Set 5 - Modelling data with Pareto's Law and Zipf's Law

Vishal Dhoriya (202101446)* and Akhil Patoliya (202101505)†
Dhirubhai Ambani Institute of Information & Communication Technology,
Gandhinagar, Gujarat 382007, India
CS302, Modelling and Simulation

In this report, we explore two prominent statistical phenomena: Pareto's Law and Zipf's Law. Pareto's Law elucidates the unequal distribution of resources, while Zipf's Law uncovers patterns in the frequency of occurrences across ranked lists.

I. THE PARETO'S LAW

A. Model

Pareto's Law, commonly known as the 80/20 rule, states that a significant portion of effects come from a small fraction of causes, reflecting the unequal distribution of outcomes in many phenomena. The Pareto's equation is given by:

$$N(x) = A + Bx^{-\alpha} \tag{1}$$

where x is the amount of wealth and $N(x)$ is the frequency distribution of wealth holders.

Here we use this equation to model the distribution of wealth in India.

B. Results

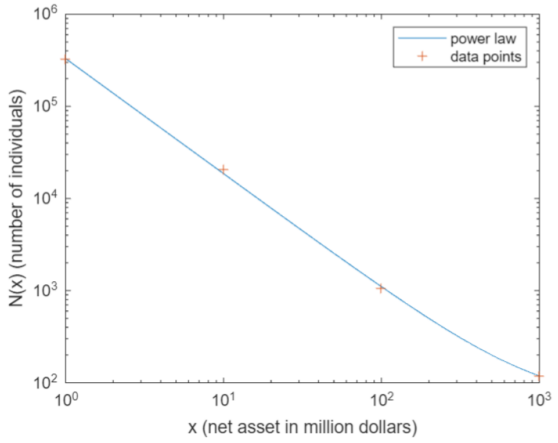

FIG. 1: Plot of $N(x)$ vs x with $A = 60$, $\alpha = \frac{5}{4}$, and $B = 300,000$.

II. THE ZIPF'S LAW

A. Model

Zipf's Law, describes a dependency observed in various datasets where the frequency of an item is inversely proportional to its rank highlighting a fundamental principle of unequal distribution in large-scale datasets.

Here we use Zipf's Law for modelling dependency distributions in Debian. The model equation used for analysis is given by:

$$\phi(x) = \left[\eta + \left(\frac{x+\lambda}{c}\right)^{-\mu\alpha}\right]^{-1/\mu} \tag{2}$$

$$\phi(x) = \eta + \left(\frac{c}{x+\lambda}\right)^{2} \tag{3}$$

where c,η and λ are constants.

B. Results

1. We plot six figures for three successful releases of Debian: Etch, Lenny, and Squeeze.
2. For each release, we plot two graphs: Distribution of Incoming and Outgoing Dependency links. The x-axis represents the number of links, and the y-axis represents the actual software packages (frequency distribution) of the links.

*Electronic address: 202101446@daiict.ac.in
†Electronic address: 202101505@daiict.ac.in

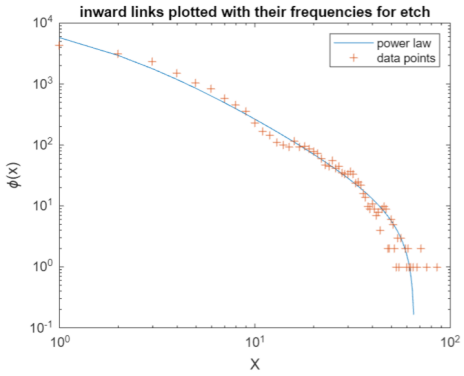

FIG. 2: Plot of degree distribution of incoming links for Etch release with $\eta = -8$, $c = 190$, and $\lambda = 1.5$.

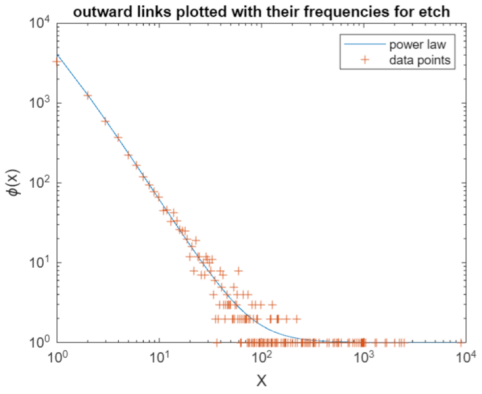

FIG. 3: Plot of degree distribution of outgoing links for Etch release with $\eta = 1$, $c = 80$, and $\lambda = 0.25$.

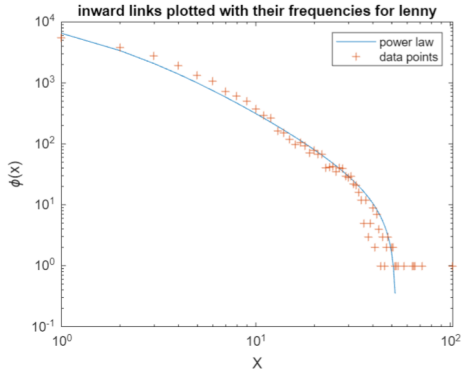

FIG. 4: Plot of degree distribution of incoming links for Lenny release with $\eta = -15$, $c = 210$, and $\lambda = 1.6$.

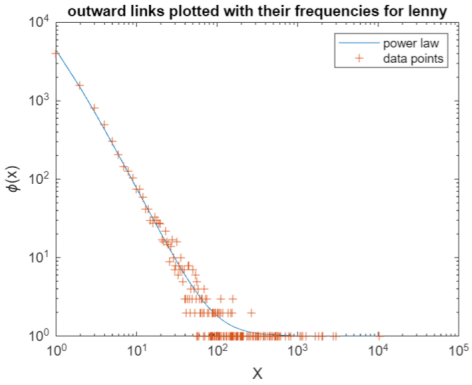

FIG. 5: Plot of degree distribution of outgoing links for Lenny release with $\eta = 1$, $c = 90$, and $\lambda = 0.35$.

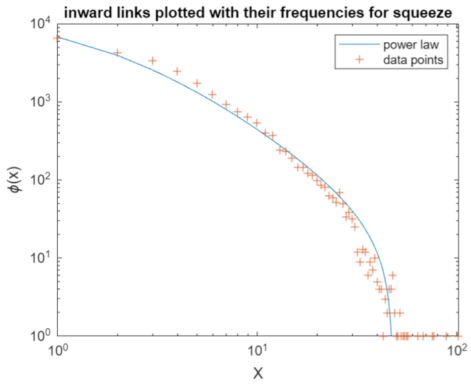

FIG. 6: Plot of degree distribution of incoming links for Squeeze release with $\eta = -28$, $c = 265$, and $\lambda = 2.2$.

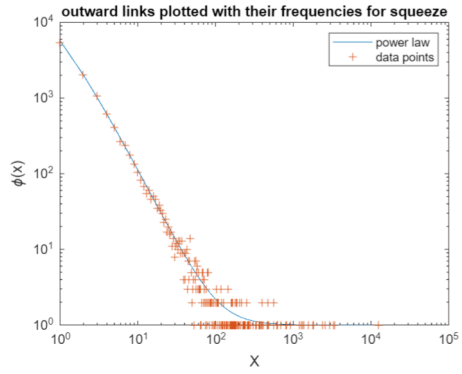

FIG. 7: Plot of degree distribution of outgoing links for Squeeze release with $\eta = 1$, $c = 110$, and $\lambda = 0.45$.

C. Conclusions

1. Pareto's law leads to the conclusion that an individual's net worth is inversely proportionate to their population size.

2. For intermediate nodes, the degree distributions obey Zipf's law; however, the degree distributions of densely and poorly linked nodes depart from this pattern and show finite-size effects.

3. With each software version, the number of nodes generating out-directed links increases, but eventually saturates since the network's semantic possibilities are finite. Studying the saturation characteristics and directional features of scale-free networks in general, like the World Wide Web, is possible with the mathematical model created in this research.