Statistical Models for Frequency Distributions of Count Data

with Applications to Scientometrics

Ruheyan Nuermaimaiti[1], Leonid Bogachev[1] and Jochen Voss[1]

[1] School of Mathematics, University of Leeds

Aims

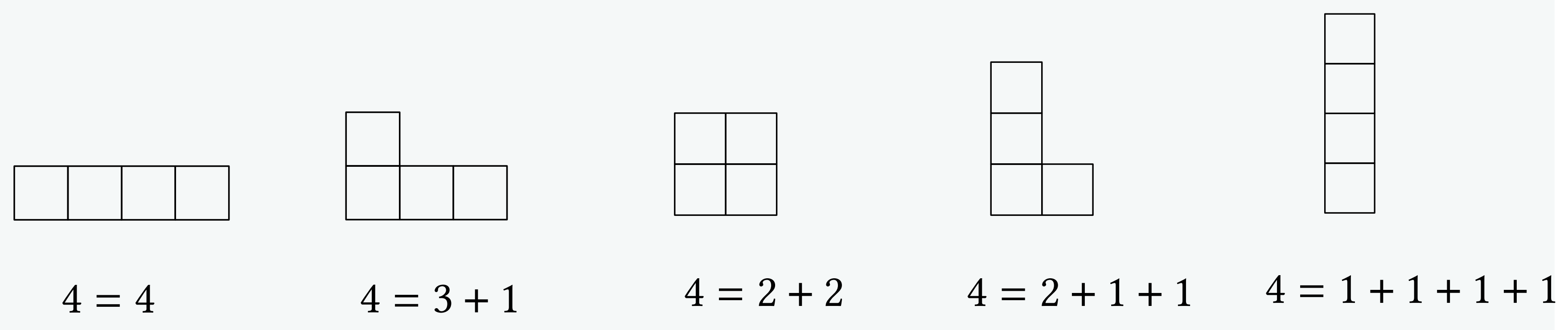
In this research, we are trying to model citation data using ideas from diverse and “unrelated” fields:

- the limit shape of large random structures
- survival analysis of the time to the first citation
- point processes capturing cumulative effects (“success breeds success”)

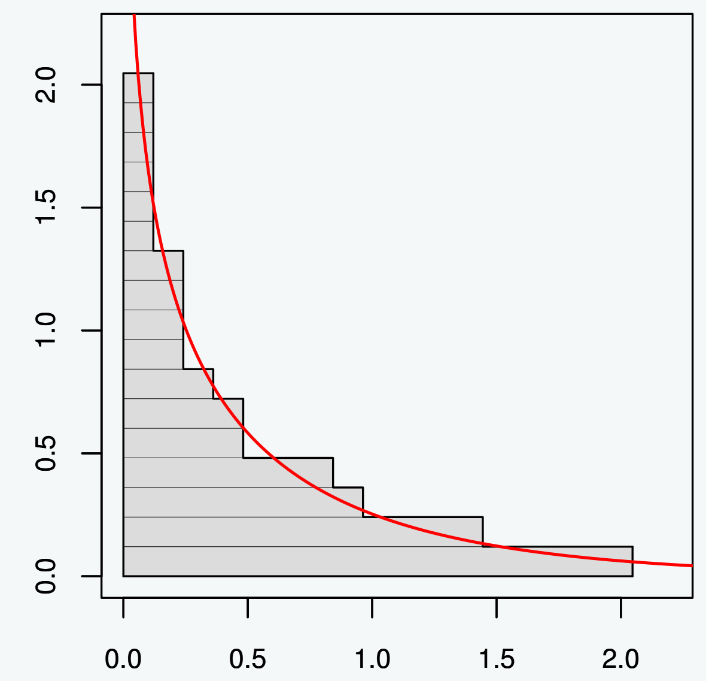
Limit shapes

- **Bridge complex mathematical concepts and scientific citations**

Integer partitions and Young diagrams



The limit shape

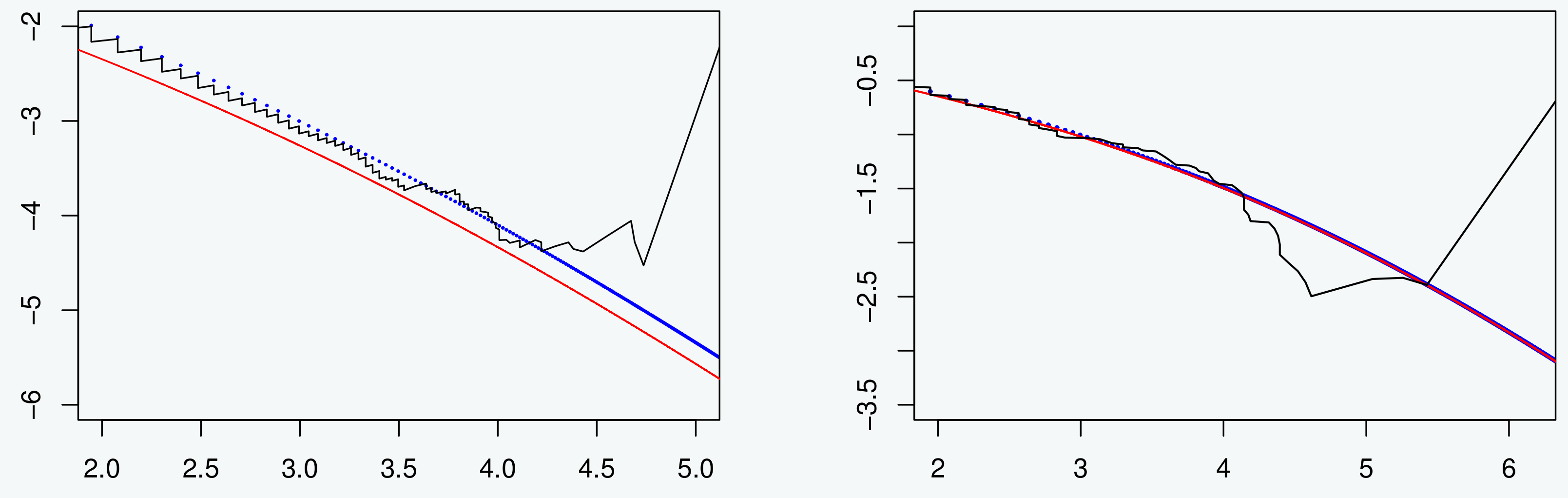


Scaled Young diagram and the limit shape of integer partitions example

- Choose an appropriate model
- Find the limit shape
- The limit shape helps to visualise the goodness of fit

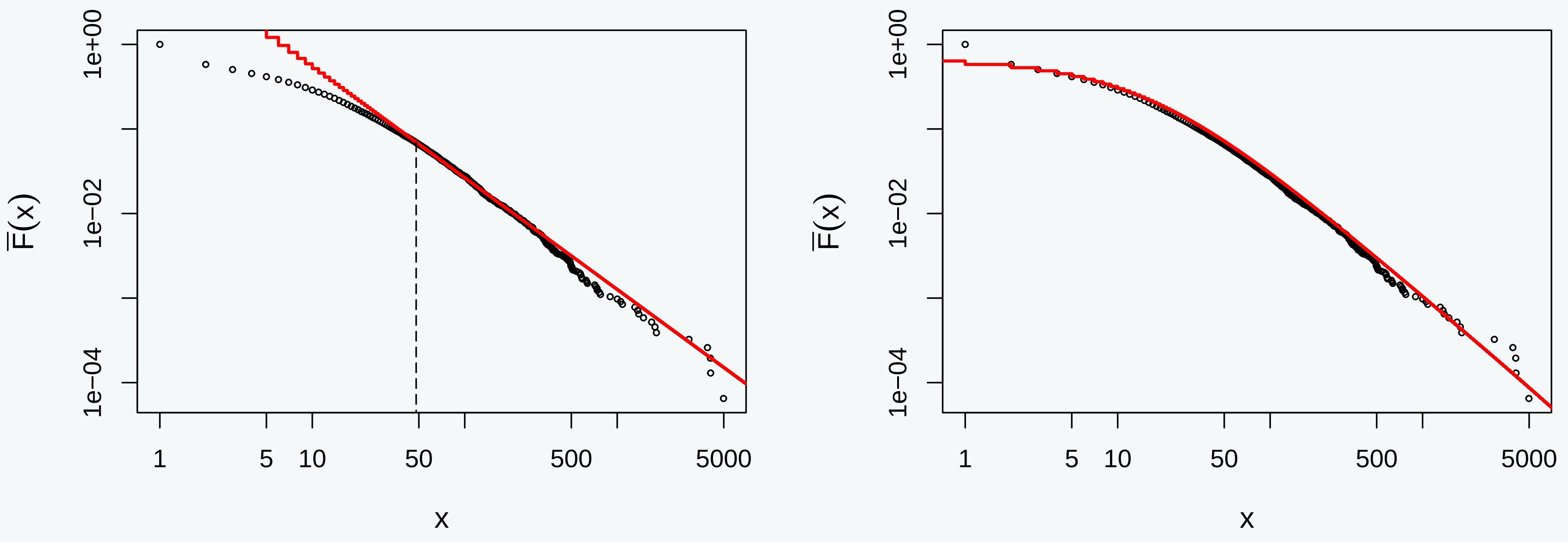
Data visualisation

To detect unusual data points:



- The left plot shows Lotka’s data set: author productivity
- The right plot shows Chen’s data set: journal use

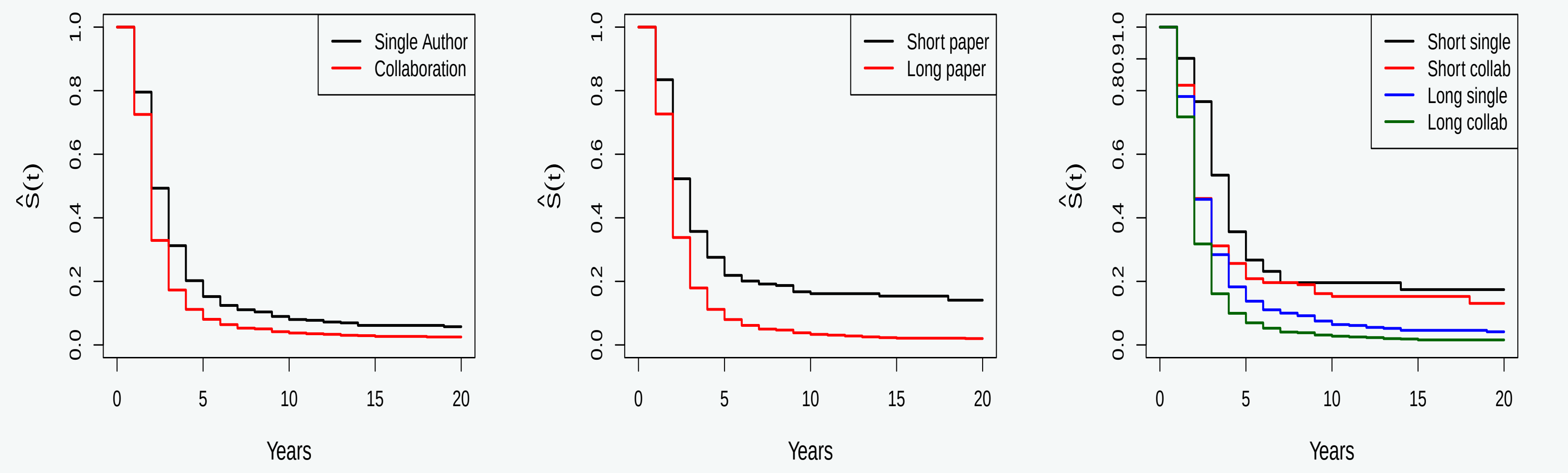
To find the best model fit to the data:



- Citations of papers data (shown in black dots) are collected using web scarping techniques.
- The conventional power law (red in the right plot) requires a truncation of lower values.
- We proposed the generalised power law (GPL) model (shown in red in the right plot) in order to provide a very good fit across the entire citation spectrum.

Survival Analysis of time to the first citation

- We apply the traditional powerful tool in medical statistics and actuarial science – survival analysis, to investigate which paper is more impactful.
- We look at the time to reveice the first citation.
- The “death” in survival analysis is interpreted as the first citation after the publication of the paper in our study.



Survival plots until first citation for papers categorised according to different covariates.

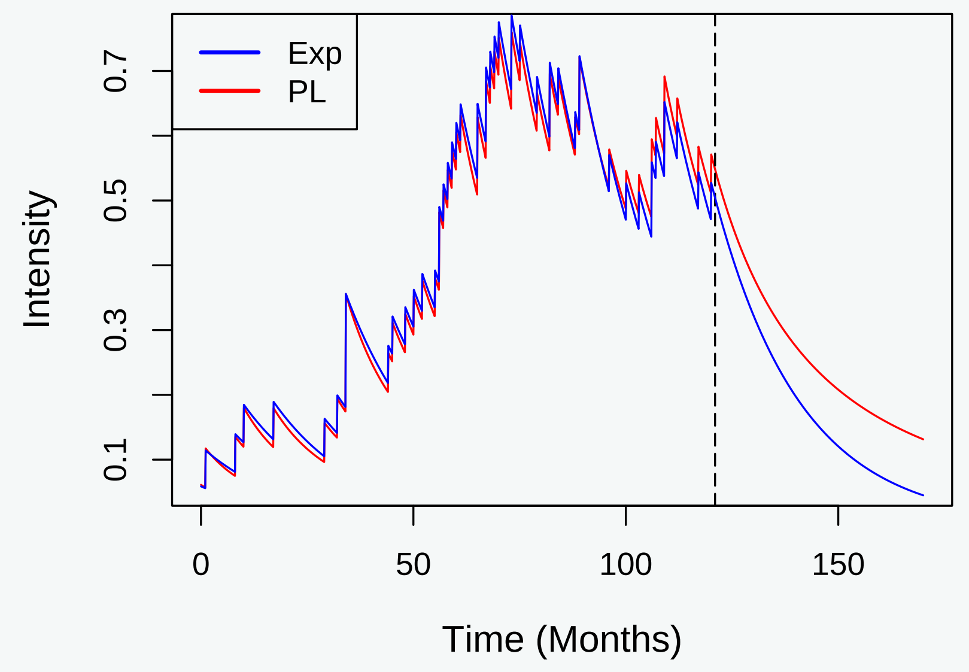
Illustrations:
The lower lines show the higher chance to be cited earlier.

Findings:

- Collaborative work receives the first citation earlier than a single-authored paper
- **(Surprisingly!)** Longer papers tend to receive their first citations earlier than shorter papers

Point processes

- We consider dynamic citations as a point process.
- The conventional Poisson process is *memoryless*, however the citation data has the property of “success breeds success”.
- The Hawkes process has the self excitement property similar to the cumulative effect of citation.
- We can set different background intensity such as power law and exponential as the kernel.



Intensity plot of the Hawkes process fitted to citations of a paper example.

These approaches can be extended to

- Butterfly species
- Followers of X
- Repeat buying
- Criminology
- and more …

References

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington academy of sciences*, 16, 317–323.

Nuermaimaiti, R., Bogachev, L.V. & Voss J. (2021). A generalized power law model of citations. *Proc. ISSI2021*, 843–848.

Nuermaimaiti, R., (2023) Statistical models for frequency distributions of count data with applications to scientometrics. PhD thesis, University of Leeds.

Bogachev, L.V., Nuermaimaiti, R. & Voss, J.(2023). Limit Shape of the Generalized Inverse Gaussian-Poisson Distribution. *arXiv preprint arXiv:2303.08139*

Ruheyan (Rukia) Nuermaimaiti

R.Nuermaimaiti@leeds.ac.uk

University of Leeds