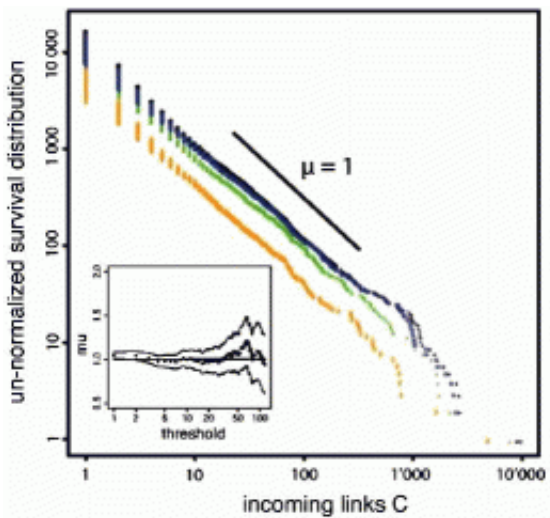# Linux Evolution Reveals Origins of Curious Mathematical Phenomenon

December 1 2008, By Lisa Zyga

When the Zipf curve is plotted on a log-log scale, it appears as a straight line with a slope of -1. This graph shows that four Debian Linux releases each follow Zipf's law: Woody (orange), Sarge (green), Etch (blue) and Lenny (black). Credit: T. Maillart, et al.

(PhysOrg.com) -- Zipf's law is a testament to the order in our world, showing that the same patterns emerge in a wide variety of situations. The linguist George Kingsley Zipf first proposed the law in 1949, when he noticed that the distribution of words in a newspaper, book, or other literary article always followed the same pattern.

Zipf counted how many times each word appeared, and found that the

probability of the occurrence of words starts high and tapers off. Specifically, the most frequent word occurs about twice as often as the second most frequent word, which occurs about twice as often as the fourth most frequent word, and so on. Mathematically, this means that the frequency of any word is inversely proportional to its rank. When the Zipf curve is plotted on a log-log scale, it appears as a straight line with a slope of -1.

Since Zipf's discovery, researchers have found that the power law describes many other natural and human phenomena, including the distribution of cities ranked by their population, the distribution of corporate wealth, and Internet traffic characteristics.

When analyzing systems that follow Zipf's law, researchers usually assume certain mechanisms to be responsible for this patterned behavior. However, no one has ever empirically demonstrated that these assumed mechanisms are indeed the origin of Zipf's law.

Now, a team of researchers from ETH Zürich (the Swiss Federal Institute of Technology Zürich) in Switzerland has confirmed that these assumed mechanisms – such as scale-free, proportional growth rates – are at the origin of Zipf's law. The researchers used four orders of magnitude of data detailing the evolution of open source software applications created for a Linux operating system to confirm the assumption.

The team studied Debian Linux, a free operating system continuously being developed by more than 1,000 volunteers from around the world. Developers create software packages, such as text editors or music players, that are added to the system. Beginning with 474 packages in 1996, Debian Linux has expanded to include more than 18,000 packages today. The packages form an intricate network, with some packages having greater connectivity than others, as defined by how many other

packages depend on a given package.

"Open source offers a unique opportunity provided by the high completeness of data concerning open source (thanks to the disclosure policy of the open source terms of license)," lead author Thomas Maillart of ETH Zürich told *PhysOrg.com*. "Debian Linux allowed us to retrieve exhaustive information from several years ago. Many other complex systems are not so well 'documented.'"

As the researchers explain, the Linux network is constantly changing: new packages enter, some disappear, and others gain or lose connectivity. Yet throughout the 12 years, the distribution of packages, as ranked by their number of incoming links from other packages, has followed Zipf's law, with a few very popular packages having much greater connectivity than most.

While many previous models of Zipf's law start with the assumption that the set of entities (e.g. packages) appeared at the same time, the Swiss researchers track the time evolution of package connectivity in the Linux network since 1996. This perspective enabled them to test for the presence of specific characteristics of the growth of the Linux network, which leads to the emergence of Zipf's law.

Using the data, they showed that the growth rates of connectivities between packages are proportional to the degree of connectivity between packages. In addition, they showed empirically that the average growth rate of the total number of links to a given package over a time interval is proportional to that time interval. Further, the variability of the total number of links to a given package increases proportionally to the square-root of time, providing a crucial test of the mechanism of stochastic proportional growth of connectivity between packages. Altogether, these characteristics are responsible for the universal distribution pattern of Zipf's law.

"We show that the distribution of connectivity of new entrants is also a power law with an exponent much bigger than 1, confirming that the proportional growth mechanism is solely responsible for the Zipf's law," Maillart said.

He explained that, while Linux data allowed the researchers to confirm the origins of Zipf's law, their results bring up more questions.

"Linux Debian gave us the opportunity to verify the 'proportional mechanism,' thanks to an important dataset and a huge investigation potential," Maillart said. "All changes (evolution) in open source software are freely available and therefore can be tracked in detail. However, model verification has brought one answer and many resulting questions we intend to give an answer to. We think particularly of mechanisms of success/failure of projects in relation with their management.

"Remember that we still do not clearly understand the reasons of the success of the open source, since it's free and based on altruist contributions by programmers," he said. "Additionally, one can bet that further research in this direction (open source and proportional growth) may raise useful questions for other systems (cities, economy, etc.) that would bring new insights to explain their evolution."

More information: T. Maillart.; D. Sornette; S. Spaeth, and G. von Krogh. "Empircal Tests of Zipf's Law Mechanism in Open Source Linux Distribution." *Physical Review Letters* 101 218701 (2008).