Ranking influential communities in networks

What can citation data reveal about the flow of influence in scientific fields?

David Selby

D.Selby@warwick.ac.uk

David Firth

D.Firth@warwick.ac.uk

Finding communities

How to group journals into fields using random walks

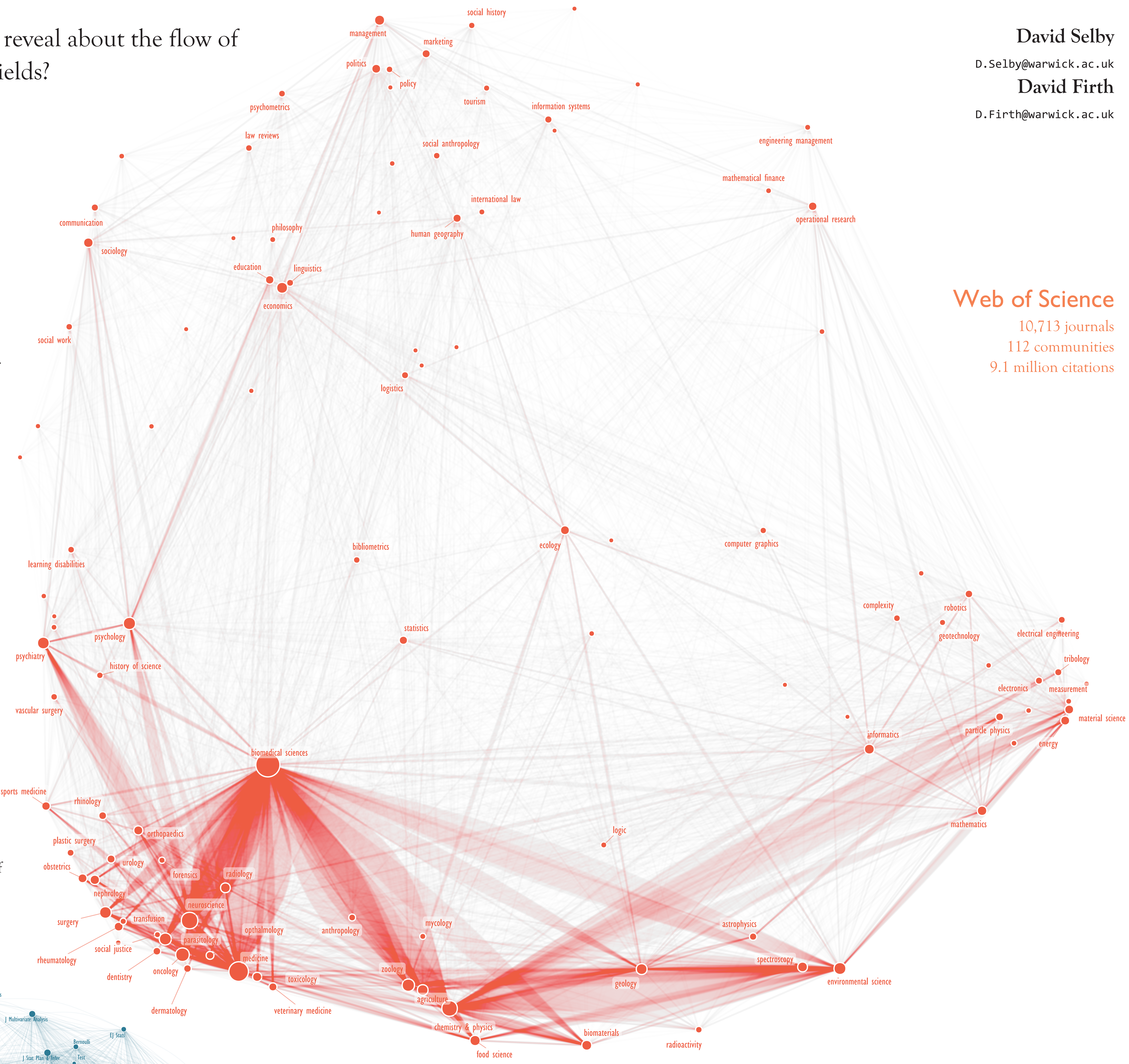
The **Infomap algorithm** (Rosvall & Bergstrom 2008; *PNAS* 105) looks for a clustering of nodes that gives the shortest possible description of a random walk around the network.

Every node has a two-level address, identifying a community and the node's position within that community. It is like a street address, comprising a street name (position) and city name (group). Street names are often re-used between cities, but every street–city pair is unique.

By aggregating the journals in each community into a single 'super-journal', we can model the exchange of citations between disciplines.

Right: the Web of Science network. Each node represents a community of journals. The edges represent interdisciplinary citations.

Below: journals and citations within the statistics community.

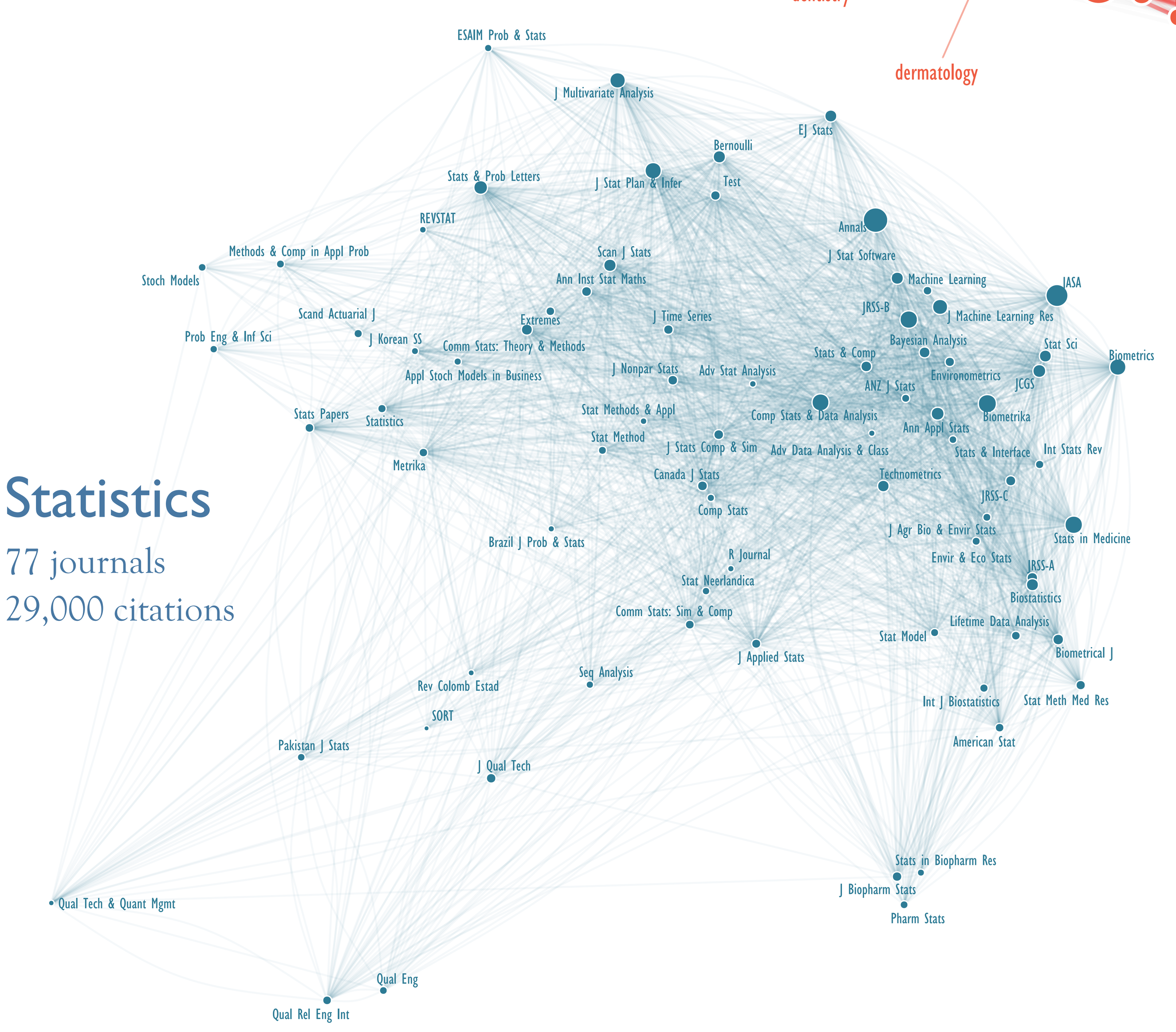


Web of Science

10,713 journals

112 communities

9.1 million citations



Statistics

77 journals

29,000 citations

Statistical model

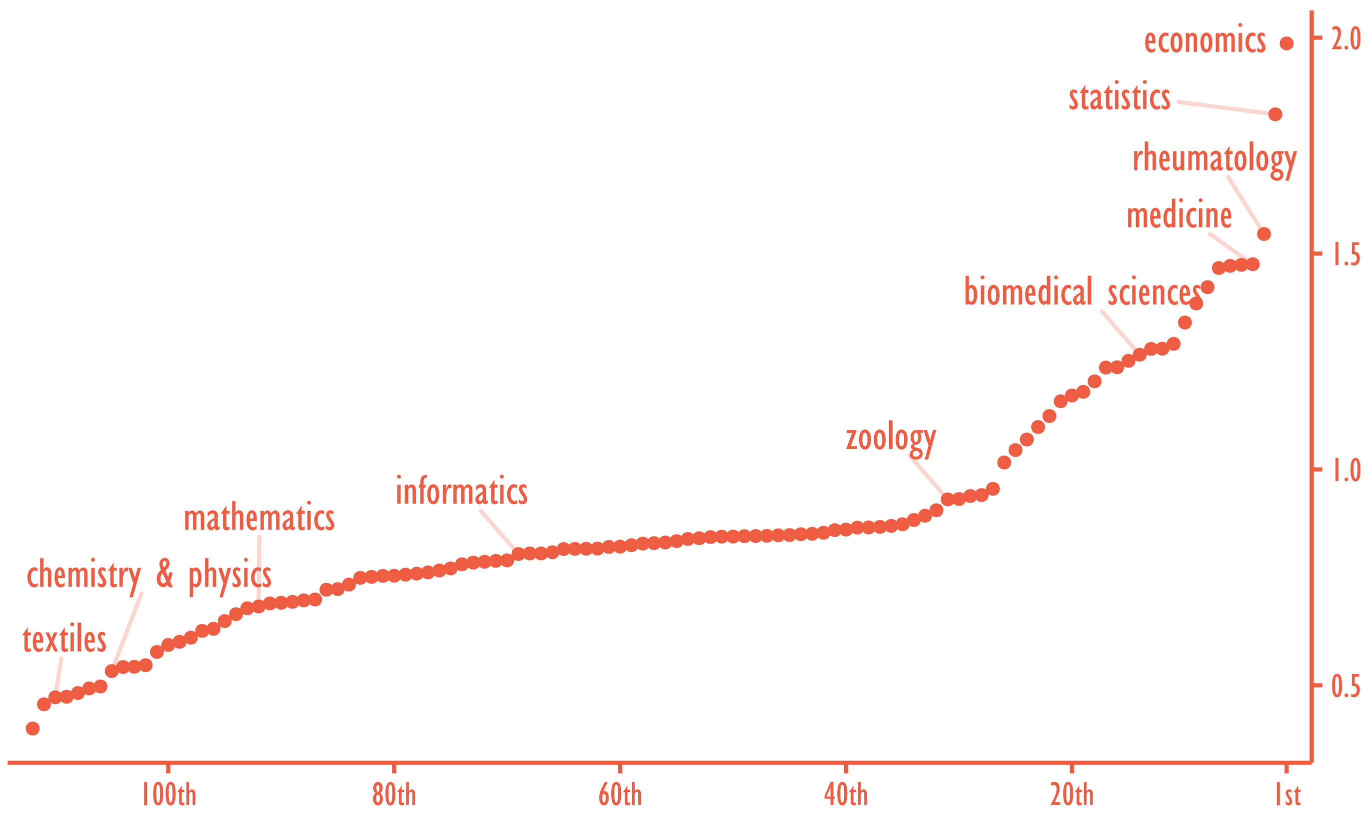
Given a set of paired comparisons, the **Bradley–Terry model** estimates an ability score for each object, such that

$$\frac{\mathrm{P}(i \text{ beats } j)}{1 - \mathrm{P}(i \text{ beats } j)} = \frac{\mu_i}{\mu_j}$$

for any pair of objects i and j.

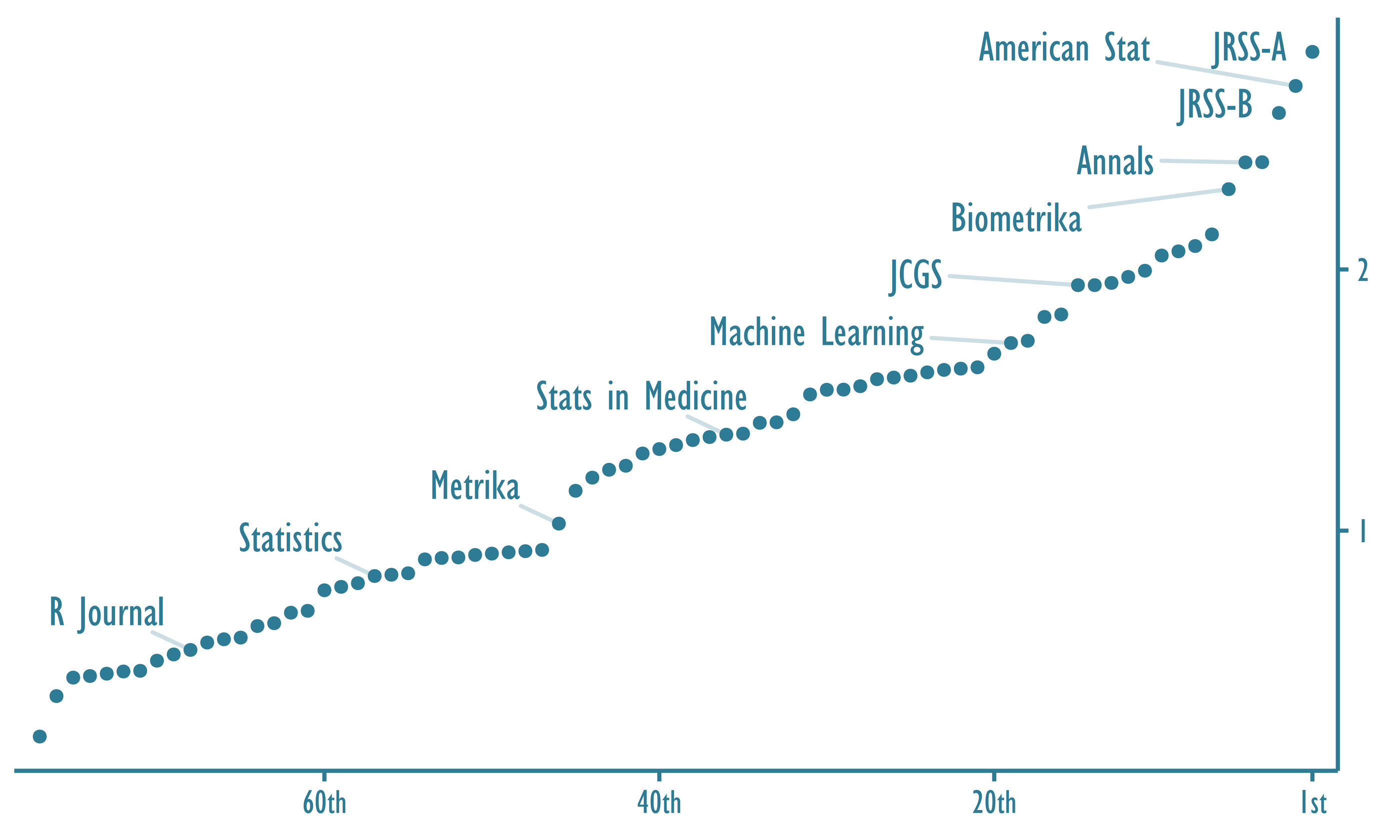
Citations between academic journals can be treated as paired comparisons: being cited means being an 'exporter of intellectual influence' (Stigler 1994; *Statistical Science*).

Using ability scores, we can predict the probability that journal i cites journal j more than j cites i. Influential journals are more likely to be cited by other influential journals.



Above: a between-fields ranking. Applied, general disciplines tend to have higher Bradley–Terry ability scores than highly specialised or theoretical ones.

Below: a within-field ranking for statistics journals. A high score implies the journal is influential within statistics, but not necessarily influential on other fields.



Ranking journals

How to measure influence using random walks

The **PageRank algorithm** models the behaviour of a typical PhD student.

1. Open a random journal.
2. Pick a random reference and open the cited journal.
3. Repeat, *ad nauseum*.

PageRank is the proportion of time spent reading each journal, i.e. the stationary distribution of an ergodic Markov chain. It is a measure of total influence.

PageRank has a size bias: bigger journals have more/longer articles in them, attracting more citations. What if we want to measure *prestige*, rather than *popularity*?

The **Scroogefactor** score, defined as PageRank *per reference*, controls for this size bias. It measures influence weight per outgoing citation.

Like the Bradley–Terry model, journals are, in effect, penalised for being generous with citations and rewarded for being miserly. When the Bradley–Terry model fits exactly, a journal's Scroogefactor score is exactly equal to its Bradley–Terry score.

Acknowledgements

David Selby is supported by EPSRC grant EP/M508184/1 and David Firth is supported by the EPSRC *i-like* programme.

Thanks to Thomson Reuters for providing Journal Citation Reports data in a convenient format.

R code and full output from the analysis is available on GitHub: `https://github.com/Selbosh/user2017`