=1}^{k-1} \frac{\Gamma\left(f'_r+1\right)}{\Gamma\left(f'_r-m_r+1\right)} \times && \text{(B.212)}\\
&(-1)^{n_r} B\left(b'_r, e'_r + m_r + n_r\right)(1-\gamma)^{m_r+n_r} \times\\
&(-1)^{p_r} B\left(a'_r, d'_r + f'_r - m_r + p_r\right) \frac{\Gamma\left(c'_r+1\right)}{\Gamma\left(p_r+1\right)\Gamma\left(c'_r-n_r-p_r+1\right)} \gamma^{f'_r-m_r+p_r}.
\end{aligned}$$

Products of constants raised to powers involving r can be converted to constants raised to powers that are sums of terms involving r,

$$\begin{aligned} G_{II} = & \sum_{m_{k-1}=0}^{f'_{k-1}} \sum_{n_{k-1}=0}^{c'_{k-1}} \sum_{p_{k-1}=0}^{c'k-1-n_{k-1}} \cdots \sum_{m_1=0}^{f'_1} \sum_{n_1=0}^{c'_1} \sum_{p_1=0}^{c'_1-n_1} \times \\ & (-1)^{\sum_{r=1}^{k-1} n_r + p_r} \gamma^{\sum_{r=1}^{k-1} f'_r - m_r + p_r} (1-\gamma)^{\sum_{r=1}^{k-1} m_r + n_r} \times \\ & \prod_{r=1}^{k-1} \frac{\Gamma\left(f'_r+1\right)}{\Gamma\left(f'_r - m_r + 1\right)} B\left(b'_r, e'_r + m_r + n_r\right) \times \\ & B\left(a'_r, d'_r + f'_r - m_r + p_r\right) \frac{\Gamma\left(c'_r+1\right)}{\Gamma\left(p_r+1\right)\Gamma\left(c'_r - n_r - p_r + 1\right)}. \end{aligned} \tag{B.213}$$

There is an opportunity to cancel some of the Γ terms in the beta functions and in the Γ functions associated with $\boldsymbol{c}_r$ across different values of r. The product is then

$$\begin{aligned} \prod_{r=1}^{k-1} B\left(b'_r, e'_r + m_r + n_r\right) = & \frac{\Gamma\left(b'_1\right)\Gamma\left(e'_1+m_1+n_1\right)}{\Gamma\left(b'_1+e'_1+m_1+n_1\right)} \frac{\Gamma\left(b'_2\right)\Gamma\left(e'_2+m_2+n_2\right)}{\Gamma\left(b'_2+e'_2+m_2+n_2\right)} \cdots \\ & \frac{\Gamma\left(b'_{k-2}\right)\Gamma\left(e'_{k-2}+m_{k-2}+n_{k-2}\right)}{\Gamma\left(b'_{k-2}+e'_{k-2}+m_{k-2}+n_{k-2}\right)} \frac{\Gamma\left(b'_{k-1}\right)\Gamma\left(e'_{k-1}+m_{k-1}+n_{k-1}\right)}{\Gamma\left(b'_{k-1}+e'_{k-1}+m_{k-1}+n_{k-1}\right)}. \end{aligned} \tag{B.214}$$

Equation B.210 can be used to expand terms so that pairs of Γ functions will cancel down the chain of expanded terms,

$$\begin{aligned} \prod_{r=1}^{k-1} B\left(b'_r, e'_r + m_r + n_r\right) = & \frac{\Gamma\left(b'_2+e'_2+m_2+n_2\right)\Gamma\left(e'_1+m_1+n_1\right)}{\Gamma\left(b'_1+e'_1+m_1+n_1\right)} \\ & \frac{\Gamma\left(b'_3+e'_3+m_3+n_3\right)\Gamma\left(e'_2+m_2+n_2\right)}{\Gamma\left(b'_2+e'_2+m_2+n_2\right)} \cdots \\ & \frac{\Gamma\left(b'_{k-1}+e'_{k-1}+m_{k-1}+n_{k-1}\right)\Gamma\left(e'_{k-2}+m_{k-2}+n_{k-2}\right)}{\Gamma\left(b'_{k-2}+e'_{k-2}+m_{k-2}+n_{k-2}\right)} \\ & \frac{\Gamma\left(b'_{k-1}\right)\Gamma\left(e'_{k-1}+m_{k-1}+n_{k-1}\right)}{\Gamma\left(b'_{k-1}+e'_{k-1}+m_{k-1}+n_{k-1}\right)}. \end{aligned} \tag{B.215}$$

Cancel the common terms between numerators and denominators to get

$$\begin{aligned} \prod_{r=1}^{k-1} B\left(b'_r, e'_r + m_r + n_r\right) = & \frac{\Gamma\left(e'_1+m_1+n_1\right)}{\Gamma\left(b'_1+e'_1+m_1+n_1\right)} \\ & \Gamma\left(e'_2+m_2+n_2\right) \cdots \\ & \Gamma\left(e'_{k-2}+m_{k-2}+n_{k-2}\right) \\ & \Gamma\left(b'_{k-1}\right)\Gamma\left(e'_{k-1}+m_{k-1}+n_{k-1}\right). \end{aligned} \tag{B.216}$$

Use Equations B.204 and B.207 to get b_1' as a sum of the counts $\boldsymbol{b}_r$ and summation terms m_r and n_r to get

$$\prod_{r=1}^{k-1} B\left(b_r', e_r' + m_r + n_r\right) = \frac{\prod_{r=1}^{k} \Gamma\left(\boldsymbol{b}_r + m_r + n_r\right)}{\Gamma\left(\sum_{r=1}^{k} \boldsymbol{b}_r + m_r + n_r\right)}. \tag{B.217}$$

A similar result holds for the product of $B\left(a_r', d_r' + f_r' - m_r + p_r\right)$:

$$\begin{aligned}\prod_{r=1}^{k-1} B\left(a_r', d_r' + f_r' - m_r + p_r\right) = {} & \frac{\Gamma\left(a_1'\right)\Gamma\left(d_1' + f_1' - m_1 + p_1\right)}{\Gamma\left(a_1' + d_1' + f_1' - m_1 + p_1\right)} \frac{\Gamma\left(a_2'\right)\Gamma\left(d_2' + f_2' - m_2 + p_2\right)}{\Gamma\left(a_2' + d_2' + f_2' - m_2 + p_2\right)} \cdots \\ & \frac{\Gamma\left(a_{k-2}'\right)\Gamma\left(d_{k-2}' + f_{k-2}' - m_{k-2} + p_{k-2}\right)}{\Gamma\left(a_{k-2}' + d_{k-2}' + f_{k-2}' - m_{k-2} + p_{k-2}\right)} \\ & \frac{\Gamma\left(a_{k-1}'\right)\Gamma\left(d_{k-1}' + f_{k-1}' - m_{k-1} + p_{k-1}\right)}{\Gamma\left(a_{k-1}' + d_{k-1}' + f_{k-1}' - m_{k-1} + p_{k-1}\right)}.\end{aligned} \tag{B.218}$$

Equation B.209 is used to repeat a similar cancellation for the product of beta functions involving a', d', and f', which results in

$$\begin{aligned}\prod_{r=1}^{k-1} B\left(a_r', d_r' + f_r' - m_r + p_r\right) = {} & \frac{\Gamma\left(d_1' + f_1' - m_1 + p_1\right)}{\Gamma\left(a_1' + d_1' + f_1' - m_1 + p_1\right)} \Gamma\left(d_2' + f_2' - m_2 + p_2\right) \cdots \\ & \Gamma\left(d_{k-2}' + f_{k-2}' - m_{k-2} + p_{k-2}\right) \\ & \Gamma\left(a_{k-1}'\right)\Gamma\left(d_{k-1}' + f_{k-1}' - m_{k-1} + p_{k-1}\right).\end{aligned} \tag{B.219}$$

Replace the ellipsis-implied product with an explicit product representation to get

$$\prod_{r=1}^{k-1} B\left(a_r', d_r' + f_r' - m_r + p_r\right) = \frac{\Gamma\left(a_{k-1}'\right) \prod_{q=1}^{k-1} \Gamma\left(d_q' + f_q' - m_q + p_q\right)}{\Gamma\left(a_1' + d_1' + f_1' - m_1 + p_1\right)}. \tag{B.220}$$

Replace the prime variables with their unprimed equivalences to get

$$\prod_{r=1}^{k-1} B\left(a_r', d_r' + f_r' - m_r + p_r\right) = \frac{\Gamma\left(\boldsymbol{a}_k\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{a}_q + \boldsymbol{c}_q - m_q + p_q\right)}{\Gamma\left(\left(\sum_{q=1}^{k} \boldsymbol{a}_q\right) + \left(\sum_{q=1}^{k-1} \boldsymbol{c}_q - m_q + p_q\right)\right)}, \tag{B.221}$$

which is also

$$\prod_{r=1}^{k-1} B\left(a_r', d_r' + f_r' - m_r + p_r\right) = \frac{\Gamma\left(\boldsymbol{a}_k\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{a}_q + \boldsymbol{c}_q - m_q + p_q\right)}{\Gamma\left(\boldsymbol{a}_k + \left(\sum_{q=1}^{k-1} \boldsymbol{a}_q + \boldsymbol{c}_q - m_q + p_q\right)\right)}. \tag{B.222}$$

The ratio of Γ terms involving c' can be cancelled in a similar manner to get

$$\prod_{r=1}^{k-1} \frac{\Gamma\left(c'_r+1\right)}{\Gamma\left(p_r+1\right)\Gamma\left(c'_r-n_r-p_r+1\right)} = \frac{\Gamma\left(c'_1+1\right)}{\Gamma\left(p_1+1\right)\Gamma\left(c'_1-n_1-p_1+1\right)} \frac{\Gamma\left(c'_2+1\right)}{\Gamma\left(p_2+1\right)\Gamma\left(c'_2-n_2-p_2+1\right)} \cdots \frac{\Gamma\left(c'_{k-2}+1\right)}{\Gamma\left(p_{k-2}+1\right)\Gamma\left(c'_{k-2}-n_{k-2}-p_{k-2}+1\right)} \frac{\Gamma\left(c'_{k-1}+1\right)}{\Gamma\left(p_{k-1}+1\right)\Gamma\left(c'_{k-1}-n_{k-1}-p_{k-1}+1\right)}. \tag{B.223}$$

Equation B.211 is used to cancel terms to get

$$\prod_{r=1}^{k-1} \frac{\Gamma\left(c'_r+1\right)}{\Gamma\left(p_r+1\right)\Gamma\left(c'_r-n_r-p_r+1\right)} = \frac{1}{\Gamma\left(p_1+1\right)\Gamma\left(c'_1-n_1-p_1+1\right)} \frac{1}{\Gamma\left(p_2+1\right)} \cdots \frac{1}{\Gamma\left(p_{k-2}+1\right)} \frac{\Gamma\left(c'_{k-1}+1\right)}{\Gamma\left(p_{k-1}+1\right)}. \tag{B.224}$$

Replace the ellipsis-implied product with an explicit product symbol to get

$$\prod_{r=1}^{k-1} \frac{\Gamma\left(c'_r+1\right)}{\Gamma\left(p_r+1\right)\Gamma\left(c'_r-n_r-p_r+1\right)} = \frac{\Gamma\left(c'_{k-1}+1\right)}{\Gamma\left(c'_1-n_1-p_1+1\right)\prod_{r=1}^{k-1}\Gamma\left(p_r+1\right)}. \tag{B.225}$$

Replace the primed variables with their unprimed equivalences and use Equation B.205 to convert $\boldsymbol{c}_1$ and related terms to a summation to get

$$\prod_{r=1}^{k-1} \frac{\Gamma\left(c'_r+1\right)}{\Gamma\left(p_r+1\right)\Gamma\left(c'_r-n_r-p_r+1\right)} = \frac{\Gamma\left(\boldsymbol{c}_k+1\right)}{\Gamma\left(\boldsymbol{c}_k-\left(\sum_{q=1}^{k} n_q+p_q\right)+1\right)\prod_{r=1}^{k-1}\Gamma\left(p_r+1\right)}. \tag{B.226}$$

The remaining Γ function ratio does not have cancellations, but can be converted to

$$\prod_{r=1}^{k-1} \frac{\Gamma\left(f'_r+1\right)}{\Gamma\left(f'_r-m_r+1\right)} = \prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{c}_r+1\right)}{\Gamma\left(\boldsymbol{c}_r-m_r+1\right)}, \tag{B.227}$$

which has terms that match the $\boldsymbol{c}_k$ term in Equation B.226.

Inserting all the revised terms results in

$$\begin{aligned}
G_{II}\left(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\right) = & \sum_{m_{k-1}=0}^{\boldsymbol{c}_{k-1}} \sum_{n_{k-1}=0}^{\boldsymbol{c}_{k}} \sum_{p_{k-1}=0}^{\boldsymbol{c}_{k}-n_{k-1}} \Bigg(\sum_{m_{k-2}=0}^{\boldsymbol{c}_{k-2}} \sum_{n_{k-2}=0}^{\boldsymbol{c}_{k}-n_{k-1}-p_{k-1}} \sum_{p_{k-2}=0}^{\boldsymbol{c}_{k}-n_{k-1}-p_{k-1}-n_{k-2}} \Bigg(\times \\
& \cdots \Bigg(\sum_{m_{1}=0}^{\boldsymbol{c}_{1}} \sum_{n_{1}=0}^{\boldsymbol{c}_{k}-\left(\sum_{r=2}^{k-1} n_r+p_r\right)} \sum_{p_{1}=0}^{\boldsymbol{c}_{k}-\left(\sum_{r=2}^{k-1} n_r+p_r\right)-n_1} \times \\
& \Bigg(\left[(-1)^{\sum_{r=1}^{k-1} n_r+p_r}\right] \gamma^{\sum_{r=1}^{k-1} \boldsymbol{c}_r-m_r+p_r} \left(1-\gamma\right)^{\sum_{r=1}^{k-1} m_r+n_r} \times \\
& \frac{\Gamma\left(\boldsymbol{a}_k\right) \prod_{q=1}^{k-1} \Gamma\left(\boldsymbol{a}_q+\boldsymbol{c}_q-m_q+p_q\right)}{\Gamma\left(\boldsymbol{a}_k+\left(\sum_{q=1}^{k-1} \boldsymbol{a}_q+\boldsymbol{c}_q-m_q+p_q\right)\right)} \frac{\Gamma\left(\boldsymbol{b}_k\right) \prod_{r=1}^{k-1} \Gamma\left(\boldsymbol{b}_r+m_r+n_r\right)}{\Gamma\left(\boldsymbol{b}_k+\sum_{r=1}^{k-1} \boldsymbol{b}_r+m_r+n_r\right)} \times \\
& \left[\prod_{r=1}^{k-1} \frac{\Gamma\left(\boldsymbol{c}_r+1\right)}{\Gamma\left(\boldsymbol{c}_r-m_r+1\right)}\right] \frac{\Gamma\left(\boldsymbol{c}_k+1\right)}{\Gamma\left(\boldsymbol{c}_k-\left(\sum_{q=1}^{k-1} n_q+p_q\right)+1\right) \prod_{r=1}^{k-1} \Gamma\left(p_r+1\right)} \Bigg)\Bigg)\cdots\Bigg)\Bigg).
\end{aligned} \tag{B.228}$$

B.5.6 General Integral III

This general integral is associated with the integration of mixing parameters for the multinomial merger and splitting probability estimates. In most cases, the mixing parameter γ has to be treated as a nuisance parameter and eliminated by integration with a prior probability distribution function, for example,

$$g_{III} = \int_0^1 \gamma^a \left(1-\gamma\right)^b p\left(\gamma\right) d\gamma. \tag{B.229}$$

An option is to use a Dirichlet distribution function for the prior distribution of γ, with parameters, α_1 and α_2, which define the shape of the prior distribution. The function can be defined as

$$p_\gamma = \frac{\gamma^{\alpha_1-1}\left(1-\gamma\right)^{\alpha_2-1}}{B\left(\alpha_1,\alpha_2\right)}. \tag{B.230}$$

With the selection of a Dirichlet distribution function for the prior, $g_{III,D}$ can be written as

$$g_{III,D} = \int_0^1 \frac{\gamma^{a+\alpha_1-1}\left(1-\gamma\right)^{b+\alpha_2-1}}{B\left(\alpha_1,\alpha_2\right)} d\gamma. \tag{B.231}$$

This is the integral equation that represents the beta function (Gradshteyn and Ryzhik 3.191.3 [13]):

$$\int_0^1 \gamma^a \left(1-\gamma\right)^b d\gamma = B\left(a+1,b+1\right). \tag{B.232}$$

The result is then

$$g_{III,D}\left(a,b,\alpha_1,\alpha_2\right) = \frac{B\left(a+\alpha_1,b+\alpha_2\right)}{B\left(\alpha_1,\alpha_2\right)}. \tag{B.233}$$

if $\alpha_1 = 1$ and $\alpha_2 = 1$, then the Dirichlet distribution function is a uniform distribution, and

$$g_{III,U} = B\left(a+1, b+1\right), \tag{B.234}$$

because $B\left(1,1\right) = 1$.

This page intentionally left blank.

GLOSSARY

LL	Lincoln Laboratory
CFB	College Football
LI	fifty-one
MIT	Massachusetts Institute of Technology
NFL	National Football League
ROC	Receiver Operating Characteristics

This page intentionally left blank.

REFERENCES

[1] J. Przyborowski and H. Wilenski, "Homogeneity of results in testing samples from Poisson series: With an application to testing clover seed for dodder," *Biometrika* 31(3/4), 313–323 (1940).

[2] H. Jeffreys, *Theory of Probability*, Oxford University Press, 2nd ed. (1948).

[3] I.J. Good, "A Bayesian significance test for multinomial distributions," *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 399–431 (1967).

[4] M. Gönen, W.O. Johnson, Y. Lu, and P.H. Westfall, "The Bayesian two-sample t test," *The American Statistician* 59(3), 252–257 (2005).

[5] J.N. Rouder, P.L. Speckman, D. Sun, R.D. Morey, and G. Iverson, "Bayesian t tests for accepting and rejecting the null hypothesis," *Psychonomic Bulletin & Review* 16(2), 225–237 (2009).

[6] Q.F. Gronau, A. Ly, and E.J. Wagenmakers, "Informed Bayesian t-tests," *arXiv preprint arXiv:1704.02479* (2017).

[7] R. Sides, D. Kahle, and J. Stamey, "Bayesian sample size determination in two-sample Poisson models," *Biometrics & Biostatistics International Journal* 2(1), 00023 (2015).

[8] Z. Zhao, N. Tang, and Y. Li, "Sample-size determination for two independent binomial experiments," *Journal of Systems Science and Complexity* 24(5), 981 (2011).

[9] H. Jeffreys, *Theory of Probability*, Oxford University Press (1961).

[10] A. Etz, E.J. Wagenmakers, et al., "JBS Haldane's contribution to the Bayes factor hypothesis test," *Statistical Science* 32(2), 313–329 (2017).

[11] R.A. Sides, *Sample size determination for two sample binomial and Poisson data models based on Bayesian decision theory.*, Ph.D. thesis, Baylor University (2013).

[12] S.H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences* 1(4), 300–307 (2007).

[13] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products*, London: Academic Press (1980).

[14] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products, Translated from the Russian, Translation edited and with a preface by Daniel Zwillinger and Victor Moll, Revised from the seventh edition*, Elsevier/Academic Press, Amsterdam (2015).

[15] E.W. Weisstein, "Pochhammer symbol, from Mathworld–A Wolfram Web Resource." Wolfram.com (2016), URL http://mathworld.wolfram.com/PochhammerSymbol.html, [Online; accessed 14 December 2018].

This page intentionally left blank.

This page intentionally left blank.