On Skewed Multi-dimensional Distributions: the FusionRP Model, Algorithms, and Discoveries

Venkata Krishna Pillutla*†, Zhanpeng Fang*‡†, Pravallika Devineni§,
Danai Koutra¶, Christos Faloutsos*, and Jie Tang‡

Abstract

How do we model and find outliers in Twitter data? Given the number of retweets of each person on a social network, what is their expected number of comments? Real-life data are often very skewed, exhibiting power-law-like behavior. For such skewed multi-dimensional discrete data, the existing models are not general enough to capture various realistic scenarios, and often need to be discretized as they often model continuous quantities. We propose FusionRP, short for Fusion Restaurant Process, a simple and intuitive model for skewed multi-dimensional discrete distributions, such as number of retweets vs. comments in Twitter-like data. Our model is discrete by design, has provably asymptotic log-logistic sum of marginals , is general enough to capture varied relationships, and most importantly, and fits the real data very well. We give an effective and scalable maximum-likelihood based fitting approach that is linear in the number of unique observed values and the input dimension. We test FusionRP on a twitter-like social network with 2.2M users, a phone call network with 1.9M call records, game data with 45M users and *Facebook* data with 2.5M posts. Our results show that FusionRP significantly outperforms several alternative methods and can detect outliers, such as bot-like behaviors in the *Facebook* data.

1 Introduction

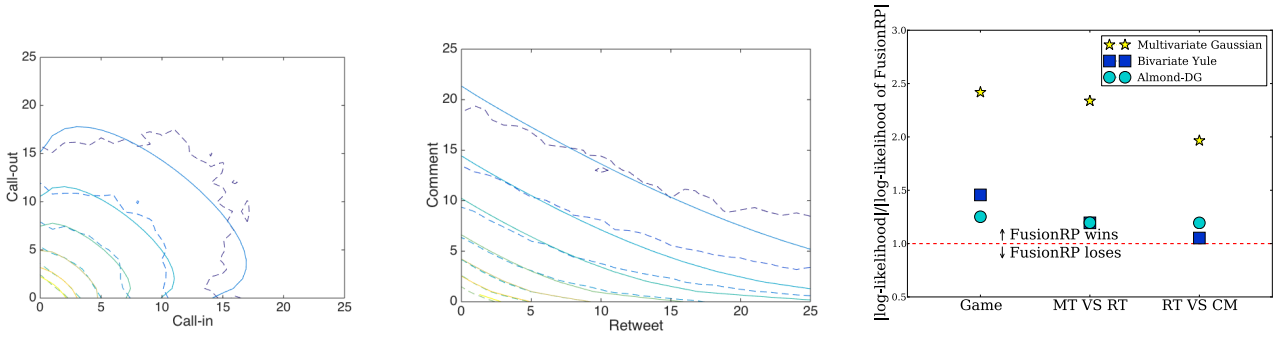
If "Alice" has 1000 wall-posts on her Facebook account, how many friends would you guess she has? If "Bob" has played our online game 200 times last month, totaling 500 minutes of play, and has spent $10 on digital items for the game, is he a human, or a bot?

The key to answering both questions is to characterize joint, multi-dimensional distributions (of friends-and-posts, for the case of Alice, of logins-minutes-dollars for the case of Bob) that are very skewed. Most real-life data are skewed. The traditional multivariate normal distribution fails miserably, even as it tries to model the marginals: none of them have negative values (that a Gaussian permits); all of them are skewed, with heavy tails, that a Gaussian fails to match. Skewed distributions, like the multivariate Pareto [17, 24, 14], and the very recent Almond-DG [10], also have shortcomings, as we discuss in Section 2.2. For example, some methods, such as the one in [25] fails to capture positive or negative correlations. An example of positive correlation would be between wall-posts and friends: the more friends one has on Facebook, the more posts he/she should see. For example, negative correlation could be between phone calls and texts: some people prefer talking over typing, and thus have many more phone-calls than texts, while others (typically, younger people) prefer the reverse - thus we may have a negative correlation.

Figure 1 illustrates the modeling power of our proposed FusionRP: Figure 1(a) illustrates the excellent fit of our FusionRP, to real phone call data: the dashed lines are iso-probability lines (contours) of the real data, while the solid lines are the iso-probability lines of our proposed FusionRP. Notice that the iso-probability lines are ellipsoid-looking, indicating "attraction" between the count of incoming, and outgoing phone calls (the more you call, the more you will be called). Figure 1(b) shows the fit for real, twitter-like data. Again, the solid lines are the iso-surfaces of real data (count of retweets vs. count of comments), and the dashed lines are the iso-surfaces of our fitted model FusionRP. Notice the "replulsion": most people will either have a lot of retweets, or a lot of comments, but not both. Further, as summarized in Table 1, most of these skewed multivariate distributions can be satisfactorily extended beyond 2-D to the question at hand, a problem we address with our FusionRP.

In summary, the contributions of this work are the

*Computer Science Department, Carnegie Mellon University, {pillutla, christos}@cs.cu.edu, zhanpenf@andrew.cmu.edu
†indicates equal contributions.
‡Department of Computer Science and Technology, Tsinghua University, jietang@tsinghua.edu.cn
§Department of Computer Science, University of California, Riverside, pdevi002@ucr.edu
¶Computer Science and Engineering, University of Michigan, Ann Arbor, dkoutra@umich.edu



(a) Call-in VS Call-out (b) Retweets VS Comments (c) Performance comparison

Figure 1: Goodness of fit of FusionRP: (a-b) Fitting the "call-in VS call-out" and the "retweets VS comments" datasets: real iso-surfaces are in dashed lines; the estimates of FusionRP are in solid lines - notice how close they are. (c) relative log-likelihood on unseen test data: all competitors lose to FusionRP- their ratio is 1.0 or higher.

following:

- **Model**: We present a simple, skewed, d-dimensional distribution. It fits real data well and is flexible enough to model a diverse range of scenarios. We test it successfully on 2 to 5-D data.
- **Analysis**: We analyze the theoretical properties of this model and present theoretically motivated and scalable fitting algorithm and outlier detection procedure.
- **Applicability**: As an immediate consequence of a good model and fitting procedure, we can flag points that do not fit the model as outliers. In our experiments, FusionRP finds **bot-like behavior** as outliers in *Facebook* wallpost data.
- **Reproducibility**: We have made our code open-source[1] for reproducibility.

The structure of the paper is typical. Next we give the survey, and then the proposed method and its analysis, algorithms for fitting and outlier detection, followed by experiments and conclusions.

2 Background and Related Work

A vast majority of work has been done on univariate skewed distributions, which are well understood both in theory and practice. Multivariate distributions, on the other hand, are well-studied in theory, but not so much in the context of skewed data [24].

2.1 1-D skewed distributions

2.1.1 The Yule distribution and Yule-Simon Process. The Yule distribution or the Yule-Simon distribution is a skewed distribution supported on the set of positive integers, and a parameter ρ has the mass function:

$$p_{\text{Yule}}(k;\rho) = \rho B(k, \rho+1) \tag{2.1}$$

where B is the Legendre Beta function. The Yule-Simon Process or preferential attachment [27, 21] is a discrete-time stochastic process. Following convention, the model is described in terms of a restaurant here.

Suppose there is a restaurant with an infinite number of tables, each with infinite capacity. N customers enter the restaurant one after another. The first customer occupies any empty table. In a scheme where "the rich get richer", customer number i, joins a table j with a probability proportional to the current size of the table, m_j or occupies a new table with probability s. If z_i is the table he/she joins, and we have k tables so far, we have,

$$P(z_i = j | z_{-i}, s) = \begin{cases} (1-s)\frac{m_j}{i-1} & \text{if } j \le k \\ s & \text{if } j = k+1 \end{cases} \tag{2.2}$$

Note that this is different from the Chinese Restaurant Process (CRP) [28] (see eq 2.3), introduced by Pitman, which is a Dirichlet Process. In the CRP, the probability of starting a new table reduces as we have more customers, but is a constant in the Yule-Simon process.

$$P(z_i = j | z_{-i}, \alpha) = \begin{cases} \frac{m_j}{i-1+\alpha} & \text{if } j \le k \\ \frac{\alpha}{i-1+\alpha} & \text{if } j = k+1 \end{cases} \tag{2.3}$$

Lemma 2.1. ([21]) *The limiting distribution of table sizes of Yule-Simon process as $N \to \infty$ is Yule($\frac{1}{1-s}$).*

The beta function has a power-law tail $B(a,b) \approx \Gamma(b)a^{-b}$ when a is large and b is fixed. Thus, the probability of observing a table of size k also has a power-law tail, where the exponent is: $\alpha = 1 + 1/1-s$. Hence, the Yule-Simon process is popular in modeling

[1] Code can be found at `https://github.com/krishnap25/FusionRP`

Data properties	Multivariate Gaussian	Multivariate Pareto	Bivariate Yule	Almond-DG	FusionRP
Multi-dimensional?	✓	✓		?	✓
Nonnegative?		✓	✓	✓	✓
Containing zero?	✓		✓	✓	✓
Discrete by design?			✓		✓
Skewed marginal?		✓	✓	✓	✓
Attraction?	✓			✓	✓
Repulsion?				✓	✓

Table 1: Superiority of FusionRP over existing approaches: It can model multi-dimensional, nonnegative, discrete data with skewed marginals and repulsion/attraction/indifference.

real-world power-law data such as word frequencies [21]. Note that vanilla CRP has exponential tails, and not power law tails.

2.1.2 Log-logistic distribution. The log-logistic distribution is a continuous distribution defined on $x \geq 0$, with CDF $F(x; \alpha, \beta) = (1 + (x/\alpha)^{-\beta})^{-1}$, $\alpha > 0, \beta > 0$. The odds-ratio of this distribution or its truncated version [23] follow a power law and hence, naturally finds application in modeling skewed data.For instance, Devineni et al [5] use PowerWall, a discrete log-logistic distribution to model the broadcasting behaviors on Facebook users using wallpost properties.

2.1.3 Pareto and Lognormal Distributions. The Pareto distribution was originally used to model allocation of wealth amongst individuals [4], which has since become famous as the 80-20 rule: that 20% of the people own 80% of the society's wealth. For any power law distribution, the CCDF, defined as $Pr(X \geq x)$, will be a straight line in log-log scale. The Pareto distribution satisfies this requirement by definition, but it has a necessarily positive minimum x_m, and this can be a drawback. A random variable is lognormal iff its logarithm is Gaussian. It is continuous and its CCDF is almost a straight line for a large part [16].

2.2 Multi-dimensional skewed distributions

2.2.1 Almond-DG. The Almond-DG [10] model combines log-logistic marginals with the concept of "copulas" in order to generate a bivariate distribution. Copulas have been successfully used to model the rainfall frequency as a joint distribution of rain characteristics (e.g., volume, duration), to capture the dependence between loss and the corresponding adjustment costs to calculate insurance premiums, and more. Although Almond-DG is proposed for modelling skewed data (e.g. counts), which are discrete by nature, it starts by modelling continuous data and continues with a discretization step. While this is not much of a problem for large numbers, it may alter the distribution of a vast majority of the observations that occur at the head (near zero). Moreover, while the original paper does not consider higher dimensions, the extension of Almond-DG to more dimensions is possible, but requires almost twice as many parameters as FusionRP.

2.2.2 Bivariate Yule, Pareto and Lognormal. Bivariate versions of log-normal and Pareto distributions have been proposed [26, 17, 24, 14] for applications such as drought and flood predictions, but these methods suffer the same drawbacks as their 1D variants. A bivariate Yule model was proposed in [25], but it cannot model attraction or repulsion and hence, cannot model real data very well.

Table 1 compares most of the existing skewed multi-dimensional distribution models and FusionRP.

2.3 Outlier Detection Outliers, as per Barnett and Lewis [2] are "observation(s) which appears to be inconsistent with the remainder of that set of data". Traditional techniques of outlier or anomaly detection are based on local density estimaties such as k-nearest neighbors. Examples include Local Outlier Factor (LOF) [3] and its numerous variants [22, 19, 20, 12].

These methods proceed in two phases: the first phase is to estimate the density in some form, and the second is to compare it with the density of its k nearest neighbors. Another approach uses angles instead [11].

But these methods are, as the names suggest, local and do not capture global patterns in data.

3 Proposed Method: FusionRP

In this section, we propose our method, Fusion Restaurant Process (FusionRP), to model skewed multi-

Symbol	Description
N	the total number of customers in the restaurant
n	the number of occupied tables
$m^{(j)}$	the number of customers on table j
$\mathbf{p}^{(j)}$	the parameter of the multinomial distribution of table j
$\boldsymbol{\alpha}$	the parameter of the Dirichlet distribution that generates $\mathbf{p}^{(j)}$ for each table j
s	the probability that a customer joins a new table, and $\rho = 1/(1-s)$
$\mathbf{x}^{(j)}$	distribution of customers on table j; more generally, the quantity we model
d	the dimensionality of $\mathbf{x}, \boldsymbol{\alpha}, \mathbf{p}$

Table 2: Notation used in the paper.

dimensional distributions such as #Retweets vs #Comments of Twitter users. We analyze the properties of FusionRP, and propose theoretically motivated and scalable algorithms for distribution fitting and anomaly detection. All proofs from this section have been provided in the appendix.

3.1 Intuition Following the restaurant metaphor, imagine a fusion restaurant that serves multiple cuisines, with an infinite number of tables each with infinite capacity, and there are two categories of dishes, e.g., Italian and Japanese. Suppose each customer has to choose a cuisine. N customers enter the restaurant one after another, and each chooses a table at random to sit down. The number of customers at a table is modeled by the Yule-Simon process. That is, the first customer comes in the restaurant and sits at the first table and the i^{th} customer, $i > 1$ sits at an occupied table, or at the next unoccupied table according to the distribution defined by Eq. 2.2. Note that the Dirichlet Process CRP of Eq. 2.3 is not used here because it has exponential tails. Yule-Simon process gives us power law tails that better agree with our data.

After all the customers have sat down, for each table j with $m^{(j)}$ customers, a biased coin c_j with probability p_j is sampled from a beta distribution i.e., $p_j \sim \text{Beta}(\alpha_1, \alpha_2)$. Each customer on table j tosses the biased coin c_j. She picks Italian dishes if the coin shows heads and Japanese dishes otherwise. The parameters α_1, α_2 determine the attraction, repulsion or indifference between customers choosing different types of dishes. Let $x_1^{(j)}$ and $x_2^{(j)}$ be the number of customers that take Italian dishes and Japanese dishes respectively on table j, where $x_1^{(j)} + x_2^{(j)} = m^{(j)}$. In keeping with the tradition of using restaurant metaphors, we name our model of the joint density of $(x_1^{(j)}, x_2^{(j)})$ as Fusion Restaurant Process or FusionRP in short. In the Twitter example, for instance, each table would represent a user, and each customer enjoying Italian (Japanese) cuisine would represent a retweet (comment).

3.2 Model Let $\mathbf{x} = (x_1, ..., x_d)^T \in {\mathbb{Z}_*}^d$ be the quantity we wish to model, where $\mathbb{Z}_*$ is the set of non-negative integers. We use parameters $s \in \mathbb{R}$ and $\boldsymbol{\alpha} \in \mathbb{R}^d$ for our model. Let $m = \mathbf{x}^T\mathbf{1} \geq 1$ be the total number of customers at a table, where $\mathbf{1} = [1, 1, ..1]^T \in \mathbb{R}^d$ is the vector of all ones. Also, let $\rho = (1-s)^{-1}$, for ease of notation.

The generative process of FusionRP is:

$$m|\rho \sim \text{Yule}(\rho) \tag{3.4a}$$

$$\mathbf{p}|\boldsymbol{\alpha} \sim \text{Dir}(\boldsymbol{\alpha}) \tag{3.4b}$$

$$\mathbf{x}|m, \mathbf{p} \sim \text{Mult}(m, \mathbf{p}) \tag{3.4c}$$

Here, 'Dir' and 'Mult' denote Dirichlet and Multinomial distributions respectively. Note that $\mathbf{x}|m, \boldsymbol{\alpha}$ is exactly the Dirichlet-Multinomial distribution [13], also known as the Polya distribution. It is commonly used in document classification to model distributions of word counts. The multinomial distribution is a natural choice to categorize customers into connoisseurs of a single cuisine (Italian or Japanese, in the above example). However, multinomial distribution with a fixed $\mathbf{p}$ cannot model all variations of real data. Hence, we draw $\mathbf{p}$ from its conjugate prior, a Dirichlet distribution. For the 2-D example above, the beta distribution is a special case of the Dirichlet.Table 2 summarizes notation used.

We can compute the probability mass function (pmf) of FusionRP in closed form.

Proposition 3.1. *The pmf of* FusionRP *is:*

$$p(\mathbf{x}|\rho, \boldsymbol{\alpha}) = \rho B(\mathbf{x}^T\mathbf{1}, 1+\rho)\frac{B(\mathbf{x}+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}\frac{\Gamma(\mathbf{x}^T\mathbf{1}+1)}{\prod_{i=1}^{d}\Gamma(x_i+1)}$$

Observation 3.1. FusionRP *can capture varied relationships in the data and this is determined by components of α relative to each other and to 1.*

For instance, for $d = 2$, $\alpha_1 = \alpha_2 = 1$ represents indifference, $\alpha_1 > 1, \alpha_2 > 1$ is attraction, $\alpha_1 <$

$1, \alpha_2 < 1$ is repulsion. Also, $\alpha_1 > 1, \alpha_2 < 1$ describes a scenario in our restaurant metaphor, where a customer ordering Italian dishes, corresponding to α_1 would be attracted to other customers, whereas a customer ordering Japanese dishes stays away from other customers. Refer to figure 2 for plots of iso-probability lines of these scenarios.

For all tables of a fixed size, attraction means that we are more likely to see customers of both cusines together on a table while repulsion means each table is more likely to be dominated by a single cuisine and indifference means that all combinations are equally likely. Note that these are not the same as statistical independence and positive/negative correlation as two independent Gaussians exhibit attraction where as independent log-logistic random variables repel each other.

3.3 Union Marginals To look at marginals, we define the notion of the union-marginals. Intuitively, in our fusion restaurant with d cuisines, if we group some cuisines into categories, the distribution of customers on these categories is a union-marginal. The univariate union-marginal gives the number of customers at a table since all cuisines are grouped into one category.

DEFINITION 3.1. *A Union-marginal of a distribution of* $\mathbf{x} = (x_1, ..., x_d)^T$ *corresponding to a partition* $I = \{I_1, ..., I_k\}$ *of dimensions is the distribution of* $\mathbf{x}_I = (\sum_{i \in I_1} x_i, ..., \sum_{i \in I_k} x_i)^T$. *In particular, the univariate union-marginal of* $\mathbf{x}$ *is the distribution of* $\mathbf{x}^T \mathbf{1}$.

The following theorem states that FUSIONRP is consistent in the sense that each union-marginal is another FUSIONRP with different parameters.

THEOREM 3.1. *Union-marginals of* FUSIONRP$(s, \boldsymbol{\alpha})$ *corresponding to the partition* $I = \{I_1, ..., I_k\}$ *are distributed according to* FUSIONRP$(s, \boldsymbol{\alpha}_I)$ *where* $\boldsymbol{\alpha}_I = (\sum_{i \in I_1} \alpha_i, ..., \sum_{i \in I_k} \alpha_i)$. *In particular, univariate union-marginal of* FUSIONRP *follows a Yule distribution and has an asymptotically log-logistic tail.*

3.4 Parameter Estimation Given a dataset $X = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$, we wish to estimate parameters $\boldsymbol{\theta} = \{s, \boldsymbol{\alpha}\}$. By observing the generative process, it can be seen that estimations of s and $\boldsymbol{\alpha}$ are decoupled.

The estimation of s of the Yule-Simon process admits a very simple solution: the moment matching estimate. Let $N = \sum_{j=1}^{n} m^{(j)}$ be the total number of customers in the restaurant. This estimate is: $\hat{s} = n/N$.

For $\boldsymbol{\alpha}$, we use Newton's method to find the Maximum Likelihood Estimator (MLE), $\hat{\boldsymbol{\alpha}}$ and this can be done efficiently [15]. We use the moment matching estimate of $\boldsymbol{\alpha}_0$ from $\mathbf{x}^1/m^1, \mathbf{x}^2/m^2, ..., \mathbf{x}^n/m^n \sim \text{Dir}(\boldsymbol{\alpha}_0)$ as a warm start. Further details of the algorithm may be found in the appendix.

PROPOSITION 3.2. *Parameter estimation of* FUSIONRP *is linear on the input size, that is, it runs in time* $\mathcal{O}(n_0 d)$, *where* n_0 *is the number of unique observations and* d *is the dimensionality.*

3.5 Outlier Detection We use the parametric form of our model to derive an outlier detection procedure. Unlike traditional outlier detection procedures such as LOF [3], we do *not* have any user-defined parameters, which affect the performance.

For a point $\mathbf{x}^i$ in $X = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n\}$, define $\lambda_i = np(\mathbf{x}^i)$ to be the expected number of times $\mathbf{x}$ appears, where $p(.)$ is the pmf of FUSIONRP.We can define a confidence interval $C_i(\delta)$ for λ_i based on the the multiplicity n_i of $\mathbf{x}^i$ i.e. the number of times each unique $\mathbf{x}^i$ is seen, as follows:

$$C_i(\delta) = \Big[\frac{1}{2}\chi^2\big(\frac{\delta}{2}; 2n_i\big), \frac{1}{2}\chi^2\big(1 - \frac{\delta}{2}; 2n_i + 2\big)\Big] \tag{3.5}$$

Here, $\chi^2(q; k)$ is the q^{th} quantile of the χ^2 distribution with k degrees of freedom.

THEOREM 3.2. *The confidence interval* $C_i(\delta)$ *of Eq. 3.5 is an approximate* $(1 - \delta)$ *confidence interval for* λ_i, *the expected number of observations at* $\mathbf{x}^i$.

We also give an *anomaly score* as the number of standard deviations of the expected count λ_i from the confidence interval $C_i(\delta) = [C_l, C_u]$: $(\lambda_i - C_u)/\sqrt{\lambda_i}$ if $\lambda_i > C_u$, $(C_l - \lambda_i)/\sqrt{\lambda_i}$ if $\lambda_i < C_l$ and zero otherwise. Points with high anomaly score lie far from the confidence interval, and may be investigated further.

Speedwise, this method is linear in the number of unique observations, n_0 and the dimensionality d. Local methods such as LOF are close to quadratic in n and may be infeasible for large applications. In practice, our algorithm is robust to the choice of δ. We demonstrate an application of this method in action in Section 4.

3.6 The model in hindsight Overall, the three-stage model and the choice of stages (Yule and Dirichlet-Multinomial) gives us flexibility to model various real, long-tailed distributions while still being able to efficiently and acceptably estimate parameters. In particular, we use the Yule distribution because it generates long, power-law tails that closely match empirical observations.

4 Experiments

Here we report experiments to answer the following questions:

Q1. **Flexibility and Generality**: What can FusionRP model?
Q2. **Applicability**: How well does FusionRP work on real data?
Q3. **Scalability**: How fast is FusionRP?
Q4. **Practicality**: What is the practical use of FusionRP?

We evaluate the proposed method on four datasets:

1. *Phone call* [6]. The dataset is made of a large collection of call records provided by a mobile communication company. It contains 1.9 million call records of 202,897 users during one month. We extract and use the number of incoming and outgoing calls for each user.
2. *Game* [7]. The game dataset comprises user activities for 45,429,334 users in a large online game during July and August 2014. For each user, we extract the number of days logged on to the game, and the number of items purchased in the game.
3. *Tencent Weibo* [18]. Tencent Weibo is a popular Twitter-like microblogging service in China. The dataset contains records for 2.1 million users, for each one of whom we extract three quantities: the number of retweets, comments, and mentions.
4. *Facebook* [5]. The dataset contains $\sim$2.5 million Facebook wallposts from over $7,000$ users during four months in $2011 - 2013$. For each user, we extract the number of links, statuses, photos, videos or other posts made, posting time and the application name, if the post was generated *via* an app.

For each dataset, we construct a two, three, four or five dimensional distribution over the users, where each user is viewed as a point in the space. To connect to our restaurant metaphor, each user here is a table, and each attribute corresponds to a cuisine.

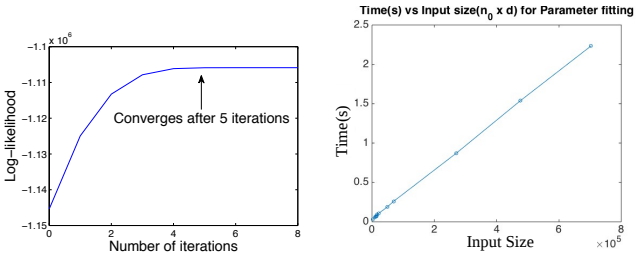
4.1 Q1 - Flexibility and Generality FusionRP is quite flexible and can fit different shapes of distributions. For instance, Fig. 2 illustrates three contour plots of 2-D distributions that were generated by FusionRP. We can see that with different configurations of model parameters, the proposed method can model distributions showing the properties of indifference (Fig. 2(a)), attraction (Fig. 2(b)), or repulsion (Fig. 2(c)).

FusionRP is general, and it can model d-dimensional distributions for an arbitrary value d. Figure 3 shows the fitting results of FusionRP to the "retweets VS comments VS mentions" distribution of the *Tencent Weibo* dataset. Figures 3(a) and (c) show the distributions of the real data, while Figs. 3(b) and (d) depict the 2-D marginal distributions generated by the fitted 3-D FusionRP. FusionRP also fits 4-D and 5-D *Facebook* data, but we do not present them due to difficulty in visualization.

Evaluating goodness of fit for skewed data is challenging, even in the univariate case [9] and only gets worse with more dimensions. We adopt the approach of [10] and qualitatively measure the goodness of fit.

4.2 Q2 - Applicability Figures 1(a), (b) and 3 show qualitatively that FusionRP provides a good fit to real world datasets. Moreover, FusionRP has been tested with up to 5-D real world data, and can work for d-dimensional data for any d, in principle. We are restricted by the unavailability of higher dimensional skewed data.

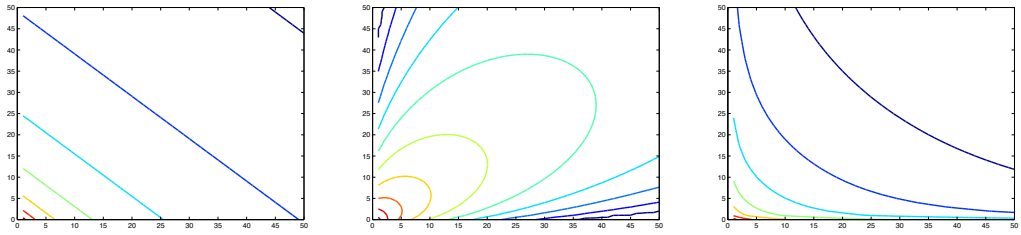
To further establish applicability, we compare our model with three other parametric distributions for multi-dimensional data: the multivariate Gaussian, the bivariate Yule [25], and the Almond-DG [10]. We do not compare with the multivariate Pareto [1] here because it cannot handle zero values, and this case exists in all three of our datasets. For multivariate Gaussian and bivariate Yule, we use maximum likelihood estimation to fit the parameters. For Almond-DG, we use the code provided by the authors of the original paper to estimate the parameters. In Figure 1(c), we report the relative ratio of the log-likelihood scores computed by five-fold cross validation. We can see that in all three datasets, the ratios of all the competing methods are greater than 1.0, which suggests that FusionRP achieves the best performance in each of the datasets, despite having fewer parameters than the Gaussian and Almond-DG models.



(a) Convergence analysis (b) Running time analysis

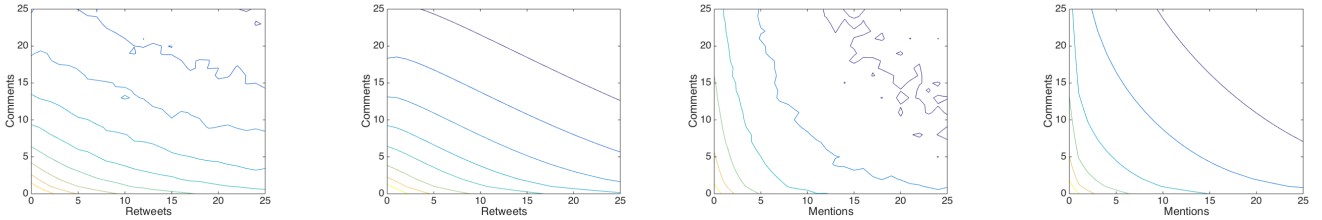
Figure 4: Scalability of FusionRP: (a) Convergence analysis of the proposed method on the *Phone call* dataset. (b) Running time analysis of the proposed method on synthetic datasets.

4.3 Q3 - Scalability We now evaluate the scalability performance of FusionRP. Figure 4(a) shows the convergence analysis results of our method on the *Phone call* dataset. We can see that our fitting procedure converges after 5 iterations. For the other datasets, the algorithm has a similar convergence rate, and converges



(a) Indifference (b) Attraction (c) Repulsion

Figure 2: Flexibility of FusionRP: Contour plots of FusionRP with parameters $s = 0.1$ and (a) $\alpha_1 = \alpha_2 = 1$, (b) $\alpha_1 = \alpha_2 = 10$, (c) $\alpha_1 = \alpha_2 = 0.1$. For a fixed table size, notice that attraction means mixed tables are more likely while repulsion means homogenous tables are more likely.



(a) Real: RT VS CM (b) FusionRP: RT VS CM (c) Real: MT VS CM (d) FusionRP: MT VS CM

Figure 3: Generality of FusionRP: Contour plots of FusionRP to the 3-D "retweets VS comments VS mentions" distribution of the *Tencent Weibo* dataset. (a) and (c) are the contour plots of the real distributions; (b) and (d) are the contour plots of the marginal distributions generated by the fitted 3-D FusionRP.

within 8 iterations. This fast convergence property with linear time complexity of FusionRP with the input size makes it efficient for large datasets. Figure 4(b) shows the runtime analysis results of the proposed method on synthetic datasets. We observe that the running time of FusionRP is linear in the input size, $n_0 \times d$, and it can process multi-dimensional datasets with hundreds of thousands of data points within seconds on a commodity laptop with 4GB memory and 2.5GHz processor.

4.4 Q4 - Practicality In this section, we show how we can leverage our proposed model to perform outlier detection, and empirically evaluate our approach on synthetic and real data. We first estimate the parameters of FusionRP on the given dataset, and use Eq. 3.5 to flag anomalies. The detection of the top outliers is robust to the value of δ which we set to 0.01 in our experiments. Finally, we sort on the anomaly score defined previously and examine further the top outliers.

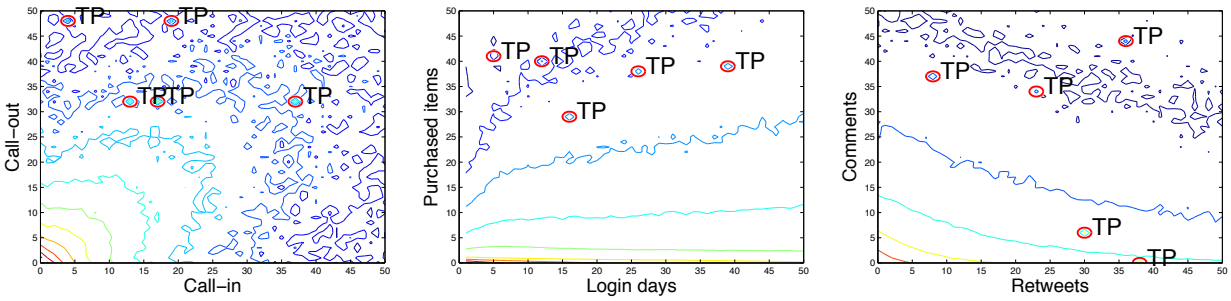
Sanity Check: Injected Outliers. First, we check whether the detected method can find artificially injected anomalies. For each dataset, we inject anomalies by randomly selecting five points in the dataset, and amplifying the count of the selected points to 20 times of their original values. Figure 5 shows the top five anomalous points detected by FusionRP on the three datasets. We can see that FusionRP perfectly detects all the injected anomalies in each of the datasets.

Real Ouliers in *Facebook* Data. We now test the method on 5-D *Facebook* wallpost data. The input is the count of link, status, photo, video and other posts made by a *Facebook* user. Further investigation of the top 40 anomalies of our algorithm revealed that 90% of these outliers share a common theme: most of the posts were links shared *via* applications that post on behalf of the users on their *Facebook* wall without active participation of the user. Some apps we found were gaming apps like Farmville, Cityville and social platforms like Tumblr and Twitter. Additionally, some of these users have a small number of posts that were not app generated. It is remarkable indeed that FusionRP is able to detect bot-like behavior amongst users using only the total number of posts made over a four month period.

To investigate ouliers flagged by our algorithm, we generated a set of metrics and associated plots and discover some interesting behavior. These plots are in Figure 6.

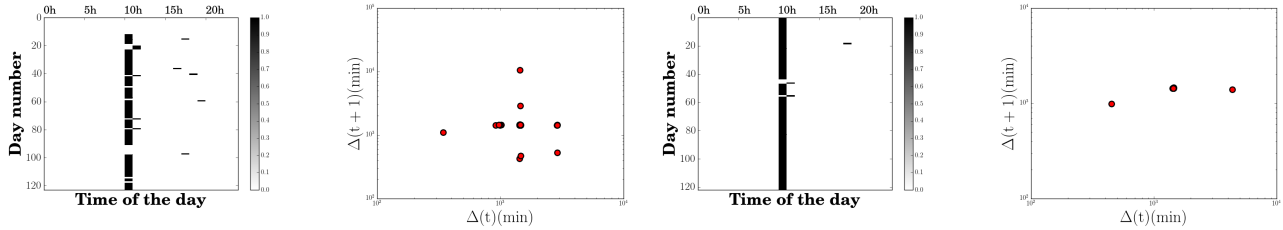
- **Time-day plot:** Figures 6(a) and (c) provide the heatmap of the number of user posts per hour (x-axis) and per day (y-axis).
- **Lag-correlation plot:** Figures 6(b) and (d) provide a check of whether a time series is random or not. Here, we plot the time difference between two consecutive pairs of posts. Since bots usually post at regular intervals, non-random behavior can be easily spotted.

We present as examples two of the outliers, ranked 5 and



(a) *Phone call* (b) *Game* (c) *Tencent Weibo*

Figure 5: Detecting artificial outliers using FusionRP: Automatic outlier detection by FusionRP on the three datasets: (a) *Phone call*, (b) *Game*, and (c) *Tencent Weibo*. "TP" stands for an injection detected as a true positive. FusionRP detects all the injected anomalies perfectly in all three datasets.



(a) Outlier 1, Hour of day (b) Outlier 1, Lag-correlation (c) Outlier 2, Hour of day (d) Outlier 2, Lag-correlation

Figure 6: FusionRP can detect meaningful outliers: Two representative outliers found by FusionRP in *Facebook* data whose posts were made almost entirely by apps such as "Giveaway of the Day" and "Astrology". Most posts were made at one particular hour of the day everyday, and the lag between successive posts is non-random, indicating bot-like behavior.

32 which stood out as obvious anomalies to a human evaluator. The first outlier has 104 posts which contain links to other webpages. Figures 6(a),(b) show that most of these posts are made during the same hour of the day throughout the four months, exactly 24 hours apart from one another. Our data indicates that these posts were generated *via* an app called "Giveaway of the day" that promises the user with daily discounts on software applications, thereby verifying our claim.

The second outlier exhibits similar behavior, which can be seen in Fig. 6(c). Outlier 2 has 120 link posts and 1 photo post in the period of four months. These link posts were generated *via* an app, "Astrology", without the active participation of the user. The lag-correlation plot (Fig. 6(d)) clearly indicates the non-random bot-like behavior.

These examples illustrate FusionRP's potential use as a crude, fast way to detect outliers.

5 Conclusions

Modeling skewed multivariate distributions is an important problem in the field of data mining. In this paper, we propose a stochastic process, FusionRP to this end. Specifically, our contributions are:

- **Model**: We propose a new model FusionRP for skewed, d-dimensional distributions for $d \geq 2$. The method is flexible enough to fit diverse real world data with various types of interactions well, and the parameters can still be efficiently estimated.
- **Analysis**: We study theoretical properties of FusionRP and describe theoretically supported and scalable methods for parameter fitting and anomaly detection.
- **Applicability**: We show that FusionRP is practical for real-world applications. It can be used not only for modelling and understanding various user-related activities, but also for quick and dirty outlier detection. In our experiments, FusionRP finds **bot-like behavior** as outliers in *Facebook* wallpost data.
- **Reproducibility**: Our code has been made available online[2], for enthusiasts to play with, or extend.

References

[1] N. Balakrishnan. *Continuous multivariate distributions*. Wiley Online Library, 2001.

[2] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley Series in Probability & Statistics. Wiley, 1994.

[2] https://github.com/krishnap25/FusionRP

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, *SIGMOD Conference*, pages 93–104. ACM, 2000. SIGMOD Record 29(2), June 2000.

[4] D. Chotikapanich. *Modeling Income Distributions and Lorenz Curves*. Economic Studies in Inequality, Social Exclusion and Well-Being. Springer New York, 2008.

[5] P. Devineni, D. Koutra, M. Faloutsos, and C. Faloutsos. If walls could talk: Patterns and Anomalies in Facebook wallposts. In *ASONAM*. IEEE/ACM, August 2015.

[6] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla. How long will she call me? distribution, social theory and duration prediction. In *Machine Learning and Knowledge Discovery in Databases*, pages 16–31. Springer, 2013.

[7] Z. Fang, X. Zhou, J. Tang, W. Shao, A. Fong, L. Sun, Y. Ding, L. Zhou, and J. Luo. Modeling paying behavior in game social networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 411–420. ACM, 2014.

[8] F. GARWOOD. (i) fiducial limits for the poisson distribution. *Biometrika*, 28(3-4):437–442, 1936.

[9] N. Johnson, S. Kotz, and N. Balakrishnan. Continuous univariate distribution, 2-nd edition, vol. 1, 1995.

[10] D. Koutra, V. Koutras, B. A. Prakash, and C. Faloutsos. Patterns amongst competing task frequencies: Super-linearities, and the almond-dg model. In *Advances in Knowledge Discovery and Data Mining*, pages 201–212. Springer, 2013.

[11] H.-P. Kriegel, A. Zimek, et al. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452. ACM, 2008.

[12] J. Y. Lee, U. Kang, D. Koutra, and C. Faloutsos. Fast anomaly detection despite the duplicates. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 195–196. International World Wide Web Conferences Steering Committee, 2013.

[13] R. E. Madsen, D. Kauchak, and C. Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 545–552, New York, NY, USA, 2005. ACM.

[14] K. V. Mardia. Multivariate pareto distributions. *The Annals of Mathematical Statistics*, pages 1008–1015, 1962.

[15] T. Minka. Estimating a Dirichlet distribution. Web, 2000.

[16] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1(2):226–251, 2003.

[17] S. Nadarajah. A bivariate pareto model for drought. *Stochastic Environmental Research and Risk Assessment*, 23(6):811–822, 2009.

[18] Y. Niu, Y. Wang, G. Sun, A. Yue, B. Dalessandro, C. Perlich, and B. Hamner. The tencent dataset and kdd-cupą́f12. In *KDD-Cup Workshop*, volume 170, 2012.

[19] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. pages 315–326, 2003.

[20] E. Schubert, A. Zimek, and H. Kriegel. Generalized outlier detection with flexible kernel density estimates. In *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pages 542–550, 2014.

[21] H. A. SIMON. On a class of skew distribution functions. *Biometrika*, 42(3-4):425–440, 1955.

[22] J. Tang, Z. Chen, A. W.-C. Fu, and D. W.-L. Cheung. Enhancing effectiveness of outlier detections for low density patterns. In *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '02, pages 535–548, London, UK, UK, 2002. Springer-Verlag.

[23] P. Vaz de Melo, L. Akoglu, C. Faloutsos, and A. Loureiro. Surprising patterns for the call duration distribution of mobile phone users. In J. BalcÃązar, F. Bonchi, A. Gionis, and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 354–369. Springer Berlin Heidelberg, 2010.

[24] J. Wesolowski. Bivariate distributions via a pareto conditional distribution and a regression function. *Annals of the Institute of Statistical Mathematics*, 47(1):177–183, 1995.

[25] E. Xekalaki. The bivariate yule distribution and some of its properties. *Statistics: A Journal of Theoretical and Applied Statistics*, 17(2):311–317, 1986.

[26] S. Yue. The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, 14(14):2575–2588, 2000.

[27] U. G. Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character:*, 213:21–87, 1925.

[28] X. Zhang. A Very Gentle Note on the Construction of Dirichlet Process. Technical report, The Australian National University, Canberra,Australia, 2008.