Indicator Random Variables in Traffic Analysis and the Birthday Problem

Phillip G. Bradford* ‡ Irina Perevalova* Michiel Smid† Charles B. Ward*

Abstract

This paper proposes using collisions of Pareto random variables in traffic analysis and in generating fictitious network traffic that follows various Pareto distributions. Pareto distributions are commonly found in network statistics, but the distributions may be truncated or overlapping, thus making it hard to estimate their sample parameters. Therefore, this paper investigates methods of computing parameters of binned collisions of Pareto random variables.

This paper explores an indicator variable approach to analyzing collisions of Pareto random variables. These collisions are initially modeled by the Birthday problem or paradox and then they are extended to understand independence of collisions. This paper's use of indicator variables simplifies the calculation of higher moments for binned collisions of Pareto random variables.

1 Introduction

Encryption alone is not sufficient for secure communication. That is, many successful security breaches are not the result of finding an opponent's encryption key. An adversary can gain significant information about communicating parties' transmissions just by observing their encrypted traffic. Currently, most direct encryption-protocol cracking attacks are far more expensive in computer and intellectual costs than some simple and straight-forward traffic analysis attacks.

Numerous researchers, for example Barabási and Albert [2], Fabrikant, *et al.* [6], Faloutsos, *et al.* [7], and Fowler [9], have found many network and computer statistics follow power-law or Pareto distributions. Specifically, numerous aspects of network traffic exhibits variations on Pareto distributions [9]. These range from low-level statistics such as the timing of packet requests, file sizes, *etc.*, to high-level aspects such as the popularity of web sites, the popularity of certain web pages in particular web sites, the number of incoming links, *etc.*

This paper extends variations of the classical birthday problem to better understand Pareto-like network traffic. This is accomplished by applying indicator random variables to counting collisions in binned network statistics. The birthday problem has been studied using indicator random variables, as an example see the second exposition of the birthday problem in Cormen, Leiserson, Rivest and Stein [3]. While Cormen *et al.* assume uniform random variables, we apply these indicator variables with a focus on Pareto-type distributions to gain a better understanding of network traffic statistics for traffic analysis.

1.1 Previous Research

Select Research on Traffic Analysis. There has been a good deal of work on traffic analysis. Here we very briefly review selected papers. We are not aware of any papers that take our approach to traffic analysis of using indicator random variables to understand Pareto distributions.

Raymond [18] surveys traffic analysis and related issues. Newman-Wolfe and Venkatraman [17] give a model for preventing traffic analysis. This model tries to make the traffic behave neutrally, thus disguising the actual traffic patterns. They base this on matrices describing neutral traffic patterns. Then they suggest traffic padding, re-routing, and traffic delays as countermeasures.

Guan, Li, Xuan, Bettati, and Zhao [14] use traffic padding and host-based re-routing to disguise network traffic. Further, they give heuristic methods that allow real-time constraints to be met while preventing traffic analysis. Guan, Fu, Xuan, Shenoy, Bettati, and Zhao [13] describe the NetCamo system which forestalls traffic analysis in real-time systems. Fu, Graham, Bettati, and Zhao [10] give an analytical framework for traffic analysis. The focus is on constant-interarrival time packets and variable interarrival time packets for countermeasures.

Fu, Graham, Bettati, Zhao, and Xuan [11] study traffic link padding with constant-interarrival time packets and variable interarrival time packets for countermeasures to

*Department of Computer Science, The University of Alabama, Box 870290, Tuscaloosa, AL 35487-0290. { pgb, iperevalova, cward } @cs.ua.edu

†School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario Canada K1S 5B6. michiel@scs.carleton.ca

‡Partially supported by a University of Alabama Research Advisory Committee (RAC) grant.

traffic analysis. They give extensive analytical and empirical analysis of these countermeasures and attacks against them. The attacks they examine are sample mean, sample variance and sample entropy. Finally, they give guidelines for system configurations to harden security.

The birthday problem. The birthday problem was originally proposed and solved by Richard von Mises [20]. Subsequently, a number of variations of the birthday problem and their solutions have appeared. For example, due to its applicability to attacking hash functions, the birthday problem is now an important part of the computer security literature [19].

Work on the birthday problem that is somewhat relevant to this paper is Flajolet, Gardy, and Thimonier [8]. Flajolet, *et al.* [8] give results on the expectation for getting j different letter k-collisions. Different from our approach, their results are expressed as truncated exponentials or gamma functions.

Challenges simulating heavy-tailed distributions. Crovella and Lipsky [4] examine challenges of simulating heavy-tailed distributions. These challenges are due to sampling large numbers of moderate (tail) values and fewer peak values. They point out that this is a particular challenge for Pareto-like distributions where $\alpha < 1.7$.

Gross, Shortle, Fischer, and Masi [12] discuss the challenges of simulating truncated Pareto distributions. The question arises as to where to truncate a Pareto distribution, and this has a large impact on properly simulating a Pareto distribution.

Our approach circumvents the issue of analyzing truncated Pareto distributions by focusing on moments and parameters of t-sized buckets of collision bins.

1.2 Structure of this Paper

Section 2 reviews useful facts about Pareto distributions. Section 3 discusses measurement of Pareto distributions by collisions of values in bins. Section 4 gives a brief review of the birthday problem. Subsection 4.1 gives examples of probabilities for birthday collisions in binned data from Pareto distributions, while subsection 4.2 gives ways to compute indicator moments. Section 4 applies indicator random variables to Pareto-based birthday problems. Section 5 concludes the paper and discusses future directions.

2 Pareto Distributions

Pareto or power-law distributions are loosely characterized by having heavy tails. Intuitively, this is a result of their density functions which are variations on geometric functions.

Let $H_{n,k}$ denote the n^{th} Harmonic number of the k^{th} order. This is defined as:

$$H_{n,k} = \sum_{i=1}^{n} \frac{1}{i^k}.$$

The Riemann Zeta function, denoted $\zeta(\cdot)$, is defined similarly:

$$\zeta(k) = \sum_{i=1}^{\infty} \frac{1}{i^k},$$

for any complex number k with real component larger than 1.

Although Johnson, *et al.* [15] give three versions of the Pareto distribution, here we examine only the (continuous) Pareto distributions of the first and second kinds. Given the parameters *location* $c > 0$ and *shape* $\alpha > 0$, the Pareto distribution of the 1-st Kind has the probability distribution function (PDF) $\frac{\alpha c^\alpha}{x^{\alpha+1}}$ for $x > c$, while the Pareto distribution of the 2-nd Kind (also known as the Lomax distribution) has the PDF $\frac{\alpha}{(x+1)^{\alpha+1}}$ for $x > 0$.

The Zeta distribution with parameter $\alpha > 0$ is a discrete distribution sometimes also called the discrete Pareto distribution or the Zipf-Estoup law, and has the PDF $\mathbb{P}[X = i] = c\, i^{-(\alpha+1)}$ for $i = 1, 2, ...$ and $c = \left(\sum_{i=1}^{\infty} i^{-(\alpha+1)}\right)^{-1} = (\zeta(\alpha+1))^{-1}$. We can also consider the case where $\alpha = 0$, though only over a finite range $[1, n]$. This is the Harmonic Zipf distribution and has the PDF $\mathbb{P}[X = i] = \frac{1}{iH_n}$ for $i = 1, 2, \ldots, n$, where H_n is the n^{th} harmonic number of the first order.

In general, for both types of Pareto distributions, the k^{th} central and raw moment are only defined for $\alpha > k$. Thus, it is worth noting that for values of the shape parameter α which are less than or equal to 2, the variance of both types of Pareto is infinite. As noted in [5], this means that the Central Limit Theorem does not hold with respect to the distribution for values of $\alpha \leq 2$, which can cause significant difficulties in analyzing Pareto-like behaviors with simulations.

The rest of this paper assumes suitably large α so the moments exist for the discussion at hand.

3 Traffic Analysis and the Birthday Problem

Many systems generate numerous overlapping data transmissions. These transmissions in effect wash-out or truncate the Pareto tails of other transmissions.

As an example, Figure 1 shows two overlapping Pareto distributions of the 2^{nd} kind, the first distribution has $\alpha = 1.5$ and the second has $\alpha = 2.5$. This was generated from simulated data. Such data is not unusual in network transmissions, see for example [1] for analysis of data from Internet sites serving the 1998 World Cup in which $\alpha = 1.37$.

	Pareto of the 1^{st} Kind Distribution	Pareto of the 2^{nd} Kind Distribution
pdf $f(x)$	$\frac{\alpha c^{\alpha}}{x^{\alpha+1}}$ $\alpha > 0, c > 0, x > c$	$\frac{\alpha}{(x+1)^{\alpha+1}}$ $\alpha > 0, x > 0$
$\mathbb{P}[X > x]$	$\left(\frac{x}{c}\right)^{-\alpha}$	$(1+x)^{-\alpha}$
$\mathbb{E}[X]$	$\frac{\alpha c}{\alpha-1}$ if $\alpha > 1$ ∞ otherwise	$\frac{1}{\alpha-1}$ if $\alpha > 1$ ∞ otherwise
$\mathbb{E}[X^2]$	$\frac{c^2\alpha}{\alpha-2}$ if $\alpha > 2$ ∞ otherwise	$\frac{2}{2-3\alpha+\alpha^2}$ if $\alpha > 2$ ∞ otherwise
	Harmonic Zipf	ζ
$\mathbb{P}[X = i]$	$\frac{1}{i\, H_n}$	$\frac{\zeta(1+\alpha)^{-1}}{i^{1+\alpha}}$ $\alpha > 0, x > 0$
$\mathbb{P}[X > x]$	$1 - \frac{H_x}{H_n}$	$1 - \frac{H_{x,\alpha+1}}{\zeta(\alpha+1)}$
$\mathbb{E}[X]$	$\frac{n}{H_n}$	$\frac{\zeta(\alpha)}{\zeta(1+\alpha)}$ $\alpha > 1$
$\mathbb{E}[X^2]$	$\frac{n(n+1)}{2H_n}$	$\frac{\zeta(\alpha-1)}{\zeta(1+\alpha)}$ $\alpha > 2$

Table 1. Main Distributions and their Characteristics with various parameters

A standard way of computing and understanding the parameters of Pareto distributions is based on binning data [16]. At a rudimentary level, binned data may be examined as histograms to determine if a data-set forms a Pareto distribution. Variable sized bins may also lead to more sophisticated analyzes. Histogram-based techniques can be used to derive sample parameters for Pareto distributions. These sample parameters are important in both performing traffic analysis as well as building counter-measures by spoofing Pareto distributions.

Due to the truncation of Pareto distribution's tails and the overlapping nature of much network traffic (see Figure 1), this paper proposes computing statistics on t-sized collisions. Data points that are put in the same buckets of Pareto distributions give identifying information to track similar network statistics. Also, if we are spoofing network traffic, then we should be careful to make sure these t-size collisions follow appropriate network statistics.

4 The Birthday Problem using Indicator Variables

Let $Y_1, Y_2, \ldots, Y_k$ be a sequence of independent and identically distributed random variables, whose range is $[n]$. We imagine this to model the following situation: There are

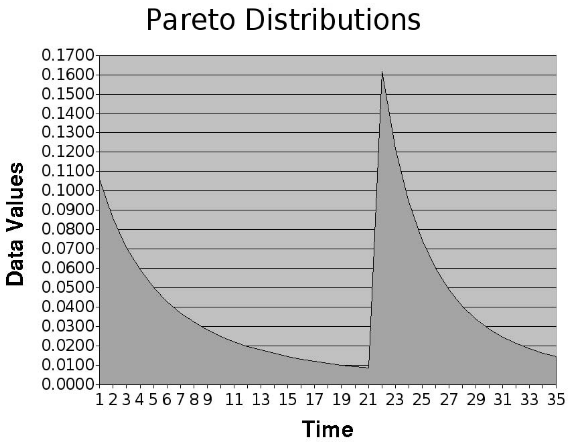

Figure 1. Two Pareto Transmissions of the 2nd kind.

n days in one year, and there are k people. The random variable Y_i, for $i \in [k]$, represents the birthday of the i-th person.

Definition 1 For any non-empty subset I of $[k]$, we define X_I to be the indicator random variable representing the event that all people in I have the same birthday. Thus, if $I = \{i_1, i_2, \ldots, i_t\}$, then

$$X_I = \begin{cases} 1 & \text{if } i_1, \cdots, i_t \text{ all have the same birthday} \\ 0 & \text{otherwise.} \end{cases}$$

For each $i \in [k]$ and $j \in [n]$, we define $p_j = \mathbb{P}[Y_i = j]$. Then, for any subset I of $[k]$ with $|I| = t$,

$$\begin{aligned} \mathbb{E}[X_I] &= \mathbb{P}[X_I = 1] \\ &= \sum_{j=1}^{n} \mathbb{P}\left[\bigwedge_{i \in I} (Y_i = j)\right] \\ &= \sum_{j=1}^{n} \prod_{i \in I} \mathbb{P}[Y_i = j] \\ &= \sum_{j=1}^{n} \prod_{i \in I} p_j \\ &= \sum_{j=1}^{n} p_j^t. \end{aligned}$$

If we define the function Q by

$$Q(x) = \sum_{j=1}^{n} p_j^{x+1},$$

then we have shown that, for any t with $1 \leq t \leq k$, and for any subset I of $[k]$ of size t,

$$\mathbb{E}[X_I] = Q(t-1).$$

4.1 Examples of Pareto Events

Suppose $t = 2$ and consider the uniform random distribution, therefore $p_i = \frac{1}{n}$ for all $i : n \geq i \geq 1$, giving,

$$\begin{aligned} Q(1) &= \sum_{i=1}^{n} \frac{1}{n^2} \\ &= \frac{1}{n}. \end{aligned}$$

That is, for any two fixed people i_1, i_2, with random and uniform birthdays, the probability these birthdays are the same is $\frac{1}{n}$.

Likewise, consider $t = 2$ people whose birthdays are distributed according to the harmonic Zipf distribution, $p_i = \frac{1}{i\,H_n}$, then

$$\begin{aligned} Q(1) &= \sum_{i=1}^{n} \frac{1}{(i\,H_n)^2} \\ &= \frac{1}{H_n^2} \sum_{i=1}^{n} \frac{1}{i^2}. \end{aligned}$$

Since $\lim_{n\to\infty}(H_n - \ln n) = \gamma$, where γ is Euler's constant, and $\sum_{i=1}^{\infty} 1/i^2 = \pi^2/6$, it follows that the probability that two people have the same birthday is $Q(1) \in \Theta(\frac{1}{\ln^2(n)})$, a rather dramatic increase compared to the uniform case.

4.2 Computing indicator Expectations

Here the focus is on the expectation of indicator variables. These variables allow us to gain a better understanding of how to distinguish different Pareto variables.

Definition 2 For any fixed $t \in [k]$,

$$\begin{aligned} X &= |\{(i_1, \cdots, i_t) : \{i_1 < i_2 < \cdots < i_t\} \subseteq [k] \\ &\quad \text{and all } i_1, \cdots, i_t \text{ have the same birthday } \}|. \end{aligned}$$

Definition 2 immediately gives

$$X = \sum_{I \subseteq [k], |I|=t} X_I.$$

The random variable X represents the number of t-size sets of k people with the same birthday. Therefore, in this case, $\mathbb{E}[X]$ is the expected number of t-size groups of k randomly and uniformly chosen people having the same birthday.

If we write $\mathbb{E}[X_I] = Q(t-1)$ as before, then we have

$$\begin{aligned} \mathbb{E}[X] &= \sum_{I \subset [k], |I|=t} \mathbb{E}[X_I] \\ &= \binom{k}{t} Q(t-1), \end{aligned}$$

by the linearity of expectation.

Lemma 1 If I and J are sets of indices whose intersection has size one, then X_I and X_J are independent.

Proof: Let $t \in [k]$ and $t > 1$ and $s \in [t]$, let us first show that

$$\begin{aligned} \mathbb{P}[X_{i_1,\cdots,i_t} = 1] &= \mathbb{P}[X_{i_1,\cdots,i_s} = 1]\, \mathbb{P}[X_{i_s,\cdots,i_t} = 1] \\ &\quad \text{for any set } \{\, i_1, \cdots, i_t \,\} \subseteq [k]. \end{aligned}$$

Let Y_j be the birthday of the j^{th} person, for $1 \leq j \leq k$.

$$\begin{aligned} &\mathbb{P}[X_{i_1,\cdots,i_t} = 1] \\ &\quad = \mathbb{P}[Y_{i_1} = y \wedge \cdots \wedge Y_{i_t} = y | Y_{i_s} = y] \\ &\qquad \text{by definition, for any } s : k \geq t > s \geq 1 \\ &\quad = \mathbb{P}[Y_{i_1} = y \wedge \cdots \wedge Y_{i_{s-1}} = y \\ &\qquad \wedge Y_{i_{s+1}} = y \wedge \cdots \wedge Y_{i_t} = y] \\ &\qquad \text{by independence of all } Y_i \\ &\quad = \mathbb{P}[Y_{i_1} = y] \cdots \mathbb{P}[Y_{i_{s-1}} = y] \\ &\qquad \cdot \mathbb{P}[Y_{i_{s+1}} = y] \cdots \mathbb{P}[Y_{i_t} = y] \\ &\quad = \mathbb{P}[Y_{i_1} = y \wedge \cdots \wedge Y_{i_{s-1}} = y] \\ &\qquad \cdot \mathbb{P}[Y_{i_{s+1}} = y \wedge \cdots \wedge Y_{i_t} = y] \\ &\quad = \mathbb{P}[Y_{i_1} = y \wedge \cdots \wedge Y_{i_s} = y | Y_{i_s} = y] \\ &\qquad \cdot \mathbb{P}[Y_{i_s} = y \wedge \cdots \wedge Y_{i_t} = y | Y_{i_s} = y] \\ &\quad = \mathbb{P}[X_{i_1,\cdots,i_s} = 1]\, \mathbb{P}[X_{i_s,\cdots,i_t} = 1] \end{aligned}$$

Now, assume without loss that $I = \{i_1, \cdots, i_s\}$ and $J = \{i_s, \cdots, i_t\}$. Hence that $I \cap J = \{i_s\}$. It must follow that

$$\mathbb{P}[X_{i_1,\ldots,i_s} = 1 \wedge X_{i_s,\ldots,i_t} = 1] = \mathbb{P}[X_{i_1,\ldots,i_t} = 1]$$

since the birthday of person i_s fixes both sets. Hence,

$$\begin{aligned} &\mathbb{P}[X_{i_1,\ldots,i_s} = 1 \wedge X_{i_s,\ldots,i_t} = 1] \\ &\quad = \mathbb{P}[X_{i_1,\ldots,i_t} = 1] \\ &\quad = \mathbb{P}[X_{i_1,\ldots,i_s} = 1] \mathbb{P}[X_{i_s,\ldots,i_t} = 1] \end{aligned}$$

completing the proof. ∎

Lemma 2 Let I and J be subsets of $[k]$, such that $I \cap J \neq \emptyset$. The random variables X_I and X_J have joint probability

$$\mathbb{P}[X_I = 1 \wedge X_J = 1] \quad = \quad \mathbb{P}[X_{I \cup J} = 1].$$

Proof: Since $I \cap J \neq \emptyset$, we have

$$X_I = 1 \wedge X_J = 1 \text{ if and only if } Y_i \quad = \quad Y_j \quad \forall i, j \in I \cup J.$$

It follows that

$$\begin{aligned}\mathbb{P}[X_I = 1 \wedge X_J = 1] &= \sum_{y \in [n]} \mathbb{P}\left[\bigwedge_{i \in I \cup J} (Y_i = y)\right] \\ &= \mathbb{P}[X_{I \cup J} = 1]\end{aligned}$$

∎

Lemma 3 Let $t \in [k]$ and $t \geq 2$ so $\{i_1, \cdots, i_t\} \subseteq [k]$ and $\{j_1, \cdots, j_t\} \subseteq [k]$. The random variables $X_{i_1,\cdots,i_t}$ and $X_{j_1,\cdots,j_t}$ are independent iff $1 \geq |\{i_1, \cdots, i_t\} \cap \{j_1, \cdots, j_t\}| \geq 0$.

Proof: Let $I = \{i_1, \cdots, i_t\}$ and $J = \{j_1, \cdots, j_t\}$.

$\Leftarrow$ If $|I \cap J| = 0$, then $X_{i_1,\cdots,i_t}$ and $X_{j_1,\cdots,j_t}$ are obviously independent, since birthdays are independent. On the other hand, if $|I \cap J| = 1$, then apply Lemma 1.

$\Rightarrow$ We will prove the contrapositive. We take $|I \cap J| > 1$, and for the sake of a contradiction, suppose the variables X_I and X_J are independent. We consider two cases.

First, assume that either $I \cap J = I$ or $I \cap J = J$, then since $t \geq 2$ gives

$$\begin{aligned}\mathbb{P}[X_I = 1 \wedge X_J = 1] &= \mathbb{P}[X_J = 1] \\ &\neq \mathbb{P}[X_I = 1]\,\mathbb{P}[X_J = 1]\end{aligned}$$

Thus, X_I and X_J are not independent, contradicting this sub case.

On the other hand, suppose $I \cap J \neq I$ so since $|I| = |J|$ it must be that $|I - (I \cap J)| > 1$ (the case where $|I| = |J|$ and $I \cap J \neq J$ and $|J - (I \cap J)| > 1$ is symmetrical). Thus considering Lemma 2 gives

$$\begin{aligned}&\mathbb{P}[X_I = 1 \wedge X_J = 1] \\ &\quad = \mathbb{P}[X_{I \cup J} = 1] \\ &\quad = Q(|I| + |J| - |I \cap J| - 1).\end{aligned}$$

Moreover, it must also be that

$$\mathbb{P}[X_I = 1]\,\mathbb{P}[X_J = 1] \quad = \quad Q(|I| - 1)Q(|J| - 1)$$

but,

$$Q(|I| - 1)Q(|J| - 1) \quad \neq \quad Q(|I| + |J| - |I \cap J| - 1)$$

for $|I \cap J| > 1$. This completes the proof.

∎

Theorem 1 Let $X = X_{i_1,i_2}$ be as per Definition 2 with $t = 2$. Then $\mathbb{E}[X] = \binom{k}{2}Q(1)$ and $\textbf{Var}[X] = \mathbb{E}[X](1 - Q(1))$.

Proof: First note that since $X_{i_1,\cdots,i_t}$ is an indicator variable, then $\mathbb{E}[X_{i_1,\cdots,i_t}] = \mathbb{E}[X^2_{i_1,\cdots,i_t}]$.

Take the set of all pairs $T = [k] \times [k]$, and the subset $U \subset T$ so that $(u_1, u_2) \in U$ iff $u_1 \neq u_2$, then

$$\begin{aligned}&\textbf{Var}[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \sum_{u \in U} \mathbb{E}[X_u^2] + 2 \sum_{\substack{v,w \in U \\ v \neq w}} \mathbb{E}[X_v X_w] - \binom{k}{2}^2 Q^2(1) \\ &= \sum_{u \in U} \mathbb{E}[X_u^2] + 2 \sum_{\substack{v,w \in U \\ v \neq w}} \mathbb{E}[X_v]\mathbb{E}[X_w] - \binom{k}{2}^2 Q^2(1) \\ &\quad \text{(by Lemma 3)} \\ &= \binom{k}{2}Q(1) + 2\binom{\binom{k}{2}}{2}Q^2(1) - \binom{k}{2}^2 Q^2(1) \\ &= \binom{k}{2}Q(1) + \binom{k}{2}\left(\binom{k}{2} - 1\right)Q^2(1) \\ &\quad -\binom{k}{2}^2 Q^2(1) \\ &= \mathbb{E}[X]\left(1 - Q(1)\right),\end{aligned}$$

completing the proof. ∎

In the case of harmonic Zipf birthday collisions, for $t = 2$ variables $Q(1) = \Theta(\frac{1}{\ln^2(n)})$ meaning $\mathbb{E}[X] = \Theta\left(\frac{k(k-1)}{2\ln^2(n)}\right)$. Theorem 1 gives $\textbf{Var}[X] = \Theta\left(\frac{k(k-1)}{2\ln^2(n)}\left(1 - \frac{1}{\ln^2(n)}\right)\right)$. Now, if X' is uniformly distributed, then $\mathbb{E}[X'] = \Theta\left(\frac{k(k-1)}{2n}\right)$ and $\textbf{Var}[X'] = \Theta\left(\frac{k(k-1)}{2n}\left(1 - \frac{1}{n}\right)\right)$. These different moments should be easily detectable.

The asymptotic notation is used here to deal with the difference between the harmonic numbers and their logarithmic representation.

Consider two Pareto distributions of the second kind with parameters $\alpha = 1.5$ and $\alpha' = 2.5$. If we consider truncating each distribution at $x = 3.6$ and using bins of size 0.1, we have that the probabilities of two birthdays colliding are approximately 0.04 and 0.06, respectively. Thus, for $t = 2$ we have $\mathbb{E}[X] \approx 0.04\binom{k}{2}$ and $\mathbb{E}[X'] \approx 0.06\binom{k}{2}$. Further, $V[X] \approx (0.96 * 0.04)\binom{k}{2}$ and $V[X'] \approx (0.94 * 0.06)\binom{k}{2}$. This means the variances of the $t = 2$ sized bin collisions for each of these different distributions are different by about 50%.

Lemma 4 Assume $t = 2$ and $k \geq t$ and let $X \equiv X_{i_1,i_2}$ then,

$$\textbf{Skew}[X] = \mathbb{E}[X](-(2\mathbb{E}[X]-1) \cdot (\mathbb{E}[X]+1) + 2Q(1)).$$

Proof: Take the set of all pairs $T = [k] \times [k]$, and the subset $U \subset T$ so that $(u_1, u_2) \in U$ iff $u_1 \neq u_2$.

Now, consider the definition of skew:

$$\begin{aligned}
&\textbf{Skew}[X] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^3] \\
&= \mathbb{E}[X^3] - 3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^2 \\
&= \sum_{u\in U} \mathbb{E}[X_u^3] + 6 \sum_{v,w,z\in U} \mathbb{E}[X_v]\mathbb{E}[X_w]\mathbb{E}[X_z] \\
&\quad -3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X] + \binom{k}{2}\left(\binom{k}{2} - 1\right)\left(\binom{k}{2} - 2\right) Q^3(1) \\
&\quad -3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X] + \mathbb{E}[X]\left(\binom{k}{2}^2 - 3\binom{k}{2} + 2\right) Q^2(1) \\
&\quad -3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X] + \mathbb{E}[X]\left(\mathbb{E}[X]^2 - 3\mathbb{E}[X]Q(1) + 2Q^2(1)\right) \\
&\quad -3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X] + \mathbb{E}[X]^3 - 3\mathbb{E}[X]^2 Q(1) + 2\mathbb{E}[X]Q^2(1) \\
&\quad -3\mathbb{E}[X]\mathbb{E}[X^2] + 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X] + \mathbb{E}[X]^3 - 3\mathbb{E}[X]^2 Q(1) + 2\mathbb{E}[X]Q^2(1) \\
&\quad -3\mathbb{E}[X]\left(\mathbb{E}[X] - \mathbb{E}[X]Q(1) + \mathbb{E}[X]^2\right) + 2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X] - 2\mathbb{E}[X]^3 - 3\mathbb{E}[X]^2 + 2\mathbb{E}[X]Q^2(1) \\
&\quad +2\mathbb{E}[X]^2 \\
&= \mathbb{E}[X]\left(1 - 2\mathbb{E}[X]^2 - \mathbb{E}[X] + 2Q^2(1)\right) \\
&= \mathbb{E}[X]\left(-(2\mathbb{E}[X]-1)(\mathbb{E}[X]+1) + 2Q(1)\right),
\end{aligned}$$

completing the proof. ∎

Lemma 5 Assume $t = 2$ and $k \geq t$ and let $X \equiv X_{i_1,i_2}$ then

$$\begin{aligned}
\textbf{Kurtosis}[X] &= \mathbb{E}[X]((3\mathbb{E}[X]-1)(\mathbb{E}[X]-1) \\
&\quad -4\mathbb{E}[X]Q^2(1) - 6Q^3(1)).
\end{aligned}$$

Proof: Take the set of all pairs $T = [k] \times [k]$, and the subset $U \subset T$ so that $(u_1, u_2) \in U$ iff $u_1 \neq u_2$.

Now, consider the definition of kurtosis:

$$\begin{aligned}
&\textbf{Kurtosis}[X] \\
&= \mathbb{E}[(X - \mathbb{E}[X])^4] \\
&= \mathbb{E}[X^4] - 4\mathbb{E}[X]\mathbb{E}[X^3] \\
&\quad +3\mathbb{E}[X]^2(2\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\
&= \sum_{u\in U} \mathbb{E}[X_u^4] + 4! \sum_{v,w,s,t\in U} \mathbb{E}[X_v]\mathbb{E}[X_w]\mathbb{E}[X_s]\mathbb{E}[X_t] \\
&\quad -4\mathbb{E}[X]\mathbb{E}[X^3] + 3\mathbb{E}[X]^2(2\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\
&= \binom{k}{2}\left(\binom{k}{2} - 1\right)\left(\binom{k}{2} - 2\right)\left(\binom{k}{2} - 3\right) Q^4(1) \\
&\quad -4\mathbb{E}[X]\mathbb{E}[X^3] + 3\mathbb{E}[X]^2(2\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\
&\quad +\mathbb{E}[X] \\
&= \mathbb{E}[X]Q^3(1)\left(\binom{k}{2}^3 - 6\binom{k}{2}^2 + 11\binom{k}{2} - 6\right) \\
&\quad -4\mathbb{E}[X]\mathbb{E}[X^3] + 3\mathbb{E}[X]^2(2\mathbb{E}[X^2] - \mathbb{E}[X]^2) \\
&\quad +\mathbb{E}[X] \\
&= \mathbb{E}[X](\mathbb{E}[X]^3 - 6\mathbb{E}[X]^2 Q(1) \\
&\quad +11\mathbb{E}[X]Q^2(1) - 6Q^3(1)) \\
&\quad -4\mathbb{E}[X](\mathbb{E}[X] - 3\mathbb{E}[X]^2 Q(1) \\
&\quad +2\mathbb{E}[X]Q^2(1) + \mathbb{E}[X]^3) \\
&\quad +3\mathbb{E}[X]^2((\mathbb{E}[X] - \mathbb{E}[X]Q(1)) \\
&\quad +\mathbb{E}[X] - \mathbb{E}[X]Q(1) + \mathbb{E}[X]^2) \\
&\quad +\mathbb{E}[X] \\
&= -4\mathbb{E}[X]^2 Q^2(1) - 6\mathbb{E}[X]Q^3(1) \\
&\quad +3\mathbb{E}[X]^3 - 4\mathbb{E}[X]^2 + \mathbb{E}[X] \\
&= \mathbb{E}[X](3\mathbb{E}[X]^2 - 4\mathbb{E}[X] + 1 \\
&\quad -4\mathbb{E}[X]Q^2(1) - 6Q^3(1)) \\
&= \mathbb{E}[X]((3\mathbb{E}[X]-1)(\mathbb{E}[X]-1) \\
&\quad -4\mathbb{E}[X]Q^2(1) - 6Q^3(1))
\end{aligned}$$

completing the proof. ∎

Lemmas 4 and 5 give moderately complex expressions for Pareto type distributions. However, these expressions are easy to program. For instance, they are exclusively dependent on linear functions of the expectation of the birthday collisions in addition to products of powers of probabilities.

Theorem 2 For the more general case, where $t \in [k]$ and for $\{i_1, \cdots, i_t\} \subseteq [k]$ and let $X \equiv X_{i_1,\ldots,i_t}$, then

$$\begin{aligned}\mathbf{Var}[X] &= \mathbb{E}[X]\,(1 - \mathbb{E}[X]) \\ &\quad + \sum_{i=2}^{t-1} \binom{k}{t-i} \binom{k-i-t}{t-i} Q(2t-i-1).\end{aligned}$$

Proof: First, let

$$T = \underbrace{[k] \times [k] \times \cdots \times [k]}_{t}$$

and take $U \subseteq T$ such that for all $(u_1, \cdots, u_t) \in U$ it must be that $u_i \neq u_j$ for $i \neq j$ and all $i, j : t \geq i, j \geq 1$. Here, $\mathbb{E}[X] = \binom{k}{t} Q(t-1)$.

$$\begin{aligned}&\mathbf{Var}[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \sum_{u \in U} \mathbb{E}[X_u^2] + 2 \sum_{\substack{v,w \in U \\ v \neq w}} \mathbb{E}[X_v X_w] \\ &\quad - \binom{k}{t}^2 Q^2(t-1) \\ &= \sum_{i=2}^{t-1} \left(\frac{k!}{(k-i)!}\right) \binom{k-i}{t-i} \binom{k-i-t}{t-i} Q(2t-i-1) \\ &\quad + \sum_{u \in U} \mathbb{E}[X_u^2] - \binom{k}{t}^2 Q^2(t-1) \\ &= \sum_{u \in U} \mathbb{E}[X_u^2] - \binom{k}{t}^2 Q^2(t-1) \\ &\quad + \sum_{i=2}^{t-1} \binom{k}{t-i} \binom{k-i-t}{t-i} Q(2t-i-1) \\ &= \binom{k}{t} Q(t-1) - \binom{k}{t}^2 Q^2(t-1) \\ &\quad + \sum_{i=2}^{t-1} \binom{k}{t-i} \binom{k-i-t}{t-i} Q(2t-i-1) \\ &= \mathbb{E}[X]\,(1 - \mathbb{E}[X]) \\ &\quad + \sum_{i=2}^{t-1} \binom{k}{t-i} \binom{k-i-t}{t-i} Q(2t-i-1)\end{aligned}$$

completing the proof. ∎

5 Conclusion and Future Directions

Indicator random variables are useful tools for giving insight into Pareto random variables. When applied to the birthday problem, indicator random variables may provide useful and sometimes easy to compute parameters. Understanding and working with Pareto distributions is important for traffic analysis since network statistics often exhibit Pareto distributions.

Left for future work is the question of whether it is possible to get a closed form for the variance expression given in Theorem 2 when t is a large integer. Furthermore, generalizations of Theorem 1, Lemmas 4 and 5 for $t > 2$ would also be of interest.

References

[1] M. Arlitt and T. Jin. Workload characterization of the 1998 world cup web site. *IEEE Network*, 14(3):30–37, May/June 2000.

[2] A.-L. Barabási and R. Albert. *Science*, 286:509, 1999.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, 2nd Edition*. MIT Press, 2001.

[4] M. E. Crovella and L. Lipsky. Long-lasting transient conditions in simulations with heavy-tailed workloads. In *Proceedings of the 1997 Winter Simulation Conference*, pages 407–415, 1997.

[5] M. E. Crovella and L. Lipsky. Long-lasting transient conditions in simulations with heavy-tailed workloads. In *Proceedings of the 1997 Winter Simulation Conference*, pages 407–415, 1997.

[6] A. Fabrikant, E. Koutsoupias, and C. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, volume 2380, pages 110–122, 2002.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of ACM SIGCOMM '99*, pages 251–262, Aug 1999.

[8] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39:207–229, 1992.

[9] T. B. Fowler. A short tutorial on fractals and internet traffic. *The Telecommunications Review*, 10:1–14, 1999.

[10] X. Fu, B. Graham, R. Bettati, and W. Zhao. On effectiveness of link padding for statistical traffic analysis attacks. In *23rd International Conference on Distributed Computing Systems (ICDCS)*, pages 340–349, 2003.

[11] X. Fu, B. Graham, R. Bettati, W. Zhao, and D. Xuan. Analytical and empirical analysis of countermeasures to traffic analysis attacks. In *In the Proceedings of the* 32^{nd} *International Conference on Parallel Processing*, pages 483–492, 2003.

[12] D. Gross, J. F. Shortle, M. J. Fischer, and D. M. B. Masi. Difficulties in simulating queues with pareto service. In *Proceedings of the 2002 Winter Simulation Conference*, pages 407–415, 2002.

[13] Y. Guan, X. Fu, D. Xuan, P. U. Shenoy, R. Bettati, and W. Zhao. Netcamo: Camouflaging network traffic for qos-guaranteed mission critical applications. *IEEE Transactions on System, Man, and Cybernetics*, 31(4), July 2001.

[14] Y. Guan, C. Li, D. Xuan, R. Bettati, and W. Zhao. Preventing traffic analysis for real-time communication networks. In *Proceedings of IEEE Military Communications Conference (MILCOM)*, 1999.

[15] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions, Second Edition, Wiley Series in Probability and Mathematical Statistics*. 1992.

[16] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[17] R. Newman-Wolfe and B. R. Venkatraman. High level prevention of traffic analysis. *Seventh Annual Computer Security and Applications Conference*, pages 102–109, 1991.

[18] J.-F. Raymond. Traffic analysis: Protocols, attacks, design issues, and open problems. *Anonymity 2000*, pages 10–30, 2001.

[19] D. R. Stinson. *Cryptography, Theory and Practice, Third Edition*. Chapman & Hall/CRC Press, 2006.

[20] R. von Mises. Uber aufteilungs–und besetzungs-wahrscheinlichkeiten. *Revue de la Faculté des Sciences de l'Université d'Istanbul, N. S*, 4:145–163, 1939.