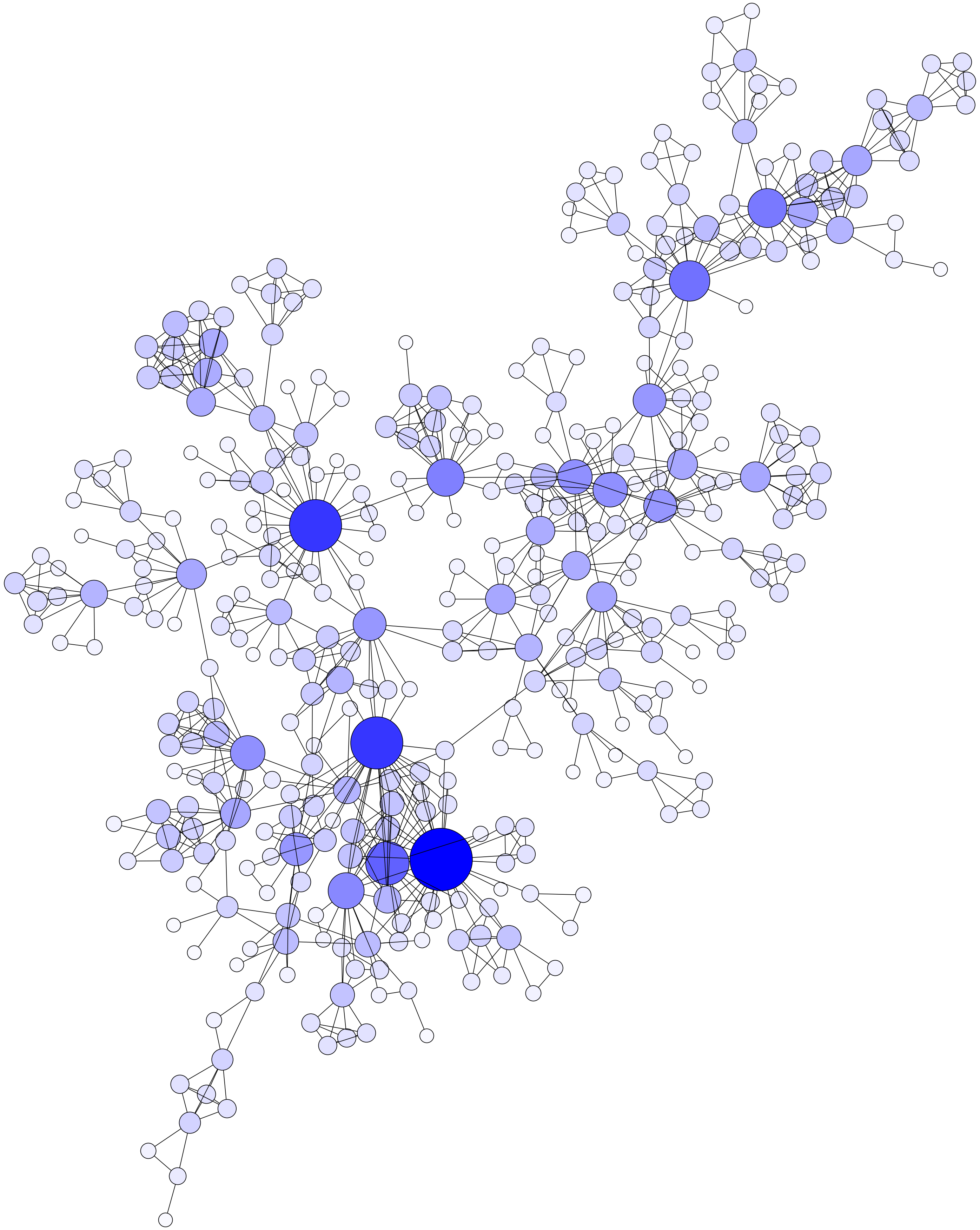
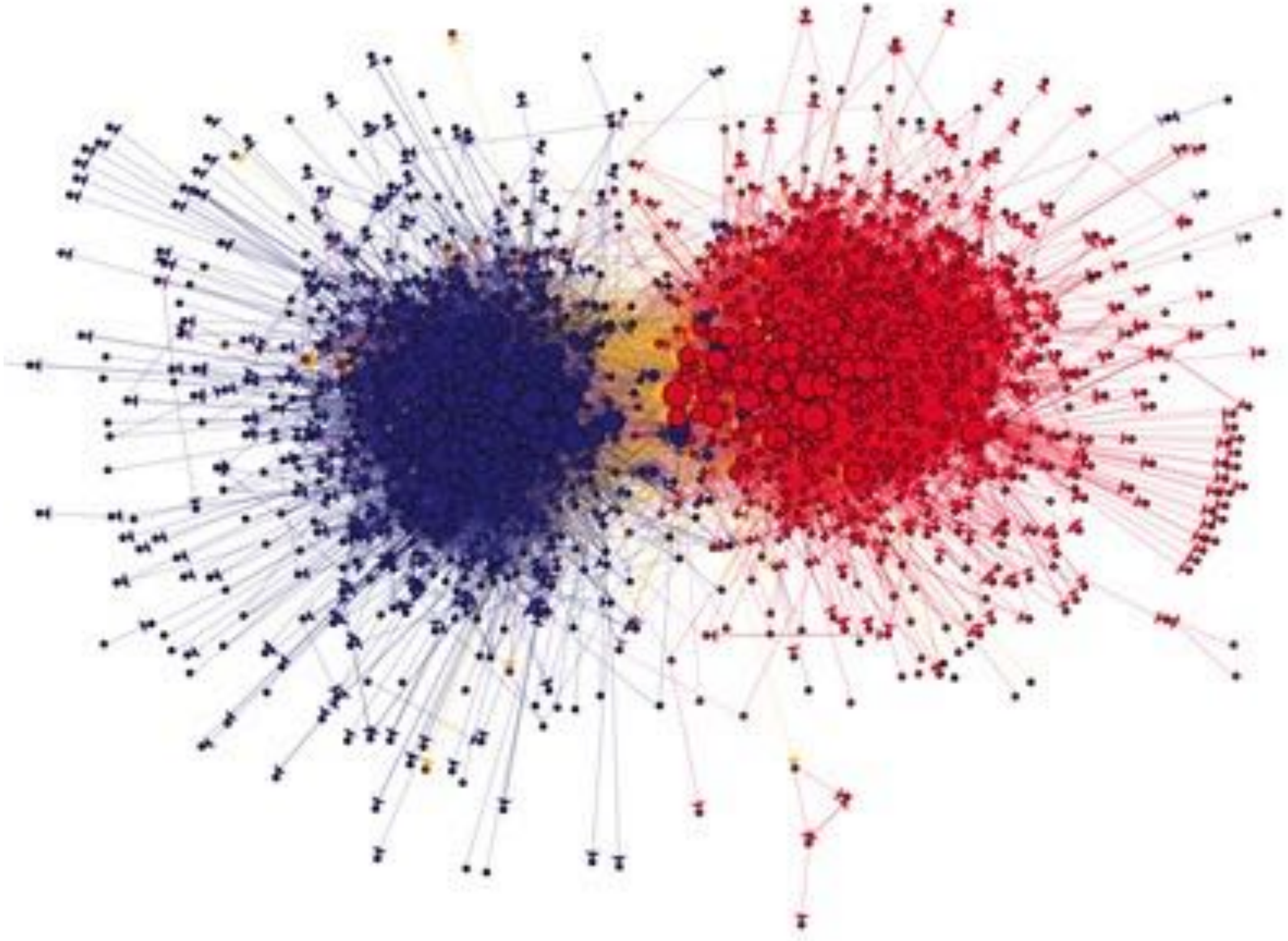


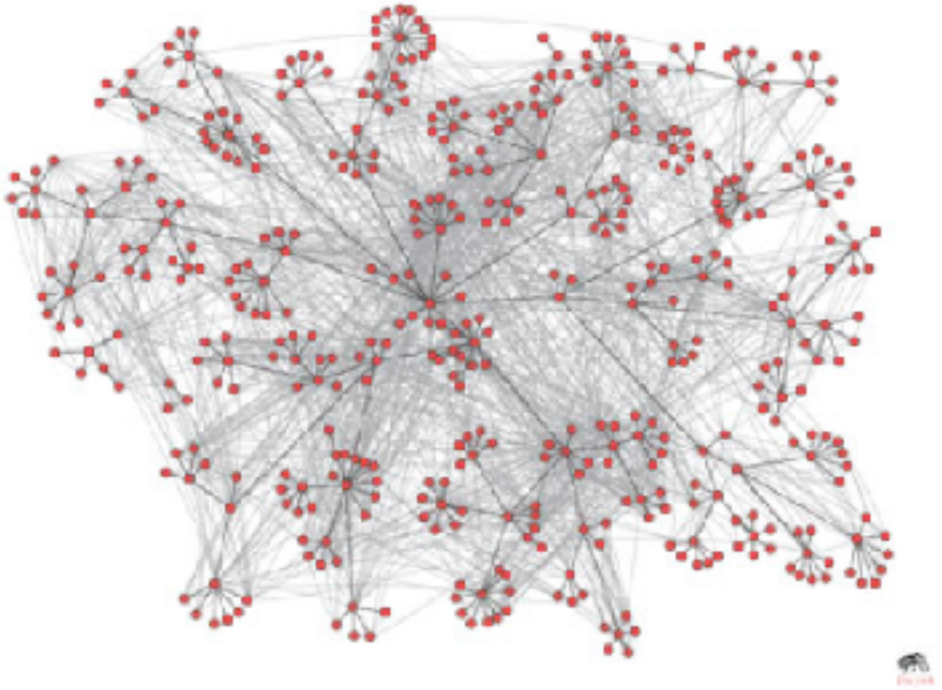
Network Data: Collection, Representation, Visualization

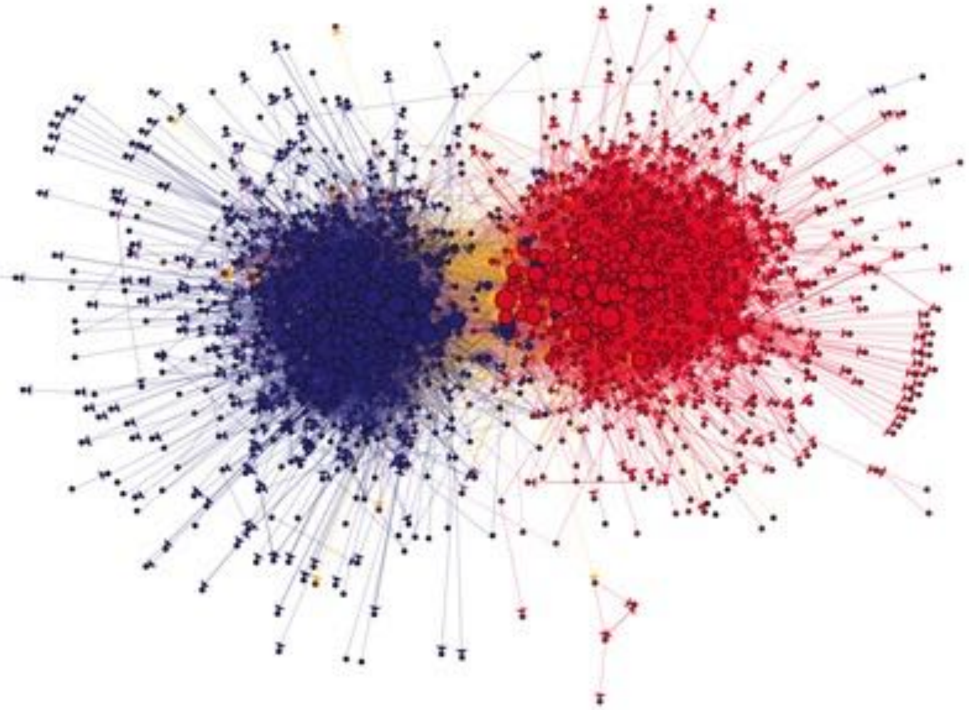
Network Data

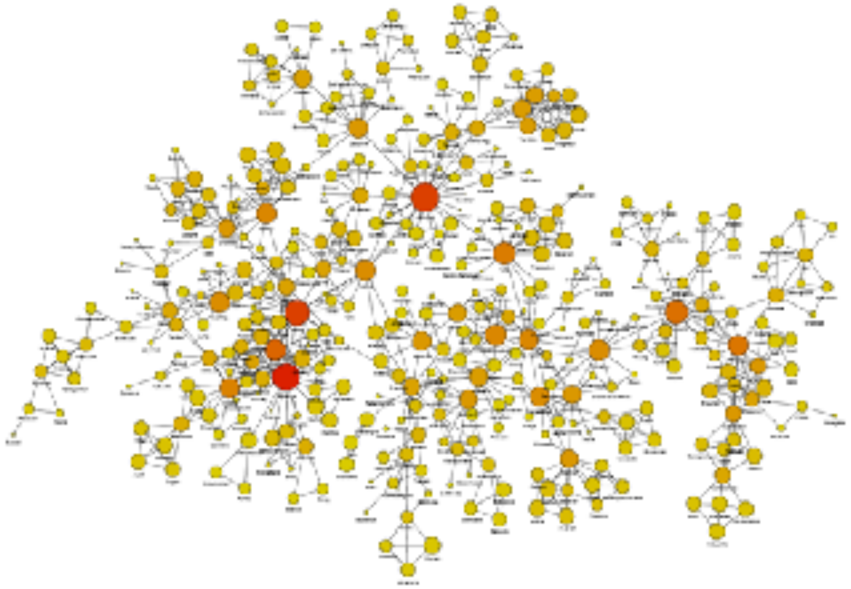


ref: Adamic political blogs

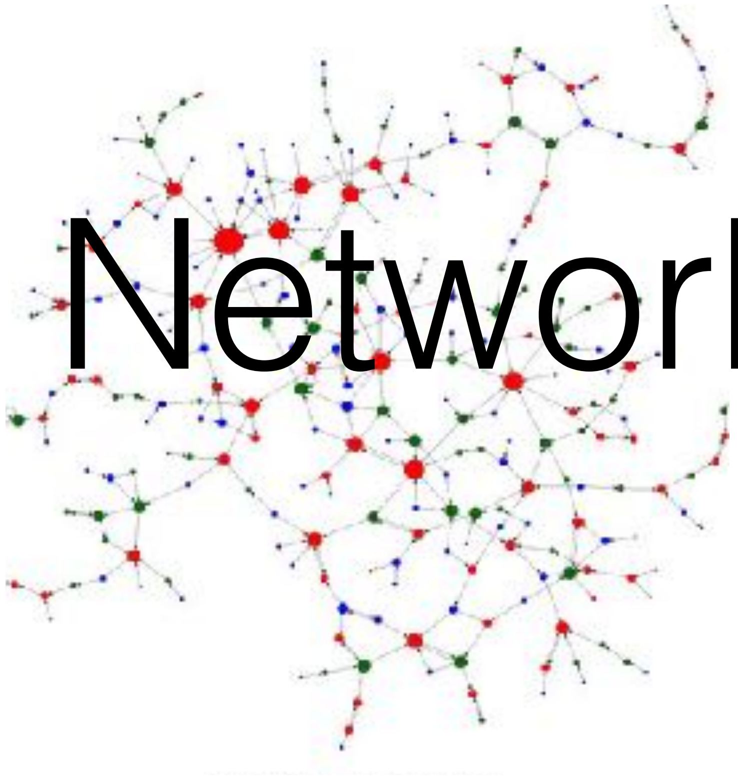
- Types of Networks
- Data Collection
- Data Representations
- Visualization
- Tools







A Taxonomy of Networks

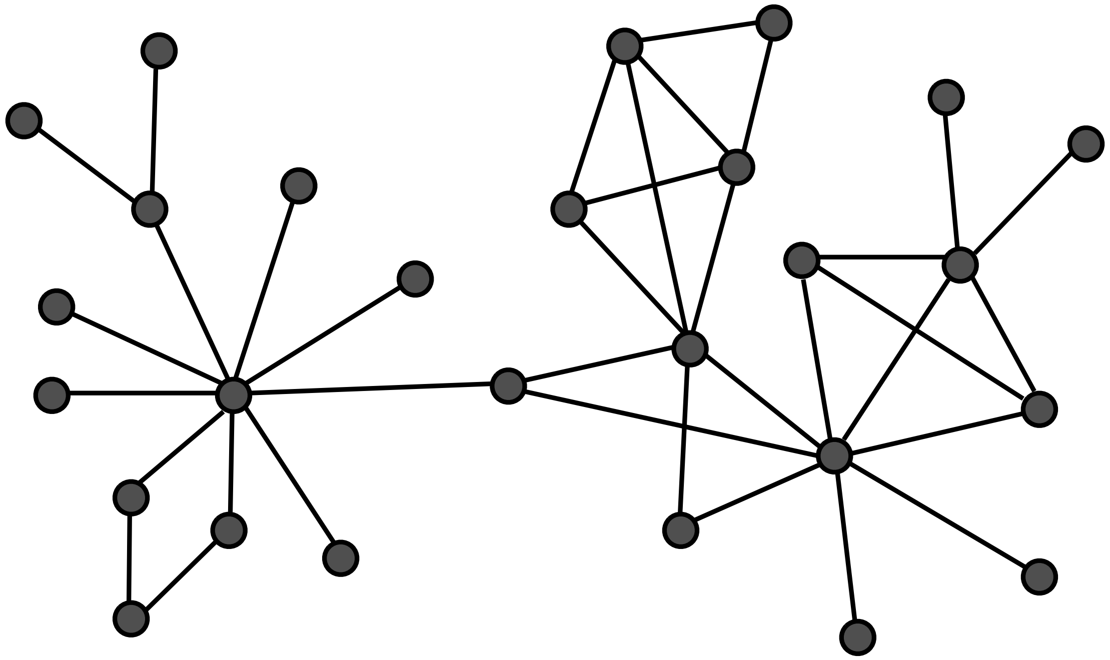


Social Networks: Parts

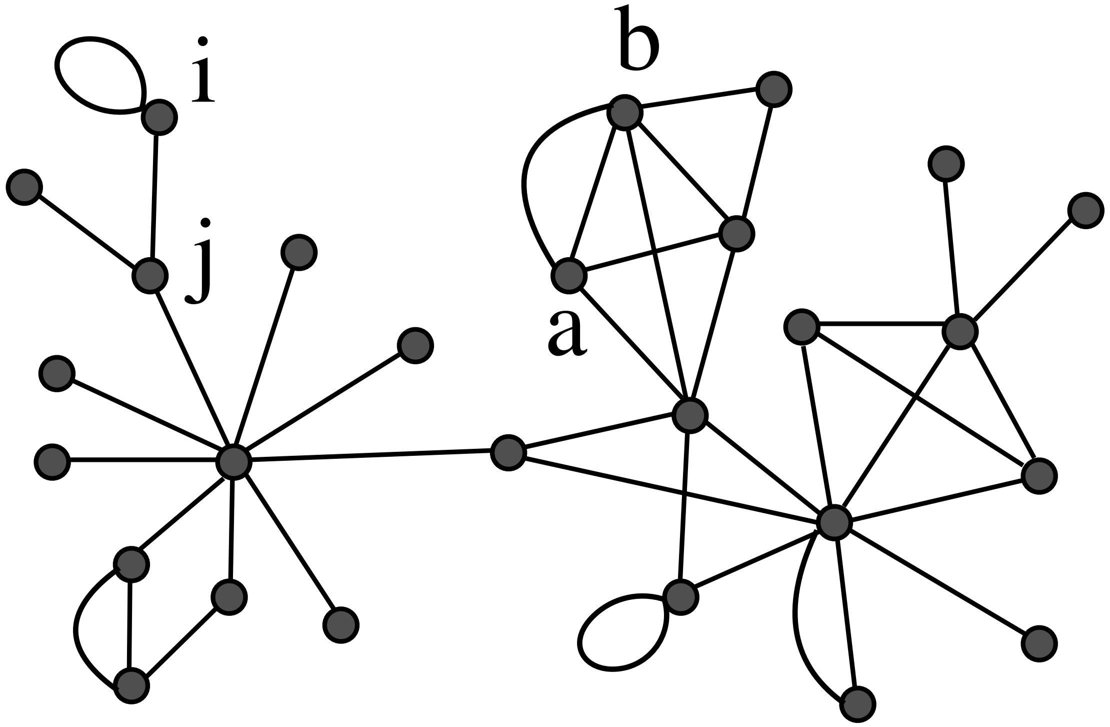
Nodes: people, firms, organizations

Edges: interactions

- Friendship
- Trust
- Cooperation
- Co-membership
- Co-location
- Trade



A bit of terminology…



Typical notation:

- $\mathbf{N}$ = number of nodes
- $\mathbf{M}$ = number of links
- referring to a particular network: $\mathbf{g}$
- referring to a node: $\mathbf{i}$ or $\mathbf{j}$
- referring to a link: $\mathbf{ij}$

- multi-edge: multiple edges between two nodes (often replaced with a link weight)
- self-edge (self-loop): a link to the same node, $\mathbf{ii}$

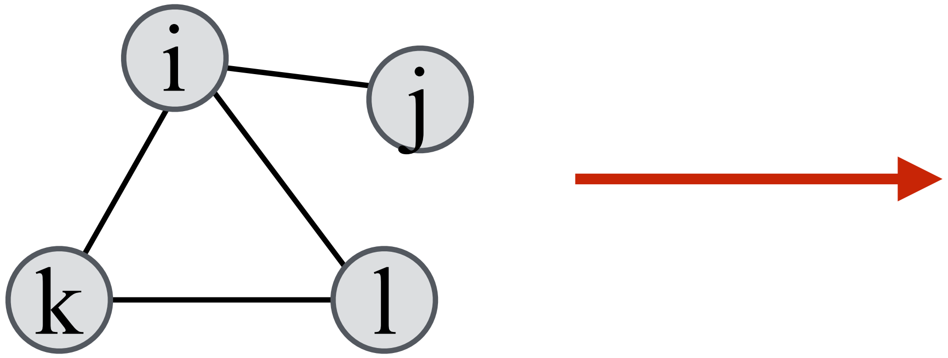
Mostly won't be dealing with these

Another way of referring to these things

We can represent a network using matrix notation:

$a_{ij} = 1$ if i and j are connected

$a_{ij} = 0$ otherwise



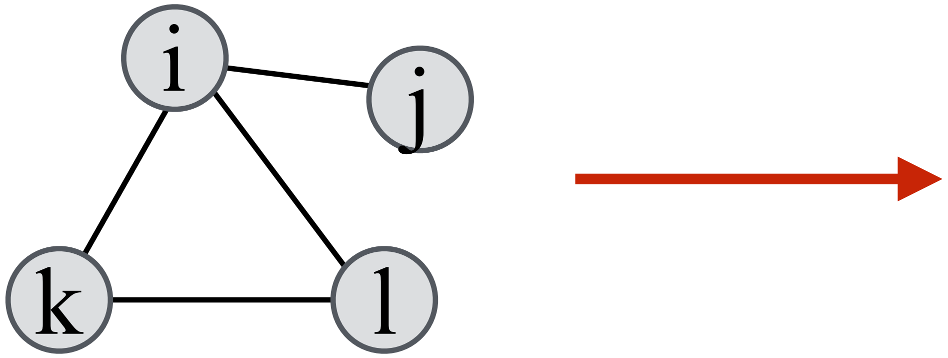
$$\begin{array}{c|cccc|} & i & j & k & l \\ \hline i & 0 & 1 & 1 & 1 \\ j & 1 & 0 & 0 & 0 \\ k & 1 & 0 & 0 & 1 \\ l & 1 & 0 & 1 & 0 \\ \hline \end{array}$$

We call that an *adjacency matrix*

Many of the measures we'll talk about can be calculated by manipulating the adjacency matrix

Another way of referring to these things

If there are no self-loops, then $a_{ii} = 0$



$$\begin{array}{c|cccc|} & i & j & k & l \\ \hline i & 0 & 1 & 1 & 1 \\ j & 1 & 0 & 0 & 0 \\ k & 1 & 0 & 0 & 1 \\ l & 1 & 0 & 1 & 0 \\ \hline \end{array}$$

Kinds of Links

Unweighted (binary): there is either an edge between two nodes, or there is not

Weighted: the edge can have a "strength"

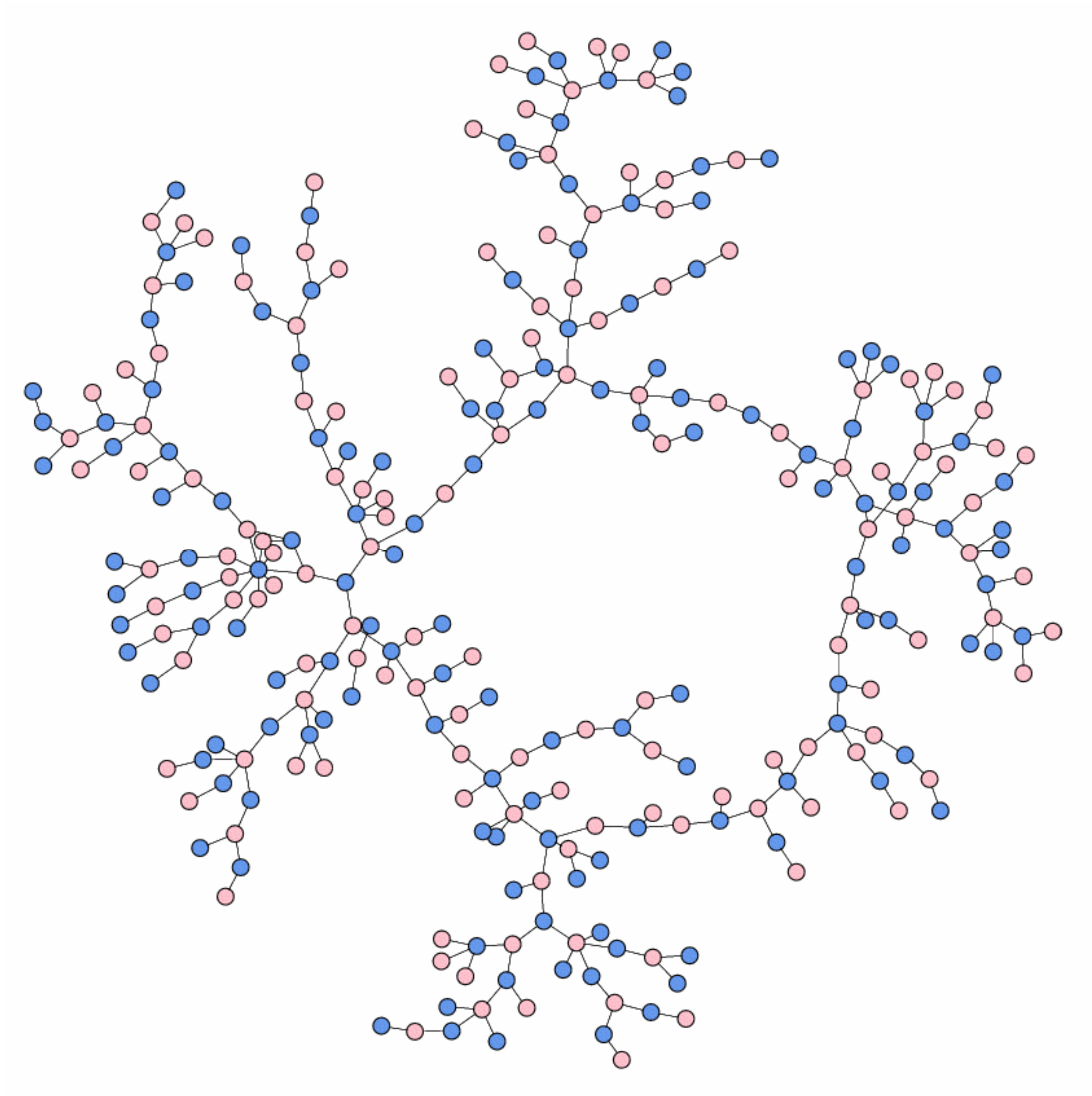
Link weight is w_{ij} for link ij, or in matrix notation:
$a_{ij} = w_{ij}$

Undirected: if there is a link from i to j, then there is one from j to i: $w_{ij} = w_{ji}$

Directed: there may be a link from i to j, but not from j to i: $w_{ij} \neq w_{ji}$

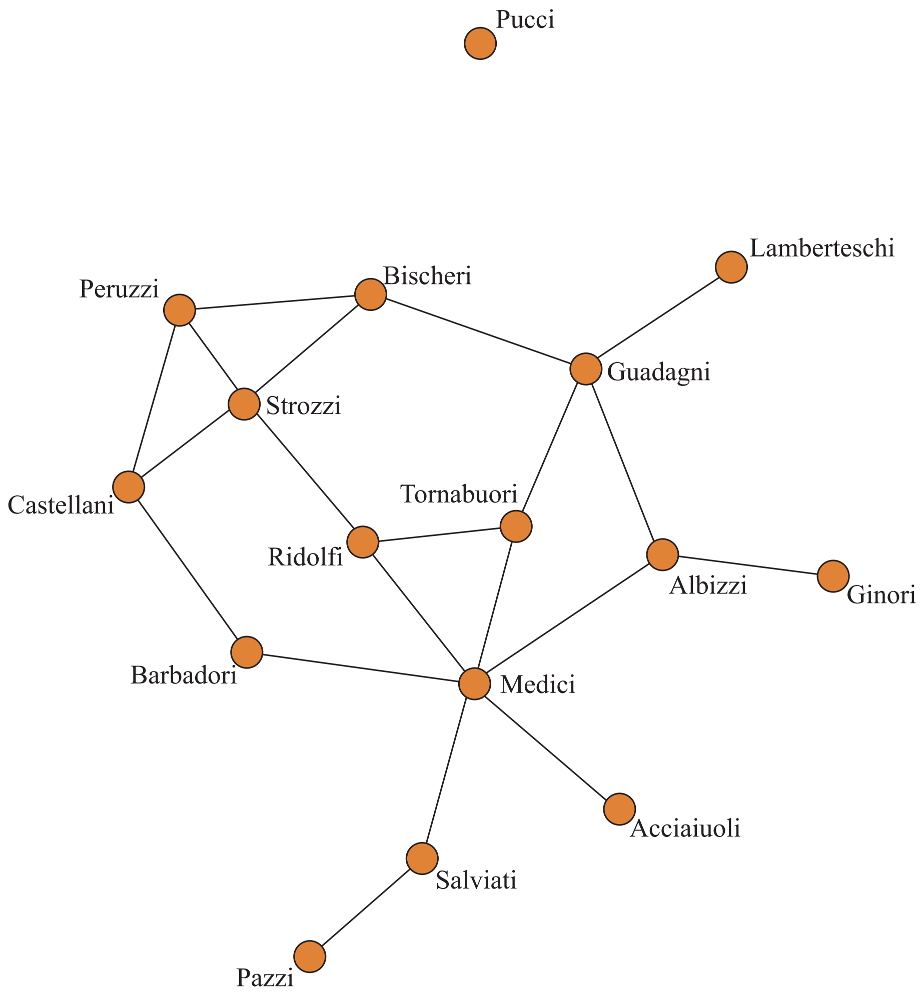
A taxonomy of networks

Unweighted and undirected: links are binary ($w_{AB} = 0$ or 1) and mutual (symmetric matrix: $w_{AB} = w_{BA}$)



High School Dating

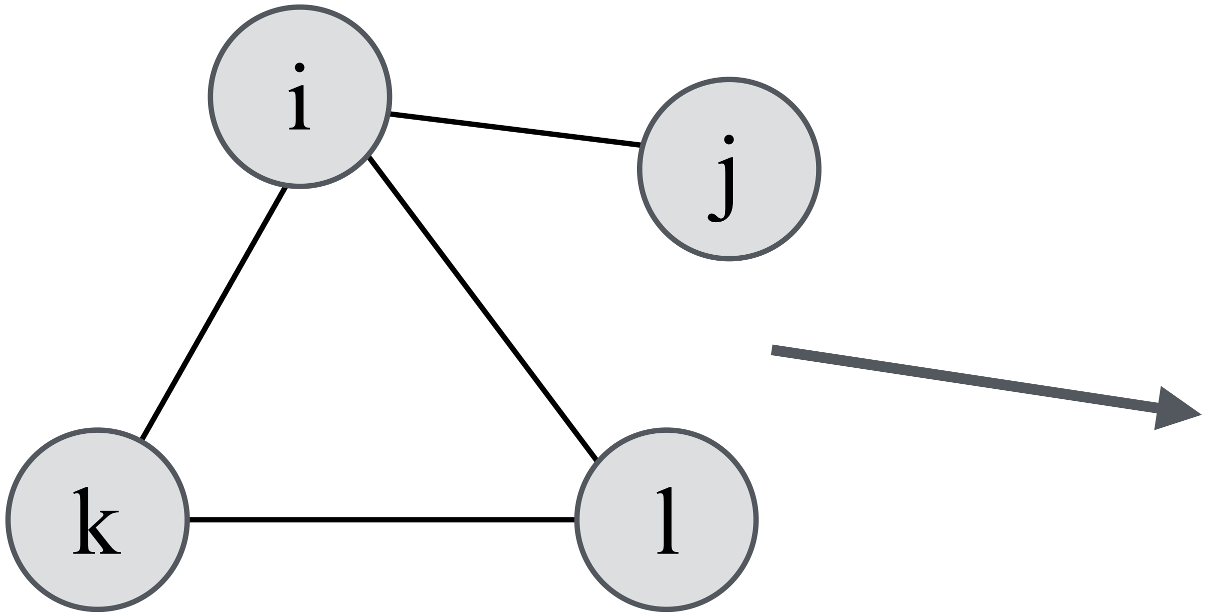
ref: Data by Bearman et al (2004)
Graphic by M.E.J. Newman



Florentine Marriage

Kinds of Links

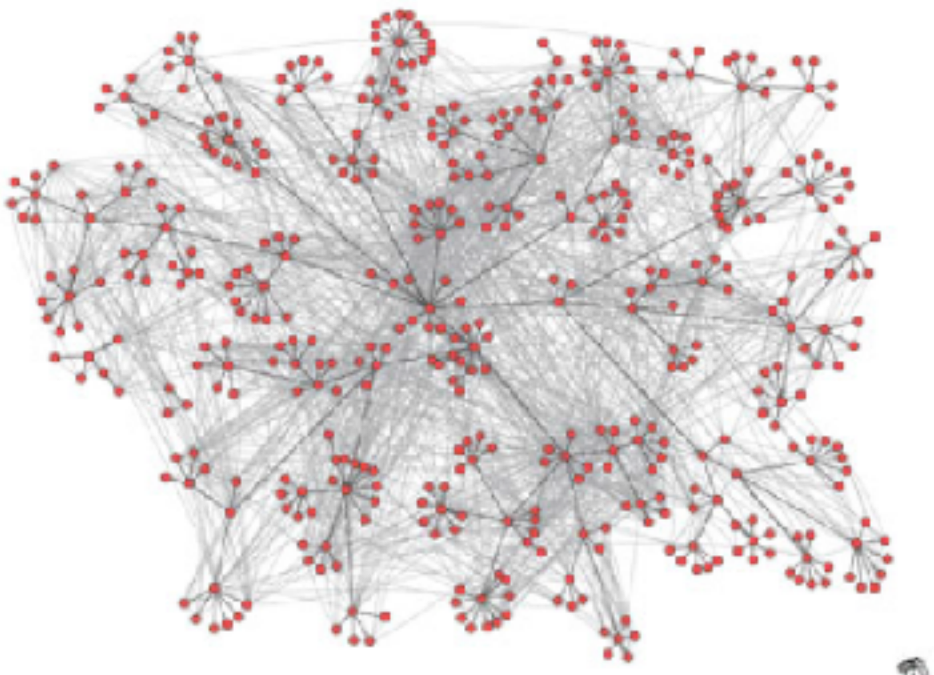
Unweighted and Undirected



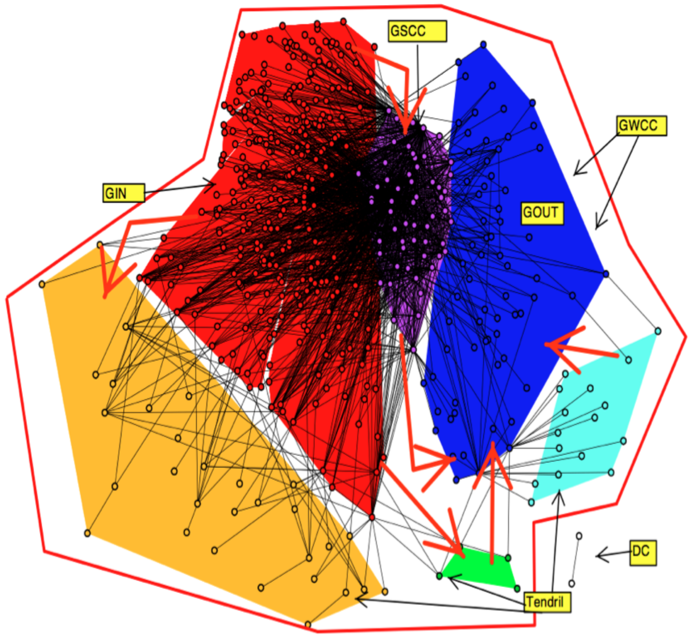
$$\begin{array}{c|cccc} & i & j & k & l \\ \hline i & 0 & 1 & 1 & 1 \\ j & 1 & 0 & 0 & 0 \\ k & 1 & 0 & 0 & 1 \\ l & 1 & 0 & 1 & 0 \end{array}$$

A taxonomy of networks

Weighted: some ties are “stronger” than others ($0 \leq w_{AB} \leq 1$)

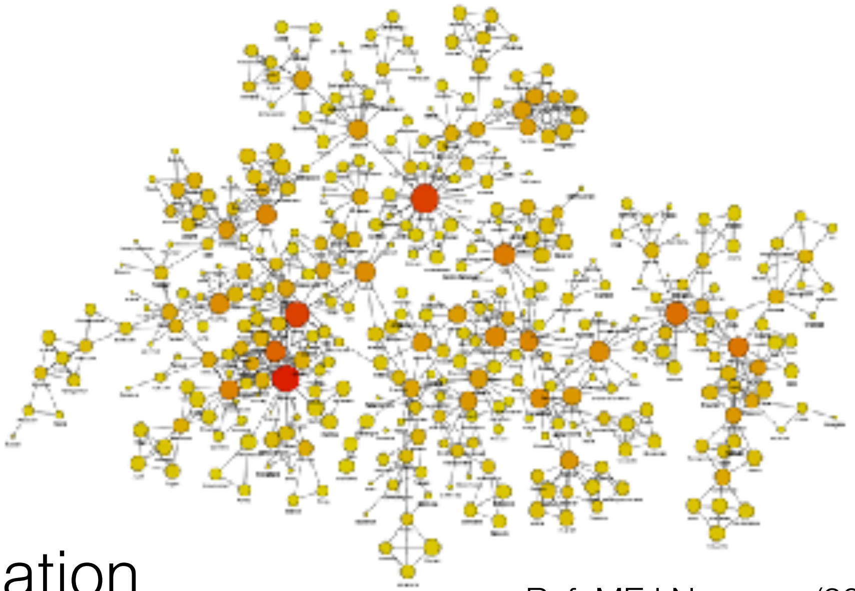


HP email ref: Adamic



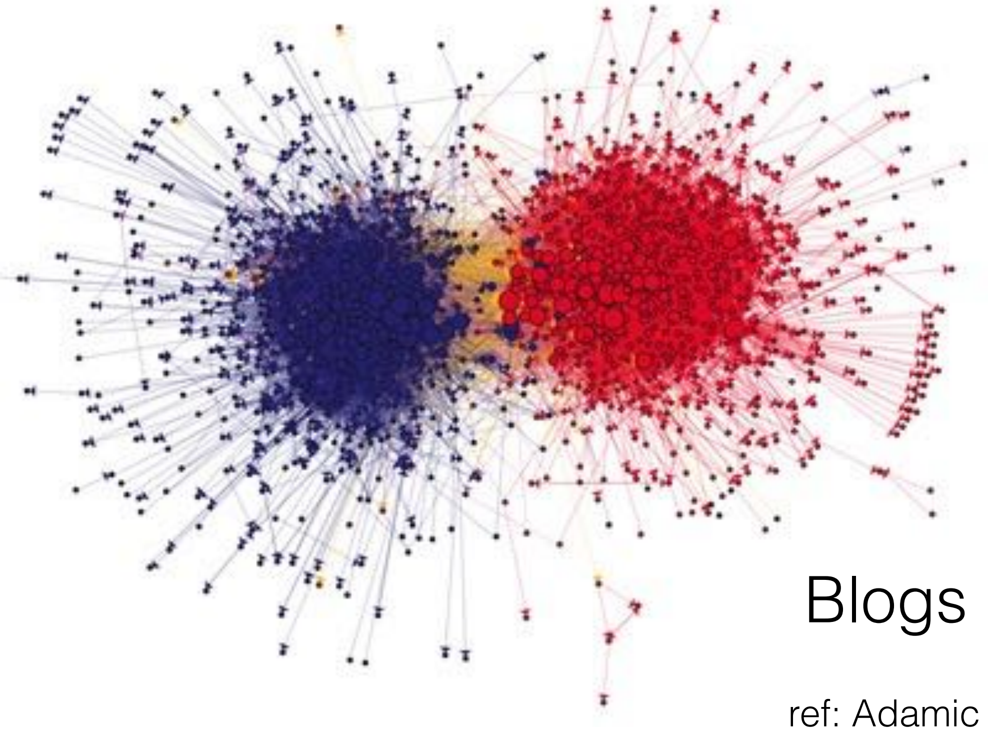
Financial

ref: Bech and Atalay 2008



Collaboration

Ref: MEJ Newman (2006)

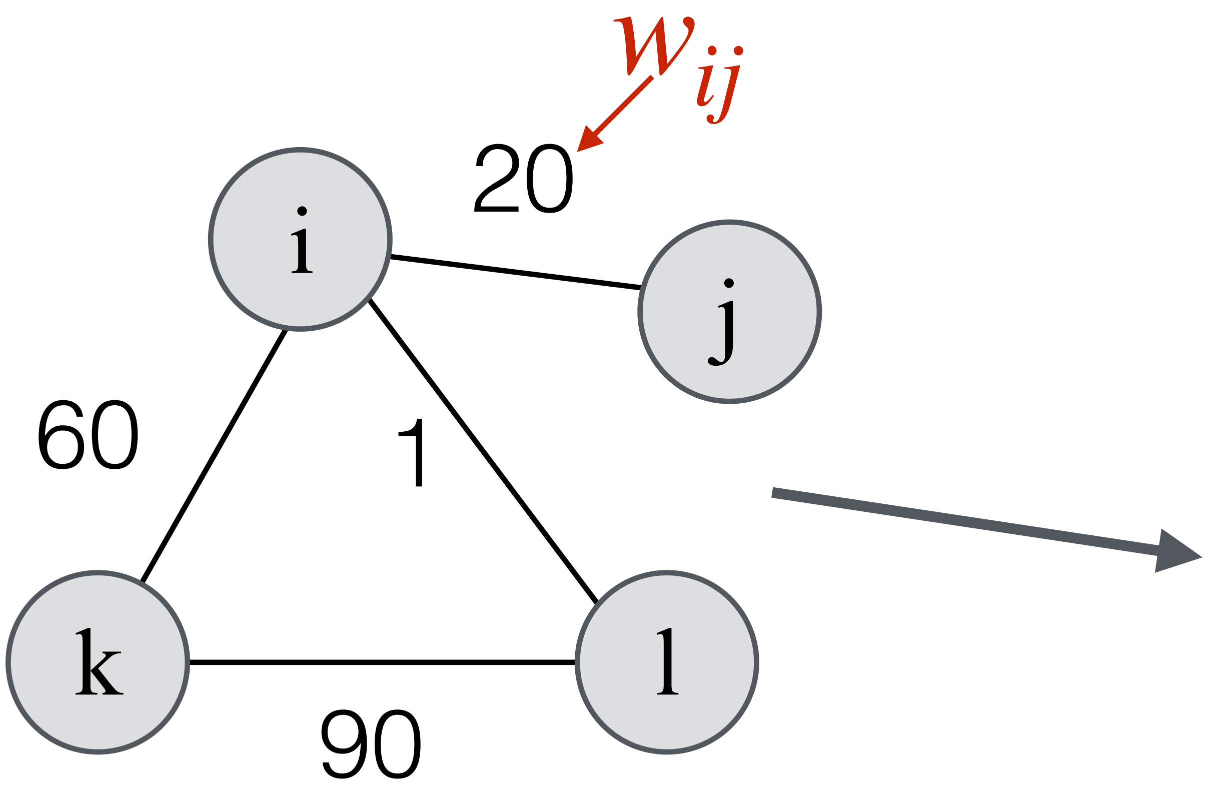


Blogs

ref: Adamic

Kinds of Links

Weighted and Undirected

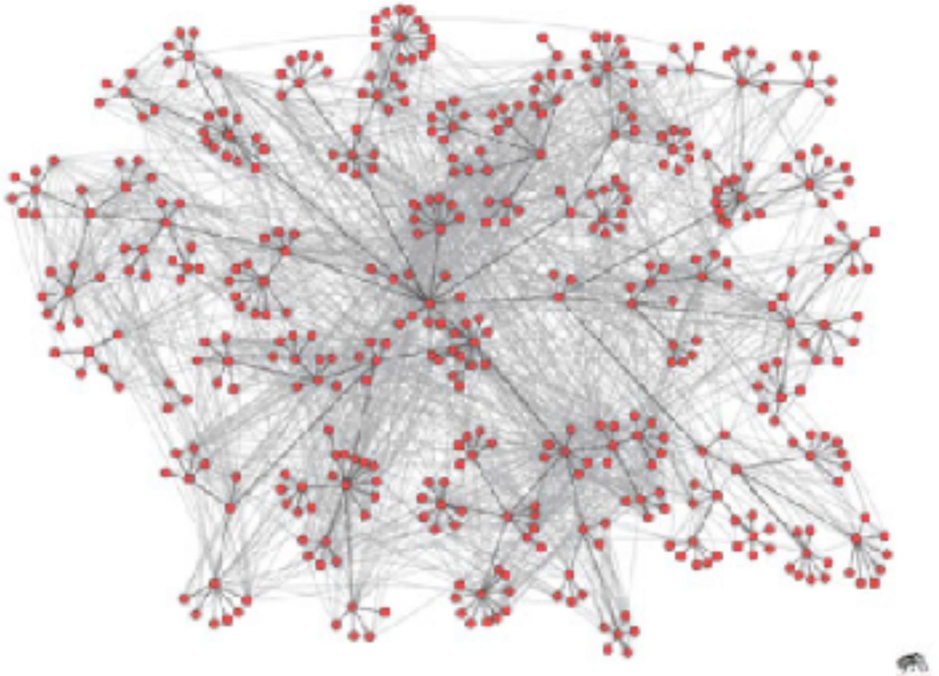


$$\begin{array}{c|cccc} & i & j & k & l \\ \hline i & 0 & 20 & 60 & 1 \\ j & 20 & 0 & 0 & 0 \\ k & 60 & 0 & 0 & 90 \\ l & 1 & 0 & 90 & 0 \end{array}$$

$$a_{ij} = w_{ij}$$

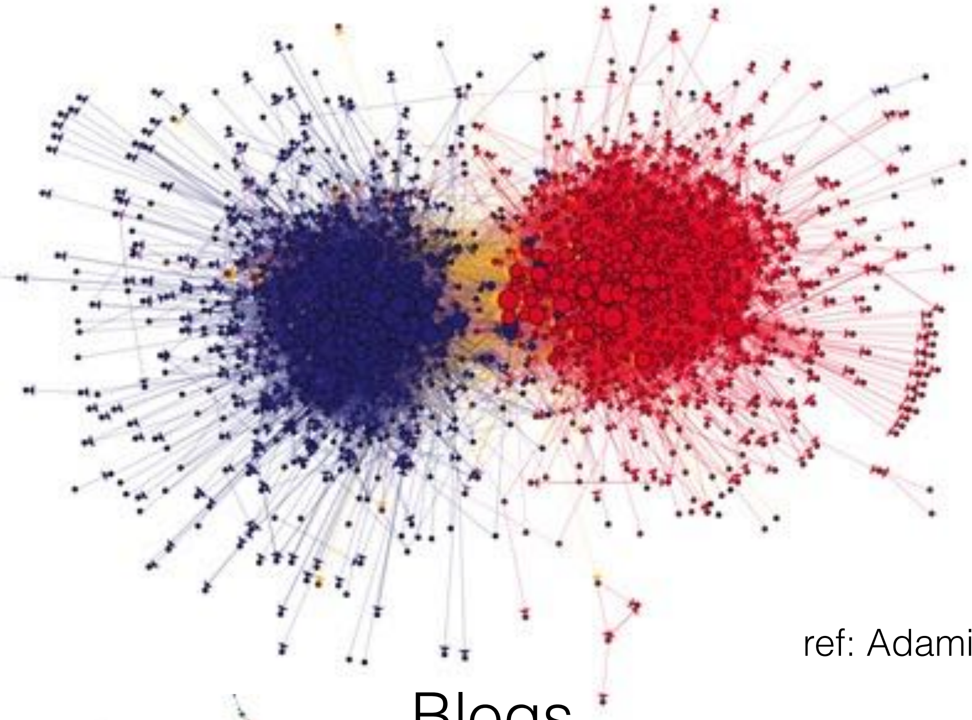
A taxonomy of networks

Directed: A linked to B ⇏ B linked to A (asymmetric adjacency matrix: $w_{AB} \neq w_{BA}$)



ref: Adamic

HP email

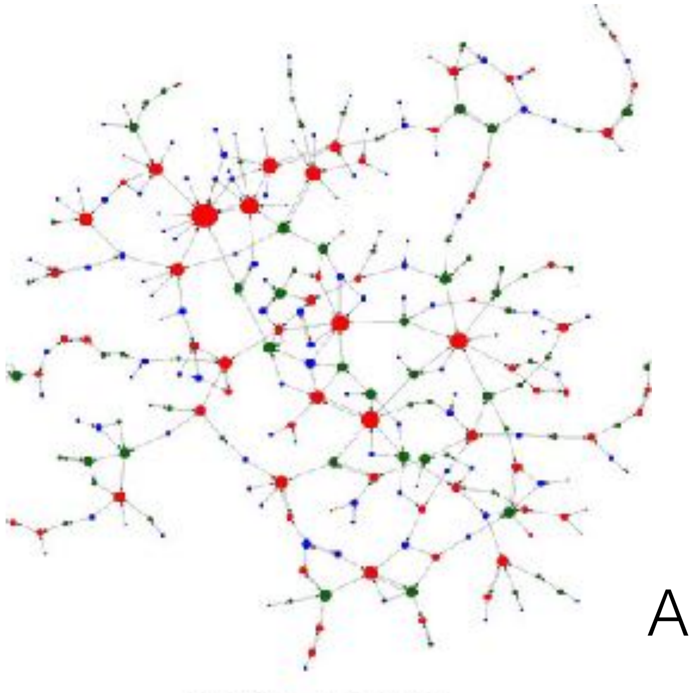


ref: Adamic

Blogs

ref: http://goo.gl/L9ars

Citations

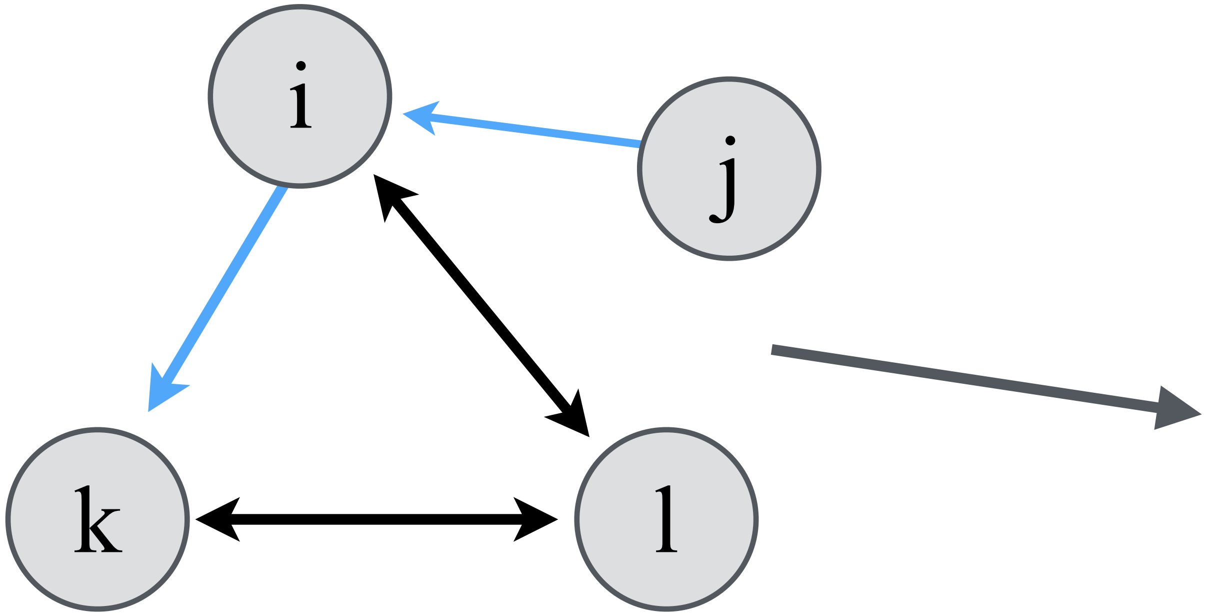


graphic: Wardil and Hauert
data: Jackson et al

Advice

Kinds of Links

Unweighted and Directed



$$\begin{array}{c|cccc} & i & j & k & l \\ \hline i & 0 & 0 & 1 & 1 \\ j & 1 & 0 & 0 & 0 \\ k & 0 & 0 & 0 & 1 \\ l & 1 & 0 & 1 & 0 \end{array}$$

$a_{ij} = 1$ if i connects to j

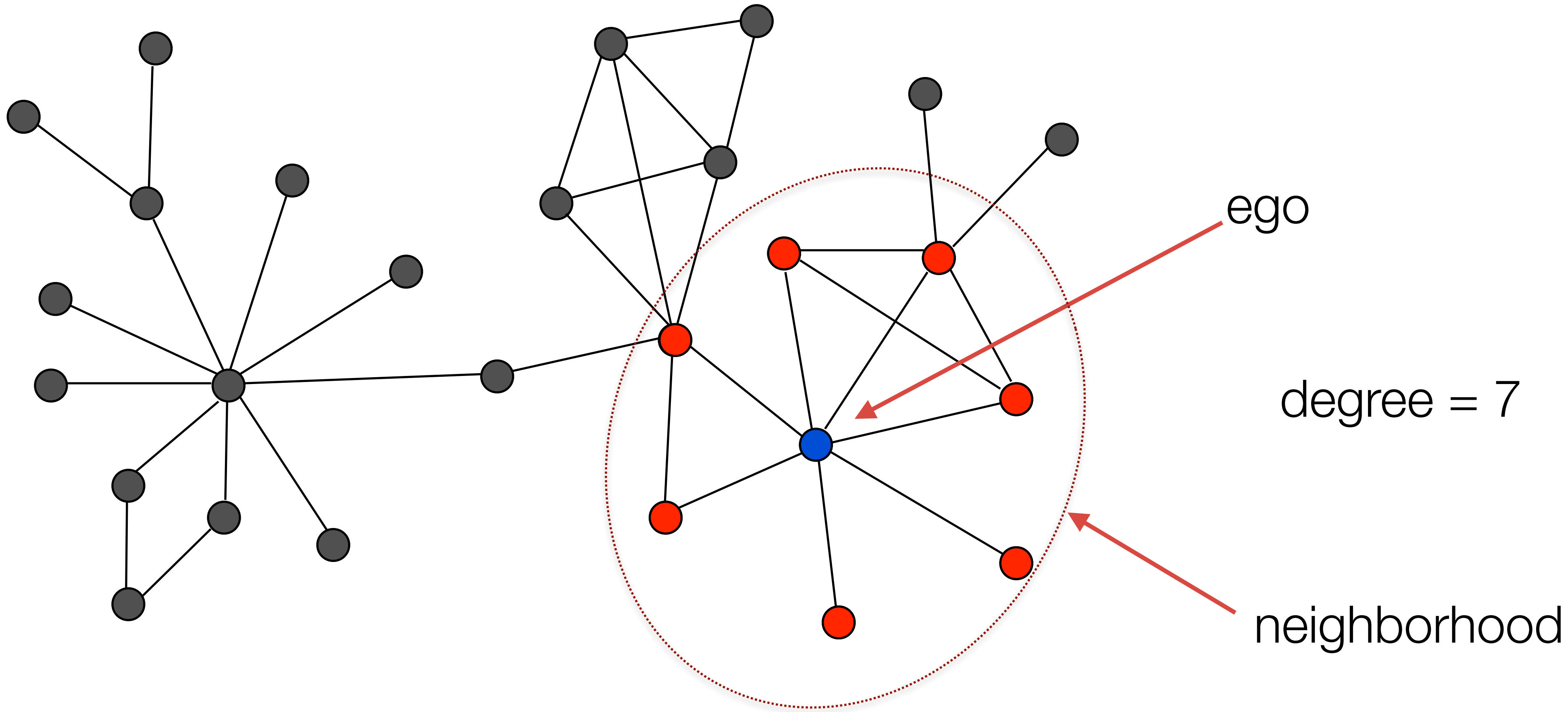
$a_{ij} = 0$ otherwise

A Node’s Neighborhood: Degree

Ego = any single node: i

Neighborhood = the set of nodes ego is connected to: n_i

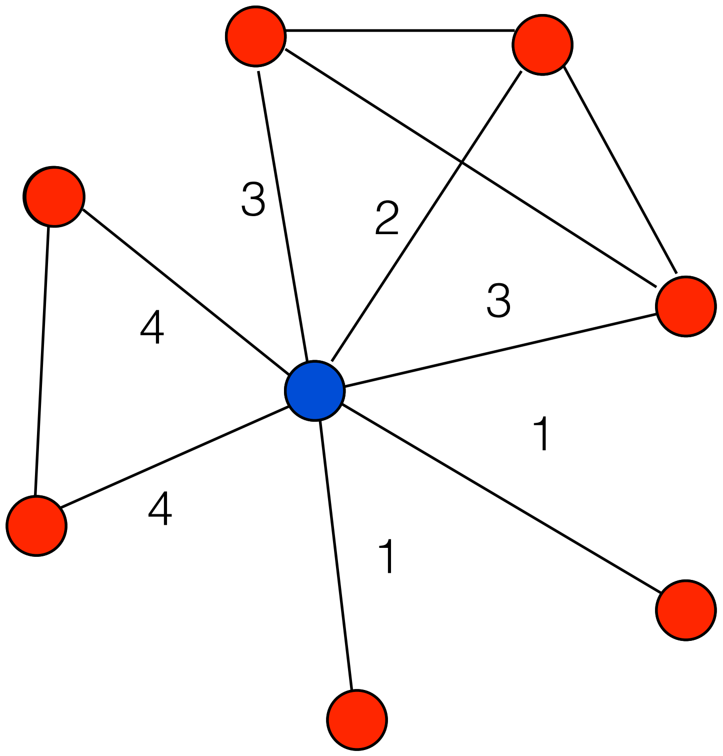
Degree = the number of nodes ego is connected to: $|n_i|$



Degree in a Weighted Network

In a weighted network, there is a second measure of degree: weighted degree:

$$d_i^W = \sum_j w_{ij}$$



weighted degree = 18

Weighted degree tells you something different about nodes than degree does

What does degree mean in an email network? Weighted degree?

Degree in a Directed Network

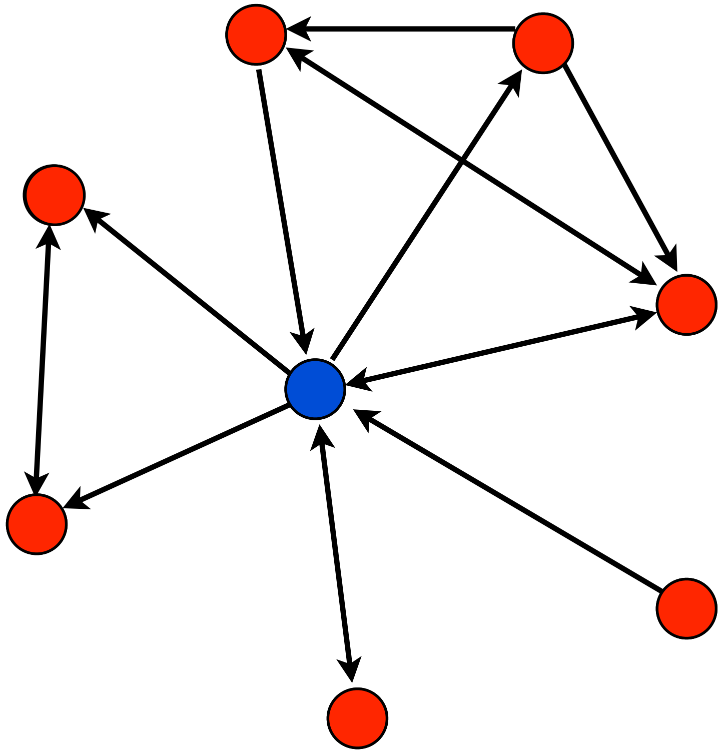
When links are directed, there are two measures of degree:

in-degree = number of nodes who link to ego

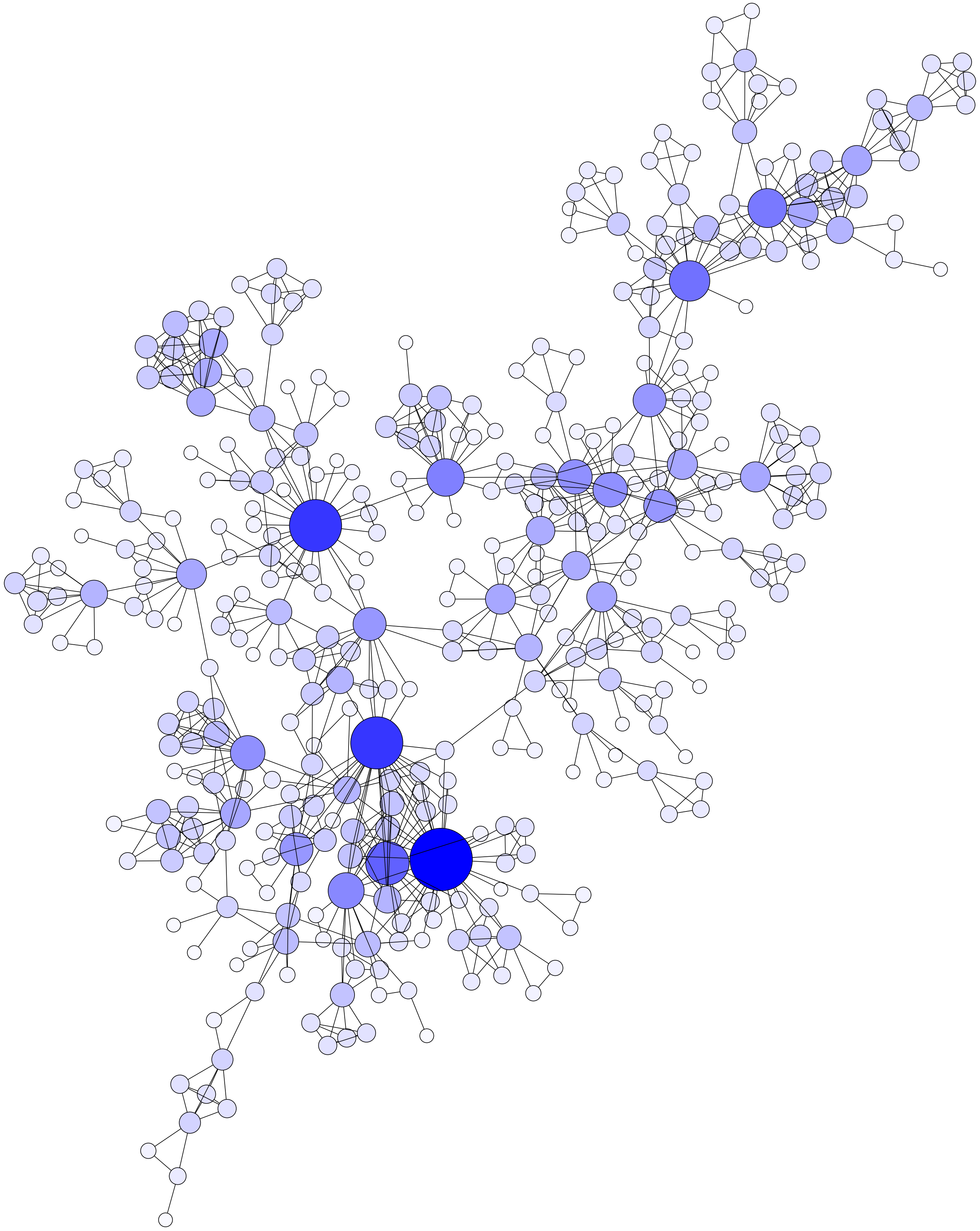
$$d_i^I = \sum_j w_{ij}$$

out-degree = number of nodes ego links to

$$d_i^O = \sum_j w_{ji}$$



in-degree = 4
out-degree = 5

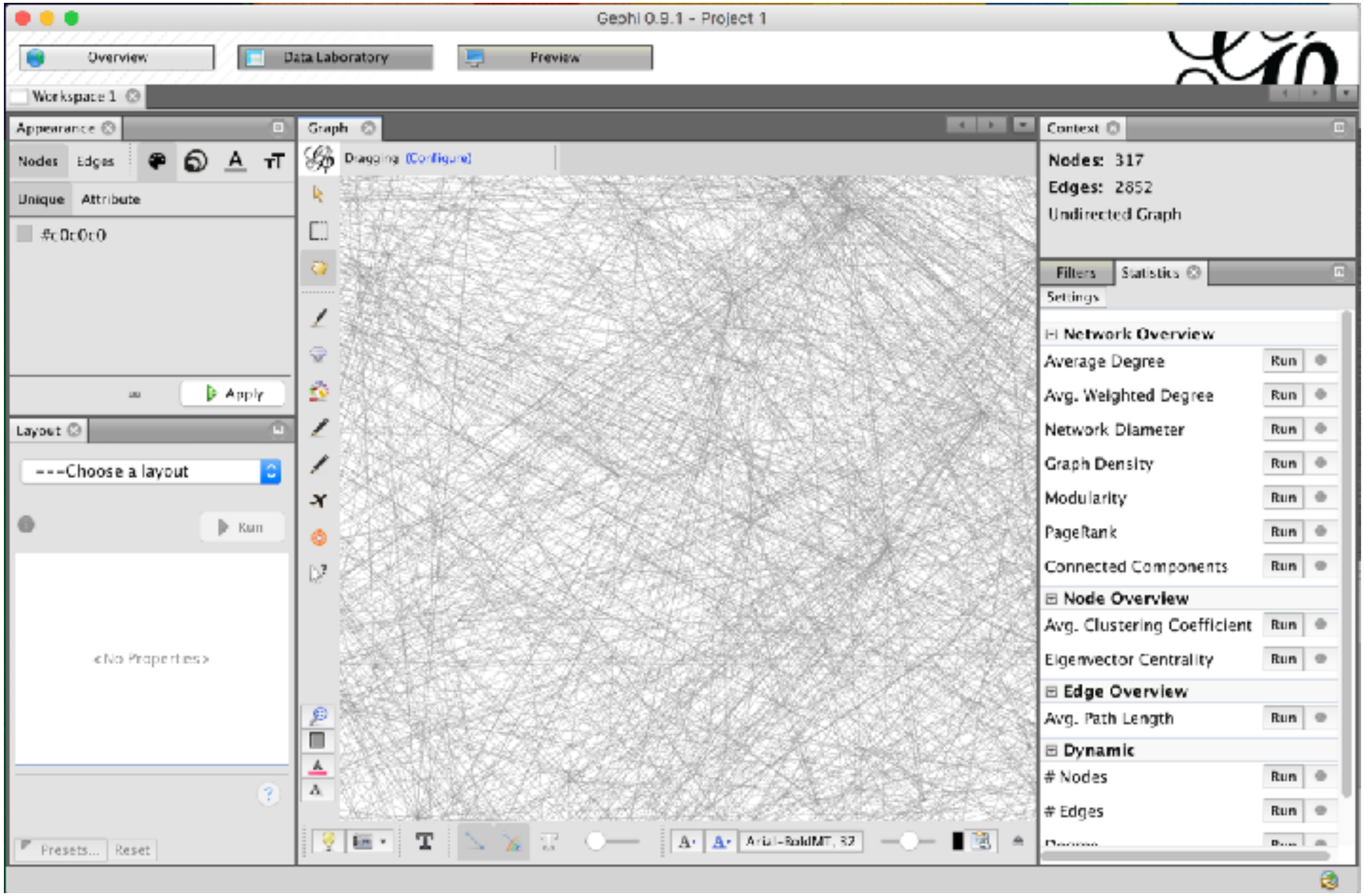


Gephi!

Instructions

- Start with “DrAFacebookWithAttributes.gml” (on Canvas)— that is data pulled from my personal Facebook account
- Open Gephi (close that start up window for now)
- Load the data file into Gephi (File > Open)

Overview Page



Gephi 0.9.1 - Project 1

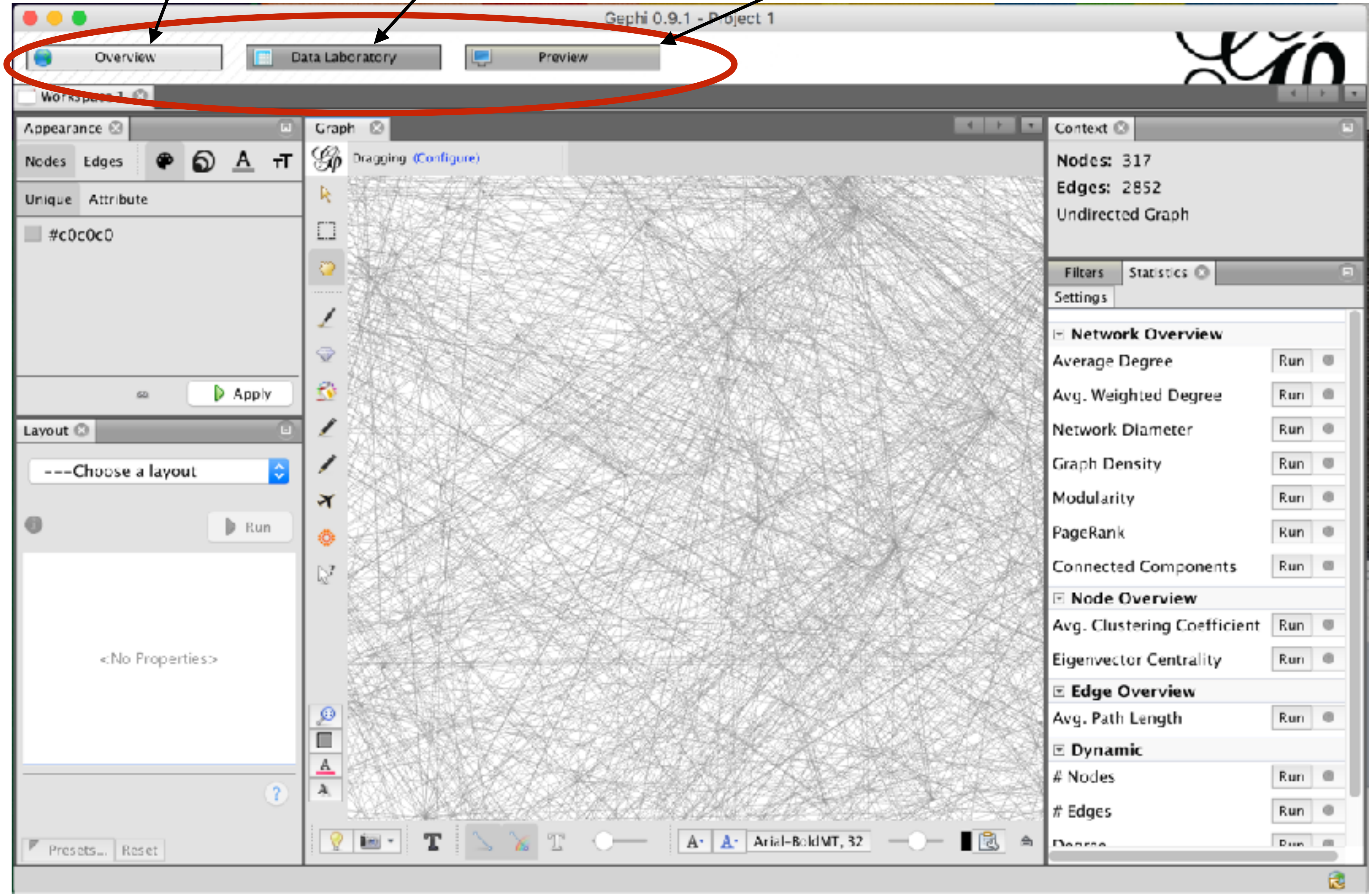
Nodes: 317
Edges: 2852
Undirected Graph

It should look roughly like this.
If not: go to >Window to add what is missing

Basics

Look at data

Create diagrams



Nodes: 317
Edges: 2852
Undirected Graph

Data

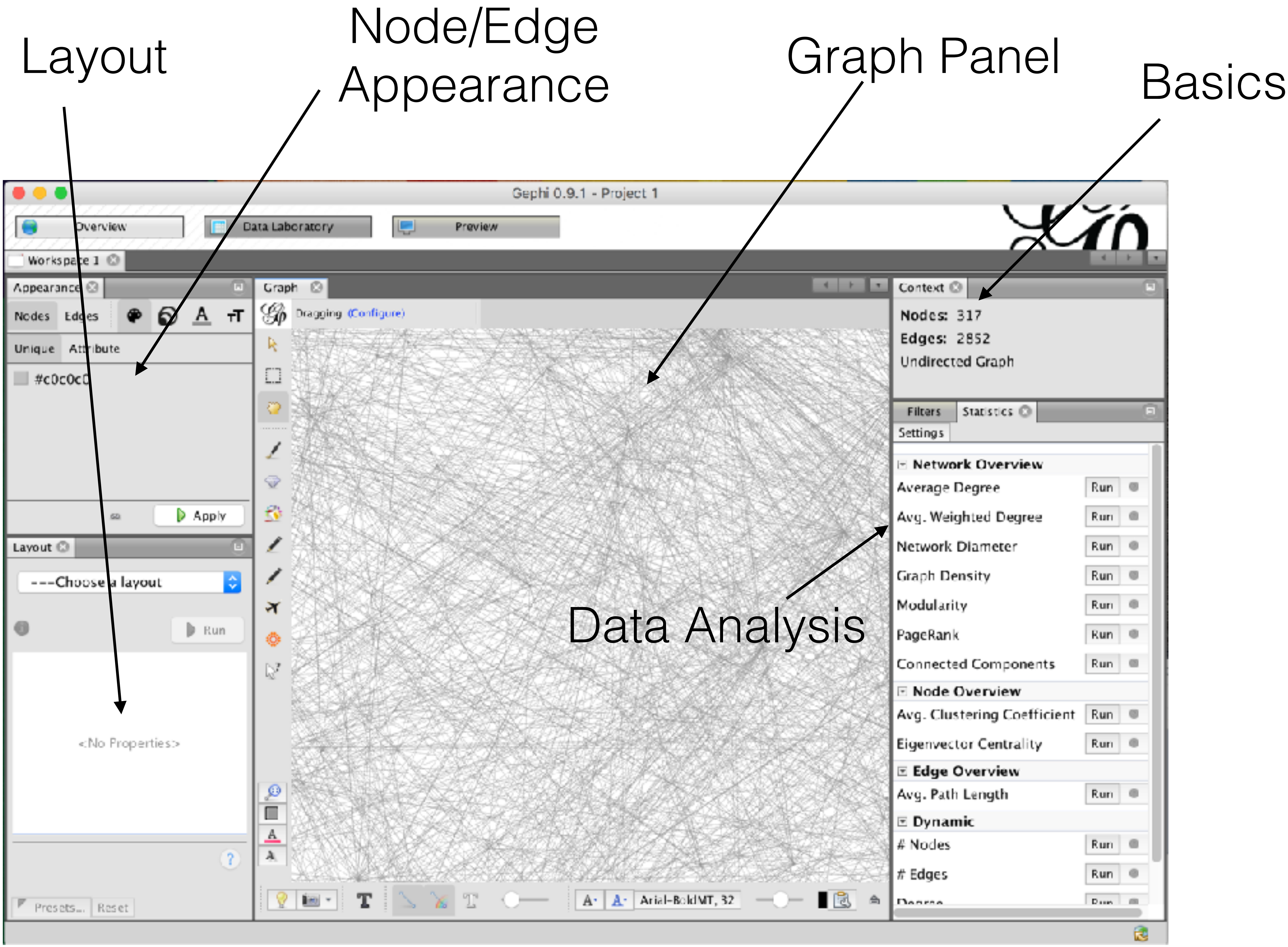
Now, go to the “data laboratory” button at the top of the window

This pane gives you a look at the underlying data (a spreadsheet)

We are looking at the nodes

This data has two extra attributes attached to the people in the data set: gender, and the country they originate from.

- gender: 1 = female, 0 = male
- countries:
    - 0 = US
    - 1 = Canada
    - 2 = Australia
    - 3 = Russia
    - 4 = Germany
    - etc…


Layout

Node/Edge Appearance

Graph Panel

Basics

Data Analysis

Gephi 0.9.1 - Project 1

Nodes: 317
Edges: 2852
Undirected Graph

#c0c0c0

Arial-BoldMT, 32

Overview Page: Graph Panel

- Graph Panel: a visualization of the network, and some basic tools to alter that:
    - Important ones:
        - fist = grabber to grab and drag nodes
        - arrow = select (default: shows the node’s neighborhood)
        - magnifying glass: reset the zoom to the center
    - At bottom:
        - Filled “T” = turn labels on/off (depending on zoom, may not be able to see the labels)
        - Change size/font of labels

Overview Page: Layout Panel

There are lots of ways to lay out a network.

Click the magnifying glass to reset the zoom on the network

- layout is currently random (square)—not very useful

Most layout algorithms are what is called a “force-directed graph drawing”. Basically, it uses physical analogies to layout nodes and edges in a visually pleasing (and hopefully informative) way.

- Nodes repel each other
- Edges pull nodes together (like springs)

Fruchterman Reingold is a good place to start (select from pull-down menu and press “play”)

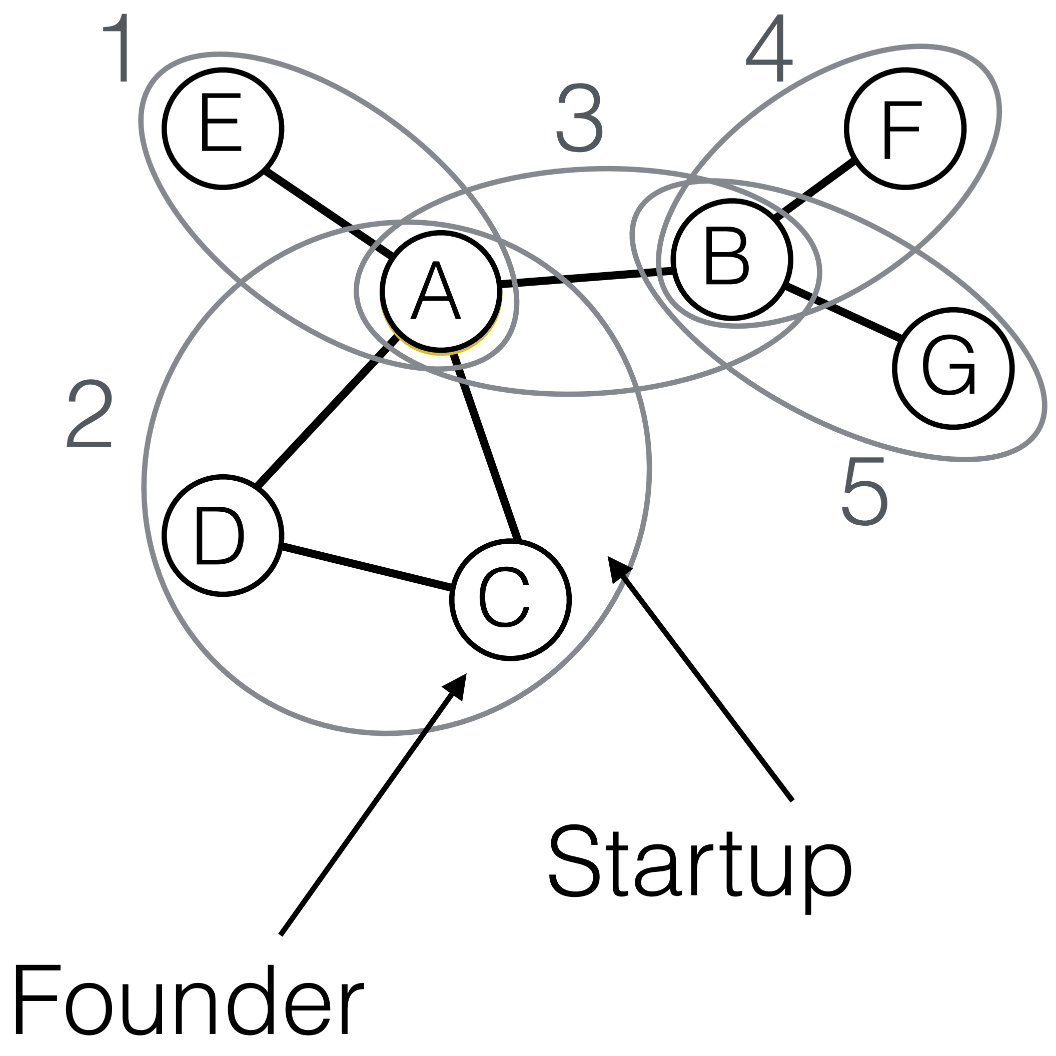
Alternatives: Force Atlas and Force Atlas 2.

Other useful manipulations: expand (>1 = expand and <1 = contract), label adjust (keeps labels from interacting)

NB: with some layout algorithms, you have to press stop

A Different Kind of Network: Bipartite Networks and Hypergraphs

In a *hypergraph*, groups of nodes are connected

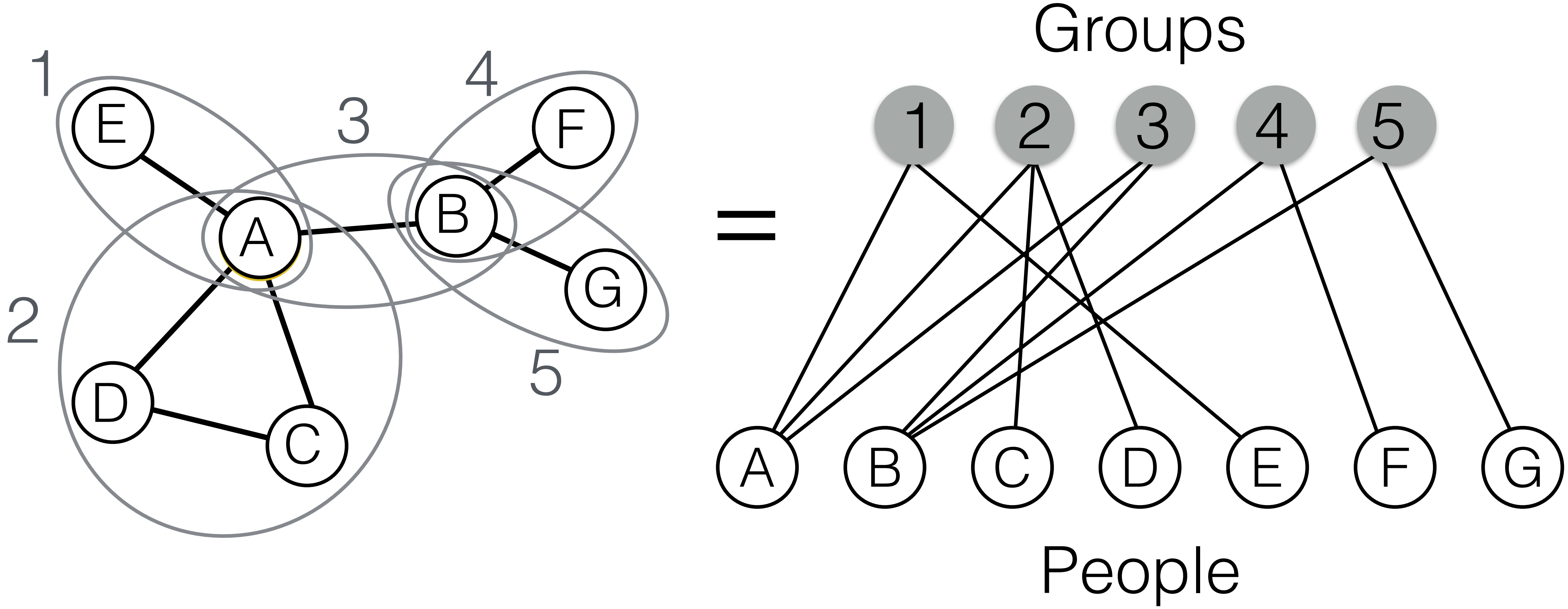


Examples:

- Coauthors on papers
- Members of clubs/ organizations
- Courses
- Teams
- Founders of startups

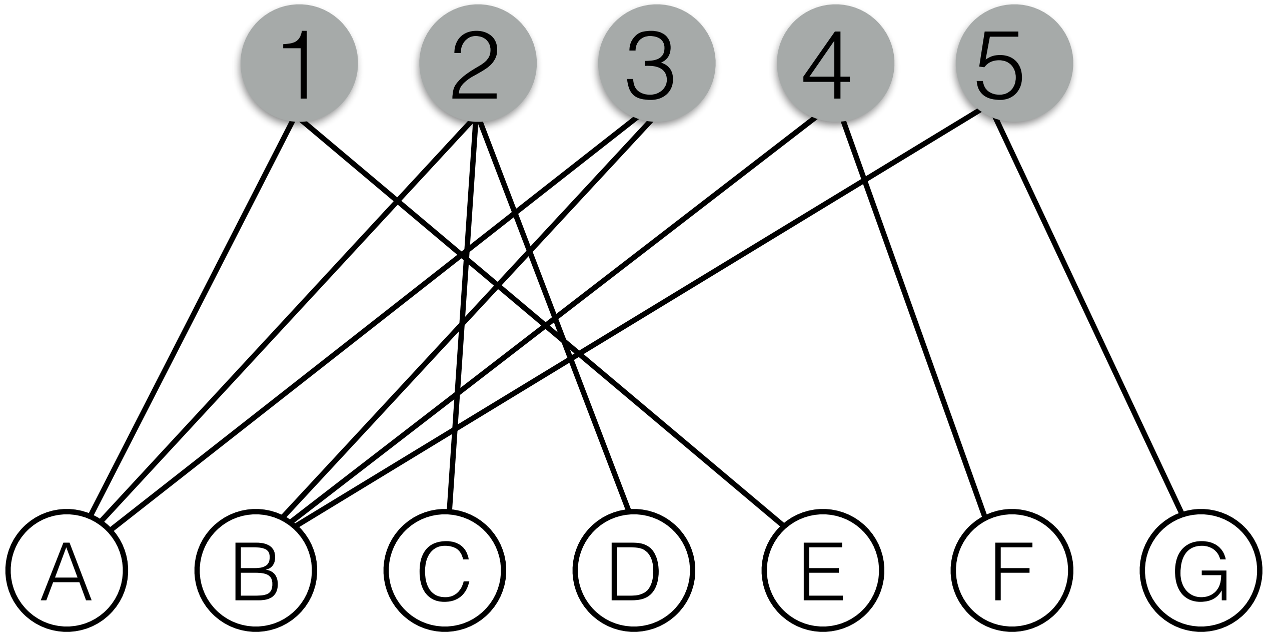
Bipartite Networks and Hypergraphs

You can represent this network as a bipartite ("two-mode") network, with two types of nodes



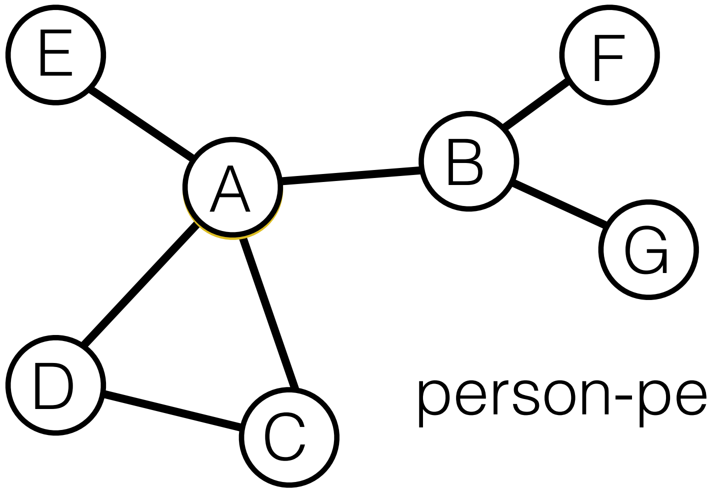
Bipartite Networks and Hypergraphs

You can project a bipartite network onto two kinds of one-mode networks



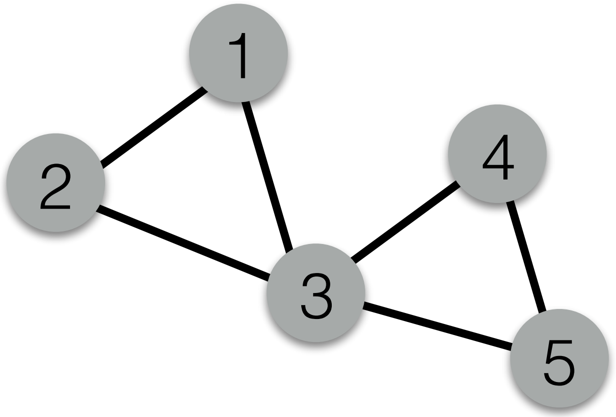
bipartite

=



person-person

=



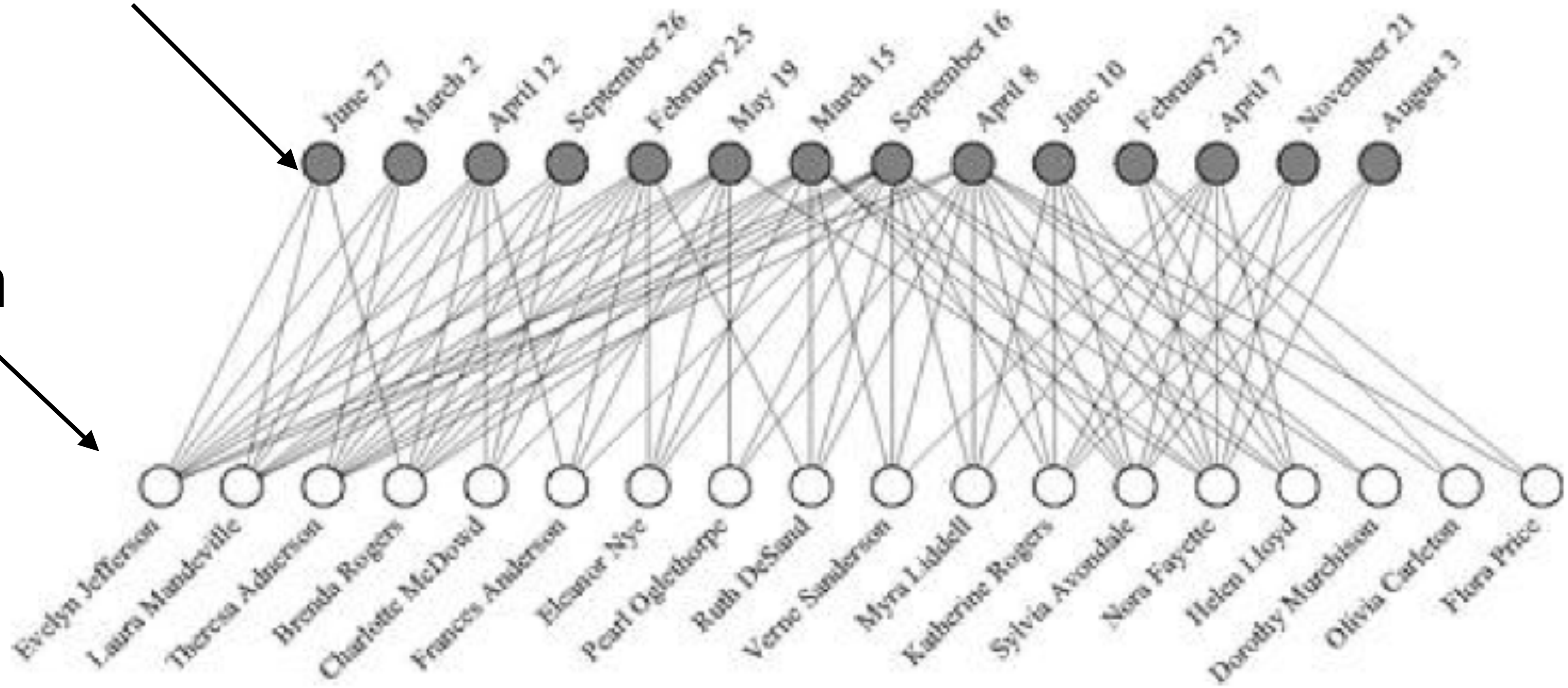
group-group

Both lose some information!

Bipartite

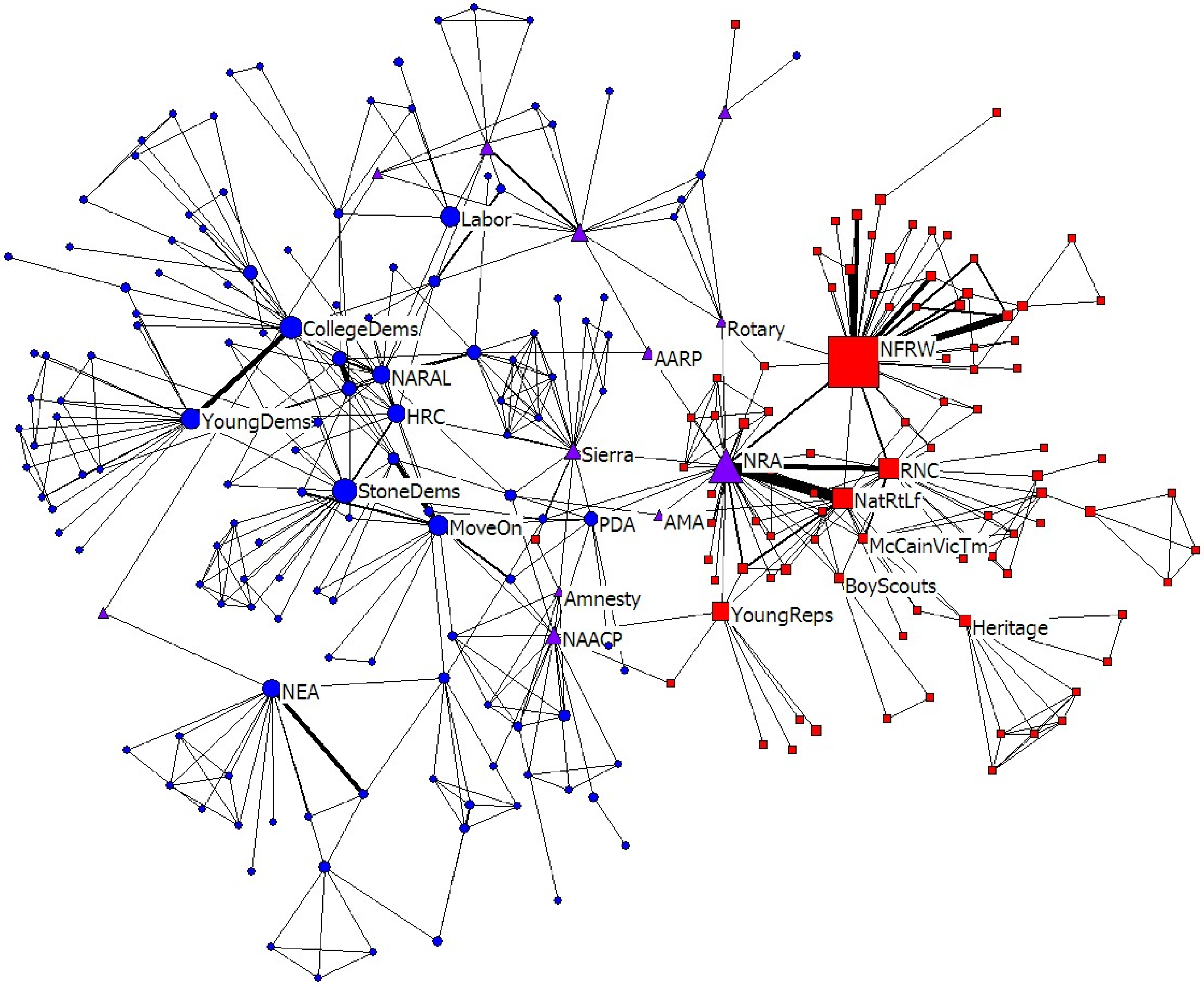
Social Events

Women



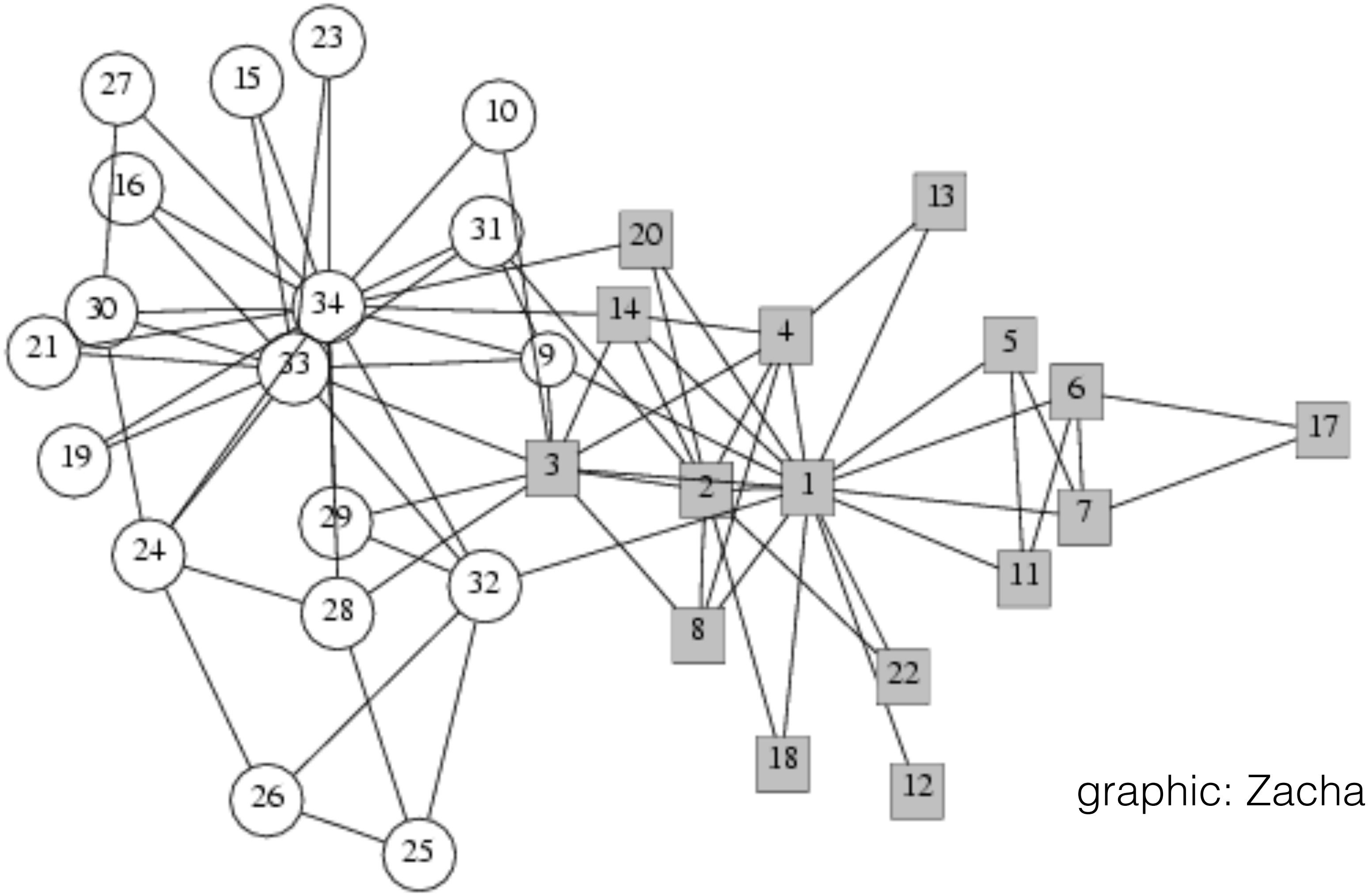
Examples: The Southern Women Network

One-mode Projection



ref: Masket et al (2009)

Examples: US delegate co-membership



graphic: Zachary

Data Collection

Collecting Data about Empirical Social Networks

Some things that make collecting data on social networks difficult

- It can be very expensive
- Links are subjective
- It can be difficult to decide where your network starts and stops
- Sometimes the interactions you care about are not the ones you can observe

Collecting Data about Empirical Social Networks

Methods:

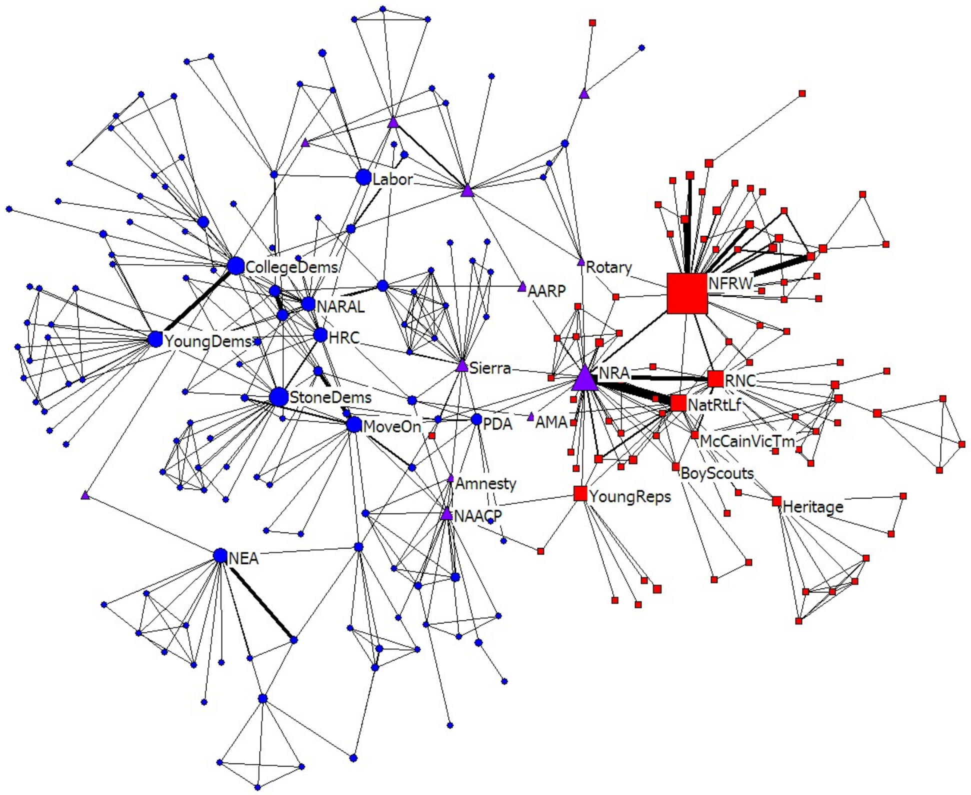
- Surveys and Interviews
- Direct Observation
- Passive Data Collection/Archival Records

Option 1: Surveys and Interviews

Just ask people about their network connections

Examples:

- Indian villages
- Political co-membership
- Friends, Dating, Sexual Contacts



ref: Masket et al (2009)

- Our class social networks

Option 1: Surveys and Interviews

Just ask people about their network connections

An example: jr. high school friendship networks

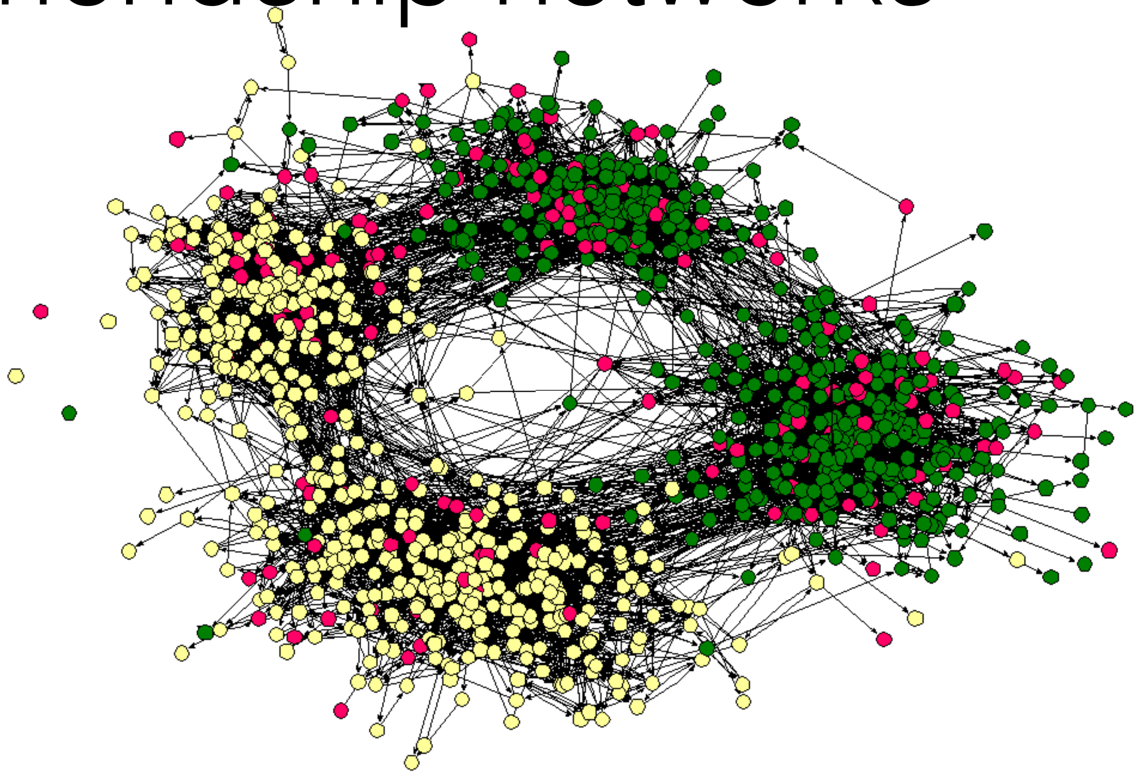
1) Who is your best friend?

2) Who is your second best friend?

3) Who is your third best friend?

…

8) Who is your eighth best friend?



ref: Moody

This kind of question is called a *name generator*

Surveys and Interviews

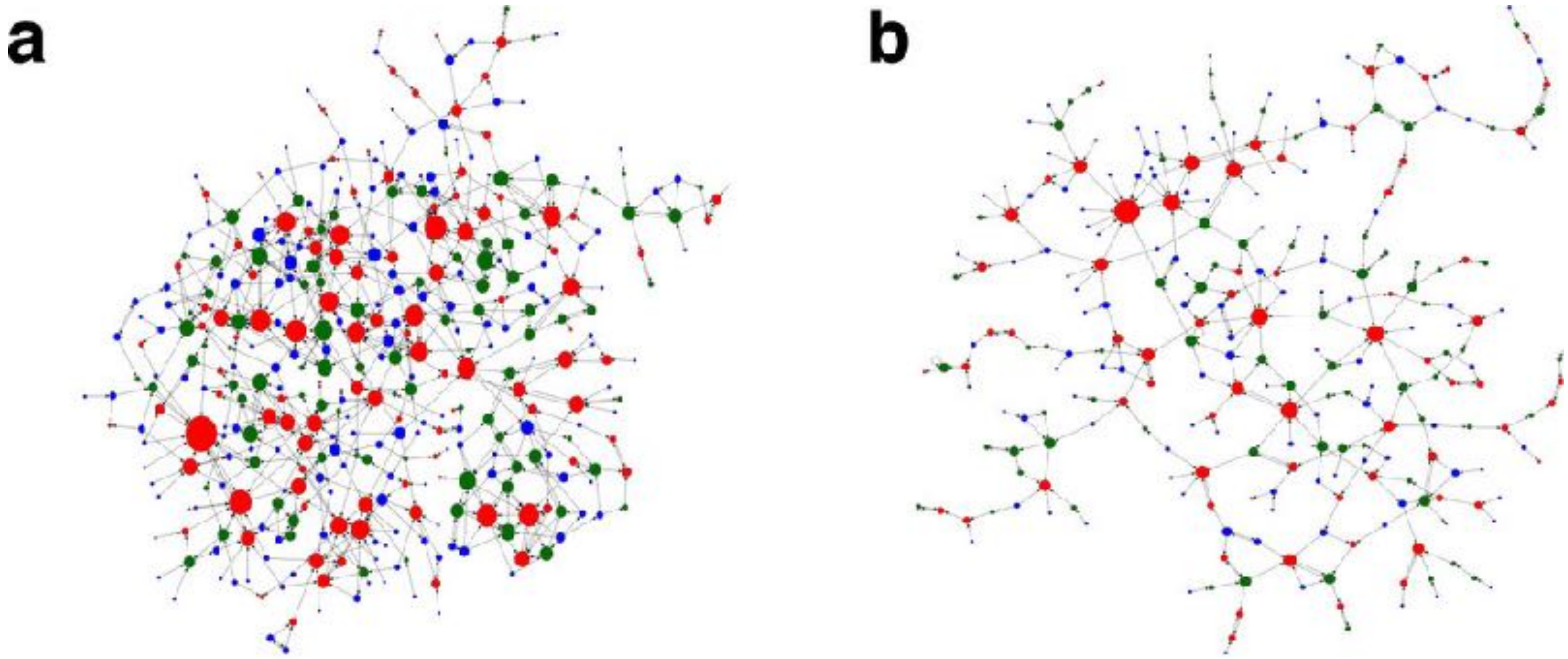
Another example: India village data

Data collected from 75 rural Indian villages

Asked about the following (among many others):

- Who do you go to visit?
- Who comes to visit you?
- Who would you borrow rice from?
- Who would you borrow money from?
- Who would you go to for advice?

Surveys and Interviews



rice/kerosene network

advice network

graphic: Wardil and Hauert
data: Jackson et al

Surveys and Interviews

Advantages:

- Can collect data on multiple kinds of connection
- Can ask about the most relevant type of connection
- Can collect demographic information

Disadvantages:

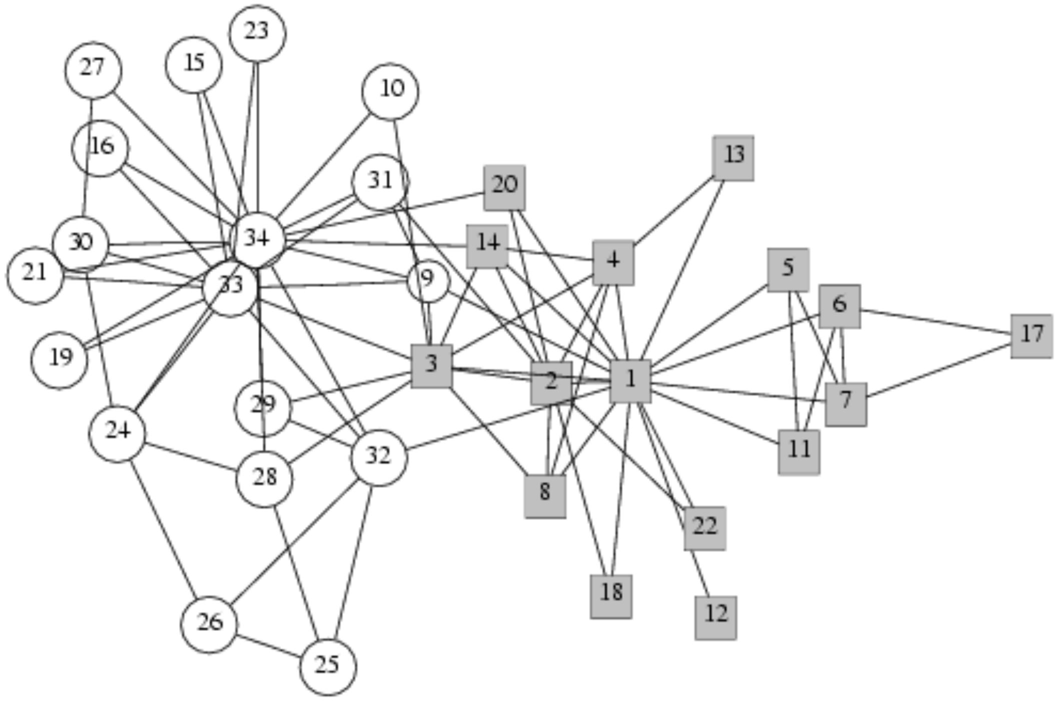
- Labor intensive and costly
- Limited to small groups or small samples
- Link definitions are subjective
- People are very bad at recalling their network connections!

Option 2: Direct Observation

Watch the individuals and note the duration/ frequency of interactions

Examples:

- Karate Club Network
- Macaque Network
- Office Communication
- Conference Interaction



graphic: Zachary

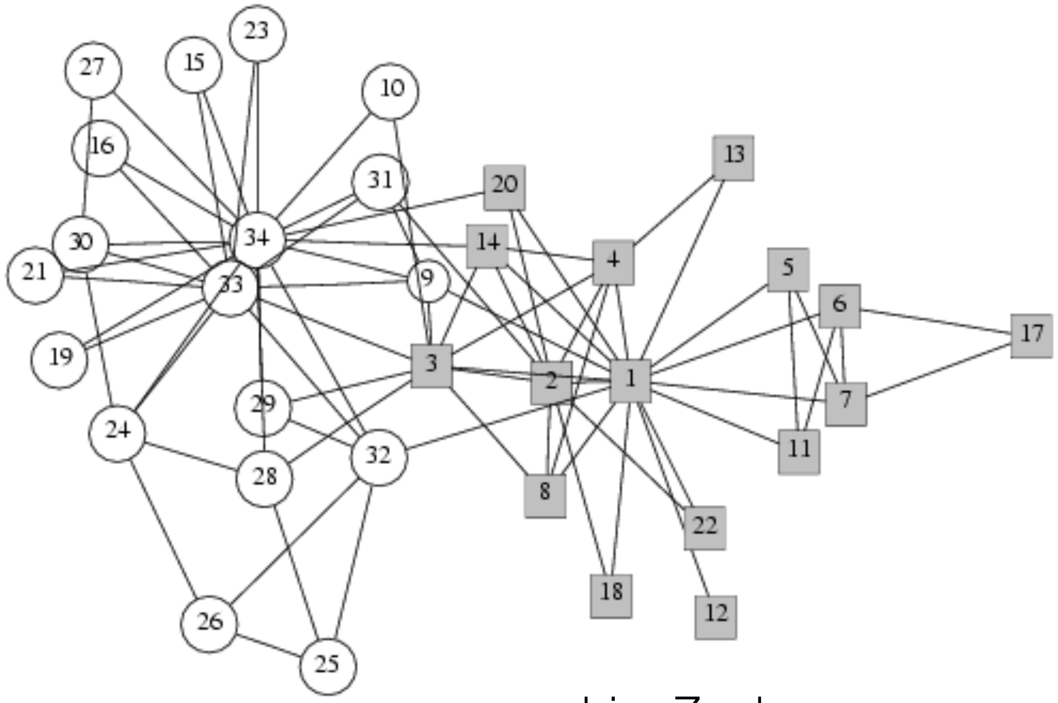
Traditionally: pen and paper

More recently: sociometric badges and other tech

Direct Observation

An example: Zachary Karate Club Network

- 1970s study by Wayne Zachary
- Direct observations of interactions within a university karate club over two years



graphic: Zachary

- Interesting fact: during the study, the members of the club had a falling out and split into two parts

Direct Observation

Another example: Macaque Networks

- Direct observation of Macaque social groups
- Behaviors observed: grooming, playing
- By physically removing an individual from the group, they intuit a third behavior: policing

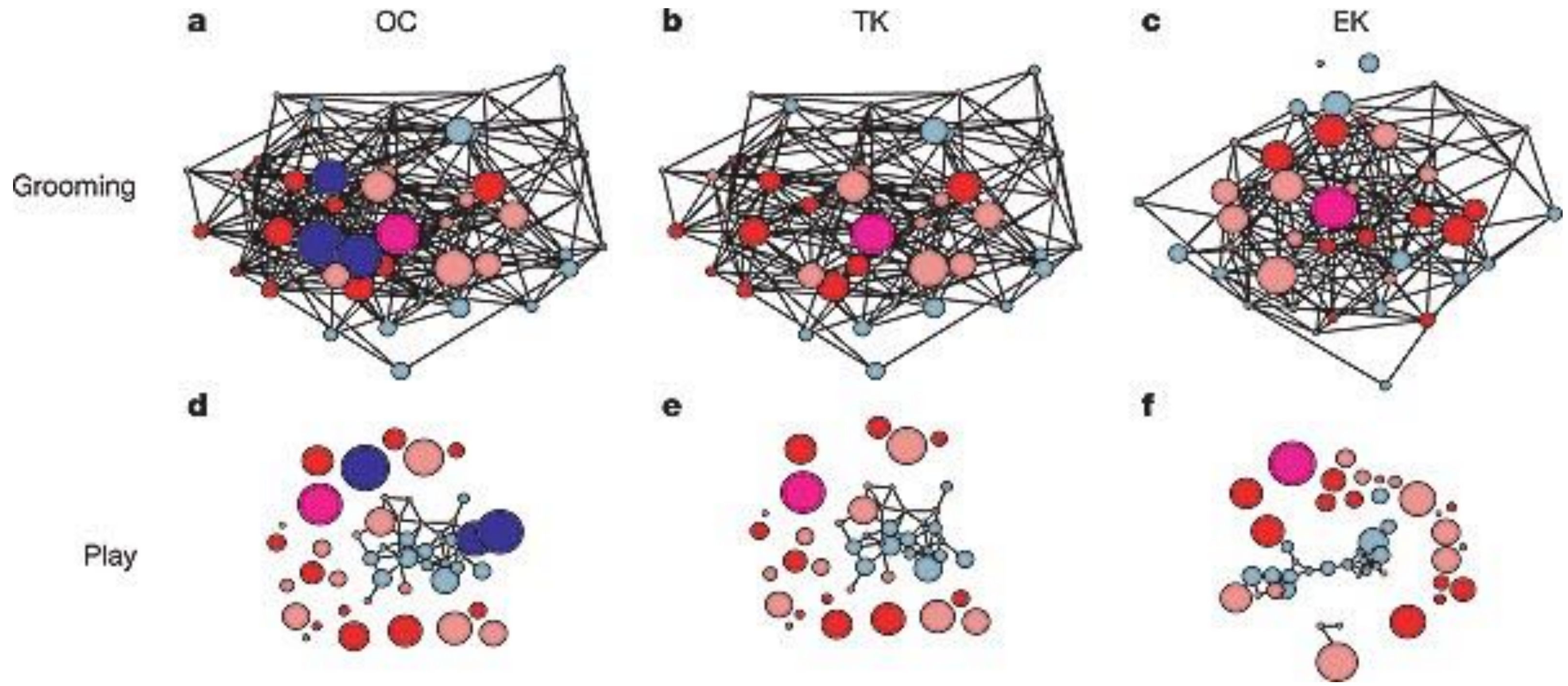
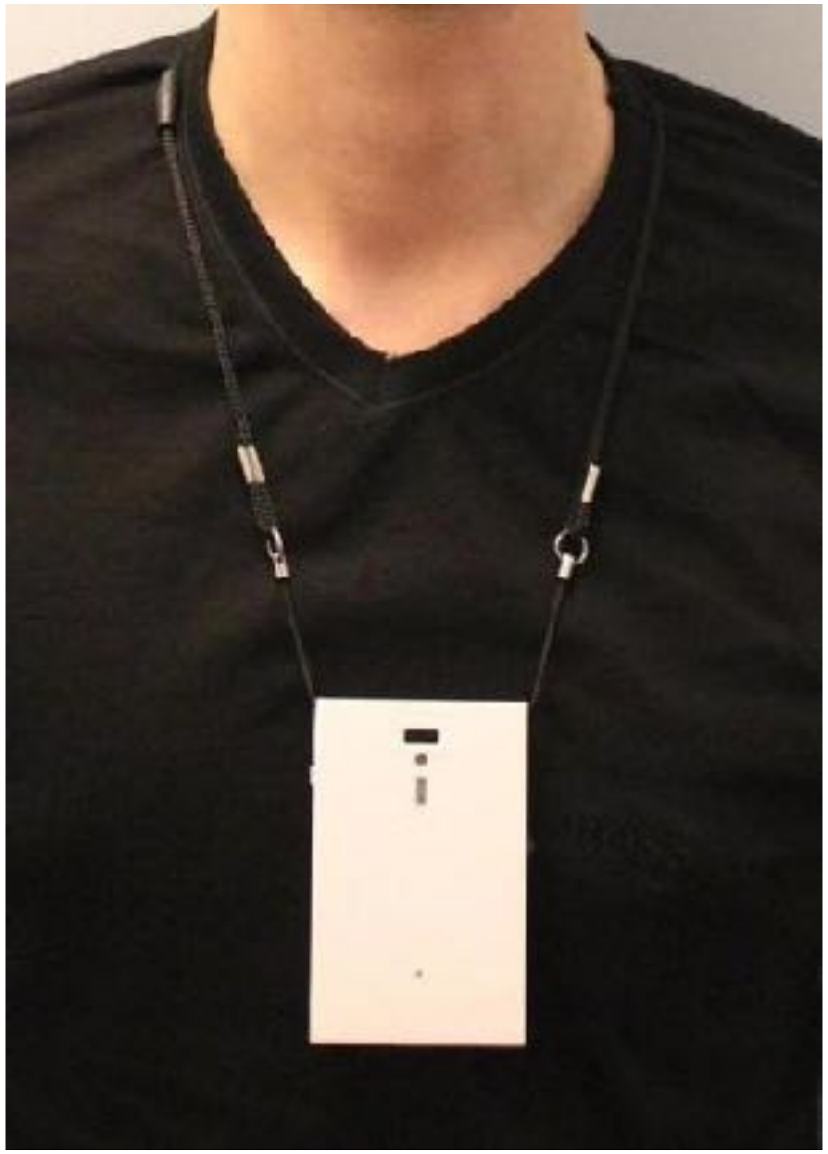


Figure: Flack et al (Nature, 2006)

Direct Observation

New Technology: Sociometric Badges



- Developed at MIT
- Record proximity to other people wearing badges
- Does not record conversations
- Does record data about time spent speaking, time spent listening, and turn-taking

Photo: Boston Globe

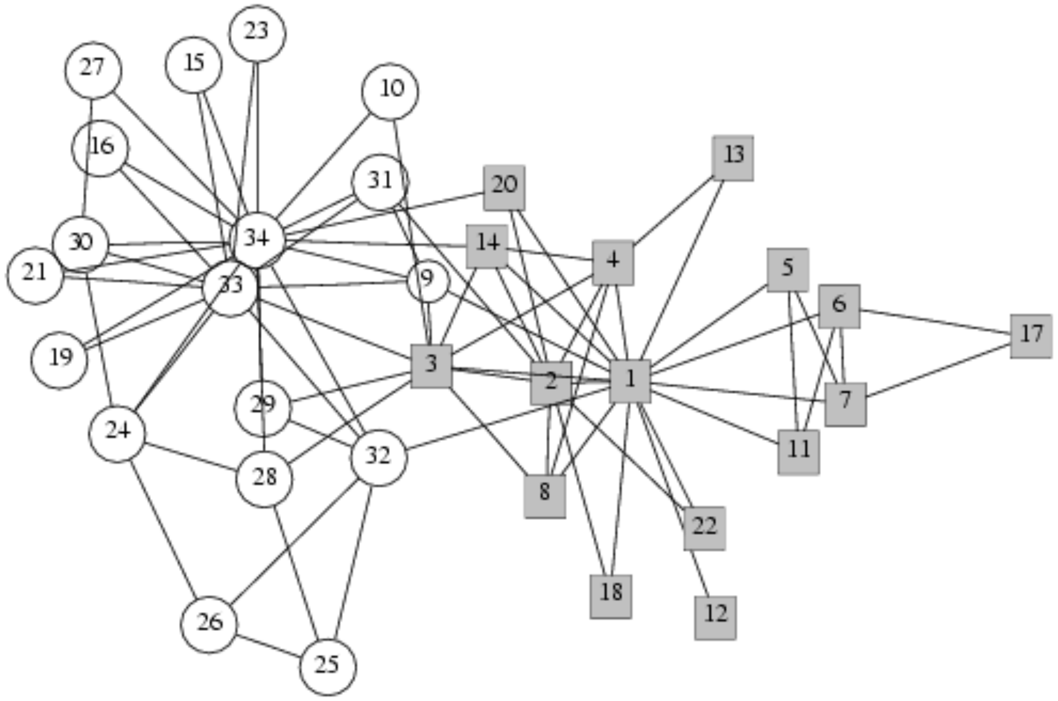
Direct Observation

Advantages:

- Link definitions are more objective than surveys
- Easier for subjects
- Can use animal data

Disadvantages:

- Very labor intensive!
- Limited to small groups
- Can be hard to interpret



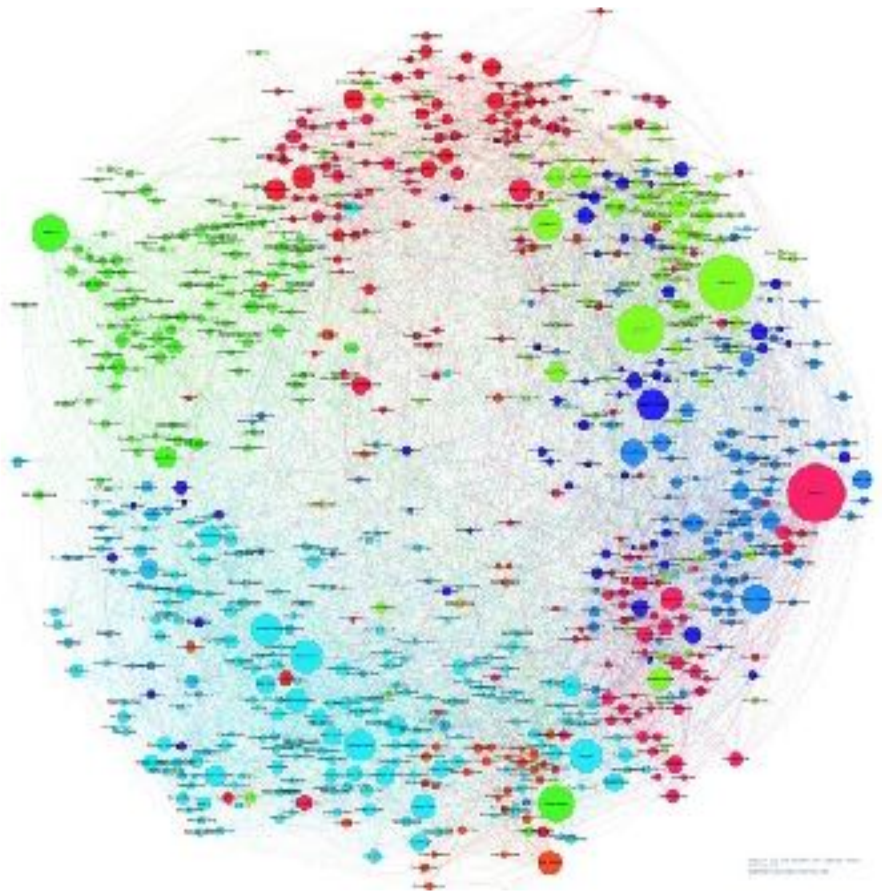
graphic: Zachary

Option 3: Passive Data Collection

Get information about interaction from a third party:

Non-electronic examples

- Florentine family network
- Southern women's study
- Six Degrees of Francis Bacon

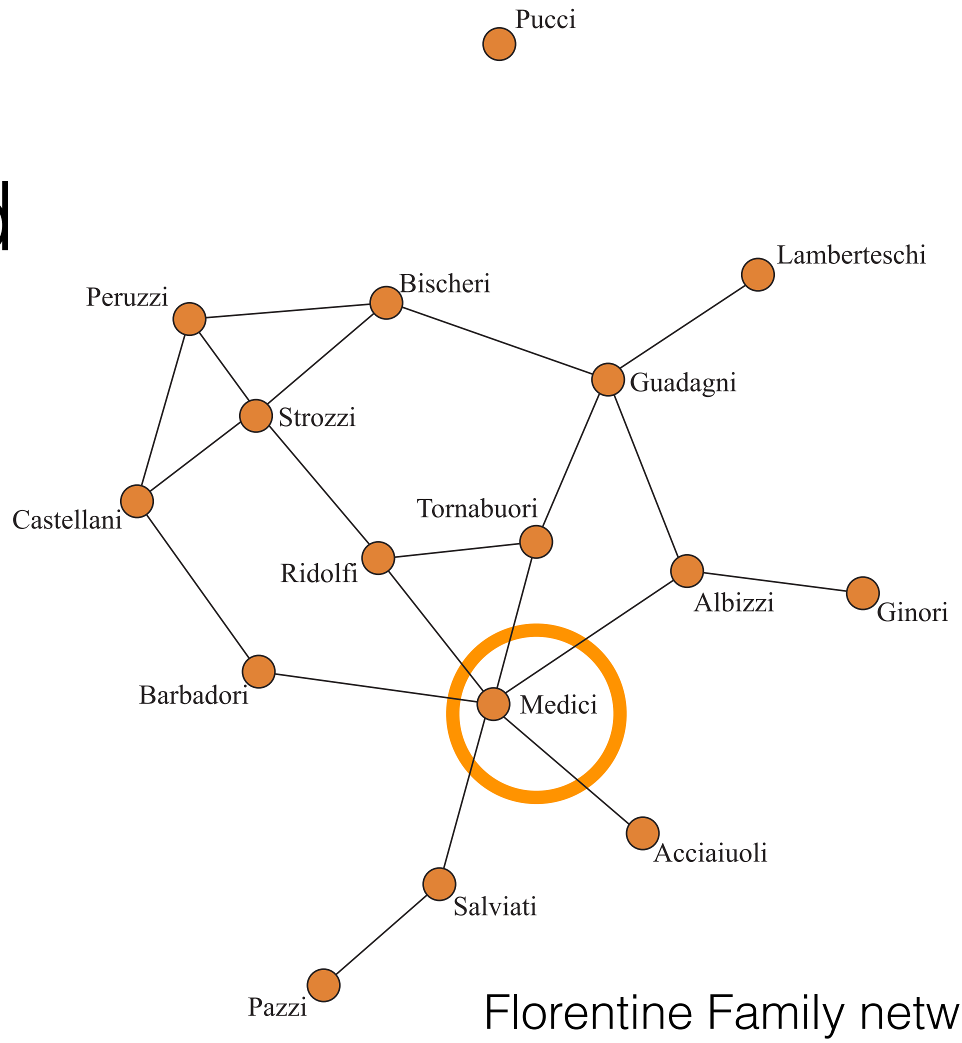


Passive Data Collection

An example: the Florentine Family network

Data from contemporaneous sources about marriages and trade ties between families

Suggested that the Medici may have married strategically to improve trade relationship



Florentine Family network

Passive Data Collection

Collection from records has become easier as more data is generated online

Examples:

- Source Forge
- Stack Overflow
- Wikipedia
- World of Warcraft
- Patents and Papers
- Newspapers
- Public Records

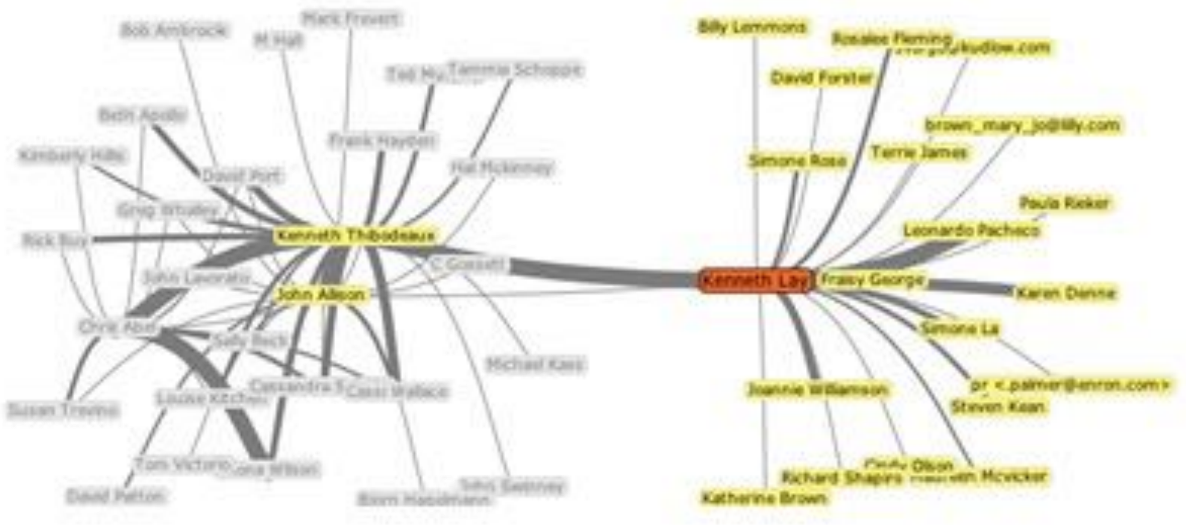
ref: Crandall et al 2009

Passive Data Collection

An example: The Enron email network

Public data release of all internal Enron emails:

- From that, you can generate a network
- Nodes = email addresses
- Links = emails

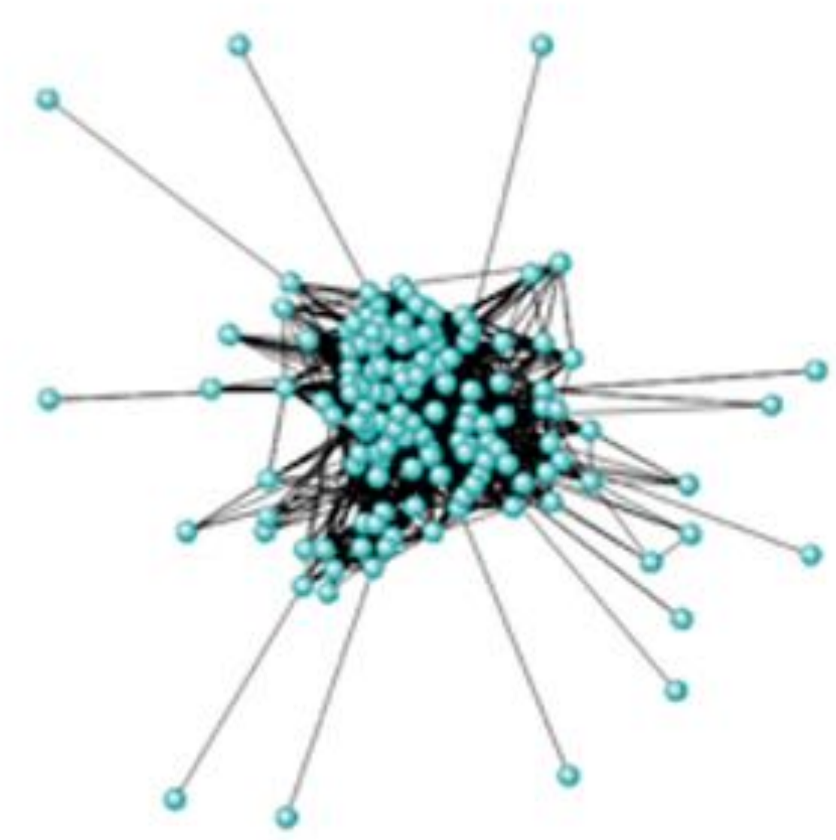


Enron email network

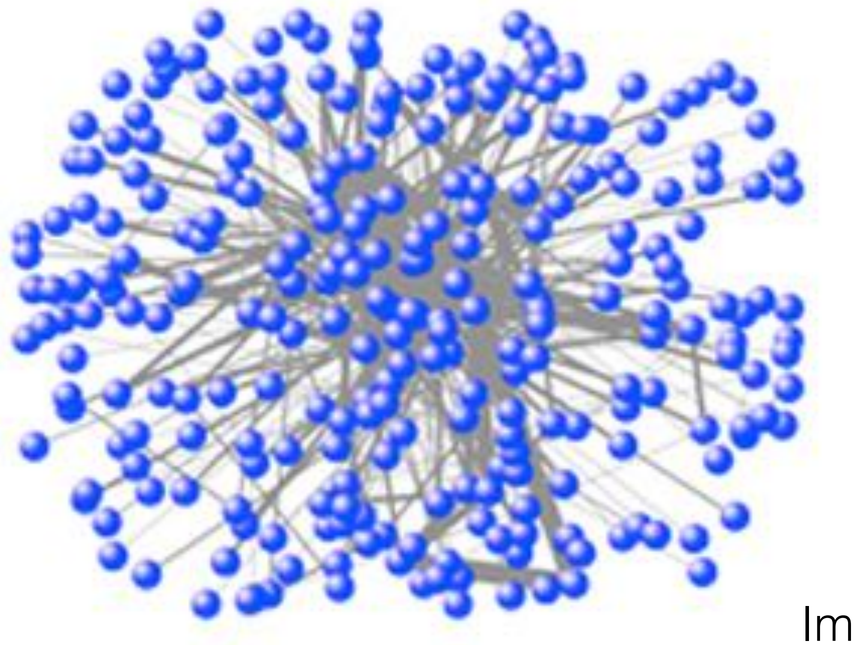
Passive Data Collection

An example: The Enron email network

- Can construct two different networks: one using illicit emails and another using ordinary emails



Illicit Emails



Ordinary Emails

Image ref: Brandy Aven

Passive Data Collection

Another example: The Twitter network

- @A is connected to @B if @A follows @B
- @A is connected to @B if @A retweets a tweet from @B

Open questions:

How does information spread?

Can we predict who will be most influential?

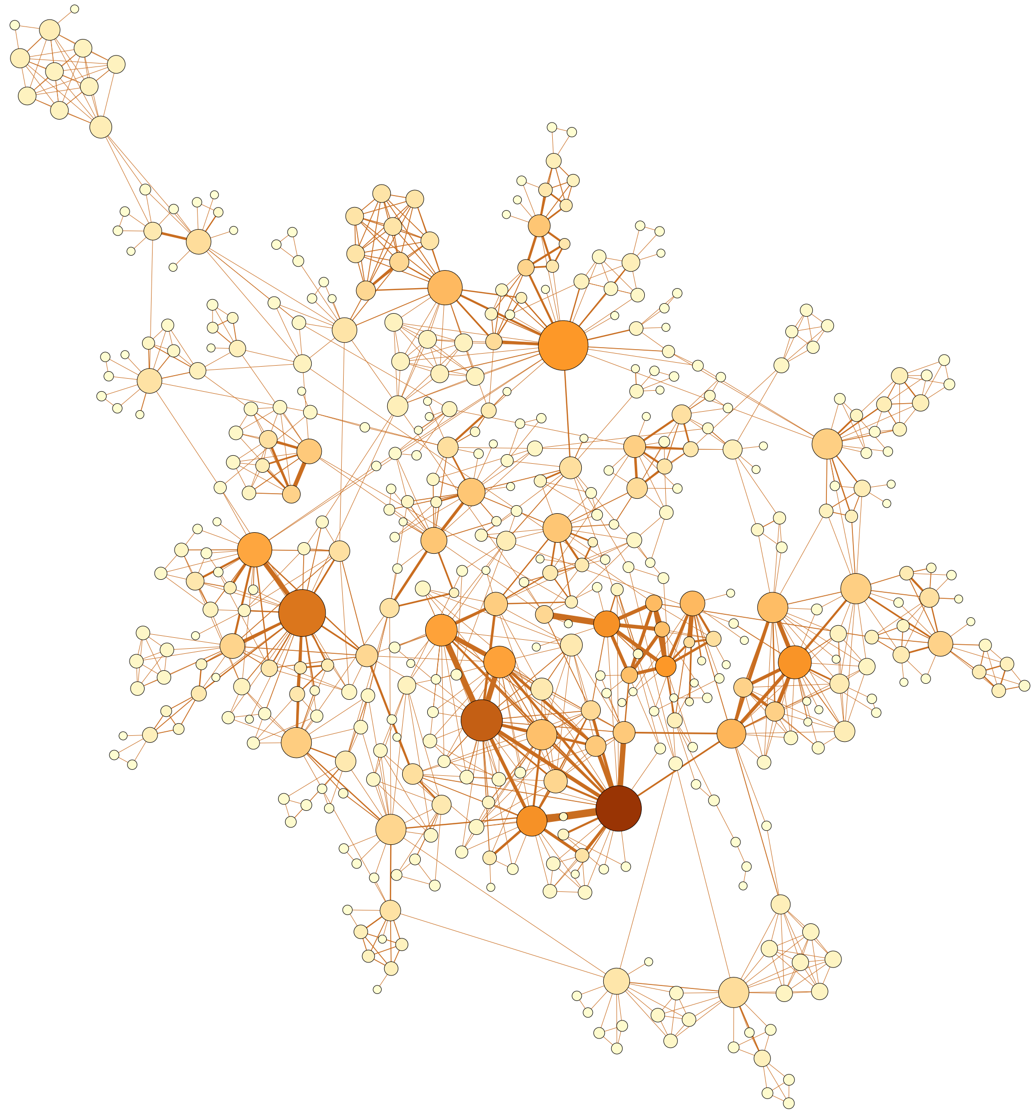
Passive Data Collection

Advantages:

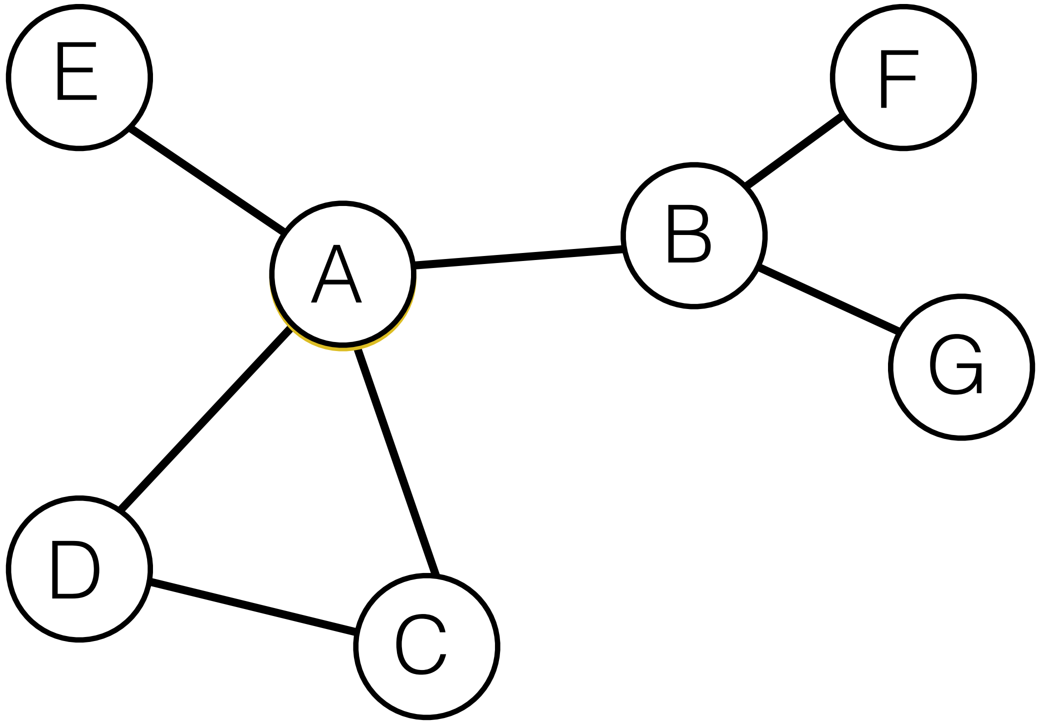
- Great diversity of data
- Digital data sources are easily collected
- Larger data sets possible

Disadvantages:

- Stuck with what you have
- Tempting to make inappropriate generalizations



source: PER Coauthorship 2000-2010



Data Representations

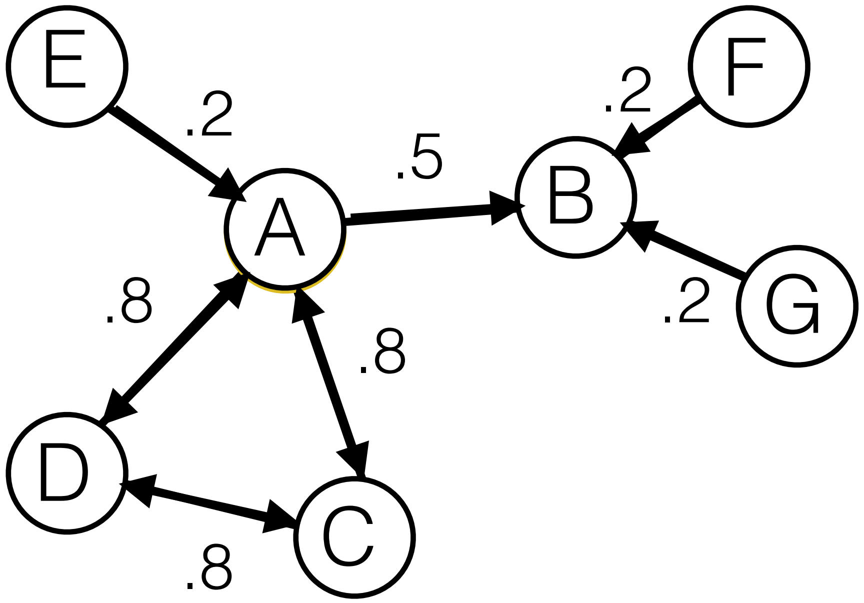
Network Data Formats

Many many formats exist…

- Adjacency matrix
- Edge list / Adjacency list
- Pajek (.net)
- GML (.gml)
- etc…

Adjacency Matrix

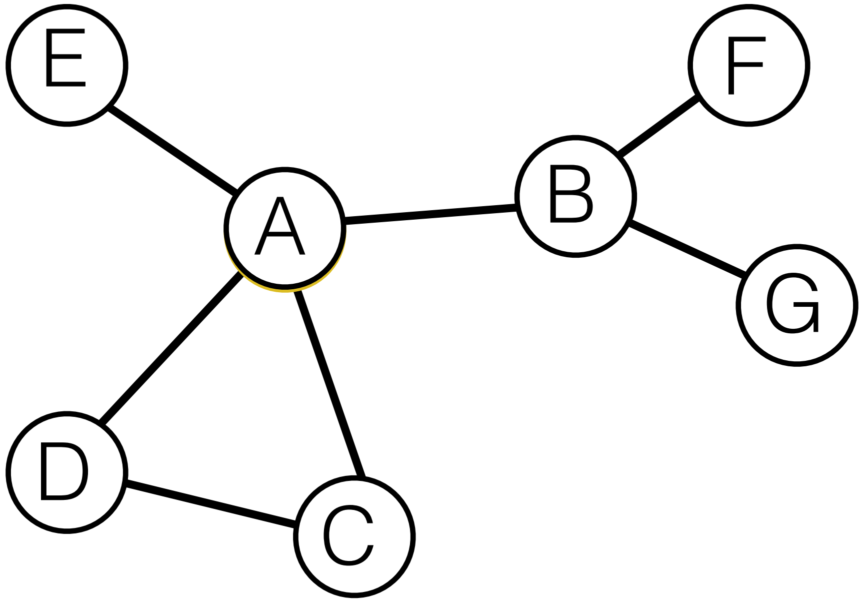
- Entry $a_{ij} = w_{ij}$ if there is a link from i to j with weight w_{ij}



$$
= \begin{array}{cccccccc}
 & A & B & C & D & E & F & G \\
A & 0 & 0 & 0.8 & 0.8 & 0.2 & 0 & 0 \\
B & 0.5 & 0 & 0 & 0 & 0 & 0.2 & 0.2 \\
C & 0.8 & 0 & 0 & 0.8 & 0 & 0 & 0 \\
D & 0.8 & 0 & 0.8 & 0 & 0 & 0 & 0 \\
E & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
F & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
G & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
$$

Edge list / Adjacency List

One row for each edge (edge list) or node (adjacency list)



=

A	B
A	C
A	D
A	E
B	A
B	F
B	G
C	A

…etc

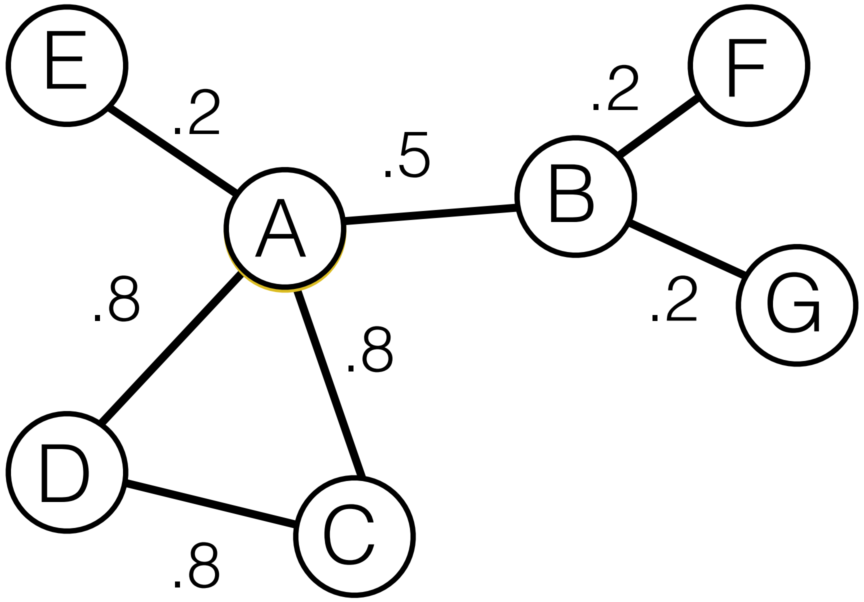
or

A	B C D E
B	A F G
C	A D
D	A C
E	A
F	B
G	B

Note that there is no way to represent an isolate (lone node) in edge list format!

Pajek (.net)

- Two sections: nodes and edges
- Both can have attributes



=

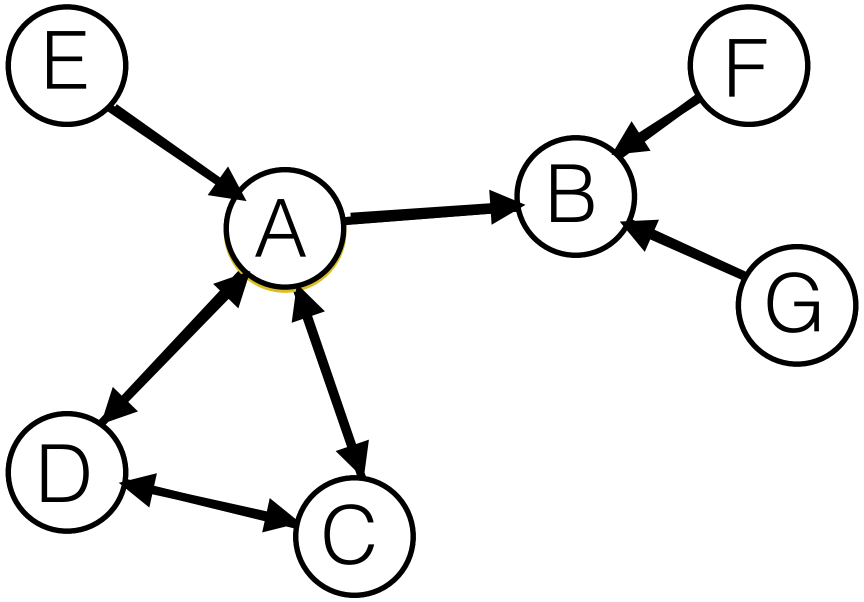
```
*Vertices 7
1 “A” “female”  0
2 “B” “female” .5
3 “C” “male”     1
4 “D” “female”  0
5 “E” “male”    .5
6 “F” “female”  0
7 “G” “male”    1

*Edges
A B .5 “blue”
A C .8 “green”
A D .8 “green”
…etc
```

GML (.gml)

A more complex format, allowing for more embedded information

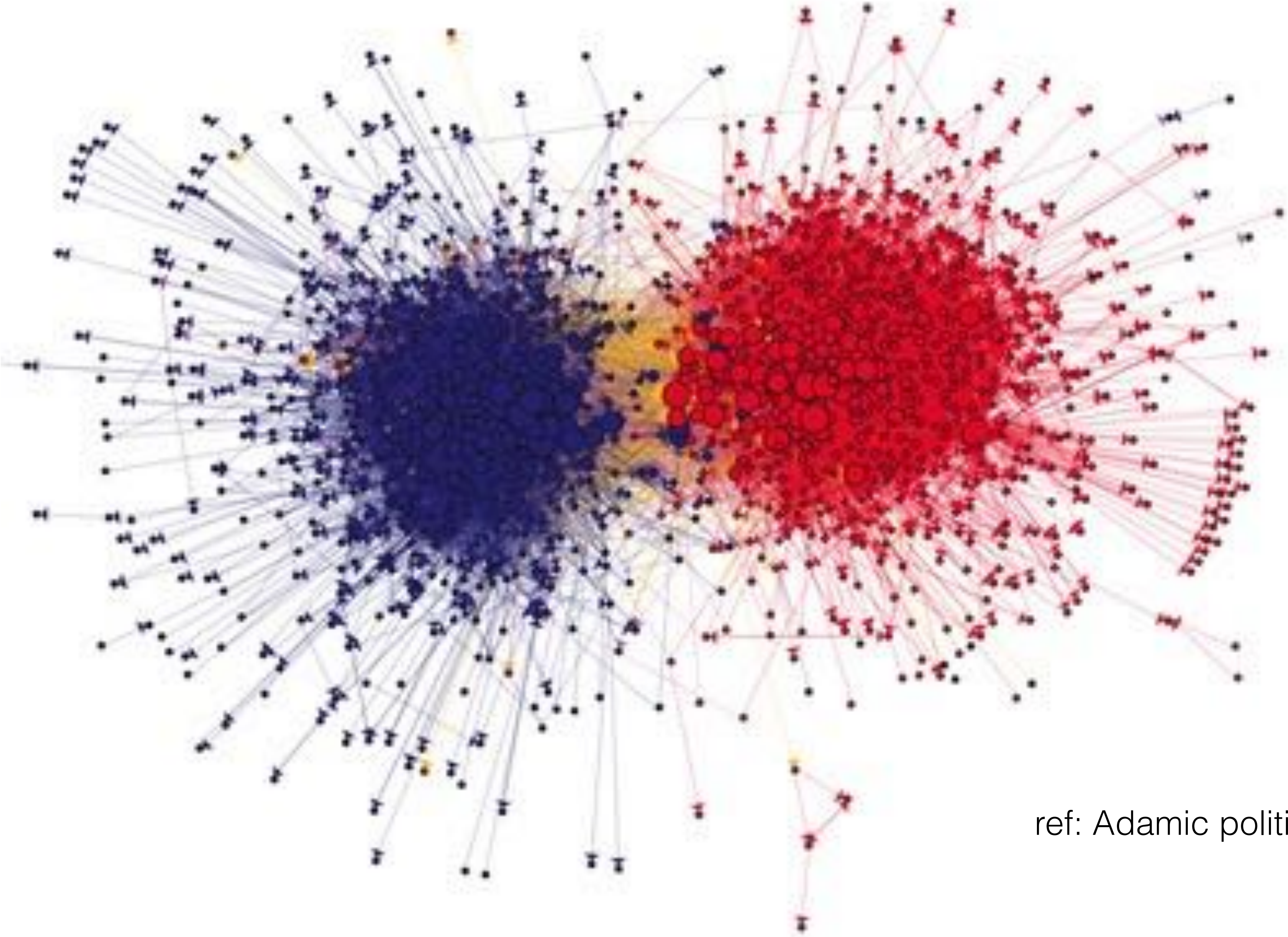
- Labels
- Colors
- Locations
- etc…



=

```
graph
[
  node
  [
   id A
   label "Node A”
    color “green”
    major “econ”
  ]
  node
  [
   id B
   label "Node B”
    color “blue”
    major “ece”
  ]
   edge
  [
   source A
   target B
   label "Edge A to B"
  ]
```

etc…

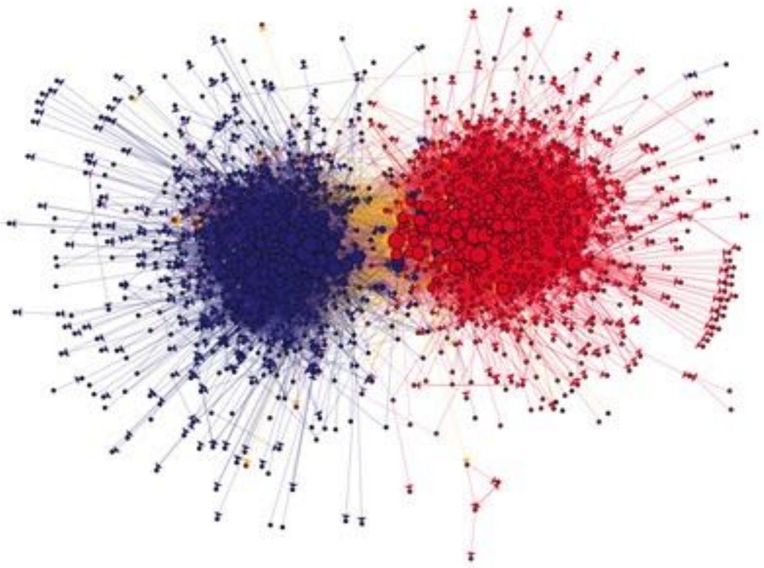


ref: Adamic political blogs

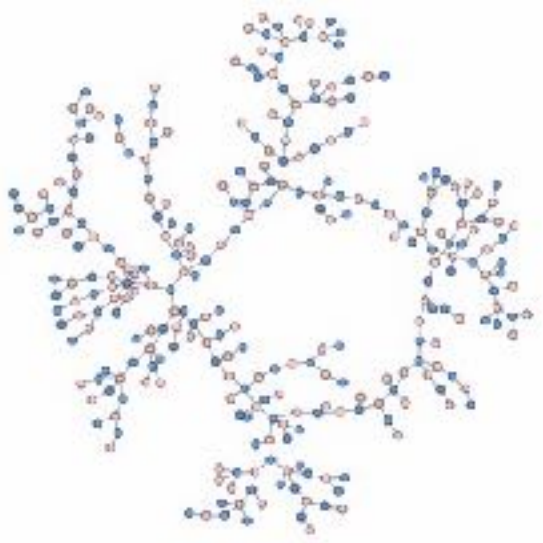
Data Visualization

Network Visualization

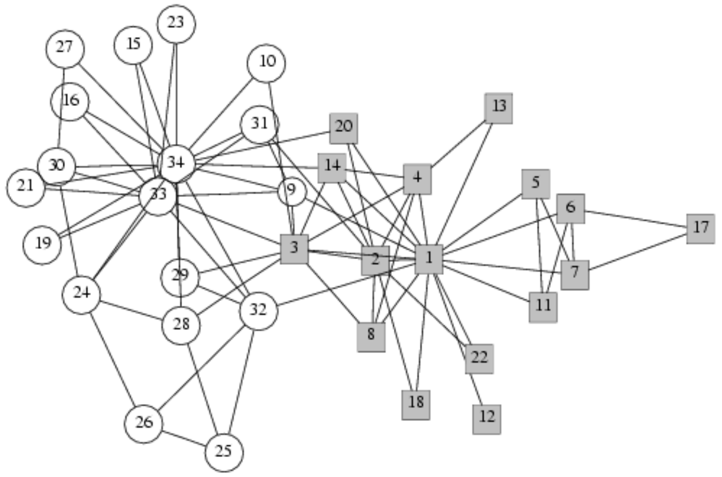
Good network visualization reveals patterns in the data



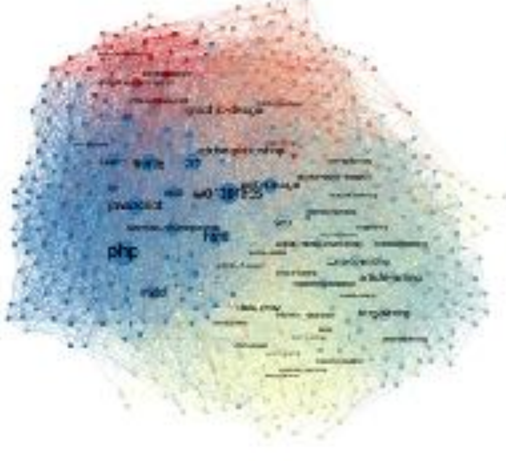
ref: Adamic political blogs



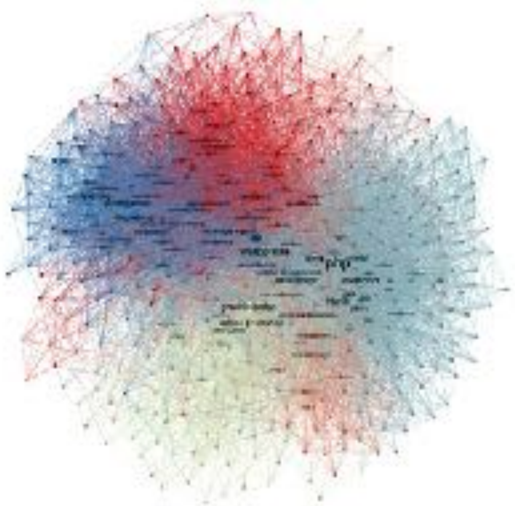
ref: Bearman et al
teen romance network



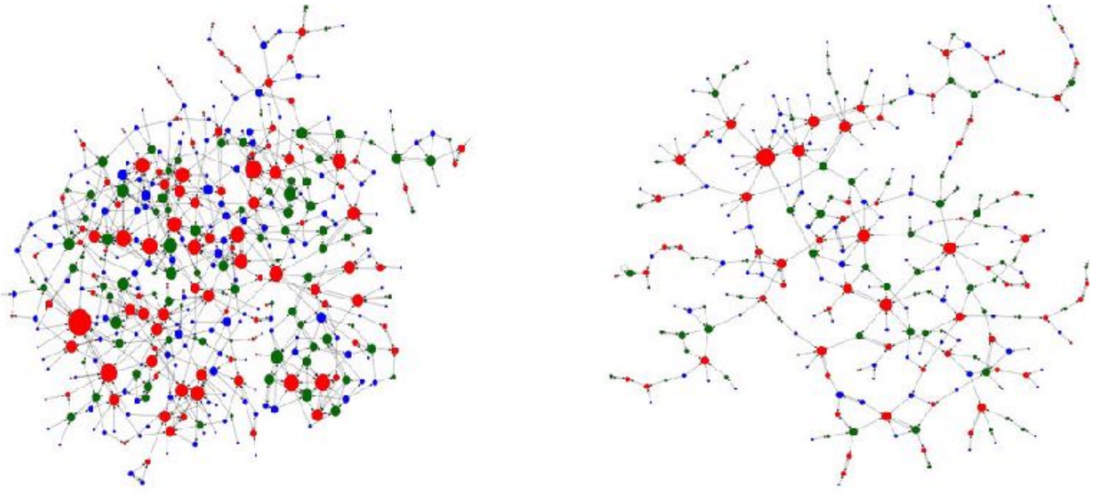
graphic: Zachary



Worker Skill Network



Job Skill Network



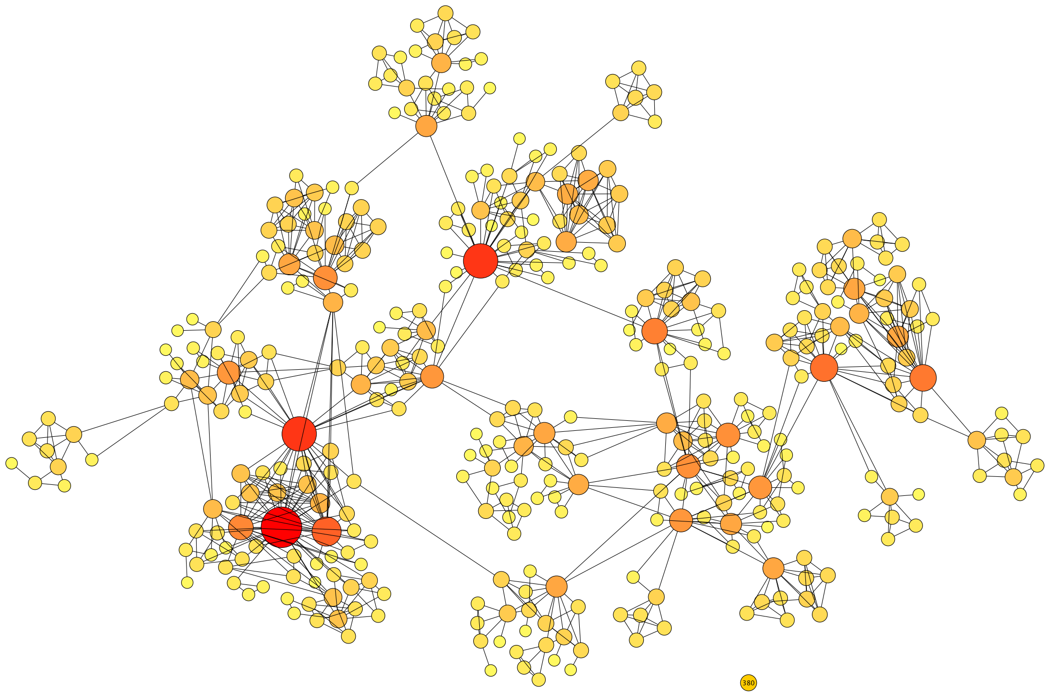
rice/kerosene network

advice network

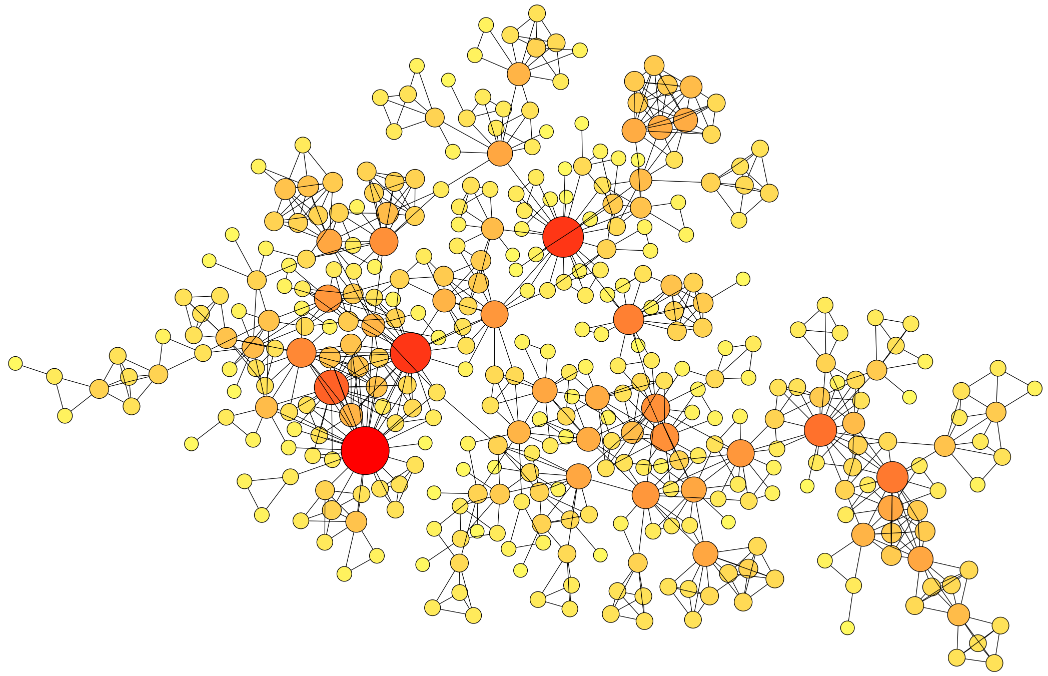
graphic: Wardil and Hauert

Network Visualization

Bad visualizations reveal patterns that just aren’t there!



=



If possible, you should back up your eye’s observations with data from the network

Gephi

Network Analysis Tools

Network Analysis Tools

Different tools for different purposes

	Ease	Viz	Metrics	max N	Biggest Strength
NodeXL	+	~	limited	~10,000	Simple Metrics and Visualization
YEd	+	+	none	~50,000	Visualization
Gephi	+	++	limited	~100,000	Simple Metrics and Visualization
Pajek	-	-	many	~5,000,000	Complex Metrics
Python and R*	-	-	many	~5,000,000	Large Scale Metrics and Statistics

*** packages: networkx and igraph**

The Statistics Panel

Way to calculate some measures.

Choose “Degree”

- press run
- we get a panel with a summary: degree, distribution of degree (number of connections vs number of people)

Choose “Graph Density”

- press run
- we get a panel with a summary

These measures are now recorded for individuals as a column in the data laboratory

You can use these measures to color/size nodes using the “Ranking” panel

The Ranking Panel

Additional control over the visualization

Can select node color, node size (can also do the same with edges)

- size = circles
  - choose “attribute”
  - from drop-down, choose attribute (select degree)
- color
  - choose “attribute”
  - choose an attribute (select degree)
  - choose a color way (small box next to spectrum)

Pro-tip: “Spline” allows you to adjust how the color/size is scaled across different values of the ranking parameter

Best way to get a feeling for this is to play around with it: try coloring by “country” and “sex”—what do you observe?

Pretty Pictures

The “Preview” window gives you a chance to output a nice-looking network picture

Note: You have to click the “preview” button after every change you make to the options

There are MANY options here.

- Things to try:
    - click the edges>rescale weight box (this keeps the width of the links from being determined by the weight on the link
    - click/unclick the edges>curved box
    - click labels on/off
        - with labels on, click/unclick the proportional size (this sizes the labels according to the size of the node)
    - adjust the width of the lines

Finishing up (for now)

You can export the picture by pressing the SVG/PDF/PNG button. That will let you export in a range of different image formats.

If you want to save the work you've done, you can save the whole workspace as what is called a .gephi file. This preserves your visualization choices, data manipulations (e.g. degree calculation, recasting the integer columns)