SOCIAL NETWORK ANALYTICS

Pareto Law and Zipf's Law

Prakash C O

Department of Computer Science and Engineering

SOCIAL NETWORK ANALYTICS

Pareto Law (80/20 rule)

Prakash C O

Department of Computer Science and Engineering

The Pareto Law or '80–20' rule

➢**There are two variants of Power law:**

1. **Pareto Law('80-20' rule/Golden ratio)** and

2. **Zipf's Law.**

The Pareto Law or '80–20' rule

What Is the Pareto Law/Principle?

➢ The Pareto Principle, named after esteemed economist Vilfredo Pareto.

➢ **Pareto Principle specifies that 80% of consequences come from 20% of the causes**, asserting an unequal relationship between inputs and outputs.

➢ This principle serves as **a general reminder that the relationship between inputs and outputs is not balanced**.

Vilfredo Pareto
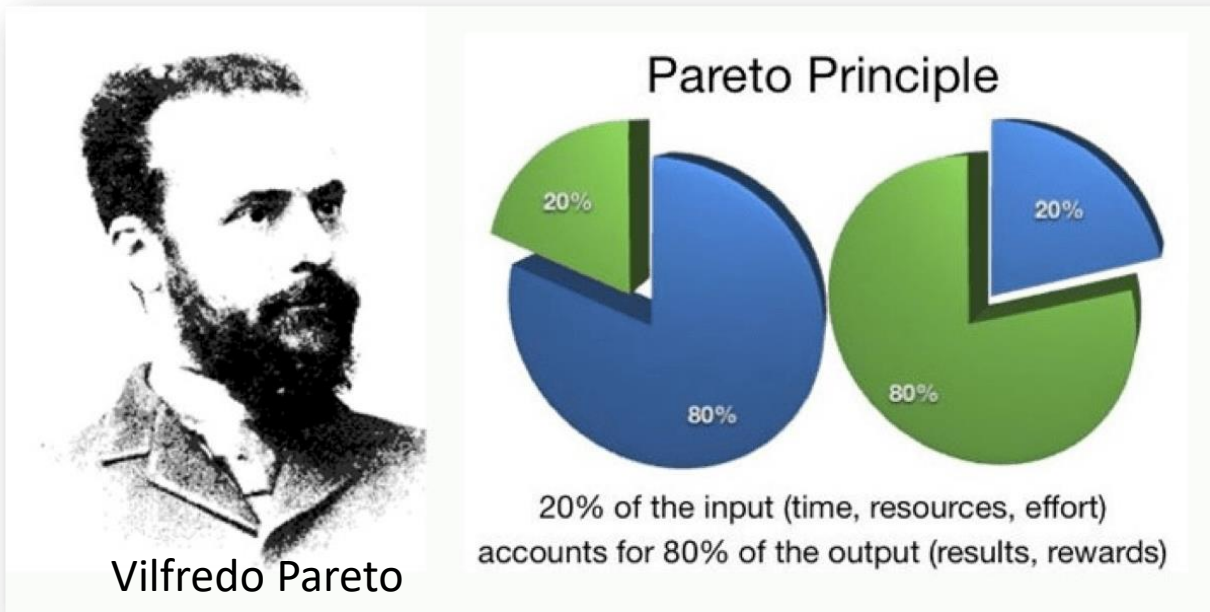
What Is the Pareto Law/Principle?

➢ Pareto Principle is **an observation that things in life are not always distributed evenly.**

• The principle, which was **derived from the imbalance of land ownership in Italy**, is commonly used to illustrate the notion that **not things are equal, and the minority owns the majority.**

• Unlike other principles, the **Pareto Principle is merely an observation, not law**. Although broadly applied, **it does not apply to every scenario.**

The Pareto Law or '80–20' rule
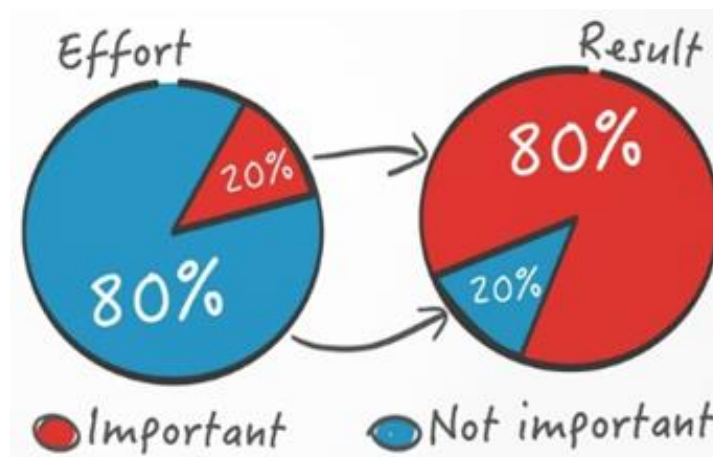
➢ **Examples**

- The 20% of the Web sites gets the 80% of the visits (actual data: 15%-85%)
- The 20% of the Internet routers handles the 80% of the total Internet traffic
- The 20% of world industries hold the 80% of the world's income
- The 20% of the world population consumes the 80% of the world's resources
- The 20% of the Italian population holds the 80% of the lands (that was true before the Mussolini fascist regime, when lands re-distribution occurred)
- The 20% of the earthquakes caused the 80% of the victims
- The 20% of the rivers in the world carry the 80% of the total sweet water
- The 20% of the proteins handles the 80% of the most critical metabolic processes

➢ Does this derive from the power law distribution? YES!

The Pareto Law or '80–20' rule

Advantages of the Pareto Principle

➢ There is a practical reason for applying the Pareto Principle.
Simply, **it can give you a window into who to reward or what to fix**.

➢ **Examples:**

- if 20% of the design flaws in a car are leading to 80% of the crashes, you can identify and fix those flaws.

- if 20% of your customers are driving 80% of your sales, you may want to focus on those customers and reward them for their loyalty.

➢ In this sense, the Pareto Principle **becomes a guide for how to allocate resources efficiently**.

Uses of the Pareto Principle

➢ The Pareto Principle helps you realize that **the majority of results come from a minority of inputs**. Knowing this, if…

- 20% of workers contribute 80% of results: Focus on rewarding these employees.

- 20% of bugs contribute 80% of crashes: Focus on fixing these bugs first.

- 20% of customers contribute 80% of revenue: Focus on satisfying these customers.

- The examples go on.

➢ The point is to realize that you can often **focus your effort on the 20% that makes a difference**, instead of the 80% that doesn't add much.

The Pareto Law or '80–20' rule

Uses of the Pareto Principle

Ref: https://www.thebalancecareers.com/pareto-s-principle-the-80-20-rule-2275148

Disadvantages of the Pareto Principle

➢ While the **80-20 split** is true for Pareto's observation, that **doesn't necessarily mean that it is always true.**

For instance, 30% of the workforce (or 30 out of 100 workers) may only complete 60% of the output.

The remaining workers may not be as productive or may just be slacking off on the job.

➢ This further reiterates that **the Pareto Principle is merely an observation and not necessarily a law.**

SOCIAL NETWORK ANALYTICS

Zipf's Law

Prakash C O

Department of Computer Science and Engineering

Zipf's Law

➢ **Zipf's law** is an empirical law formulated using mathematical statistics that refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution

➢ Both natural and artificial systems often exhibit a surprising degree of statistical regularity.

➢ One such regularity is Zipf's law. Originally formulated for word frequency.

➢ **Zipf's law has since been observed in a broad range of domains**, including city size, firm size, mutual fund size, amino acid sequences, and neural activity.

Zipf's Law

➢ **Zipf's law is a relation between rank order and frequency of occurrence.**

➢ Zipf's law states that when observations (e.g., words) are ranked by their frequency, the frequency of a particular observation is inversely proportional to its rank,

$$\text{Frequency} \propto \frac{1}{\text{Rank}}$$

Or the frequency of the *k*th most frequent word is proportional to 1/*k*.

Note: Zipf's law is not an exact law, but a statistical law and therefore does not hold exactly but only on average.

	frequency (per milln)	cumulative frequency	frequency rank	alphabet rank
the	68351.63	68351.63	1	318525
of	33008.66	101360.29	2	212425
and	28651.11	130011.40	3	11331
to	27599.22	157610.62	4	322312
a	23160.48	180771.10	5	1
in	20670.81	201441.91	6	149032
is	10571.15	212013.06	7	156934
that	10549.02	222562.08	8	318470
was	9939.26	232501.34	9	356587
it	9882.90	242384.23	10	157771
for	9309.44	251693.67	11	114281
on	7636.66	259330.33	12	213645
with	7171.07	266501.39	13	361235
he	7167.84	273669.23	14	134413
be	7153.17	280822.40	15	27945
I	7036.88	287859.28	16	146205
by	5866.89	293726.17	17	44040
as	5793.35	299519.52	18	19178
at	5154.12	304673.64	19	20631
you	5043.27	309716.91	20	364651
are	5000.14	314717.05	21	17618
his	4963.47	319680.52	22	139433
had	4922.27	324602.79	23	131212
not	4899.77	329502.56	24	209444
this	4789.41	334291.97	25	319827
have	4685.82	338977.79	26	134106
from	4625.21	343603.01	27	117354
but	4616.26	348219.26	28	43732
which	4131.11	352350.37	29	358956

Zipf's Law

➢**Zipf's law usually refers to the 'size' y of an occurrence of an event relative to it's rank r.**

➢George Kingsley Zipf, a Harvard linguistics professor, sought to determine the 'size' of the 3rd or 8th or 100th most common word. Size here denotes the frequency of use of the word in English text.

➢**Zipf's law states that the size of the r'th largest occurrence of the event is inversely proportional to it's rank:**

$$y \sim r^{-b}, \text{ with } b \text{ close to unity}$$

Zipf's Law

Summary:

➤ Cumulative distributions with a power-law form are sometimes said to follow Zipf 's law or a Pareto distribution.

➤ **Zipf's law and the Pareto distribution differ from one another in the way the cumulative distribution is plotted**—

- Zipf made his plots with rank on the horizontal axis and frequency on the vertical one;

- Pareto did it the other way around.

Zipf's Law

➢**Assignment: Reading Paper:**

1. Zipf's law and the Internet - Lada A. Adamic, Bernardo A. Huberman

2. Identification of Suspicious Patterns in Social Network using Zipf's law

   - GeethikaSarna, Dr. MPS Bhatia

Summary

➢**All three terms("power-law", "Zipf", "Pareto") are used to describe phenomena where large events are rare, but small ones quite common.**

➢**Examples:**

    ➢There are few large earthquakes but many small ones. There are a few mega-cities, but many small towns.

    ➢There are few words, such as 'and' and 'the' that occur very frequently, but many which occur rarely.

References

➢ Wikipedia – Current Literature

➢ https://www.thebalancecareers.com/pareto-s-principle-the-80-20-rule-2275148

➢ https://www.investopedia.com/terms/p/paretoprinciple.asp

THANK YOU

Prakash C O

Department of Computer Science and Engineering

coprakasha@pes.edu

+91 98 8059 1946

The Pareto Law or '80–20' rule

➤ It postulates things like: 20% of the worlds population own 80% of the wealth.

➤ The fraction W of the wealth in the hands of the richest P of the the population is given by

$$W = P^{(\alpha-2)/(\alpha-1)}$$

➤ Example: US wealth: **α** = 2.1

- richest 20% of the population holds 86% of the wealth