Exploring the World of CRAN: Collaboration and Dependencies

David Schoch (https://mr.schochastics.net)

Keywords — CRAN, RStats, Quarto

Figure 1: R logo

I. CRAN collaboration network

The CRAN collaboration network consists of R package developers who are connected if they appear together as authors of an R package in the DESCRIPTION file. The data was crawled on September 5th 2023. The code is available on GitHub.

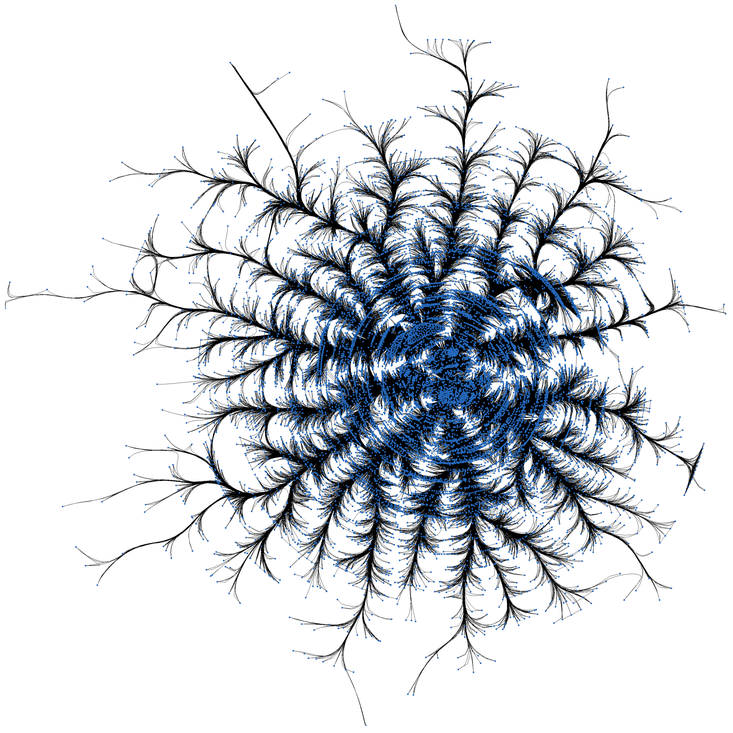

Figure 2: CRAN collaboration network (September 2023)

The network consists of 15419 R developers and 126,988 collaborative ties. Note that the graph only shows the *biggest component* of the network.

II. Six Degrees of Hadley Wickham

If you are familiar with the Erdős number number and/or the Bacon number then you know where this is going. The "Hadley number" is defined as the distance of R developers to Hadley Wickham in the collaboration graph. Someone ("A") who contributed to a package that Hadley has contributed to has a Hadley number of 1. Someone who contributed to a package that A has contributed to but not Hadley has Hadley number 2, and so on. Hadley himself is the only person with Hadley number 0.

The distribution of Hadley numbers is shown in Figure 1.

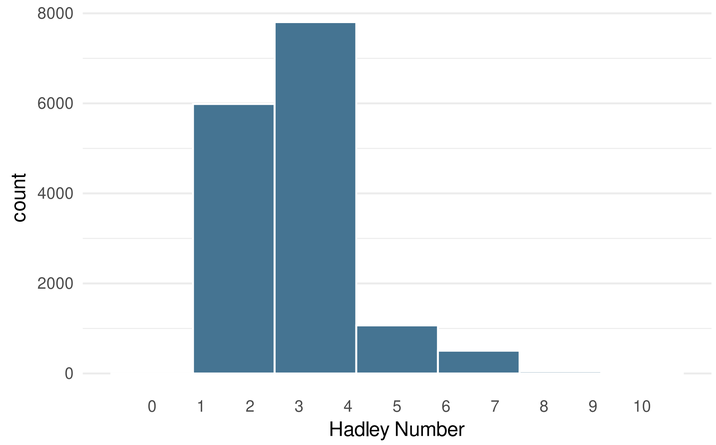

Figure 1: Distribution of the Hadley Number of R package developers

To get your own Hadley number, head to GitHub, download the network and run

```
library(igraph)
g <- readRDS("coauthor-biggest_comp.RDS")
me <- "David Schoch"
idx <- which(V(g)$name==me)
V(g)$dist2HW[idx]
```

III. The center of the collaboration network

The center of the collaboration network is defined as the developer who's average distance to all other developers is the lowest. The top ten developers in that regard are shown below.

Table 1: Most central developers in the CRAN collaboration network

Developer	centrality
Hadley Wickham	2.98178
Ben Bolker	3.10481
Dirk Eddelbuettel	3.13269
Martin Maechler	3.17355
Romain Francois	3.17375
Michael Friendly	3.18030
R Core Team	3.19534
Jim Hester	3.20585
Posit Software	3.21376
Kevin Ushey	3.23419

IV. R package dependency network

The R package dependency network is based on the Imports field in the DESCRIPTION file.

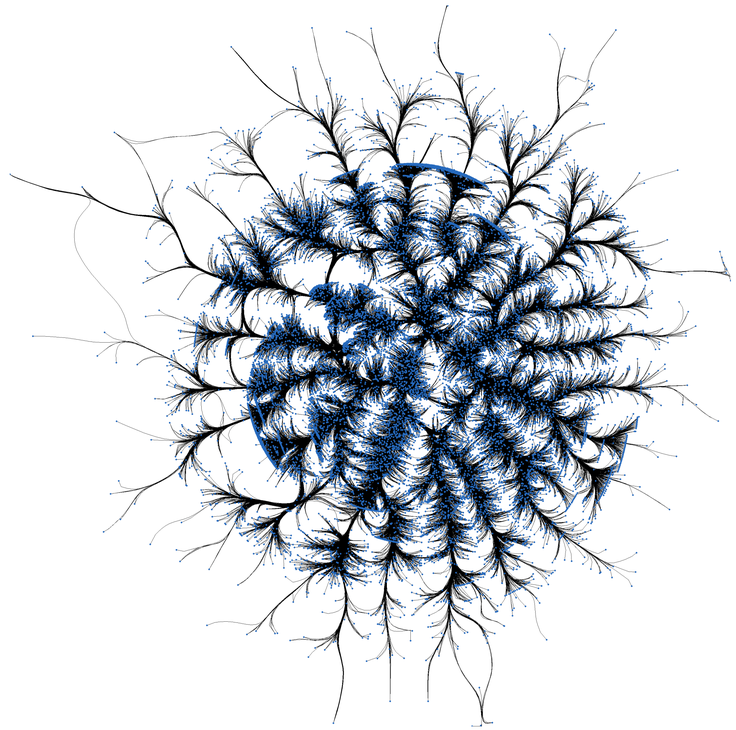

Figure 3: R package dependency network (based on Imports, September 2023)

The dependency network consists of 16855 packages with 94267 imports.

Table 2: Most frequently imported R packages

package	imported
stats	5236
utils	3311
methods	3146
dplyr	3100
ggplot2	3000
Rcpp	2477
graphics	2148
rlang	2017
magrittr	1892
stringr	1613