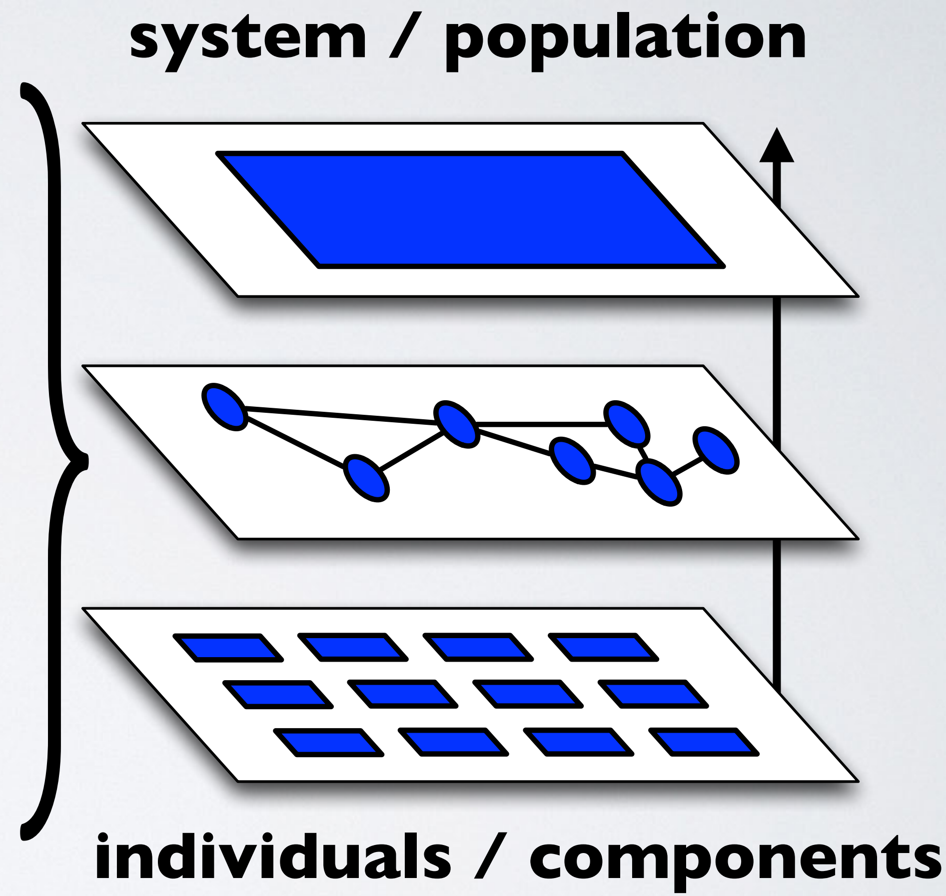
Introduction to Networks

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

what are networks?

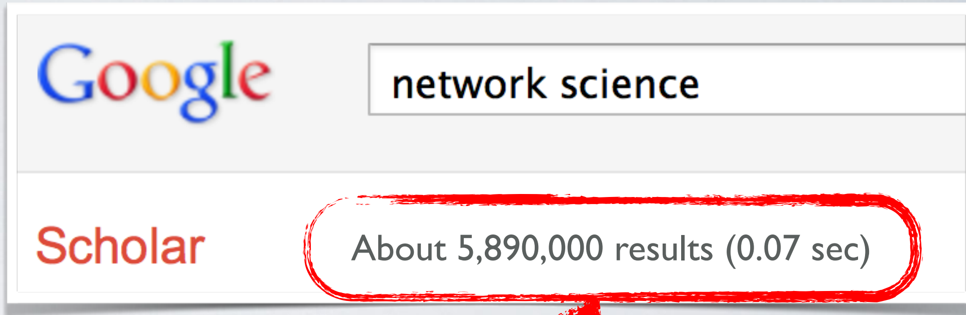
what are networks?

- an approach
- a mathematical representation
- provide structure to complexity
- *structure above* individuals / components
- *structure below* system / population



this lecture

- build intuition
- highlight a few concepts & questions
- provide some examples
- pointers to further study
- not a substitute for technical coursework



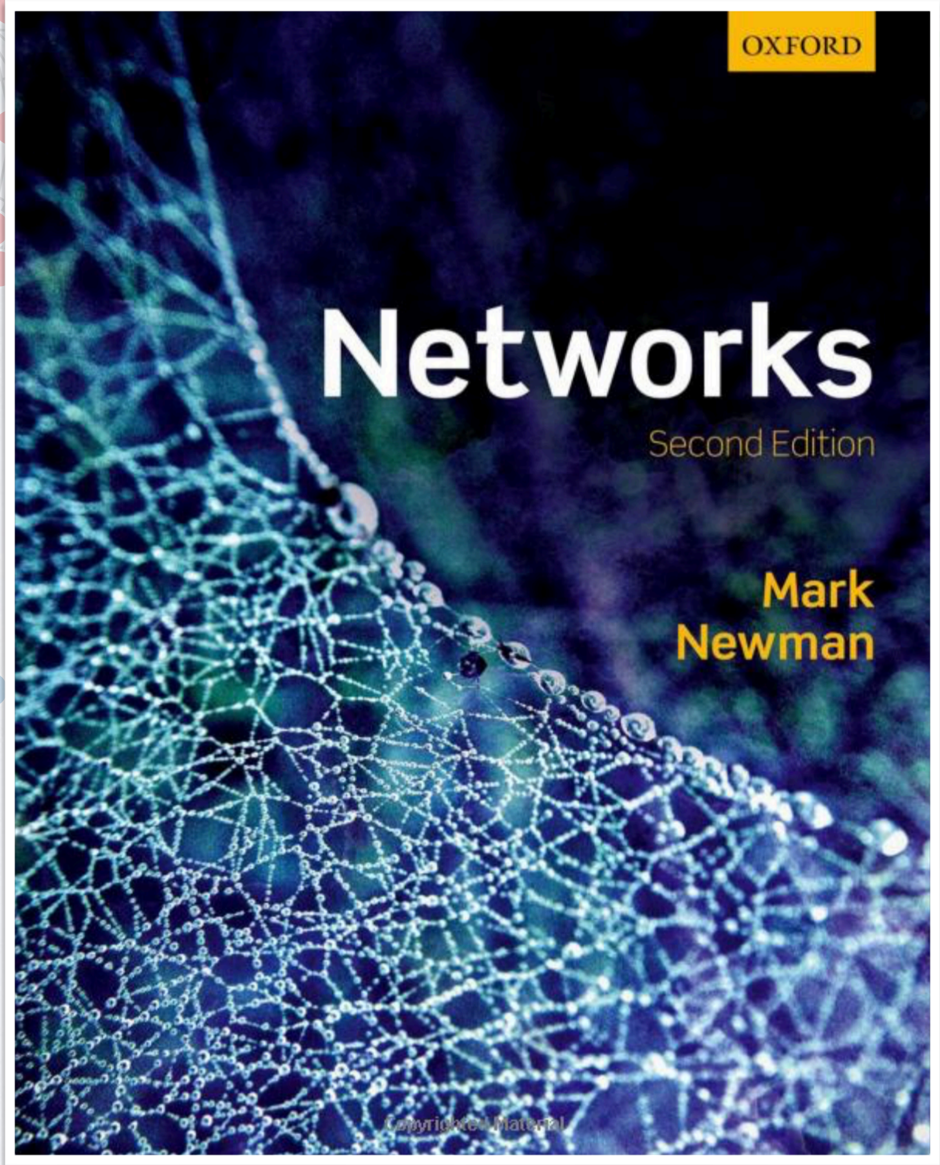
it's a big field now

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)

https://aaronclauset.github.io/courses/5352/

University of Colorado **Boulder**

Biological Networks

Instructor: Aaron Clauset

This undergraduate-level course examines the computational representation and analysis of biological phenomena through the structure and dynamics of networks, from molecules to species. Attention focuses on algorithms for clustering network structures, predicting missing information, modeling flows, regulation, and spreading-process dynamics, examining the evolution of network structure, and developing intuition for how network structure and dynamics relate to biological phenomena.

Full lectures notes online (~150 pages in PDF)

https://aaronclauset.github.io/courses/3352/

Software

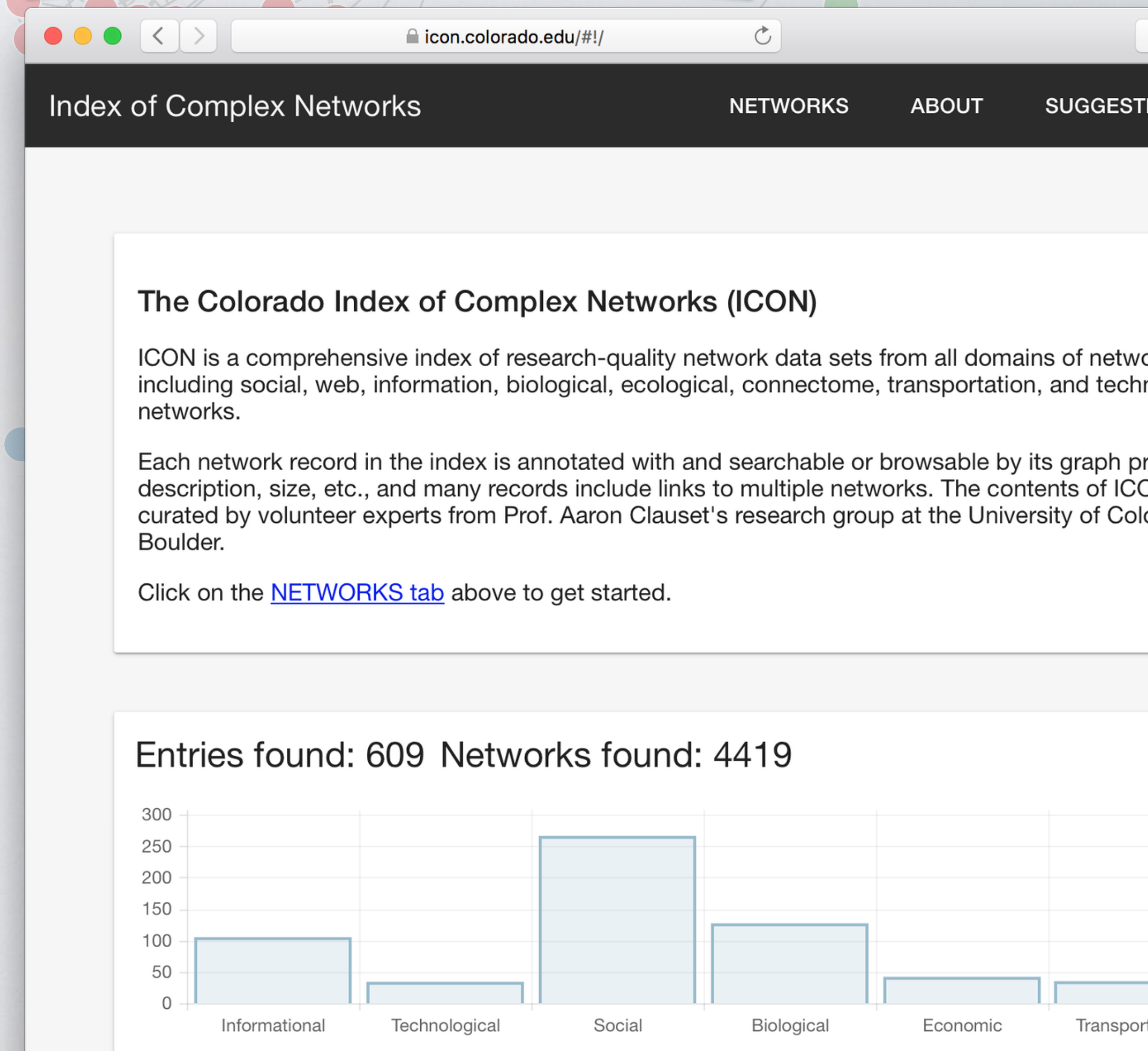
[R](#)
[Python](#)
Matlab
[NetworkX](#) [python]
[graph-tool](#) [python, c++]
[GraphLab](#) [python, c++]

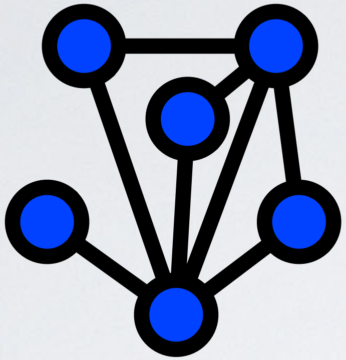
Standalone editors

[UCI-Net](#)
[NodeXL](#)
[Gephi](#)
[Pajek](#)
[Network Workbench](#)
[Cytoscape](#)
[yEd graph editor](#)
[Graphviz](#)

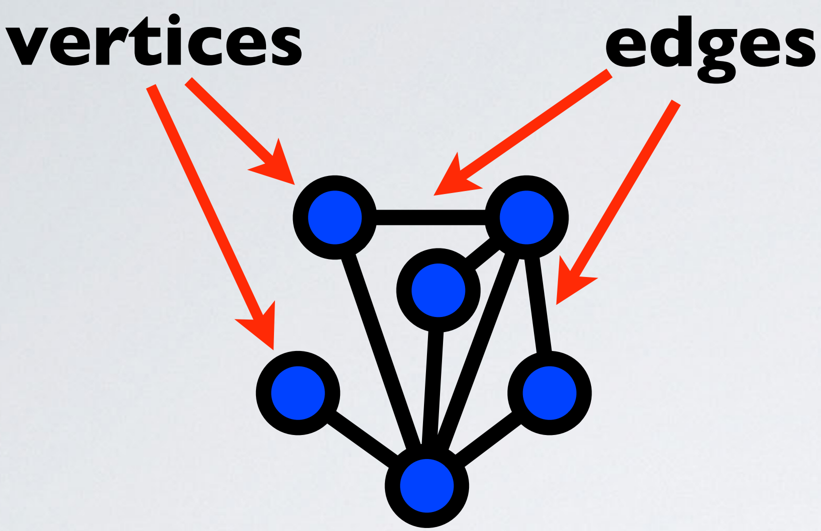
Network data sets

[Colorado Index of Complex Networks](#)





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



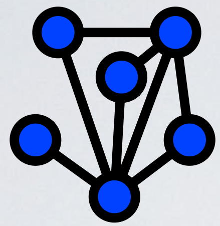
what is a vertex?

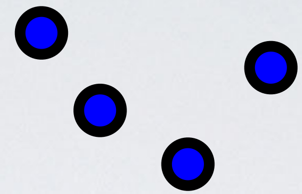
V distinct objects (vertices / nodes / actors)

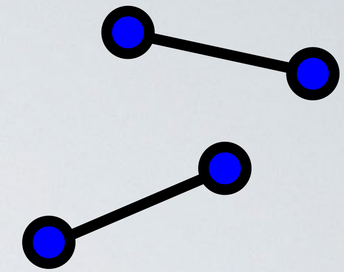
when are two vertices connected?

$$E \subseteq V \times V$$

pairwise relations (edges / links / ties)





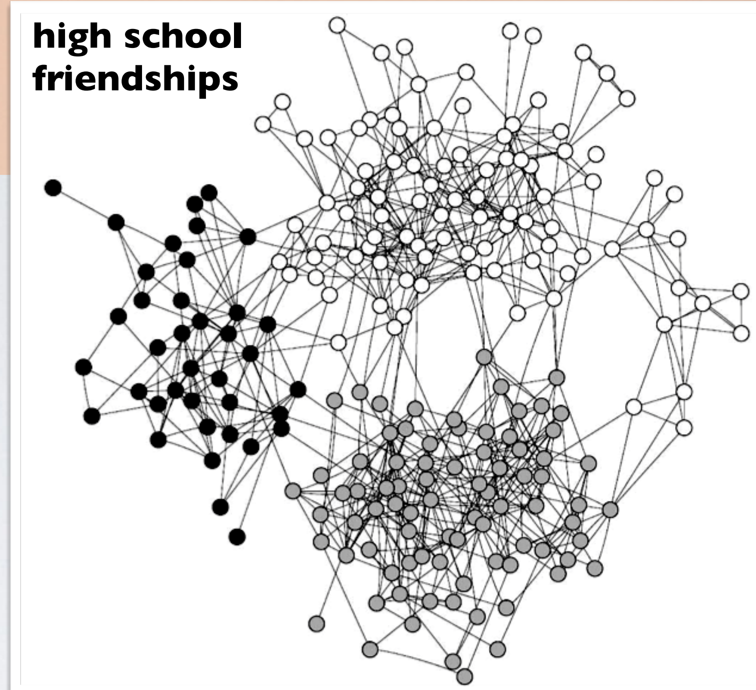


	network	vertex	edge
informationaltelecommunications	Internet(1)	computer	IP network adjacency
	Internet(2)	autonomous system (ISP)	BGP connection
	software	function	function call
	World Wide Web	web page	hyperlink
	documents	article, patent, or legal case	citation
transportation	power grid transmission	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
	protein-interaction network	protein	bonding
	gene regulatory network	gene	regulatory effect
	neuronal network	neuron	synapse
	food web	species	predation or resource transfer

social networks

vertex: a person

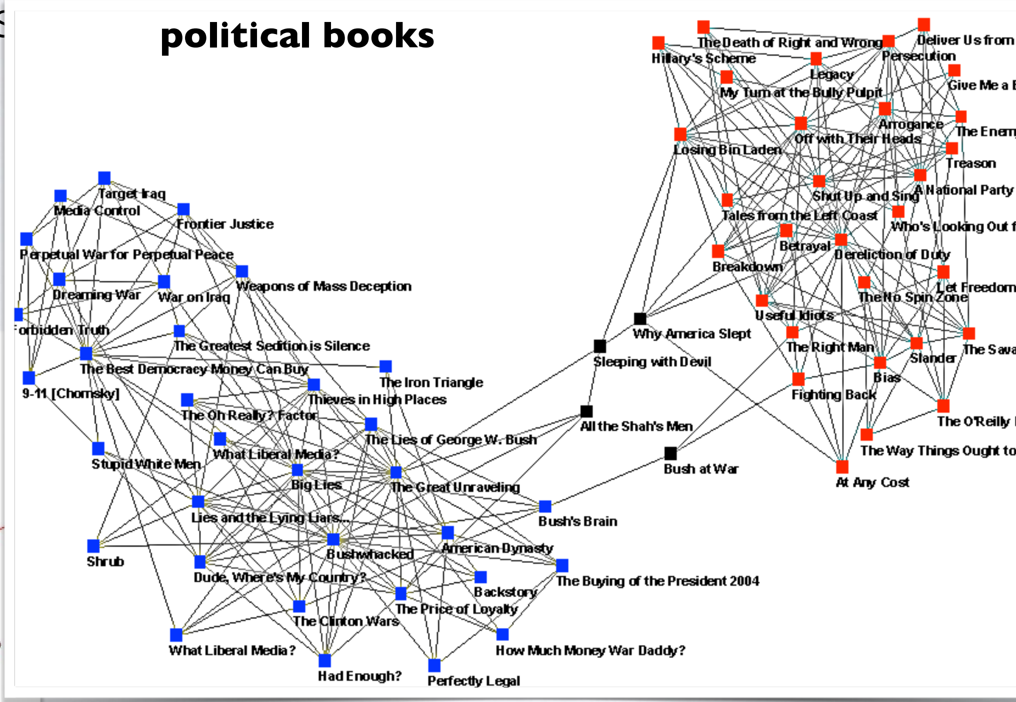
edge: friendship, collaborations, sexual contacts, communication, authority, exchange, etc.

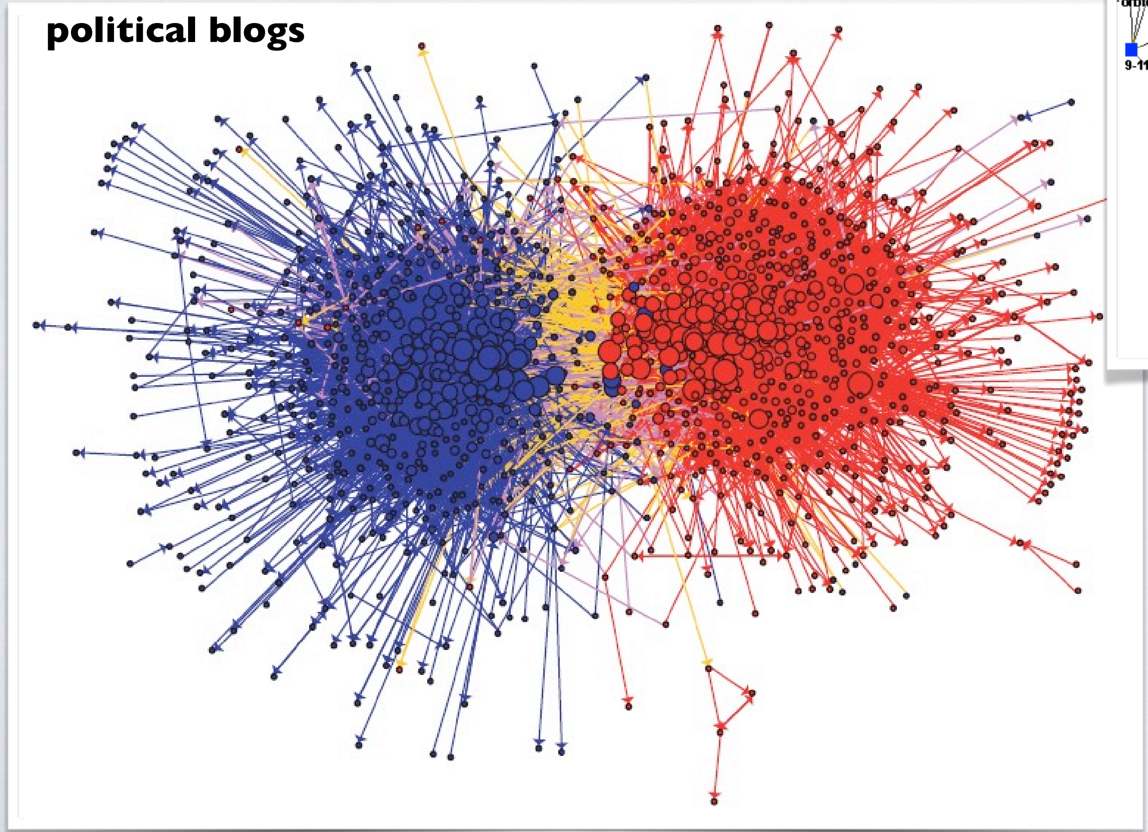


information networks

vertex: books, blogs, webpages, etc.

edge: citations, hyperlinks, recommendations, similarity, etc.

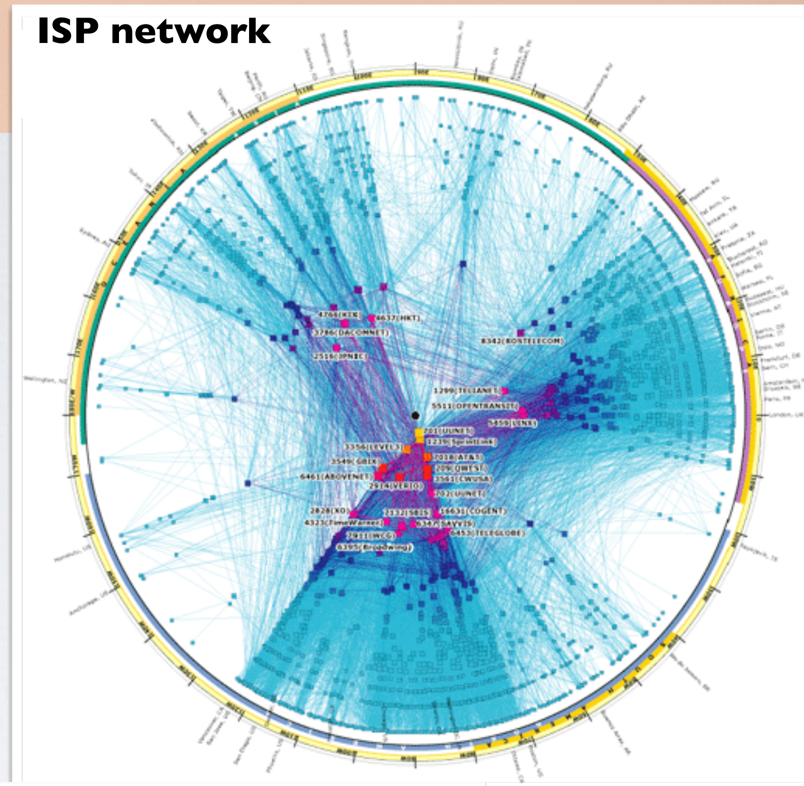


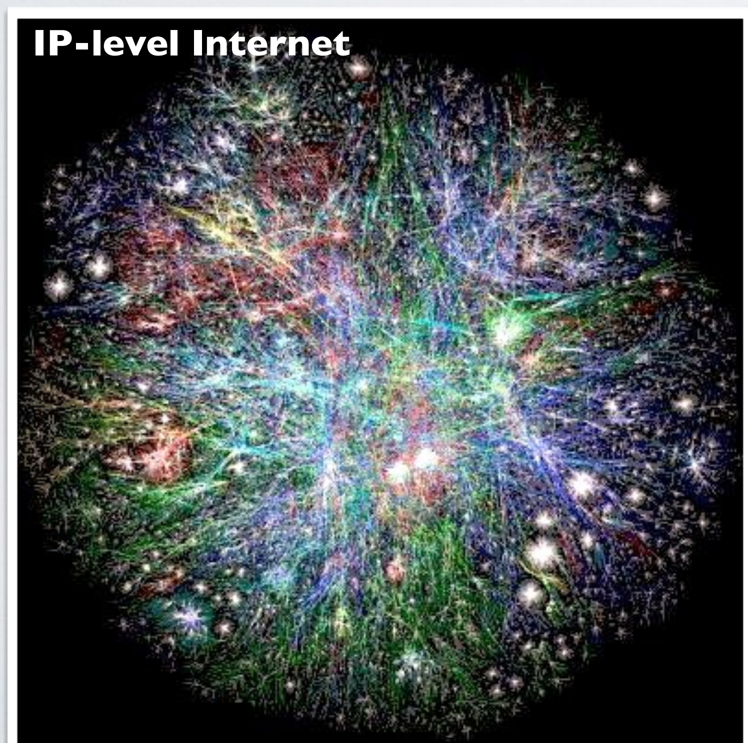


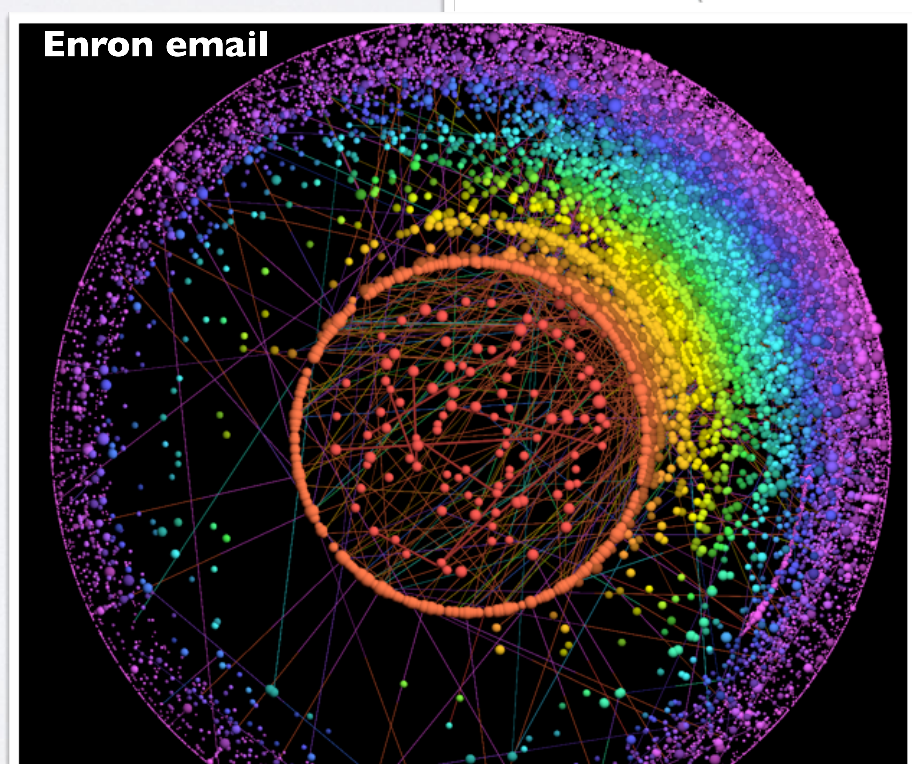
communication networks

vertex: network router, ISP, email address, mobile phone number, etc.

edge: exchange of information





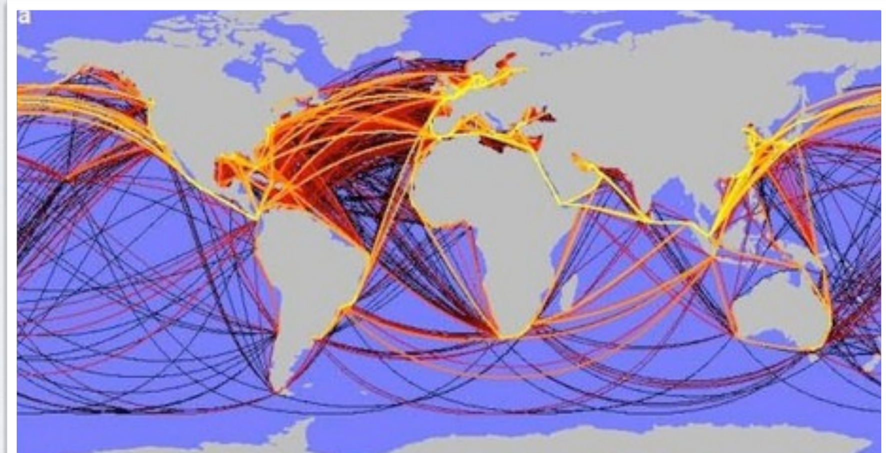


transportation networks

vertex: city, airport, junction, railway station, river confluence, etc.

edge: physical transportation of material

global shipping



global air traffic

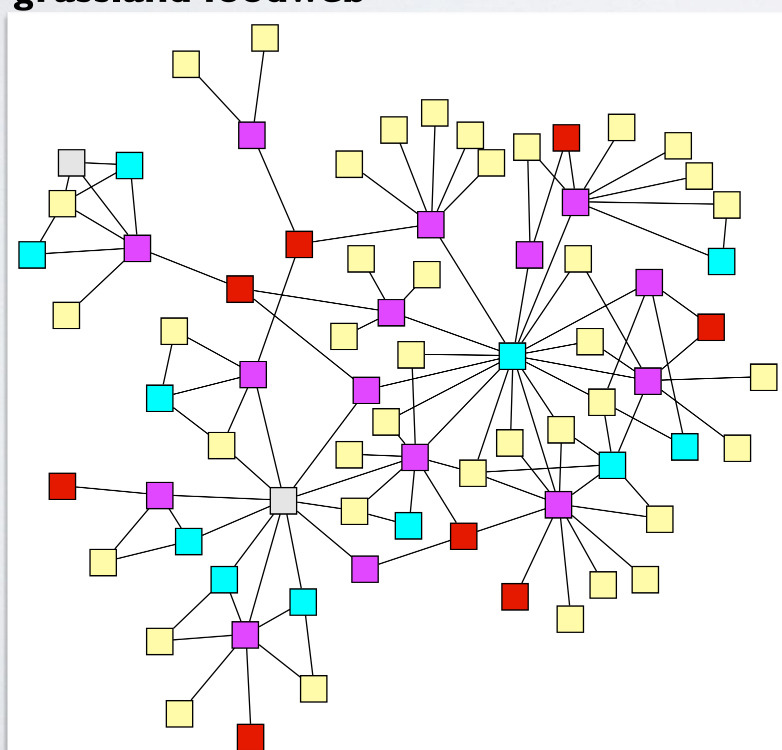
US Interstates

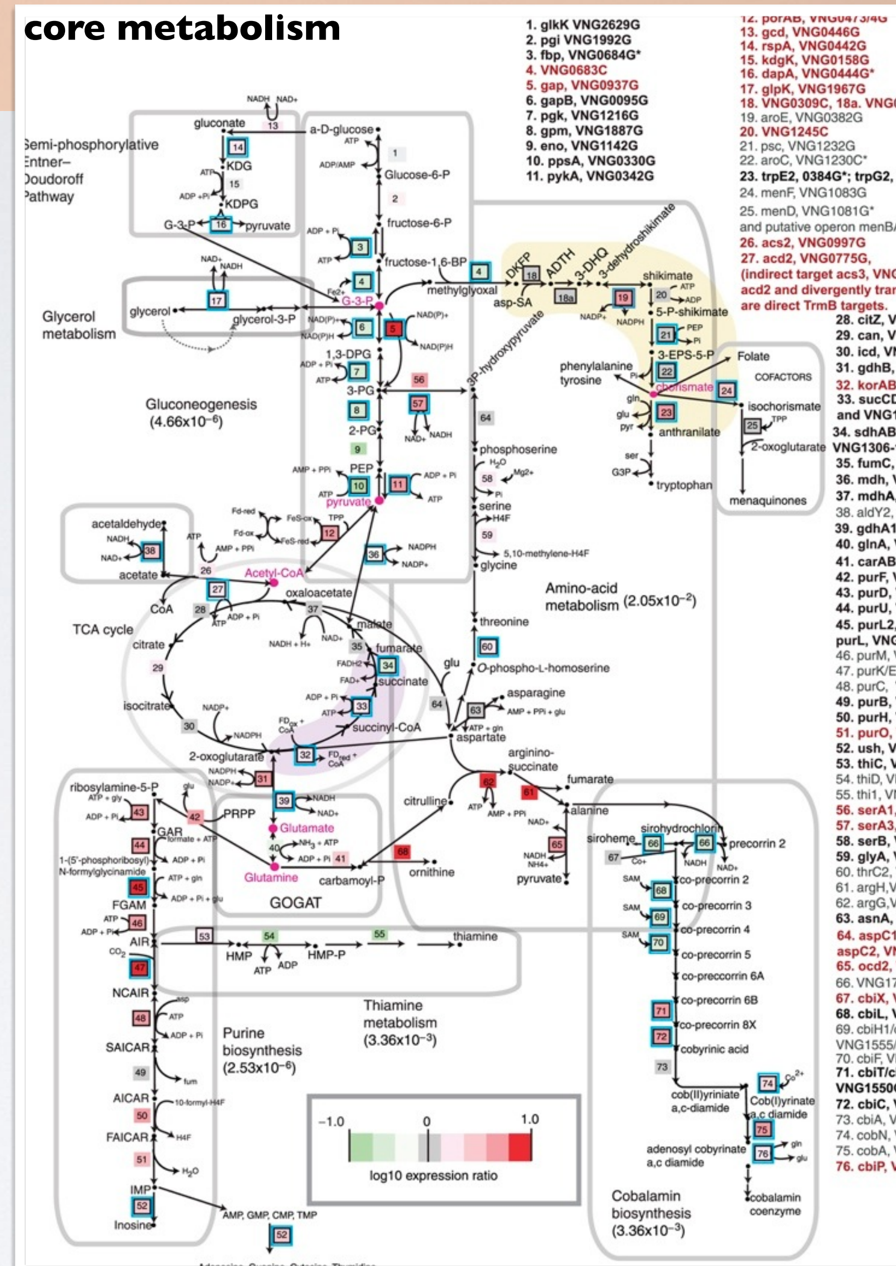
biological networks

vertex: species, metabolic, protein, gene, neuron, etc.

edge: predation, chemical reaction, binding, regulation, activation, etc.

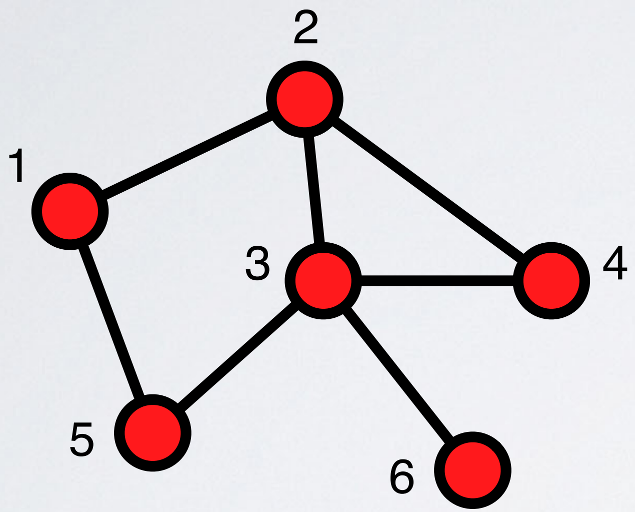
grassland foodweb





representing networks

a *simple* network

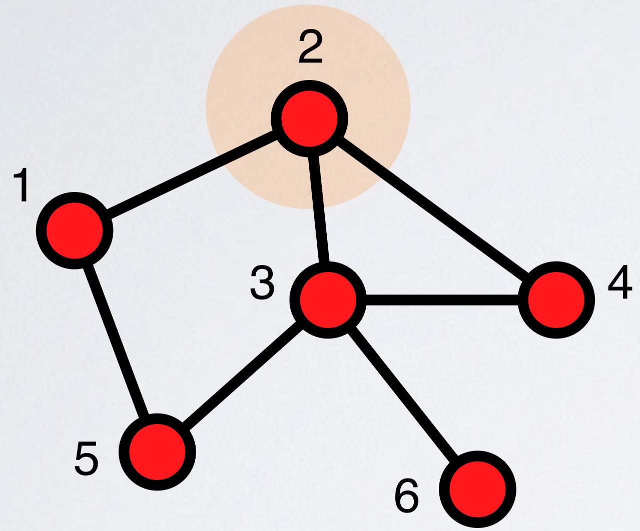


undirected

unweighted

no self-loops

a *simple* network



undirected

unweighted

no self-loops

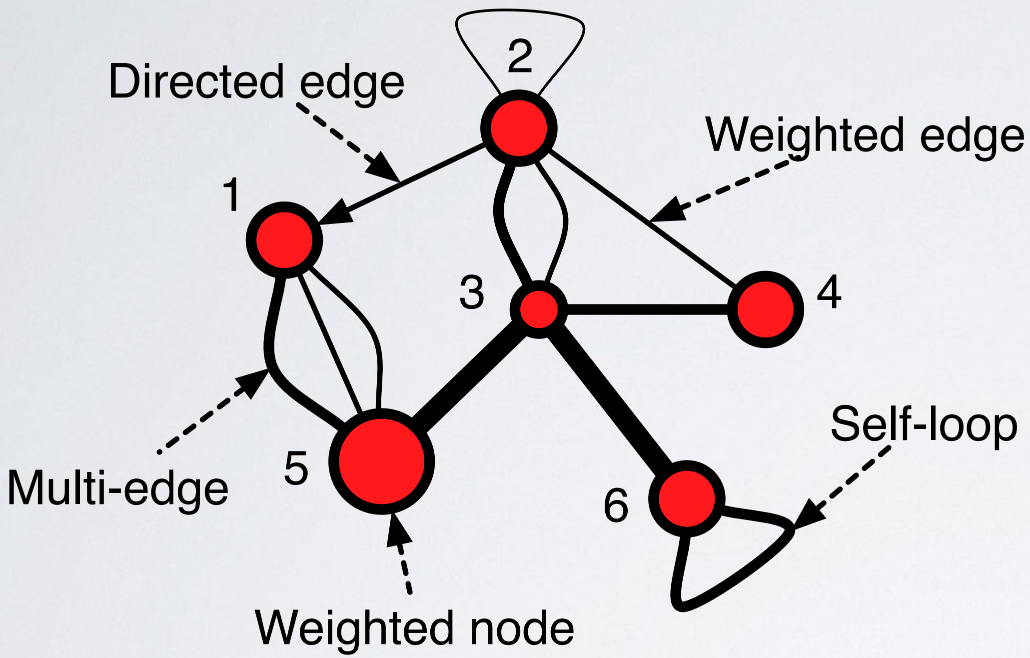
adjacency matrix

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

a less simple network

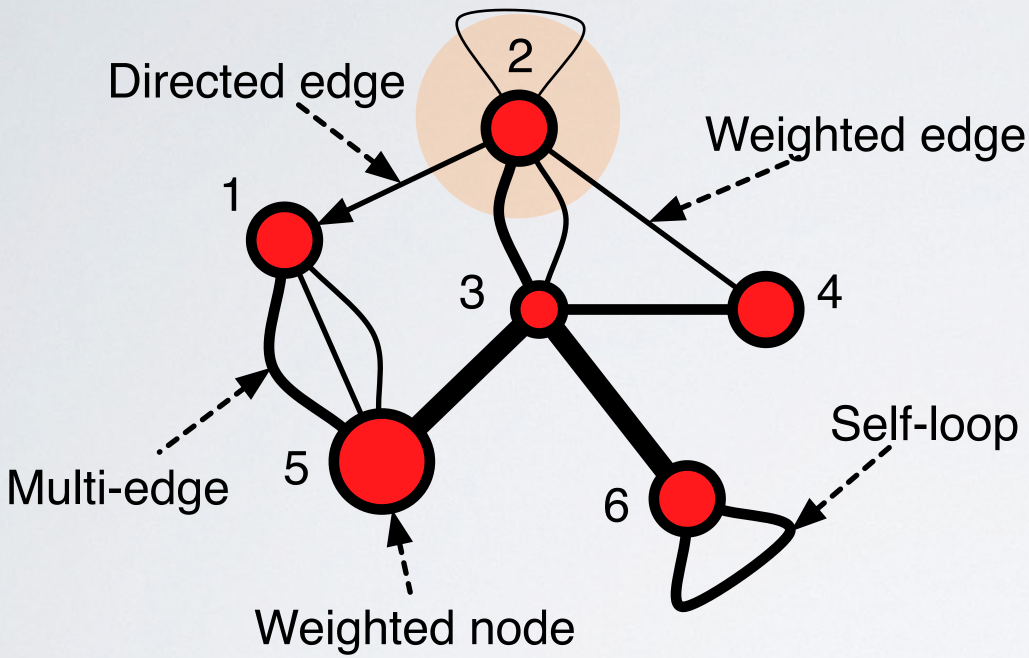


~~undirected~~

~~unweighted~~

~~no self-loops~~

a less simple network



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	$\rightarrow \{(3,4),(6,2)\}$

directed networks

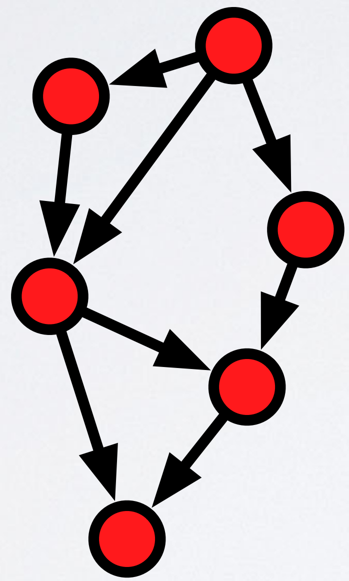
$$A_{ij} \neq A_{ji}$$

citation networks

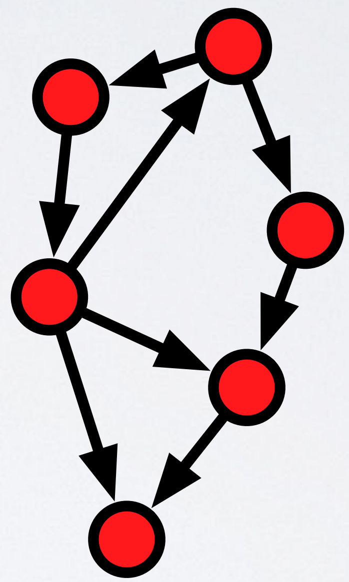
foodwebs*

epidemiological

others?



directed acyclic graph



directed graph

WWW

friendship?

flows of goods, information

economic exchange

dominance

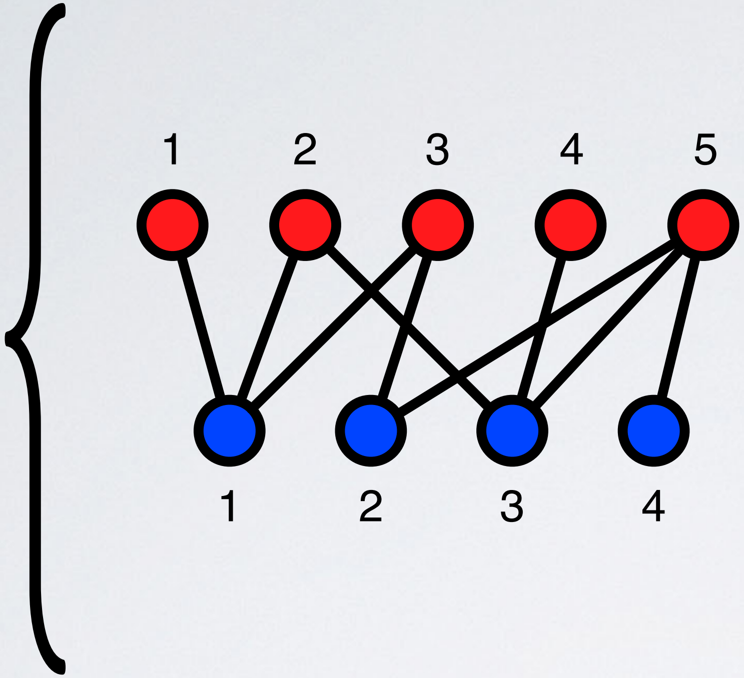
neuronal

transcription

time travelers

bipartite networks

bipartite network



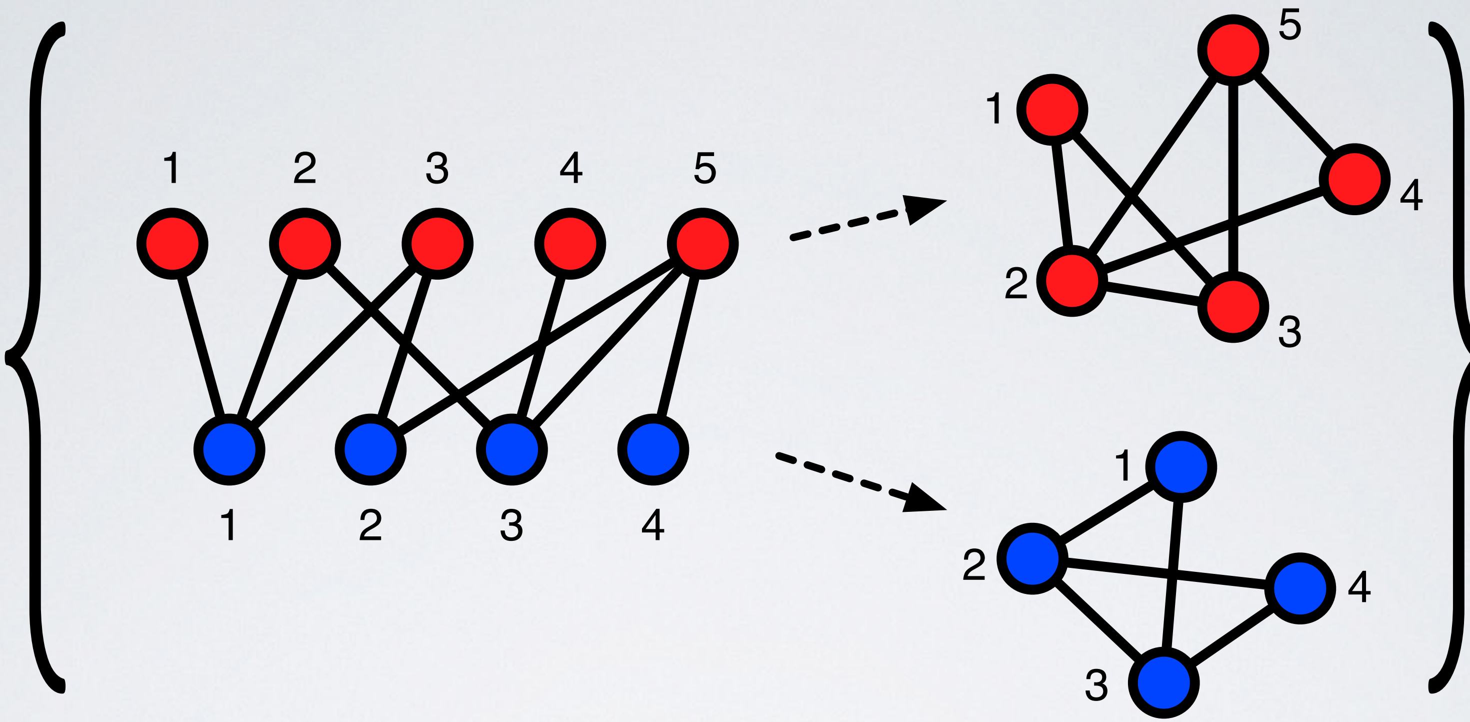
no within-type edges

authors & papers
actors & movies/scenes
musicians & albums
people & online groups
people & corporate boards

people & locations (checkins)
metabolites & reactions
genes & substrings
words & documents
plants & pollinators

bipartite networks

bipartite network



one-mode projections

no within-type edges

one type only

authors & papers

actors & movies/scenes

musicians & albums

people & online groups

people & corporate boards

people & locations (checkins)

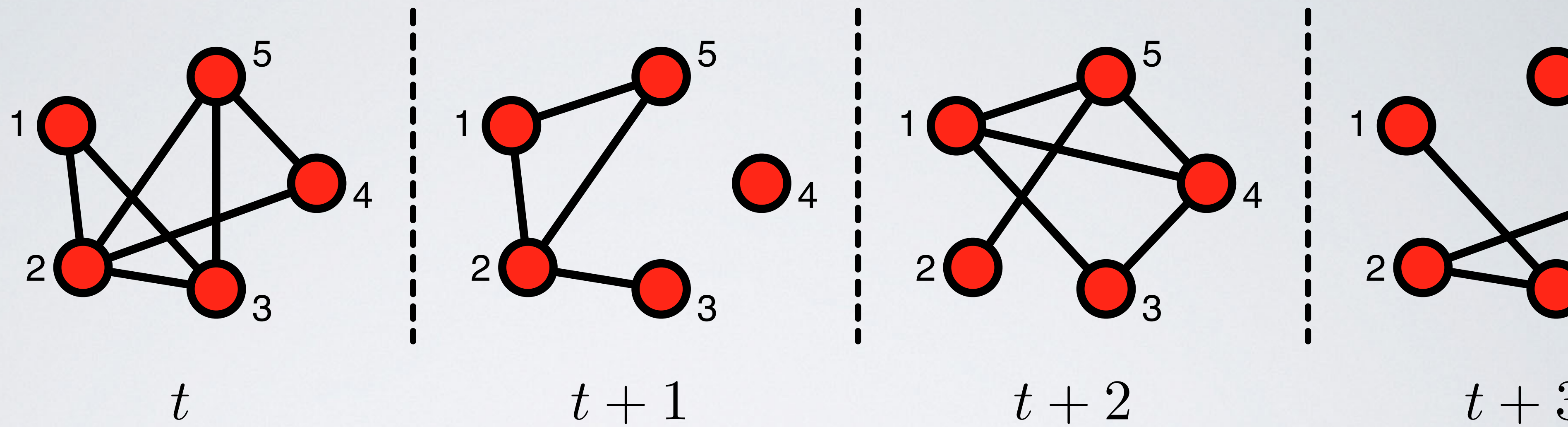
metabolites & reactions

genes & substrings

words & documents

plants & pollinators

temporal networks



any network over time; comes in two flavors

1. discrete time (snapshots), edges (i, j, t)
2. continuous time, edges $(i, j, t_s, \Delta t)$

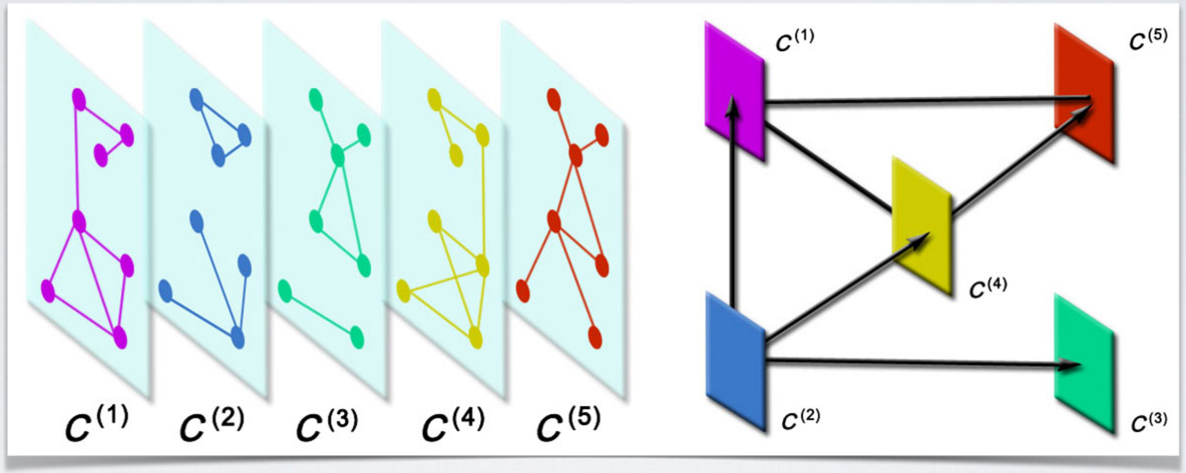
physical proximity over time

transportation connections over time

social interactions over time

Holme & Saramäki, Phys. Rep. 519, 97-125 (2012)

multiplex or multilayer networks



multiple network "layers"

each layer has same set of nodes V

but different sets of edges $\{E_1, E_2, \ldots, E_\ell\}$

different types of transportation within a city

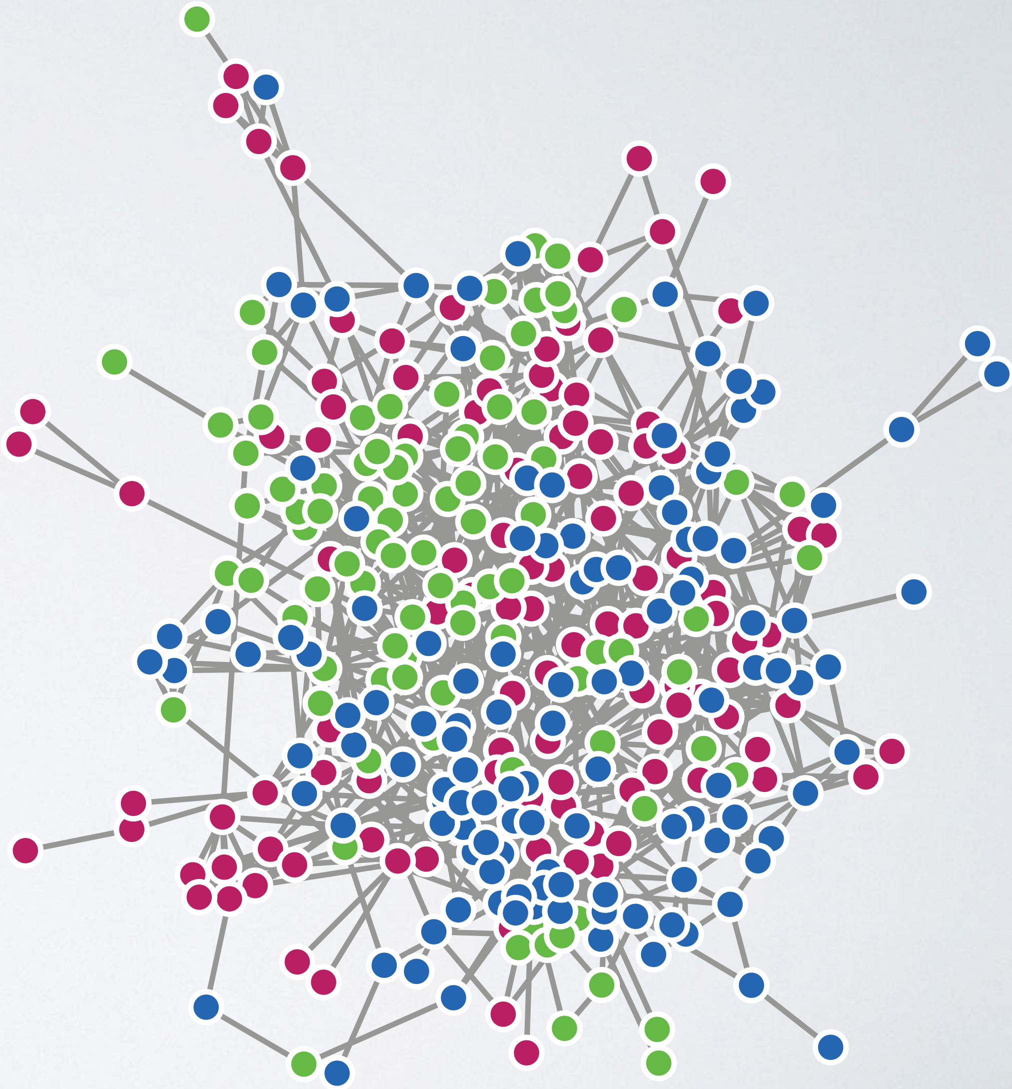
different types of social interactions (trust, socializing, co-located, etc.)

interactions on different social media platforms

De Domenico et al., Phys. Rev. X 3, 041022 (2013)
Kivelä et al., J. Complex Networks 2(3), 203-271 (2014)

analyzing networks

what real networks look like…



analyzing networks

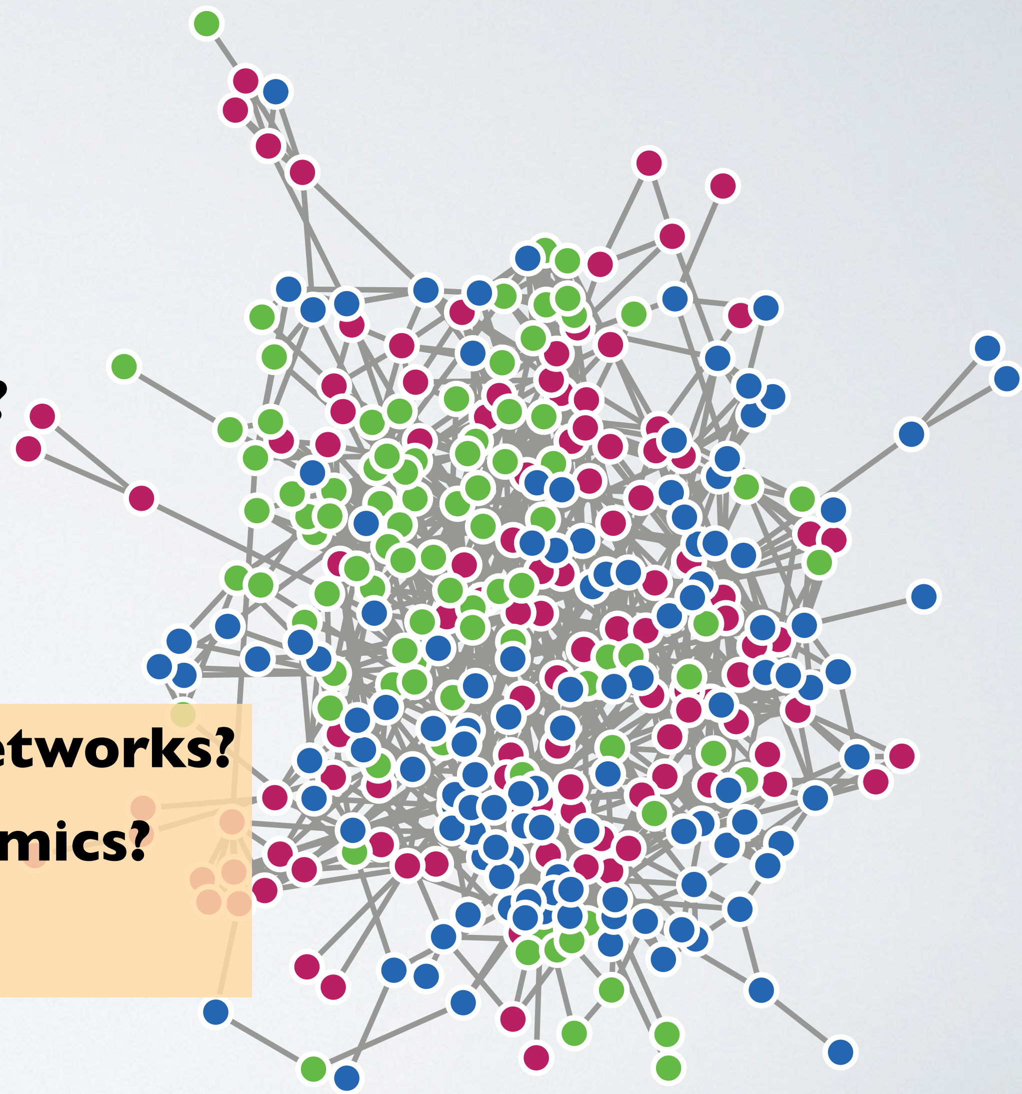
what real networks look like…

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?**

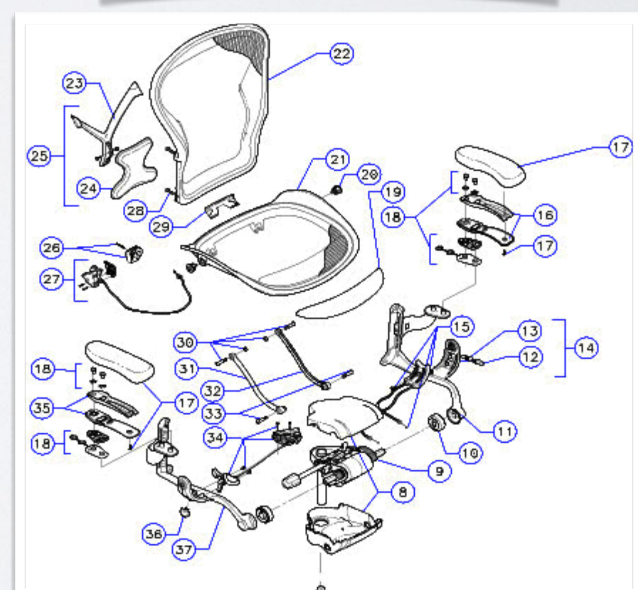


analyzing networks

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

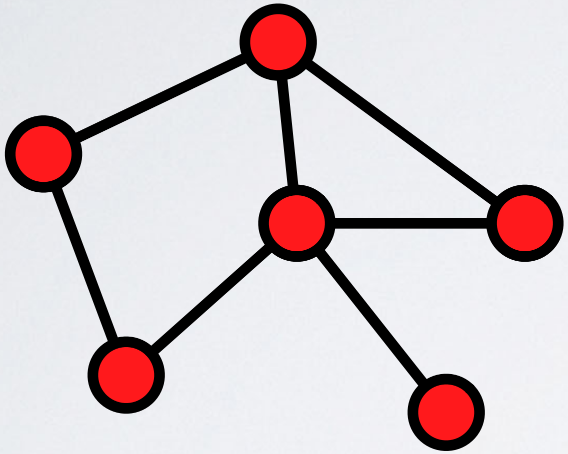
1. **exploratory data analysis:** count & compare all the things (degree distributions, centrality scores, community detection, etc.)
2. **simple regressions:** convert network structure into node-level features, and do traditional explanatory modeling
3. **null models:** use some kind of random graph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

analyzing networks

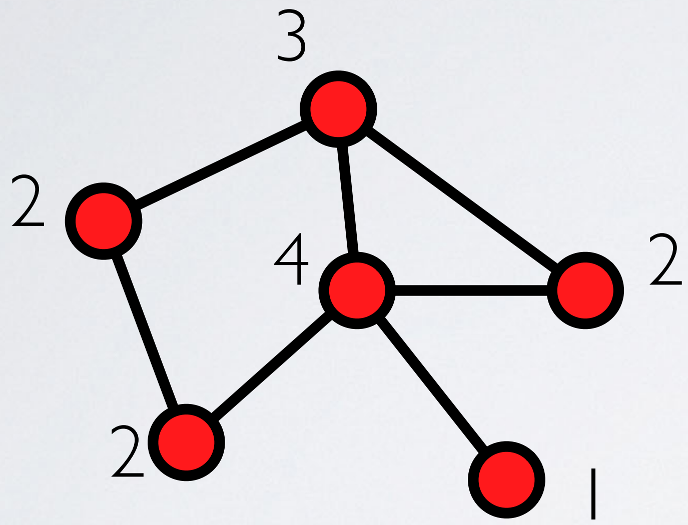
- degrees & distributions
- network position & centrality scores
- some applications

degree distributions

degree



degree distributions

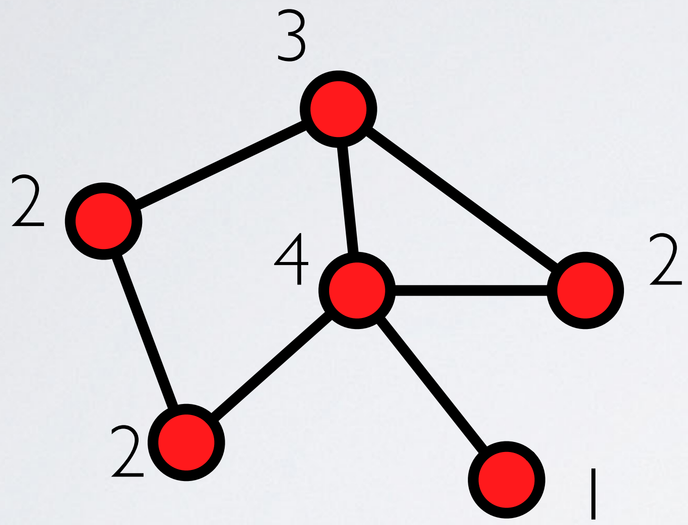


degree:

number of connections k

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

degree distributions



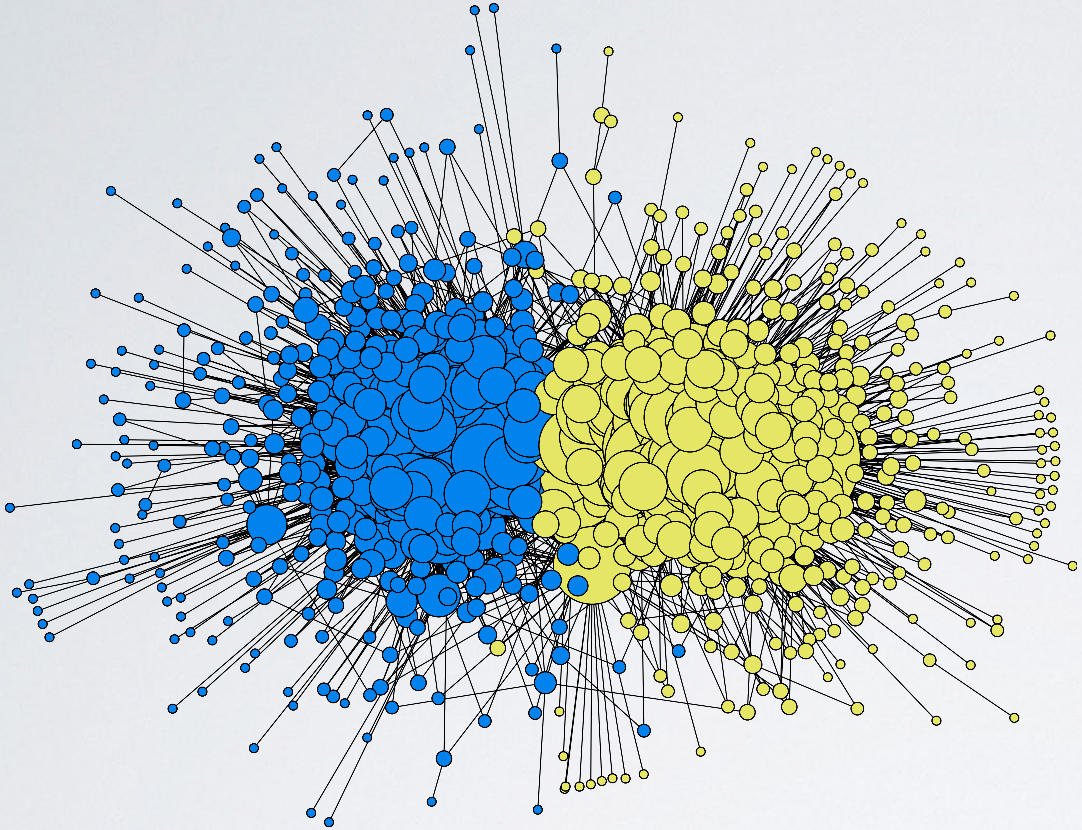
degree:
number of connections k

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

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

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

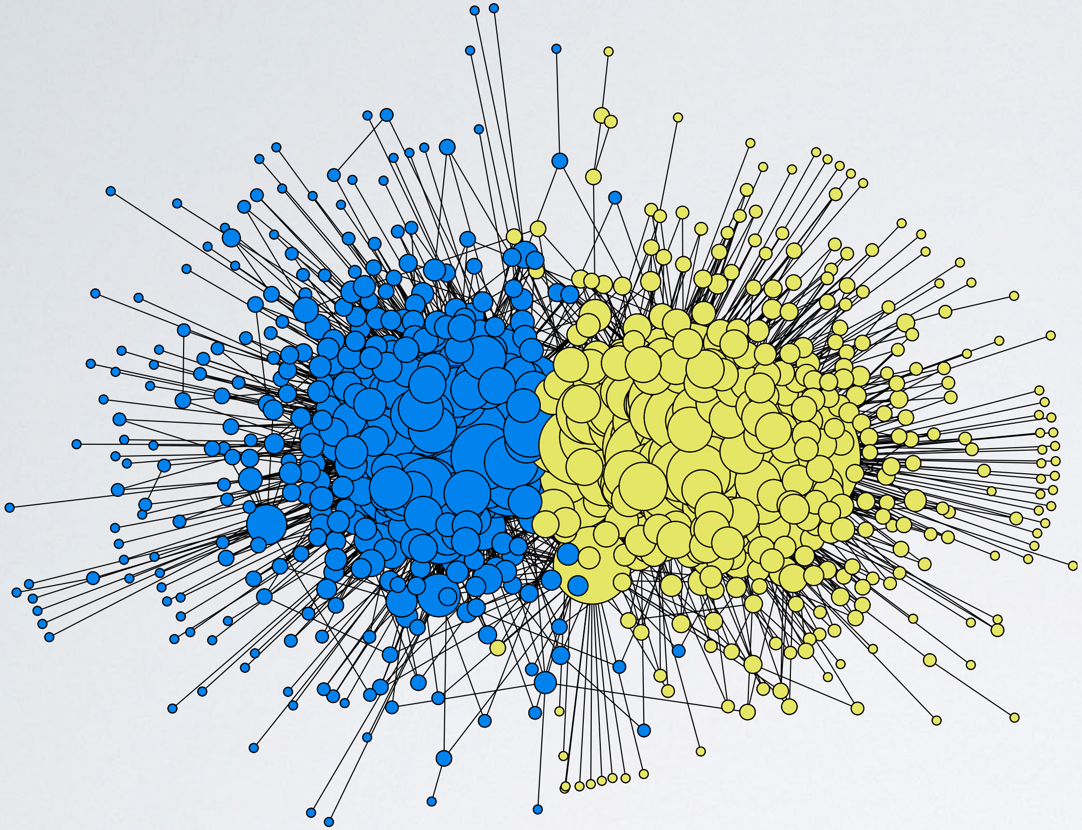
degree distributions



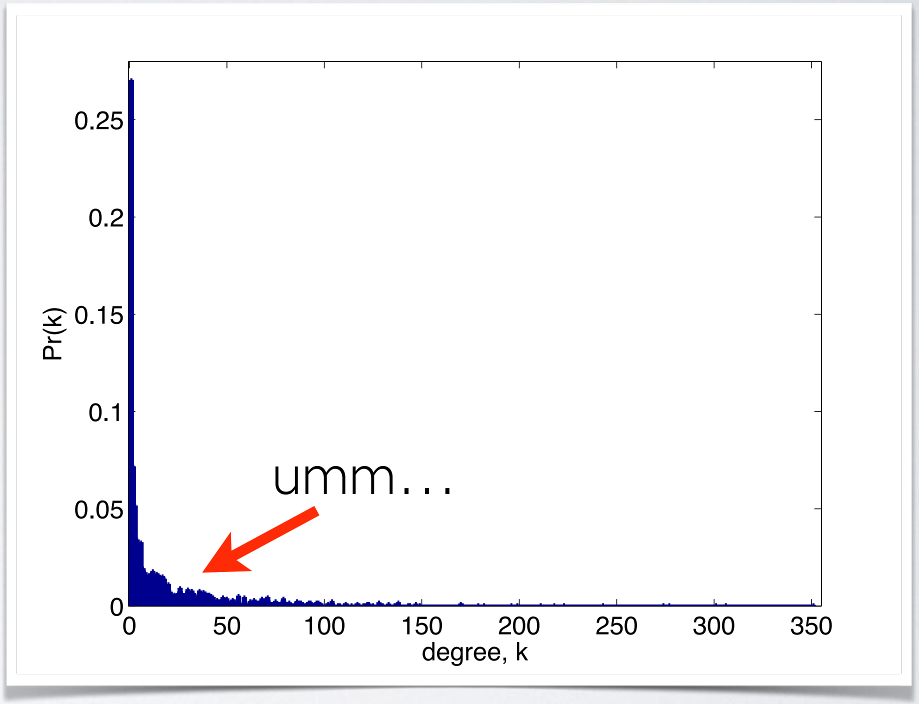
political blogs

Adamic & Glance, WWW Workshop on the Weblogging Ecosystem (2005)

degree distributions

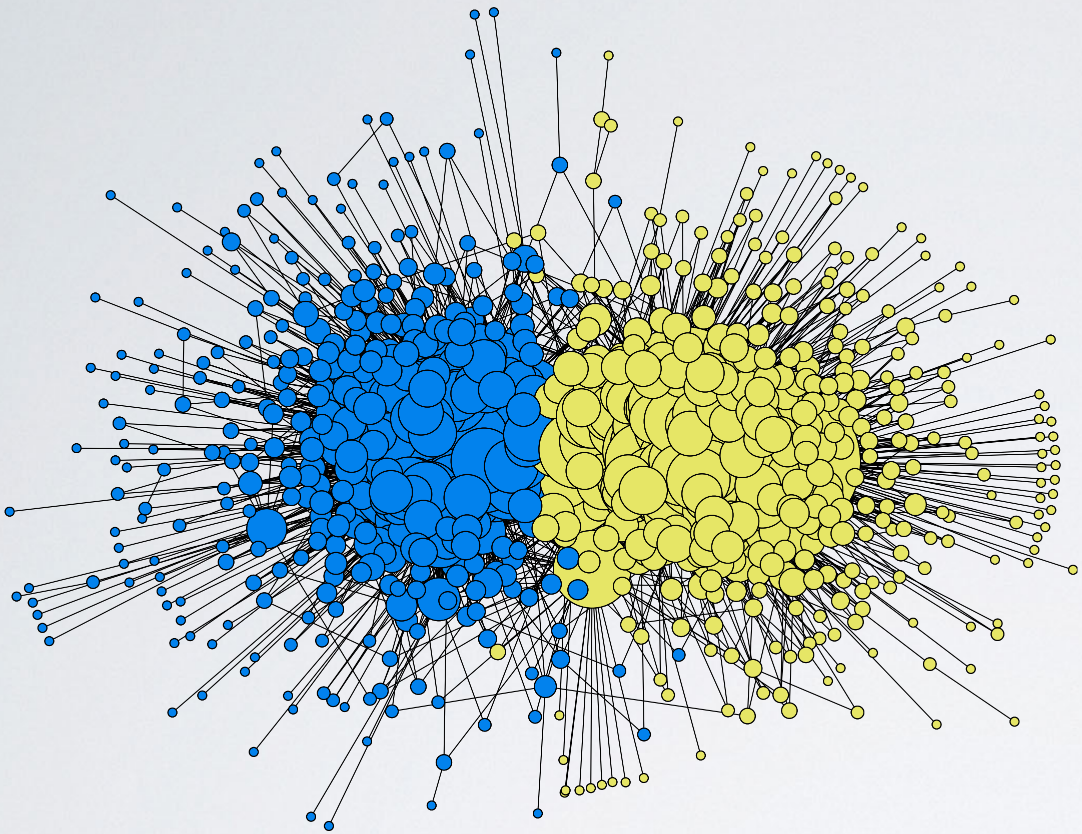


political blogs

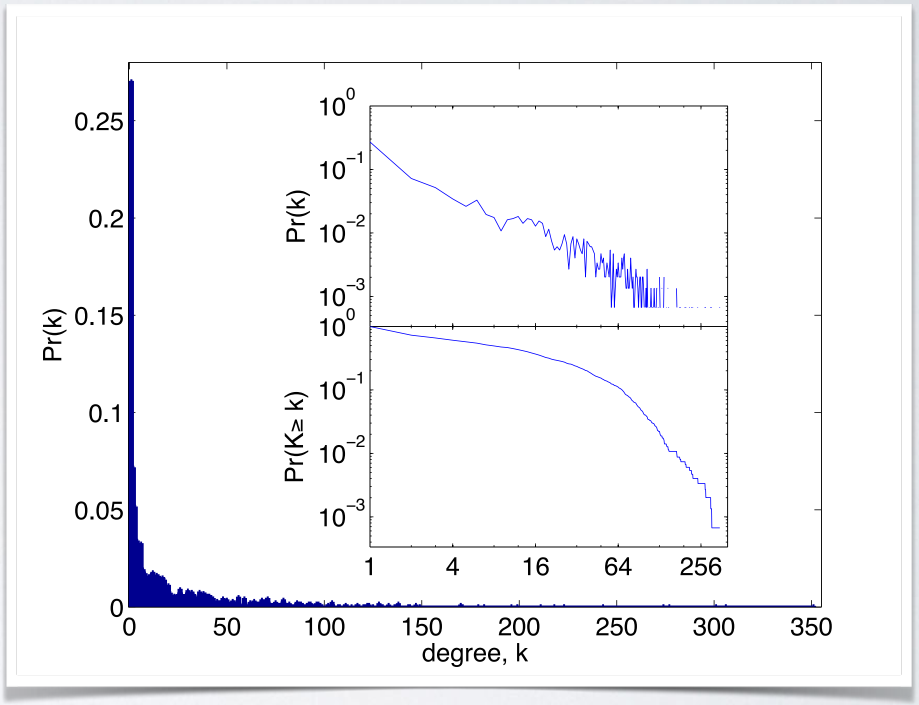


Adamic & Glance, WWW Workshop on the Weblogging Ecosystem (2005)

degree distributions

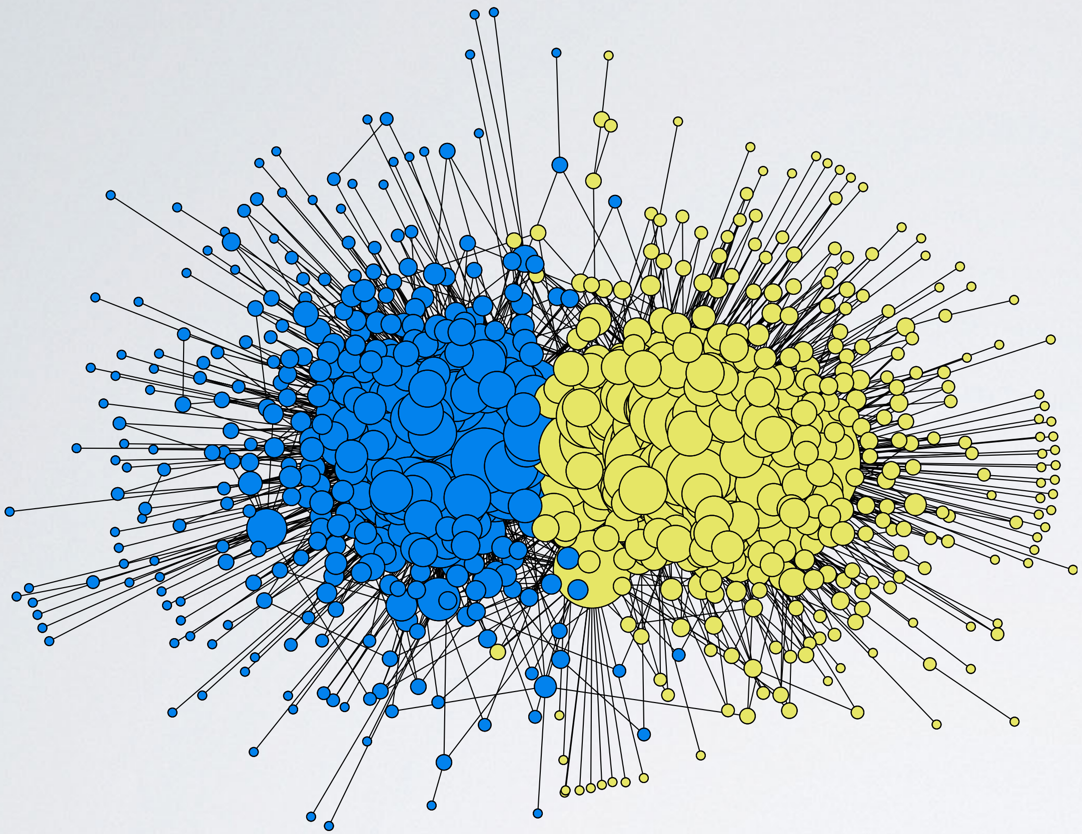


political blogs

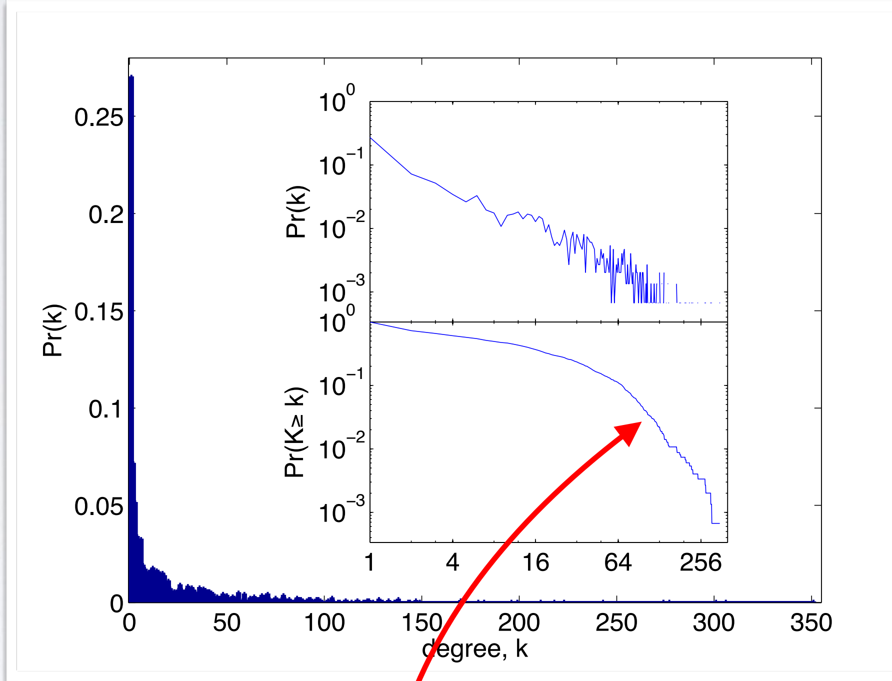


Adamic & Glance, WWW Workshop on the Weblogging Ecosystem (2005)

degree distributions



political blogs



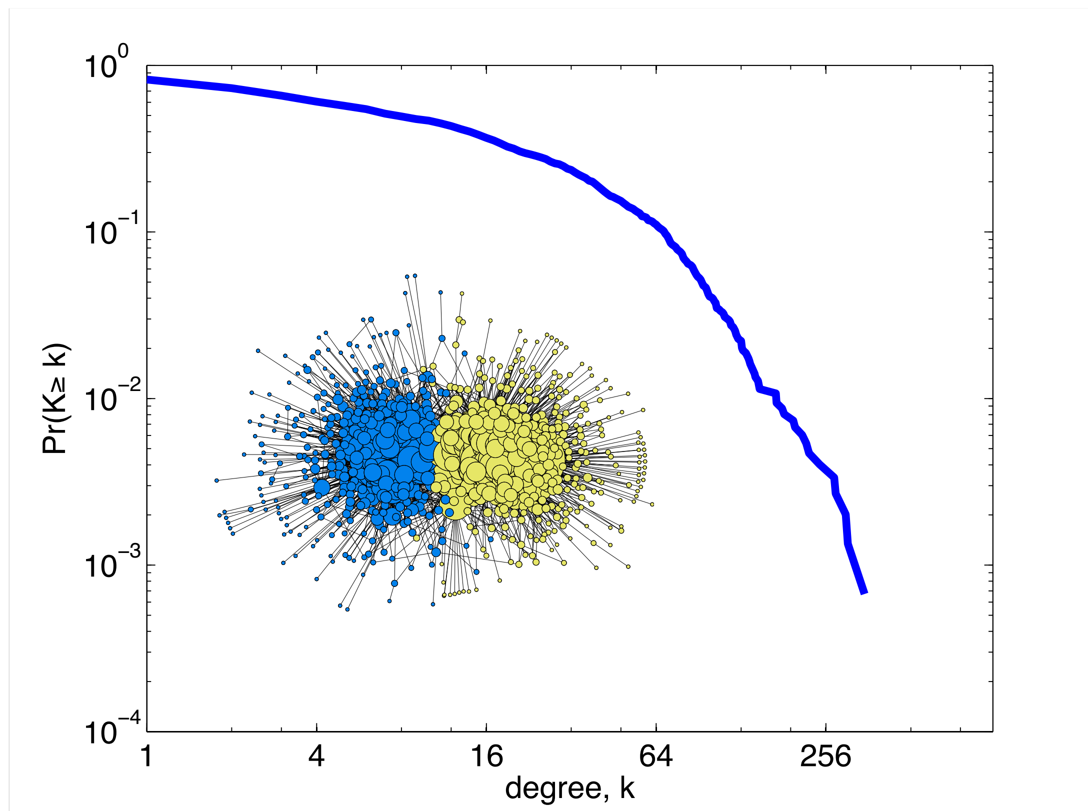
🤔 **is this a power law?**

Adamic & Glance, WWW Workshop on the Weblogging Ecosystem (2005)

degree distributions

🤔 fun facts 1:

- 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 identifying them in data (and not confusing them for other things) ***requires statistics***



Newman, Contemporary Physics 46, 323-351 (2005)
Clauset et al. SIAM Review 51, 661–703 (2009)
Broido & Clauset, Nat. Comm. 10 (2019)

degree distributions

🤔 fun facts 2:

- ***degree distribution*** is the first-order description of network structure
- ***degree heterogeneity*** alone drives interesting phenomena ("friendship paradox", spreading dynamics, etc.)
- ***degree heterogeneity*** alone explains many *other* network patterns (various centralities, disassortativity, etc.)
- the ***configuration model*** is how to tell : random graph model with specified degree sequence

airport network & global epidemics

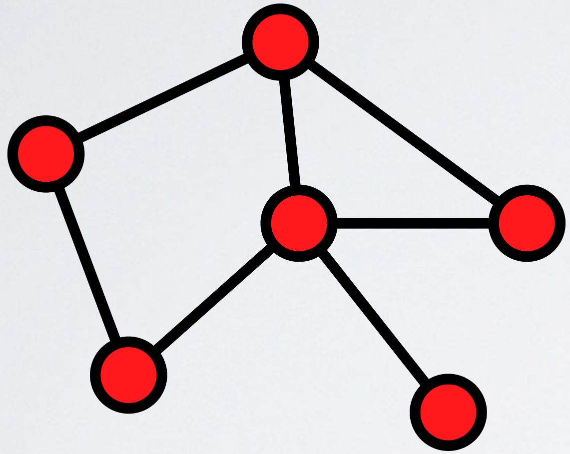
SIAM REVIEW
Vol. 60, No. 2, pp. 315–355

Configuring Random Graph Models with Fixed Degree Sequences*

Bailey K. Fosdick[†]
Daniel B. Larremore[‡]
Joel Nishimura[§]
Johan Ugander[¶]

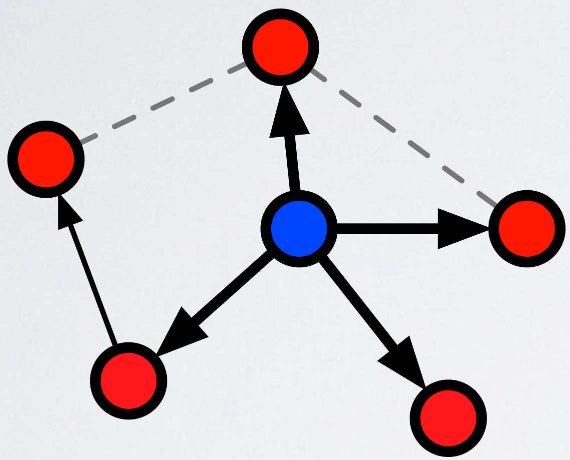
Fosdick et al., SIAM Review 60(2), 315-355 (2018)

network position



position

network position



position = centrality:

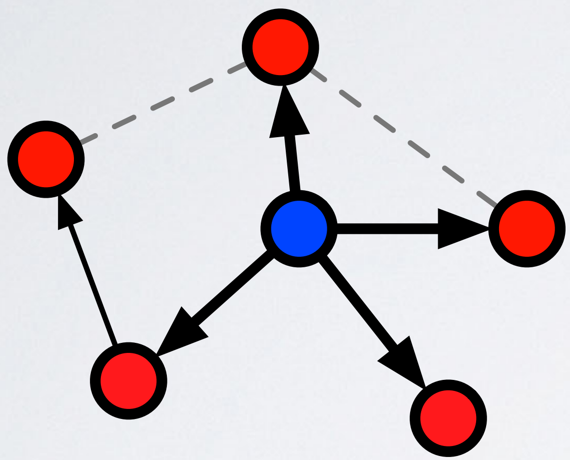
structural vs. dynamical importance

geometric	harmonic centrality
	closeness centrality
	betweenness centrality
connectivity	degree centrality
	eigenvector centrality
	PageRank
	Katz centrality
	many many more…

🤔 **structural importance = cheap estimate of dynamical importance (aka "influence")**

Boldi & Vigna, arXiv:1308.2140 (2013)
Borgatti, Social Networks 27, 55–71 (2005)

network position



position = centrality:

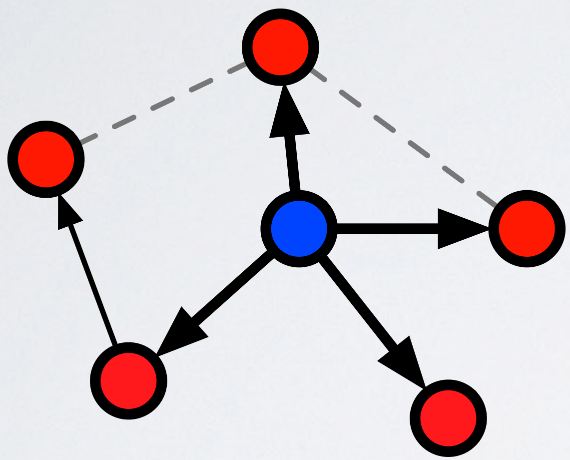
structural vs. dynamical importance

centrality = unsupervised node ranking

$$f : G \rightarrow \overrightarrow{v}$$

there are an infinite number of choices of f!
most are correlated
choose f that is most meaningful for downstream analysis

network position



position = centrality:

harmonic, closeness centrality

importance = being in "center" of the network

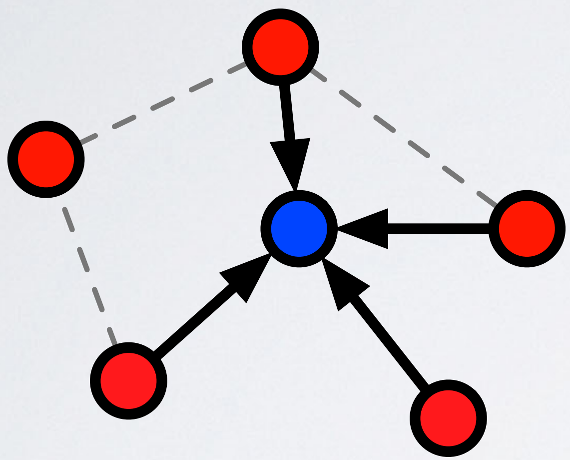
harmonic $$c_i = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$$

length of shortest path

distance: $$d_{ij} = \begin{cases} \boxed{\ell_{ij}} & \text{if } j \text{ reachable from } i \\ \infty & \text{otherwise} \end{cases}$$

Boldi & Vigna, arXiv:1308.2140 (2013)
Borgatti, Social Networks 27, 55–71 (2005)

network position



position = centrality:
PageRank, Katz, eigenvector centrality

importance = sum of importances of nodes that point at you*

$$I_i = \sum_{j \to i} \frac{I_j}{k_j}$$

or, the right eigenvector of

$$\mathbf{A}\mathbf{x} = \lambda\,\mathbf{x}$$

Boldi & Vigna, arXiv:1308.2140 (2013)
Borgatti, Social Networks 27, 55–71 (2005)

*modulo several technical details

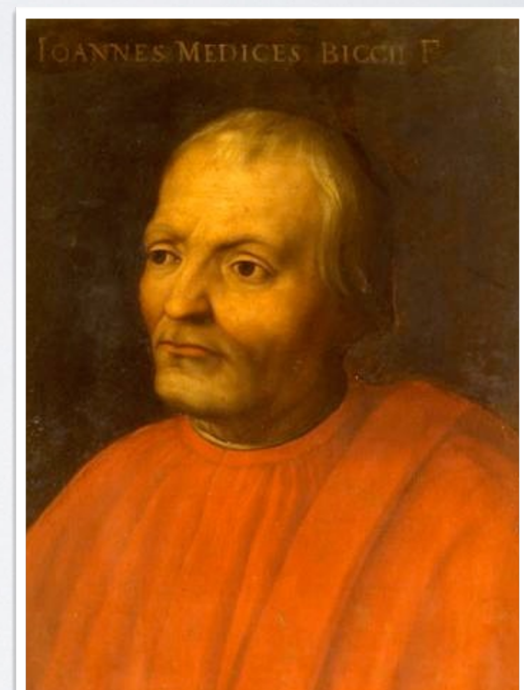
network position

Robust Action and the Rise of the Medici, 1400–1434

John F. Padgett and Christopher K. Ansell
University of Chicago

Duomo

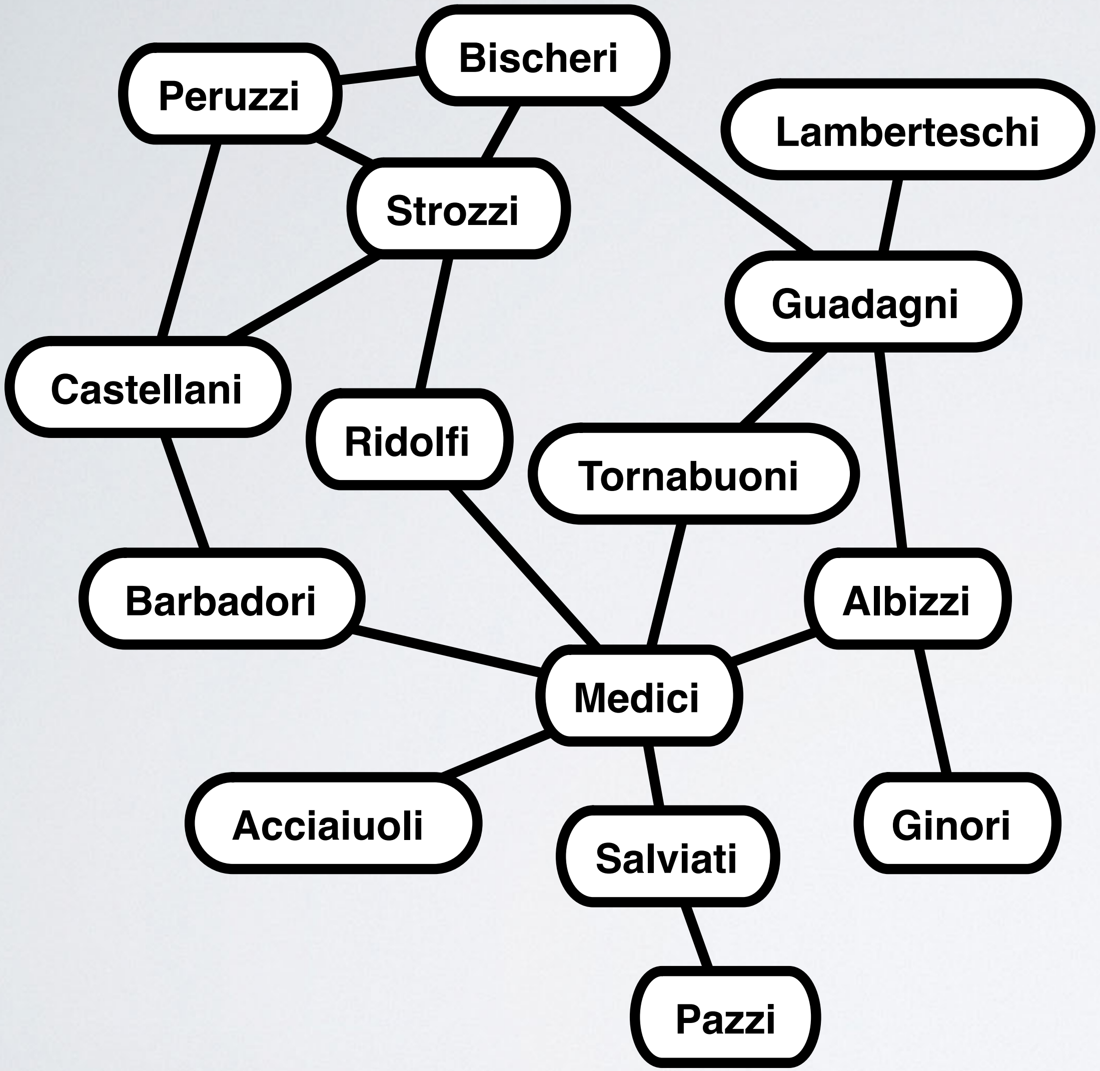
Palazzo Medici

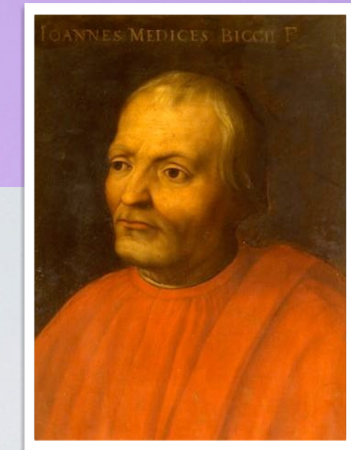


Giovanni de Medici

Padgett & Ansell, Am. J. Soc. 98(6), 1259-1319 (1993)

network position

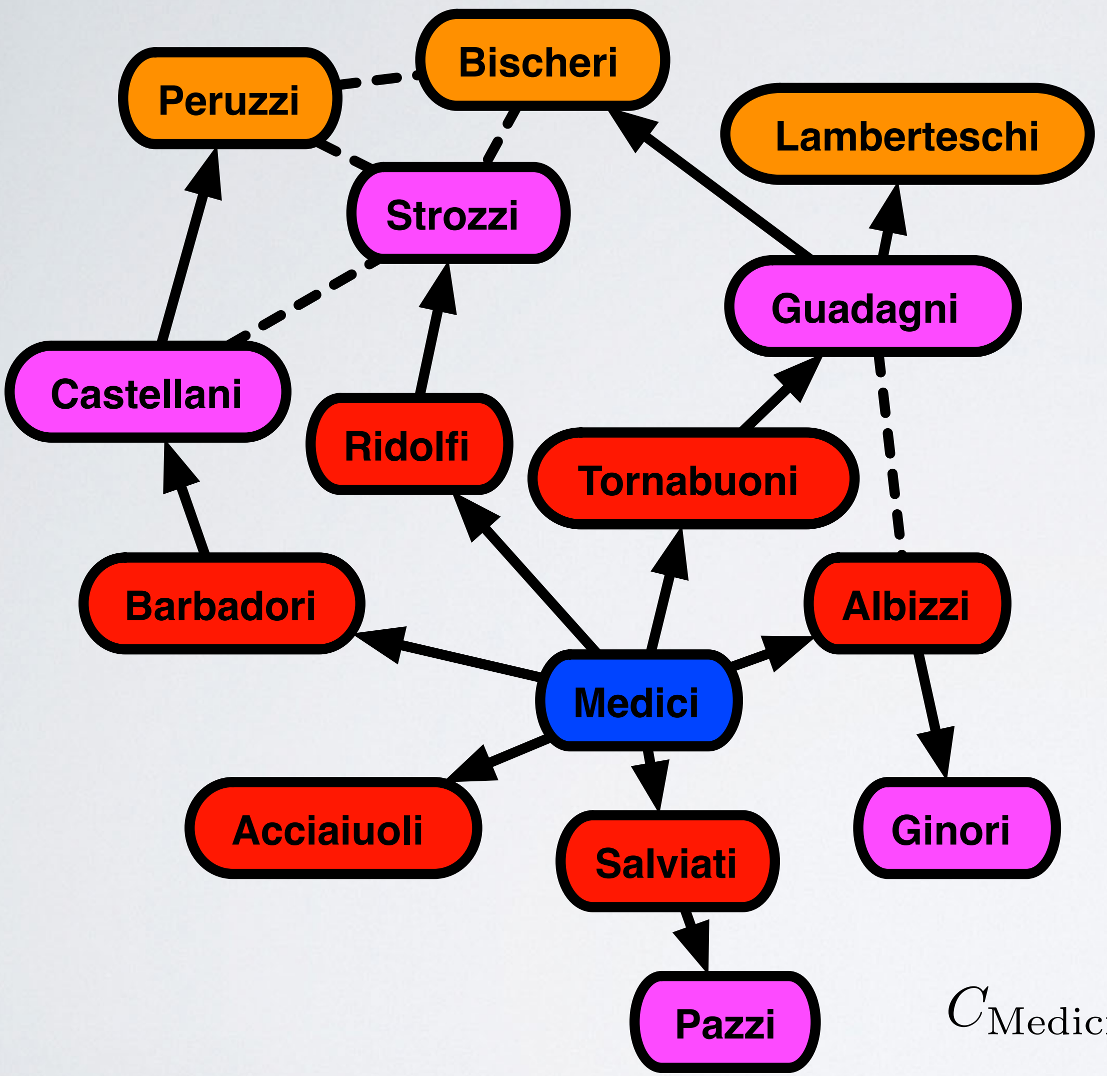


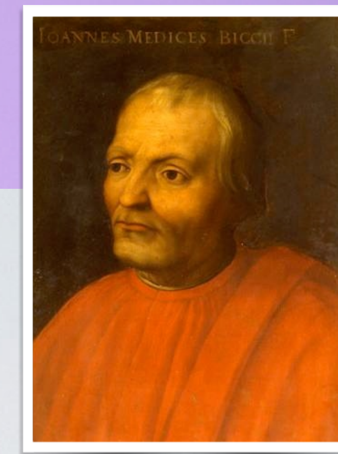


nodes: Florence families
edges: inter-family marriages

which family is most central?

network position



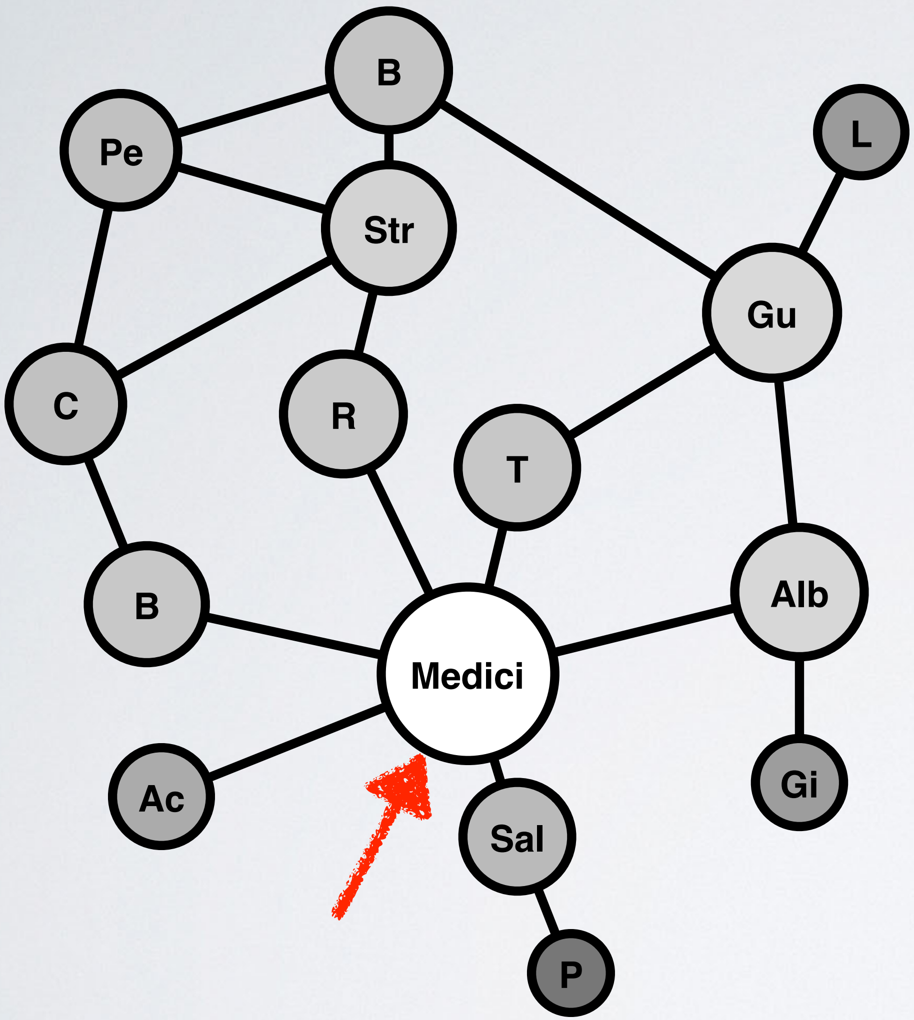


nodes: Florence families
edges: inter-family marriages

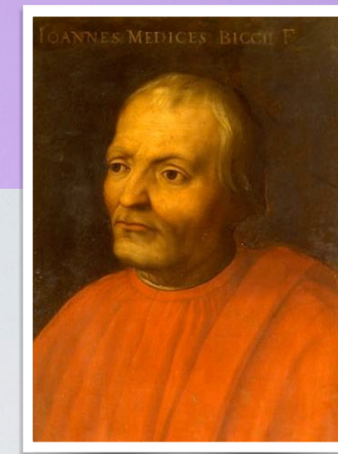
which family is most central?
Medici.

$$C_{\text{Medici}} = 6\left(\frac{1}{1}\right) + 5\left(\frac{1}{2}\right) + 3\left(\frac{1}{3}\right)$$
$$= 9.5$$

network position

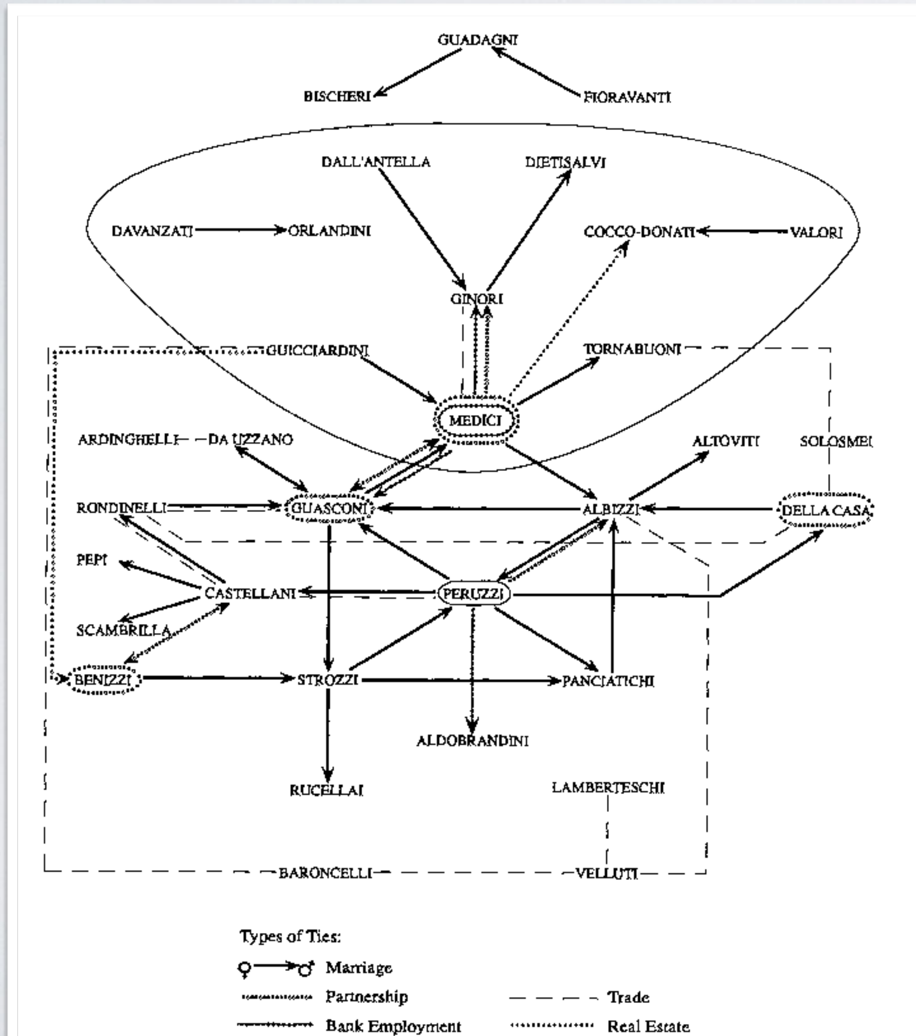


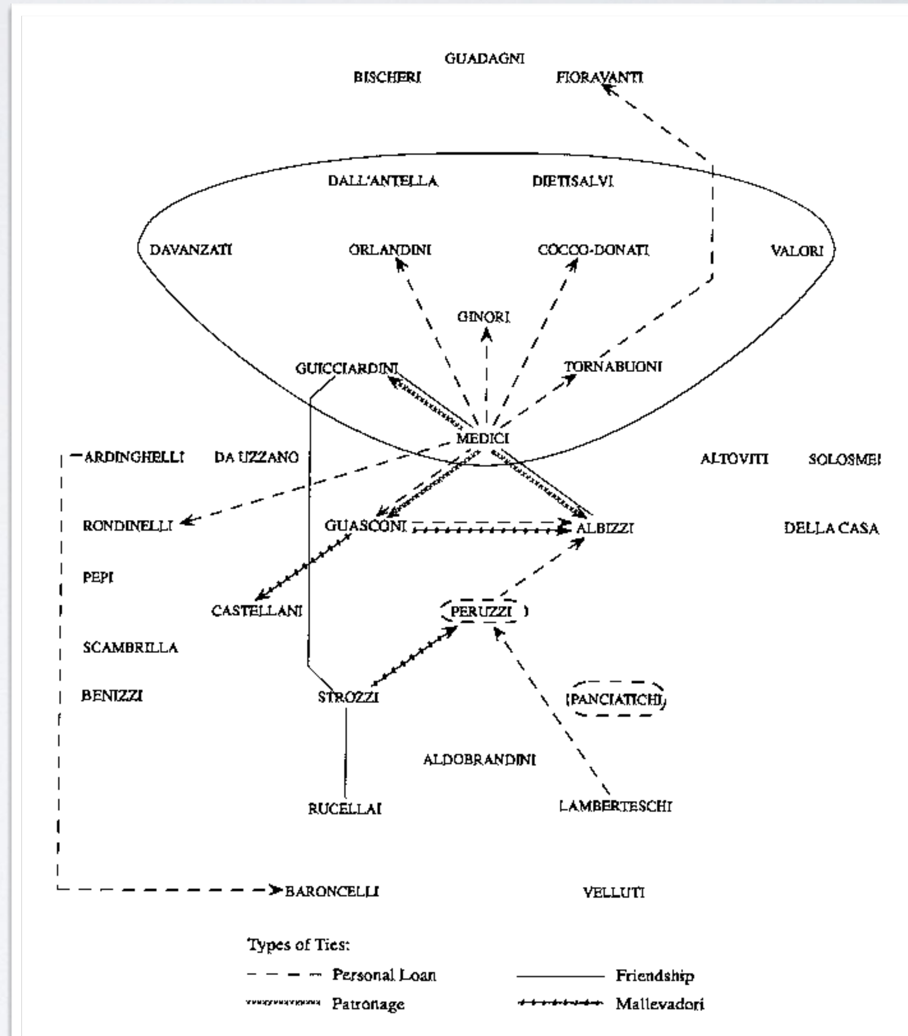
Medici	9.5
Guadagni	7.92
Albizzi	7.83
Strozzi	7.67
Ridolfi	7.25
Bischeri	7.2
Tornabuoni	7.17
Barbadori	7.08
Peruzzi	6.87
Castellani	6.87
Salviati	6.58
Acciaiuoli	5.92
Ginori	5.33
Lamberteschi	5.28
Pazzi	4.77



network position

actually, it's complicated…





[1] Marriage edges were only one type of inter-family interaction; hence, centrality on them alone is a simplification, and the deeper questions are about *dynamics* (how did the full network assemble over time?) and about *function* (was network position causally related to Medici dominance?).

let's apply these network concepts

RESEARCH ARTICLE

NETWORK SCIENCES

Systematic inequality and hierarchy in faculty hiring networks

Aaron Clauset,[1,2,3]* Samuel Arbesman,[4] Daniel B. Larremore[5,6]

10.1126/sciadv.1400005

let's apply these network concepts

RESEARCH ARTICLE

NETWORK SCIENCES

Systematic inequality and hierarchy in faculty hiring networks

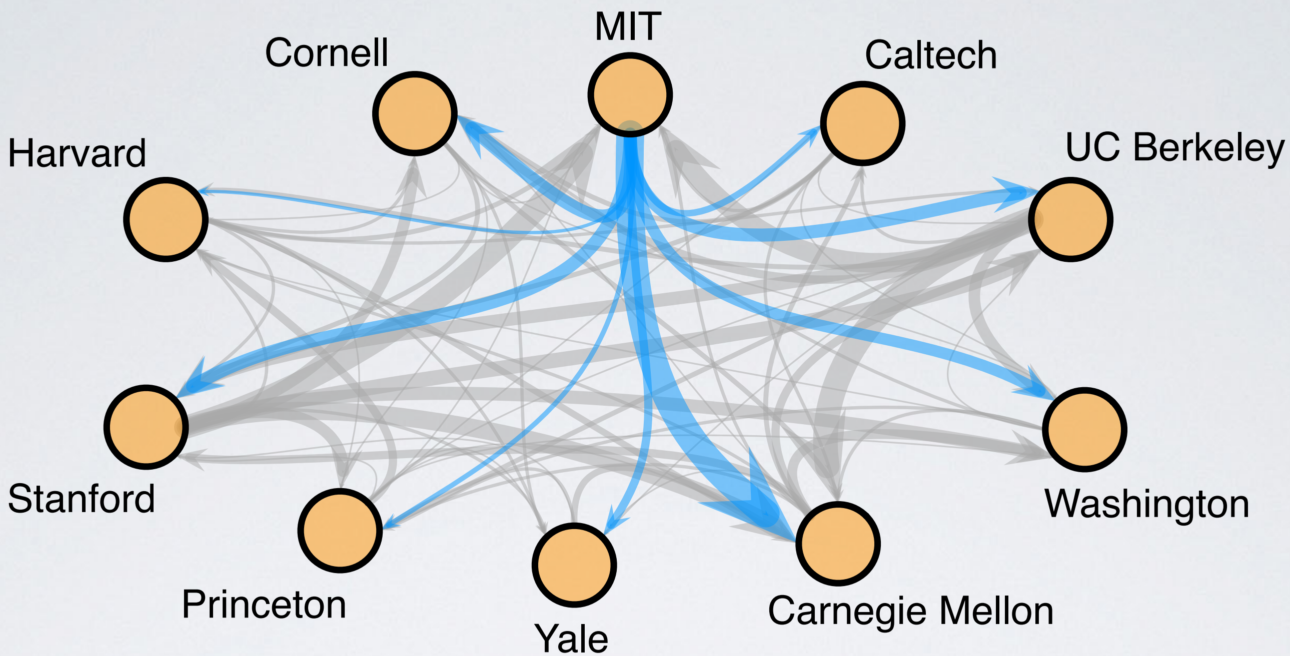
Aaron Clauset,[1,2,3]* Samuel Arbesman,[4] Daniel B. Larremore[5,6]

10.1126/sciadv.1400005

faculty market is a *network*

- vertices are PhD-granting universities
- consumers $\leftrightarrow$ producers
- v hires from u, add an edge $u \rightarrow v$

Clauset et al., Sci. Adv. 1(1), e1400005 (2015)



faculty market is a *network*

- vertices are PhD-granting universities
- consumers $\leftrightarrow$ producers
- v hires from u , add an edge $u \rightarrow v$

[1] actual exchanges of 267 faculty among 10 elite CS departments, from our 2011 data, without self-hires

collecting the data

complete, hand-curated data for 19,000 tenure-track faculty across 461 departments in

- Computer Science (205 depts)
- Business (112)
- History (144)

roughly 5000 hours of manual data collection

```
>>> record 1059
# facultyName  : James H. Martin
# email        : [redacted]
# sex          : M
# department   : Computer Science
# place        : University of Colorado, Boulder
# current      : Full Professor
# [Education]
# degree       : BS
# place        : Columbia University
# field        : Computer Science
# years        : ????-1981
# [Education]
# degree       : PhD
# place        : UC Berkeley
# field        : Computer Science
# years        : ????-1988
# [Faculty]
# rank         : Assistant Professor
# place        : University of Colorado, Boulder
# years        : 1989-1995
# [Faculty]
# rank         : Associate Professor
# place        : University of Colorado, Boulder
# years        : 1995-2007
# [Faculty]
# rank         : Full Professor
# place        : University of Colorado, Boulder
# years        : 2007-2011
# recordDate   : 7/4/2011
```

[1] CS data from 2011, Business schools from 2012, History from 2013
[2] all data from public sources, mainly faculty CVs and homepages
[3] data collected by a team of 12 students over 3 years, using a random 20% re-collection protocol, with script-based post-processing to detect errors & inconsistencies, which were then corrected by hand

collecting the data

complete, hand-curated data for 19,000 tenure-track faculty across 461 departments in

- Computer Science (205 depts)
- Business (112)
- History (144)

roughly 5000 hours of manual data collection

```
>>> record 1059
# facultyName : James H. Martin
# email       : 
# sex         : M
# department  : Computer Science
# place       : University of Colorado, Boulder
# current     : Full Professor
# [Education]
# degree      : BS
# place       : Columbia University
# field       : Computer Science
# years       : ????-1981
# [Education]
# degree      : PhD
# place       : UC Berkeley
# field       : Computer Science
# years       : ????-1988
# [Faculty]
# rank        : Assistant Professor
# place       : University of Colorado, Boulder
# years       : 1989-1995
# [Faculty]
# rank        : Associate Professor
# place       : University of Colorado, Boulder
# years       : 1995-2007
# [Faculty]
# rank        : Full Professor
# place       : University of Colorado, Boulder
# years       : 2007-2011
# recordDate  : 7/4/2011
```

[1] CS data from 2011, Business schools from 2012, History from 2013
[2] all data from public sources, mainly faculty CVs and homepages
[3] data collected by a team of 12 students over 3 years, using a random 20% re-collection protocol, with script-based post-processing to detect errors & inconsistencies, which were then corrected by hand

what's in the data

	Computer Science	Business	History
institutions	205	112	144
tenure-track faculty	5032	9336	4556
mean size	25	83	32

$\sum = 18,924$

what's in the data

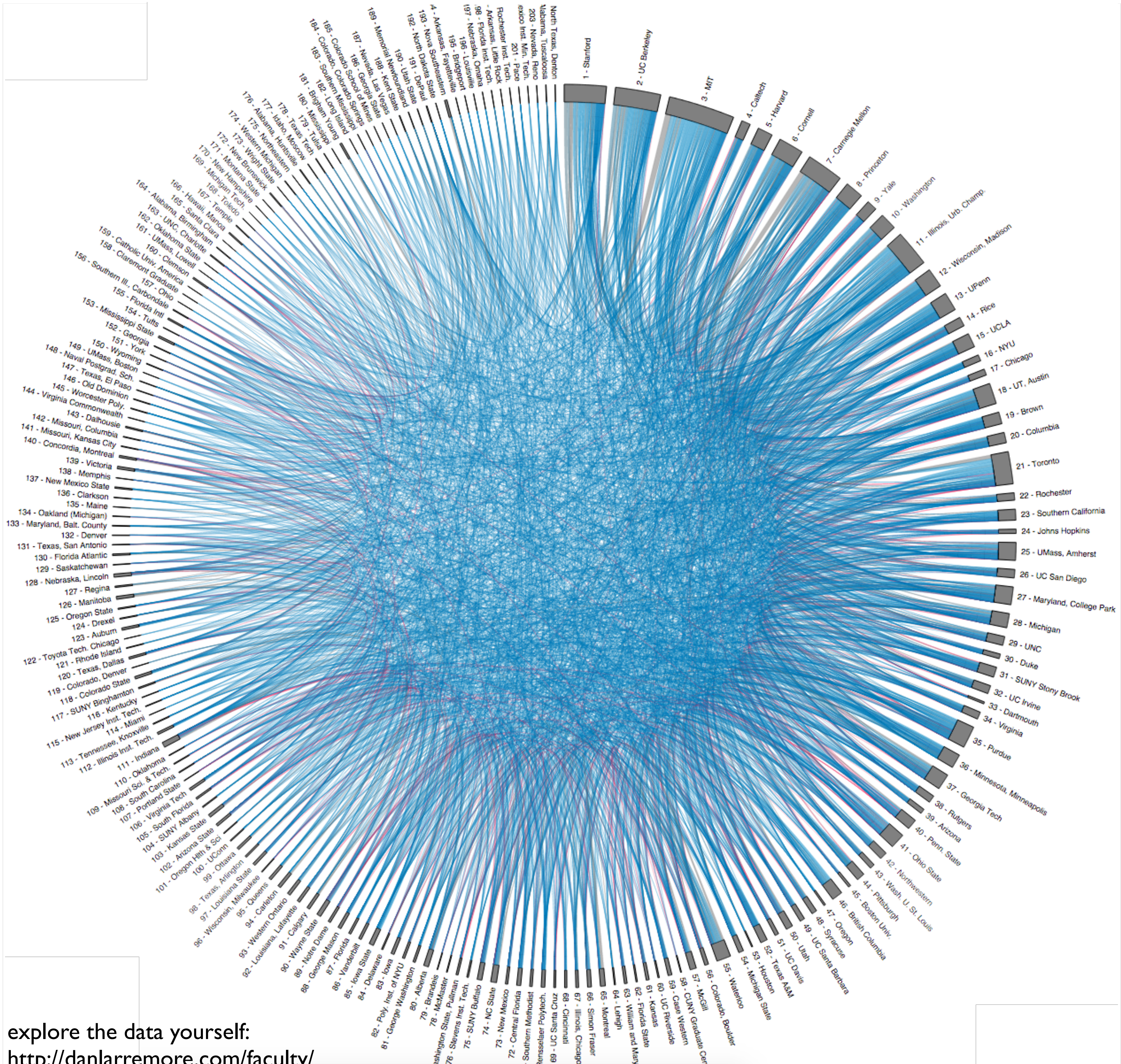
	Computer Science	Business	History	
institutions	205	112	144	
tenure-track faculty	5032	9336	4556	$\sum = 18,924$
mean size	25	83	32	
Full Professors	2400 (48%)	4294 (46%)	2097 (46%)	
Associate Prof.	1772 (35%)	2521 (27%)	1611 (35%)	
Assistant Prof.	860 (17%)	2521 (27%)	848 (19%)	
female	15%	22%	36%	

what's in the data

	Computer Science	Business	History
institutions	205	112	144
tenure-track faculty	5032	9336	4556
mean size	25	83	32
Full Professors	2400 (48%)	4294 (46%)	2097 (46%)
Associate Prof.	1772 (35%)	2521 (27%)	1611 (35%)
Assistant Prof.	860 (17%)	2521 (27%)	848 (19%)
female	15%	22%	36%
PhDs in-sample	87%	84%	89%

$\sum = 18,924$

nearly closed hiring systems



explore the data yourself:
http://danlarremore.com/faculty/

huge inequalities in faculty production

huge inequalities in faculty production

Gini coefficients (out-degree)

- 0.69, 0.62, 0.72

50% of faculty from

- 18, 16, 8 universities

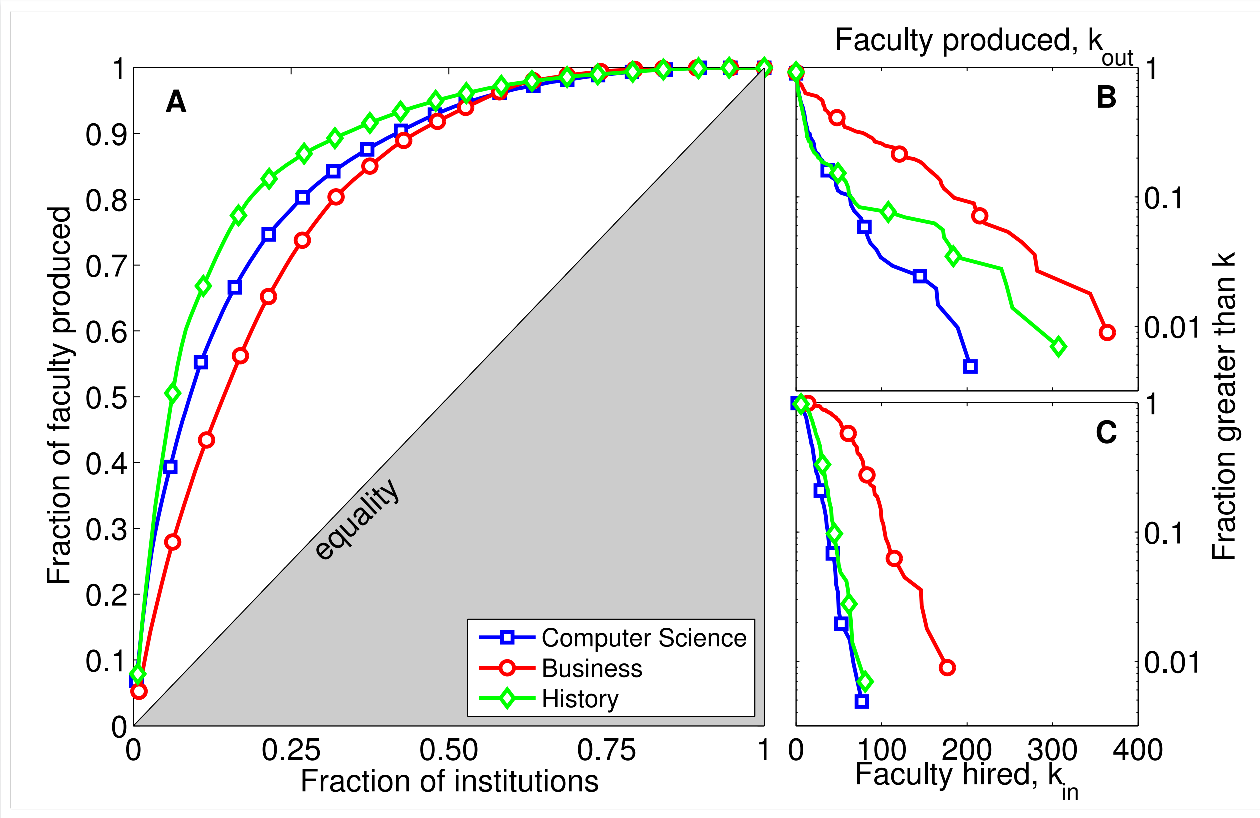
net producers $k_{\text{out}}/k_{\text{in}} > 1$

- 24%, 36%, 18%

1-10 producers vs.

- 11-20 : 1.6, 2.1, 3.0x more
- 21-30 : 3.1, 2.3, 5.6x more

CS Business History

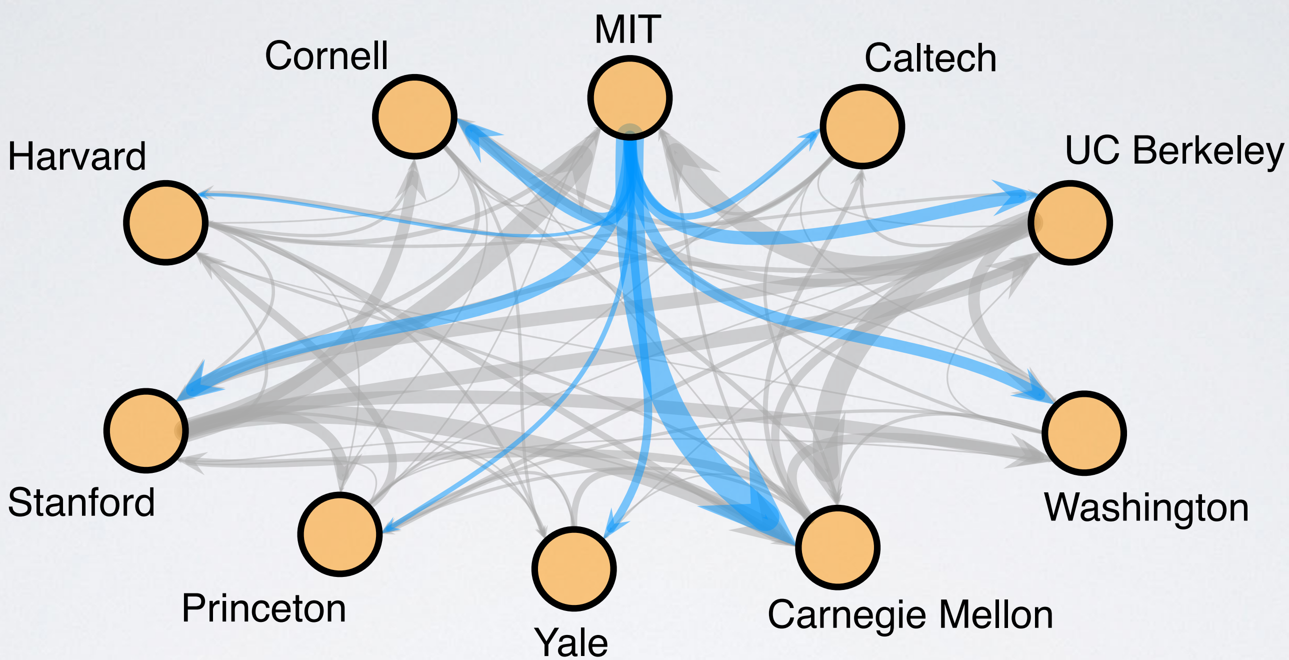


[1] order: CS, Business, History
[2] U.S. Income Gini coefficient = 0.45

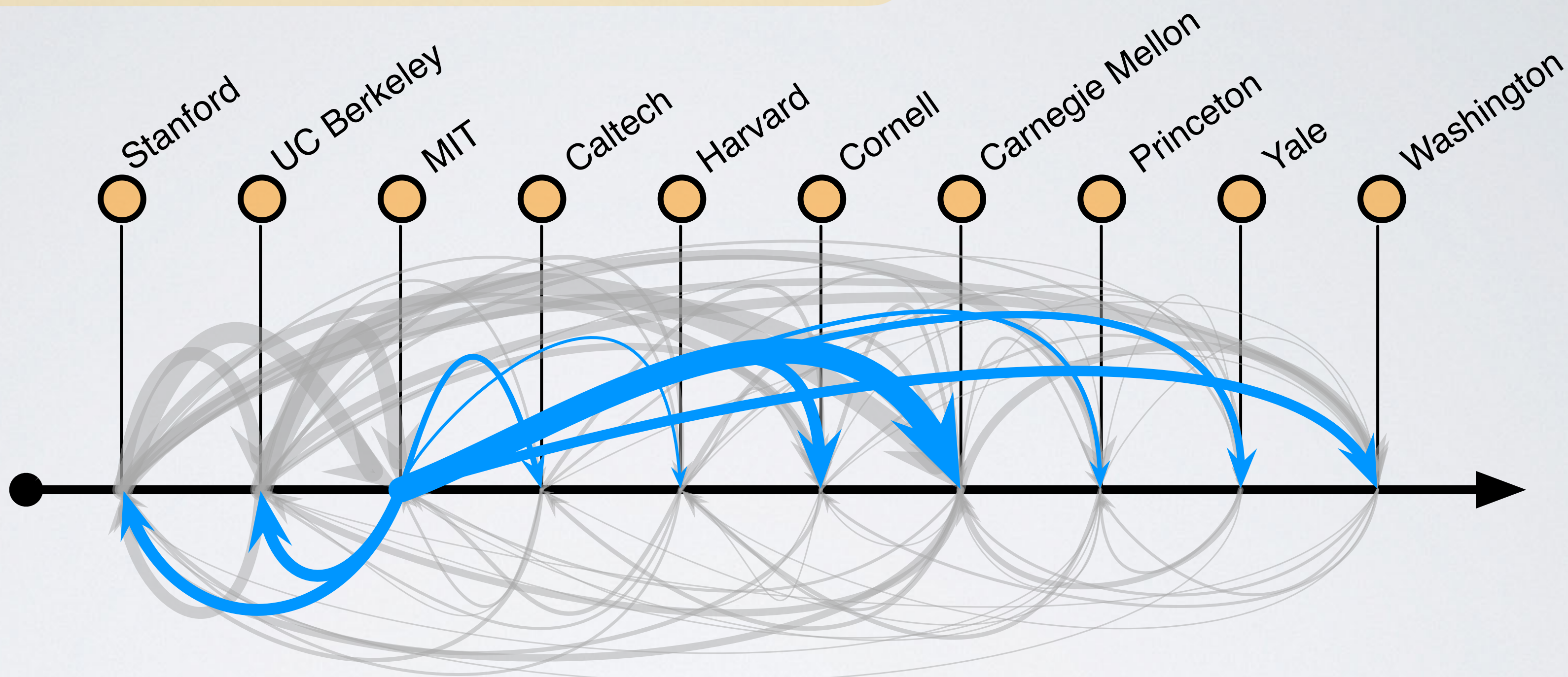
a prestige hierarchy

- difficult to talk about inequalities in academia without talking about rankings (centralities!)
- let's extract a data-driven ranking from the network

a prestige hierarchy



a prestige hierarchy

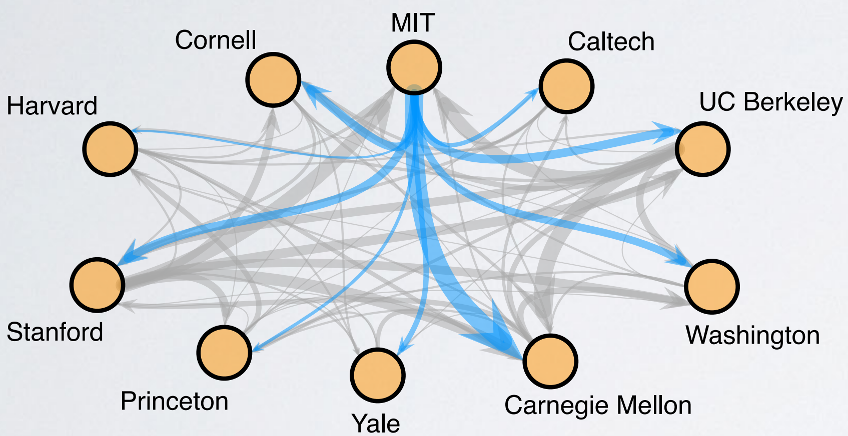


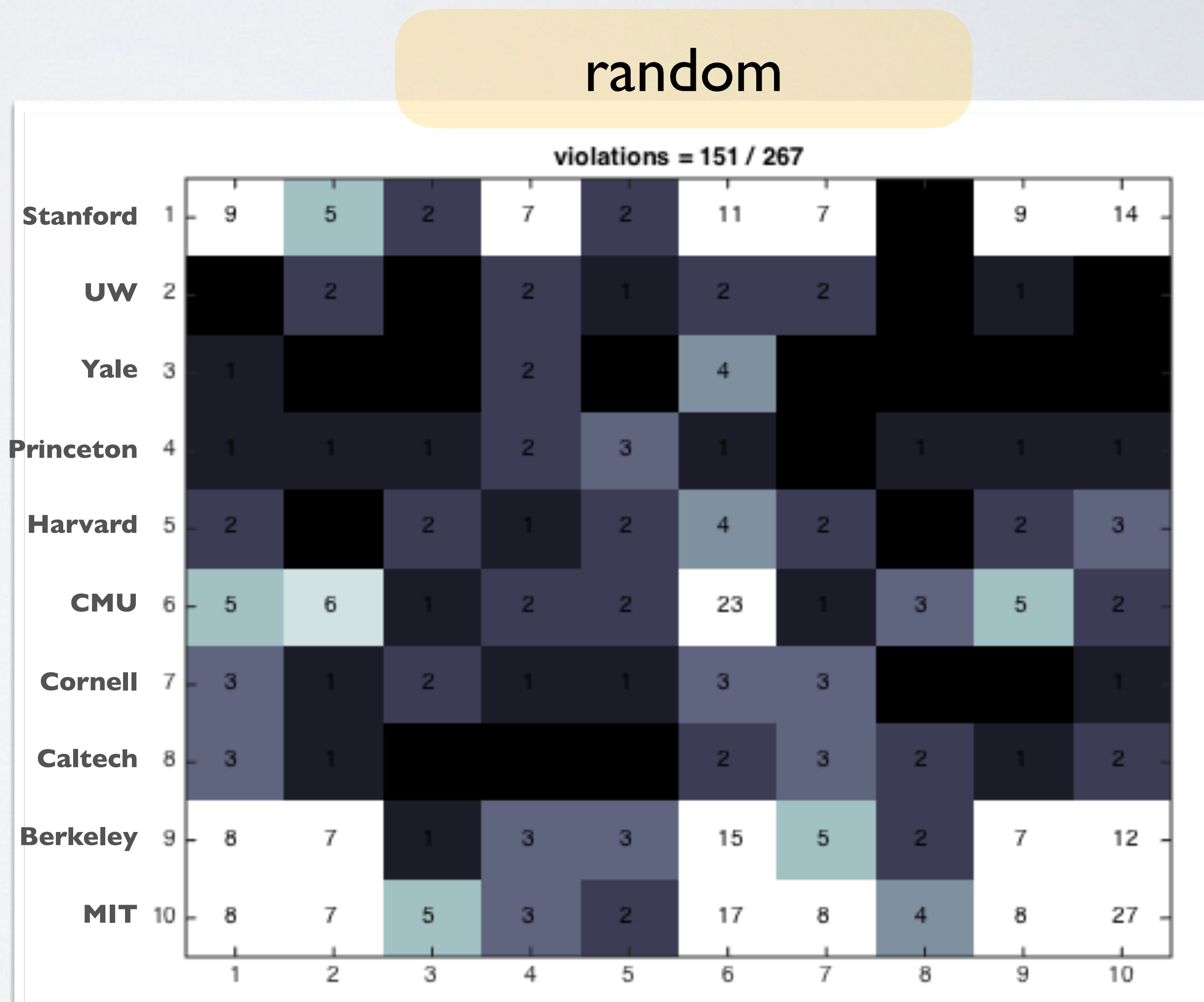
- select permutation (a ranking) π that minimizes the number of "rank violations" : edges (u, v) where $\pi_v < \pi_u$
- higher-ranked nodes have greater "placement power"
- equivalent to *minimum feedback arc set* problem (NP-hard)

[1] these "MVR"s have a deep history in social theory for extracting dominance or prestige hierarchies from data, especially in animal behavior
[2] *MFAS*: find the set of arcs of minimum cardinality whose removal converts a directed graph G into a directed acyclic graph
[3] there are many equivalent MVRs for our network. we sample these using a zero-temp MCMC, and average across them to obtain $\langle\pi\rangle$

simple zero-temperature MCMC sampler:

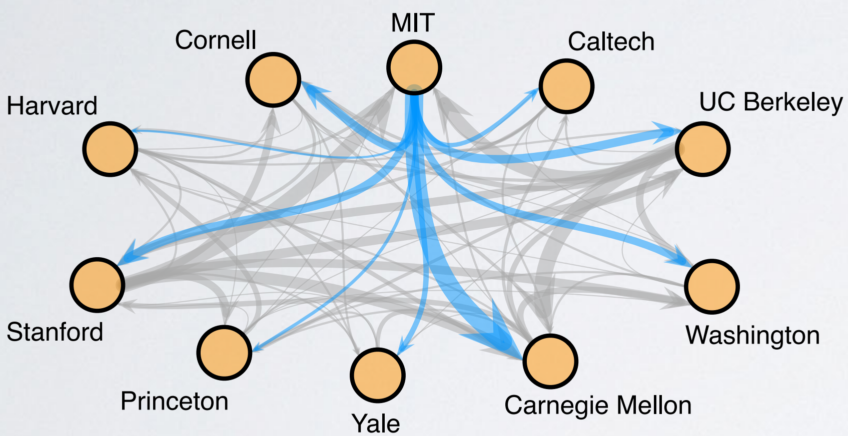
- given an ordering π with $\psi(\pi, A)$ rank violations on network A
- repeat *ad infinitum*: choose a pair (u, v), swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain π', compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$
- for instance:



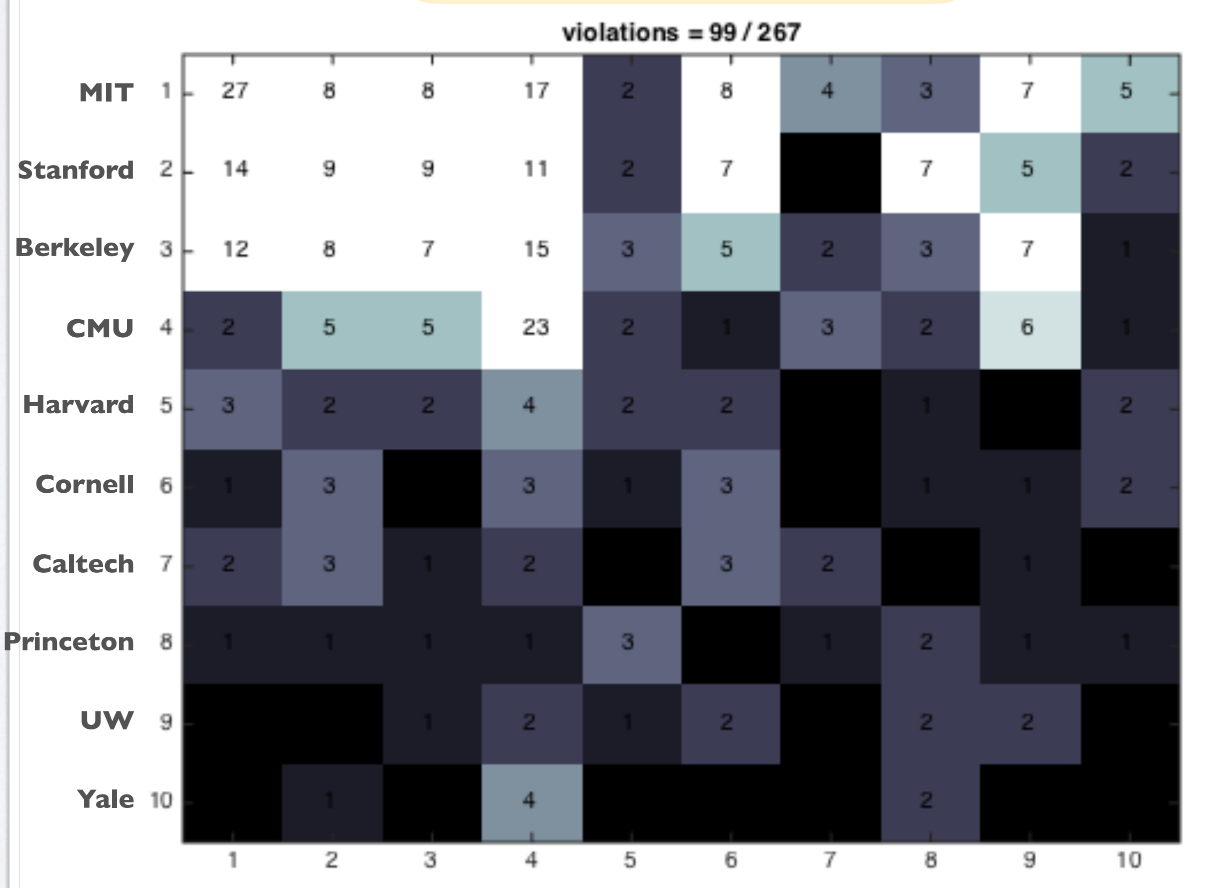


simple zero-temperature MCMC sampler:

- given an ordering π with $\psi(\pi, A)$ rank violations on network A
- repeat *ad infinitum*: choose a pair (u, v), swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain π', compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$
- for instance:

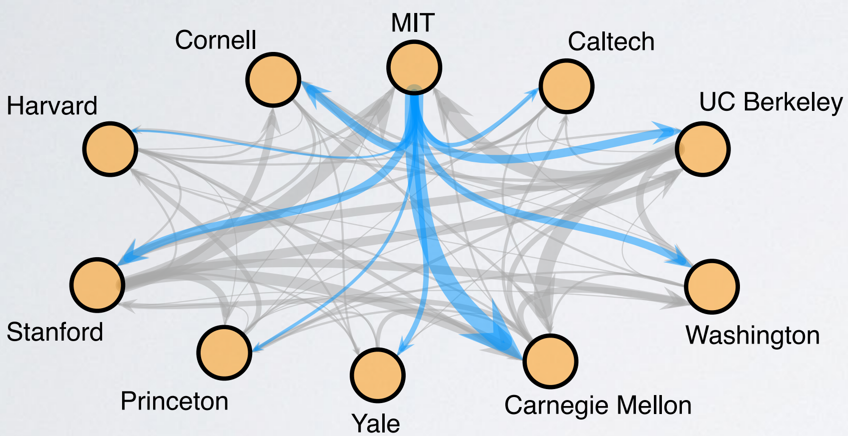


sort by degree

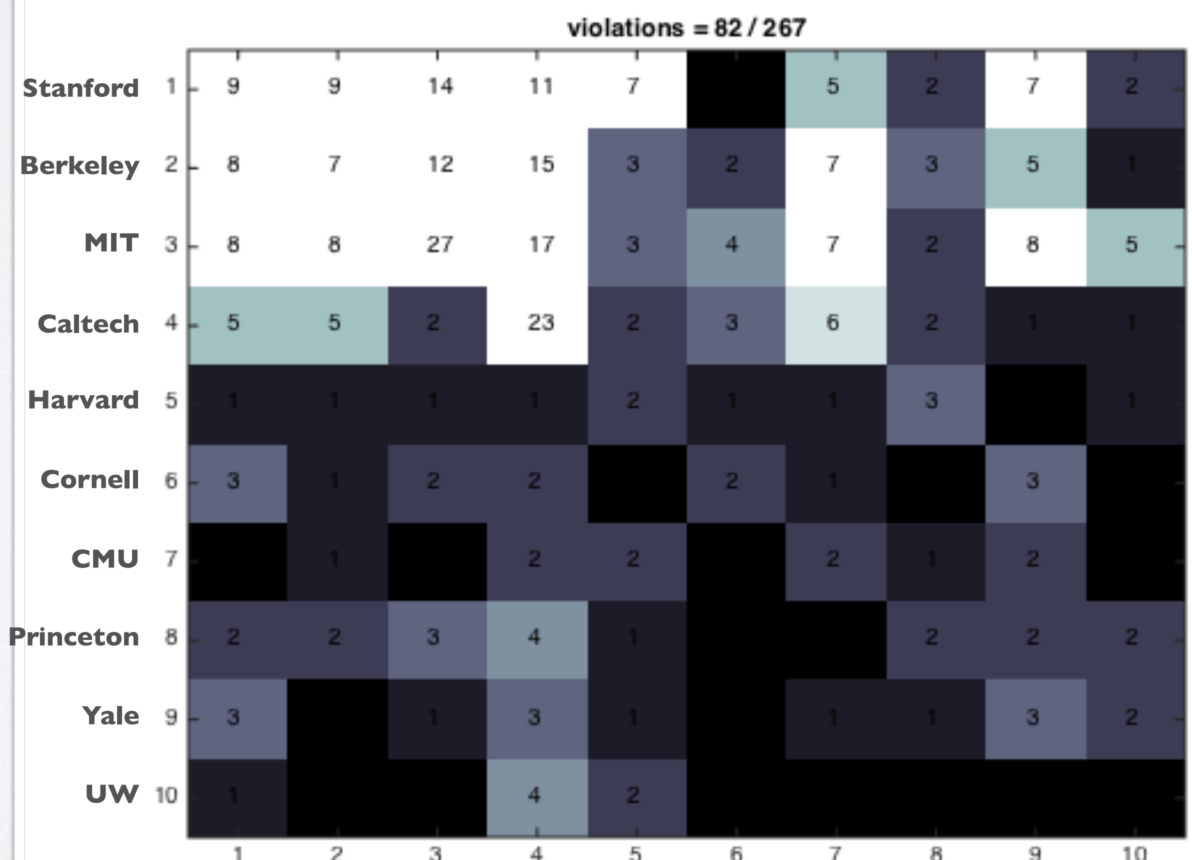


simple zero-temperature MCMC sampler:

- given an ordering π with $\psi(\pi, A)$ rank violations on network A
- repeat *ad infinitum*: choose a pair (u, v), swap their ranks $\pi_u \leftrightarrow \pi_v$ to obtain π', compute $\psi(\pi', A)$, accept change if $\psi(\pi', A) \geq \psi(\pi, A)$
- for instance:

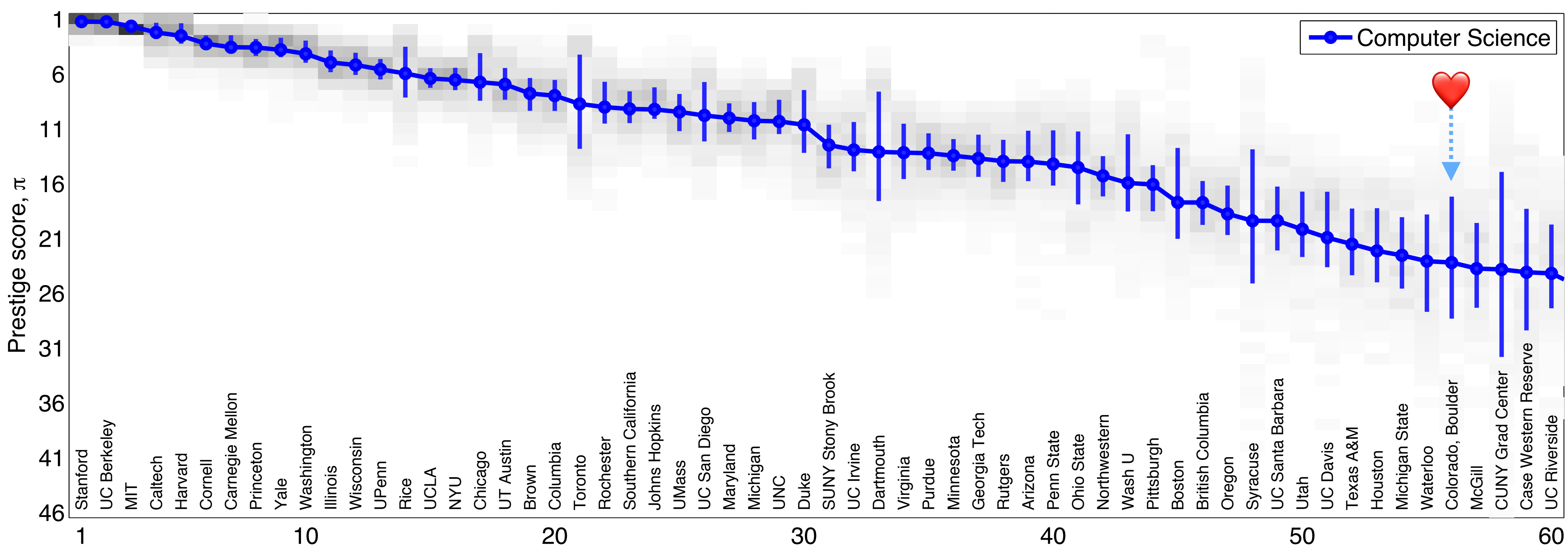


MVR



a prestige hierarchy

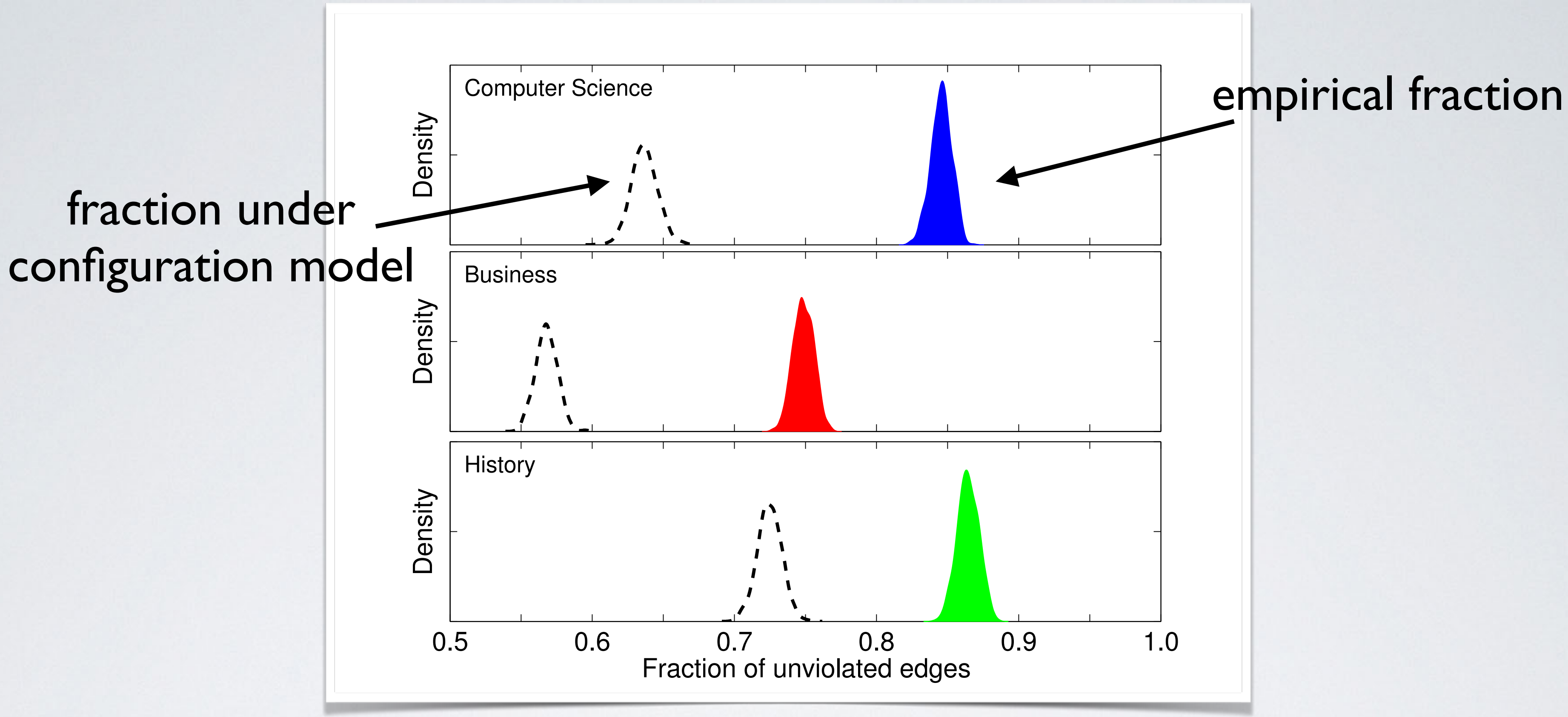
- what do these prestige hierarchies look like?
- what do they tell us about the structure of faculty hiring?
- what predicts placement?



prestige rankings correlate with USNews and NRC

- here, prestige π quantifies *placement power*
- uncertainty increases as prestige decreases
- similar results, but different orderings for Business and History

[1] uncertainties derived from the set of MVRs



degrees alone do not explain all of the hierarchy

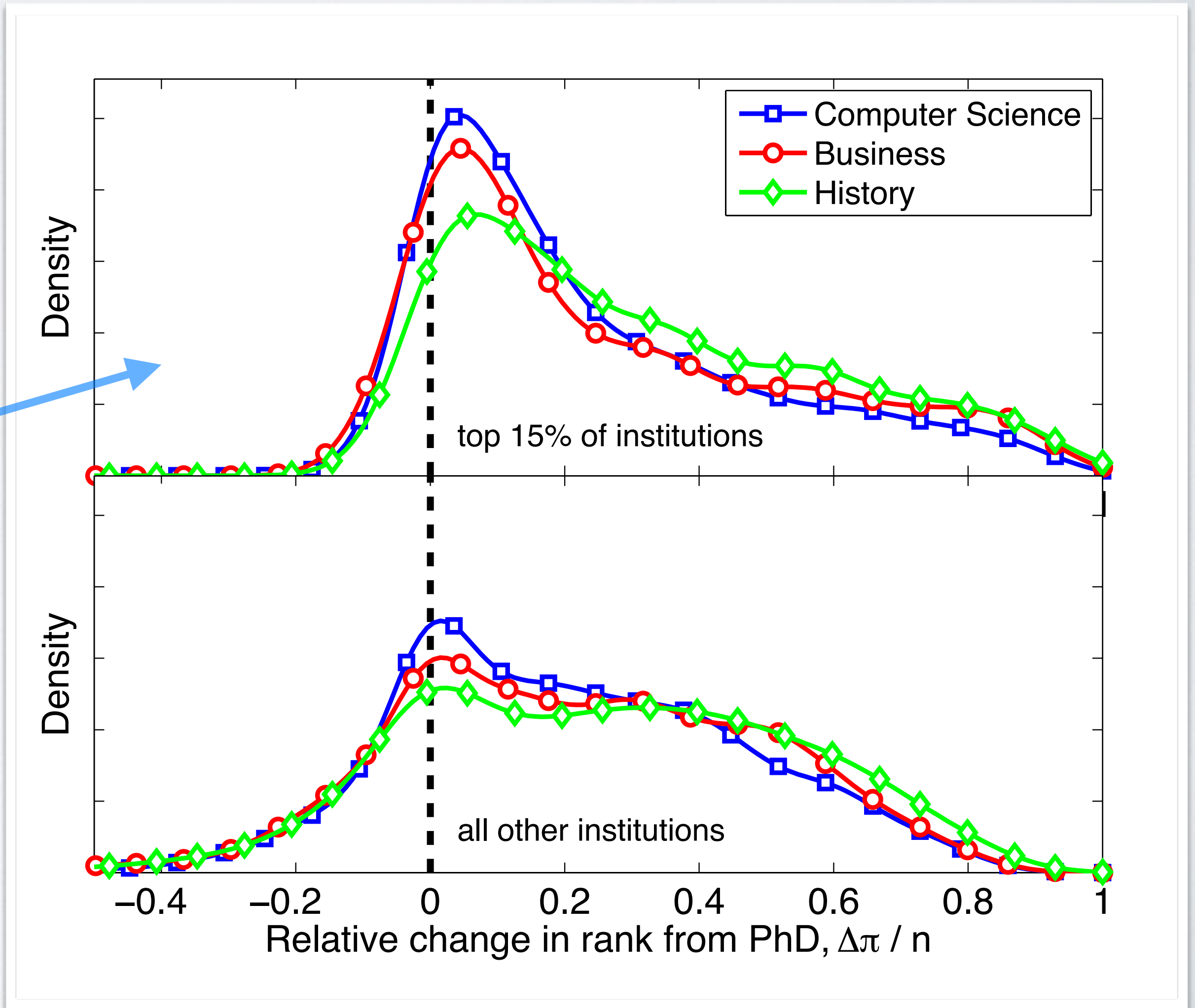
- use configuration model with *observed* degree sequence
- extract MVRs for random graphs & measure fraction of unviolated edges
- compare that null distribution to empirical fraction
- the gap is prestige effect beyond faculty production alone

most placements are down the hierarchy

most placements are down the hierarchy

- down : 88%, 86%, 91%
- up : 12%, 14%, 9%
- $\langle \Delta\pi \rangle$ = 47, 27, 42 steps down
- CS: top 15% of departments produce 68% of their own faculty

  and hire 7% from outside top 25% of departments



[1] order: CS, Business, History

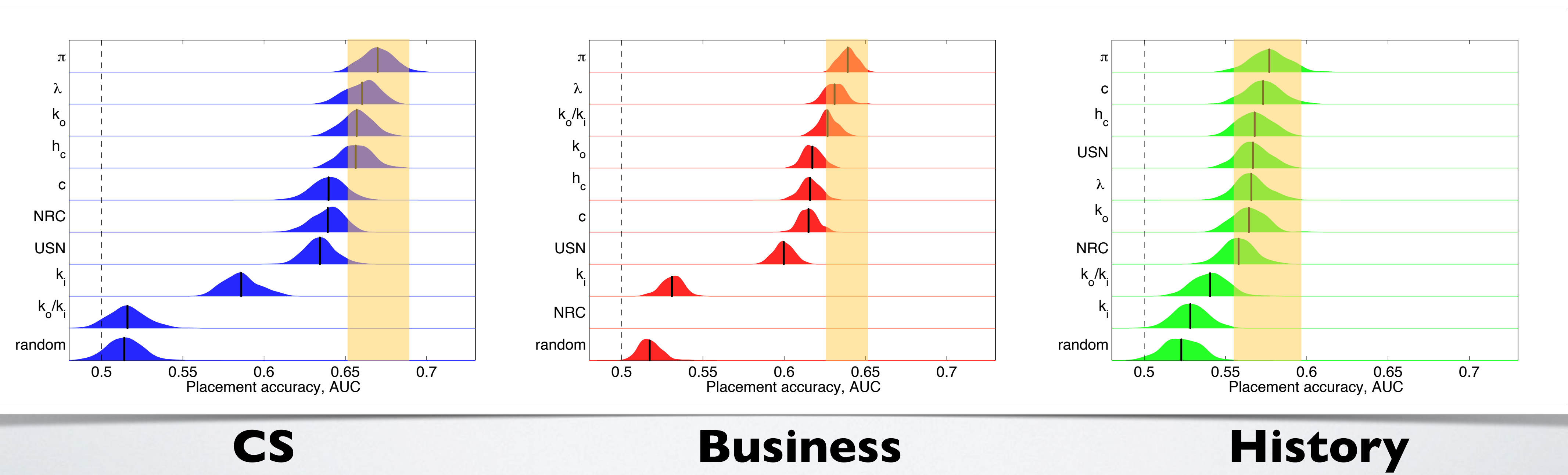
what predicts placement?

- compare 10 node-level features ("importances"):

  prestige
  US News rank
  NRC rank
  out-degree
  in-degree
  out/in degree
  eigenvector centrality
  harmonic centrality
  closeness centrality
  random

what predicts placement?

- prestige best *single* predictor in all 3 fields
- order of other features varies by field
- AUCs all below 0.67 = plenty of room for improvement

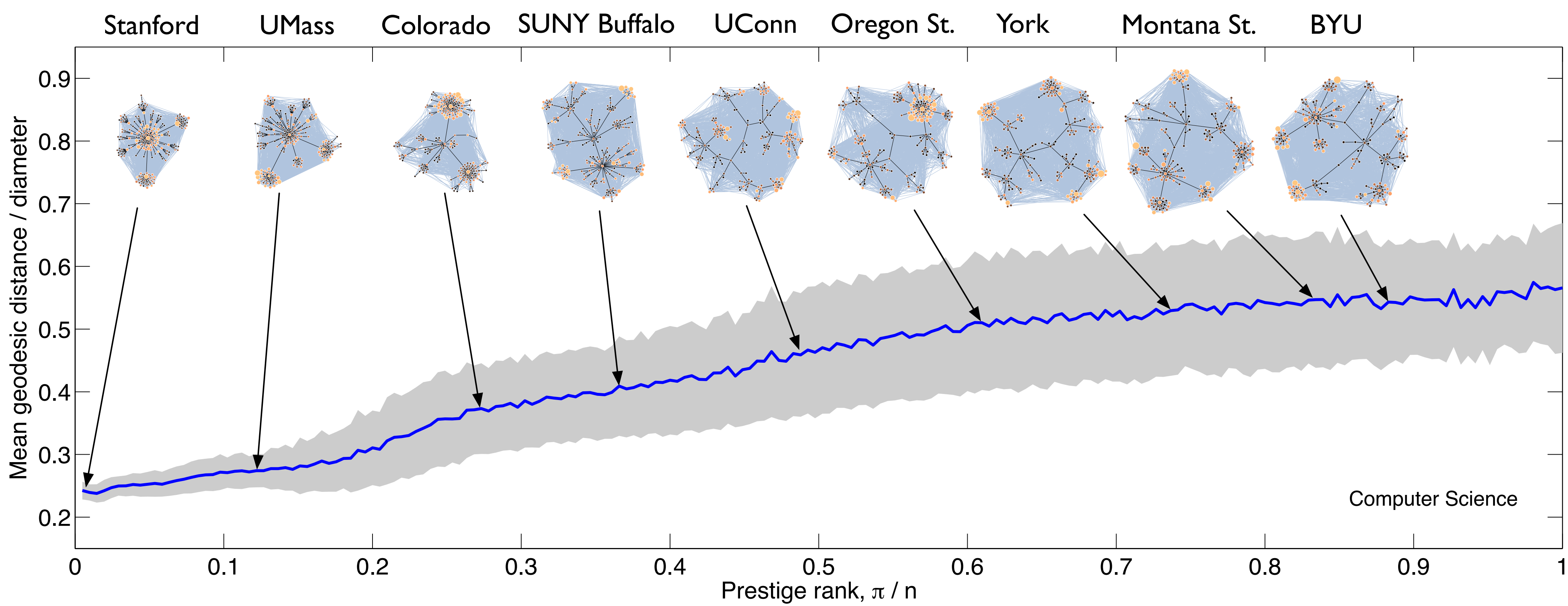


prestige correlates with network position

- core and periphery

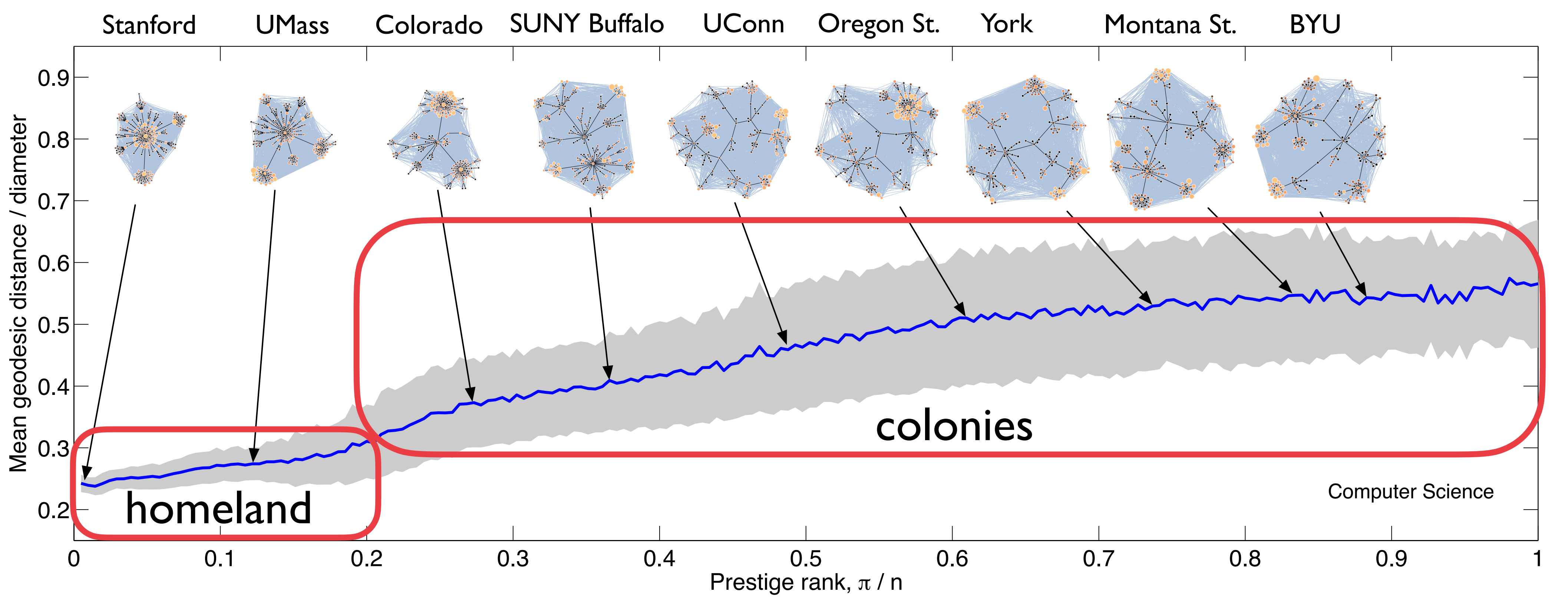
prestige correlates with network position

- core and periphery



prestige correlates with network position

- ~~core and periphery~~ homeland and colonies
- prestige is *influence*, via doctoral placement, over research agendas, research communities, and departmental norms across the discipline



inequality and prestige hierarchies

- prestige is *influence*, encoded in faculty hiring network
- faculty flow out of network core, into periphery ("the colonies")
- small fraction stay inside core
- only ~10% of hires flow "upstream"

future work

- how to measure cultural influence of core departments?
- what is different about "upstream" hires?
- what role for other inequalities : gender, ethnicity/race, SES, neighborhood effects, productivity, etc.?

conclusions and outlook

conclusions and outlook

networks are cool

[obviously, right?]

conclusions and outlook

networks are cool

[obviously, right?]

powerful window into *structure of complex systems*

[structure + dynamics = function]

network methods for exploiting rich data

[connectivity + node annotations + edge weights + temporal information | link or label prediction | etc.]

abundance of interesting science applications

[dynamics of social influence | emergence of hierarchy | online social network assembly | etc.]

but be careful: network methods have fundamental limits:

- networks are themselves a *model* of underlying system
- centralities and community detection typically *un*supervised
- some supervision & auxiliary data = *better* inferences / predictions
- formulate your mechanism in terms of nodes, edges, and attributes
- have fun!

more material here

University of Colorado **Boulder**

Network Analysis and Modeling

Instructor: Aaron Clauset or Daniel 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/