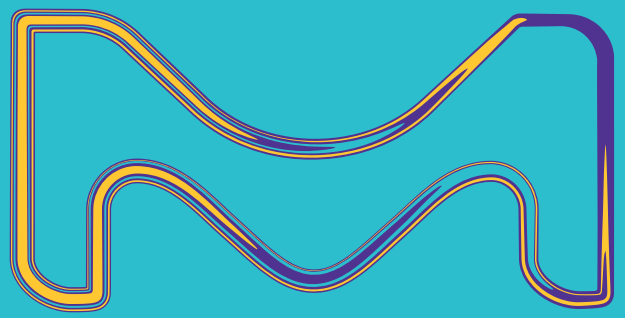
improve installation sequences for R package cohorts

A statistical analysis of the R package dependency network structures

Juliane Manitz[1]
with contributions by Martin Gregory[2] *and Ed Lauzier*[3]

[1] *EMD Serono* - A Subsidiary of Merck KGaA, Darmstadt, Germany
[2] *Merck KGaA Darmstadt, Germany*
[3] *Merck & Co.*

EMD Serono is a business of Merck KGaA, Darmstadt, Germany



EMD SERONO

Aim & Objectives of this Talk

Outline:

1. Background/Motivation
2. Statistical Analysis of
   a. CRAN Network and
   b. Package Dependency Trees
3. Reduce packages to a sufficient shortlist
   a. Algorithm for improved installation sequence
   b. Details for auditing and change control in regulated workflows
4. Summary and Conclusions

Abstract: The installation of a cohort of R packages can constitute a challenge; especially considering different dependency types, package versions, overlapping namespaces and varying risks assigned to each of the packages. At the same time, the number of R packages to be installed grows exponentially with each new package added. Their complex dependencies may create conflicts. In this context, the R admin is often confronted with a cohort of packages without knowing the package of interest.

We use statistical analysis techniques from the field of complex network analysis in order to shed light into the non-trivial dependency structures of package cohorts. Furthermore, we simplify the network graph to find improved installation sequences for a pre-selected cohorts of R packages. We reduce large package cohorts to a sufficient shortlist of packages, whose installation automatically pulls in other packages via dependencies without causing conflicts. The build time of a library may be greatly reduced. As a byproduct, we generate a graph of the build on the exact dependency tree and actual versions used for auditing and change control in the regulated workflows. This strategy also allows for the identification of high-risk packages and their importance in the dependency tree.

Some Background: It depends – A dialog about dependencies

- Consider differences in package dependencies:
  - type of the dependency, number of upstream dependencies, already fulfilled dependencies
  - time taken to (build and) compile the package
  - additional system dependencies
- Adding (or removing) dependencies come with trade-offs:
  - Provides additional features, bug fixes, and real-world testing,
  - Costs of increased installation time, disk space and maintenance of dependency

➢ Instead of striving for a minimal number of dependencies, it has been suggested a more holistic, balanced, and quantitative approach

- **Don’t underestimate the maintenance if the dependencies**
  - Two packages depend on different version of a third package
  - If you install packages sequentially new packages may create version conflicts
  - If you want to create local repo of package cohort, you need to decide on one version

Source: https://www.tidyverse.org/articles/2019/05/itdepends/

2a statistical analysis of CRAN network

R Dependency Tree as Complex Network

- Download CRAN data base and extract network structure using `cranly` package

- Represent R package dependency structured as a complex network, i.e. a collection of nodes connected by links
  - Nodes = packages:

    Base, Recommended, Trusted Source, Add-on, etc.
  - Links = Dependencies:

    Depends, Imports, Suggests, LinkingTo, Enhances

➢ Analyze dependency network using statistical techniques for complex networks

CRAN Network:

Timestamp July 24, 2019

Packages: 14,055

Links:	Frequency
imports	50,436
depends	11,547
suggests	29,724
enhances	473
linkingto	2,989
Total	**95,161**

Data Preparation & Network Analysis Measures

CRAN Network (Timestamp: July 24, 2019)						
	#Nodes	#Links	Density (10^4)	⌀Degree	Diameter	Transitivity (10^3)
Complete CRAN database	14,056	95,169	4.8	13.5	18	8.3
Optional Packages	14,027	74,890	3.8	10.7	19	8.9
Suggests removed	14,027	43,878	2.2	6.3	9	7.4
Connected component	12,074	43,793	3.0	7.3	9	7.4

Network summary measures

- Number of nodes and links
- Connected component: maximal set of nodes such that each pair of nodes is connected by a path
- Density: number of existent links compared possible edges
- Average Degree: mean number of links has a node
- Diameter: longest shortest path connecting any two nodes
- Transitivity: relative number of triangles

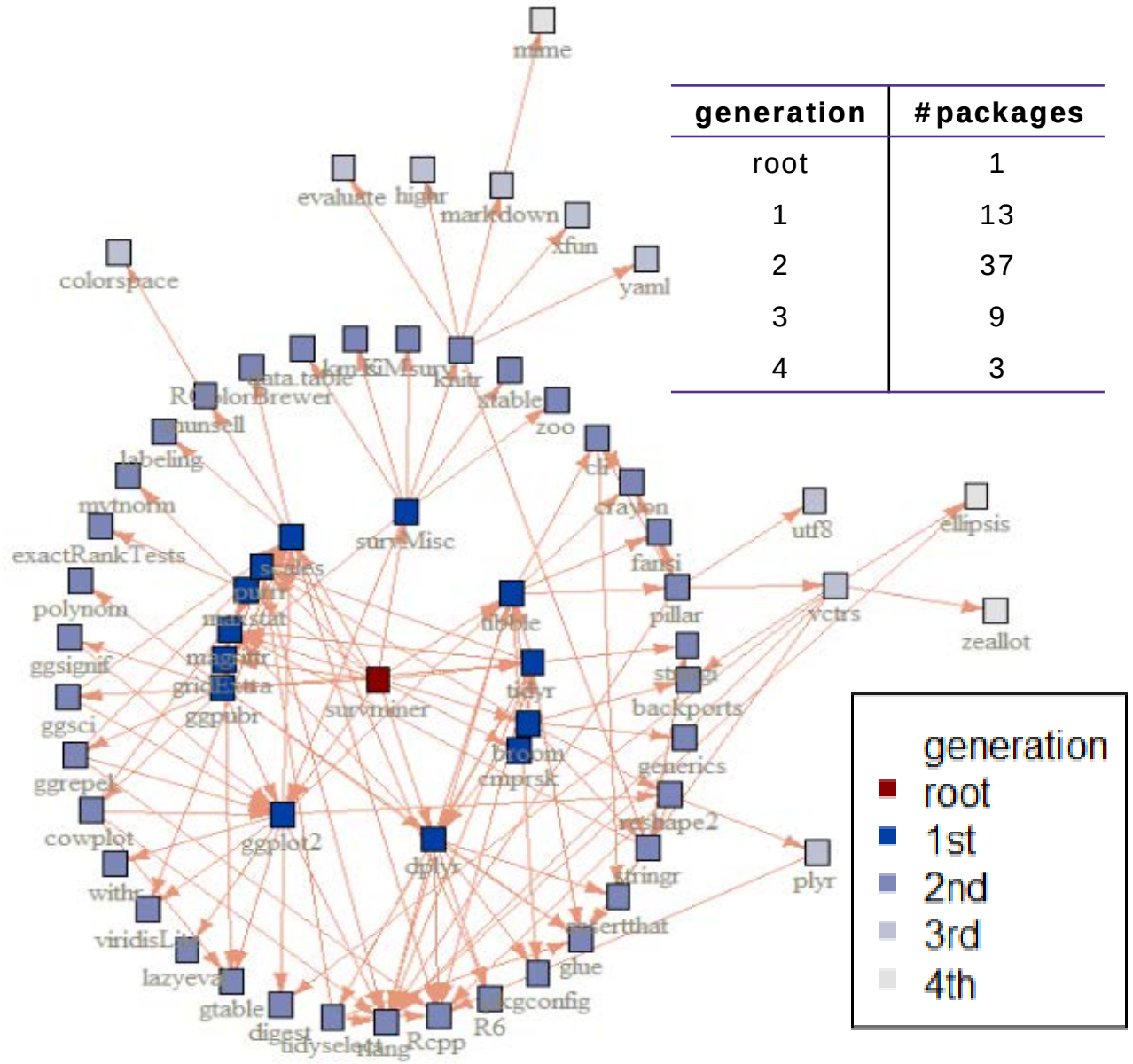
2b statistical analysis of dependency trees

Example Package {survminer}

```
Title: Drawing Survival Curves using 'ggplot2'
Maintainer: Alboukadel Kassambara
Version: 0.4.4
License: GPL-2
NeedsCompilation: no
Published: 2019-05-21
```

Network Characteristics

	Full Network	MST
# Nodes	63	63
# Links	130	62
Density	3.3%	1.6%
⌀ Degree	4.12	1.97
Diameter	5	3
Transitivity	26.0%	0%



generation	#packages
root	1
1	13
2	37
3	9
4	3

Example Cohort {haven, Hmisc, survminer}

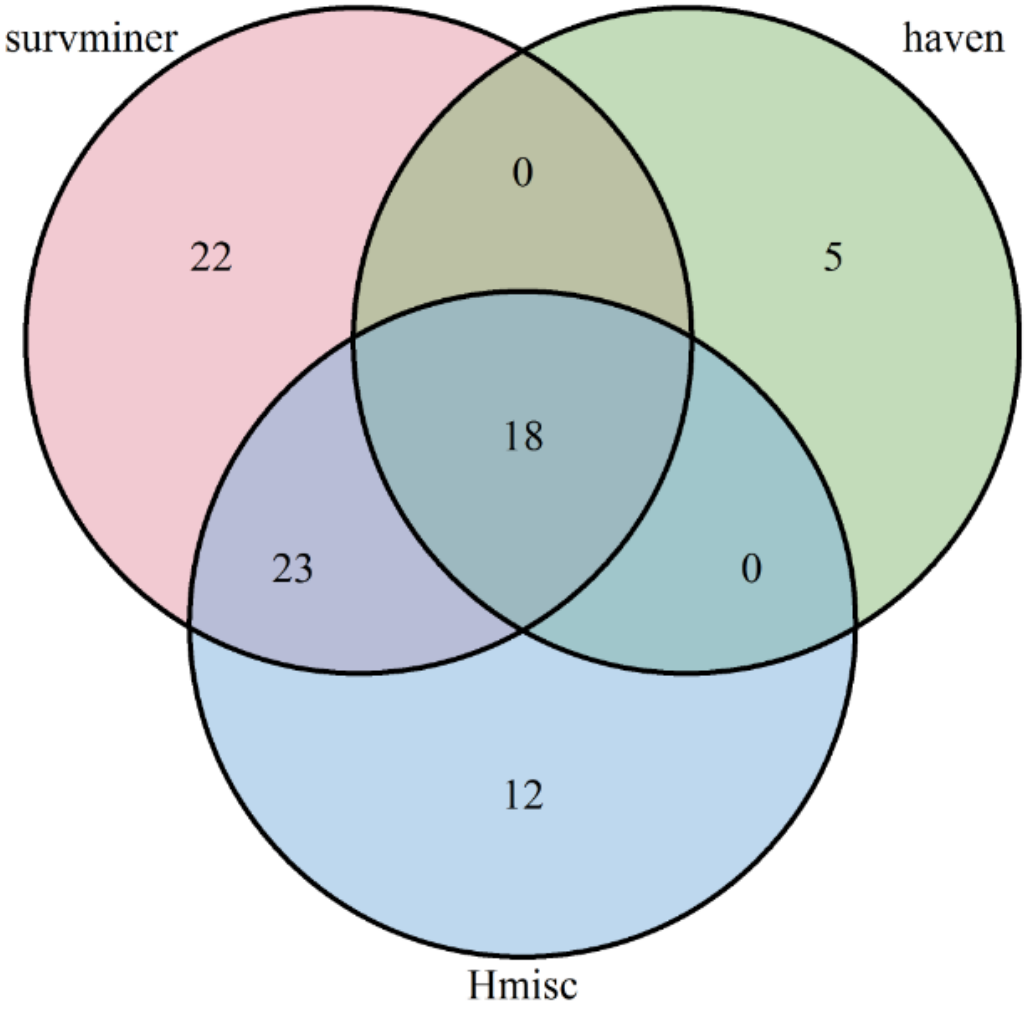
```
## mini ‘survival’ analysis
require(haven, Hmisc, survminer)

dat <- read_sas("analysis_data.sas7bdat")
bystats(y=dat$AAGE, dat$SEX)

km_os <- survfit(Surv(DUR_OS, 1-CENS_OS) ~ 1, data=dat)
ggsurvplot(km_os, data=dat, title=‘KM analysis for OS’)
```

Dependencies

		Direct	Total
haven	Import and Export 'SPSS', 'Stata' and 'SAS' Files	5	23
Hmisc	Harrell Miscellaneous	11	53
survminer	Drawing Survival Curves using 'ggplot2'	13	63
Cohort	**{haven, Hmisc, survminer}**	**26**	**80**

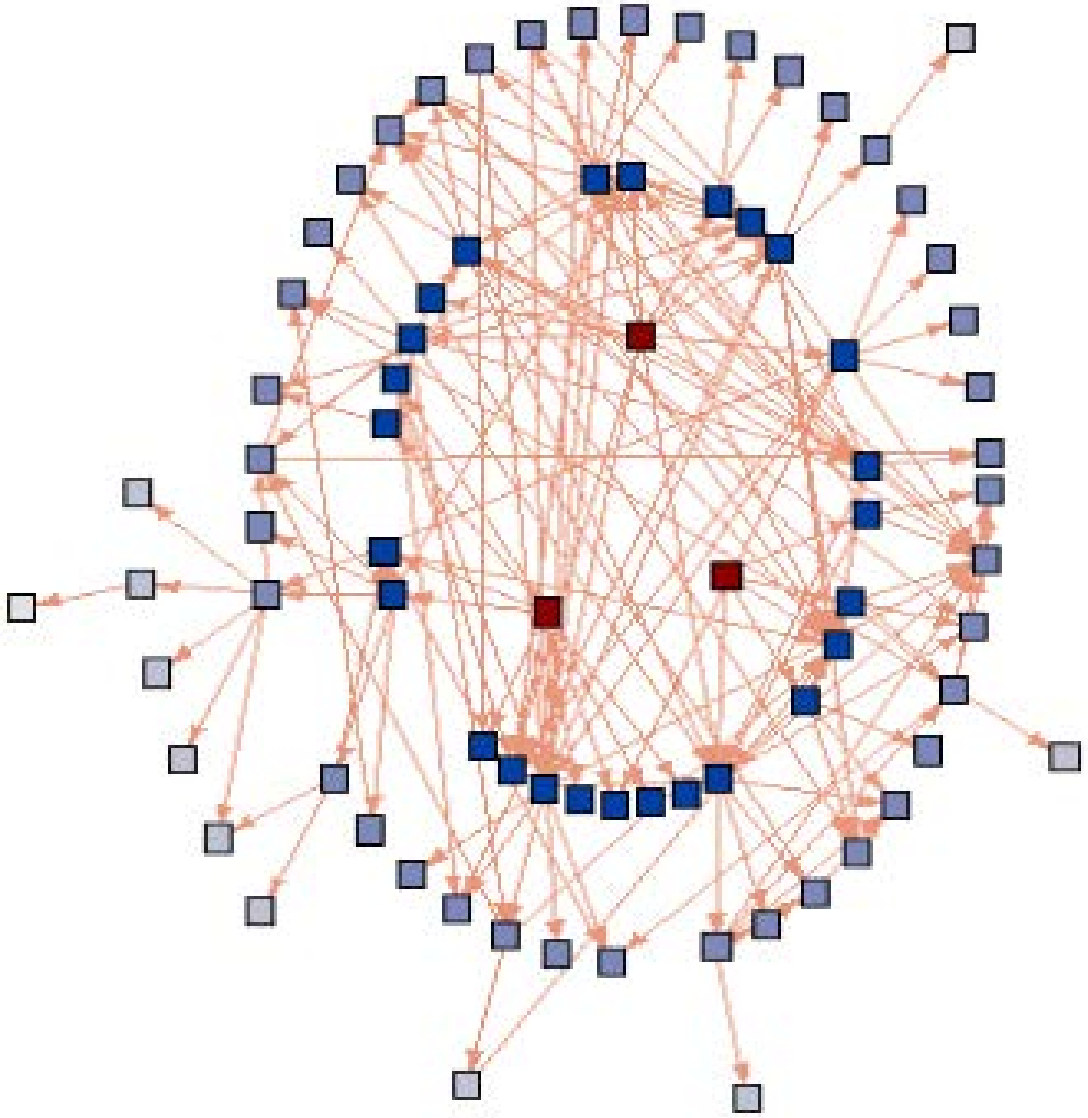


Example Cohort {haven, Hmisc, survminer}

```
## mini ‘survival’ analysis
require(haven, Hmisc, survminer)
dat <- read_sas("analysis_data.sas7bdat")
bystats(y=dat$AAGE, dat$SEX)
km_os <- survfit(Surv(DUR_OS, 1-CENS_OS) ~ 1, data=dat)
ggsurvplot(km_os, data=dat, title=‘KM analysis for OS’)
```

Network Characteristics

	Full Network	MST
# Nodes	80	80
# Links	176	79
Density	2.8%	1.2%
⌀ Degree	4.40	1.98
Diameter	5	4
Transitivity	23.1%	0%



Network Characteristics for Different Cohorts

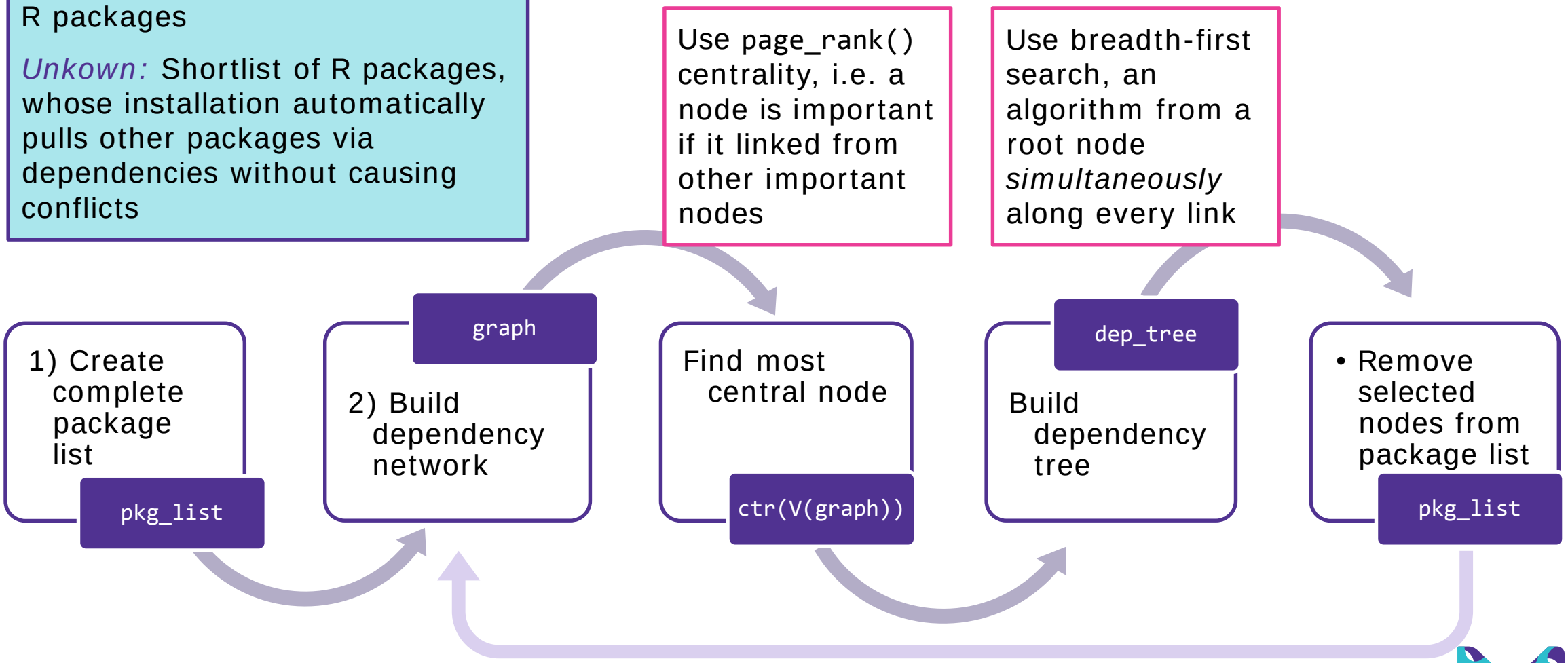
	# nodes	# links	density	ø degree	diameter	transitivity
Package {survminer}	63	130	3.3%	4.12	5	26%
Full network {haven, Hmisc, survminer}	80	176	2.8%	4.4	5	23.1%
MST {haven, Hmisc, survminer}	80	79	1.2%	1.98	4	0%
Tidyverse {ggplot2, dplyr, tidyr, readr, purr, tibble, stringr, forcats}	40	81	5.2%	4.05	4	23.9%
Quick list of useful R packages (R Studio)	162	403	1.5%	4.98	6	18.1%

03 reducing R package list

Iterative Algorithm to Reduce List of R packages

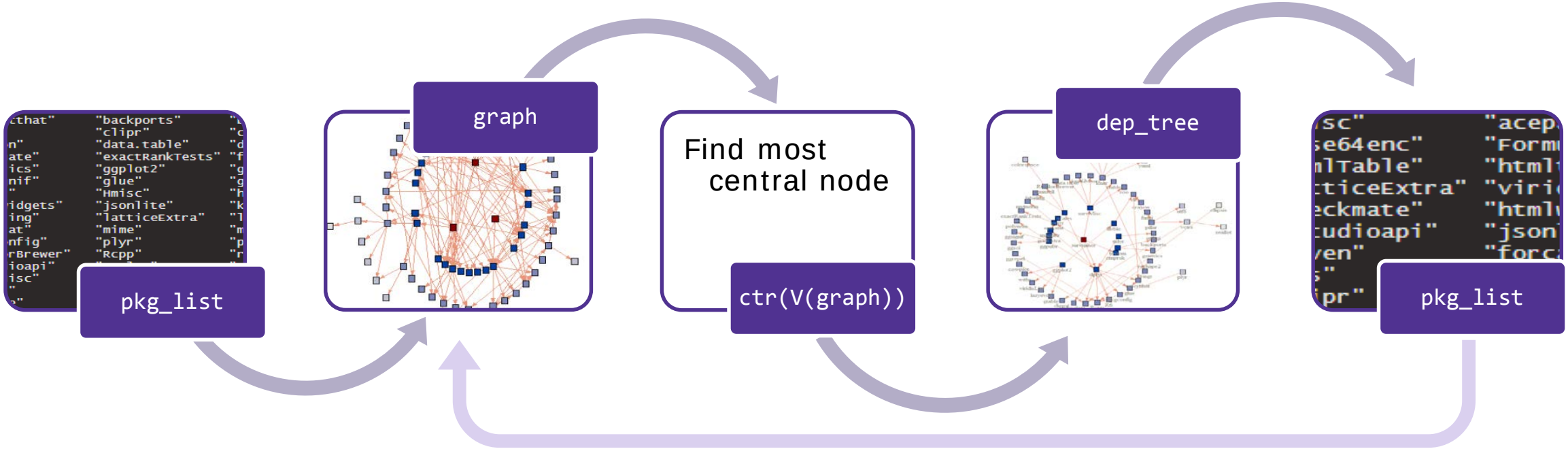
Given: Comprehensive cohort of R packages

Unkown: Shortlist of R packages, whose installation automatically pulls other packages via dependencies without causing conflicts



Testing Cohort {haven, Hmisc, survminer}

```
> ex1 <- reduce(cran_net=pkg_net, cohort=c("survminer","Hmisc","haven"))
step 1 : 80 packages left - survminer  selected
step 2 : 17 packages left - Hmisc  selected
step 3 : 5 packages left - haven  selected
```



Example Cohort: Quick list of useful R packages (R Studio)

Recommended List of R Packages

To load data: DBI, Odbc, RMySQL, RPostgresSQL, RSQLite, XLConnect, xlsx, foreign, haven

To manipulate data: Dplyr, tidyr, stringr, lubridate

To visualize data: ggplot2, ggvis, rgl, htmlwidgets, leaflet, dygraphs, DT, diagrammeR, network3D, threeJS, googleVis

To model data: car, mgcv, lme4/nlme, randomForest, multcomp, vcd, glmnet, survival, caret

To report results: shiny, rmarkdown, xtable

For Spatial data: sp, maptools, maps, ggmap

For Time Series and Financial data: zoo, xts, quantmod

To write high performance R code: Rcpp, data.table, parallel

To work with the web: XML, jsonlite, httr

To write your own R packages: devtools, testthat, roxygen2

```
step 1 : 167 packages left - devtools  selected
step 2 : 115 packages left - car  selected
step 3 : 84 packages left - caret  selected
step 4 : 56 packages left - rgl  selected
step 5 : 38 packages left - ggmap  selected
step 6 : 32 packages left - RSQLite  selected
step 7 : 27 packages left - multcomp  selected
step 8 : 22 packages left - leaflet  selected
step 9 : 18 packages left - quantmod  selected
step 10 : 15 packages left - XLConnect  selected
step 11 : 12 packages left - rmarkdown  selected
step 12 : 10 packages left - vcd  selected
step 13 : 8 packages left - xlsx  selected
step 14 : 6 packages left - DT  selected
step 15 : 5 packages left - dygraphs  selected
step 16 : 4 packages left - ggvis  selected
glmnet googleVis RMySQL  remain(s)
```

Installation List of R Packages

devtools, car, caret, rgl, ggmap, RSQLite, multcomp, leaflet, quantmod, XLConnect" rmarkdown, vcd, xlsx, DT, dygraphs, ggvis, glmnet, googleVis, RMySQL

Source: https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages

Details for Auditing and Change Control in Regulated Workflows

	title	maintainer	version	published	Needs compil.	license
devtools	Tools to Make Developing R Packages Easier	Jim Hester	2.1.0	7/6/2019	no	GPL (>= 2)
car	Companion to Applied Regression	John Fox	3.0-3	5/27/2019	no	GPL (>= 2)
caret	Classification and Regression Training	Max Kuhn	6.0-84	4/27/2019	yes	GPL (>= 2)
rgl	3D Visualization Using OpenGL	Duncan Murdoch	0.100.26	7/8/2019	yes	GPL
ggmap	Spatial Visualization with ggplot2	David Kahle	3.0.0	2/5/2019	no	GPL-2
RSQLite	'SQLite' Interface for R	Kirill Müller	2.1.2	7/24/2019	yes	LGPL (>= 2)
multcomp	Simultaneous Inference in General Parametric Models	Torsten Hothorn	1.4-10	3/5/2019	no	GPL-2
leaflet	Create Interactive Web Maps with the JavaScript 'Leaflet' Lib.	Joe Cheng	2.0.2	8/27/2018	no	GPL-3
quantmod	Quantitative Financial Modelling Framework	Joshua M. Ulrich	0.4-15	6/17/2019	no	GPL-3
XLConnect	Excel Connector for R	Martin Studer	0.2-15	4/5/2018	no	GPL-3
rmarkdown	Dynamic Documents for R	Yihui Xie	1.14	7/12/2019	no	GPL-3
vcd	Visualizing Categorical Data	David Meyer	1.4-4	12/6/2017	no	GPL-2
xlsx	Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files	Adrian A. Dragulescu	0.6.1	6/11/2018	no	GPL-3
DT	A Wrapper of the JavaScript Library 'DataTables'	Yihui Xie	0.7	6/11/2019	no	GPL-3 | file
dygraphs	Interface to 'Dygraphs' Interactive Time Series Charting Library	Petr Shevtsov	1.1.1.6	7/11/2018	no	MIT + file
ggvis	Interactive Grammar of Graphics	Winston Chang	0.4.4	9/28/2018	no	GPL-2 | file
glmnet	Lasso and Elastic-Net Regularized Generalized Linear Models	Trevor Hastie	2.0-18	5/20/2019	yes	GPL-2
googleVis	R Interface to Google Charts	Markus Gesmann	0.6.4	5/16/2019	no	GPL (>= 2)
RMySQL	Database Interface and 'MySQL' Driver for R	Jeroen Ooms	0.10.17	3/4/2019	yes	GPL-2

05 summary & outlook

Summary and Outlook

Summary

- Statistical analysis of network structures helps to understand dependency structures of packages and cohorts
- Improved installation sequence by reducing packages to a sufficient shortlist
- Generate and extract installation list with package versions that can be used for auditing and change control in the regulated workflows

Outlook

- Expand beyond CRAN to Bioconductor, github, …
- Deal with system dependencies, external software, etc.
- Integrate with risk assessment package and check for centrality of high-risk packages
- Time-dynamics comparison of CRAN networks with different timestamps
- Export Function names (from namespace) to make sure function names are not overlapping, and check for isolated names spaces,
- Shiny app?

Thanks for your attention!

Juliane Manitz

EMD Serono - - A Subsidiary of
Merck KGaA, Darmstadt, Germany

Global Biostatistics – Oncology
Billerica MA, USA

Juliane.manitz@emdserono.com