Working with network data

Jim Bagrow
james.bagrow@uvm.edu
bagrow.com

Complex Networks
Winter Workshop
2019-12-16

The University of Vermont

VERMONT
COMPLEX SYSTEMS CENTER

About myself

Understanding networks from data

Community detection

Link communities reveal multiscale complexity in networks

Yong-Yeol Ahn[1,2]*, James P. Bagrow[1,2]* & Sune Lehmann[3,4]*

Ahn *et al.* (2010)

A Local Method for Detecting Communities

James P. Bagrow[1] and Erik M. Bollt[2,1]
[1] *Department of Physics, Clarkson University, Potsdam, NY 13699-5820, USA.*
[2] *Department of Math and Computer Science, Clarkson University, Potsdam, NY 13699-5815, USA.*

May 16, 2006

Bagrow & Bollt (2005)

PHYSICAL REVIEW E **85**, 066118 (2012)

Communities and bottlenecks: Trees and treelike networks have high modularity

James P. Bagrow*
Department of Engineering Sciences and Applied Mathematics, Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois 60208, USA

(Received 2 January 2012; published 15 June 2012)

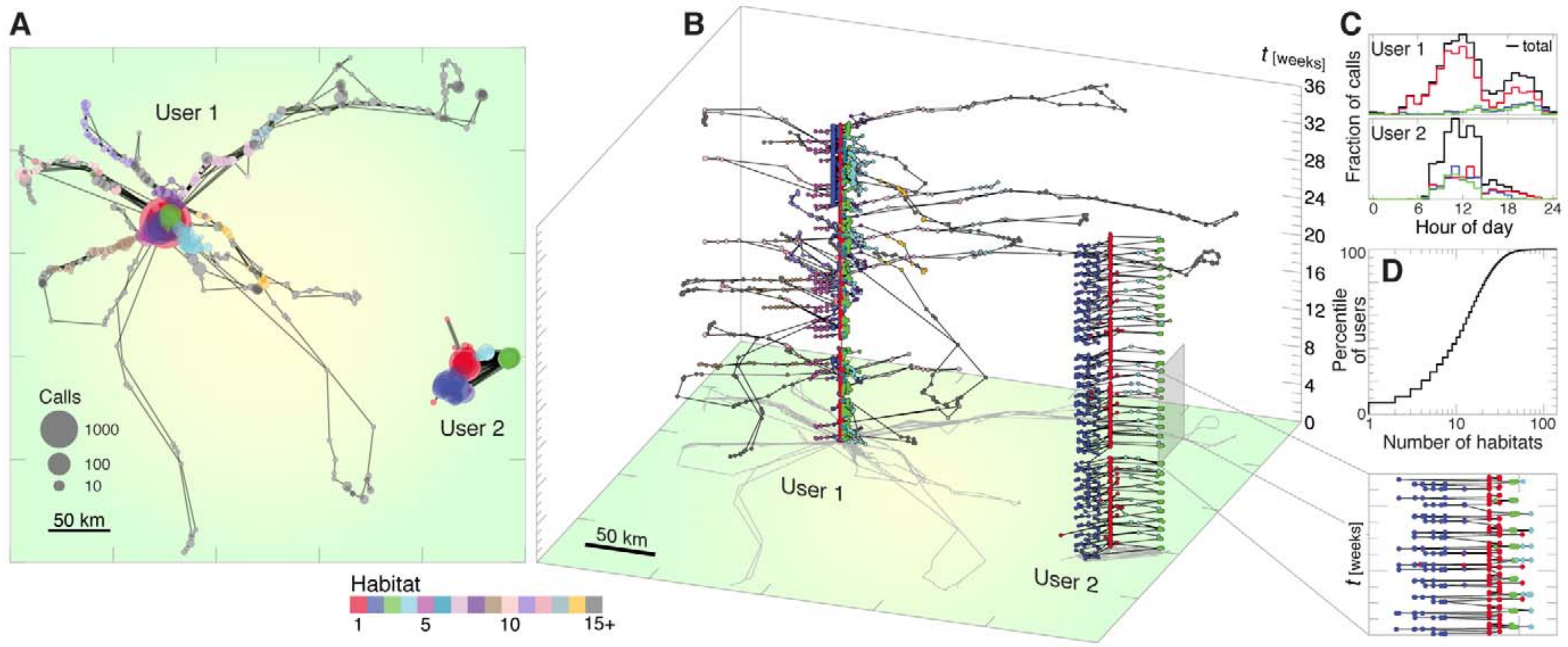
Bagrow (2012)

Applied to data

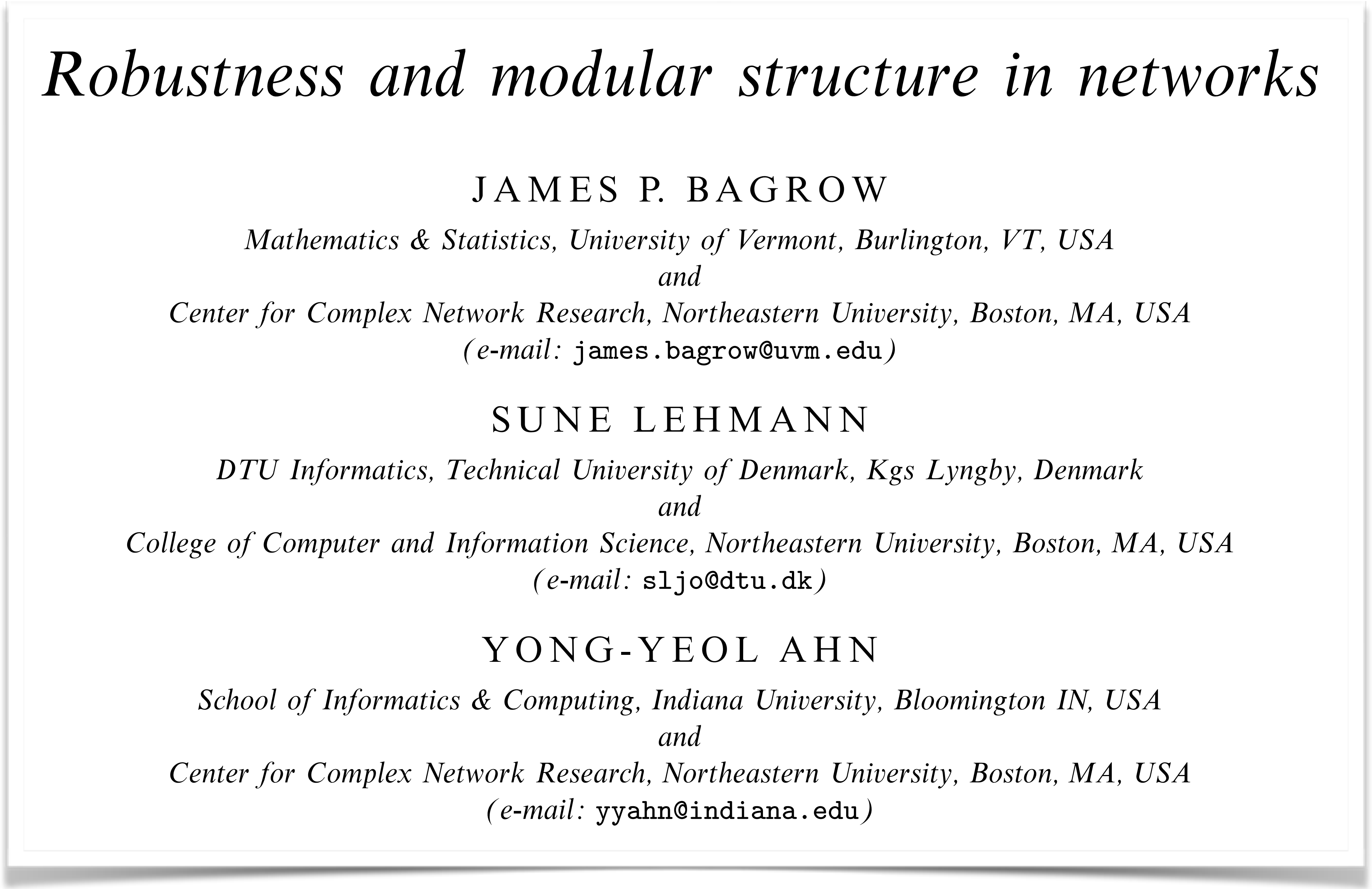
Mesoscopic Structure and Social Aspects of Human Mobility

James P. Bagrow[1,2]*, Yu-Ru Lin[3,4]

Bagrow & Lin (2012)



Data + Models


Robustness and modular structure in networks

JAMES P. BAGROW

Mathematics & Statistics, University of Vermont, Burlington, VT, USA
and
Center for Complex Network Research, Northeastern University, Boston, MA, USA
(*e-mail:* james.bagrow@uvm.edu)

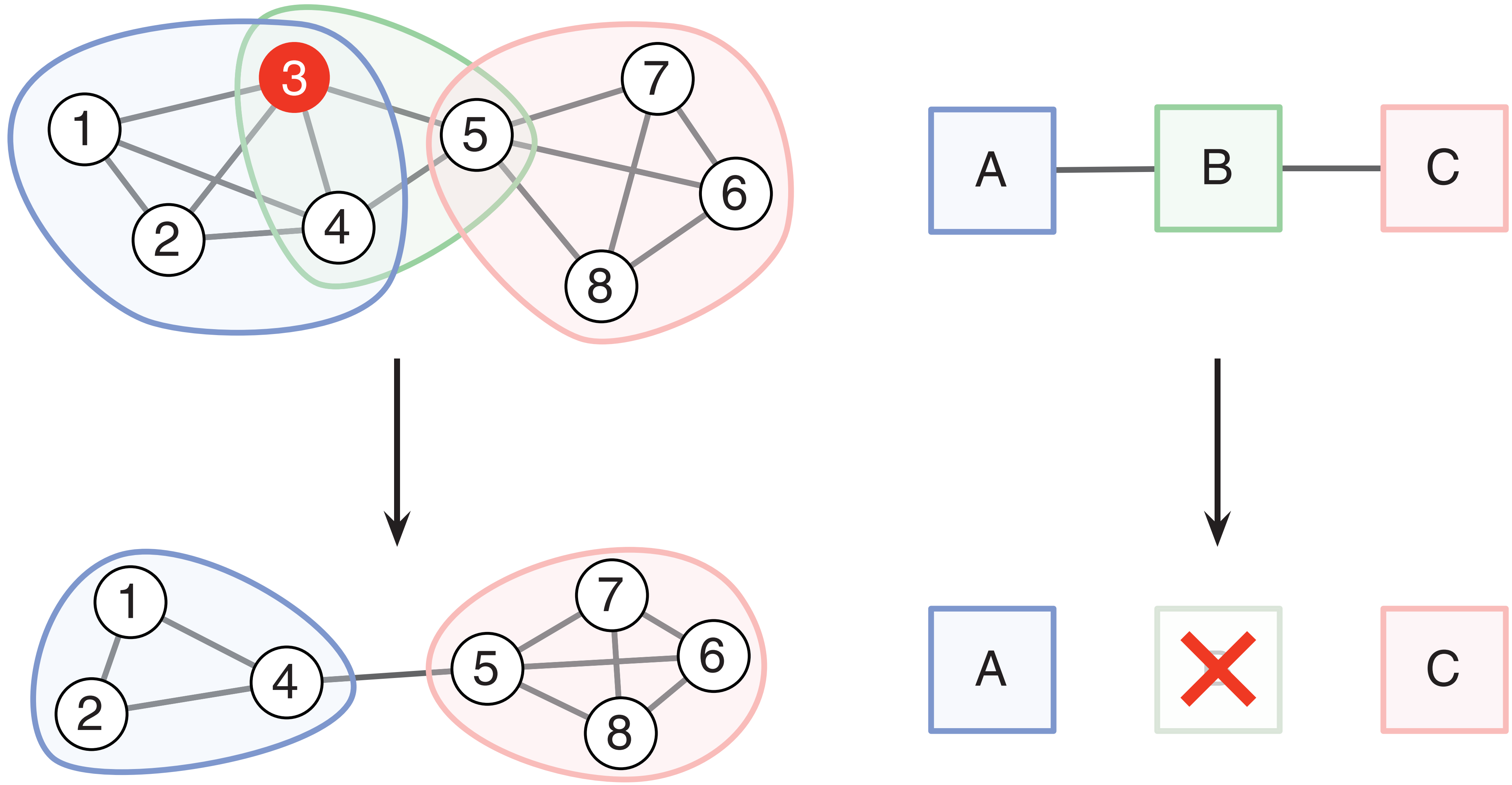
SUNE LEHMANN

DTU Informatics, Technical University of Denmark, Kgs Lyngby, Denmark
and
College of Computer and Information Science, Northeastern University, Boston, MA, USA
(*e-mail:* sljo@dtu.dk)

YONG-YEOL AHN

School of Informatics & Computing, Indiana University, Bloomington IN, USA
and
Center for Complex Network Research, Northeastern University, Boston, MA, USA
(*e-mail:* yyahn@indiana.edu)

How does missing data change the appearance of communities?



Bagrow *et al.* (2015)

Data + Models

The quoter model: A paradigmatic model of the social flow of written information

James P. Bagrow[1,a)] and Lewis Mitchell[2,b)]
[1]Department of Mathematics and Statistics, University of Vermont, Burlington, Vermont 05405, USA
[2]School of Mathematical Sciences, North Terrace Campus, The University of Adelaide, Adelaide, South Australia 5005, Australia

(Received 31 October 2017; accepted 23 February 2018; published online 11 July 2018)

Measuring the flow of information between individuals

nature human behaviour
LETTERS
https://doi.org/10.1038/s41562-018-0510-5

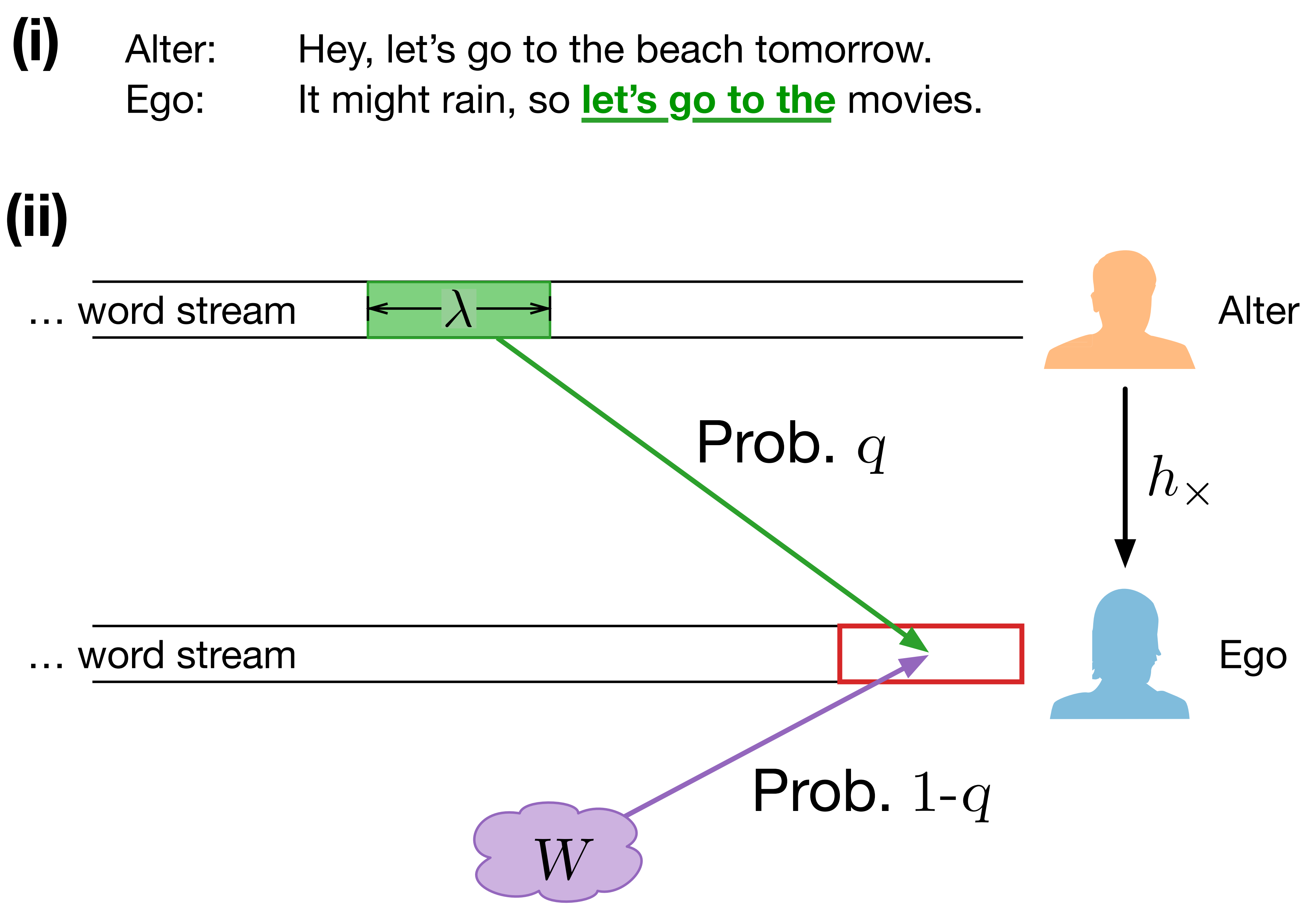
Information flow reveals prediction limits in online social activity

James P. Bagrow[1,2*], Xipei Liu[1,2] and Lewis Mitchell[1,2,3*]

Modern society depends on the flow of information over online social networks, and users of popular platforms generate substantial behavioural data about themselves and their social ties[1–5]. However, it remains unclear what fundamental limits exist when using these data to predict the activities and interests of individuals, and to what accuracy such predictions can be made using an individual's social ties. Here, we show

postings to online social platforms present a unique opportunity to explore the textual content of messages in conjunction with their timings, giving a richer understanding of social ties.

Information theory allows us to mathematically quantify the information contained in data and is well suited to data in the form of online written communication. Although the mathematical definition of information is somewhat distinct from our com-

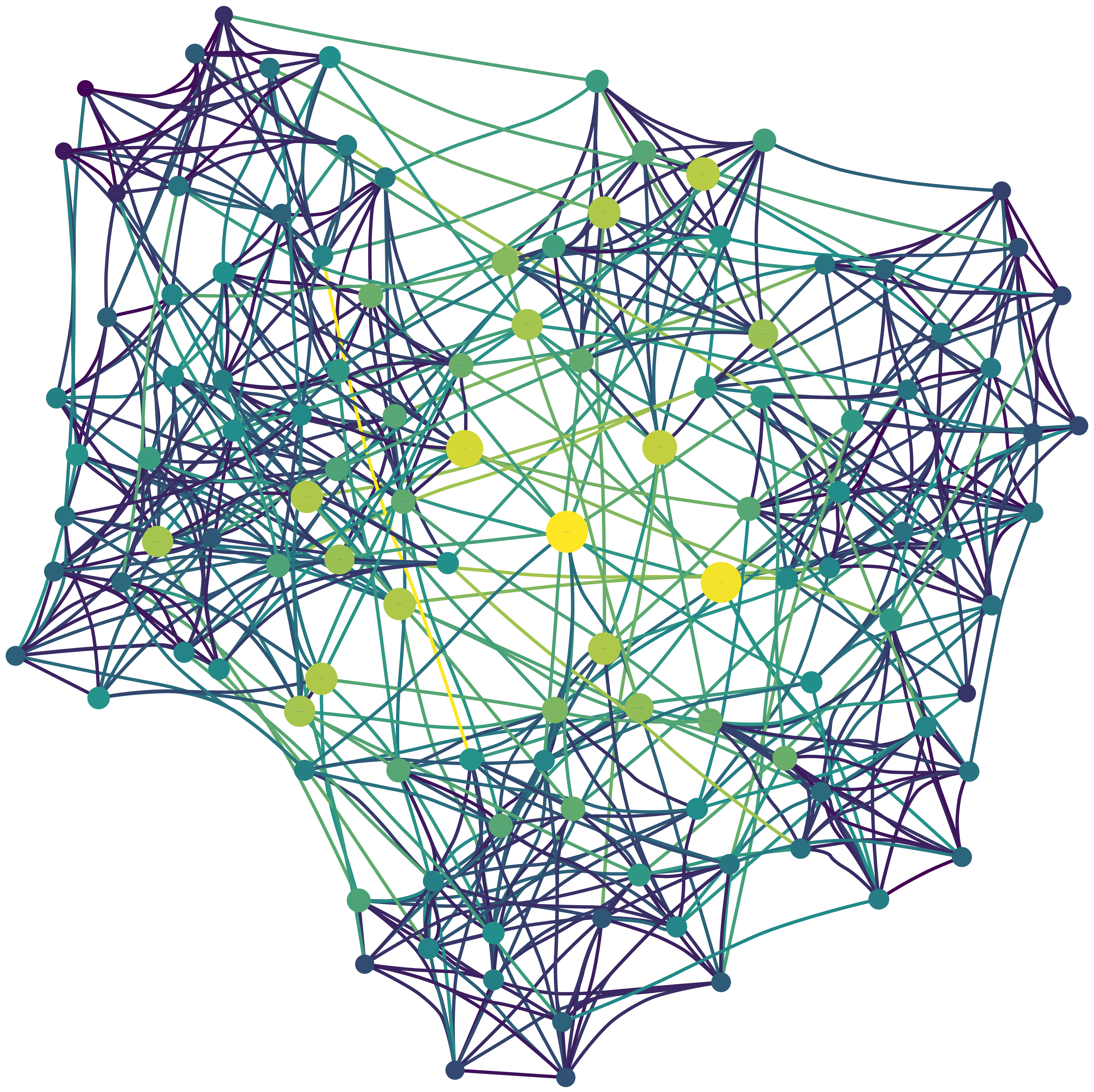


Bagrow & Mitchell (2018)
Bagrow *et al* (2019)

Rough Outline

- Basics
  - file formats, code, databases
- Networks from data
  - common tasks and good practices
- Case studies and examples
- Machine learning for data and networks
- Visualization (*time permitting*)

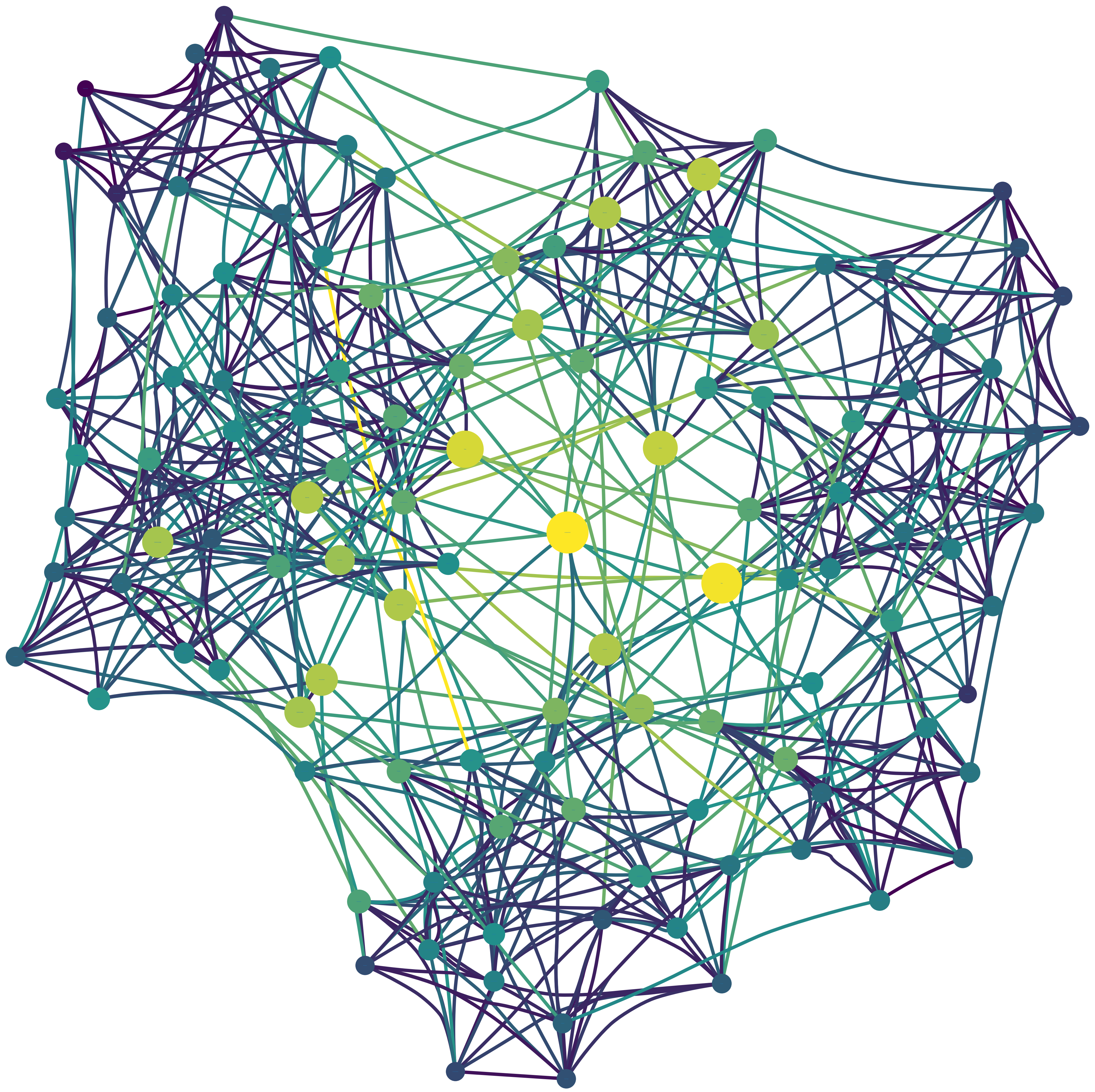
Network data are simple



- Looks like a complicated object
- Lots of measures, metrics, and algorithms to quantify and understand it

- But from a data perspective, very little to implement

Network data are simple



Store graph topology → need to define the nodes (vertices) and the links (edges):

$G = (V, E)$, $|V| = N$, $|E| = M$

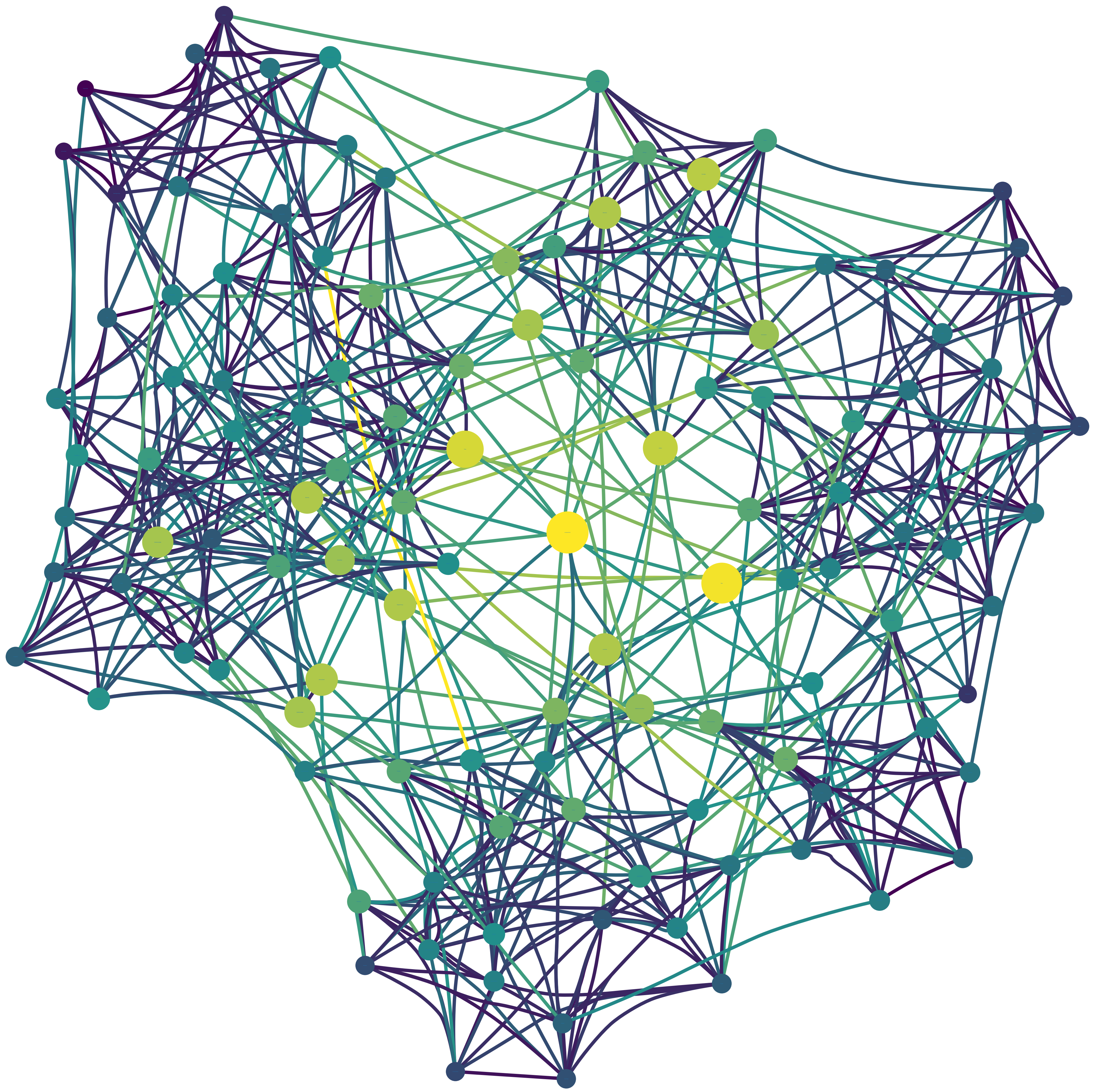
Edgelist:

M x 2 matrix

Alice	Bob
Bob	Carol
Bob	Dani
⋮	⋮

Need identifiers for nodes and *two delimiter symbols*

Network data are simple



Store graph topology → need to define the nodes (vertices) and the links (edges):

$G = (V, E)$, $|V| = N$, $|E| = M$

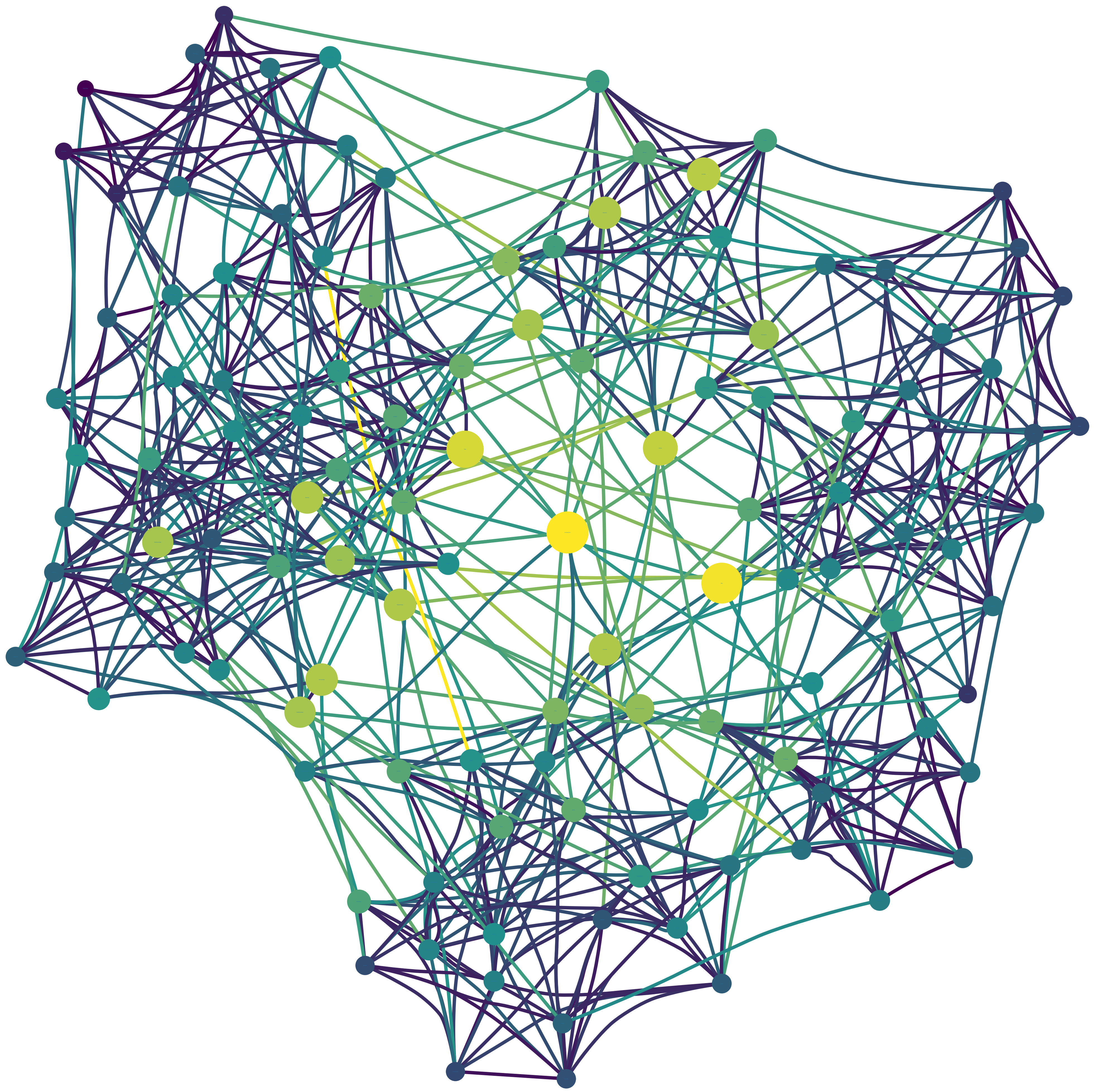
Adjacency list:

(ragged)

Alice	Bob		
Bob	Carol	Dani	
Carol	Bob	Erik	Fan
⋮			

May be harder to process in some programming languages

Network data are simple



Store graph topology → need to define the nodes (vertices) and the links (edges):

$G = (V, E)$, $|V| = N$, $|E| = M$

Adjacency Matrix:

$$\begin{matrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{matrix}$$

Network data are simple

Store graph topology → need to define the nodes (vertices) and the links (edges):

$G = (V, E)$, $|V| = N$, $|E| = M$

GraphML

Complex but more flexible

```xml
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
     http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <graph id="G" edgedefault="undirected">
    <node id="n0"/>
    <node id="n1"/>
    <node id="n2"/>
    <node id="n3"/>
    <node id="n4"/>
    <node id="n5"/>
    <node id="n6"/>
    <node id="n7"/>
   <edge source="n0" target="n2"/>
    <edge source="n1" target="n2"/>
    <edge source="n2" target="n3"/>
    <edge source="n3" target="n5"/>
    <edge source="n3" target="n4"/>
    <edge source="n4" target="n6"/>
    <edge source="n6" target="n5"/>
    <edge source="n5" target="n7"/>
 </graph>
</graphml>
```

Data surrounding network

What about **extra attributes**?

$G = (V, E, X)$

X = **attributes,** node labels or colors, timestamps

Can also have *edge* attributes

Edgelist

Alice	Bob	e1
Bob	Carol	e2
Bob	Dani	e3
⋮	⋮	

attributes

Node attribute list

Alice	x11	x12
Bob	x21	x22
Carol	x31	x32
⋮	⋮	⋮

attributes

Data surrounding network

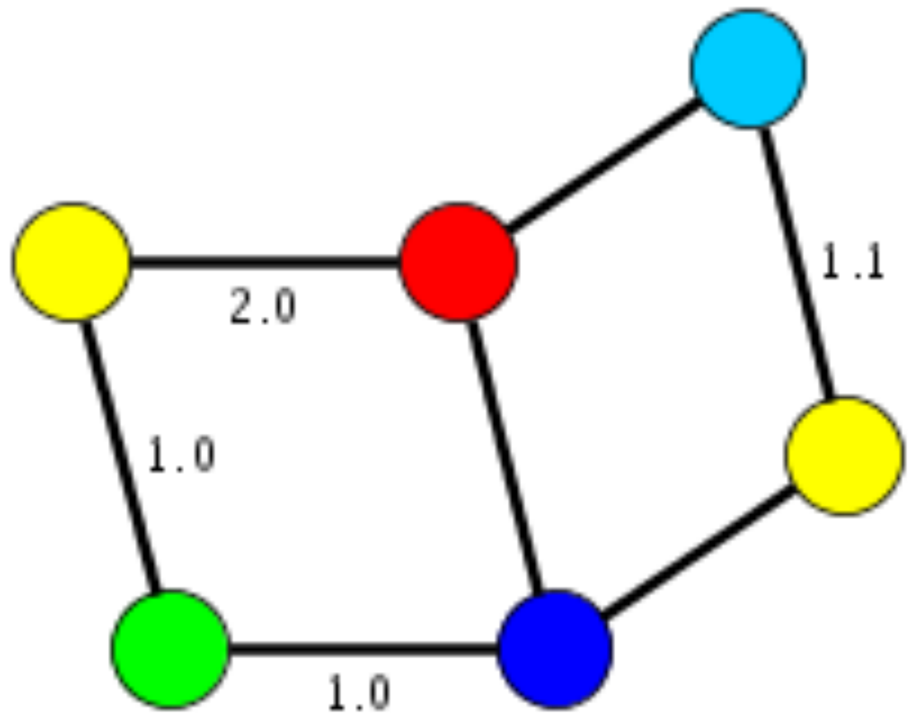
What about **extra attributes**?

$G = (V, E, X)$

X = **attributes**, node labels or colors, timestamps

Can also have *edge* attributes

GraphML

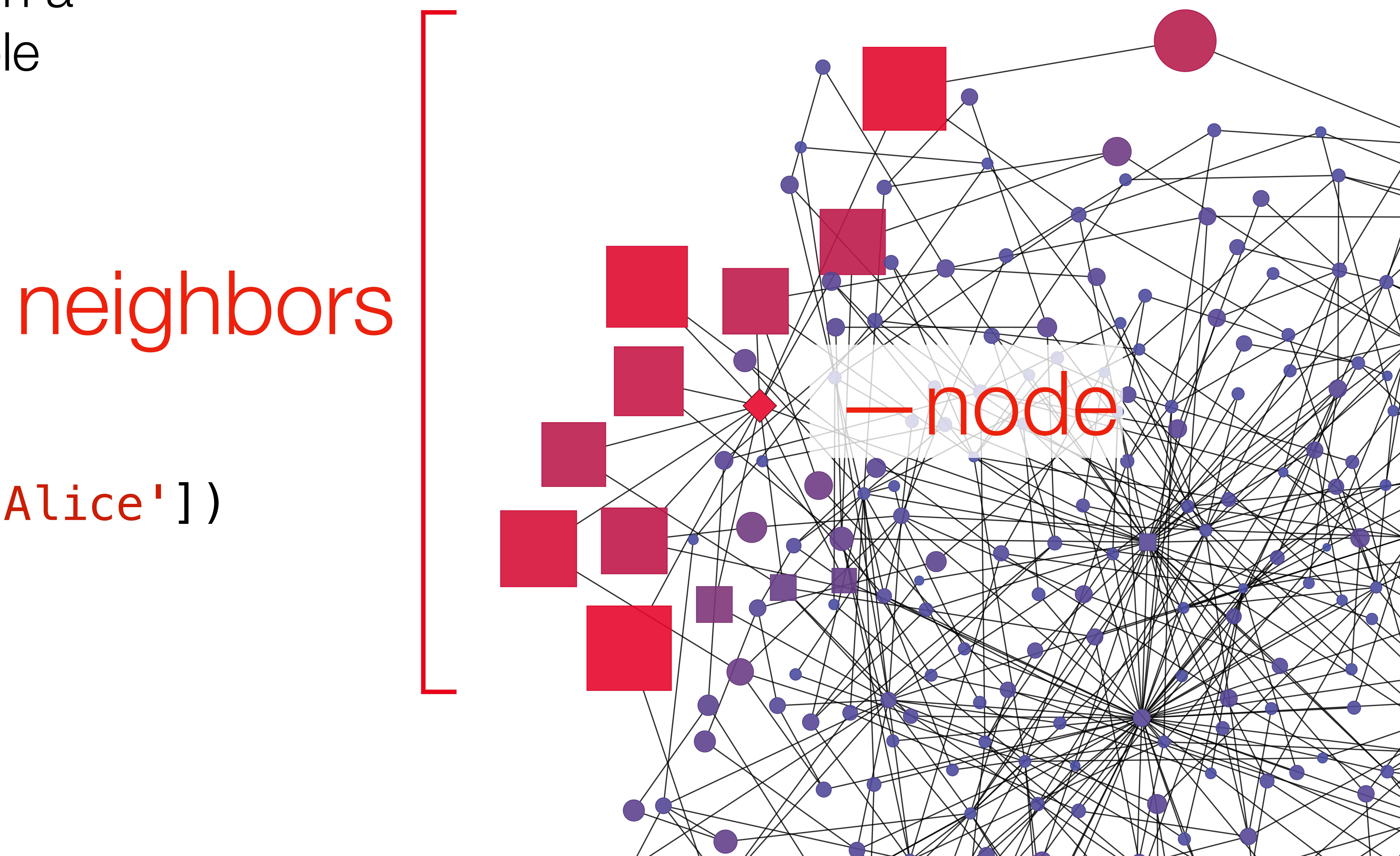



```xml
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
        http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <key id="d0" for="node" attr.name="color" attr.type="string">
    <default>yellow</default>
  </key>
  <key id="d1" for="edge" attr.name="weight" attr.type="double"/>
  <graph id="G" edgedefault="undirected">
    <node id="n0">
      <data key="d0">green</data>
    </node>
    <node id="n1"/>
    <node id="n2">
      <data key="d0">blue</data>
    </node>
    <node id="n3">
      <data key="d0">red</data>
    </node>
    <node id="n4"/>
    <node id="n5">
      <data key="d0">turquoise</data>
    </node>
    <edge id="e0" source="n0" target="n2">
      <data key="d1">1.0</data>
    </edge>
    <edge id="e1" source="n0" target="n1">
      <data key="d1">1.0</data>
```

Network data structures

To perform computations on a network, need a computable representation

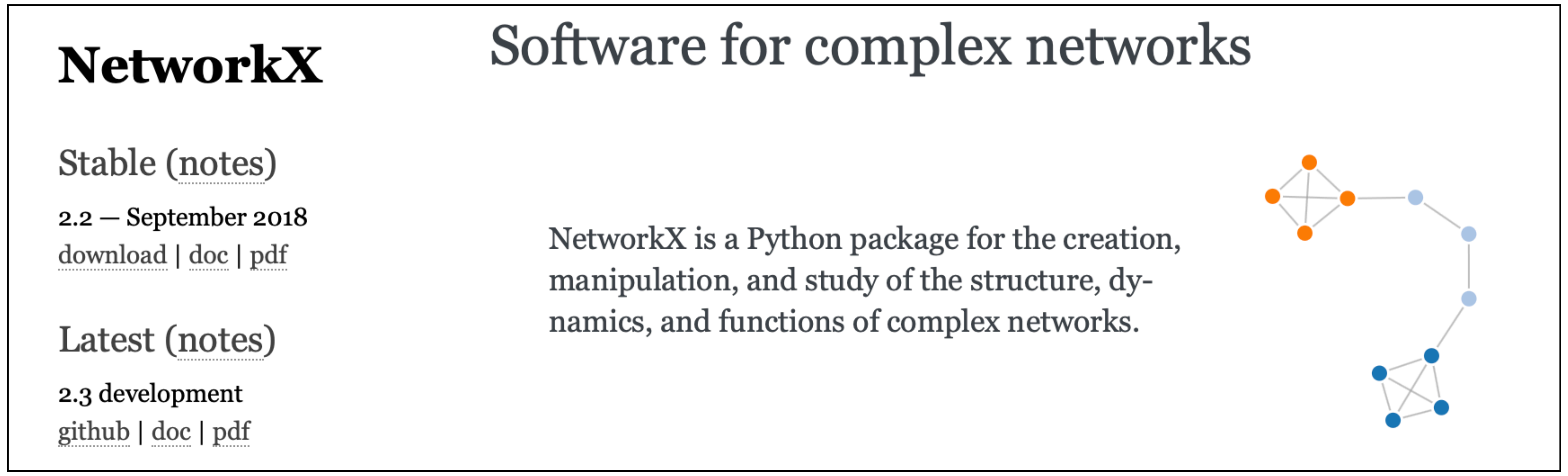
```
node2neighbors = …

print(node2neighbors['Alice'])
{'Bob','Carol'}
```

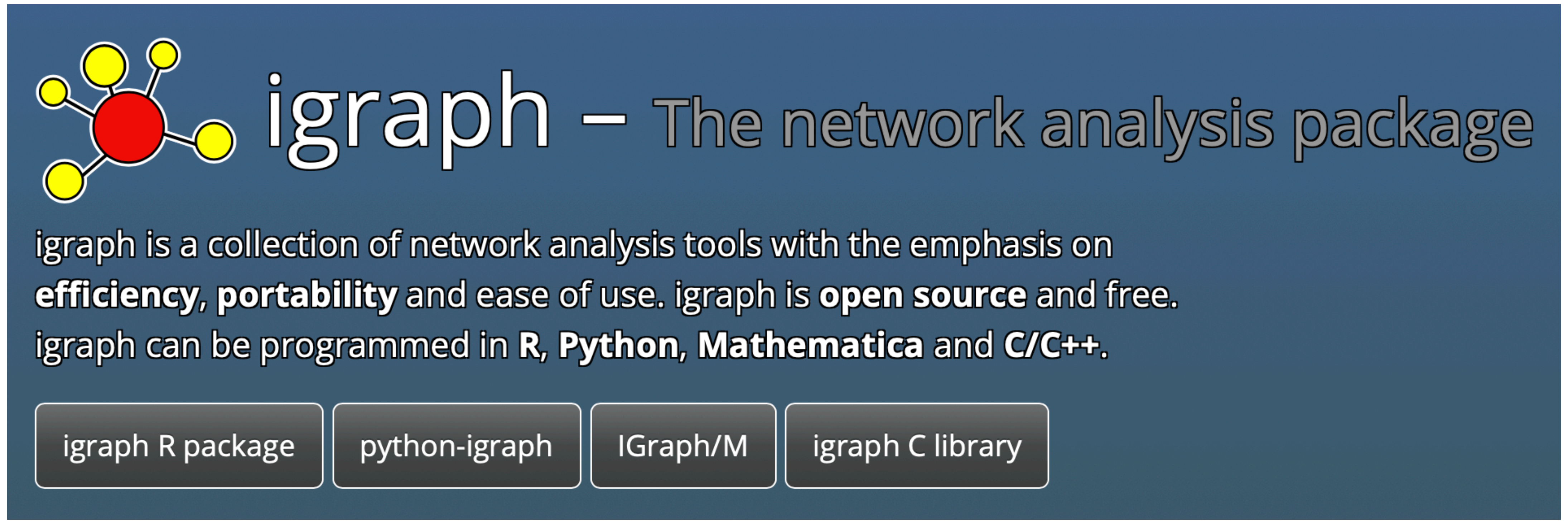


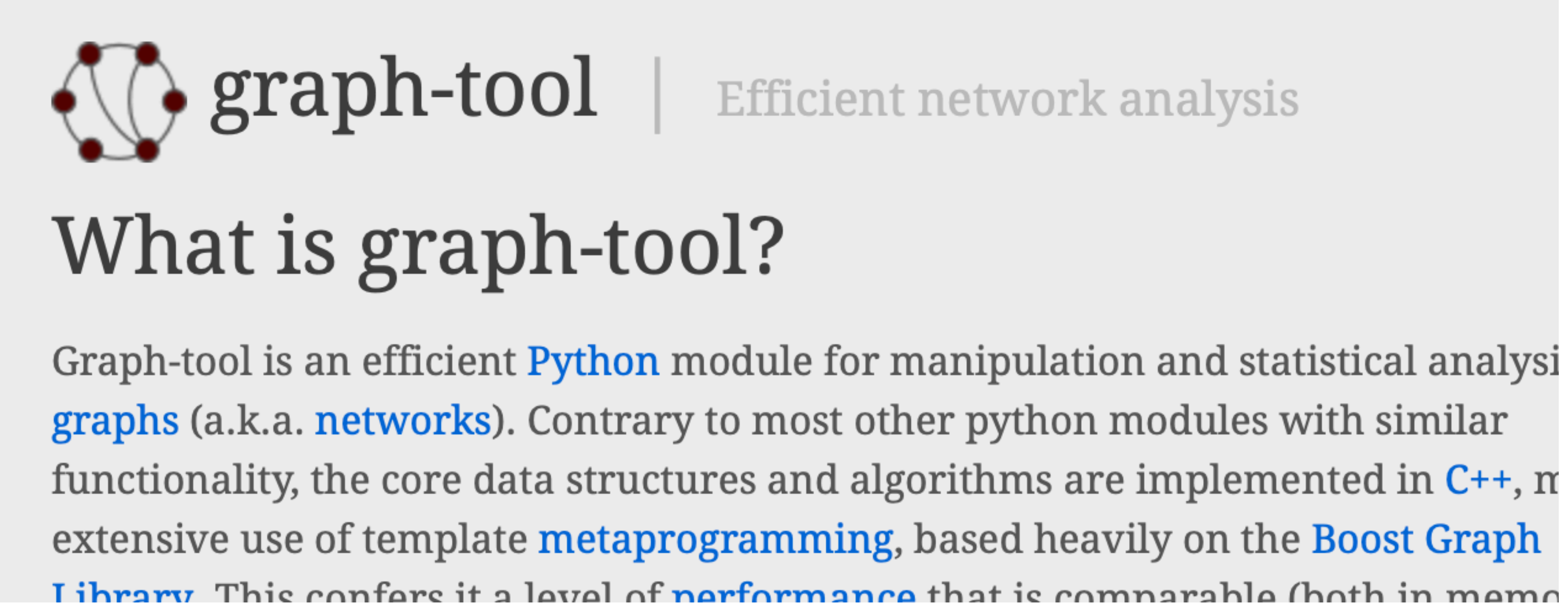
Network libraries

It's a good exercise to build your own data structures or even library, but in practice: lots of existing libraries

https://networkx.github.io
https://igraph.org
https://graph-tool.skewed.de







MATLAB

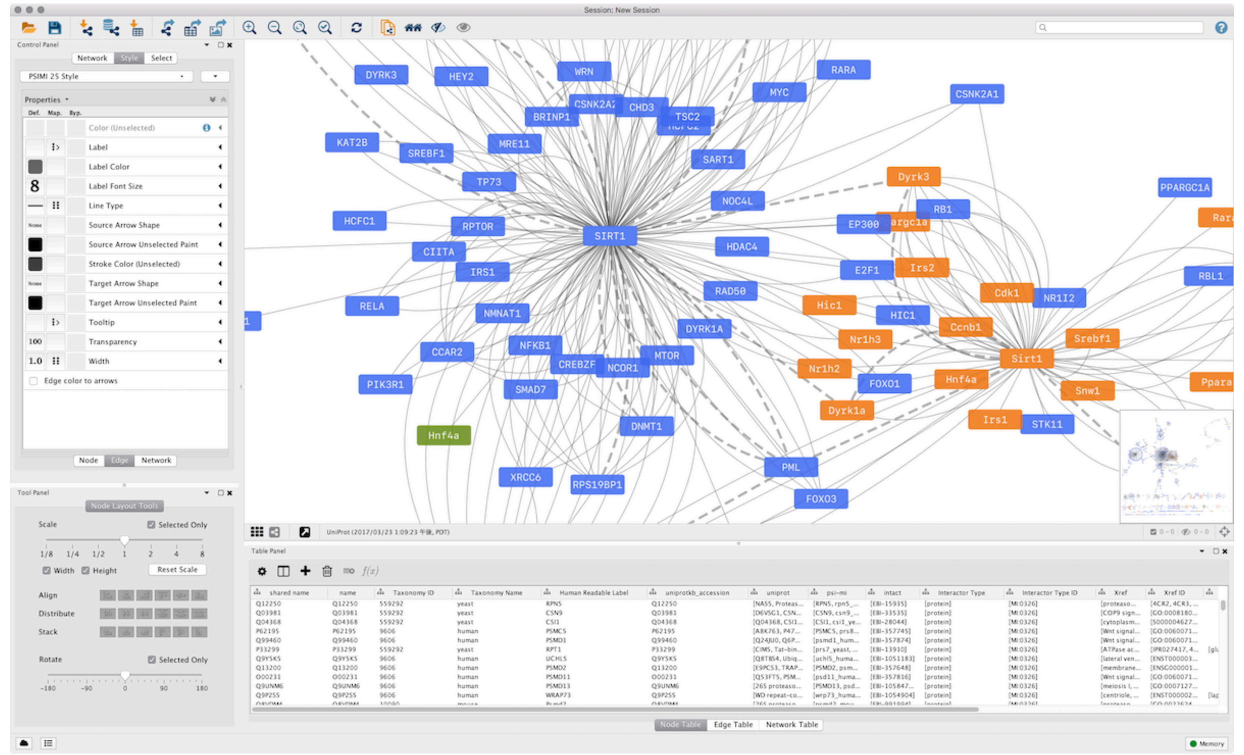
recent versions have graph algorithms (+ always have adjacency matrix)

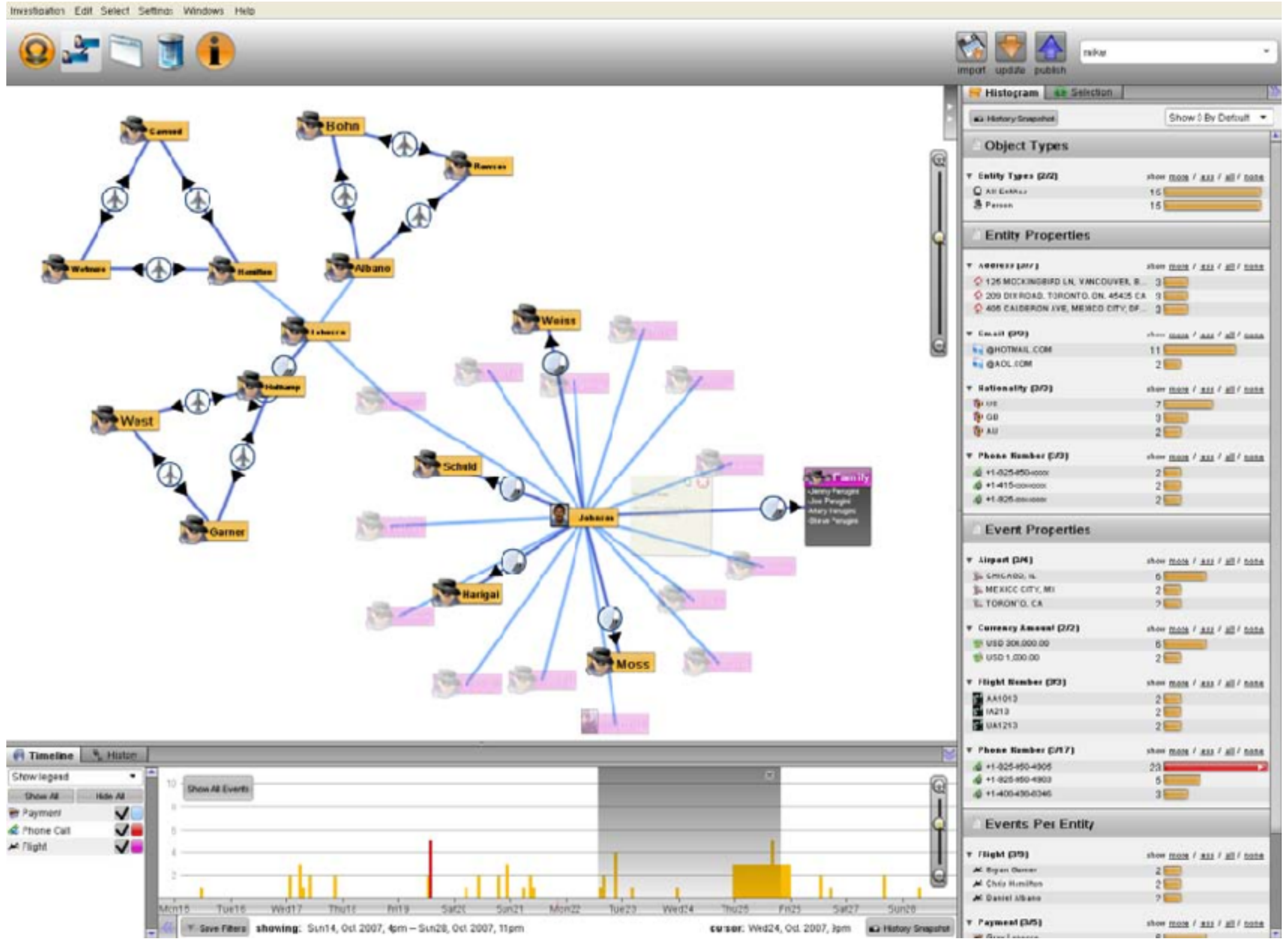
Graphical Interfaces and dashboards

I prefer to handle networks computationally, writing and running code—expressive, provenance

Interactive interfaces easier to get started but then you max out quickly!

Can be good for visualizations





Palantir

Graph databases—Big Data

neo4j

GraphDB

GraphQL

APACHE GIRAPH

SPARQL

Jena

(semantic web)

APACHE Spark GraphX

databases: relational
key-value
document
graph
⋮

https://neo4j.com
https://jena.apache.org/
http://graphdb.ontotext.com
https://graphql.org

Graph databases—Big Data

Applications of Graph DBs:

Knowledge graphs — semantic web
Fraud detection — real time
Recommendations (Netflix, Amazon)

Graph DBs best for real-time, high-volume, *local* operations

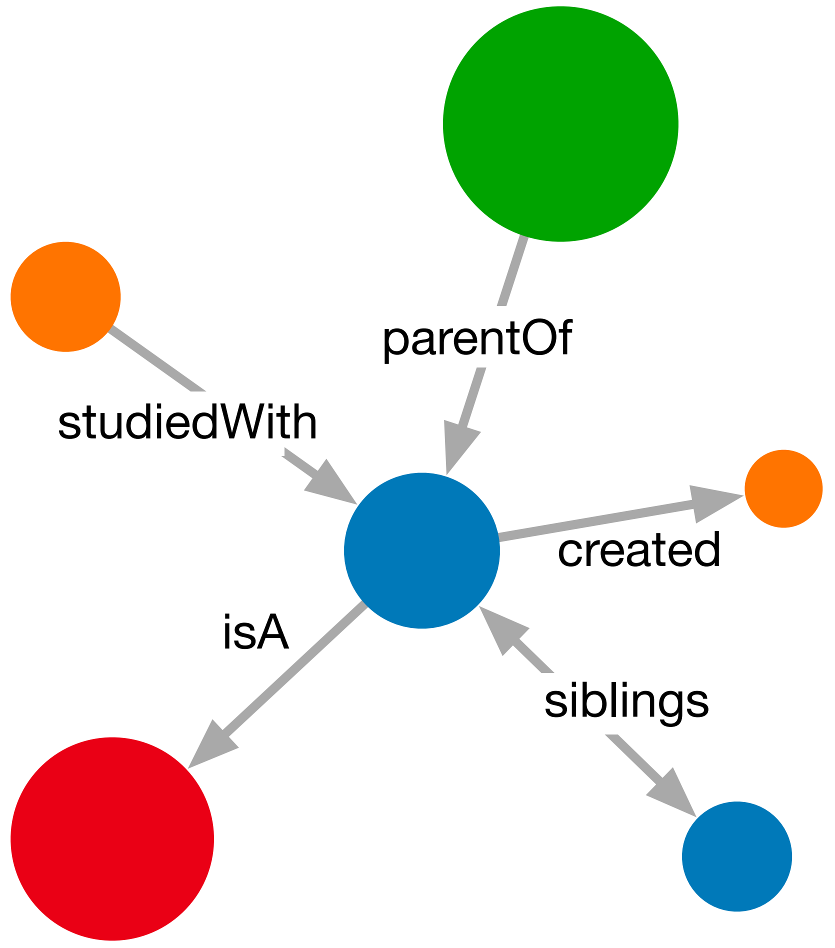
Graph databases—Big Data

Applications of Graph DBs:

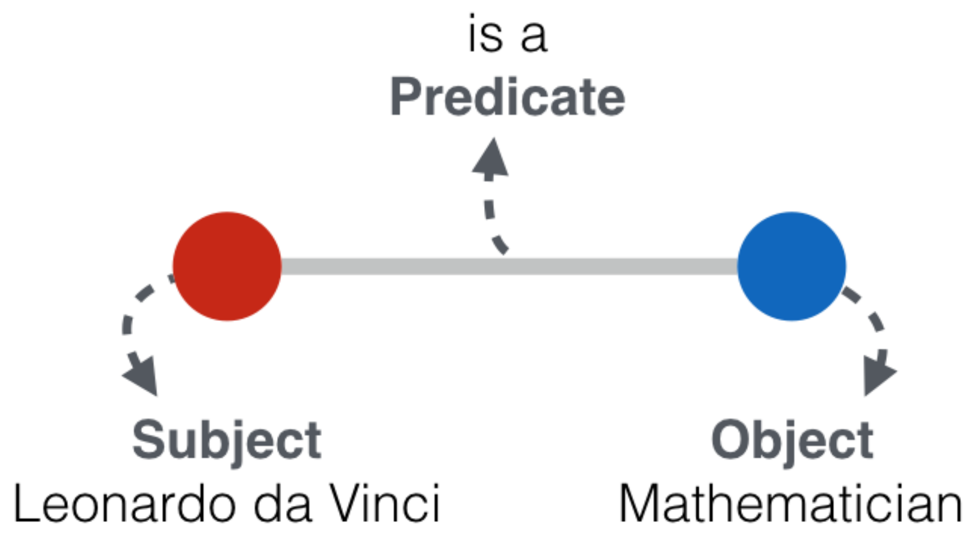
Knowledge graphs — semantic web
Fraud detection — real time
Recommendations (Netflix, Amazon)

Graph DBs best for real-time, high-volume, *local* operations

Knowledge Graph



Triplestore/RDF:



<**Leonardo da Vinci**> <is a> <**Mathematician**>
<**Diabetes Genes**> <encodes> <**Leptin**>
<**Visigoths**> <conquered> <**Ostrogoths**>
<**Barack Obama**> <born in> <**Hawaii**>
<**Harry Potter**> <is a> <**Fictional Character**>
<**Mount Everest**> <elevation> <**8,848 meters**>
<**Magic**> <is> <**Real**>

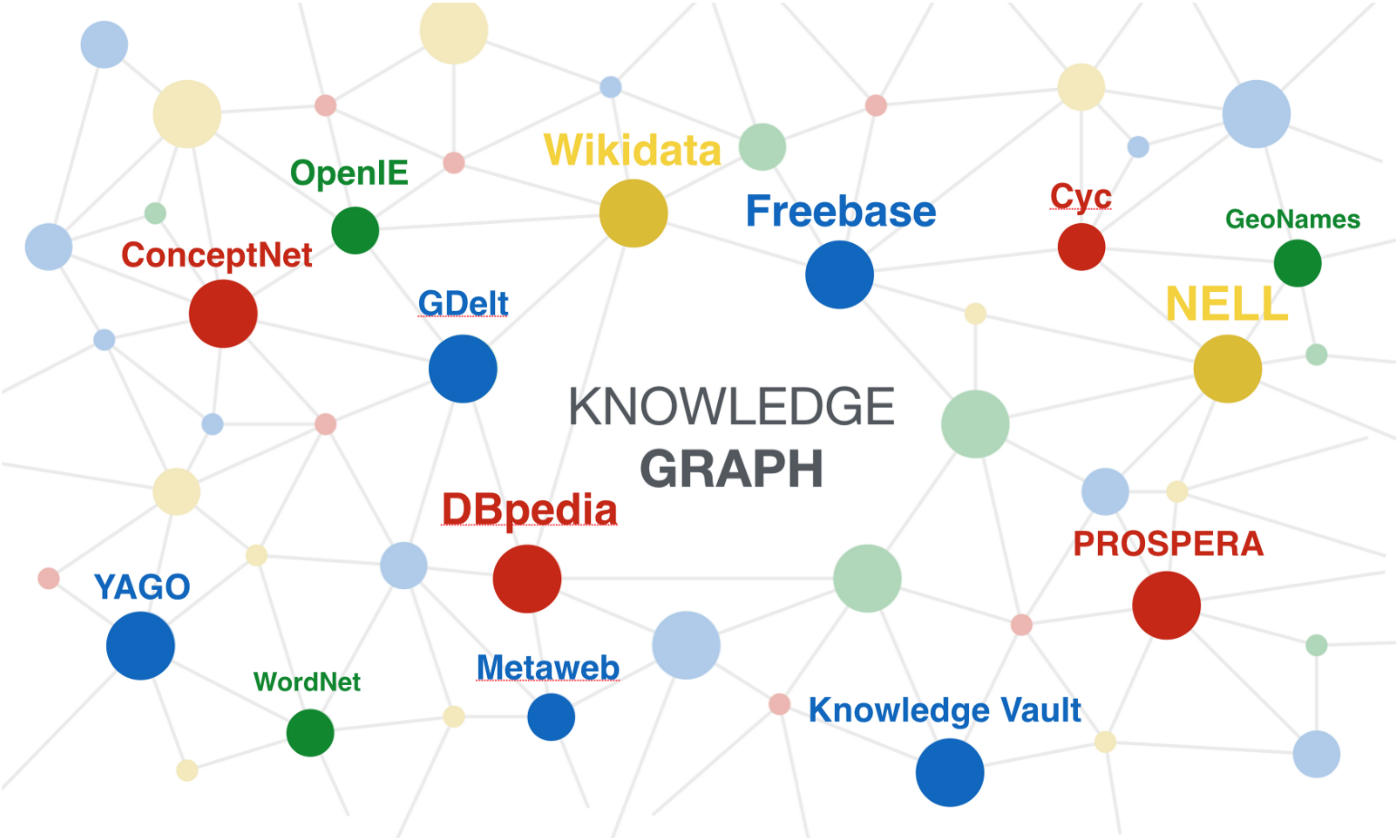
Courtesy: Sebastian Dery

Graph databases—Big Data

Some Knowledge Graphs

Dataset	Triples	Size
Wikidata (2018-09-11)	7.2B	28GB
DBPedia 2016-04 English	1B	13GB
DBLP 2017	882M	1GB
Freebase	2B	11GB
YAGO2s Knowledge Base	159M	903MB
WordNet 3.1	5.5M	23MB

Courtesy: rdfhdt.org

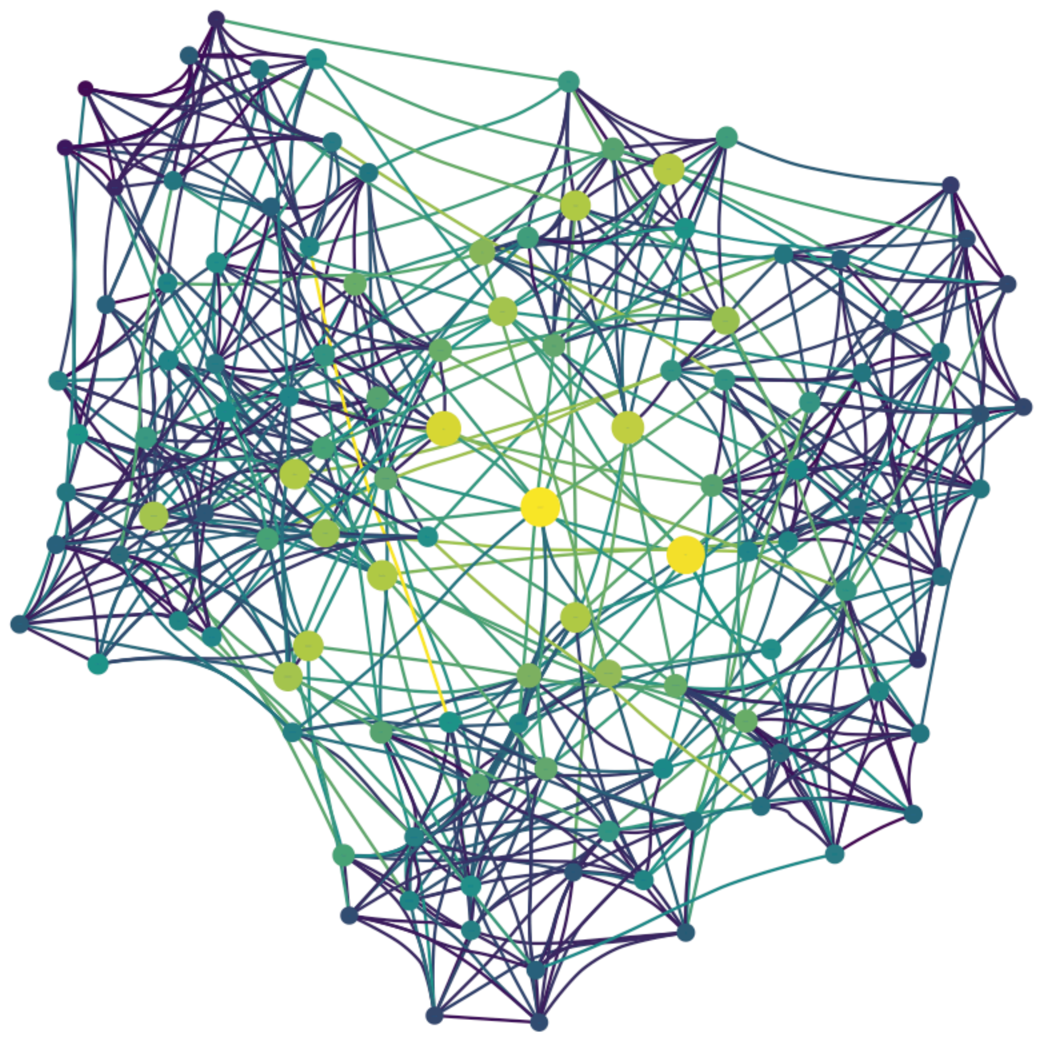


Courtesy: Sebastian Dery

Network data are not simple

There is an upstream task

Network data are simple



- Looks like a complicated object
- Lots of measures, metrics, and algorithms to quantify and understand it

- But from a data perspective, very little to implement

What defines your network?

Criteria for nodes?
Criteria for links?

Only simple *after* addressing these questions (if you need to)

Example: social network from mobile phone data

OPEN ACCESS Freely available online PLoS one

Collective Response of Human Populations to Large-Scale Emergencies

James P. Bagrow[1,2*], Dashun Wang[1,2], Albert-László Barabási[1,2,3]

Link communities reveal multiscale complexity in networks

Yong-Yeol Ahn[1,2*], James P. Bagrow[1,2*] & Sune Lehmann[3,4*]

OPEN ACCESS Freely available online PLoS one

Mesoscopic Structure and Social Aspects of Human Mobility

James P. Bagrow[1,2*], Yu-Ru Lin[3,4]

Example: social network from mobile phone data

OPEN ACCESS Freely available online PLoS one

Collective Response of Human Populations to Large-Scale Emergencies

James P. Bagrow[1,2]*, Dashun Wang[1,2], Albert-László Barabási[1,2,3]

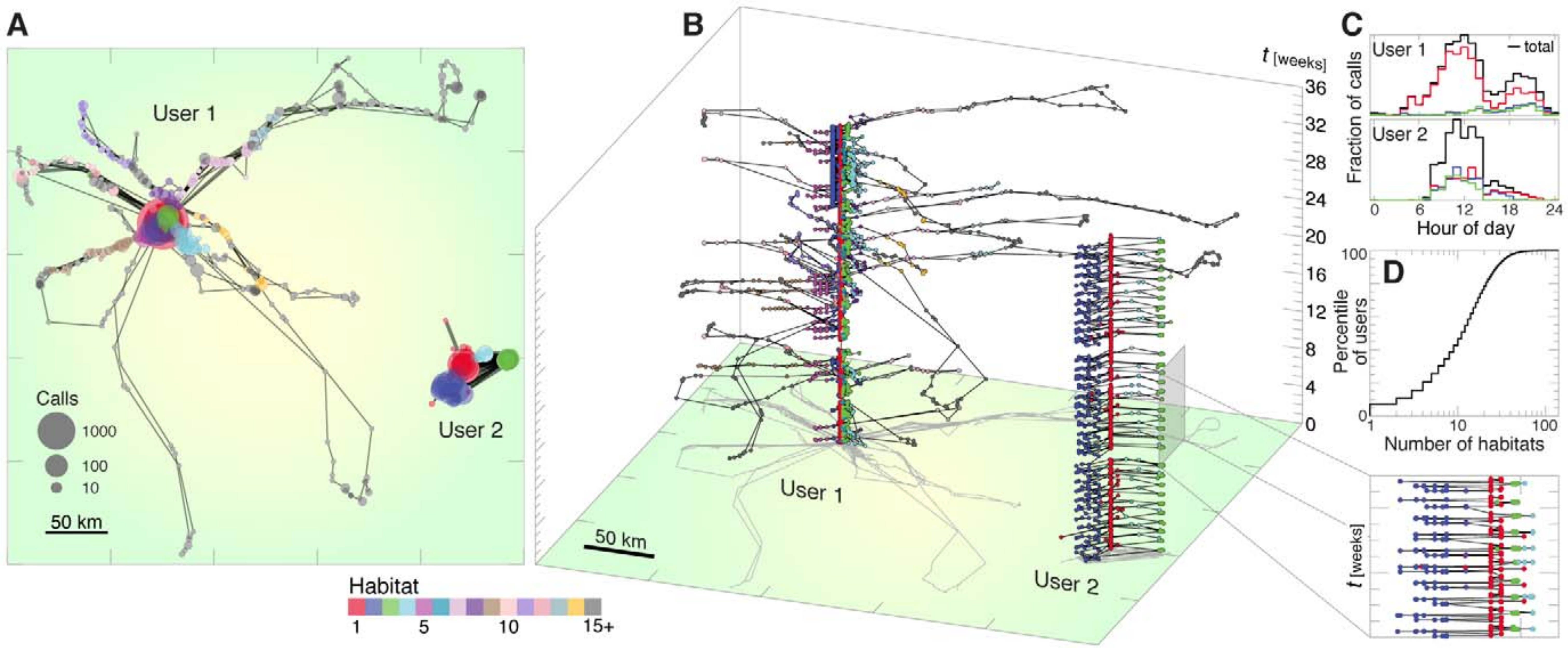
OPEN ACCESS Freely available online PLoS one

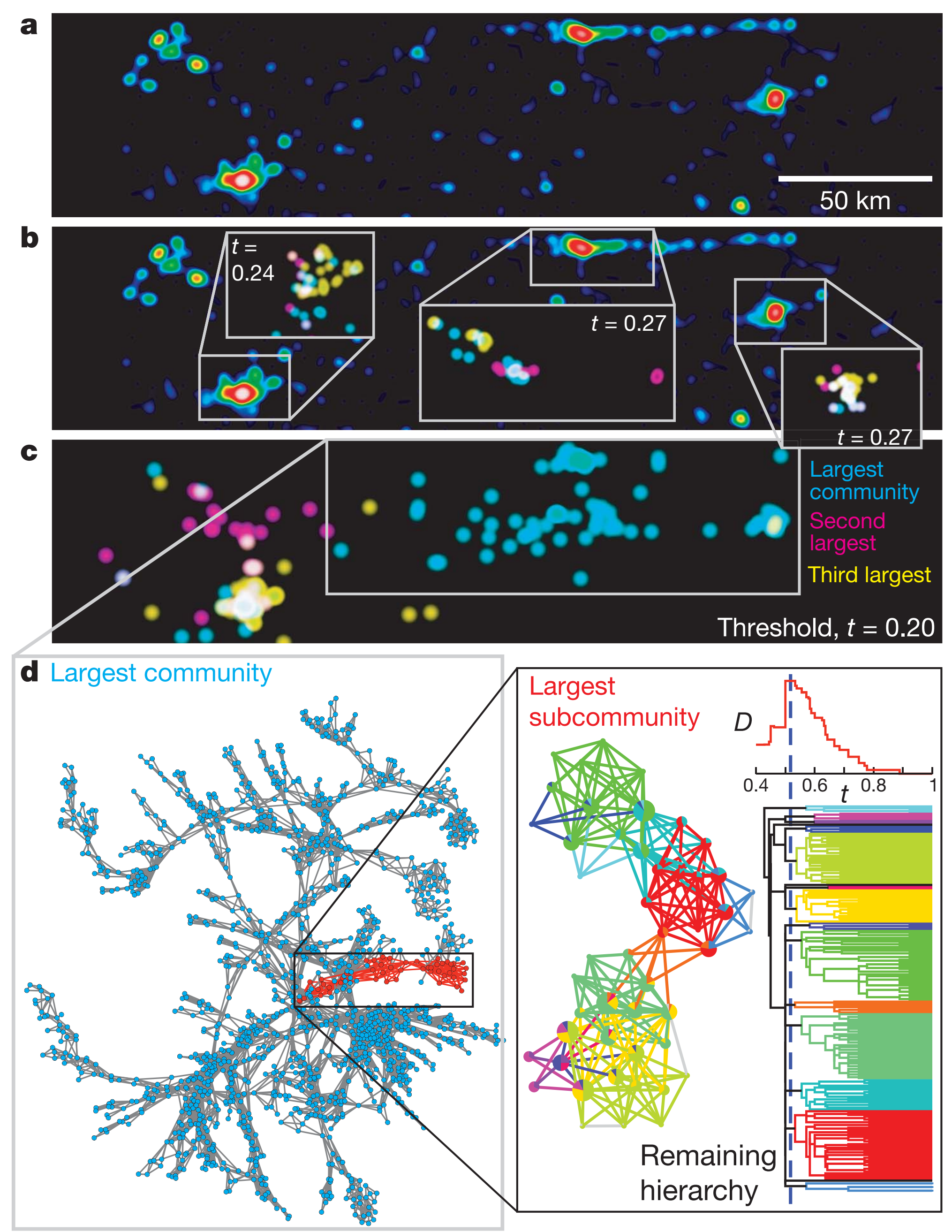
Mesoscopic Structure and Social Aspects of Human Mobility

James P. Bagrow[1,2]*, Yu-Ru Lin[3,4]

Link communities reveal multiscale complexity in networks

Yong-Yeol Ahn[1,2]*, James P. Bagrow[1,2]* & Sune Lehmann[3,4]*





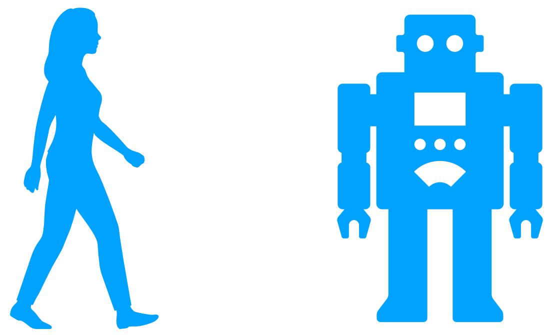
spatial social network

Example: social network from mobile phone data

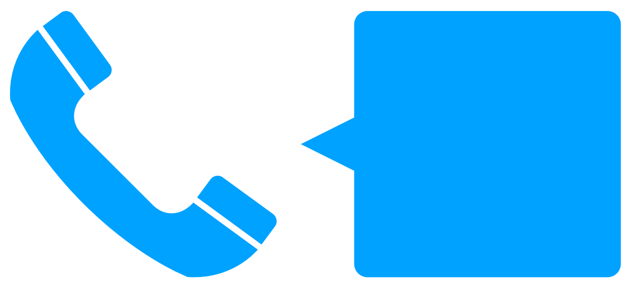
Extracted from deidentified Call Detail Record (CDR) files

What defines your network?

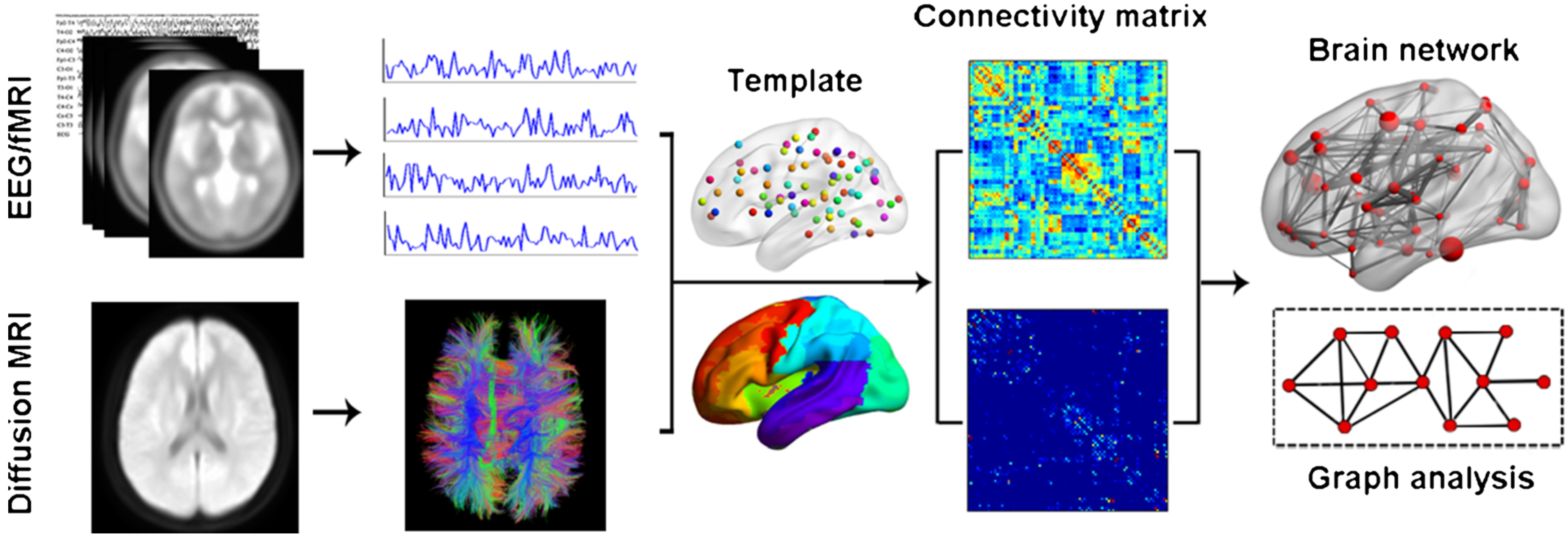
Criteria for nodes?



Criteria for links?

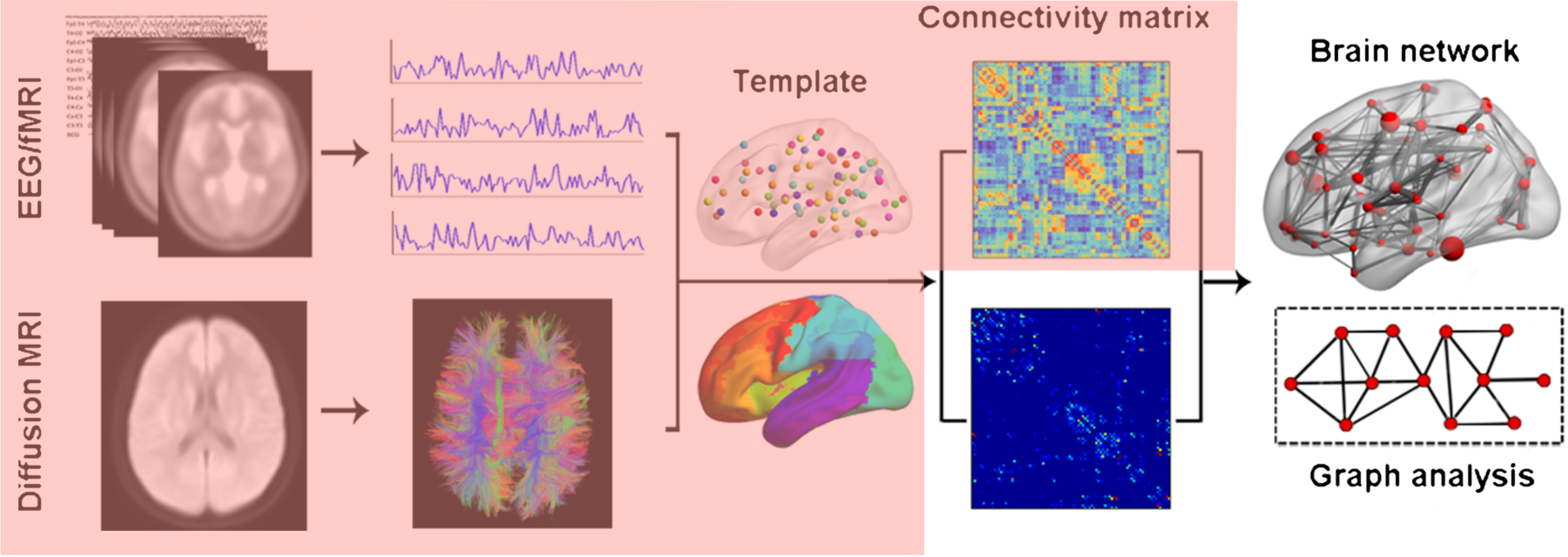


Example: brain networks



Cao, et al. *Mol. Neurobiol.* (2014)

Example: brain networks

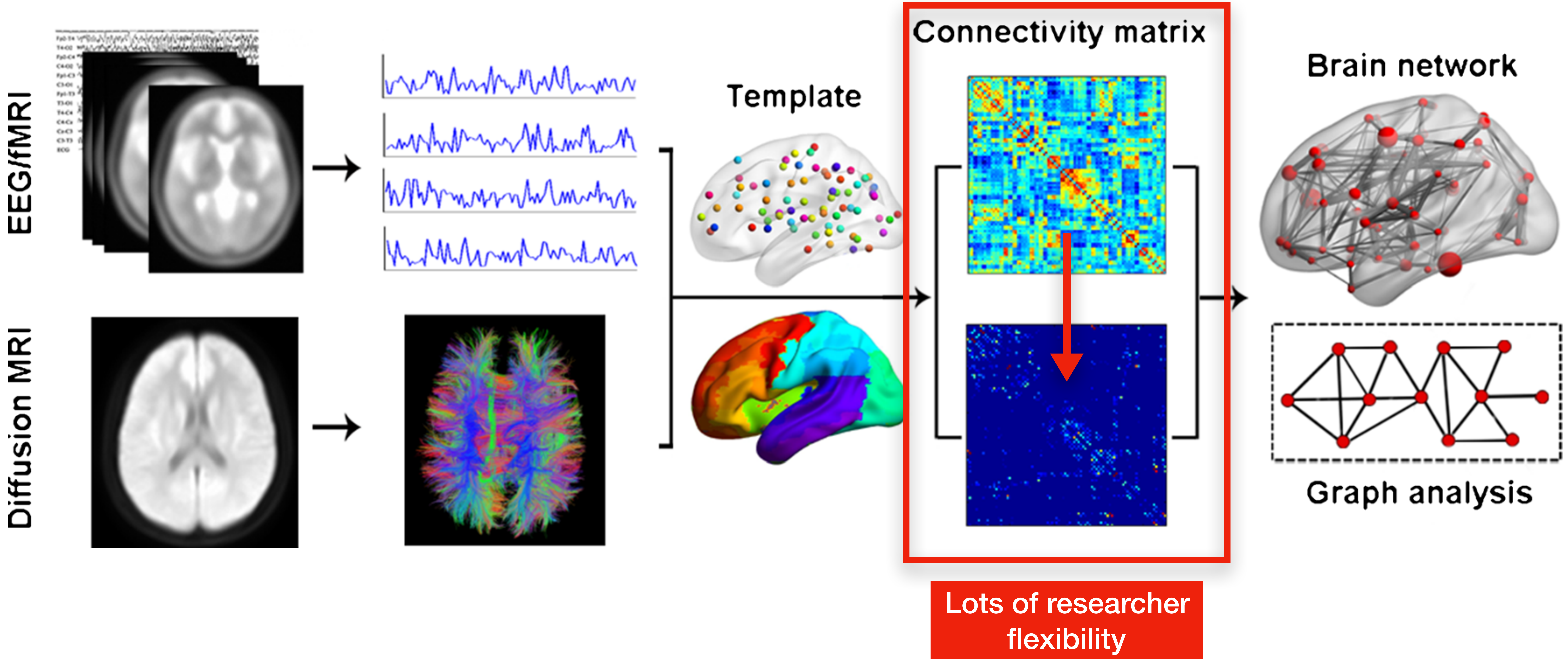


Upstream

Network

Cao, et al. *Mol. Neurobiol.* (2014)

Example: brain networks



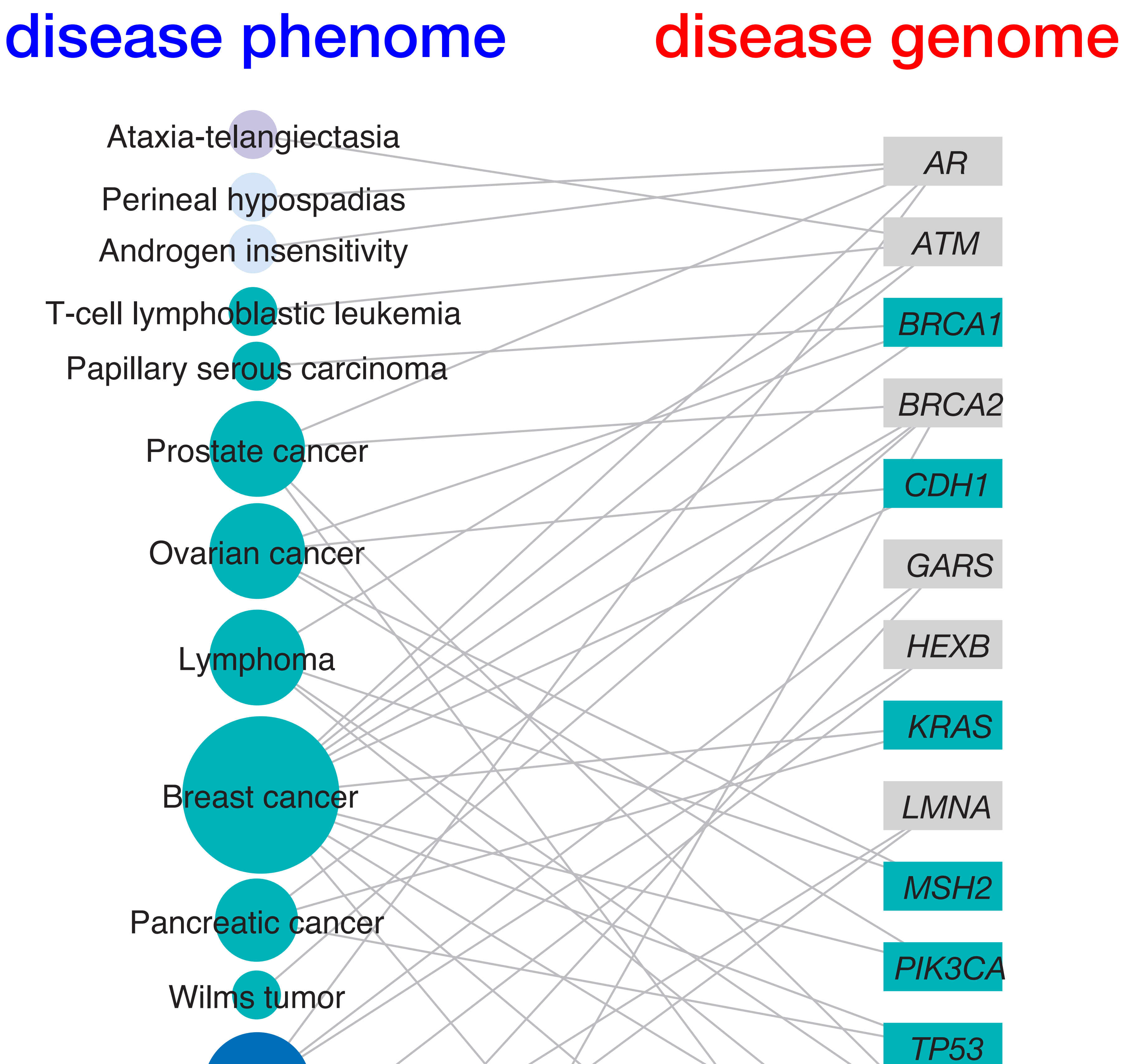
Cao, et al. *Mol. Neurobiol.* (2014)

There is an upstream task

What's the best network (there may be more than one)?

Define nodes
Define edges (hyper-edges?)
Directed?
Weighted?
Use a bipartite representation or project down?

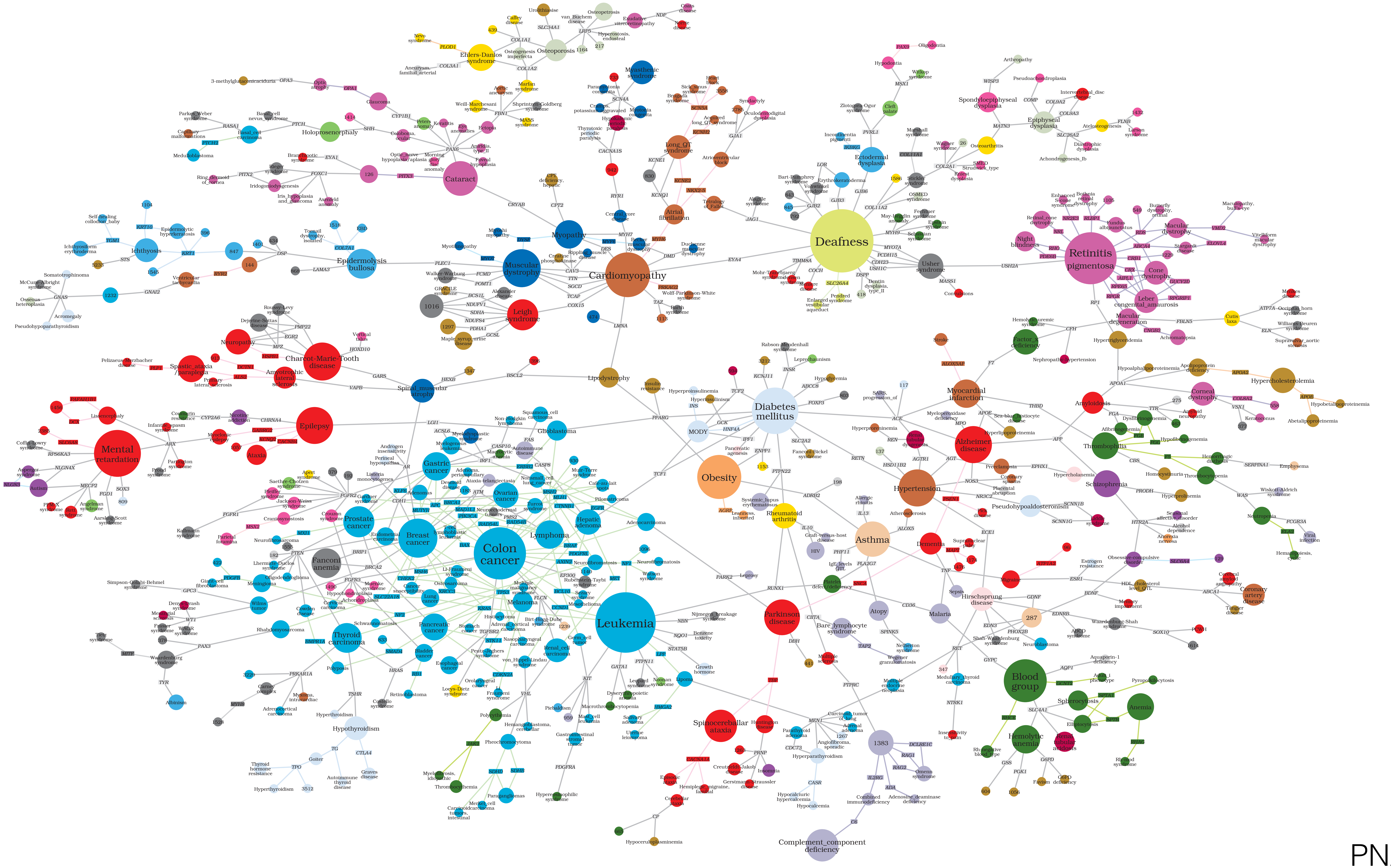
"Diseaseome"



Goh *et al.* PNAS (2007)

The human disease network

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) *Proc Natl Acad Sci USA* 104:8685-8690



Goh *et al.*
PNAS (2007)

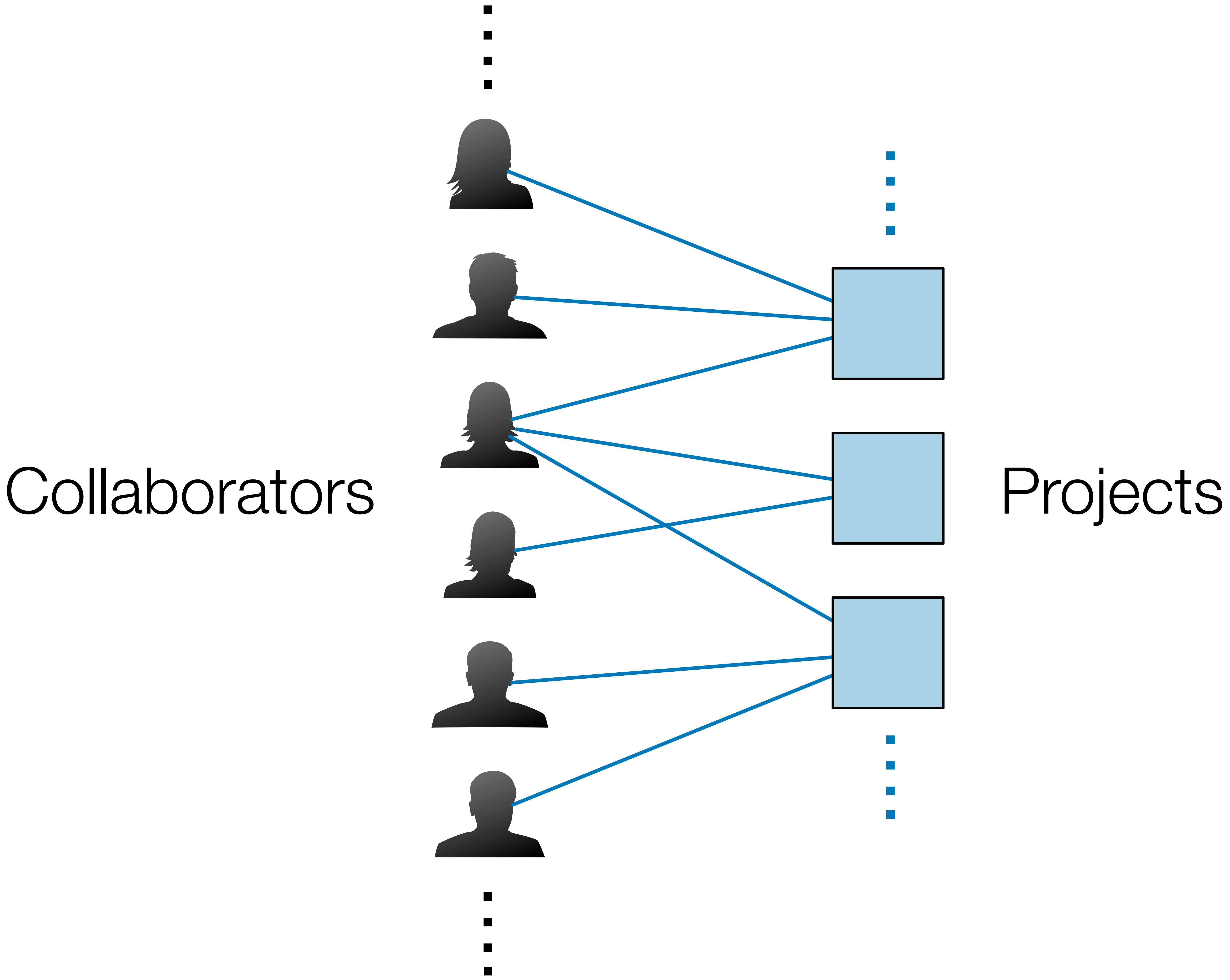
Do you have a bipartite network? Keep it that way?

Understanding the group dynamics and success of teams

Michael Klug[1] and James P. Bagrow[1,2,3]

teams of collaborators

GitHub



Klug and Bagrow (2016)

Building this network from the data

GitHub provides an API that lets you access the activities (events) of users as they make changes to code, join different teams, etc.

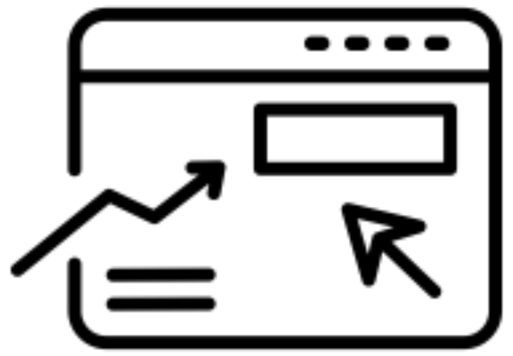
Building this network from the data

GitHub provides an API that lets you access the activities (events) of users as they make changes to code, join different teams, etc.

user

https://api.github.com/repos/bagrow/linkcomm/events

project



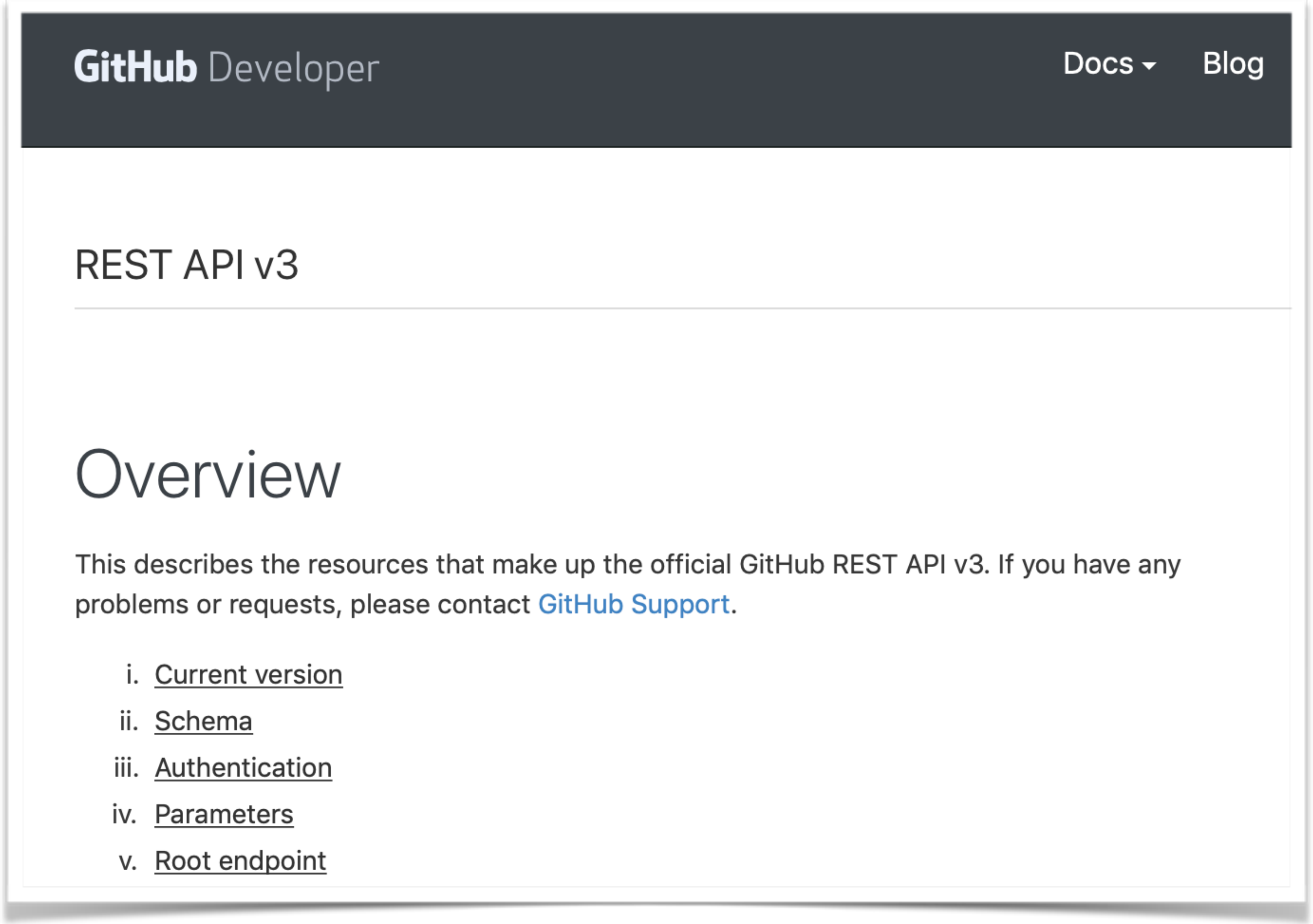
(actually, write a program)


```
⋮
{
   "id": "8401895651",
   "type": "PushEvent",
   "actor": {
     "login": "bagrow",
     "display_login": "bagrow",
     "gravatar_id": "",
     "url": "https://api.github.com/users/bagrow",
  },
   "repo": {
     "id": 904212,
     "name": "bagrow/linkcomm",
     "url": "https://api.github.com/repos/bagrow/linkcomm"
   },
   "payload": {
     "action": "started"
   },
   "public": true,
   "created_at": "2018-10-11T03:33:42Z"
}
⋮
```

Building this network from the data

GitHub provides an API that lets you access the activities (events) of users as they make changes to code, join different teams, etc.

JSON data


GitHub Developer Docs Blog

REST API v3

Overview

This describes the resources that make up the official GitHub REST API v3. If you have any problems or requests, please contact GitHub Support.

i. Current version
ii. Schema
iii. Authentication
iv. Parameters
v. Root endpoint

https://developer.github.com/v3/


```
⋮
{
    "id": "8401895651",
    "type": "PushEvent",
    "actor": {
      "login": "bagrow",
      "display_login": "bagrow",
      "gravatar_id": "",
      "url": "https://api.github.com/users/bagrow",
  },
    "repo": {
      "id": 904212,
      "name": "bagrow/linkcomm",
      "url": "https://api.github.com/repos/bagrow/linkcomm"
    },
    "payload": {
      "action": "started"
    },
    "public": true,
    "created_at": "2018-10-11T03:33:42Z"
}
⋮
```

Building this network from the data

GitHub provides an API that lets you access the activities (events) of users as they make changes to code, join different teams, etc.

Code updated

User

Project


```
⋮
{
  "id": "8401895651",
  "type": "PushEvent",
  "actor": {
    "login": "bagrow",
    "display_login": "bagrow",
    "gravatar_id": "",
    "url": "https://api.github.com/users/bagrow",
  },
  "repo": {
    "id": 904212,
    "name": "bagrow/linkcomm",
    "url": "https://api.github.com/repos/bagrow/linkcomm"
  },
  "payload": {
    "action": "started"
  },
  "public": true,
  "created_at": "2018-10-11T03:33:42Z"
}
⋮
```

Building this network from the data

GitHub provides an API that lets you access the activities (events) of users as they make changes to code, join different teams, etc.

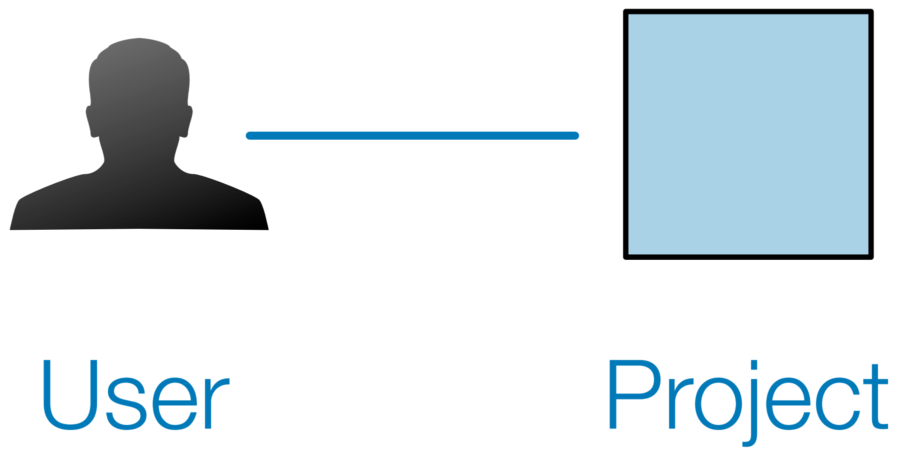
Code updated — "type"

User — "actor"

Project — "repo"

```
⋮
{
  "id": "8401895651",
  "type": "PushEvent",
  "actor": {
    "login": "bagrow",
    "display_login": "bagrow",
    "gravatar_id": "",
    "url": "https://api.github.com/users/bagrow",
  },
  "repo": {
    "id": 904212,
    "name": "bagrow/linkcomm",
    "url": "https://api.github.com/repos/bagrow/linkcomm"
  },
  "payload": {
    "action": "started"
  },
  "public": true,
  "created_at": "2018-10-11T03:33:42Z"
}
⋮
```

Insert link into bipartite graph



Building this network from the data

GitHub provides an API that lets you access the activities (events) of users as they make changes to code, join different teams, etc.

🤔 Are "PushEvents" meaningful?

🤔 Are node IDs?

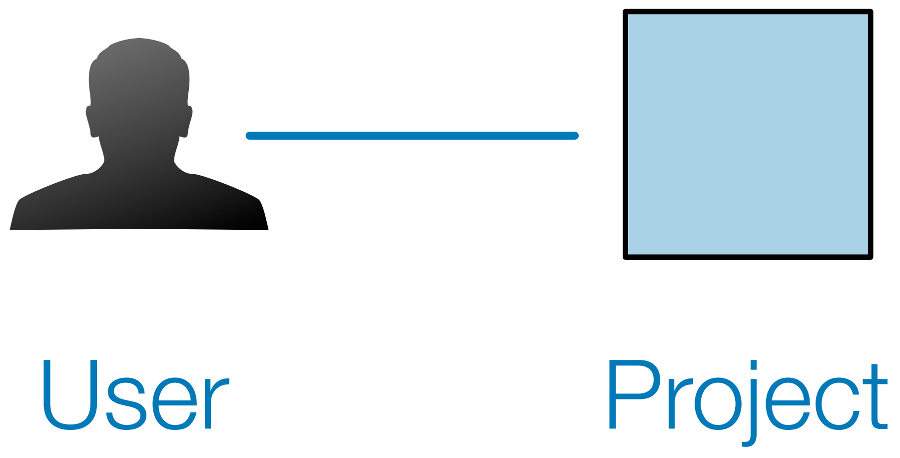
Code updated

User

Project

```
⋮
{
  "id": "8401895651",
  "type": "PushEvent",
  "actor": {
    "login": "bagrow",
    "display_login": "bagrow",
    "gravatar_id": "",
    "url": "https://api.github.com/users/bagrow",
  },
  "repo": {
    "id": 904212,
    "name": "bagrow/linkcomm",
    "url": "https://api.github.com/repos/bagrow/linkcomm"
  },
  "payload": {
    "action": "started"
  },
  "public": true,
  "created_at": "2018-10-11T03:33:42Z"
}
⋮
```

Insert link into bipartite graph



Building this network from the data

To build the entire network requires scraping their API:

- probably too slow
- API provider will probably block you

Solutions:

- Give up on getting the entire network and **work locally**; snowball sample?
- Find another source of data:

Building this network from the data

To build the entire network requires scraping their API:

- probably too slow
- API provider will probably block you

Solutions:

- Give up on getting the entire network and **work locally**; snowball sample?
- Find another source of data:

git GH **Archive**

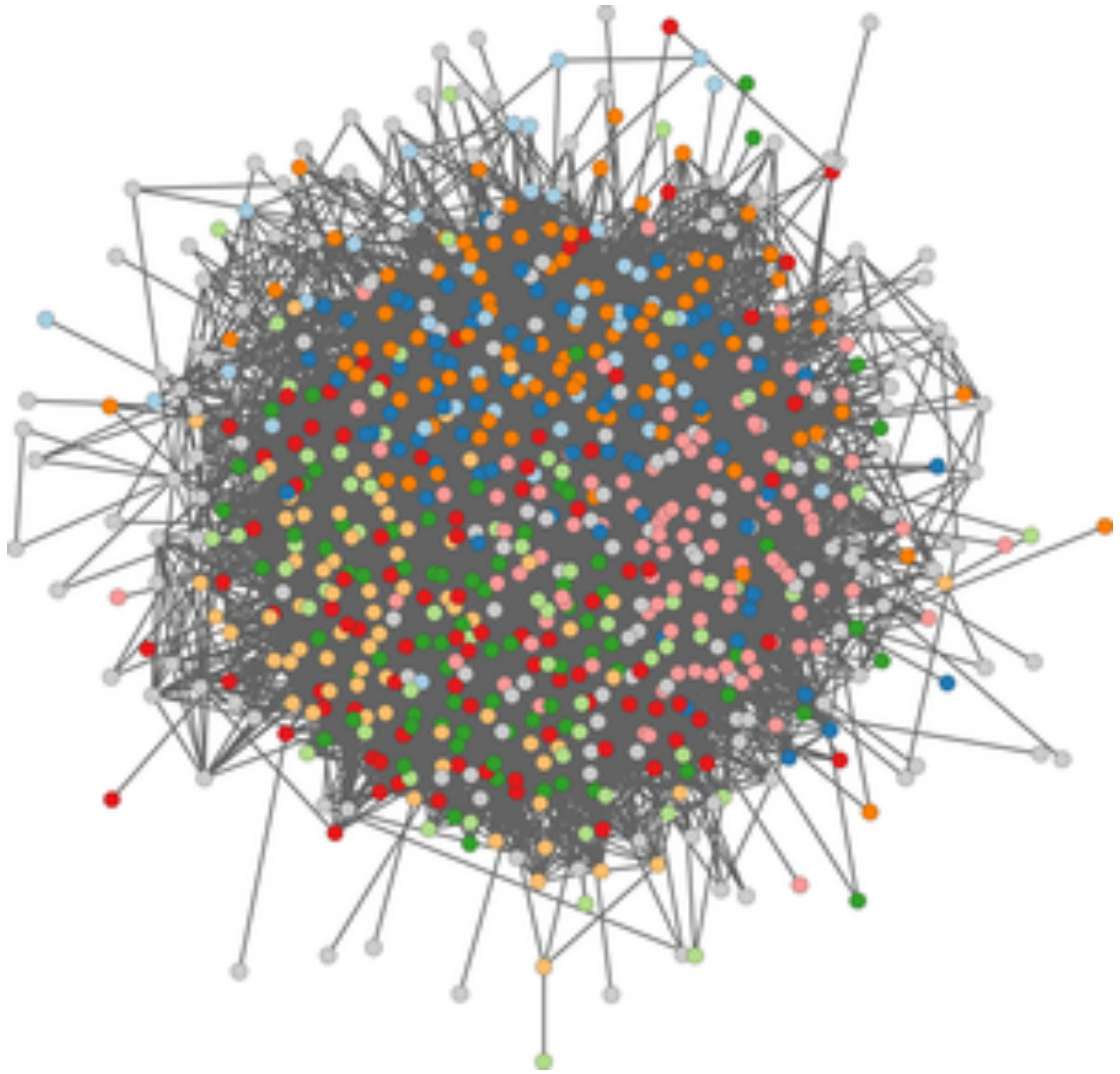
GitHub provides 20+ event types, which range from new commits and fork events, to opening new tickets, commenting, and adding members to a project. These events are aggregated into hourly archives, which you can access with any HTTP client:

Query	Command
Activity for 1/1/2015 @ 3PM UTC	`wget http://data.gharchive.org/2015-01-01-15.json.gz`
Activity for 1/1/2015	`wget http://data.gharchive.org/2015-01-01-{0..23}.json.gz`
Activity for all of January 2015	`wget http://data.gharchive.org/2015-01-{01..31}-{0..23}.json.gz`

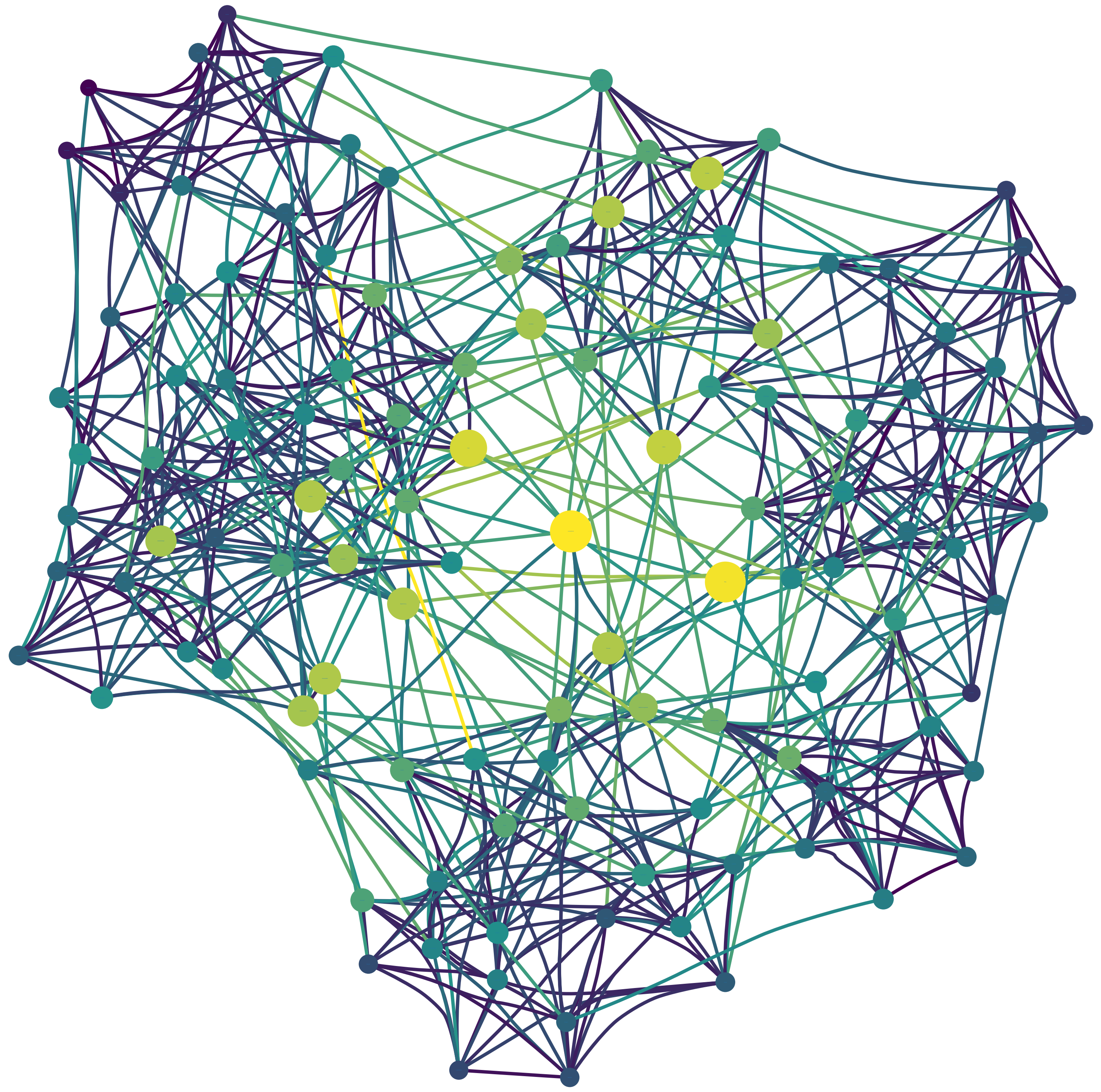
https://www.gharchive.org

Common task: thinning

Common task: thinning



vs.



Common task: thinning (subsetting)

Sometimes necessary to remove spurious links and/or nodes

Remove singleton nodes?

Remove nodes with degree < *k*
→ k-cores

Common task: thinning (subsetting)

Sometimes necessary to remove spurious links and/or nodes

Remove singleton nodes?

Remove nodes with degree < k
→ k-cores

⌛ Temporal network?

- Keep nodes/links of a certain age
- Consider a certain time window
- But how to pick? 🤔

Common task: thinning (subsetting)

Sometimes necessary to remove spurious links and/or nodes

Remove singleton nodes?

Remove nodes with degree < *k*
→ k-cores

⌛ Temporal network?

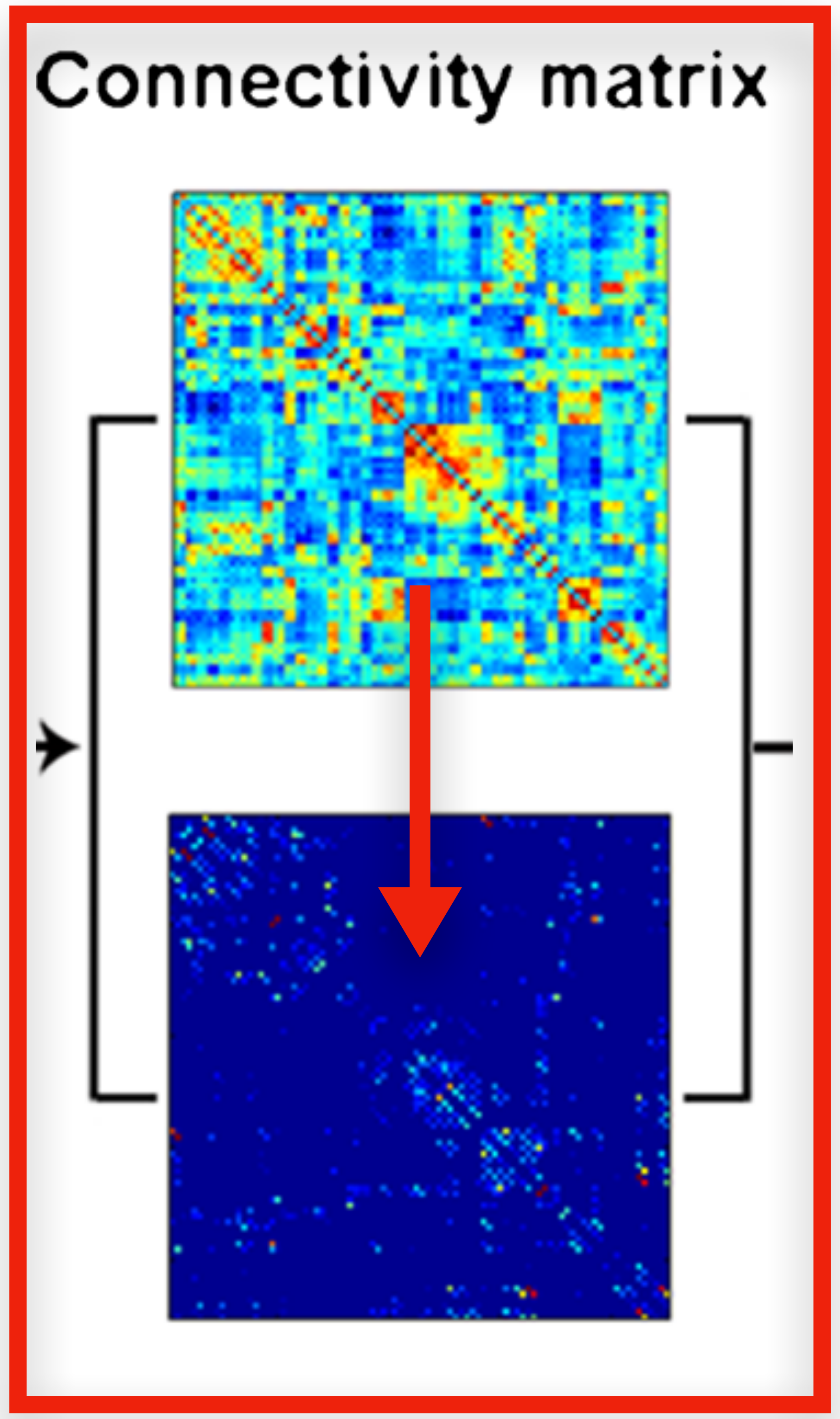
- Keep nodes/links of a certain age
- Consider a certain time window
- But how to pick? 🤔

Choices depend on problem area, type of data, and your scientific goals

Common task: thinning

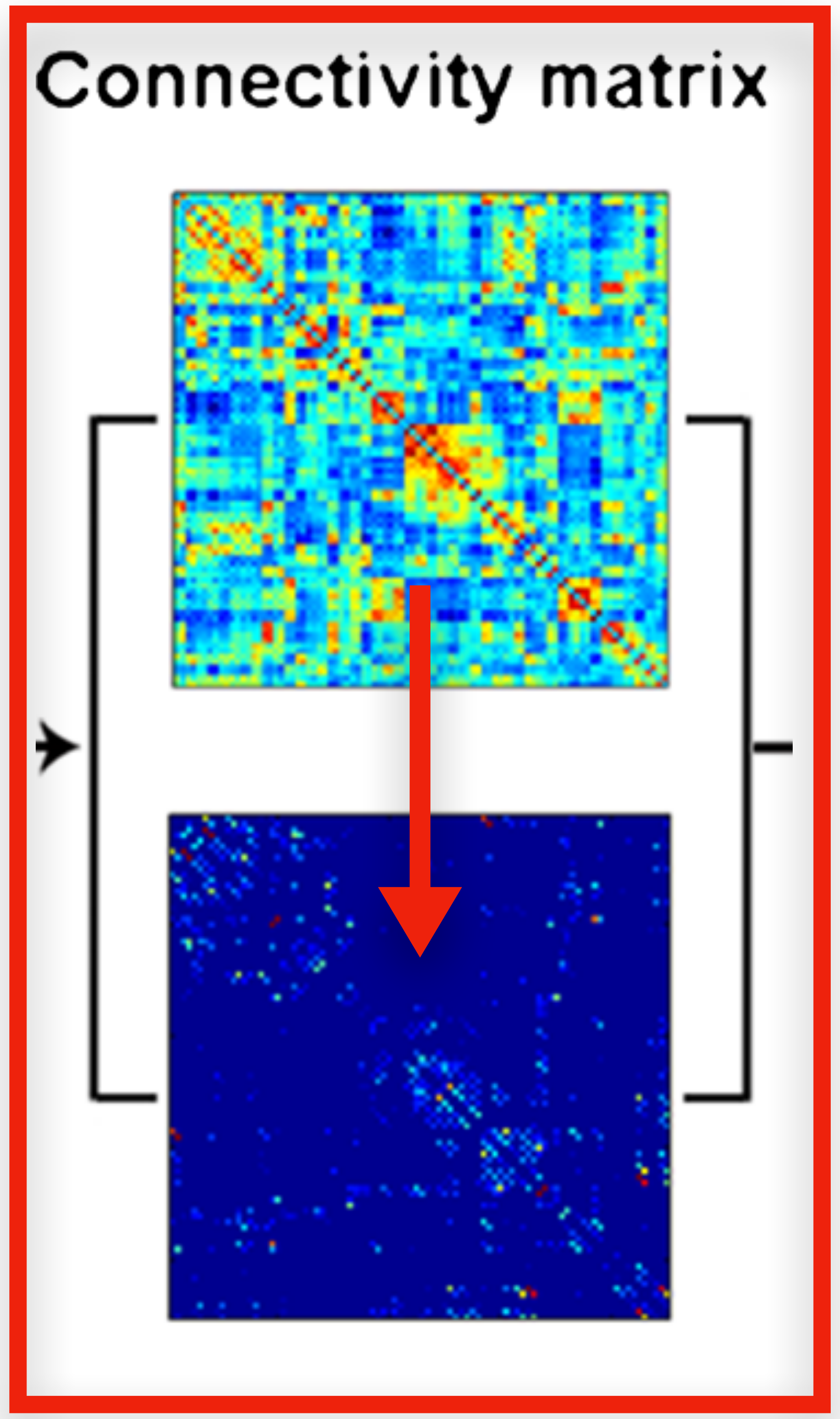
Network is very dense, lots of potentially spurious edges

How to sparsify?



Common task: thinning

Network is very dense, lots of potentially spurious edges
How to sparsify?



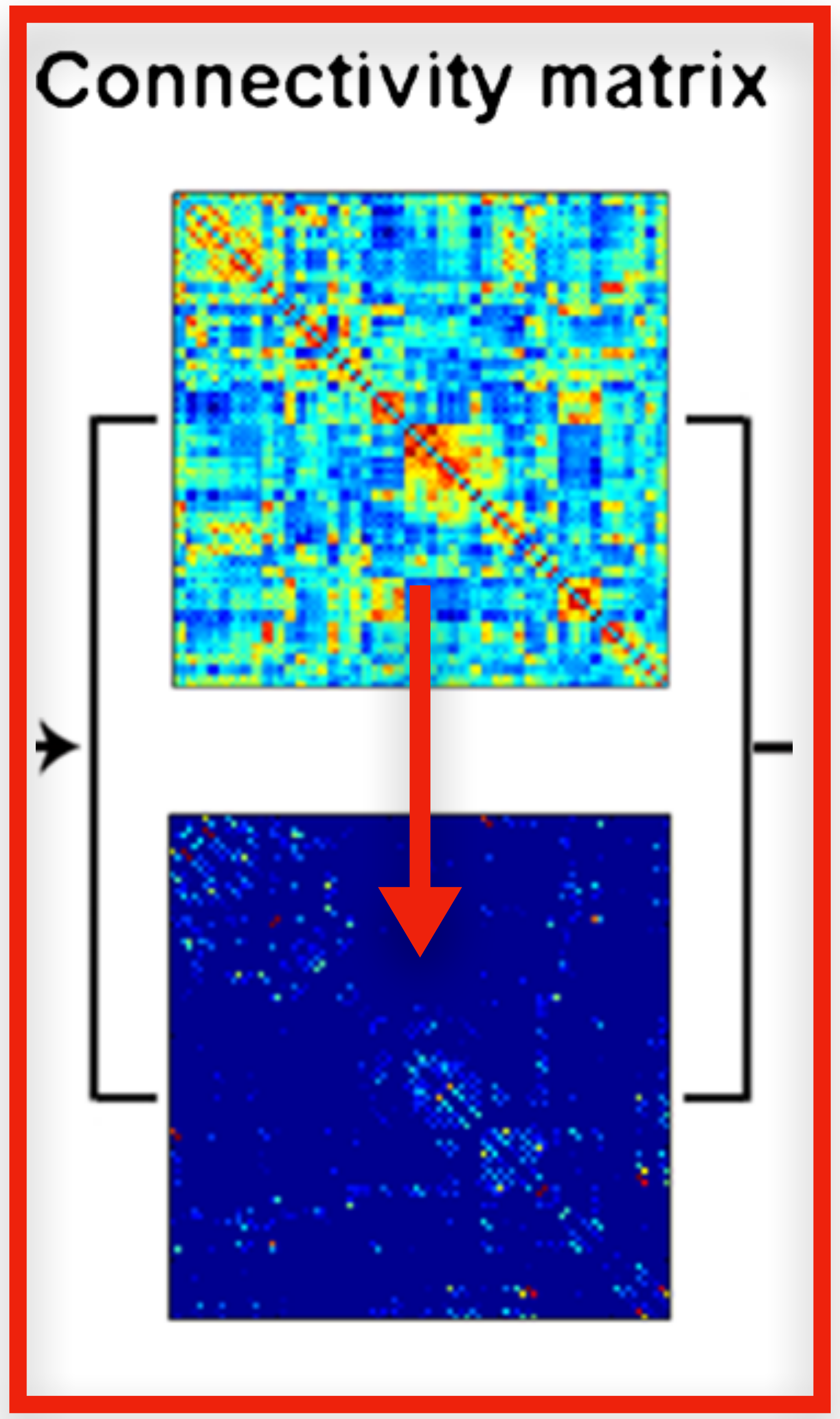
Threshold this matrix?

Edge (i,j) exists if w_{ij} > cutoff

weighted network

Common task: thinning

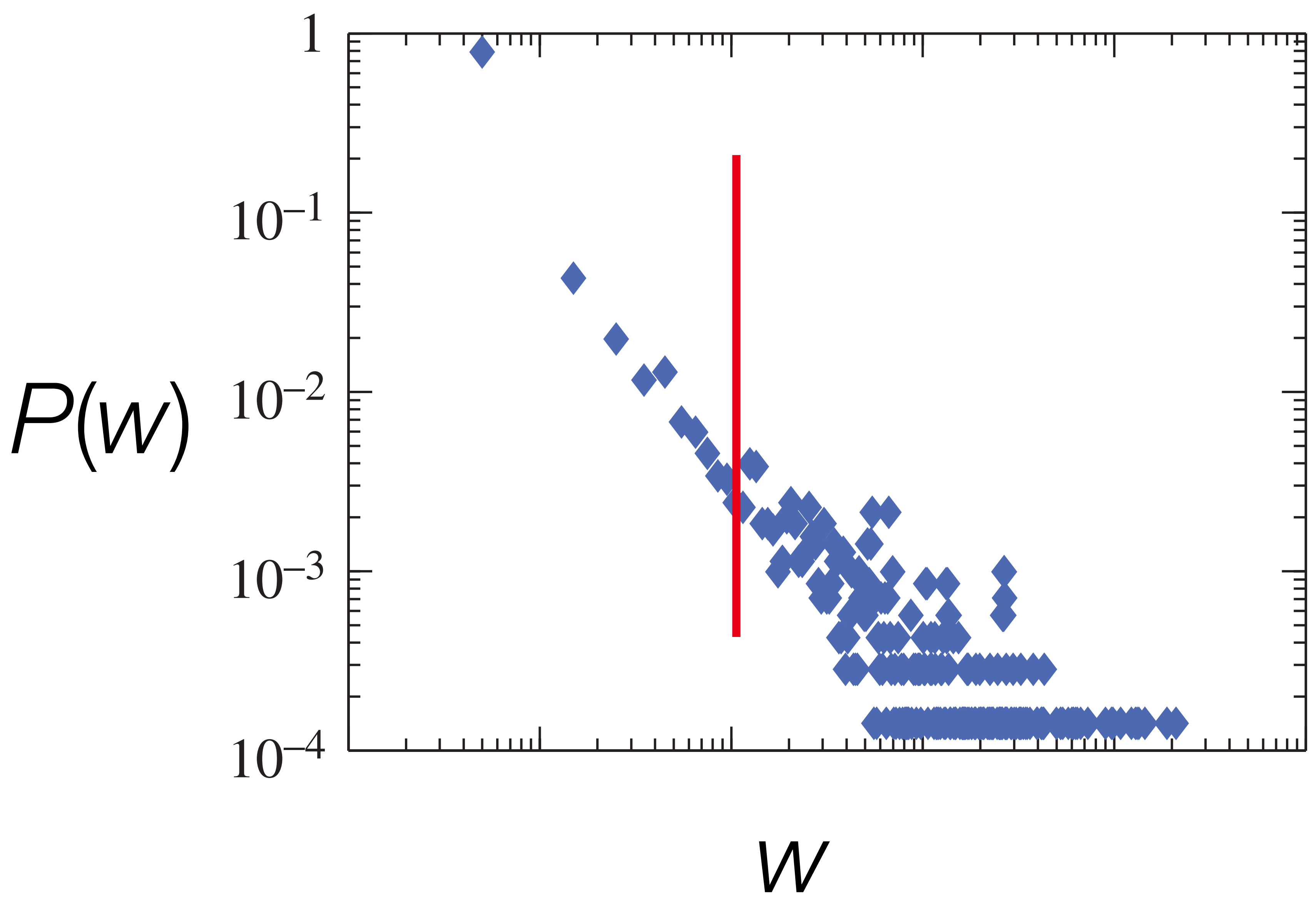
Network is very dense, lots of potentially spurious edges
How to sparsify?



Threshold this matrix?

Edge (i,j) exists if w_{ij} > cutoff

weighted network



Common task: thinning

Idea: Use a local threshold

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Serrano *et al*, *PNAS* (2009)

Common task: thinning

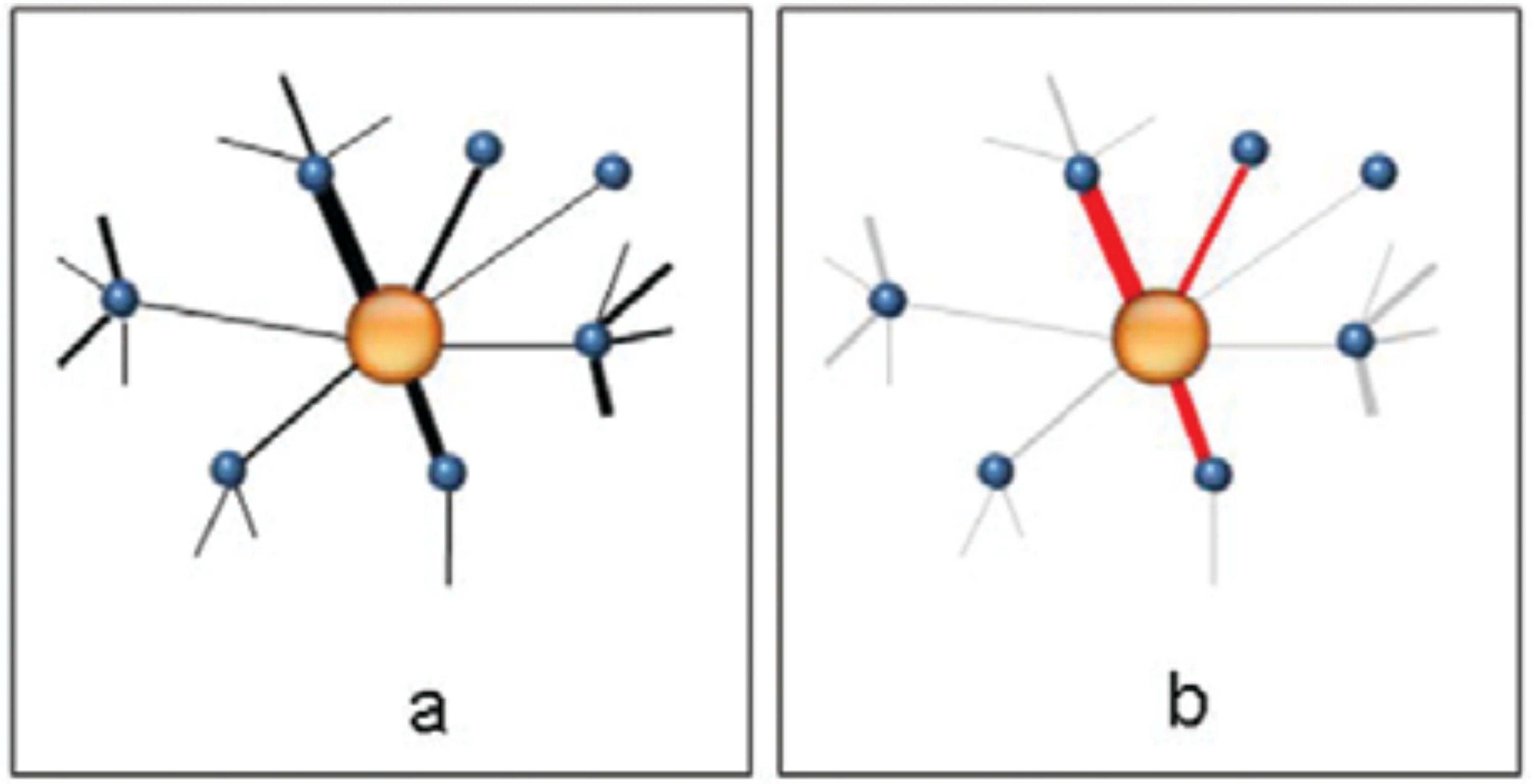
Idea: Use a local threshold

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Normalize weights in the neighborhood of a node:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$



Serrano *et al*, *PNAS* (2009)

Common task: thinning

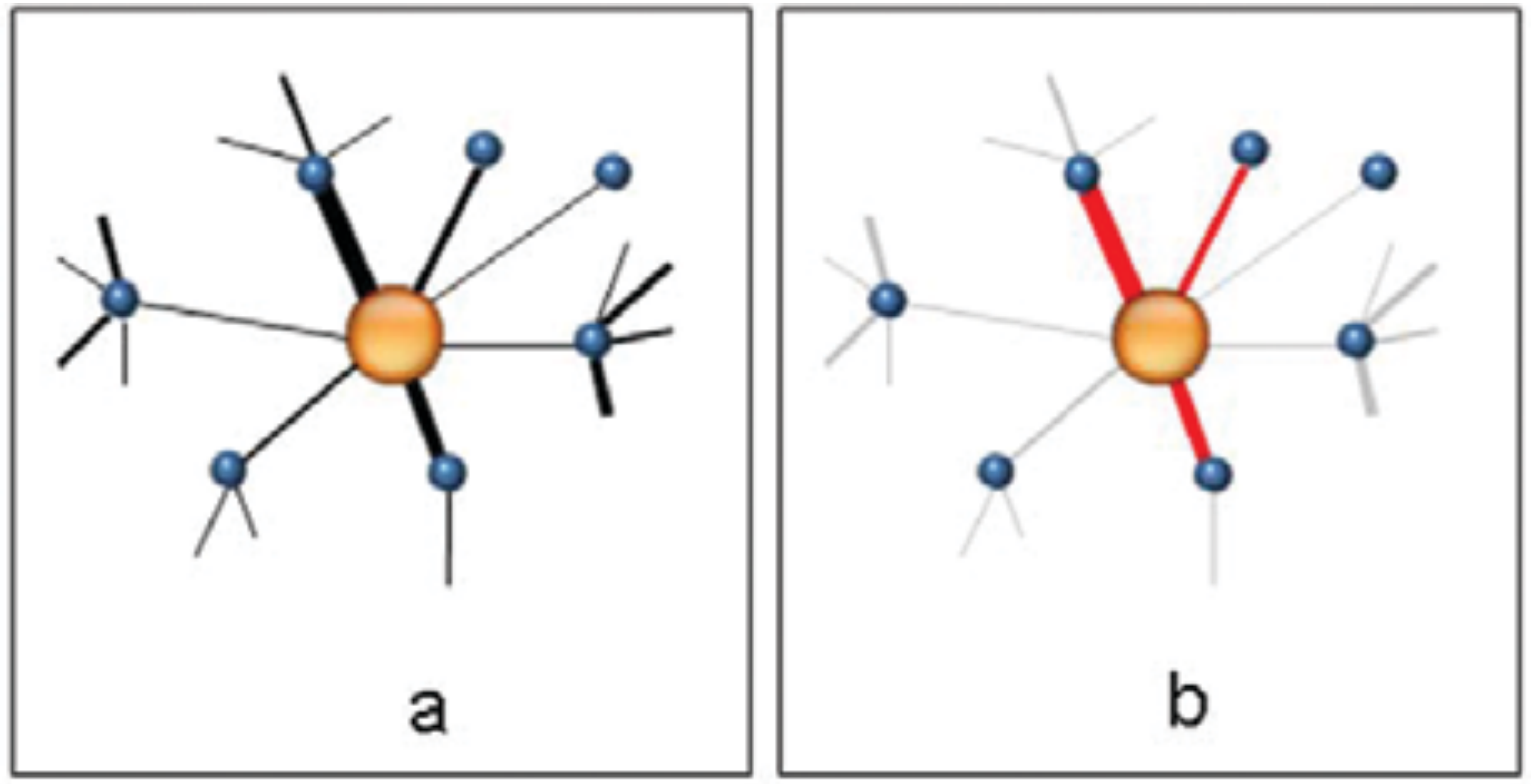
Idea: Use a local threshold

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Normalize weights in the neighborhood of a node:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$



Keep (*i*,*j*) with statistically significant values p_{ij}

How?

Serrano *et al*, *PNAS* (2009)

Common task: thinning

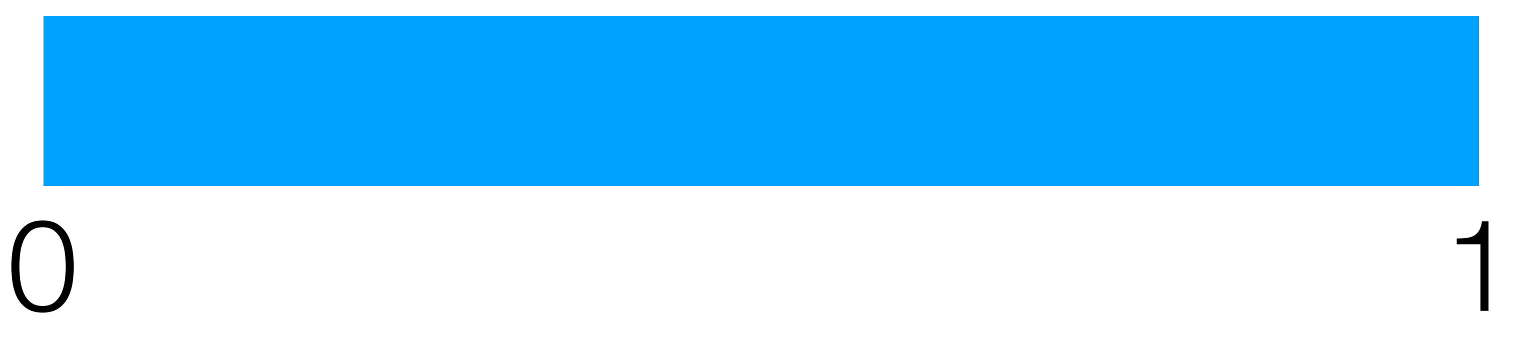
Idea: Use a local threshold

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Normalize weights in the neighborhood of a node:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$



Because p_{ij} sum to 1, imagine dropping k_i-1 points uniformly at random onto [0,1]

Serrano *et al*, *PNAS* (2009)

Common task: thinning

Idea: Use a local threshold


Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Normalize weights in the neighborhood of a node:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$

Because p_{ij} sum to 1, imagine dropping k_i-1 points uniformly at random onto [0,1]

Serrano *et al*, *PNAS* (2009)

Common task: thinning

Idea: Use a local threshold

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Normalize weights in the neighborhood of a node:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$

What's the prob of getting a gap between points at least as big as the observed p_{ij}?

Serrano *et al*, *PNAS* (2009)

Common task: thinning

Idea: Use a local threshold

Extracting the multiscale backbone of complex weighted networks

M. Ángeles Serrano[a,1], Marián Boguñá[b], and Alessandro Vespignani[c,d]

Normalize weights in the neighborhood of a node:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}$$

Keep edges where:

$$1 - (k_i - 1)\int_0^{p_{ij}} (1-x)^{k_i-2}\,dx = (1-p_{ij})^{k_i-1} < \alpha$$

Serrano *et al*, *PNAS* (2009)

Common task: thinning

Easy to implement! 😀

```
import networkx # http://networkx.github.io

def extract_backbone(G, weights, alpha):
    keep_graph = networkx.Graph()
    for i in G:
        neighbors = G[i]
        k = len(neighbors)
        if k > 1:
            W = sum( weights[i,j] for j in neighbors )
            for j in neighbors:
                pij = 1.0*weights[i,j]/W
                if (1-pij)**(k-1) < alpha: # edge significant
                    keep_graph.add_edge( i,j )
    return keep_graph
```

Common task: thinning

Robustness and modular structure in networks

JAMES P. BAGROW

Mathematics & Statistics, University of Vermont, Burlington, VT, USA
and
Center for Complex Network Research, Northeastern University, Boston, MA, USA
(e-mail: james.bagrow@uvm.edu*)*

SUNE LEHMANN

DTU Informatics, Technical University of Denmark, Kgs Lyngby, Denmark
and
College of Computer and Information Science, Northeastern University, Boston, MA, USA
(e-mail: sljo@dtu.dk*)*

YONG-YEOL AHN

School of Informatics & Computing, Indiana University, Bloomington IN, USA
and
Center for Complex Network Research, Northeastern University, Boston, MA, USA
(e-mail: yyahn@indiana.edu*)*

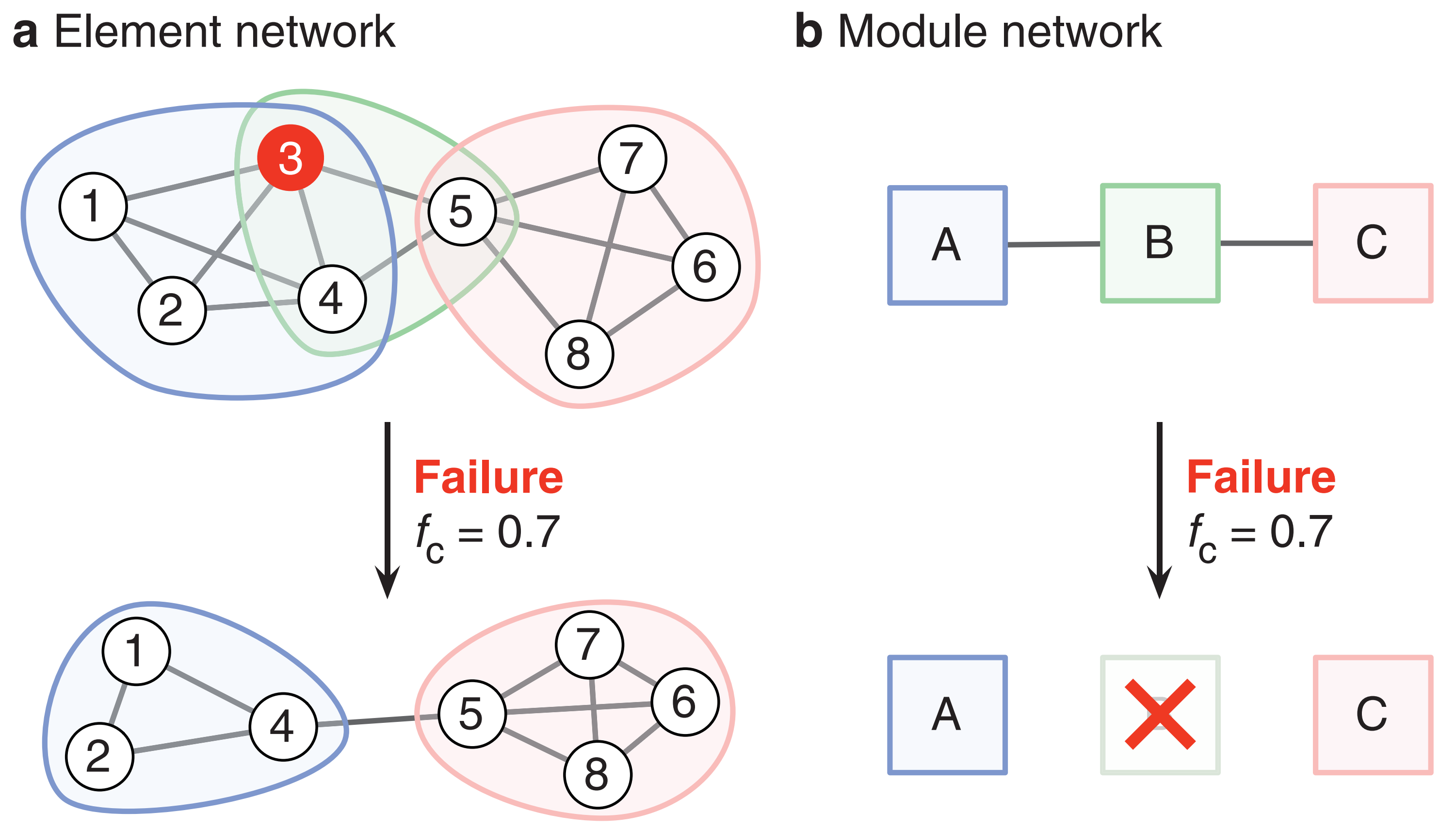
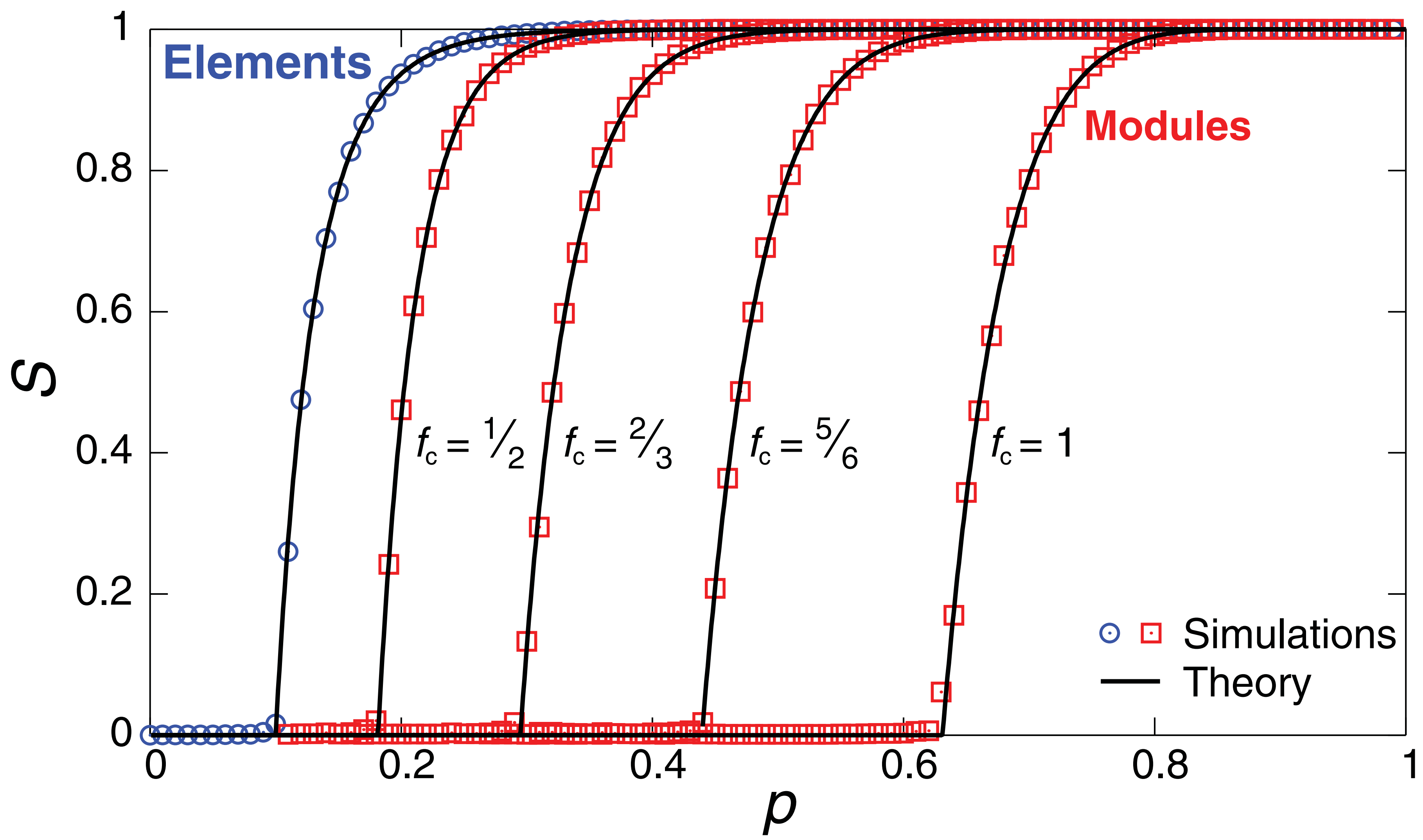


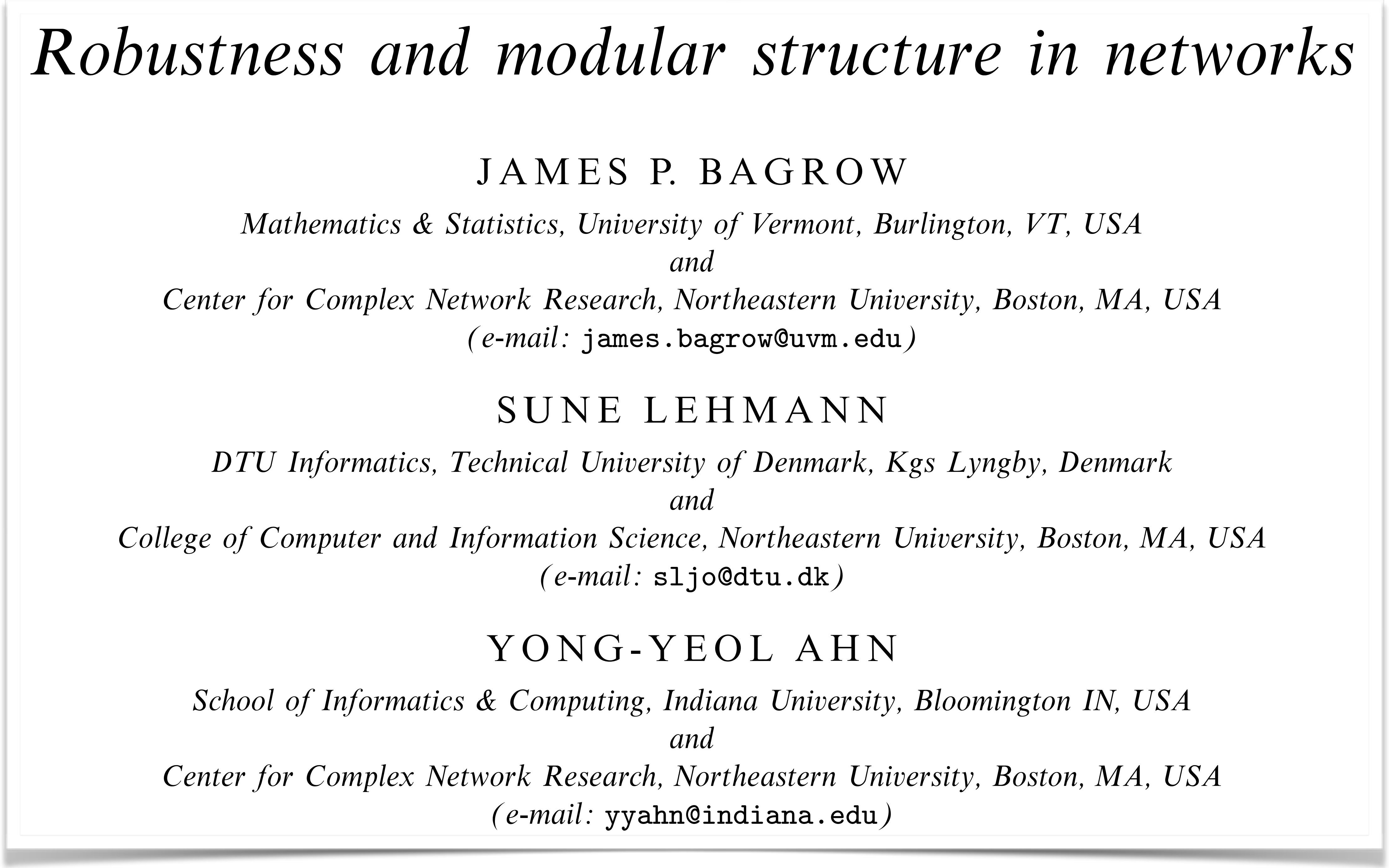
Fig. 1. Modeling failures in modular networks. We analyze two networks, one representing



Example where I used the method

Bagrow *et al.* (2015)

Common task: thinning



Robustness and modular structure in networks

JAMES P. BAGROW

Mathematics & Statistics, University of Vermont, Burlington, VT, USA
and
Center for Complex Network Research, Northeastern University, Boston, MA, USA
(e-mail: james.bagrow@uvm.edu*)*

SUNE LEHMANN

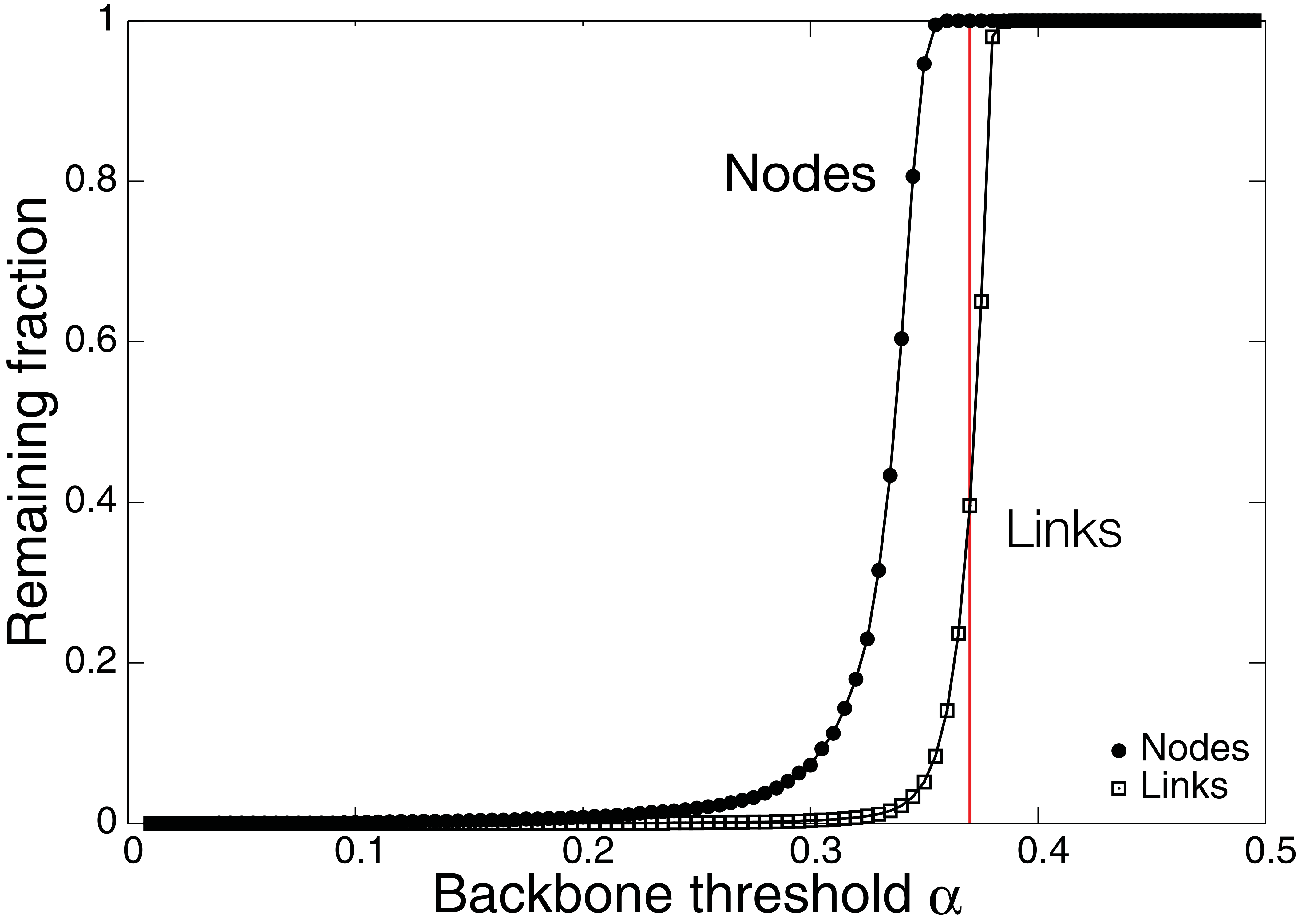
DTU Informatics, Technical University of Denmark, Kgs Lyngby, Denmark
and
College of Computer and Information Science, Northeastern University, Boston, MA, USA
(e-mail: sljo@dtu.dk*)*

YONG-YEOL AHN

School of Informatics & Computing, Indiana University, Bloomington IN, USA
and
Center for Complex Network Research, Northeastern University, Boston, MA, USA
(e-mail: yyahn@indiana.edu*)*

Example where I used the method

Applied to fMRI data



Bagrow *et al.* (2015)

Case study:
Nodes are ambiguous

Inferring the size of the causal universe: features and fusion of causal attribution networks

Daniel Berenberg[1,2] and James P. Bagrow[3,2,*]

[1]Department of Computer Science, University of Vermont, Burlington, VT, United States
[2]Vermont Complex Systems Center, University of Vermont, Burlington, VT, United States
[3]Department of Mathematics & Statistics, University of Vermont, Burlington, VT, United States
[*]Corresponding author. Email: james.bagrow@uvm.edu, Homepage: bagrow.com

December 14, 2018

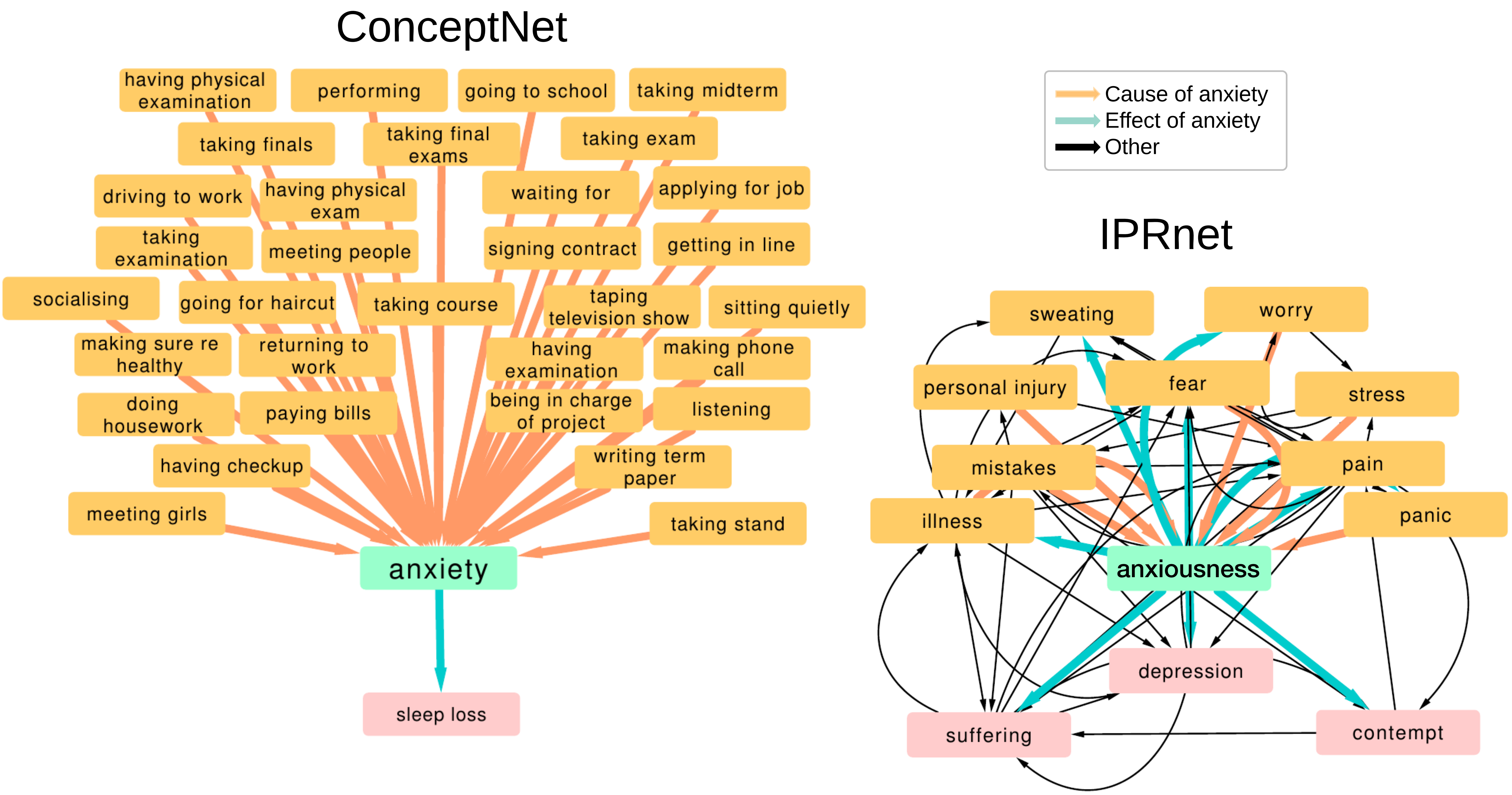
Crowdsourced knowledge graphs

Berenberg & Bagrow (2018)

Knowledge graphs

$s_i =$ "anxiety"

$s_j =$ "sleep loss"

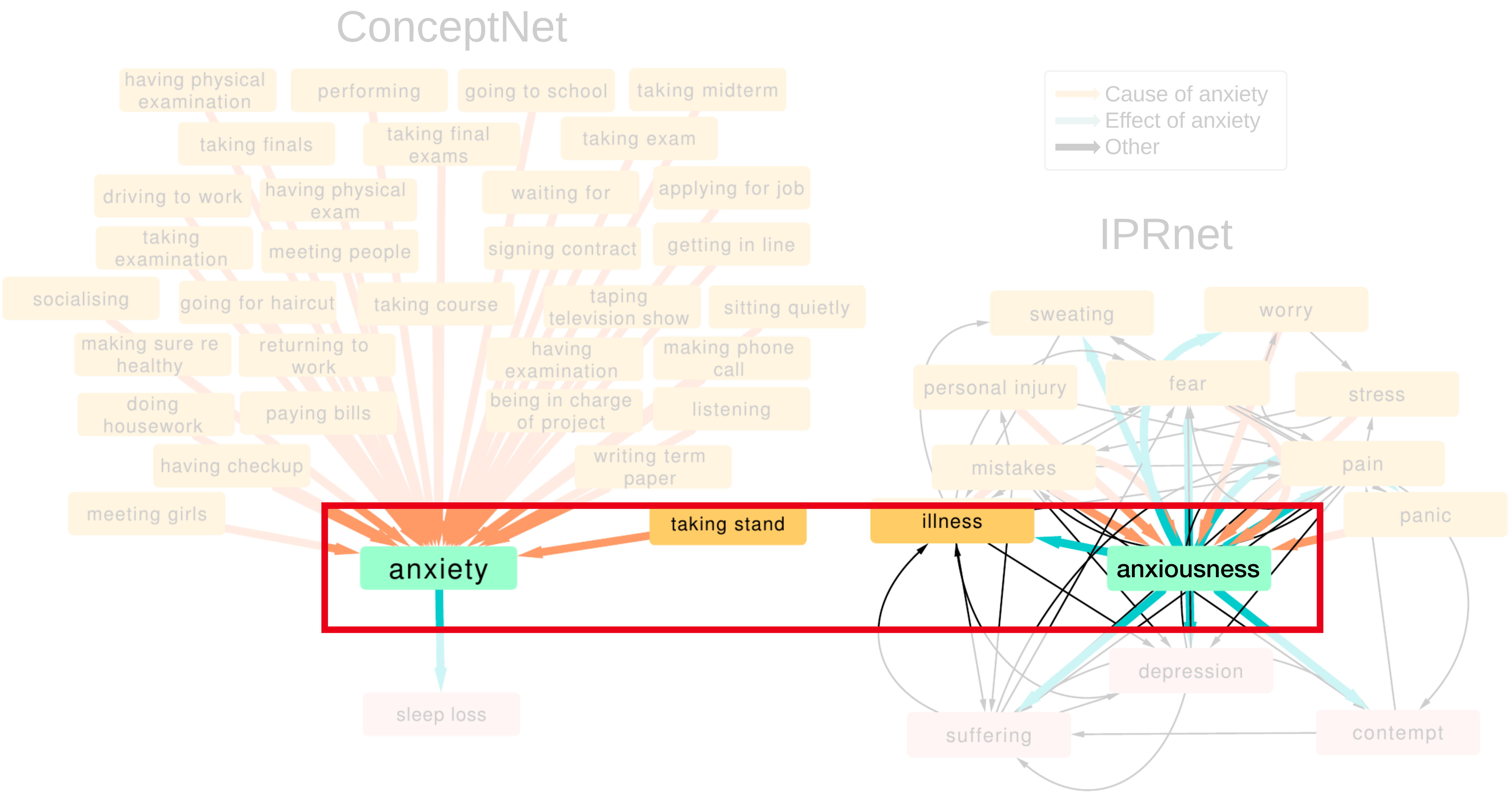


Berenberg & Bagrow (2018)

Knowledge graphs

$s_i =$ "anxiety"

$s_j =$ "sleep loss"



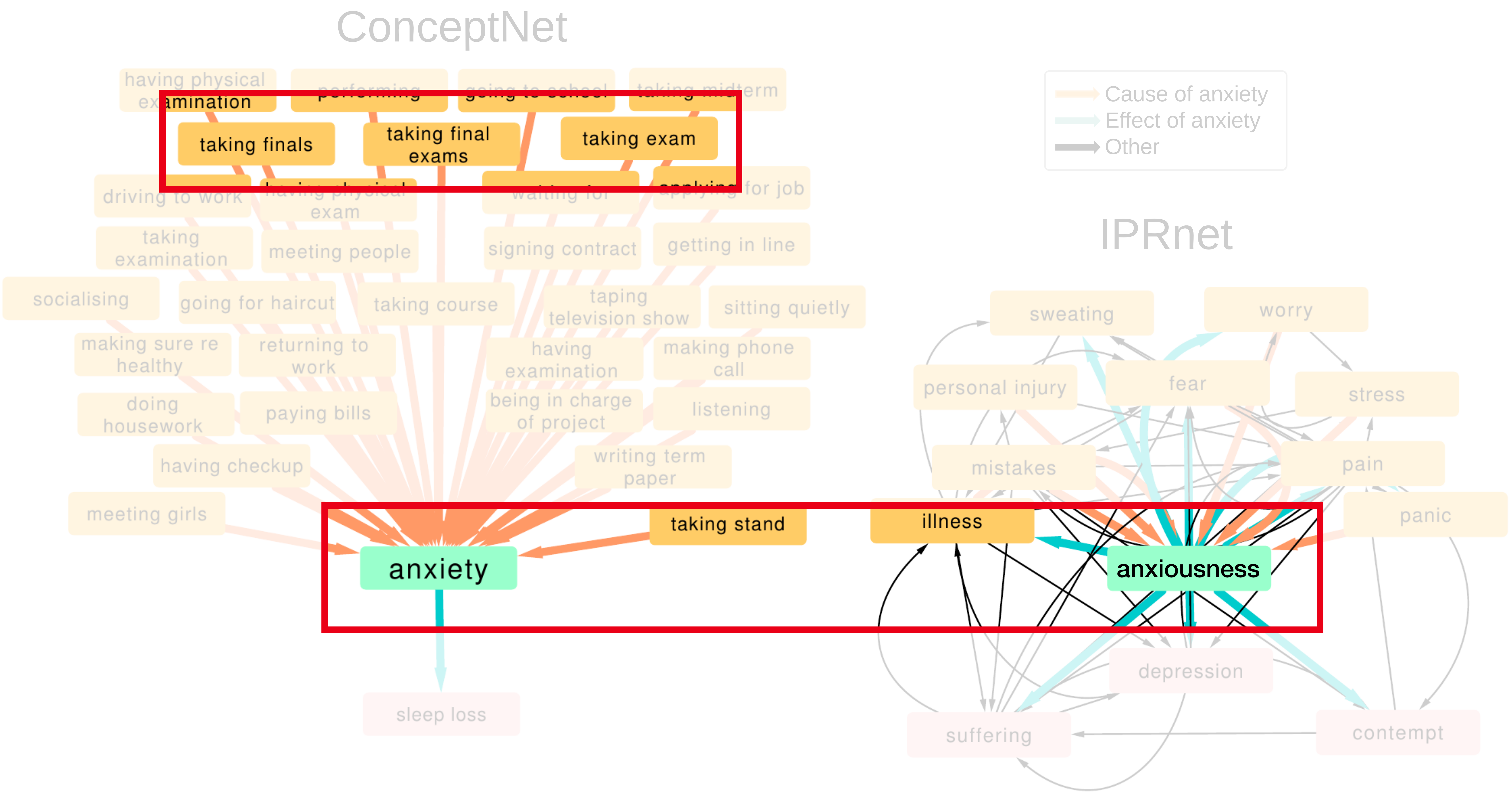
Nodes are identified *only* by these text…
Could be ambiguous, even within one network…

Berenberg & Bagrow (2018)

Knowledge graphs

s_i = "anxiety"

s_j = "sleep loss"



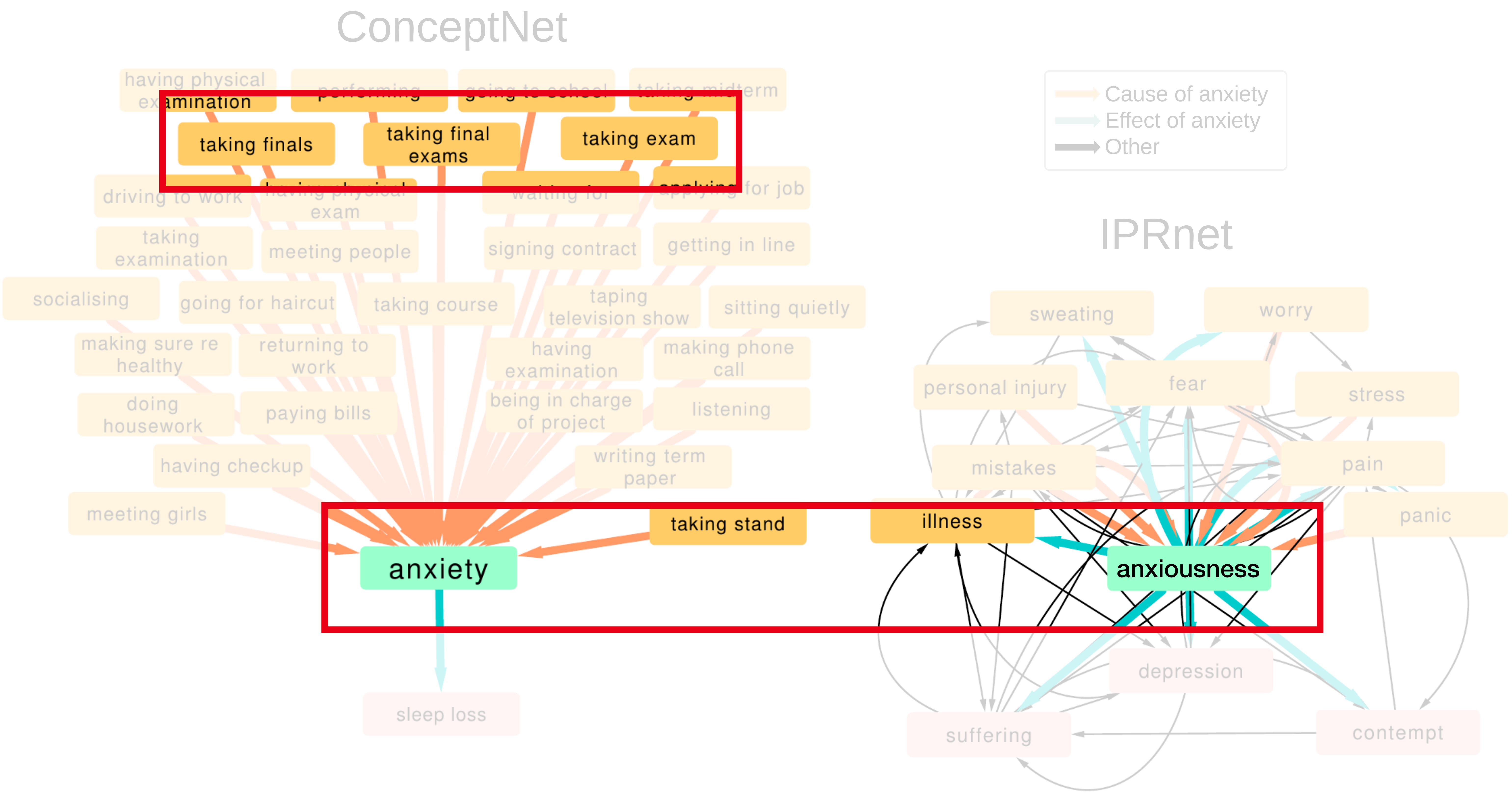
Nodes are identified *only* by these text…
Could be ambiguous, even within one network…

Berenberg & Bagrow (2018)

Knowledge graphs

$s_i =$ "anxiety"

$s_j =$ "sleep loss"



Nodes are identified *only* by these text…
Could be ambiguous, even within one network…

Can we combine these different networks together?

Berenberg & Bagrow (2018)

NetFUSES: Network FUsion with SEmantic Similarity

s_i = "anxiety"

s_j = "sleep loss"

Define a semantic similarity S between sentences:

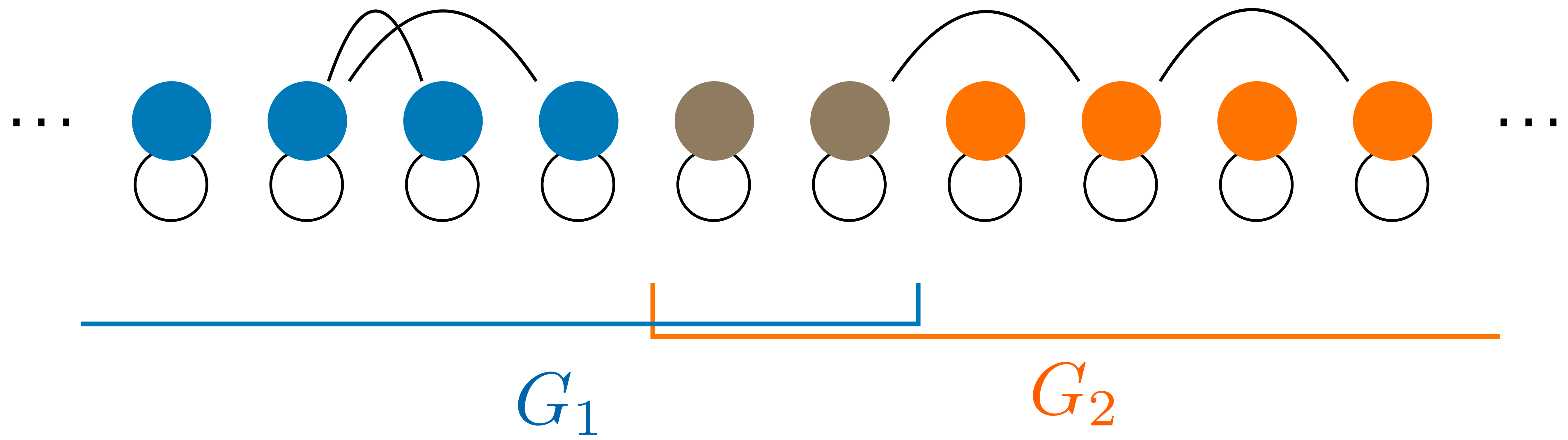
$$S(s_i, s_j) \leq 1$$

$$S(s_i, s_i) = 1$$

$$S(s_i, s_j) = S(s_j, s_i)$$

Threshold $\quad S(s_i, s_j) \geq t \quad i, j \in V_1 \cup V_2$

edges of a *fusion indicator graph*:



Fuse nodes using connected components

Berenberg & Bagrow (2018)

NetFUSES: Network FUsion with SEmantic Similarity

s_i = "anxiety"

s_j = "sleep loss"

Define a semantic similarity S between sentences:

$$S(s_i, s_j) \leq 1$$

$$S(s_i, s_i) = 1$$

$$S(s_i, s_j) = S(s_j, s_i)$$

Threshold $S(s_i, s_j) \geq t$ $i, j \in V_1 \cup V_2$

edges of a *fusion indicator graph*:

How to measure semantic similarity of text?

Fuse nodes using connected components

Berenberg & Bagrow (2018)

Machine Learning

(How to measure semantic similarity of text?)

Measuring semantic similarity with neural networks

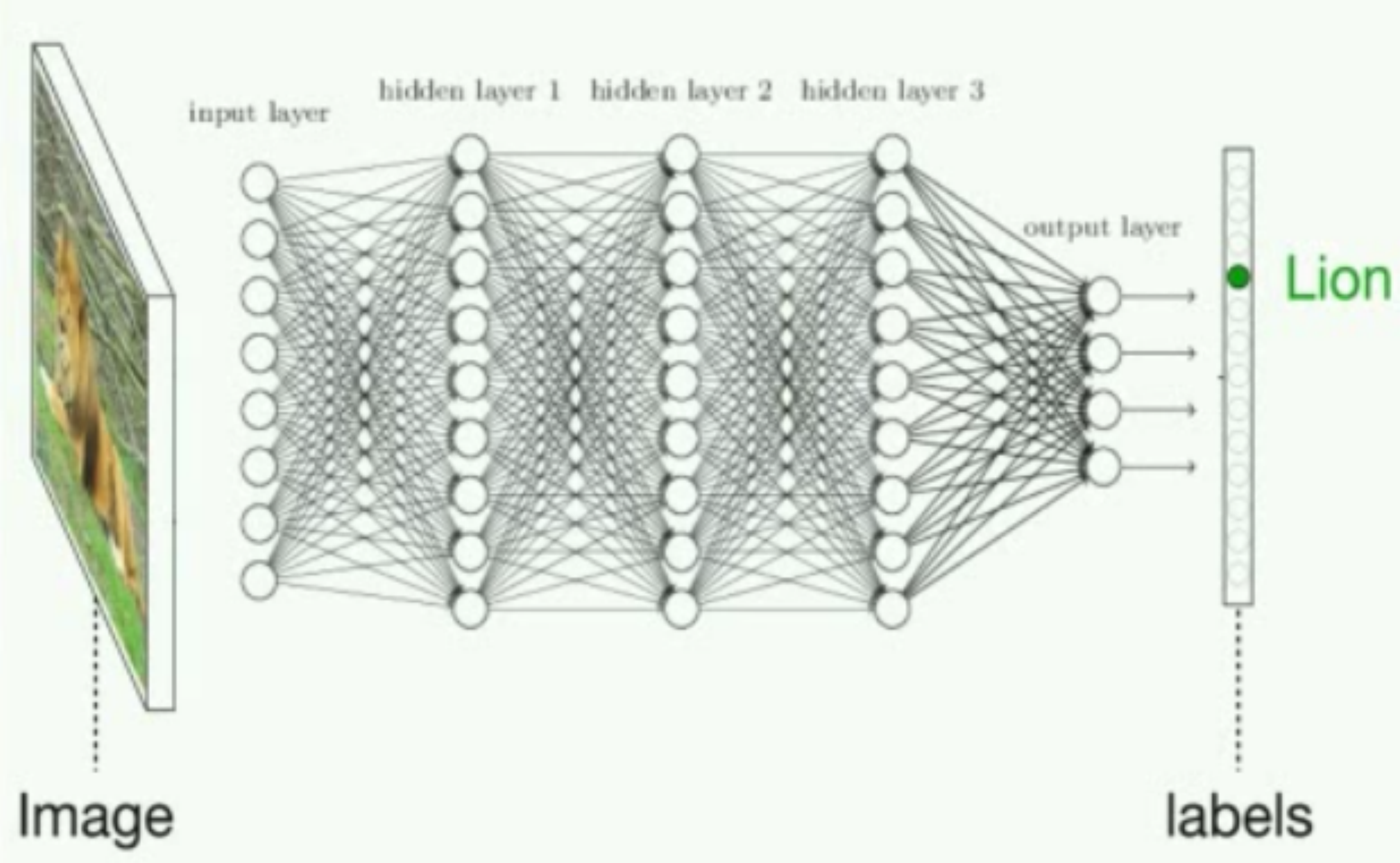


Image: Jeff Clune

Example: image classification

using **training data:** labeled images

…, ( ,'lion'), …

Measuring semantic similarity with neural networks

Example: image classification

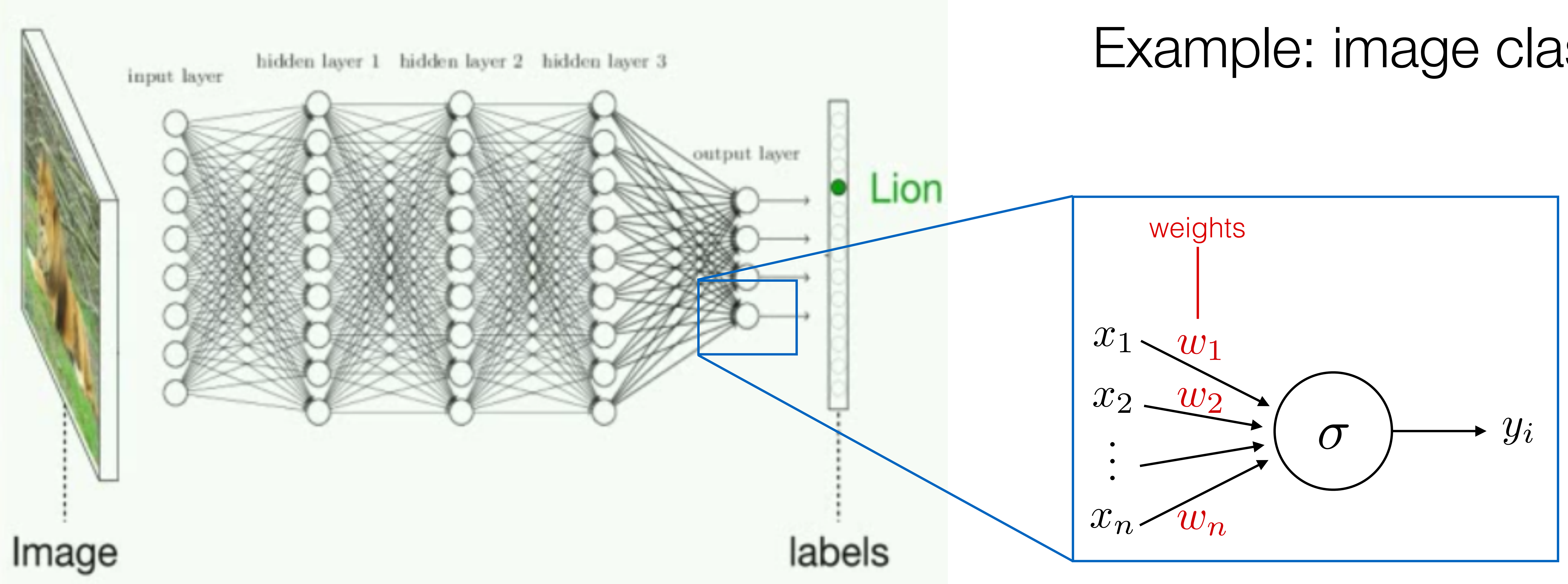


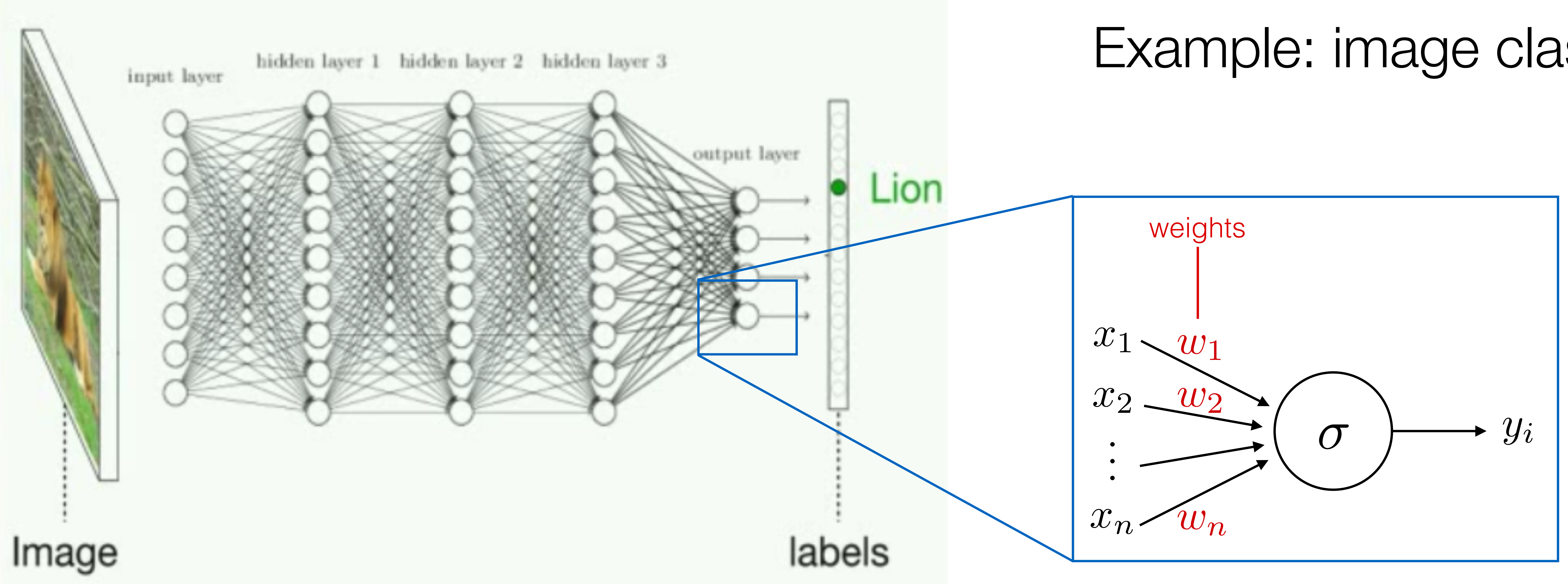
Image: Jeff Clune

$$y = \sigma(\mathbf{w}^\top \mathbf{x})$$

using **training data:** labeled images

…, ( ,'lion'), …

Measuring semantic similarity with neural networks



[Image: Jeff Clune](#)

Example: image classification

$$y = \sigma(\mathbf{w}^\top \mathbf{x})$$

using **training data**: labeled images

…, (

,'lion'), …

What training data can we use for **text**?

"You shall know a word by the company it keeps."

–JR Firth

Distributional Semantics

"You shall know a word by the company it keeps."

–JR Firth

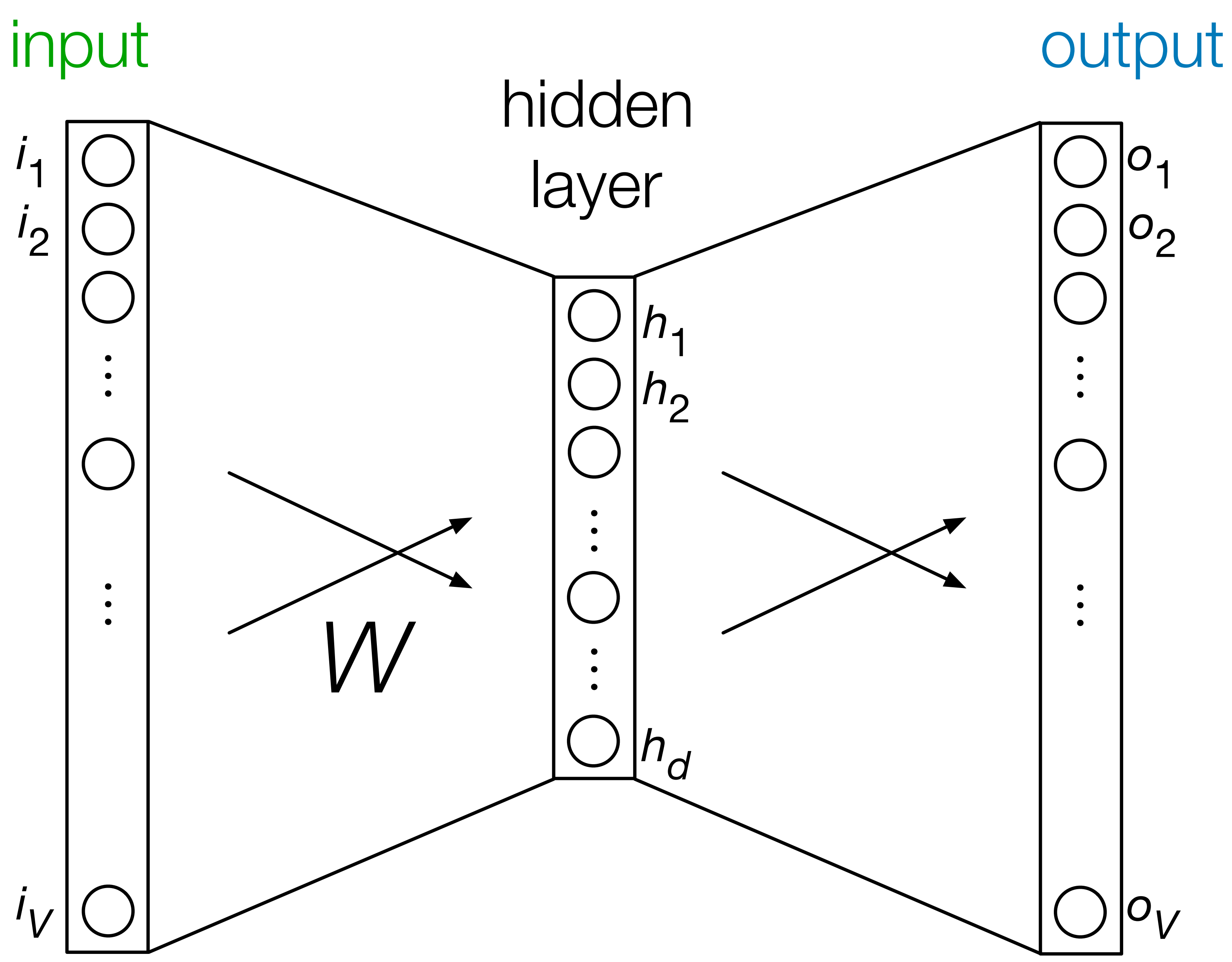
Distributional Semantics

… worlds are yours except europa attempt no landings there …

Turn *large* text corpus into collection of word-context pairs

Predict word
from context

Training data

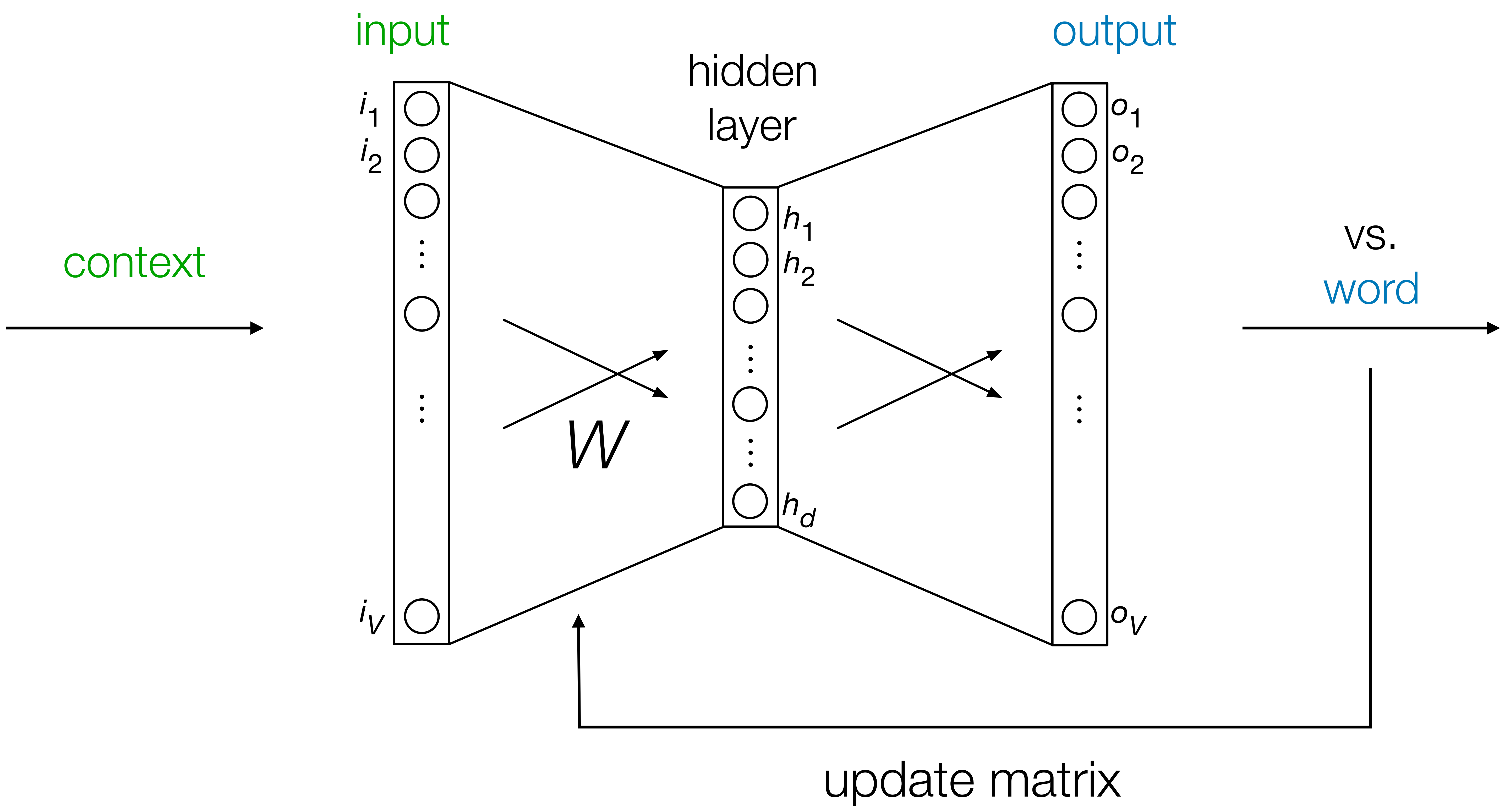


(Or predict context
from word!)

Bengio *et al.* JMLR (2003)
Mikolov *et al.* *NIPS* (2013)

Predict word
from context

Training data

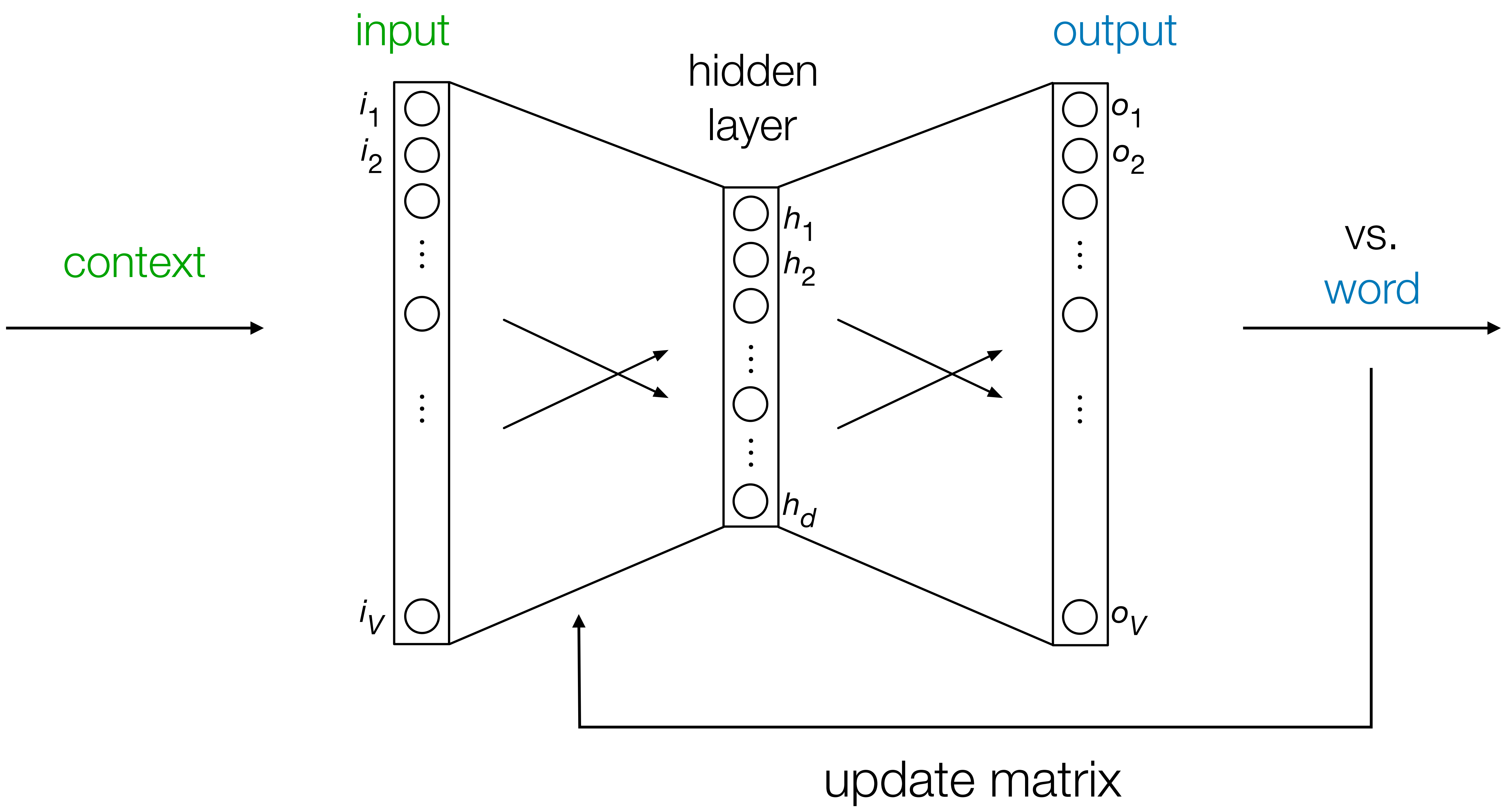


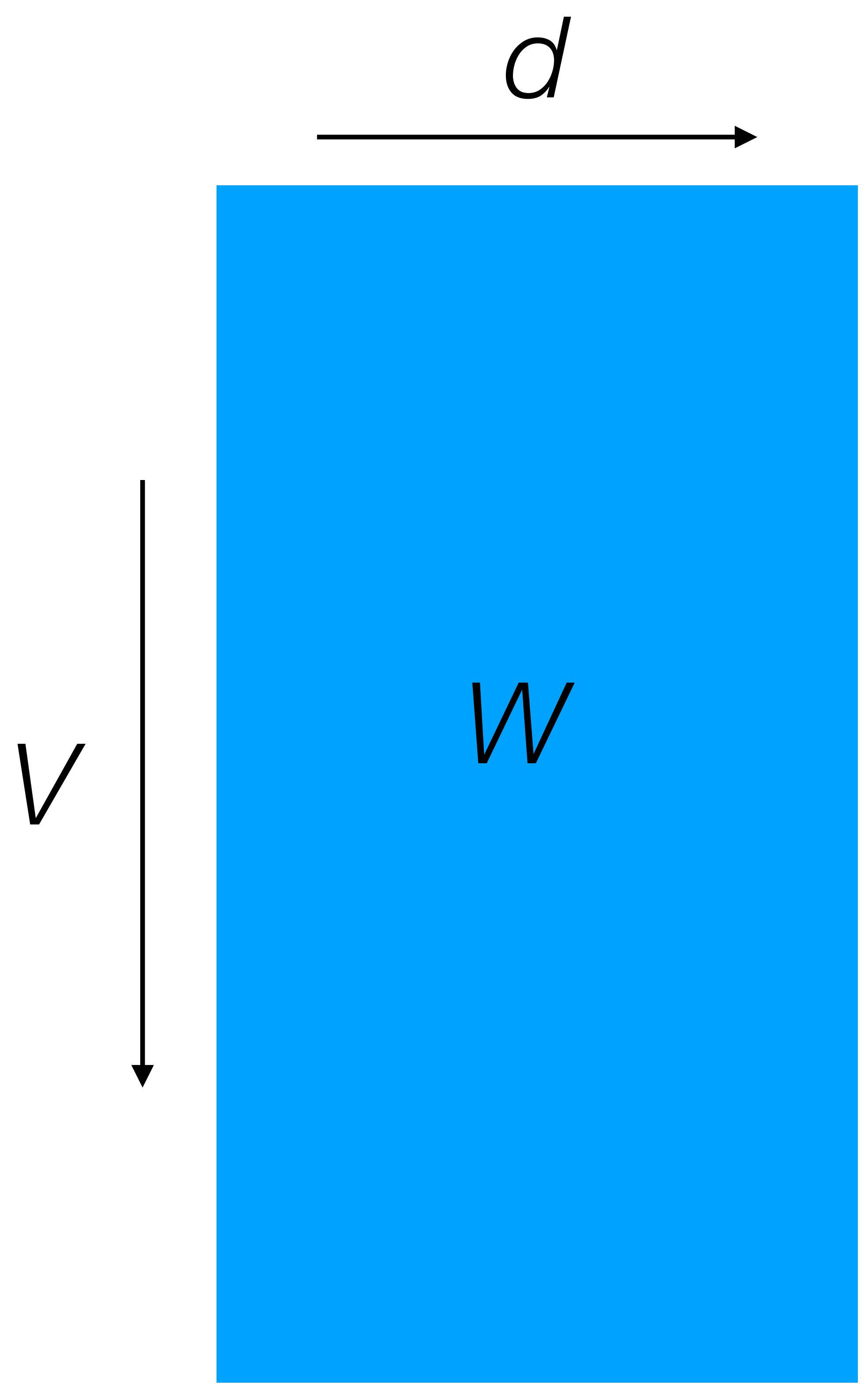
(Or predict context
from word!)

Bengio *et al.* JMLR (2003)
Mikolov *et al.* *NIPS* (2013)

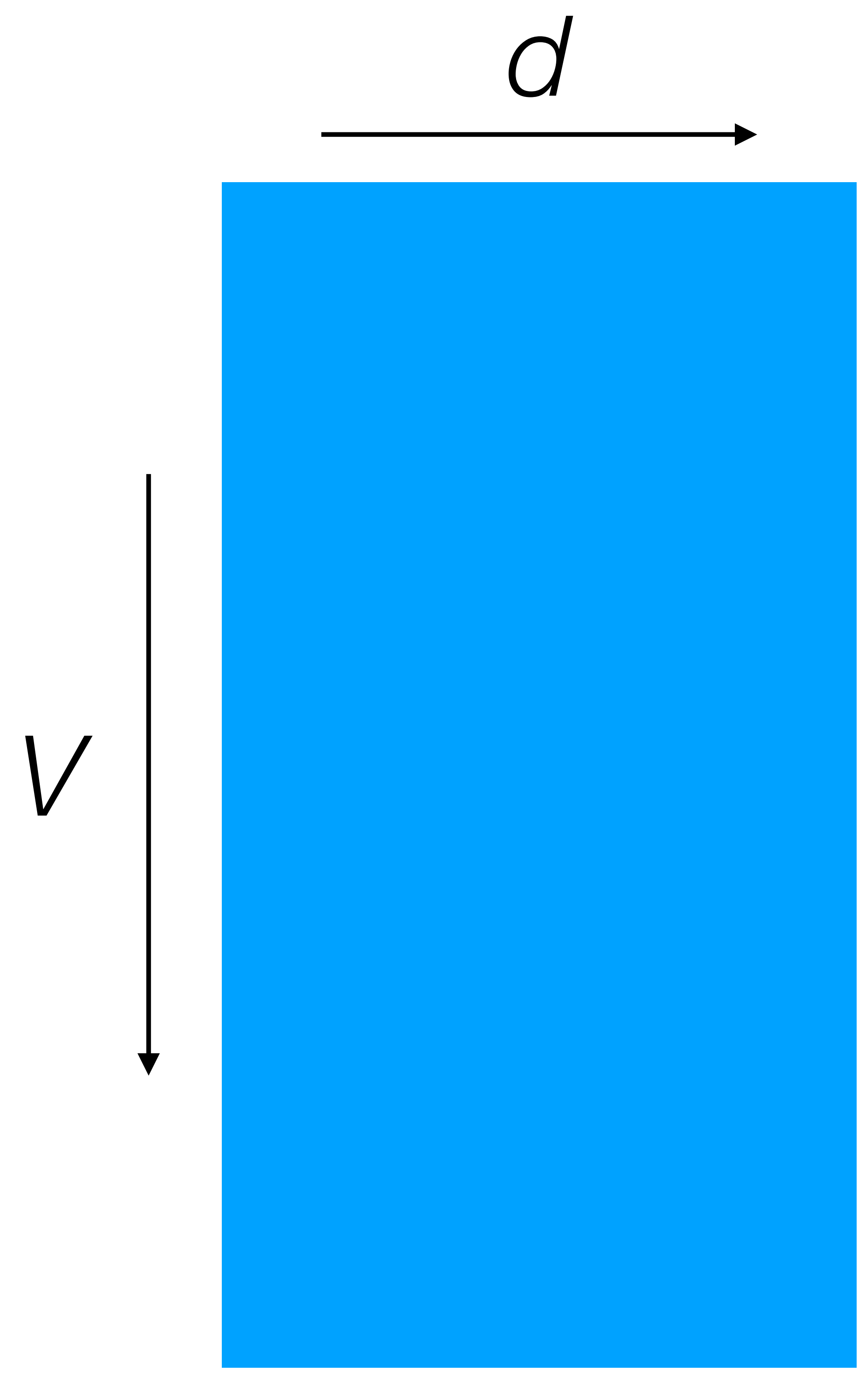
Predict word
from context

Training data



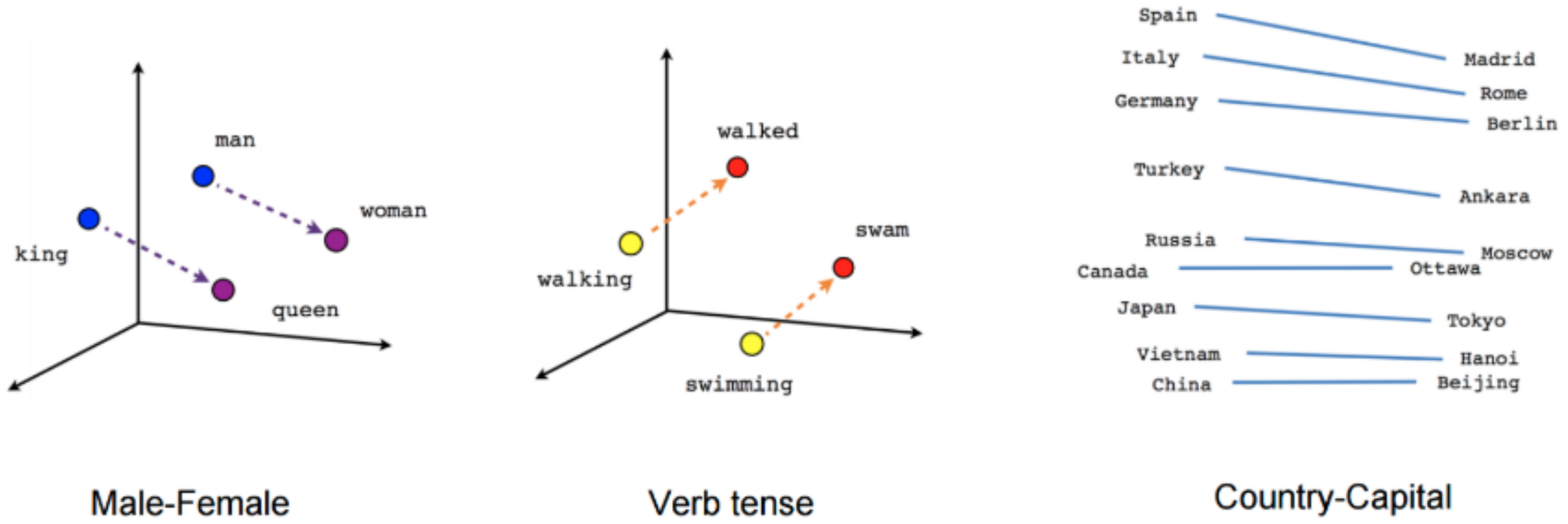


Bengio *et al.* JMLR (2003)
Mikolov *et al.* *NIPS* (2013)



— Each row is an *d*-dimensional *word vector*

vectors encode semantics



Bengio *et al.* JMLR (2003)
Mikolov *et al.* *NIPS* (2013)

If this sounds like SVD, you're not crazy….

$$M \sim \log \frac{P(w, c)}{P(w)P(c)}$$

$$M = U\Sigma V^\top$$

$$M \approx M_d = U_d \Sigma_d V_d^\top$$

$$\boxed{W^{\text{SVD}} = U_d \Sigma_d}$$

Neural Word Embedding as Implicit Matrix Factorization

Omer Levy
Department of Computer Science
Bar-Ilan University
omerlevy@gmail.com

Yoav Goldberg
Department of Computer Science
Bar-Ilan University
yoav.goldberg@gmail.com

Neural network implicitly performs *weighted* factorization of *M*

Levy & Goldberg, *NIPS* (2014)

Embedding words in vector spaces has taken the world by storm

Google Scholar

Distributed representations of words and phrases a
T Mikolov, I Sutskever, K Chen, GS Corrado… - Advances in neural
The recently introduced continuous Skip-gram model is an efficient m
quality distributed vector representations that capture a large number
and semantic word relationships. In this paper we present several imp
Cited by 16889 Related articles All 32 versions Import

word vectors, sentence vectors, *thought* vectors…

Lots of natural language processing applications including semantic similarity:

$$S(s_i, s_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$$

Embedding words in vector spaces has taken the world by storm

Google Scholar

Distributed **representations** of **words** and phrases a

T **Mikolov**, I Sutskever, K Chen, GS Corrado… - Advances in neural

The recently introduced continuous Skip-gram model is an efficient m
quality distributed vector representations that capture a large number
and semantic word relationships. In this paper we present several imp

☆ 99 Cited by 16889 Related articles All 32 versions Import

word vectors, sentence vectors, *thought* vectors…

Lots of natural language processing applications including semantic similarity:

$$S(s_i, s_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$$

Must be approached with caution

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1*] Joanna J. Bryson,[1,2*] Arvind Narayanan[1*]

"[…] word vectors contain stereotypes matching those documented with the [Implicit Association Test]"

Caliskan *et al.* Science (2017)

Must be approached with caution

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan,[1*] Joanna J. Bryson,[1,2*] Arvind Narayanan[1*]

"[…] word vectors contain stereotypes matching those documented with the [Implicit Association Test]"

ConceptNet Numberbatch

ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge

Robyn Speer   Joanna Lowry-Duda

Neural language representations predict outcomes of scientific research

James P. Bagrow[1,2,*], Daniel Berenberg[3,2], and Joshua Bongard[3,2]

[1]Department of Mathematics & Statistics, University of Vermont, Burlington, VT, United States
[2]Vermont Complex Systems Center, University of Vermont, Burlington, VT, United States
[3]Department of Computer Science, University of Vermont, Burlington, VT, United States
[*]Corresponding author. Email: james.bagrow@uvm.edu, Homepage: bagrow.com

May 17, 2018

Caliskan *et al.* Science (2017)

Speer & Lowry-Duda SemEval-2017 (2017)
Bagrow *et al.* (2018)

Machine Learning for Networks

DeepWalk: Online Learning of Social Representations

Bryan Perozzi
Stony Brook University
Department of Computer Science

Rami Al-Rfou
Stony Brook University
Department of Computer Science

Steven Skiena
Stony Brook University
Department of Computer Science

{bperozzi, ralrfou, skiena}@cs.stonybrook.edu

word embedding (word2vec):

… worlds are yours except europa attempt no landings there …

Perozzi *et al.* KDD (2014)

DeepWalk: Online Learning of Social Representations

Bryan Perozzi
Stony Brook University
Department of Computer Science

Rami Al-Rfou
Stony Brook University
Department of Computer Science

Steven Skiena
Stony Brook University
Department of Computer Science

{bperozzi, ralrfou, skiena}@cs.stonybrook.edu

word embedding (word2vec): … worlds are yours except europa attempt no landings there …

DeepWalk: $[\ldots, v_{i-2}, v_{i-1}, v_i, v_{i+1}, v_{i+2}, \ldots]$

- Take a short random walk on the graph
- record the visited sequence of vertices
- treat vertices as words and do embedding!

Perozzi *et al.* KDD (2014)

DeepWalk: Online Learning of Social Representations

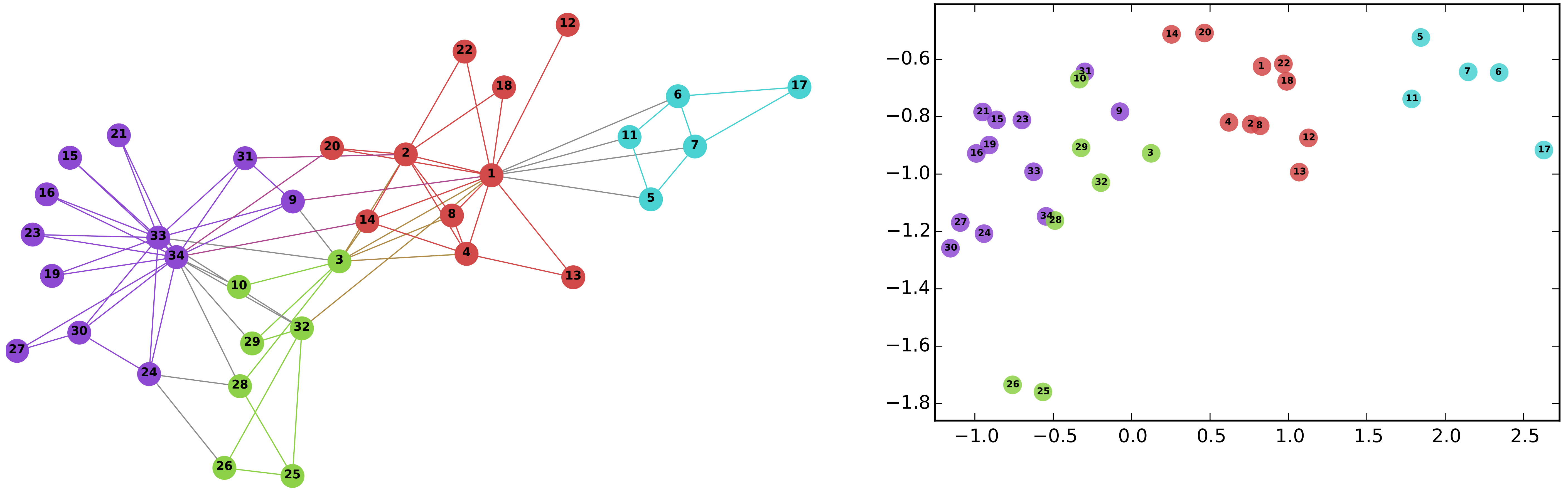
Bryan Perozzi
Stony Brook University
Department of Computer Science

Rami Al-Rfou
Stony Brook University
Department of Computer Science

Steven Skiena
Stony Brook University
Department of Computer Science

{bperozzi, ralrfou, skiena}@cs.stonybrook.edu

DeepWalk:



(a) Input: Karate Graph (b) Output: Representation

Perozzi *et al.* KDD (2014)

Is it also a matrix factorization problem? Yes!

Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec

Jiezhong Qiu[†*], Yuxiao Dong[‡], Hao Ma[‡], Jian Li[♯], Kuansan Wang[‡], and Jie Tang[†]

Qiu *et al.*, *WSDM'18* (2018)

Is it also a matrix factorization problem? Yes!

Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec

Jiezhong Qiu[†*], Yuxiao Dong[‡], Hao Ma[‡], Jian Li[♯], Kuansan Wang[‡], and Jie Tang[†]

Table 1: The matrices that are implicitly approximated and factorized by DeepWalk, LINE, PTE, and node2vec.

Algorithm	Matrix
DeepWalk	$\log\left(\mathrm{vol}(G)\left(\frac{1}{T}\sum_{r=1}^{T}(\boldsymbol{D}^{-1}\boldsymbol{A})^r\right)\boldsymbol{D}^{-1}\right)-\log b$
LINE	$\log\left(\mathrm{vol}(G)\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{D}^{-1}\right)-\log b$
PTE	$\log\left(\begin{bmatrix}\alpha\,\mathrm{vol}(G_{\mathrm{ww}})(\boldsymbol{D}_{\mathrm{row}}^{\mathrm{ww}})^{-1}\boldsymbol{A}_{\mathrm{ww}}(\boldsymbol{D}_{\mathrm{col}}^{\mathrm{ww}})^{-1}\\ \beta\,\mathrm{vol}(G_{\mathrm{dw}})(\boldsymbol{D}_{\mathrm{row}}^{\mathrm{dw}})^{-1}\boldsymbol{A}_{\mathrm{dw}}(\boldsymbol{D}_{\mathrm{col}}^{\mathrm{dw}})^{-1}\\ \gamma\,\mathrm{vol}(G_{\mathrm{lw}})(\boldsymbol{D}_{\mathrm{row}}^{\mathrm{lw}})^{-1}\boldsymbol{A}_{\mathrm{lw}}(\boldsymbol{D}_{\mathrm{col}}^{\mathrm{lw}})^{-1}\end{bmatrix}\right)-\log b$
node2vec	$\log\left(\frac{\frac{1}{2T}\sum_{r=1}^{T}\left(\sum_u \boldsymbol{X}_{w,u}\underline{\boldsymbol{P}}_{c,w,u}^{r}+\sum_u \boldsymbol{X}_{c,u}\underline{\boldsymbol{P}}_{w,c,u}^{r}\right)}{(\sum_u \boldsymbol{X}_{w,u})(\sum_u \boldsymbol{X}_{c,u})}\right)-\log b$

Notations in DeepWalk and LINE are introduced below. See detailed notations for PTE and

— Many methods besides DeepWalk

Qiu *et al.*, *WSDM'18* (2018)

Random walks are a fundamental concept when studying networks

Random walks and diffusion on networks

Naoki Masuda [a,*], Mason A. Porter [b,c,d], Renaud Lambiotte [c]

[a] *Department of Engineering Mathematics, University of Bristol, Bristol, UK*
[b] *Department of Mathematics, University of California Los Angeles, Los Angeles, USA*
[c] *Mathematical Institute, University of Oxford, Oxford, UK*
[d] *CABDyN Complexity Centre, University of Oxford, Oxford, UK*

ARTICLE INFO

ABSTRACT

Masuda *et al.*, *Physics Reports* (2017)

Random walks are a fundamental concept when studying networks

Random walks and diffusion on networks

Naoki Masuda [a,*], Mason A. Porter [b,c,d], Renaud Lambiotte [c]

[a] *Department of Engineering Mathematics, University of Bristol, Bristol, UK*
[b] *Department of Mathematics, University of California Los Angeles, Los Angeles, USA*
[c] *Mathematical Institute, University of Oxford, Oxford, UK*
[d] *CABDyN Complexity Centre, University of Oxford, Oxford, UK*

ARTICLE INFO ABSTRACT

Maps of random walks on complex networks reveal community structure

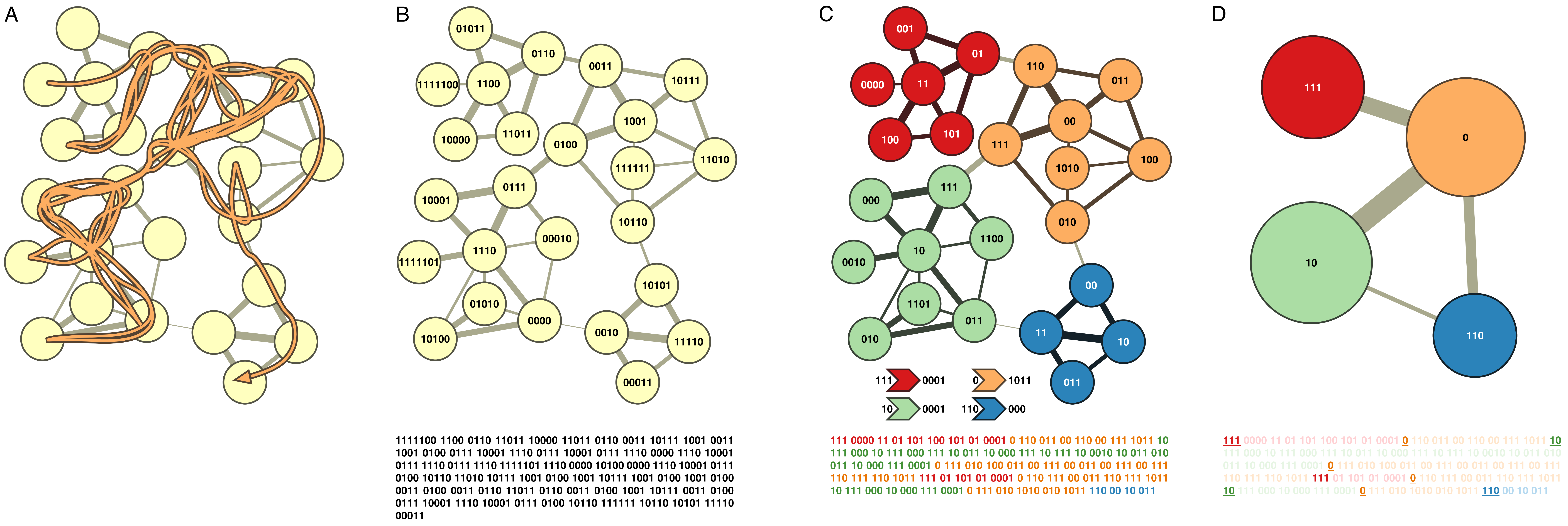
Martin Rosvall*† and Carl T. Bergstrom*‡

Masuda *et al.*, *Physics Reports* (2017)

Random walks are a fundamental concept when studying networks

Maps of random walks on complex networks reveal community structure

Martin Rosvall*† and Carl T. Bergstrom*‡



Rosvall and Bergstrom, *PNAS* (2008)

Graph neural networks

Recall: Node attribute list

Alice	x11	x12
Bob	x21	x22
Carol	x31	x32
⋮	⋮	⋮

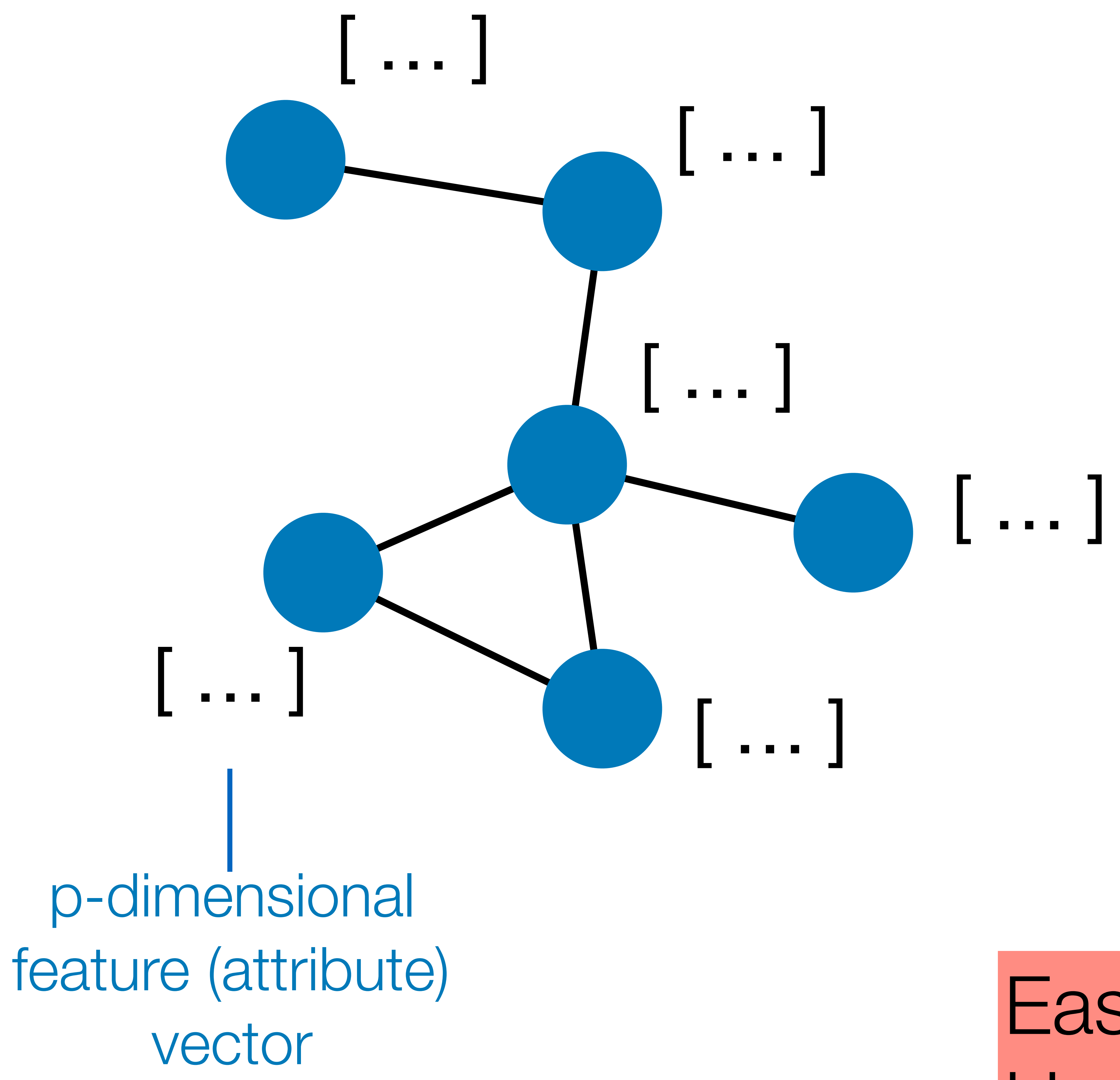
$\underbrace{\qquad\qquad}_{p \text{ features (attributes)}}$

Supervised learning

$$y = f(X)$$

N x *p* matrix of features or predictors each row is an observation, each column is a feature

Graph neural networks



Supervised learning

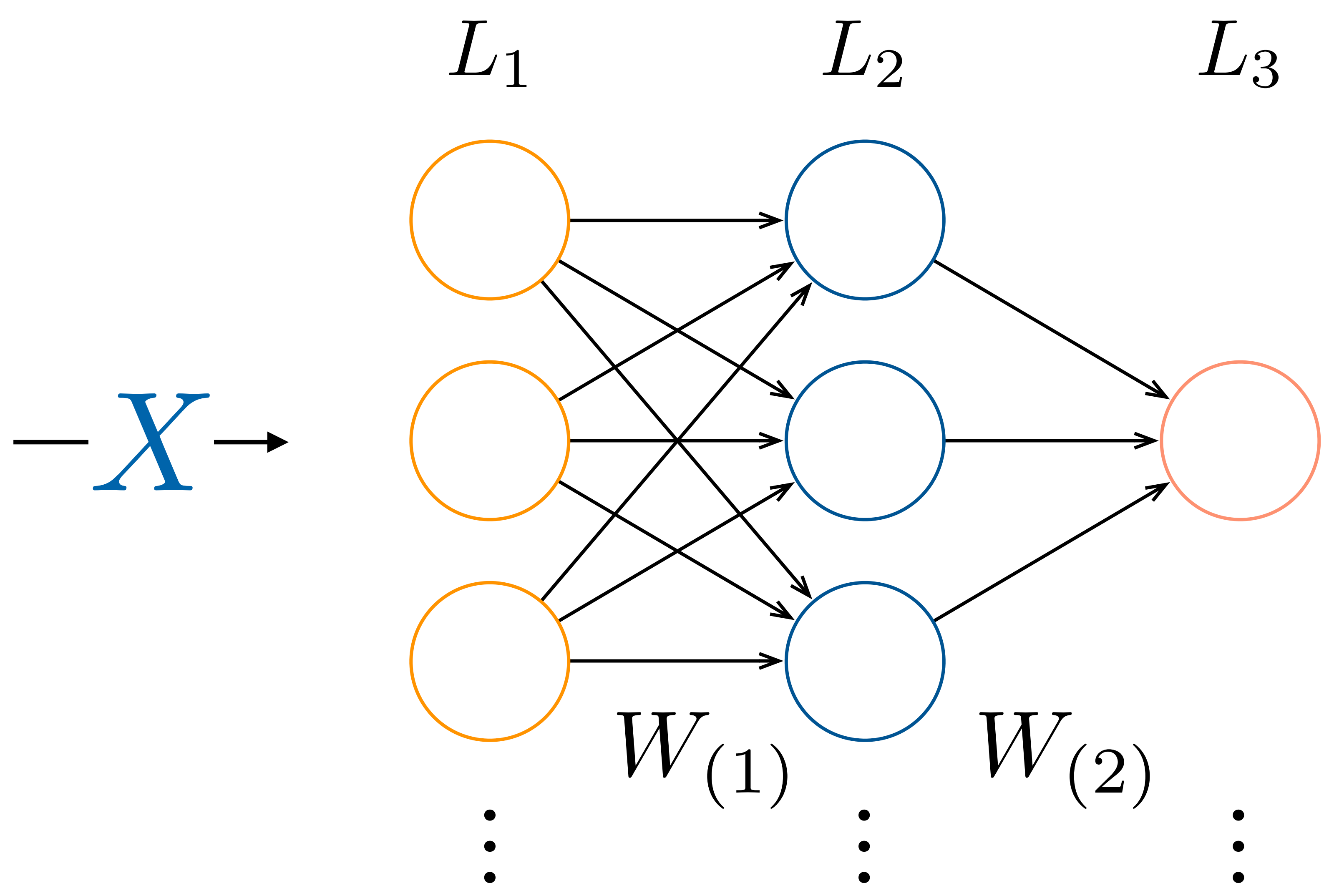
$$y = f(X)$$

N x p matrix of features or predictors each row is an observation, each column is a feature

Easy enough when observations are independent
How to incorporate the network?

neural networks

Idea: propagate your data through the neural network



$$H_{(0)} = X$$

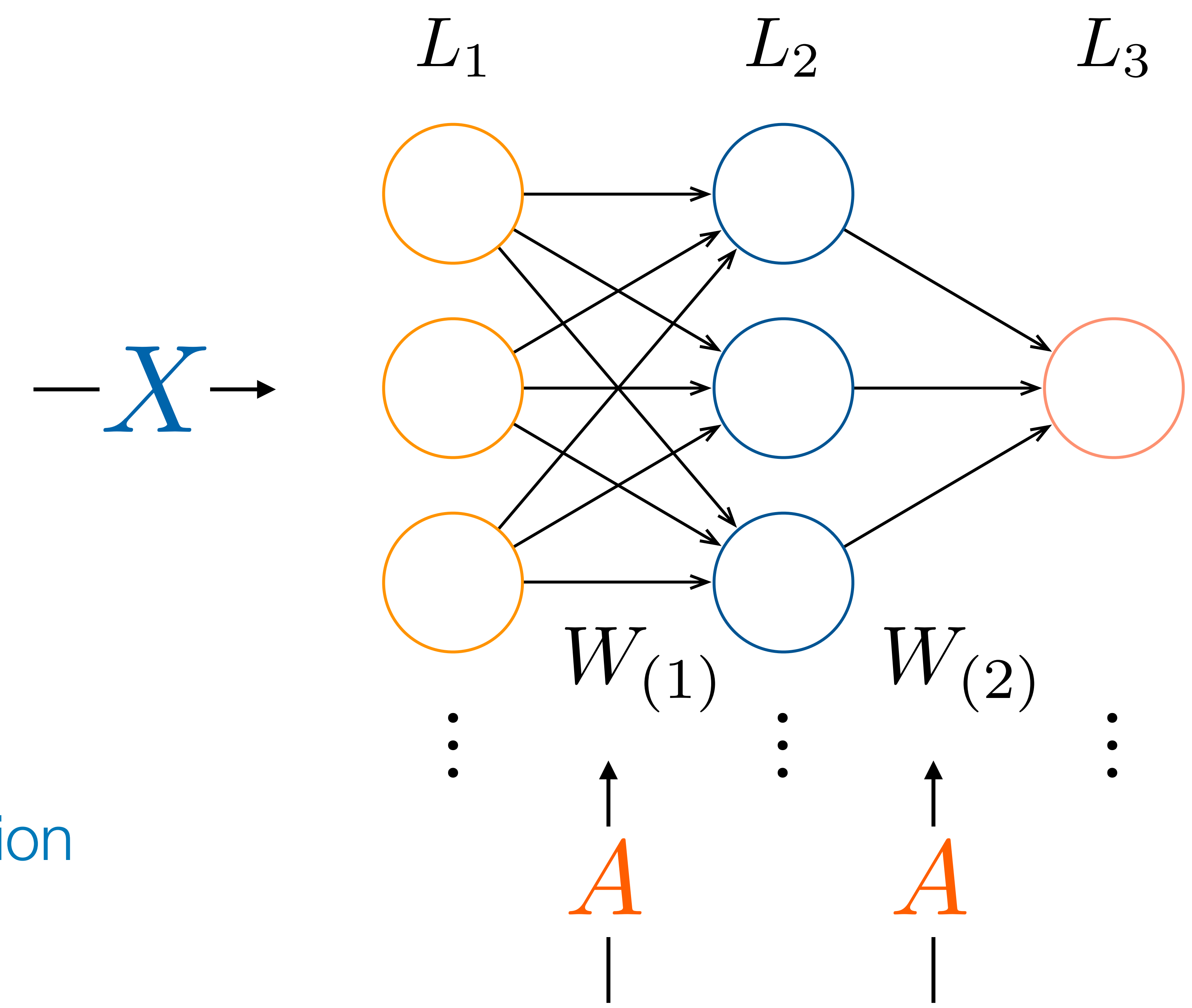
NN: $$H_{(\ell+1)} = \sigma\left(H_{(\ell)} W_{(\ell)}\right)$$

σ —activation function

e.g. Duvenaud *et al.* NIPS (2015)
Kipf *et al.* ICLR (2017)

Graph neural networks

Idea: propagate your data through the neural network, but hit it with the graph at each layer



$$H_{(0)} = X$$

NN: $H_{(\ell+1)} = \sigma\left(H_{(\ell)} W_{(\ell)}\right)$ σ —activation function

GNN: $H_{(\ell+1)} = \sigma\left(\tilde{A} H_{(\ell)} W_{(\ell)}\right)$ $\tilde{A}$ —*preprocessed* adjacency matrix

e.g. Duvenaud *et al.* NIPS (2015)
Kipf *et al.* ICLR (2017)

Graph neural networks

Idea: propagate your data through the neural network, but hit it with the graph at each layer

Applications

- classifying nodes
- predicting links
- comparing networks

$$H_{(0)} = X$$

NN: $H_{(\ell+1)} = \sigma\left(H_{(\ell)} W_{(\ell)}\right)$ σ —activation function

GNN: $H_{(\ell+1)} = \sigma\left(\tilde{A} H_{(\ell)} W_{(\ell)}\right)$ $\tilde{A}$ —*preprocessed* adjacency matrix

e.g. Duvenaud *et al.* NIPS (2015)
Kipf *et al.* ICLR (2017)

Designing visualizations

Visualization ⊂ Communication

Visualizations are one tool to tackle the larger problem of communicating your results

Designing visualizations

Which kind of door handle is better?

Designing visualizations

Which kind of door handle is better?

Better? Easier to open!

Designing visualizations

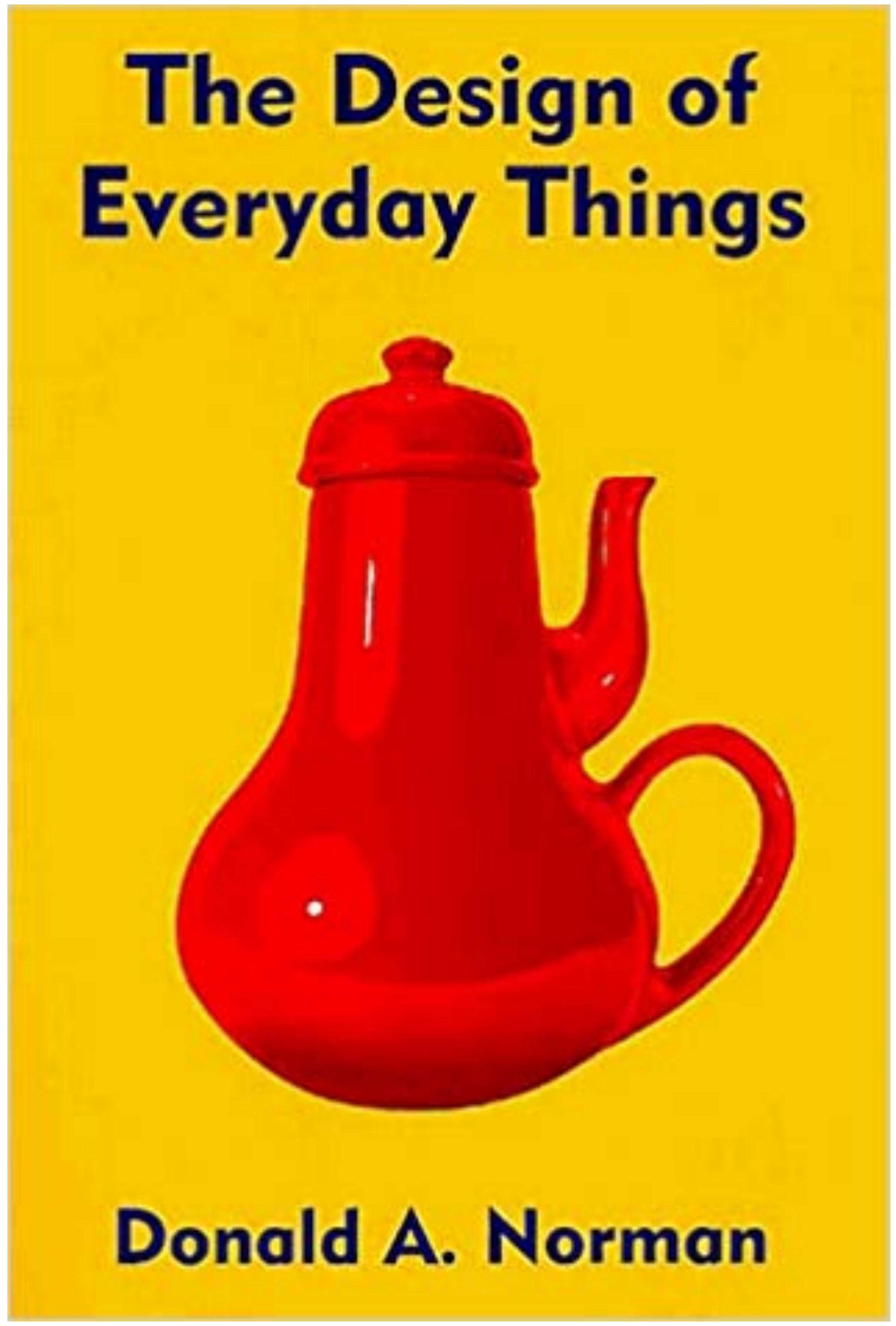
"Design is how it works"
–Steve Jobs

Designing visualizations

"Design is how it works"
–Steve Jobs

Designing visualizations

"Design is how it works"
–Steve Jobs



Designing visualizations

Which kind of door handle is better?

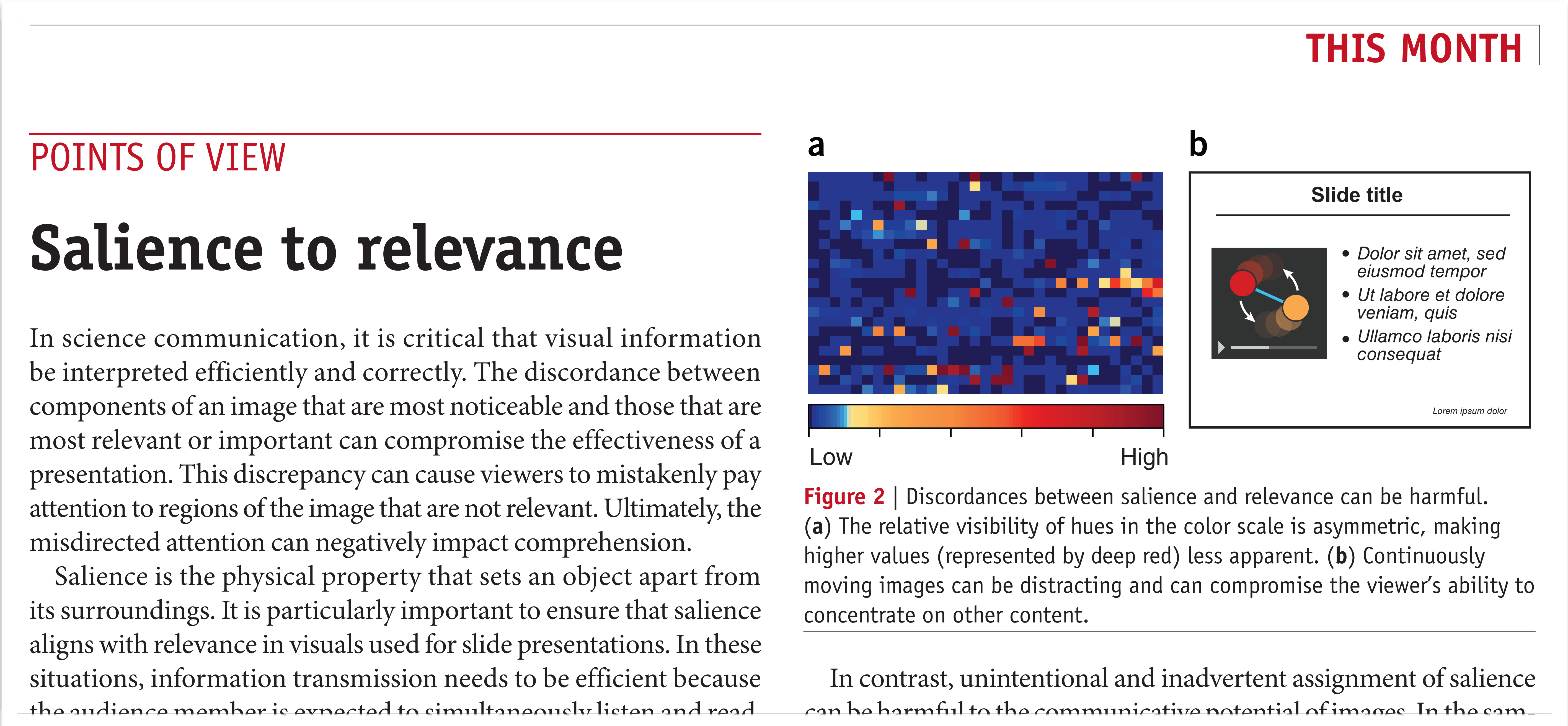
Better? Easier to open!

Visualizations: better = easier to understand

Designing visualizations

- Know your <u>message</u>
- Know your <u>medium</u>
- Know your <u>audience</u>
- Account for strengths and weaknesses of <u>human perception</u>
- Keep it <u>simple</u>

Great info: series of articles published in Nature Methods during 2010-2015 called "Points of View"



THIS MONTH

POINTS OF VIEW

Salience to relevance

In science communication, it is critical that visual information be interpreted efficiently and correctly. The discordance between components of an image that are most noticeable and those that are most relevant or important can compromise the effectiveness of a presentation. This discrepancy can cause viewers to mistakenly pay attention to regions of the image that are not relevant. Ultimately, the misdirected attention can negatively impact comprehension.

Salience is the physical property that sets an object apart from its surroundings. It is particularly important to ensure that salience aligns with relevance in visuals used for slide presentations. In these situations, information transmission needs to be efficient because

Figure 2 | Discordances between salience and relevance can be harmful. (**a**) The relative visibility of hues in the color scale is asymmetric, making higher values (represented by deep red) less apparent. (**b**) Continuously moving images can be distracting and can compromise the viewer's ability to concentrate on other content.

In contrast, unintentional and inadvertent assignment of salience

http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

The challenge

Six months of work

↓

~ 1000 plots

↓

5-10 figures

The challenge

Six months of work

↓

~ 1000 plots

↓

5-10 figures

Hard to remember
being a **beginner**

Know your message

A figure/visualization has a **goal**: what do you want the reader to learn?

Know your message

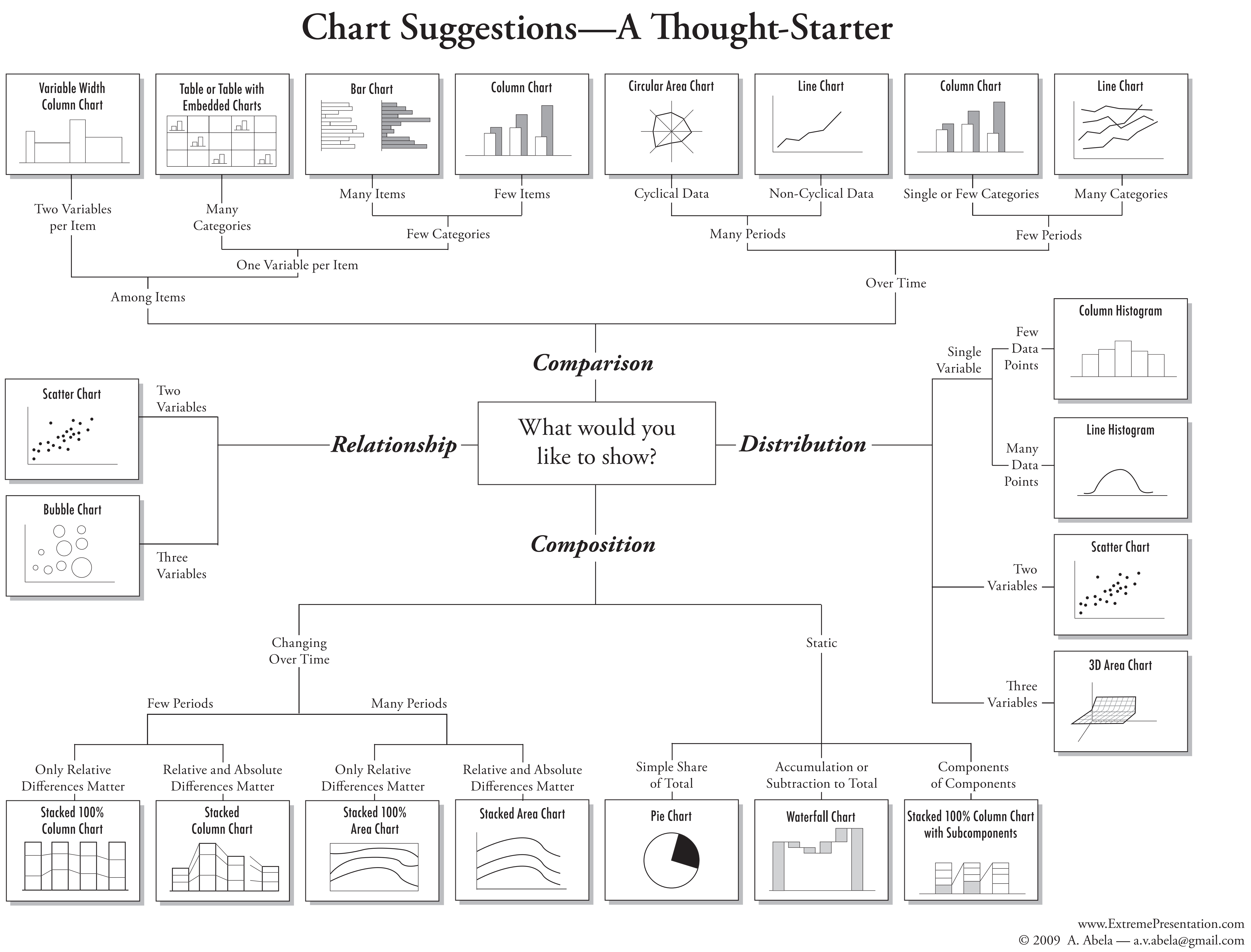
A figure/visualization has a **goal**: what do you want the reader to learn?

Summary sentence:

> "Cancer deaths are down, but mostly due to decreased smoking rates."
>
> "Algorithm B converges faster than A."
>
> "Bats spread Ebola, not rodents."
>
> "The rate of text messages increased after approximately day 45."

Build your figure(s) with this goal in mind.

Use your summary sentence to guide the kind of visualization(s) you use

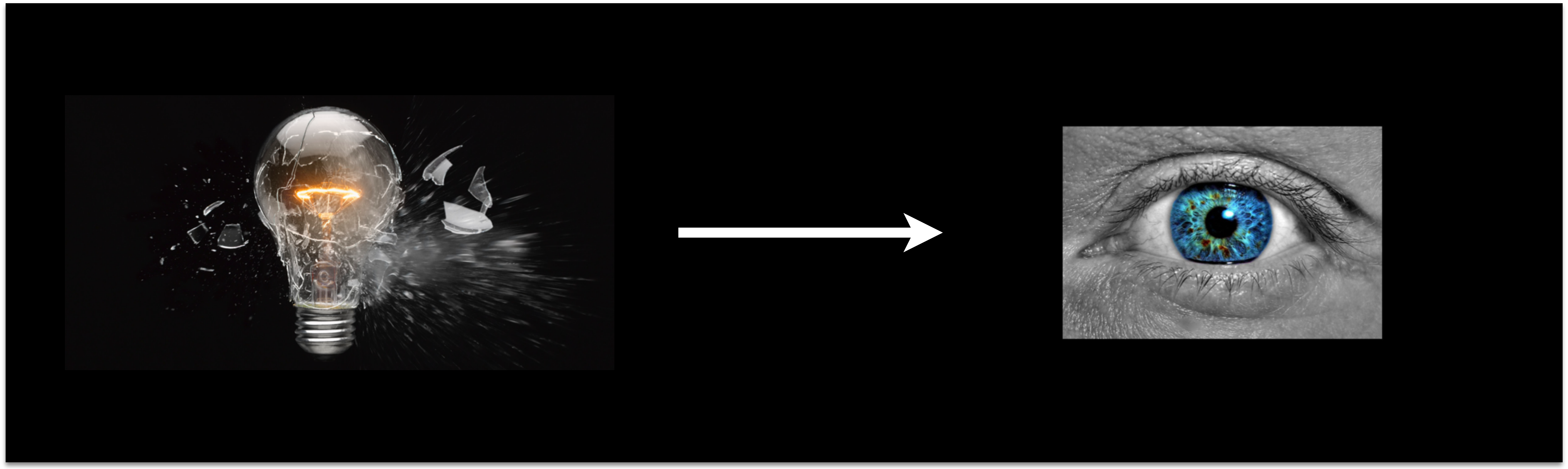


http://extremepresentation.com

Zelazny, Say it with Charts, 2001

Know your medium

Print? Web? Slides?



Know your audience

Human perception

Parsing a figure or visualization requires performing **visual tasks**
Humans are better at some tasks and worse at others

	Aspect to compare
easiest	Positions on a common scale
	Positions on the same but nonaligned scales
	Lengths
	Angles, slopes
	Area
	Volume, color saturation
hardest	Color hue

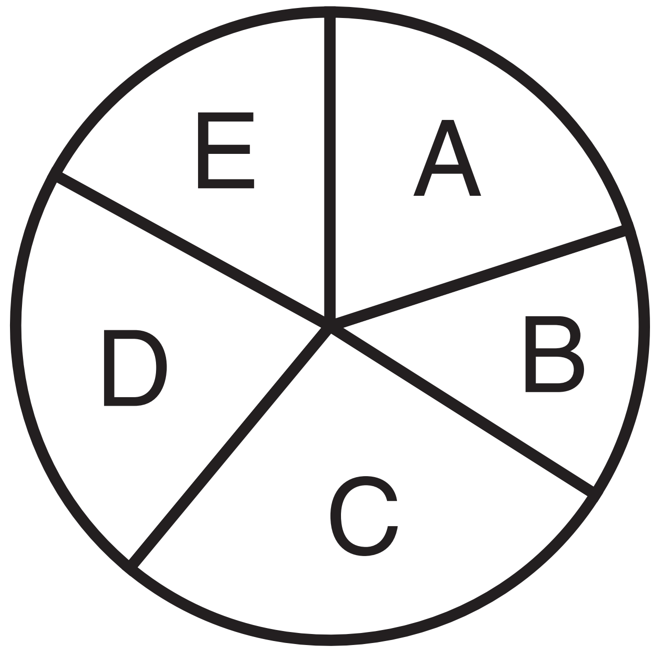
Graphical Perception and Graphical Methods for Analyzing Scientific Data

William S. Cleveland and Robert McGill

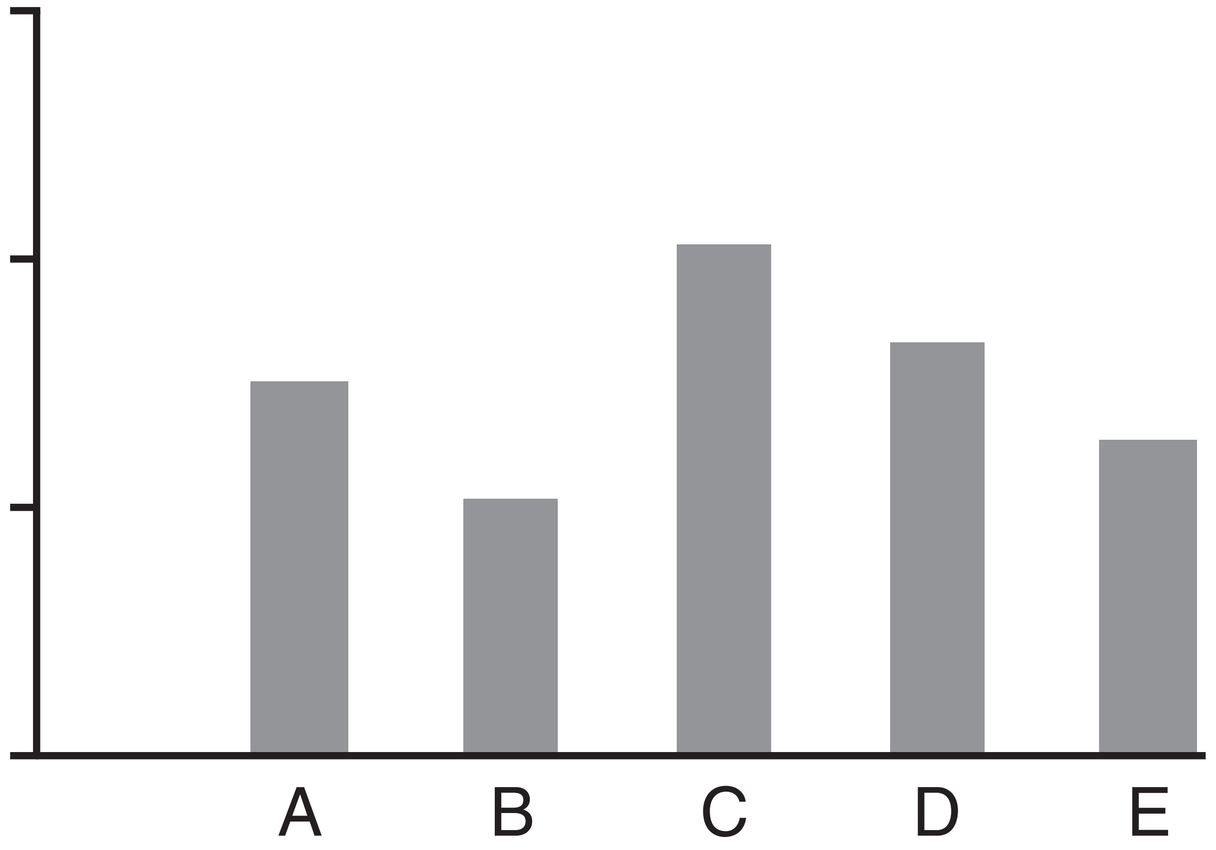
Science (1985)

Human perception

Example: Comparing **areas** vs. **lengths**

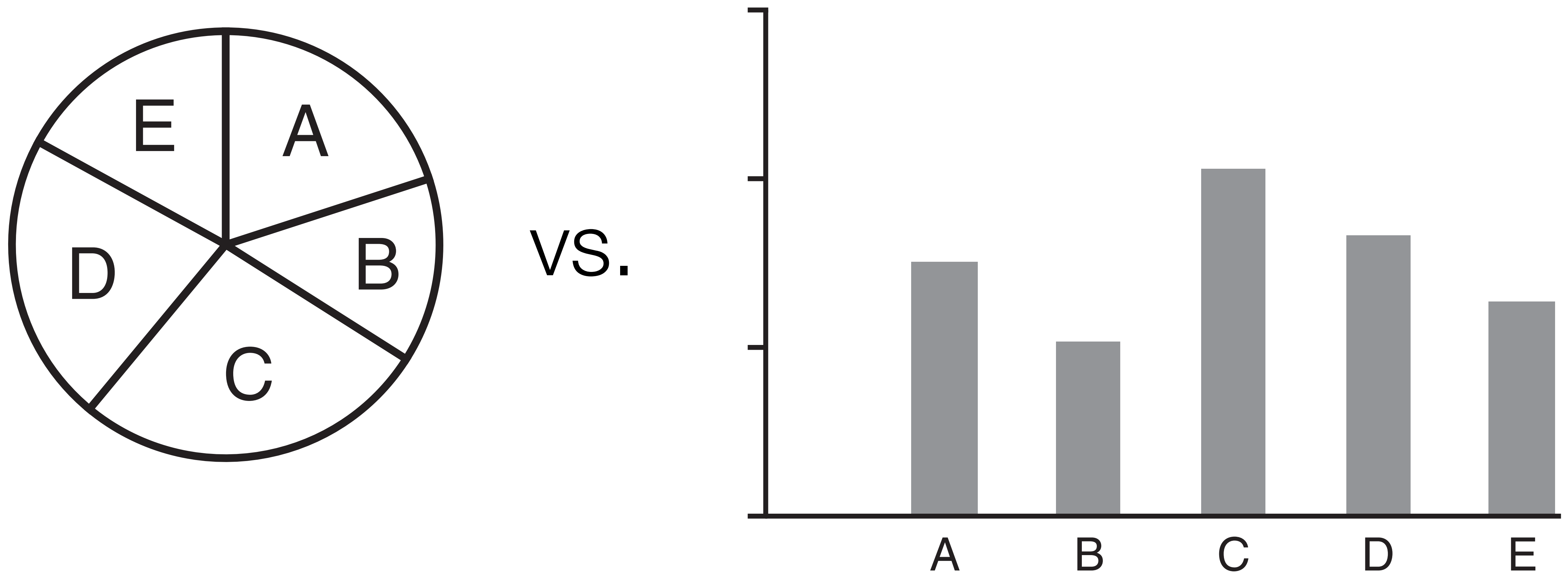


vs.



Human perception

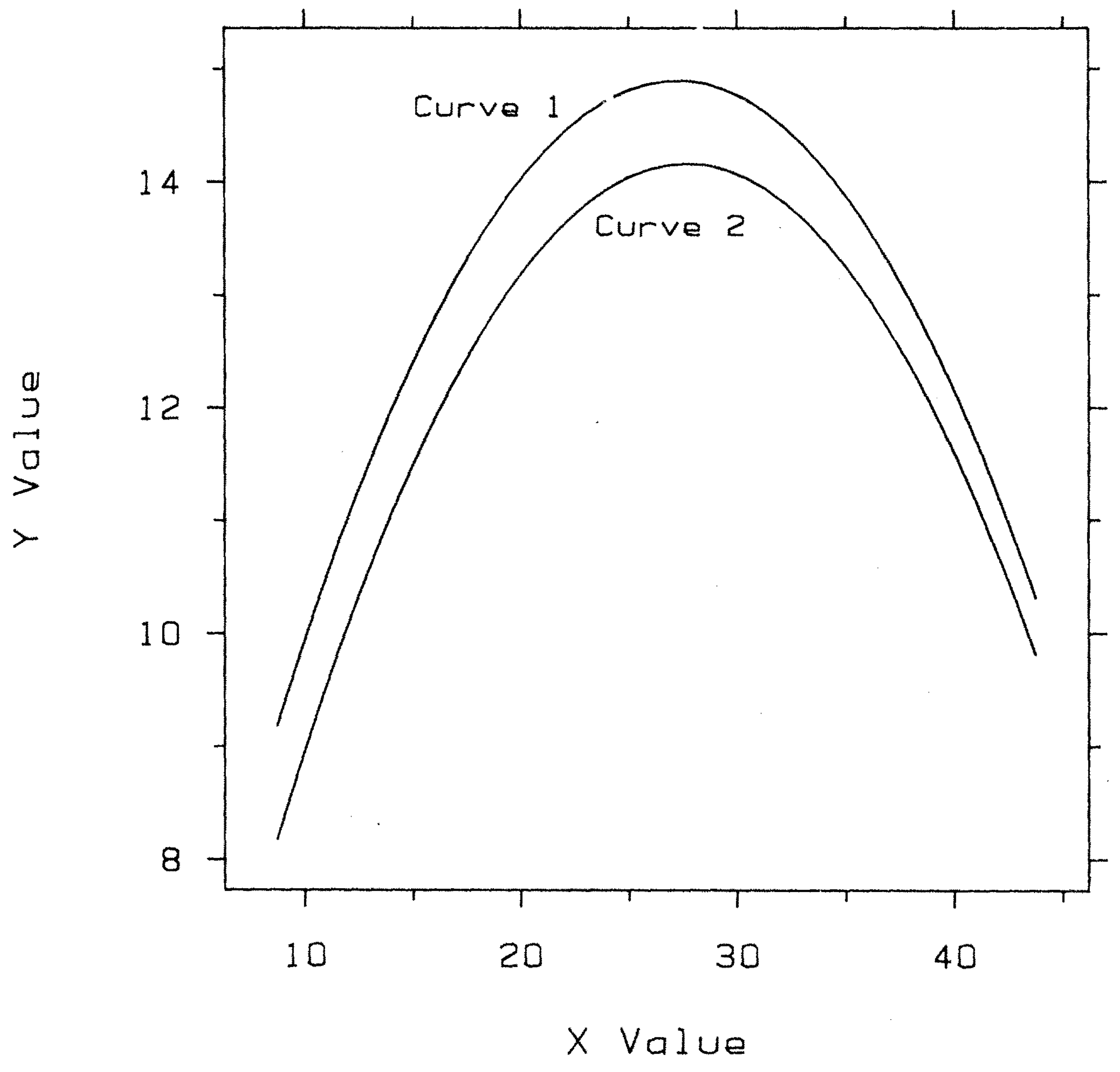
Example: Comparing **areas** vs. **lengths**



Avoid Pie Charts!

Human perception

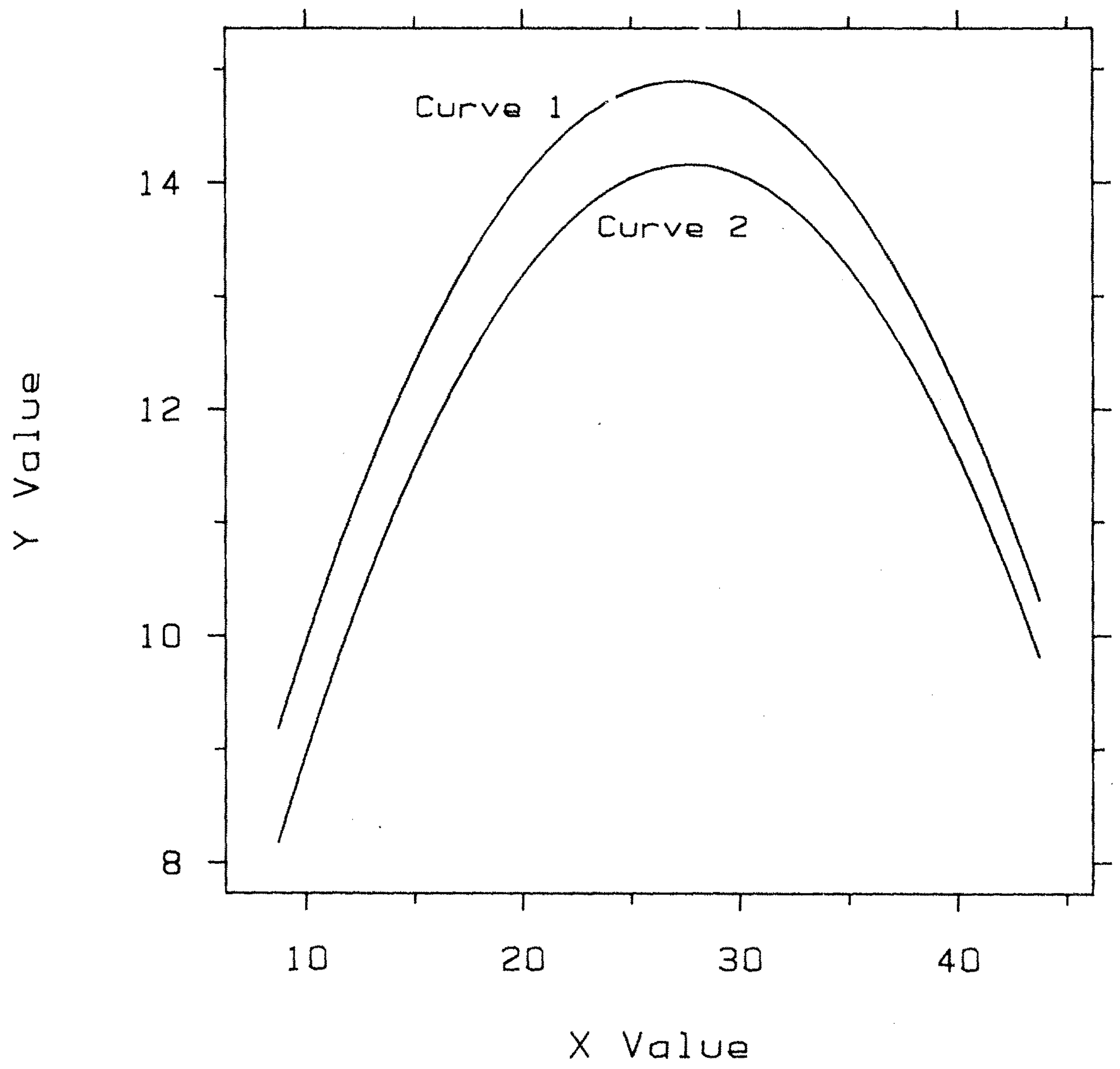
Perceptual biases plague even basic graphics

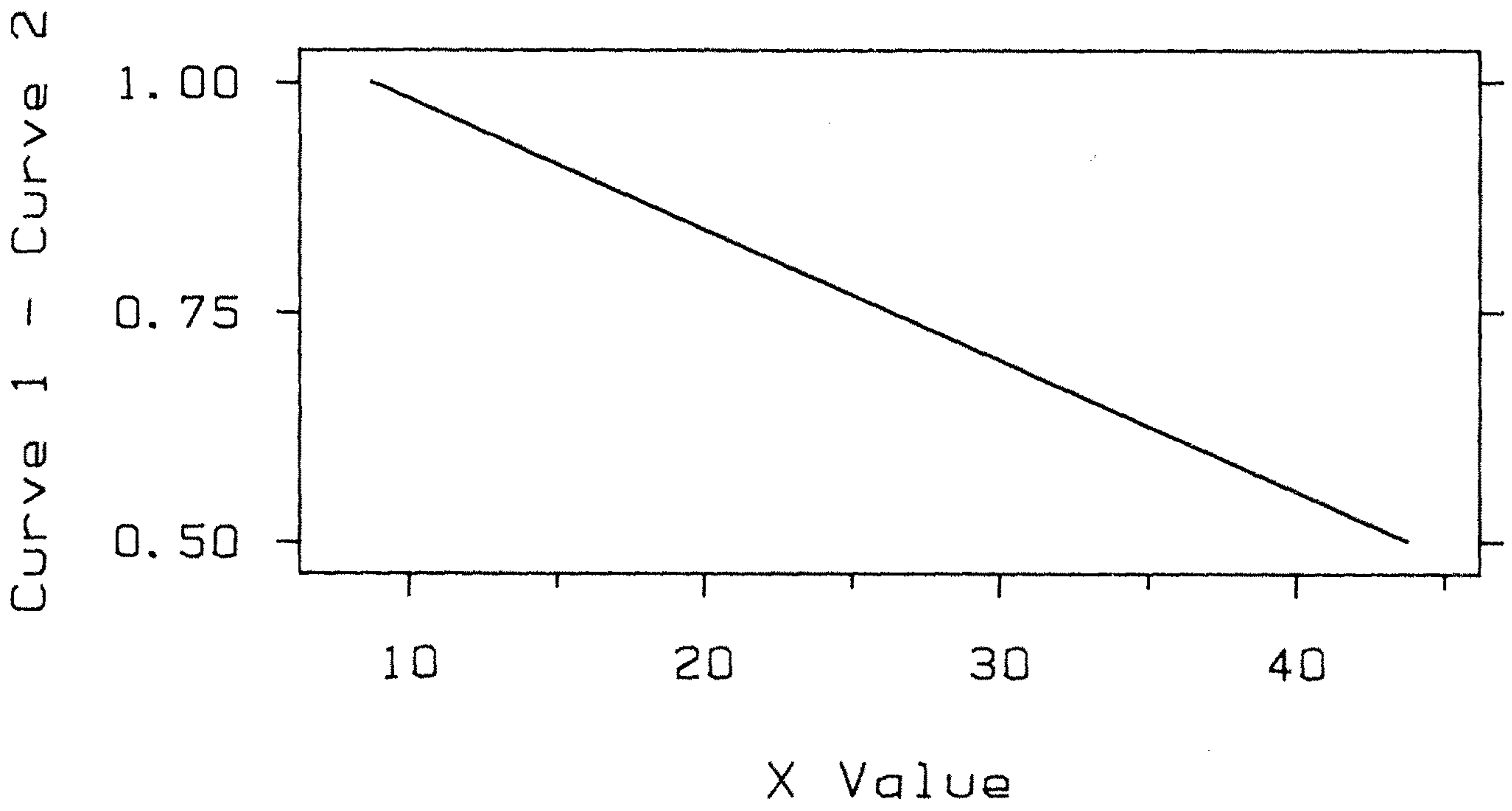


Cleveland & McGill (1985)

Human perception

Perceptual biases plague even basic graphics





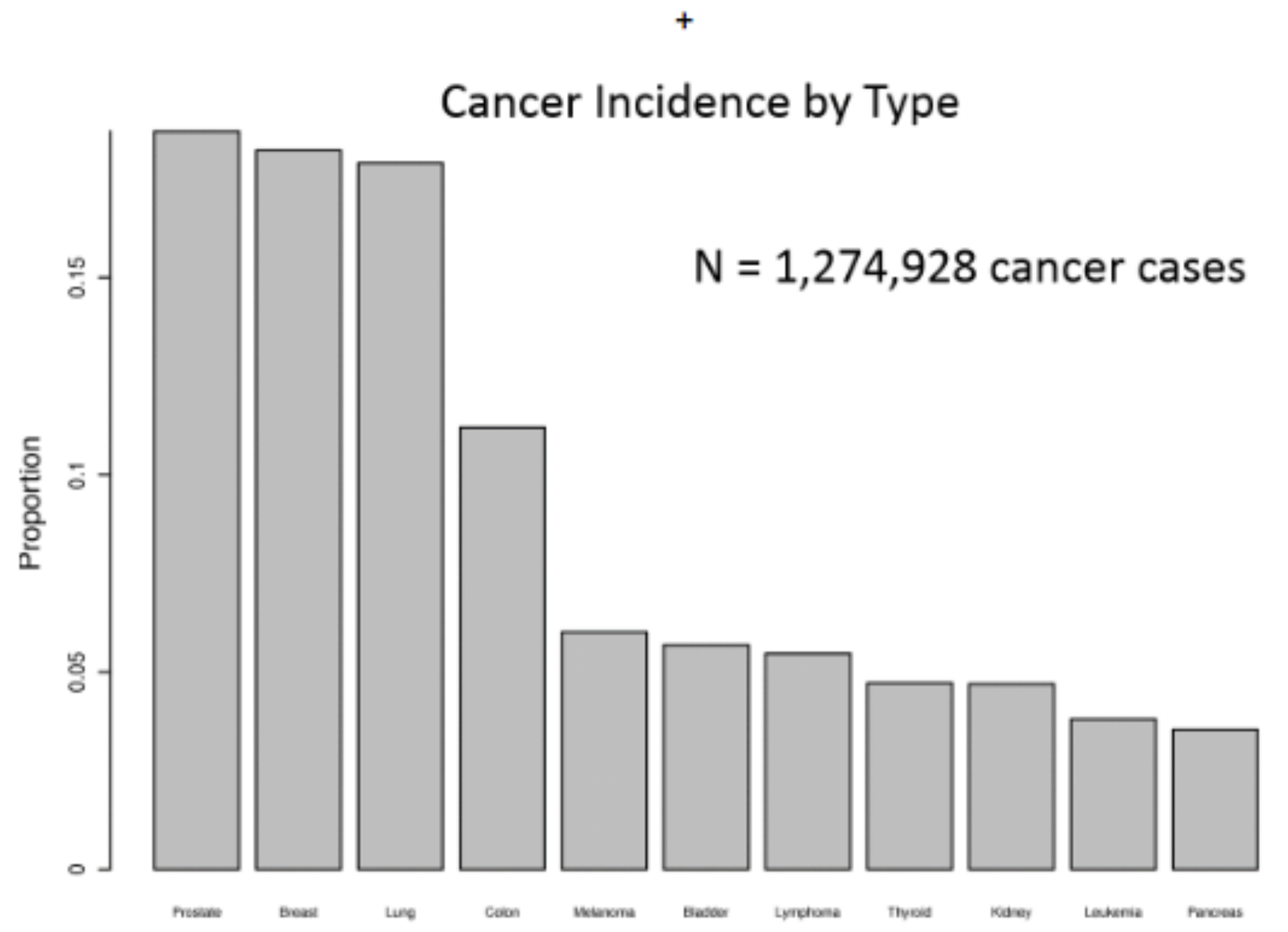
Cleveland & McGill (1985)

Getting it right takes time

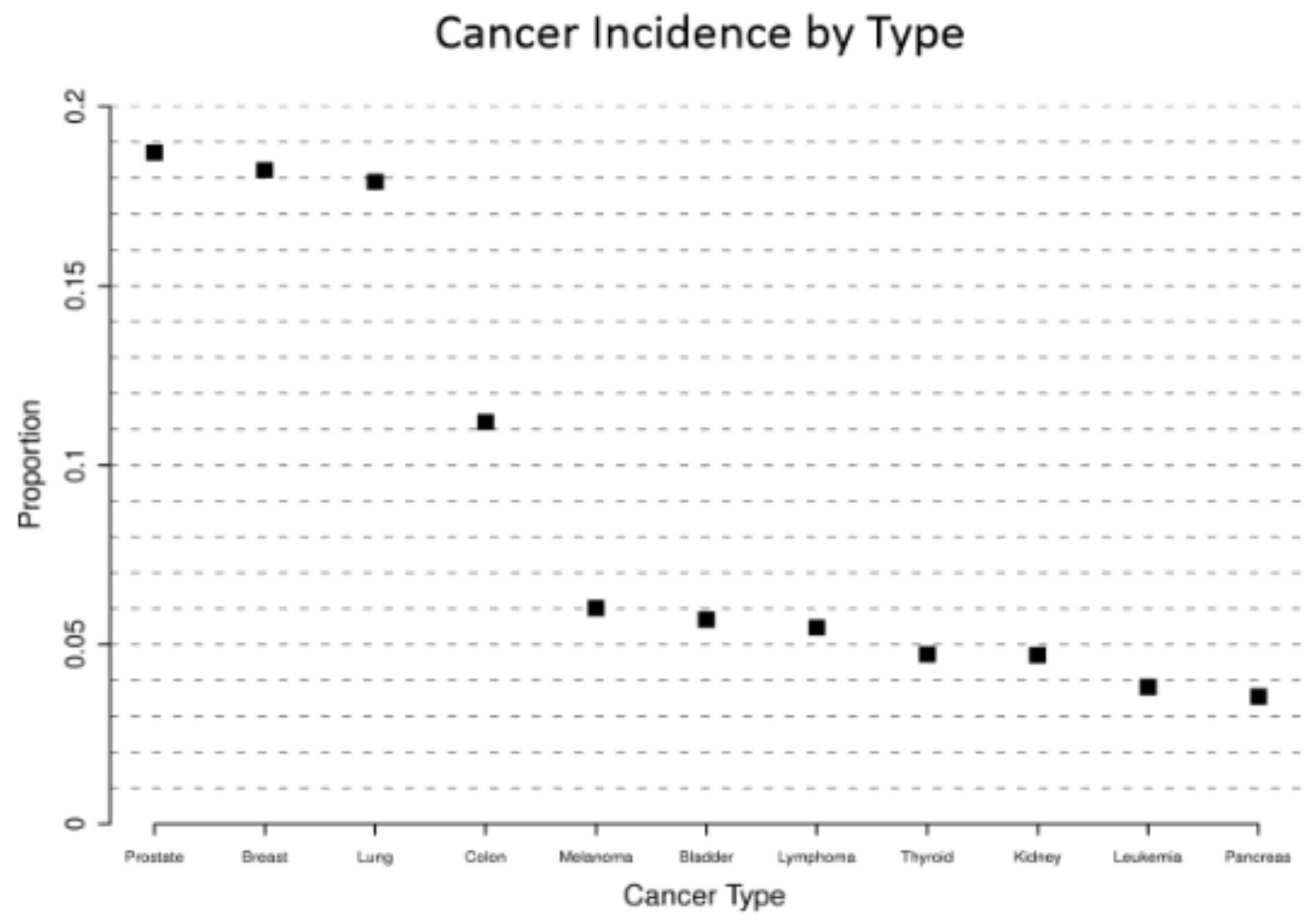
Iterate!

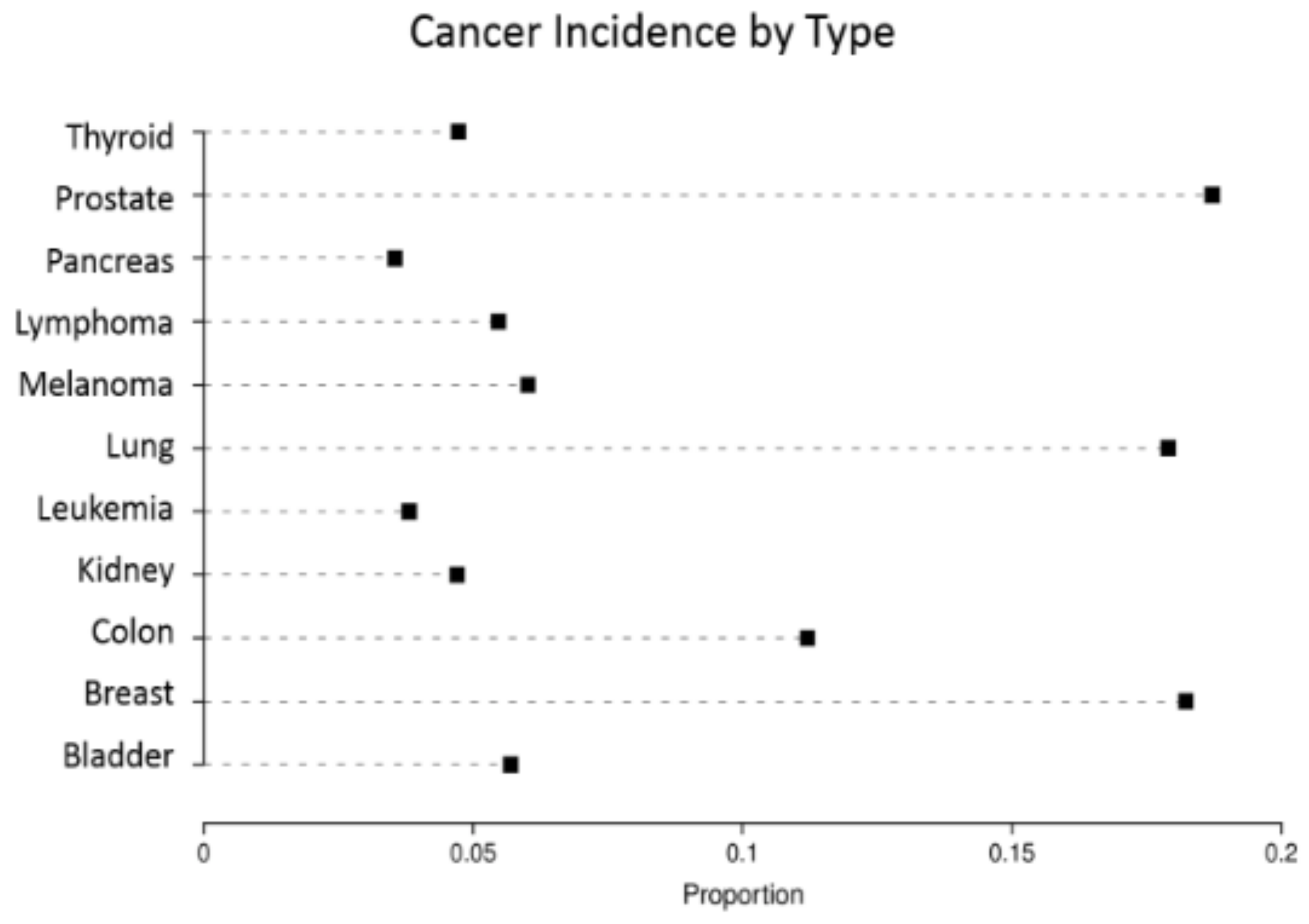
Readability is the most important goal!

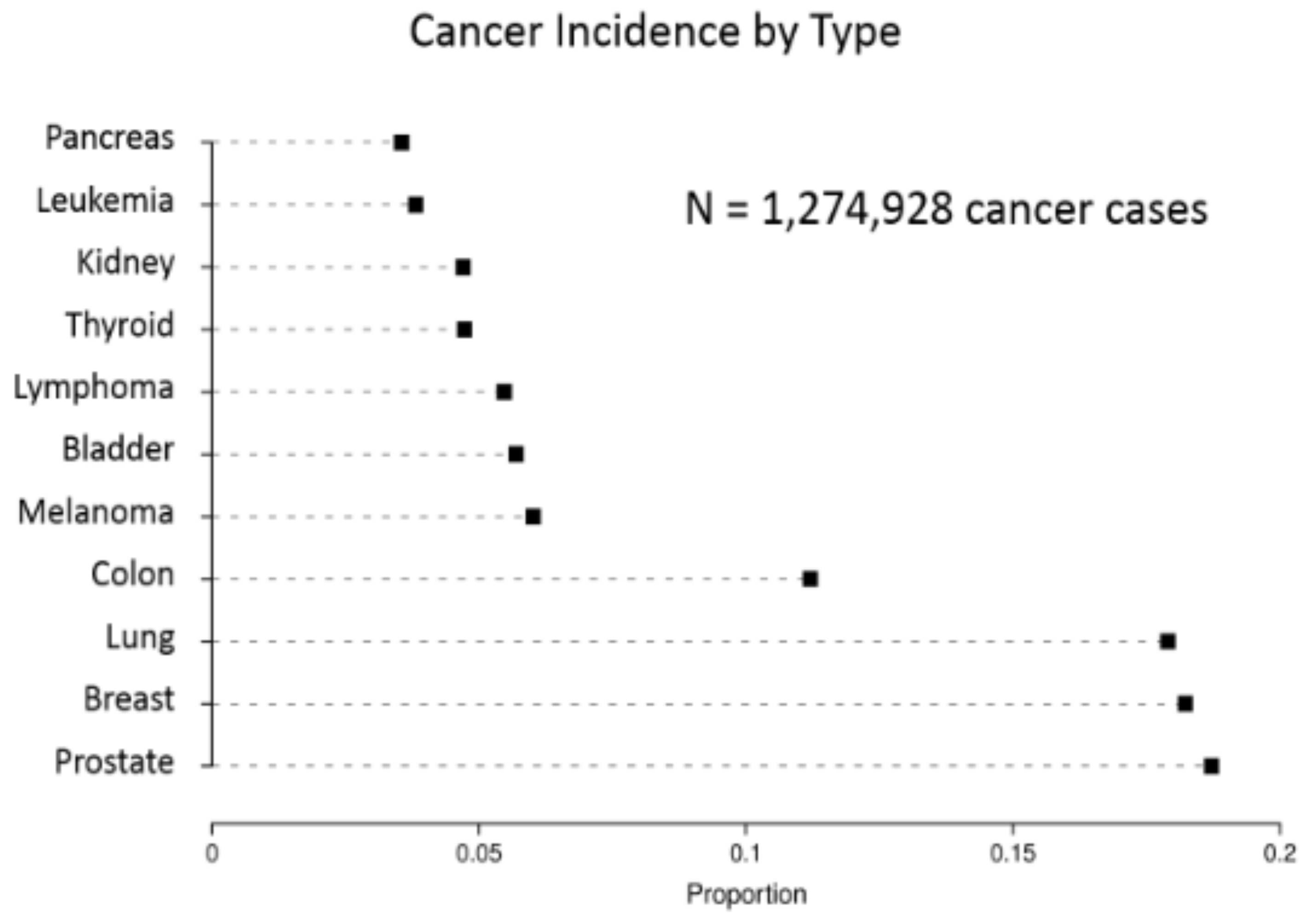
Ver. 1



Ver. 2







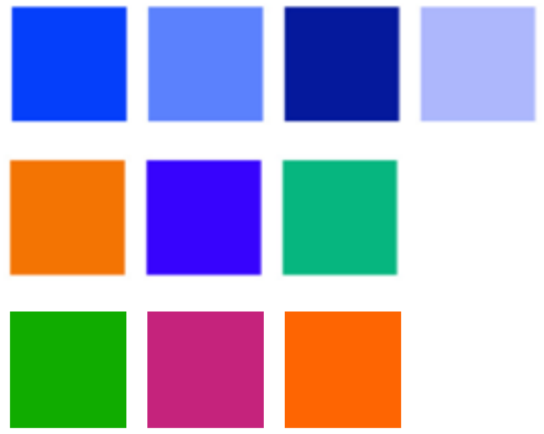
Ver. 3

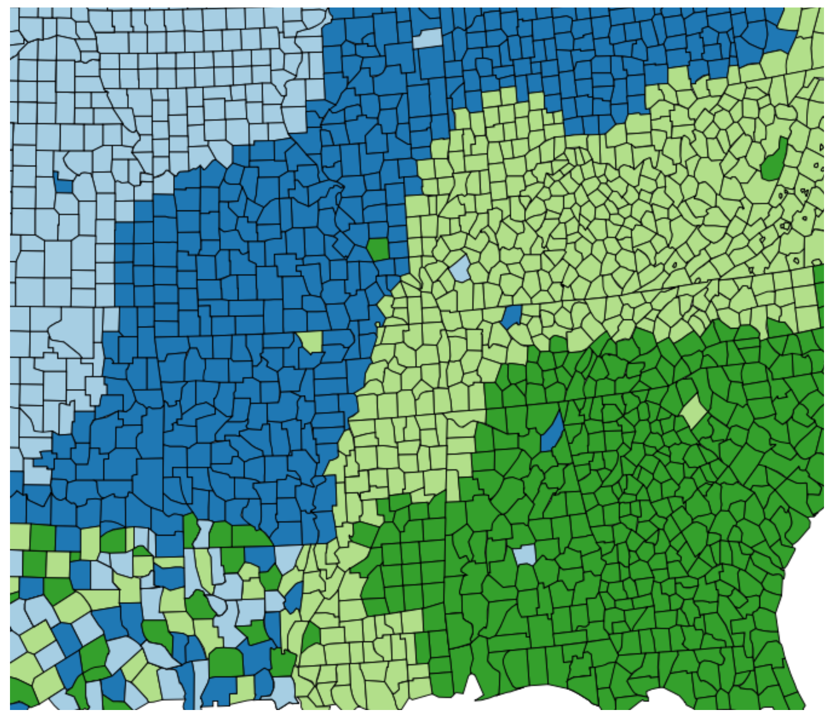
Ver. 4


color

Color schemes

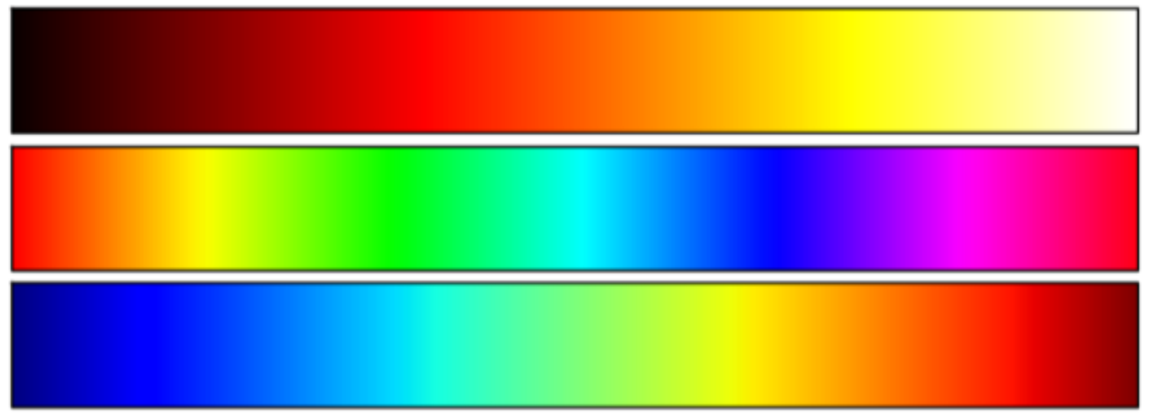
discrete palette

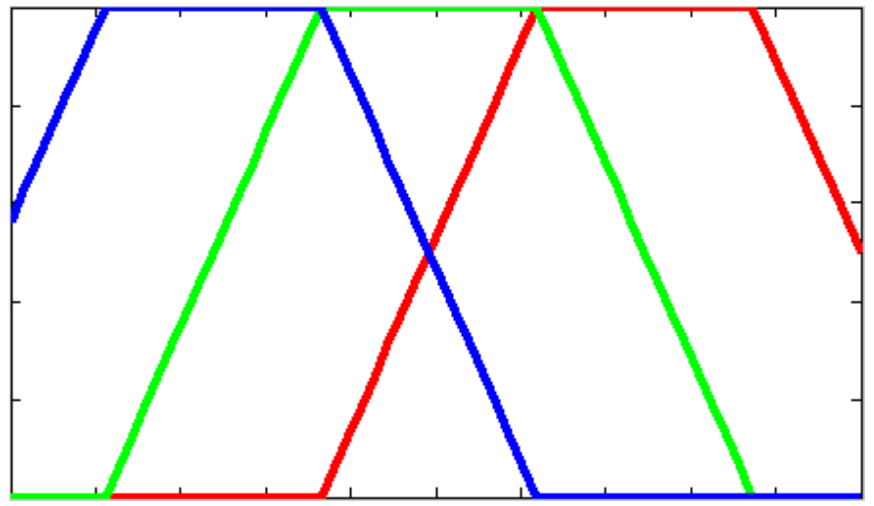


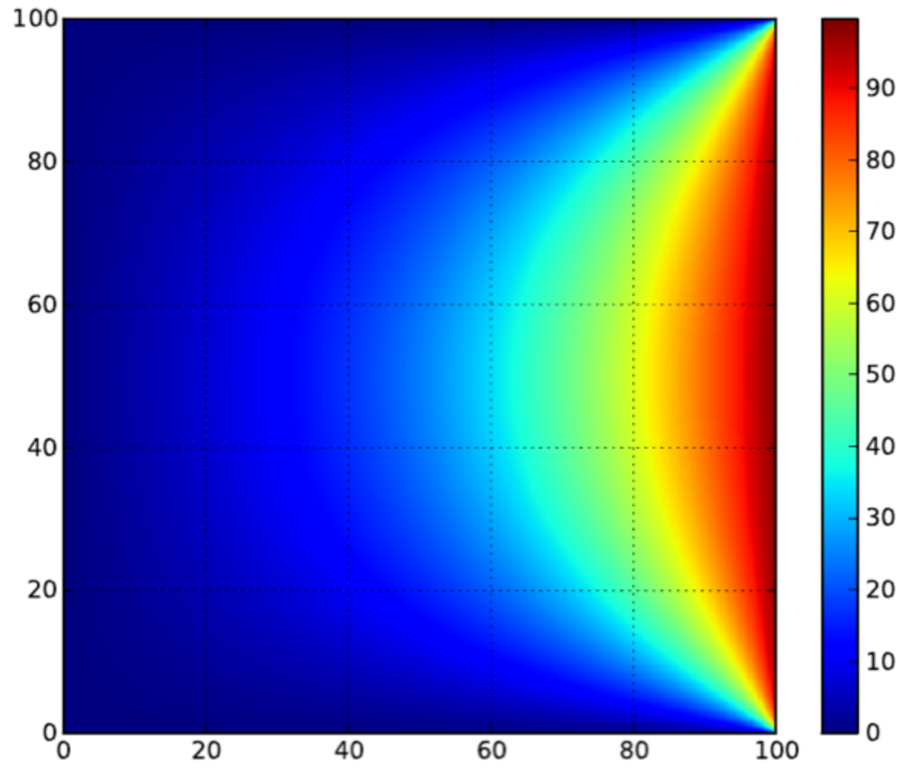


Colormaps

continuous (function)





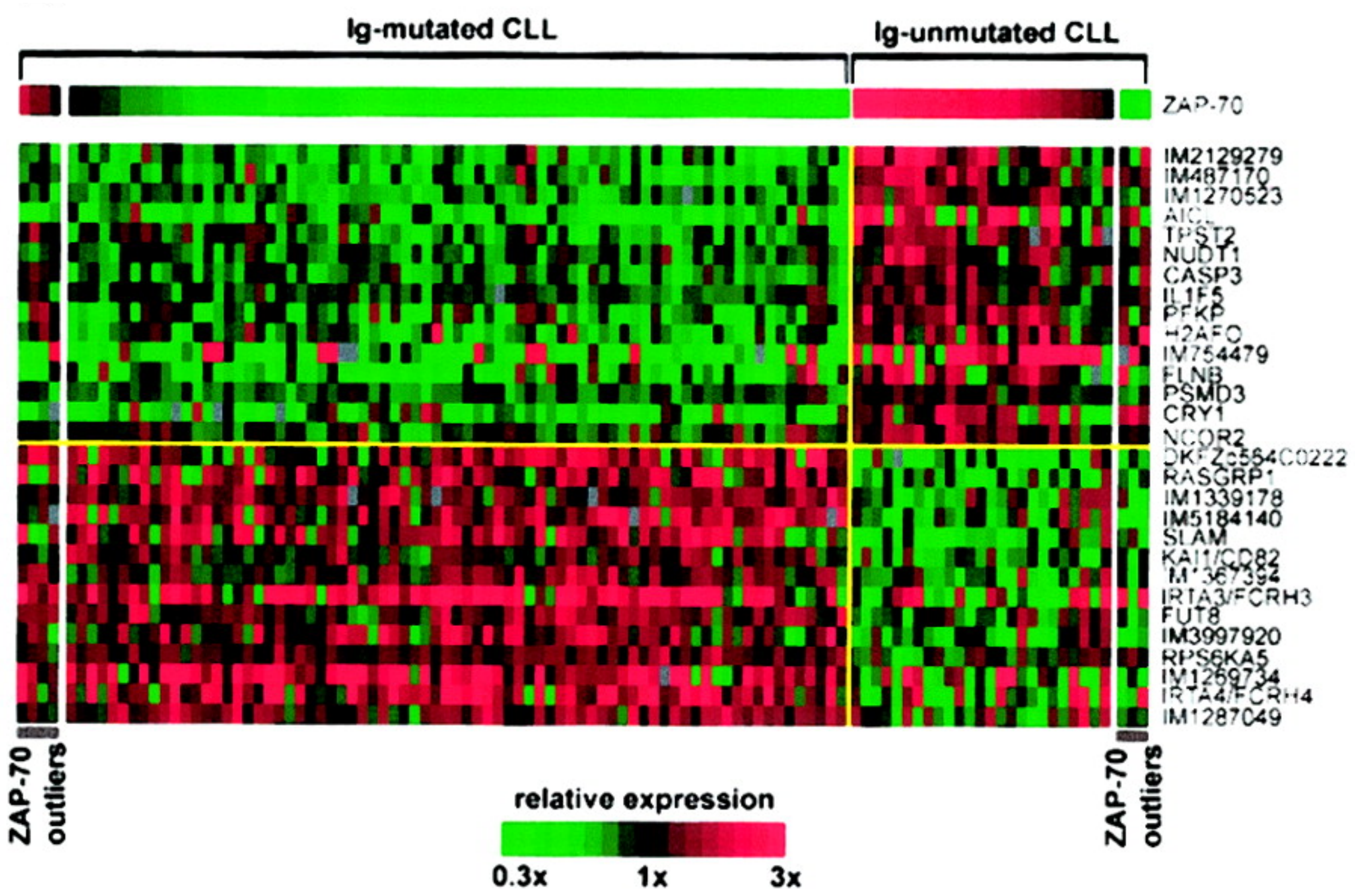


Good idea to lean on existing, evidence-based palettes (*Tableau 10*) and maps (*Viridis*)

tableau.com: How we designed the new color palettes

SciPy 2015: A Better Default Colormap for Matplotlib [YouTube]

Color blindness: the **eye** is a noisy channel



Red/Green blindness is most common → avoid it

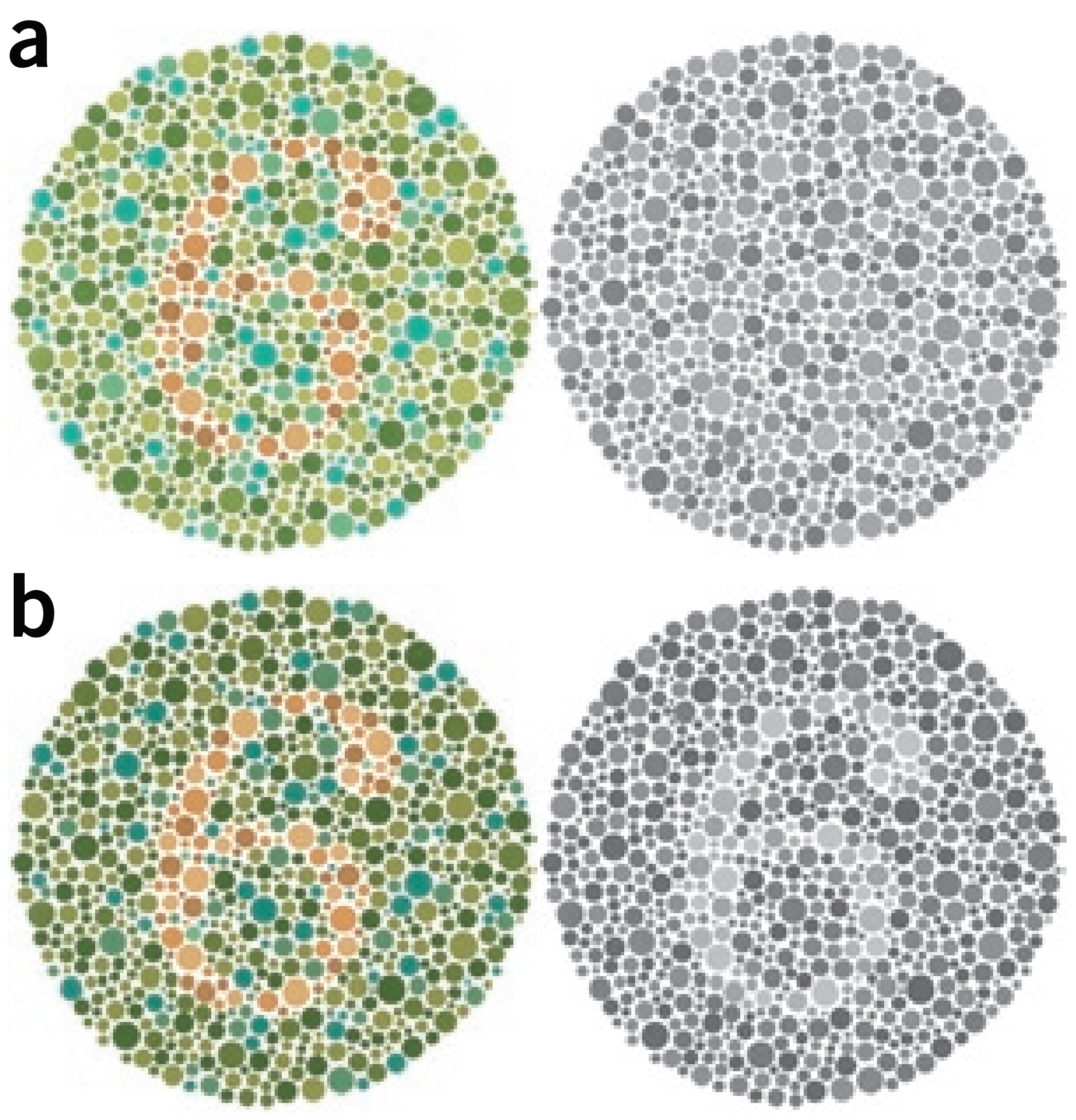
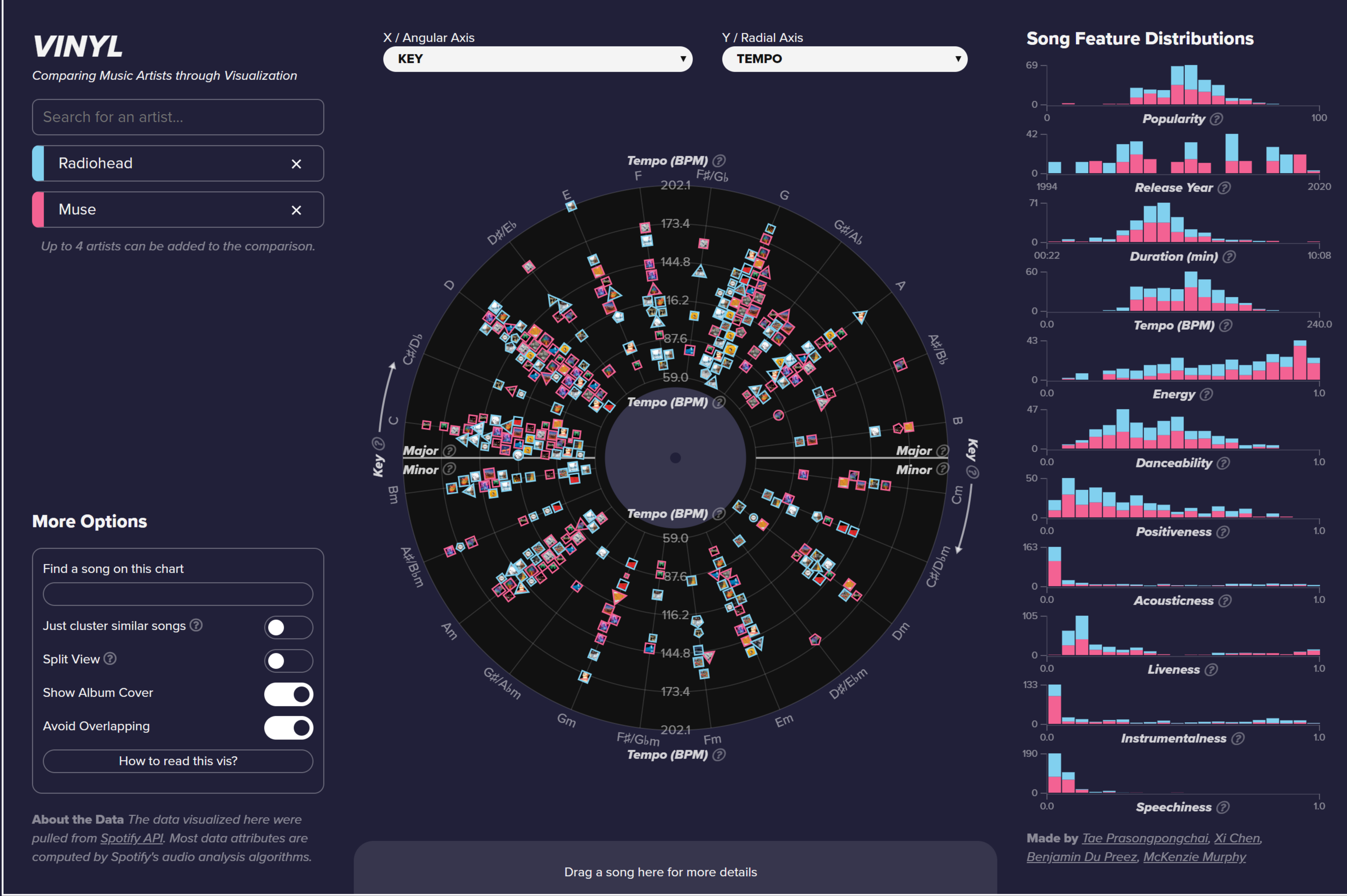


Figure 1 | Ishihara color-vision test plate. (**a**) Viewers with normal color vision should see the numeral ‘6’. (**b**) Changing lightness of background improves contrast.

Don't rely completely on color— tweak hue/saturation to improve contrast

Put it all together:

Keep it simple?



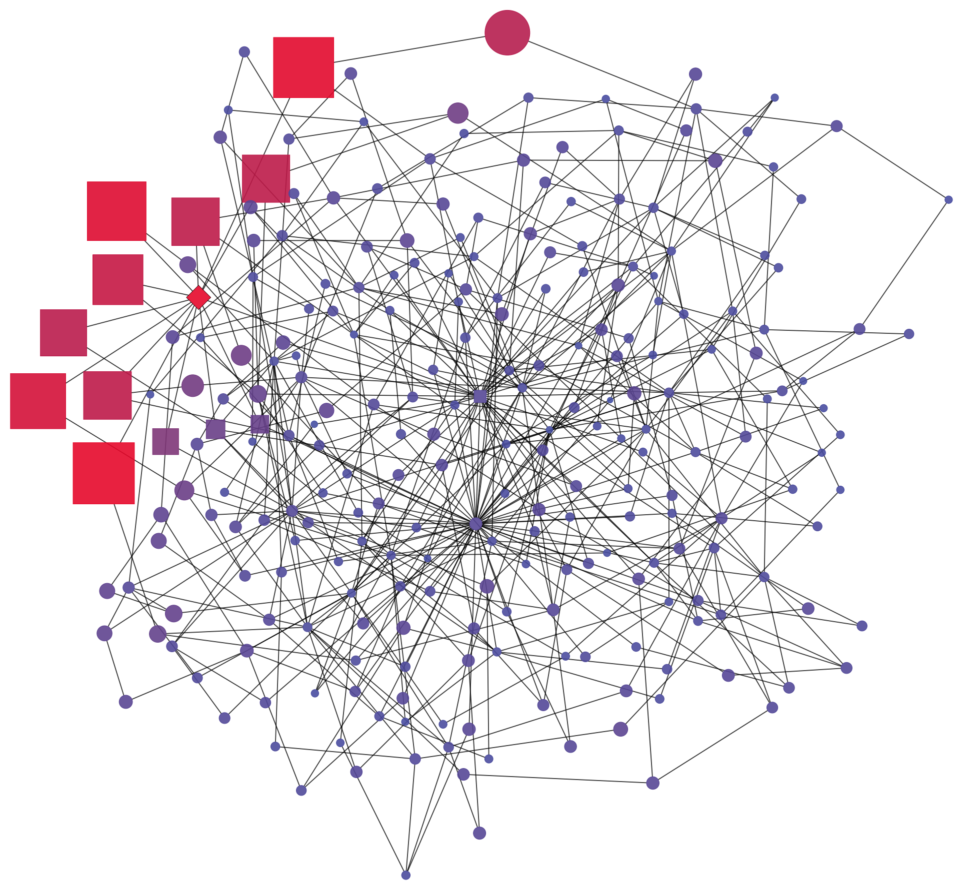
kenziemurphy.github.io/vinyl/

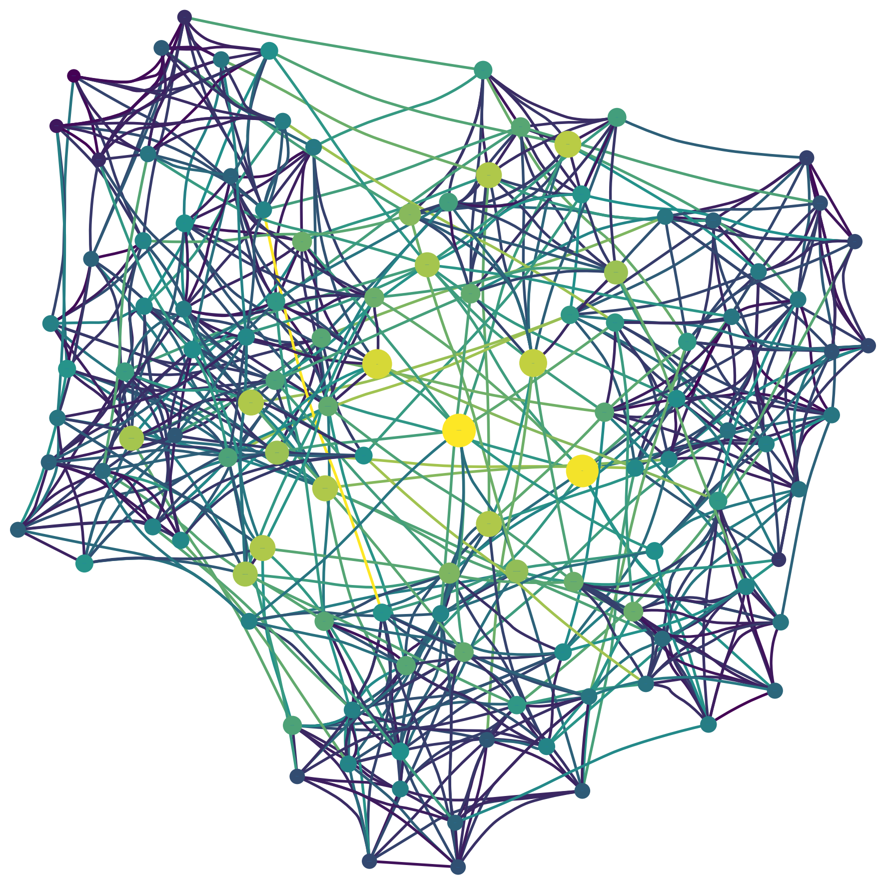
Network Visualizations

Network visualizations

Before we begin, a tough question: is a network visualization appropriate?

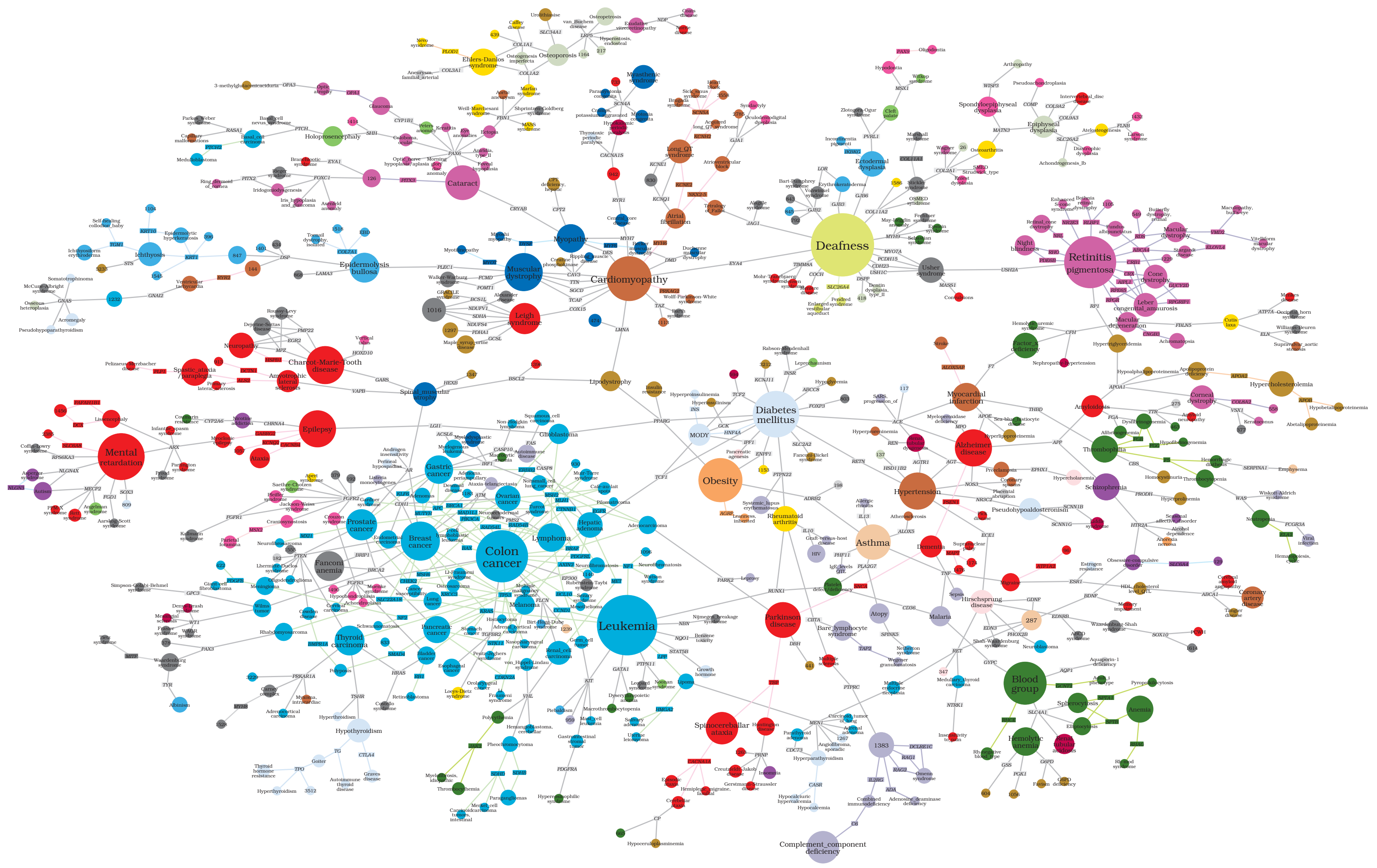
Ghoniem *et al*. InfoVis'04 (2004)
Foucault Welles & Meirelles (2015)
Foucault Welles & Xu (2018)





Alternative approaches

Bagrow *et al.* EPL (2008)
Bagrow & Bollt (2019)
Schulman *et al.* (2011)



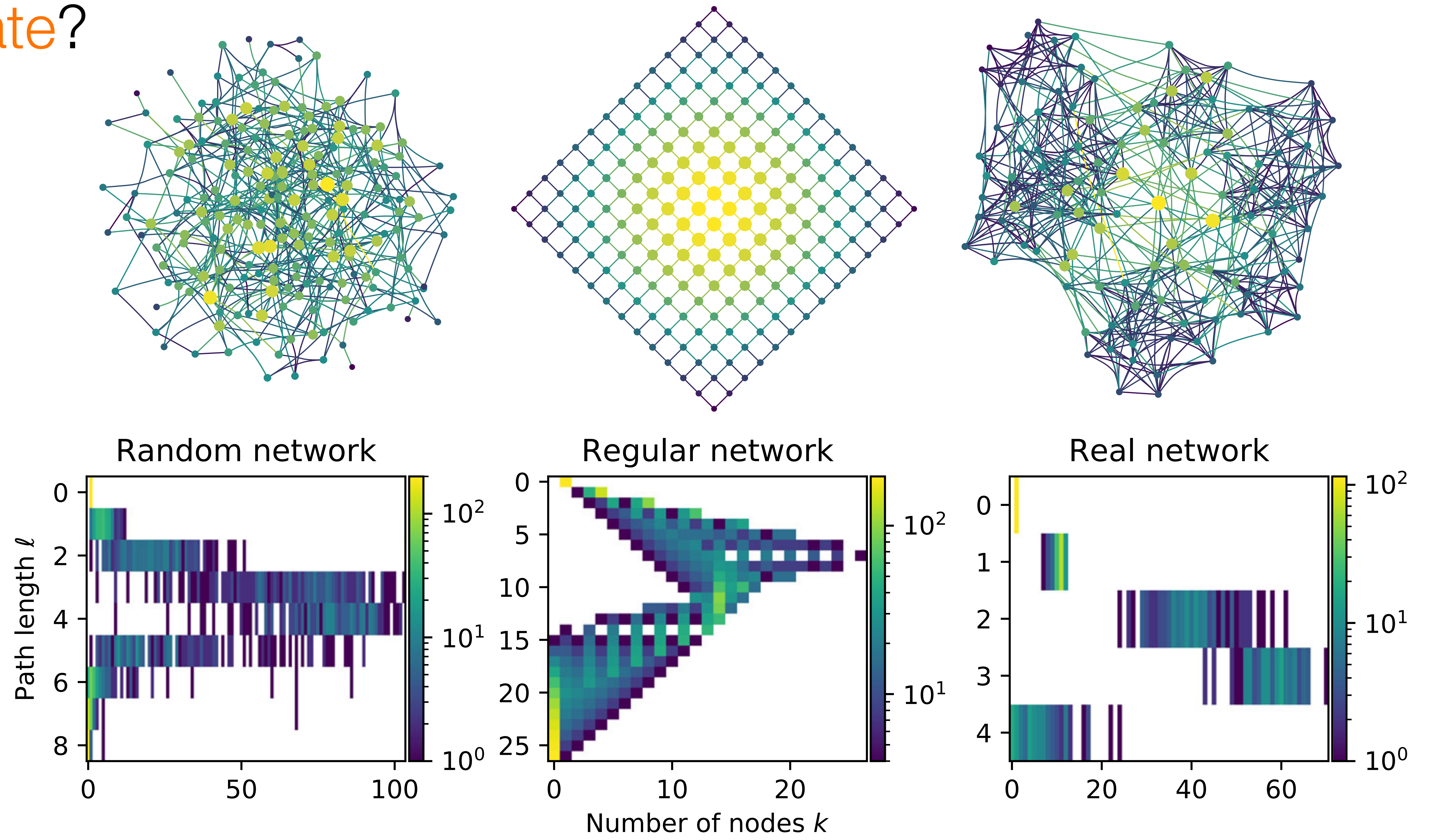
Network visualizations

Before we begin, a tough question: is a network visualization appropriate?

Ghoniem *et al*. InfoVis'04 (2004)
Foucault Welles & Meirelles (2015)
Foucault Welles & Xu (2018)

Alternative approaches

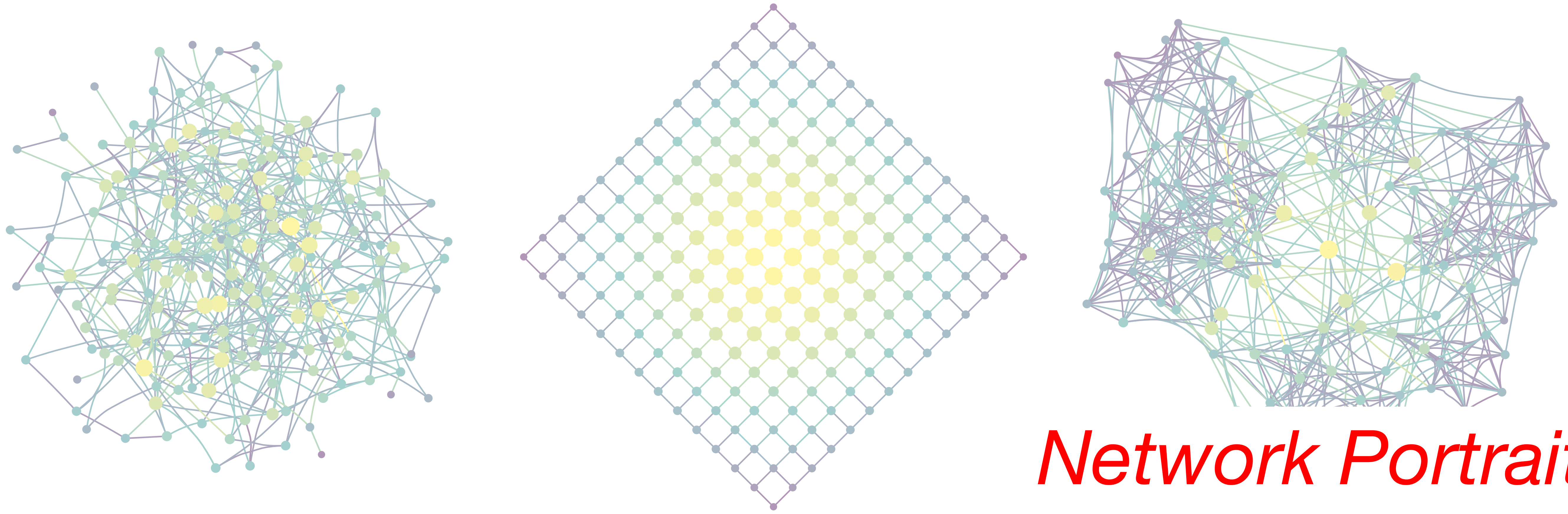
Bagrow *et al.* EPL (2008)
Bagrow & Bollt (2019)
Schulman *et al.* (2011)



Network visualizations

Before we begin, a tough question: is a network visualization appropriate?

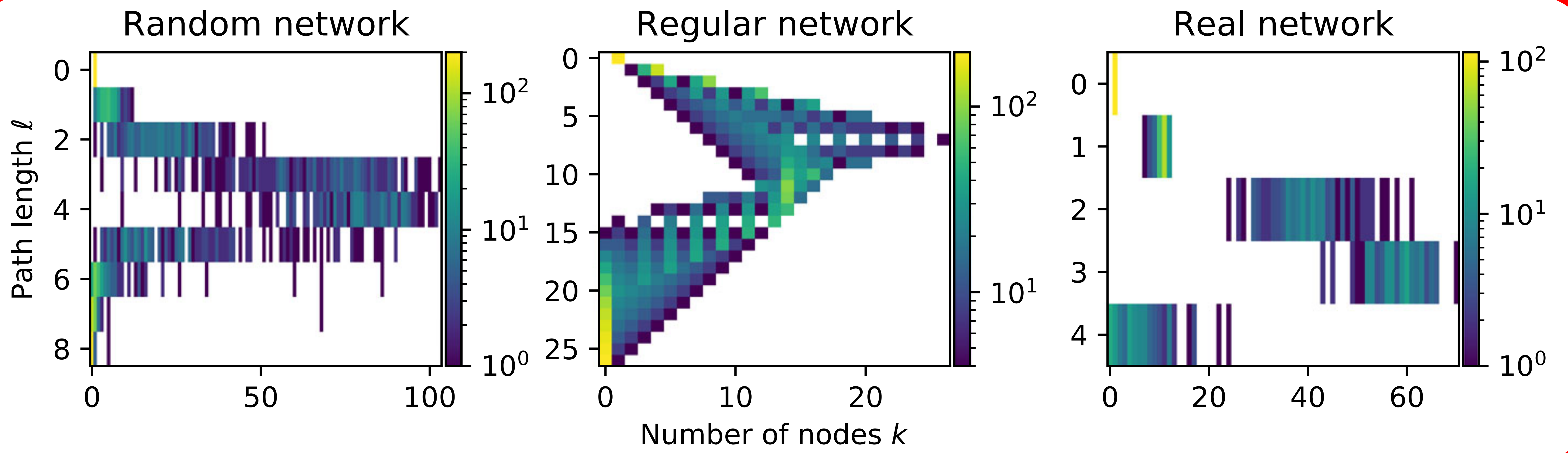
Ghoniem *et al*. InfoVis'04 (2004)
Foucault Welles & Meirelles (2015)
Foucault Welles & Xu (2018)



Alternative approaches

Bagrow *et al.* EPL (2008)
Bagrow & Bollt (2019)
Schulman *et al.* (2011)

Network Portraits



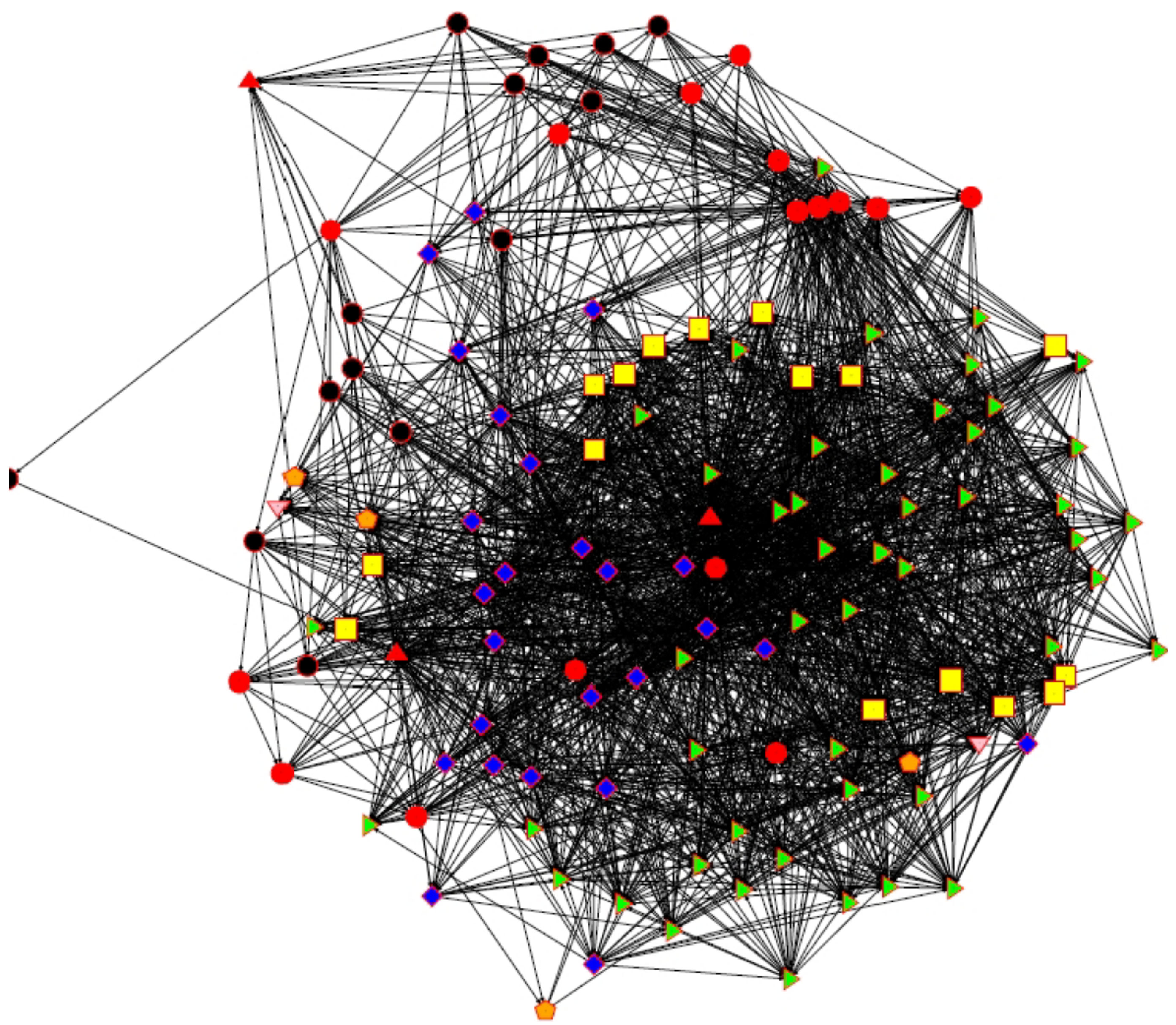
Network visualizations

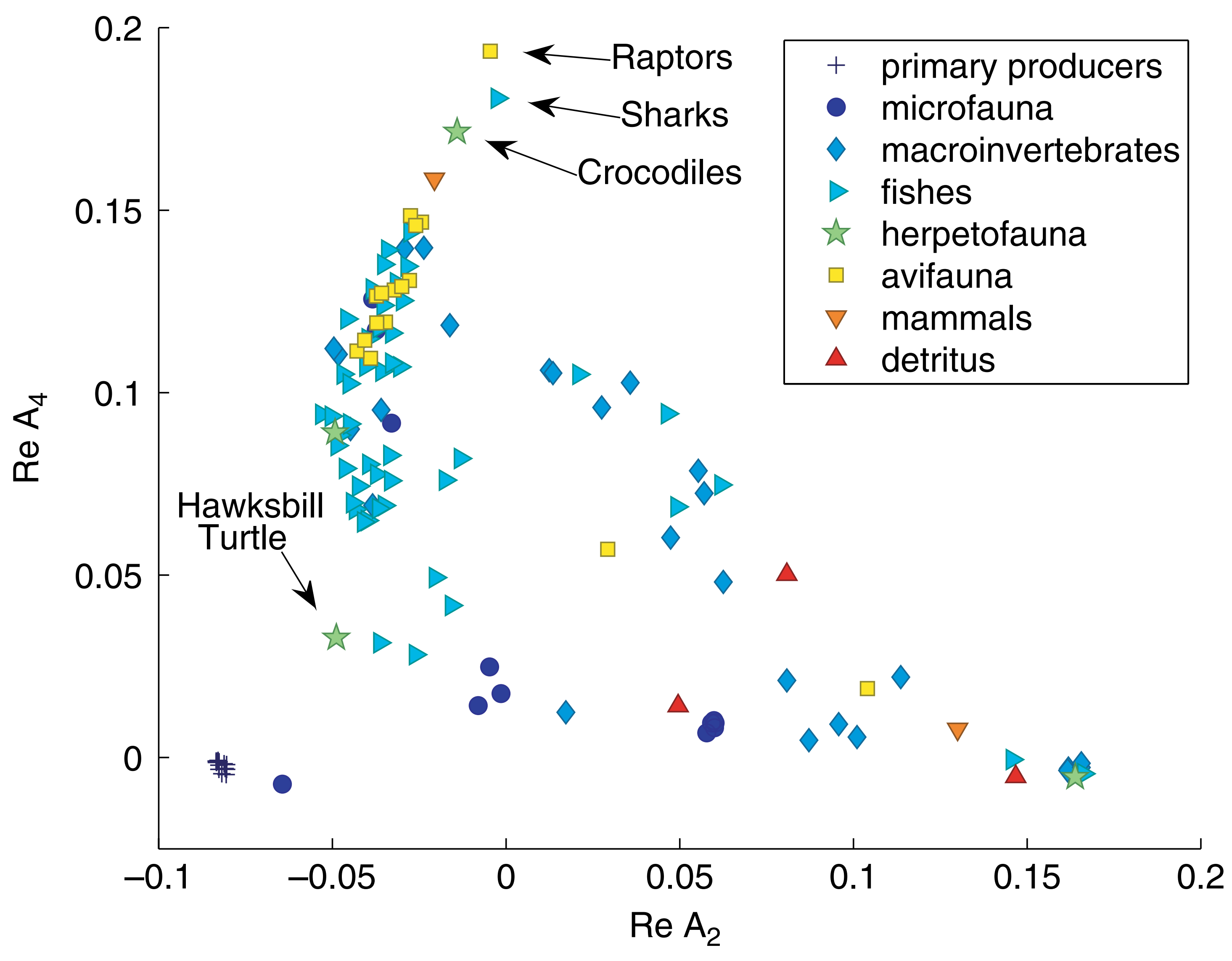
Before we begin, a tough question: is a network visualization appropriate?

Ghoniem *et al*. InfoVis'04 (2004)
Foucault Welles & Meirelles (2015)
Foucault Welles & Xu (2018)

Alternative approaches

Bagrow *et al.* EPL (2008)
Bagrow & Bollt (2019)
Schulman *et al.* (2011)





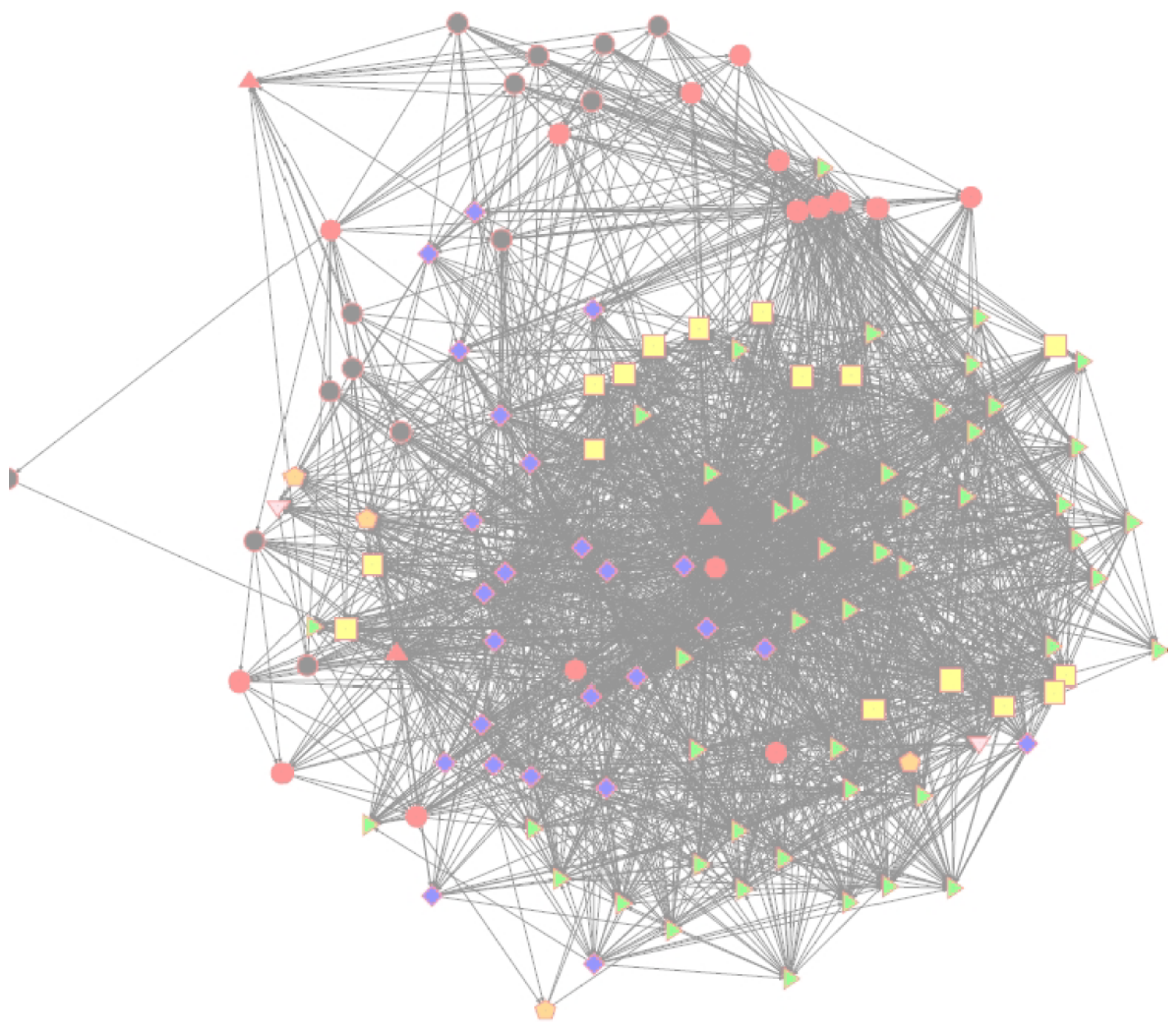
Network visualizations

Before we begin, a tough question: is a network visualization appropriate?

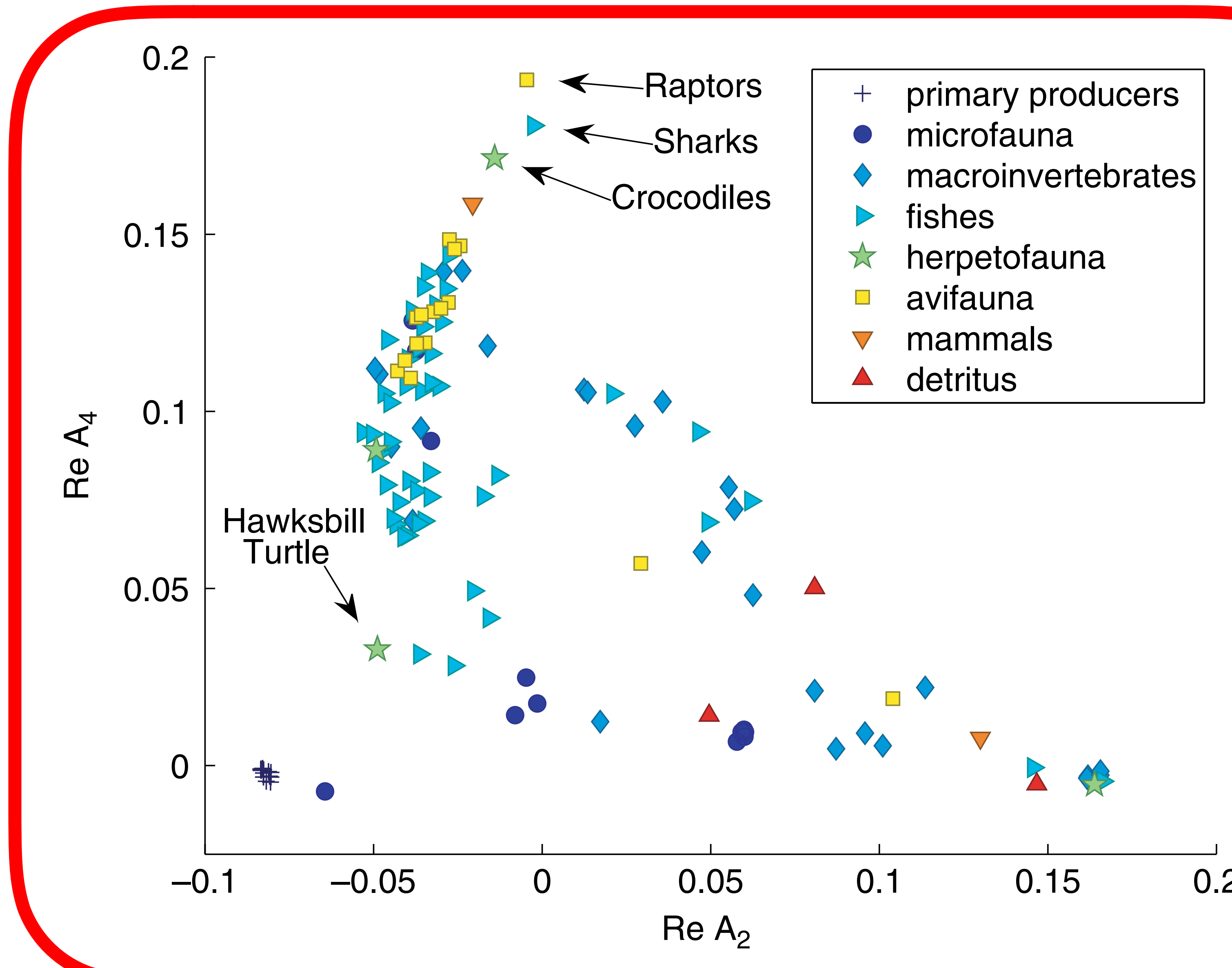
Ghoniem *et al*. InfoVis'04 (2004)
Foucault Welles & Meirelles (2015)
Foucault Welles & Xu (2018)

Alternative approaches

Bagrow *et al.* EPL (2008)
Bagrow & Bollt (2019)
Schulman *et al.* (2011)



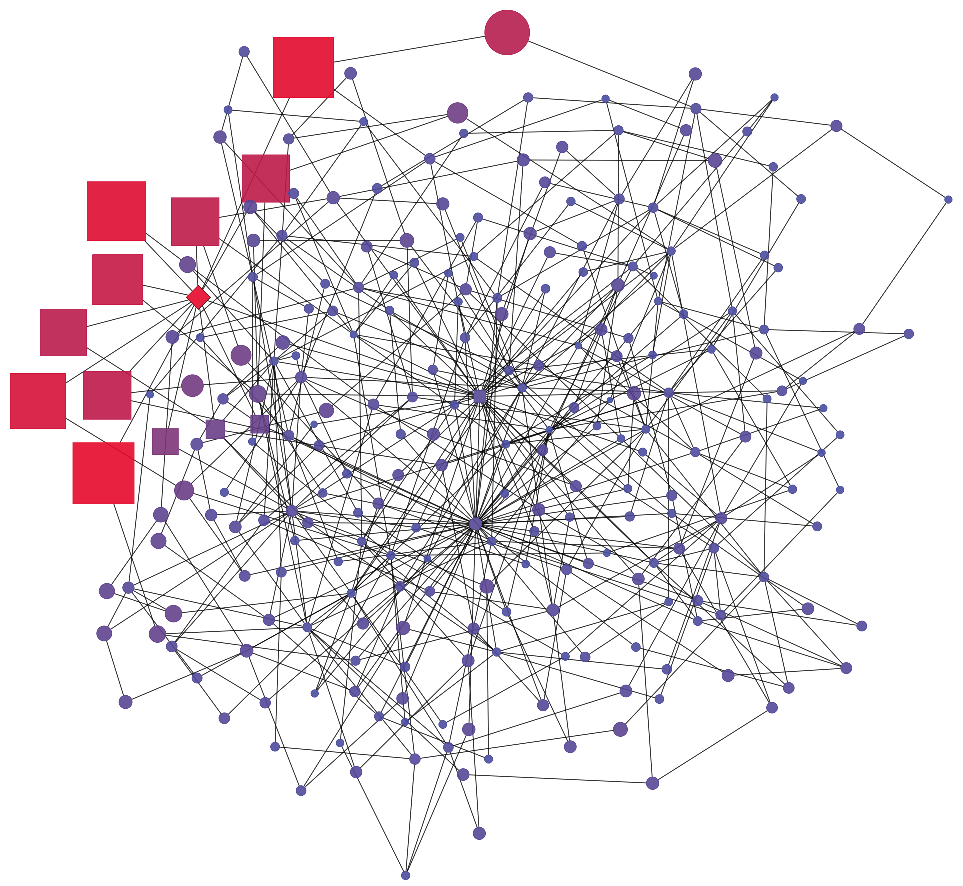
Observable Representation

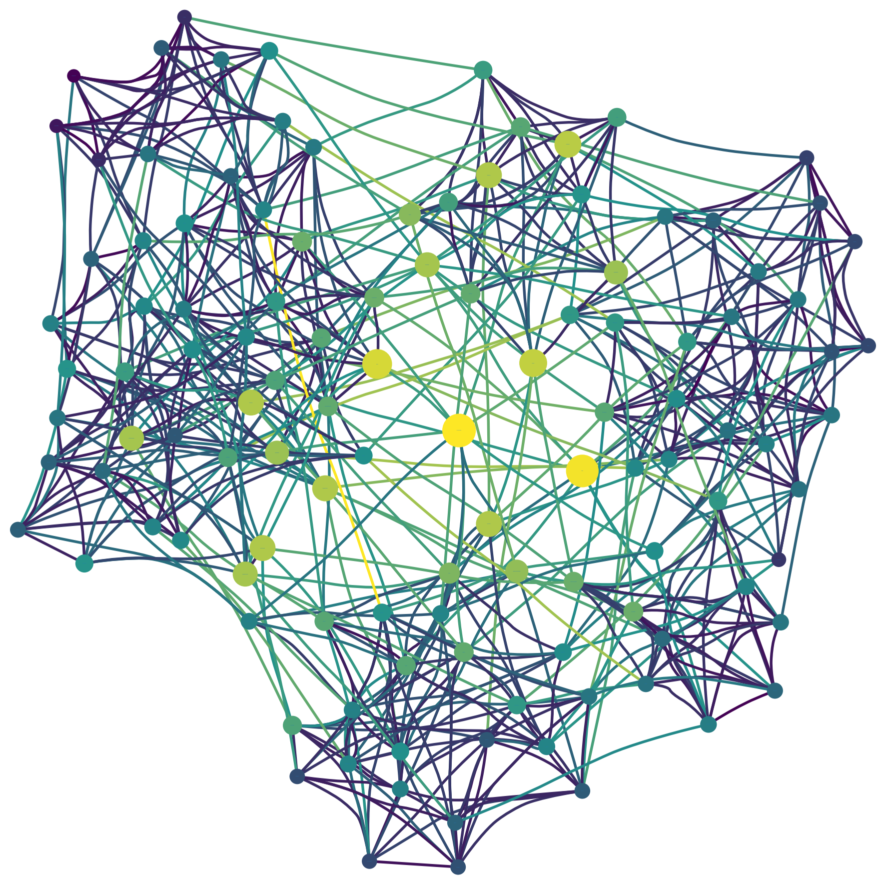


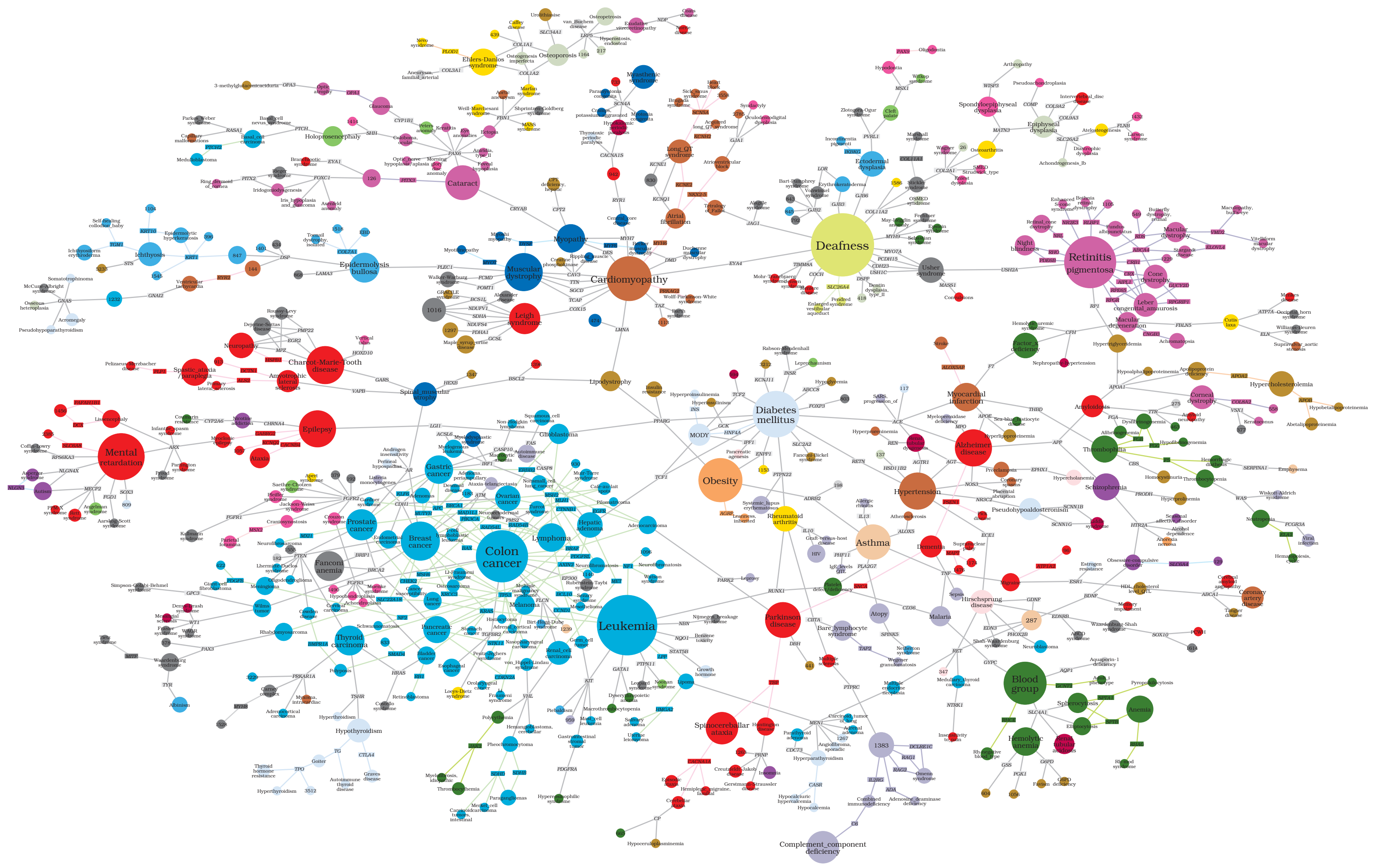
Network visualizations

- Know your <u>message</u>
- Know your <u>medium</u>
- Know your <u>audience</u>
- Account for strengths and weaknesses of <u>human perception</u>
- Keep it <u>simple</u>

All these points still hold for *visualizing networks*







Aspects of a network visualization

1. Layout (node coordinates)
2. Node "mapper"
3. Link "mapper"

Aspects of a network visualization

0. Preprocessing
   - Project if bipartite?
   - Thin the network
   - Retain only subgraph(s)
   - Group nodes, network of communities?
   - …

1. Layout (node coordinates)
2. Node "mapper"
3. Link "mapper"

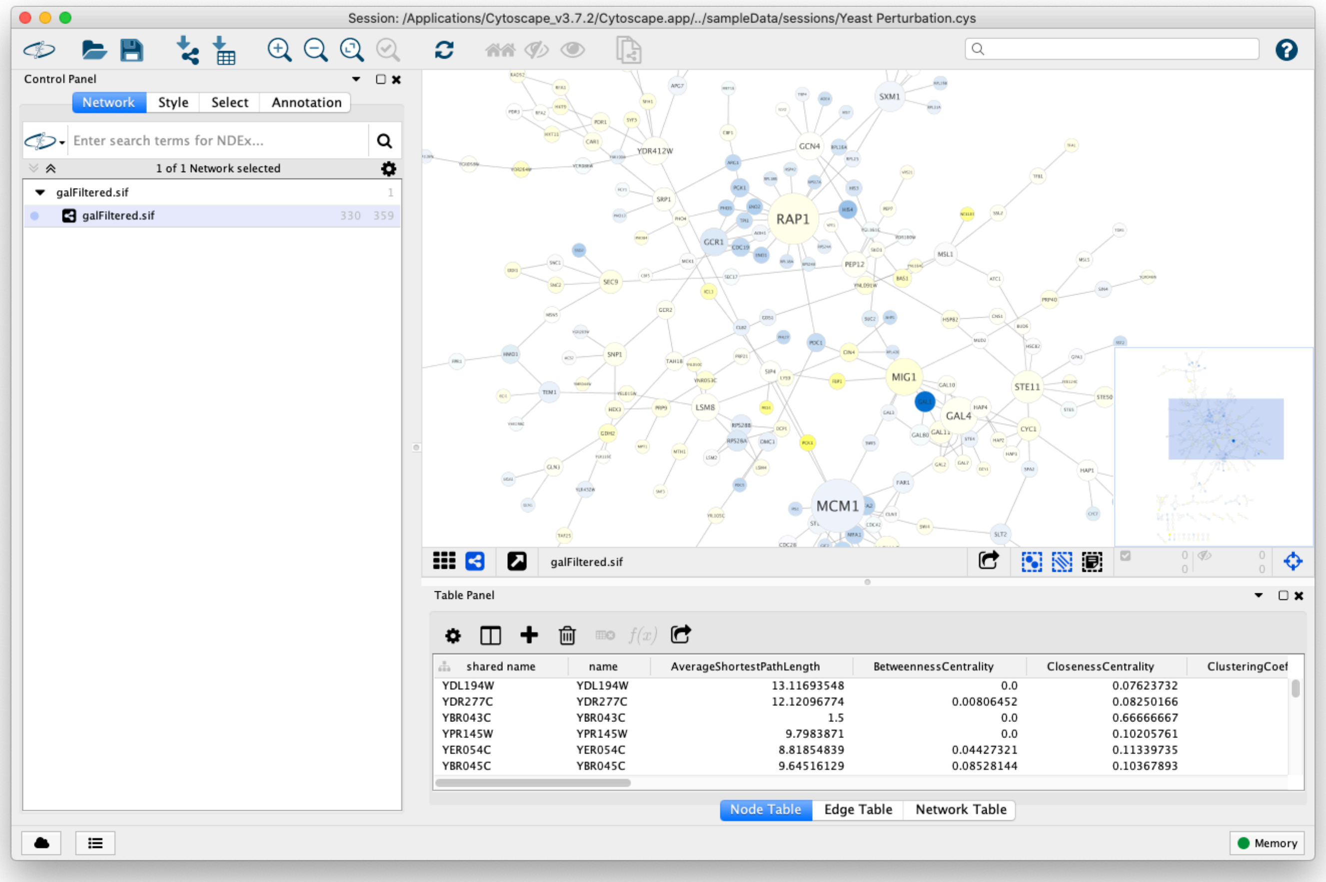
Aspects of a network visualization

0. Preprocessing
   - Project if bipartite?
   - Thin the network
   - Retain only subgraph(s)
   - Group nodes, network of communities?
   - …

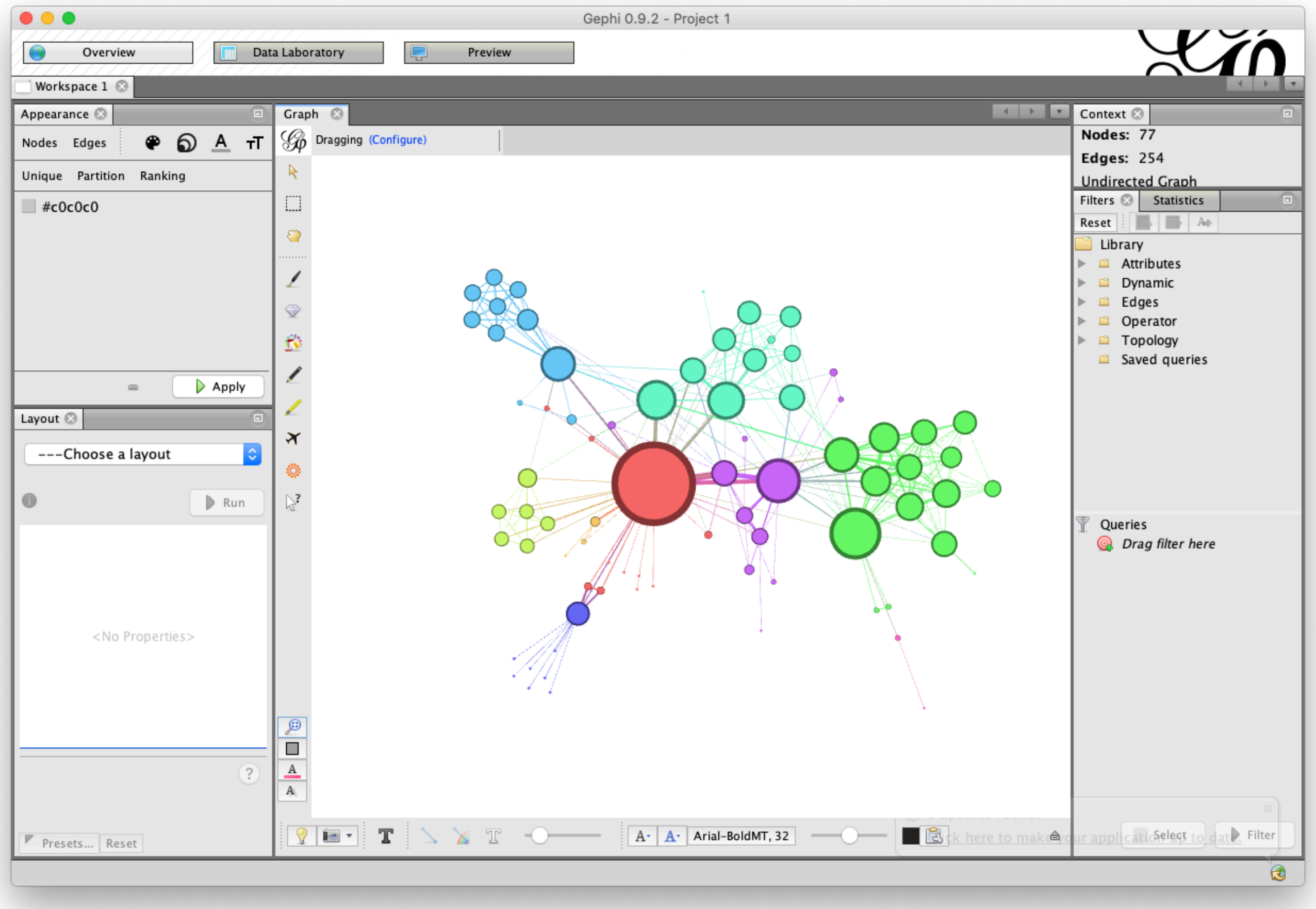
1. Layout (node coordinates)
2. Node "mapper"
3. Link "mapper"

Apps

Cytoscape



Gephi



1. Layout (node2xy)

Place nodes in a visually meaningful way
Minimize link length and crossing…

Graph drawing — many algorithms

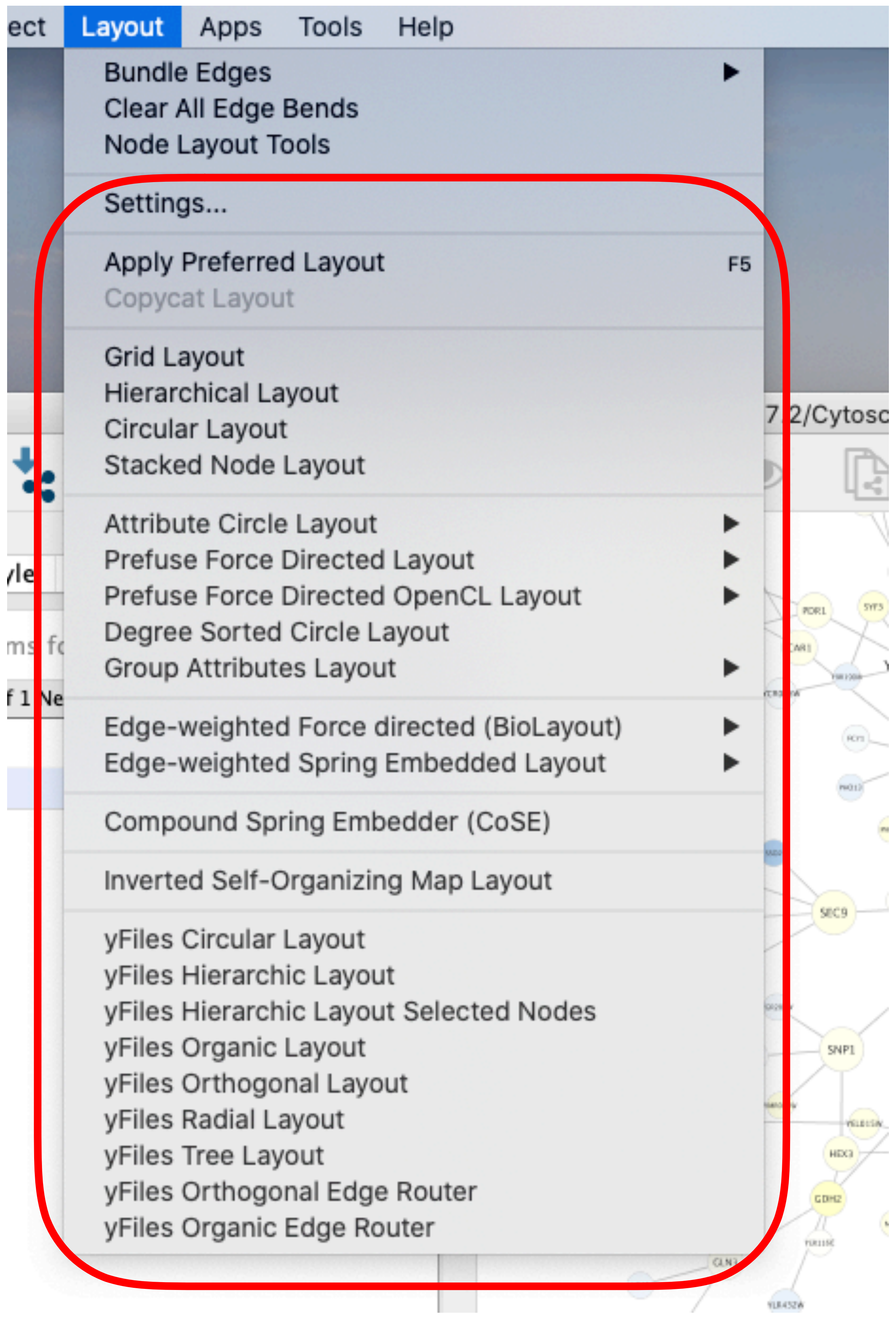
Can be slow for dense/large networks… should large networks even be visualized?

1. Layout (node2xy)

Place nodes in a visually meaningful way
Minimize link length and crossing…

Graph drawing — many algorithms

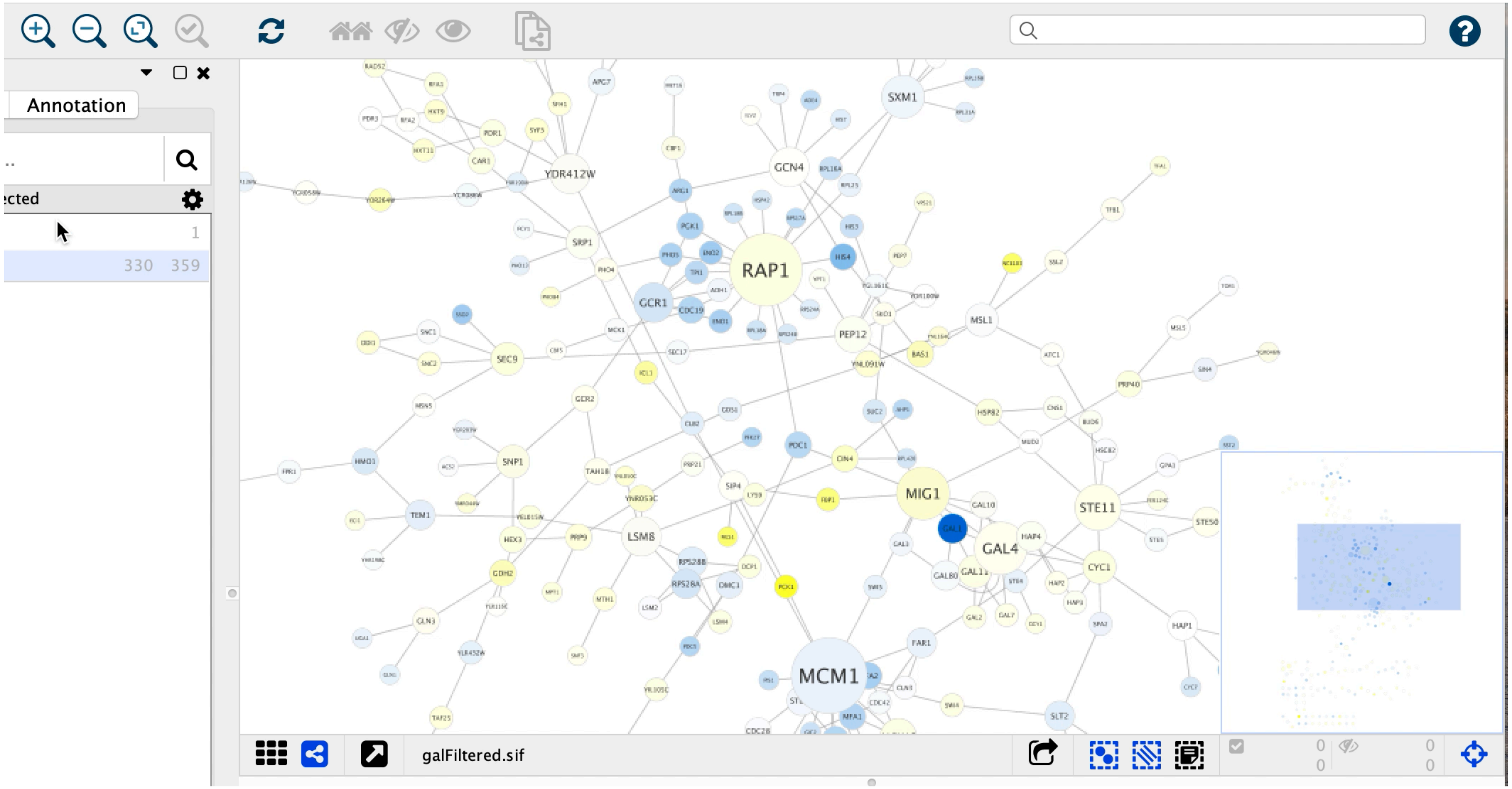
Cytoscape →



Can be slow for dense/large networks… should large networks even be visualized?

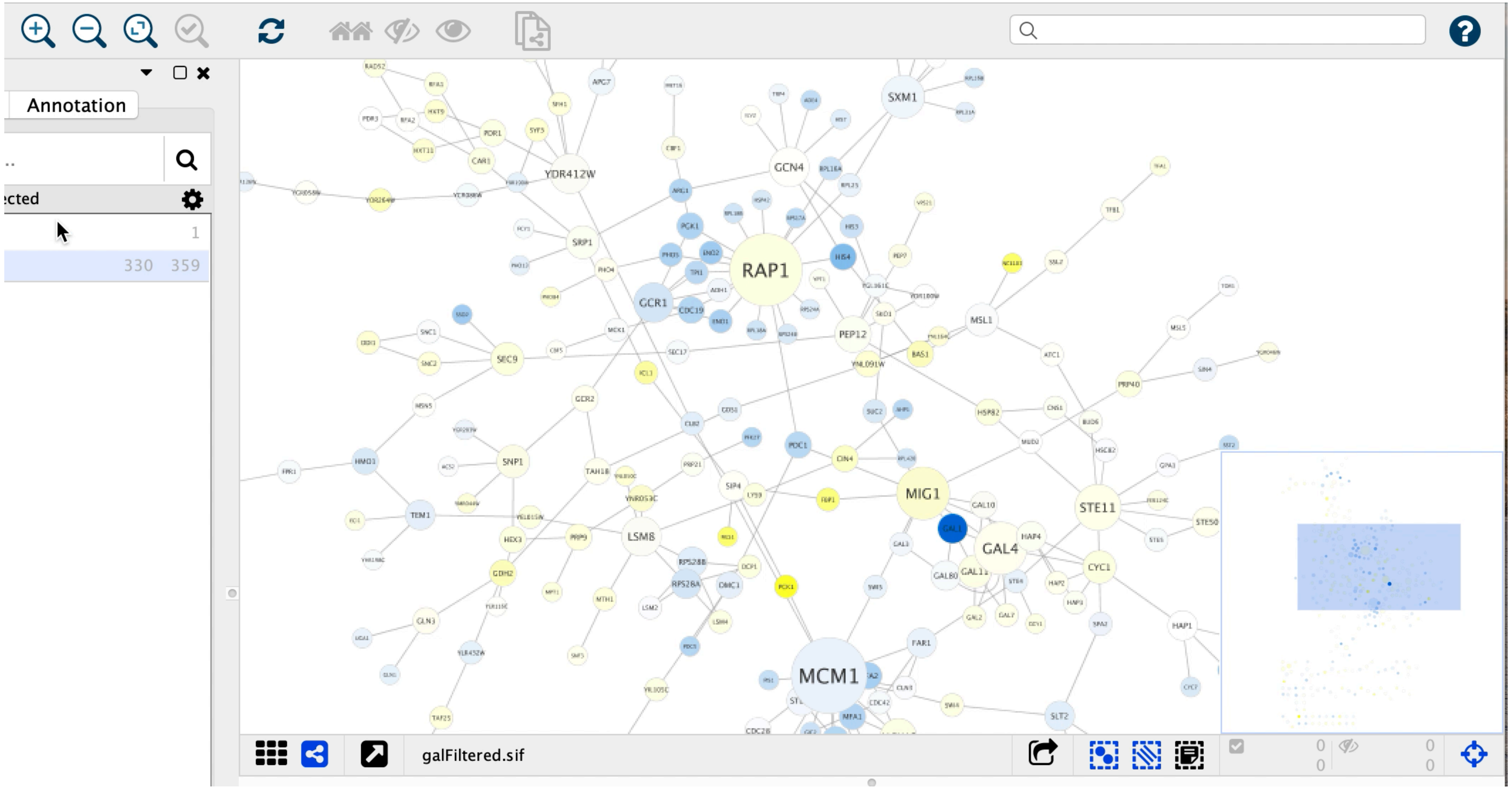
1. Layout (node2xy)

Tip: Algorithms are not perfect, *fine-tune by hand!* (for static visualizations)



1. Layout (node2xy)

Tip: Algorithms are not perfect, *fine-tune by hand!* (for static visualizations)



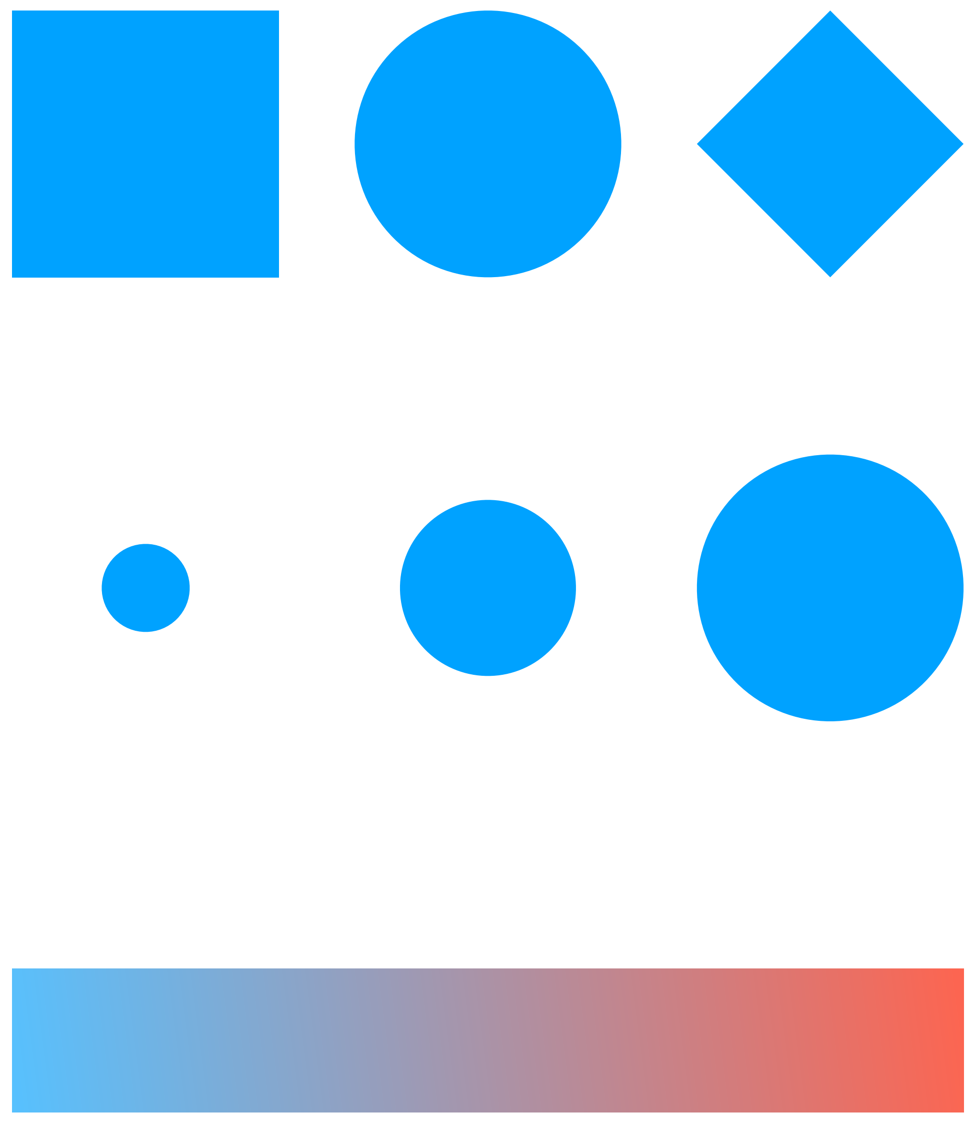
2. Node mapper (node2viz)

How to draw nodes?

Shape(s)

Size(s)

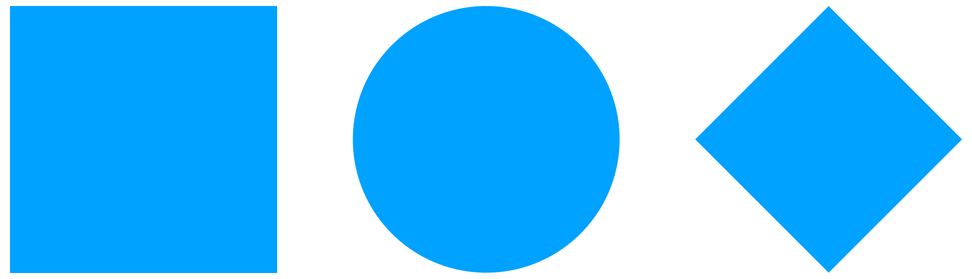
Color(s)



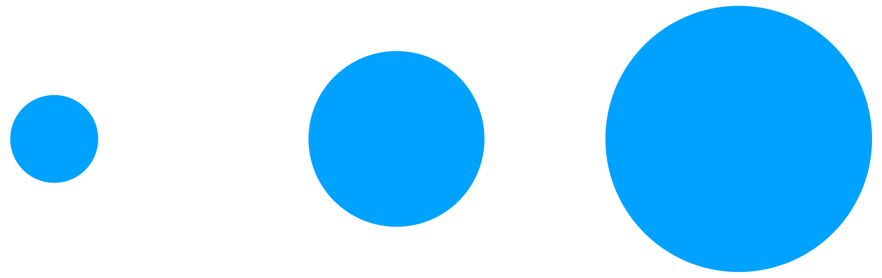
2. Node mapper (node2viz)

How to draw nodes?

Shape(s)

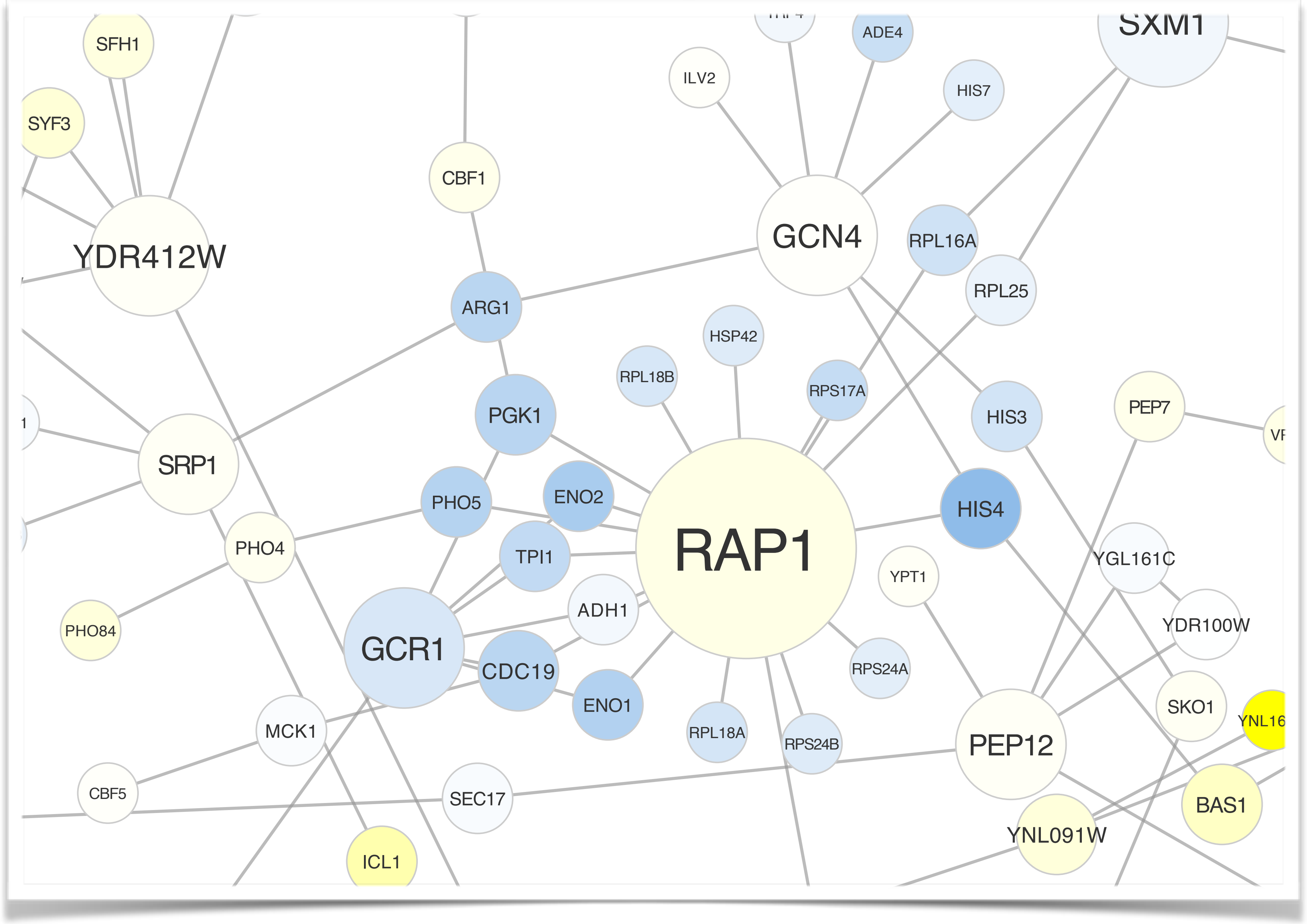


Size(s)



Color(s)

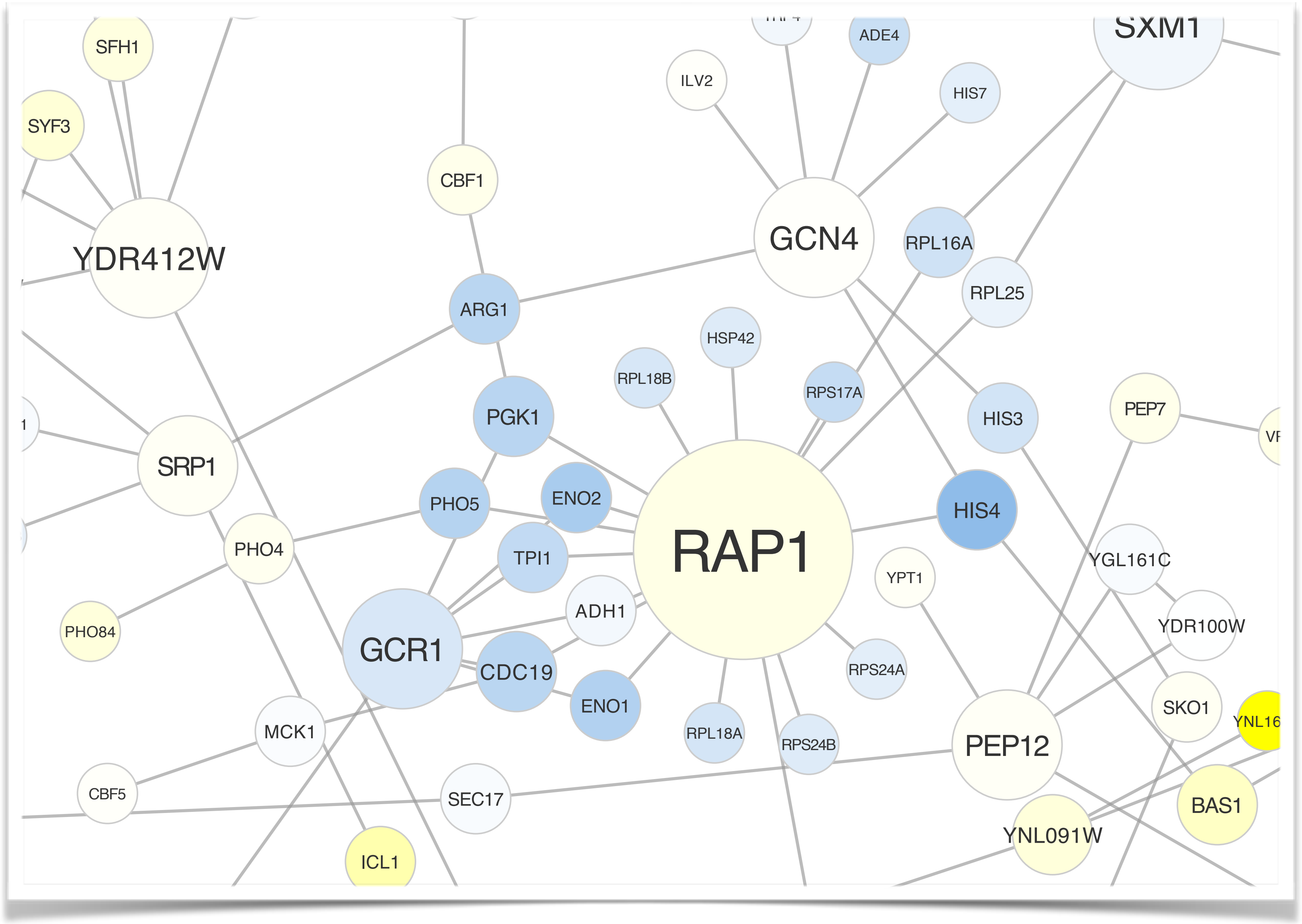
Tip: represent attributes by varying graphics



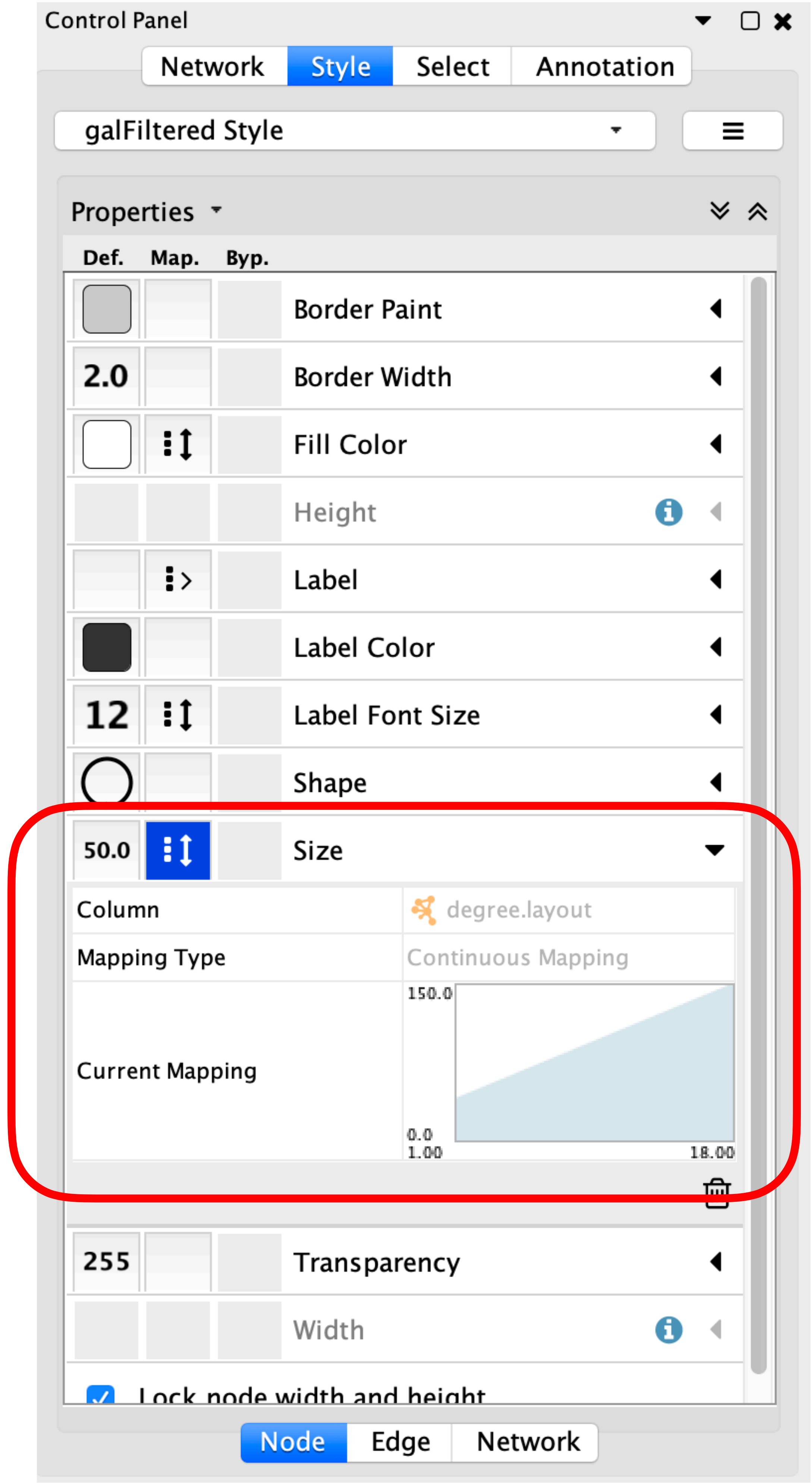
2. Node mapper (node2viz)

Tip: represent attributes by varying graphics

Cytoscape



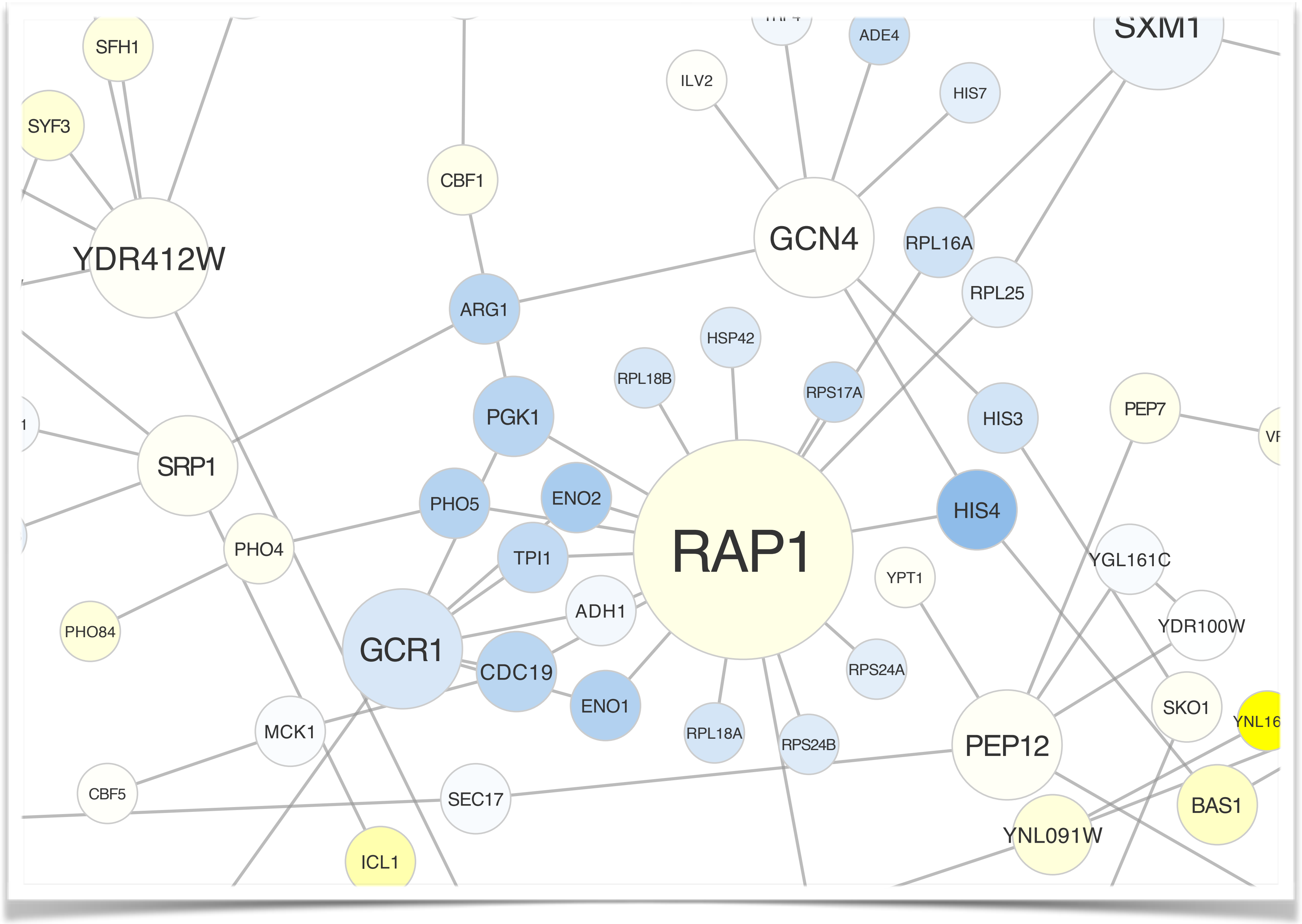
Node size ~ degree



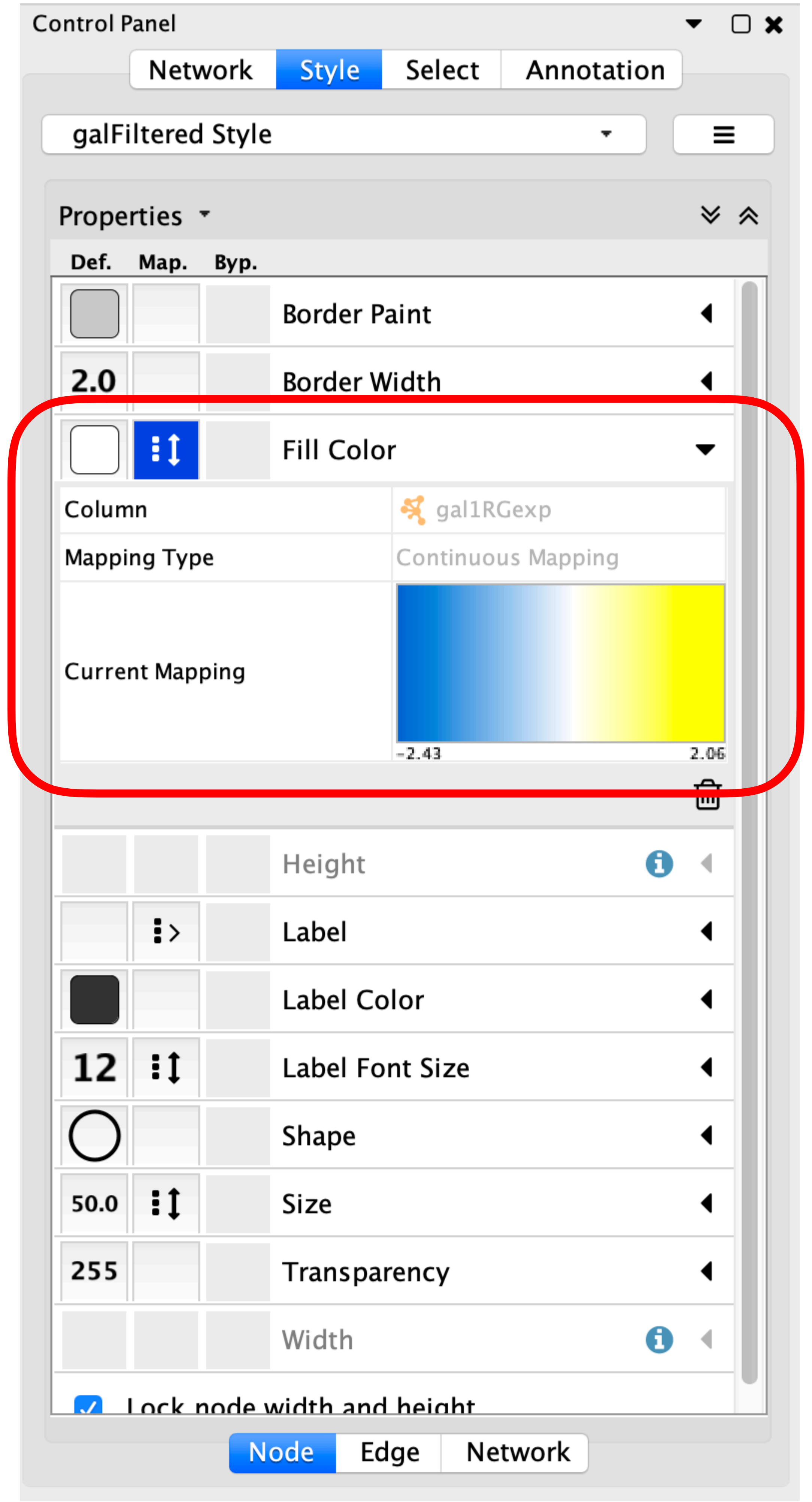
2. Node mapper (node2viz)

Cytoscape

Tip: represent attributes by varying graphics



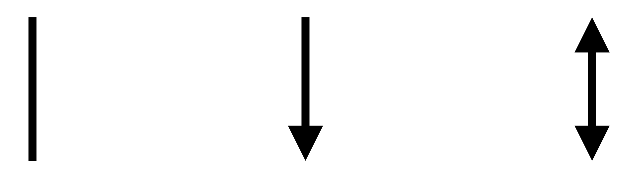
Node color ~ gene expression level



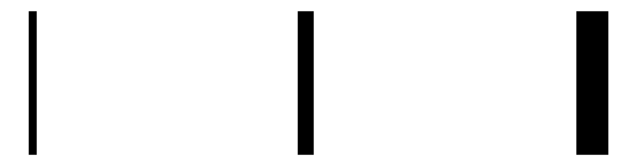
3. Link mapper (link2viz)

How to draw links?

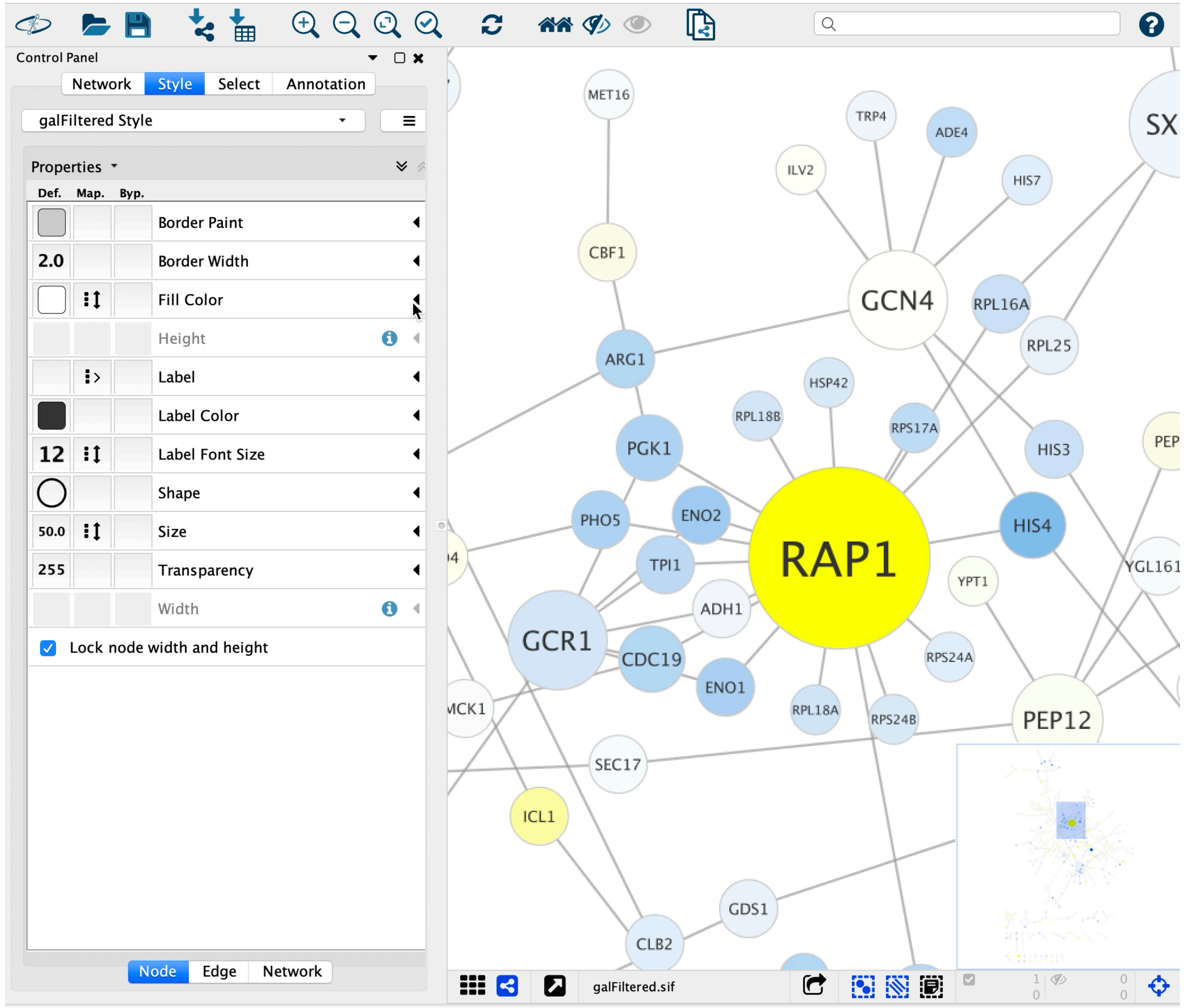
Shape(s)



Thickness(s)

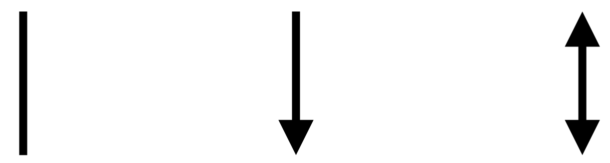


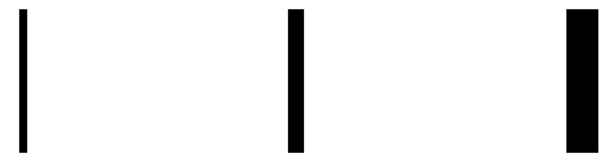
Color(s)



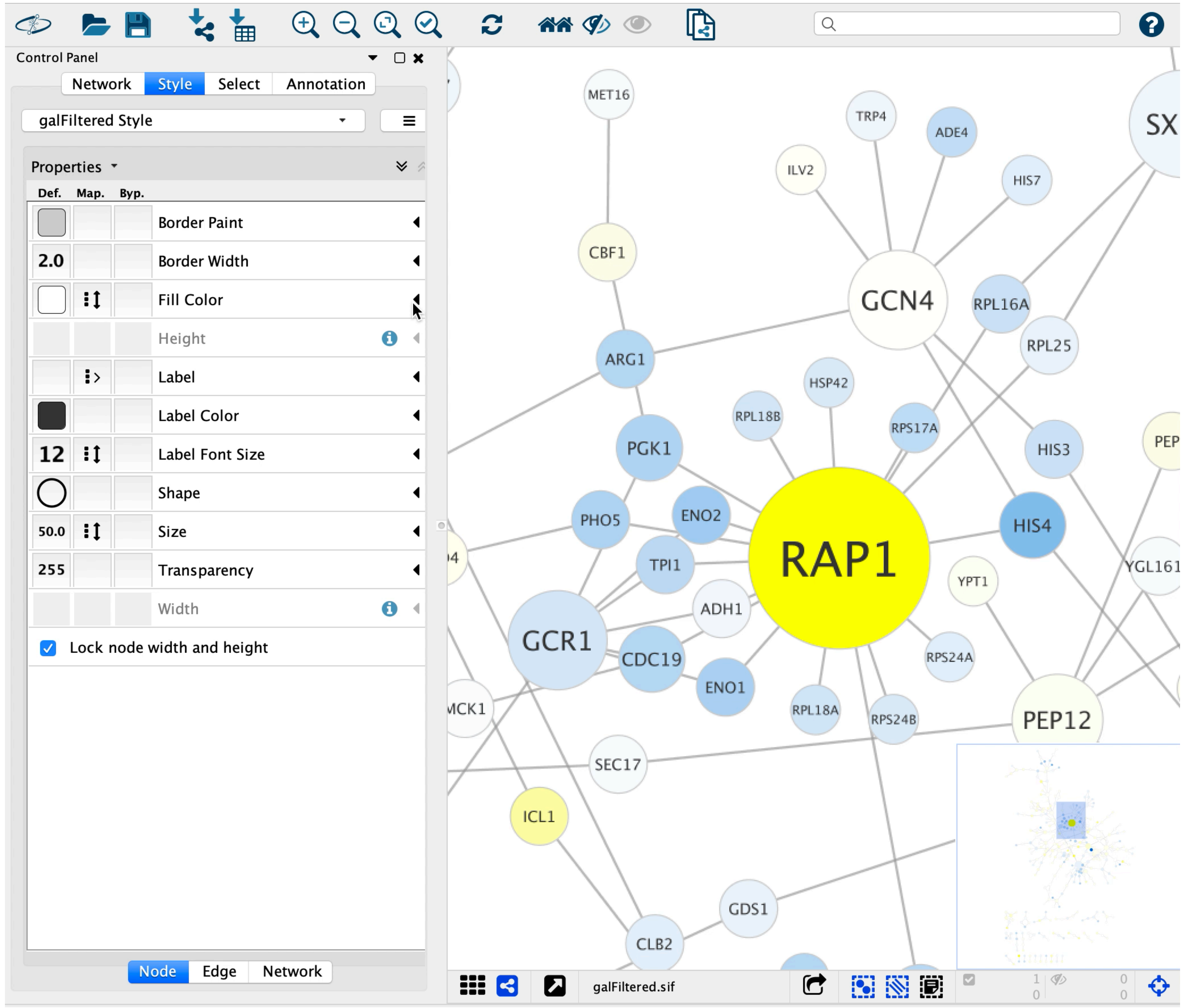
3. Link mapper (link2viz)

How to draw links?

Shape(s) 

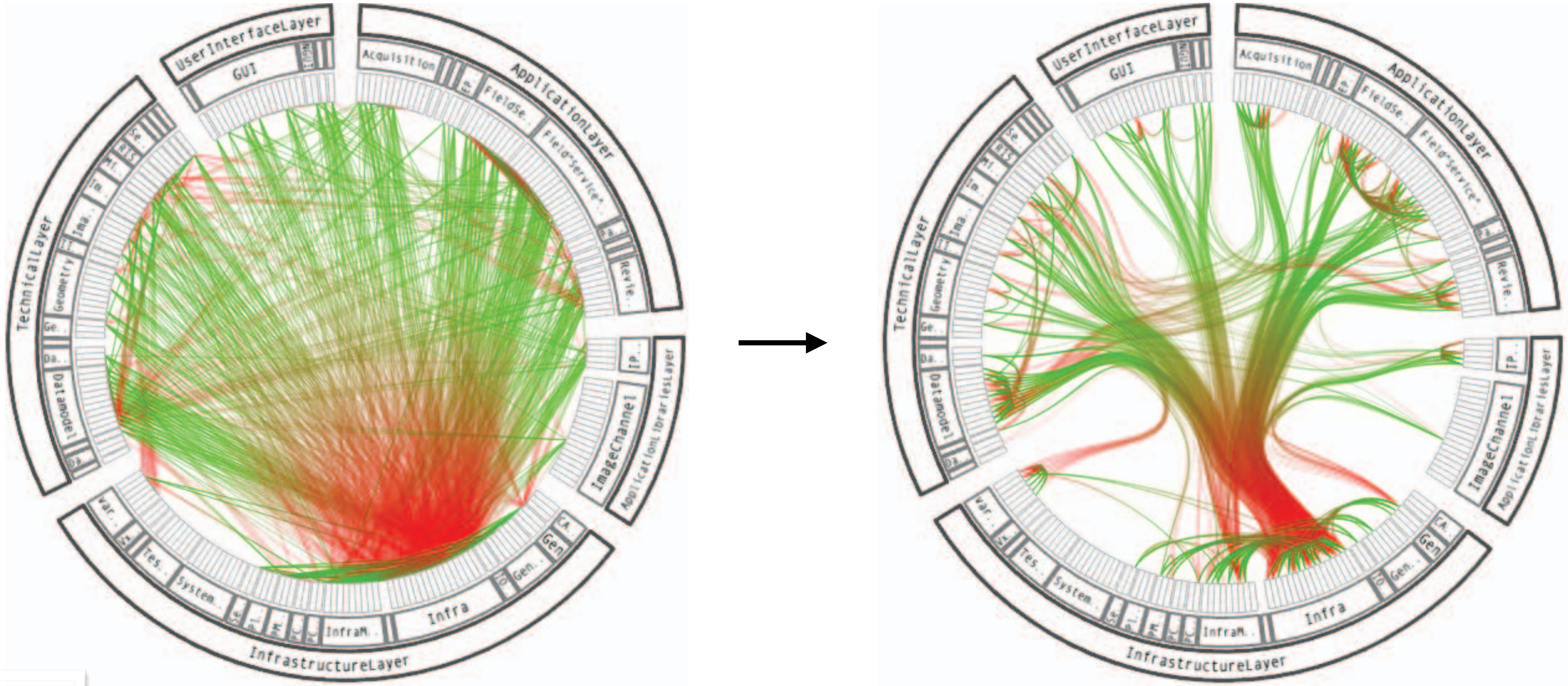
Thickness(s) 

Color(s) 



Tip: edges don't need to be straight lines

Edge bundling



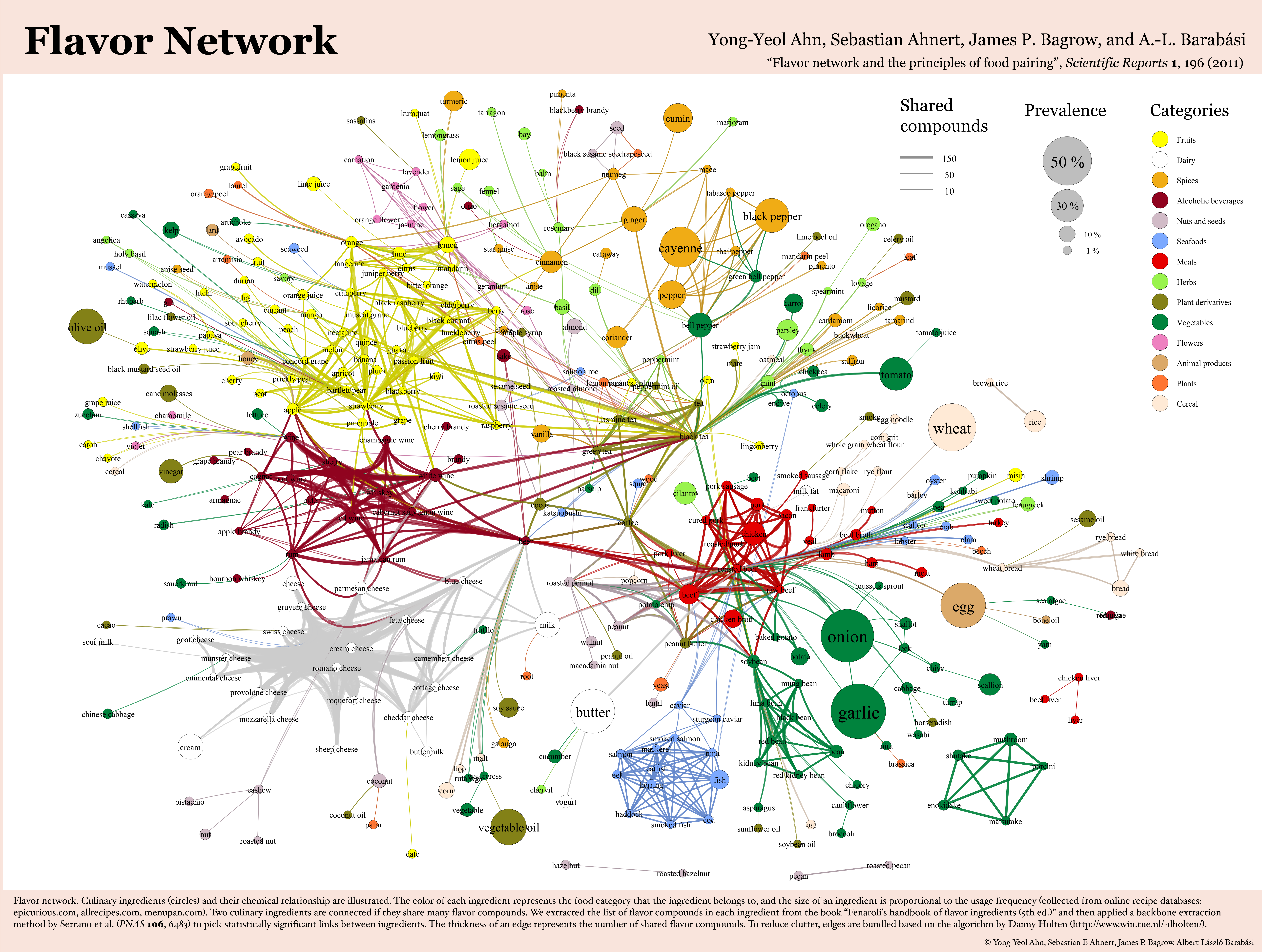
Hierarchical Edge Bundles:
Visualization of Adjacency Relations in Hierarchical Data

Danny Holten

Holton (2006)

Tip: edges don't need to be straight lines

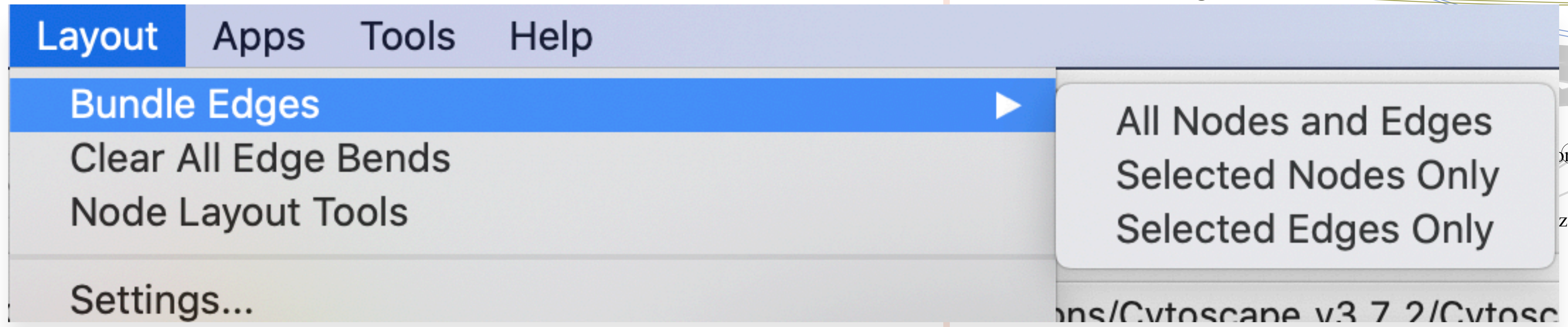
Edge bundling

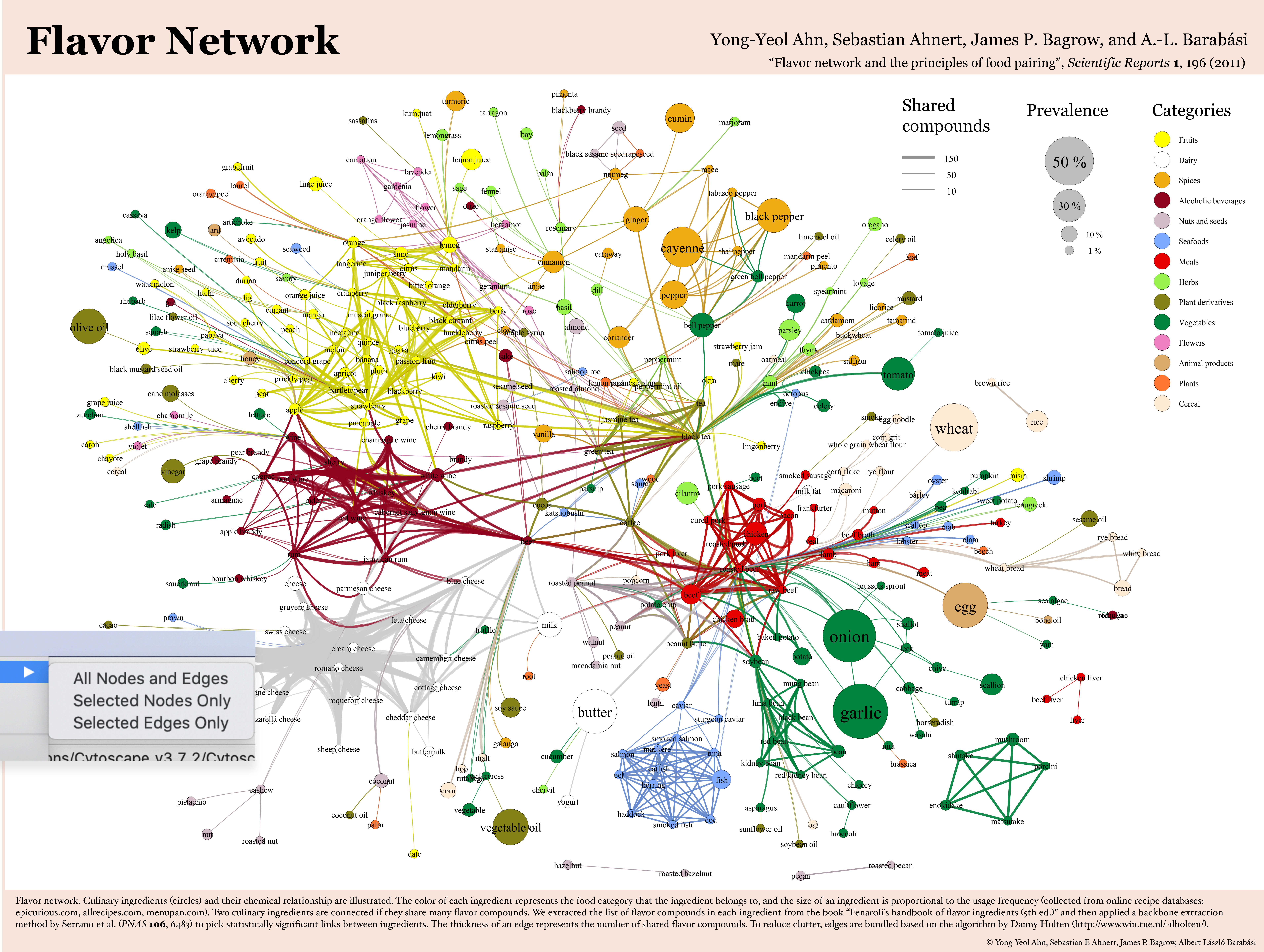


Tip: edges don't need to be straight lines

Edge bundling

Cytoscape:





Summary

- Basics
  - file formats, code, databases
- Networks from data
  - common tasks and good practices
- Case studies and examples
- Machine learning for data and networks
- Visualization (*time permitting*)

Challenges

- Hard to automate, generalize data analysis
  - upstream tasks defining the network
  - different fields have different needs
- Many tools, statistics, and algorithms—what to choose? standardize?
- Gap between models and data?
- Error analysis / Uncertainty quantification
- Big data:
  - Gap between research and industry needs
  - Graph databases—tech moving too quickly
  - Visualizations (at scale)

Working with network data

Jim Bagrow
james.bagrow@uvm.edu
bagrow.com

Complex Networks
Winter Workshop
2019-12-16

The University of Vermont

VERMONT
COMPLEX SYSTEMS CENTER

Working with network data

Jim Bagrow
james.bagrow@uvm.edu
bagrow.com

Complex Networks
Winter Workshop
2019-12-16

THANK YOU

The University of Vermont

VERMONT
COMPLEX SYSTEMS CENTER