Incentives and Rewards in Scientific Software Communities

Edzer Pebesma

ifgi
Institute for Geoinformatics
University of Münster

Software and Services for Science (S3),
Second Conference on Non-Textual Information,

TIB Hannover, May 10-11, 2017

Who am I?

1997- contributor of open source software,

2003- active developer in, and member of, the R community

2007- professor at the Institute for Geoinformatics, Münster

2014- co-editor-in-chief of *Computer & Geosciences*

2015- co-editor-in-chief of *Journal of Statistical Software*

2015- associate editor for *Spatial Statistics*

2016- co-PI in a DFG-funded project *Opening Reproducible Research* `http://o2r.info`

2016- blogger on `http://r-spatial.org`, active twitter user

Scientists...

- try to discern *facts* from *false facts*
- try to find concensus about this,
- do this by a public discourse,
- use methods about which a shared understanding exists
- (should) strive, in communication, for ultimate transparency

$\Rightarrow$ Communication is a key activity for scientists

Scientists...

- try to discern *facts* from *false facts*
- try to find concensus about this,
- do this by a public discourse,
- use methods about which a shared understanding exists
- (should) strive, in communication, for ultimate transparency

$\Rightarrow$ Communication is a key activity for scientists

Successful Scientists...

are those

- who other people listen to
  $\Rightarrow$ attention: publications, citations, grants
  $\Rightarrow$ reputation: cumulative attention
- whose work is being reused a lot
  $\Rightarrow$ proxies: # citations, *h*-index

The most cited papers...

according to Van Noorden et al.,[1], the most cited paper is

- Lowry et al., 1951, *Protein measurement with the folin phenol reagent*

and most cited papers are often

- not the ones decribing discoveries, or big scientific break-throughs
- papers that describe methods, or *tools* that everyone uses
  - first sequenced human genome
  - a particular method/tool used by a large domain
  - software tools that make things possible, and are understood

[1] `http://www.nature.com/news/the-top-100-papers-1.16224`

R history

History:

1976-1988 S (AT&T)

1988-2007 S-Plus (Lucent, Insightful, Tibco)

1997- R

2013- TERR, ...

Parallel history: ftp sites like `netlib`, `StatLib`
`https://www.r-project.org/`:
"R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS."

R history

History:

1976-1988 S (AT&T)

1988-2007 S-Plus (Lucent, Insightful, Tibco)

1997- R

2013- TERR, ...

Parallel history: ftp sites like `netlib`, `StatLib`

`https://www.r-project.org/`:

"R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS."

R history

History:

1976-1988 S (AT&T)

1988-2007 S-Plus (Lucent, Insightful, Tibco)

1997- R

2013- TERR, ...

Parallel history: ftp sites like `netlib`, `StatLib`
`https://www.r-project.org/`:
"R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS."

R, the R community

- R started off by computer scientists/statisticians (who needed it most)
- S's original goal: interact with data, programatically
- R evolved from a group of people using (extending) S-Plus, into a group of people who believed they didn't need S-Plus for this
- R is statistics oriented, but domain agnostic (empirical sciences)
- R started as a research project – "can we do this?"

R, the R community

- R started off by computer scientists/statisticians (who needed it most)
- S's original goal: interact with data, programatically
- R evolved from a group of people using (extending) S-Plus, into a group of people who believed they didn't need S-Plus for this
- R is statistics oriented, but domain agnostic (empirical sciences)
- R started as a research project – "can we do this?"

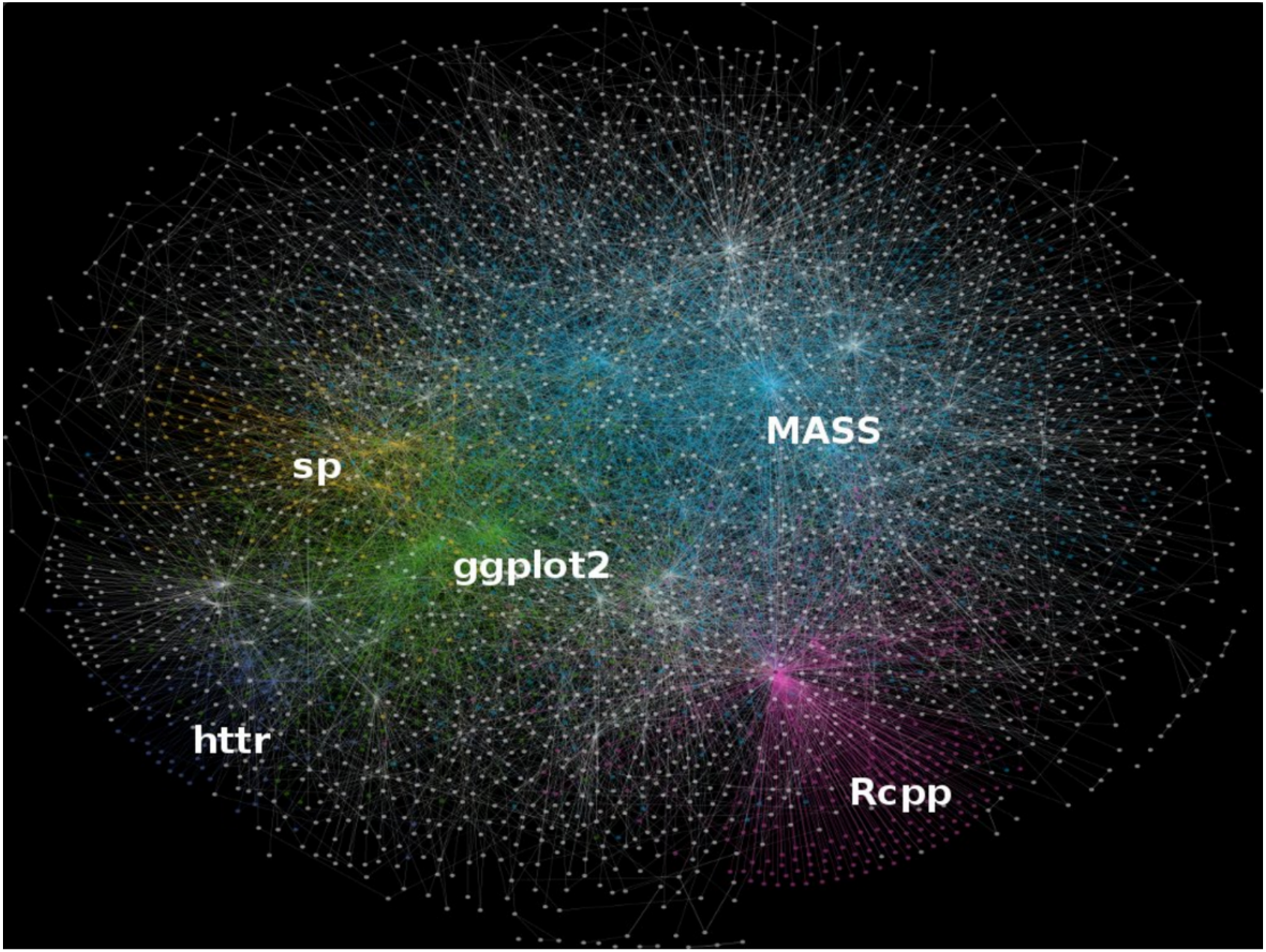
R packages

- R can be easily extended by *R packages*, software libraries for all kind of purposes; *methods*, *classes*, *interfaces*.
- CRAN, the Comprehensive R Archive Network, is a network of 50+ mirrored servers serving R source and binary distros (currently maintained by 21 authors), and now over 10,000 R packages, maintained by around 8000 authors.
- CRAN only accepts R packages in source code form, and keeps an archive of source code of all accepted versions
- CRAN compiles binary R packages (containing e.g. C or C++ code) for Windows and Mac OS-X platforms
- binary packages contain statically linked external dependencies
- when R changes, package maintainers may have to update their package; if they don't, after some time, packages are "archived": no longer visible or offered in binary form.
- unresponsive authors may cause packages to become orphaned; these may be adopted by new maintainers.

Package dependencies

- many packages reuse other packages, esp. those that
  - provide basic infrastructure (time series, spatial, omics, plotting, web services)
  - give access to a popular analysis method
  - interface e.g. databases, web services, file formats, other programming languages
  - make life easier
- my package A can depend on your package B
- making my package depend on someone else's is an expression of trust, similar to citing a paper as being a foundation for a certain idea, but with more dynamic risks:
  - package B might change its interface
  - changes in R may cause package B to fail
  - the author of package B might stop maintaining it

  all potentially causing my package A to fail
- CRAN lists reverse dependencies, and gives access to the dependency graph



Colin Gillespie @csgillespie · Apr 18

Updated #rstats dependencies map of CRAN (original by @RevoAndrie see blog.revolutionanalytics.com/2015/07/the-ne…)
pic.twitter.com/4hXpnu8O4A

7 23

Reproducibility

Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible[2].

the statistical community is quite apt to warrant reproducibility:

- methods underpin arguments, underpin decisions
- it helps argument this is about science, not engineering
- R, R scripts, and data files, are a way to secure this

Is data + R script (with R & package versions) enough?
paper + frozen versions: `http://www.JStatSoft.org`
Will the script still run, 10 years from now?

[2] Eos, Vol. 93, No. 16, 17 April 2012

Reproducibility

Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible[2].
the statistical community is quite apt to warrant reproducibility:

- methods underpin arguments, underpin decisions
- it helps argument this is about science, not engineering
- R, R scripts, and data files, are a way to secure this

Is data + R script (with R & package versions) enough?
paper + frozen versions: `http://www.JStatSoft.org`
Will the script still run, 10 years from now?

[2] Eos, Vol. 93, No. 16, 17 April 2012

Reproducibility

Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible[2].
the statistical community is quite apt to warrant reproducibility:

- methods underpin arguments, underpin decisions
- it helps argument this is about science, not engineering
- R, R scripts, and data files, are a way to secure this

Is data + R script (with R & package versions) enough?
paper + frozen versions: `http://www.JStatSoft.org`
Will the script still run, 10 years from now?

[2] Eos, Vol. 93, No. 16, 17 April 2012

Reproducibility

Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible[2].
the statistical community is quite apt to warrant reproducibility:

- methods underpin arguments, underpin decisions
- it helps argument this is about science, not engineering
- R, R scripts, and data files, are a way to secure this

Is data + R script (with R & package versions) enough?
paper + frozen versions: `http://www.JStatSoft.org`
Will the script still run, 10 years from now?

[2] Eos, Vol. 93, No. 16, 17 April 2012

Reproducibility

Reproducibility is an important aspect of scientific research, because the credibility of science is at stake when research is not reproducible[2].
the statistical community is quite apt to warrant reproducibility:

- methods underpin arguments, underpin decisions
- it helps argument this is about science, not engineering
- R, R scripts, and data files, are a way to secure this

Is data + R script (with R & package versions) enough?
paper + frozen versions: `http://www.JStatSoft.org`
Will the script still run, 10 years from now?

[2] Eos, Vol. 93, No. 16, 17 April 2012

“Opening Reproducible Research”

2-year DFG-funded, 2016-2017, LIS “Open Access Transformation”; Kray, ULB, me[3].

- can papers be made executable? $\Rightarrow$ executable research compendium, ERC
- how can data, software and procedures be encapsulated? $\Rightarrow$ docker containers
- how can data + software + scripts be handled in the publication cycle?
- how can a library offer a service for validating and archiving ERCs?
- which *interactions* would scientists like to make available, or have, with ERCs?
- how can we make it attractive to publish reproducibly?

`http://o2r.info`

[3] `http://www.dlib.org/dlib/january17/nuest/01nuest.html`

Sustainability

Will R and CRAN exist, 10 or 20 years from now?

- 20 maintainers have write R access to R, largely academics
- R foundation has 37 members; manages copyrights, legal, financial
- R community is keen on cooperation and communication
- yearly UseR! conferences, many domain specific conferences
- strong increase in submissions to JStatSoft and *The R Journal*
- strong increase in number of R related books
- R consortium (industry) funds/supports local R user groups, satuRdays, Rladies, community infrastructure projects
- rise of "data science": chairs, and study programs

Referencing scientific software

Default citation entry:

```
> citation("rgdal")
To cite package 'rgdal' in publications use:
Roger Bivand, Tim Keitt and Barry Rowlingson (2017). rgdal: Bindings for the Geospatial Data Abstrac-
tion Library. R package version 1.2-7. https://CRAN.R-project.org/package=rgdal

A BibTeX entry for LaTeX users is ...
```

Custom citations (author added):

```
> citation("gstat")
To cite package gstat in publications use:
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences,
30: 683-691.
Benedikt Gräler, Edzer Pebesma and Gerard Heuvelink, 2016. Spatio-Temporal Interpolation using gstat.
The R Journal 8(1), 204-218
```

Default package citations end up in google scholar.

What are the requirements to a software paper?

JORS, JOSS, R Journal, JStatSoft, ...

Referencing scientific software

Default citation entry:

```
> citation("rgdal")
To cite package 'rgdal' in publications use:
Roger Bivand, Tim Keitt and Barry Rowlingson (2017). rgdal: Bindings for the Geospatial Data Abstrac-
tion Library. R package version 1.2-7. https://CRAN.R-project.org/package=rgdal

A BibTeX entry for LaTeX users is ...
```

Custom citations (author added):

```
> citation("gstat")
To cite package gstat in publications use:
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences,
30: 683-691.
Benedikt Gräler, Edzer Pebesma and Gerard Heuvelink, 2016. Spatio-Temporal Interpolation using gstat.
The R Journal 8(1), 204-218
```

Default package citations end up in google scholar.

What are the requirements to a software paper?

JORS, JOSS, R Journal, JStatSoft, ...

Trends in R programming practice

- byte compiler; large objects, distributed computing
- `stringsAsFactors = TRUE`
- R. Peng: ... R is [now] being used a by a very wide variety of people doing all kinds of things the creators of R never envisioned.`http://simplystatistics.org/2015/07/24/stringsasfactors-an-unauthorized-biography/`
- H. Wickham, tidyverse: ... These days, making factors automatically is no longer so helpful, so packages in the tidyverse never create them automatically.[4]
- base graphics vs. plotting using packages.

[4] http://forcats.tidyverse.org/

Software sharing and legal aspects.

- documentation, tracing of OS licenses (important esp. for commercial R runtime providers)
- CRAN repository policy[5]:
  $\Rightarrow$ (implicit) contract between CRAN and authors
  $\Rightarrow$ "*The ownership of copyright and intellectual property rights of all components of the package must be clear and unambiguous*"
  $\Rightarrow$ "*The package's DESCRIPTION file must show both the name and email address of a single designated maintainer (a person, not a mailing list).*"
- contributed packages use weak authentication (confirmation by email): similar to journals, ORCID etc; discussions on code signing (X.509 or PGP?)

[5] `https://cran.r-project.org/web/packages/policies.html`

Conclusions

- The R community is a healthy, growing community that fills lots of demands that scientists have
- it stimulates to work reproducibly, by offering a sustainable infrastructure
- tensions between progressives and conservatives are here too, naturally
- there's still a lot to do to make scientists
  - share data, scripts, workflows along with publications
  - work reproducibly
  - properly cite the software they used
  - write (better) software
- we now address lots of these challenges at the educational (BSc, MSc) level