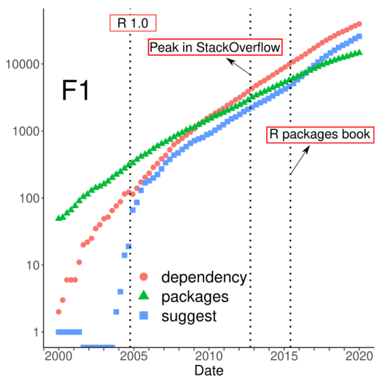
The R Network evolution: characterization of a collaborative network of software

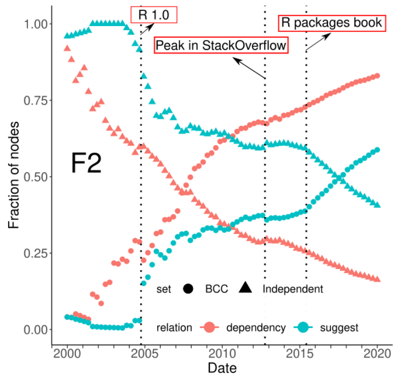
A. Salgado[1], *I. Caridi*[1]. *(1) Instituto de Cálculo UBA-CONICET, Buenos Aires, Argentina, asalgado@df.uba.ar*

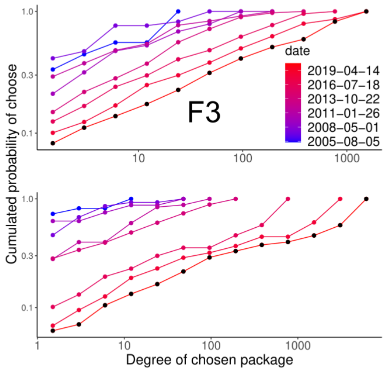
We present the evolution of the *R* project's collaborative software network, based on the official site data where *R*-users upload their package contributions (Comprehensive R Archive Network, CRAN). *R* is a free open source programming language created in the late '90s from *S*, a statistical analysis-oriented software. Nowadays, *R* packages cover a wide range of topics, including statistics, economics, machine-learning, biology, ecology, physics, geography, and many others. A package consists of a set of functions designed to provide a specific tool and can take advantage of other available packages. CRAN has grown from a few packages in 2000 to more than 14,000 interrelated packages today (F1). This growth results from an ever-increasing collaboration between developers from different fields and the emergence of a community of *R*-users worldwide, which develop packages and build other resources on *R* like books, tutorials, FAQ, etc.

We study the evolution of this package network made up of three elements: the nodes (packages) and two types of links among them (dependencies and suggestions) (F1). Dependencies are software-directed relationships, similar to paper citations, where a package depends on another if it uses its functionalities to work. Suggestions indicate the existence of examples or tutorials using both packages. While the network was sparse in its origin, the number of connections surpasses the number of packages nowadays, making the network mostly connected. We characterize the network's changes in time using macroscopic measures accounting for the size of the biggest connected component (BCC), and the fraction of independent packages (*BCC* and *independent* packages in F2). We analyze how the relationships of both types are distributed between packages, finding long-tailed distributions. The process of package addition is characterized using preferential attachment (PA) notions, finding sub and superlinear PA in dependencies and suggestions, respectively (F3, upper and lower panel, respectively). The number of relationships (dependencies and suggestions) of a new package is described with a unique distribution, using the BCC in each network as its unique parameter.

We associate the network increase in connectivity with external "R events" like book publications, journal creations, and interest in StackOverflow. Changes in the network connectivity can be related to both changes in the CRAN guidelines for development and the availability of resources for learning *R*. F1 and F2 show as an example the publication of R 1.0, most *R*-related questions in StackOverflow, and the publication of the *R packages* book.


F1
R 1.0
Peak in StackOverflow
R packages book
Date
dependency
packages
suggest


F2
R 1.0
Peak in StackOverflow
R packages book
Fraction of nodes
Date
set
BCC
Independent
relation
dependency
suggest


F3
Cumulated probability of choose
Degree of chosen package
date