Lesson3:
Modelling the Web with Advanced Statistical Descriptive Text Models
Unit3:
Fitting a curve on a (log-log) plot

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST

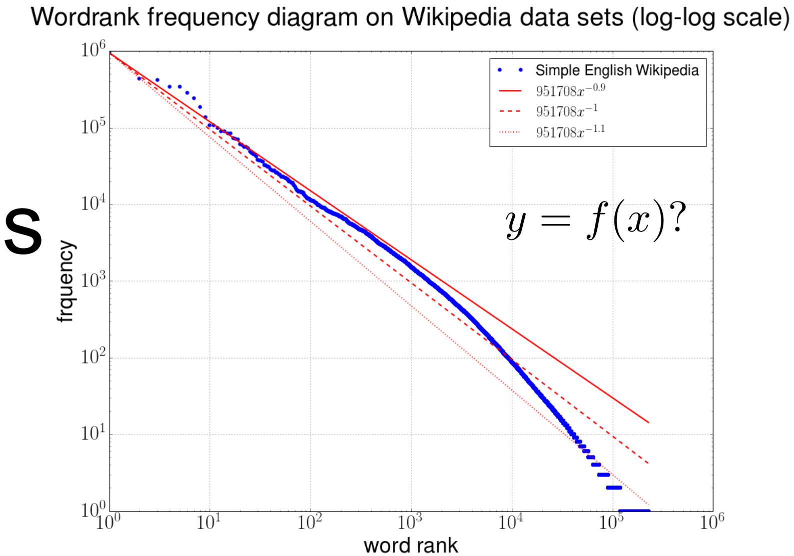
People and Knowledge Networks

Completing this unit you should

- Know the axioms for a distance measure and how they relate to norms.
- Know at least two distance measures on functions spaces.
- Understand why changing to the CDF makes sense when looking at distance between functions.
- Understand the principle of the Kolomogorov-Smirnov test for fitting curves

Can we fit a function to this data?

- On a log log plot the rank / frequency diagram appears roughly as a straight line

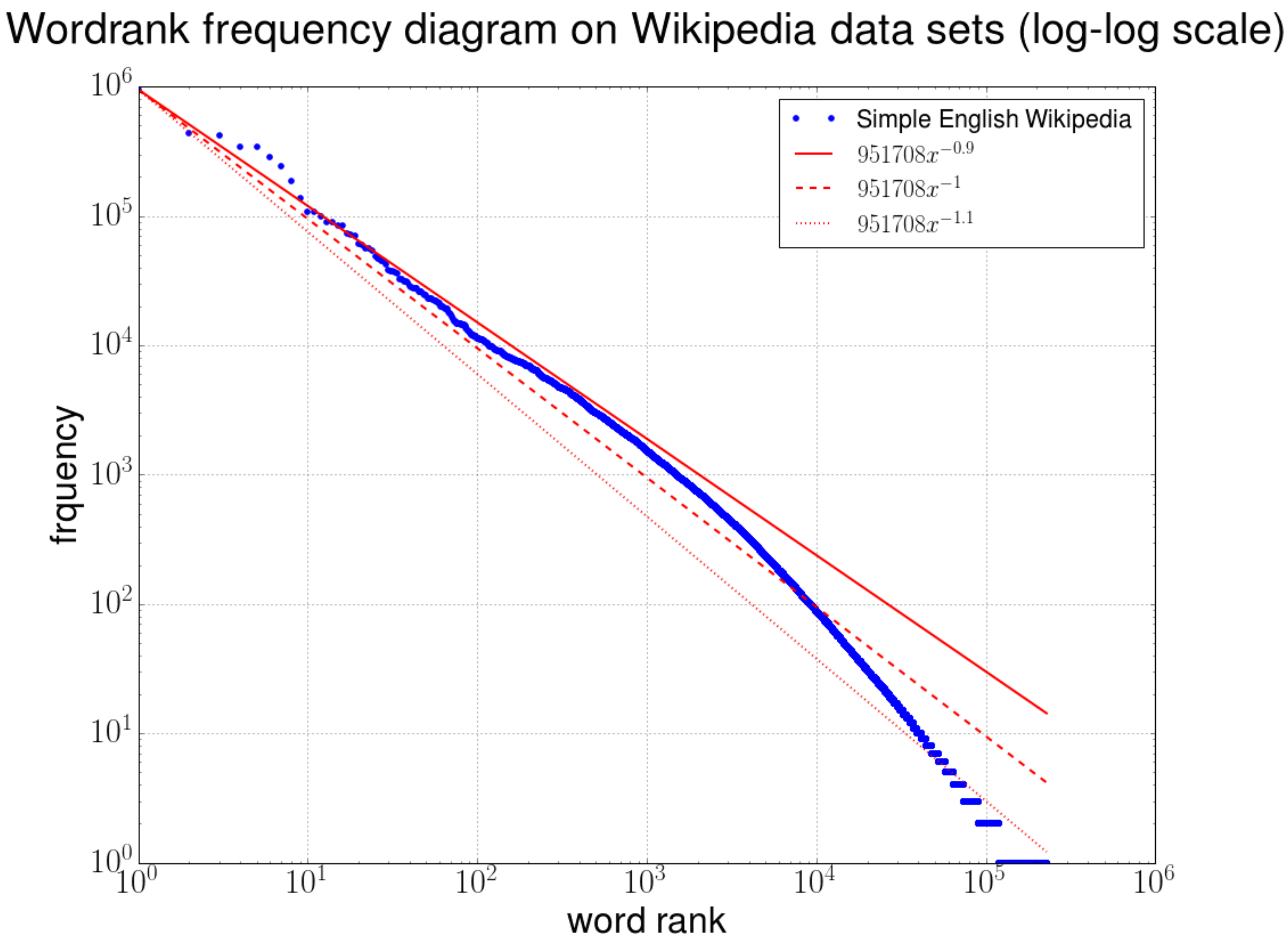


1. Power functions appear as straight lines on log log plots

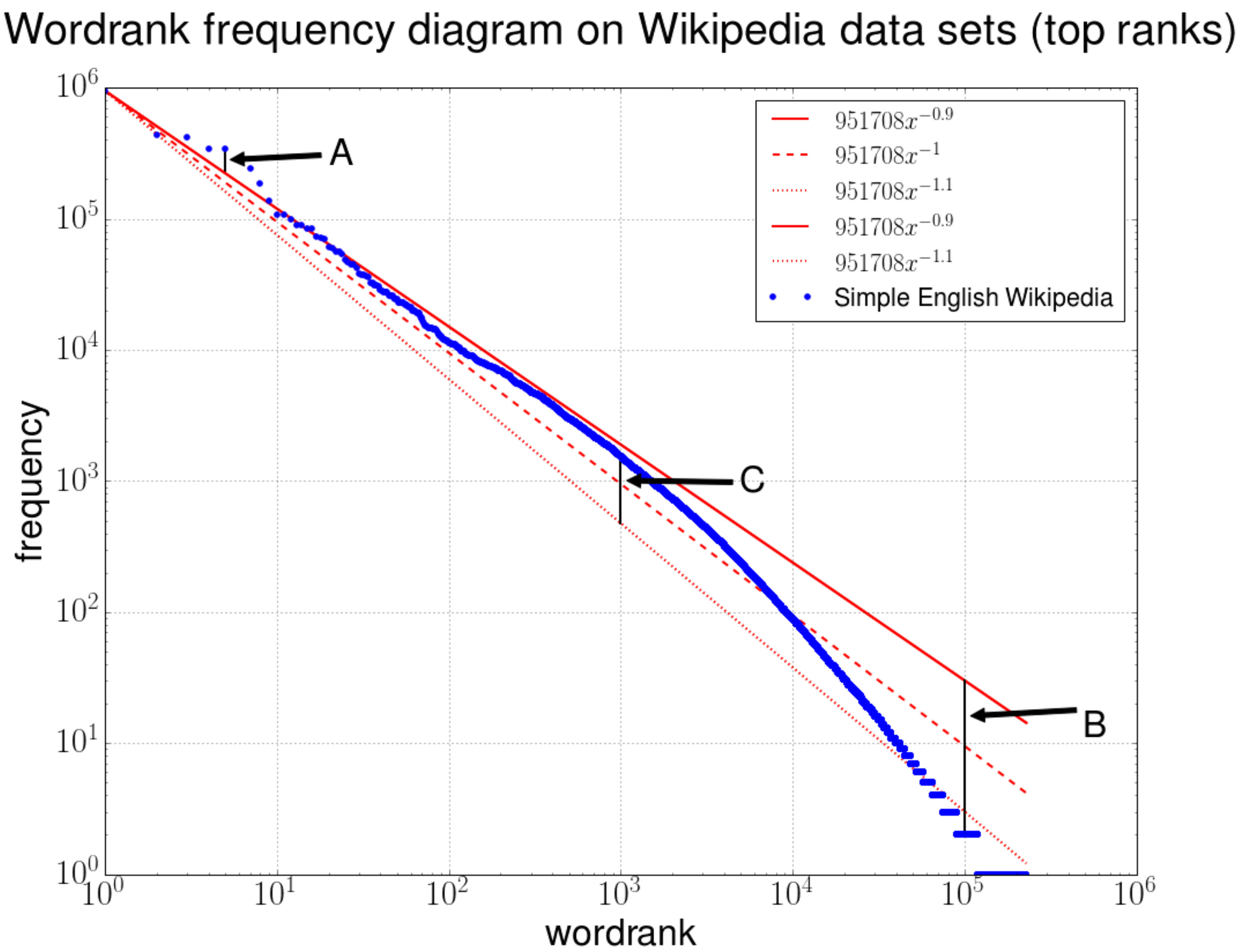
2. Distance of both functions should be smaller than c

$$||f_{fit} - f_{observed}|| < c$$

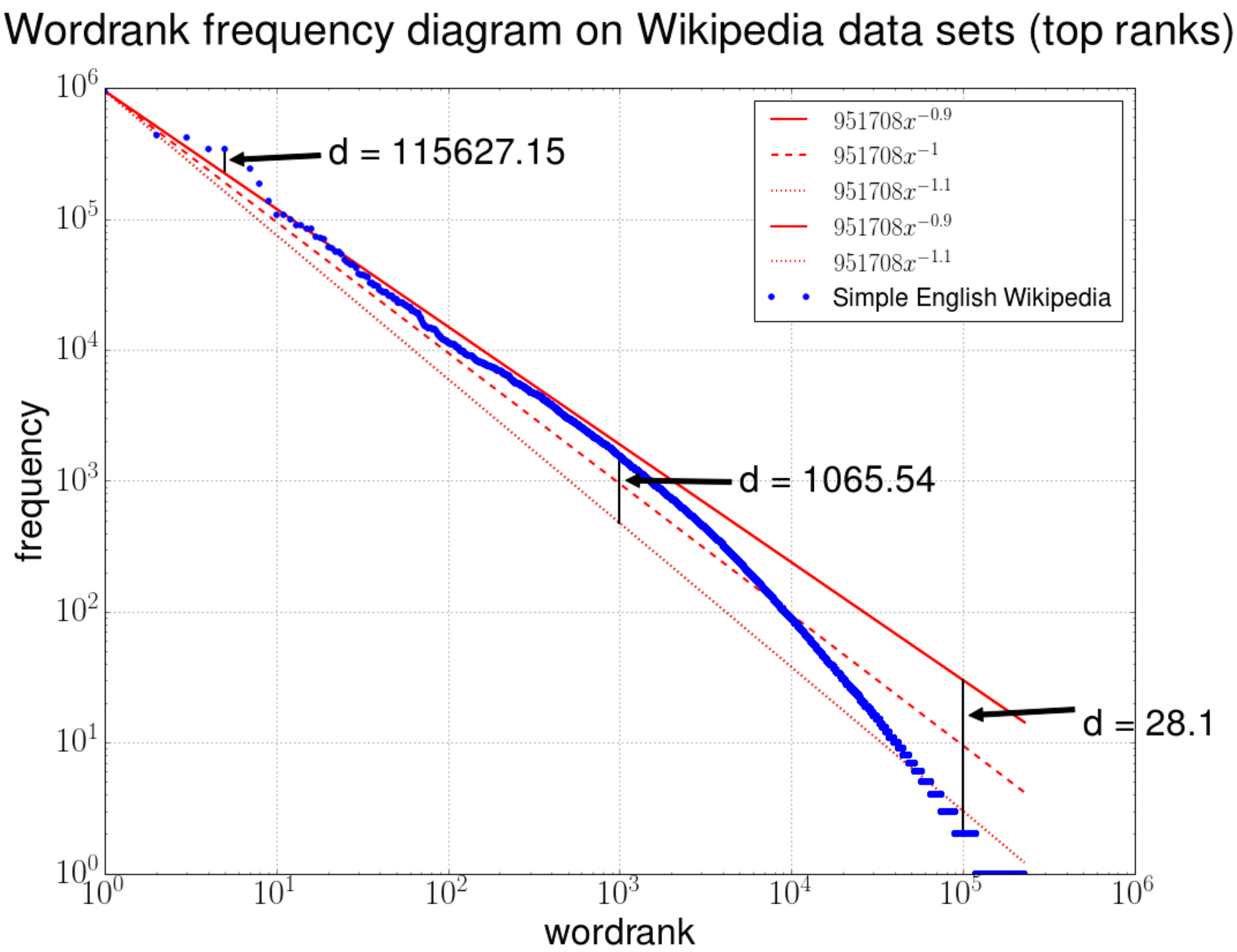
Which curve is fitting the data best? Why?



Which of the black lines is longest – Why?

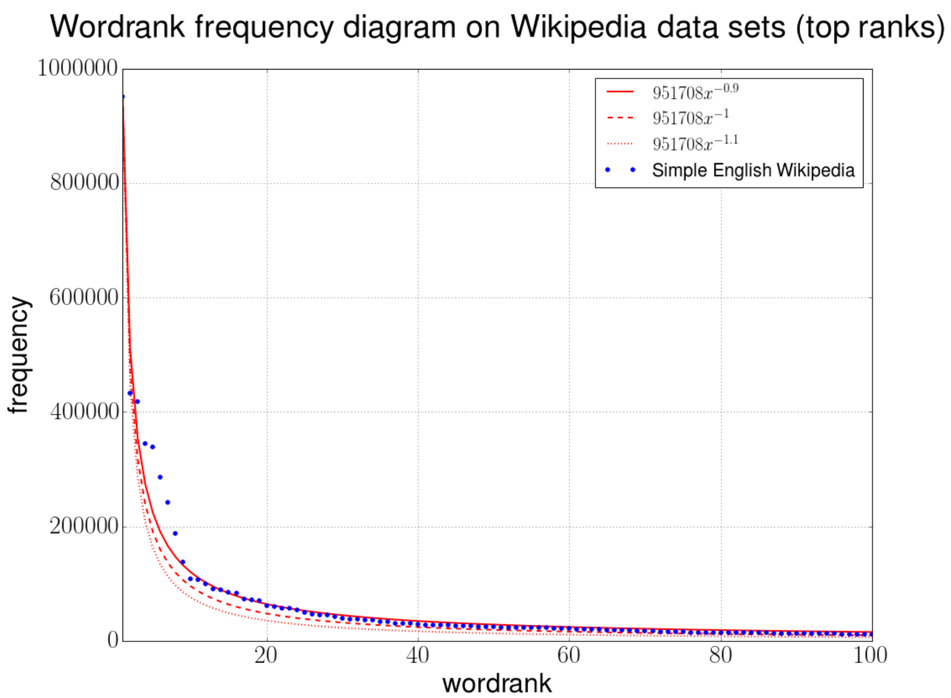


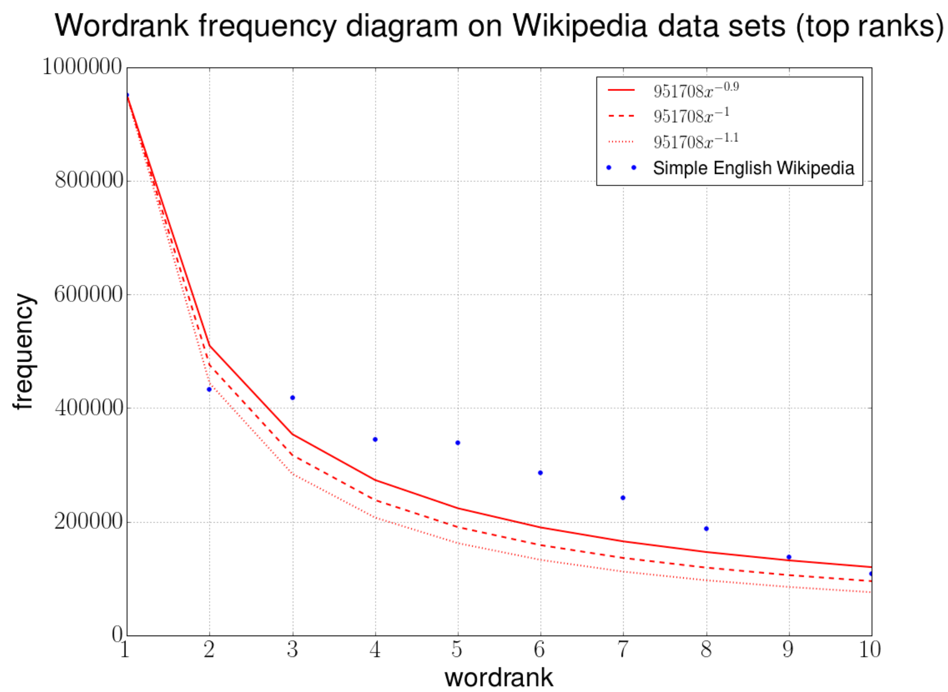
Again! Do not get fooled by log

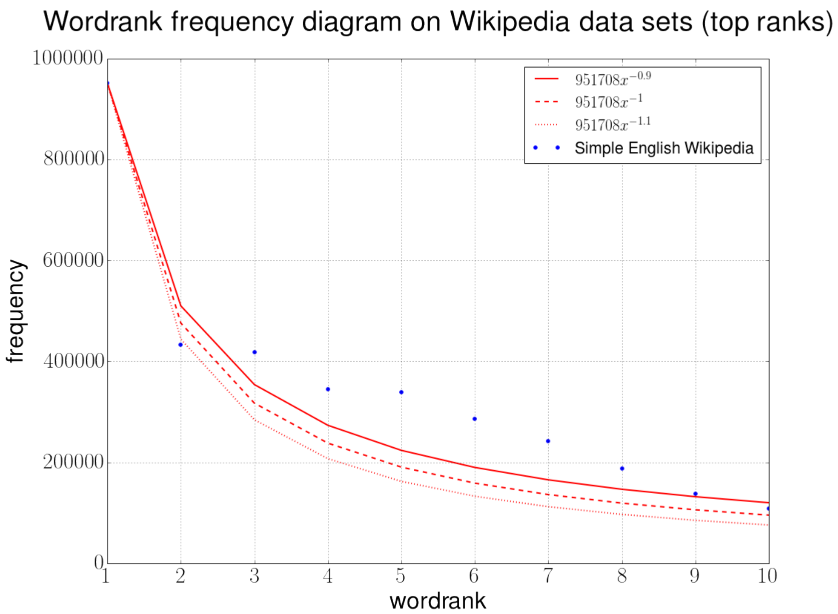


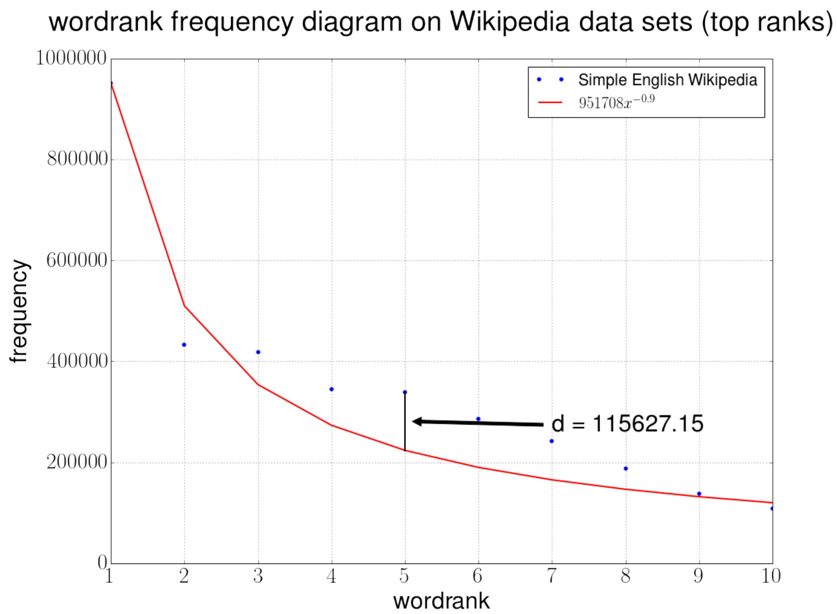
Distances are best seen on linear plots

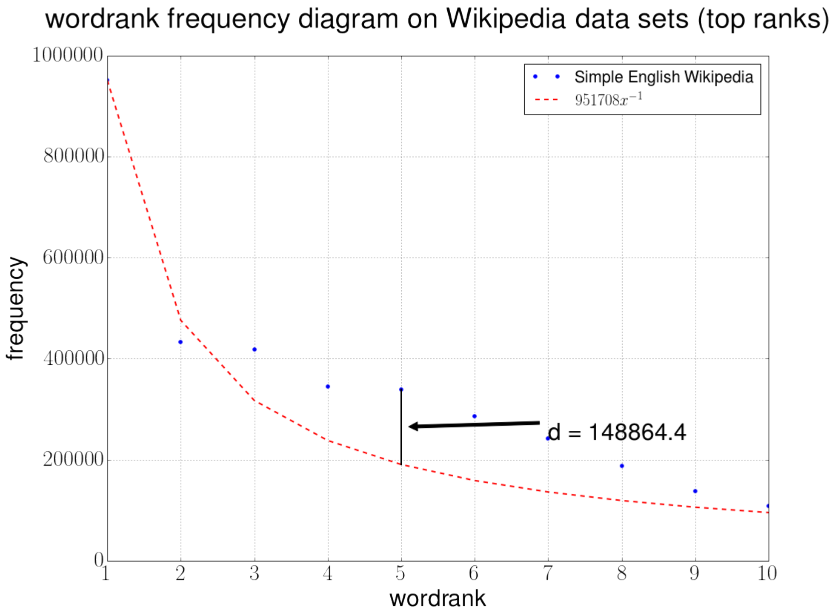
- It makes sense to look at the top ranked words to find the greatest distance

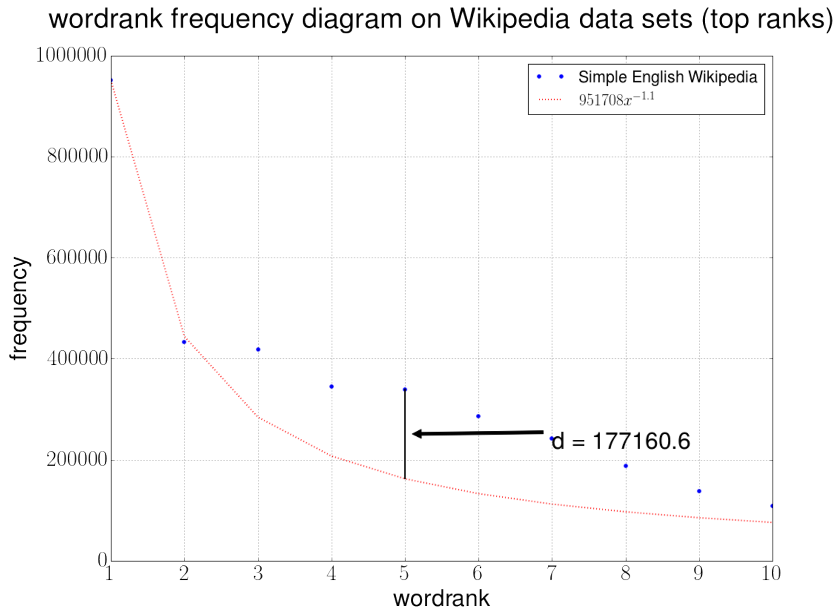





Wordrank frequency diagram on Wikipedia data sets (top ranks)
$951708x^{-0.9}$
$951708x^{-1}$
$951708x^{-1.1}$
Simple English Wikipedia
frequency
wordrank


wordrank frequency diagram on Wikipedia data sets (top ranks)
Simple English Wikipedia
$951708x^{-0.9}$
d = 115627.15
frequency
wordrank


wordrank frequency diagram on Wikipedia data sets (top ranks)
Simple English Wikipedia
$951708x^{-1}$
d = 148864.4
frequency
wordrank


wordrank frequency diagram on Wikipedia data sets (top ranks)
Simple English Wikipedia
$951708x^{-1.1}$
d = 177160.6
frequency
wordrank

Let's be a little more systematic

- We looked at maximum point wise distance
- We used this as a distance measure between functions
- Are there others / better distance measures for functions?
- How can distance measures be characterized anyway?

How to define distance between functions?

- Recall our goal:
  - Find f_{fit} such that $||f_{fit} - f_{observed}|| < c$

- We define the distance of two functions as:

$$d(f_{fit}, f_{observed}) := ||\underbrace{f_{fit} - f_{observed}}_{g}||$$

- But how to calculate $||g||$ for some function?

The uniform norm (aka sup norm)

- Let $f : M \longrightarrow \mathbb{R}$ be a function

$$\|f\|_\infty := \sup_{x \in M} \|f(x)\|_{\mathbb{R}} = \sup\{\,|f(x)| : x \in M\,\}$$

- $\|f\|_\infty$ is a norm i.e. it has the following properties
  - Positive definite
  - Homogeneous
  - Triangle inequality

Positive definite ($\|f\|_\infty = 0 \Rightarrow f = 0$)

- Let $f : M \longrightarrow \mathbb{R}$ be a function

$$\|f\|_\infty := \sup_{x \in M} \|f(x)\|_\mathbb{R}$$

Proof:

$$\|f\|_\infty = 0 \Leftrightarrow \sup_{x \in M} \|f(x)\|_\mathbb{R} = 0$$

$$\Rightarrow \|f(x)\|_\mathbb{R} = 0 \, \forall x$$

$$\Rightarrow f(x) = 0 \, \forall \qquad \Rightarrow \qquad f = 0$$

Homogeneous ($\|\alpha f\|_\infty = \alpha\|f\|_\infty, \alpha \in \mathbb{R}$)

- Let $f : M \longrightarrow \mathbb{R}$ be a function

$$\|f\|_\infty := \sup_{x \in M} \|f(x)\|_\mathbb{R} = \sup\{\,|f(x)| : x \in M\,\}$$

- Proof:

$$\begin{aligned}\|\alpha f\|_\infty &= \sup_{x \in M} \|\alpha f(x)\|_\mathbb{R} \\ &= \sup_{x \in M} |\alpha|\|f(x)\|_\mathbb{R} \\ &= |\alpha| \sup_{x \in M} \|f(x)\|_\mathbb{R} = |\alpha|\,\|f\|_\infty\end{aligned}$$

Triangle inequality $\|f+g\|_\infty \le \|f\|_\infty + \|g\|_\infty$

$$\|f+g\|_\infty = \sup_{x\in M} \|f(x)+g(x)\|_{\mathbb{R}}$$

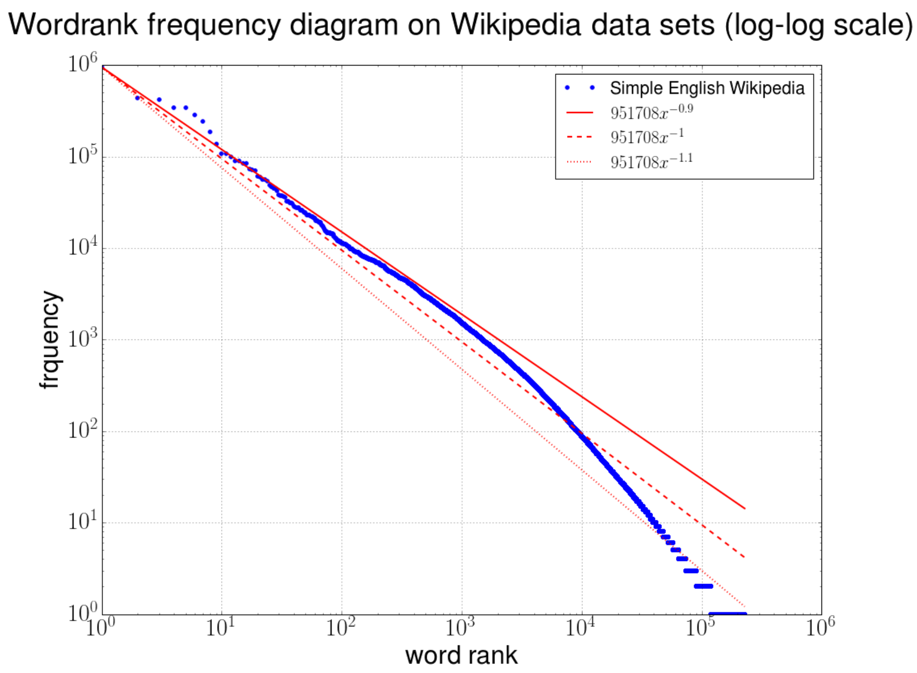
$$\le \sup_{x\in M} \|f(x)\|_{\mathbb{R}} + \|g(x)\|_{\mathbb{R}}$$

$$\le \sup_{x\in M} \|f(x)\|_{\mathbb{R}} + \sup_{x\in M} \|g(x)\|_{\mathbb{R}}$$

$$= \|f\|_\infty + \|g\|_\infty$$

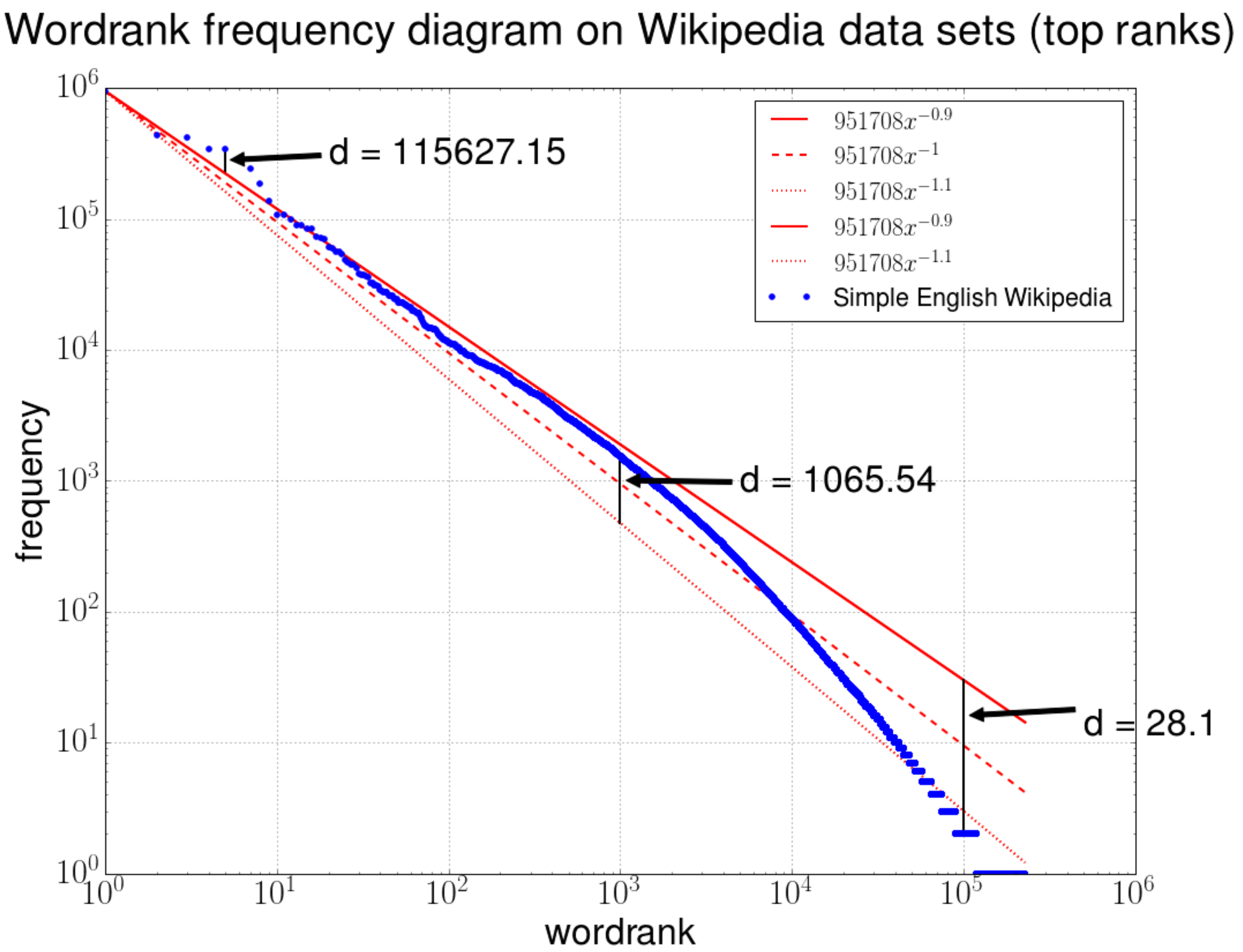
"-0.9" now seems to be the best exponent

f_{fit}	$d(f_{obs}, f_{fit})$
$C/x^{0.9}$	115 k
$C/x^{1.0}$	148 k
$C/x^{1.1}$	177 k

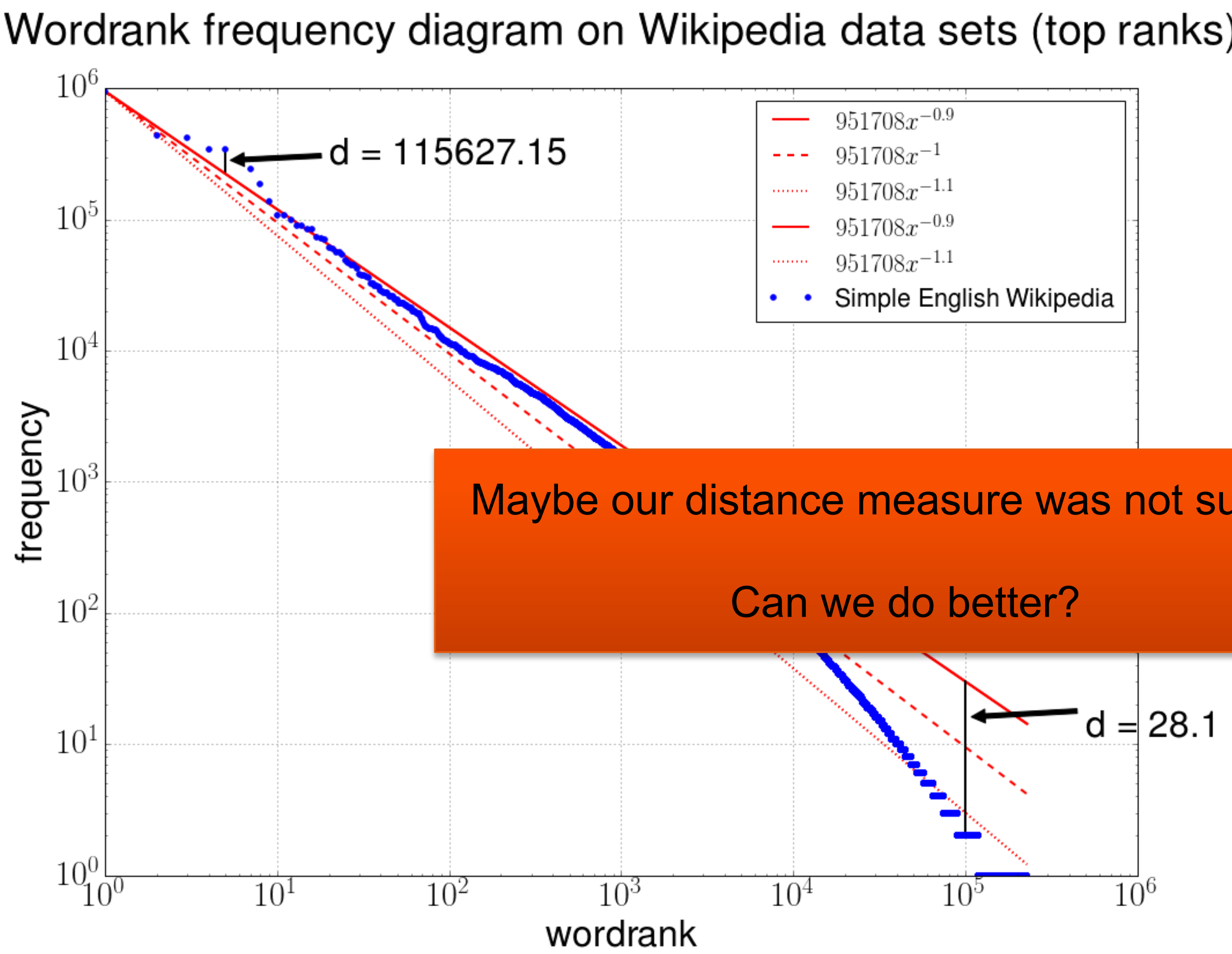


$$C/x^a = C * x^{-a}$$

Biggest distance occurs at rank 5 for all fits



Biggest distance occurs at rank 5 for all fits

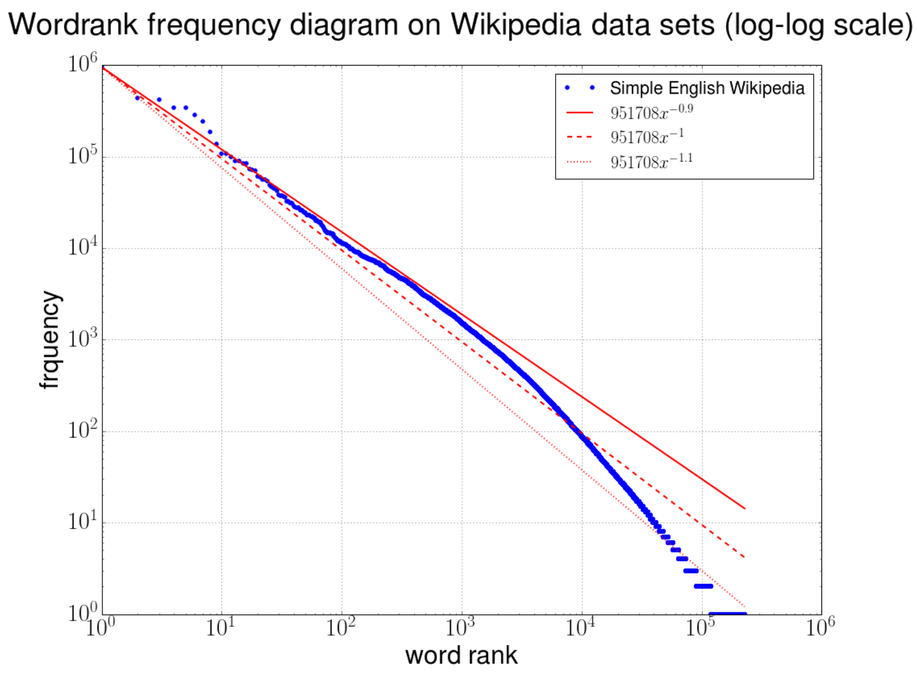


Maybe our distance measure was not suitable

Can we do better?

Problems with our 1^{st} approach

- We measured the largest **point wise** distance between observed data and fit.
- One outlier enough to skew our result
- Millions of low rank distances will not contribute to the result even if they are all off.



L1 Norm (integral) cumulate point wise error

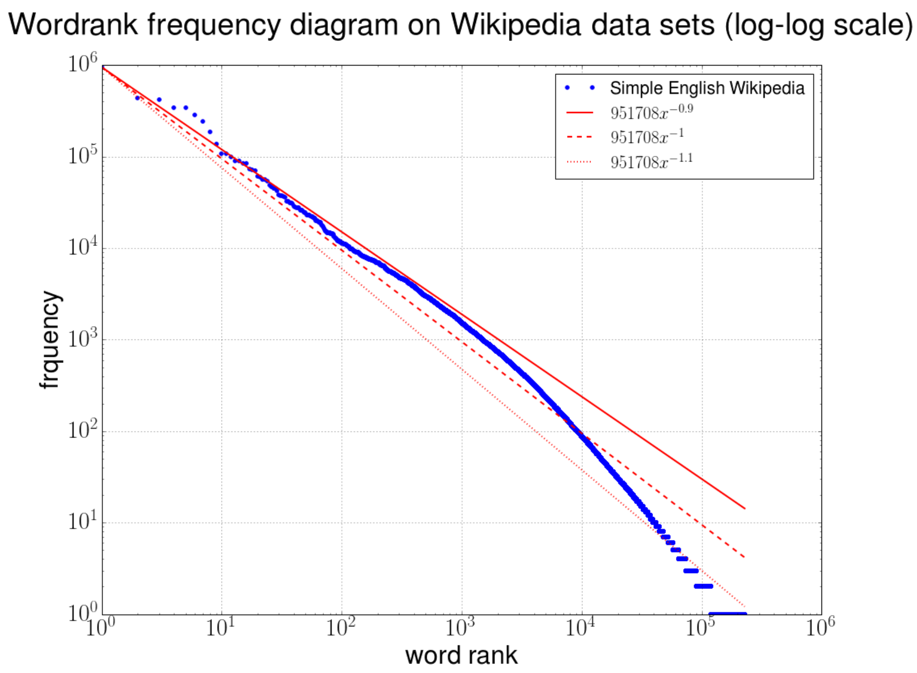
$$\|f\|_1 = \int_\Omega |f(x)|\,\mathrm{d}\mu(x)$$

L1-Norm of f in our case:

$$\|f\|_1 = \sum_{x\in\Omega} |f(x)|$$

Let us define a distance:

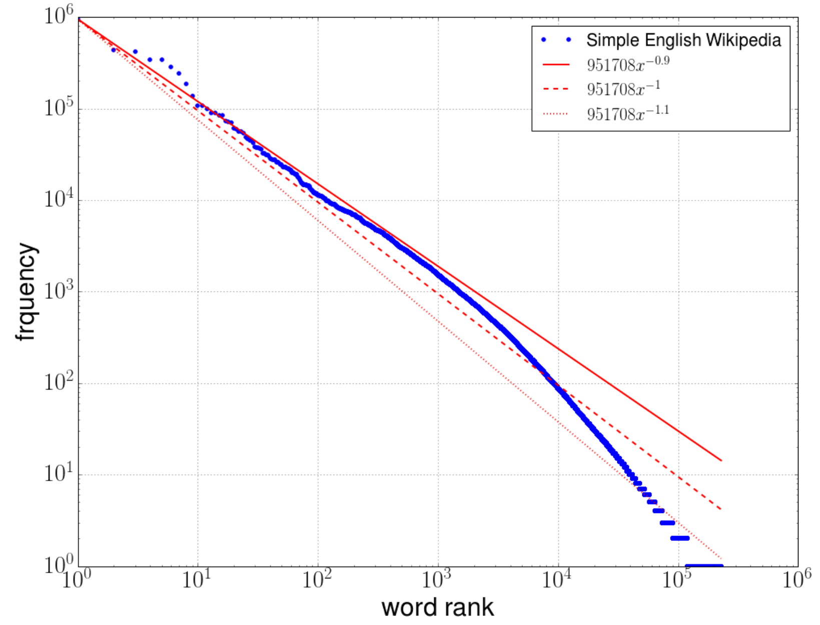
$$d_1(f_{obs}, f_{fit}) := \|f_{obs} - f_{fit}\|_1$$



Now "-1.0" seems to be the best exponent

f_{fit}	$d_1(f_{obs}, f_{fit})$
$C/x^{0.9}$	11 M
$C/x^{1.0}$	4.9 M
$C/x^{1.1}$	6.7 M

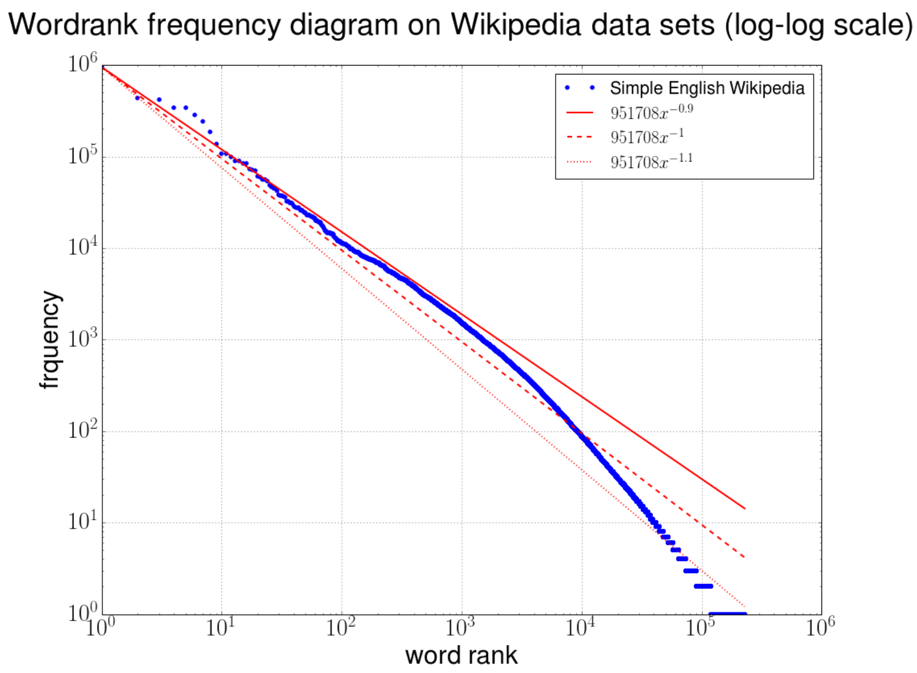
Wordrank frequency diagram on Wikipedia data sets (log-log scale)



$$C/x^a = C * x^{-a}$$

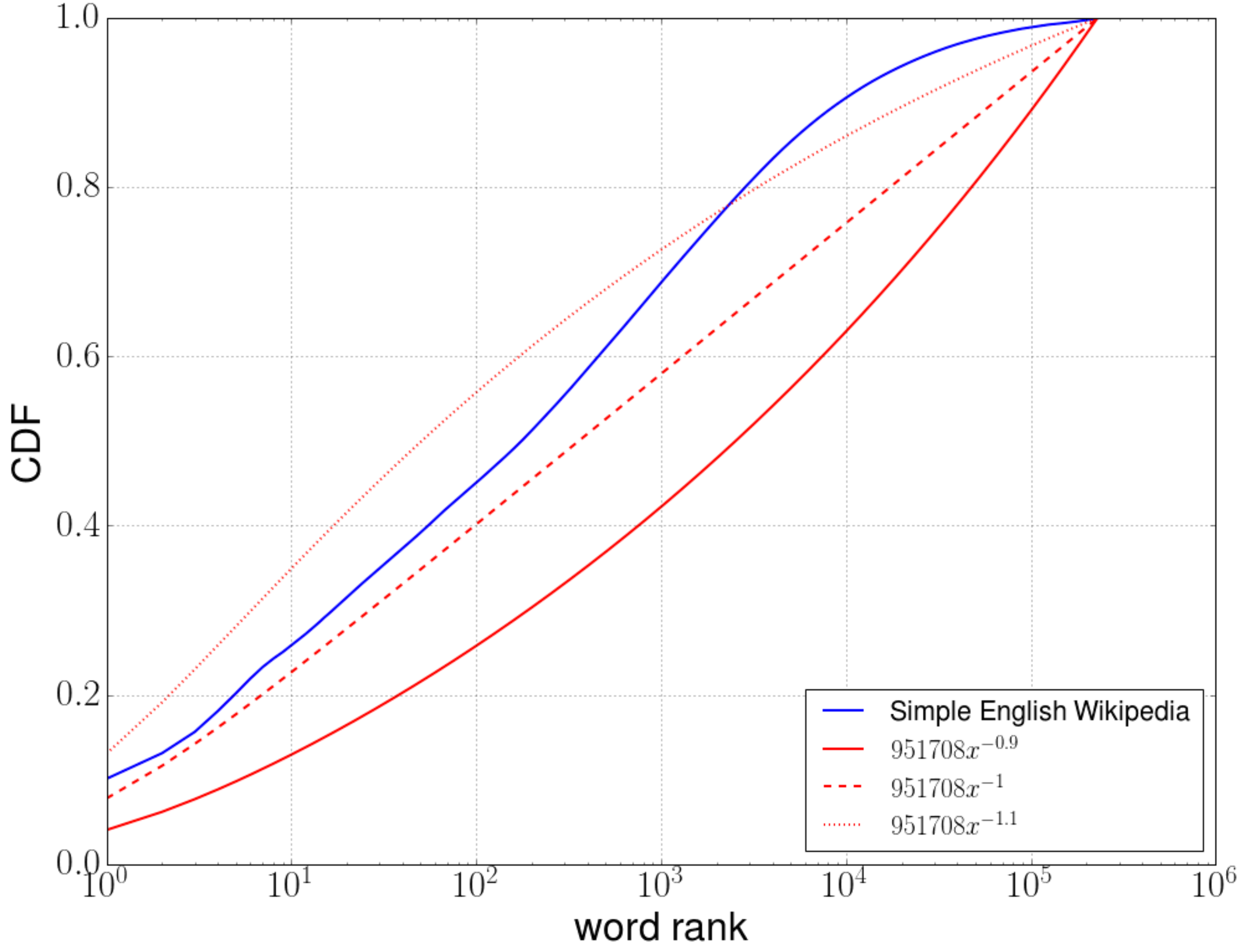
Problems with our 2nd approach

- One outlier is still enough to skew our result
- Result is not normalized
  - It can be an arbitrary large number
  - We don't know is 720M a good fit
  - Will better fits exist?

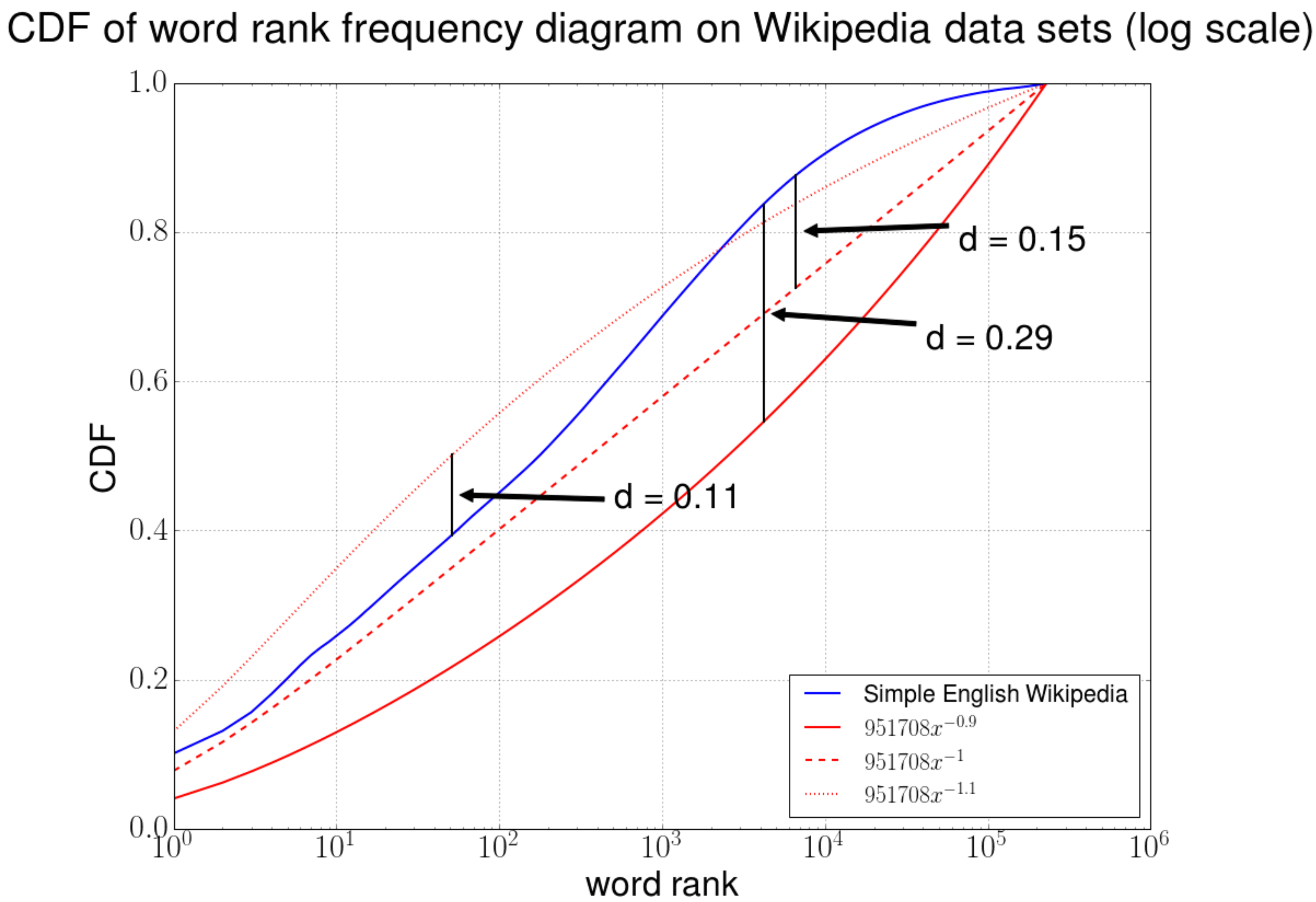


3^{rd} approach: Study the cumulative plots

CDF of word rank frequency diagram on Wikipedia data sets (log scale)



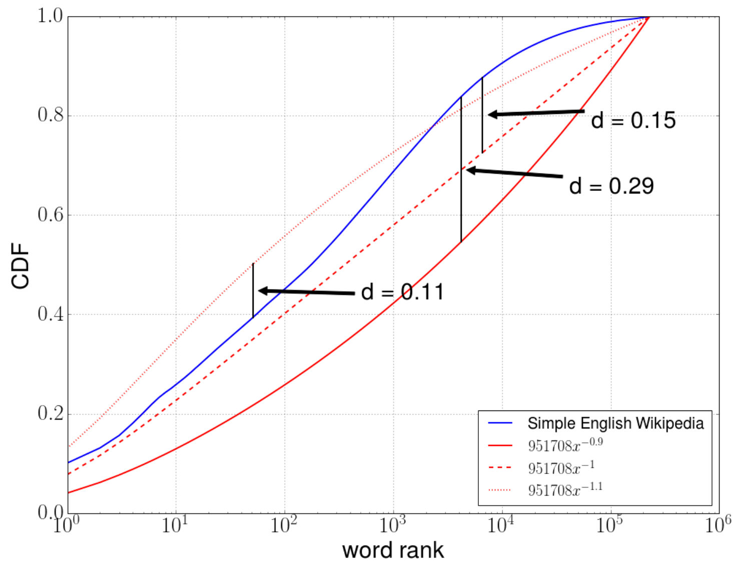
“-1.1” is now the best exponent we can find



Now “-1.1” seems to be the best exponent

f_{fit}	$d_{ks}(f_{obs}, f_{fit})$
$C/x^{0.9}$	0.29
$C/x^{1.0}$	0.15
$C/x^{1.1}$	0.11

CDF of word rank frequency diagram on Wikipedia data sets (log scale)



$$C/x^a = C * x^{-a}$$

3rd way is called Kolmogorov Smirnov Test

- Cancelling out positive and negative errors
- Wide spread statistical test for fitting tasks
- Implemented in many fitting libraries
- Even though it is wide spread it is still a modelling choice

Comparing the results of the thee distance measures

f_{fit}	**Uniform norm** (point wise distance)	**L1-norm** (cumulated error)	**Kolmogorov Smirnov** (uniform norm on CDF) – Mix of 1 and 2
$C/x^{0.9}$	**115 k**	11 M	0.29
$C/x^{1.0}$	148 k	**4.9 M**	0.15
$C/x^{1.1}$	177 k	6.7 M	**0.11**

We can characterize our data with the help of the Zipf parameter

- The **exponent** of the best fitting function is called the **Zipf parameter**

- Obviously the parameter depends on the choice of distance measure in our "**meta-model**"
  - Beware: Modelling choices change results!

- **Non trivial task** to find the best parameter
  - We just guessed and tested 3 values
  - Next unit: Estimate the parameter directly without guessing

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: