packdep: network abstractions of CRAN and Bioconductor

Radhakrishnan Nagarajan1*, Marco Scutari2

1. Division of Biomedical Informatics, Department of Biostatistics, University of Kentucky, Lexington, USA
2. UCL Genetics Institute, University College London, London, UK
*Contact author: ragnarjan@uky.edu

Keywords: Network Analysis, CRAN, Bioconductor

Objective: The objective of packdep is to model the associations between the R packages in CRAN and Bioconductor as networks and subsequently investigate their topological/statistical properties.

Rationale: While it is clear that the number of R packages in CRAN and Bioconductor is growing steadily with time, and that packages evolve significantly during their development, little attention has been given to understanding the associations between them. Such an understanding can provide novel system-level insights that capture their intricate wiring patterns. They may also be useful in identifying critical packages whose perturbation can challenge the stability of CRAN and Bioconductor. While we present some preliminary findings, we expect packdep to mature into a useful surveillance tool for CRAN and Bioconductor.

Network Abstraction of CRAN and Bioconductor: The exponential growth of R packages has largely been attributed to the active user community and special interest groups that span a spectrum of disciplines including Bioinformatics. Reuse of existing packages and/or functionalities minimizes redundancies and is characteristic of open-source communities where user contributions are voluntary. However, reuse introduces dependencies/associations between packages that warrant a detailed investigation, since user contributed packages are often subject to significant changes during their development. In R, associations fall under three broad categories of decreasing significance: (i) depends (ii) imports and (iii) suggests and can be modeled as a directed graph.

CRAN: As expected, analysis of CRAN network generated using the category (depends) revealed the core packages to be highly connected and dominant mediators. Removing the core packages and repeating the exercise on the weakly connected network component revealed the degree and betweenness distribution to be positively skewed reflecting a few packages are highly connected and dominant mediating comprising the tail of the distribution. Such distributions implicitly reflect the inherent non-random association patterns in CRAN. Top ten packages with maximal impact on others (i.e. out-degree) were mvtnorm, Rcpp, coda, sp, ggplot2, rgl, XML, plyr, igraph, rJava. Statistical packages (mvtnorm, sp, coda), graphical packages (ggplot2, rgl) are routinely used across a number of disciplines while (XML, Rcpp, rJava) integrate R to other popular software environments. The high rank igraph being may be attributed to the increasing emphasis on system and network science. Top ten dominant mediators in CRAN consisted of gplots, Deducer, RCurl, hdrcde, ROCR, spdep, fda, geiger, distr, ks. These packages while not highly wired by themselves, they do facilitate cross-talk across the highly wired packages establishing their critical role.

Bioconductor: A similar analysis on Bioconductor revealed the highly connected packages (out-degree) to be AnnotationDbi, Biobase, oligoClasses, oligo, org.Hs.eg.db, BiocGenerics, affy, BSgenome, IRanges, graph and the dominant mediators to be AnnotationDbi, oligo, Biobase, GenomicFeatures, oligoClasses, affy, ShortRead, Biostrings, Rsamtools. The high ranks of the annotation packages (AnnotationDbi, Biobase, GenomicFeatures) may be attributed to their critical role in the integration across diverse high-throughput molecular assays and the increasing emphasis on translational research (org.Hs.eg.db). The widespread use of microarrays and recent shift towards high-throughput sequencing and GWAS may also explain the high ranks of (oligo, Biobase, oligoClasses, affy) and (BSgenome, ShortRead, RSamtools). Unlike CRAN there was a significant overlap between highly wired packages and dominant mediators in Bioconductor.

A more detailed investigation of these network abstractions is currently under investigation.