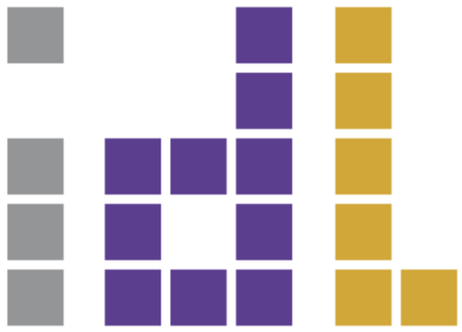
Document Exploration with Topic Models

Designing Interactive Visualizations to Support Effective Analysis Workflows



UW Interactive Data Lab

Jason Chuang U. Washington
Yuening Hu U. Maryland, College Park
Ashley Jin Stanford University
John D. Wilkerson U. Washington
Daniel A. McFarland Stanford University
Christopher D. Manning Stanford University
Jeffrey Heer U. Washington

Topical Analysis

Mapping 30 Years of Academic Discourse

in collaboration with Daniel Ramage, Christopher D. Manning, Daniel A. McFarland

Topic Models as a Tool

Research Question

Dissemination of theories, methods, techniques

Analysis Tool

Track transfer of language using topic models

Topic Modeling

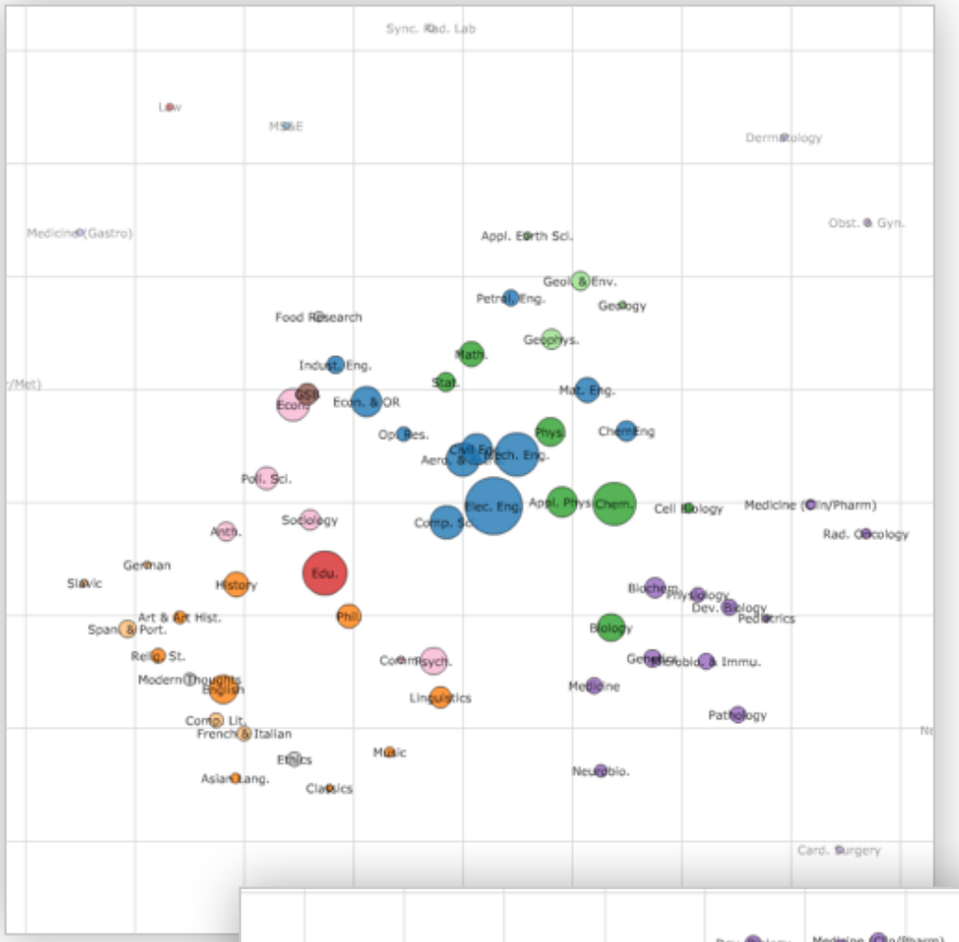
Dataset

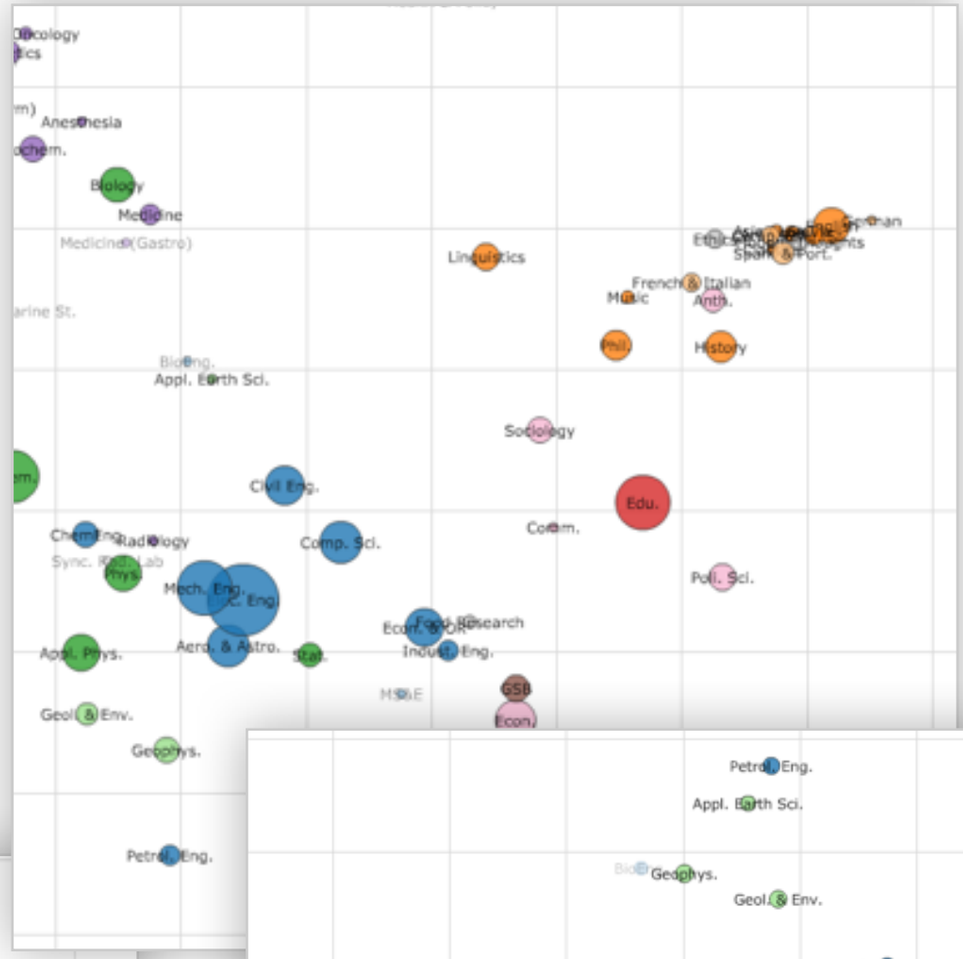
1.05 million PhD dissertations over 30 years

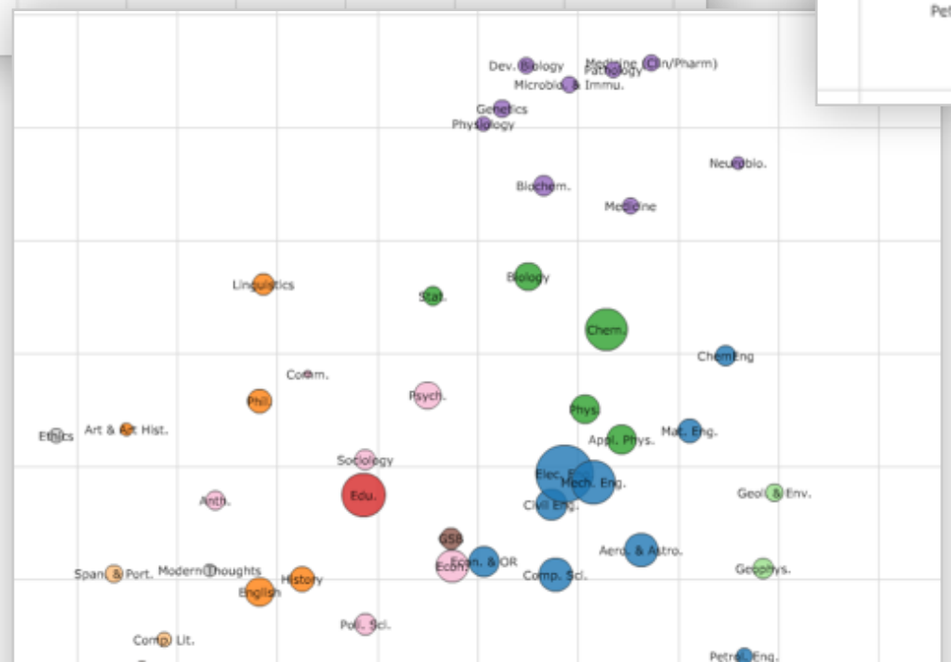
Model

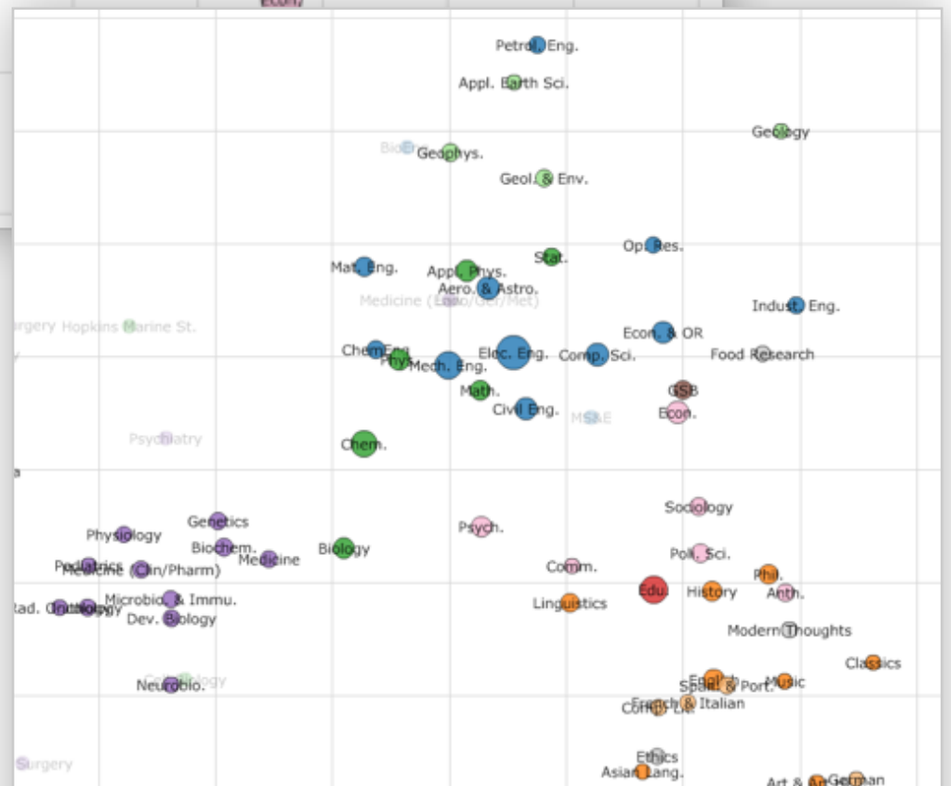
Disciplines as latent topics (Partially Labeled LDA)

It’s just one data point!









Mapping Academic Discourse

Additional Analysis

Citation graphs, Movement of researchers

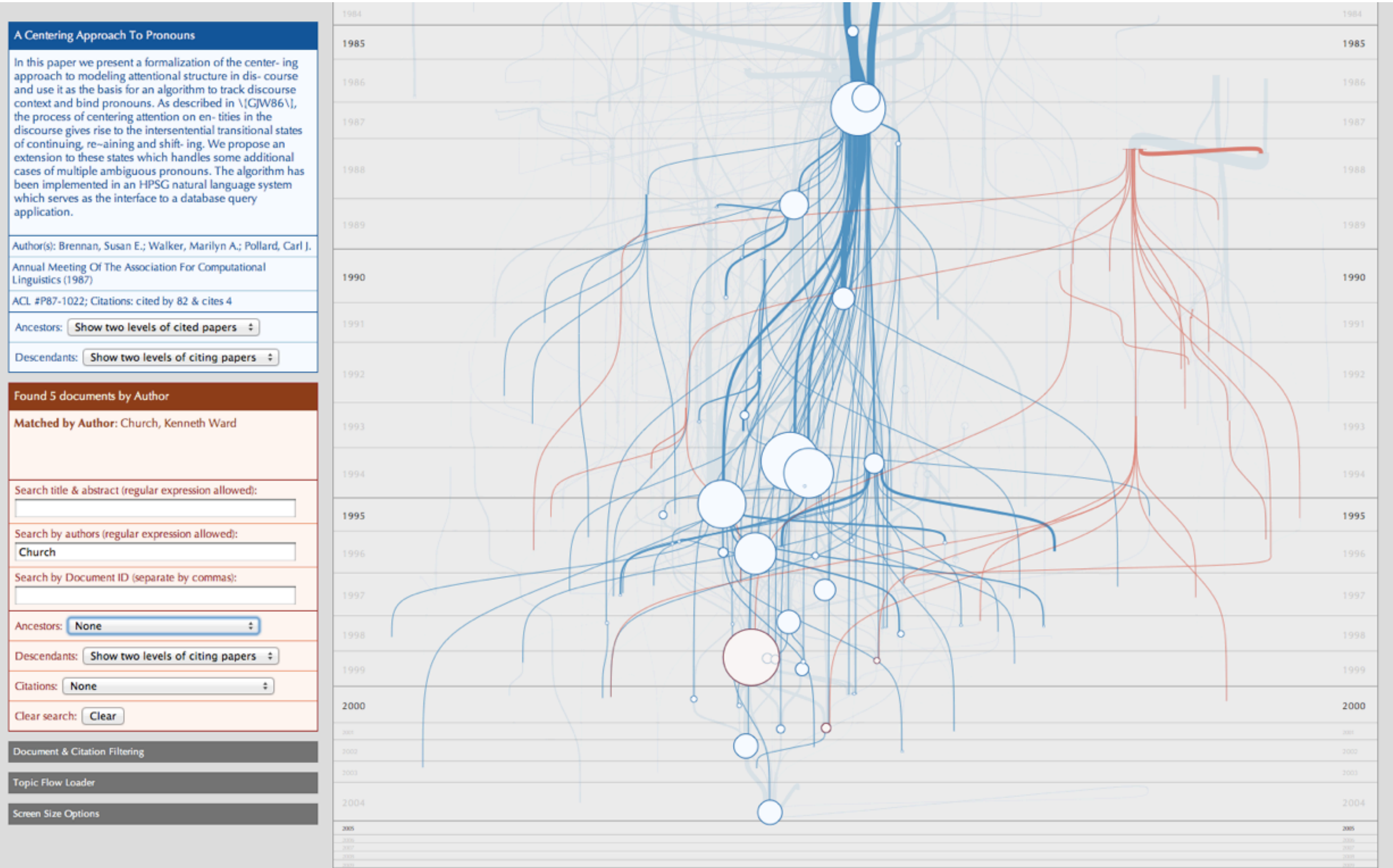
Topic Models + Interactive Visualizations

Interpretable

Accurate

Verifiable

Visualizing Topic Flow



Termite | Topic Model Visualization

in collaboration with Christopher D. Manning, Jeffrey Heer

Making Sense of Topic Models

Topic Words: vegf angiogenesis vascular_endothelial_growth_factor angiogenic end
antiangiogenic anti_angiogenic vegf_a tumor_angiogenesis vegfr2 growth signaling
Title Words: angiogenesis, vegf, vascular_endothelial_growth_factor, angiogenic, tu
neovascularization, angiopoietin, signaling, vegfr, vascular, human
Phrases: vascular_endothelial_growth_factor vegf, vegf angiogenesis, vegf receptor,

[Talley et al. 2011]

Anaphora Resolution	resolution anaphora pronoun discourse antecedent pronouns coref
Automata	string state set finite context rule algorithm strings language symb
Biomedical	medical protein gene biomedical wkh abstracts medline patient cl
Call Routing	call caller routing calls destination vietnamese routed router desti
Categorial Grammar	proof formula graph logic calculus axioms axiom theorem proofs
Centering*	centering cb discourse cf utterance center utterances theory coher
Classical MT	japanese method case sentence analysis english dictionary figure
Classification/Tagging	features data corpus set feature table word tag al test
Comp. Phonology	vowel phonological syllable phoneme stress phonetic phonology
Comp. Semantics*	semantic logical semantics john sentence interpretation scope log

[Hall et al. 2008]

Topic 14: [chat nil lig terminator png tactile praep snap taglet ng reentrant
anomalous autosem npr translat jsl ttc iim]

Topic 15: [cp cx bd cccwct cqct bp cqdd dbct cy bc dbcwcxcrcw bt cz cfct bf db
ctcpcrcw cwcpdact]

Topic 16: [bottle pf netgram viewpoint basilisk coloured fsv rufolo viewpoints
nursing georgei maryi amplifier td ipf circuit postcondition usrs]

Topic 17: [system language data user information text research project develop
database time processing machine work computer interface application natural p
tools]

DISCOURES SEGMENTATION Topic 18: [segment segmentation segments boundaries bou
cohesion topic chains seg chain hearst countability lexical false countable su
texttiling cues segmentations segmenting]

Topic 19: [ds dw composite jacobson deduction umts reseda deverbal ew board bs
che preverbs dv bankrupt thc umt sval kb]

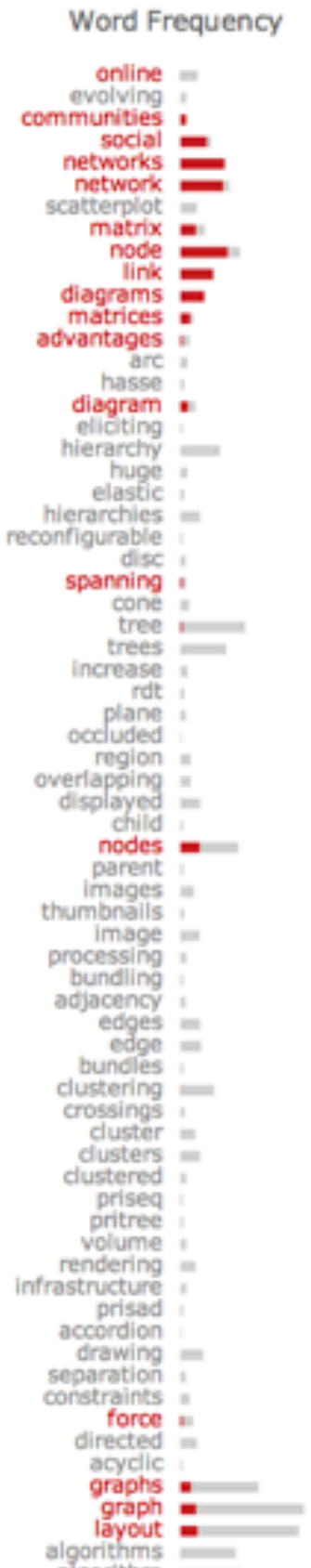
GENRE AND AUTHORSHIP DETECTION Topic 20: [genre stylistic genres style biber f
registers register humor stylistics postmodification primitive gender prose au
concordant discord greek discriminant bn]

SYNTACTIC STRUCTURE Topic 21: [verb syntactic noun semantic phrase verbs sent
object np structure case lexical frame phrases pp role argument head clause]

Topic 22: [shape unl bracketleftbig bracketleftbigg mins neg bracketrightbig h
bracketrightbigg image maxs lrs uce kallmeyer scrambling spitz dl portuguese]

Topic 23: [pi afips cc uc vat spmt soviet privacy vocnet urdu fcc ehlert cckg
anc melchuk spatter feeding alpp localities]

Termite | Topic Model Visualization



Representative Docu

A Comparison of the Readability of Graphs Using Node-Link and Ma
Mohammad Ghoniem Jean-Daniel Fekete Philippe Castagliola

Using Multilevel Call Matrices in Large Software Projects
Frank van Ham

Improving the Readability of Clustered Social Networks using Node
Nathalie Henry Anastasia Bezerianos Jean-Daniel Fekete

MatrixExplorer: a Dual-Representation System to Explore Social Ne
Nathalie Henry Jean-Daniel Fekete

NodeTrix: a Hybrid Visualization of Social Networks
Nathalie Henry Jean-Daniel Fekete Michael J. McGuffin

The need to visualize large social networks is growing as hardware
and many new data sets become available. Unfortunately, the visu
resolve the basic dilemma of being readable both for the global str
of local communities. To address this problem, we present NodeTri
combines the advantages of two traditional representations: node-l
of a network, while arbitrary portions of the network can be shown
analysis of communities. A key contribution is a set of interaction t
NodeTrix visualization by dragging selections to and from node-link
NodeTrix representation to explore the dataset and create meaning
Finally, we present a case study applying NodeTrix to the analysis
illustrate the capabilities of NodeTrix as both an exploration tool an

Visualizing Causal Semantics using Animations
Nivedita R. Kadaba Pourang P. Irani Jason Leboe

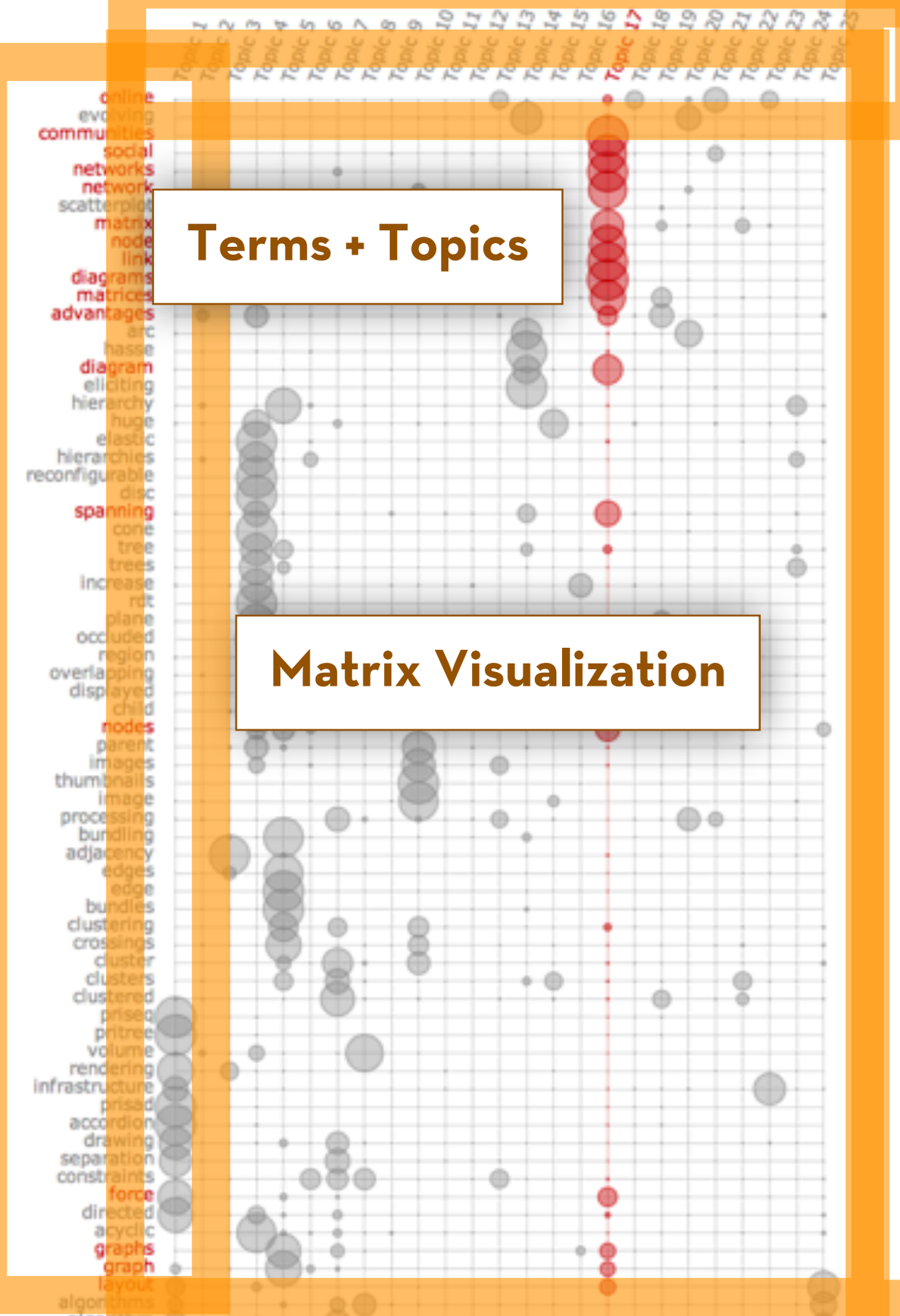
Balancing Systematic and Flexible Exploration of Social Netw
Adam Perer Ben Shneiderman

Social network analysis (SNA) has emerged as a powerful method
networks. However, interactive exploration of networks is currently
patterns and comprehend the structure of networks with many nod
medley of statistical methods and overwhelming visual output whic
explore in an orderly manner. This results in exploration that is lar
to help structural analysts understand social networks more effecti
attribute ranking and coordinated views to help users systematicall
flexibly iterate through visualizations of measures to gain an overv
networks using link structure, find cohesive subgroups, and focus o
networks by viewing different link types separately, or find pattern
overview. For each operation, a stable node layout is maintained in
comparisons. SocialAction offers analysts a strategy beyond opport
techniques for exploring social networks.

Causality Visualization Using Animated Growing Polygons
Niklas Elmqvist Philippas Tsigas

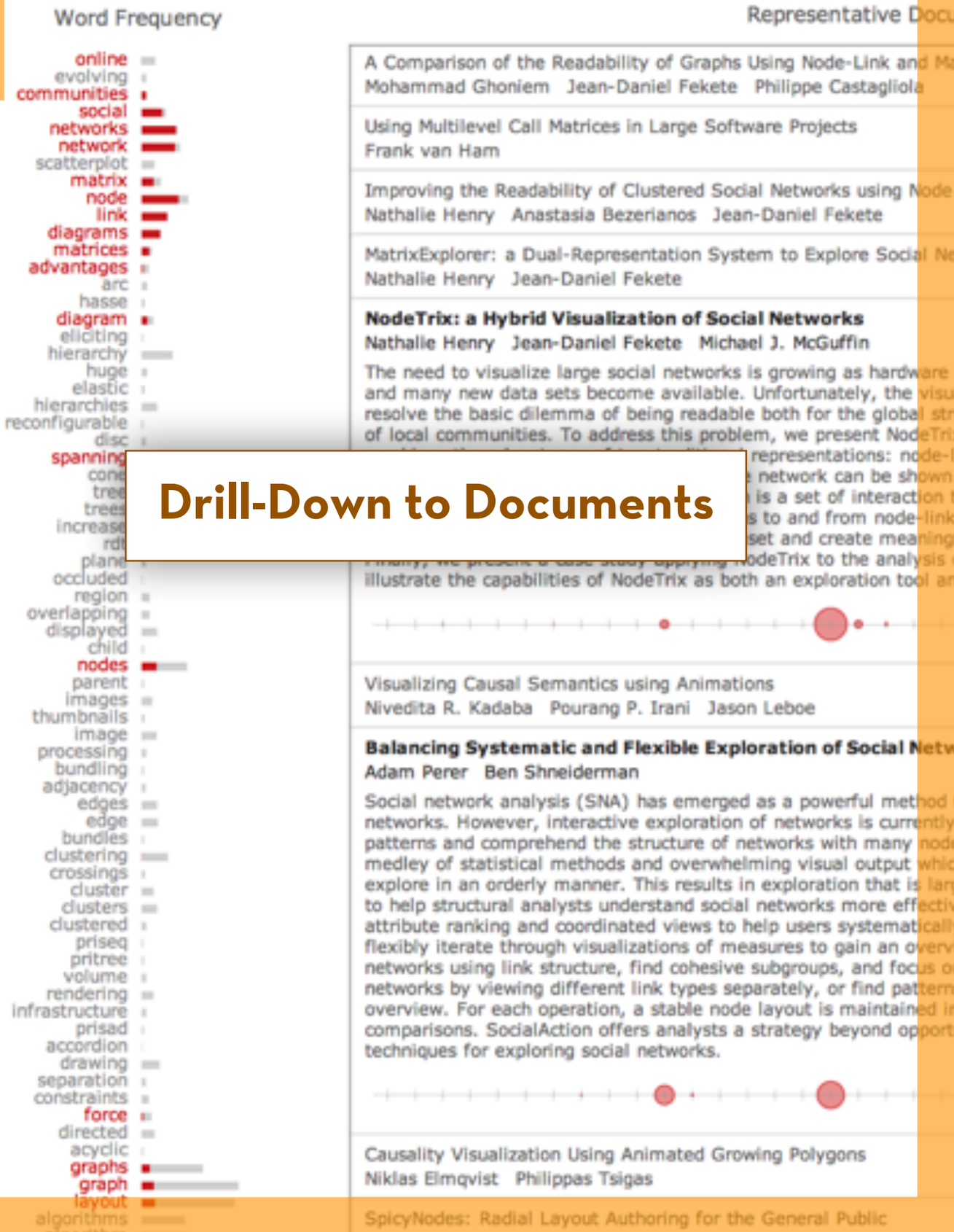
SpicyNodes: Radial Layout Authoring for the General Public

Termite | Topic Model Visualization

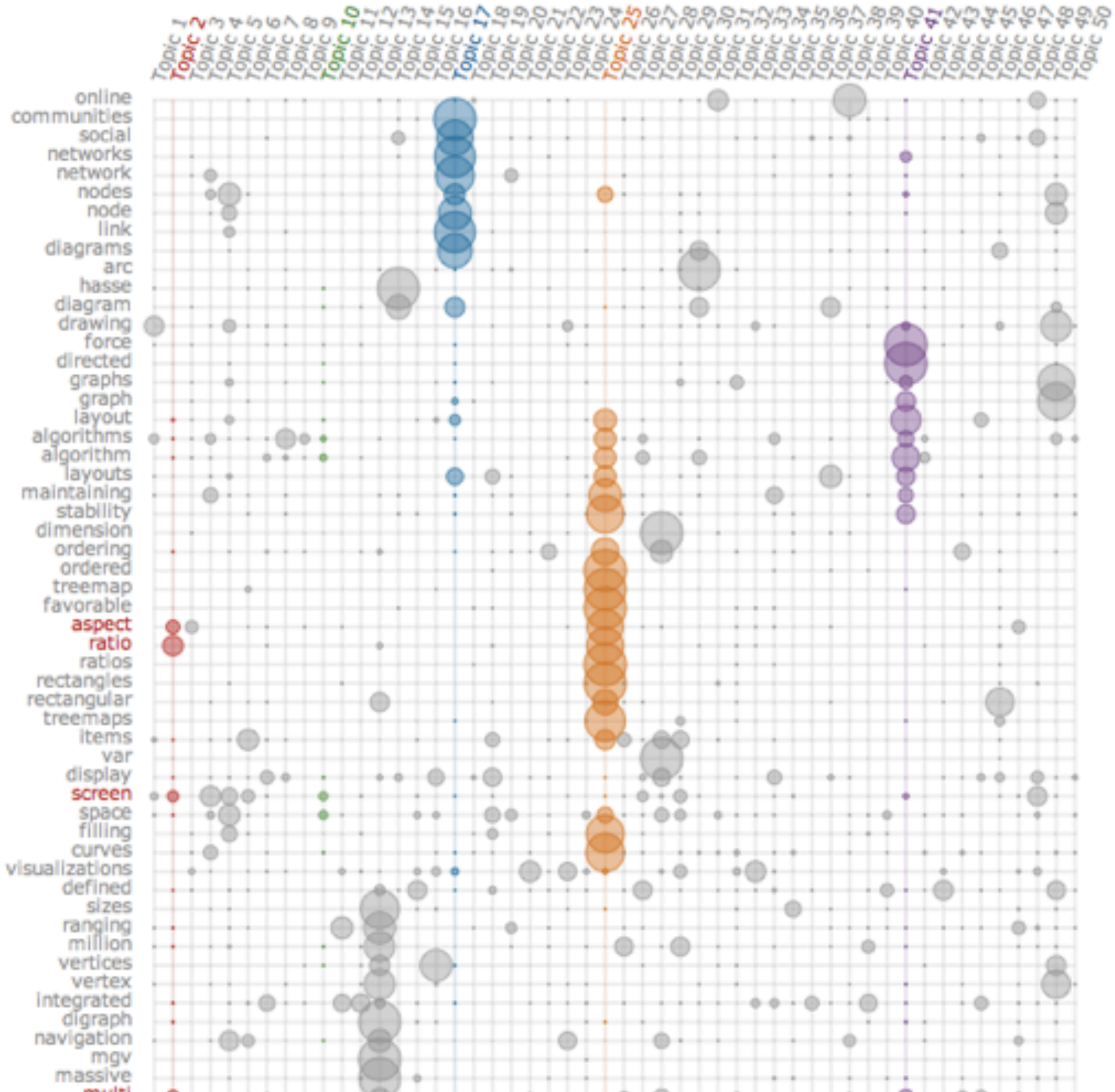


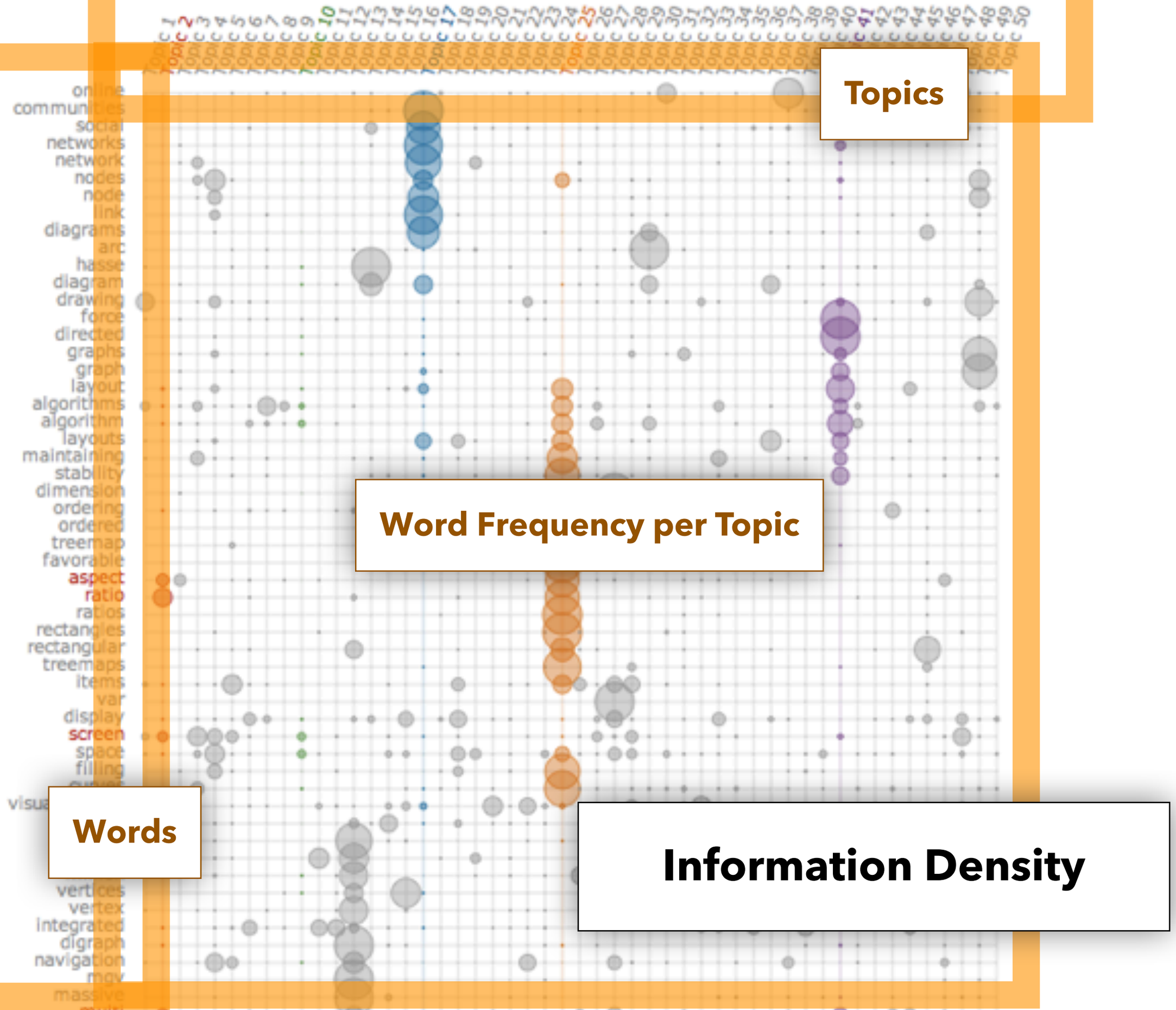
Terms + Topics

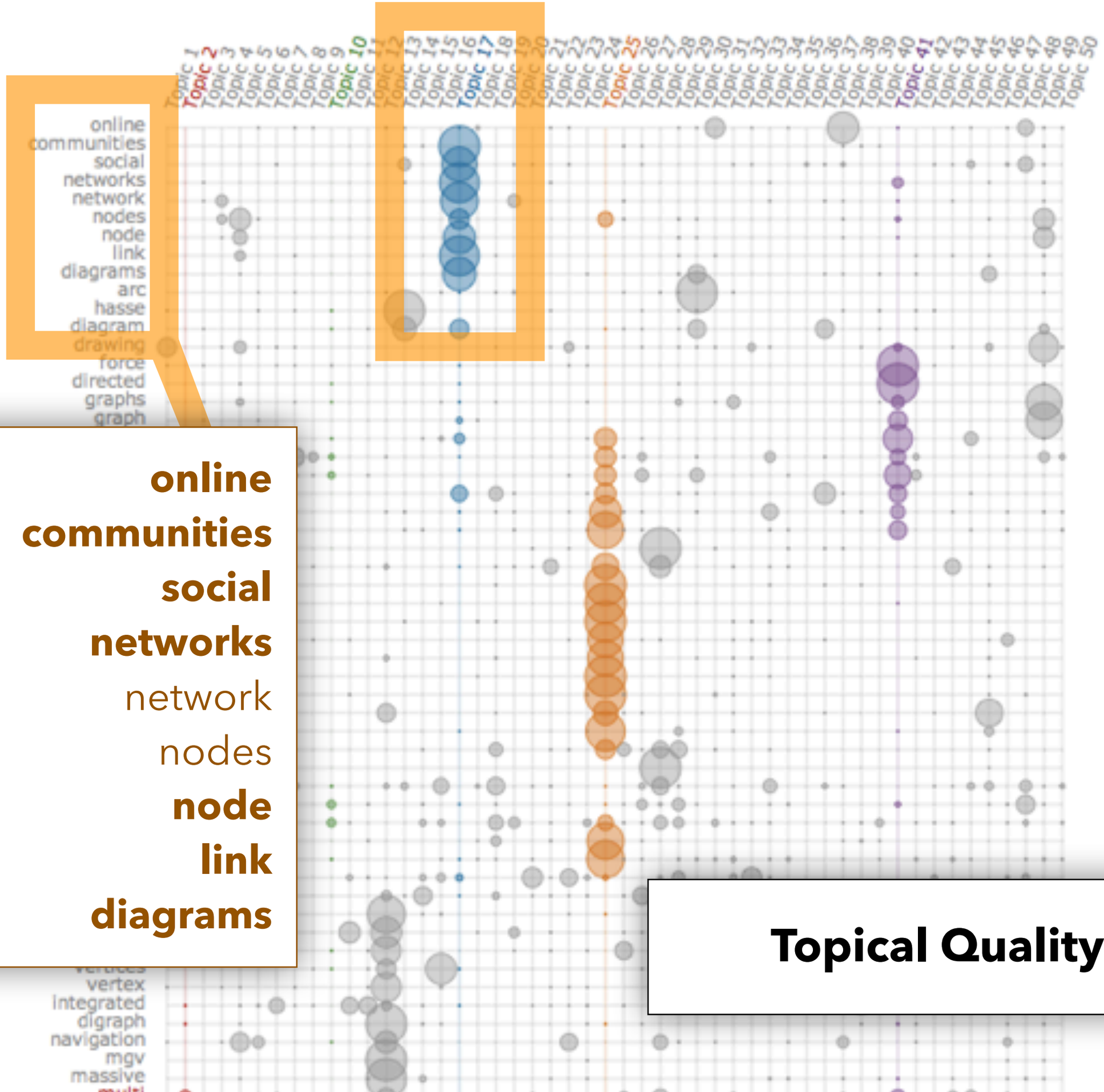
Matrix Visualization



Drill-Down to Documents

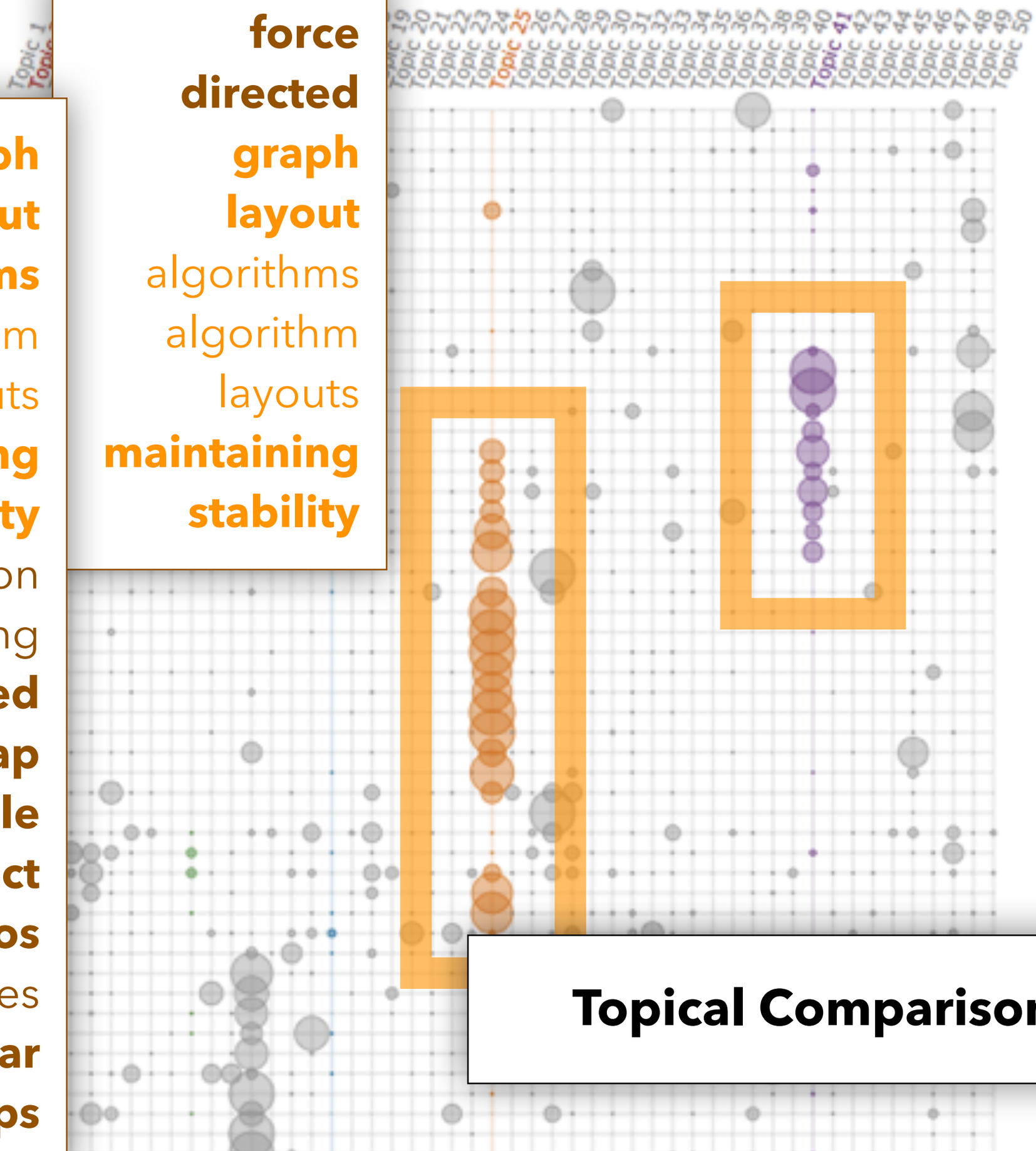

Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
Topic 9
Topic 10
Topic 11
Topic 12
Topic 13
Topic 14
Topic 15
Topic 16
Topic 17
Topic 18
Topic 19
Topic 20
Topic 21
Topic 22
Topic 23
Topic 24
Topic 25
Topic 26
Topic 27
Topic 28
Topic 29
Topic 30
Topic 31
Topic 32
Topic 33
Topic 34
Topic 35
Topic 36
Topic 37
Topic 38
Topic 39
Topic 40
Topic 41
Topic 42
Topic 43
Topic 44
Topic 45
Topic 46
Topic 47
Topic 48
Topic 49
Topic 50
online
communities
social
networks
network
nodes
node
link
diagrams
arc
hasse
diagram
drawing
force
directed
graphs
graph
layout
algorithms
algorithm
layouts
maintaining
stability
dimension
ordering
ordered
treemap
favorable
aspect
ratio
ratios
rectangles
rectangular
treemaps
items
var
display
screen
space
filling
curves
visualizations
defined
sizes
ranging
million
vertices
vertex
integrated
digraph
navigation
mgv
massive


Topics
Word Frequency per Topic
Words
Information Density
online
communities
social
networks
network
nodes
node
link
diagrams
arc
hasse
diagram
drawing
force
directed
graphs
graph
layout
algorithms
algorithm
layouts
maintaining
stability
dimension
ordering
ordered
treemap
favorable
aspect
ratio
ratios
rectangles
rectangular
treemaps
items
var
display
screen
space
filling
vertices
vertex
integrated
digraph
navigation
massive


Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
Topic 9
Topic 10
Topic 11
Topic 12
Topic 13
Topic 14
Topic 15
Topic 16
Topic 17
Topic 18
Topic 19
Topic 20
Topic 21
Topic 22
Topic 23
Topic 24
Topic 25
Topic 26
Topic 27
Topic 28
Topic 29
Topic 30
Topic 31
Topic 32
Topic 33
Topic 34
Topic 35
Topic 36
Topic 37
Topic 38
Topic 39
Topic 40
Topic 41
Topic 42
Topic 43
Topic 44
Topic 45
Topic 46
Topic 47
Topic 48
Topic 49
Topic 50
online
communities
social
networks
network
nodes
node
link
diagrams
arc
hasse
diagram
drawing
force
directed
graphs
graph
vertex
integrated
digraph
navigation
mgv
massive
online
communities
social
networks
network
nodes
node
link
diagrams
Topical Quality

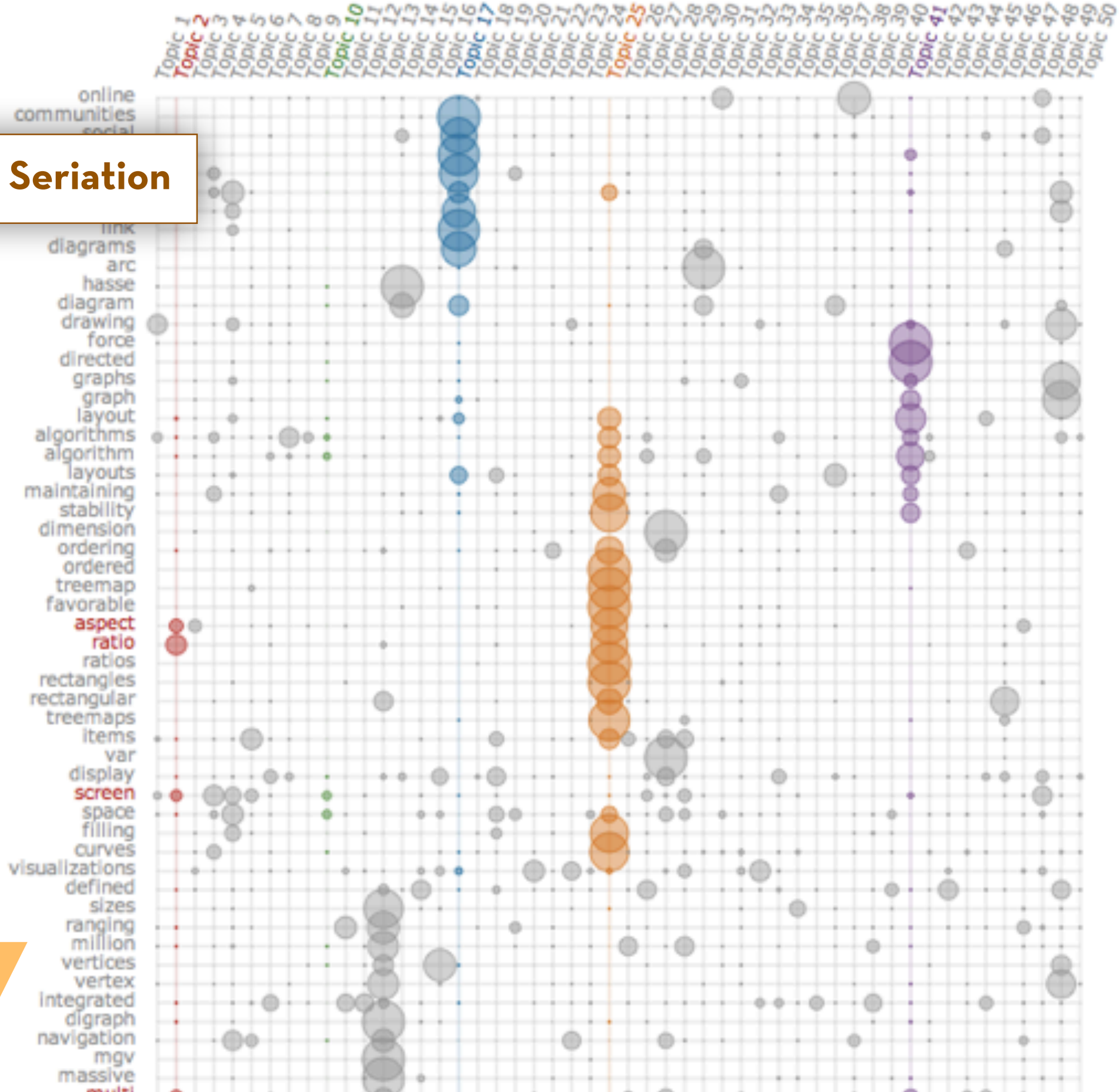
graph
layout
algorithms
algorithm
layouts
maintaining
stability
dimension
ordering
ordered
treemap
favorable
aspect
ratios
rectangles
rectangular
treemaps

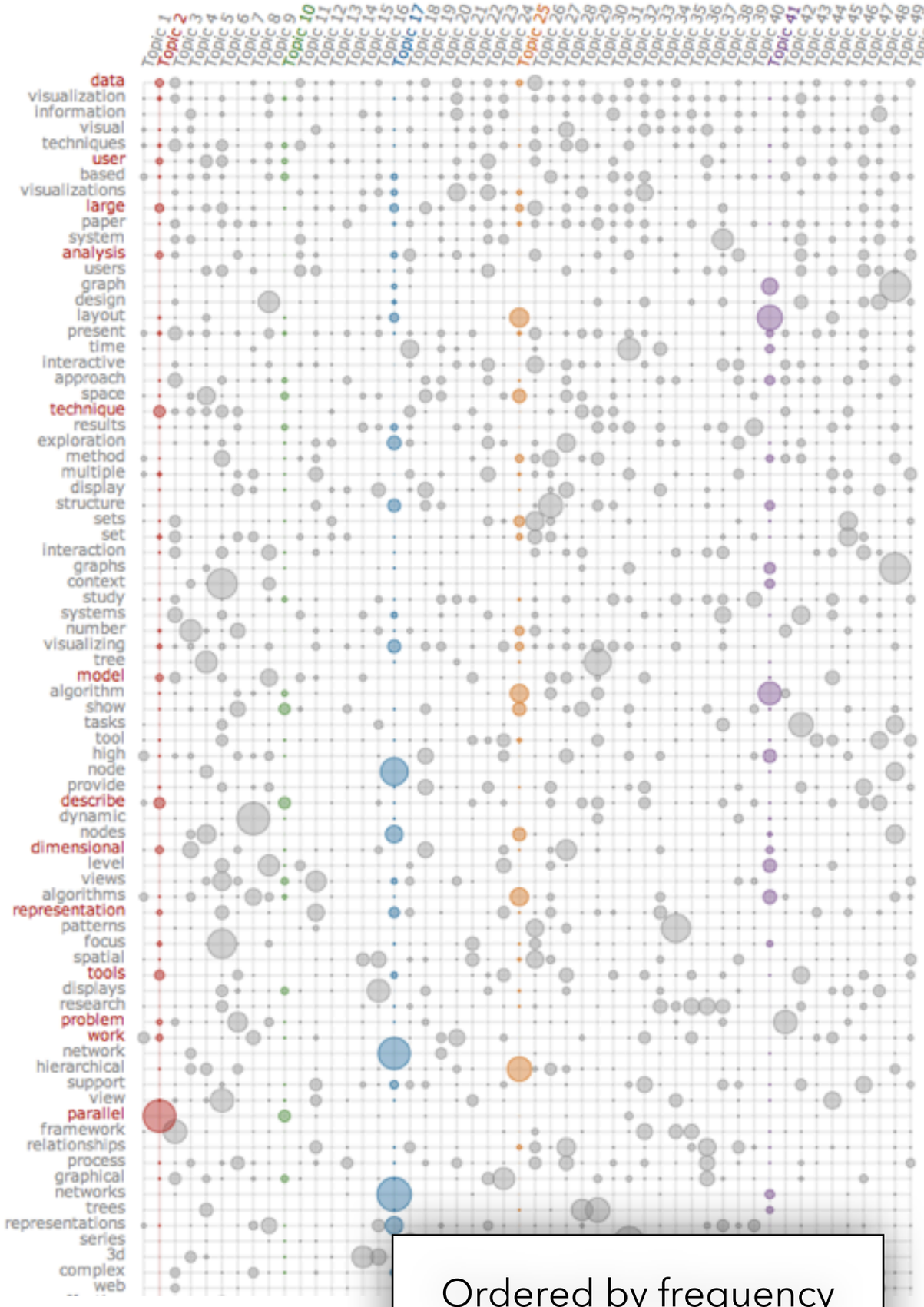
force
directed
graph
layout
algorithms
algorithm
layouts
maintaining
stability



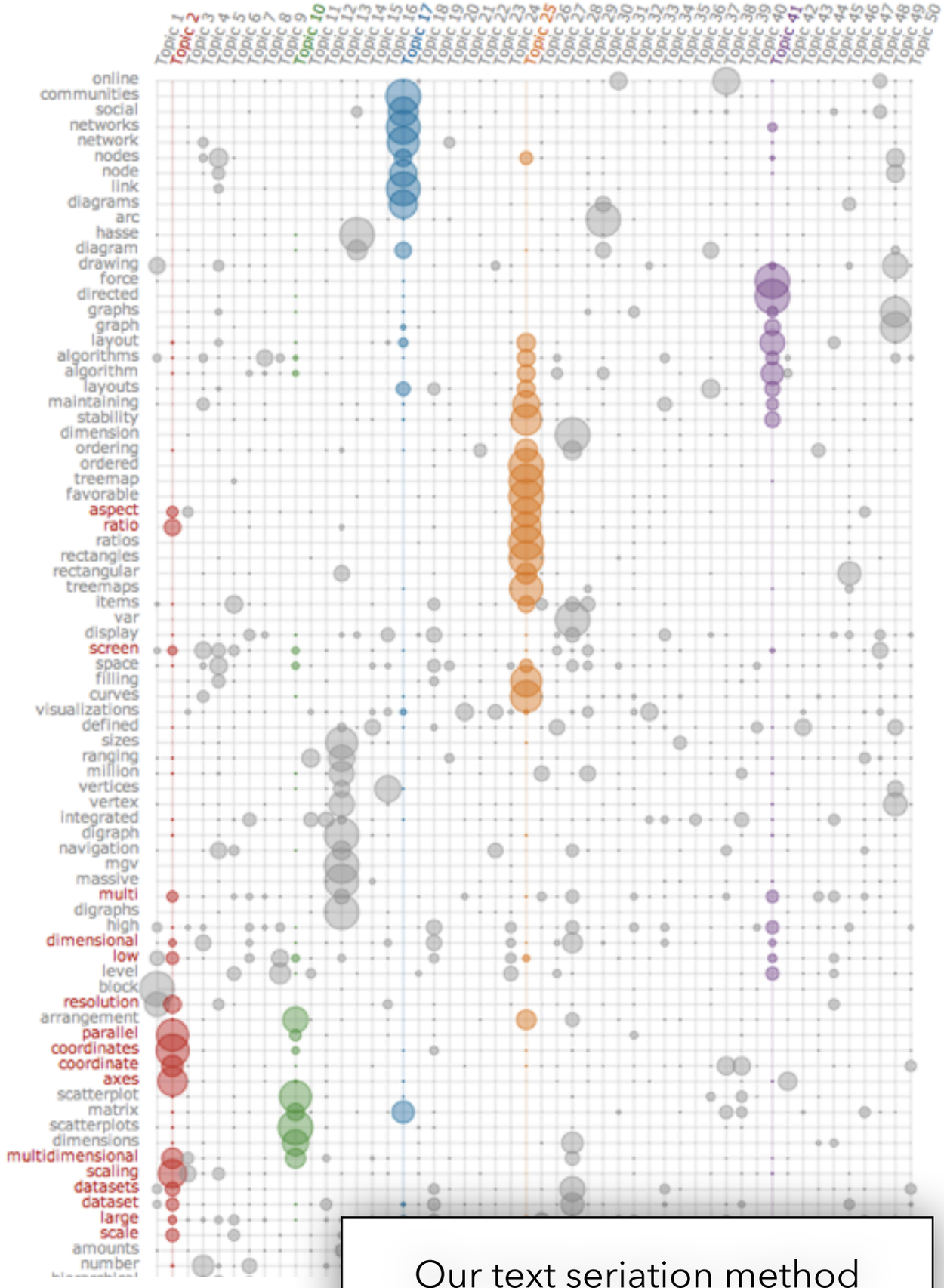
Topical Comparisons

Term Seriation

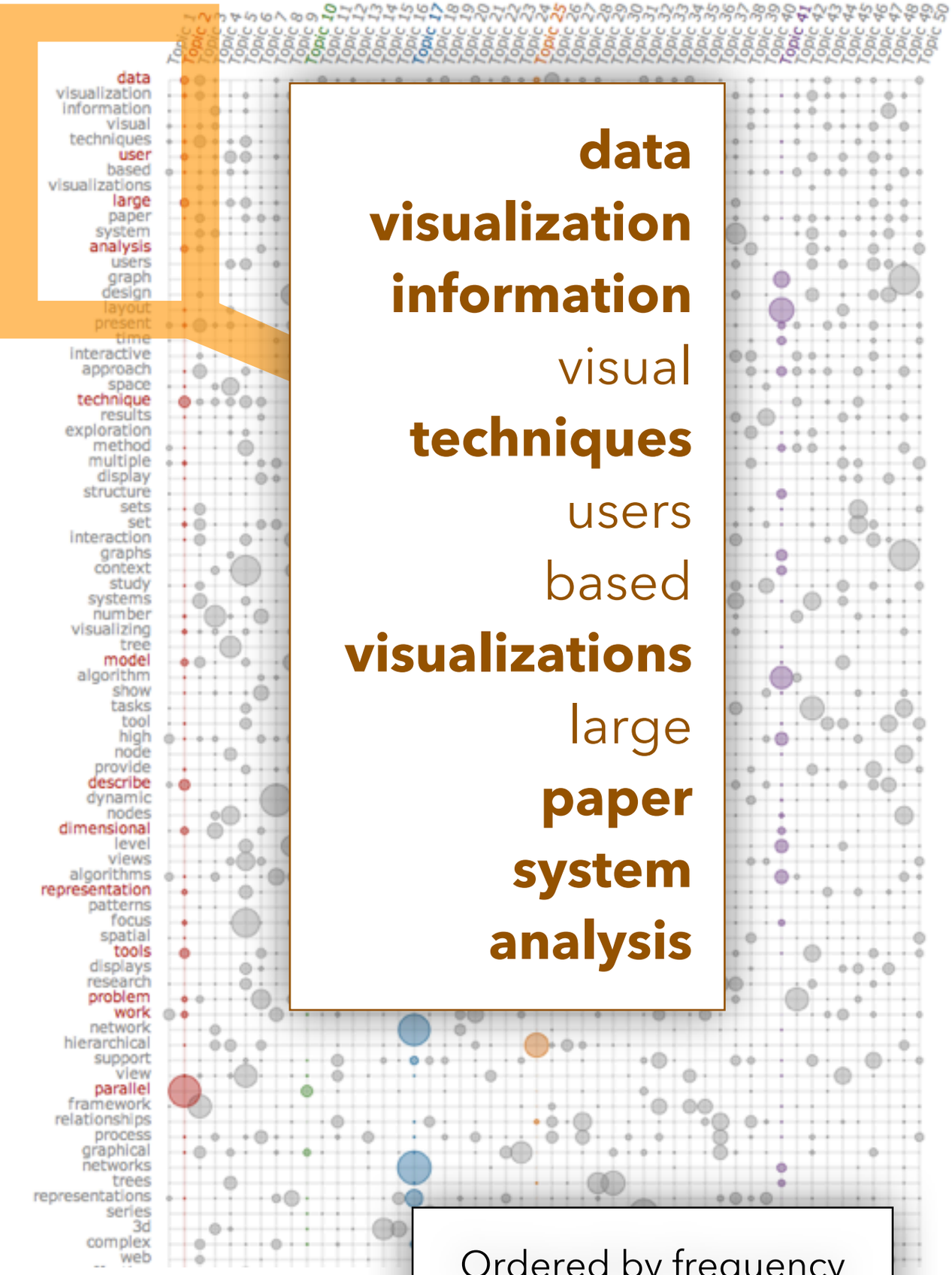




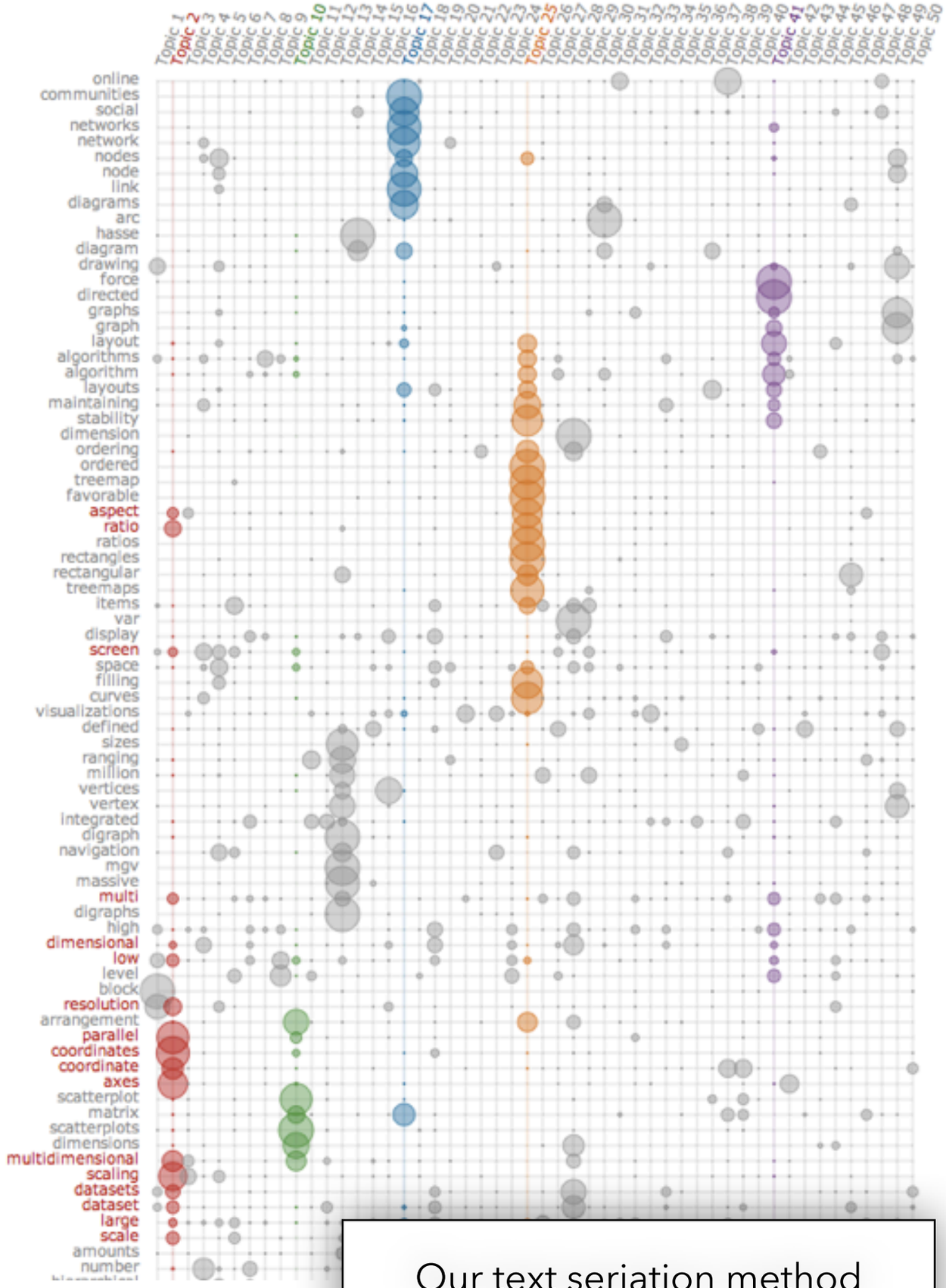
Ordered by frequency



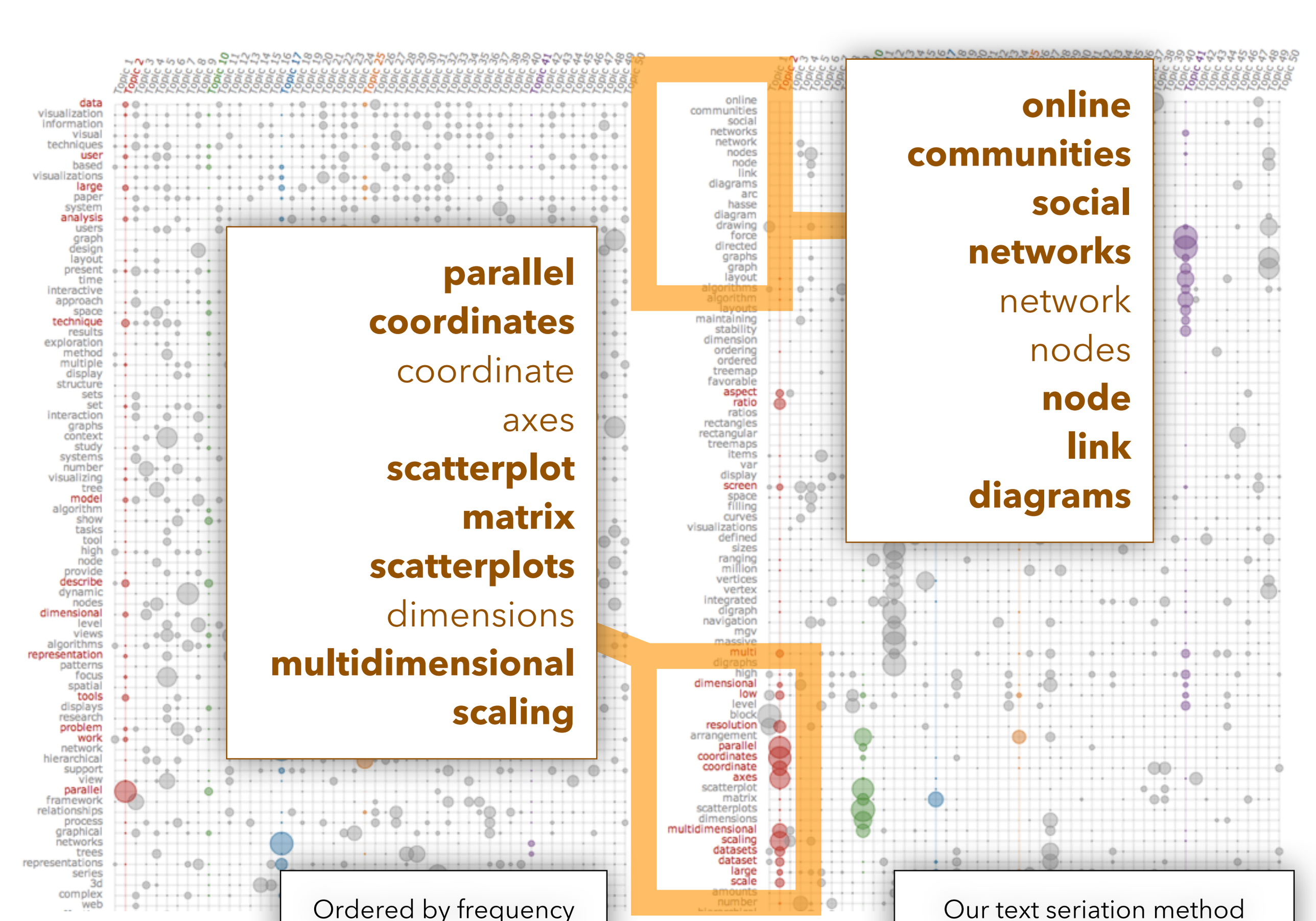
Our text seriation method



Ordered by frequency



Our text seriation method


parallel
coordinates
coordinate
axes
scatterplot
matrix
scatterplots
dimensions
multidimensional
scaling
online
communities
social
networks
network
nodes
node
link
diagrams
Ordered by frequency
Our text seriation method

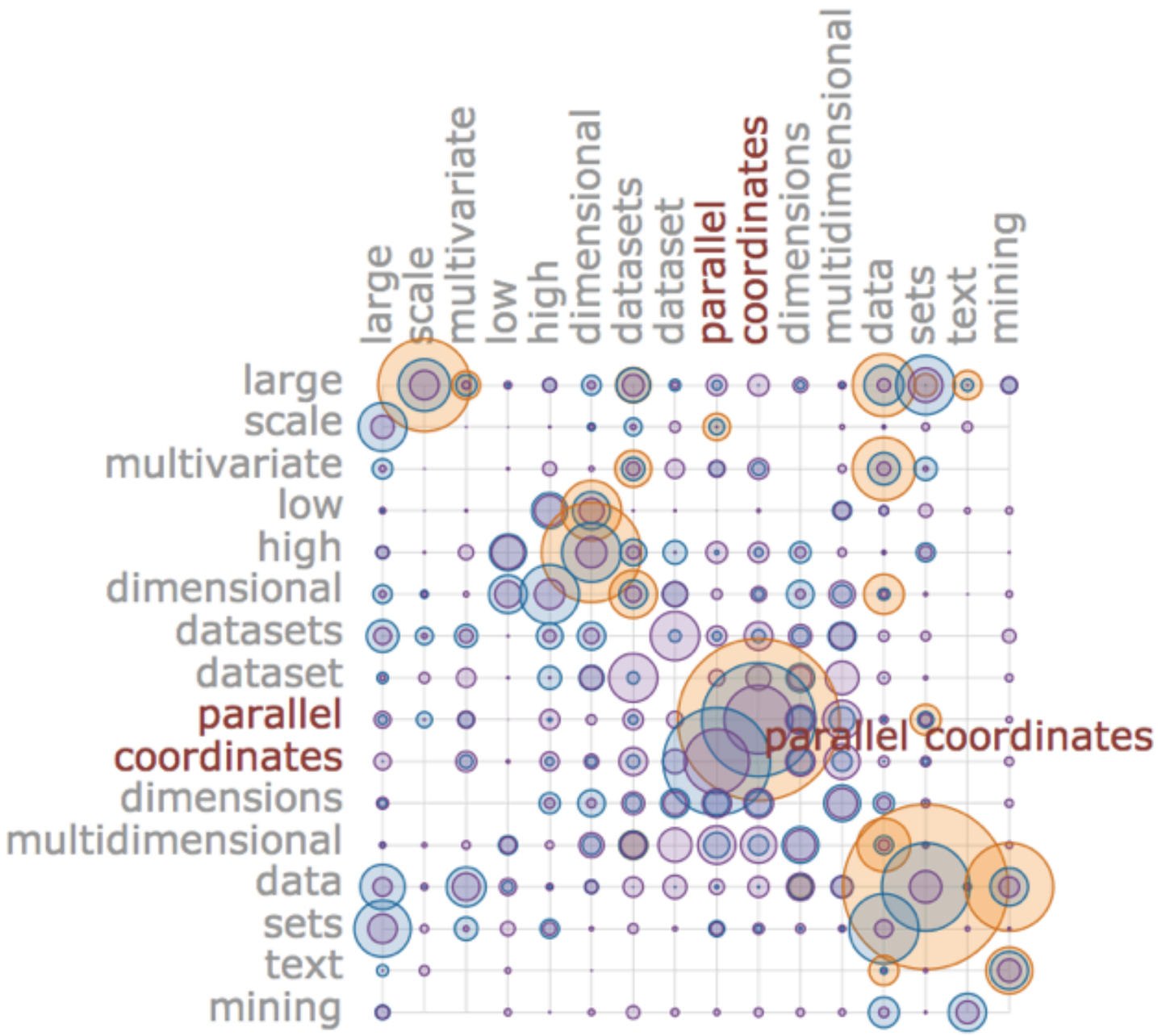
parallel
coordinates
coordinate
axes
scatterplot
matrix
scatterplots
dimensions
multidimensional
scaling

online
communities
social
networks
network
nodes
node
link
diagrams

Ordered by frequency

Preservation of Reading Order to Produce Multi-Word Phrases

Word Similarity & Seriation



G2 statistics [Dunning 1993]

document co-occurrence

sentence co-occurrence

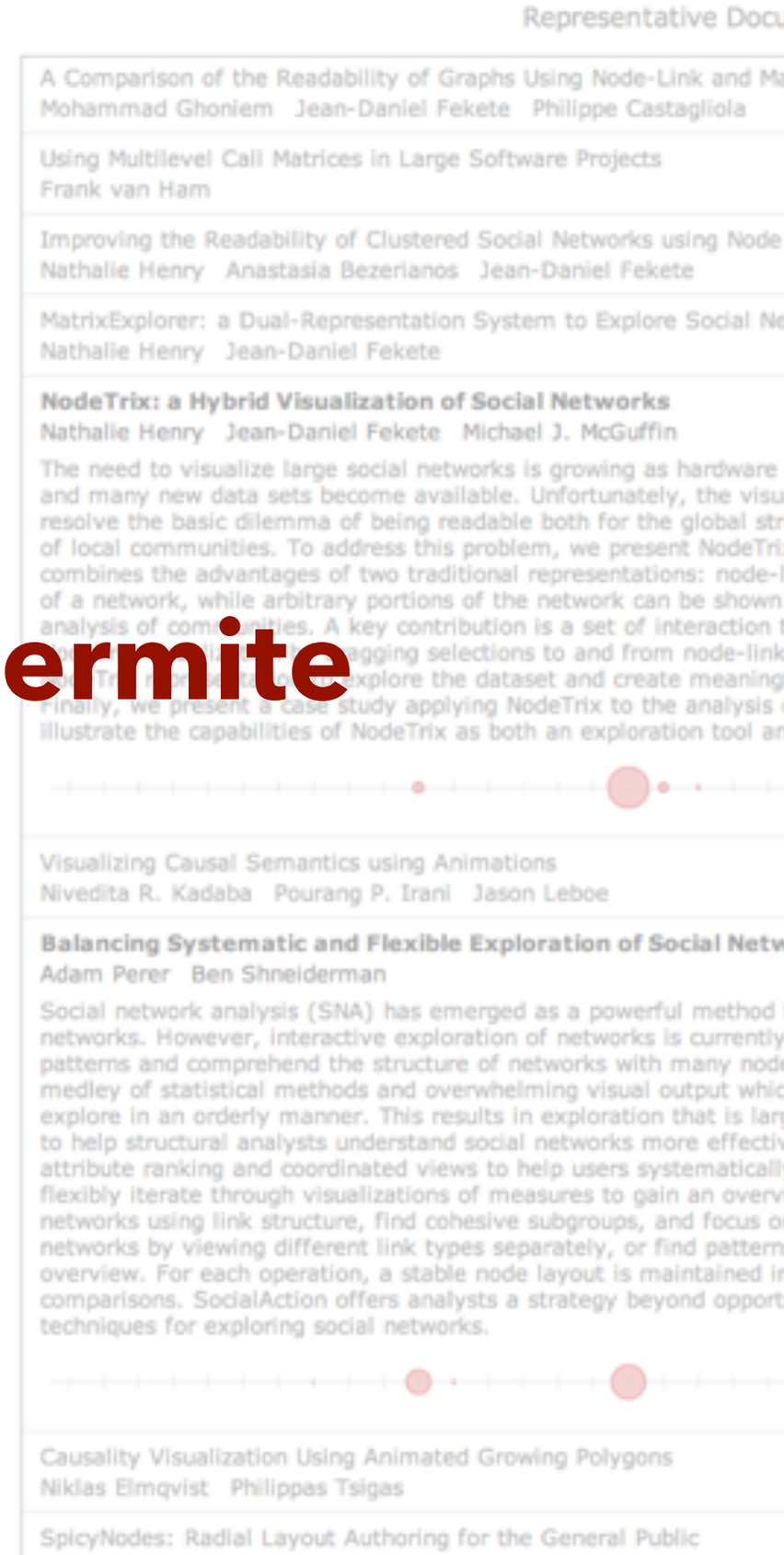
collocation likelihood

Bond energy [McCormick et al. 1972]

asymmetric similarities

greedy algorithm

Termite: Topic Model Visualization

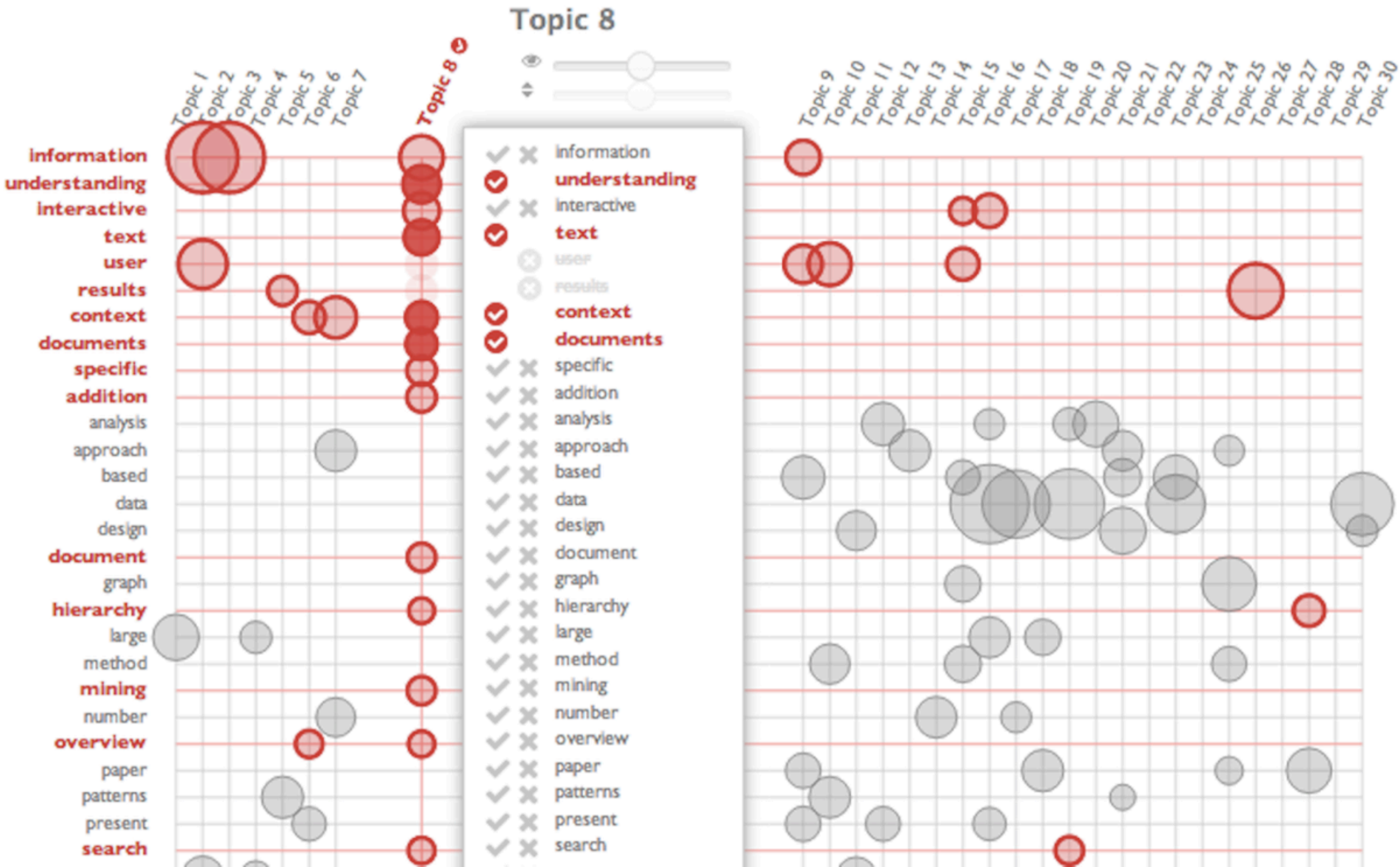


github.com/uwdata/termite

Supporting Topical Analysis Workflow

in collaboration with Yuening Hu, Ashley Jin, John D. Wilkerson,
Daniel A. McFarland, Christopher D. Manning, Jeffrey Heer

Termite + Interactive Topic Modeling



Visualization + Interaction Design

How are topics defined?

Words, people, papers, ...

Similar terms to construct a topic

Multiple levels of details

More specific terms to split a topic

More general terms to refine a topic or merge topics

Interactive Topic Modeling

Improvements

User feedback

Incremental construction of topics

“This is better than anything we have!”

Model designs

Split + merge topics

Insert phrases and metadata

“All models are wrong, but some are useful.”

- George E. P. Box

Analysis Workflow

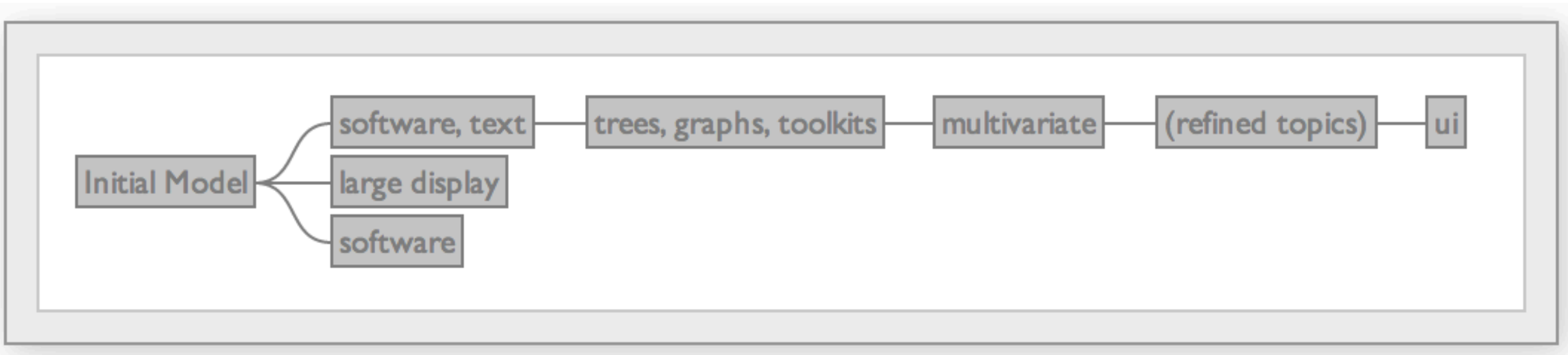
Keeping a "Lab Notebook"

"promoted a topic about wall street"

"Removed 'November' event"

"subcommittee related to congressional hearing"

"Lab Notebook"



Analysis Workflow

Keeping a "Lab Notebook"

"promoted a topic about wall street"

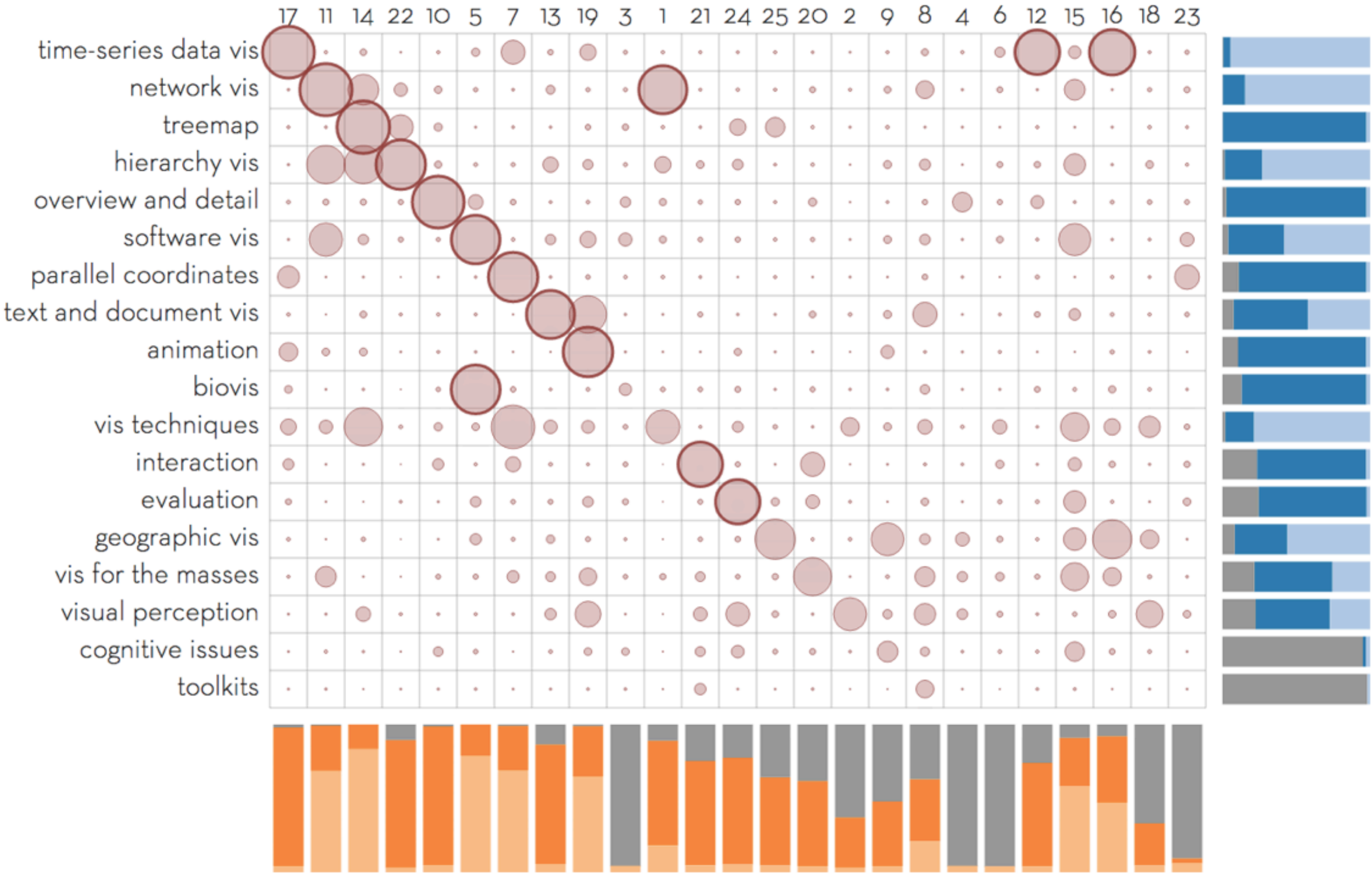
"Removed 'November' event"

"subcommittee related to congressional hearing"

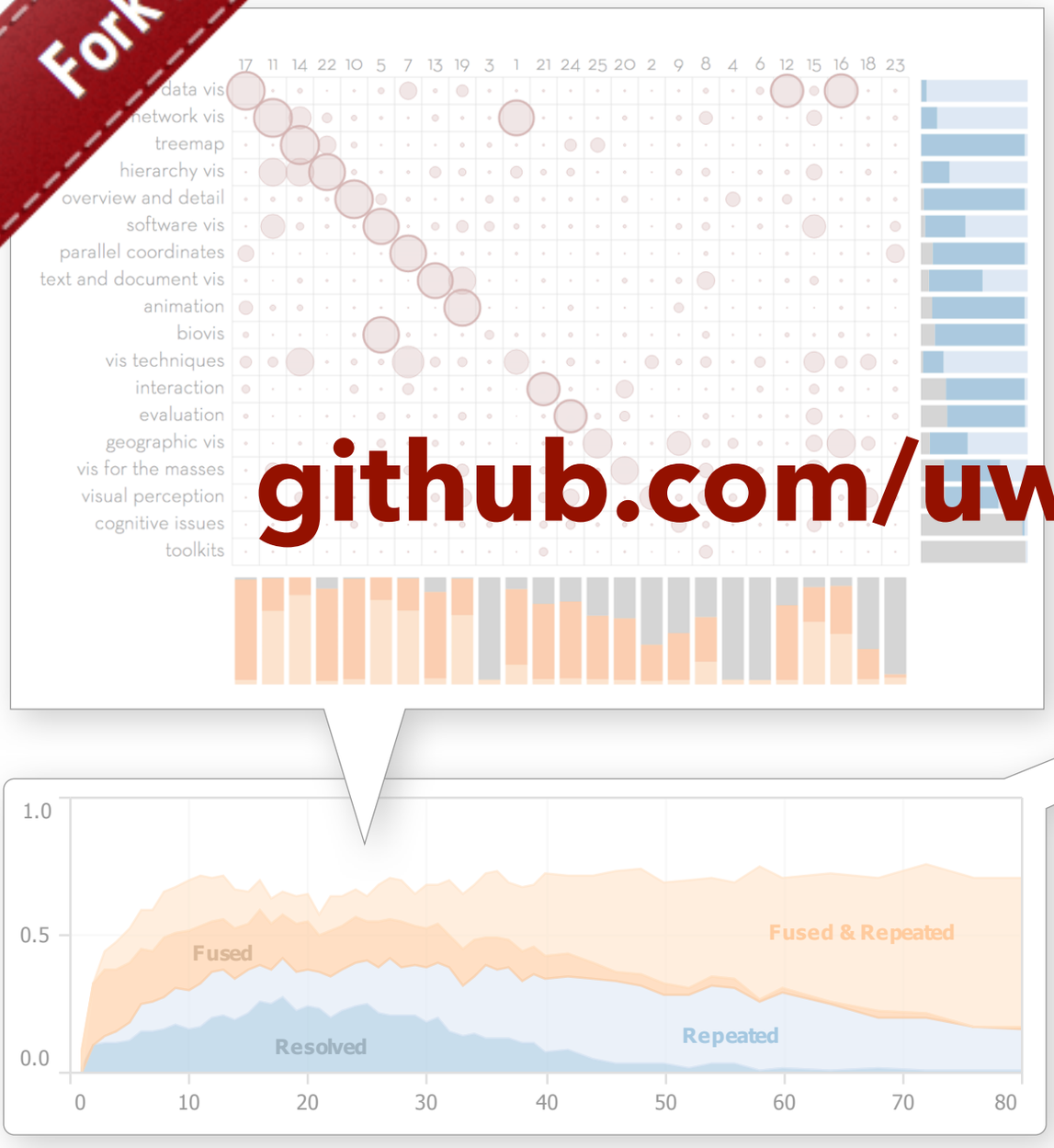
Model diagnostics

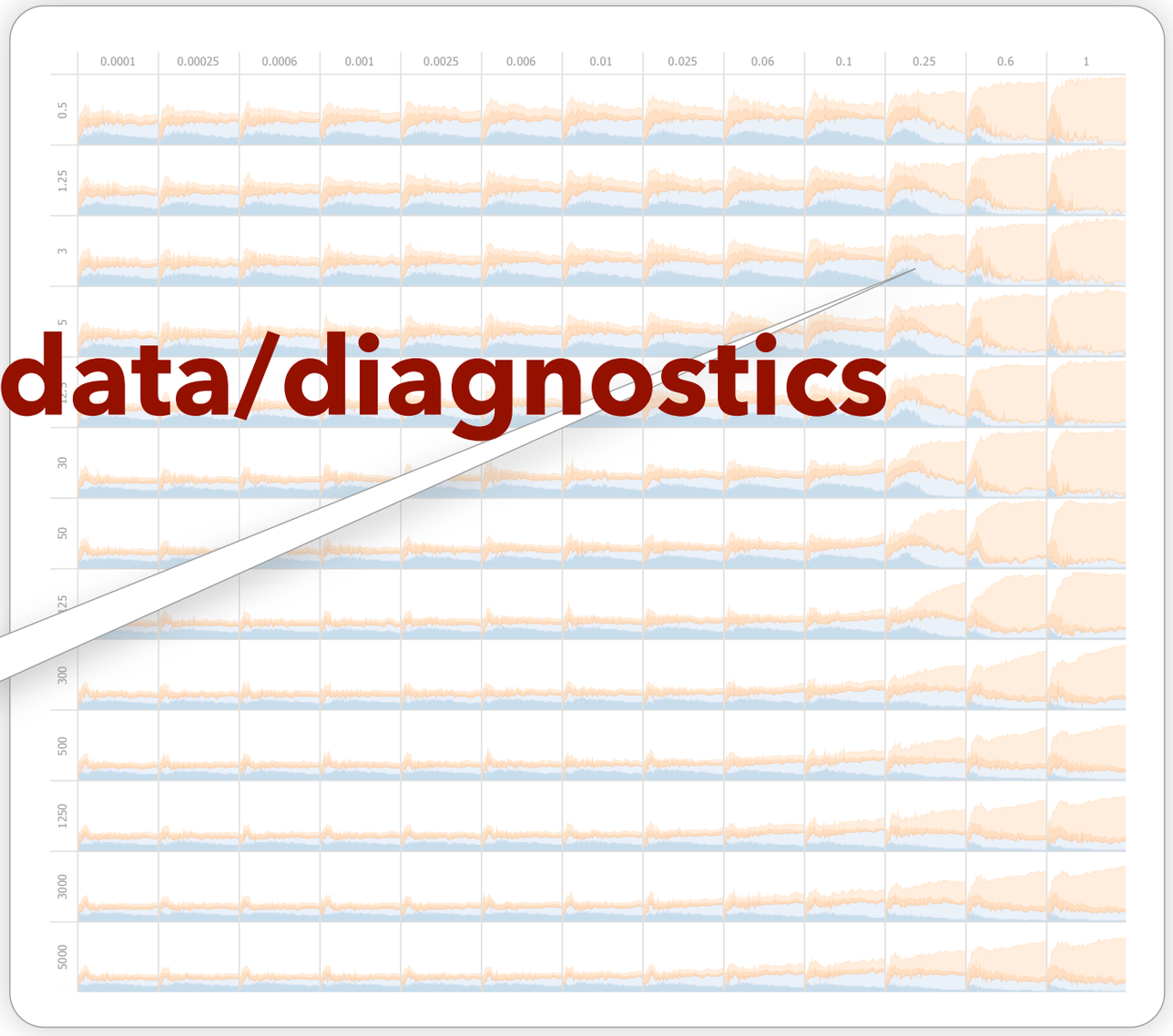
Large-scale model comparison + evaluation

Correspondence Chart



Diagnostics Framework





github.com/uwdata/diagnostics

Analysis Workflow ^and Publication

Keeping a "Lab Notebook"

"promoted a topic about wall street"

"Removed 'November' event"

"subcommittee related to congressional hearing"

Model diagnostics

Large-scale model comparison + evaluation

Interpretable, Accurate, Verifiable

External review of topic models

Data (and model) for academic publications

Thank you.

Questions or comments?

Document Exploration with Topic Models

Designing Interactive Visualizations to Support Effective Analysis Workflows

UW Interactive Data Lab

Jason Chuang
jcchuang@cs.washington.edu

Yuening Hu
ynhu@cs.umd.edu

jason.chuang.ca/research