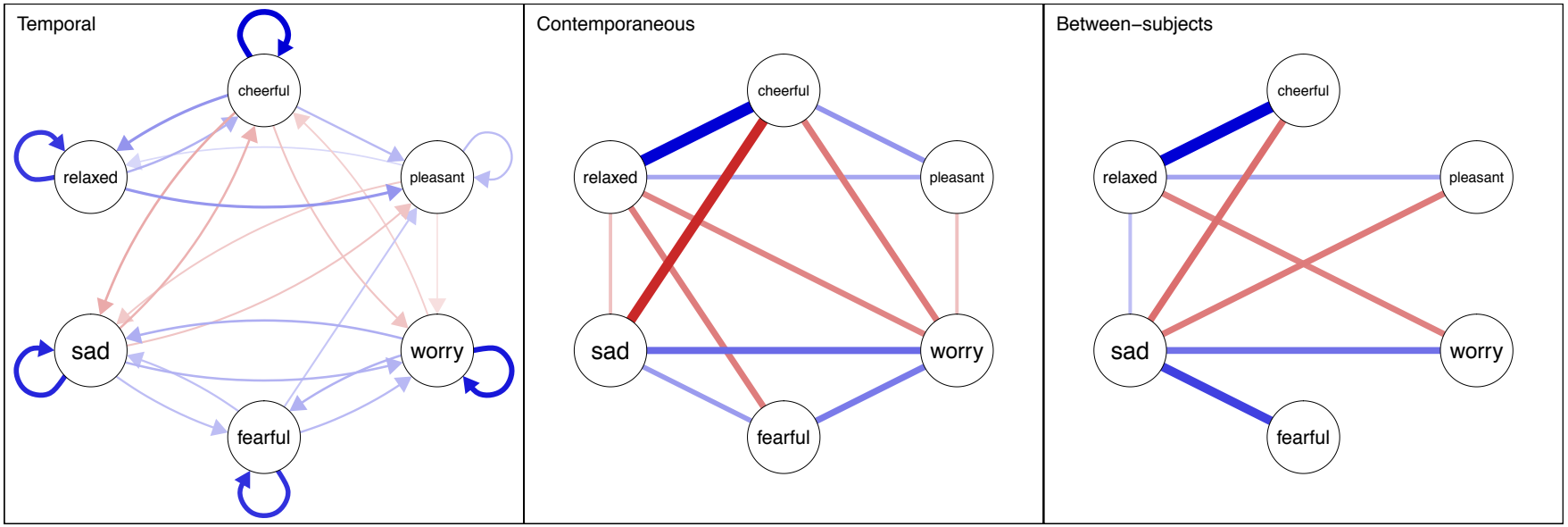
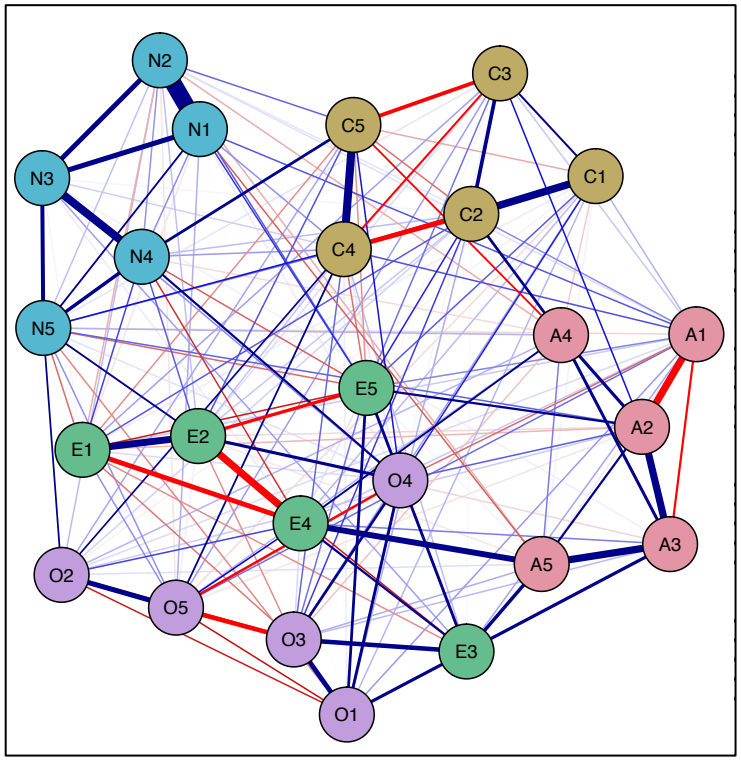
Reproducibility and Replicability in a Fast-paced Methodological World

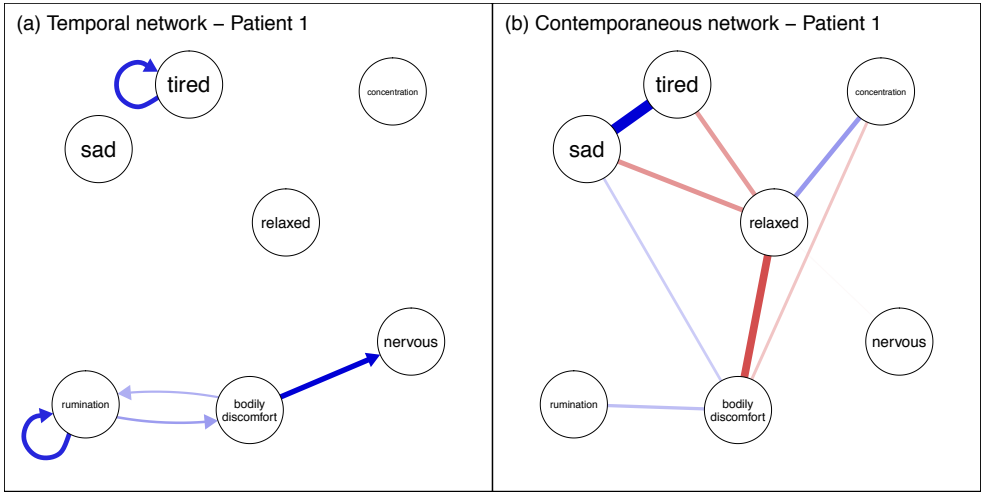
Sacha Epskamp

A long time ago

Developed together with Michele Nuijten in 2011-2012, after which I moved on to…







Network Psychometrics

http://sachaepskamp.com/Dissertation

Sacha Epskamp

SachaEpskamp

Add a bio

University of Amsterdam

Amsterdam

mail@sachaepskamp.com

http://www.sachaepskamp.c...

Organizations

Overview | Repositories 23 | Stars 6 | Followers 56 | Following 1

Pinned repositories

Customize your pinned repositories

semPlot

Path diagrams and visual analysis of various SEM packages' output

R ★27 8

qgraph

Developmental version of qgraph

R ★22 4

lvnet

Latent variable network models

R ★5 2

bootnet

Bootstrap methods for various network estimation routines

R ★3 4

graphicalVAR

R ★2 2

mlVAR

R ★1 2

qgraph: Network visualizations of relationships in psychometric data — 322 — 2012
S Epskamp, AOJ Cramer, LJ Waldorp, VD Schmittmann, D Borsboom
Journal of Statistical Software 48 (4), 1-18

State of the aRt personality research: A tutorial on network analysis of personality data in R — 116 — 2015
G Costantini, S Epskamp, D Borsboom, M Perugini, R Mõttus, LJ Waldorp, ...
Journal of Research in Personality 54, 13-29

Estimating psychological networks and their accuracy: A tutorial paper — 78 — 2017
S Epskamp, D Borsboom, EI Fried
Behavior Research Methods, 1-18

A Tutorial on Regularized Partial Correlation Networks — 35 * — 2016
S Epskamp, EI Fried
arXiv preprint arXiv:1607.01367

Generalized Network Pschometrics: Combining Network and Latent Variable Models — 33 — 2017
S Epskamp, M Rhemtulla, D Borsboom
Psychometrika 82 (4), 904-927

Reproducibility and Replicability

- Reproducibility

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

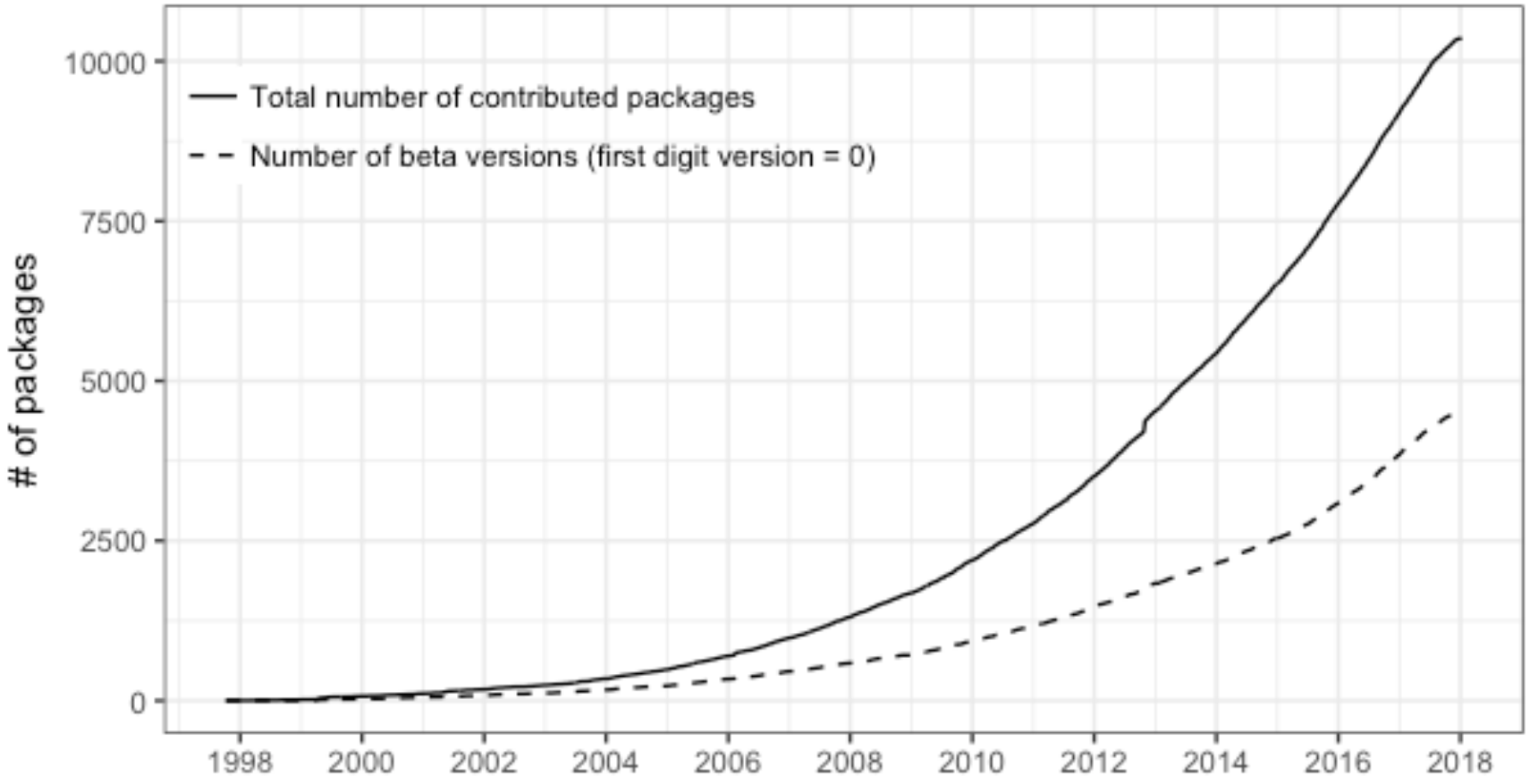
Estimating the reproducibility of psychological science

Open Science Collaboration*

-

them to be replicable

Fast-paced Methodological World



Rapidly growing number of statistical software packages

Fast-paced Methodological World

arXiv.org

Rapidly growing number of methodological pre-prints

Goal

While the age of fast-paced methodology greatly facilitates reproducibility, it also undermines it in ways not often realized by researchers. I will discuss:

- Sources of Limited Reproducibility and Replicability in Novel Methodology
- Sources of Limited Reproducibility and Replicability in Software
- Recommendations for improving reproducibility
  - Methodologists
  - Empirical Researchers

SOURCES OF LIMITED REPRODUCIBILITY AND REPLICABILITY IN NOVEL METHODOLOGY

Researcher Degrees of Freedom

- Usually many choices made in performing some analysis
  - Prior in Bayes Factors
  - Standardization
  - …
- May impact replicability
  - Results may only replicate using very specific design choices
  - Results may not replicate using future state-of-the-art
- Multiverse analysis remediates the first, but not the second
  - And may be complicated…

An investigation of emotion dynamics in major depressive disorder patients and healthy persons using sparse longitudinal networks

Stijn de Vos, Klaas J. Wardenaar, Elisabeth H. Bos, Ernst C. Wit, Mara E. J. Bouwmans, Peter de Jonge

Published: June 1, 2017 • https://doi.org/10.1371/journal.pone.0178586

Conclusions

The sparse network approach can be useful for the estimation of networks with multiple nodes, where overparameterization is an issue, and for individual-level networks. However, its current inability to model random effects makes it less useful as a population-level approach in case of large heterogeneity. Different preprocessing strategies appeared to strongly influence the results, complicating inferences about network density.

Lack of Validation Studies

- Novel methods may be only validated in certain settings, or proven under certain conditions
  - Multivariate normality
  - Sparsity
- Such conditions may be violated in empirical datasets
  - Not clear how well method performs
  - Results may not replicate using future state-of-the-art

Pre-print Reports

It used to take a long time before methodological papers were made public…

	Quantile				
	0%	25%	50%	75%	100%
Review time	0	287	408	542	4310
Online lag	38	90	121	165	551
Print lag	53	148	218	291	874
Total time	55	428	570	706	4392

Time in days for the journal Psychological Methods

Source: http://persistentastonishment.blogspot.nl/2015/07/median-publication-delays-at-38-apa.html

Pre-print Reports

- Now, methodological reports can be made public instantly using pre-print reports
- These reports may not have been checked by external experts
- The pre-print report may contain errors, or may change severely throughout the review process
  - That is the whole purpose of peer-review to begin with!
- Empirical work based on earlier versions of the pre-print may find their analyses now outdated or their conclusions now no longer warranted

A Tutorial on Regularized Partial Correlation Networks

Sacha Epskamp, Eiko I. Fried

(Submitted on 5 Jul 2016 (v1), last revised 1 Dec 2017 (this version, v9))

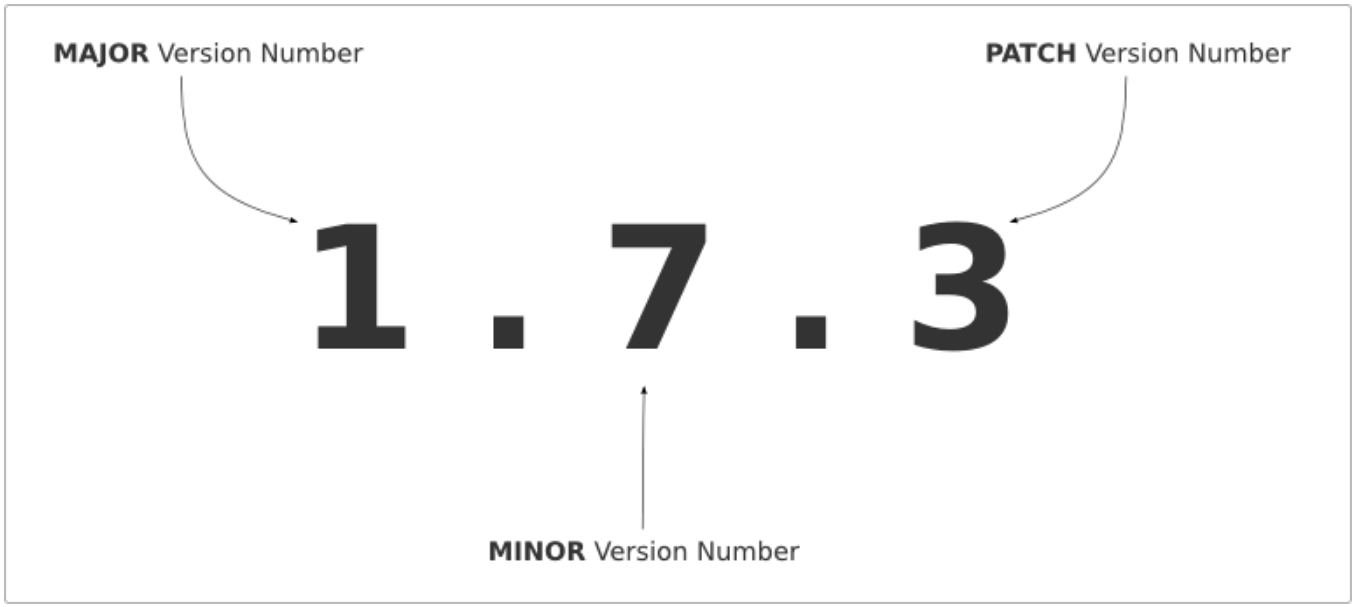
Submission history

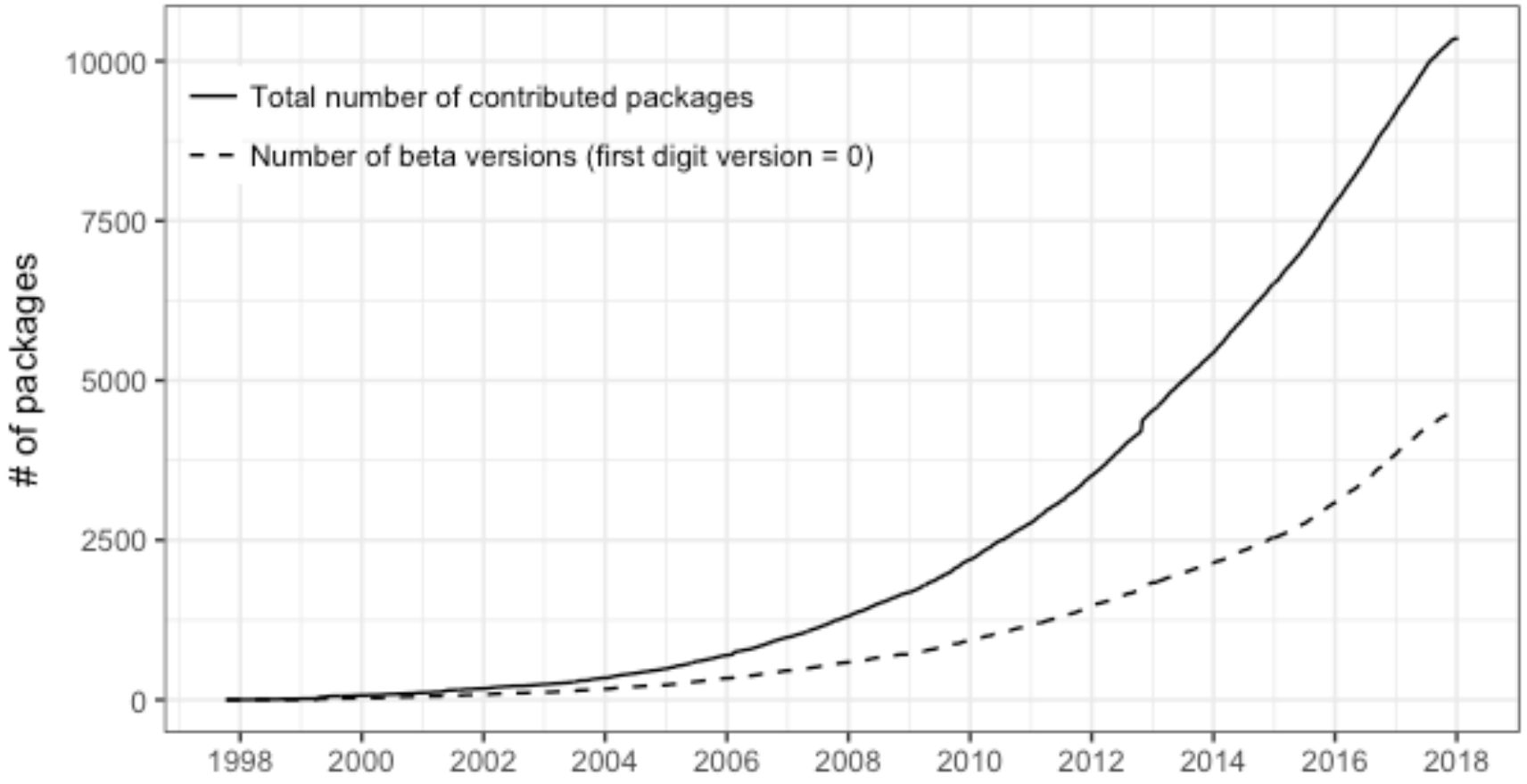
From: Sacha Epskamp [view email]
[v1] Tue, 5 Jul 2016 18:55:19 GMT (1113kb)
[v2] Tue, 13 Sep 2016 20:01:32 GMT (128kb,D)
[v3] Sun, 25 Sep 2016 10:24:17 GMT (150kb,D)
[v4] Mon, 3 Oct 2016 15:34:25 GMT (132kb,D)
[v5] Sun, 30 Apr 2017 18:29:41 GMT (221kb,D)
[v6] Wed, 28 Jun 2017 16:15:15 GMT (223kb,D)
[v7] Mon, 11 Sep 2017 20:46:08 GMT (181kb,D)
[v8] Thu, 14 Sep 2017 11:09:41 GMT (181kb,D)
[v9] Fri, 1 Dec 2017 13:35:57 GMT (181kb,D)

SOURCES OF LIMITED REPRODUCIBILITY AND REPLICABILITY IN SOFTWARE

Beta Versions

- Developers typically indicate software is still unstable / in development by denoting the version number as 0 (beta version)
- Using such software comes at a risk




Total number of contributed packages
Number of beta versions (first digit version = 0)
# of packages
10000
7500
5000
2500
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

lavaan: Latent Variable Analysis

Fit a variety of latent variable models, including confirmatory factor analysis, structural equation modeling and latent growth curve models.

Version:	0.5-23.1097
Depends:	R (≥ 3.1.0)
Imports:	methods, stats4, stats, utils, graphics, MASS, mnormt, pbivnorm, quadprog, numDeriv
Suggests:	testthat
Published:	2017-02-24
Author:	Yves Rosseel [aut, cre], Daniel Oberski [ctb], Jarrett Byrnes [ctb], Leonard Vanbrabant [ctb], Victoria Savalei [ctb], Ed Merkle [ctb], Michael Hallquist [ctb], Mijke Rhemtulla [ctb], Myrsini Katsikatsou [ctb], Mariska Barendse [ctb], Michael Chow [ctb], Terrence Jorgensen [ctb]
Maintainer:	Yves Rosseel <Yves.Rosseel at UGent.be>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	http://lavaan.org
NeedsCompilation:	no
Citation:	lavaan citation info
Materials:	README
In views:	Econometrics, OfficialStatistics, Psychometrics
CRAN checks:	lavaan results

mlVAR: Multi-Level Vector Autoregression

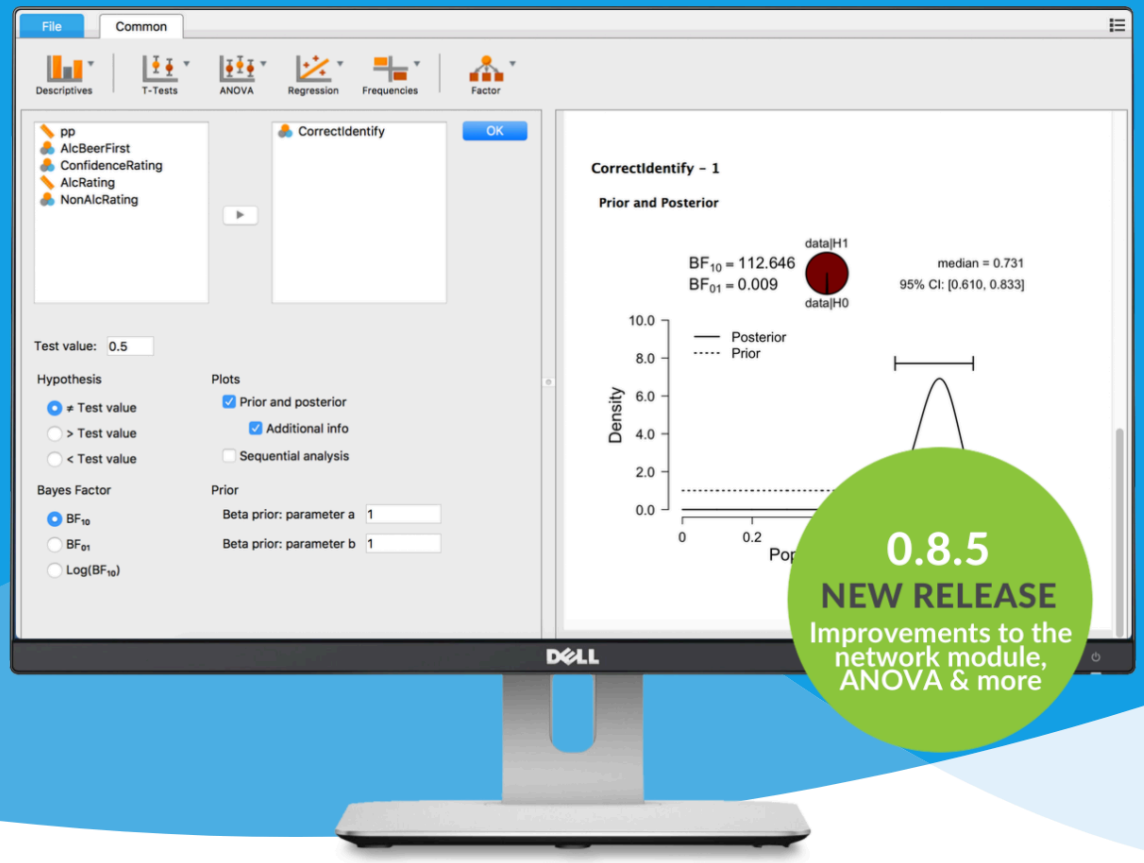
Estimates the multi-level vector autoregression model on time-series data. Three network structures are obtained: temporal networks, contemporaneous networks and between-subjects networks.

Version:	0.4
Depends:	R (≥ 3.3.0)
Imports:	lme4, arm, qgraph, dplyr, clusterGeneration, mvtnorm, corpcor, plyr, abind, methods, parallel, MplusAutomation
Published:	2017-09-02
Author:	Sacha Epskamp, Marie K. Deserno and Laura F. Bringmann
Maintainer:	Sacha Epskamp <mail at sachaepskamp.com>
License:	GPL-2
Copyright:	see file COPYRIGHTS
NeedsCompilation:	no
Materials:	NEWS
In views:	TimeSeries
CRAN checks:	mlVAR results

JASP

A Fresh Way to Do Statistics

Download



Are beta versions bad?

- No, it simply indicates the developers do not yet deem the software stable enough to warrant a release version
- Input, output, argument names, default options etcetera may still change
  - This may lead to reproducibility problems!
- There may be bugs
  - But then again, every program has bugs
- At least report the version number! A 0 implies beta software
- Note that especially open-source developers sometimes *never* set the version number > 0

Bugs and Unexpected Results

- Typically, methodologists are not trained in software engineering
- As a result, methodological software may not be optimally coded or thoroughly checked
- Using unexpected input may lead to unexpected output, and there may be bugs
- Methodological software may also not be well maintained, leading to new bugs to be introduced in the future
  - Many R packages only maintained by one person, who may change jobs or have other duties

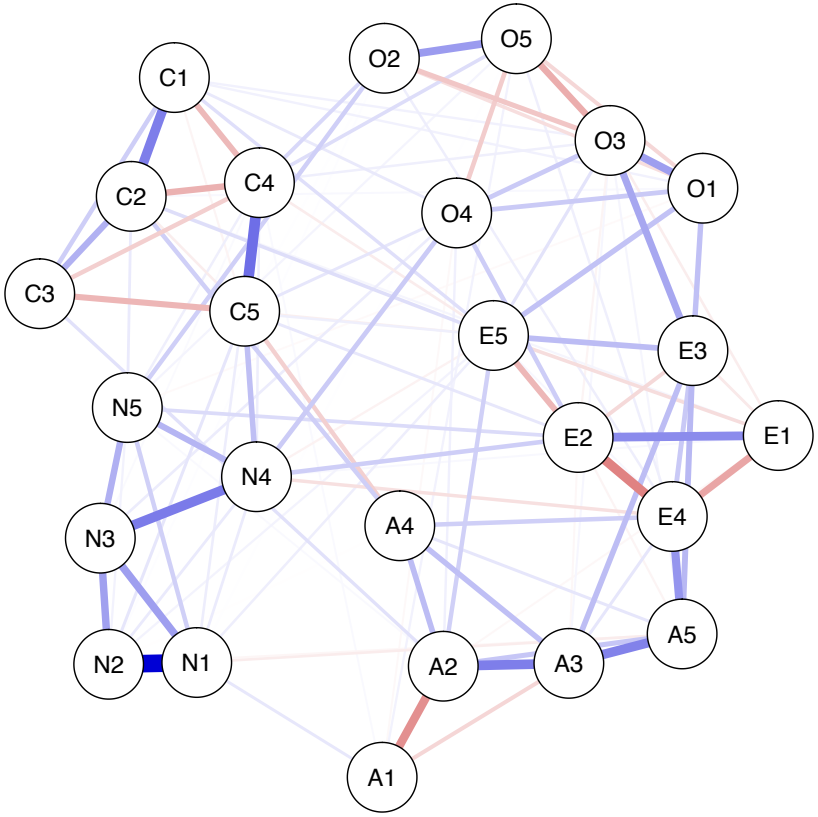
Updates and Changes

- Even crystalized software may change
  - Defaults may change due to user feedback
  - Methodological advances may warrant updates to estimation algorithms
  - Bugs may be fixed
- This leads to a risk that codes will not reproduce using later versions of the software
- Unfortunately, especially R has virtually no backward compatibility!

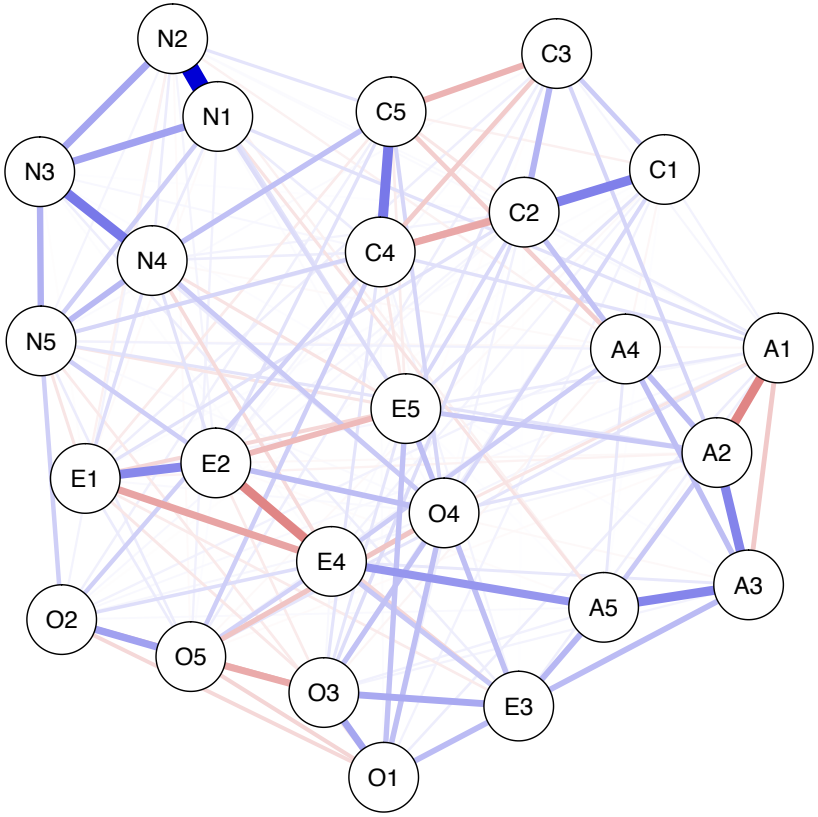
qgraph_1.0.0.tar.gz	2012-01-18 07:52	133K
qgraph_1.0.1.tar.gz	2012-02-19 21:58	133K
qgraph_1.0.2.tar.gz	2012-03-08 15:22	133K
qgraph_1.0.3.tar.gz	2012-04-19 10:02	133K
qgraph_1.0.4.tar.gz	2012-05-08 16:22	134K
qgraph_1.0.5-2.tar.gz	2012-07-19 17:49	134K
qgraph_1.0.5.tar.gz	2012-05-25 08:19	134K
qgraph_1.1.0.tar.gz	2012-09-16 08:03	140K
qgraph_1.2.1.tar.gz	2013-01-30 08:13	152K
qgraph_1.2.2.tar.gz	2013-05-08 18:38	166K
qgraph_1.2.3.tar.gz	2013-08-29 22:33	164K
qgraph_1.2.4.tar.gz	2014-02-11 17:04	179K
qgraph_1.2.5.tar.gz	2014-05-22 06:54	191K
qgraph_1.2.tar.gz	2013-01-24 17:05	152K
qgraph_1.3.1.tar.gz	2015-03-10 16:57	206K
qgraph_1.3.2.tar.gz	2016-01-19 15:15	208K
qgraph_1.3.3.tar.gz	2016-04-29 12:11	210K
qgraph_1.3.4.tar.gz	2016-06-30 18:16	211K
qgraph_1.3.5.tar.gz	2016-08-31 20:47	213K
qgraph_1.3.tar.gz	2014-12-02 12:44	204K
qgraph_1.4.0.tar.gz	2016-10-29 10:48	217K
qgraph_1.4.1.tar.gz	2016-11-22 16:57	222K
qgraph_1.4.2.tar.gz	2017-01-21 23:16	223K
qgraph_1.4.3.tar.gz	2017-03-27 16:41	224K

```
Changes in Version 1.3.3
  o The 'lambda.min.ratio' of 'EBICglasso' now defaults to 0.01 instead of
0.1, similar to earlier versions of qgraph. This will cause EBICglasso and
qgraph(..., graph = "glasso") to potentially return less sparse models!
```

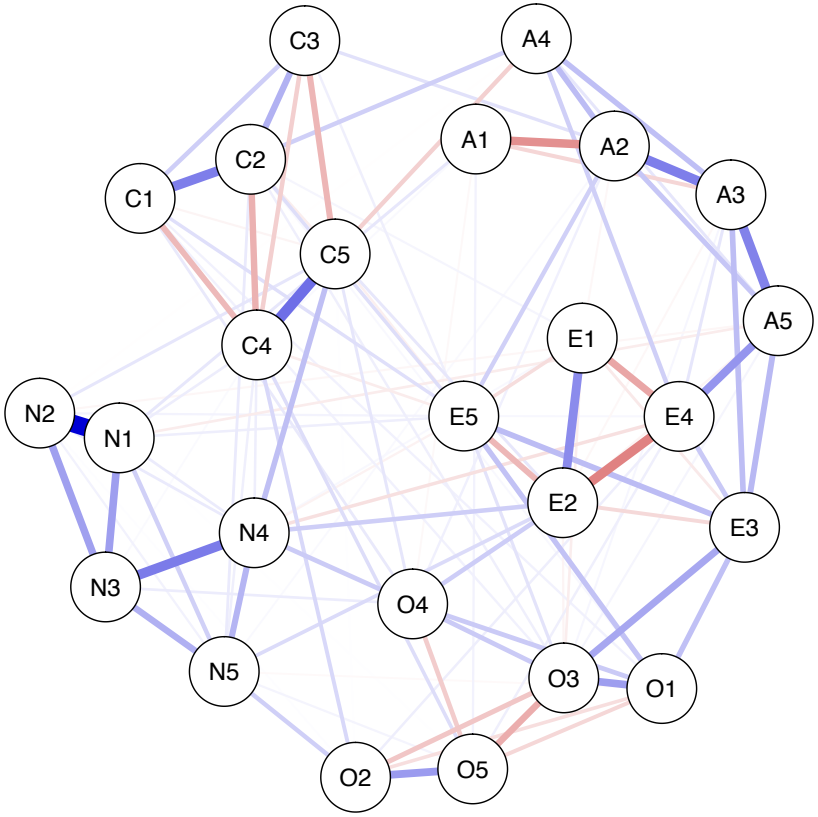
Before 1.3.3



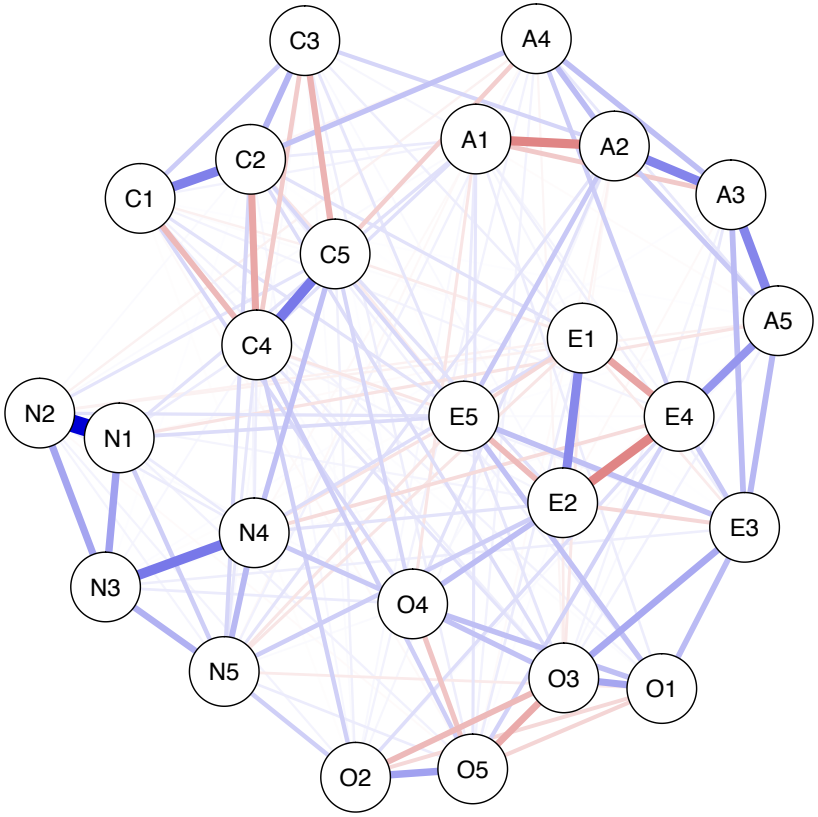
After 1.3.3



Before 1.3.3

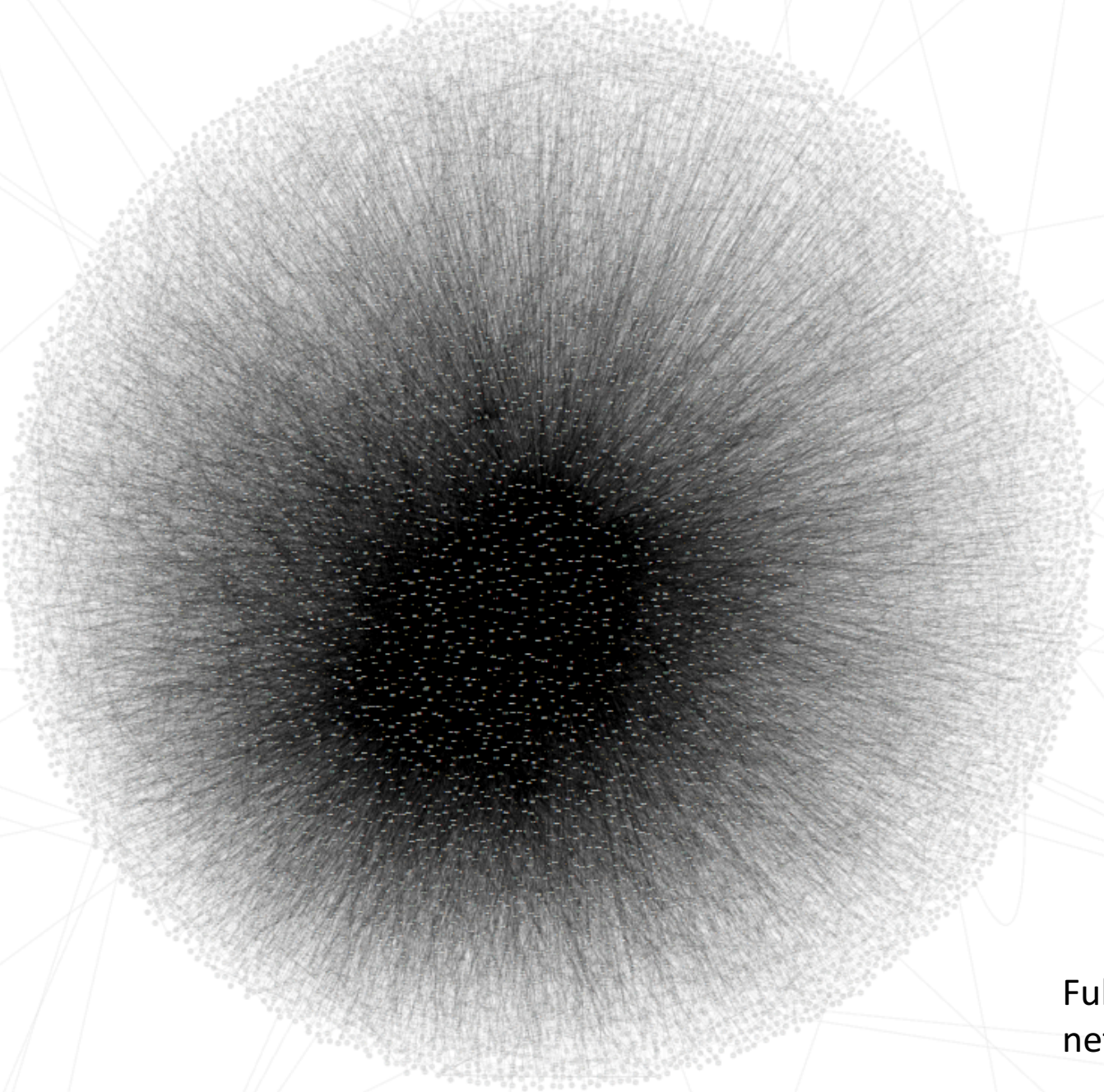


After 1.3.3

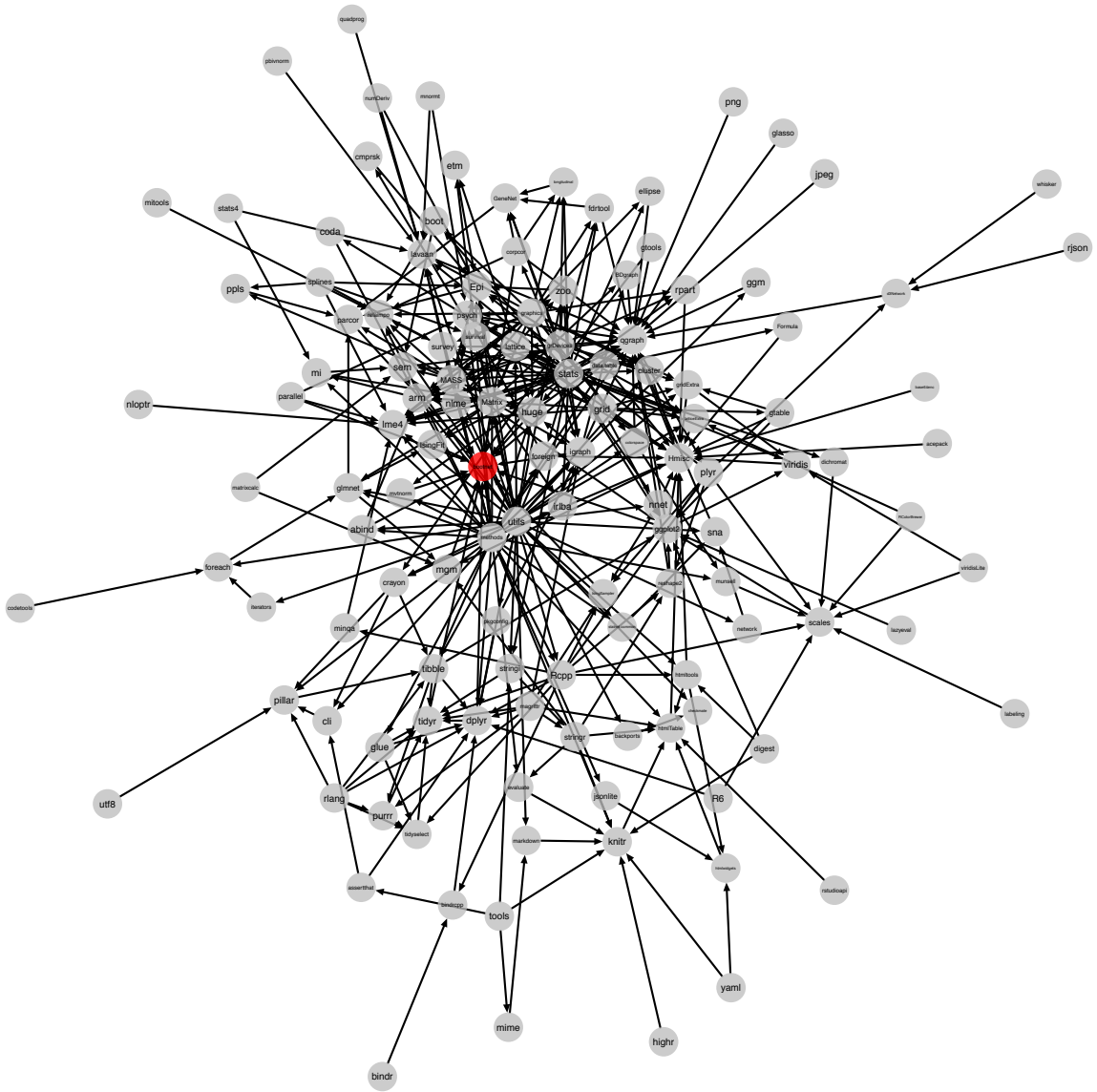


Dependencies

- All problems mentioned are greatly inflated by huge dependency trees between software packages
- Codes may not reproduce because one package changed, or because any of the depended packages changed
- Virtually impossible to trace back the source of the problem
- Changes in depending packages may also introduce new bugs!



Full CRAN dependency network



Bootnet dependency network

RECOMMENDATIONS FOR IMPROVING REPRODUCIBILITY

Distribute codes together with methodological publications

Distribute codes together with methodological publications

- Methodological studies without supplemented codes are hardly not reproducible
  - Especially not by empirical researchers
- If results do not reproduce, without codes you don't know if that is due to:
  1. Problems with the method itself
  2. Inconsistencies in implementation of the method
- If you ran a simulation study, you have codes!
- The more flexible your codes, the better
  - Do not hard-code variable names, sample size, etcetera

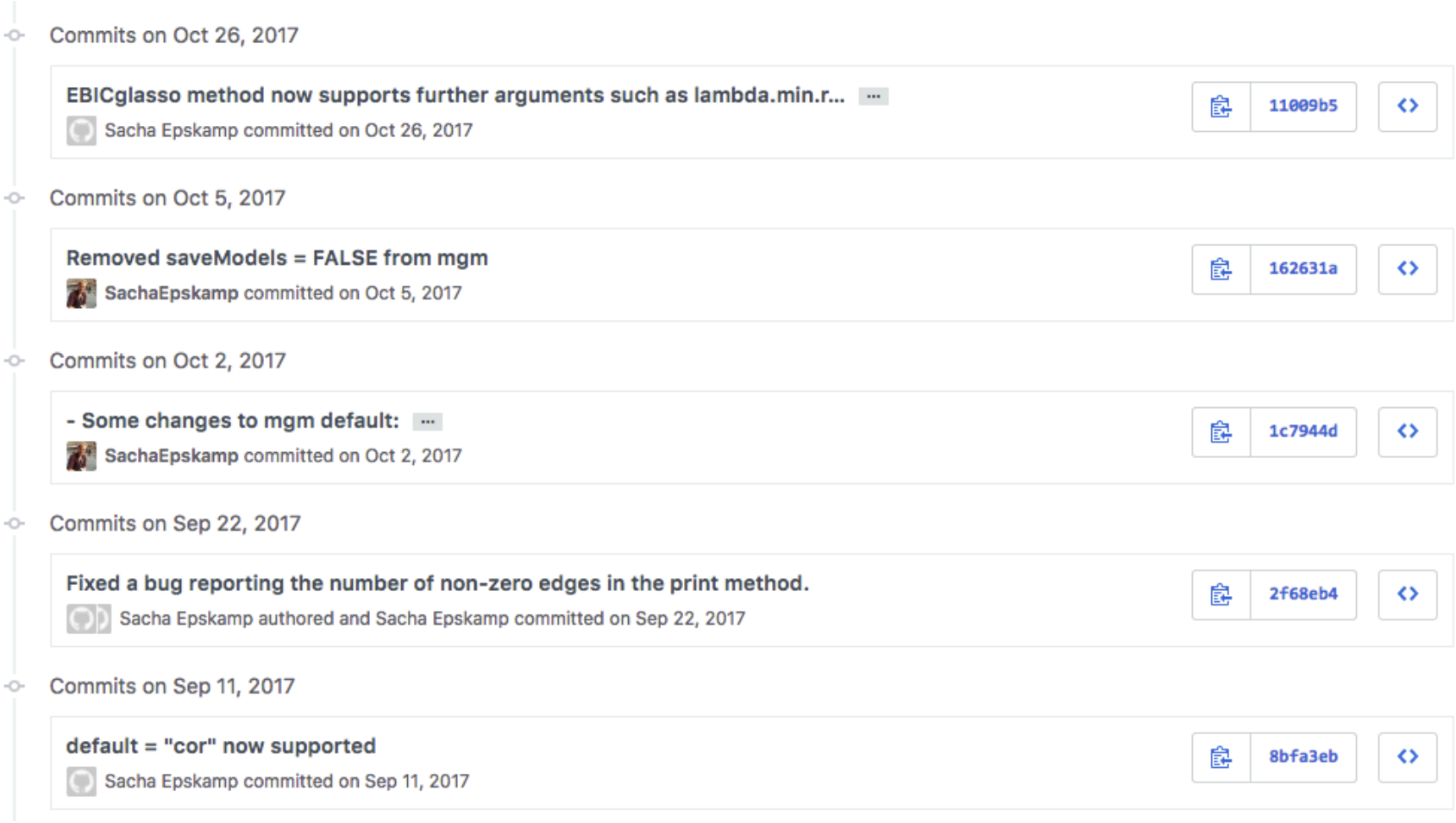
Provide documentation to the codes

- Codes are useless if nobody understands how to use them
- In a script: Add comments indicating what codes do and where users can change something
- In a function: Provide a description of expected input and the output
- For codes written in R, an ideal format in presenting codes is the form of an R package to be submitted to the Comprehensive R Archive Network (CRAN)
  - Requires you to pass numerous checks designed to force you to provide basic documentation

Use a clear coding style

- If your codes are not readable, then others cannot figure out what you did
- Thus, unreadable codes are almost not reproducible by others in different programming languages
- Some tips to make your code better:
  - Use a style guide
    - E.g., https://google.github.io/styleguide/Rguide.xml
  - Use many clarifying comments
  - Use informative object names
  - Use spaces / indents!

Use a version-control system

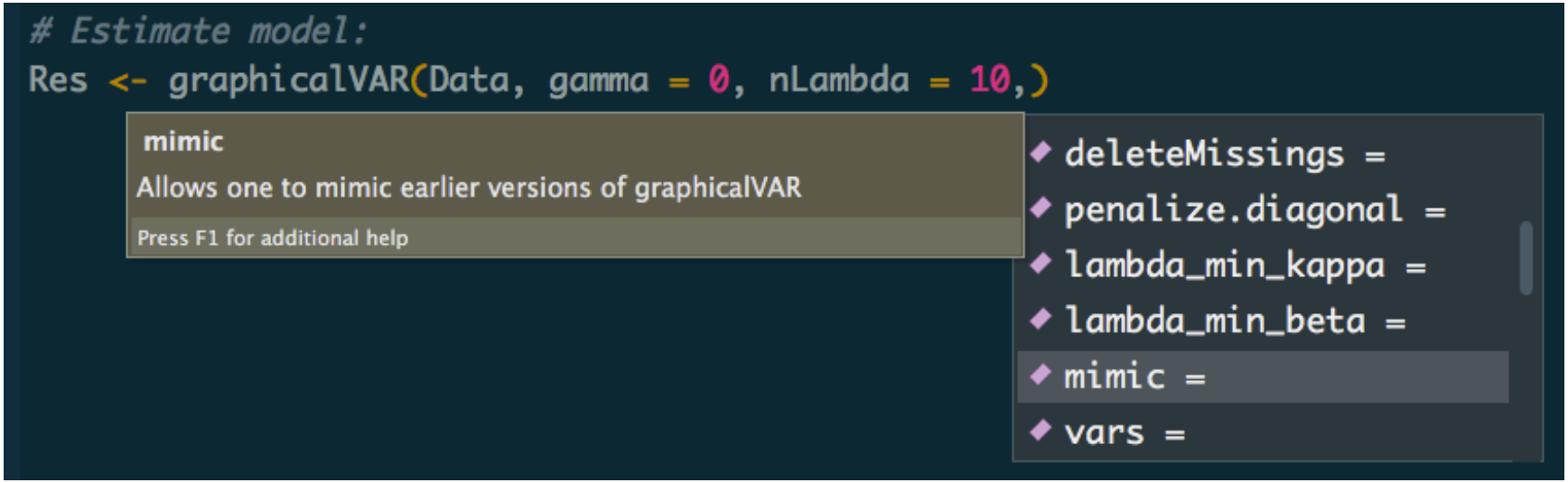


Use a version-control system

- Github is a great online repository!
  - Integrated with git
  - Clearly shows when what changes are made
  - Allows users to report bugs / ask questions
  - Allows others to suggest code changes
    - Pull requests

Offer backward compatibility

- When software defaults change, you want to make it possible to reproduce earlier analyses
- At least include arguments that can be set to mimic the earlier version
- Another option is to add a “mimic” argument that you can assign a version number
- Note: if the outcome change is due to a depending package changing, there is not much you can do…



Recommendations for Empirical Researchers

- Understand the methodology
  - If you do not understand the methodology, you do not know when something goes wrong
  - Your results may not replicate, simply because they rely on errors
- The more *crystallized* the software, the more likely codes will reproduce
  - Crystalized software is less likely to change and more likely to be validated
  - Note: many software packages may not be regarded as fully crystalized!

Journal of Affective Disorders
Volume 199, 15 July 2016, Pages 99-105

Research paper

Taxonicity and network structure of generalized anxiety disorder and major depressive disorder: An admixture analysis and complex network analysis

Table 1
Demographic and clinical measures.

	GAD ($n=70$)	MDD ($n=41$)	Statistic	p
Gender				
Male	19 (27%)	16 (39%)	$\chi^2(1)=1.69$	0.19
Female	51 (73%)	25 (61%)		
Ethnicity				
Caucasian	61 (87%)	31 (76%)	$\chi^2(1)=2.42$	0.12
Non-Caucasian	9 (13%)	10 (24%)		
Comorbidity			$\chi^2(1)=0.29$	0.59
With Comorbid	14 (20%)	10 (24%)		
Without Comorbid	56 (80%)	31 (76%)		
Age	35.17 (12.39)	33.27 (14.41)	$t(109)=-0.71$	0.48
BDI-II	18.69 (10.29)	22.52 (11.89)	$t(109)=1.71$	0.09
STAI-T	53.59 (9.97)	54.34 (9.40)	$t(109)=-0.39$	0.69
CSR	5.47 (1.30)	5.00 (1.43)	$t(109)=-1.72$	0.09

Note. BDI-II = Beck Depression Inventory-II; STAI-T=State Trait Anxiety Inventory, Trait Scale; CSR = Clinical Severity Rating.

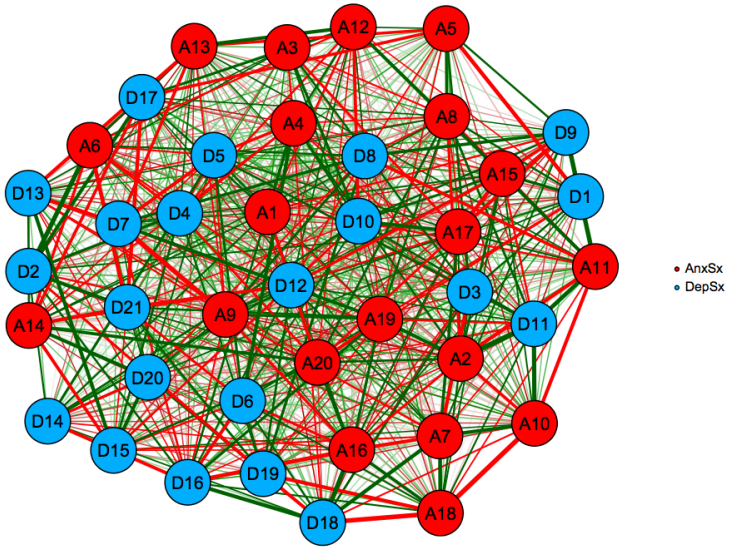


Fig. 3. GAD Complex Network.

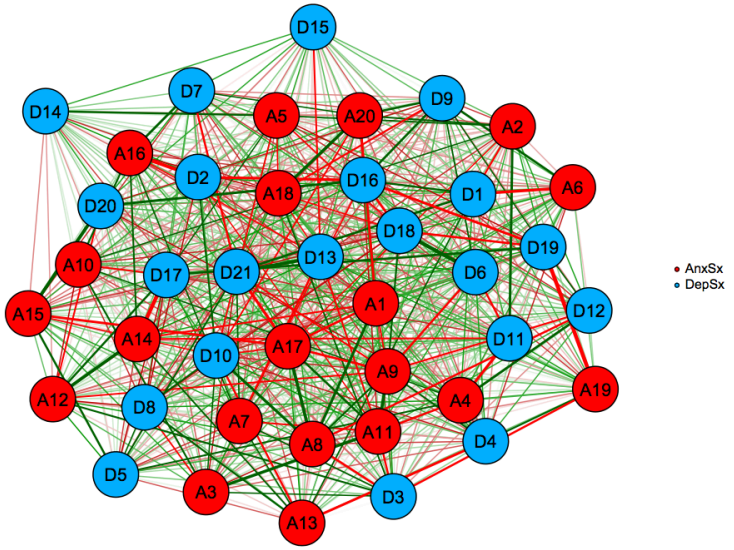
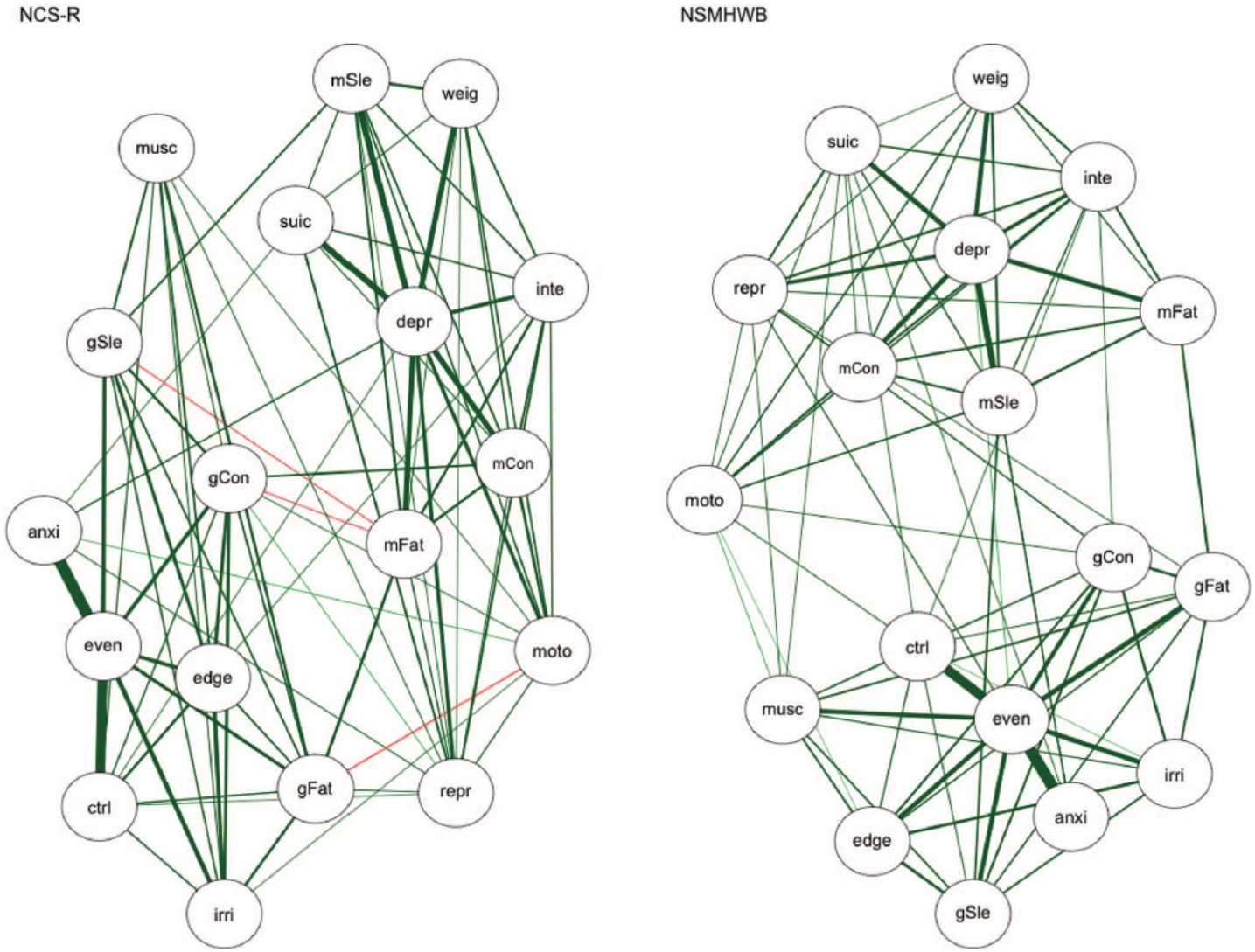


Fig. 4. MDD Complex Network.

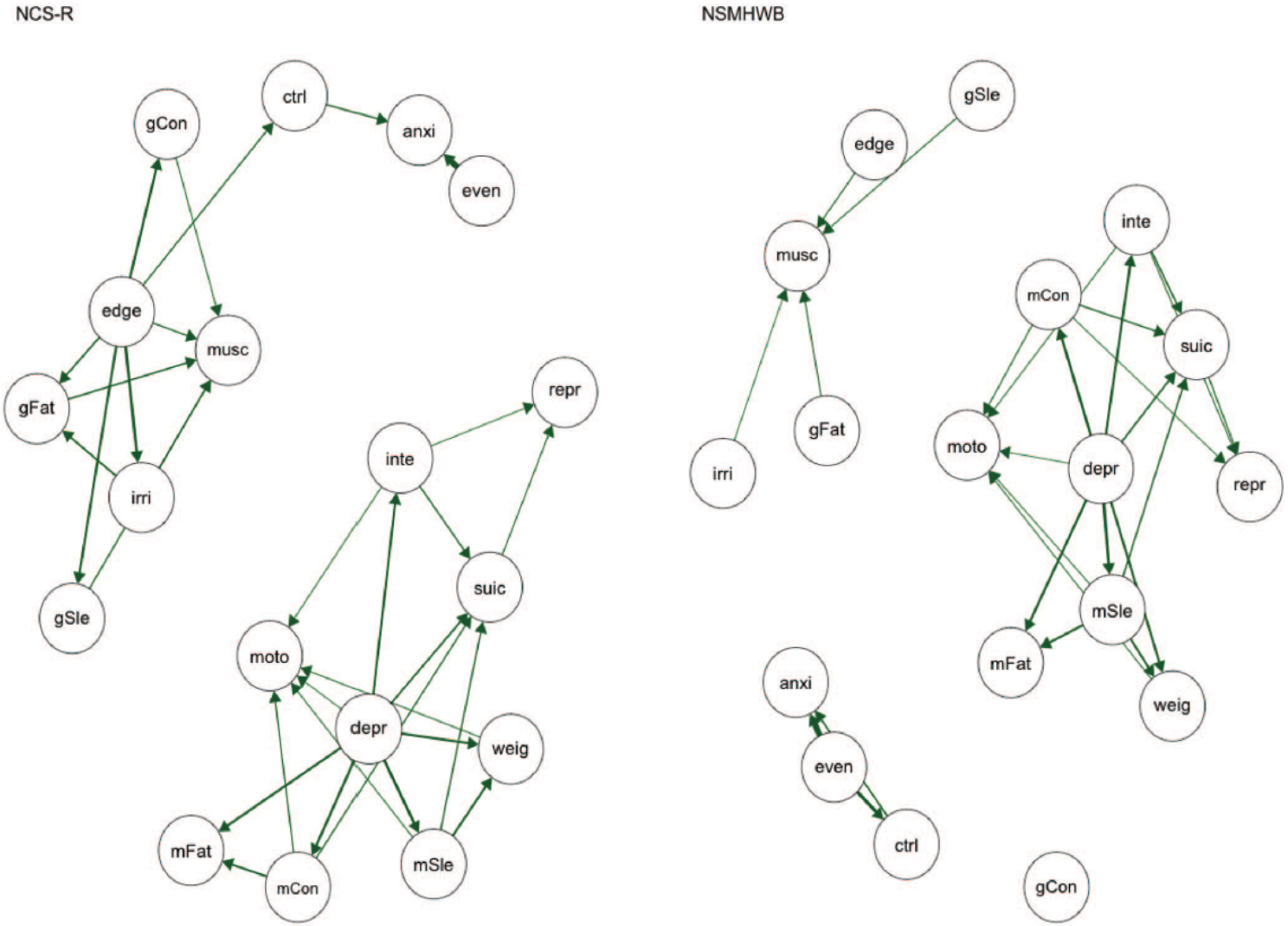
Evidence that Psychopathology Symptom Networks do not Replicate

Miriam K. Forbes[1], Aidan G. C. Wright[2], Kristian E. Markon[3], & Robert F. Krueger[4]

Ising Models

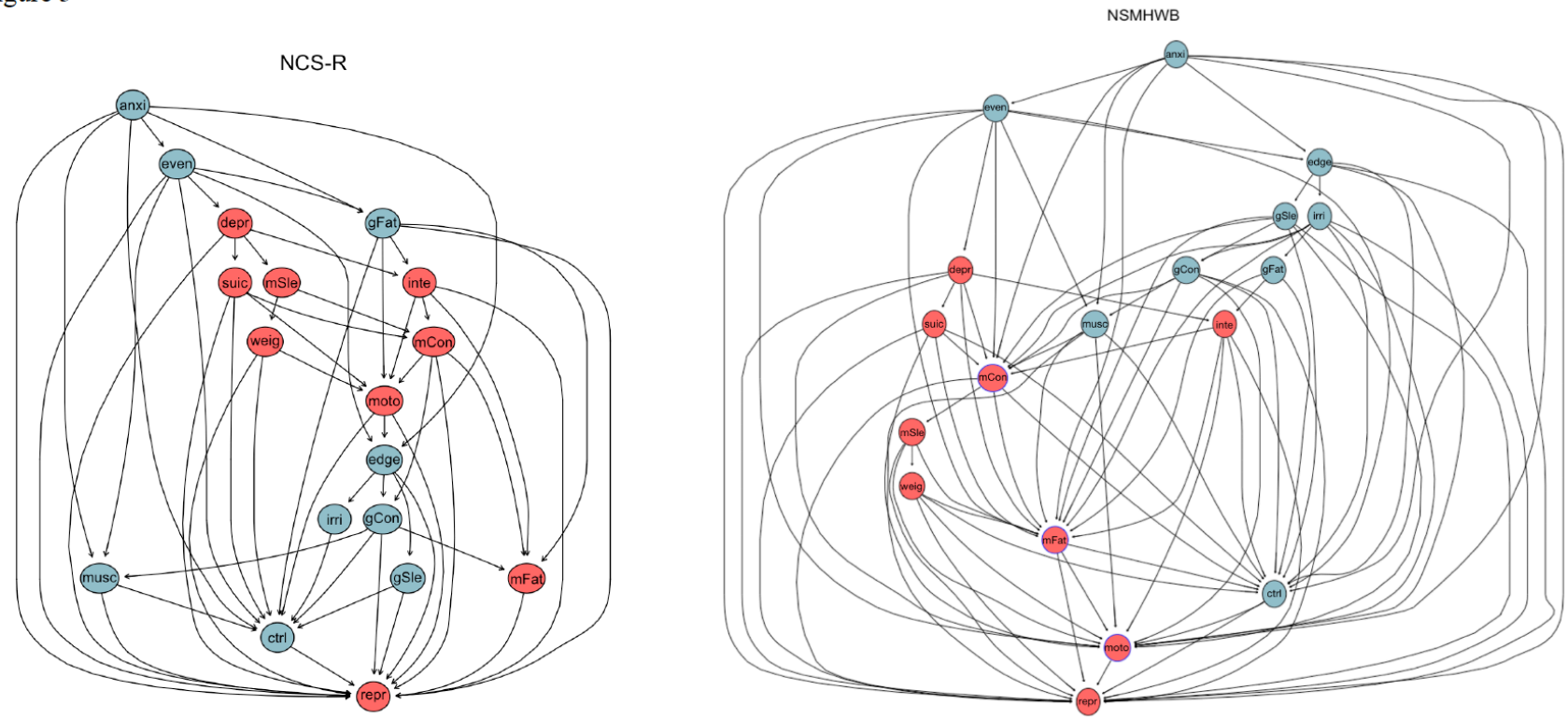


“Relative importance networks”

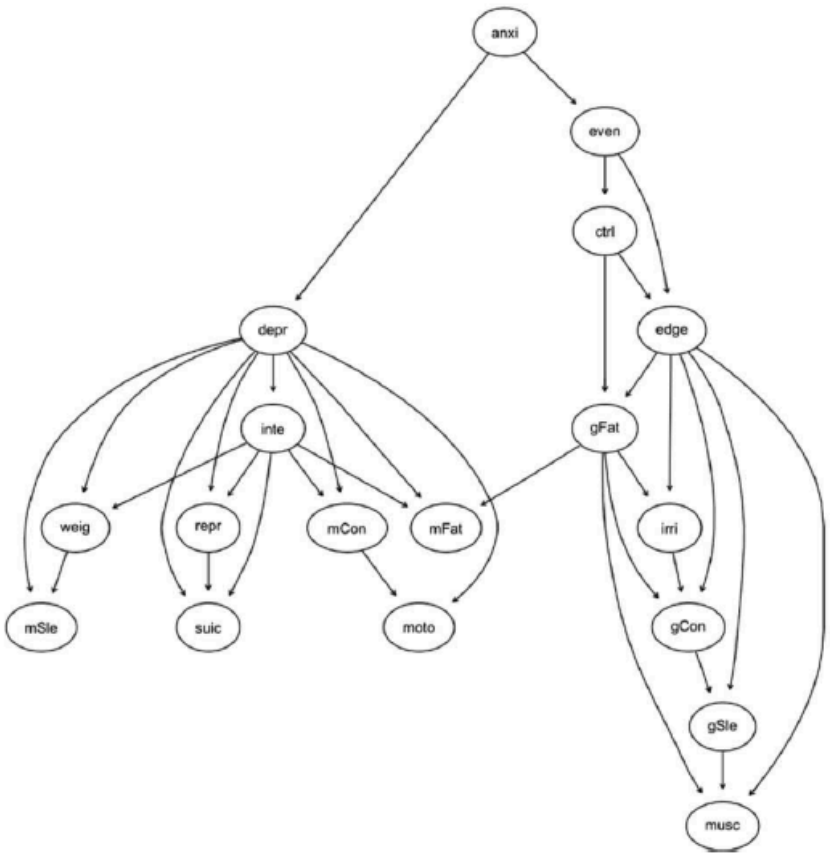


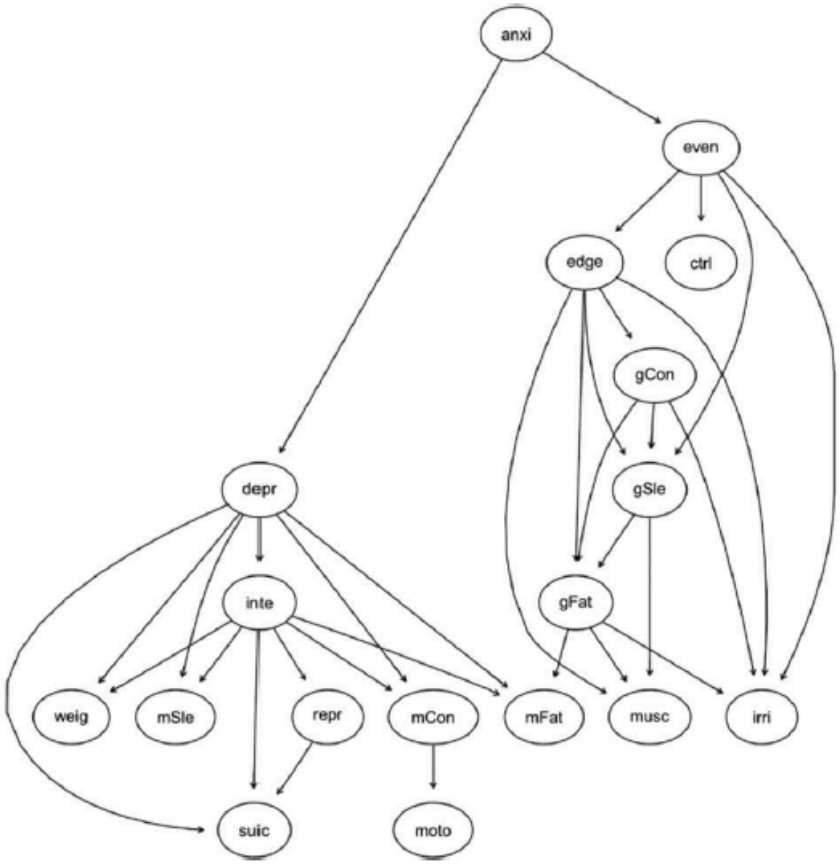
“Directed Acyclic Graphs”

Figure 5



Directed Acyclic Graphs





Changed after acceptance and without new peer-review

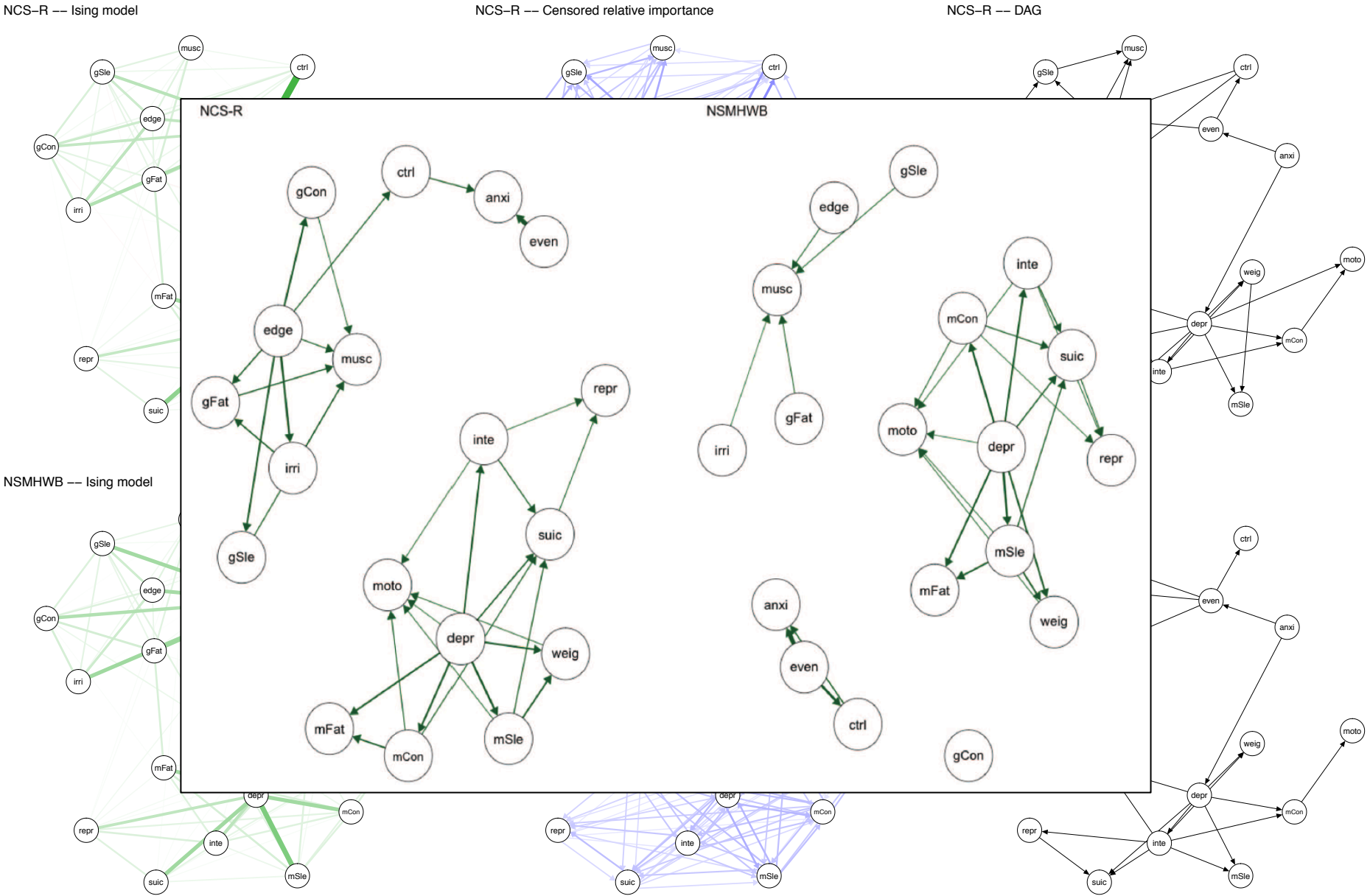
Forbes et al. results

Table 2
Summary of the Comparisons Among the Networks Between the National Comorbidity Survey—Replication (NCS-R) and the National Survey of Mental Health and Wellbeing (NSMHWB)

	Association networks		Regularized Ising models		Censored relative importance networks		Directed acyclic graphs	
Network characteristics	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB
Comparing global characteristics								
Connectivity (% possible edges)	153 (100%)	153 (100%)	80 (52.3%)	79 (51.6%)	31 (10.1%)	26 (8.5%)	34 (22.2%)	33 (21.6%)
Density (average edge strength)	.74	.79	1.10	1.17	.11	.10	N/A	N/A
Changes in estimated edges								
Mean absolute % change in edge weights of replicated edges	8.3%		30.4%		8.0%		N/A	
Proportion of edges that replicated from NCS-R to NSMHWB	153 (100%)		69 (86.3%)		23 (74.2%)		27 (79.4%)	
Edges that failed to replicate from NCS-R	0 (0%)		11 (13.8%)		8 (25.8%)		7 (20.6%)	
Edges unique to NSMHWB	0 (0%)		10 (12.7%)		3 (11.5%)		6 (18.2%)	
Node centrality								
Most central node	*mCon*	*mSle*	*depr*	*anxi*	*mSle*	*N/A*[a]	*N/A*[a]	*depr*
Rank-order	Correlation	Matches in rank-order	Correlation	Matches in rank-order	Correlation	Matches in rank-order	Correlation	Matches in rank-order
Strength/out strength/out degree	.79	4 (22.2%)	.69	3 (16.7%)	.73	6 (33.3%)	.75	14 (77.8%)
In strength/in degree	N/A	N/A	N/A	N/A	.57	3 (16.7%)	.57	16 (88.9%)
Closeness	.70	5 (27.8%)	.71	3 (16.7%)	N/A[b]	18 (100%)[b]	1.00[c]	18 (100%)[c]
Betweenness	N/A[b]	18 (100%)[b]	.77	10 (55.6%)	.46[c]	16 (88.9%)[c]	.66	10 (55.6%)

Note. Tied ranks (i.e., duplicate values) were common within each centrality index, and enabled nodes to have multiple possible ranks; a match was counted if a symptom could have an identical unique rank-order (e.g., fifth) in both samples. Mismatches were only counted if there was no possible combination of rank-orders that simultaneously facilitated a match and maintained a numerically ordered set of values. See Table 1 for node abbreviations.
[a] No node ranked as most central for at least two centrality indices. [b] All of the nodes had an estimated centrality index of 0. [c] At least 16 (89%) of nodes had an estimated centrality index of 0.

Our re-analysis



Our re-analysis

Variable	Ising models		Relative importance networks (censored)		Relative importance networks (uncensored)		DAGs	
	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB
Network characteristics[a]								
No. of edges (% of possible)	80 (52.3)	79 (51.6)	118 (38.6)	124 (40.5)	306 (100)	306 (100)	34 (22.2)	33 (21.6)
Density (as in Forbes et al.)	1.08	1.17	.13	.12	.06	.06		
Quality of replication								
Correlation between all edges	.95		.98		.99			
Correlation for nonzero edges	.97		.98		.99			
Jaccard index[b]	.77		.92		1.00		.68	
Change in edge weights (%)[a]	30.4		8.3		22.2			
Replicated edges (%)[a]	69 (86.3)		116 (98.3)		306 (100)		27 (79.4)	
Nonreplicated edges (%)[a]	11 (13.8)		2 (1.7)		0 (0)		7 (20.6)	
Edges unique to replication set (%)[a]	10 (12.7)		8 (6.5)		0 (0)		6 (18.2)	
Node centrality correlations								
Strength/outstrength/outdegree	.94		.94		.98		.87	
Instrength/indegree			.76				.62	
Closeness	.76				.98		1.00	
Betweenness	.94		.84		.92		.79	
Most central nodes[c]								
Strength/outstrength/outdegree	even	even	depr	depr	depr	depr	depr	depr, inte
Instrength/indegree			inte	weig	tie (15 nodes)	mFat	tie (4 nodes)	irri
Closeness	depr	mFat			mFat	mSle	anxi	anxi
Betweenness	depr	even	ctrl	even	gFat	gFat	edge	depr
	Correlation (τ_b)	Matches (%)	Correlation (τ_b)	Matches (%)	Correlation (τ_b)	Matches (%)	Correlation (τ_b)	Matches (%)
Rank-order correspondence[a]								
Strength/outstrength/outdegree	.69	3 (16.7)	.82	9 (50)	.8	4 (22.2)	.75	14 (77.8)
Instrength/indegree			.39	2 (11.1)			.57	16 (88.9)
Closeness	.71	3 (16.7)			.87	6 (33.3)	1.00	18 (100)
Betweenness	.77	11 (61.1)	.84	14 (77.8)	.57	9 (50)	.66	10 (55.6)

Crystalized-software checklist

1. The software is not in beta version
2. The software is validated in a methodological study
3. The software is well documented and/or accompanied by tutorials
4. The source codes are openly available

Note: very few software routines will pass all these checks!
Not all are required; the more the better!

Summary

While the age of fast-paced methodology greatly facilitates reproducibility, it also undermines it in ways not often realized by researchers. I will discuss:

- Sources of Limited Reproducibility and Replicability in Novel Methodology
- Sources of Limited Reproducibility and Replicability in Software
- Recommendations for improving reproducibility
  - Methodologists
  - Empirical Researchers