Scaling Laws for Word-Adjacency Networks

Francesc Font-Clos & Álvaro Corral

Universitat Autònoma de Barcelona

CENTRE DE RECERCA MATEMÀTICA

Given a text, we construct a directed, weighted network of word adjacency, where

Given a text, we construct a directed, weighted network of word adjacency, where

- each different word is a node
- two adjacent words form a link, with a weight $\omega_{i \to j}$ counting their frequency

$S \longrightarrow$ text length $\qquad$ $W \longrightarrow$ number of links

We are interested in the weight distribution

$$\mathcal{P}(\omega) = \mathrm{Prob}[\omega_{ij} = \omega]$$

We are interested in the weight distribution

$$\mathcal{P}_t(\omega) = \mathrm{Prob}[\omega_{ij}(t) = \omega]$$

and in its evolution over time

We are interested in the weight distribution

$$\mathcal{P}_t(\omega) = \mathrm{Prob}[\omega_{ij}(t) = \omega]$$

and in its evolution over time

We propose* the following scaling law
for the weight distribution

*(based on some reasonable assumptions and after some derivations)

We propose* the following scaling law for the weight distribution

$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{W_t S_t}$$

where the function h does NOT depend on time

*(based on some reasonable assumptions and after some derivations)

We propose* the following scaling law for the weight distribution

$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{W_t S_t}$$

where the function h does NOT depend on time

*(based on some reasonable assumptions and after some derivations)

To check our scaling law, we use google's n-grams dataset

8 million books =
0.86 trillion words =
more than 6% of all
books ever printed

with time information

To check our scaling law, we use google's n-grams dataset

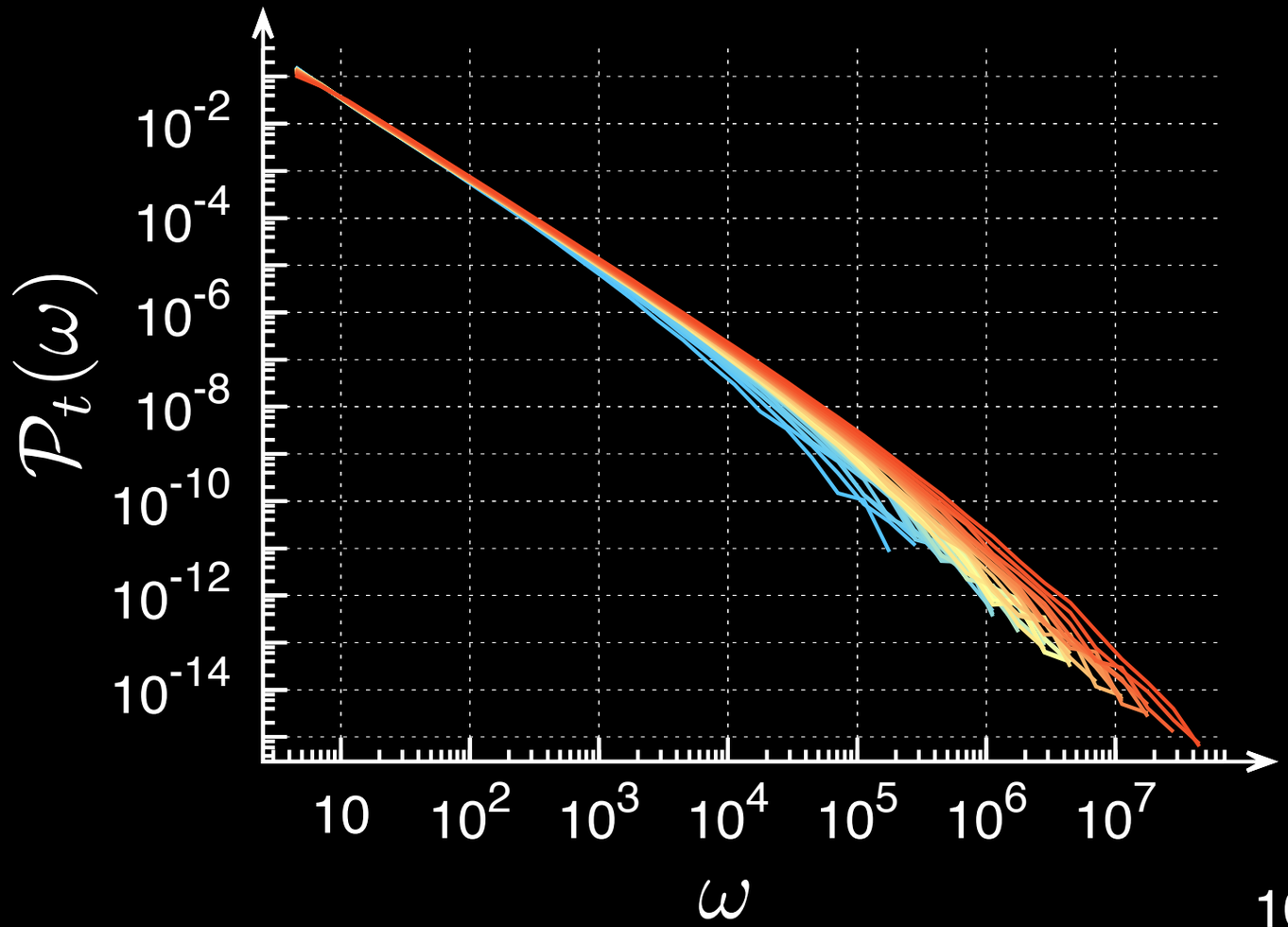
8 million books =
0.86 trillion words =
more than 6% of all
books ever printed

with time information

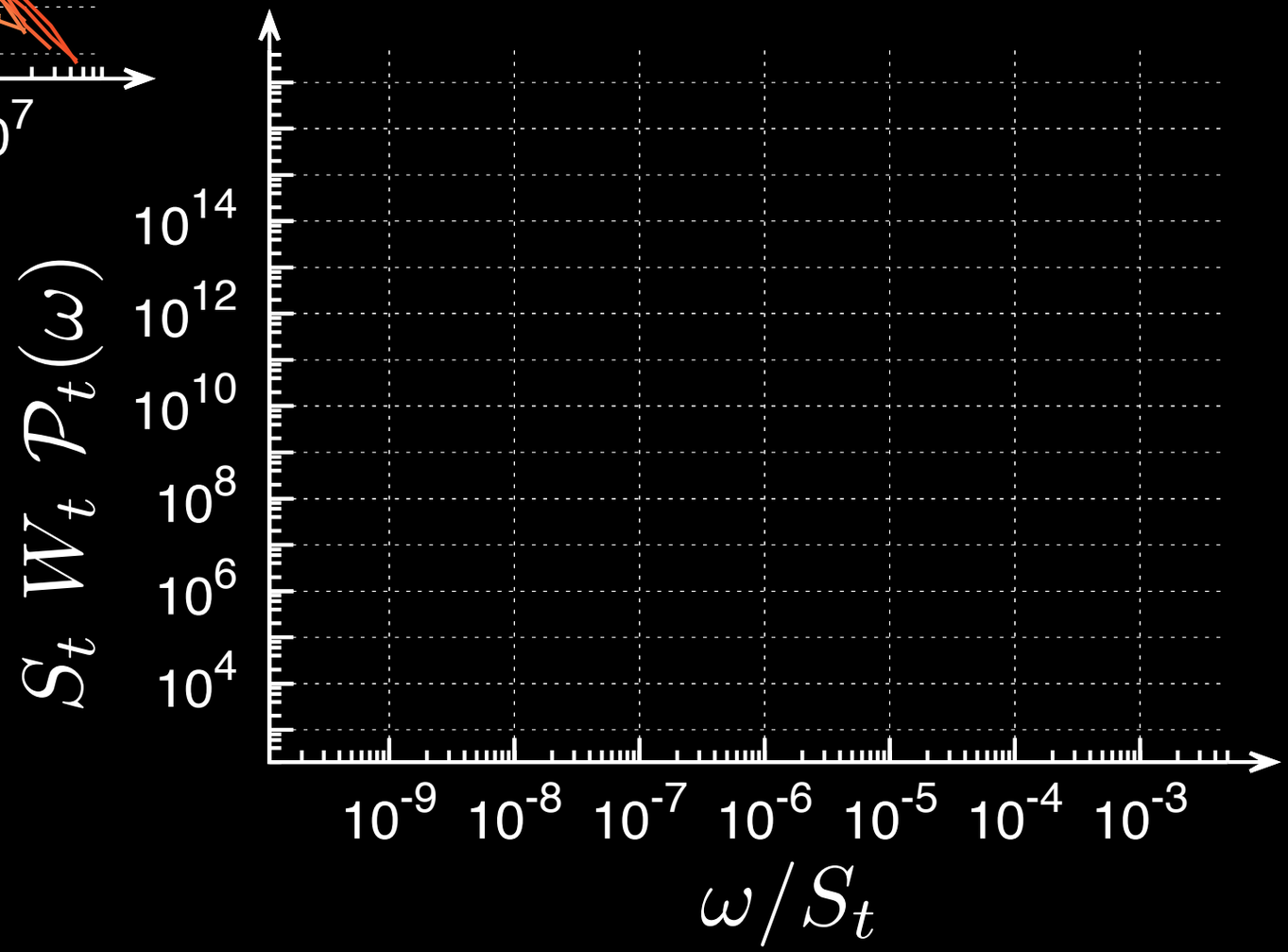
This provides data on
$\omega_{ij}(t)$ for the last
two centuries

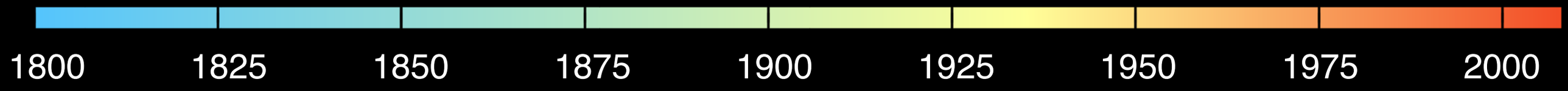
QUESTION:

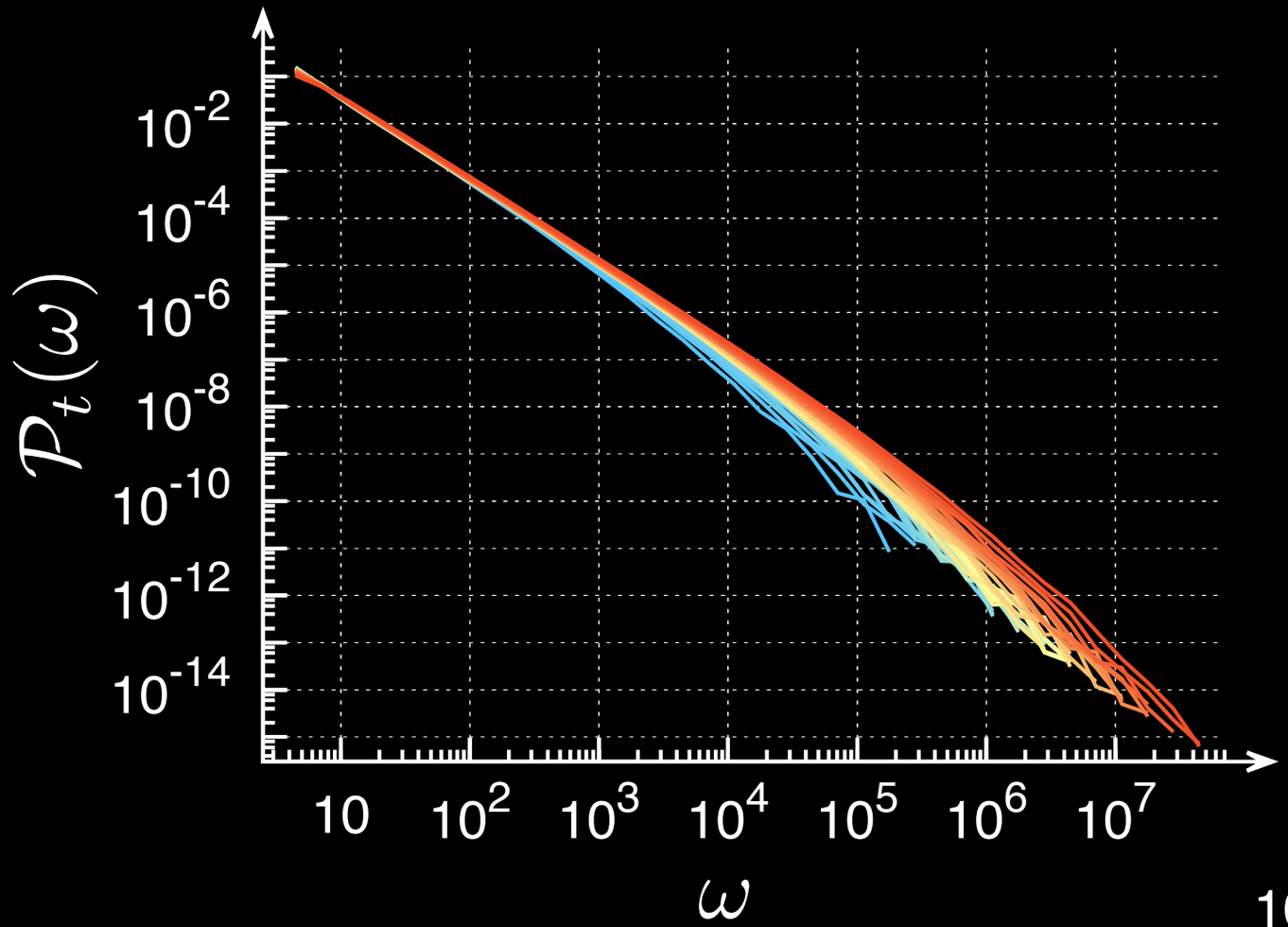
is the distribution of weights
$\mathcal{P}_t(\omega)$ robust over time?



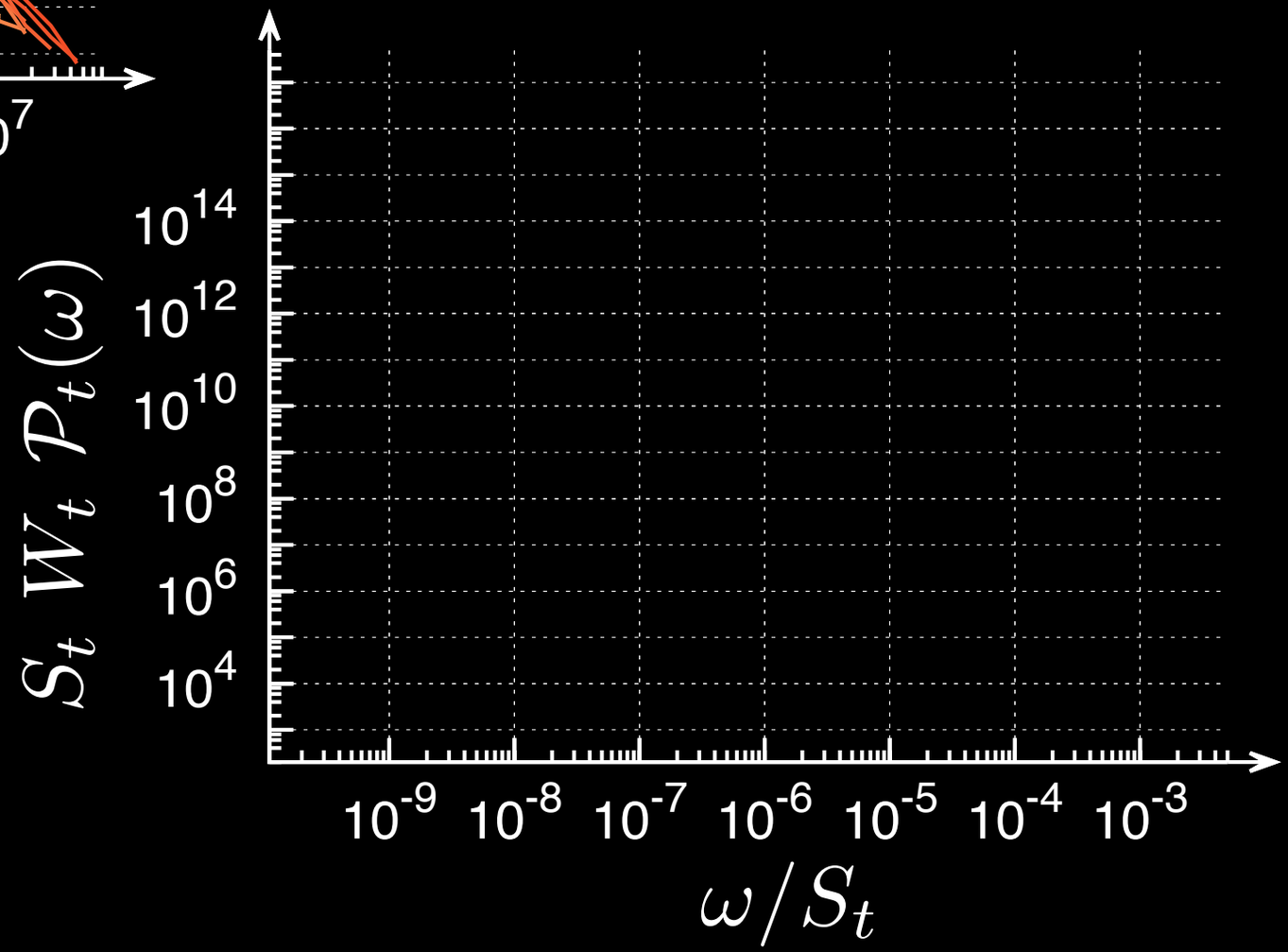
$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{S_t W_t}$$

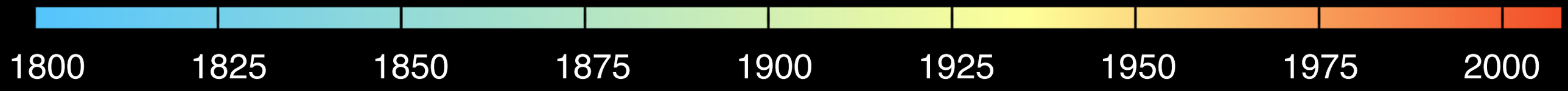


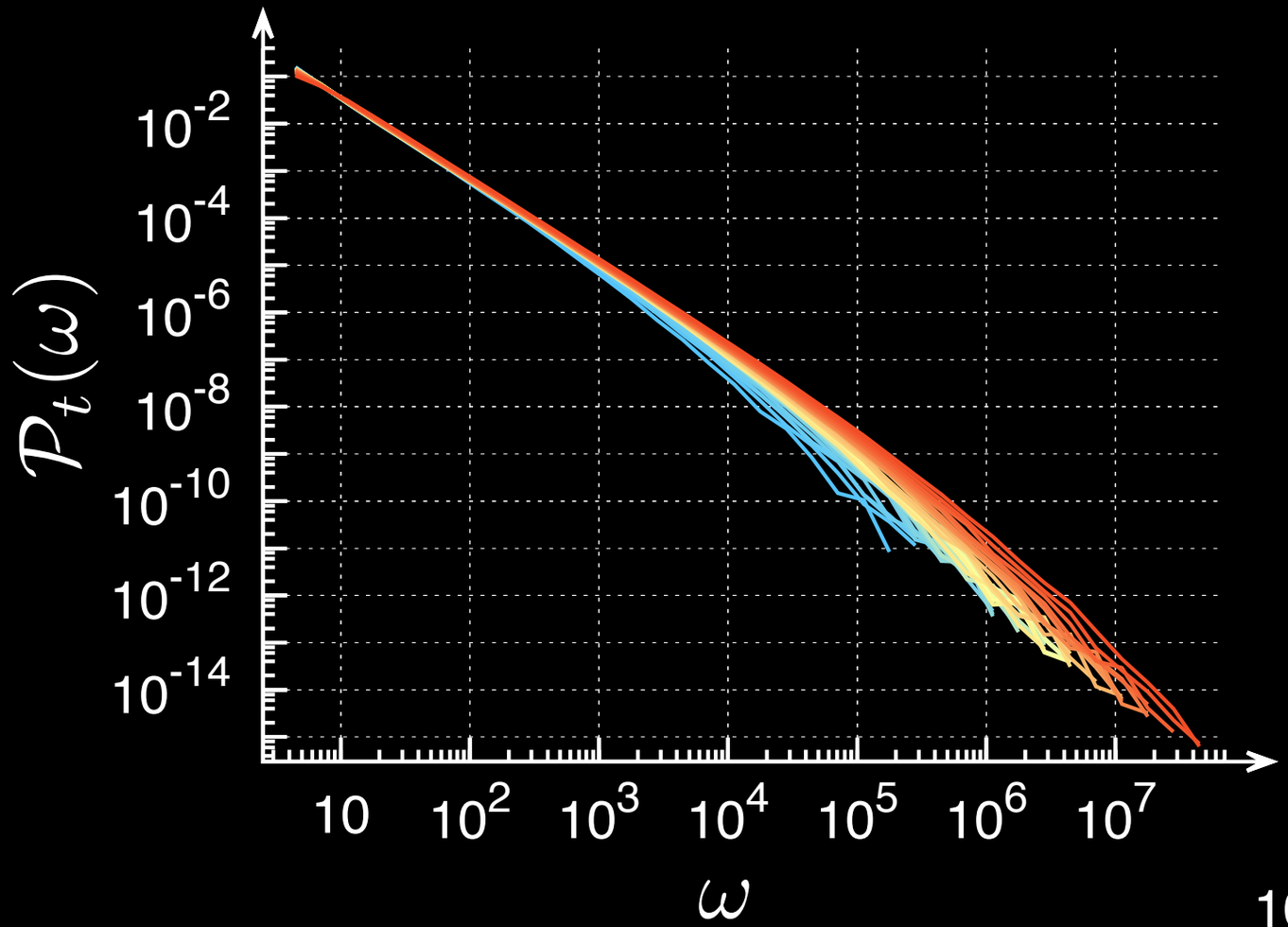




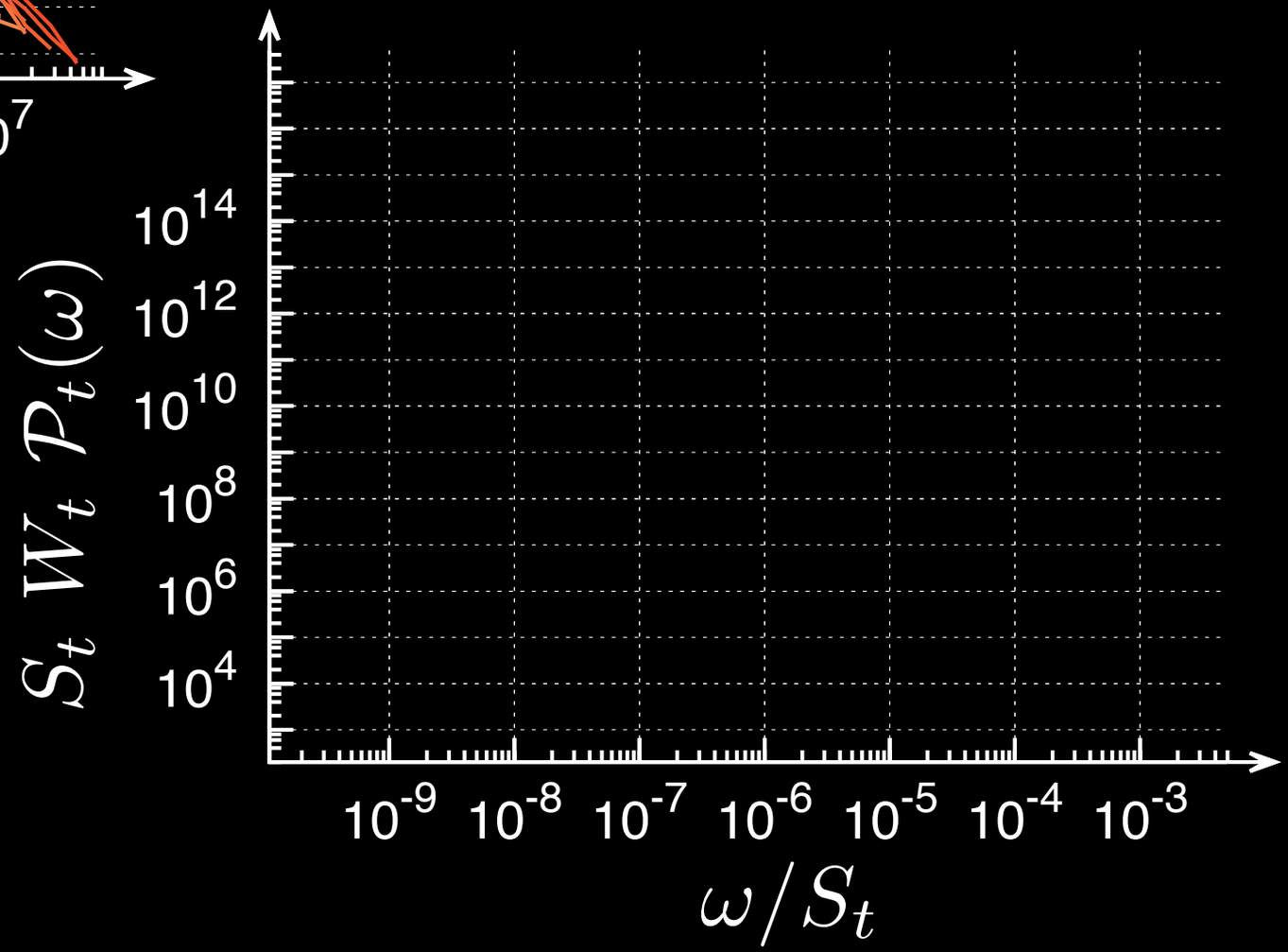
$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{S_t W_t}$$

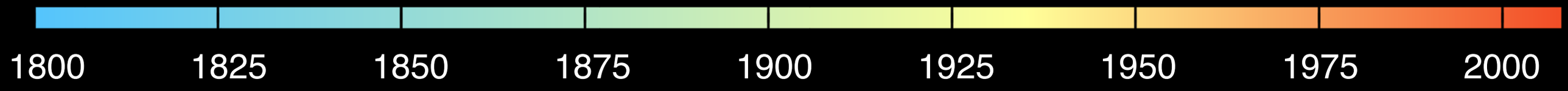


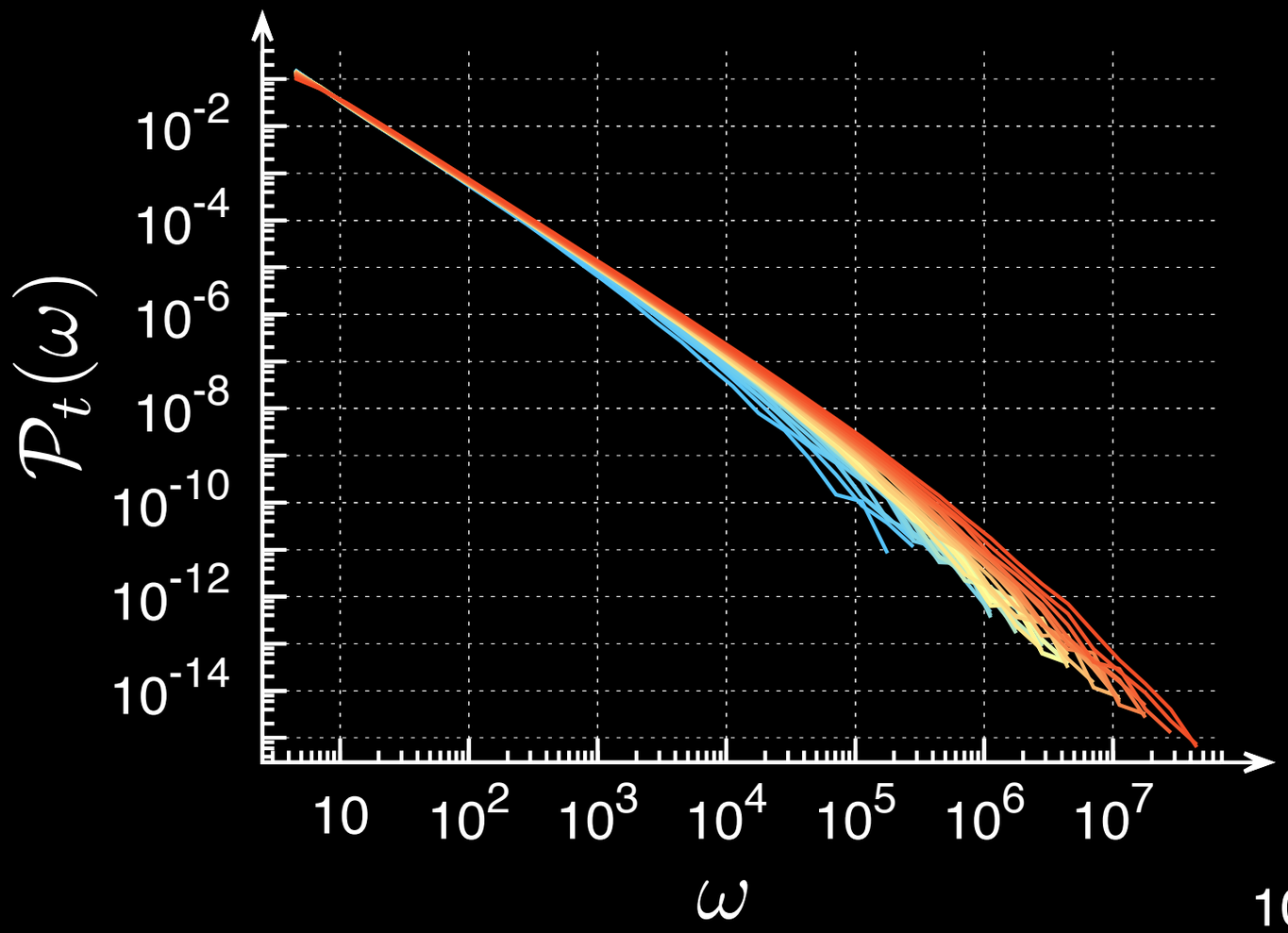




$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{S_t W_t}$$

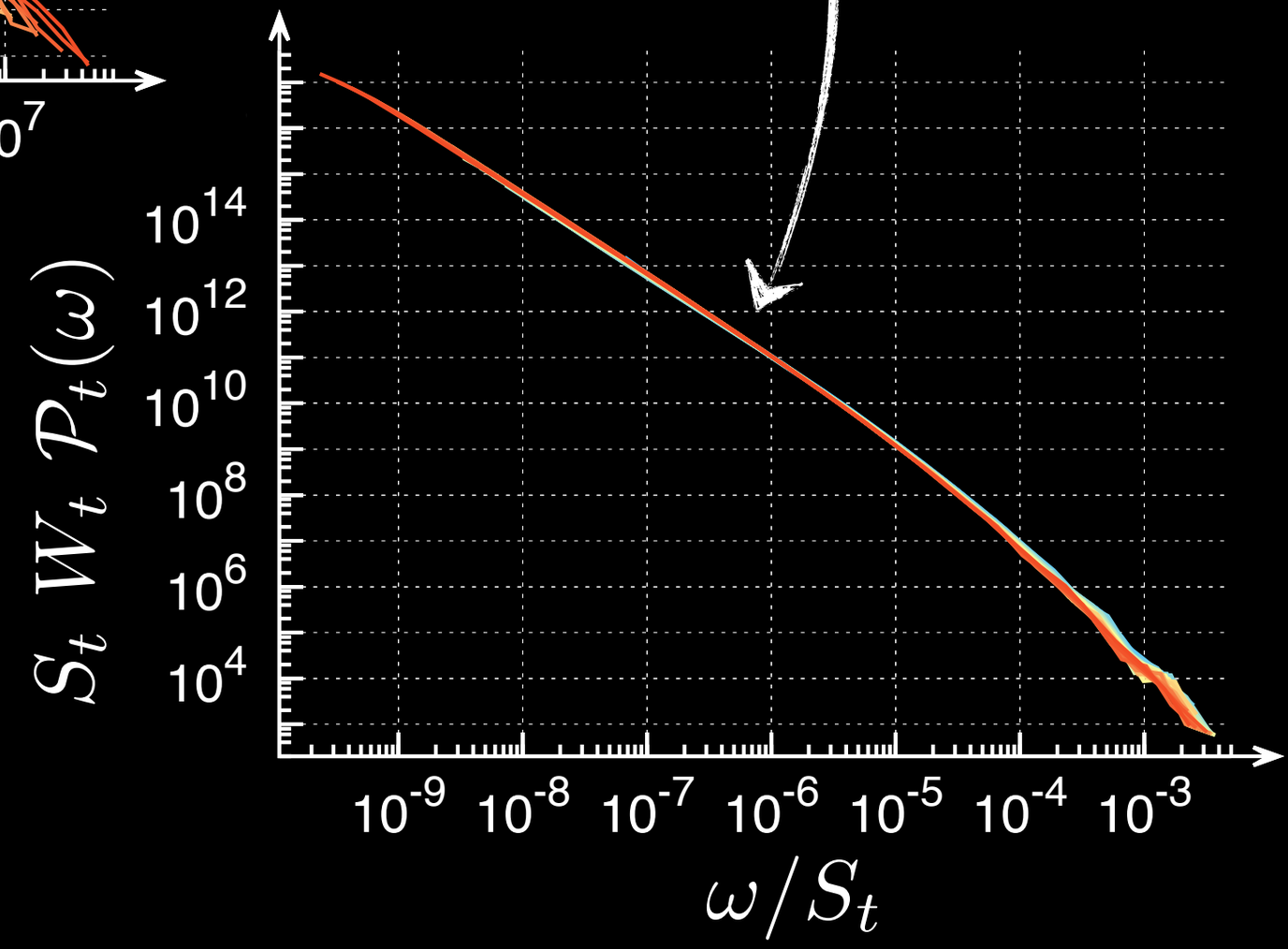


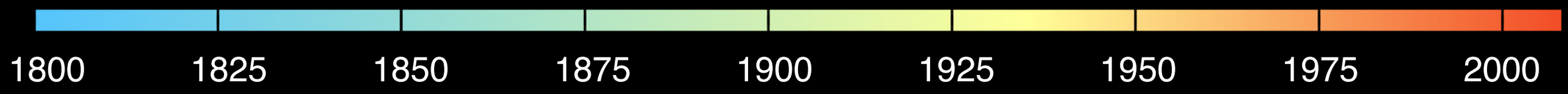


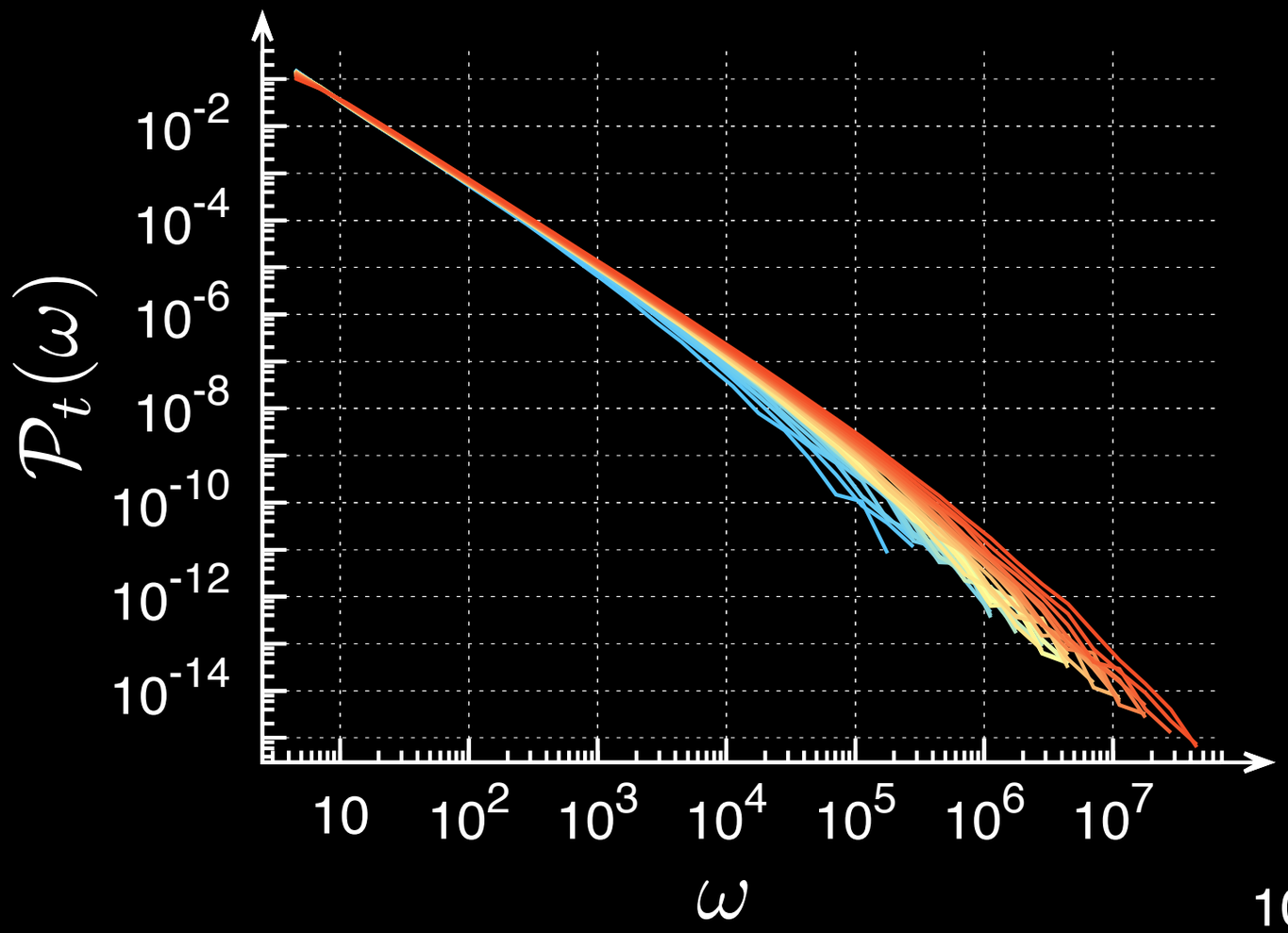


The data collapse reveals the time-independent function h

$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{S_t W_t}$$

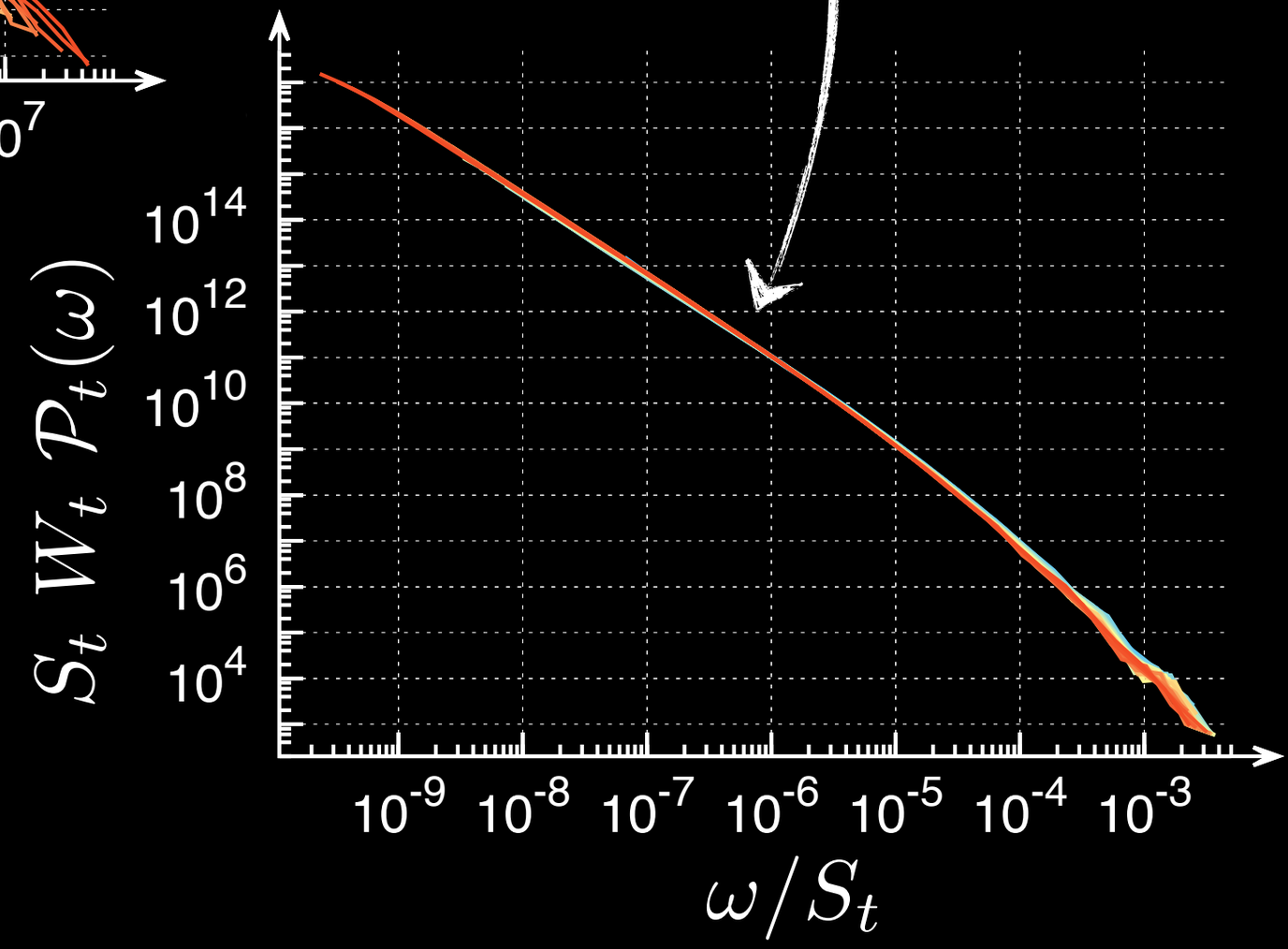


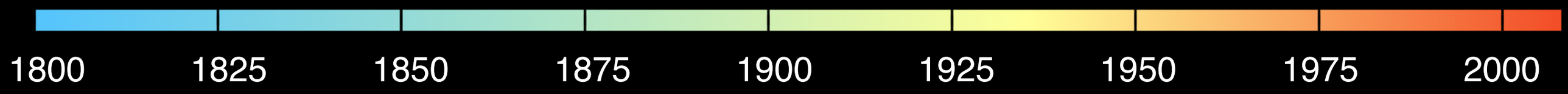




The data collapse reveals the time-independent function h

$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{S_t W_t}$$





So, in just 5 minutes, we...

- considered time-evolving language networks
- defined $\omega_{i \to j}(t)$ as frequency of transitions from word i to word j during year t
- proposed a scaling law for $\mathcal{P}_t(\omega)$
- verified it using the largest ever corpus

...but there was no time for

- the robustness of the strength distribution
- the network's growth as a function of h
- many other things... just talk to me!

At least in some sense, language is not evolving!

$$\mathcal{P}_t(\omega) = \frac{h(\omega/S_t)}{W_t S_t}$$

At least in some sense, language is not evolving!

Thanks for listening

http://www.crm.cat/researchers/fontclos
fontclos@crm.cat
@francescfont

Universitat Autònoma de Barcelona

CENTRE DE RECERCA MATEMÀTICA