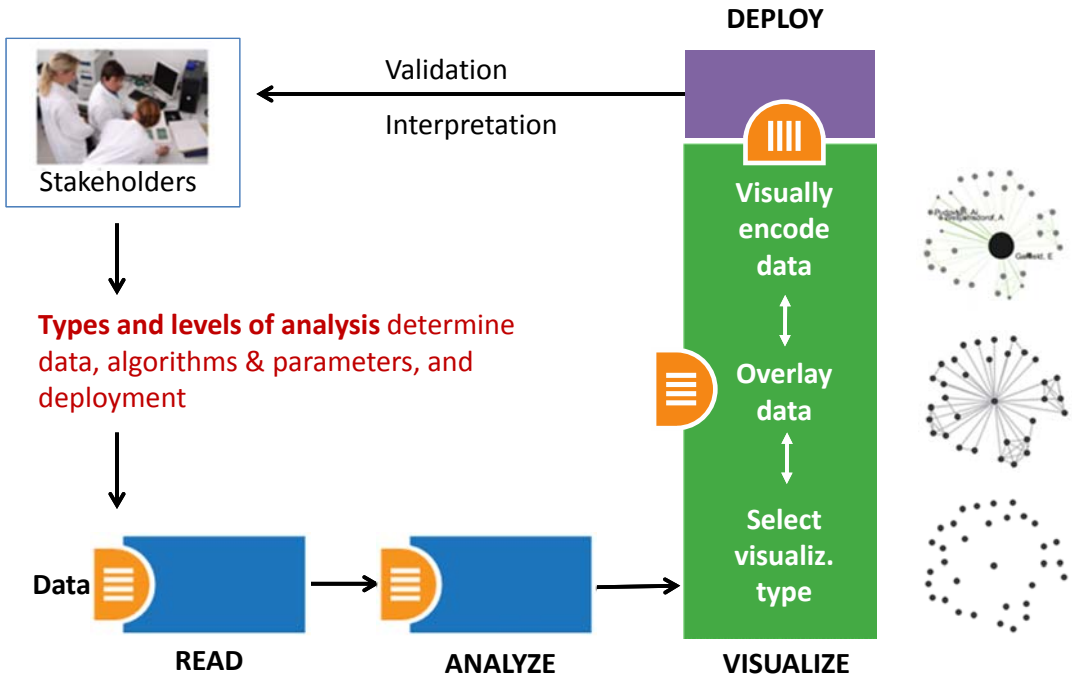
Needs-Driven Workflow Design



Information Visualization MOOC

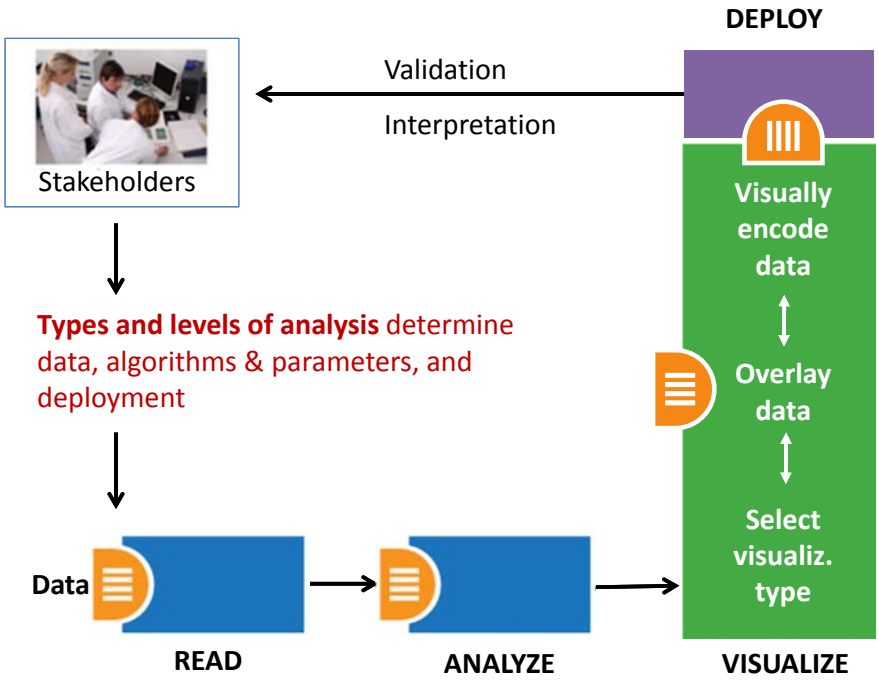
Unit 6 – "With Whom:" Networks

Workflow Design

Reference:

Börner, Katy, Soma Sanyal, and Alessandro Vespignani. 2007. "Network Science." Chap. 12 in *Annual Review of Information Science & Technology*, edited by Blaise Cronin, 537-607. Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Needs-Driven Workflow Design

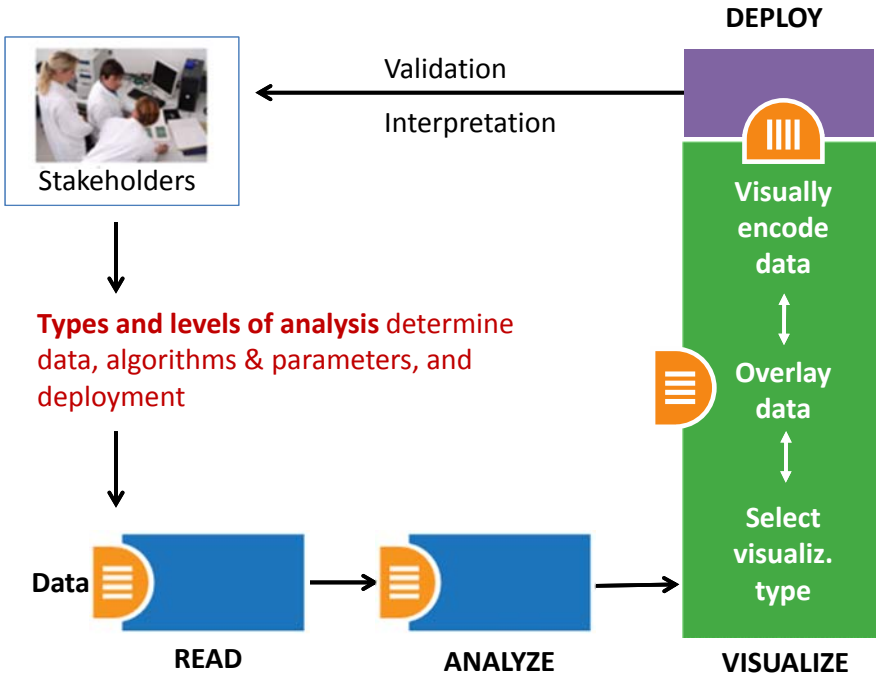


Graphic Variable Types

Modify reference system, add records & links

Visualization Types (reference systems)

Needs-Driven Workflow Design



Read Data

Sample Data:

- UCINet datasets, https://sites.google.com/site/ucinetsoftware/datasets
- Pajek Datasets, http://pajek.imfm.si/doku.php?id=data:index
- Gephi datasets, http://wiki.gephi.org/index.php/Datasets
- CASOS Datasets, http://www.casos.cs.cmu.edu/computational_tools/datasets
- Stanford Large Network Dataset Collection, http://snap.stanford.edu/data/
- Tore Opsahl's Datasets, http://toreopsahl.com/datasets/
- Sci2 Datasets, http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.5+Sample+Datasets and general data sources, http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/8.1+Datasets

Major Data Formats:
See next slide

Input:

Network Formats

- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net)
- NWB (*.nwb)

Scientometric Formats

- ISI (*.isi)
- Bibtex (*.bib)
- Endnote Export Format (*.enw)
- Scopus csv (*.scopus)
- NSF csv (*.nsf)

Other Formats

- Pajek Matrix (*.mat)
- TreeML (*.xml)
- Edgelist (*.edge)
- CSV (*.csv)

Output:

Network File Formats

- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)

Image Formats

- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps)

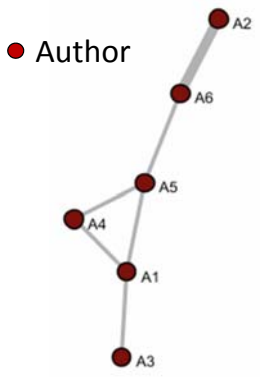
Preprocess Data

- Unification, see Unit 4
- Network extraction
- Delete isolate nodes
- Remove self-loops
- Threshold—e.g., extract nodes/edges above or below value
- Merge two networks
- Pathfinder network scaling, see Backbone Identification

Preprocessing—Network Extraction

Weighted, undirected co-occurrence network.

	A	B	C	D
1	Paper	Authors	References	Year
2	P1	A1		1970
3	P2	A2;A6	P1	1980
4	P3	A1;A3	P1;P2	1990
5	P4	A1;A4;A5	P2	1995
6	P5	A5;A6	P1;P2;P3;P4	1995
7	P6	A2;A6	P5	2000

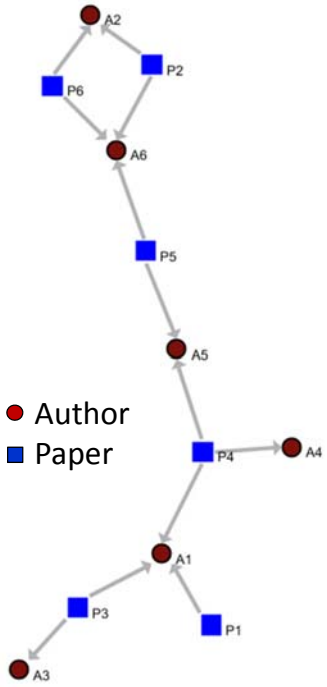



```
*Vertices 6
1 A1
2 A6
3 A2
4 A3
5 A5
6 A4
*Edges 6
2 3 2
1 4 1
1 5 1
5 6 1
1 6 1
2 5 1
```

Preprocessing—Network Extraction

Unweighted, directed bimodal network.

	A	B	C	D
1	Paper	Authors	References	Year
2	P1	A1		1970
3	P2	A2;A6	P1	1980
4	P3	A1;A3	P1;P2	1990
5	P4	A1;A4;A5	P2	1995
6	P5	A5;A6	P1;P2;P3;P4	1995
7	P6	A2;A6	P5	2000



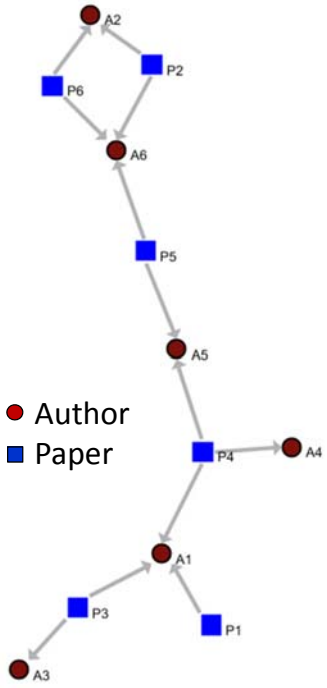

```
*Vertices 12
1 P1 bipartitetype "Paper"
2 A1 bipartitetype "Authors"
3 P2 bipartitetype "Paper"
4 A2 bipartitetype "Authors"
5 A6 bipartitetype "Authors"
6 P3 bipartitetype "Paper"
7 A3 bipartitetype "Authors"
8 P4 bipartitetype "Paper"
9 A4 bipartitetype "Authors"
10 A5 bipartitetype "Authors"
11 P5 bipartitetype "Paper"
12 P6 bipartitetype "Paper"
*Arcs
1 2
3 4
3 5
6 2
6 7
8 2
8 10
8 9
11 5
11 10
12 4
12 5
```


Preprocessing—Network Extraction

Unweighted, directed network of two types.

Calculate node degrees.

	A	B	C	D
1	Paper	Authors	References	Year
2	P1	A1		1970
3	P2	A2;A6	P1	1980
4	P3	A1;A3	P1;P2	1990
5	P4	A1;A4;A5	P2	1995
6	P5	A5;A6	P1;P2;P3;P4	1995
7	P6	A2;A6	P5	2000



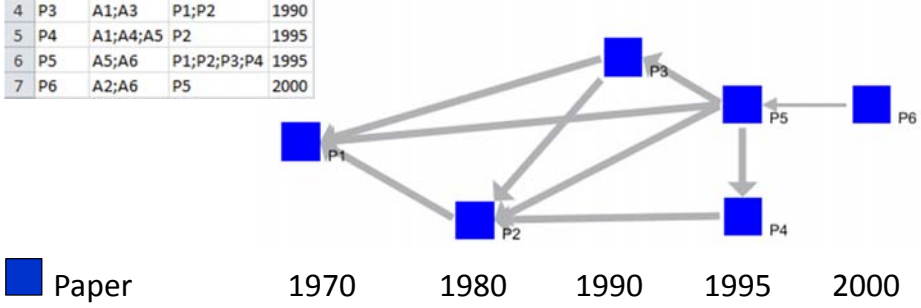

```
*Vertices 12
1 P1 indegree 0
2 A1 indegree 3
3 P2 indegree 0
4 A2 indegree 2
5 A6 indegree 3
6 P3 indegree 0
7 A3 indegree 1
8 P4 indegree 0
9 A4 indegree 1
10 A5 indegree 2
11 P5 indegree 0
12 P6 indegree 0
*Arcs
1 2
3 4
3 5
6 2
6 7
8 10
8 2
8 9
11 10
11 5
12 4
12 5
```

Preprocessing—Network Extraction

Unweighted, directed paper-citation network.

Arcs go from papers to references.

	A	B	C	D
1	Paper	Authors	References	Year
2	P1	A1		1970
3	P2	A2;A6	P1	1980
4	P3	A1;A3	P1;P2	1990
5	P4	A1;A4;A5	P2	1995
6	P5	A5;A6	P1;P2;P3;P4	1995
7	P6	A2;A6	P5	2000

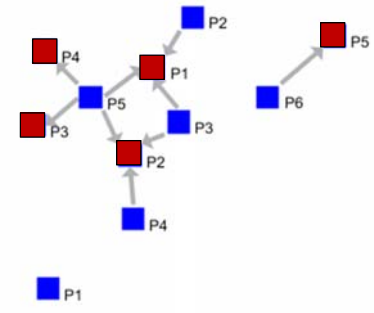



```
*Vertices 6
1 P1
2 P2
3 P3
4 P4
5 P5
6 P6
*Arcs
2 1
3 1
3 2
4 2
5 4
5 3
5 1
5 2
6 5
```


Preprocessing—Network Extraction

Unweighted, directed <u>bipartite</u> network.

	A	B	C	D
1	Paper	Authors	References	Year
2	P1	A1		1970
3	P2	A2;A6	P1	1980
4	P3	A1;A3	P1;P2	1990
5	P4	A1;A4;A5	P2	1995
6	P5	A5;A6	P1;P2;P3;P4	1995
7	P6	A2;A6	P5	2000



WRONG!!!

Paper

```
*Vertices 11
1 P1 bipartitetype "Paper"
2 P2 bipartitetype "Paper"
3 P1 bipartitetype "References"
4 P3 bipartitetype "Paper"
5 P2 bipartitetype "References"
6 P4 bipartitetype "Paper"
7 P5 bipartitetype "Paper"
8 P4 bipartitetype "References"
9 P3 bipartitetype "References"
10 P6 bipartitetype "Paper"
11 P5 bipartitetype "References"
*Arcs
2 3
4 3
4 5
6 5
7 3
7 9
7 5
7 8
10 11
```

Network Analysis

Calculate

- Node and link properties
- Network properties
- Statistical properties
- Network types

- Extract relevant subtrees

- Calculate error and attack tolerance
- Compute clusters and backbones, see later parts in this unit

Visualization Goals

Representing hierarchical data

- Structural information
- Content information

Objectives

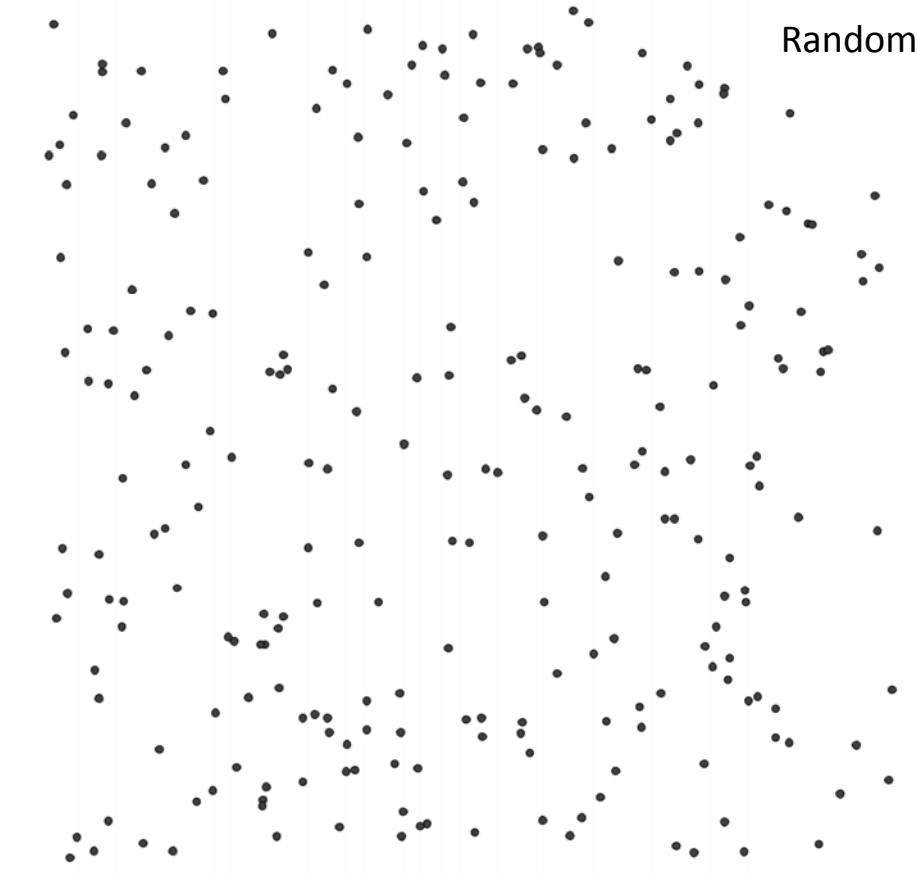
- Efficient space utilization
- Comprehension
- Interactivity
- Aesthetics

Example: Collaboration Network

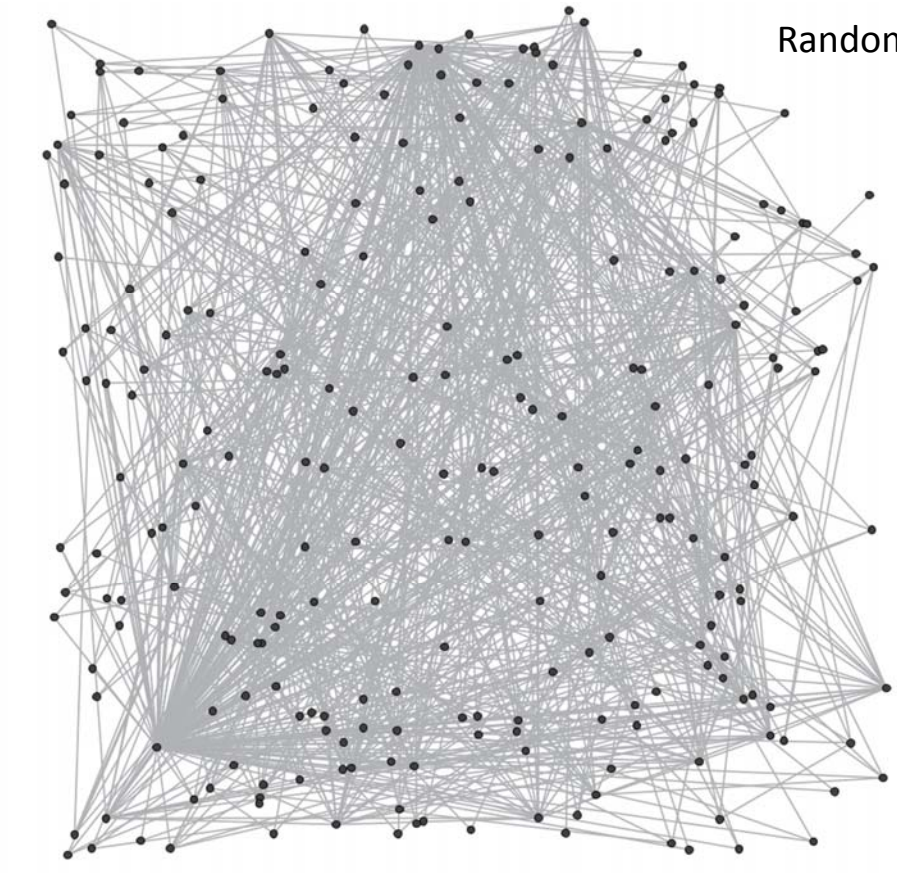
- Random layout
- Circular layout
- Generalized Expectation-Maximization (GEM) layout

First, determine node layout
Then add nodes and linkages

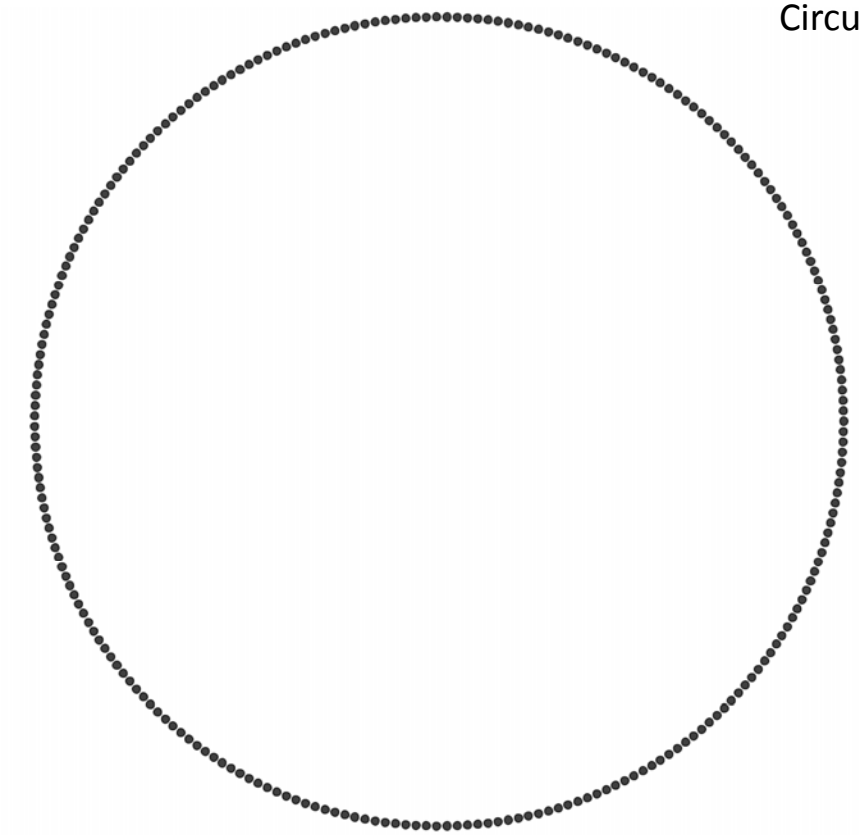
For GEM layout, color- and size-code nodes and links and add a legend.



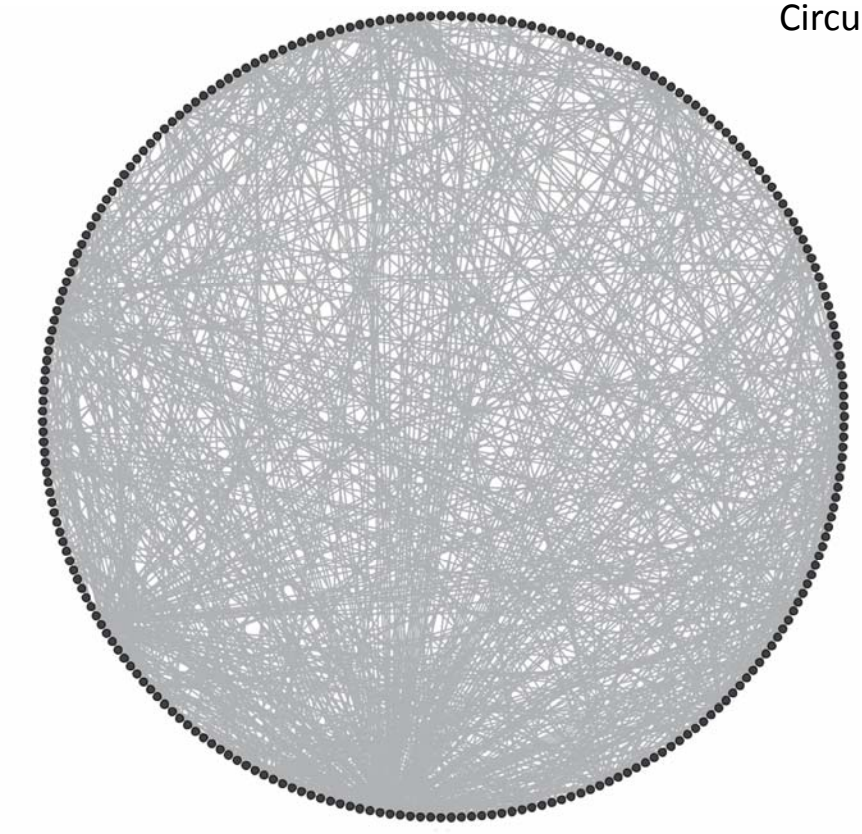
Random layout



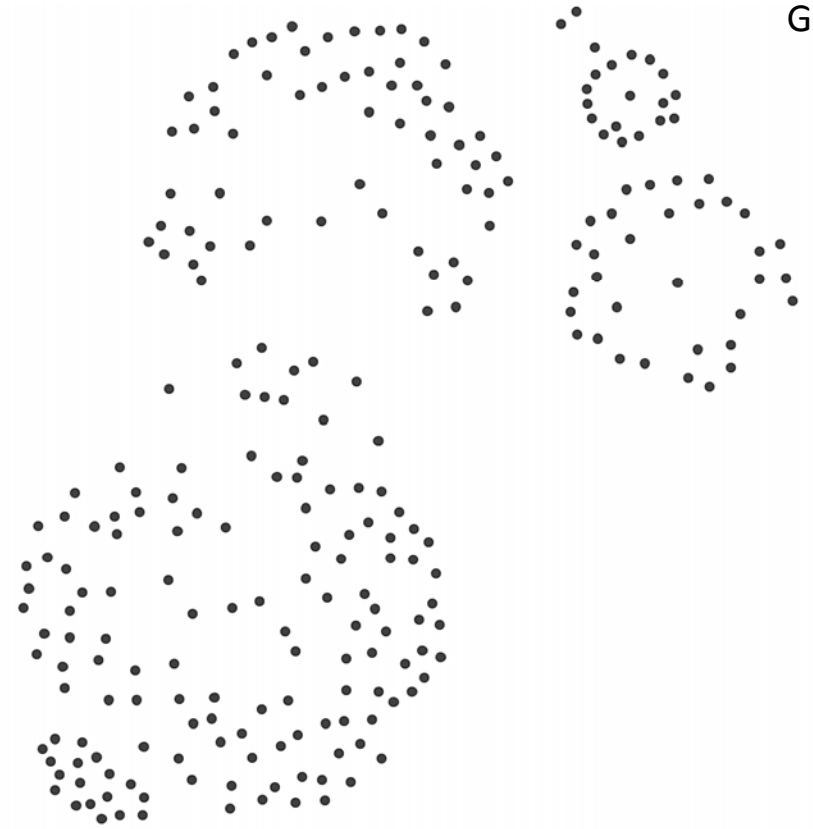
Circular layout



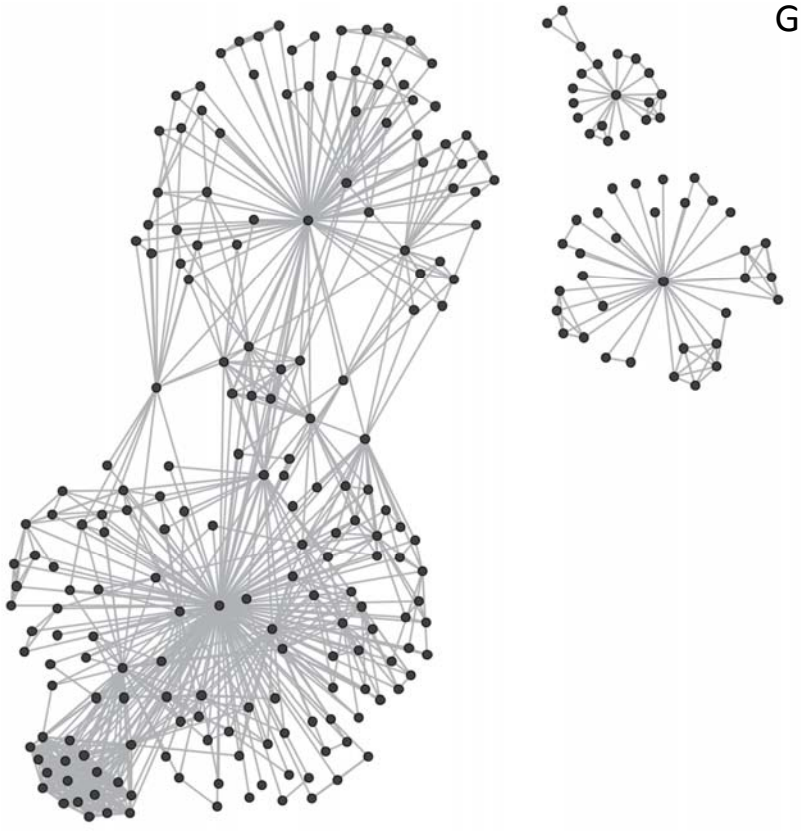
Circular layout



GEM layout

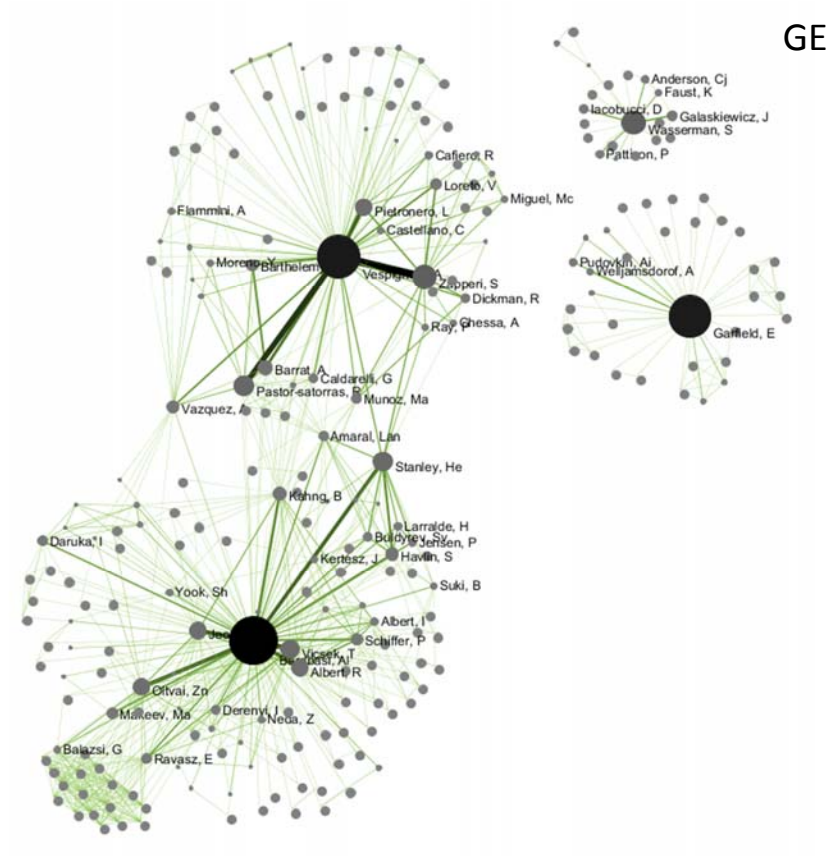


GEM layout

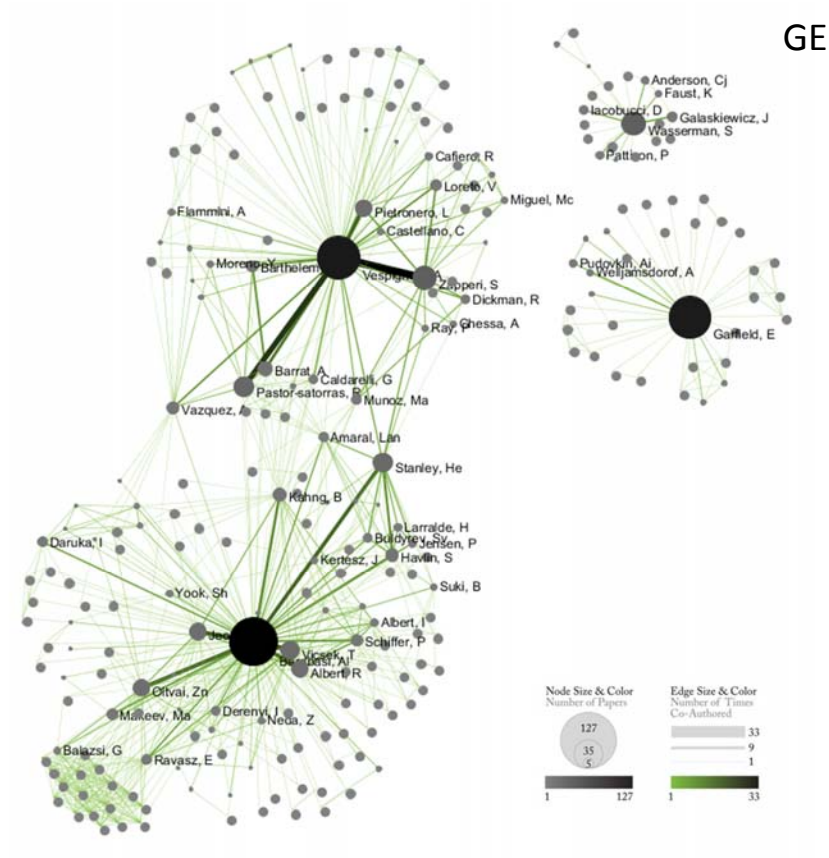


GEM layout

GEM layout



GEM layout



Aesthetic Criteria for Graph Drawing

- Maximize symmetry
- Evenly distributed nodes
- Uniform edge lengths
- Minimized edge crossings
- Orthogonal drawings
- Minimize area, bends, slopes, angles
- Maximize consistent flow direction (in directed networks)

Optimization criteria may be relaxed to speed up layout process.

Visualizing Large Networks

Discover landmark nodes based on

- Existing node attributes—e.g., frequency of access
- Connectivity (hubs & authorities)
- Depth in a hierarchy

Identify strong (and weak) links

Identify backbone, see Backbone Identification

Show clusters, see Clustering

Interacting With Networks

Modify focusing parameters while continuously providing visual feedback and updating display (fast computer response).

- Conditioning: filter, set background variables and display foreground parameters
- Identification: highlight, color, shape code
- Parameter control: line thickness, length, color legend, time slider, and animation control
- Navigation: bird's-eye view, zoom, and pan
- Information requests: mouse over or click on a node to retrieve more details or collapse/expand a subnetwork