Red de paquetes de R: CRAN, estructura y evolución temporal

Ariel Salgado[1], Inés Caridi[1]

[1]Instituto de Cálculo, FCEN, UBA-CONICET

Palabras Claves: redes complejas, redes empíricas, datos abiertos, paquetes

1 Introducción

Hoy en día existen más de 12000 paquetes o librerías de R. Los mismos son el resultado de un proyecto colaborativo sostenido por personas de distintos lugares del mundo y disciplinas. El proyecto continúa creciendo tanto en cantidad de paquetes como en áreas de conocimiento, como estadística, finanzas, genética, análisis de redes y data mining, entre muchas otras. Sin embargo, poco se ha hecho en torno al análisis de su evolución temporal [1][2]. Para estudiar su crecimiento, nos enfocamos en los paquetes y sus relaciones de dependencias y sugerencias. A partir de cada tipo de relación construimos una red dirigida entre paquetes. Observamos la evolución de estas redes desde sus comienzos en 1999 hasta la actualidad. Extrajimos la información de CRAN, empleando paquetes como *XML* y *stringr* para construir una base de datos a partir de la cual generar las redes. Empleamos los paquetes *Matrix e igraph* para realizar el análisis, caracterizando la estructura mediante magnitudes como la cantidad de paquetes y conexiones, el promedio de conexiones (grado medio), el tamaño del conjunto de paquetes interconectados más grande (componente gigante) y la cantidad de paquetes sin dependencias, entre otras.

2 Resultados

Comenzamos el análisis con la red formada por las relaciones de dependencia (se establece una conexión dirigida entre dos paquetes cuando uno depende del otro). En la evolución de la red de dependencias, se pueden identificar puntos en los cuales ocurren cambios bruscos en el crecimiento, que definen tres etapas según la conecti-

vidad de la red: una de aumento, una con poca variación, y otra de decrecimiento. Durante la primera, el número de conexiones crece más rápido que el de paquetes, generando una componente gigante en la red. Durante la segunda, la fracción que representa la componente gigante y el grado medio cambian poco. En la tercera, el grado medio empieza a disminuir, sin que la componente gigante pierda paquetes. Lo que ocurre es una fuerte incorporación de paquetes sin dependencias.

La red construida a partir de las relaciones de sugerencia muestra en la primera etapa un coincidencia total con la red de dependencias, comenzando a diferenciarse en la segunda. En la tercera, esta coincidencia disminuye aún más, haciendo las relaciones más complementarias entre sí. En esta red aparece una cuarta etapa (ausente en la red de dependencias) donde aumenta la conectividad, señalando una intensificación del uso de la sugerencia de paquetes en R. En la disminución del número de paquetes desconectados se ve que paquetes viejos empiezan a incorporar sugerencias en sus descripciones.

La apariencia final de la red de dependencias es de una componente gigante y un conjunto de paquetes sueltos. La red de sugerencias, en cambio, tiende a disminuir el número de paquetes sueltos, sumándolos a la componente gigante.

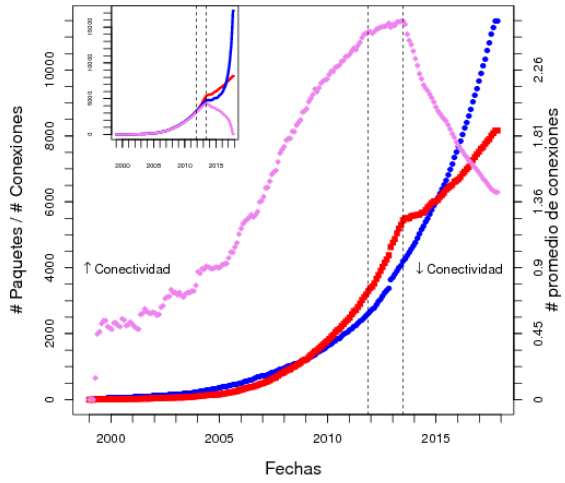

Figura: Número de nodos (símbolos azules), número de dependencias (rojos) y número promedio de dependencias (violetas) en función de la fecha. En el inset, número de dependencias (rojo), número de sugerencias (azul) y solapamiento entre ambos (violeta).

Referencias

[1] blog.revolutionanalytics.com/2016/04/cran-package-growth.html

[2] es.slideshare.net/RevolutionAnalytics/jsm-r-pkgs-2015-0809/7