Lesson3:
Modeling the Web with Advanced Statistical Descriptive Text Models
Unit4:
Zipf's law, Powerlaw or Pareto law – What's the difference?!?

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST
People and Knowledge Networks

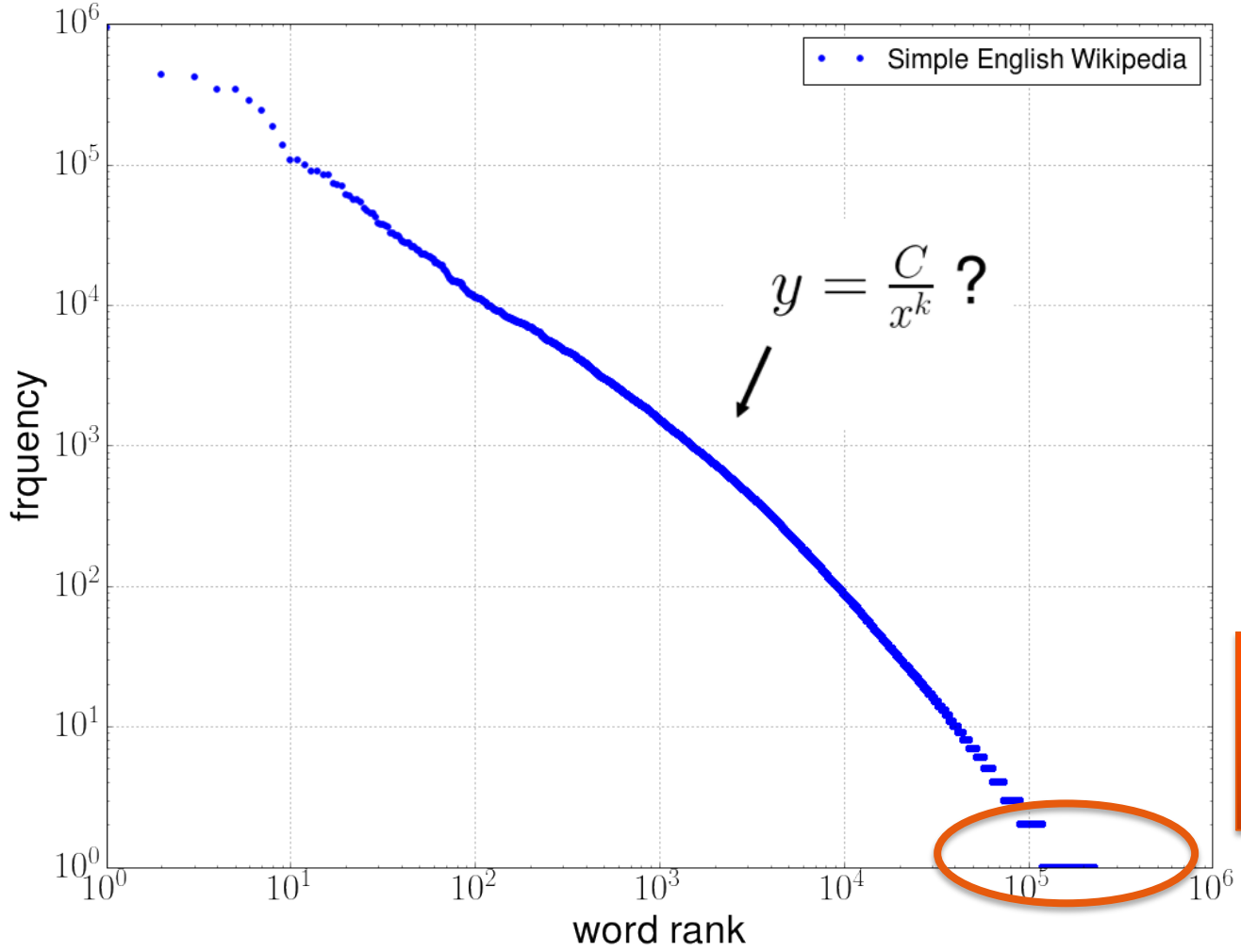
Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

Completing this unit you should

- Know how to transform a rank frequency diagram to a powerlaw plot.
- Understand how powerlaw and pareto plots relate to each other.
- Be able to explain why a pareto plot is just and inverted rank frequency diagram
- Be able to transform the zipf coefficient to the powerlaw and pareto coefficient and vice versa.
- Understand that building the CDF is basically like building the integral.

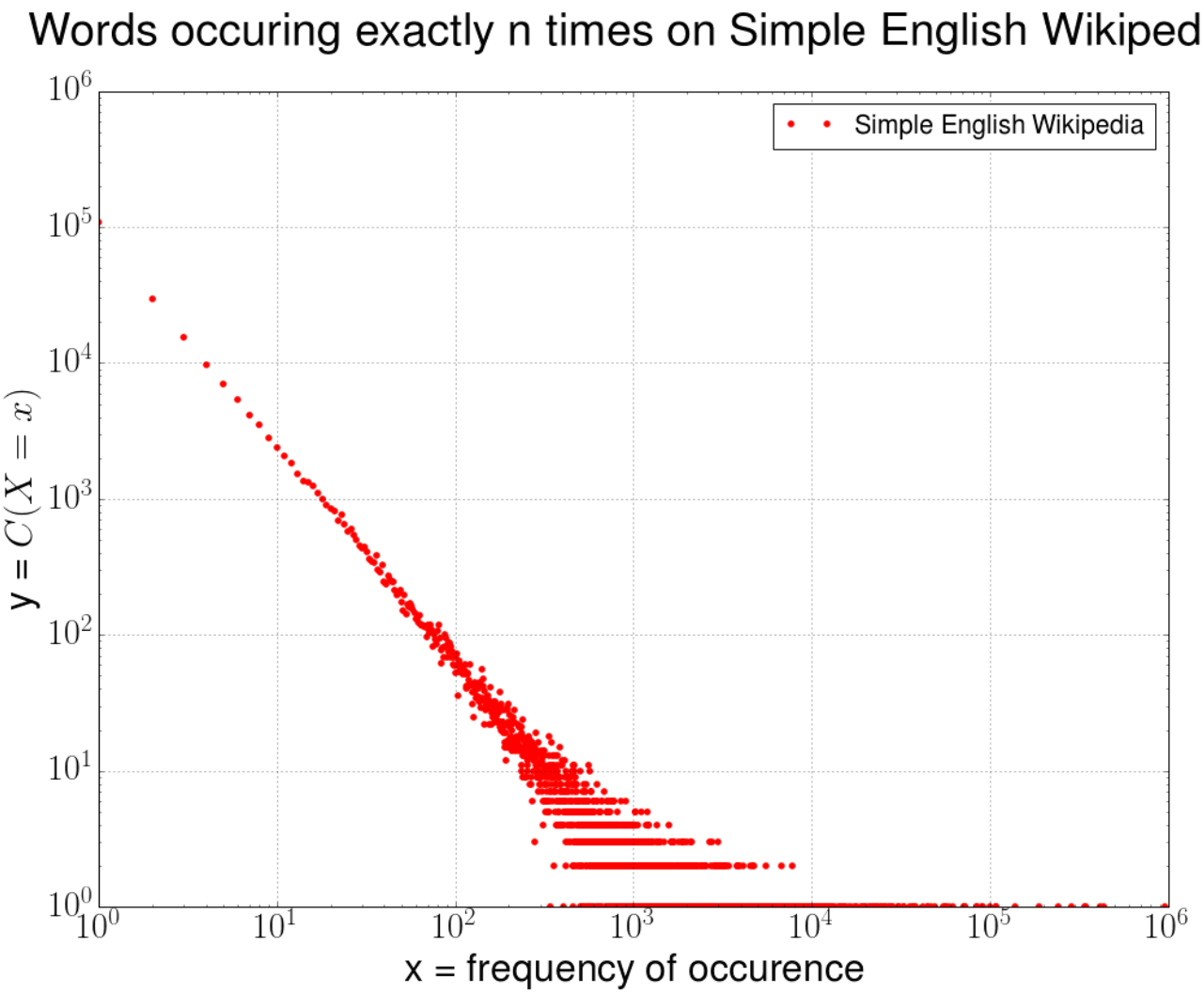
Is there a better way to estimate the zipf Parameter?

Wordrank frequency diagram on Wikipedia data sets (log-log scale)

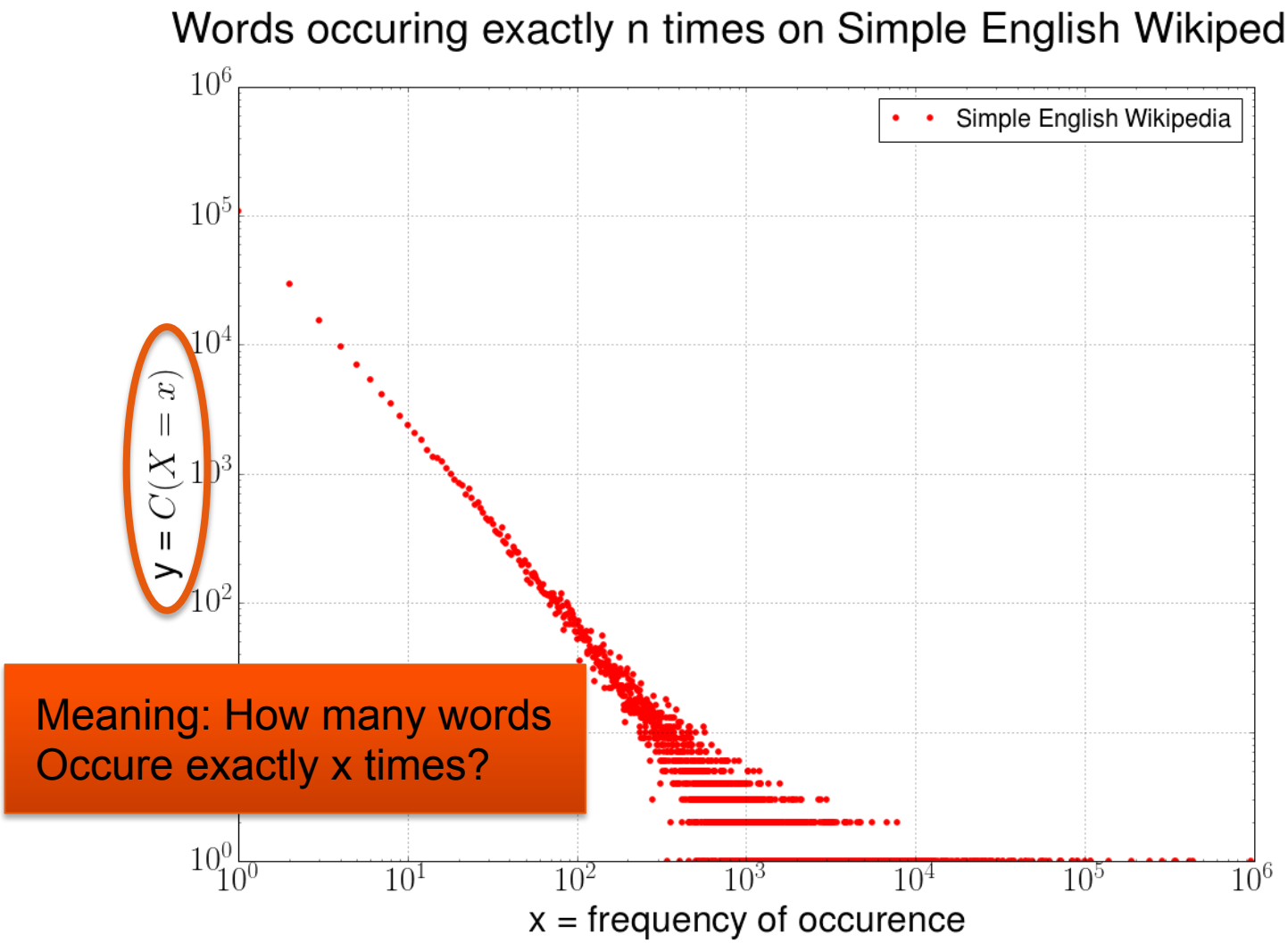


Remember: most words are here

Yet another way of displaying the data: The Power law plot

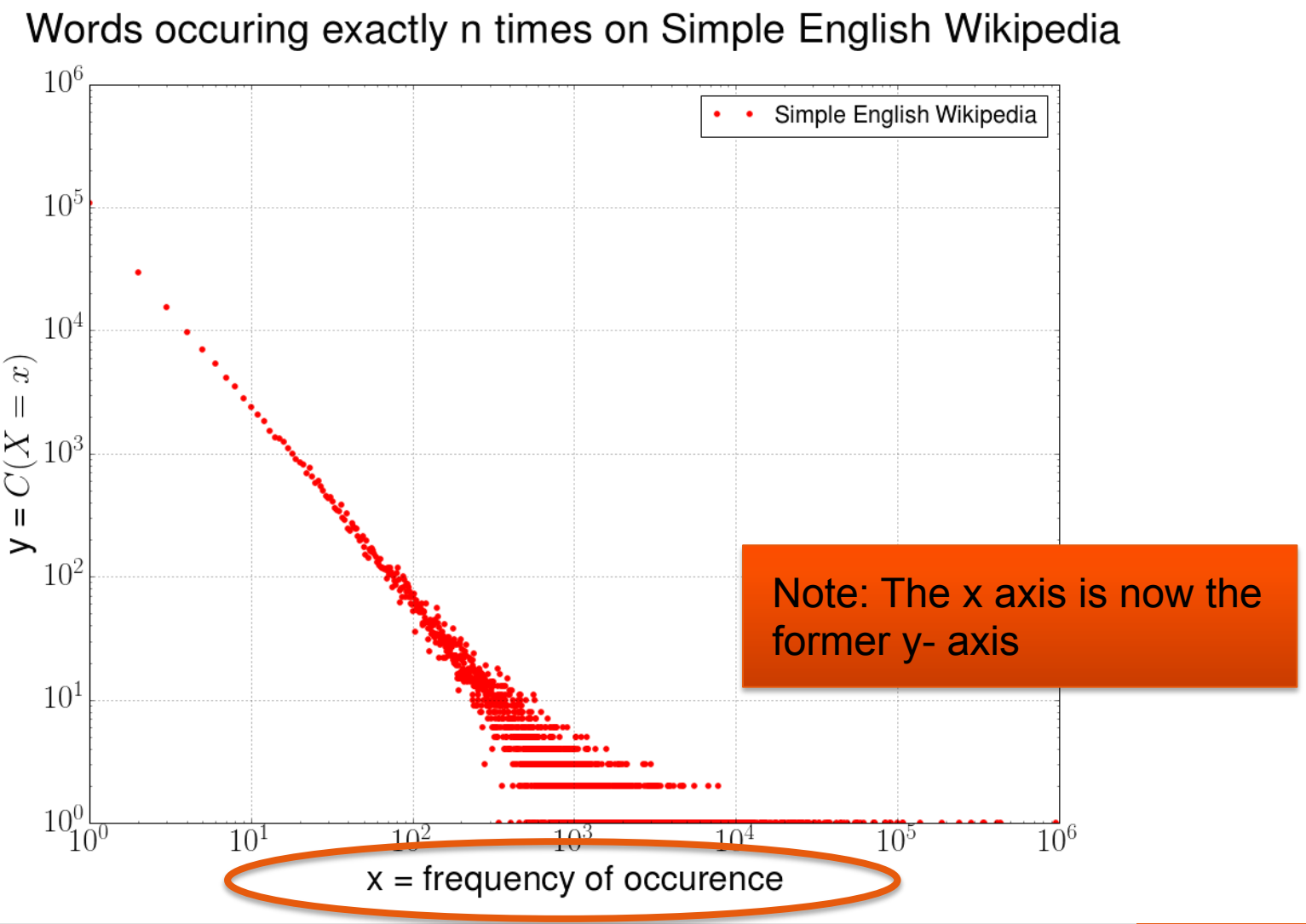


Yet another way of displaying the data: The Power law plot



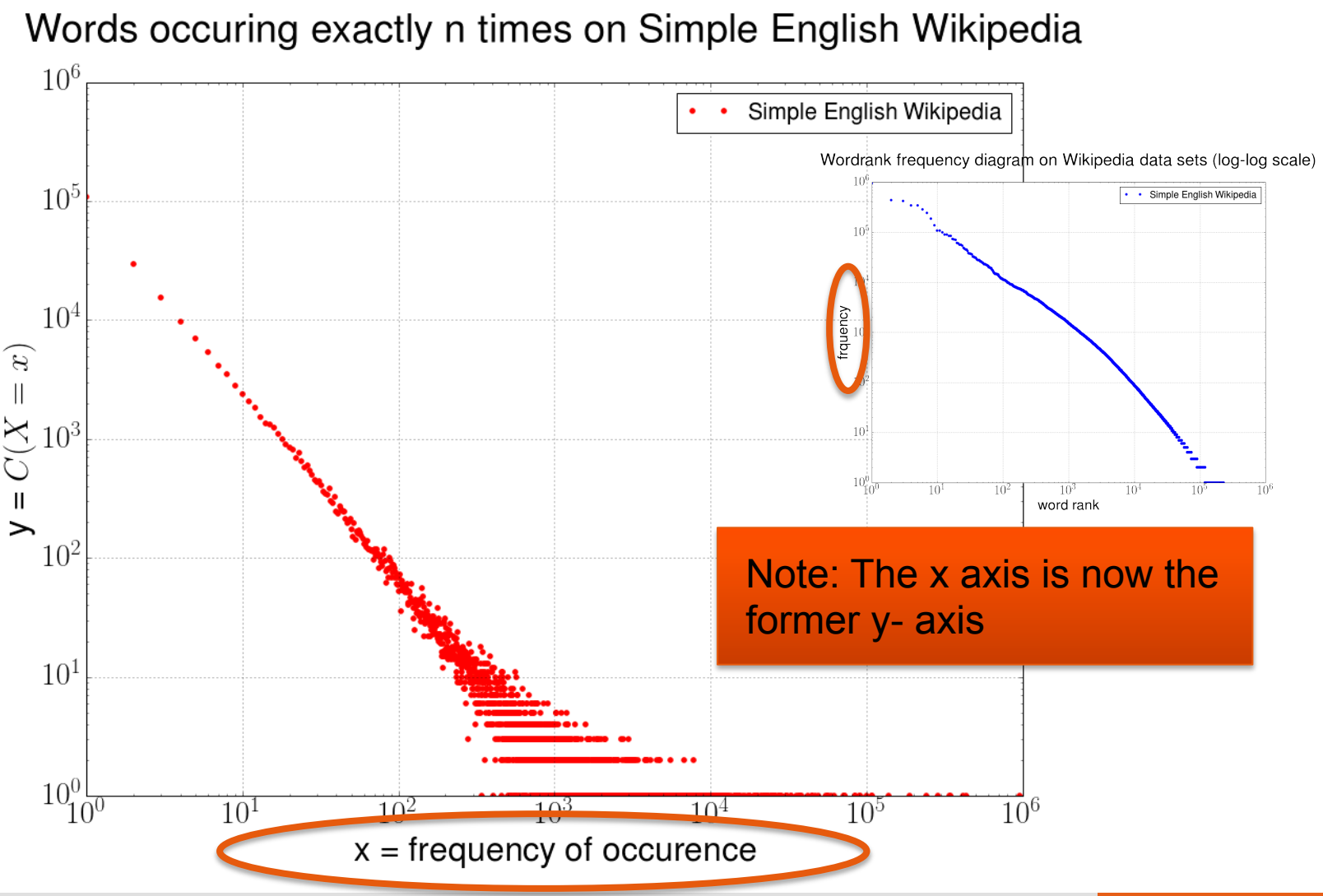
Meaning: How many words Occure exactly x times?

Yet another way of displaying the data: The Power law plot

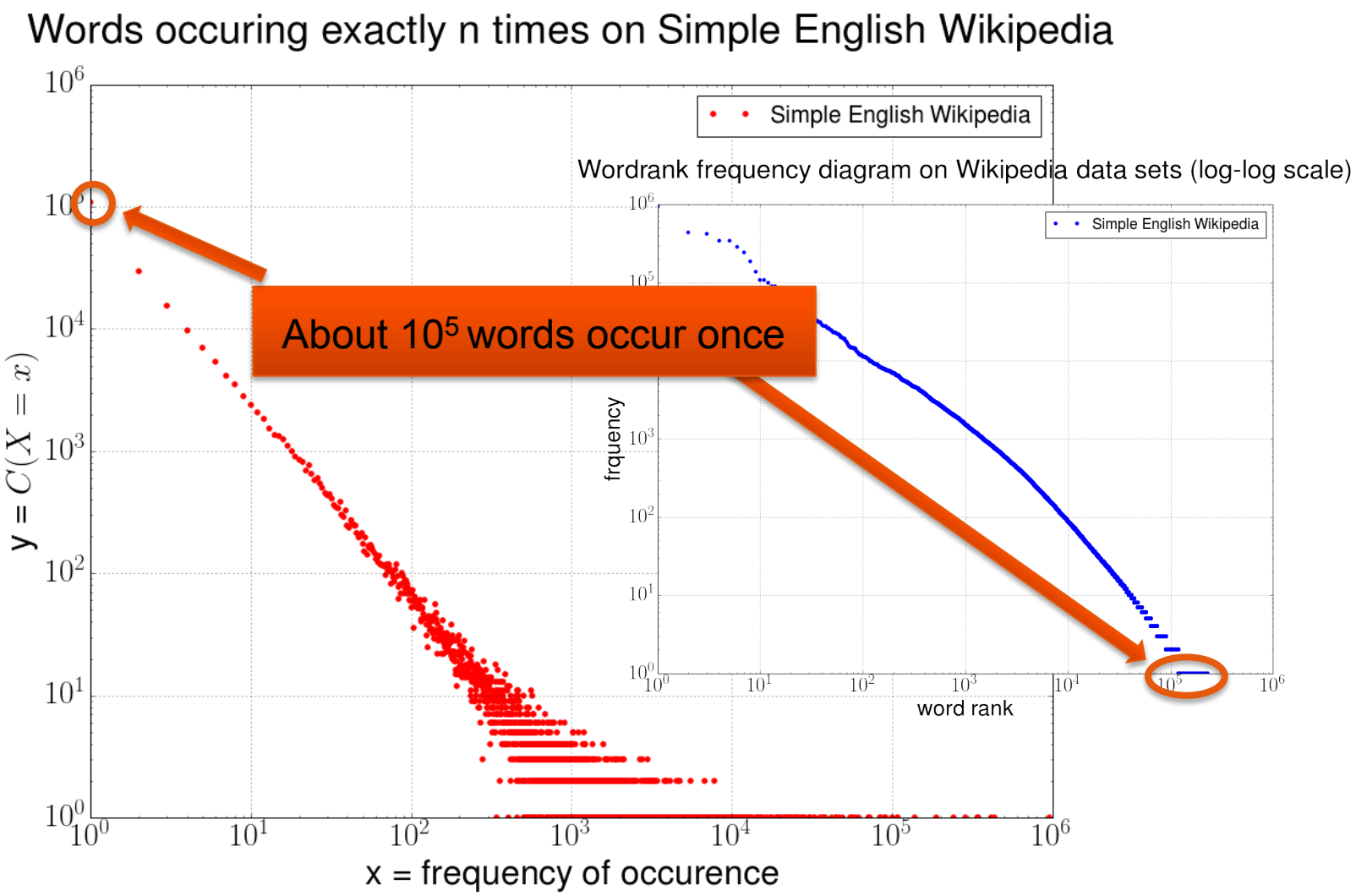


Note: The x axis is now the former y- axis

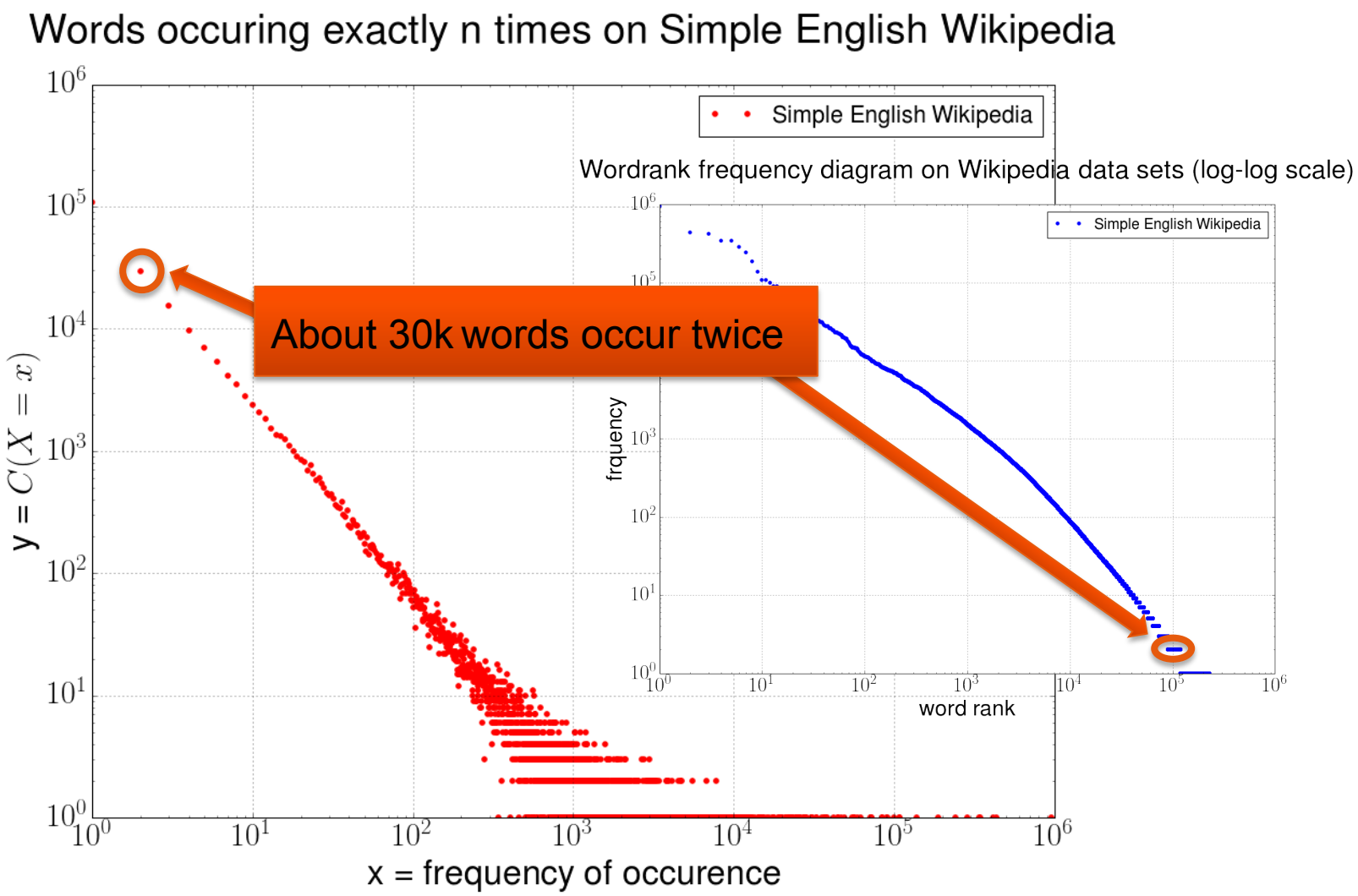
Different visualization but same information



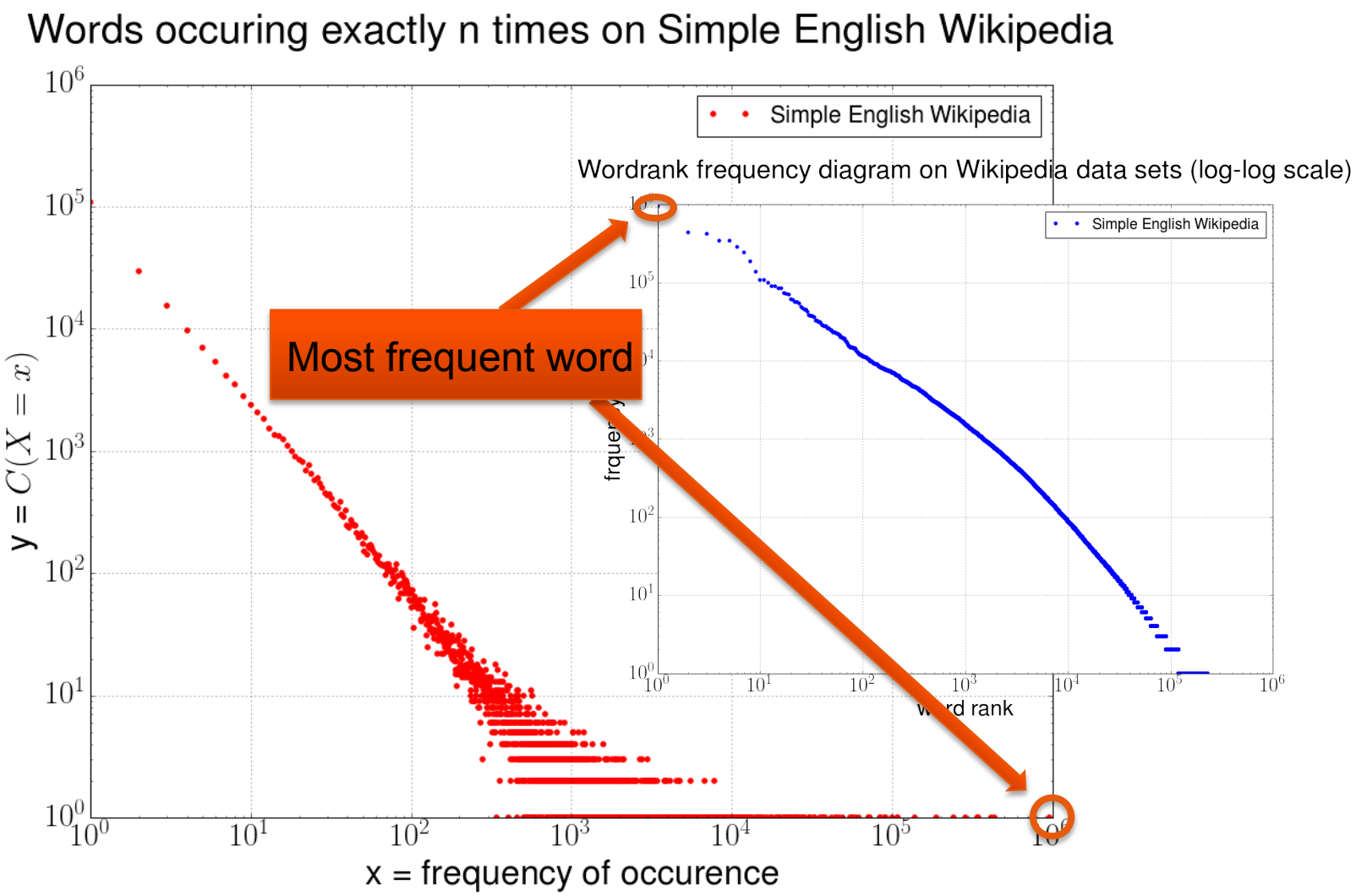
High amount of words occur only once!



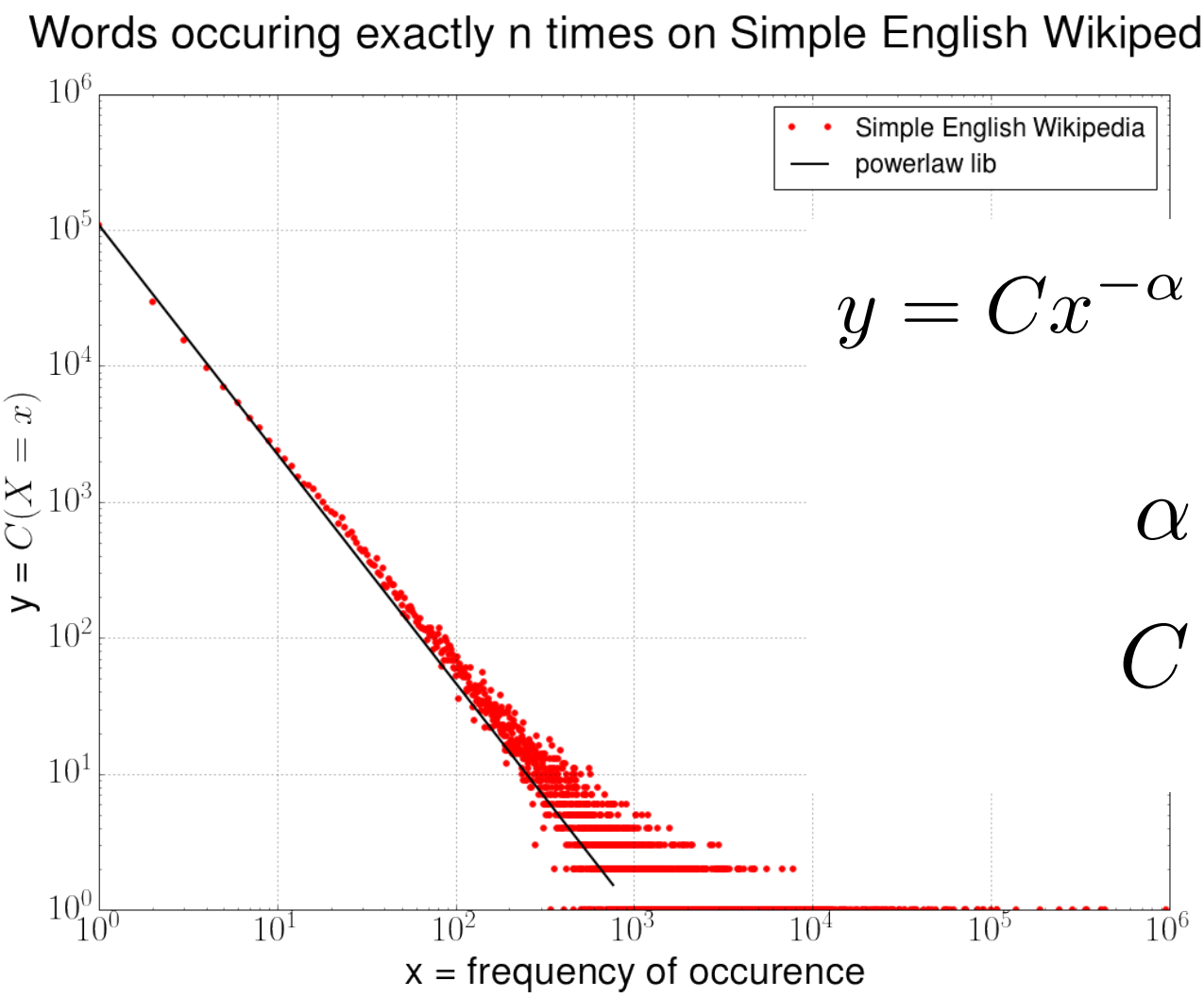
Fewer (but still a lot) words occur twice



Top frequent words



Use a fitting library to estimate alpha



$$y = Cx^{-\alpha} = \frac{C}{x^{\alpha}}$$

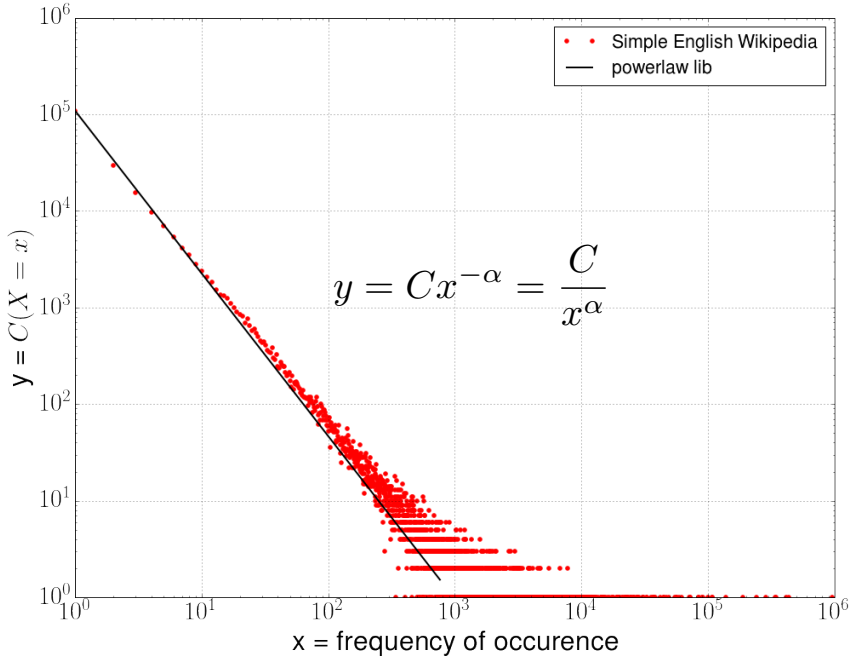
$$\alpha = 1.68$$

$$C = 109061$$

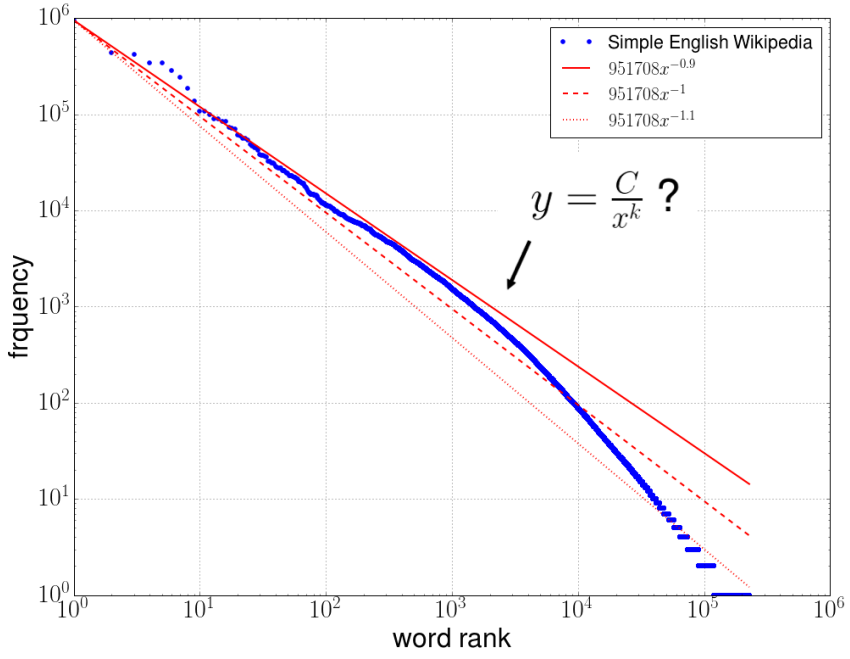
Why did we go from “Zipf” “to Power law”?

- High values for C(X=x) are more stable
- Fitting result is more reliable
- Is there a connection between k and α ?

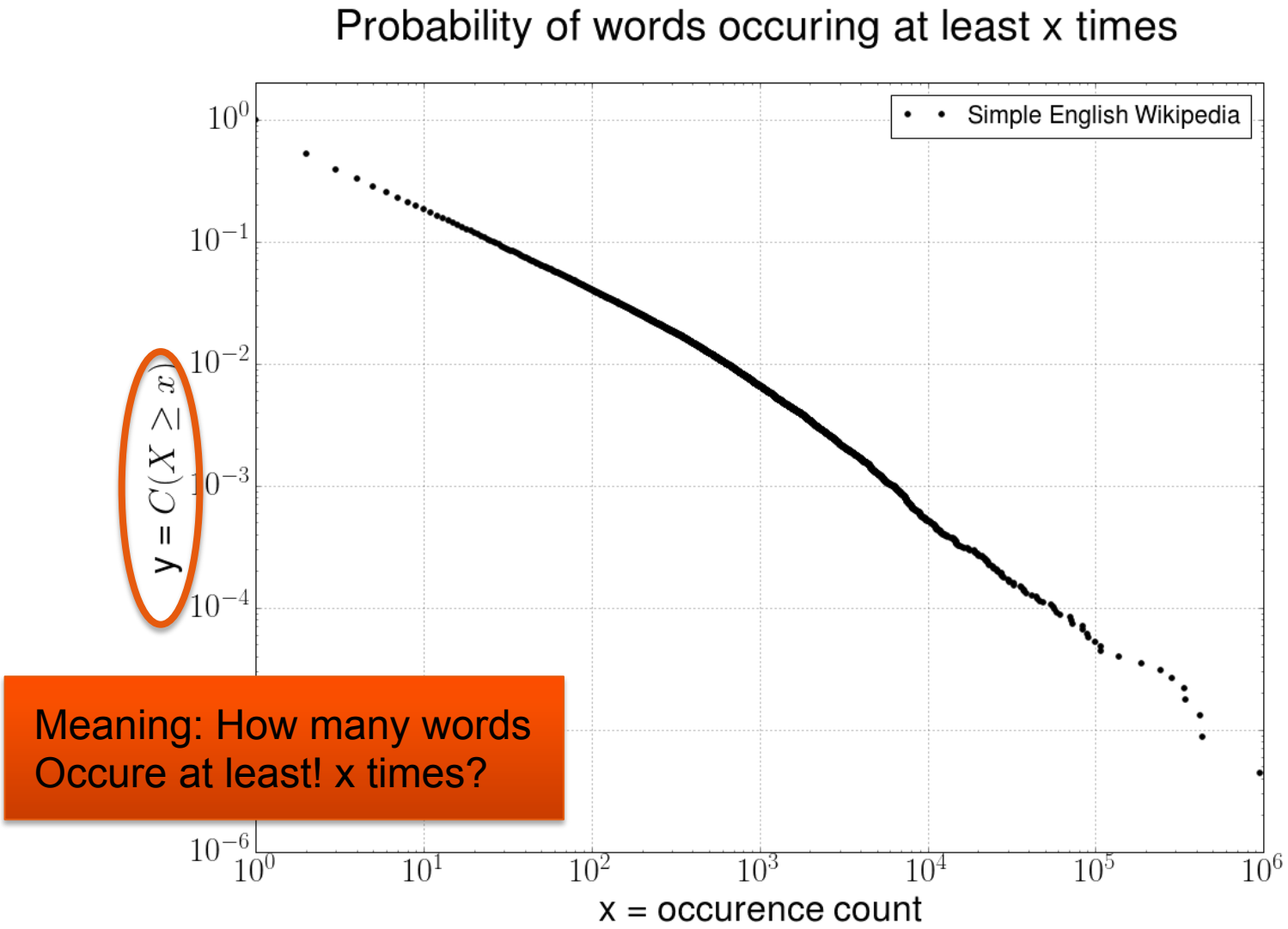
Words occuring exactly n times on Simple English Wikipedia



Wordrank frequency diagram on Wikipedia data sets (log-log scale)



The Pareto plot – Visualizing the same data

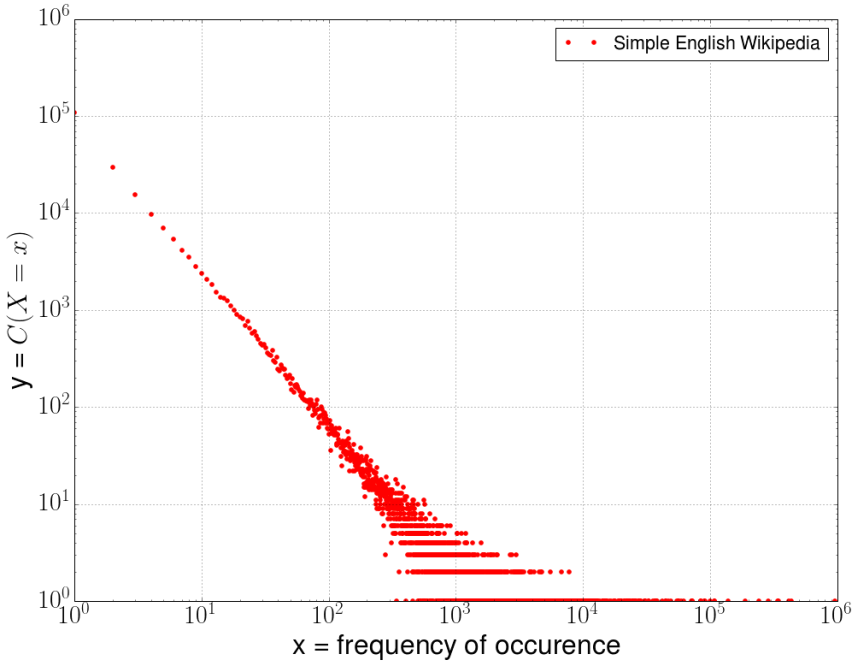


Meaning: How many words Occure at least! x times?

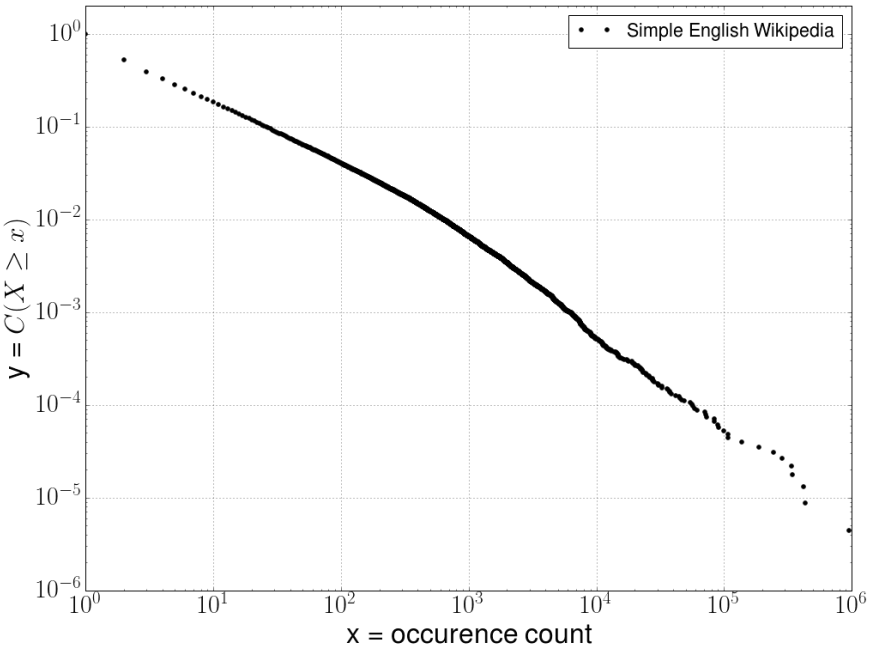
This is basically building an integral

$$p(x) \sim \int_x^\infty pl(r)dr$$

Words occuring exactly n times on Simple English Wikipedia



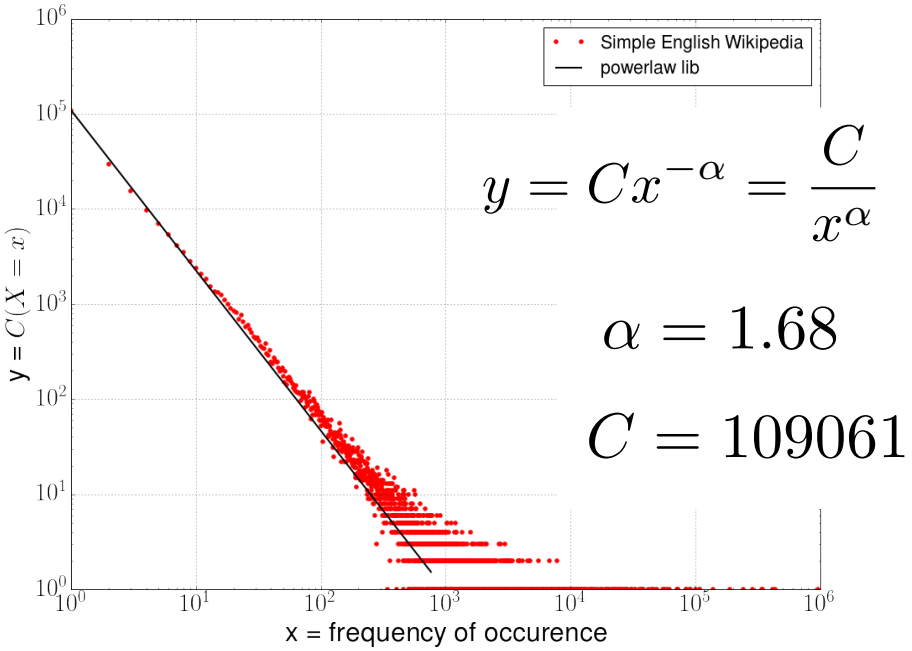
Probability of words occuring at least x times



Get a fit for the new plot by integrating

$$\int \frac{C}{x^{\alpha}} \sim \frac{C'}{x^{\alpha-1}} = \frac{C'}{x^{\gamma}}$$

Words occuring exactly n times on Simple English Wikipedia

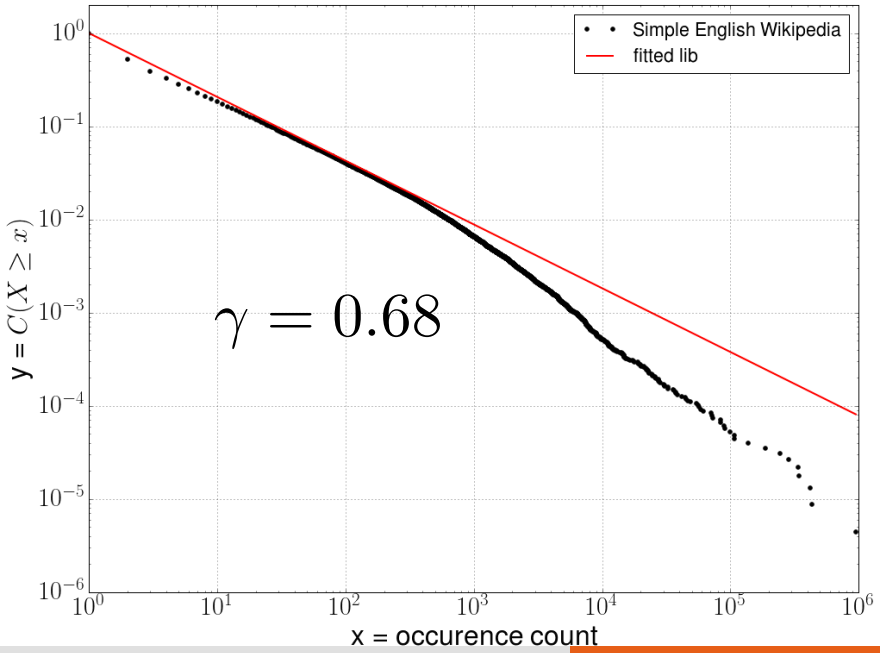


$$y = Cx^{-\alpha} = \frac{C}{x^{\alpha}}$$

$$\alpha = 1.68$$

$$C = 109061$$

Probability of words occuring at least x times

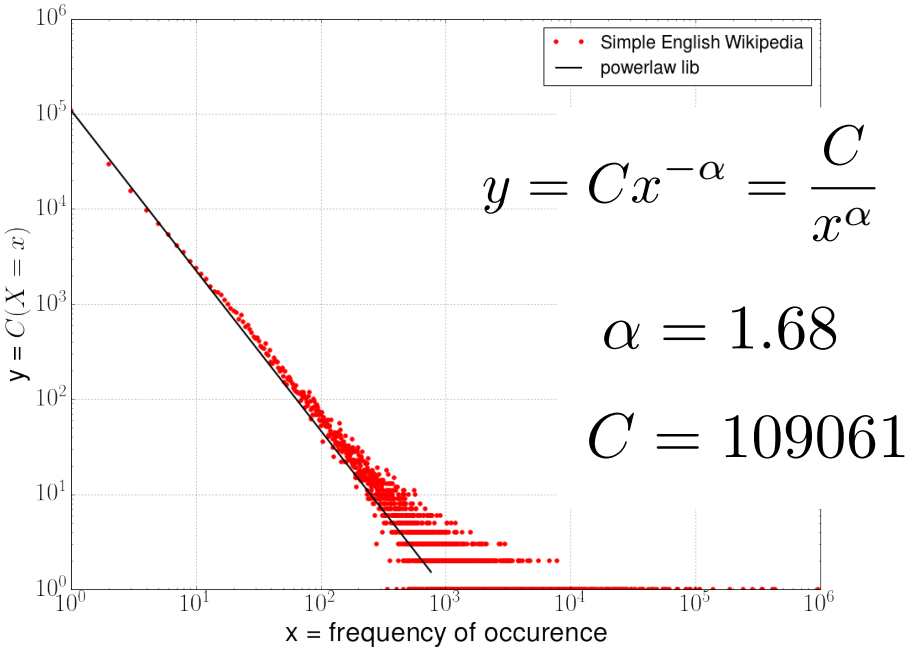


$$\gamma = 0.68$$

Get a fit for the new plot by integrating

$$\int \frac{C}{x^{\alpha}} \sim \frac{C'}{x^{\alpha-1}} = \frac{C'}{x^{\gamma}}$$

Words occuring exactly n times on Simple English Wikipedia

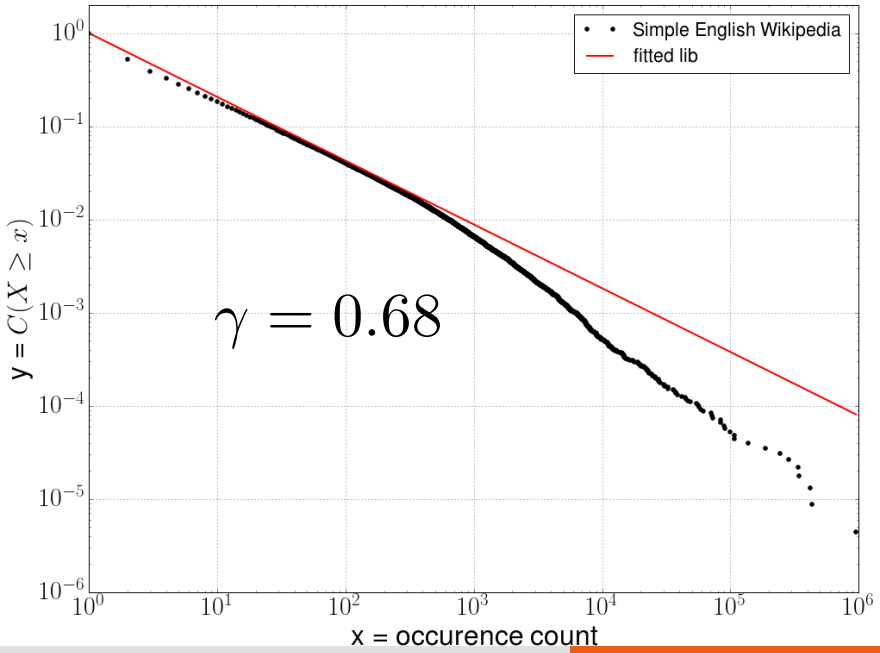


$$y = Cx^{-\alpha} = \frac{C}{x^{\alpha}}$$

$$\alpha = 1.68$$

$$C = 109061$$

Probability of words occuring at least x times

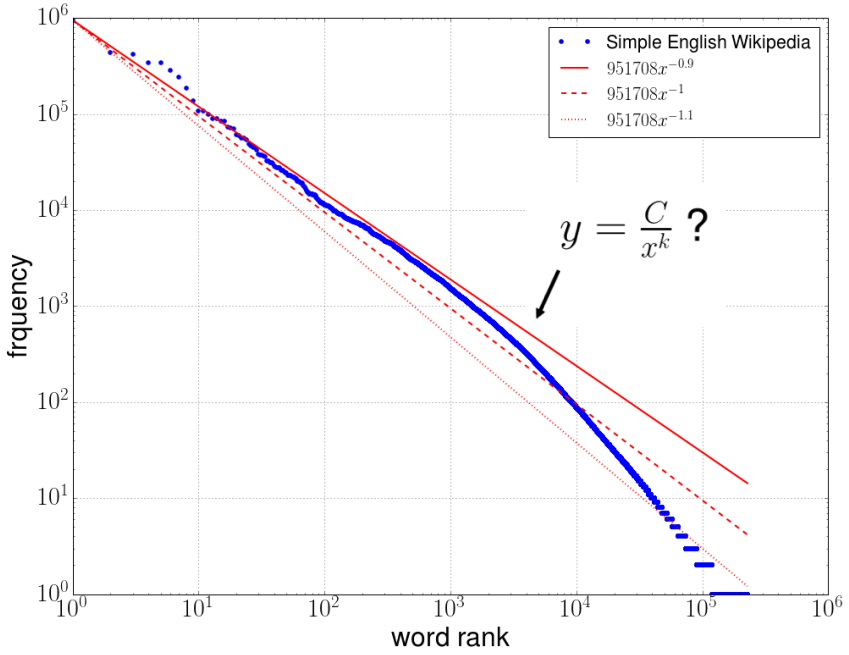


$$\gamma = 0.68$$

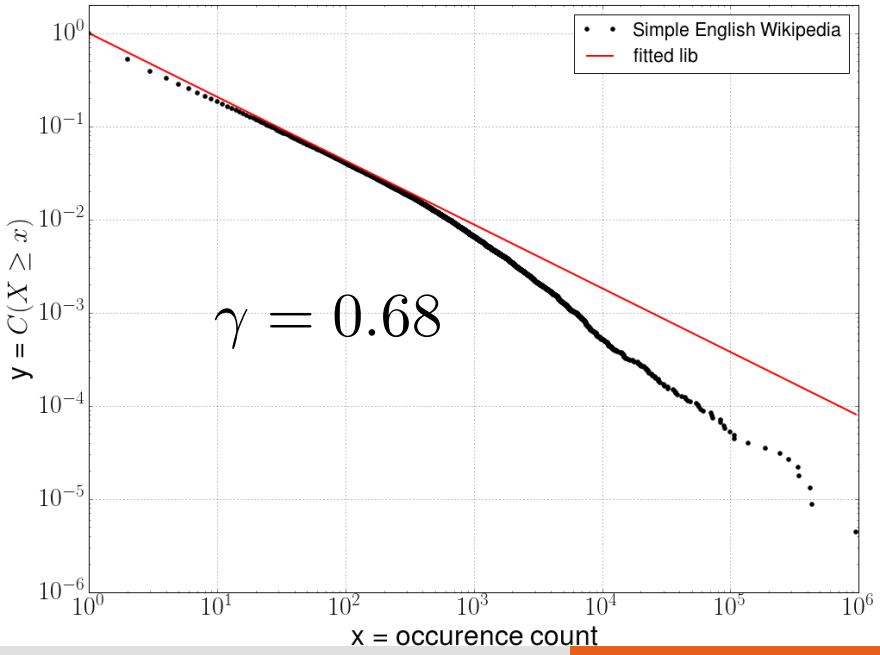
Compare Pareto and Zipf plot

- Pareto plot is a “flipped” Zipf plot
- Pareto has frequency at x axis and Zipf has it at the y-axis
- Vice versa with the rank

Wordrank frequency diagram on Wikipedia data sets (log-log scale)



Probability of words occuring at least x times

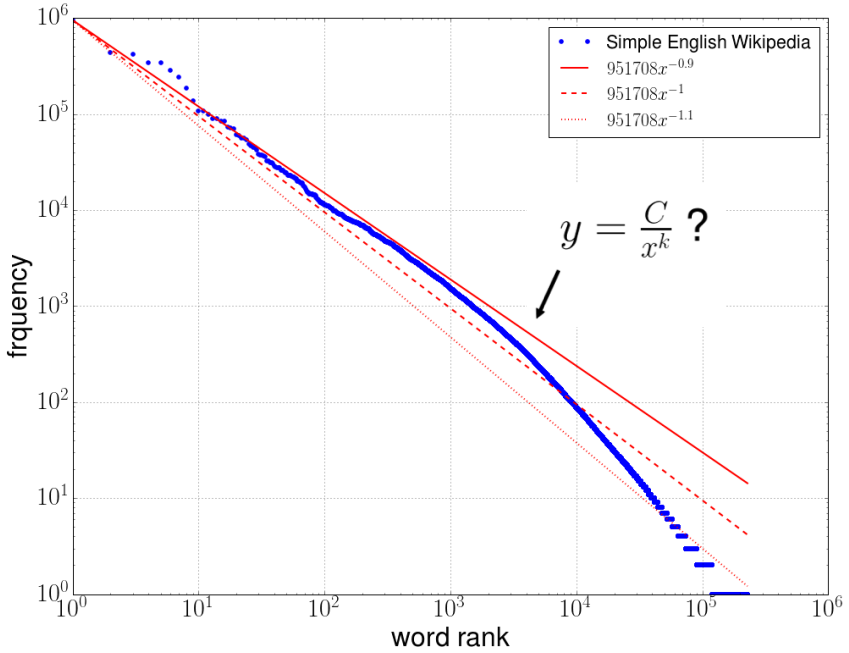


Compare Pareto and Zipf plot

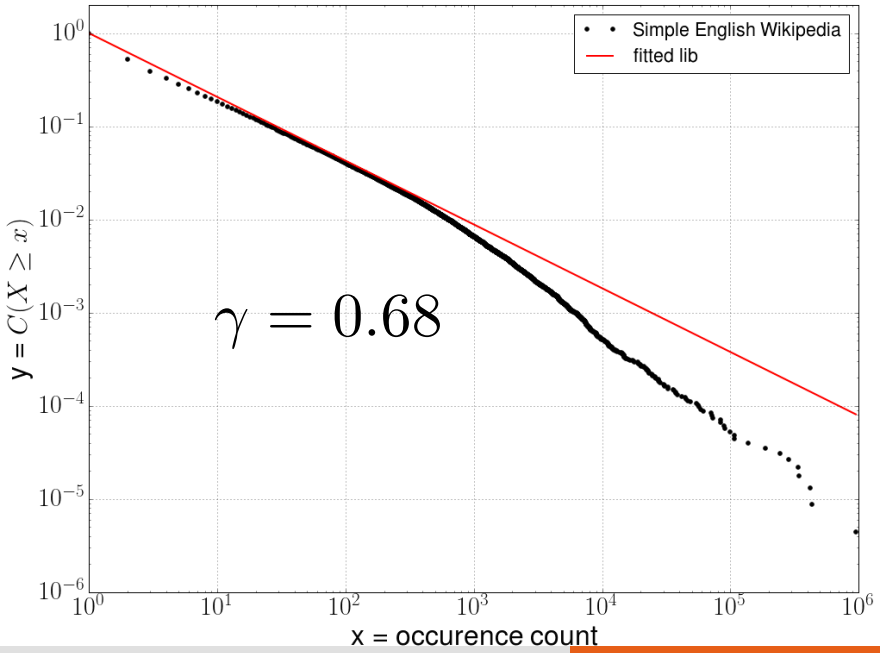
- Pareto plot is a “flipped” Zipf plot

$$k = \frac{1}{\gamma} \Leftrightarrow \gamma = \frac{1}{k}$$

Wordrank frequency diagram on Wikipedia data sets (log-log scale)



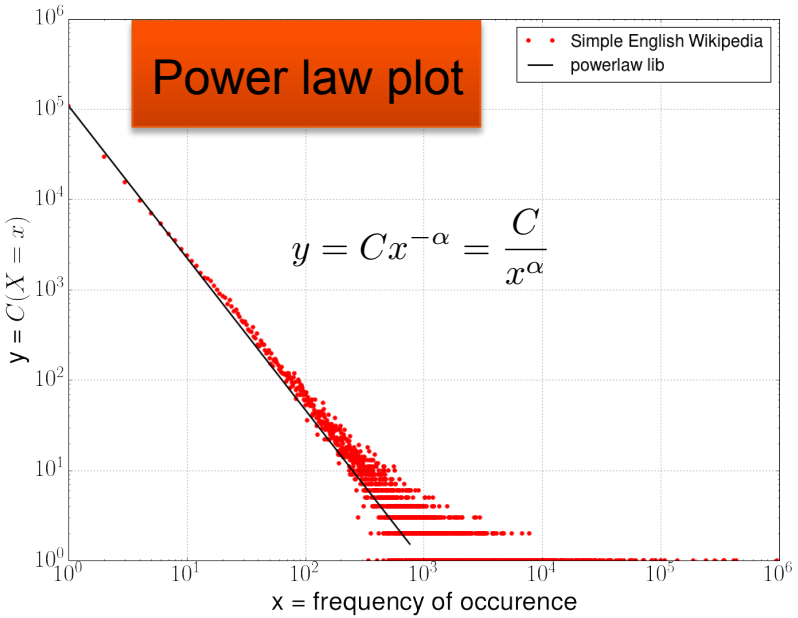
Probability of words occuring at least x times



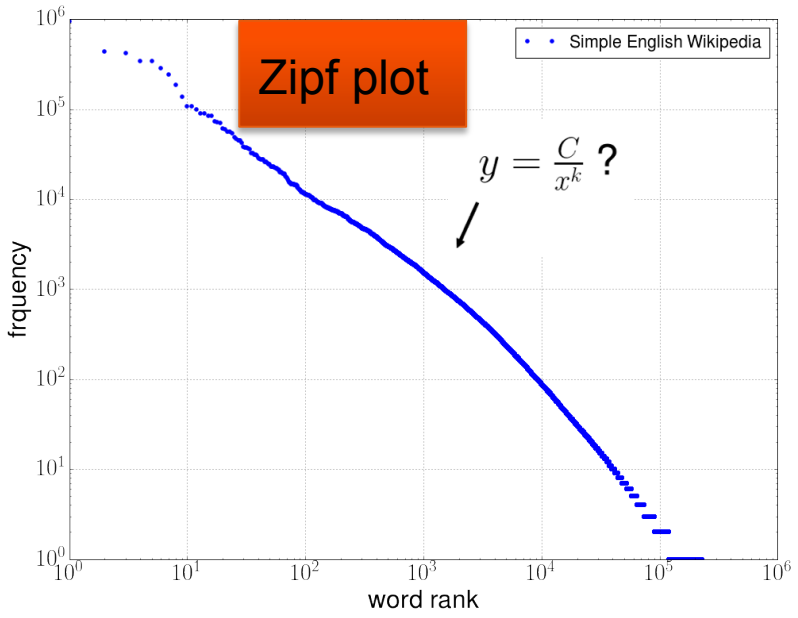
Beware the naming convention

- Both $y_1 = \frac{C_1}{x^\alpha}$ and $y_2 = \frac{C_2}{x^k}$ are power functions
- They are connected with $\alpha = 1 + \frac{1}{k} \Leftrightarrow k = \frac{1}{\alpha - 1}$
- While plotting one is called power law and the other is called Zipf plot

Words occuring exactly n times on Simple English Wikipedia



Wordrank frequency diagram on Wikipedia data sets (log-log scale)



Conclusion

- Pareto, Zipf and powerlaw plots are are equivalent views on the same data
- That is why in practise they are often exchanged (or confused)
- By the end of the day it is all about modelling and carefully reading diagrams
  - This unit should give you a chance to practise and review your skills of working with diagrams
- Why did we always look at the exponent and not at the Constant?
  - Find the answer in the next unit

Thank you for your attention!

Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

Copyright: