The network structure of CRAN

Andrie de Vries

adevries@microsoft.com
@RevoAndrie

UseR!2015, Aalborg
A network of package dependencies

- Imports
- Depends
- Suggests
- Enhances
- LinkingTo
Using miniCRAN to create the network graph

```r
pdb <- available.packages()
tag <- "chron"
library("miniCRAN")
dg <- makeDepGraph(tags, availPkgs = pdb)
plot(dg)
```
How to visualize a network of 6,791 CRAN packages?
CRAN

Note: Fruchterman Reingold layout (force-directed algorithm)
library("igraph")
write.graph(dg,
file = "example.GraphML",
format = "graphml")
Which nodes (packages) are more “important”? 
Reverse dependencies influence importance

- Imagine that each package dependency is a vote of confidence, cast by the community of package authors.

- How can we compute the importance of a package to downstream packages?
Page rank

- Page rank is famous as one of the criteria in Google search results.
- You can use page rank on any graph to compute "importance".

`igraph::page.rank()`
## Pagerank results

<table>
<thead>
<tr>
<th>Package</th>
<th>PageRank (Dec 2014)</th>
<th>PageRank (Jun 2015)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rcpp</td>
<td>0.0166</td>
<td>0.0197</td>
<td>Interface to use C++ code in R</td>
</tr>
<tr>
<td>MASS</td>
<td>0.021</td>
<td>0.0196</td>
<td>Functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S' (4th edition, 2002).</td>
</tr>
<tr>
<td>Matrix</td>
<td>0.01</td>
<td>0.0095</td>
<td>Sparse matrix engine</td>
</tr>
<tr>
<td>ggplot2</td>
<td>0.0073</td>
<td>0.0086</td>
<td>Graphics engine</td>
</tr>
<tr>
<td>lattice</td>
<td>0.0096</td>
<td>0.0085</td>
<td>Base R package for lattice (trellis) graphics</td>
</tr>
<tr>
<td>mvtnorm</td>
<td>0.0088</td>
<td>0.0083</td>
<td>Multivariate normal distributions</td>
</tr>
<tr>
<td>survival</td>
<td>0.0083</td>
<td>0.0079</td>
<td>Time-to-event analysis</td>
</tr>
<tr>
<td>plyr</td>
<td>0.0067</td>
<td>0.0072</td>
<td>Group-by operations</td>
</tr>
<tr>
<td>igraph</td>
<td>0.0047</td>
<td>0.0049</td>
<td>Analyse graph structures</td>
</tr>
<tr>
<td>XML</td>
<td>0.0047</td>
<td>0.0047</td>
<td>Parse and manipulate documents in XML format</td>
</tr>
</tbody>
</table>

CRAN
Can we detect clusters of packages?
CRAN
Community detection

• Communities are densely connected groups of nodes in a graph

• Several implemented algorithms in igraph:
  • Fast greedy
  • Walktrap
  • Spinglass
  • Leading eigenvector
  • Edge betweenness
The walktrap algorithm

• `walktrap.community`

  • This function tries to find densely connected subgraphs, also called communities in a graph via random walks. The idea is that short random walks tend to stay in the same community.

• Efficient algorithm for large, densely connected graphs*

Is BioConductor different?
Summary
library(miniCRAN)
makeDepGraph()

library(igraph)
pag.rank()
walktrap.community()
write.graph()
Scripts available at:
https://github.com/andrie/cran-network-structure

adervies@microsoft.com
@RevoAndrie