The "DGX" Distribution for Mining Massive, Skewed Data

Zhiqiang Bi

May, 2002

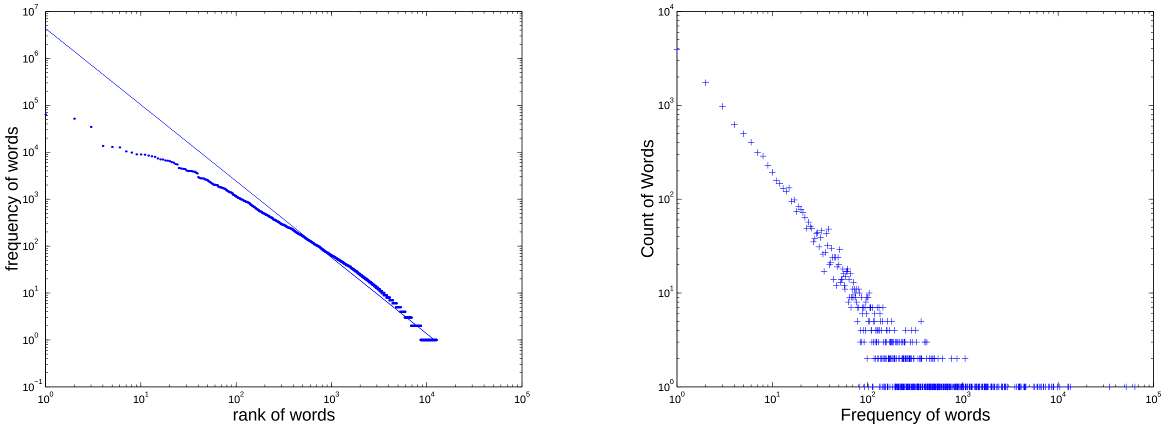
Abstract

Skewed distributions appear very often in practice. Unfortunately, the traditional Zipf distribution often fails to model them well. In this paper, we propose a new probability distribution, the Discrete Gaussian Exponential (DGX), to achieve excellent fits in a wide variety of settings; our new distribution includes the Zipf distribution as a special case. We present a fast and statistically sound method for estimating the DGX parameters based on maximum likelihood estimation (MLE). We applied DGX to a wide variety of real world data sets, such as sales data from a large retailer chain, service usage data from a telecommunication company, and Internet clickstream data; in all cases, DGX fits these distributions very well, with almost a 99% correlation coefficient in quantile-quantile plots. Our algorithm also scales very well because it requires only a single pass over the data. We also show how to generalize DGX to 2 dimensions, obtaining a good fit on real, two dimensional distributions.

Contents

1 Introduction

In countless cases we encounter skewed distributions, where a few products (or vocabulary words, or customers) are responsible for most of the revenue (or occurrences, or sales), while the rest have very little individual contributions. Zipf, in his milestone book [Zip49], proposed the distribution $frequency \propto 1/rank$ for vocabulary words (and city populations, length of articles, income distributions and so on). Although a significant step to the correct direction, the Zipf distribution often fails to model real datasets well. For example, in Figure (1), we make the "frequency-rank plot" and the "count-frequency plot" of words in the Bible. As explained in the survey section, the Zipf (or generalized-Zipf) distribution would expect the plots to be straight lines in logarithmic-logarithmic scales. However, we observe a clear tilting in Figure (1). Zipf himself had observed this deviation, he even had a name for it ("top concavity"), and he devoted several paragraphs in his book to justify it, whenever it appeared in a dataset. Similar deviations are observed in many other cases, as we will see in the experiment section.



(a) Rank-frequency plot of words in the Bible. We fit the straight line using least square method.

(b) Count-frequency plot of words in the Bible.

Figure 1: Rank-frequency plot and count-frequency plot of words in English Bible. Although they both show Zipf-like skewed behavior, they clearly do not follow Zipf's law exactly.

Our goal in this work is to do away with this recurring nuisance: We want a distribution that would have the following attractive properties:

1. it should include the "Zipf" and "generalized Zipf" as special cases

2. it should fit well all the datasets that Zipf fits, and many many more.

3. it should be parsimonious (i.e., few parameters), and
4. it should be fast to compute its parameters, even if the given datasets are huge.

Moreover, we want a two dimensional distribution that models well skewed two dimensional data. The rest of the paper is organized as follows: Section 2 describes the Zipf distribution and gives the literature survey. Section 3 presents our proposed method, the one dimensional DGX distribution, along with the proofs, the algorithms and results. In Section 5 we discuss the skewed distribution in two dimensions and Section 6 lists the conclusions and future research directions.

2 Background - Survey

First, we start with the description of the Zipf distribution, and then we describe related work.

2.1 Background: Zipf and generalized Zipf distributions

We describe the Zipf distribution and the two Zipf "laws": the rank-frequency one and the frequency-count one. The laws are best described with an example, such as words in a book (or the Bible, as we show in Figure (1)) Let V be the vocabulary size, and let f_1 be the occurrence frequency of the most frequent vocabulary word, f_2 for the second most frequent, and so on.

Definition 1 *The* rank-frequency *plot is the plot of the occurrence frequency f_r versus the rank r, in logarithmic-logarithmic scales*

The rank-frequency version of Zipf's law states that

$$f_r \propto 1/r \tag{1}$$

This is typically referred to as the *Zipf's law* or the *Zipf distribution.* In log-log scales, the Zipf distribution gives a straight line with slope -1.

The *generalized* Zipf distribution (or "Zipf-like" distribution) is defined as

$$f_r \propto 1/r^{\theta} \tag{2}$$

where the slope in log-log scales can be different than -1.

The second 'law', also known as the discrete Pareto distribution [Par97], involves the "count-frequency" plot: let c_f be the count of vocabulary words that appear f times in the document. The second Zipf's law states that

$$c_f \propto 1/f^{\phi} \tag{3}$$

There are three observations:

- The count-frequency plot actually corresponds to the PDF (probability density function) of the occurrence frequency of a word in a document
- It is a mathematical consequence of the first law. It can be shown, for example in [Hil74] or [Ada], that $\phi = 1 + 1/\theta$
- In log-log scales, the count-frequency plot of a Zipf distribution will be a straight line, with slope ϕ

Despite the success and fame of the Zipf distribution, we note that, eg. in Figure (1), the words in the Bible do not follow the Zipf distribution exactly, but instead they have the "top concavity".

For the rest of this work, we only report the 'count-frequency' (= PDF) plots for all the upcoming datasets, since the PDF is a more familiar concept than the "rank-frequency" plot, and since the two "laws" are in fact sides of the same coin.

2.2 Survey

There are significant past attempts to model skewed distributions. They form two classes: Discrete, and continuous distributions.

Discrete distributions : This is the class that we are most interested in, since most of the data of interest are either inherently integer-valued, or rounded-off to integers: salaries and dollar amounts are down to pennies, products sell integer counts ("1 loaf of bread"), and so on. In this front, there is of course the Zipf distribution with its variations ("generalized Zipf"), the Yule distribution [Yul23] , and the Pareto distribution [Par97]. . Among these distributions, Zipf's law is most widely used because of its simple form. Zipf's law has been observed in many fields. For example, the population of cities and the rank of the population [BG58], the number of articles in rth largest journal versus the rank of the journal [Sim55], the surnames of 4794 people in an area in England [FL83] all follow Zipf's law. Recently, Zipf's law has been applied to research on web caching. Studies [Gla94, CCC95, VAdO96, BCF$^+$99] show that the number of requests the server of rank r receives versus r also has the Zipf-like behavior.

Continuous distributions : Although not directly applicable, we mention them, mainly because of the "lognormal" distribution, which is extremely successful in modeling continuous datasets. The lognormal distribution [Gal79] takes positive values, and can be generated as e^X where X is a Gaussian variable. It has been used to model particle sizes in natural aggregates, dust concentration in industrial atmospheres, in geological applications, concentration of minerals in deposits, flood flows, weights of children, automobile insurance claims, the weight distribution of U.S. adult males and females (Page 238-239, [NJB94]). Gibrat found the distribution useful to represent the distribution of size for varied kinds of "natural" economic units.(Page 238-239, [NJB94])

In several cases, there are even theoretical arguments supporting the lognormal distribution [Hal44,Her60,NJB94] : For example, if we break a stick in two at a random point, and continue recursively, the length of the resulting pieces will follow a lognormal distribution. It is also considered a serious competitor to the Weibul distribution for lifetime distributions of manufactured products. In fact, it can also approximate the Gaussian distribution. (Page 238-239, [NJB94])

There have been some attempts to fit this kind of skewed data with other probability distributions, such as parabolic fractal [Lah] or stretched exponentials [LS98]. These works, however, are based on continuous probability distribution function, which is not appropriate for a lot of real world data which can only take discrete values, such as the visits to web sites, number of certain products sold in a supermarket, etc. Secondly, they estimated the parameters by fitting a curve on the rank-count plot in log-log scale, which we believe is statistically *ad hoc.*

3 DGX in One Dimension

Our goal is to find a discrete distribution, that will fit the PDF (a.k.a. frequency-count plot) of many, real datasets.

However, it is unclear where we should start from: Should we try to fit a parabola in the rank-frequency plot? Or, maybe, a third degree polynomial? or a Gaussian, a sinusoid, a spline? or something else? Or should we try all these functions on the frequency-count plot?

A deeper question is: even if one of these functions fit in a few cases, do we have "a-priori" reasons to believe that it will fit well, in multiple settings?

The answer to all this questions is our proposed DGX distribution. Judging from the success of the lognormal (also referred to as "anti-lognormal") distribution for continuous data, we propose the following thought experiment: Consider a random variable, say, the duration of a web-surfing session. This is a continuous variable, and, most likely, might follow a lognormal distribution. However, we need to store it with finite accuracy, and thus turn it into an integer (number of minutes, or seconds, or hours). This is exactly the motivation behind DGX: Consider a lognormal random variable (by creating a Gaussian variable, and exponentiating it); then, digitize it to the nearest integer. The same is true for everything else: salaries (digitized to penny accuracy), duration of hospital stays (rounded to days), body height (inches), body weight (pounds) and so on.

There is a subtle, but important point: If the lognormal random variable becomes zero after the rounding, we *omit it.* This is necessary, since, e.g., we don't know how many vocabulary words have *not* appeared in our document. Notice that this omission leads to the so-called "truncated" or "veiled" random variables, which are *notoriously* difficult with respect to their parameter estimations, in the continuous case.

3.1 Probability Distribution Function

After this motivation, we are ready to present our proposed discrete PDF. We propose to use the distribution with the following PDF:

$$P(x=k) = \frac{A(\mu,\sigma)}{k} \exp\left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right] \qquad k = 1, 2, \ldots \tag{4}$$

where

$$A(\mu,\sigma) = \left\{\sum_{k=1}^{\infty} \frac{1}{k} \exp\left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right]\right\}^{-1}$$

is a normalization constant depending on μ and σ.

This PDF has the following characteristics

- It is discrete, which means it is suitable to model many real discrete distributions.
- It is a discretized version of a known continuous distribution, the lognormal distribution. As we know, the PDF of a lognormal distribution is a parabola in log-log plot, which is next simplest model beyond a straight line.
- This model has only two parameters to estimate, so it is not difficult to compute.
- The parameters μ and σ provide sufficient statistics for this distribution.
- As we will show in next section, DGX includes Zipf's law as a special case.

3.2 Zipf's law as a special case

Lemma 1 *The Discrete Gaussian Exponential (DGX) as defined by Eq.(4) reduces to Zipf's law as $\mu \to -\infty$.*

Proof: We first rewrite Eq.(4) as

$$P(x=k) \propto \frac{1}{k}\exp\left(-\frac{\ln k(\ln k - 2\mu)}{2\sigma^2}\right)$$

Assume that $\ln k << |\mu|$, the PDF becomes

$$P(x=k) \propto \frac{1}{k}\exp\left(\frac{\mu \ln k}{\sigma^2}\right) \propto k^{-1+\mu/\sigma^2}$$

which reduces to generalized Zipf distribution (See Eq.(3)) . **QED**

Corrolary 1 *When $\mu \to -\infty$, the DGX distribution is reduced to Zipf's law with slope $\phi = 1 - \mu/\sigma^2$.*

As we will see later from the results of our experiments, DGX works well on real datasets both when their PDF has a clear curvature and when the PDF is straight in log-log plot.

3.3 Estimation of parameters

Two major methods have been used to fit the skewed data with proposed models. One is to fit the frequency-rank plot with linear or nonlinear regression [FJ92], while the other is to fit the PDF with maximum likelihood estimation (MLE). We believe the second method is statistically sound, therefore we choose to use MLE to estimate parameters, μ and σ, in DGX. If the data are $x_1, \ldots, x_n$, the likelihood is

$$L(\mu,\sigma) = \prod_{i=1}^{n} P(x_i) = A(\mu,\sigma)^n \prod_{i=1}^{n} \frac{1}{x_i} \exp\left[-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right] \tag{5}$$

and its logarithm, the log-likelihood is

$$l(\mu,\sigma) = n\ln A(\mu,\sigma) - \sum_{i=1}^{n}\left[\ln x_i + \frac{(\ln x_i - \mu)^2}{2\sigma^2}\right] \tag{6}$$

We then maximize $l(\mu,\sigma)$ numerically to estimate parameters, μ and σ. For clarity, we describe DGX on the count-frequency of words in documents, but, of course, the same algorithm applies to any setting. The full algorithm is as follows,

```
**Algorithm** *DGX_Estimator*
**Input:** A sequence ("multiset") of $N$ words $w(i), i = 1, \ldots, N$ appeared in a document
**Output:** Estimated parameters, $\mu$ and $\sigma$
1.  Create an associative array word_count (as in Perl) to store the count of distinct
    words
2.  Create another associative array y to store distinct word frequencies.
```

```
3.  for i ←1 to N
4.          w_id ←w(i)
5.          word_count(w_id) ←word_count(w_id) + 1
6.  (* $V$ is vocabulary size, i.e., the number of distinct words. *)
7.  V ←size(word_count)
8.  for i ←1 to V
9.          key ←word_count(i)
10.         y(key) ←y(key)+1
11. (* para is used to pass parameters to loglikelihood function. *)
12. (* $\mu_0$ and $\sigma_0$ are initial values for $\mu$ and $\sigma$ *)
13. (* $y$ is the count-frequency data *)
14. (* tolerance is used to set the stopping criterion *)
15. para ←($\mu_0$, $\sigma_0$, y, tolerance)
16. (* Call a maximization routine to find the optimal parameters, $\mu$ and $\sigma$. Here
    the loglikelihood_func is a function which evaluates the loglikelihood (as defined
    in Eq.(6)) given a certain ($\mu$, $\sigma$) pair. *)
17. $[\mu, \sigma]$ = Maximization(loglikelihood_func, para)
18. Output $[\mu, \sigma]$ and exit.
```

We notice that we only need to go over the dataset once (step 3 to 5) to obtain the frequency vector, `word_count`, and go over the frequency vector once to obtain the count vector, y. then the estimation was carried out only on the count vector. This computation can be done fast.

In Step 16, we called an optimization function, e.g. fminsearch [Inc] in Matlab, to which we pass the function *loglikelihood_func* and its parameters *para*=(tolerance, μ_0, σ_0).

3.4 Experiments

Datasets Description The DGX is designed to fit a wide variety of datasets. We thus applied it to three datasets from completely different fields:

- Text: the English Bible. There are totally about 800000 words and the size of the vocabulary is $V \approx 12500$.
- Sales data from a large retail chain, in which there are hundreds of branches. This dataset, which includes all sales information of the store in one week, is about $10GB$ large. We studied the count-frequency relation of the products. Here the products play the role of vocabulary words and the sales of products correspond

to the count of vocabulary words.

- Telecommunications data - customer data from an AT&T service of monthly usage volumes, broken down by customer. We used three instances of this data, each from a different geographic region, which we refer to as Region A, Region B, and Region C.
- Clickstream data: This dataset is obtained from an ISP which collects information about Internet users' browsing behavior. We studied this dataset from two angles, the count-frequency relation of website traffic (the distribution of websites versus the number of visits they receive) and the count-frequency relation of user sessions (The distribution of users versus the number of websites they visit). In the first case, the web sites play the role of vocabulary words and the number of visits they receive corresponds to the count of vocabulary words; in the second case, the web users play the role of vocabulary words, and the number of web sites they visit corresponds to the count of vocabulary words.

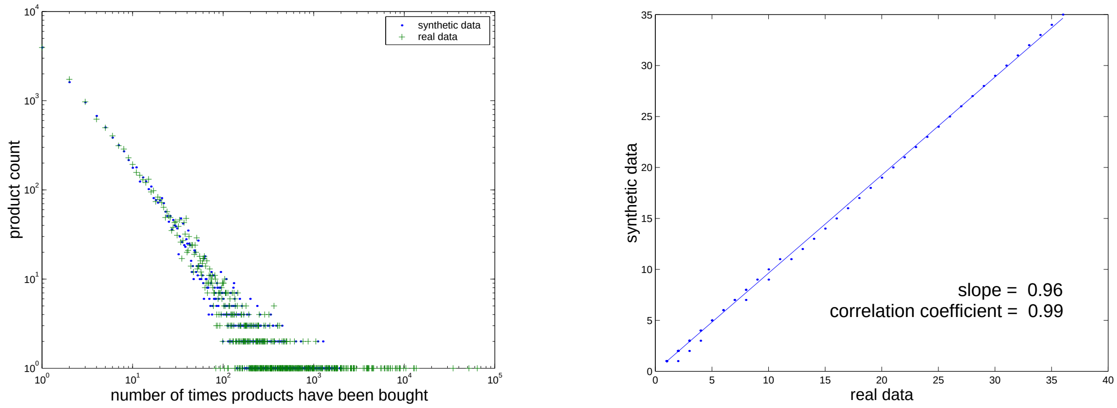
All these datasets show extremely skewed behavior, i.e., we expect to see that very few products, or websites are really popular, while most products have low sales, and most websites have low traffic. Therefore, it is meaningless to talk of the mean, median or variance of these data. To characterize these data, we need to use some skewed distribution. We also observe that Zipf's Law often fails, i.e., the PDF in log-log scale shows a clear curving trend. However, DGX gives excellent fits in all cases we tested, including when the dataset is very Zipf-like as well as when it deviates from Zipf's law very much.

Goodness of Fit The technique we used to test the goodness-of-fit is the traditional quantile-quantile plot (qqplot). The qqplot compares the quantiles of two datasets. If the two datasets are from the same distribution, the qqplot should be linear and the slope should be one. We first use the original data to estimate the parameters of DGX. We then use DGX and the estimated parameters to generate a synthetic dataset. Next, we make a qqplot between the real and the synthetic datasets. Then, we fit the qqplot with a straight line and compute the slope and correlation coefficient. If both are close to one, we can claim that the real data and the synthetic data are from the same distribution.

3.5 Results

3.5.1 Text data

We first apply DGX to text data from the Bible. The results are shown in Figure (2). We notice that the real data and the synthetic data are in good agreement. The slope and the correlation coefficient of the qqplot are both very close to one, which indicates we obtain an excellent fit.



Count-frequency plot of words in Bible.

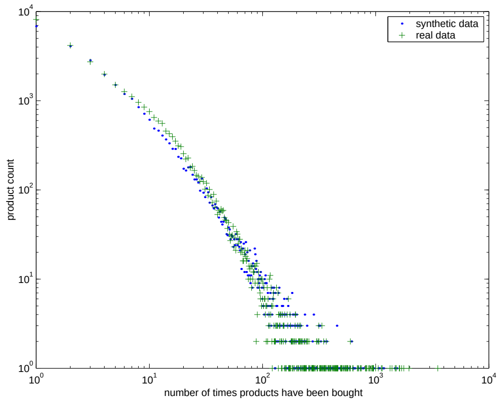
qqplot of real and synthetic data for words in Bible

Figure 2: Count-frequency plots of real data and synthetic data for words in Bible. Here, $\mu = -2.106$ and $\sigma = 3.23$. We find that the synthetic data match the real data very well. The qqplot is practically linear, the slope and the correlation coefficient are close to unity. All indicate that DGX gives an excellent fit.

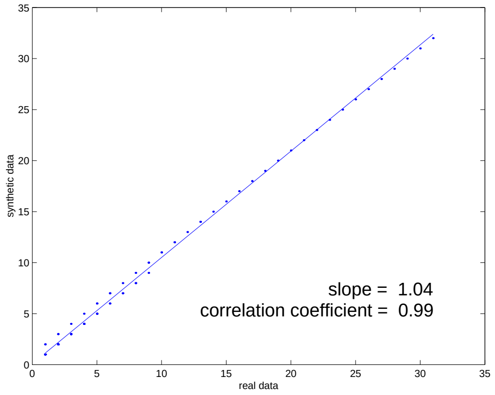
3.5.2 Sales data

We then applied DGX to sales data from the three largest branches of a retail chain. For each store, we use the sales data to estimate parameters μ and σ in our distribution. With the estimated parameters we generate a set of synthetic data. Then, for each branch, we make the count-frequency plot and the qqplot as in Figure (3). We notice they have similar count-frequency plots. Their parameters, μ and σ, have similar values. As we will see later, some other stores have very different parameters and their count-frequency plots have different shapes.

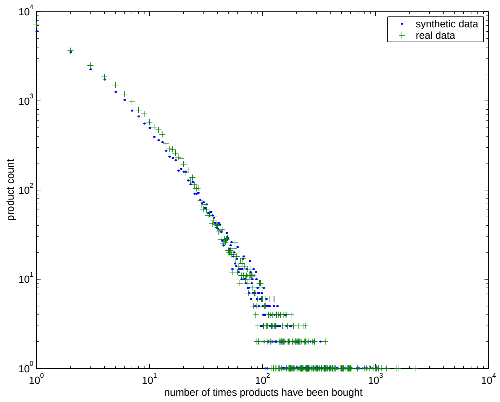
From the Figure (3), we observe an excellent fit between the synthetic data and the real data. In the count-frequency plot which is clearly not a straight line, DGX gives a nice fit. We also observe that the slope and correlation coefficient of the qqplot are very close to one, which also indicates the data is expressed with DGX very well.



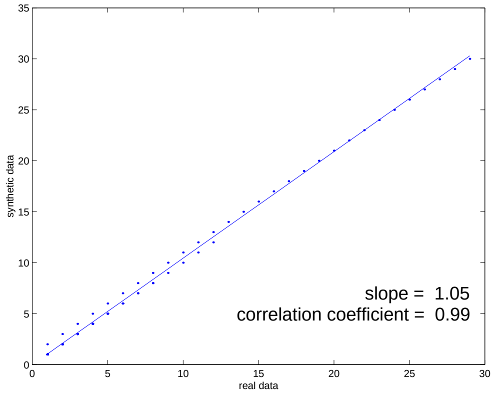
Count-frequency plot for store no. 96. $\mu = 0.999$ and $\sigma = 1.682$



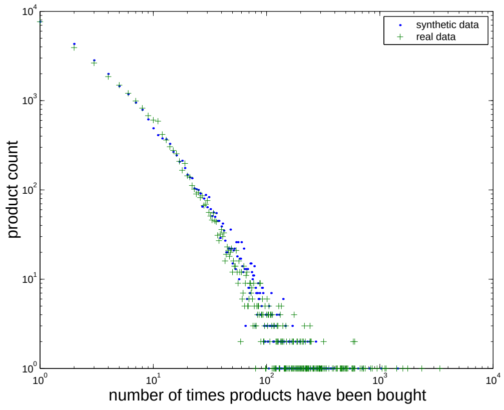
qqplot of real and synthetic data for store no. 96



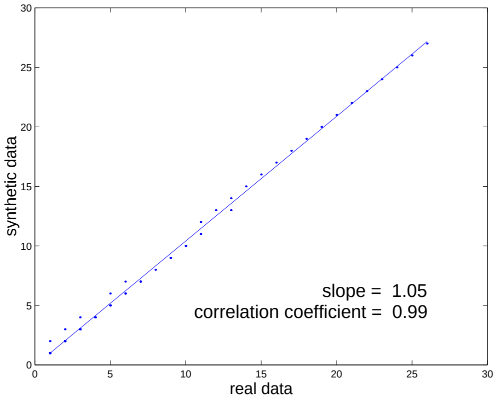
Count-frequency plot for store no. 82. $\mu = 0.905$ and $\sigma = 1.601$



qqplot of real and synthetic data for store no. 82



Count-frequency plot for store no. 101. $\mu = 0.788$ and $\sigma = 1.542$



qqplot of real and synthetic data for store no. 101.

Figure 3: Count-frequency plots of real data and synthetic for store 96, 82 and 101 and their qqplots. We notice that the real and the synthetic data are in good agreement. The qqplot is almost linear, the slope and the correlation coefficient are close to unity. All indicate that DGX gives an excellent fit.

3.5.3 Telecommunication Data

We then apply DGX to customer data from an AT&T service of monthly usage volumes, broken down by customer. We used three instances of this data, each from a different geographic region, which we refer to as Region A, Region B, and Region C. Following the same procedures, we obtain the results shown in Figure (4). Again, this dataset is fit very well with DGX.

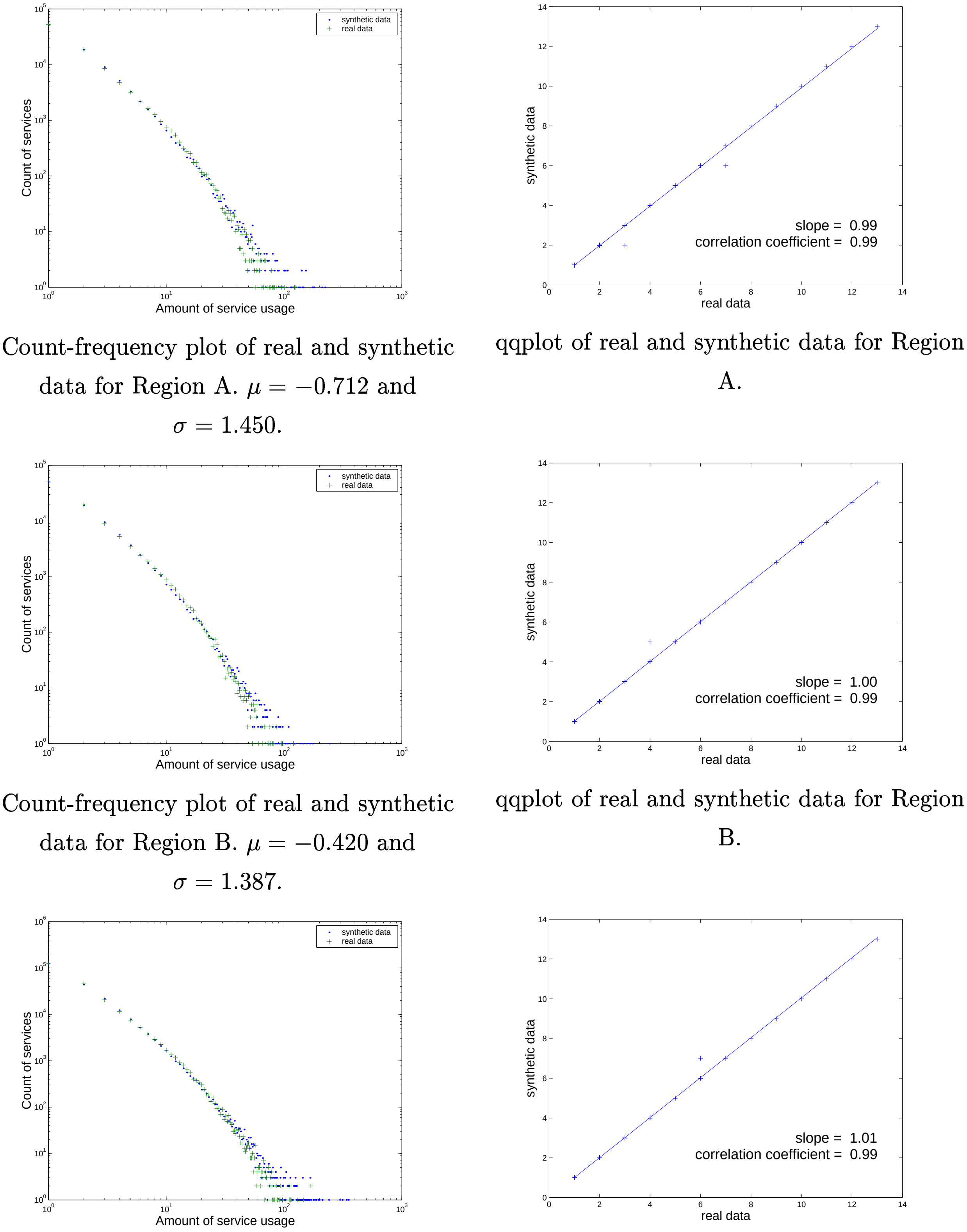
3.5.4 Clickstream Data

We also apply DGX to the clickstream data. We study the count-frequency relation of the websites and the users. The count-frequency plot of website traffic, as shown in Figure (5-(a)) shows a clear Zipf-like behavior, while the count-frequency plot of users deviates significantly from Zipf's law. However, both distributions can be fit well with our DGX.

3.6 Discussion

Skewed distributions, like the count-frequency data as we described above, exist in many fields of natural and social sciences. They can not be represented well with the "usual" features like the mean, median, maximum or minimum. For example, most words appear only once in Bible while a few common words appear very often. The mean is 63.0, the median is 3, the maximum is 63924, and the minimum is 1. They are all meaningless about this very skewed distribution. We therefore propose a new discrete distribution, DGX, which seems to be an excellent tool to model skewed data. The features we obtain with DGX are μ and σ, which can be used for datamining, such as clustering or outlier detection.

To illustrate the datamining power of DGX, we apply it to the sales data of all branches of the retail chain and obtain a (μ, σ) pair for every store. In Figure (6), we make a scatter plot of (μ, σ) pairs and mark a few outlier stores according to the parameters. Notice that store No. 4 and No. 31 are outliers in (μ, σ) plane; Figure (7) gives the count-frequency plots for these two as well as some other "mainstream" stores. It is clear that No. 4 and No. 31 have more linear plots while the others have curving plots. Moreover, closer inspection shows that these two have smaller sales volume. This shows that the outlier detection in (μ, σ) indeed successfully discovered some stores with "abnormal" distribute sales data.



Count-frequency plot of real and synthetic data for Region A. $\mu = -0.712$ and $\sigma = 1.450$.

qqplot of real and synthetic data for Region A.

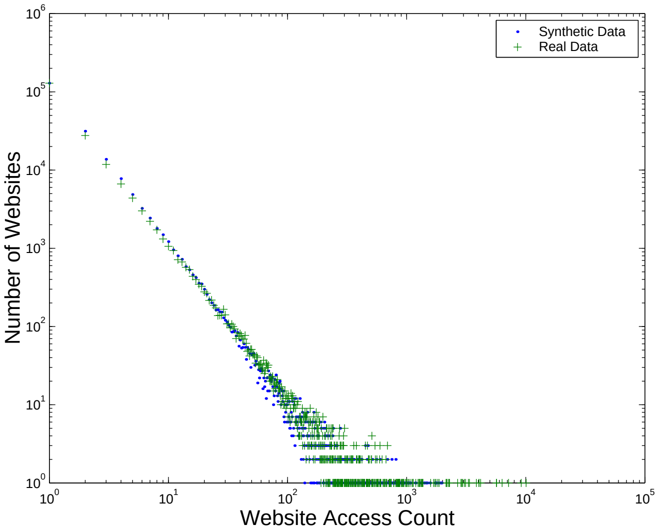
Count-frequency plot of real and synthetic data for Region B. $\mu = -0.420$ and $\sigma = 1.387$.

qqplot of real and synthetic data for Region B.

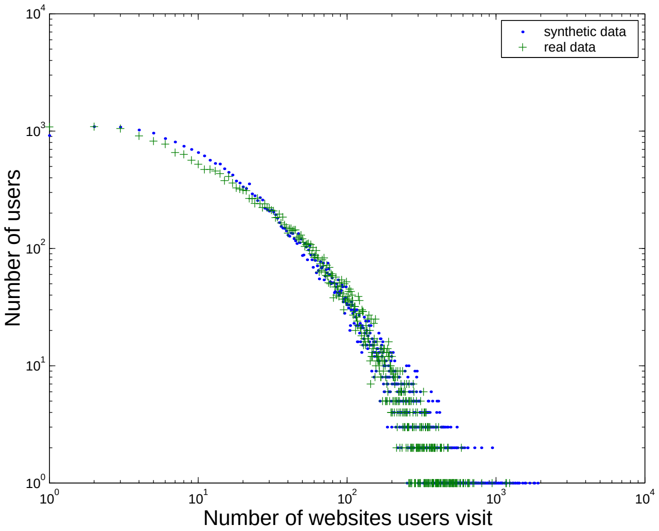
Count-frequency plot of real and synthetic data for Region C. $\mu = -0.64$ and $\sigma = 1.418$.

qqplot of real and synthetic data for Region C.

Figure 4: Count-frequency plots of service usage data from AT&T. We show real data and synthetic for three regions and their qqplots. We notice that the real and the synthetic data are in good agreement. The qqplot is almost linear, the slope and the correlation coefficient



(a) Count-frequency plot of website visits. Estimated parameters are $(\mu = -60.35, \sigma = 7.68)$. We observe that this distribution is very Zipf-like and it has a large negative μ. This seems to agree with Lemma 1.



(b) Count-frequency plot of user sessions. Estimated parameters are $(\mu = 2.86, \sigma = 1.42)$. We observe that this dataset deviates significantly from Zipf's law, but it can still be modeled well with DGX.

Figure 5: Count-frequency plots of website visitors and user sessions. They show very different behavior, but both can be modeled well with DGX

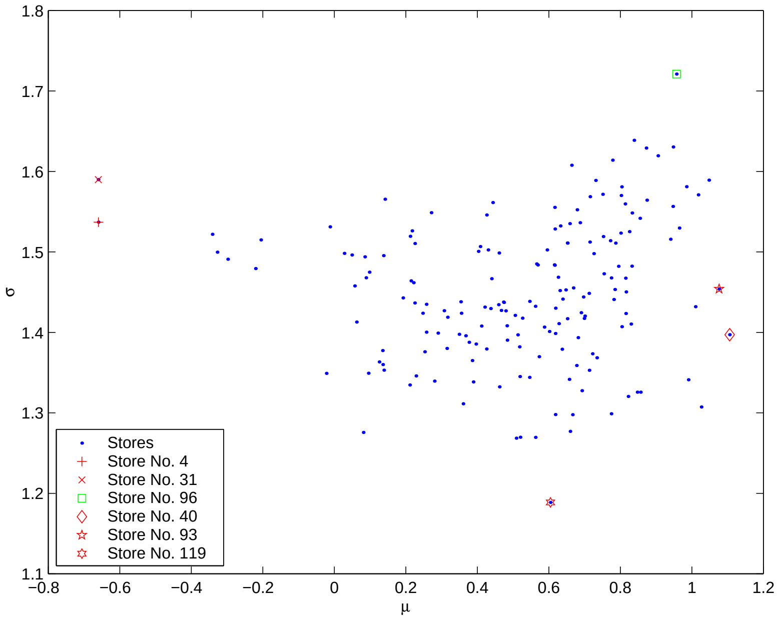


Figure 6: Scatter plot of μ and σ of all branches of a retail chain. We clear see stored No. 4 and No. 31 are outliers compared to the majority.

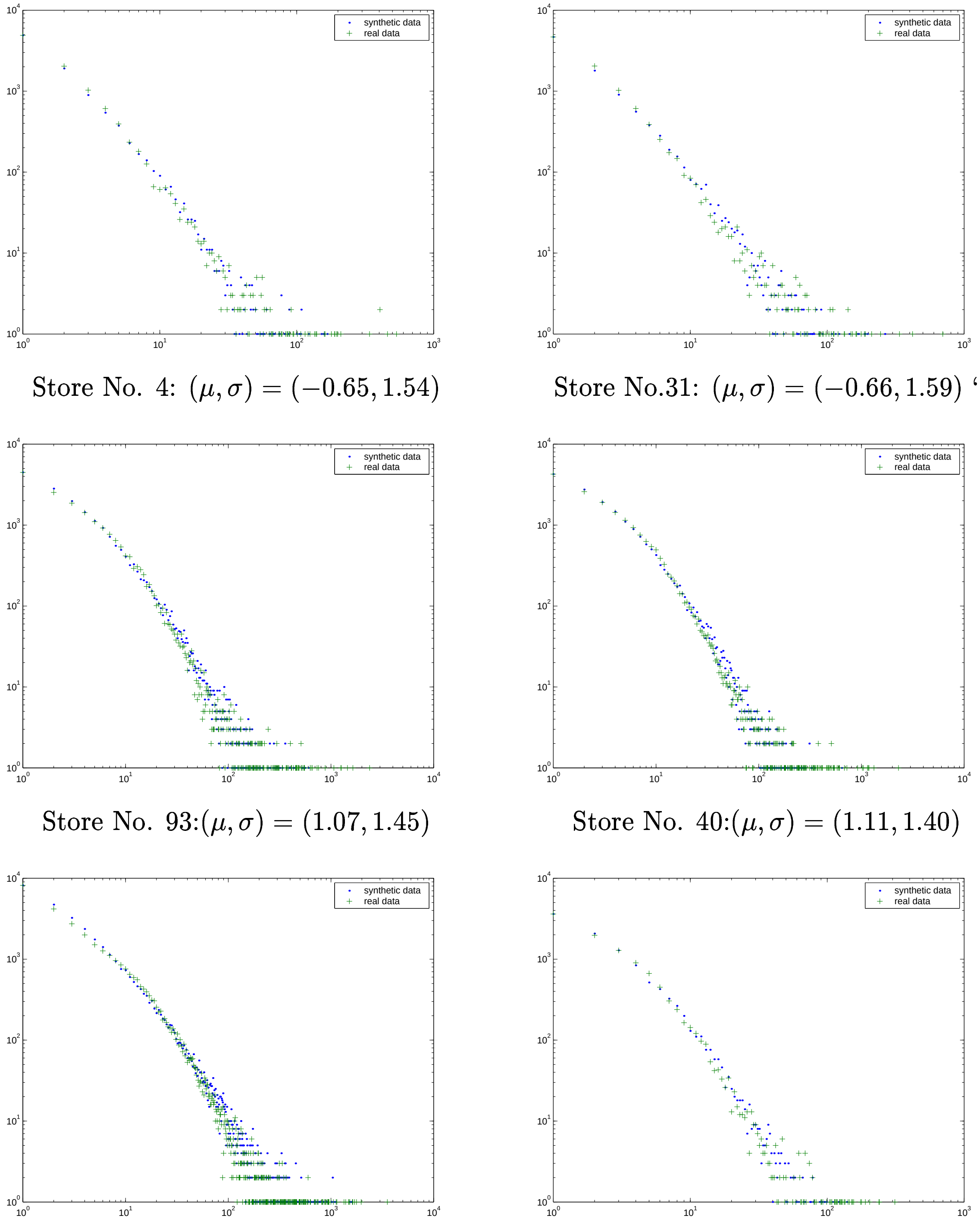
4 Skewed Distribution in Two Dimensions

We saw that many one dimensional skewed data are modeled well by DGX, the question is whether a similar disctribution can be used to model skewed data sets in two dimensions. For example, the distribution of phone callers in an area depend on both the number of calls and the duration of phone calls, and the distribution is skewed in both dimensions. A plot of phone call data is show in Fig 8. Another example is the clickstream data. The number of users is skewly distributed on both the duration they are online and the number of sites they visit. If the two variables are correlated, learning the two marginal distributions does not reveal the overall distribution. Therefore, we need an overall two dimensional skewed distribution.

4.1 Two Dimensional Zipf's Law

In order to model skewed distributions in two dimensions, a natural extension is to assume that the maginals of these data follow Zipf's law and independent. Therefore, we first use a two dimensional Zipf's law, which has a PDF of the form

$$P(k,d) = A(\alpha,\beta)n^{-\alpha}d^{-\beta} \quad n,d = 1,2,\ldots \tag{7}$$



Store No. 4: $(\mu, \sigma) = (-0.65, 1.54)$

Store No.31: $(\mu, \sigma) = (-0.66, 1.59)$ ‘

Store No. 93:$(\mu, \sigma) = (1.07, 1.45)$

Store No. 40:$(\mu, \sigma) = (1.11, 1.40)$

Store No. 96:$(\mu, \sigma) = (0.96, 1.72)$

Store No. 119:$(\mu, \sigma) = (0.60, 1.19)$

Figure 7: Count-frequency plots of outlier stores according to the (μ, σ) pair in the DGX. We notice that the distributions for Store No.4 and Store. No. 31, which have small μ values are more Zipf-like than the others. In real world, these two stores are two of the smallest stores in terms of sales.

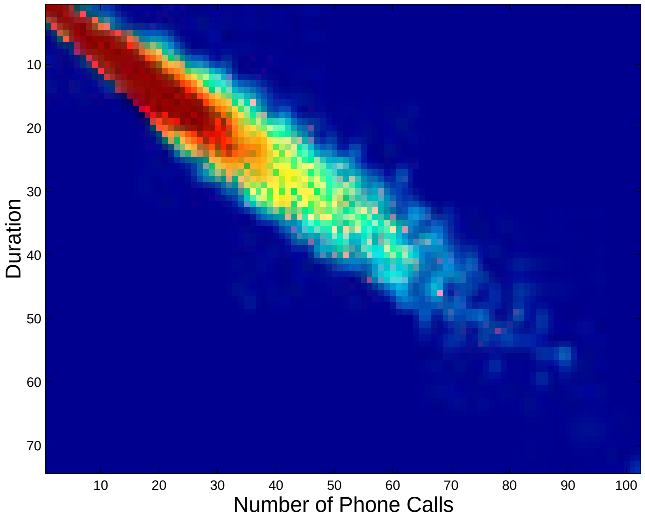


Figure 8: An example of two dimensional skewed data from one of AT&T's service usage. The darker

where $A(\alpha, \beta)$ is a normalization constant, which satisfies

$$A(\alpha, \beta) = \frac{1}{\sum_{n=1}^{\infty} \sum_{d=1}^{\infty} n^{-\alpha} d^{-\beta}} = \frac{1}{\sum_{n=1}^{\infty} n^{-\alpha} \sum_{d=1}^{\infty} d^{-\beta}} = \frac{1}{\zeta(\alpha)\zeta(\beta)} \tag{8}$$

where $\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$ is the Riemann Zeta function. Obviously, this PDF is a plane in log-log scale.

The likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{N} A(\alpha, \beta) n_i^{-\alpha} d_i^{-\beta} = A^N(\alpha, \beta) \prod_{i=1}^{N} n_i^{-\alpha} d_i^{-\beta} \tag{9}$$

and the loglikelihood function is

$$\begin{aligned} l &= N\ln A(\alpha, \beta) - \sum_{i=1}^{N} (\alpha \ln n_i + \beta \ln d_i) \\ &= -N\ln\zeta(\alpha) - N\ln\zeta(\beta) - \sum_{i=1}^{N} (\alpha \ln n_i + \beta \ln d_i) \end{aligned} \tag{10}$$

To maximize the loglikelihood function, we set the following partial derivatives to zeroes:

$$\frac{\partial l}{\partial \alpha} = -N \frac{\zeta'(\alpha)}{\zeta(\alpha)} - \sum_{i=1}^{N} \ln n_i = 0 \tag{11}$$

$$\frac{\partial l}{\partial \beta} = -N \frac{\zeta'(\beta)}{\zeta(\beta)} - \sum_{i=1}^{N} \ln d_i = 0 \tag{12}$$

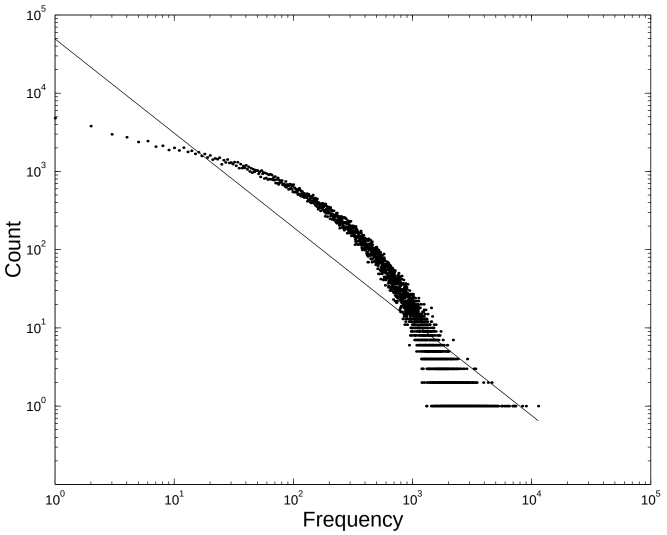


Figure 9: Frequency-Count plot of number of phone calls using Zipf's Law

Because the PDF implies that the two random variables,n and d are independent of each other, we can treat this distribution as the product of two indenpendent 1-d Zipf's law. Therefore, we can fit the two marginal with Zipf's distribution. We do an experiment on a dataset from AT&T and obtain results as shown in Fig.(9, 10)

From Fig. 9 and 10, we see their maginal distributions on each dimension does not fit very well with Zipf's law, and they look more like DGX distribution. This prompts us to try a DGX distribution in two dimensions.

4.2 DGX in Two Dimensions

Since we realize the two dimensional Zipf's law can not successfully model skewed data in two dimensions because it lacks the ability to address the correlation of the two variables, we extend DGX to two dimensions. We start with the following PDF:

$$P(x,y) = \frac{A}{x\,y} \exp\left\{ -(\ln x - \mu_x, \ln y - \mu_y) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \ln x - \mu_x \\ \ln y - \mu_y \end{pmatrix} \right\} \tag{13}$$

where μ_x, μ_y, σ_{11}, σ_{12} and σ_{22} are parameters of this distribution, A is a normalization constant satisfying

$$A(\mu_x, \mu_y, \sigma_{11}, \sigma_{12}, \sigma_{22}) = \left[\sum_{x=1}^{\infty} \sum_{y=1}^{\infty} \frac{1}{x\,y} \exp\left\{ -(\ln x - \mu_x, \ln y - \mu_y) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \ln x - \mu_x \\ \ln y - \mu_y \end{pmatrix} \right\} \right]^{-1} \tag{14}$$

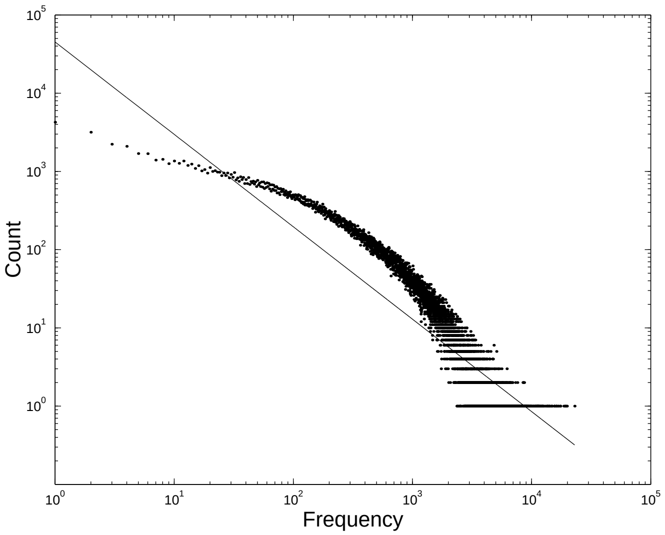


Figure 10: Frequency-Count plot of duration of phone calls using Zipf's law

and $x, y = 1, 2, \dots, \infty$. The likelihood function is

$$L = A^n \prod_{i=1}^{n} \frac{1}{x_i \; y_i} \exp \left\{ -(\text{ln}x_i - \mu_x, \text{ln}y_i - \mu_y) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \text{ln}x_i - \mu_x \\ \text{ln}y_i - \mu_y \end{pmatrix} \right\} \tag{15}$$

and loglikelihood function is

$$l = n\text{ln}A - \sum_{i=1}^{n} \left\{ \text{ln}x_i + \text{ln}y_i + (\text{ln}x_i - \mu_x, \text{ln}y_i - \mu_y) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \text{ln}x_i - \mu_x \\ \text{ln}y_i - \mu_y \end{pmatrix} \right\} \tag{16}$$

The questions are

1. how to estimte parameters in this two dimensional distribution
2. how to evaluate the goodness of fit
3. how this fit works for real and synthetic data.

4.2.1 Fitting and parameter estimation

We tried several methods to do the fitting. The first one is the normal maximum likelihood estimation. Unfortunately, we couldn't find an efficient way to accurately evaluate the normalization constant, A. Another approach is to fit the data with function

$$h = ax^2 + bxy + y^2 + dx + ey + f$$

where h is the logarithm of count, and x and y are the logarithms of two variables, such as duration and frequency. We can also write h as

$$h = \begin{pmatrix} x - \mu_x & y - \mu_y \end{pmatrix} \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}$$

The relations between two sets of coefficients are

$$\begin{aligned} \sigma_{xx} &= a && (17) \\ \sigma_{xy} &= b && (18) \\ \sigma_{yy} &= c && (19) \\ \mu_x &= \frac{-d - \sigma_{xy}}{\sigma_{xx}} && (20) \\ \mu_y &= \frac{-e - \sigma_{xy}}{\sigma_{yy}} && (21) \end{aligned}$$

4.2.2 Goodness of fit

Since qqplot can not be directly used to evaluate the goodness of fit in two dimensions, we instead first find the marginals on x direction, y direction and the 45 degree diagonal between x and y direction, where x, y refer to the two variables in the two dimensional distribution. To test the goodness of fit, we first generate a synthetic data set using the estimated parameters, then we project the synthetic data set onto three directions mentioned above and make qqplots for these three marginals. If they all show good fit, we can claim our model with estimated parameters fit the original data well.

4.3 Experiments

We did experiments to answer the following questions

1. Sanity check: Does our method work on synthetic two dimensional data, recovering the parameters μ and σ correctly?
2. How well does the two dimensional DGX fit the real data?

4.3.1 Sanity check with synthetic data

To test our model, we first generate two synthetic data sets "SYNTH1" and "SYNTH2". SYNTH1 is generated using parameters $\mu = (1, 1)^T$ and

$$\sigma = \begin{pmatrix} -0.6 & 0.4 \\ 0.4 & -0.8 \end{pmatrix}$$

SYNTH1 contains 4574 data points. After we fit it with our model, we estimated the parameters with reasonable accuracy, $\hat{\mu} = (0.999, 1.002)^T$ and

$$\hat{\sigma} = \begin{pmatrix} -0.601 & 0.406 \\ 0.406 & -0.812 \end{pmatrix}$$

In Fig.11, we plot the marginals of the real data and synthetic data projected in x direction, y direction and the 45 degree diagonal between x and y direction. and their qqplots. We can see the real and synthetic data match very well in all directions. Their qqplots are straight lines with slope and correlation coefficient close to unity, which also indicates out model fits the data very well.

For SYNTH2, we generate a synthetic data set using parameters $\mu = (-1, -1)^T$ and

$$\begin{pmatrix} -0.6 & 0.4 \\ 0.4 & -0.8 \end{pmatrix}$$

SYNTH2 contains 4574 data points. After we fit these synthetic data with our model, we estimated the parameters with reasonable accuracy, $\hat{\mu} = (-1.028, -1.024)^T$ and

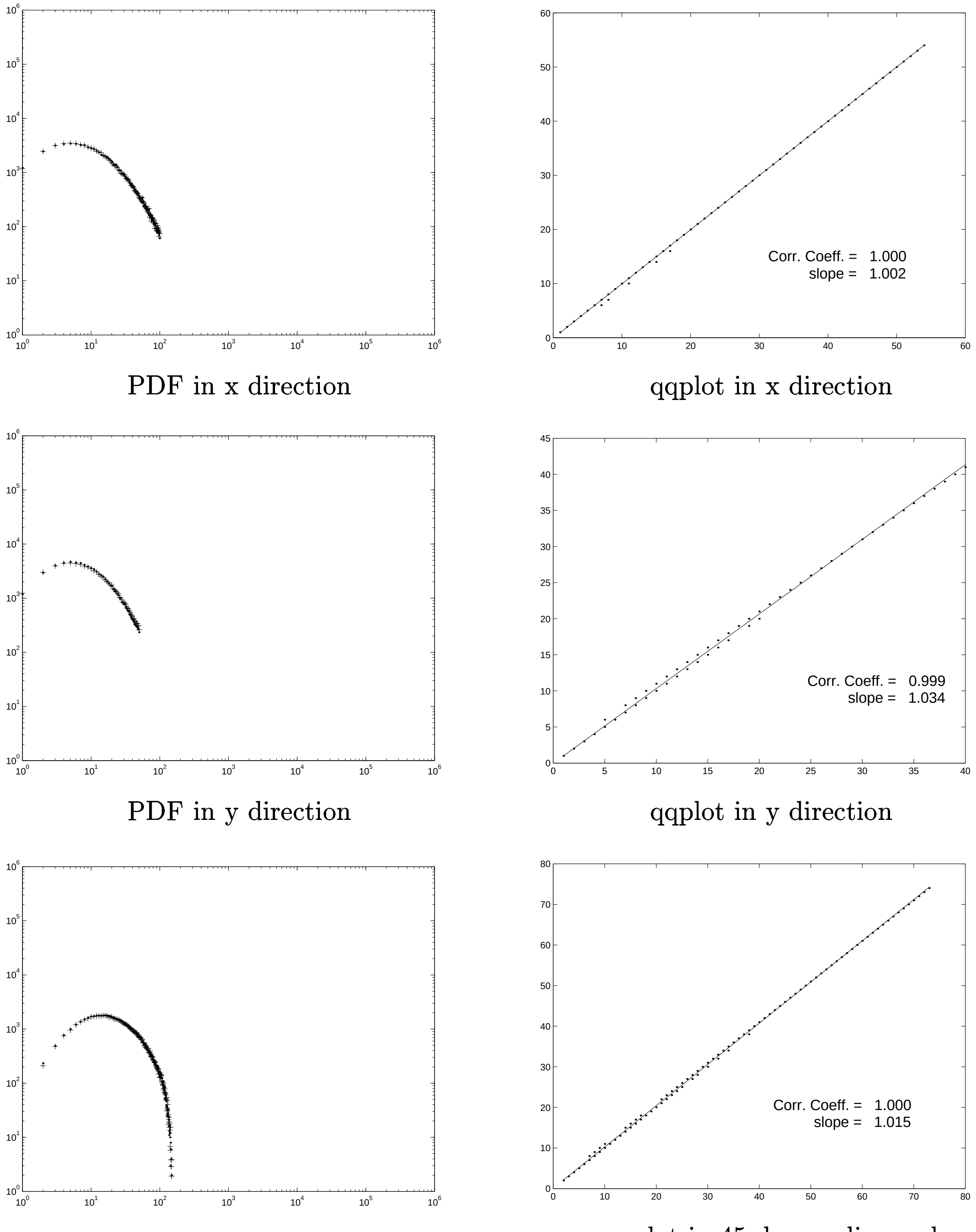
$$\hat{\sigma} = \begin{pmatrix} -0.589 & 0.395 \\ 0.395 & -0.763 \end{pmatrix}$$

The plots for SYNTH2 is shown in Fig. 12. Again, a good fit is obtained.

4.3.2 Experiments with real data

We also experiment with two dimensional DGX on two real world data sets. The first data set is the usage data of a telecommunication service from AT&T from one service area. The distribution of customers is very skew depending on the number of the of the service sessions and the duration of the service sessions. Another data set is a clickstream data obtained from an Internet service provider. Again the distribution of Internet users is skewly distributed on the number of sites they visit and the total duration of their online time.

For these two real datasets, we experience great difficulty estimating the parameters, mainly because of a problem of convergence: we are not able to accurately compute the normalization constant. We therefore use a crude method manually looking for parameters μ and σ which give good fits graphically. In Fig. 13, we plot the maginals projected on x direction, y direction and in 45 degree between x and y directions. Clearly, our model achieves a good fit, because the synthetic data and the real data overlap nicely and the slope and the correlation coefficient of the qqplot are close to one in all cases. The results for the clickstream data is plot in Fig. 14.



PDF in x direction

qqplot in x direction

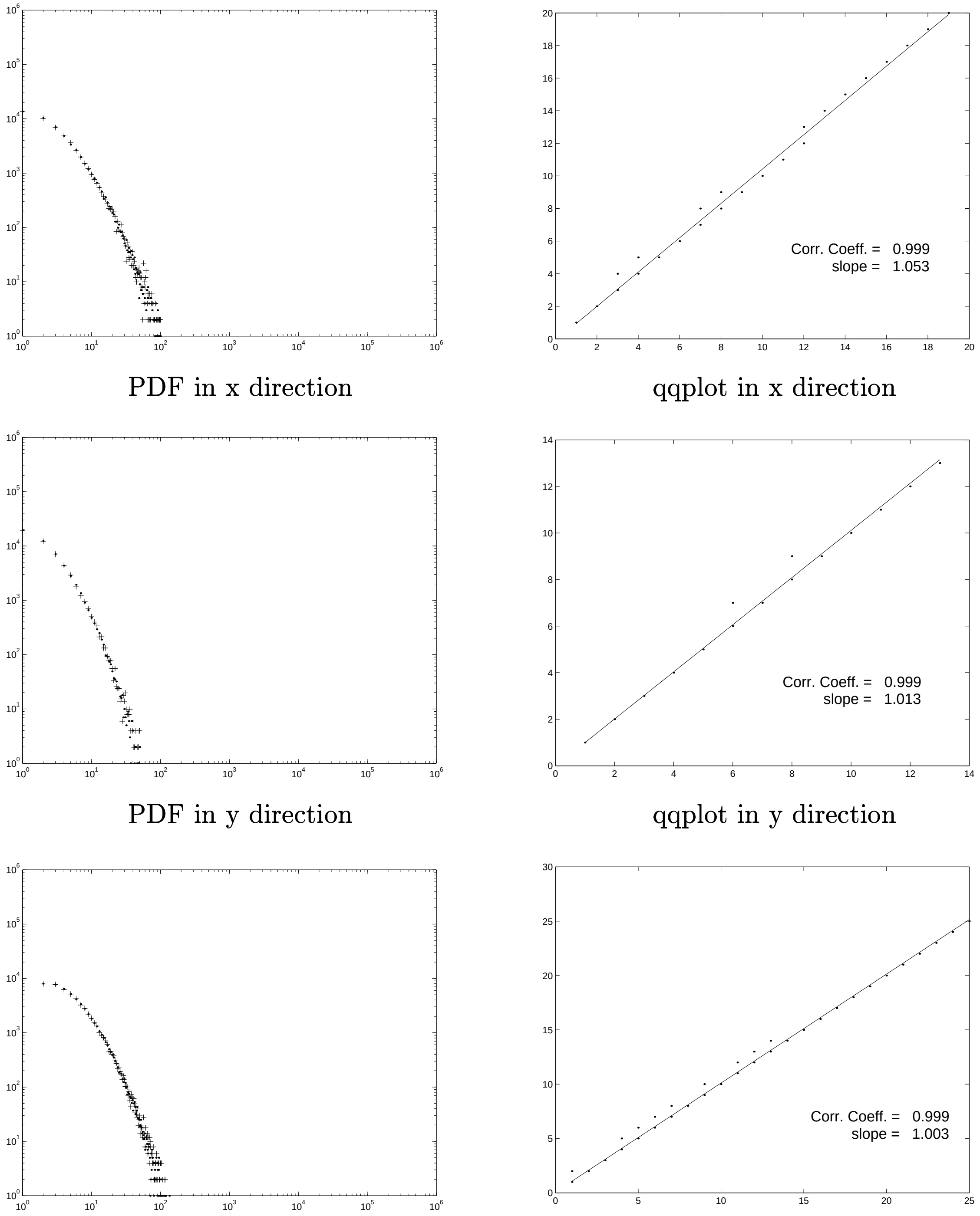
PDF in y direction

qqplot in y direction

PDF in 45 degree diagonal

qqplot in 45 degree diagonal

Figure 11: Marginals of PDF in x direction, y direction and 45 degree diagonal and their qqplots for a synthetic data set



PDF in x direction

qqplot in x direction

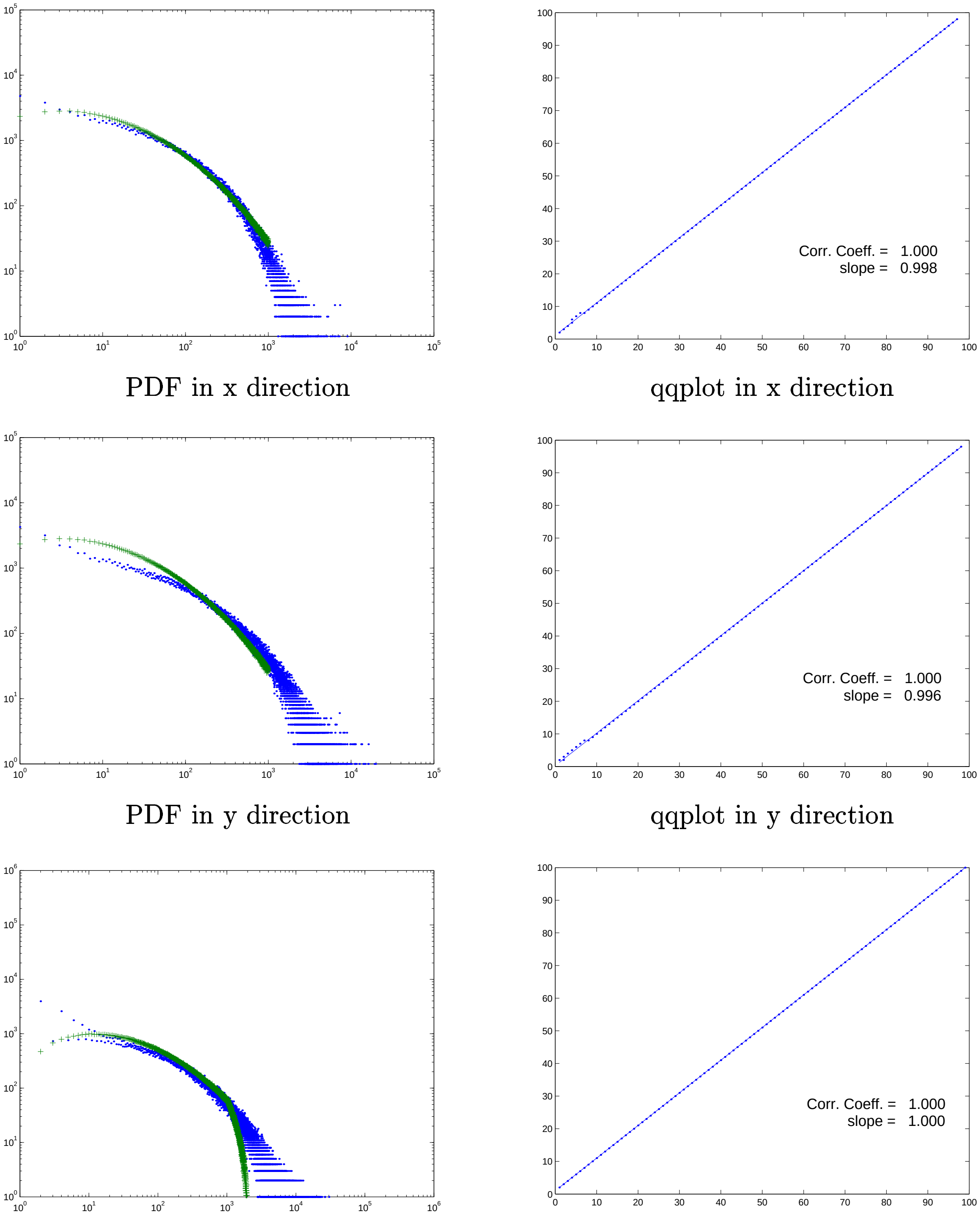
PDF in y direction

qqplot in y direction

PDF in 45 degree diagonal

qqplot in 45 degree diagonal

Figure 12: Marginals of PDF in x direction, y direction and 45 degree diagonal and their qqplots for a synthetic data set



PDF in x direction

qqplot in x direction

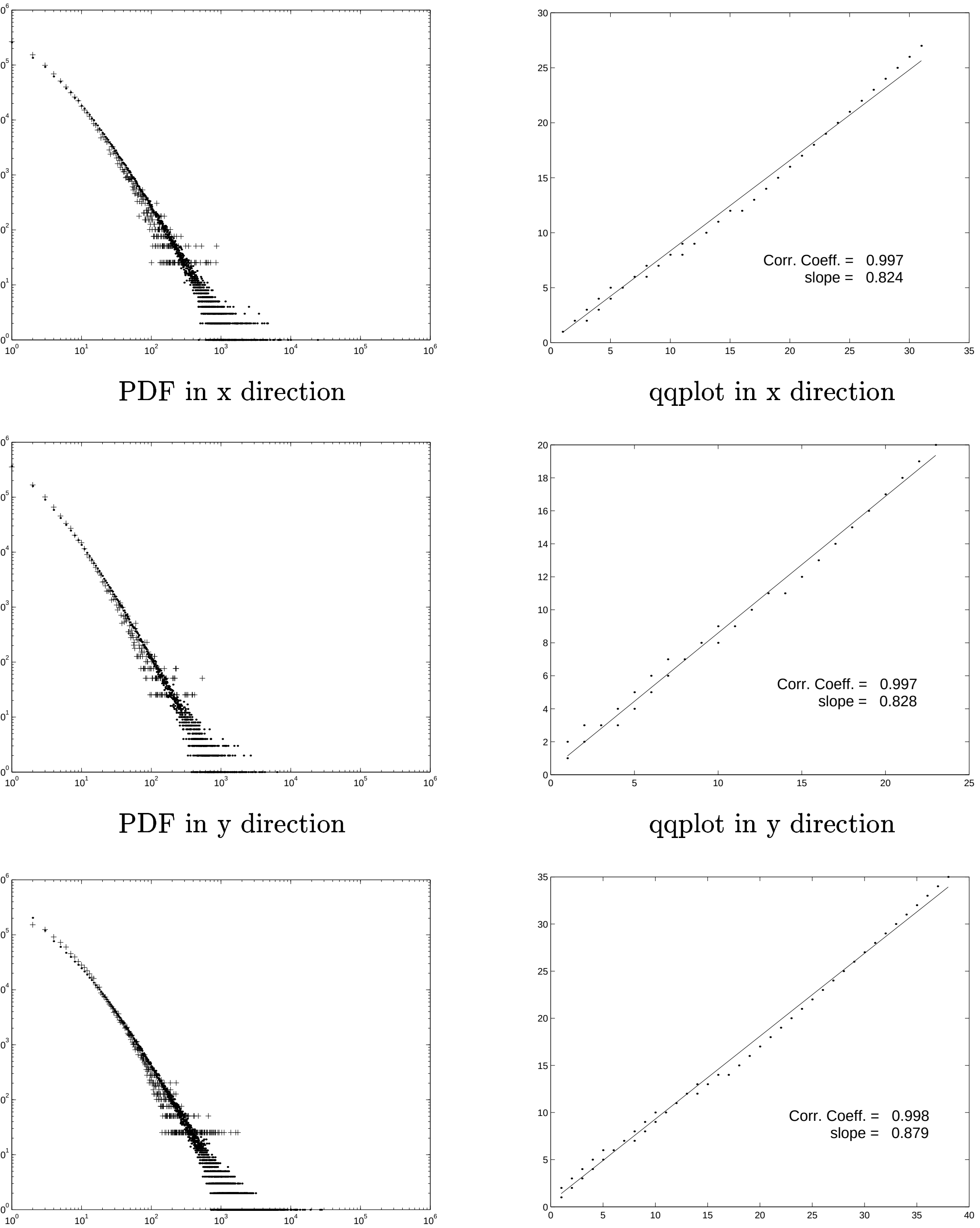
PDF in y direction

qqplot in y direction

PDF in 45 degree diagonal

qqplot in 45 degree diagonal

Figure 13: Marginals of PDF in x direction, y direction and 45 degree diagonal and their qqplots for AT&T data. Here, $\mu = (-2\ \ -2)'$ and $\sigma = ((-0.4\ 0.3)'\ (0.3\ \ -0.4)')$



PDF in x direction

qqplot in x direction

PDF in y direction

qqplot in y direction

PDF in 45 degree diagonal

qqplot in 45 degree diagonal

Figure 14: Marginals of PDF in x direction, y direction and 45 degree diagonal and their qqplots for clickstream data. Here, $\mu = (-6\ -6)'$ and $\sigma = ((-0.38\ 0.3)'\ (0.3\ -0.42)')$

5 Conclusions

Skewed distributions appear very often in practice. They are often modeled well by "power" laws or the (generalized) Zipf distribution. However, they often suffer from deviations, like the 'top-concavity'.

The main contribution of this work is to draw attention *away* from power laws, and into DGX, a new, discrete distribution, that has amazing modeling properties:

- It includes the Zipf and generalized Zipf distributions as special cases - thus, it is applicable to all the numerous settings that Zipf works well.
- It is related to the "lognormal" distribution, which models well a *huge* number of continuous distributions; it can also be derived from 'first principles', like the principle of "proportional effects" in economics ([NJB94], page 210)
- It models very well several discrete, real-life distributions, from retailer sales data to telecommunication data to web-hits, with practically perfect correlation coefficient in the traditional quantile-quantile ("qq") plots.
- It is parsimonious, requiring only two parameters (μ and σ), to describe the distribution nearly perfectly.
- Its parameters can be estimated with a *single* pass over the dataset
- It can be naturally extended to two dimensions, where we showed that it models well real two dimensional data.

We provided a statistically sound method to estimate the parameters, using the Maximum Likelihood, and we showed how to use DGX to find patterns and outliers in a collection of many skewed distributions, like branches that have clearly different patterns than the rest.

The μ and σ parameters of DGX are valuable for data mining, clustering and outliers, because they constitute consice, but accurate "features" of a discrete distribution. In contrast, for skewed distributions, the obvious 'features' of mean, median, minimum, maximum, and variance are practically useless: The minimum value is almost always '1'; the maximum value (eg., the salary of the Queen of England in a dataset with salaries) is so large and so unrelated to the rest of the data that it is useless as a feature; the mean is 'high-jacked' by the few outliers; the standard deviation tends to infinity, because of the so-called "heavy-tail" property of the Pareto-like distributions; and the median is low, but it still fails to convey much information about the rest of the distribution.

6 Acknowledgement

The authors would like to thank the anonymous retailer chain, the anonymous Internet Service Provider and AT&T for providing us their properly anonymized data. We also like to thank Dr. Bill Eddy for some useful discussions and suggestions.

References

[Ada] L.A. Adamic. Zipf, power-laws, and pareto - a ranking tutorial. http://www.parc.xerox.com/istl/groups/iea/papers/ranking/ranking.html.

[BCF$^+$99] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *IEEE Infocom'99*, New York, NY, March 1999.

[BG58] B. Berry and W. Garrison. Alternate explanations of urban rank size relations. *Annals of the Association of American Geographers*, 48:83–91, 1958.

[CCC95] A. Bestavros C. Cunha and M. Crovella. Characteristics of www client-based traces. Tr-95-010, Boston University, April 1995. http://www.cs.bu.edu/groups/oceans/papers/Home.html.

[FJ92] Christos Faloutsos and H.V. Jagadish. On B-tree indices for skewed distributions. In *18th VLDB Conference*, pages 363–374, Vancouver, British Columbia, Aug. 23-27 1992.

[FL83] W.R. Fox and W. Lasker. The distribution of surname frequencies. *International Statistical Review*, 51:81–87, 1983.

[Gal79] Galton. The geometric mean in vital and social statistics. *Proceedings of the Royal Society of London*, 29:365–367, 1879.

[Gla94] S. Glassman. A caching relay for the world wide web. In *Proc. of First International Conference on the World Wide Web*, CERN, Geneva, Switzerland, May 1994. http://www1.cern.ch/WWW94/PrelimProcs.html.

[Hal44] P.R. Halmos. Random alms. *Annals of Mathematical Statistics*, 15:182–189, 1944.

[Her60] G. Herdan. *Small Particle Statistics.* Butterworth's, London, 2 edition, 1960.

[Hil74] B.M. Hill. The rank-frequency form of zipf's law. *Journal of the American Statistical Association*, 69(348):1017–1026, 1974.

[Inc] The MathWorks Inc. Matlab user's guide.

[Lah] J. Laherrère. "parabolic fractal" distributions in nature. http://www.hubbertpeak.com/laherrere/fractal.htm.

[LS98] J. Laherrère and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *European Physical Journal*, B(2):525–539, 1998.

[NJB94] S. Kotz N.I. Johnson and N. Balakrishnan. *Continuous Univariate Distributions Volume 1.* John Wiley & Sons, Inc., U.S.A, 1994.

[Par97] V. Pareto. *Cours d'Economie Politique.* Rouge and Cie, Lausanne and Paris, 1897.

[Sim55] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[VAdO96] M. Crovella V. Almeida, A. Bestavros and A. de Oliveira. Characterizing reference locality in the www. In *Proc. of IEEE International Conference in Parallel and Distributed Information Systems*, Miami Beach, Florida, U.S.A., December 1996. http://www.cs.bu.edu/groups/oceans/papers/Home.html.

[Yul23] G.U. Yule. A mathematical theory of evolution, based on conclusions of dr. j.c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London*, 213:21–87, 1923.

[Zip49] G.K. Zipf. *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology.* Addison Wesley, Cambridge, Massachusetts, 1949.