PageRank and the Bradley–Terry model
Measuring influence with the Scroogefactor

David Selby
D.Selby@warwick.ac.uk
David Firth
D.Firth@warwick.ac.uk

Ranking fields

Which academic fields export the most intellectual influence, through recent research, to other fields?

This **chord diagram** shows the flow of 20m citations between the fields of biology, chemistry, computer science, engineering, medicine, mathematics, multidisciplinary sciences, physics, psychology and statistics & probability.

Citation data was collected by Thomson Reuters from 2003–2012.

PageRank algorithm

Do a random walk around the graph:

1 Select a random journal and open it
2 Select a random reference
3 Open the cited journal
4 Repeat 2–3, forever

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

A similar approach was first proposed as the "total influence" metric for ranking physics journals (Pinski & Narin 1976; *Inf Process Manag* 12:5).

This was later adapted by Google in the 1990s for their search engine, with an added "damping factor" to allow teleportation around the graph.

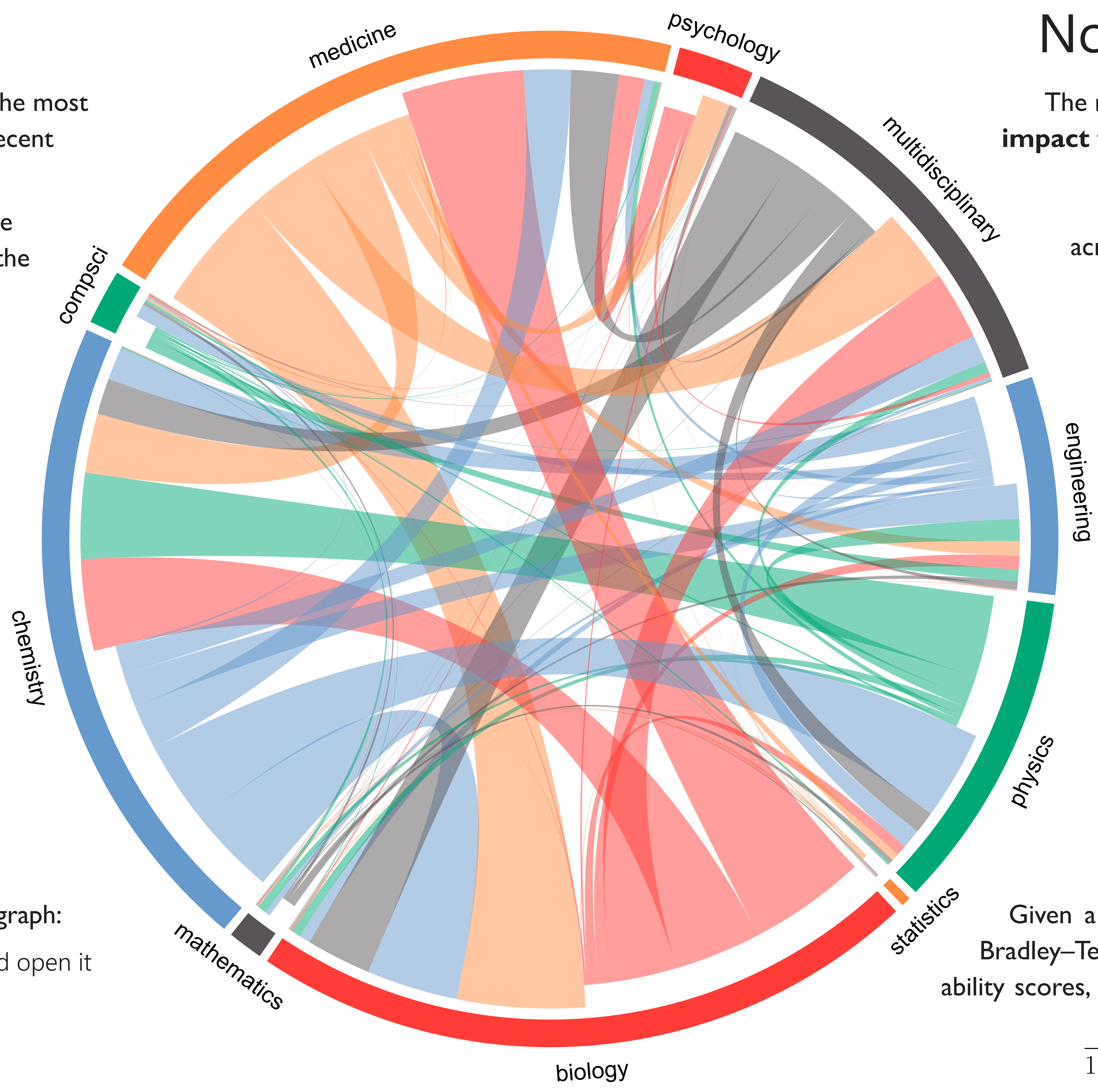
The Scroogefactor

PageRank has a size bias: larger journals or fields receive more citations overall, so will have larger PageRank.

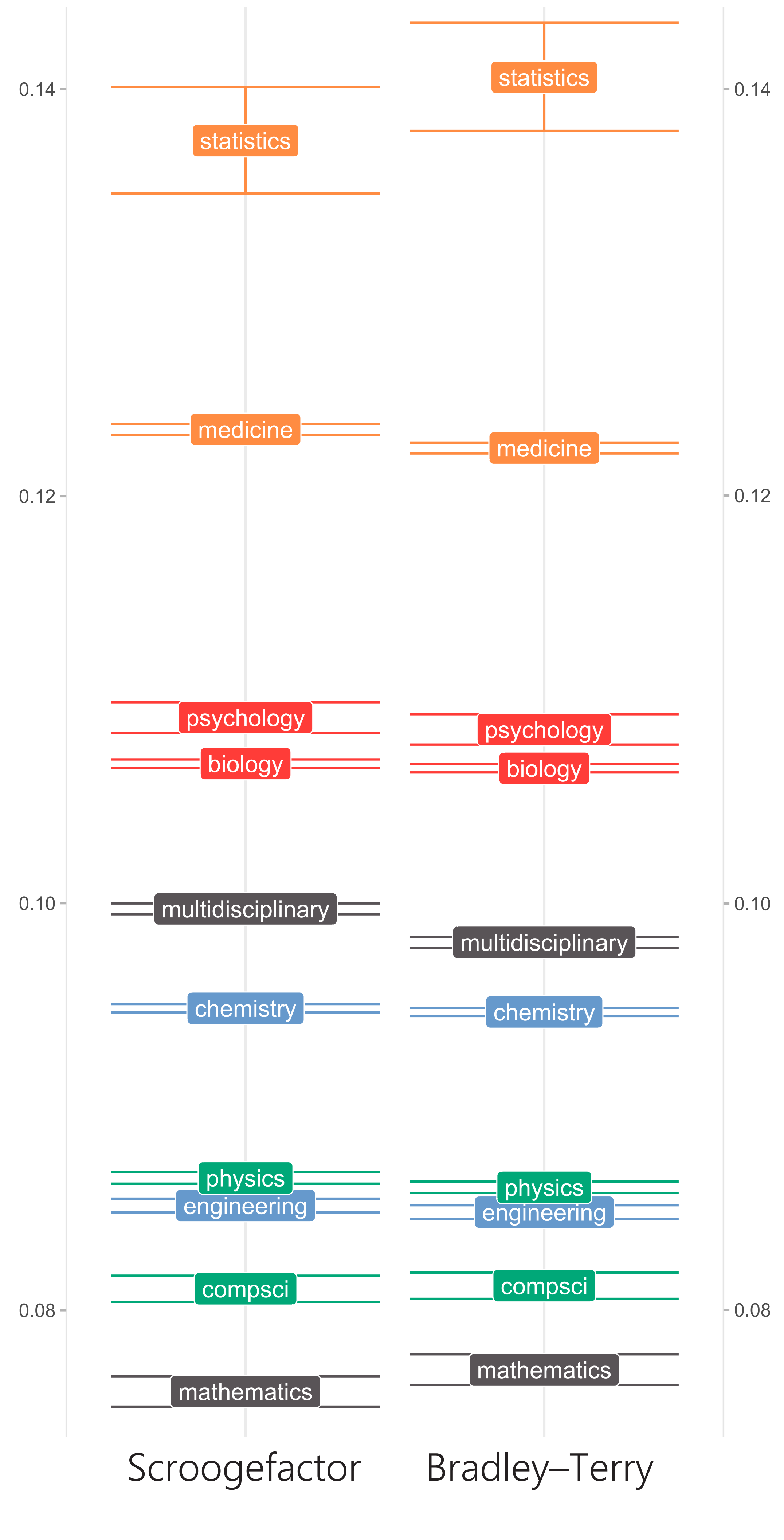
This is a problem if we want to measure *prestige*, rather than *popularity*.

The Scroogefactor, defined as **PageRank per outgoing citation**, controls for this: journals are, in effect, penalised for being generous with citations and rewarded for being miserly.

When the citation matrix is quasi-symmetric, the **Scroogefactor is exactly equal to the Bradley–Terry scores**.



Academic field rankings



Not impact factor

The most popular citation metric, the **impact factor** (Garfield 1972; *Science*) is controversial and widely abused. Citation behaviour varies greatly across disciplines, making inter-field comparisons difficult.

More recent approaches based on the PageRank algorithm improve upon the impact factor by taking the source of citations into account.

However, none of these methods includes any way of quantifying uncertainty, so are they really *statistical*?

Bradley–Terry model

Given a set of paired comparisons, the Bradley–Terry model estimates a vector of ability scores, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, 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*).

The Bradley–Terry model is equivalent to the **quasi-symmetry model**.

A square matrix **X** is called *quasi-symmetric*, or *symmetrizable*, if it can be decomposed in the form **X** = **DS**, where **D** is diagonal and **S** is symmetric. The Bradley–Terry model estimates the elements of **D**, the inverse of the *symmetrizer* matrix.

Statistics ranked top!

By aggregating all publications in each field into a single "super-journal", we can model the exchange of citations between disciplines.

Both Scroogefactor and the Bradley–Terry model give the same overall ranking: "purer" fields, such as mathematics and computer science, have lower scores, while more "applied" fields (e.g. medicine) have more interdisciplinary influence.

The error bars are 95% comparison intervals calculated using a quasi-variance approximation (Firth & de Menezes, 2005; *Biometrika*).

David Selby is supported by an EPSRC PhD studentship and David Firth is supported by the EPSRC i-like programme. We are very grateful to Thomson Reuters for providing Journal Citation Reports data in a convenient format.