Zipf's Law's Application and It's Connection with Network Theory

Jingjin Wei
Department of Physics, Boston University, Boston, Massachusetts 02215, USA
(Dated: April 23, 2016)

Zipf's Law is a kind of power law that describes many types of data studied in the physical and social sciences. We usually use it when we are looking at ranking vs frequency or ranking vs resources data. In general it describes this phenomena that the more you have, the more you will get. In network science, there is a type of network, scale-free network where the degree distribution is in power law. This paper discuss about the similarity between two field and discuss a general phenomena behind it.

I. INTRODUCTION

Zipf's law is named after the American linguist George Kingsley Zipf, who popularized it and sought to explain it, though he did not claim to have originated it. He found that, in the English language, the probability of encountering the rth most common word is given roughly by $P(r) = 0.1/r$ for r up to 1000 or so. Since then, people has tried to apply this law to many different data. It is found that it's not limited to language studies but also for other ranking data. For example, people have studied the size with ranking, like city size [1–3], web page visits [4–6], firm size and other economics data [7]. While it's almost a inverse relationship in language, in other fields it's not exactly but always close to inverse, which means the exponential index would vary depending on the field we are looking at.

A scale-free network[8] in network theory, is a network with a power law degree distribution. That is, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as $P(k) \sim k^{-\gamma}$.where γ is a parameter whose value is typically in the range $2 < \gamma < 3$. In the past few years, many network has been recognized. Many of them we can see in our daily life: air traffic networks, banking networks, chemical bonds networks, data, communications networks, ecosystems networks, interstate highways, journal citations, material structures, nervous systems, oil pipelines, organizational networks, power grids, social, structures, transportation networks, voice communication networks, water supply networks, and Web URLs.

In this paper, we first test the generality of Zipf's law in various field: transportation, economy, population and etc. After that, we will take a look at network theory on scale-free network, and see if there is some idea we can borrow from that to explain what we see in our data.

II. ZIPF'S LAW

A. Data Visualization;The universality of Zipf's law

With the data from CIA the world fact book, we do plots on various fields. Beware that here we are using only the first 30% of the countries, while they own over 90% of the resources we are talking about here.

In the six subfields we chose to plot, two come from transportation (railways and roadways), one communication (internet users), one economy (GDP), the other two basic resources of a country. As mentioned before, they all come from very different field, but from the graph we can see that the slope of the fitting curves are all close to -1. Especially for the data points other than top 5 in the lists, they conforms to a inversely linear relationship quite well.

B. Preferential attachment and Zipf's law

What exactly does Zipf's law tell us? In the beginning it's only about languages. Afterwards, people started to realize that it could be applied to many fields, and it does. What's the reason for that? Does this law come from some more basic underlying theorem of the nature? From my understanding of network theory I sense a similarity.

First let's think of the true meaning of Zipf's law. The most popular use of it is to describe the frequency as a function of rank. In our case, the frequency is how much a country has, which is obviously an equivalency as the frequency using in language study. If you think of it, this relationship actually involves three parties: *rank*, *frequency*, *'size'*. Here we use the word 'size' to represent the amount of the 'resources', for example, in language study it means the number of time the word appears, in paper citation case it means the number of citations a paper gets. Intuitively we can see why is that. Imagine we have a market where different people own different amount of apples, and we rank them by the number of apples they have. Then: 1) No doubt people higher in ranking have more apples, which comes from our premise; 2) For those who have more apples, if a new apple is coming to this market, then they have higher chance to

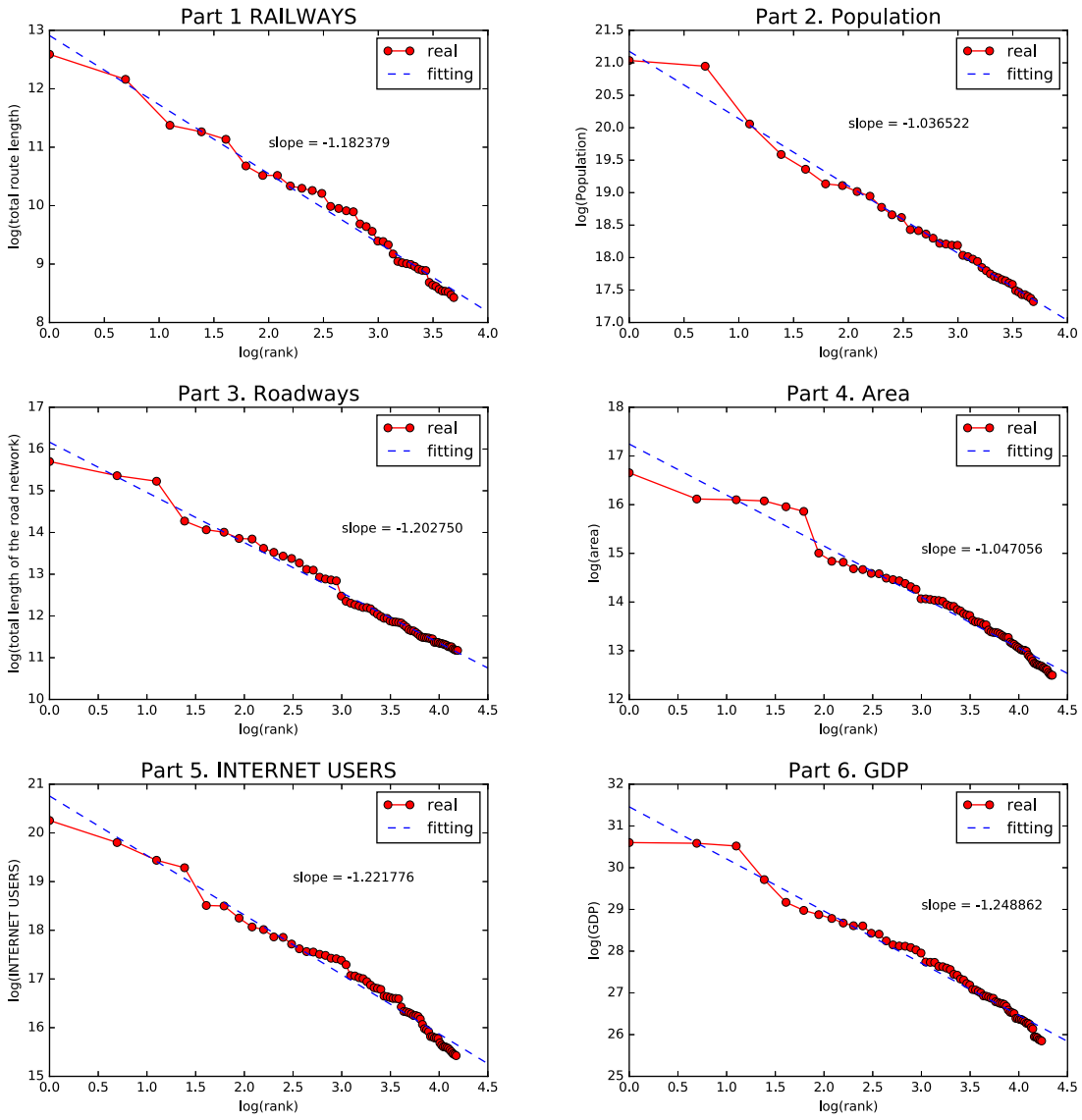

FIG. 1: Data visualization. (1)-(6): Railways, poplation, roadways, areas,Internet users, GDP.

get the apples. Because the fact 'having more apples' itself, in some sense, imply that these people having more resources and thus more attractive to the new coming resources.

To sum up, for the three parties triangle, rank is inversely proportional to frequency and 'size', frequency and 'size' are proportional. That is exactly the idea of '*preferential attachment*'. As we have mentioned in Introduction, scale-free network also comes from this idea, and its degree distribution also has a log-log relationship: we start from an initial state, but then every time step we have a new link. The probability for this link connects to node A is proportional to the degree of node A, which is how many links node A already has, and is very similar to the 'size' we have in our previous analysis.

III. NETWORK THEORY AND ZIPF'S LAW

Given the many similarities we can see between Zipf's law's application in various cases and many real-world scale-free network, we could see if network theory could be applied to the data that we studied using Zipf's law. Let's first take a look at the slope, since in scale-free network the exponential is sensitive and important.

In general, the relationship between frequency and rank could be described as:

$$\begin{aligned} P_i &= c_1 \cdot {rank_i}^{-\alpha} \\ &= c_2 \cdot {k_i}^{\alpha} \end{aligned} \tag{1}$$

In which c_1, c_2 are some constants. Here denotes the dependency of frequency on rank. It is clear that when

field	slope
Railways	-1.18238
Population	**-1.03652**
Roadways	-1.20275
Area	**-1.04706**
Internet Users	-1.22178
GDP	-1.24886

TABLE I: Slope of log(freq)-log(rank) present in fig.1

$\alpha = 1$, that is a inversely proportional relationship, and is our start point; When $\alpha = 0$, that means they are independent, the distribution of the new coming resources has no preference at all. Then take a look at the table, for sure we don't have α exactly equals 1. Here comes the question: If the exponential, or the slope in the log-log plot, is not 1, what does that mean?

The study of scale-free network[9] shed light upon this question. In his book, Barabási pointed out that for scale-free network, with different exponential value, we can divide them into different groups. For $\alpha = 0$, no dependency, no preferential attachment. For $0 < \alpha < 1$, which is called *sublinear regime*, has fewer and smaller hubs than in a scale-free network. At this point we start to have some preference, but it is not very biased. For $\alpha = 1$, that is scale-free network. It is also a line dividing stable and unstable network. For $\alpha > 1$, which is called *superlinear regime*. This kind of network could be described as 'winner-takes-all'. It features with several super hubs.

When we look at table.1, there are something that we could notice: 1) All the slopes are around -1, which shows that in many aspects of our daily life, '*the rich gets richer*' is the general case. 2) None of them are much bigger than 1, since as predict, when it is too big, the society as a whole will become unstable.

IV. CONCLUSION

In this paper we first reproduce the universality of the Zipf's law in various fields. After that, we point out part of the underlying nature of this law. Then given the similar feature, we try to apply the theory and discoveries in network theory to go a step further, and confirm their similarities. As for further research, if we take a closer look at the table, we can see that some of the slopes are around -1.2, which are subfields from transportation, communication and economy; but some are around -1, which are area and population. More research is needed if we want to determine if this is universal or what we see here are just some special cases.

[1] Xavier Gabaix. Zipf's law for cities: an explanation. *Quarterly journal of Economics*, pages 739–767, 1999.

[2] Kwok Tong Soo. Zipf's law for cities: a cross-country investigation. *Regional science and urban Economics*, 35 (3):239–263, 2005.

[3] Juan-Carlos Córdoba. On the distribution of city sizes. *Journal of urban Economics*, 63(1):177–197, 2008.

[4] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, pages 126–134. IEEE, 1999.

[5] Lada A Adamic and Bernardo A Huberman. Zipf's law and the internet. *Glottometrics*, 3(1):143–150, 2002.

[6] Mark Levene, José Borges, and George Loizou. Zipf's law for web surfers. *Knowledge and Information Systems*, 3 (1):120–129, 2001.

[7] Robert L Axtell. Zipf distribution of us firm sizes. *Science*, 293(5536):1818–1820, 2001.

[8] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[9] ALBERT-LÁSZLÓ BARABÁSI. Network science.