# Three Lectures on Networks

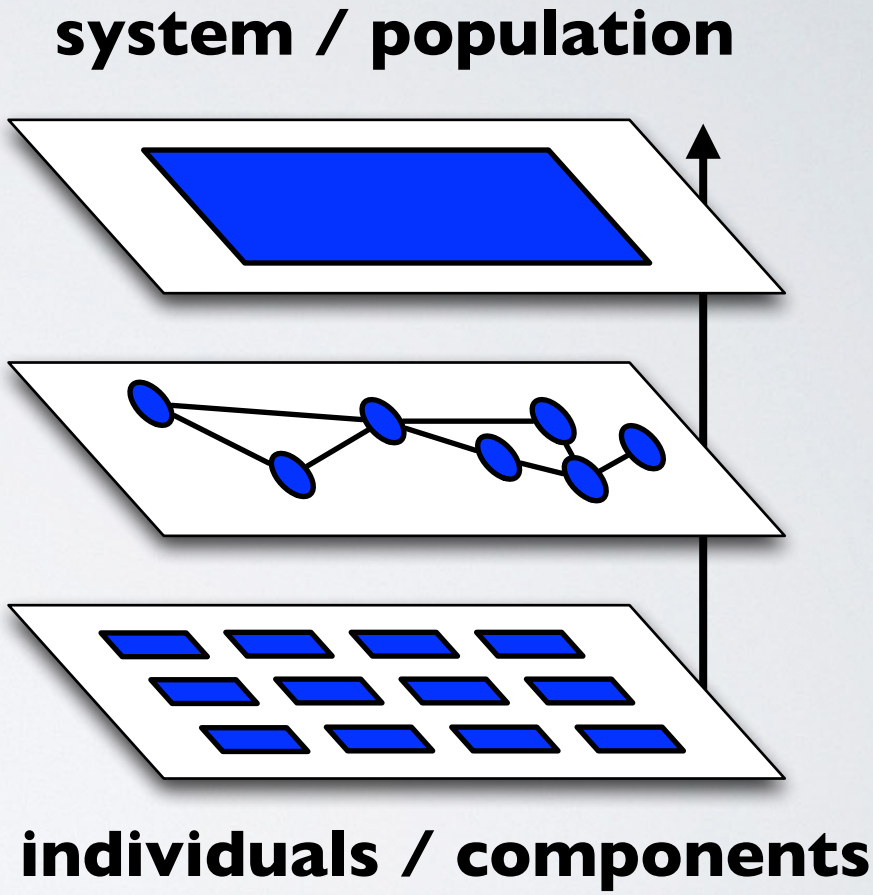
Aaron Clauset
@aaronclauset
Associate Professor of Computer Science
University of Colorado Boulder
External Faculty, Santa Fe Institute

**lecture 1: what are networks and how do we talk about them?**

# what are networks?

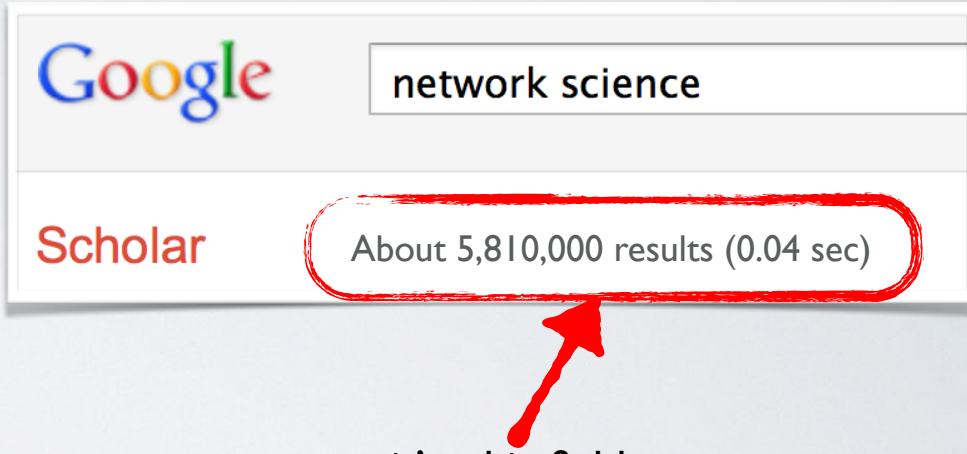
## what are networks?

- an approach
- a representation of complexity
- connect "micro" to "macro"
- *structure above*
  individuals / components
- *structure below*
  system / population



## these lectures

- build intuition
- expose key concepts
- highlight some big questions
- teach a little math
- provide some examples
- give pointers to further study
- prep for other CSSS lectures
- not a substitute for technical coursework

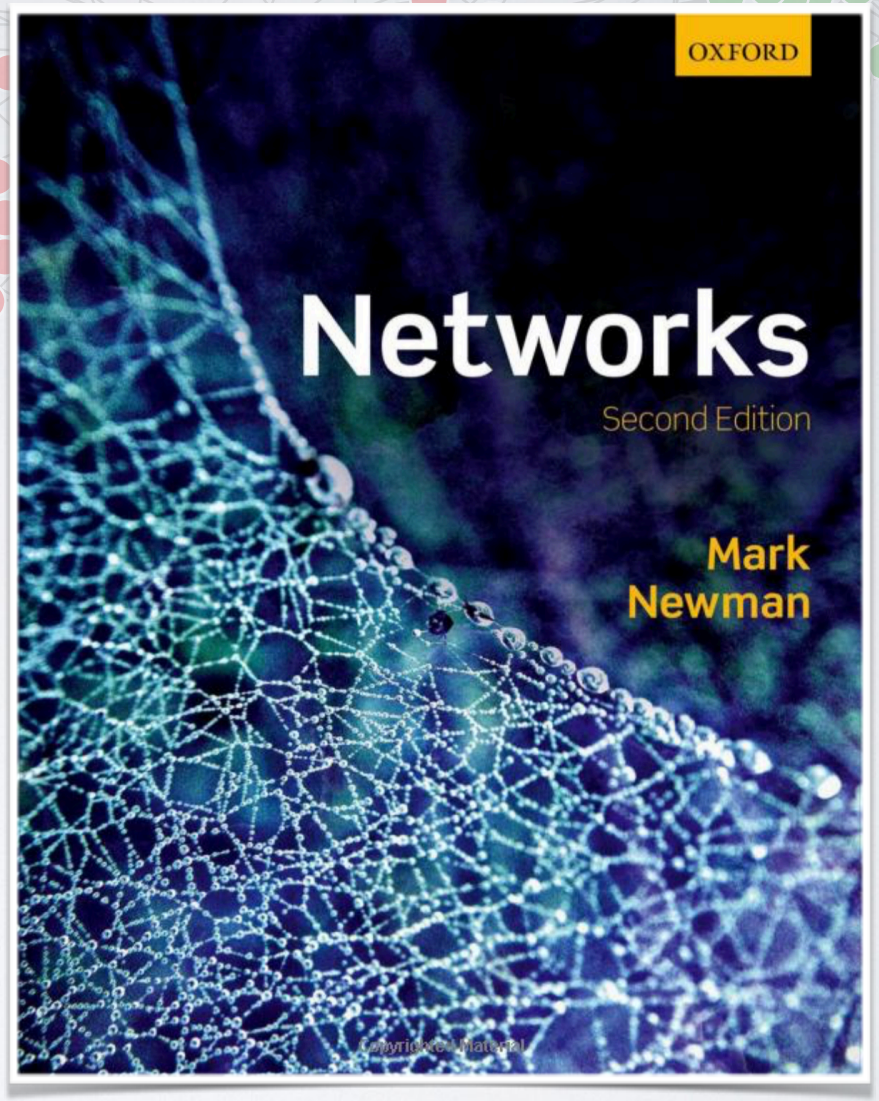


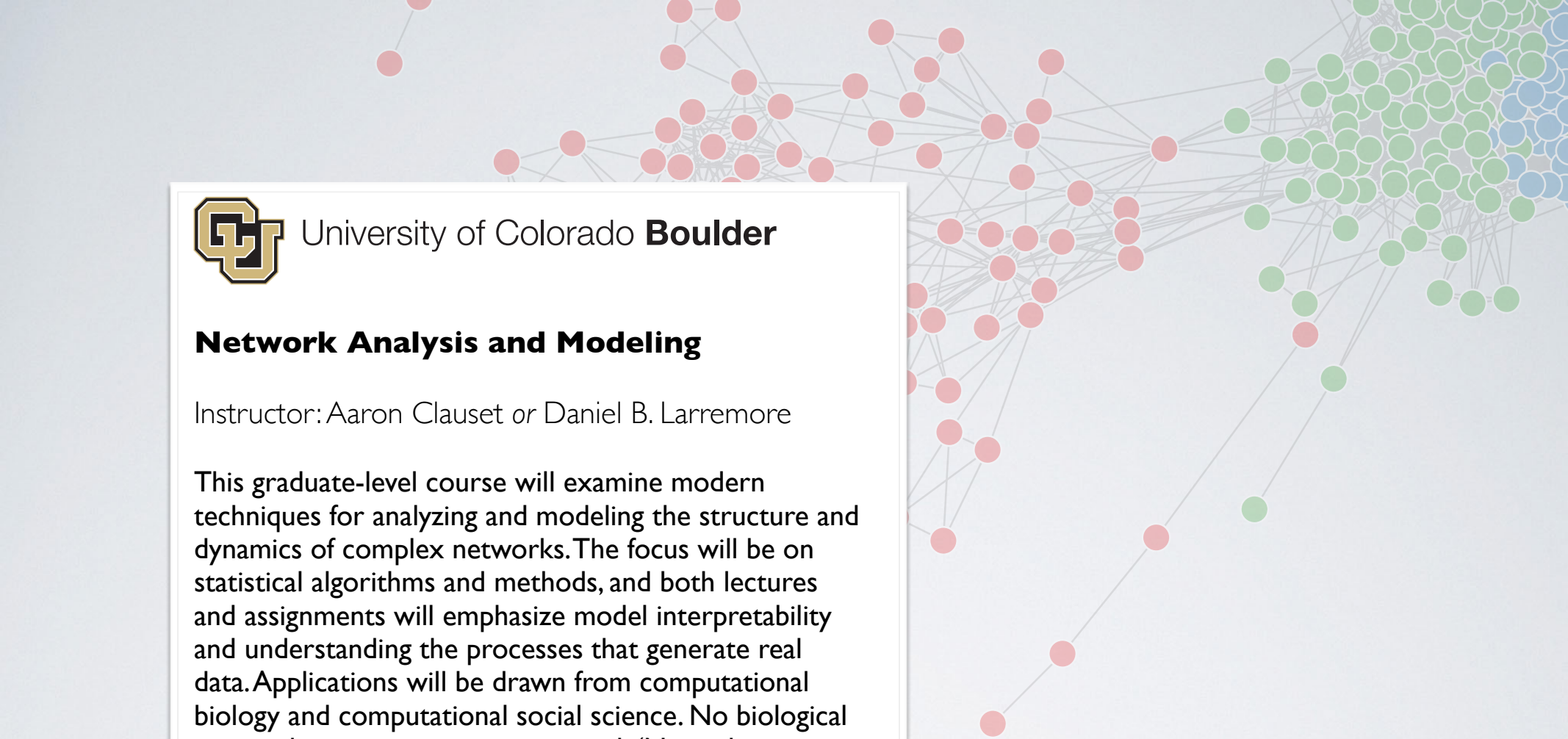
Mark Newman

Professor of Physics
University of Michigan

External Faculty
Santa Fe Institute

http://www-personal.umich.edu/~mejn/





University of Colorado **Boulder**

## Network Analysis and Modeling

Instructor: Aaron Clauset *or* Daniel B. Larremore

This graduate-level course will examine modern techniques for analyzing and modeling the structure and dynamics of complex networks. The focus will be on statistical algorithms and methods, and both lectures and assignments will emphasize model interpretability and understanding the processes that generate real data. Applications will be drawn from computational biology and computational social science. No biological or social science training is required. (Note: this is not a scientific computing course, but there will be plenty of computing for science.)

*Full lectures notes online (~150 pages in PDF)*

http://santafe.edu/~aaronc/courses/5352/

## Software

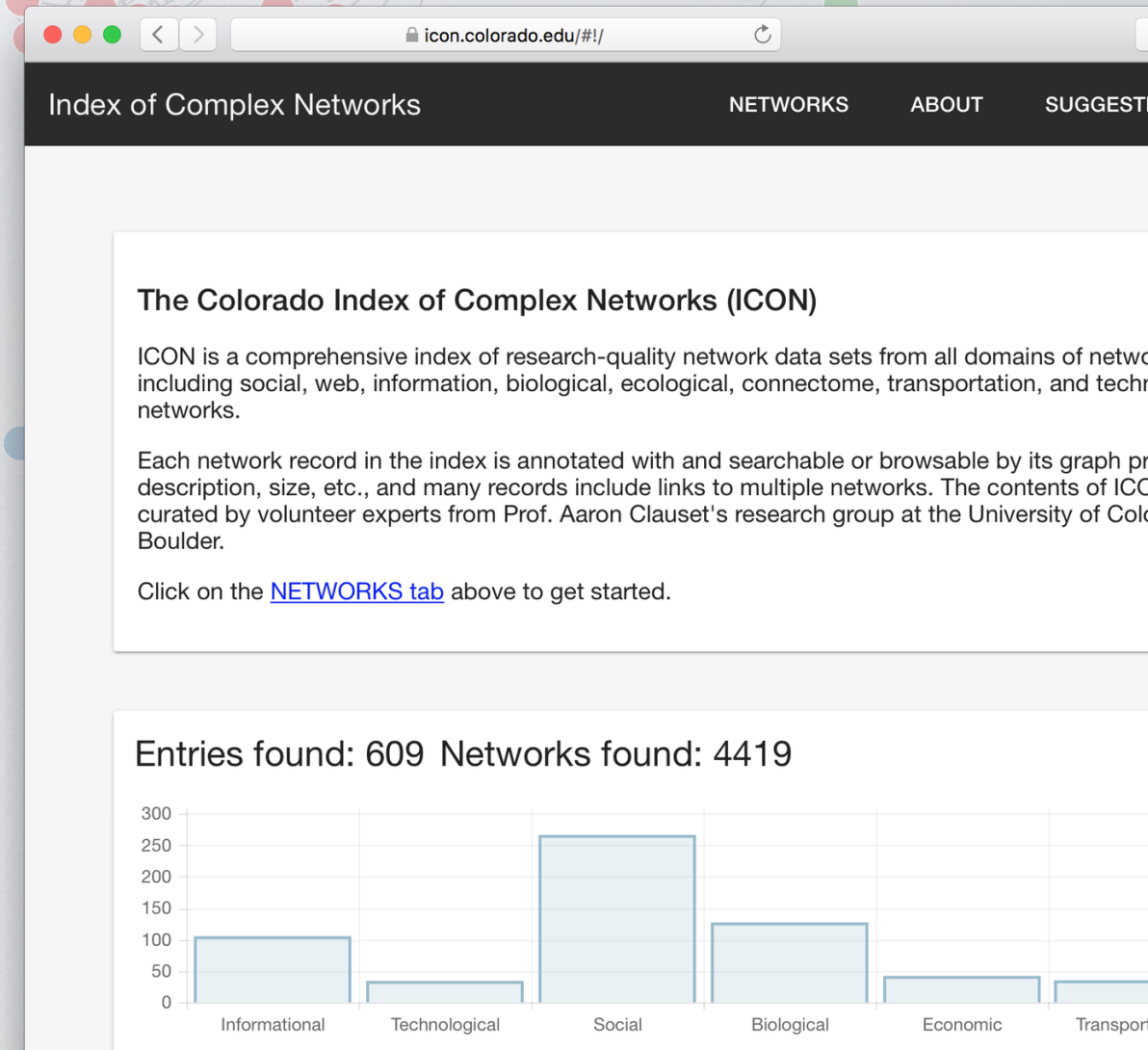
R
Python
Matlab
NetworkX [python]
graph-tool [python, c++]
GraphLab [python, c++]

## Standalone editors

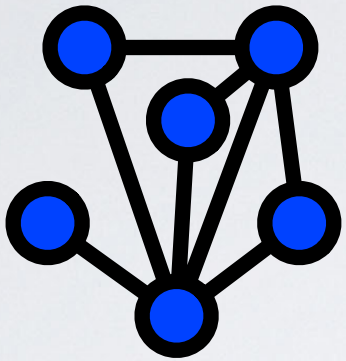
UCI-Net
NodeXL
Gephi
Pajek
Network Workbench
Cytoscape
yEd graph editor
Graphviz

## Network data sets

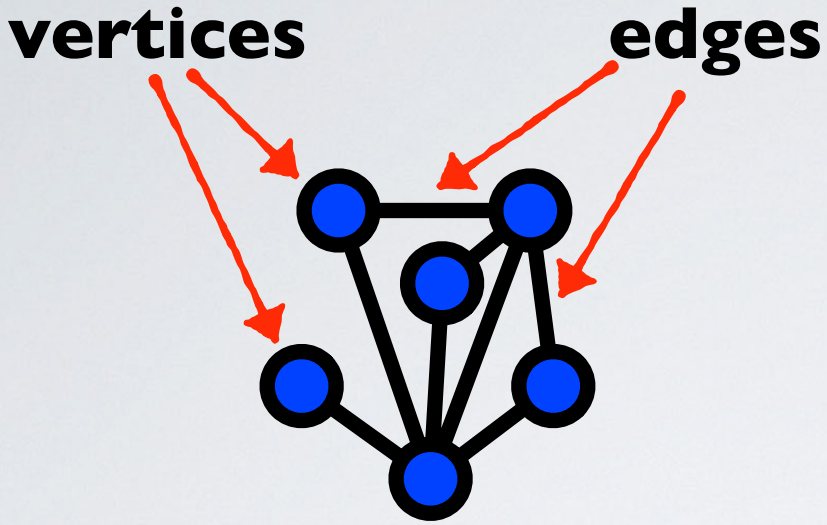
Colorado Index of Complex Networks



1. **defining a network**
2. **describing a network**
3. null models and statistical inference for networks



🤔 **the two most fundamental questions in network science**




**what is a vertex?**

$V$ distinct objects (vertices / nodes / actors)

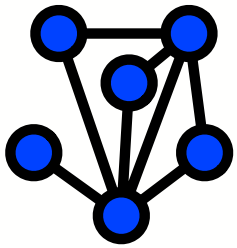
**when are two vertices connected?**

$E \subseteq V \times V$

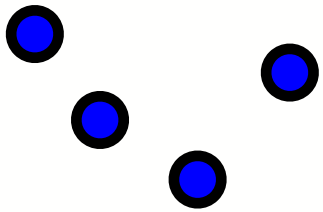
pairwise relations (edges / links / ties)

**6 major classes of networks**

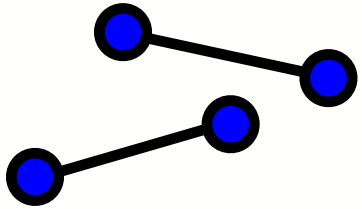
- technological
- information
- transportation
- social
- biological
- economic



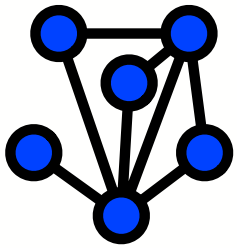
network

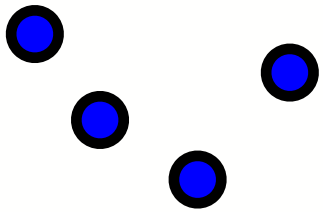


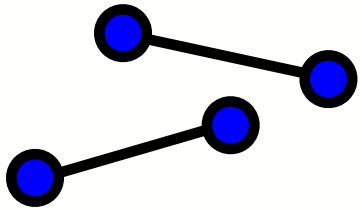
vertex



edge

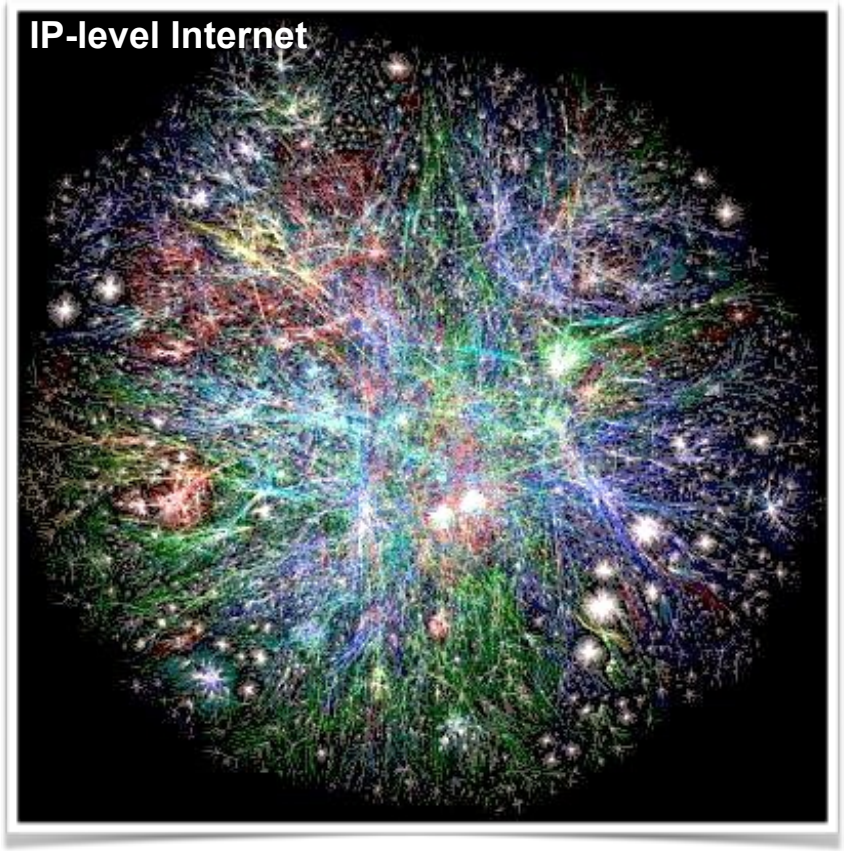


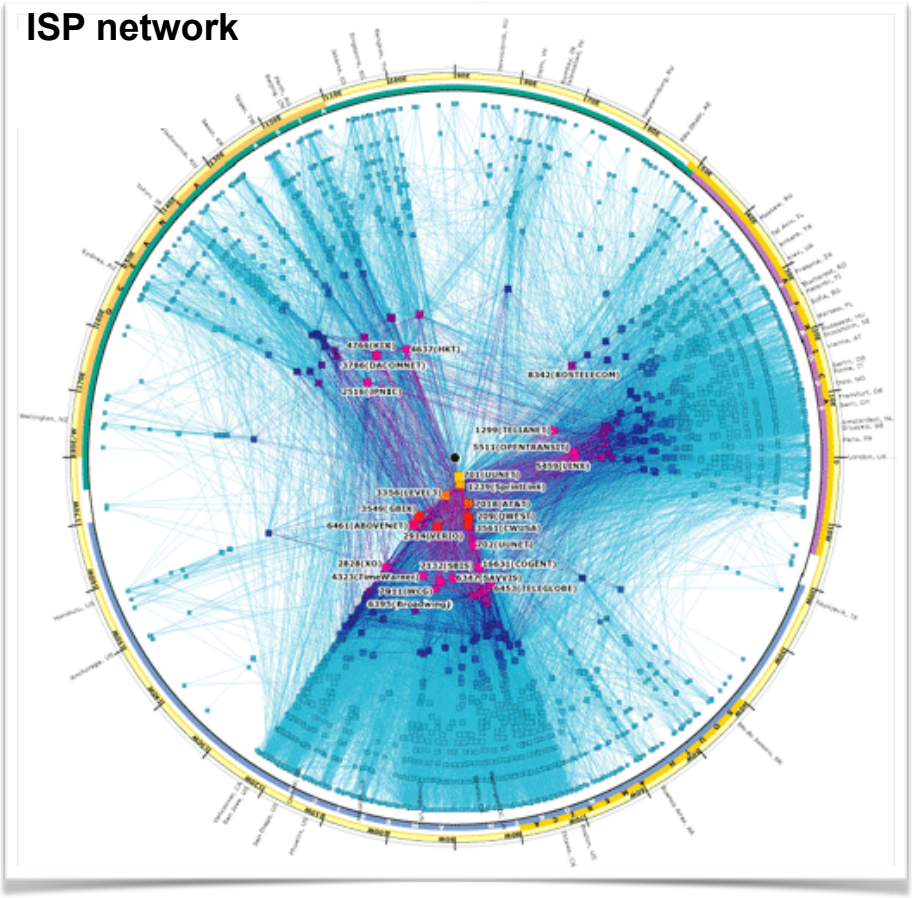


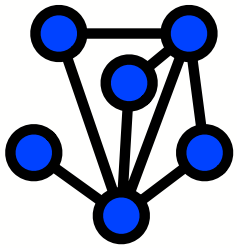


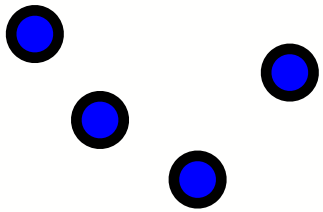
technological

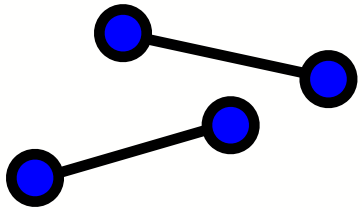
| network | vertex | edge |
|---|---|---|
| Internet(1) | computer | IP network adjacency |
| Internet(2) | autonomous system (ISP) | BGP connection |



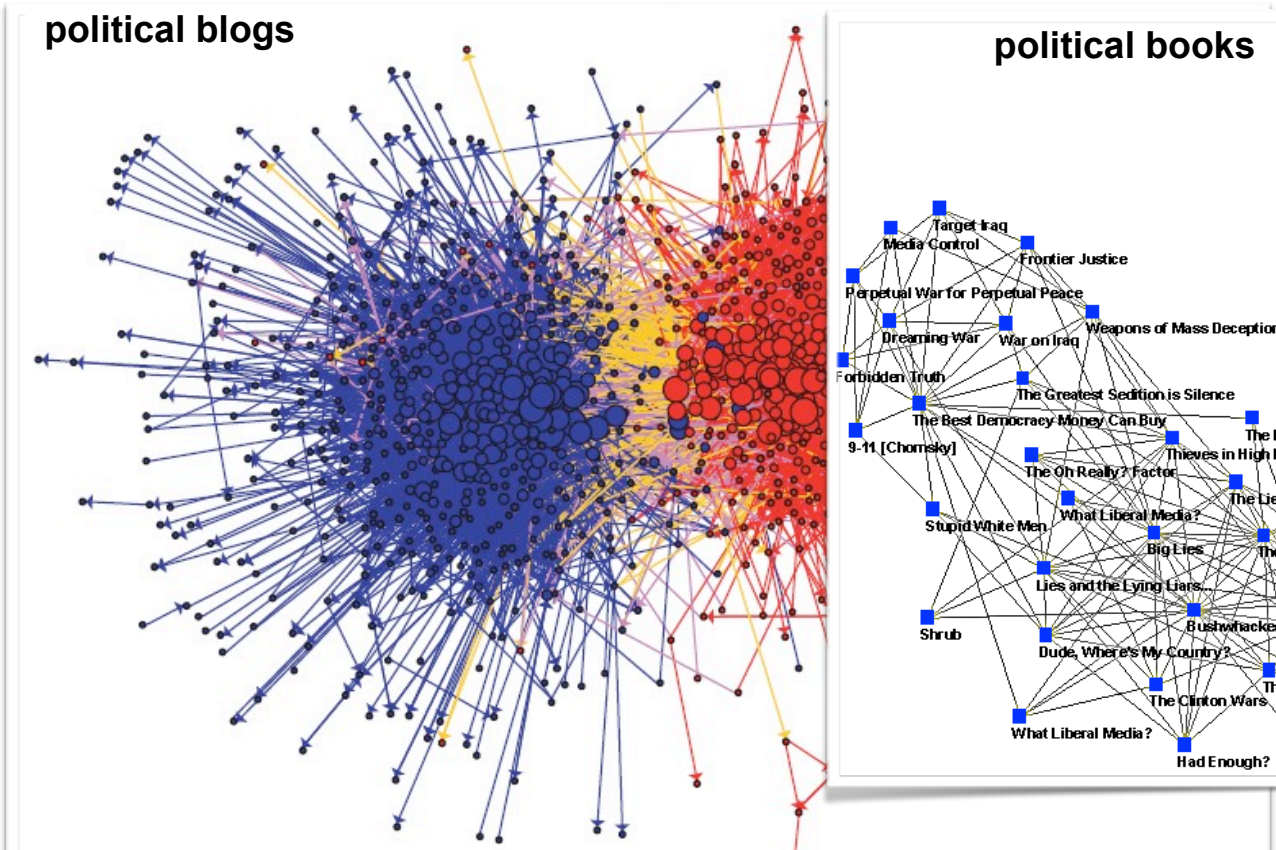



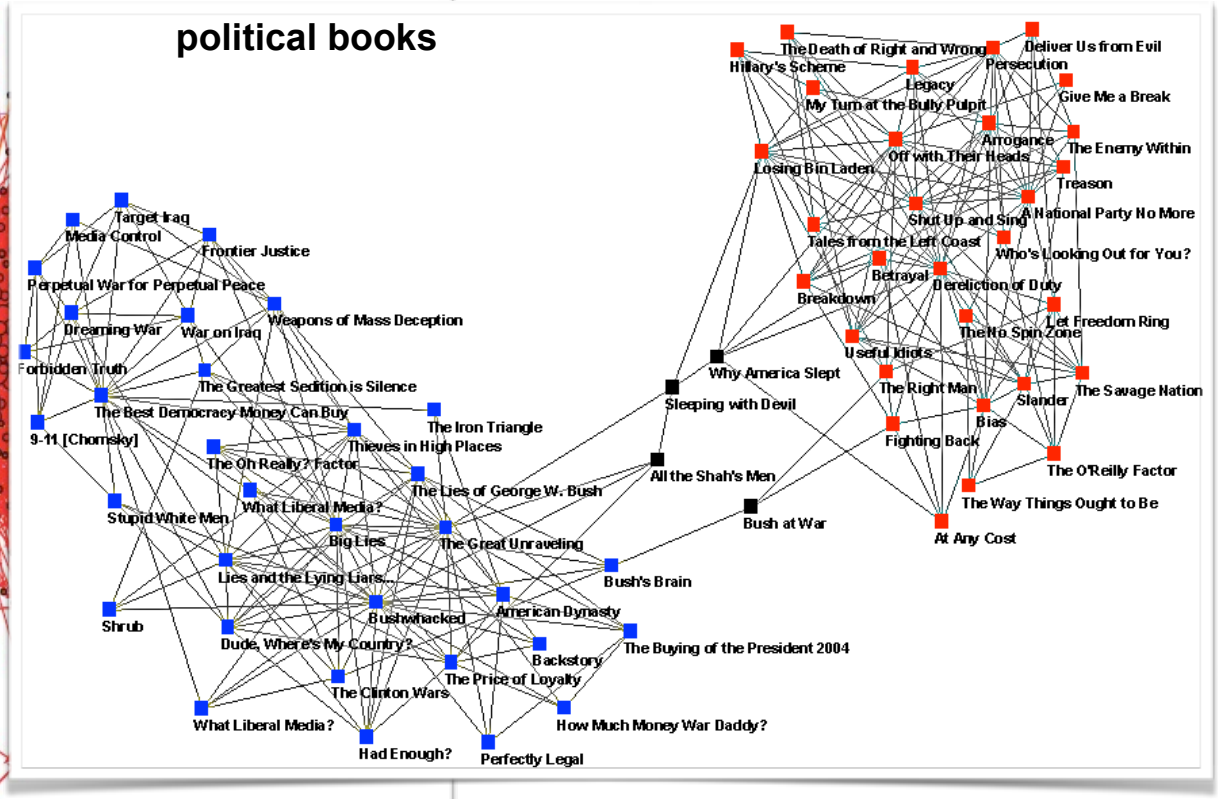


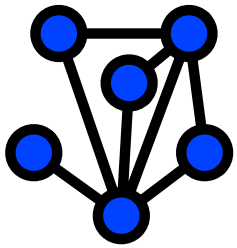


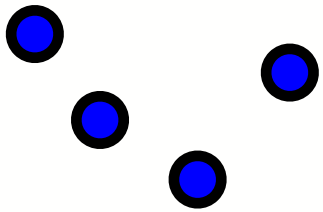


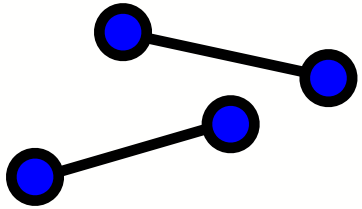
| | network | vertex | edge |
|---|---|---|---|
| technological | Internet(1) | computer | IP network adjacency |
| technological | Internet(2) | autonomous system (ISP) | BGP connection |
| information | software | function | function call |
| information | World Wide Web | web page | hyperlink |
| information | documents | article, patent, or legal case | citation |



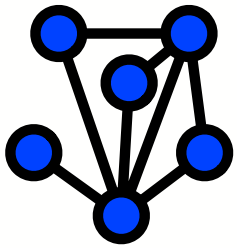



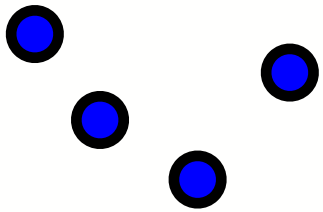


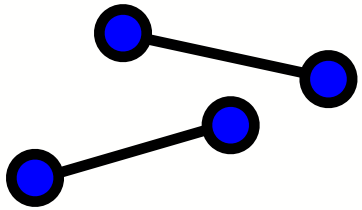




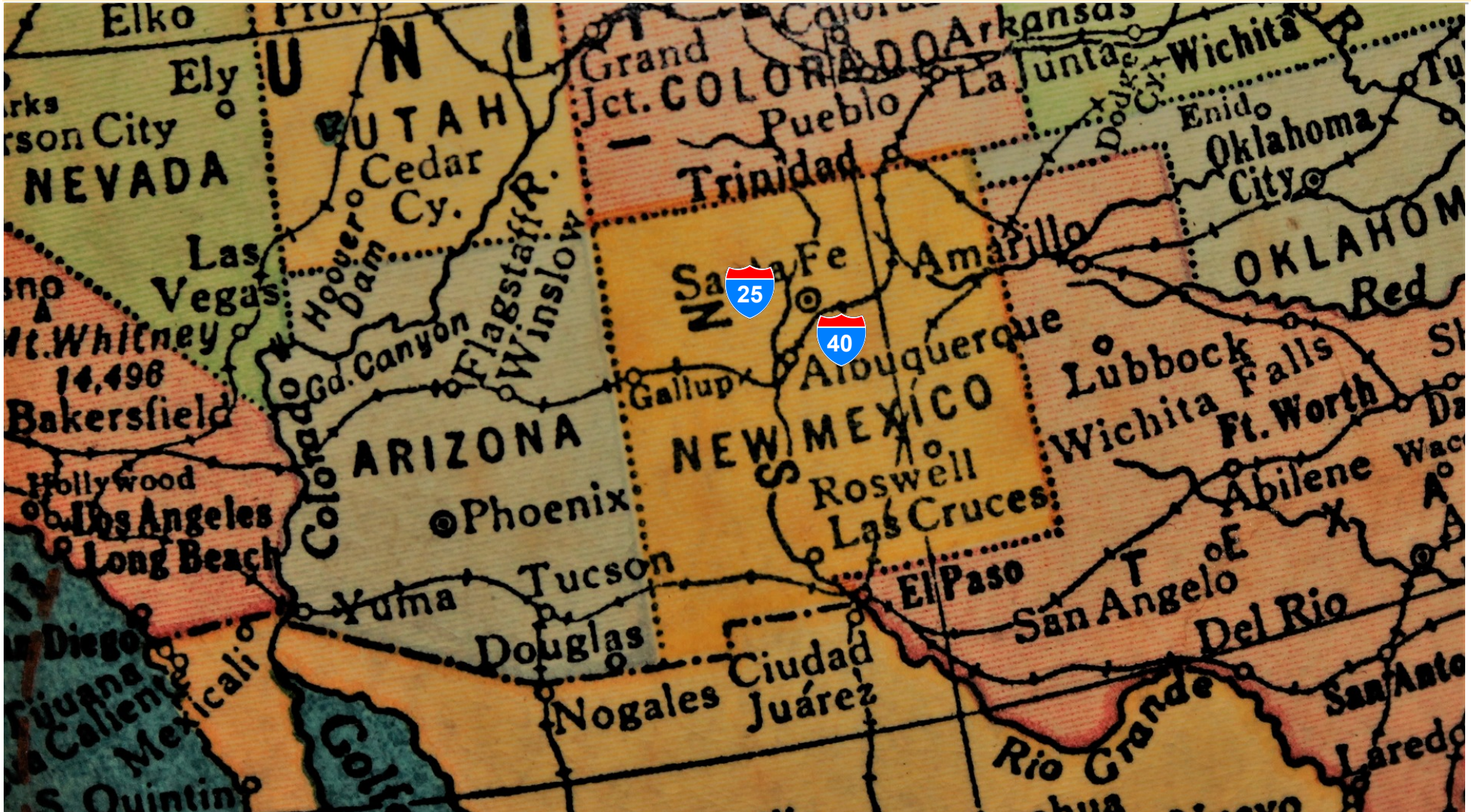
| | network | vertex | edge |
|---|---|---|---|
| technological | Internet(1) | computer | IP network adjacency |
| | Internet(2) | autonomous system (ISP) | BGP connection |
| information | software | function | function call |
| | World Wide Web | web page | hyperlink |
| | documents | article, patent, or legal case | citation |
| transportation | power grid transmission | generating or relay station | transmission line |
| | rail system | rail station | railroad tracks |
| | road network(1) | intersection | pavement |
| | road network(2) | named road | intersection |
| | airport network | airport | non-stop flight |

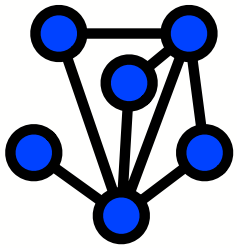


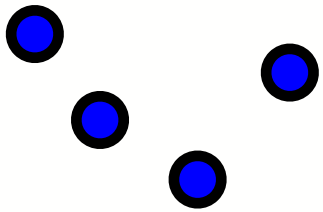


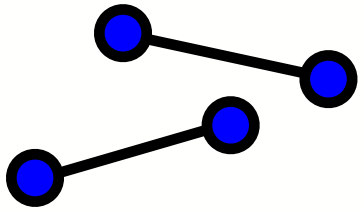


| network | vertex | edge |
|---|---|---|
| road network(1) | intersection | pavement |
| road network(2) | named road | intersection |

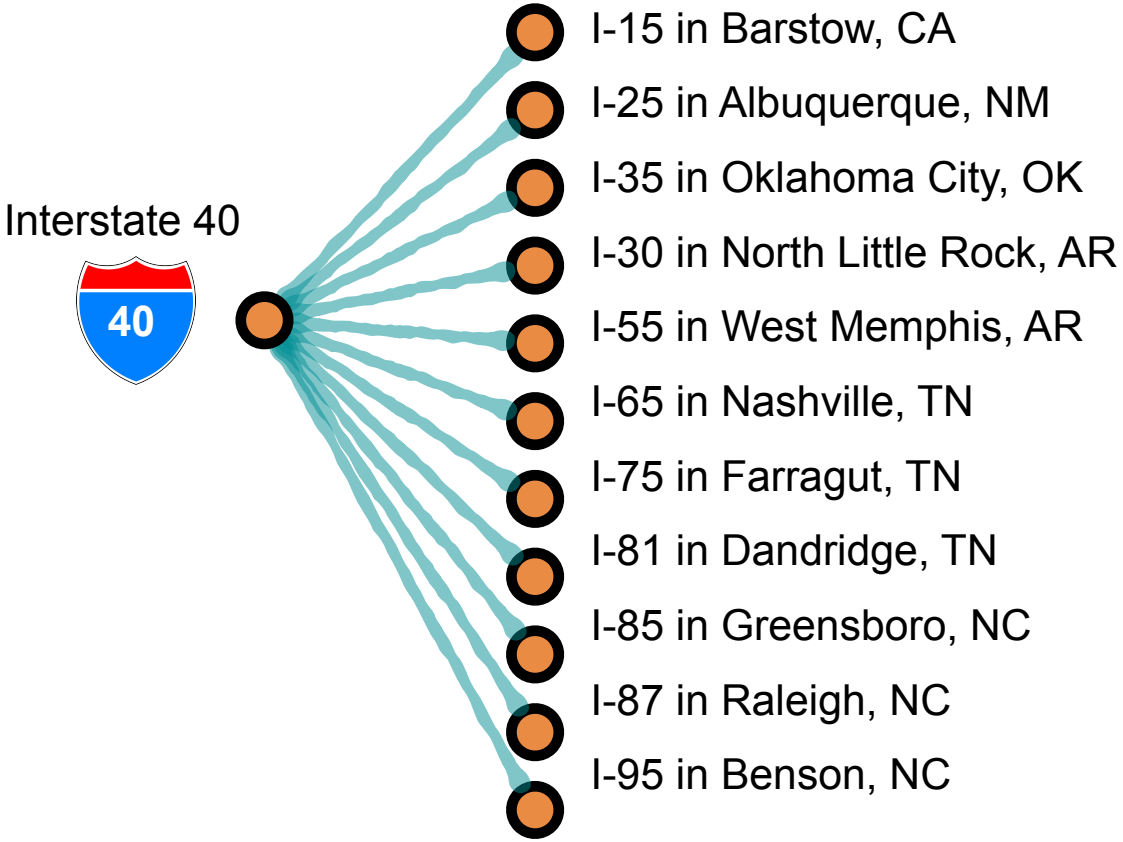


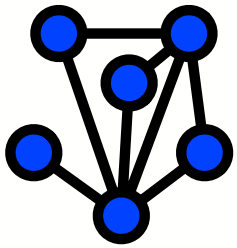
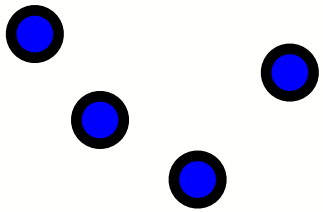
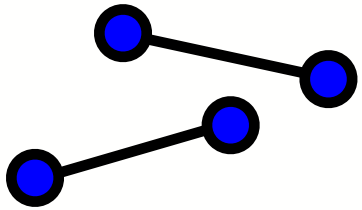




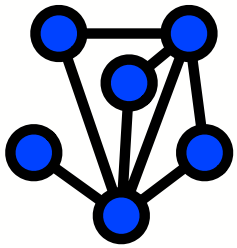


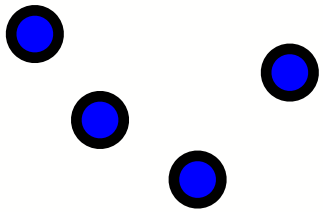
| network | vertex | edge |
|---|---|---|
| road network(1) | intersection | pavement |
| road network(2) | named road | intersection |

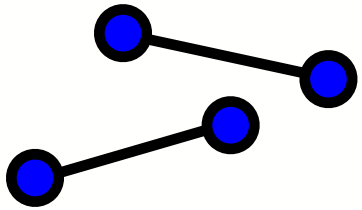


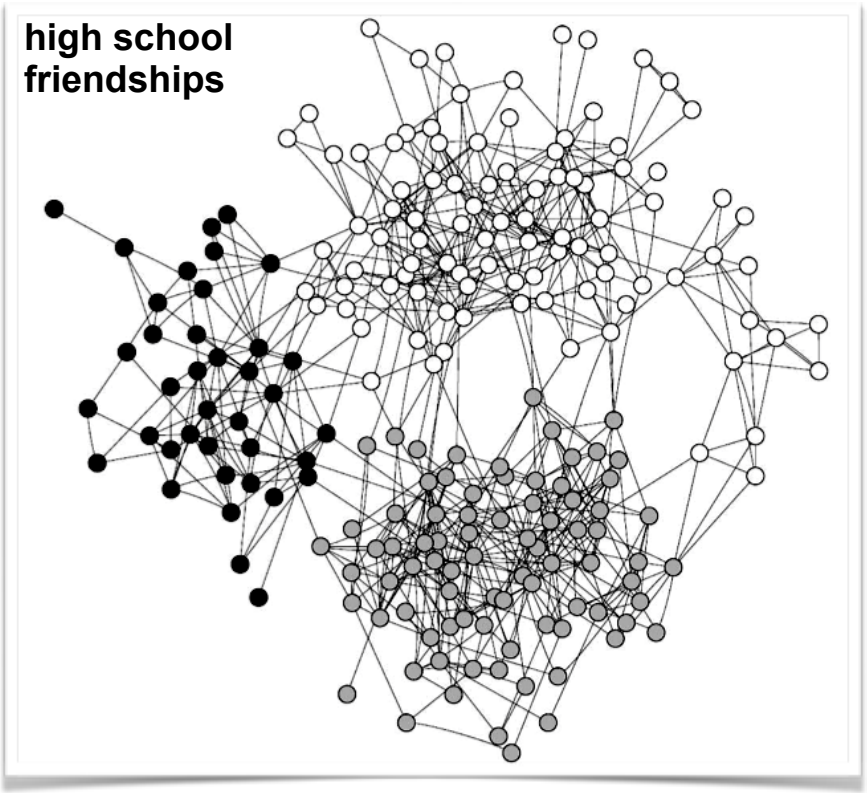
| | network | vertex | edge |
|---|---|---|---|
| technological | Internet(1) | computer | IP network adjacency |
| | Internet(2) | autonomous system (ISP) | BGP connection |
| information | software | function | function call |
| | World Wide Web | web page | hyperlink |
| | documents | article, patent, or legal case | citation |
| transportation | power grid transmission | generating or relay station | transmission line |
| | rail system | rail station | railroad tracks |
| | road network(1) | intersection | pavement |
| | road network(2) | named road | intersection |
| | airport network | airport | non-stop flight |
| social | friendship network | person | friendship |
| | sexual network | person | intercourse |

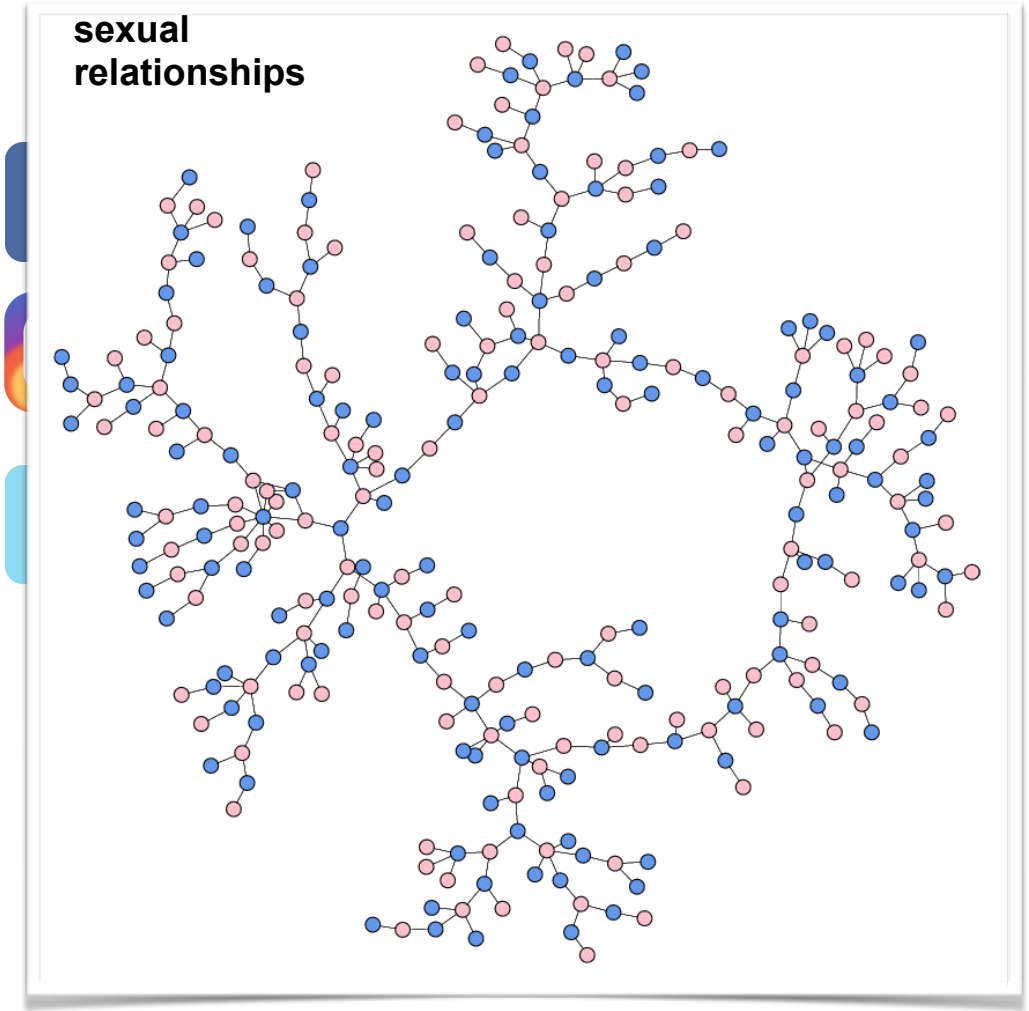


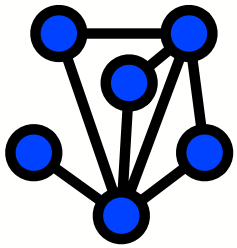
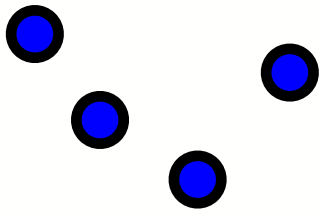
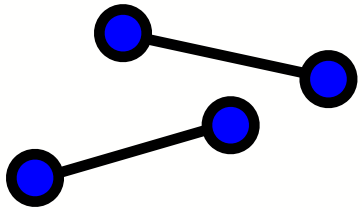




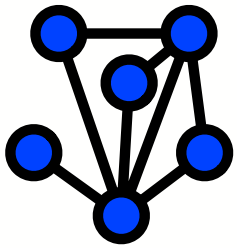
| network | vertex | edge |
|---|---|---|
| friendship network | person | friendship |
| sexual network | person | intercourse |

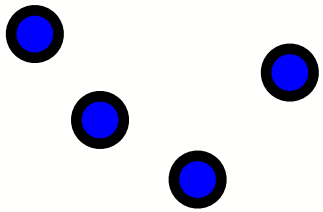


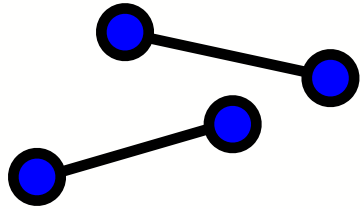



| | network | vertex | edge |
|---|---|---|---|
| technological | Internet(1) | computer | IP network adjacency |
| | Internet(2) | autonomous system (ISP) | BGP connection |
| information | software | function | function call |
| | World Wide Web | web page | hyperlink |
| | documents | article, patent, or legal case | citation |
| transportation | power grid transmission | generating or relay station | transmission line |
| | rail system | rail station | railroad tracks |
| | road network(1) | intersection | pavement |
| | road network(2) | named road | intersection |
| | airport network | airport | non-stop flight |
| social | friendship network | person | friendship |
| | sexual network | person | intercourse |
| biological | metabolic network | metabolite | metabolic reaction |
| | gene regulatory network | gene | regulatory effect |
| | neuronal network | neuron | synapse |
| | food web | species | predation or resource transfer |

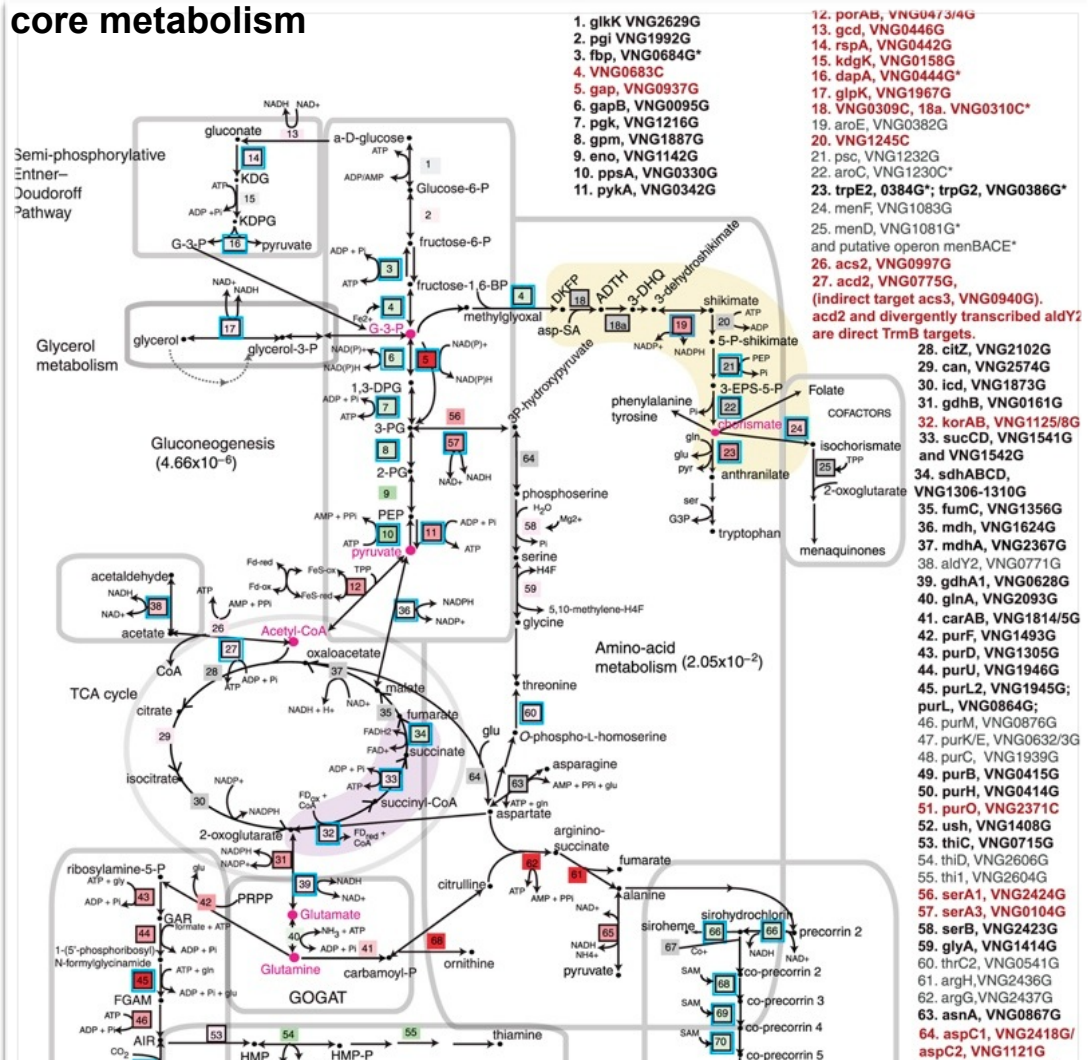




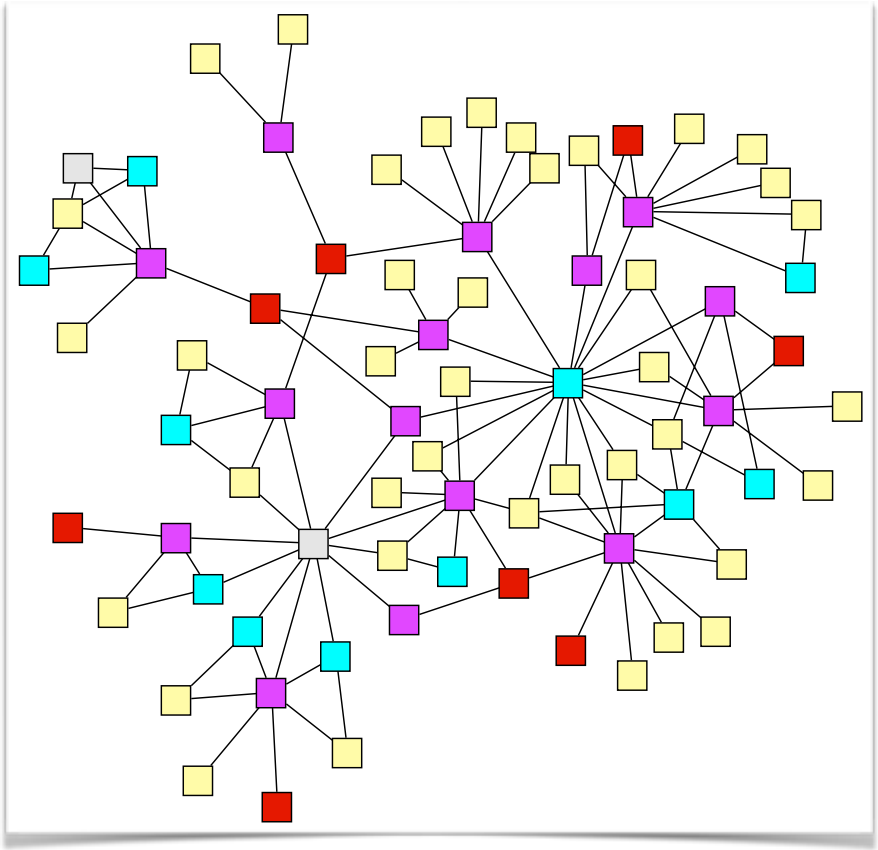


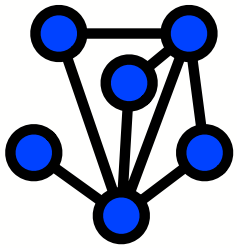
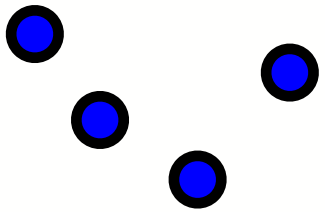
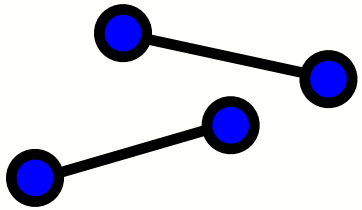
| network | vertex | edge |
|---|---|---|
| metabolic network | metabolite | metabolic reaction |
| food web | species | predation or resource transfer |

**core metabolism**

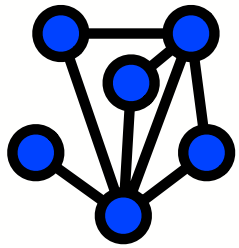


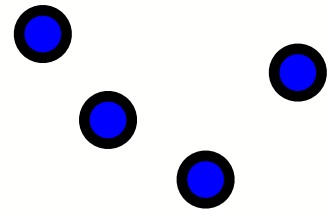

**grassland foodweb**

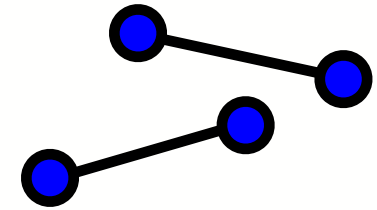


| | network | vertex | edge |
|---|---|---|---|
| technological | Internet(1) | computer | IP network adjacency |
| | Internet(2) | autonomous system (ISP) | BGP connection |
| information | software | function | function call |
| | World Wide Web | web page | hyperlink |
| | documents | article, patent, or legal case | citation |
| transportation | power grid transmission | generating or relay station | transmission line |
| | rail system | rail station | railroad tracks |
| | road network(1) | intersection | pavement |
| | road network(2) | named road | intersection |
| | airport network | airport | non-stop flight |
| social | friendship network | person | friendship |
| | sexual network | person | intercourse |
| biological | metabolic network | metabolite | metabolic reaction |
| | gene regulatory network | gene | regulatory effect |
| | neuronal network | neuron | synapse |
| economic | food web | species | predation or resource transfer |
| | faculty hiring | universities | faculty hiring |

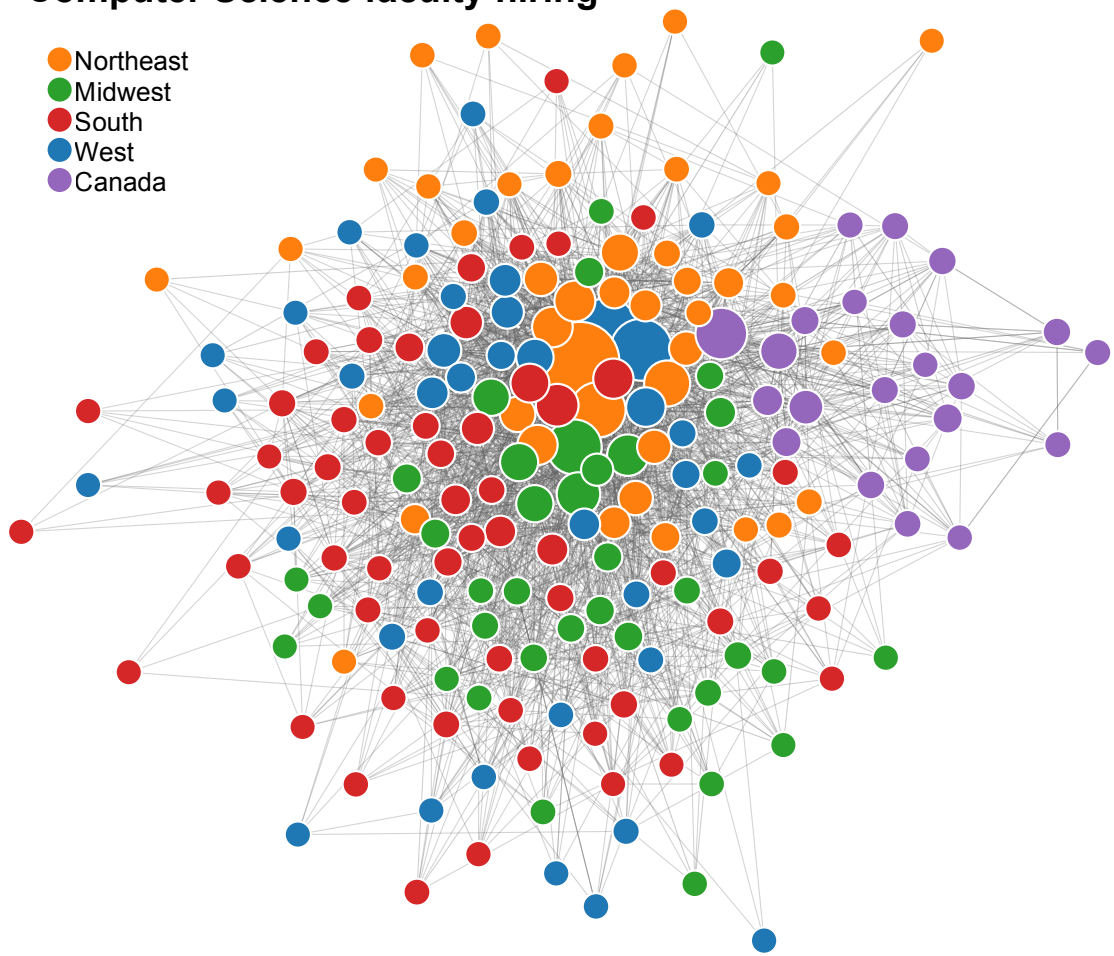






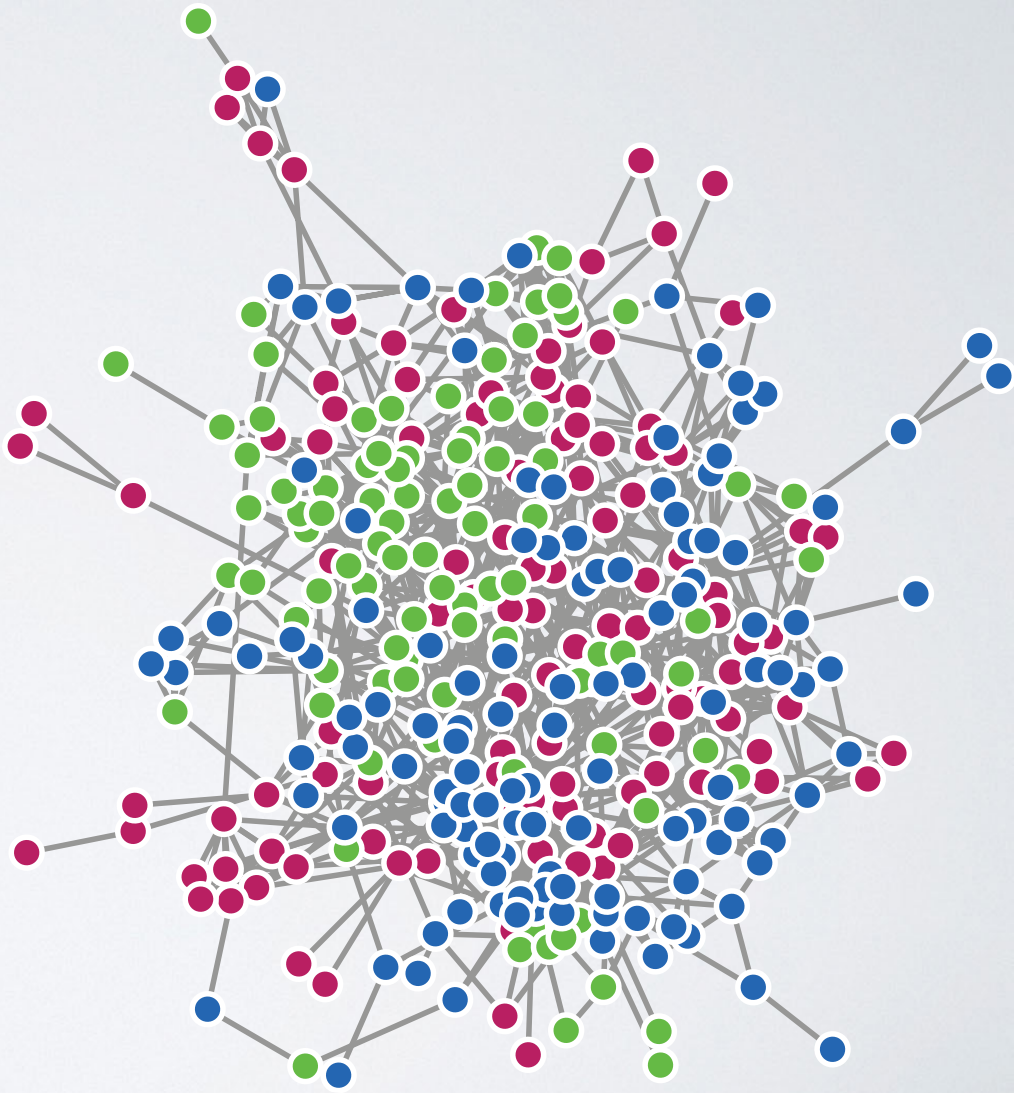
| network | vertex | edge |
|---|---|---|
| faculty hiring | universities | faculty hiring |

**Computer Science faculty hiring**



# analyzing networks

what real networks look like…



# analyzing networks

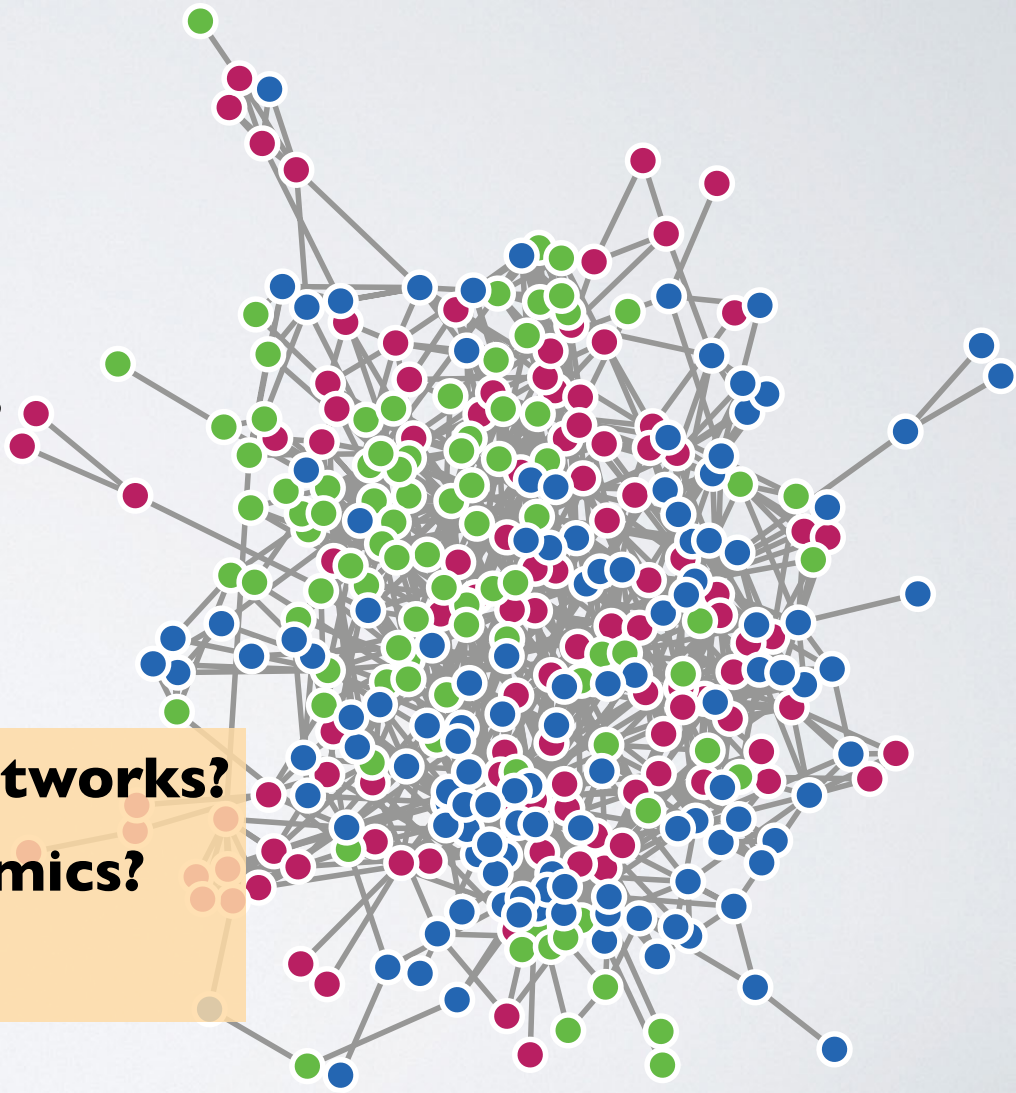
what real networks look like…

questions:

- **how are the edges organized?**
- **how do vertices differ?**
- **does network location matter?**
- **are there underlying patterns?**

what we want to know

- **what processes shape these networks?**
- **how does network shape dynamics?**
- **how can we tell?**



# analyzing networks

what we want : **understand its structure**

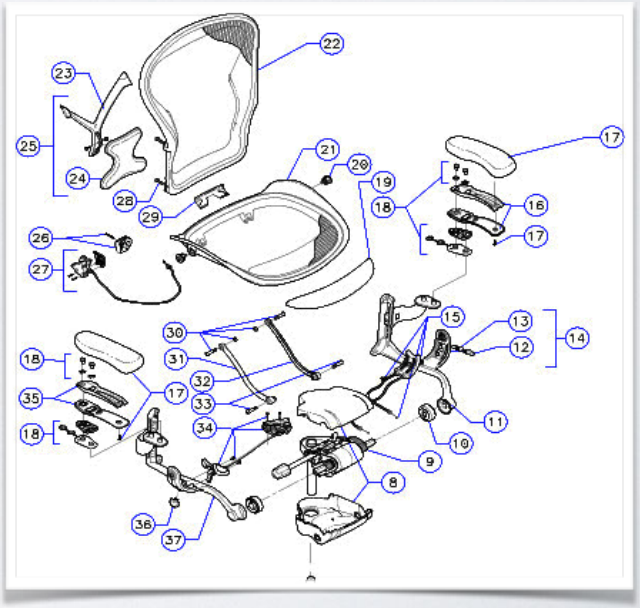
$$f : \text{object} \rightarrow \{\theta_1, \ldots, \theta_k\}$$

- **what are the fundamental parts?**
- **how are these parts organized?**
- **where are the degrees of freedom $\vec{\theta}$?**
- **how can we define an abstract class?**
- **structure — dynamics — function?**

what does ***local-level structure*** look like?

what does ***large-scale structure*** look like?

how does ***structure constrain*** function?



# analyzing networks

6 major approaches

1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)

# analyzing networks

6 major approaches

1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)
2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling

# analyzing networks

6 major approaches

1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)
2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling
3. **null models:** use some kind of random graph to identify non-random patterns as deviations from the null

# analyzing networks

6 major approaches

1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)
2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling
3. **null models:** use some kind of random graph to identify non-random patterns as deviations from the null
4. **mechanisms / simulations:** explain structural or dynamical patterns as caused by specific process

# analyzing networks

6 major approaches

1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)
2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling
3. **null models:** use some kind of random graph to identify non-random patterns as deviations from the null
4. **mechanisms / simulations:** explain structural or dynamical patterns as caused by specific process
5. **predictive models:** fit parametric model of network structure & use it to predict missing or future data (edges, labels, etc.)

# analyzing networks

6 major approaches

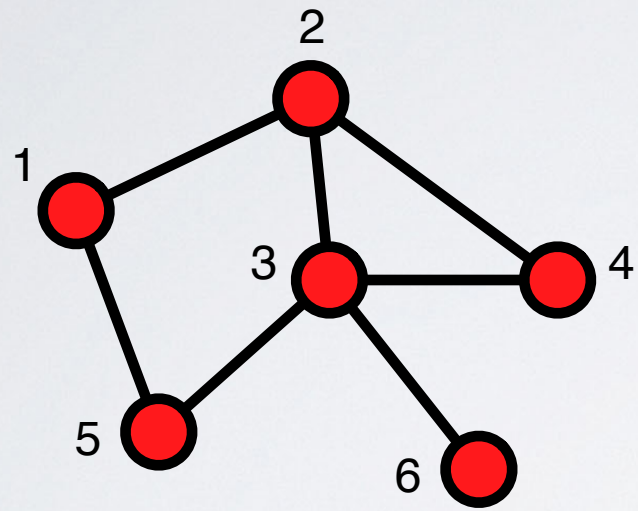
1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)
2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling
3. **null models:** use some kind of random graph to identify non-random patterns as deviations from the null
4. **mechanisms / simulations:** explain structural or dynamical patterns as caused by specific process
5. **predictive models:** fit parametric model of network structure & use it to predict missing or future data (edges, labels, etc.)
6. **network experiments:** manipulate structure and measure node-level or graph-level behavior as function of changes

# representing networks

## 4 representations

- ridiculogram — nice pictures, best for small networks
- ⭐ adjacency matrix — mathematically convenient & useful mental model
- ⭐ adjacency list — efficient computation
- edge list — efficient storage

# a *simple* network

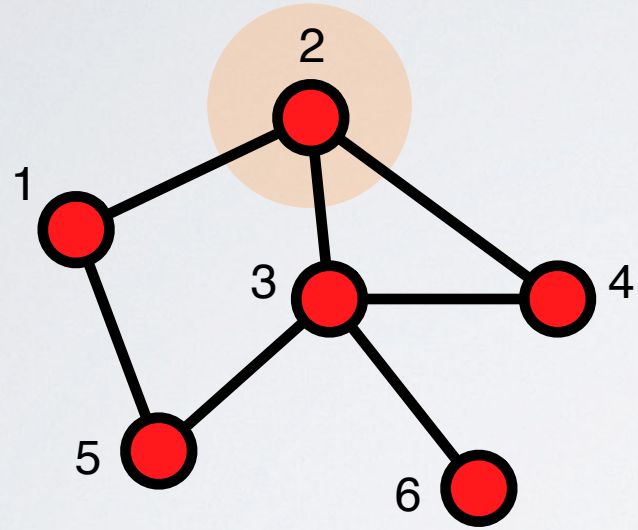



undirected

unweighted

no self-loops

# a *simple* network




undirected

unweighted

no self-loops

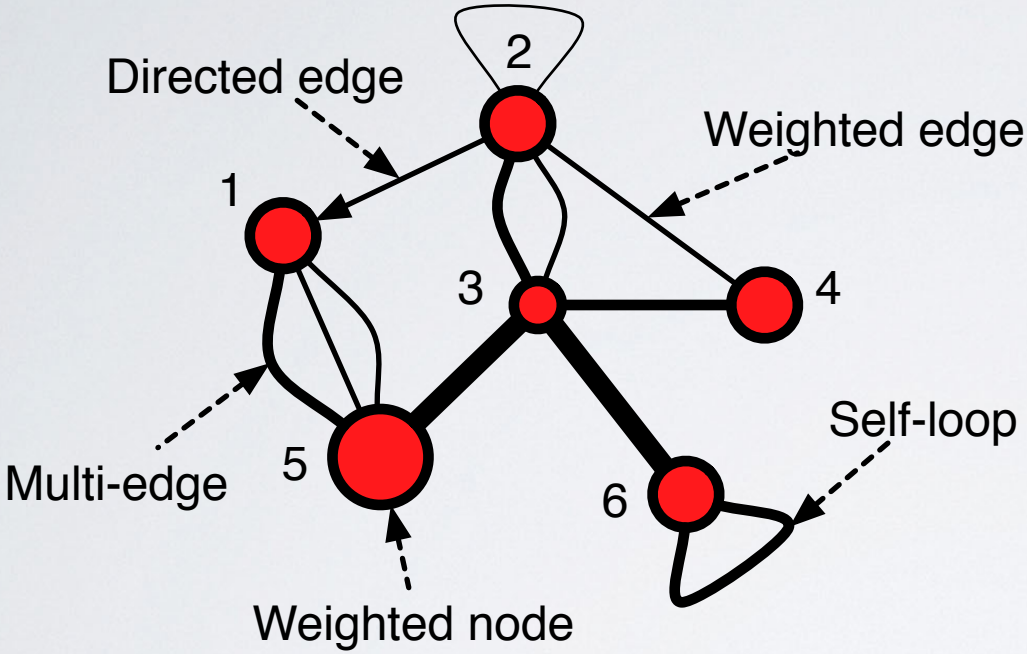
adjacency matrix

| $A$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 2 | 1 | 0 | 1 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 1 | 1 |
| 4 | 0 | 1 | 1 | 0 | 0 | 0 |
| 5 | 1 | 0 | 1 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 0 | 0 |

adjacency list

| $A$ |
|---|
| $1 \rightarrow \{2, 5\}$ |
| $2 \rightarrow \{1, 3, 4\}$ |
| $3 \rightarrow \{2, 4, 5, 6\}$ |
| $4 \rightarrow \{2, 3\}$ |
| $5 \rightarrow \{1, 3\}$ |
| $6 \rightarrow \{3\}$ |

# beyond simple graphs

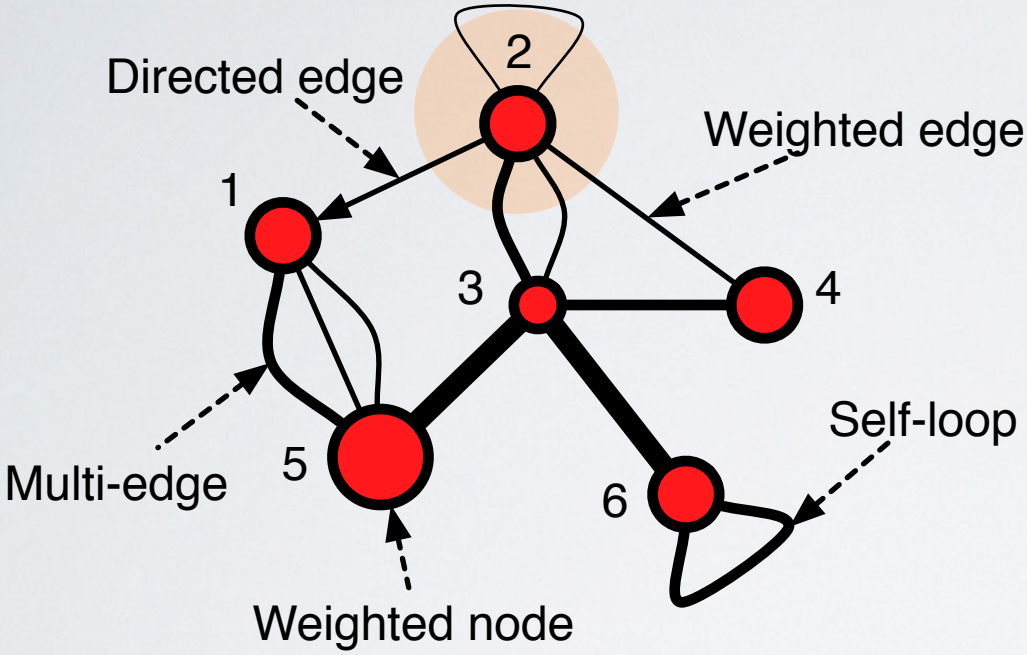



~~undirected~~

~~unweighted~~

~~no self-loops~~

# beyond simple graphs




## adjacency matrix

| $A$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | $\{1,1,2\}$ | 0 |
| 2 | 1 | $\frac{1}{2}$ | $\{2,1\}$ | 1 | 0 | 0 |
| 3 | 0 | $\{2,1\}$ | 0 | 2 | 4 | 4 |
| 4 | 0 | 1 | 2 | 0 | 0 | 0 |
| 5 | $\{1,1,2\}$ | 0 | 4 | 0 | 0 | 0 |
| 6 | 0 | 0 | 4 | 0 | 0 | 2 |

## adjacency list

| $A$ | |
|---|---|
| 1 | $\rightarrow \{(5,1),(5,1),(5,2)\}$ |
| 2 | $\rightarrow \{(1,1),(2,\frac{1}{2}),(3,2),(3,1),(4,1)\}$ |
| 3 | $\rightarrow \{(2,2),(2,1),(4,2),(5,4),(6,4)\}$ |
| 4 | $\rightarrow \{(2,1),(3,2)\}$ |
| 5 | $\rightarrow \{(1,1),(1,1),(1,2),(3,4)\}$ |
| 6 | $\rightarrow \{(3,4),(6,2)\}$ |

# beyond simple graphs

## attributes of

| edges | nodes | network |
| --- | --- | --- |
| unweighted | | |
| weighted | | |
| signed | | |
| undirected | | |
| directed | | |
| multigraph | | |
| timestamps | | |

# beyond simple graphs

## attributes of

| edges | nodes | network |
| --- | --- | --- |
| unweighted | metadata | |
| weighted | attributes | |
| signed | locations | |
| undirected | state variables | |
| directed | | |
| multigraph | | |
| timestamps | | |

# beyond simple graphs

## attributes of

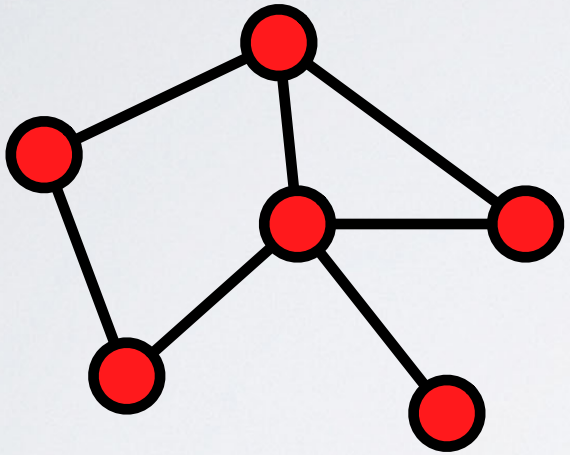
| edges | nodes | network |
|---|---|---|
| unweighted | metadata | sparse |
| weighted | attributes | dense |
| signed | locations | bipartite |
| undirected | state variables | projection |
| directed | | acyclic |
| multigraph | | temporal |
| timestamps | | multiplex |

# describing networks

**aka,** summarizing a network's structure

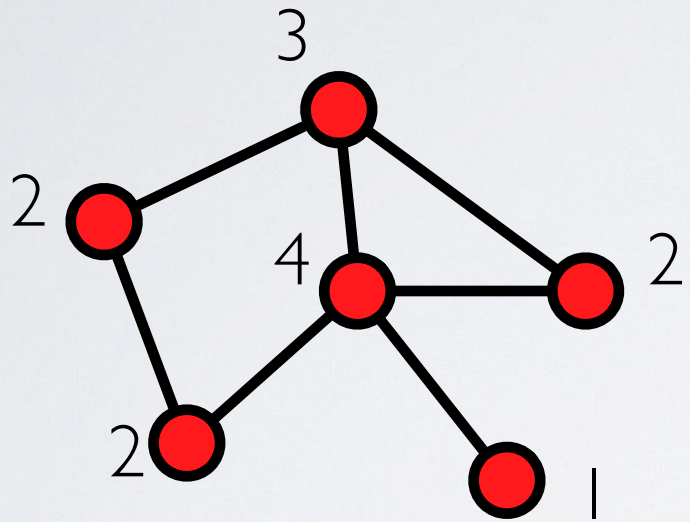
$$f : G \rightarrow \underbrace{\{x_1, \ldots, x_k\}}_{\text{summary statistics}}$$

# describing networks



**degree:**
the first order description
of a network

# describing networks

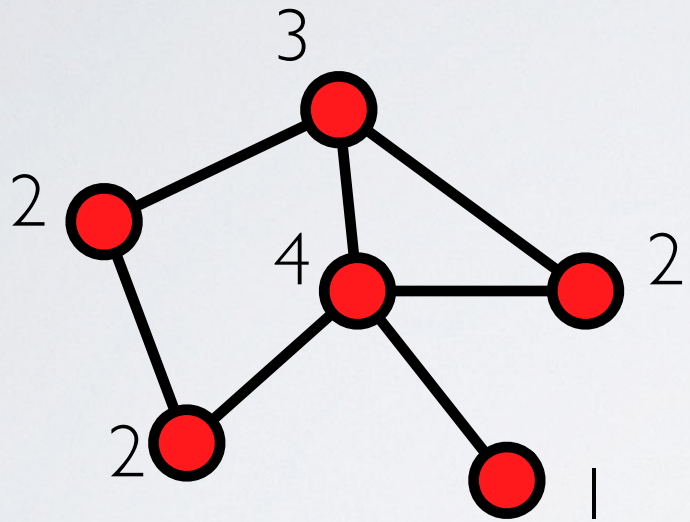



**degree:**

number of connections $k$

$$k_i = \sum_j A_{ij}$$

# describing networks

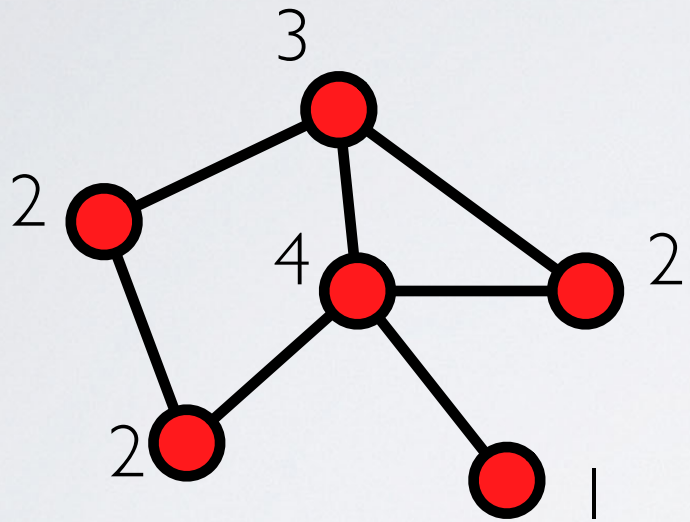



**degree:**
number of connections $k$

$$k_i = \sum_j A_{ij}$$

number of edges $$m = \frac{1}{2}\sum_{i=1}^{n} k_i = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} A_{ji}$$

# describing networks



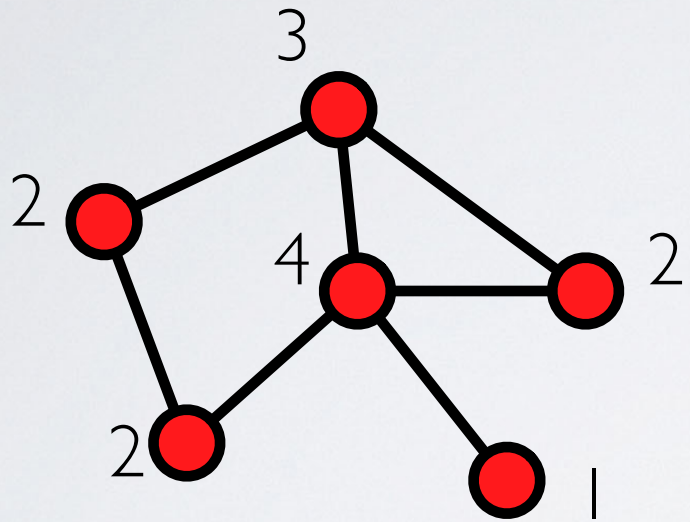

**degree:**
number of connections $k$

$$k_i = \sum_j A_{ij}$$

number of edges $$m = \frac{1}{2}\sum_{i=1}^{n} k_i = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} = \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} A_{ji}$$

mean degree $$\langle k \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i = \frac{2m}{n}$$

# describing networks



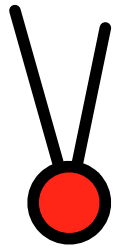

**degree:**
number of connections $k$

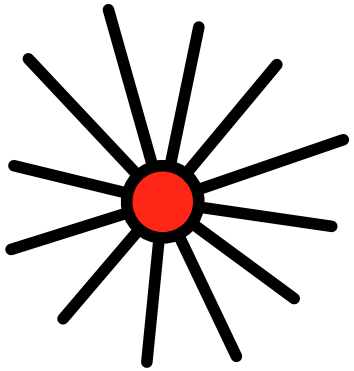
$$k_i = \sum_j A_{ij}$$

degree sequence $\{1, 2, 2, 2, 3, 4\}$

degree distribution $\Pr(k) = \left[\left(1, \frac{1}{6}\right), \left(2, \frac{3}{6}\right), \left(3, \frac{1}{6}\right), \left(4, \frac{1}{6}\right)\right]$

# node degrees




"low" degree

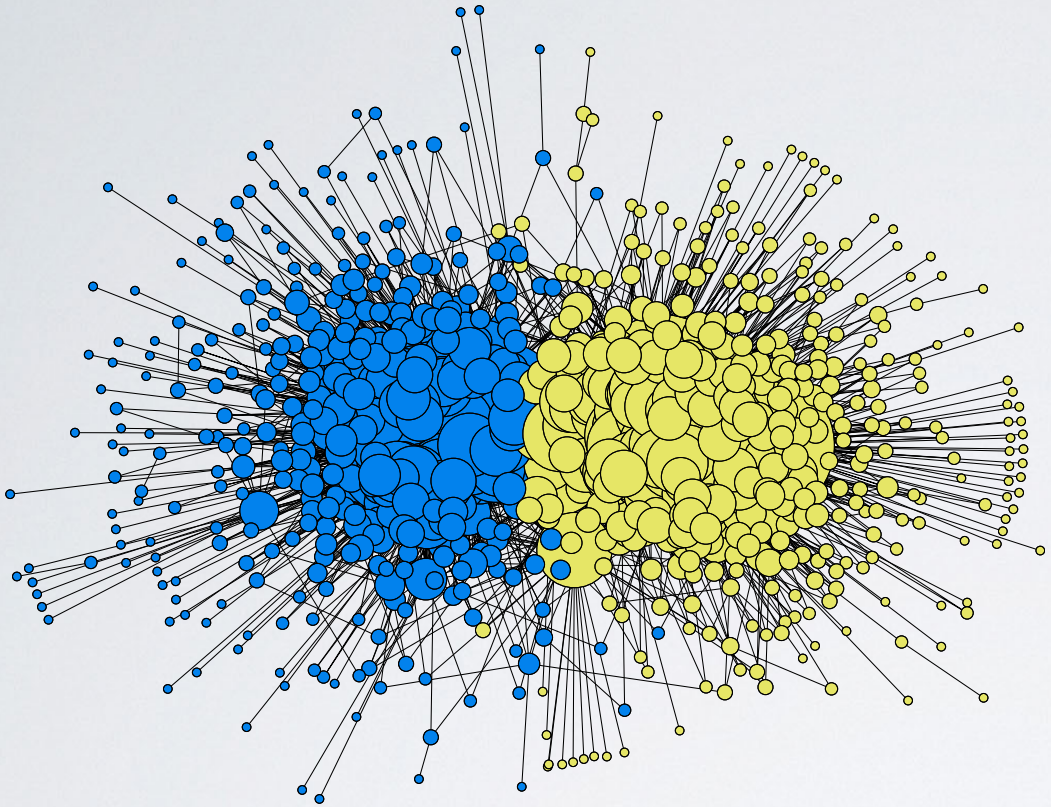
"high" degree

**node degrees :** what impact does having fewer or more connections have?

more information? more exposure to disease? more robust connectivity? more influence? less bandwidth? etc.

* scare quotes because 'low' and 'high' are relative terms

# degree distributions



political blogs (2004)

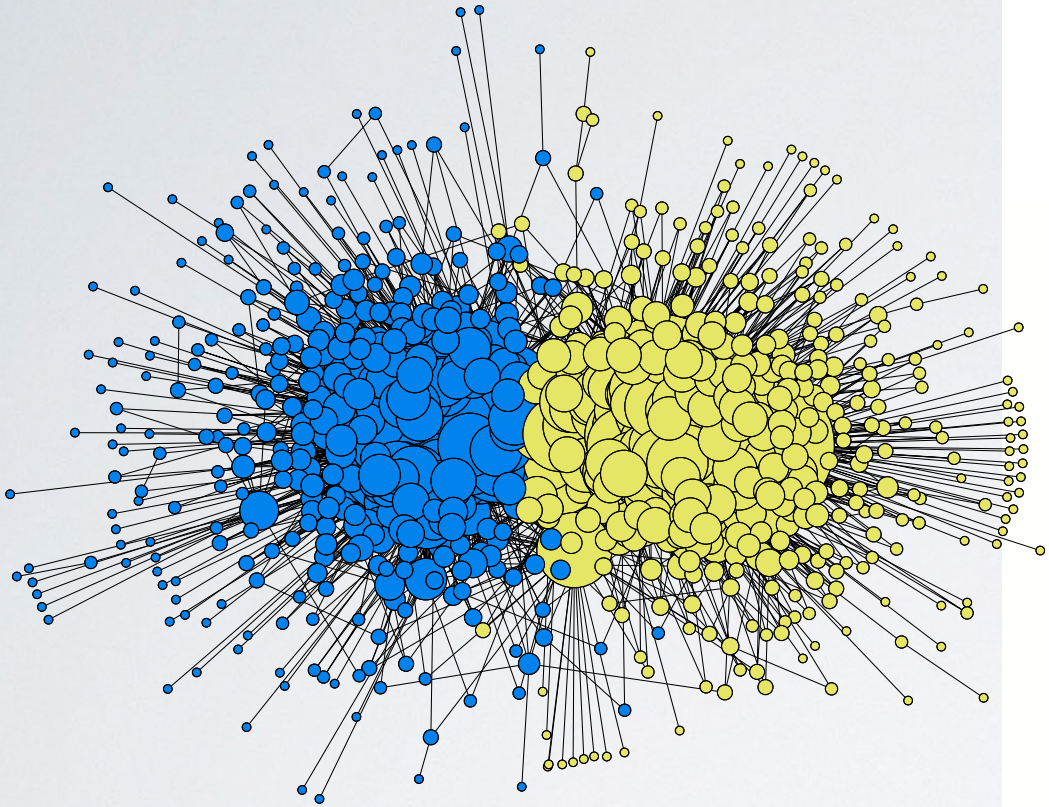
$$n = 1490$$
$$m = 19090$$
$$\langle k \rangle = 25.6$$

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\Pr(k)$ | 0.271 | 0.072 | 0.052 | 0.034 | 0.026 | 0.033 | 0.020 | 0.017 | 0.011 | 0. |

this network available via icon.colorado.edu

* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

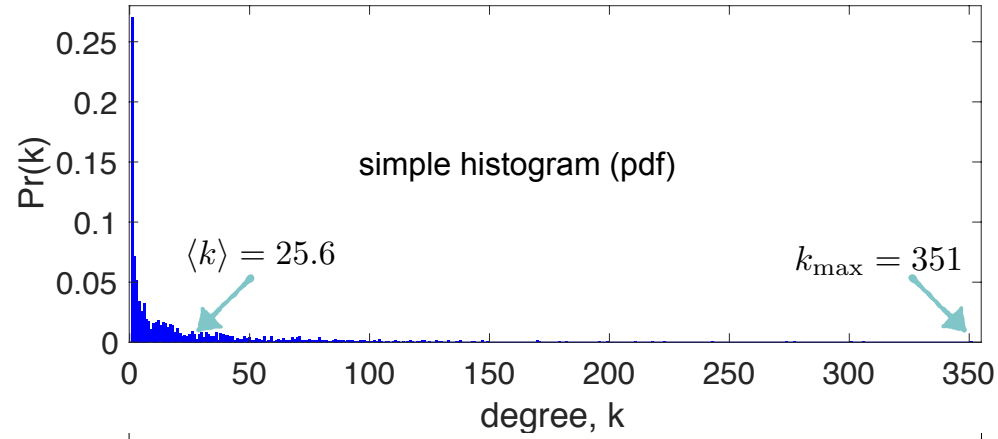
# degree distributions



political blogs (2004)

simple pdf:

$$\Pr(K = k) \text{ vs. } k$$




semilog x pdf

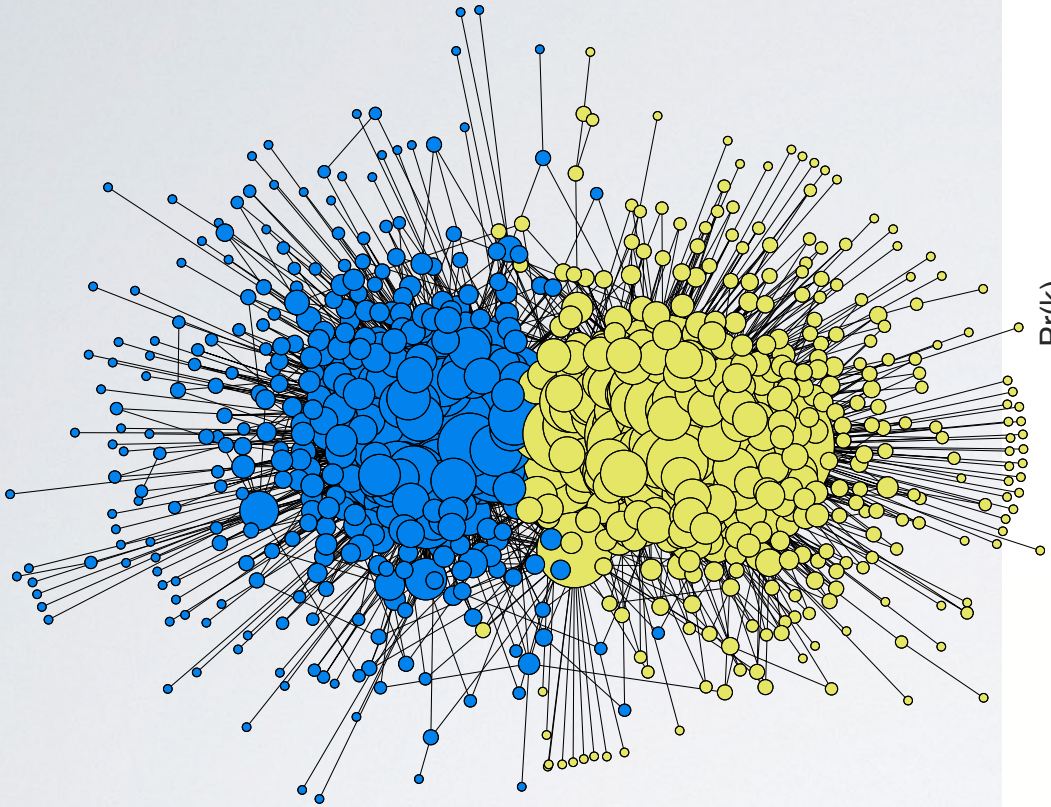
loglog pdf

loglog ccdf

this network available via icon.colorado.edu

* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

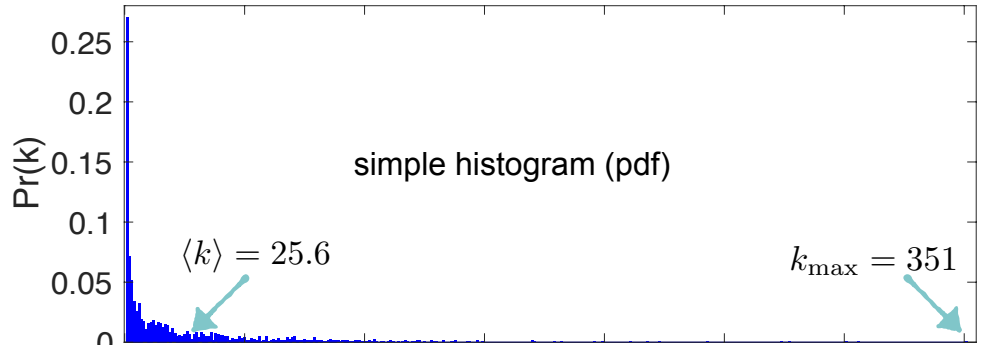
# degree distributions

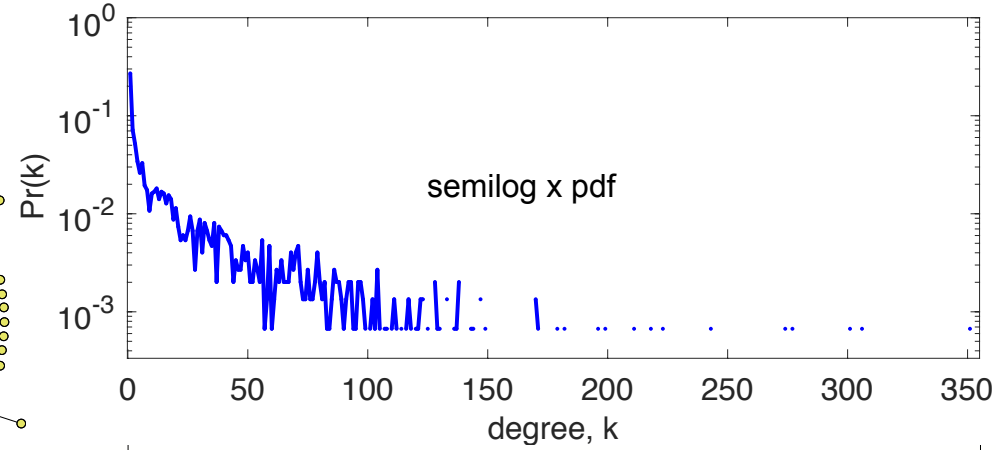


political blogs (2004)

simple pdf:

$\log_{10} \Pr(K = k)$ vs. $k$



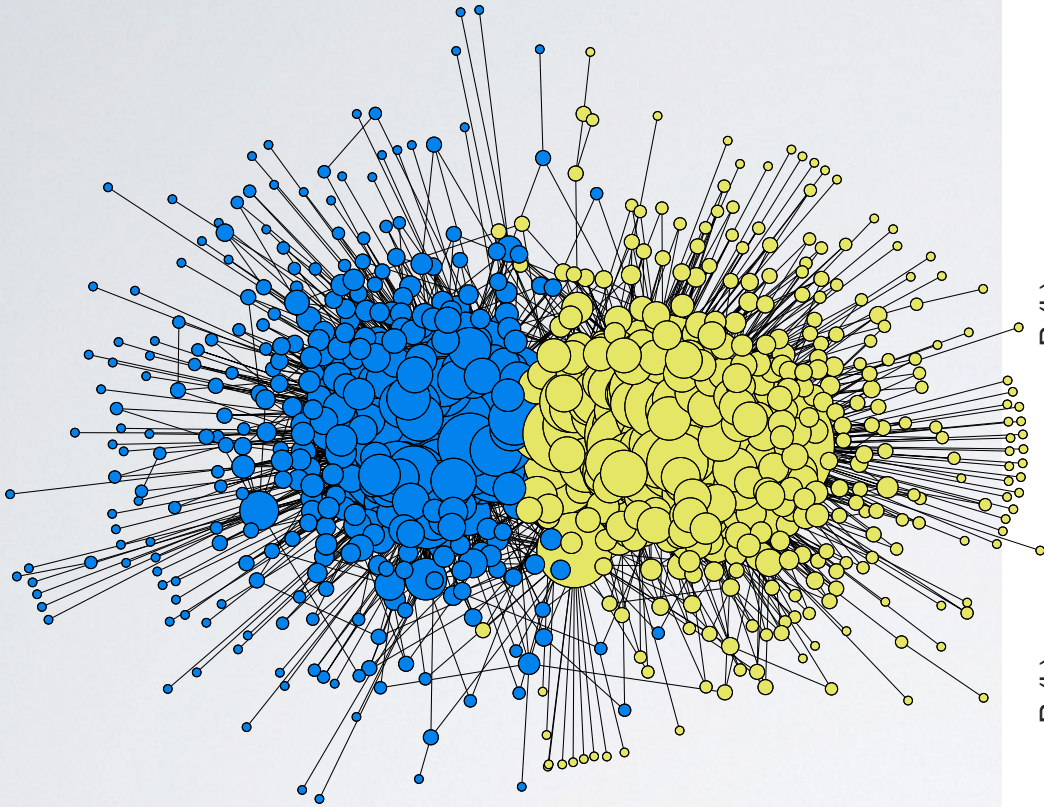

loglog pdf

loglog ccdf

this network available via icon.colorado.edu

* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

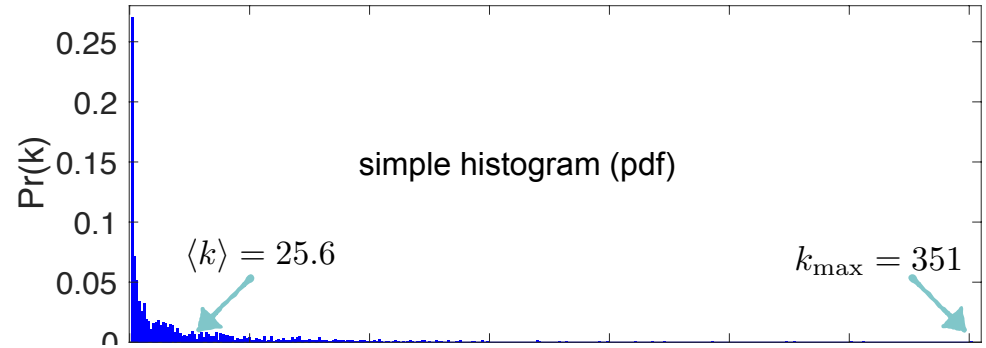
# degree distributions

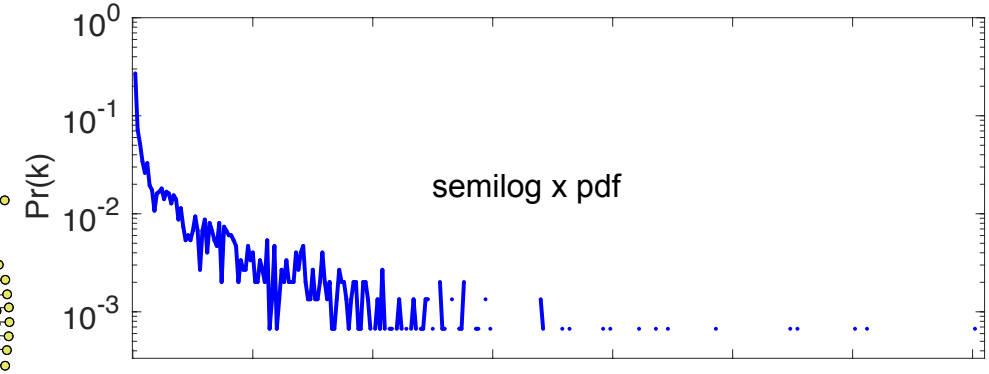


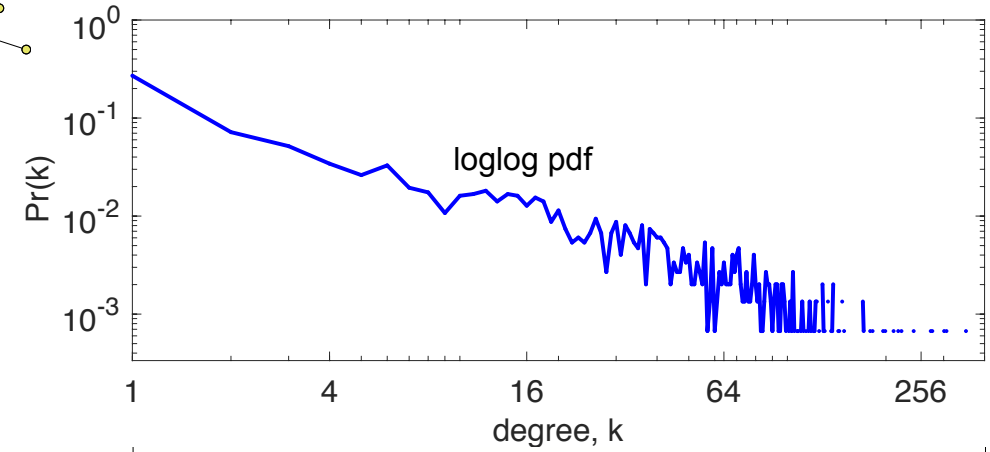
political blogs (2004)

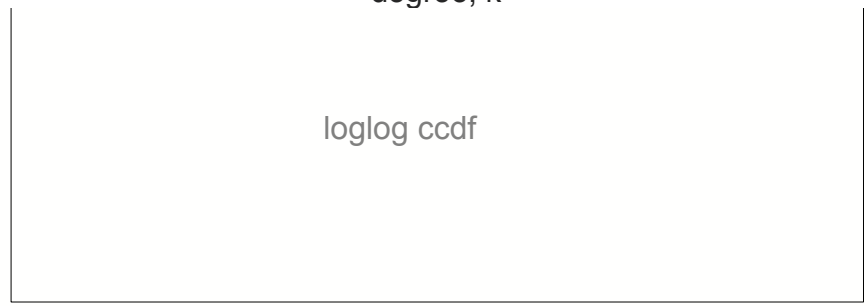
simple pdf:

$\log_{10} \Pr(K = k)$ vs. $\log_{10} k$



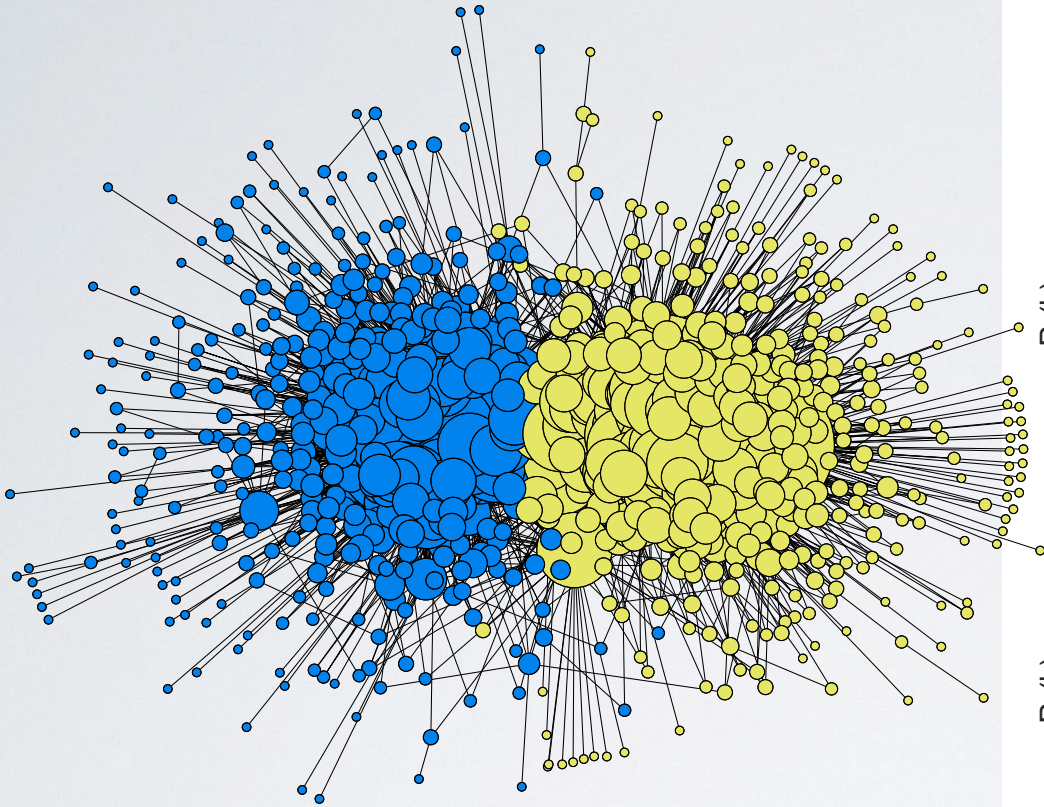




this network available via icon.colorado.edu

* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

# degree distributions



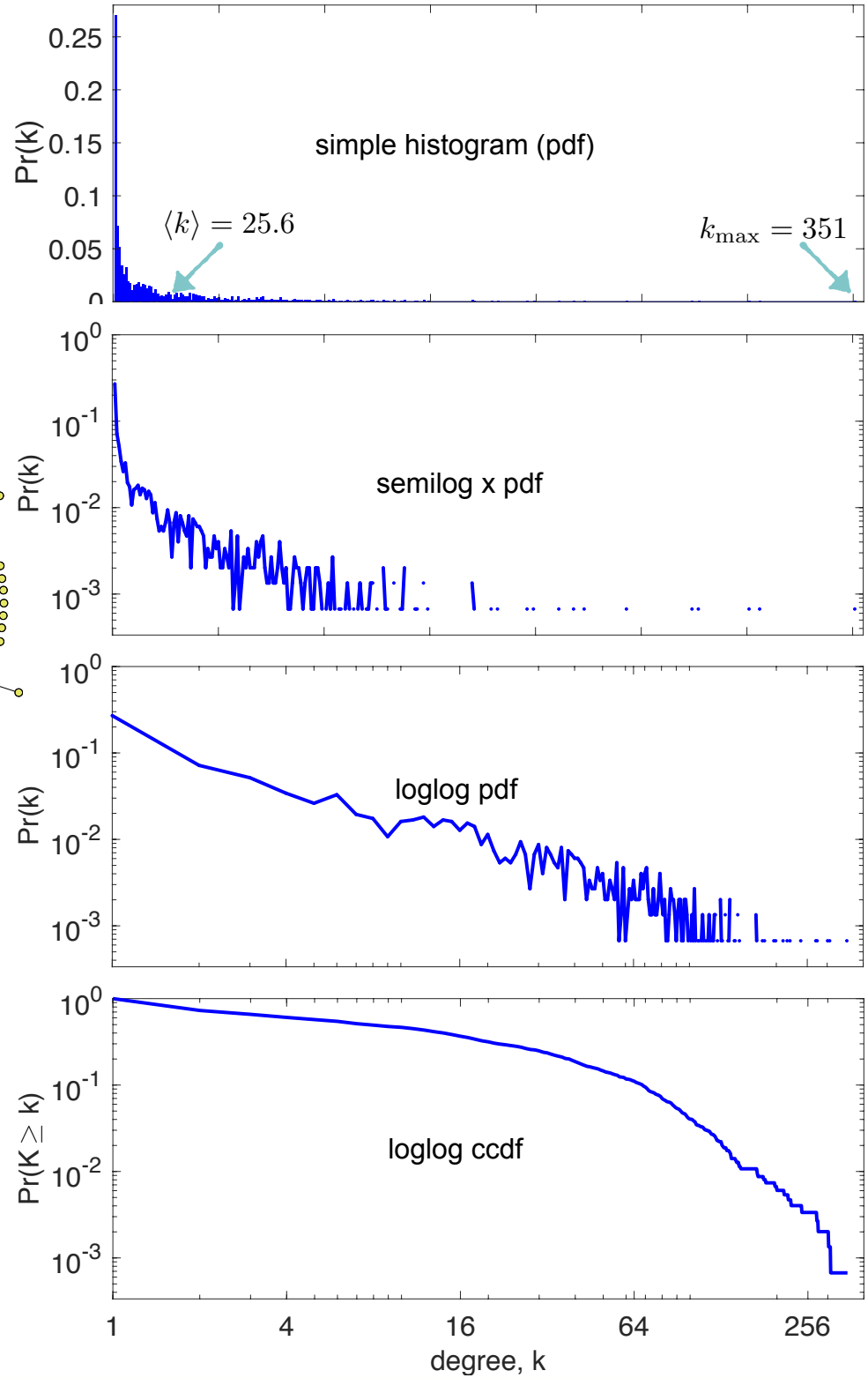
political blogs (2004)

complementary cdf:

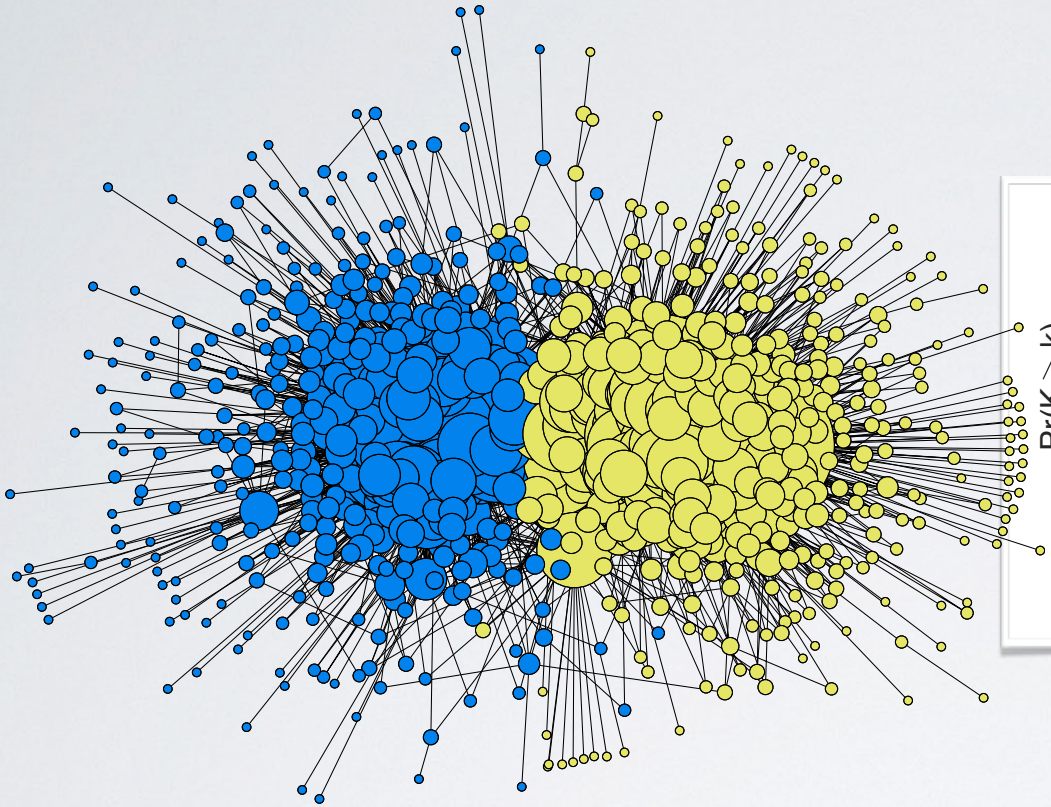
$$\Pr(K \geq k) = \sum_{j=k}^{n} \Pr(K = j)$$

this network available via icon.colorado.edu

* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

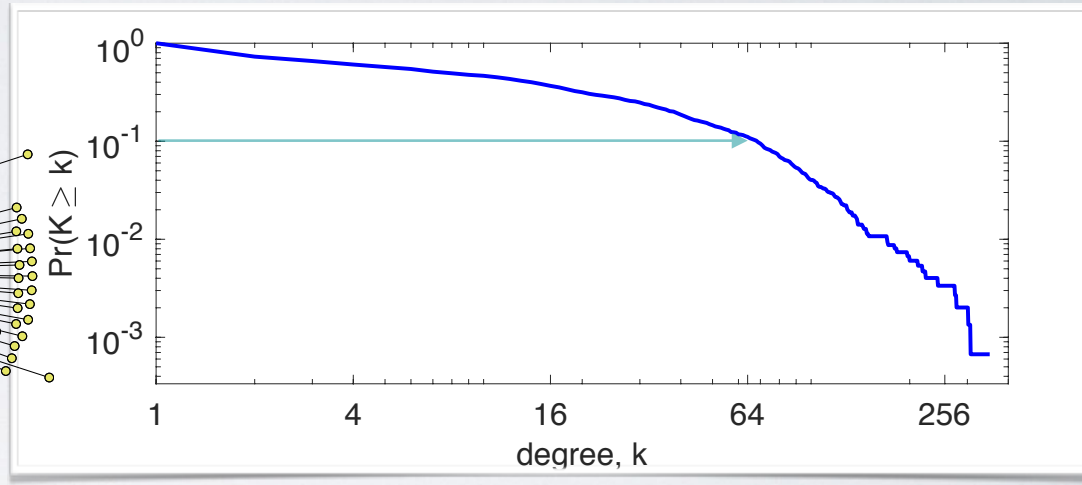


# degree distributions



political blogs (2004)

$$n = 1490$$
$$m = 19090$$
$$\langle k \rangle = 25.6$$

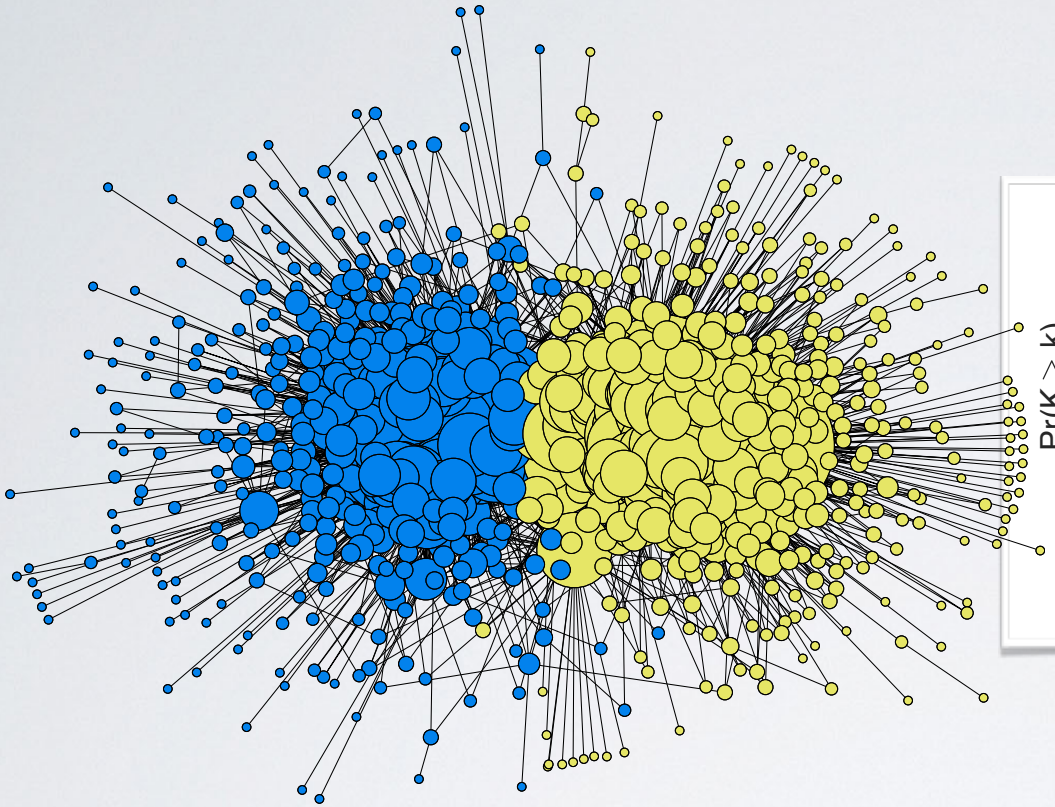



- 90% (1349) have $k \leq 67$ connecting to 53% of all $m$

this network available via icon.colorado.edu

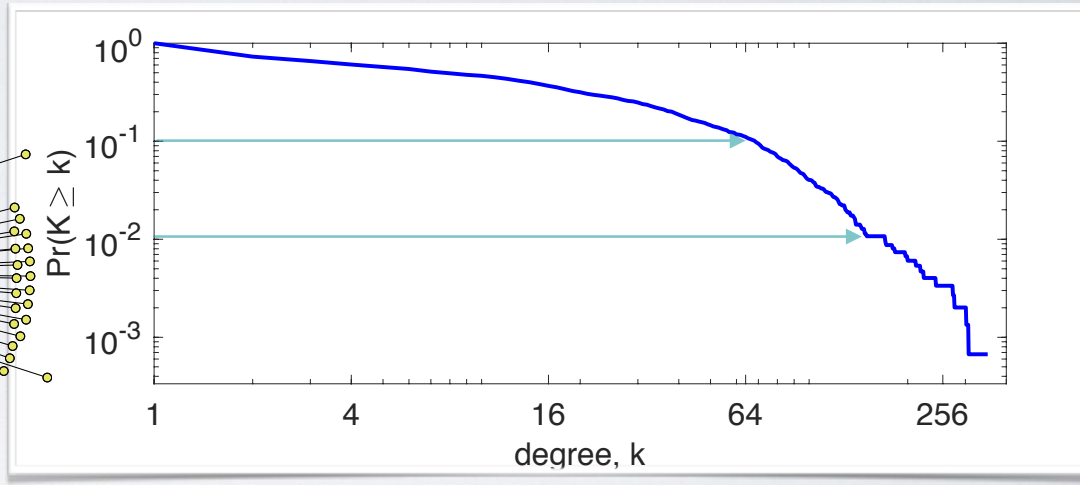
* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

# degree distributions



political blogs (2004)

$$n = 1490$$
$$m = 19090$$
$$\langle k \rangle = 25.6$$

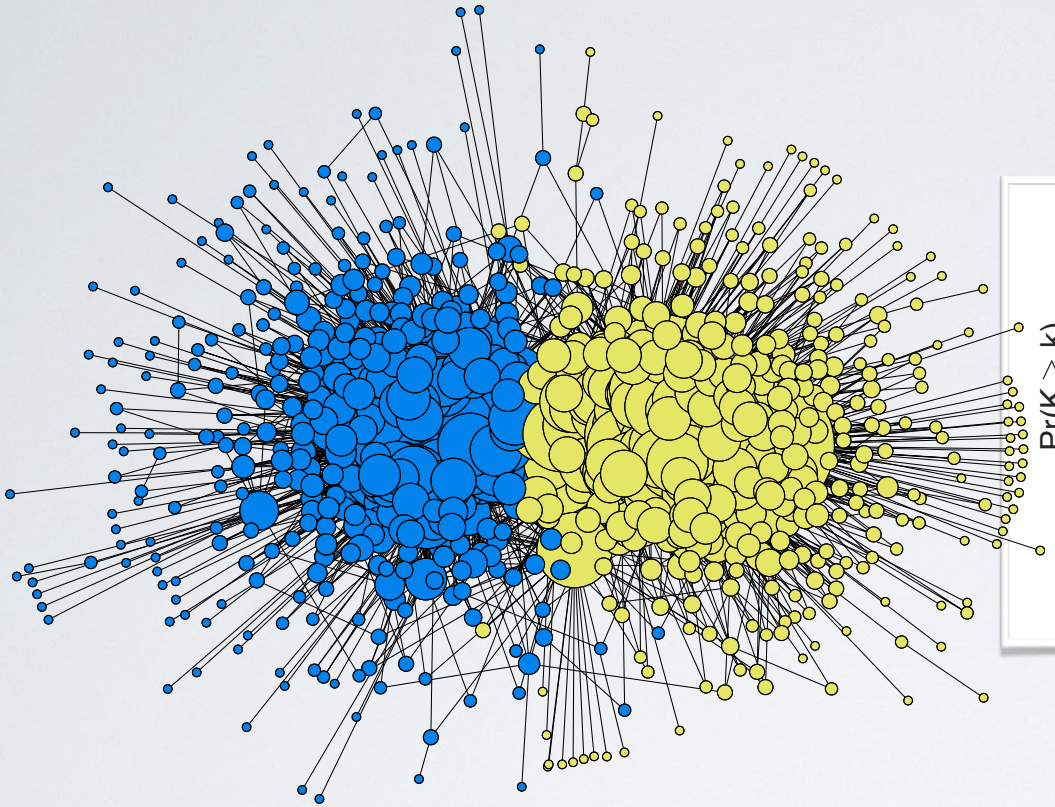



- 90% (1349) have $k \leq 67$ connecting to 53% of all $m$
- only 1% (14) have $k > 169$ connecting to 10% of all $m$

this network available via icon.colorado.edu

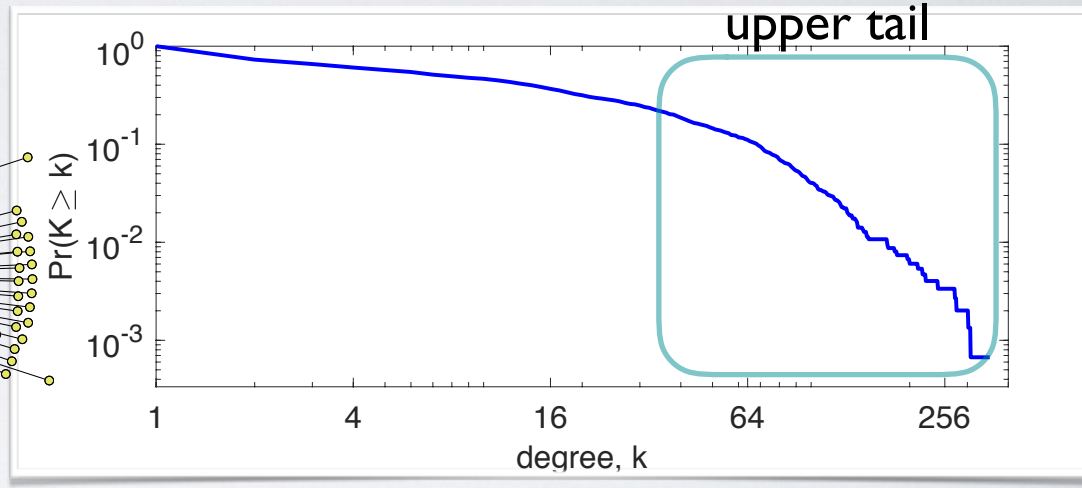
* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

# degree distributions



political blogs (2004)

$$n = 1490$$
$$m = 19090$$
$$\langle k \rangle = 25.6$$

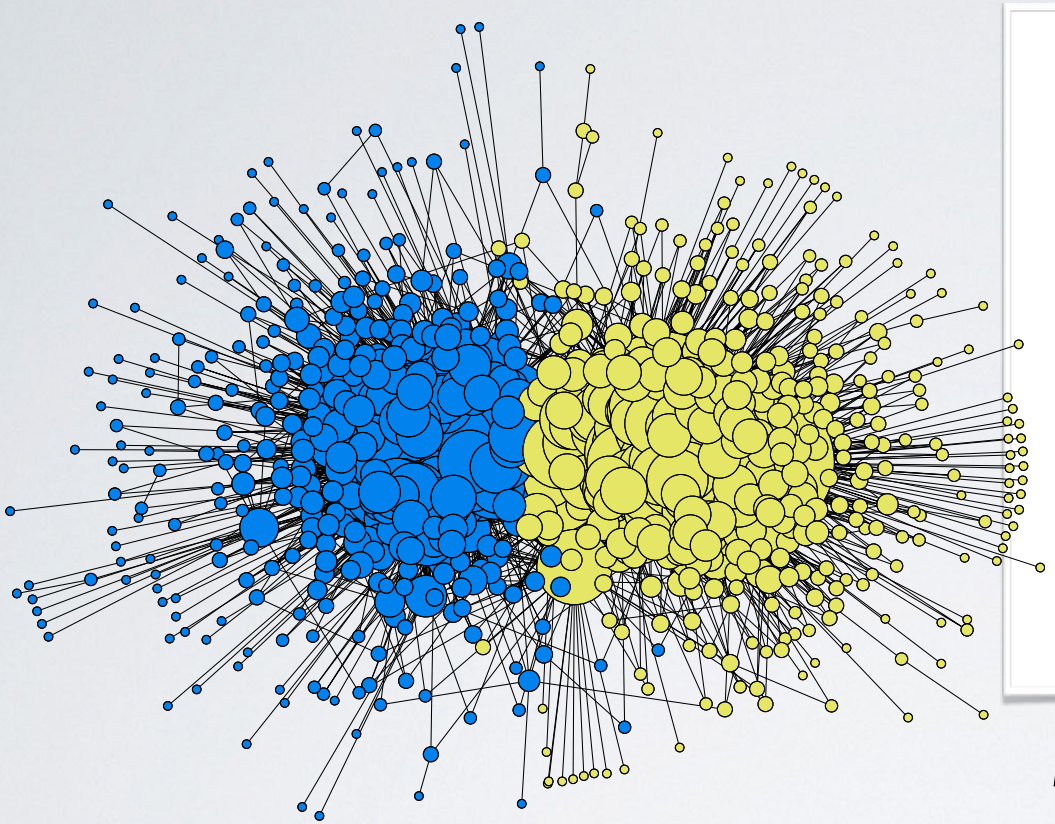



- 90% (1349) have $k \leq 67$ connecting to 53% of all $m$
- only 1% (14) have $k > 169$ connecting to 10% of all $m$

this network available via icon.colorado.edu

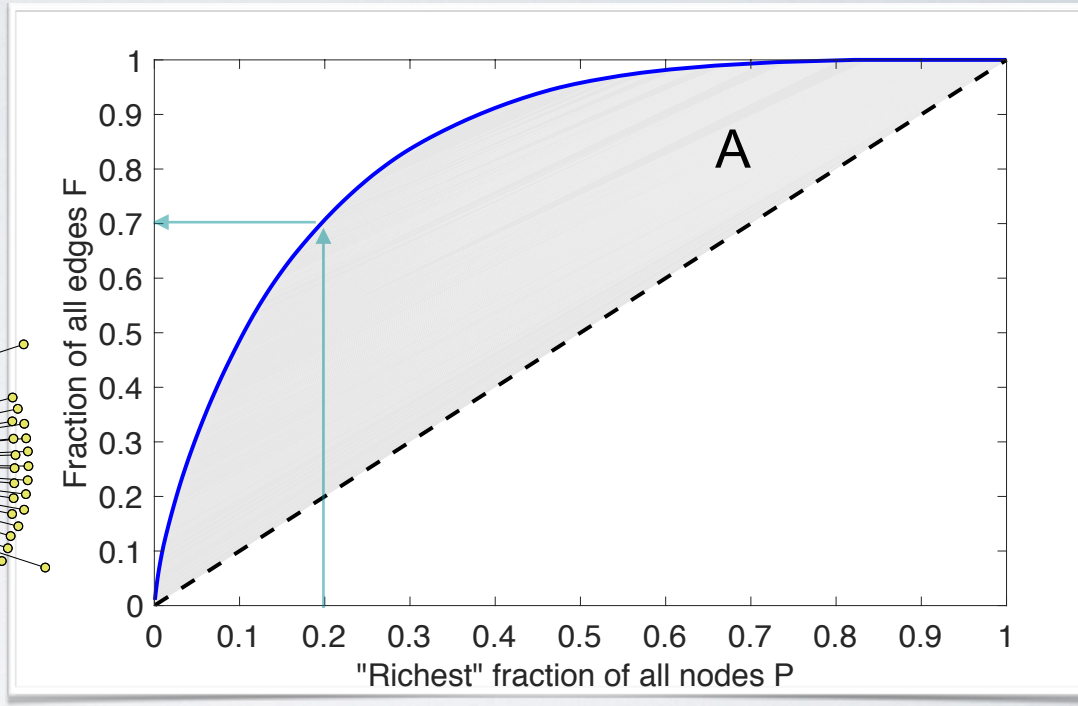
* Karrer & Newman, Phys. Rev. E. 83, 016107 (2011)

# degree distributions



political blogs (2004)

$$n = 1490$$
$$m = 19090$$
$$\langle k \rangle = 25.6$$




*Lorenz curve*

- fraction of all edges *F* held by "richest" fraction of all nodes *P*
- Gini coefficient: $G = 2A$

$$= 0.69$$

* the 80-20 rule: 80% of "wealth" held by 20% of "population"

# exploring degree distributions

the complementary CDF:

$$\Pr(K \geq k) = 1 - \Pr(K < k)$$

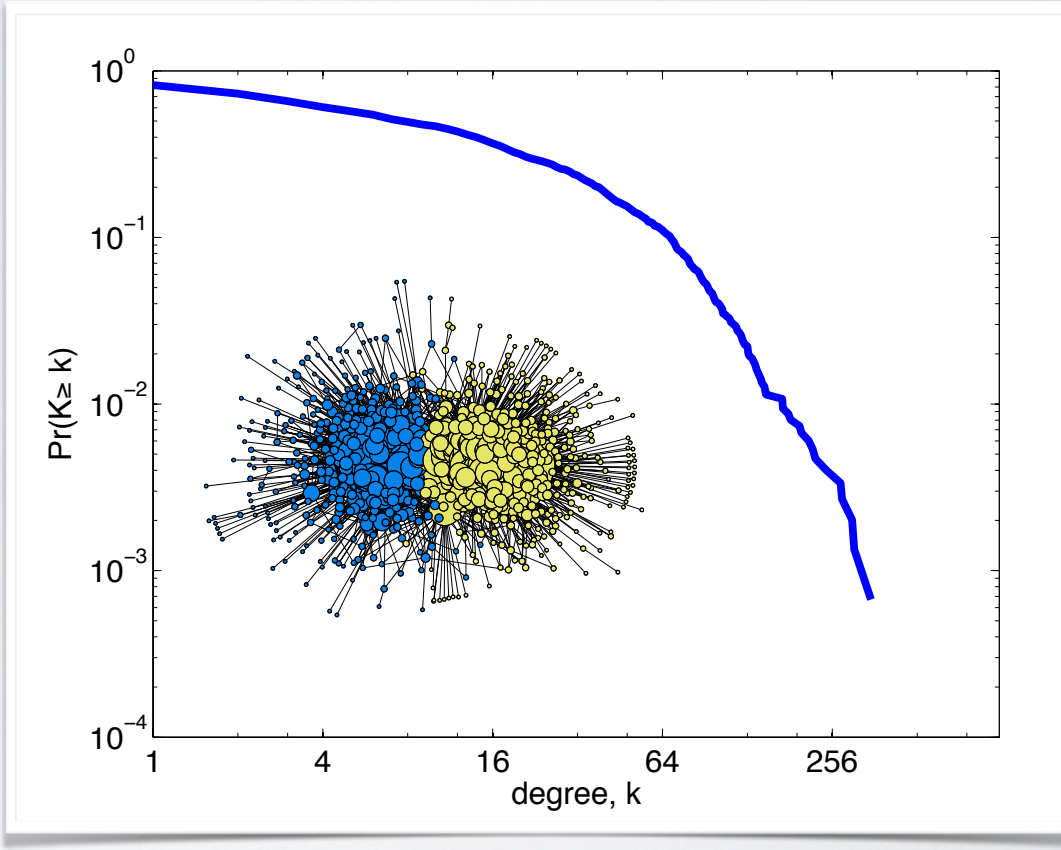
- fraction with degree $K \geq k$
- monotonic
- smoother than PDF
- better reveals curvature

"loglog" plots:

- good for high variance quantities

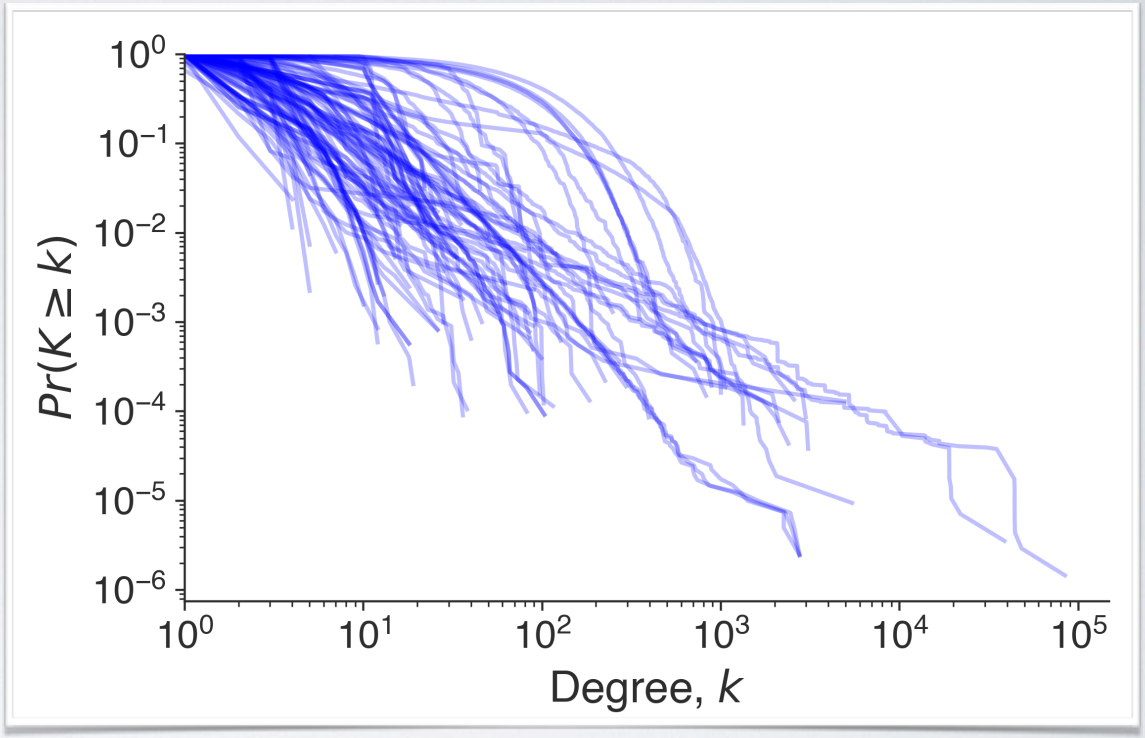
Lorenz curves & Gini:

- captures level of inequality



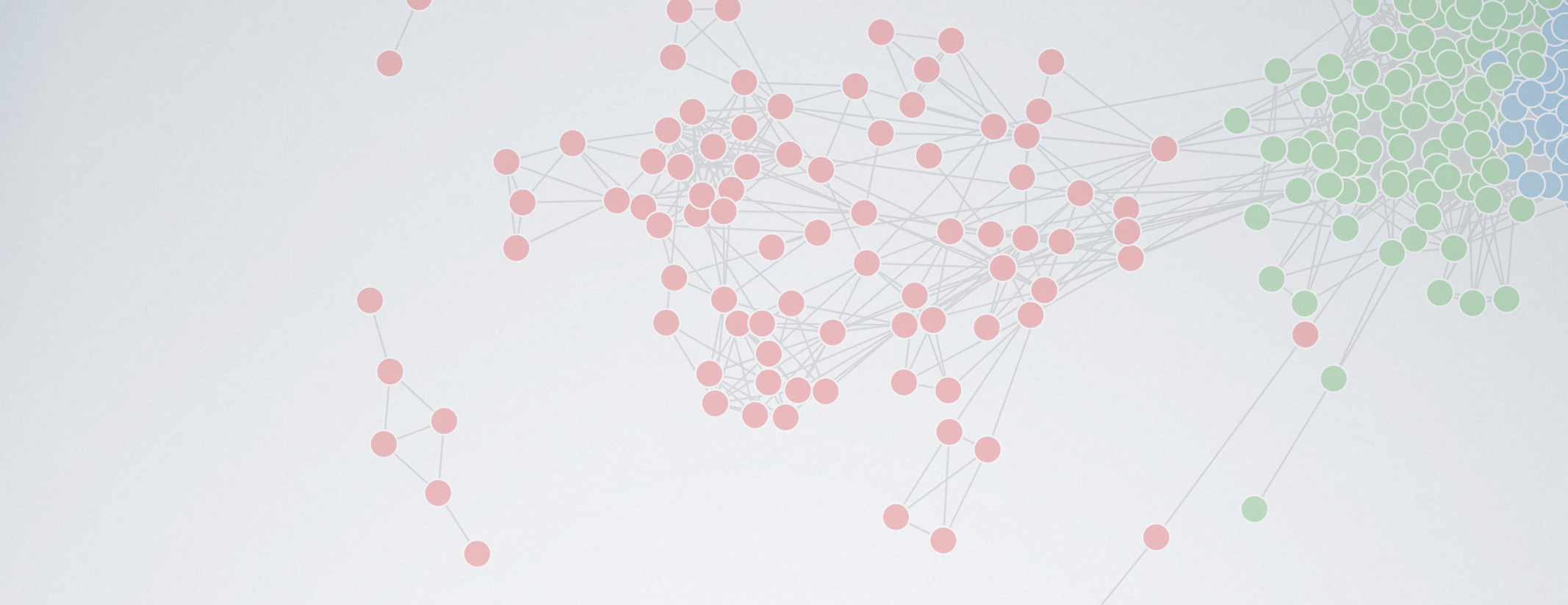
# exploring degree distributions

- nearly all real networks exhibit a ***heavy-tailed degree distribution***
- ***very few*** networks exhibit perfect power-law degree distributions
- ***some*** distributions exhibit power-law tails
- power laws are cool!
  but knowing one from garbage ***requires statistics*** *




- ***does the specific distributional form matter?***
  think carefully about whether it does (it may not).

* data from 100 networks from 4 scientific domains
* Clauset et al. SIAM Review 51, 661–703 (2009)



**end of lecture 1**

**lecture 2 : describing network structure**

**lecture 3 : null models & inference for networks**

# directed networks

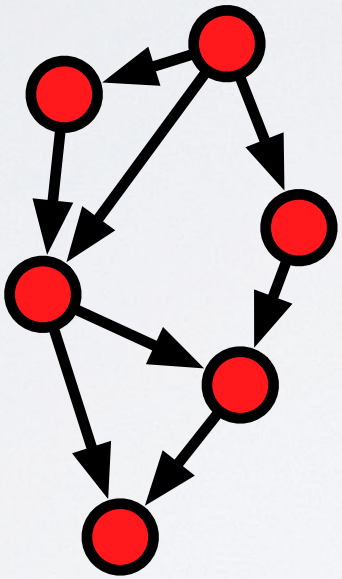
$$A_{ij} \neq A_{ji}$$

citation networks

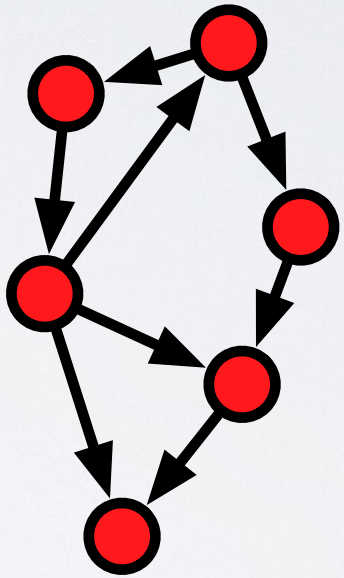
foodwebs*

epidemiological

others?



directed acyclic graph



directed graph

WWW

friendship?

flows of goods, information

economic exchange

dominance

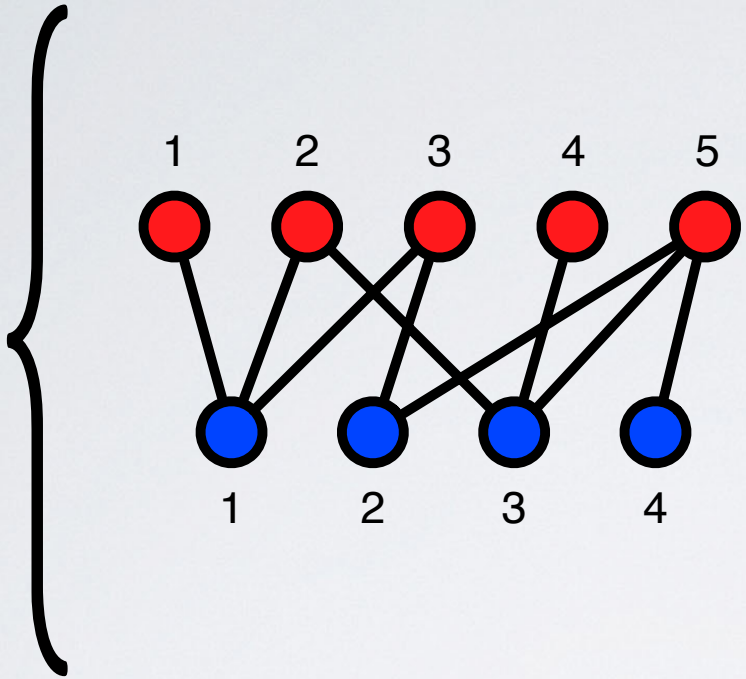
neuronal

transcription

time travelers

# bipartite networks

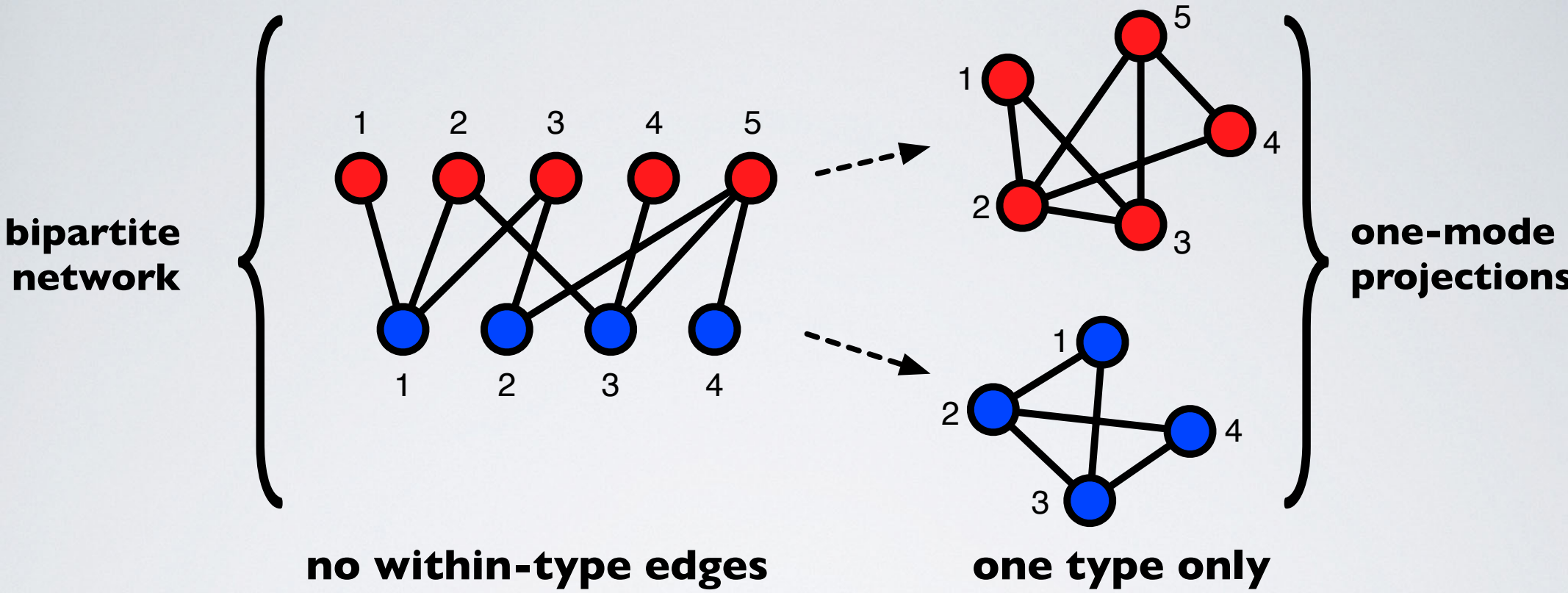
**bipartite network**




**no within-type edges**

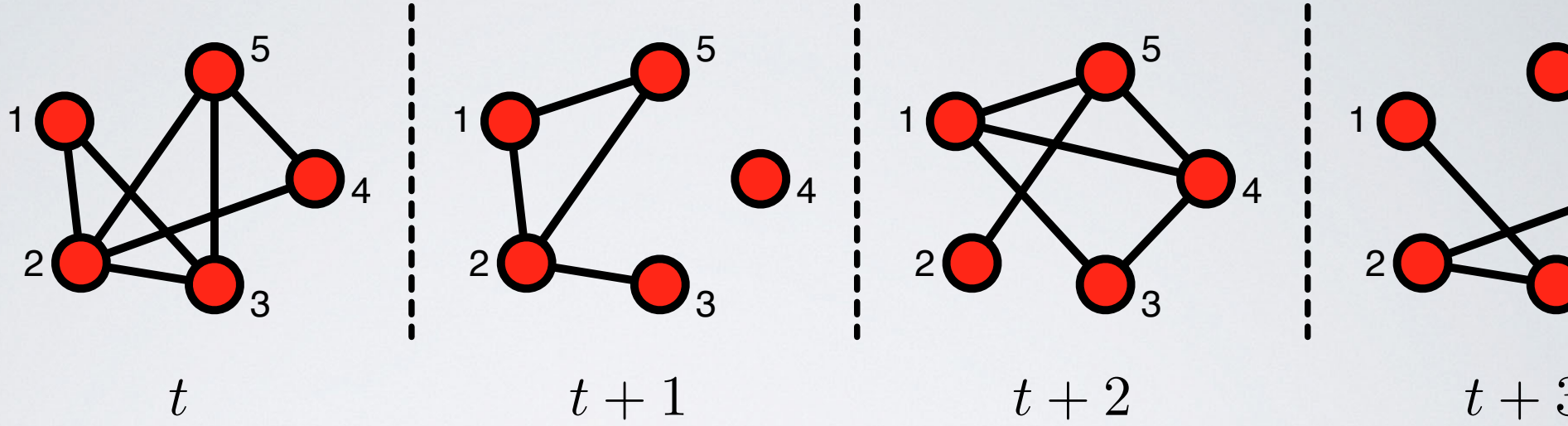
| | |
|---|---|
| authors & papers | people & locations (checkins) |
| actors & movies/scenes | metabolites & reactions |
| musicians & albums | genes & substrings |
| people & online groups | words & documents |
| people & corporate boards | plants & pollinators |

# bipartite networks




| | |
|---|---|
| authors & papers | people & locations (checkins) |
| actors & movies/scenes | metabolites & reactions |
| musicians & albums | genes & substrings |
| people & online groups | words & documents |
| people & corporate boards | plants & pollinators |

# temporal networks




**any network over time**

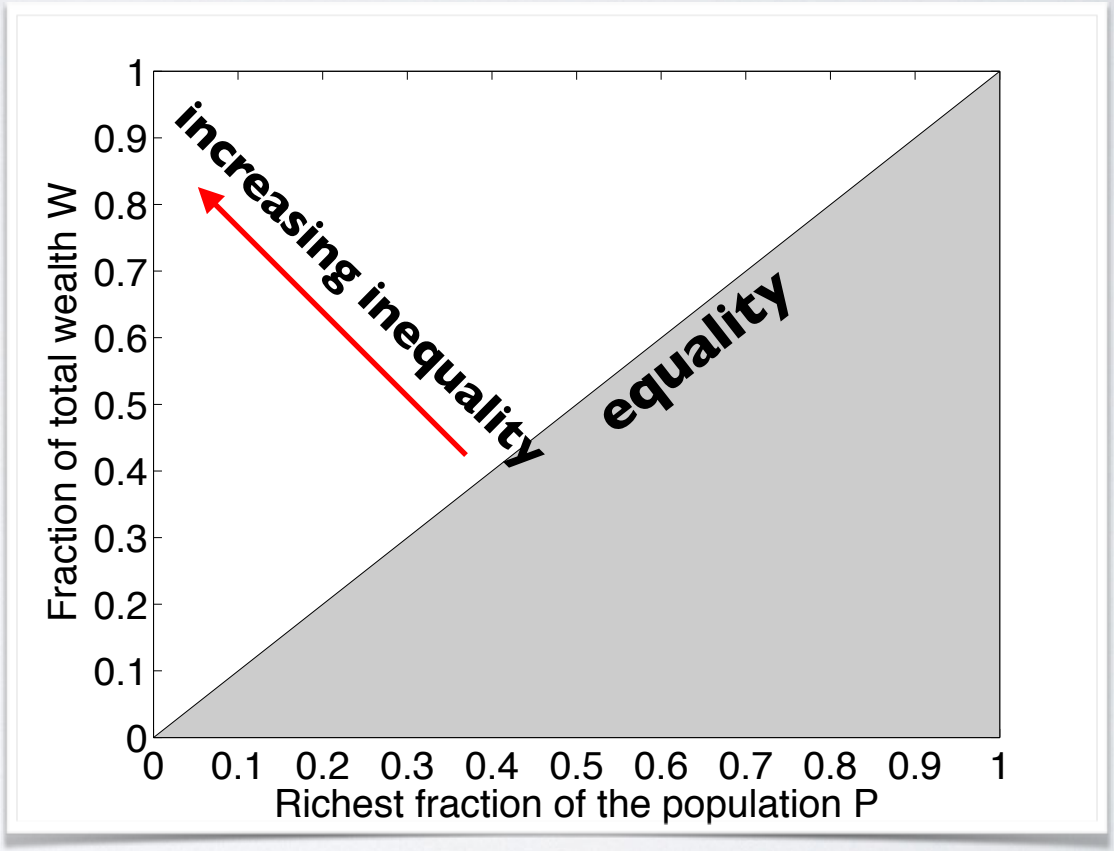
discrete time (snapshots), edges $(i, j, t)$

continuous time, edges $(i, j, t_s, \Delta t)$

# degree distributions

## degree "wealth"

what fraction of total wealth *W* is owned by richest fraction *P*




**Lorenz curve**

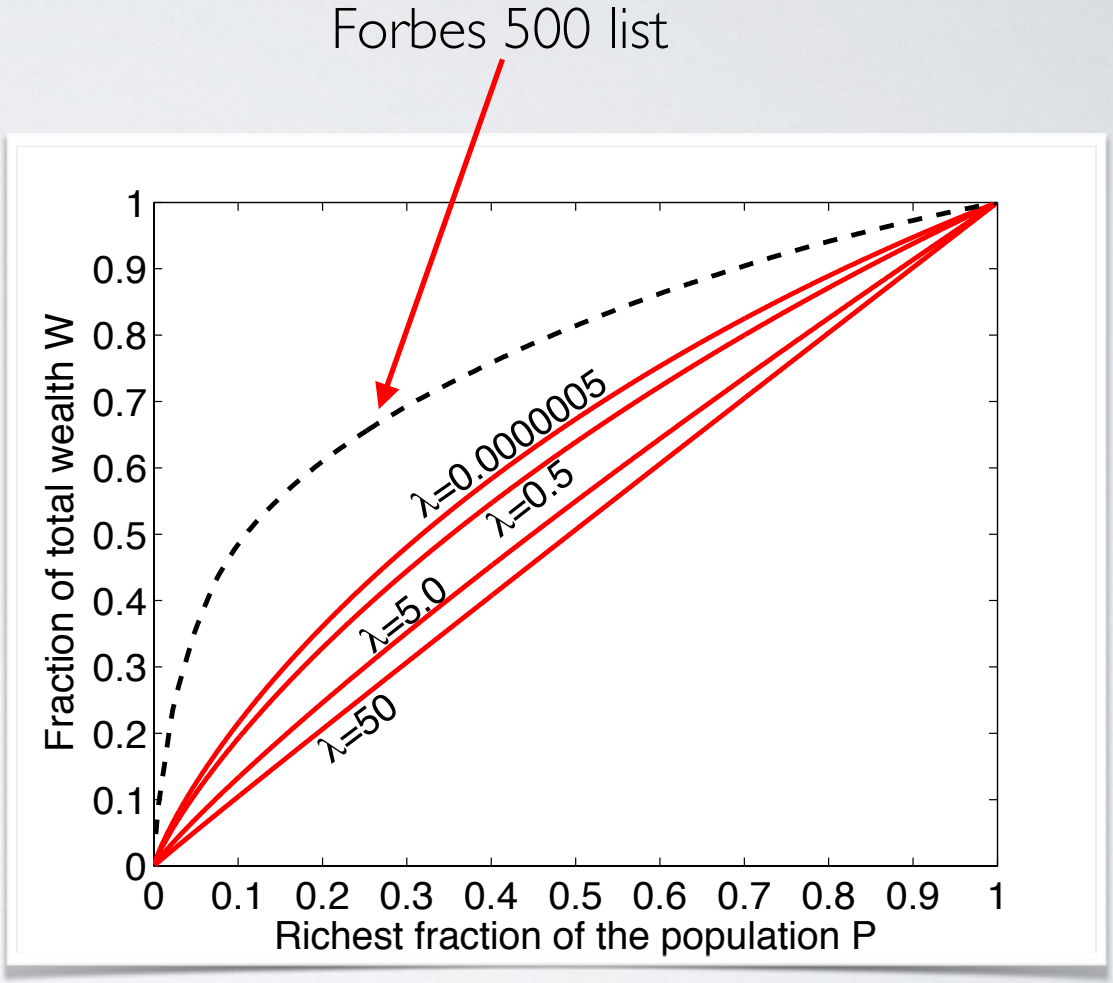
# degree distributions

## degree "wealth"

what fraction of total wealth *W* is owned by richest fraction *P*

$$\Pr(k) \propto \mathrm{e}^{-\lambda k}$$

exponential distribution




**Lorenz curve**

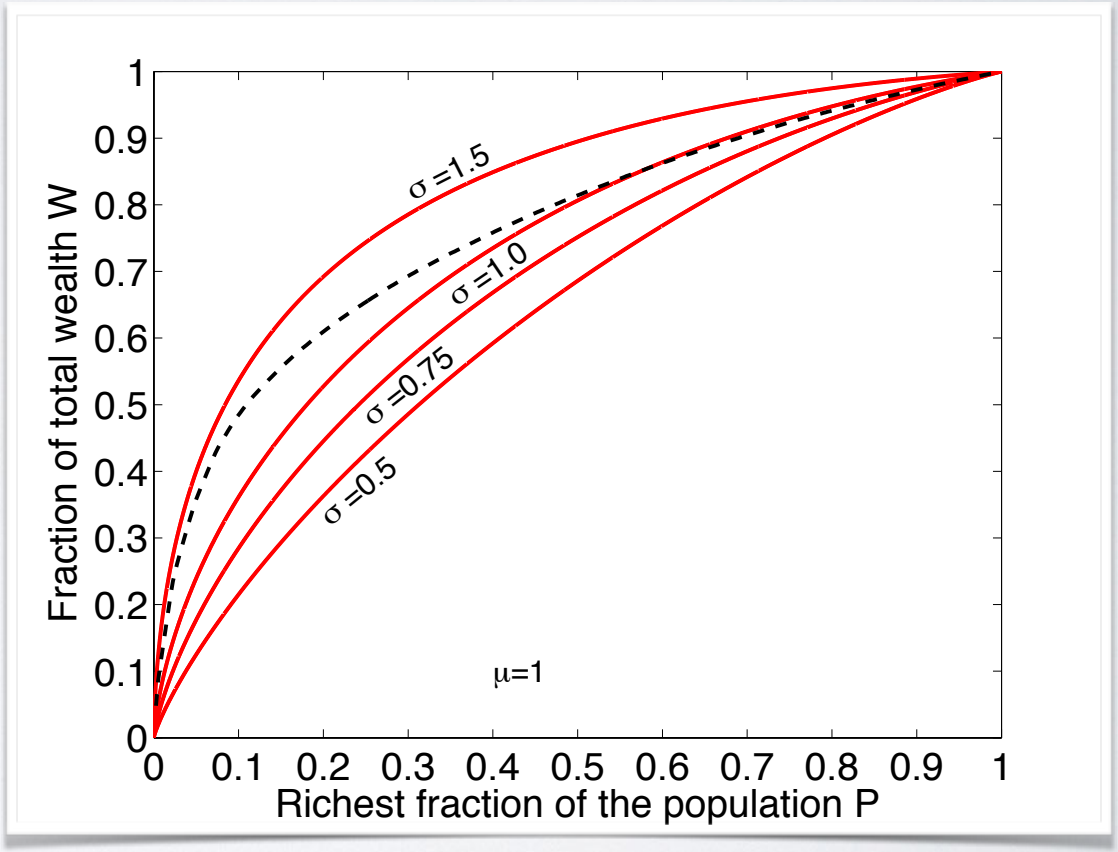
# degree distributions

## degree "wealth"

what fraction of total wealth *W* is owned by richest fraction *P*

$$\Pr(k) \propto \frac{1}{k} e^{-\left(\frac{\ln k - \mu}{\sigma\sqrt{2}}\right)^2}$$

log-normal distribution




**Lorenz curve**

# degree distributions

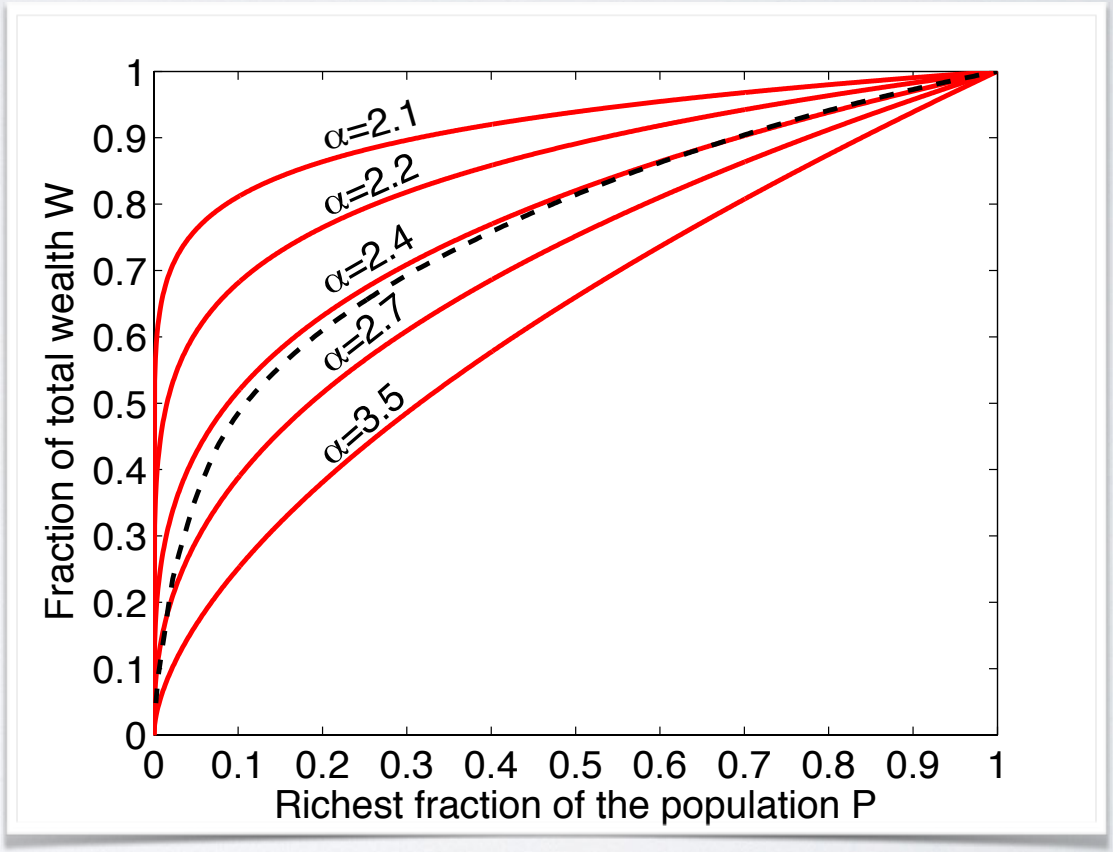
## degree "wealth"

what fraction of total wealth *W* is owned by richest fraction *P*

$$\Pr(k) \propto k^{-\alpha}$$

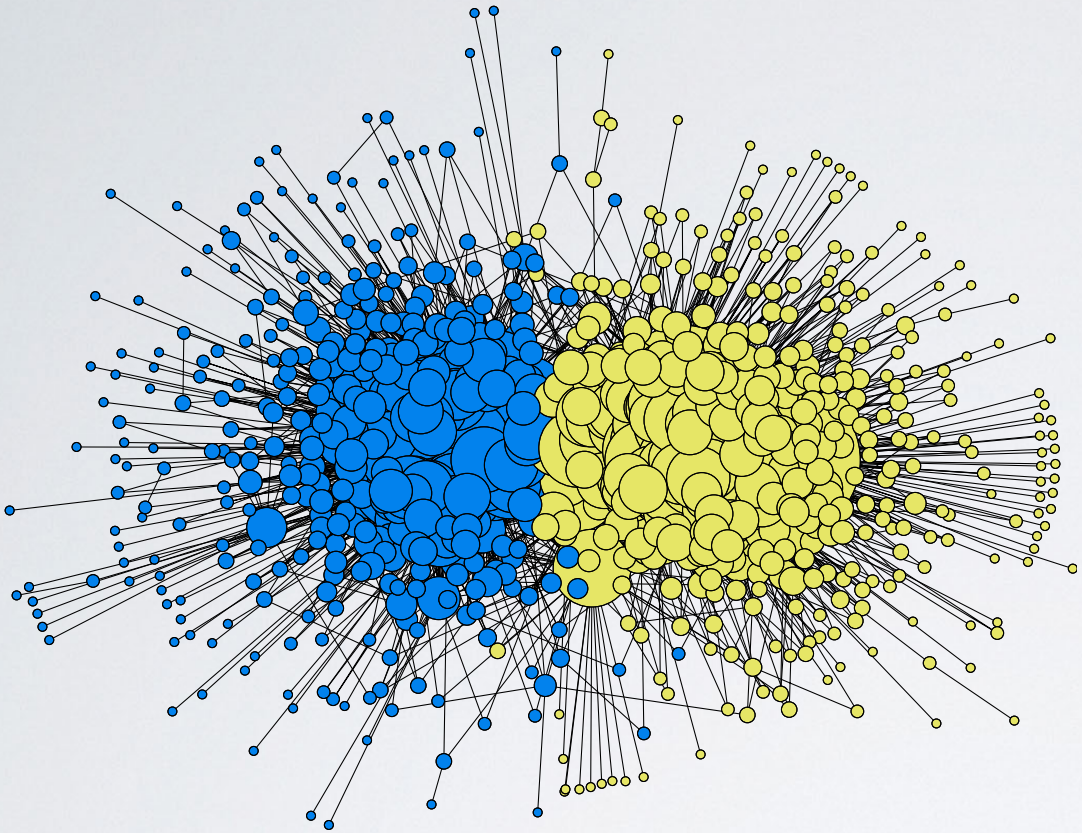
power-law distribution

**80/20 rule**

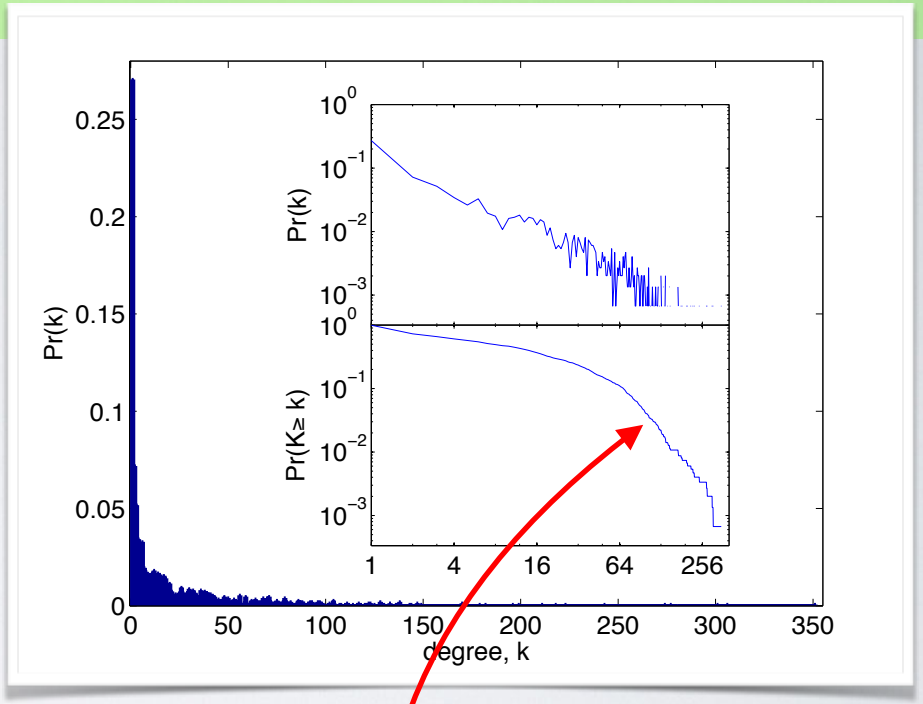



**Lorenz curve**

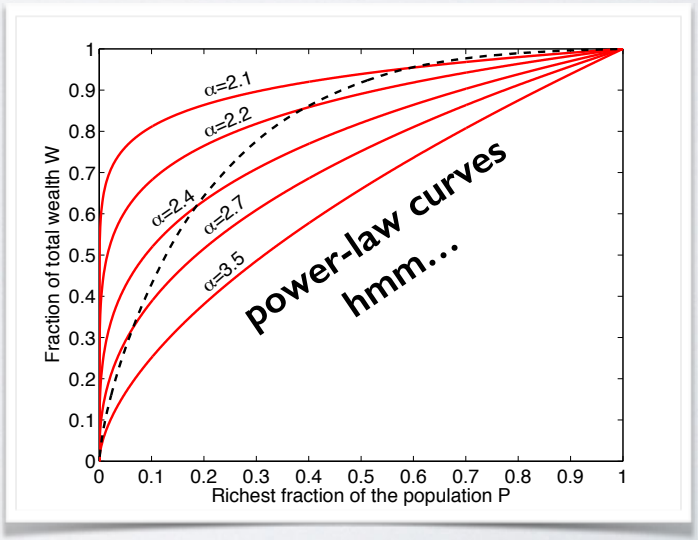
# degree distributions



political blogs*




**is this a power law?**




* Adamic and Glance, WWW Workshop on the Weblogging Ecosystem (2005).

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- let's do some math
- (a nice warm up for other things, later)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- normalization (probability density function)

$$1 = \int_{k_{\min}}^{\infty} \Pr(k)\mathrm{d}k \longrightarrow$$ **pdf**

- complementary cumulative distribution function

$$P(k) = \int_{k}^{\infty} \Pr(y)\mathrm{d}y \longrightarrow$$ **ccdf**

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- normalization (probability density function)*

$$1 = \int_{k_{\min}}^{\infty} \Pr(k)\mathrm{d}k \longrightarrow \Pr(k) = \frac{\alpha - 1}{k_{\min}} \left( \frac{k}{k_{\min}} \right)^{-\alpha}$$ **pdf**

- complementary cumulative distribution function

$$P(k) = \int_{k}^{\infty} \Pr(y)\mathrm{d}y \longrightarrow P(k) = \left( \frac{k}{k_{\min}} \right)^{-\alpha+1}$$ **ccdf**

- power laws have unusual properties, imply unusual underlying mechanisms

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

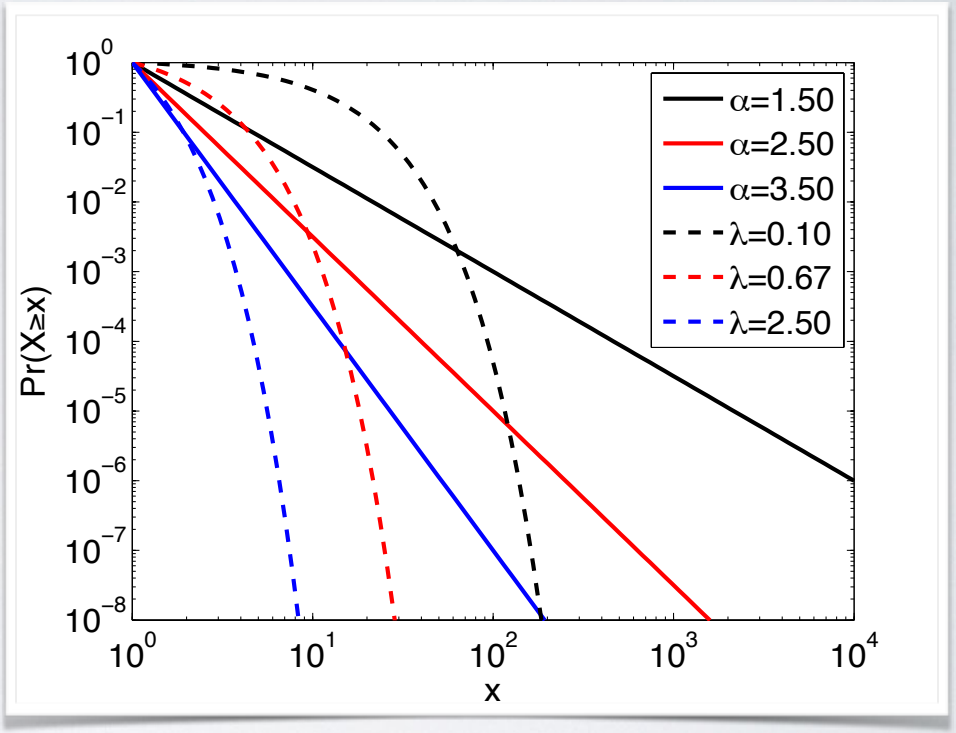
* the math here is easier for the continuous variables, but qualitatively similar results hold for discrete variables. also, yes, vertex degree is discrete not continuous.

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- high-variance

$$\langle k^m \rangle = \int_{k_{\min}}^{\infty} k^m \Pr(k) \mathrm{d}k$$




Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

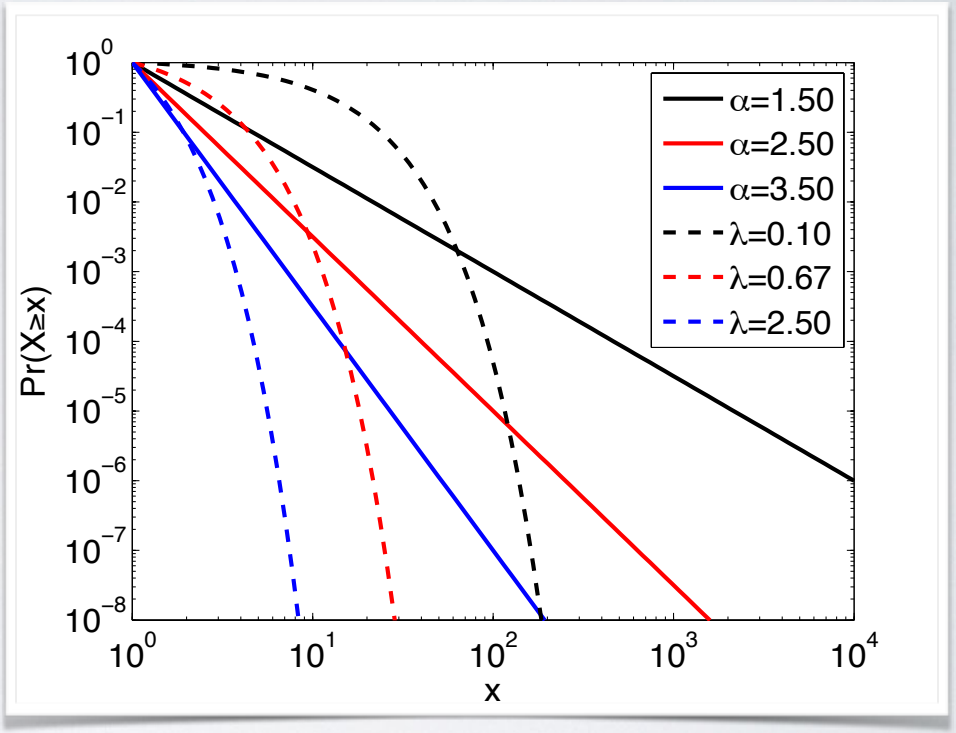
# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- high-variance

$$\langle k^m \rangle = \int_{k_{\min}}^{\infty} k^m \Pr(k) \mathrm{d}k$$
$$= k_{\min}^m \left( \frac{\alpha - 1}{\alpha - 1 - m} \right)$$

- ***infinite mean*** $1 < \alpha < 2$
- ***infinite variance*** $2 < \alpha < 3$
- much, much heavier tails than exponential, normal, etc.
- heavier than log-normal (asymptotically)




Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- "scale invariance" (aka "scale free")

$$\Pr(c\,k) = (\alpha - 1)k_{\min}^{\alpha-1}(c\,k)^{-\alpha}$$

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- "scale invariance" (aka "scale free")

$$\begin{aligned}\Pr(c\,k) &= (\alpha-1)k_{\min}^{\alpha-1}(c\,k)^{-\alpha} \\ &= c^{-\alpha}\left[(\alpha-1)k_{\min}^{\alpha-1}k^{-\alpha}\right] \\ &\propto \Pr(k)\end{aligned}$$

- power law is *only distribution* with this property

- implies no natural "scale" of distribution
- implies signature form: straight line on log-log plot

$$\ln \Pr(k) = \ln C - \alpha \ln k$$

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- exotic mechanisms
  - preferential attachment [Yule 1925, Simon 1955, Price 1976, etc.]
  - combinations of exponentials [Miller 1957, Reed & Hughes 2002]
  - phase transitions [many]
  - self-organized criticality (SOC) [Bak et al. 1988]
  - highly optimized tolerance (HOT) [Carlson and Doyle, 1999]
  - fragmentation [many]
  - multiplicative random walks (with lower limit) [many]
  - many, many others

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- how do you know? statistics.
- estimating $\alpha$ from data $\{k_i\}$ via maximum likelihood

$$\ln \mathcal{L}(\{k_i\} \,|\, \theta) = \ln \prod_{i=1}^{n} \Pr(k_i \,|\, \theta)$$

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- how do you know? statistics.
- estimating $\alpha$ from data $\{k_i\}$ via maximum likelihood

$$\ln \mathcal{L}(\{k_i\} \,|\, \theta) = \ln \prod_{i=1}^{n} \Pr(k_i \,|\, \theta)$$

- for the power-law distribution (log-likelihood)

$$\ln \mathcal{L}(\{k_i\} \,|\, \alpha, k_{\min}) = n \ln \left( \frac{\alpha - 1}{k_{\min}} \right) - \alpha \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\min}} \right)$$

- solving $\partial \mathcal{L} / \partial \alpha = 0$ , yields MLE

$$\hat{\alpha} = 1 + n \Big/ \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\min}} \right)$$

with standard error

$$\hat{\sigma} = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n)$$

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- how do you know? statistics.
- estimating $\alpha$ from data $\{k_i\}$ via maximum likelihood

$$\ln \mathcal{L}(\{k_i\} \,|\, \theta) = \ln \prod_{i=1}^{n} \Pr(k_i \,|\, \theta)$$

- for the power-law distribution (log-likelihood)

$$\ln \mathcal{L}(\{k_i\} \,|\, \alpha, k_{\min}) = n \ln \left( \frac{\alpha - 1}{k_{\min}} \right) - \alpha \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\min}} \right)$$

- solving $\partial \mathcal{L} / \partial \alpha = 0$ , yields MLE

$$\hat{\alpha} = 1 + n \Big/ \sum_{i=1}^{n} \ln \left( \frac{k_i}{k_{\min}} \right)$$

with standard error

$$\hat{\sigma} = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n)$$

umm… we don't know this value

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

# power-law distributions

$$\Pr(k) = C\,k^{-\alpha} \quad \text{for} \quad k \geq k_{\min}$$

- we can choose $k_{\min}$ smartly [see *SIAM Review* **51**; code is here *]
- but how do we know if the model is good? fitting is *easy*

**moral: always check your model's goodness-of-fit**

- ways to do this:
  1. compute a *p*-value relative to a *reasonable* null model
  2. compare your model against *reasonable* alternatives
  3. compare synthetic data drawn from your model with your empirical data
  4. use your model to predict something *reasonable*

Newman, Contemporary Physics 46, 323–351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)

* http://santafe.edu/~aaronc/powerlaws/