PRINCIPLES OF COMPLEX SYSTEMS

HW02 WRITE-UP

David W. Landay
University of Vermont
Graduate Student, Comp. Systems and Data Sci.

0.1 Problem 1

Consider a random variable X with a probability distribution given by:

$$P(x) = cx^{-\gamma}$$

where c is a normalization constant, and $0 < a \leq x \leq b$. (a and b are the lower and upper cuttoffs respectively.) Assume that $\gamma > 1$.

Given the semi-closed interval $(a,b]$ is continuous and there is no upper threshold for b, we know that X is a continuous random variable. Hence, $P(x)$ describes a probability distribution over a set of (continuous) real numbers, and is thus a probability density function. The integral of $P(x)$, then, defines the cumulative distribution of X between two points in the sample space. This is equivalent to calculating the zeroeth moment of X, or the expectation that a point within the given constraints for x exists in the sample space. We can use the fact that the area under the curve of the probability density function (the zeroeth moment) must equal 1 between a and b (the minimum and maximum values in X) to find the value of the normalization constant c:

$$\int_a^b P(x)dx = \int_a^b cx^{-\gamma}dx = 1$$

$$\implies 1 = c\frac{1}{1-\gamma}x^{1-\gamma} \Big|_a^b$$

$$= \frac{c(b-a)(b-a)^{-\gamma}}{1-\gamma}$$

$$= c\Big(\frac{b}{b^\gamma - b^\gamma\gamma} - \frac{a}{a^\gamma - a^\gamma\gamma}\Big)$$

$$\implies 1 = c\frac{ab^\gamma - ba^\gamma}{b^\gamma a^\gamma(\gamma-1)}$$

$$\implies c = \frac{b^\gamma a^\gamma(\gamma-1)}{ab^\gamma - ba^\gamma} \therefore \tag{1.a}$$

The area under the curve of $P(x)$ helps to show why the parameter γ must be greater than 1; namely, because we would have division by zero. Mathematically, if we take the integral from a to b where $\gamma = 1$,

If $\gamma = 1$, then

$$\int_a^b cx^{-1}dx$$

$$= c\int_a^b \frac{1}{x}dx$$

$$= c\ln x \,|_a^b$$

$$= c\big(\ln b - \ln a\big)$$

$$\approx \ln b \text{ given } b \to \infty$$

$$\lim_{b\to\infty} \ln b = \infty$$

and as b became very large, the integral would diverge; i.e. it would no longer be the case that $\int_a^b P(x)dx = 1$ for all values of b. Likewise, if $\gamma < 1$, then γ approaches 0; indicating that the integral will approach the value x and not 1, and diverge faster.

$$\frac{c(b-a)(b-a)^{-\gamma}}{1-\gamma}$$

$$\approx b^{1-\gamma} \text{ since } \; b \to \infty$$

$$\lim_{b\to\infty} \frac{b}{b^{\gamma}} = \infty, \text{ for } \; \gamma < 1$$

(1.b)

0.2 Problem 2

Compute the n_{th} moment of X.

The moment of a distribution is a measure that helps quantify, or characterize, the distribution itself centered about a point in the sample space. Typically, the point chosen to take the moment about is the mean. In general, the n_{th} moment of a probability distribution is defined by

$$\mu_n = \int_{-\infty}^{\infty} (x-c)^n f(x)\,dx$$

where c is a point in the sample space; typically set equal to he observed mean or 0. In our case, $f(x) = P(x)$ is a probability distribution, so the expected n_{th} moment of X centered about 0 is

$$\mathbf{E}[X^n] = \int_a^b x^n P(x)\,dx$$

with bounds a and b, as defined in part 1. We can now solve this integral to find the n_{th} moment of X:

$$\mathbf{E}[X^n] = \int_a^b x^n c x^{-\gamma} dx$$

$$= c \int_a^b x^n x^{-\gamma} dx$$

$$= c \int_a^b x^{n-\gamma} dx$$

$$= c\Big(\frac{1}{1+n-\gamma} x^{1+n-\gamma}\Big)\Big|_a^b$$

$$= c\Big(\frac{b^{1+n-\gamma} - a^{1+n-\gamma}}{1+n-\gamma}\Big)$$

$$= -c\Big(\frac{a^{1+n-\gamma} - b^{1+n-\gamma}}{\gamma-n-1}\Big)$$

$$= \frac{b^\gamma a^\gamma(\gamma-1)}{ba^\gamma - ab^\gamma}\Big(\frac{a^{1+n-\gamma} - b^{1+n-\gamma}}{\gamma-n-1}\Big) \therefore$$

Calculating the n_{th} moment automatically gives the constraints on γ under the assumption that $b \to \infty$. We calculated that the zeroeth moment gives

$$c\Big(\frac{a^{1+n-\gamma} - b^{1+n-\gamma}}{\gamma - n - 1}\Big), \text{ for } n = 0$$

$$= c\Big(\frac{a^{1+0-\gamma} - b^{1+0-\gamma}}{\gamma - 0 - 1}\Big)$$

$$\implies \gamma - 1 > 0$$

$$\implies \gamma > 1$$

and in general, for any n,

$$\gamma - n - 1 > 0$$

$$\implies \gamma - n > 1$$

$$\implies \gamma > n + 1$$

So, taking the first moment about the distribution (the mean) and second moment about the distribution (the variance) respectively, we get:

$$\gamma > 2$$

and,

$$\gamma > 3$$

for mean and finite variance (meaning that the tail of the distribution must show convergence). This will be shown in the calculation for σ in question 5.

0.3 Problem 3

In the limit $b \to \infty$, how does the n_{th} moment behave as a function of γ?

To try and highlight the behavior of the the n_{th} moment as the upper threshold for x, b, approaches infinity, we will examine changes in values of γ. As hinted above, we will want to explain why the n_{th} moment will diverge if $\gamma \leq 1+n$.

If $\gamma = 1+n$, then we have division by zero. If $\gamma < n+1$ then:

$$c\Big(\frac{a^{1+n-\gamma} - b^{1+n-\gamma}}{\gamma - n - 1}\Big) \approx b^{1+n-\gamma}, \text{ given } b \to \infty$$

$$\implies \lim_{b\to\infty} b^{1+n-\gamma} = \infty$$

Since $\gamma < 1+n$, then b^{1+n} dominates $b^{-\gamma}$ as $b \to \infty$ and the n_{th} moment will diverge. However, if $\gamma > 1+n$, then:

$$\lim_{b\to\infty} b^{1+n-\gamma} = 0$$

This time, since $\gamma > 1+n$, we get $b^{1+n-\gamma} = 1 \,/\, b^{\beta}$, where $\beta = 1+n-\gamma$. Since $b \to \infty$, $1 \,/\, b^{\beta}$ will converge to 0.

0.4 Problem 4

For finite cutoffs a and b with $a \ll b$, which cutoff dominates the expression for the n_{th} moment as a function of γ and n?

We know via the constraints that for finite cutoffs, $a \leq b$. We have also just shown for large b that when $\gamma > 1+n$, then b converges to 0. So, when a and b are finite,

$$c\Big(\frac{a^{1+n-\gamma} - b^{1+n-\gamma}}{\gamma - n - 1}\Big)$$

is dominated by a for $\gamma > +1+n$ because as n grows, γ grows faster to comply with the inequality. So, $a \leq b \implies$ the b terms will shrink (tend to 0 or small fractions) and be

subtracted from a terms which increase.

When $\gamma < 1 + n$, the larger b terms dominate as a result of a large term in the numerator. a terms will shrink and converge to small values.

0.5 Problem 5

The standard deviation, σ, is equal to $\sqrt{\mathbf{Var}[X]}$. The variance of distribution can be determined by calculating the difference between the expectation of X^2 and the expectation of X squared: $\mathbf{Var} = \mathbf{E}[X^2] - \mathbf{E}[X]^2$. This is the same as calculating the difference between the second moment about X and the first moment about X squared:

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \sigma^2$$

$$= c\Big(\frac{a^{3-\gamma} - b^{3-\gamma}}{\gamma - 3}\Big) - c^2\Big(\frac{a^{2-\gamma} - b^{2-\gamma}}{\gamma - 2}\Big)^2$$

$$= c\Big(\frac{a^{3-\gamma} - b^{3-\gamma}}{\gamma - 3}\Big) - c^2\Big(\frac{(a^{2-\gamma} - b^{2-\gamma})^2}{(\gamma - 2)^2}\Big)$$

$$= c\Big(\frac{(a^{3-\gamma} - b^{3-\gamma})(\gamma - 2)^2}{(\gamma - 3)(\gamma - 2)^2}\Big) - c^2\Big(\frac{(\gamma - 3)(a^{2-\gamma} - b^{2-\gamma})^2}{(\gamma - 3)(\gamma - 2)^2}\Big)$$

For finite a and b, $\sigma = \sqrt{\dfrac{c(a^{3-\gamma} - b^{3-\gamma})(\gamma - 2)^2 - c^2(\gamma - 3)(a^{2-\gamma} - b^{2-\gamma})^2}{(\gamma - 3)(\gamma - 2)^2}}$ $\therefore$

Hence, we have shown that for finite a and b, $\gamma > 2$ for a finite mean to exist, and $\gamma > 3$ for finite variance. In the limiting form of σ , where $b \to \infty$, we note that for $\mathbf{E}[x^n]$, $\gamma > 3$ and for $\mathbf{E}[x]$, $\gamma > 2$. Hence, the "b" terms in the equation will converge to 0 and can be neglected. We are left with:

$$\sigma^2 = \frac{c(a^{3-\gamma})(\gamma - 2)^2 - c^2(\gamma - 3)(a^{2-\gamma})^2}{(\gamma - 3)(\gamma - 2)^2}$$

Turning to the normalization constant c, we can also eliminate "b" terms:

$$c = \frac{(\gamma-1)}{a^{1-\gamma}},\ c^2 = \frac{(\gamma-1)^2}{a^{2-2\gamma}}$$

We are left with:

$$\frac{a^2(\gamma-1)(\gamma-2)^2 - (\gamma-1)^2(\gamma-3)a^2}{(\gamma-3)(\gamma-2)^2}$$

$$= \frac{a^2(\gamma-1)\left((\gamma-2)^2 - (\gamma-1)(\gamma-3)\right)}{(\gamma-3)(\gamma-2)^2}$$

$$= \frac{a^2(\gamma-1)\left(\gamma^2-4\gamma+4-\gamma^2+4\gamma-3\right)}{(\gamma-3)(\gamma-2)^2}$$

$$\sigma^2 = \frac{\gamma-1}{(\gamma-3)(\gamma-2)^2}a^2$$

$$\sigma = \sqrt{\frac{(\gamma-1)a^2}{(\gamma-3)(\gamma-2)^2}}\ \therefore \tag{5.a}$$

One may observe that for any $\gamma = 1$, there will be zero deviation from the mean. We will continue with the assumption that $a > 0$. If $\gamma < 1$, then $\lim_{\gamma\to-\infty}\sigma = 0$, because the product, $(\gamma-3)(\gamma-2)^2$ dominates $\gamma-1$, and both the numerator and denominator will have the same sign. Note, however, that the numerator will dominate so long as $|\,(\gamma-1)a^2\,| > (\gamma-3)(\gamma-2)^2$ so there forms a tailed distribution to the left of $(\gamma,\sigma) \to (1,0)$.

If $1 < \gamma \leq 2$, then σ is undefined, indicating no second moment exists. Likewise, if $2 < \gamma \leq< 3$, then σ is also undefined. When $\gamma > 3$, $(\gamma-1)a^2$ dominates $(\gamma-3)(\gamma-2)^2$ within some distance ϵ of $\gamma = 3$. Otherwise, $\lim_{\gamma\to\infty}\sigma = 0$.

0.6 Problem 6

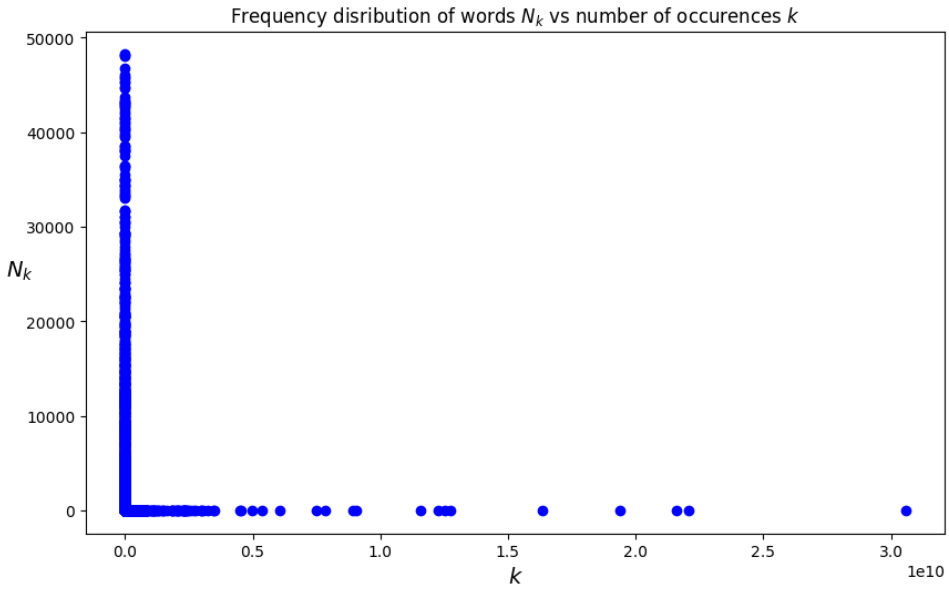

Figure 1: Frequency distribution N_k representing how many distinct words appear k times in a Google text corpus.

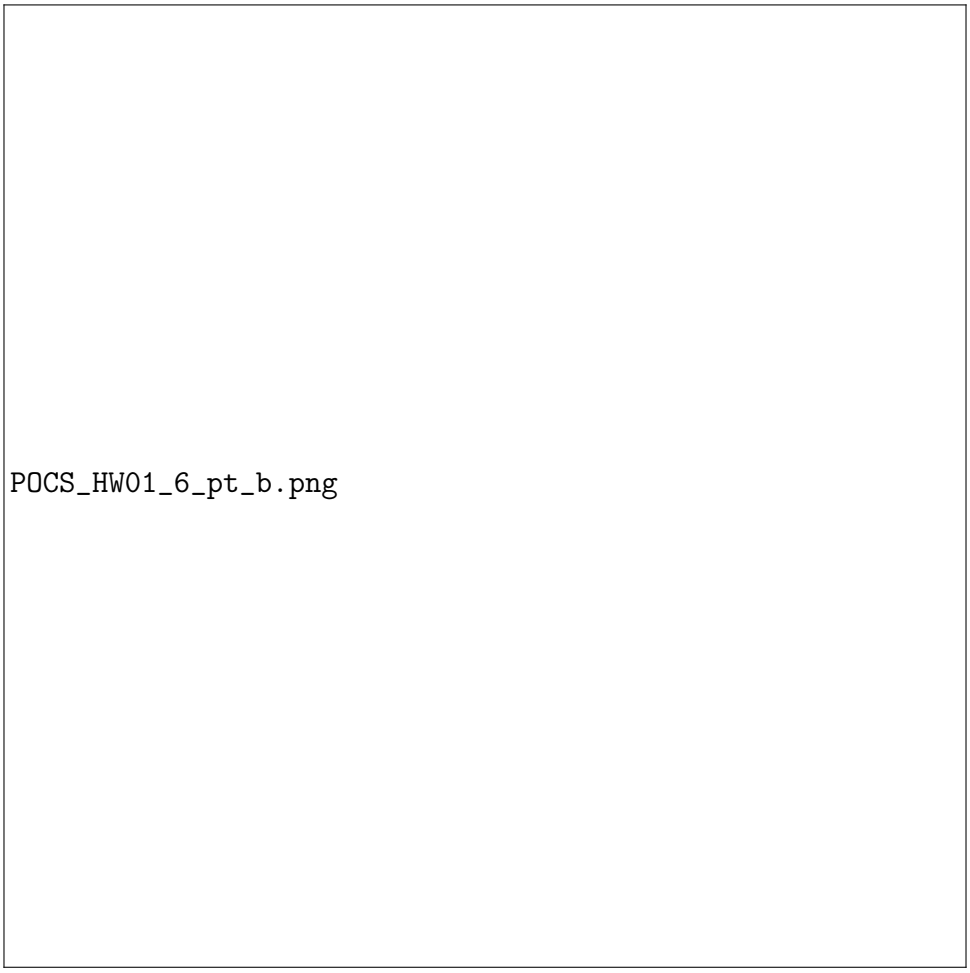

Figure 2: Log-log Frequency distribution N_k representing how many distinct words appear k times in a Google text corpus.

0.7 Problem 7

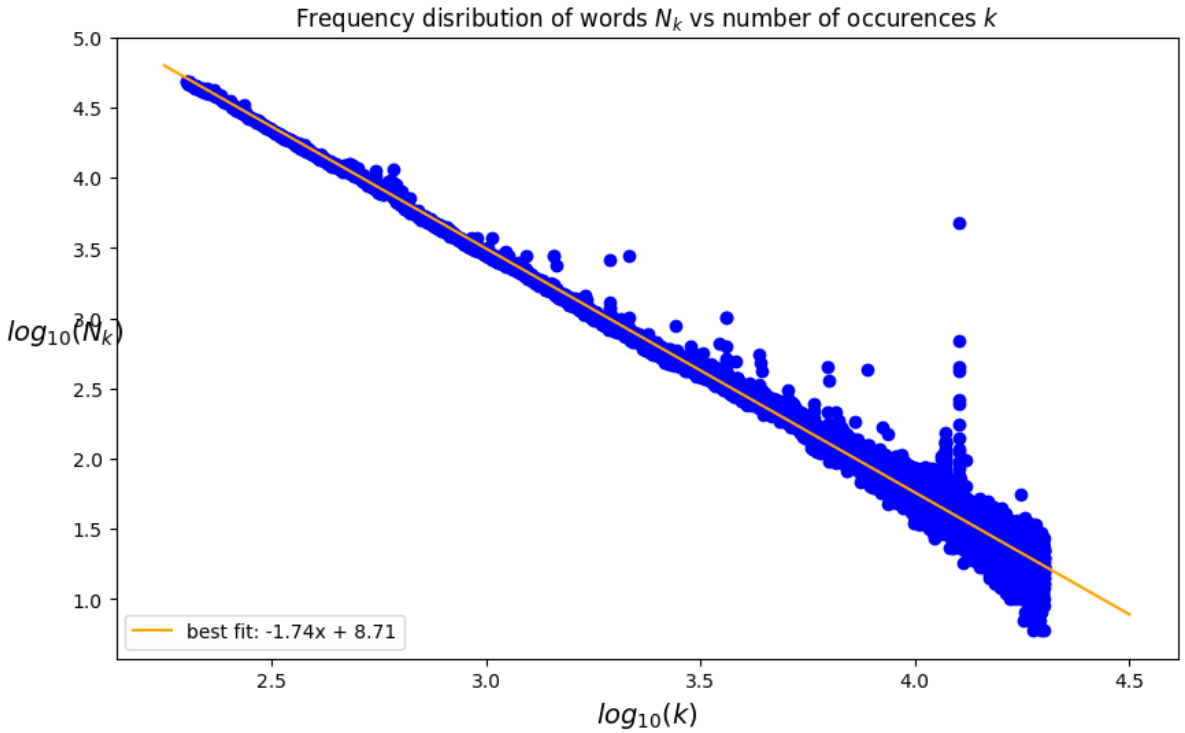

Figure 3: Best fit line $\forall N_k \in (10^{2.5}, 10^{4.4})$

0.8 Problem 8

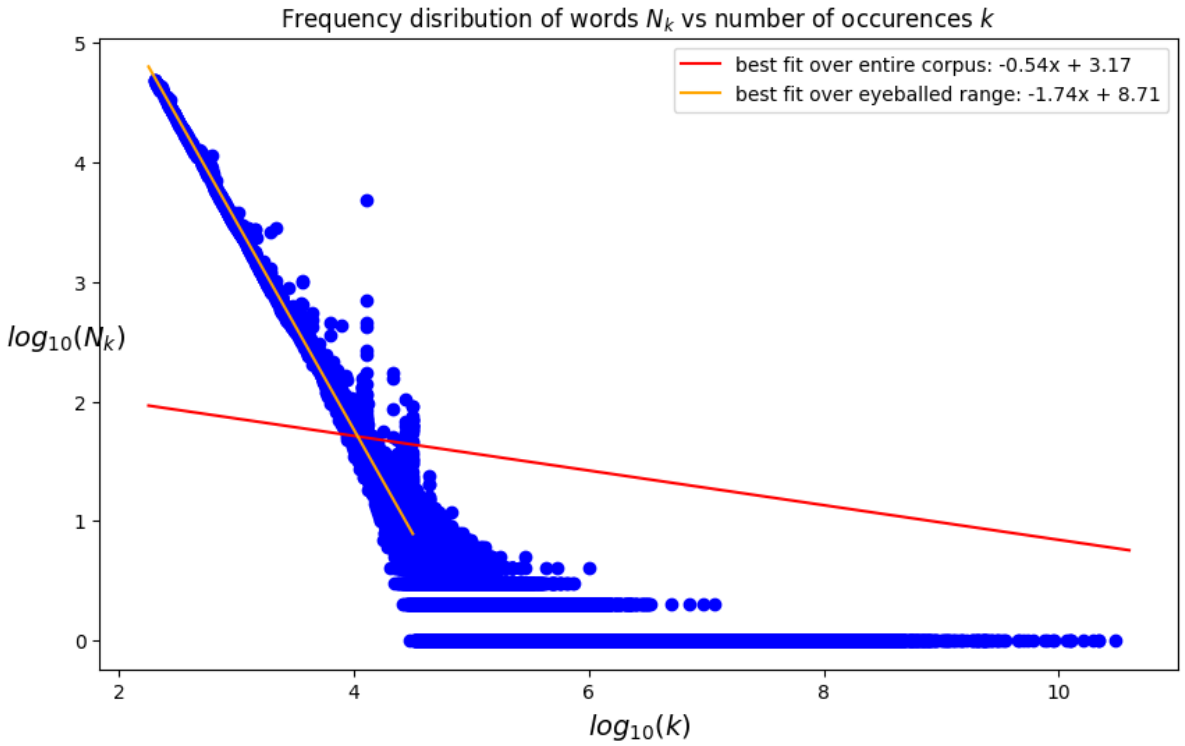

Figure 4: Best fit line over entire Google text corpus

After analyzing the raw frequency data, I determined that the mean of the distribution, $\mathbf{E}[N_k]$, is 3363719.22 occurrences, having a variance, $\sigma^2_{N_k} = (126384425.38)^2$ occurrences. The value we calculated for the mean conforms well to our estimate of $\gamma = -1.74$. First, we calculate the mean of the data and the standard deviation. Then, we use the minimum and maximum values of x in the sample space to find $\mathbf{E}[N_k]$ and $\mathbf{E}[N_k^2]$, substituting $\gamma = -1.74$. Finally, we can find our estimated σ and mean, and report std.$_{err}$. (I did something wrong here).

0.9 Problem 9

A parent has two children, not twins, and one is a girl born on a Tuesday. What's the probability that both children are girls? See if you can produce both a calculation of probabilities and a visual explanation with shapes (e.g., discs and pie pieces). Once you have the answer, can you improve our intuition here? Why does adding the more detailed piece of information of the Tuesday birth change the probability from 1/3? (Assume 50/50 birth probabilities.)

In the original problem, we wanted to know the probability that two children were girls. Hence, our sample space looked like:

B	G
B	B
G	B
G	G

and the probability that the second child is a girl, given we know the first is a girl, is 1/3. When we add the information that the child is a girl born on Tuesday, then we not only care about the order and gender of the birth, but also the day. For every day, there are two pairs of possible girls: a girl is born first on that day, and a girl is born second on Tuesday, or a girl is first born Tuesday, and then second on that day. However, there is only one way to have a pair of two girls on Tuesday: a girl is born first, then a girl is born second. Hence, there are only 13 acceptable combinations.

Boys are a bit different, there is no constraint for boys born on Tuesday like there were for girls. For every boy: a boy can be born first one day and then a girl on Tuesday, or a girl can be born on Tuesday and then a boy can be born second on some other day. Hence, There are 14 possible ways to pair a boy with a girl born on Tuesday. Thus, the probability of having two girls given that one is born Tuesday, is $\frac{13}{13+14} = \frac{13}{27}$

The Problem's sample space is not so large, so we can model the problem in a diagram. Below, is a directed graph where each node represents a girl or boy, the day they were born, and which order they were born (1st or 2nd). The nodes are of either set $B = \{b_{ij}\}$, or $G = \{g_{ij}\}$, where i is the order of birth, and j is the day of birth. Note that both g_{1T} and g_{2T} have been removed from the set G. The edges between nodes represent a pair of siblings in the sample space. The direction of the arrows correspond to the first child and the second child respectively. Since we are given that one of the children is a girl who was born on a

Tuesday, arrows leave from the two girl nodes (g_{1T}, and g_{2T}) that represent a Tuesday birth.

The red edges correspond to allowable pairs of girl pairs given the two possible Tuesday births.The black edges represent non-allowable pairs (i.e girl→boy and boy→girl links). Either we may construct an adjacency matrix and divide the sum of the number of edges that g_{1T} has and g_{2T} has by the number of possible edges, or we may simply divide the red edges by the total number of edges (they are equivalent). To determine red edges, a simple check can be done for each of the sets B and G over the nodes g_{1T}, and g_{2T}: if $i = 1$, create an edge with g_{2T}. Else, if $i = 2$, form an edge with g_{1T}. If an edge has two nodes from the set G, then color the edge between them red. This also scales, say, if the sample space expands to be the probability that a girl was born on December 31st. In that case, we would see a larger annulus of nodes; like an expanding disk.

In this problem, the answer moves away from 1/3 because there are fewer possible edges that can be formed for children born on a Tuesday. Now, there are 27 allowable connections, only 13 of which are viable solutions.

Hence the answer is $\frac{13}{27}$ ∴.

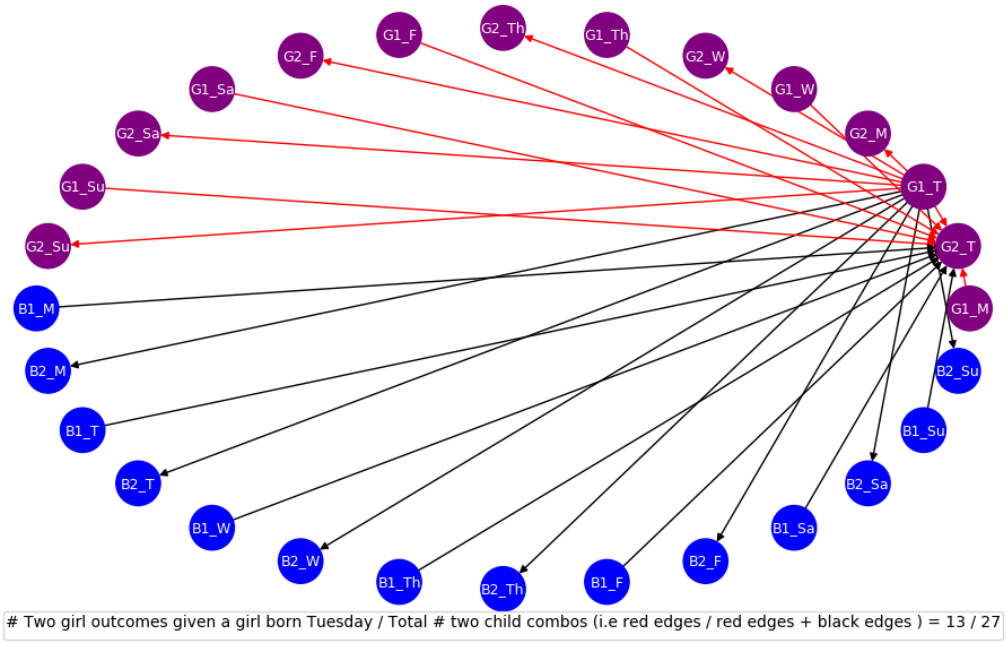

Figure 5: Proportion of two child combos that contain two girls, given a girl was born on a Tuesday

If the problem changes, and now we know that a girl is born on December 31st, then

the same methodology applies as before. For every day, each boy has 2 ways to pair with a Tuesday girl so there are $2 \times 365 = 730$ possible pairings. However, there is still only one way to pair girls born on the 31st. Hence, our possible pairings is reduced to 729. Thus, the probability of having a girl born on December 31st and then having another girl is $\frac{2\times730-1}{2\times730+2\times730-1} = \frac{729}{1459}$.

In general we can let D_t be the set of all days, where t denotes what subset of days we are taking (i.e: weeks means that $t = 7$) in which we have two girls. We can say that:

$$P(t) = \frac{2|D_t| - 1}{4|D_t| - 1}$$

0.10 Extra Credit: Modeling the Monty Hall Problem

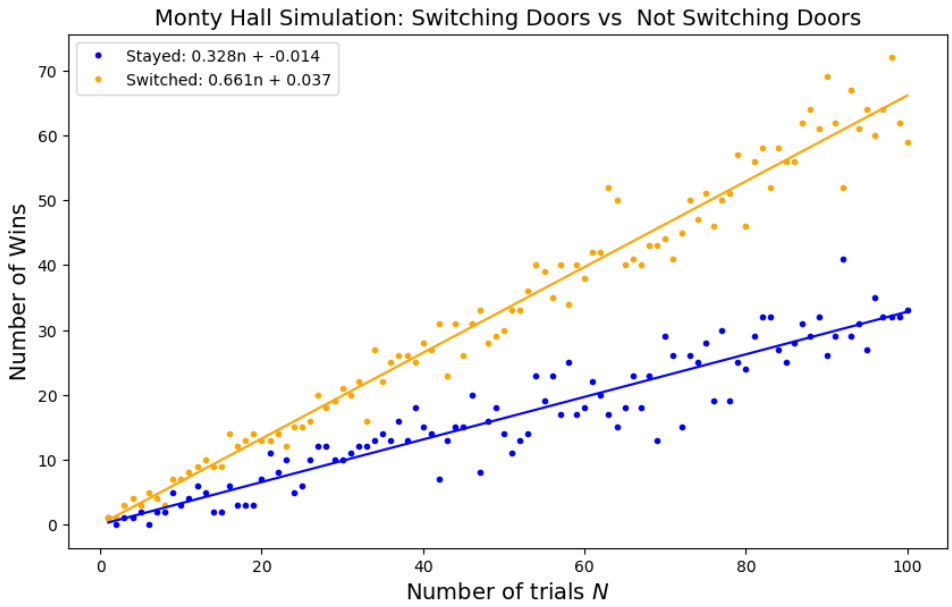

Figure 6: Monty Hall simulation demonstrating that it is more advantageous to switch doors when a goat is revealed.