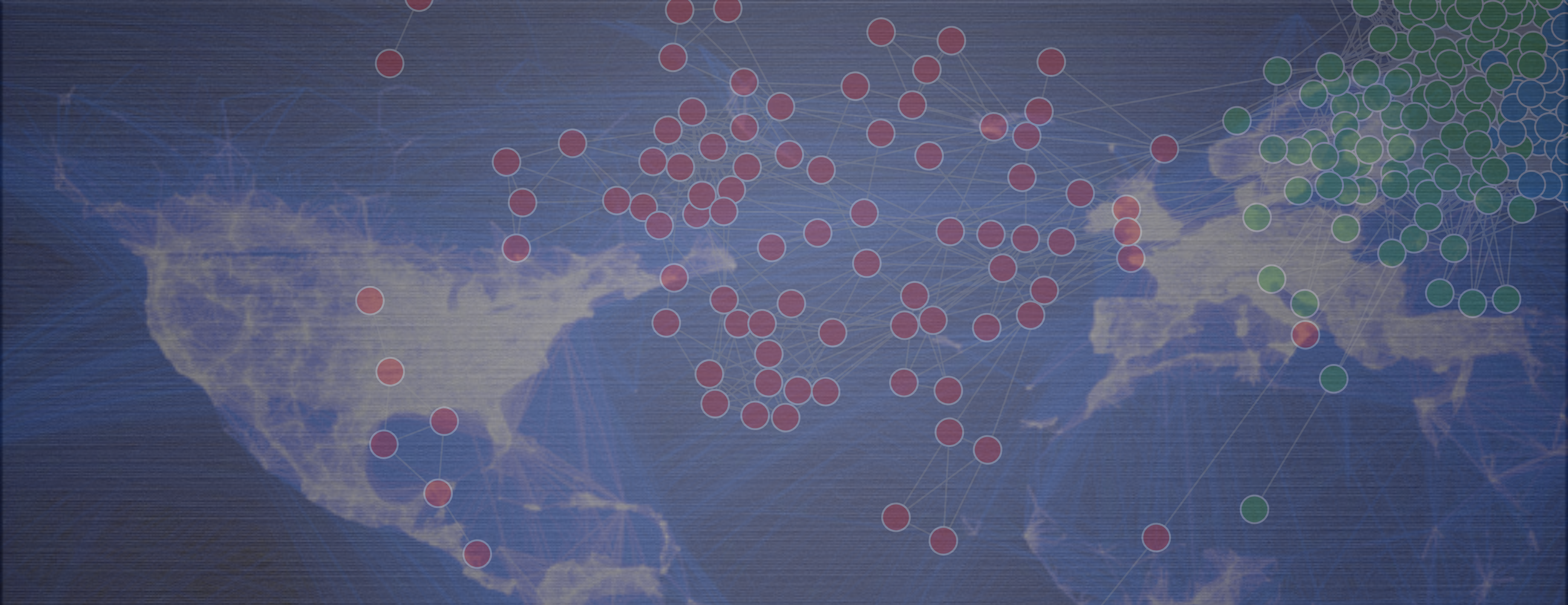


Five Lectures on Networks

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

lecture 2: describing network structure & its impact on network flows

University of Colorado **Boulder**

Network Analysis and Modeling

Instructor: Aaron Clauset

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

Data sets

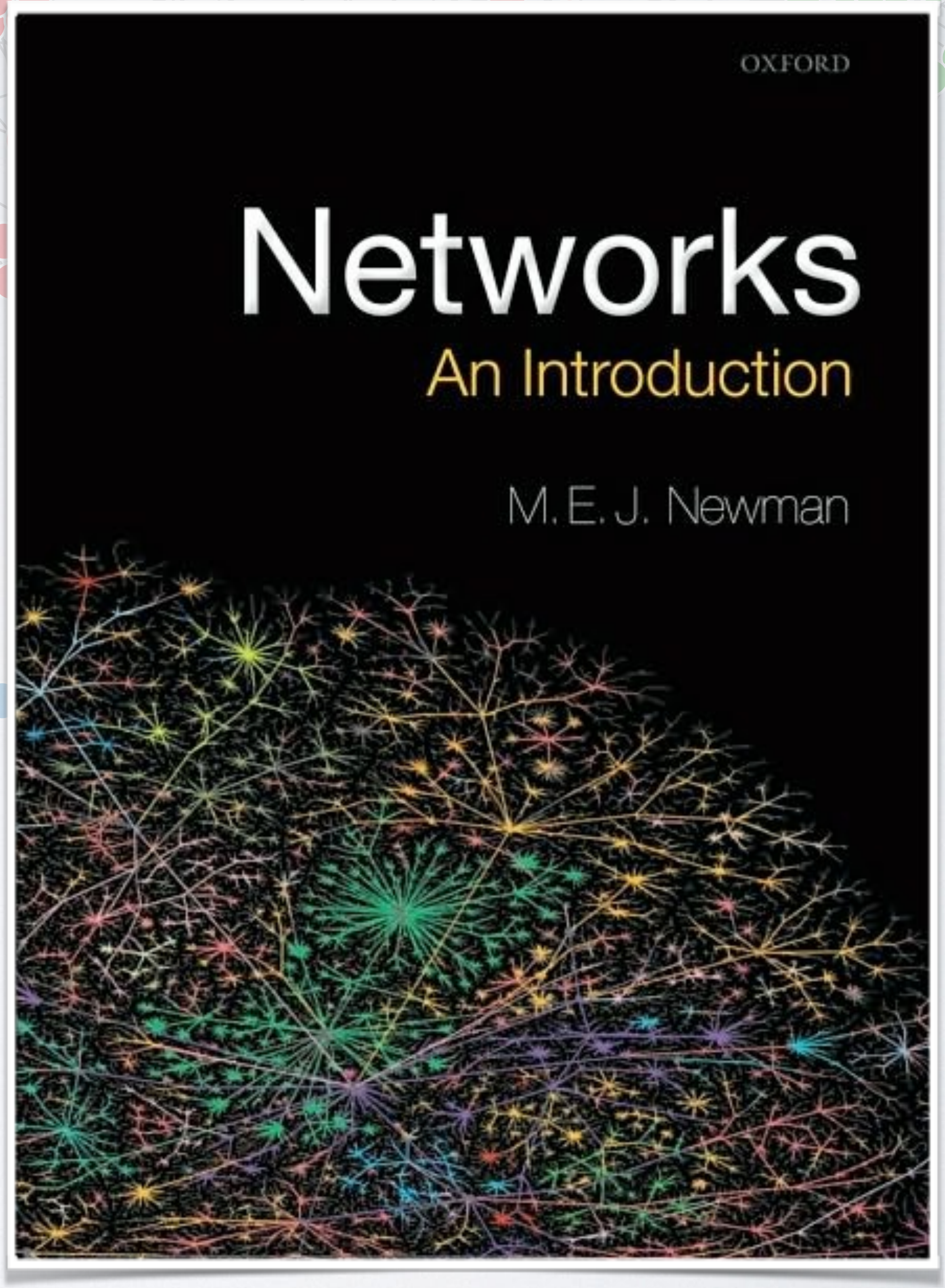
Mark Newman's network data sets
Stanford Network Analysis Project
Carnegie Mellon CASOS data sets
NCEAS food web data sets
UCI NET data sets
Pajek data sets
Linkgroup's list of network data sets
Barabasi lab data sets
Jake Hofman's online network data sets
Alex Arenas's data sets

Mark Newman

Professor of Physics
University of Michigan

External Faculty
Santa Fe Institute

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



coursera

UNIVERSITY OF MICHIGAN

Social Network Analysis

Lada Adamic

This course will use social network analysis, both its theory and computational tools, to make sense of the social and information networks that have been fueled and rendered accessible by the internet.

Workload: 5-7 hours/week (8-10 if completing additional programming exercises)

Lada Adamic
University of Michigan

https://www.coursera.org/course/sna

coursera

Stanford

Social and Economic Networks: Models and Analysis

Matthew O. Jackson

Learn how to model social and economic networks and their impact on human behavior. How do networks form, why do they exhibit certain patterns, and how does their structure impact diffusion, learning, and other behaviors? We will bring together models and techniques from economics, sociology, math, physics, statistics and computer science to answer these questions.

Workload: 3-6 hours/week

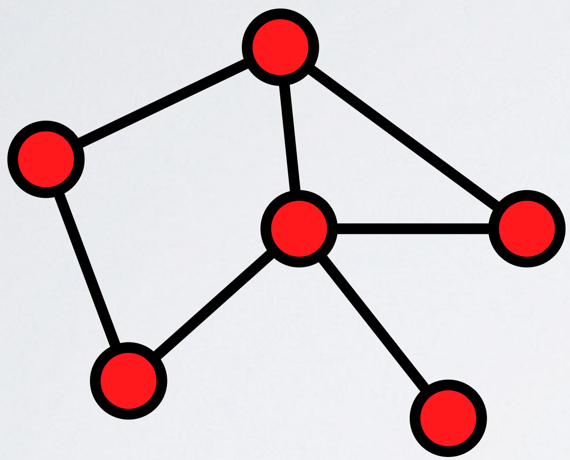
Matthew O. Jackson
Stanford University

https://www.coursera.org/course/networksonline

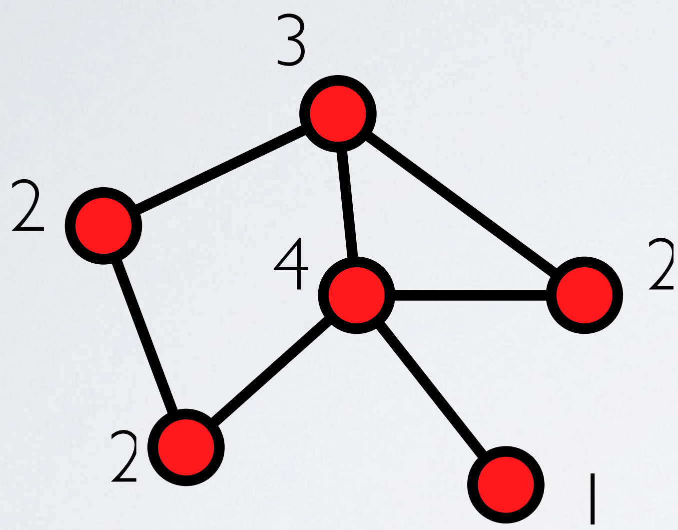
1. defining a network
2. **describing a network**
3. null models for networks
4. statistical inference

describing networks

degree



describing networks

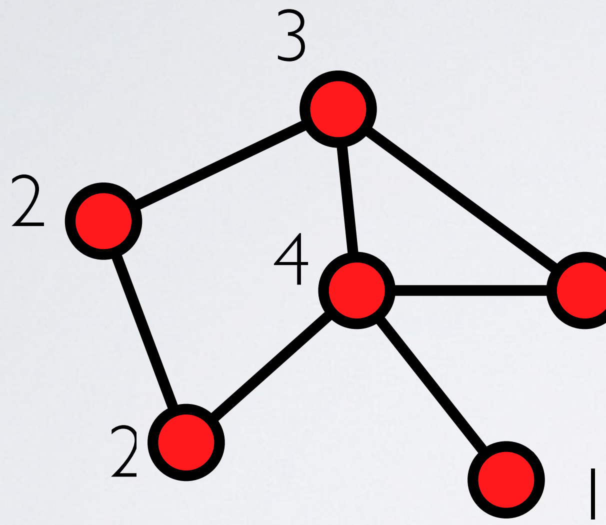


degree:

number of connections k

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

describing networks

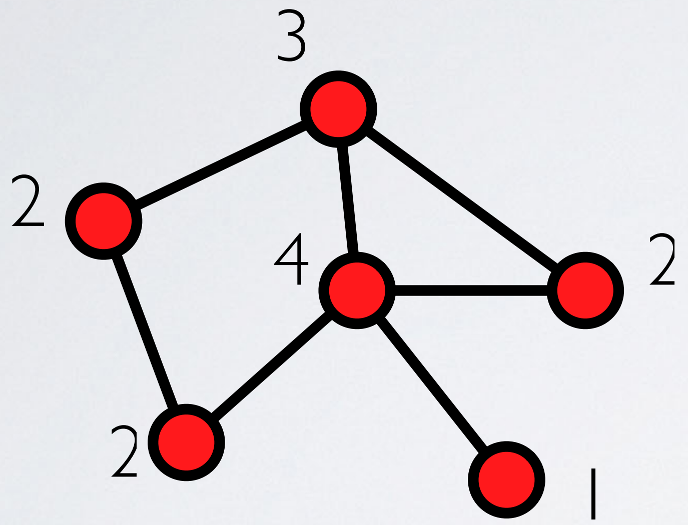


degree:
number of connections k

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

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:

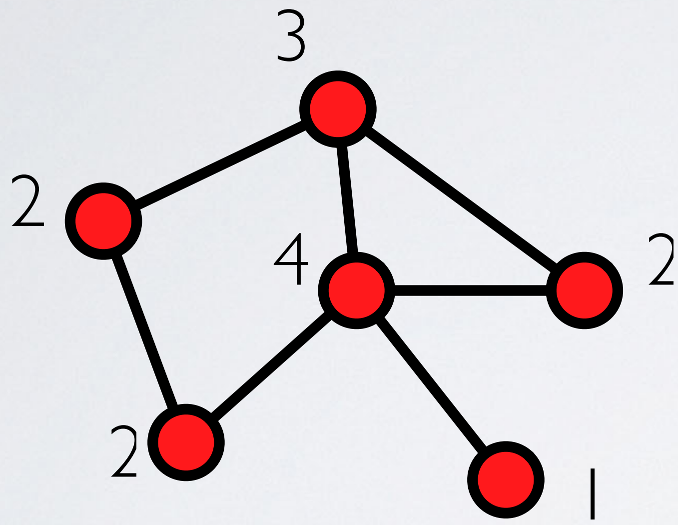
number of connections k

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

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



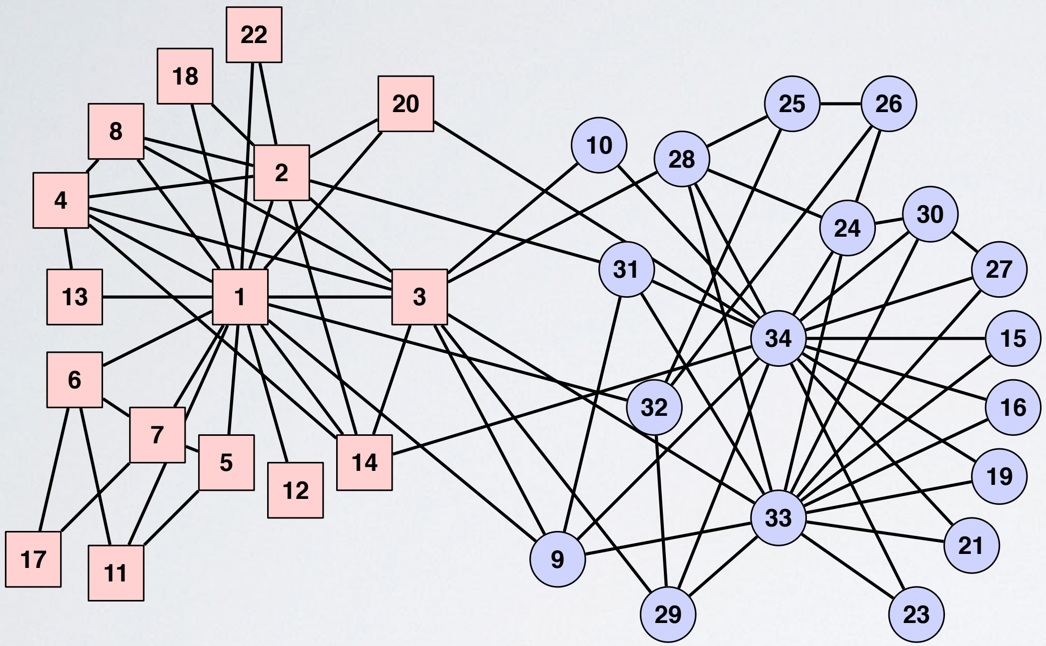
degree:
number of connections k

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

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

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

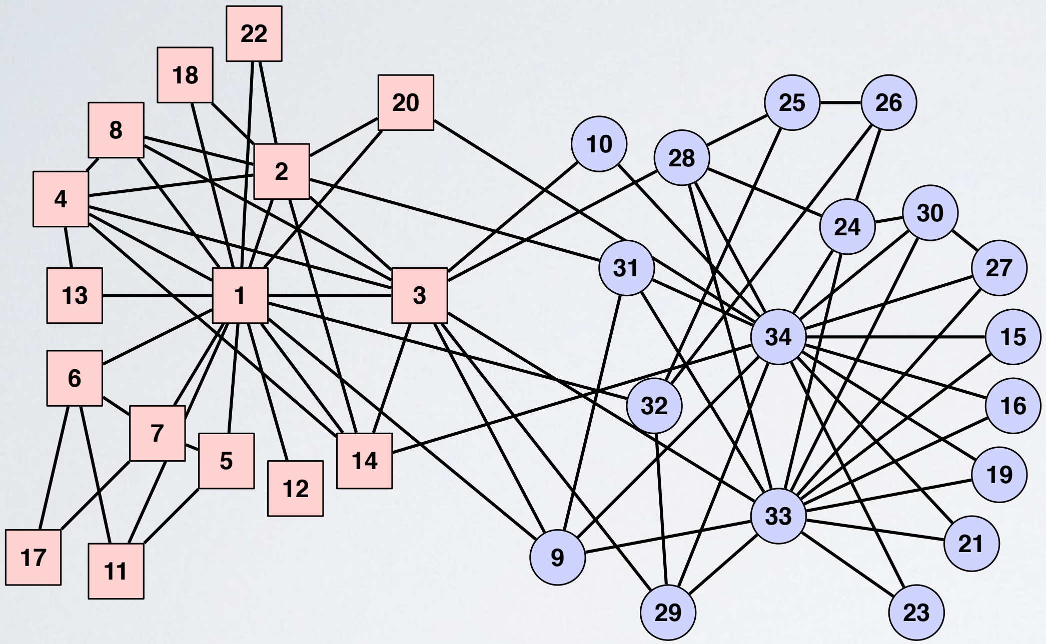
degree distributions



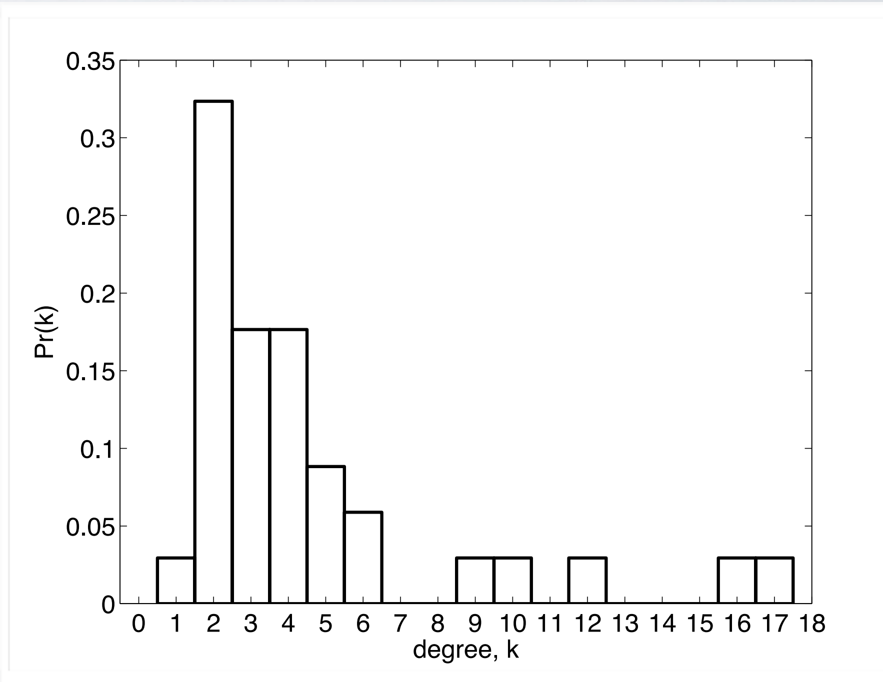
Zachary karate club*

* Zachary, *J. Anthropological Research* **33**, 452–473 (1977).

degree distributions

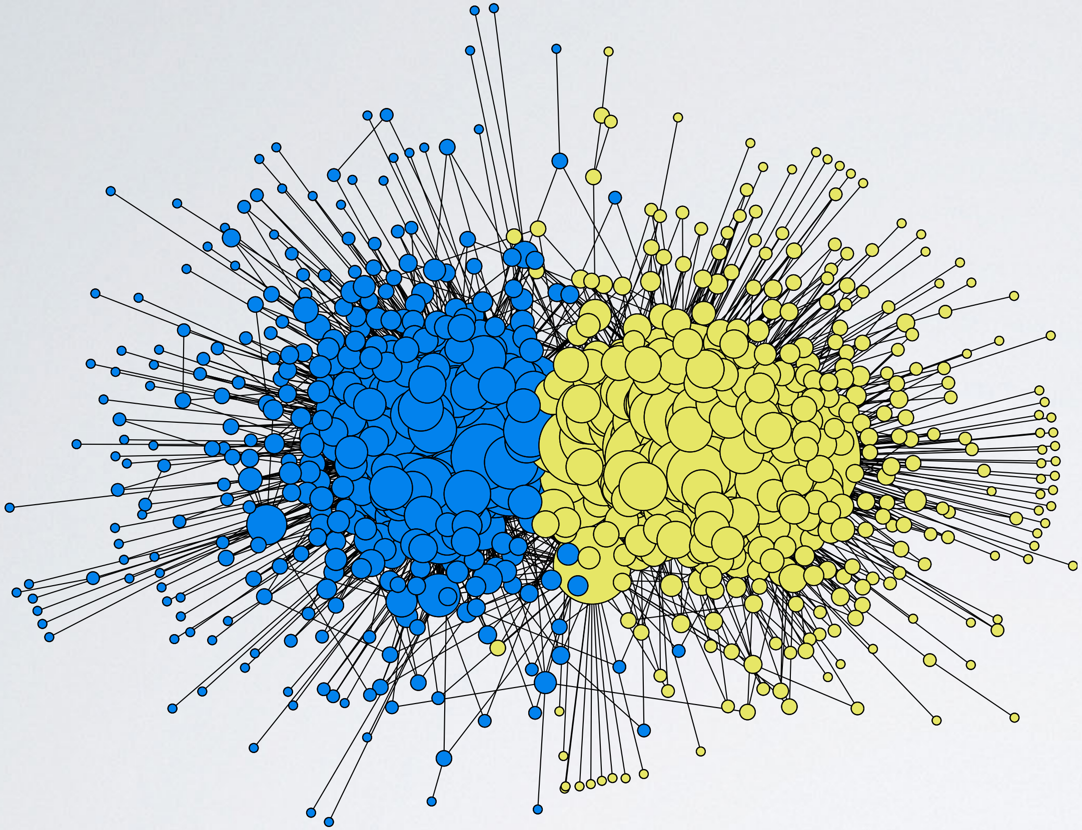


Zachary karate club*



* Zachary, *J. Anthropological Research* **33**, 452–473 (1977).

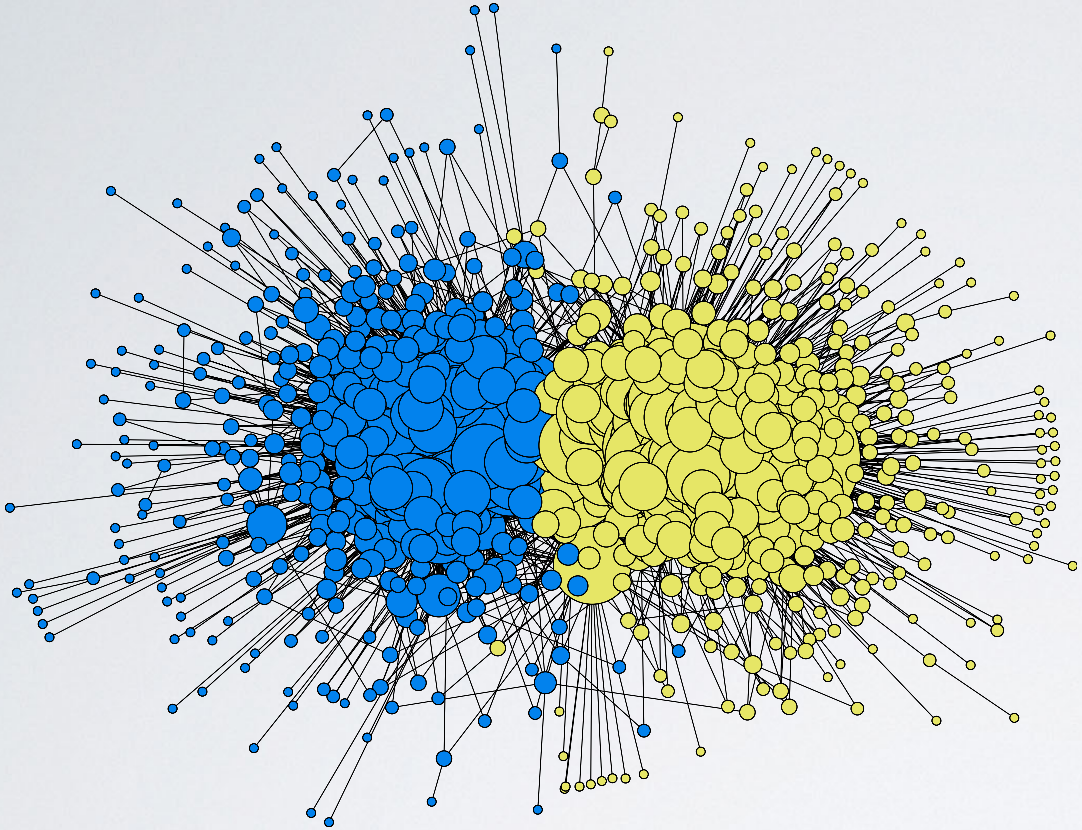
degree distributions



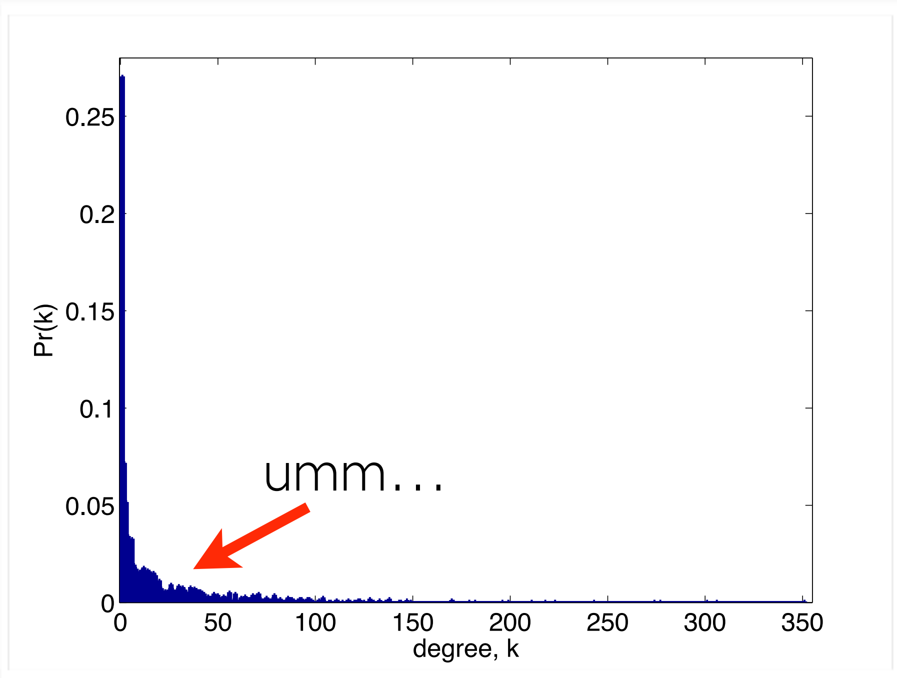
political blogs*

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

degree distributions

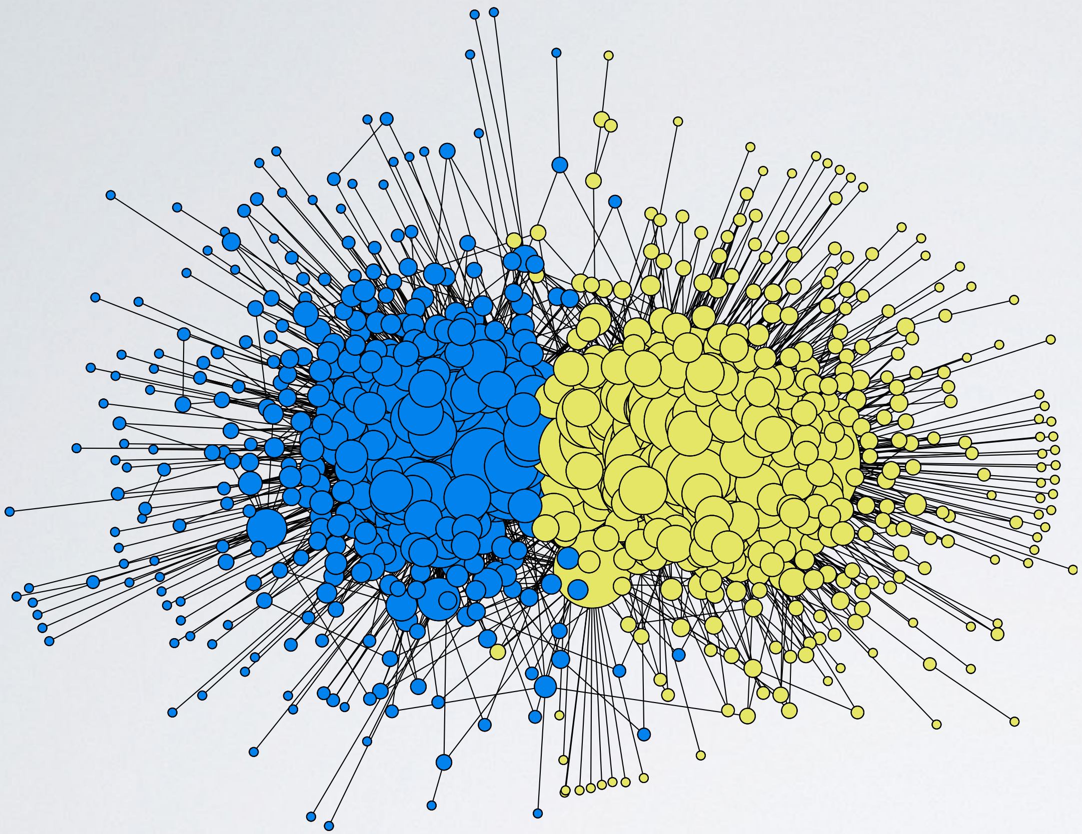


political blogs*

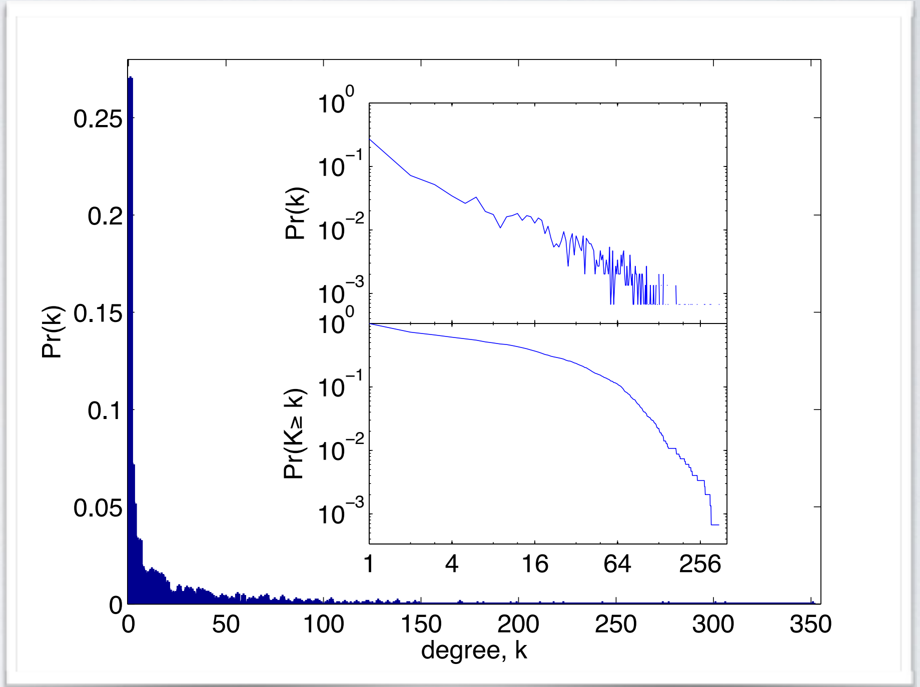


* Adamic and Glance, *WWW Workshop on the Weblogging Ecosystem* (2005).

degree distributions



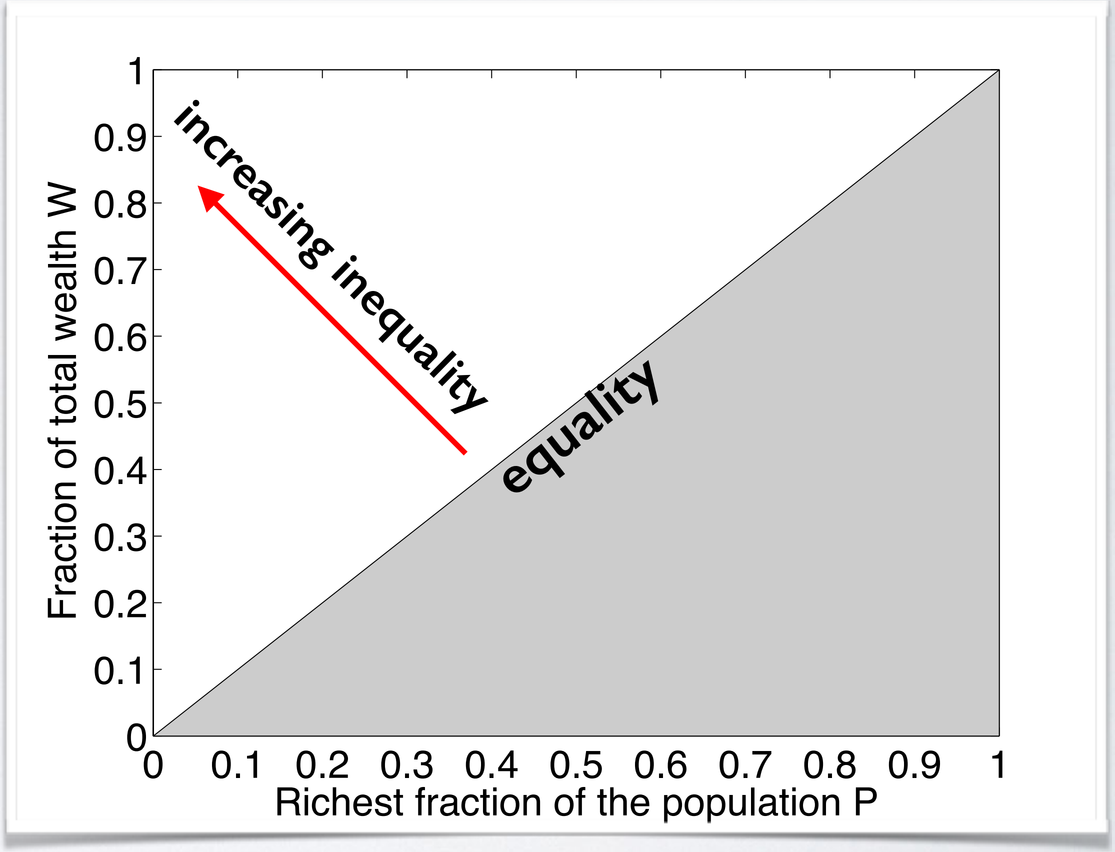
political blogs*



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

degree "wealth"

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



Lorenz curve

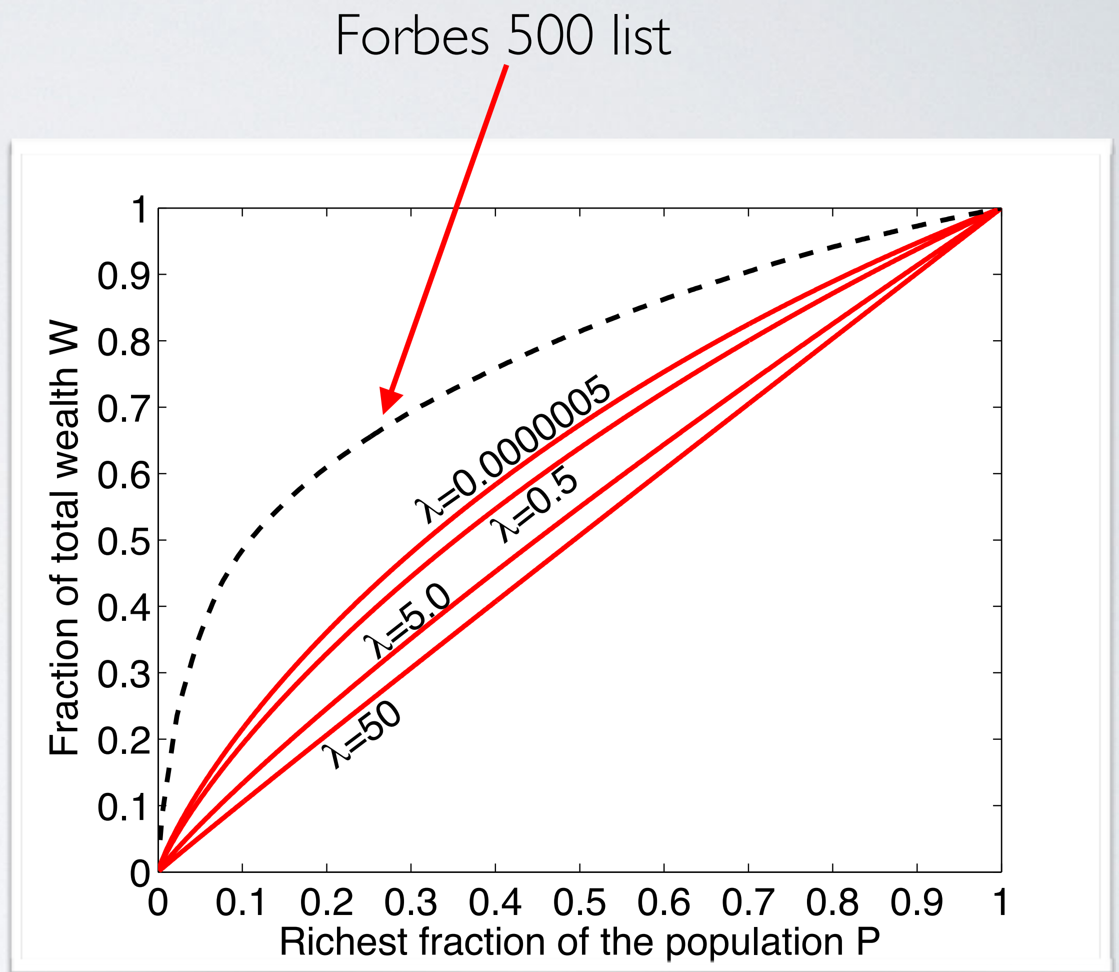
degree distributions

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

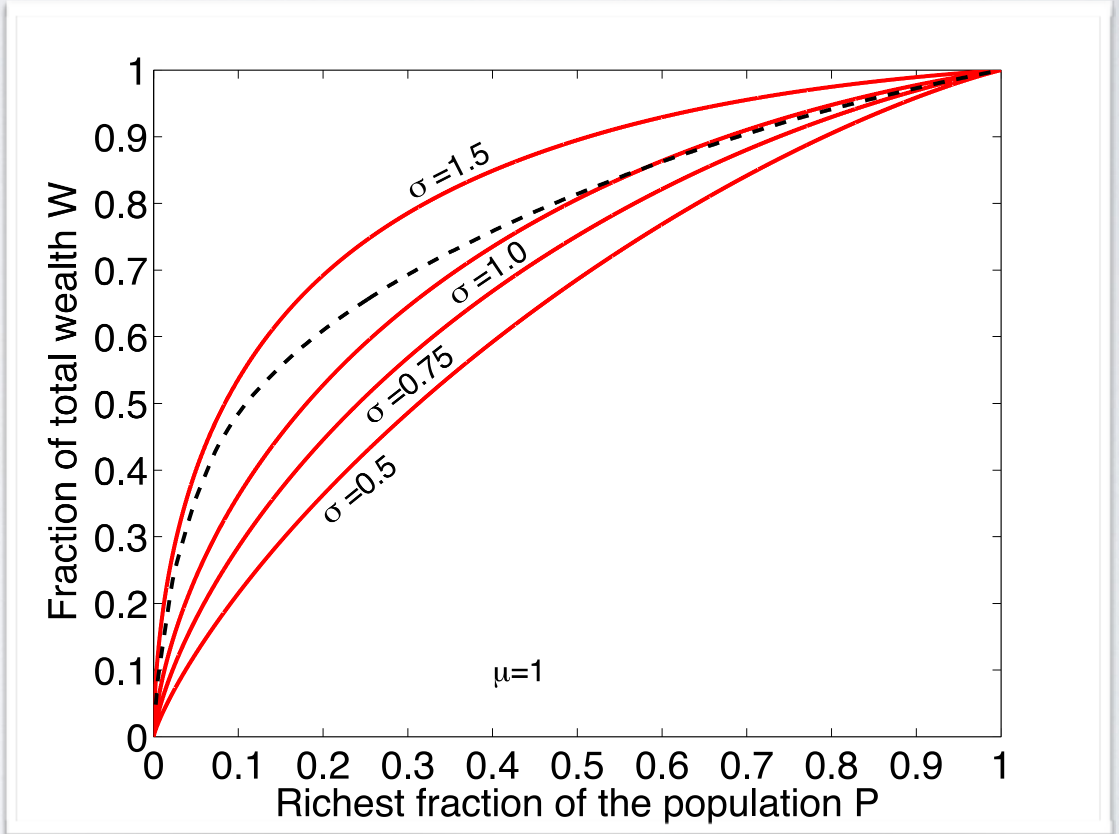
degree distributions

degree "wealth"

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

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

log-normal distribution



Lorenz curve

degree distributions

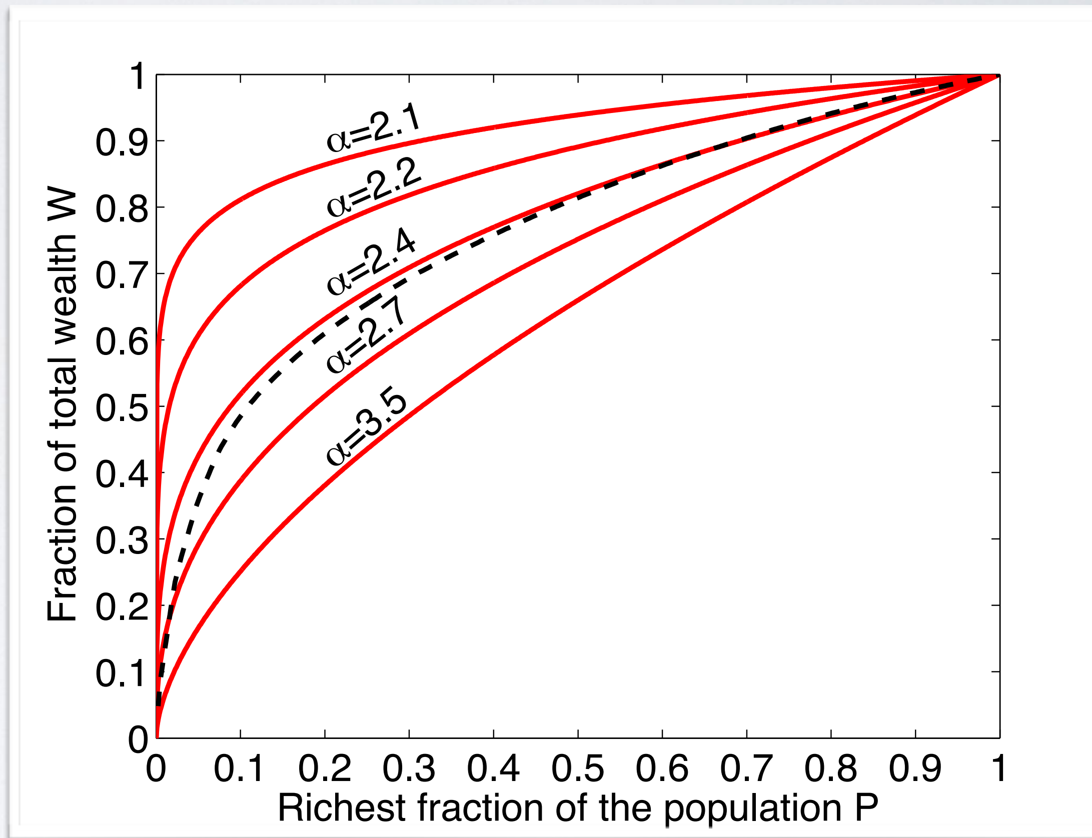
degree "wealth"

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

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

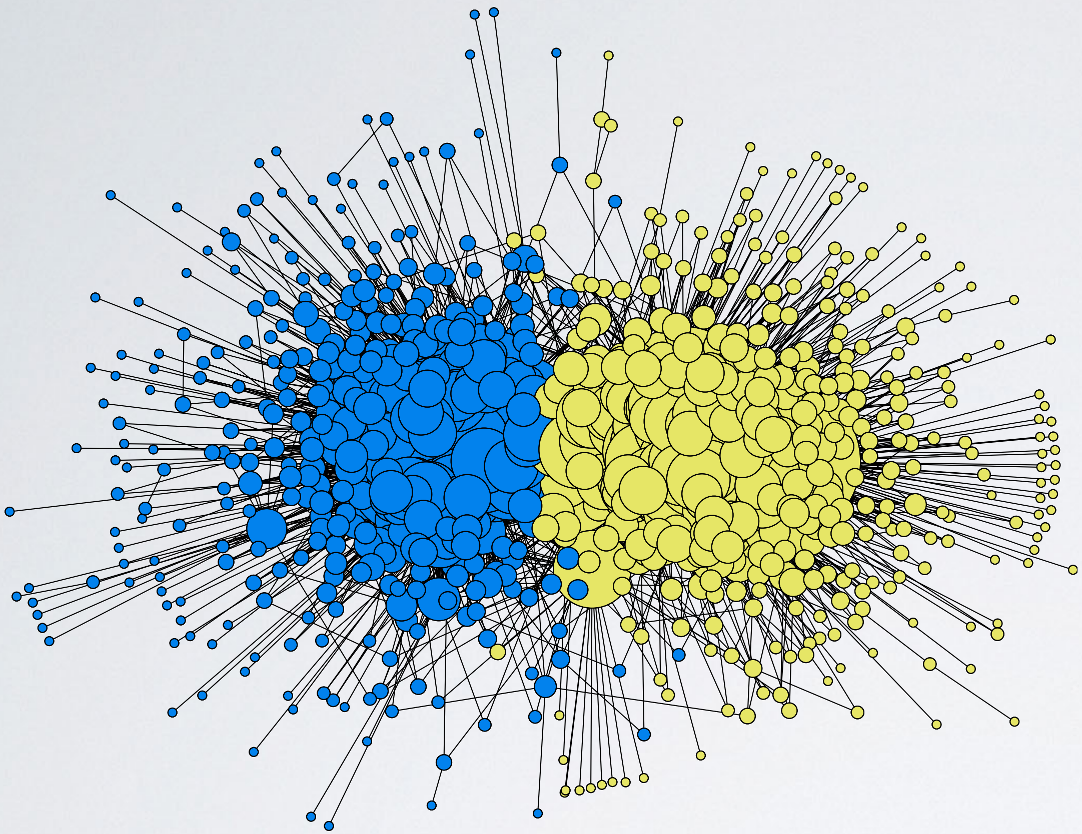
power-law distribution

80/20 rule

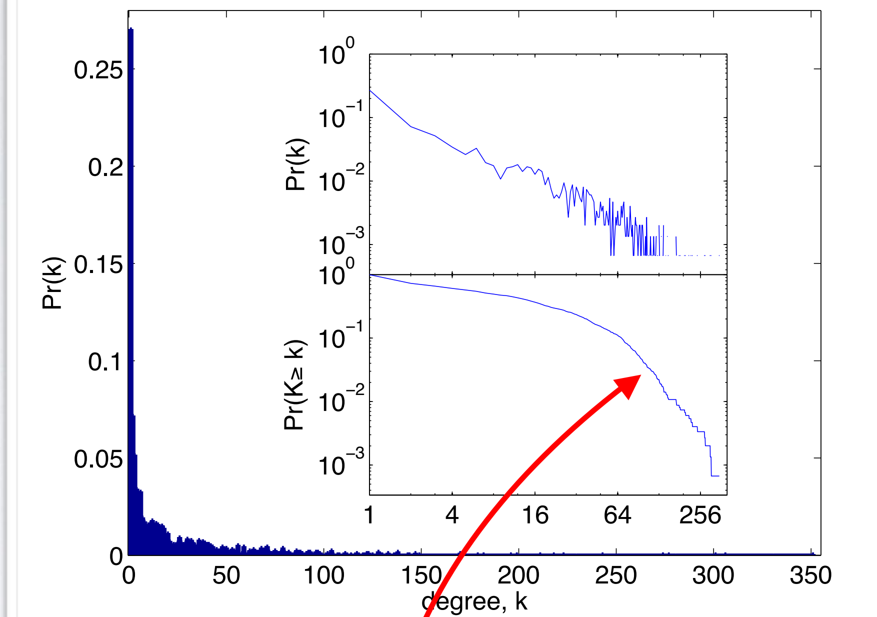


Lorenz curve

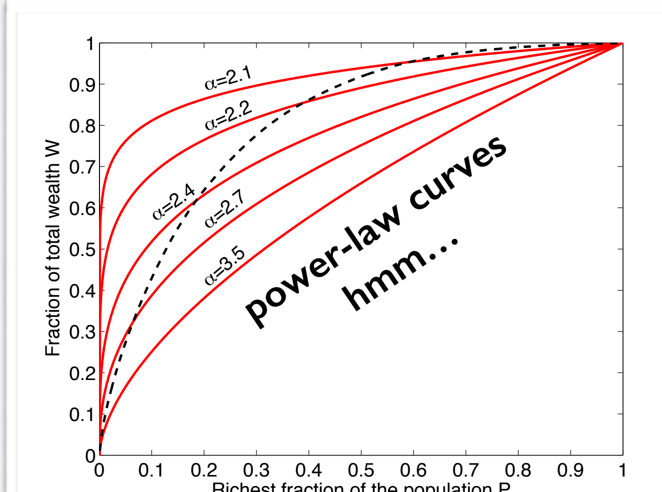
degree distributions



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

* 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.

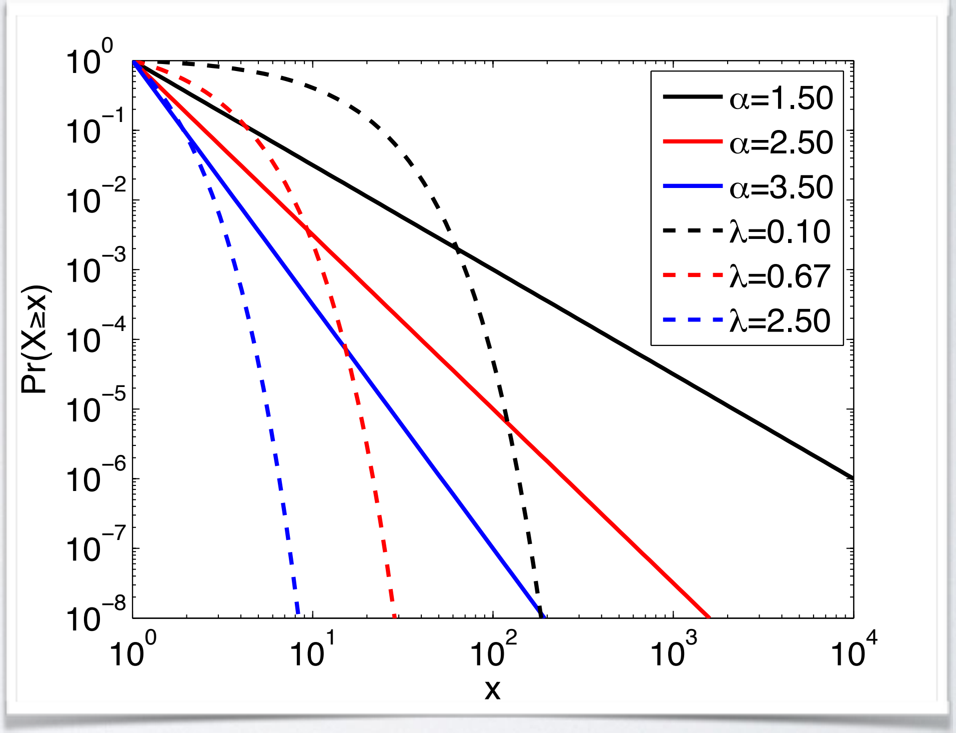
Newman, *Contemporary Physics* **46**, 323–351 (2005)
Clauset et al. *SIAM Review* **51**, 661–703 (2009)

power-law distributions

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

- high-variance

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



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

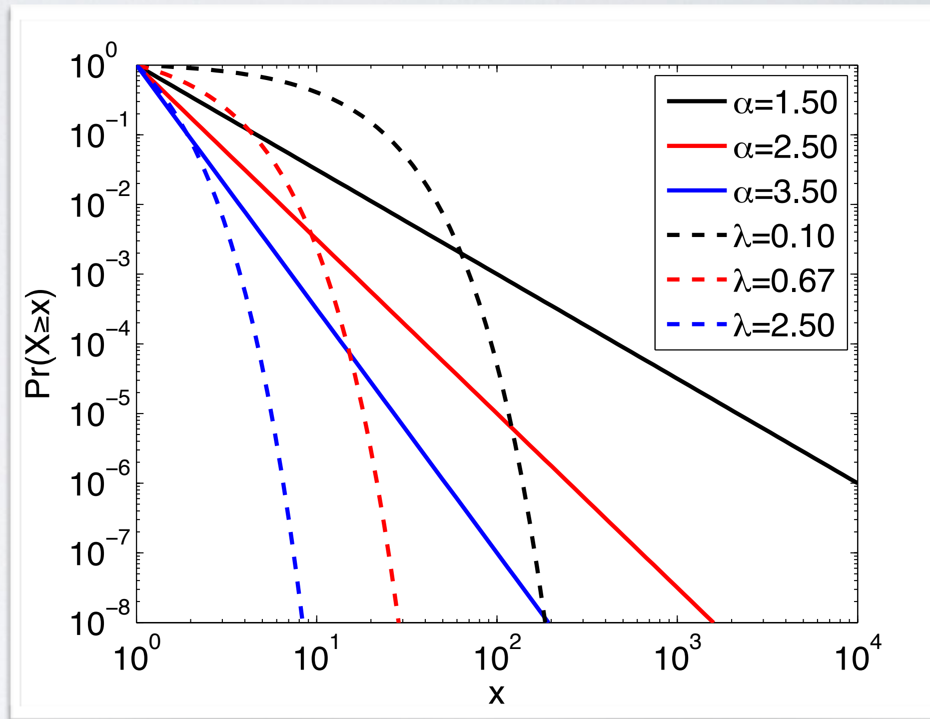
power-law distributions

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

- 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 α 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 α 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

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

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 α 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 with standard error

$$\hat{\alpha} = 1 + n \Bigg/ \sum_{i=1}^{n} \ln\left(\frac{k_i}{\boxed{k_{\min}}}\right) \qquad \hat{\sigma} = \frac{\hat{\alpha}-1}{\sqrt{n}} + O(1/n)$$

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*

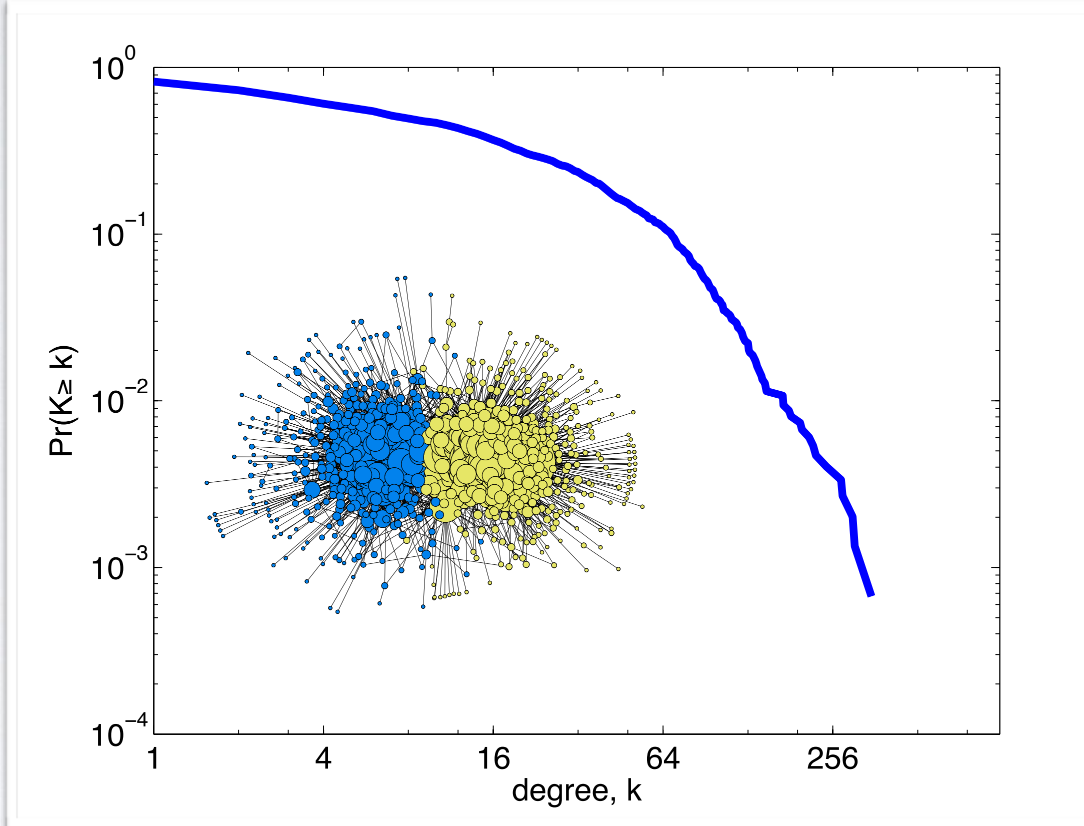
Newman, *Contemporary Physics* **46**, 323–351 (2005)
Clauset et al. *SIAM Review* **51**, 661–703 (2009)

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

degree distributions

fun facts:

- 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****

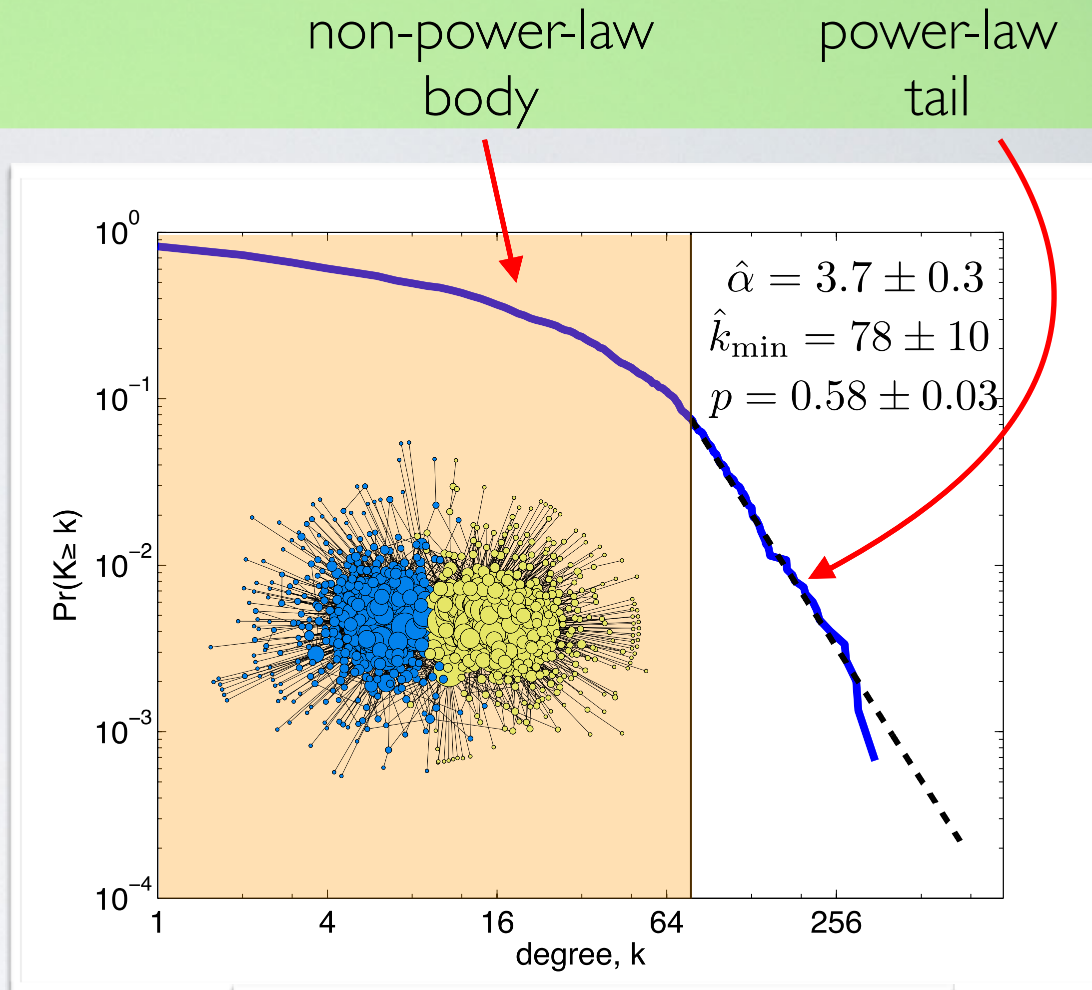


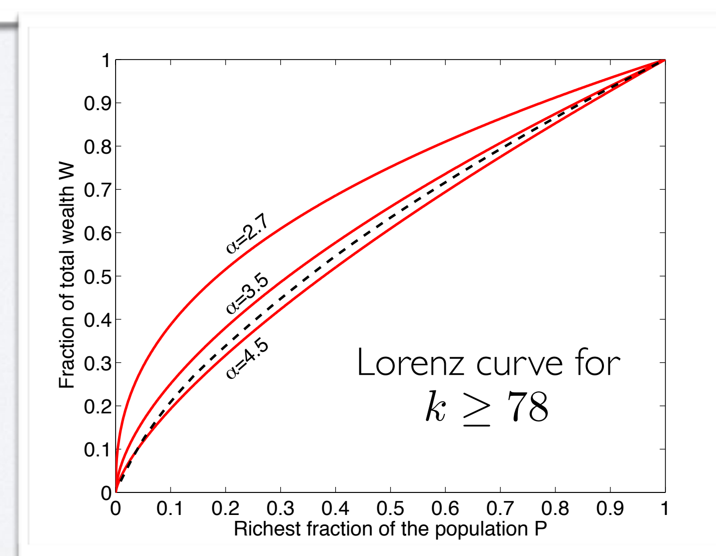
* Clauset et al. *SIAM Review* **51**, 661–703 (2009)

degree distributions

fun facts:

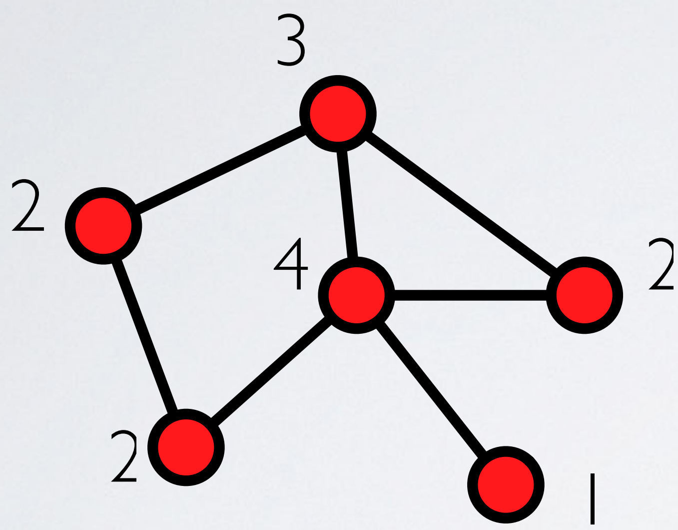
- nearly all real networks exhibit a ***heavy-tailed degree distribution***
- ***very few*** networks exhibit perfect power-law degree distributions
- ***some*** distributions exhibit power-law tails
- power laws are cool!
  but knowing one from garbage ***requires statistics****





* Clauset et al. *SIAM Review* **51**, 661–703 (2009)

describing networks



degree:

number of connections k

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

when does node degree matter?

network degrees

spreading processes on networks

biological (diseases)

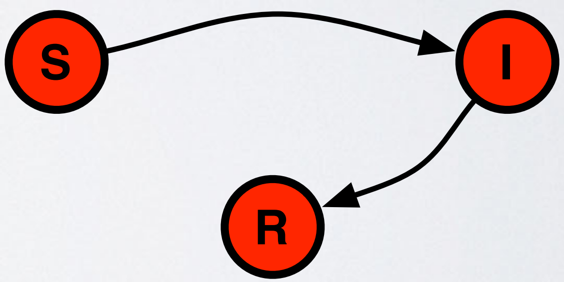
- SIS and SIR models

social (information)

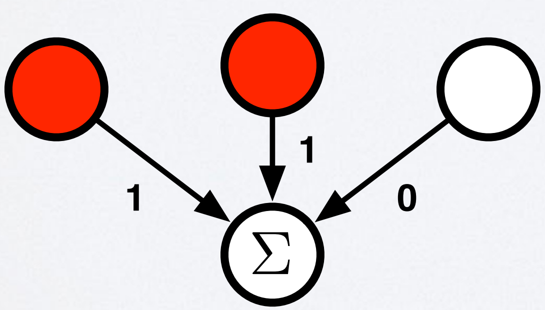
- SIS, SIR models
- threshold models

susceptible-infected-susceptible

susceptible-infected-recovered



threshold



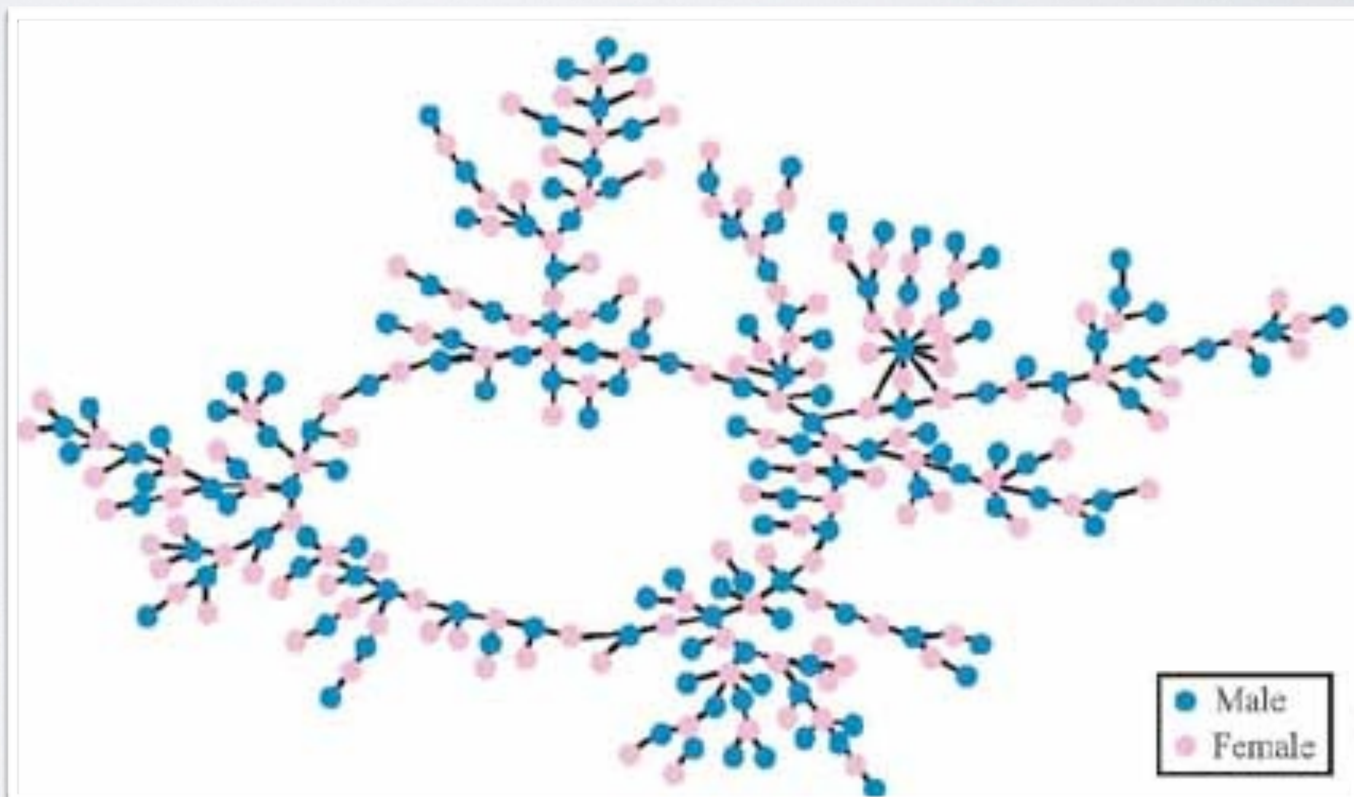
network degrees

Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks

2004

Peter S. Bearman — *Columbia University*
James Moody — *Ohio State University*
Katherine Stovel — *University of Washington*

- relationship network in "Jefferson High"
- this subgraph is 52% of school
- who are most important disease spreaders?

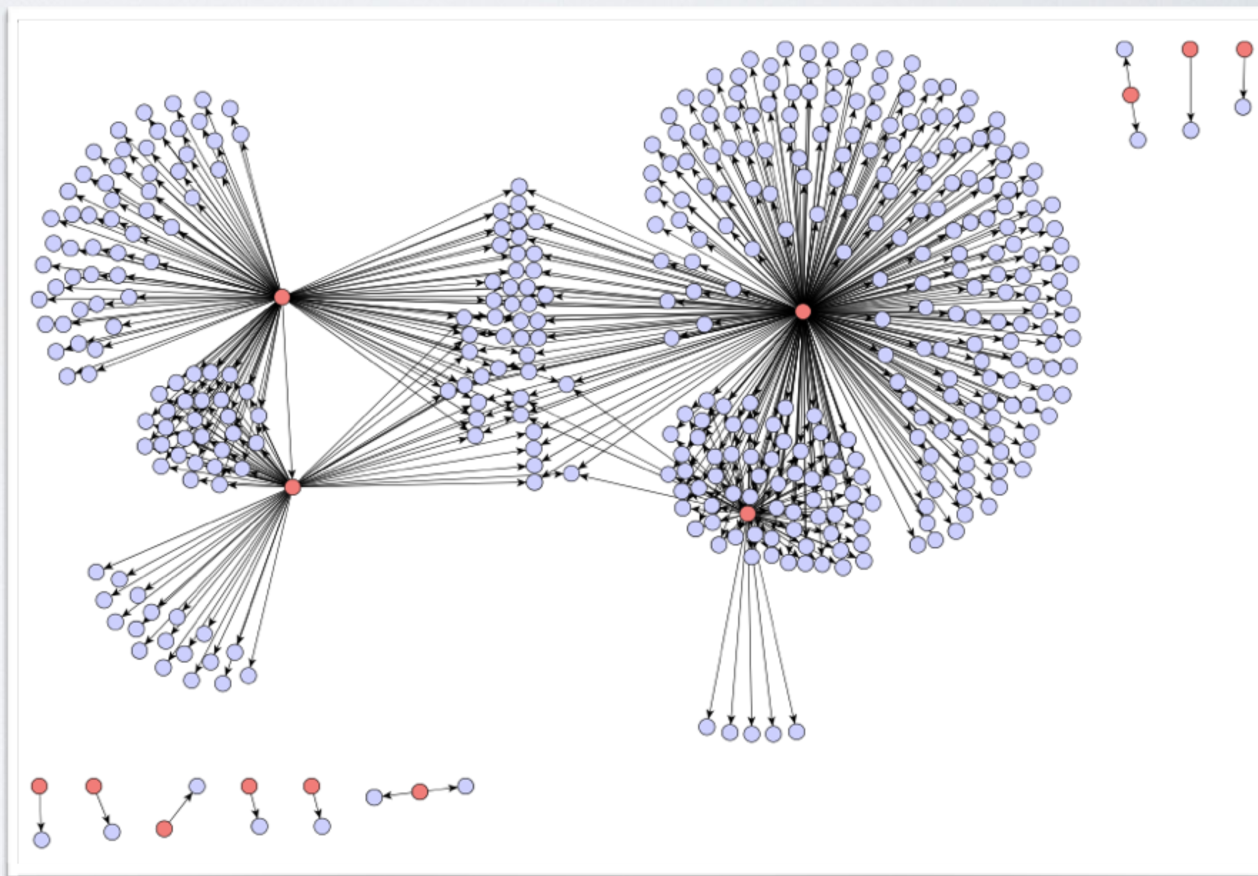


network degrees

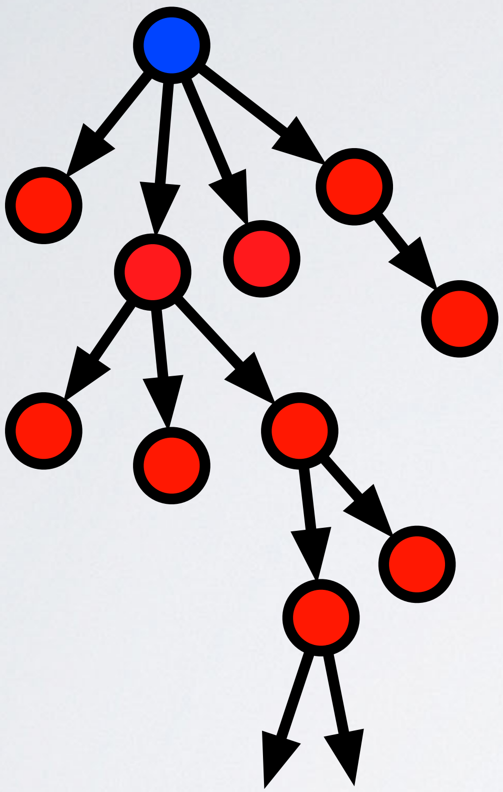
The Dynamics of Viral Marketing 2007

JURE LESKOVEC LADA A. ADAMIC BERNARDO A. HUBERMAN

- amazon.com viral marketing
- viral trace for "*Oh my Goddess!*" community
- very high degrees!
- most attempts to "influence" fail



network degrees



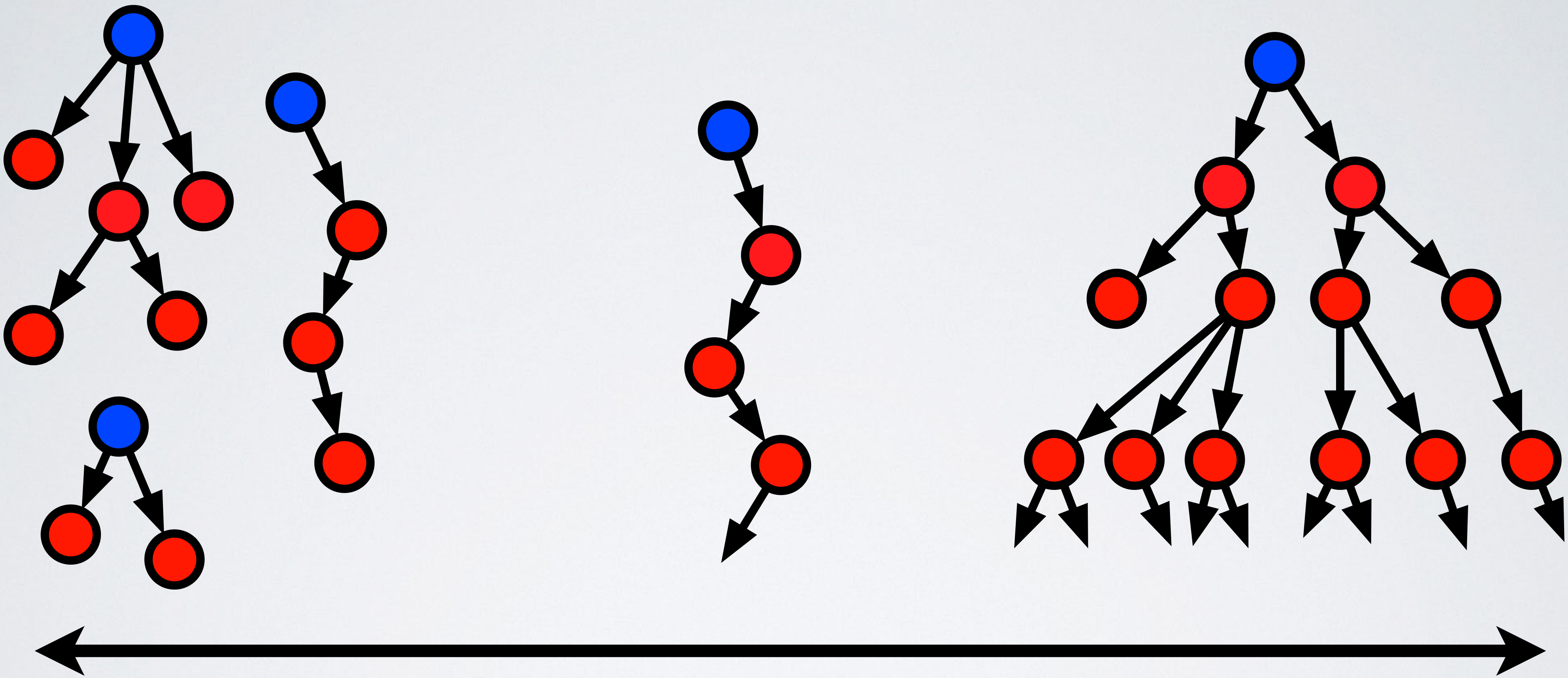
$R_0 = 0.923\ldots$

cascade
epidemic
branching process
spreading process

$$\begin{aligned} R_0 &= \text{net reproductive rate} \\ &= \text{average degree } \langle k \rangle \end{aligned}$$

caveat:
ignores network structure, dynamics, etc.

network degrees



$R_0 < 1$

"sub-critical"
small outbreaks

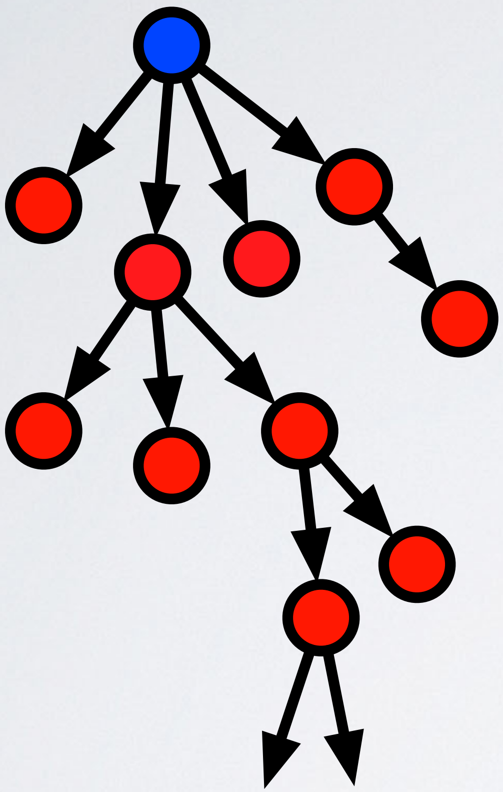
$R_0 = 1$

"critical"
outbreaks of all sizes

$R_0 > 1$

"super-critical"
global epidemics

network degrees

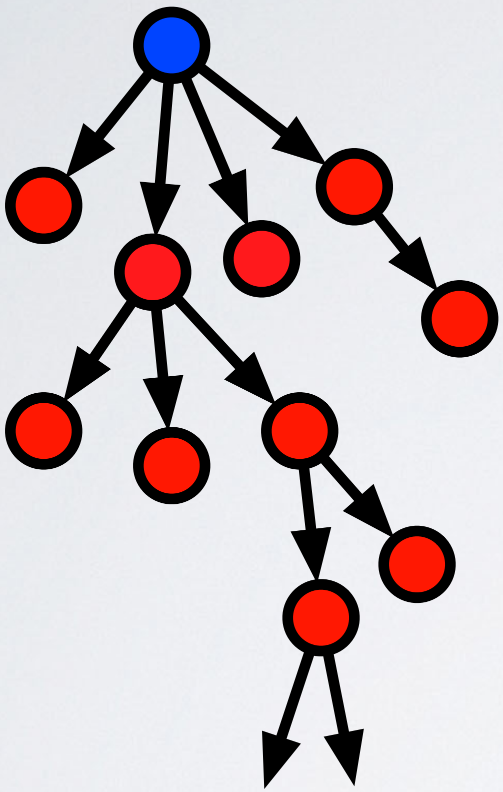


disease	R0	vaccination minimum
Measles	5-18	90-95%
Chicken pox	7-12	85-90%
Polio	5-7	82-87%
Smallpox	1.5-20+	70-80%
H1N1 influenza	1.0-3.0	

all super-critical

data from Lauren Ancel Meyers (UT Austin)

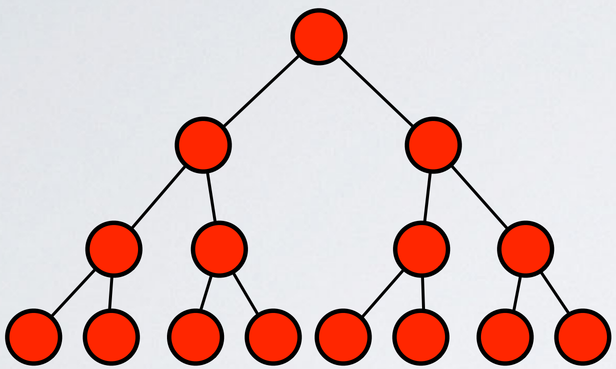
network degrees



disease	R0	vaccination minimum
Measles	5-18	90-95%
Chicken pox	7-12	85-90%
Polio	5-7	82-87%
Smallpox	1.5-20+	70-80%
H1N1 influenza	1.0-3.0	

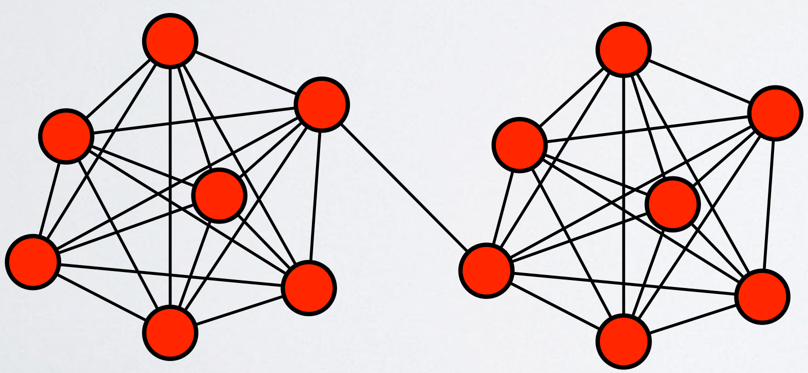
data from Lauren Ancel Meyers (UT Austin)

network degrees



bigger cascades

- smaller overlap among neighbors
- more *expander*-like
  [more like a random graph]
- higher transmission probability
- lower activation threshold



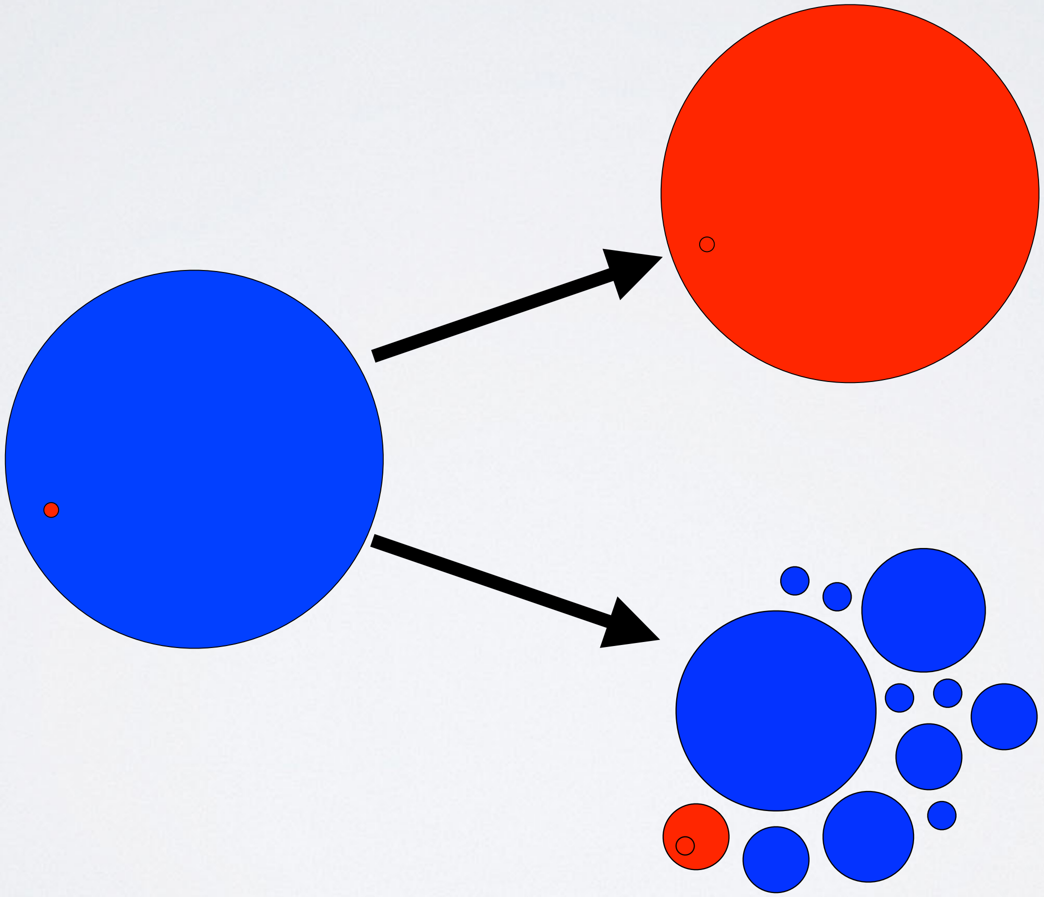
smaller cascades

- larger overlap among neighbors
- more triangles
- smaller "communities"
- more spatial-like organization
- lower transmission probability
- higher activation threshold

Volz, *J. Math. Bio.* **56**, 293–310 (2008)
Bansal et al., *J. Royal Soc. Interface* **4**, 879–891 (2007)
Karrer and Newman, *Phys. Rev. E* **82**, 016101 (2010)
Salathe and Jones, *PLoS Comp. Bio.* **6**, e1000736 (2010)

how could we halt the spread?

- break network into disconnected pieces



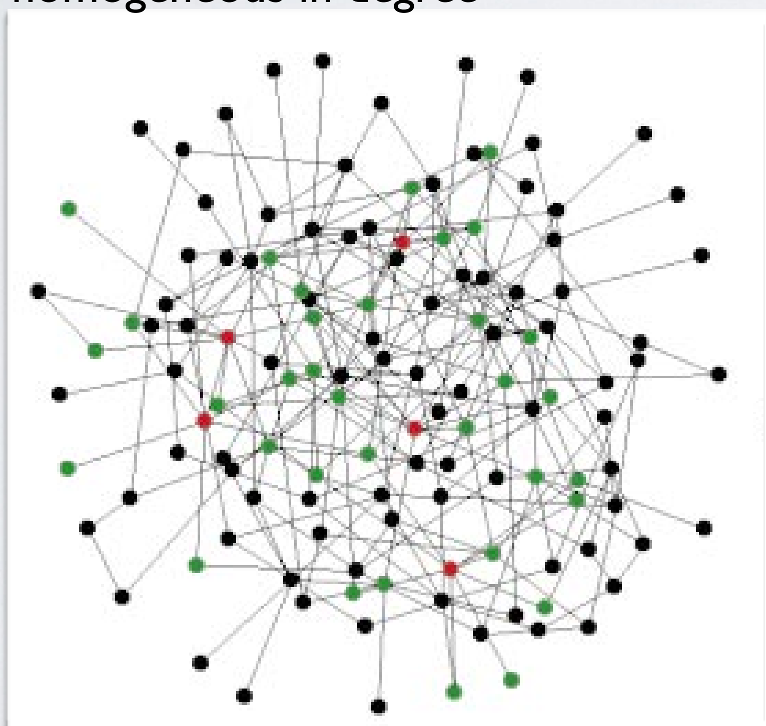
network degrees

two networks

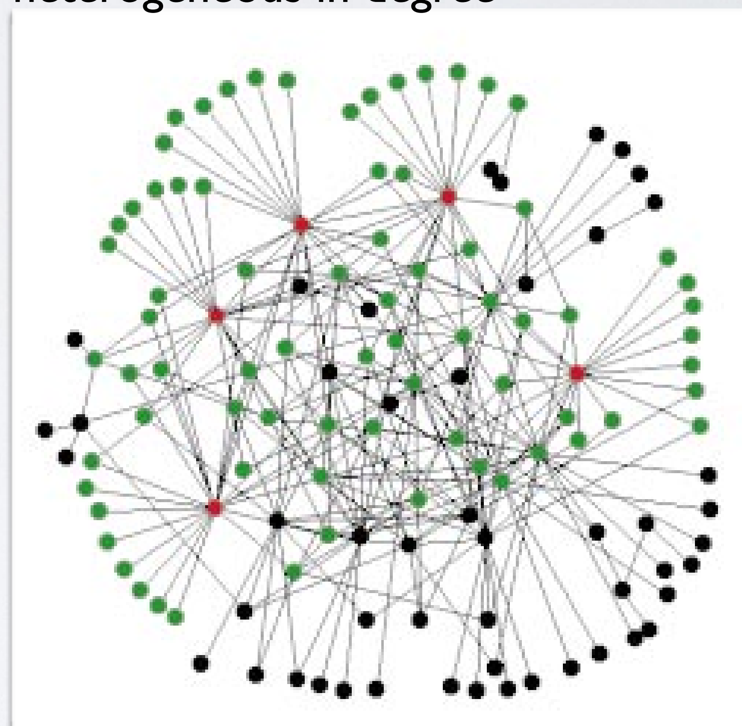
Error and attack tolerance of complex networks 2000

Réka Albert, Hawoong Jeong & Albert-László Barabási

homogeneous in degree

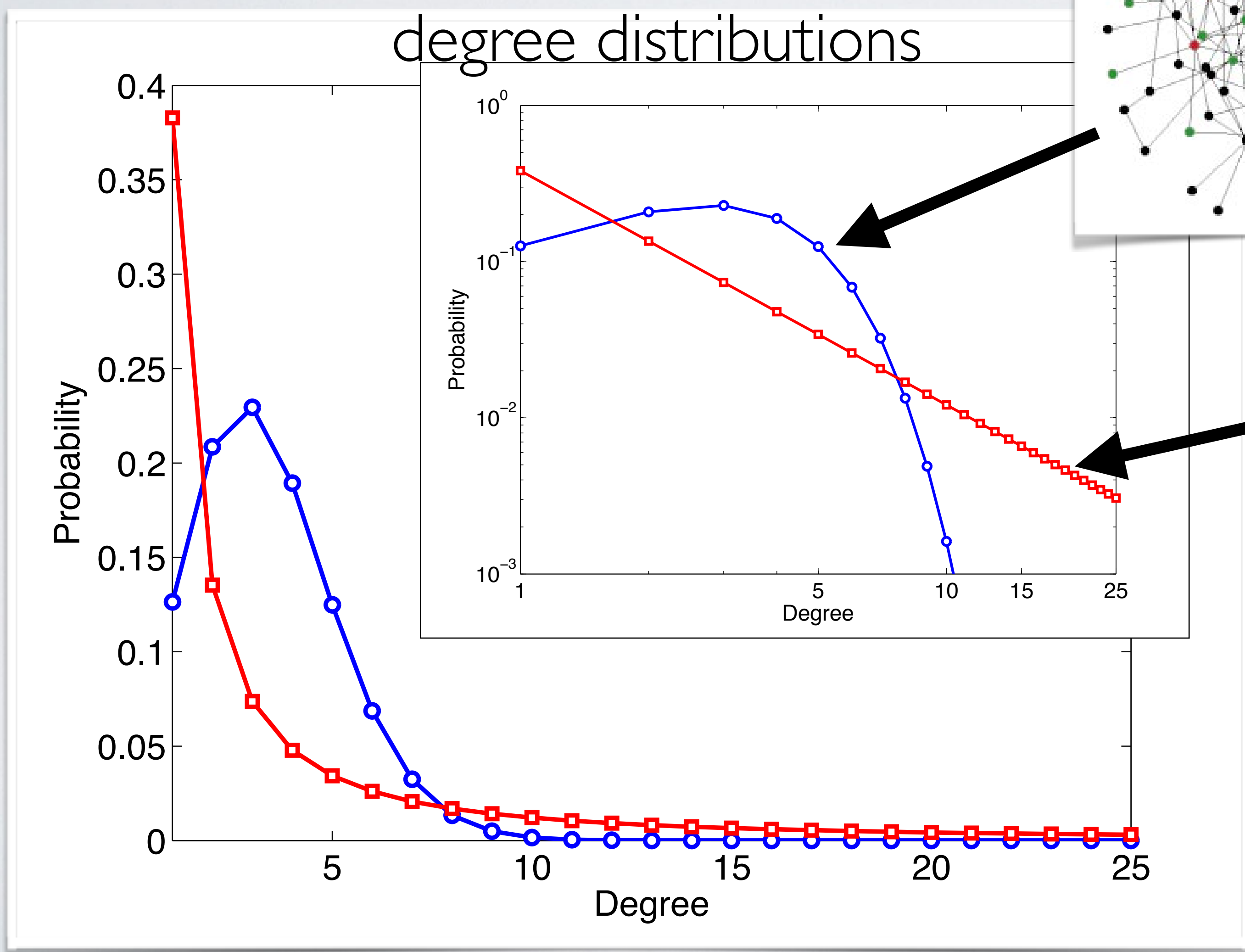


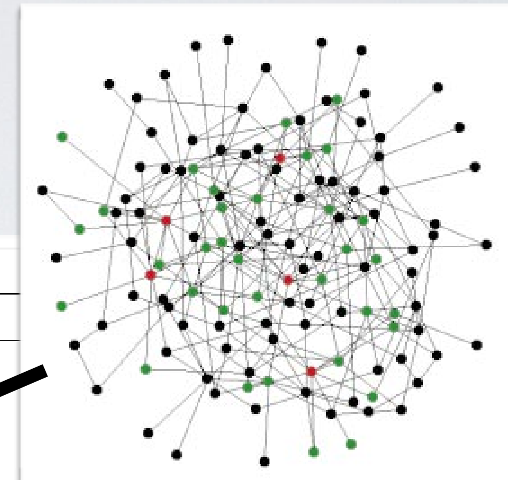
heterogeneous in degree

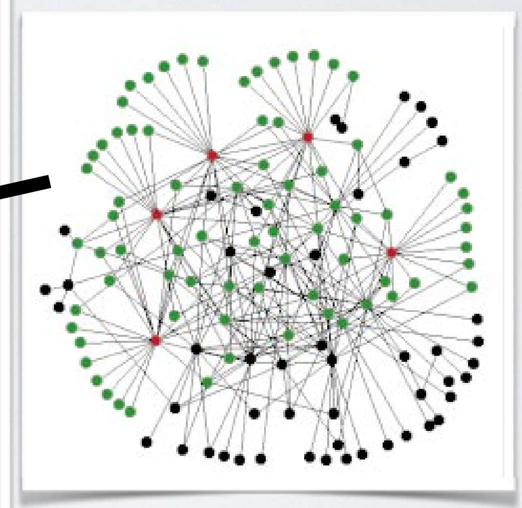


network degrees

two networks



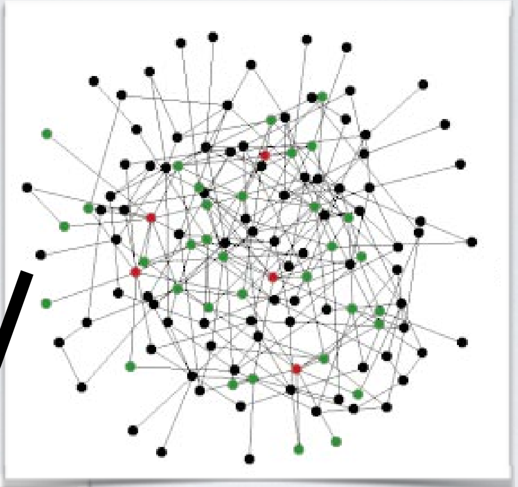


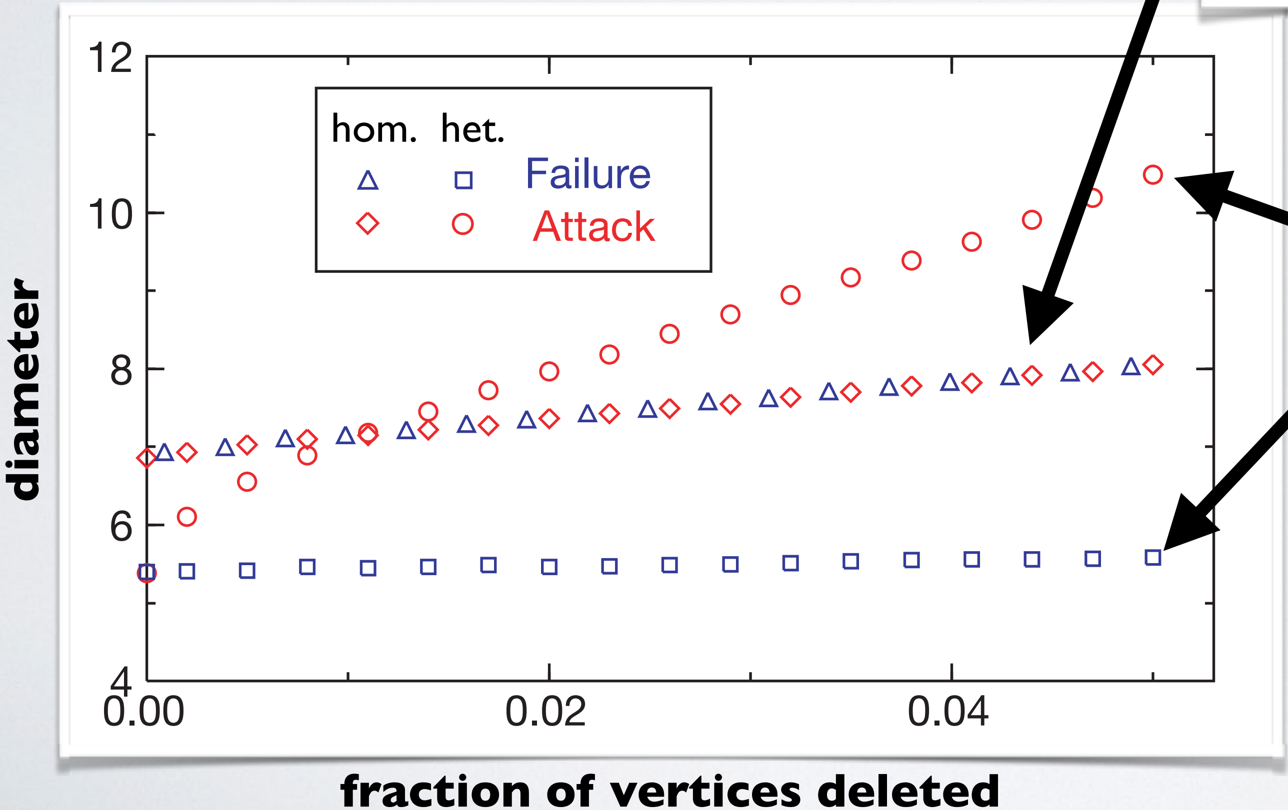


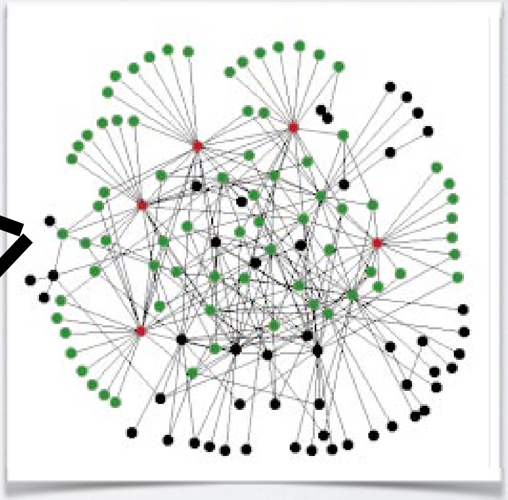
network degrees

strategy: delete vertices

1. uniformly at random ("failure")
2. in order of degree ("attack")





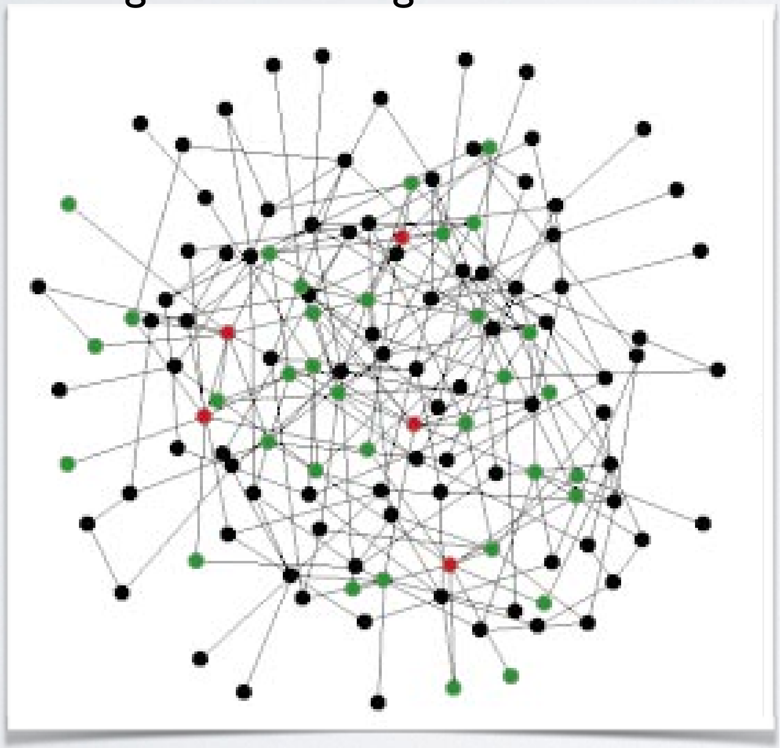


network degrees

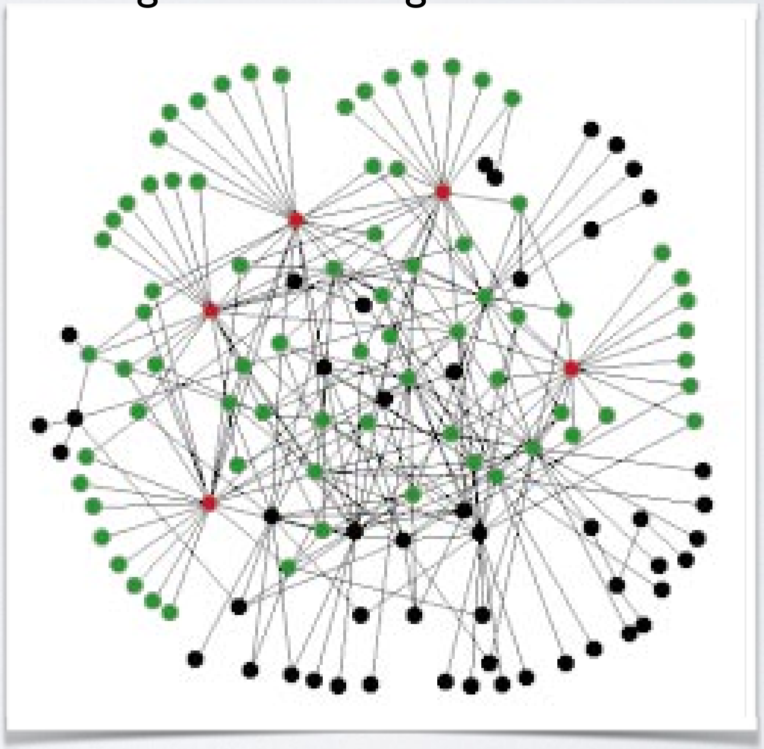
what promotes spreading?

- high-degree vertices*
- centrally-located vertices

homogeneous in degree



heterogeneous in degree

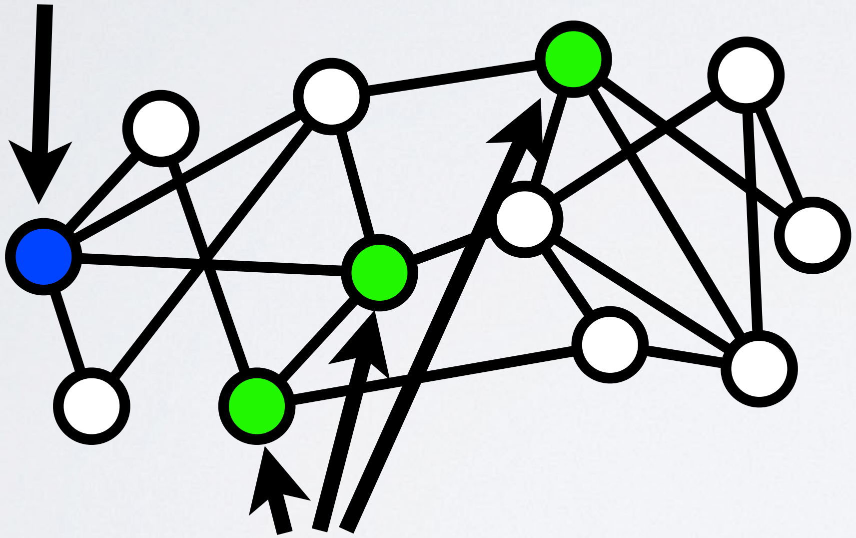


network degrees

strategy: delete vertices

3. build "fire breaks"

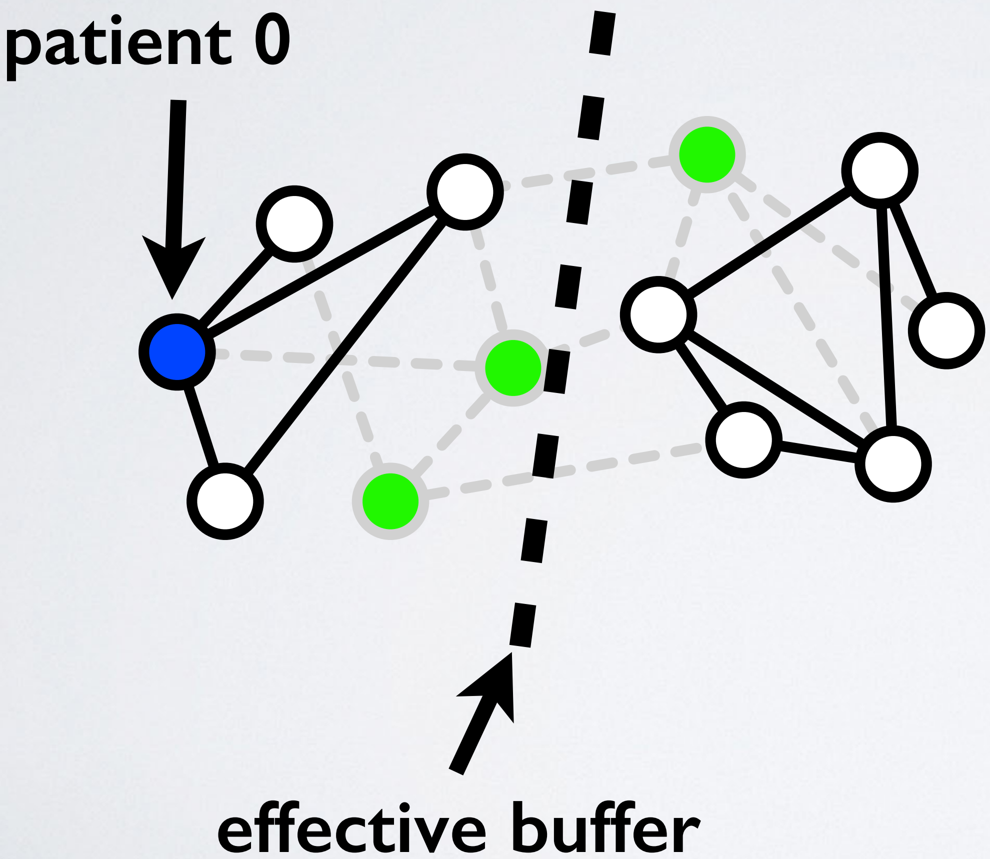
patient 0



vaccinated = deleted

("fire break")

network degrees



- **vaccination strategies**
  - the "front line" (hospitals)
  - high degree nodes
  - the vulnerable (old/young)

network degrees

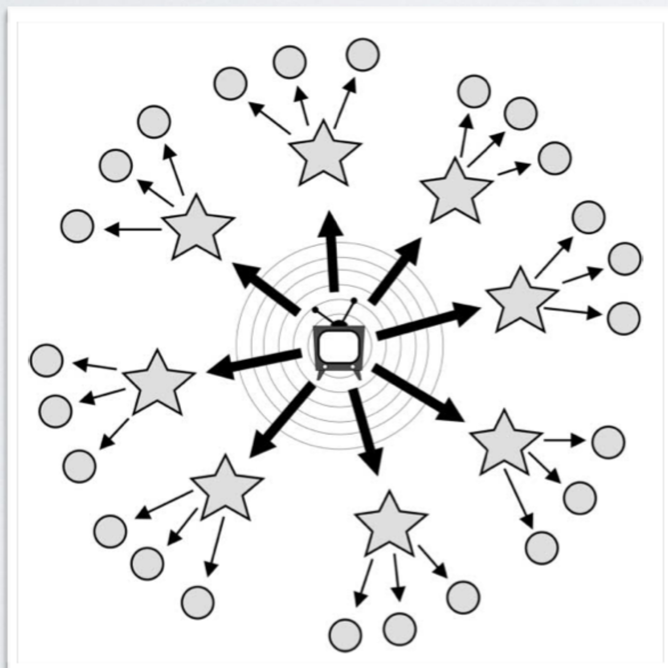
but, in social networks…

network degrees

Influentials, Networks, and Public Opinion Formation

DUNCAN J. WATTS
PETER SHERIDAN DODDS*

2007



broadcast influence

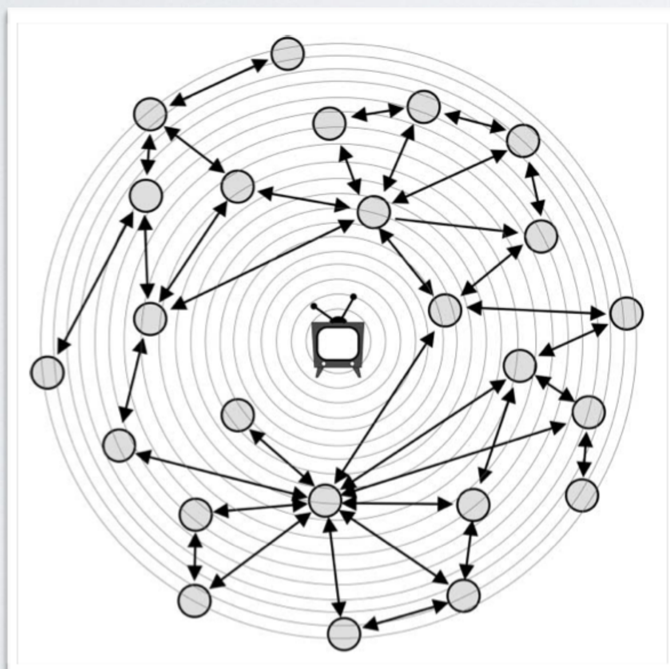
- classic information marketing
- message saturation
- **degree** is most important

network degrees

Influentials, Networks, and Public Opinion Formation

DUNCAN J. WATTS
PETER SHERIDAN DODDS*

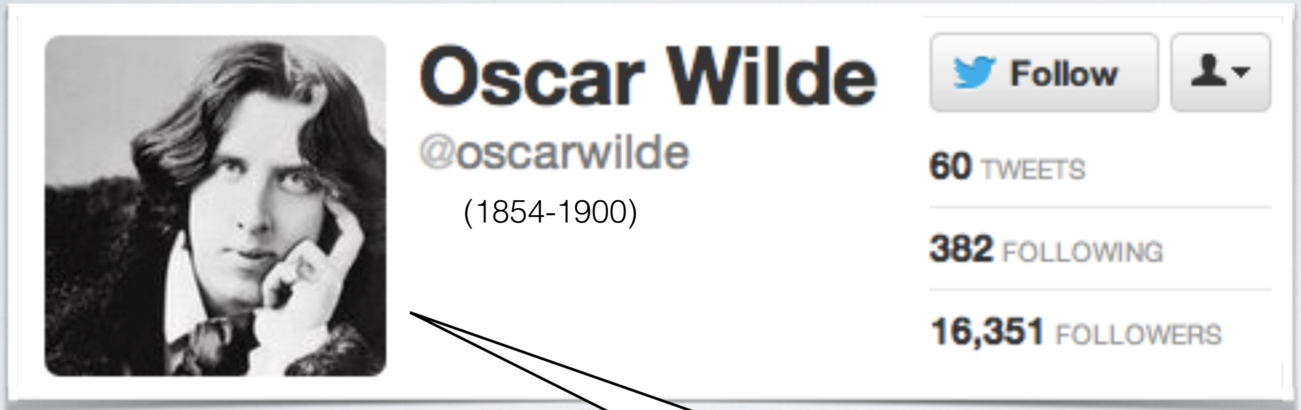
2007



network influence

- "network" (decentralized) marketing
- high-degree = "opinion leader"
- high-degree alone = **irrelevant**
- a cascade requires a legion of *susceptibles* (a system-level property)

network degrees



"The only thing worse than being talked about is not being talked about."

- "influence" not really about the influencer
- as much about the susceptibles

network degrees

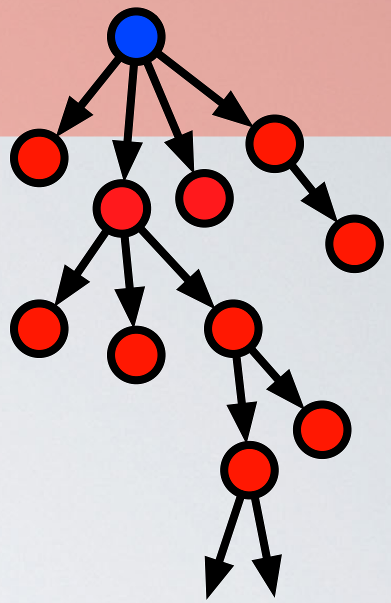
how to start a **social movement**?

network degrees

how to start a **social movement**?

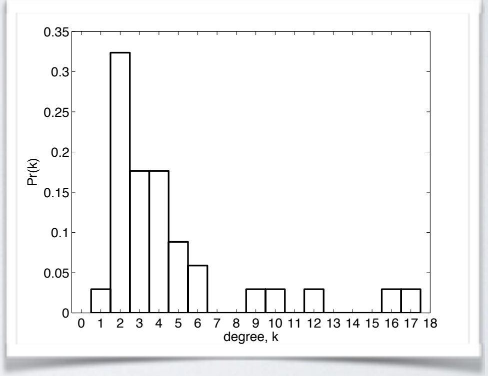
http://sivers.org/ff

network degrees



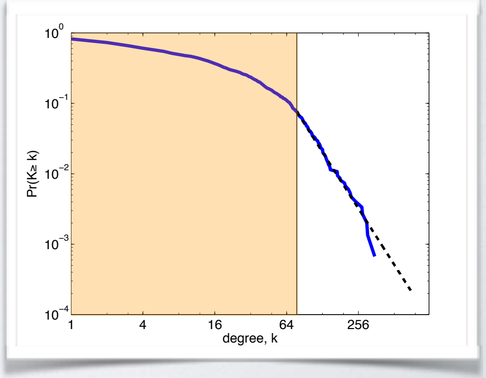
degrees:

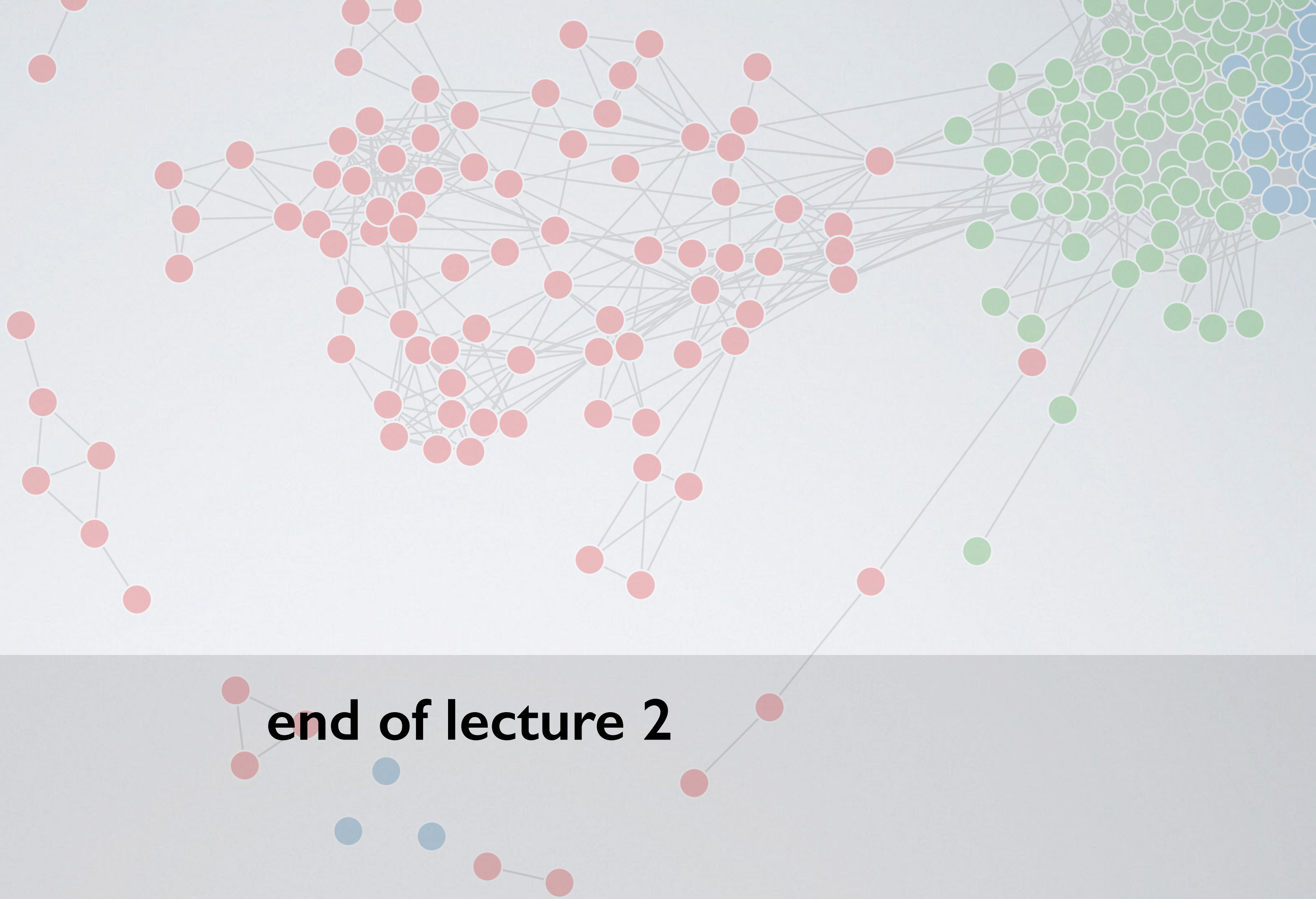
- *first-order* description of network structure
- direct implications for spreading processes
- cascades require both susceptible population *and* spreaders



open questions:

- impact of degrees on other dynamics
- feedback from dynamics to degree
  [adaptive behaviors like self-quarantine, evangelism]
- when does degree *not* matter





end of lecture 2

selected references

- The structure and function of complex networks. M. E. J. Newman, *SIAM Review* **45**, 167–256 (2003).
- *The Structure and Dynamics of Networks*. M. E. J. Newman, A.-L. Barabási, and D. J. Watts, Princeton University Press (2006).
- Hierarchical structure and the prediction of missing links in networks. A. Clauset, C. Moore, and M. E. J. Newman, *Nature* **453**, 98–101 (2008).
- Modularity and community structure in networks. M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
- Why social networks are different from other types of networks. M. E. J. Newman and J. Park, *Phys. Rev. E* **68**, 036122 (2003)
- Random graphs with arbitrary degree distributions and their applications. M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).
- Comparing community structure identification. L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas. *J. Stat. Mech.* P09008 (2005).
- Characterization of Complex Networks: A Survey of measurements. L. daF. Costa, F. A. Rodrigues, G. Travieso and P. R. VillasBoas. arxiv:cond-mat/050585 (2005).
- Evolution in Networks. S.N. Dorogovtsev and J. F. F. Mendes. *Adv. Phys.* **51**, 1079 (2002).
- Revisting "scale-free" networks. E. F. Keller. *BioEssays* **27**, 1060-1068 (2005).
- Currency metabolites and network representations of metabolism. P. Holme and M. Huss. arxiv:0806.2763 (2008).
- Functional cartography of complex metabolic networks. R. Guimera and L. A. N. Amaral. *Nature* **433**, 895 (2005).
- Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. J. Leskovec, J. Kleinberg and C. Faloutsos. *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* 2005.
- The Structure of the Web. J. Kleinberg and S. Lawrence. *Science* 294, 1849 (2001).
- Navigation in a Small World. J. Kleinberg. *Nature* 406 (2000), 845.
- Towards a Theory of Scale-Free Graphs: Definitions, Properties and Implications. L. Li, D. Alderson, J. Doyle, and W. Willinger. *Internet Mathematics 2(4), 2006*.
- A First-Principles Approach to Understanding the Internet's Router-Level Topology. L. Li, D. Alderson, W. Willinger, and J. Doyle. *ACM SIGCOMM* 2004.
- Inferring network mechanisms: The Drosophila melanogaster protein interaction network. M. Middendorf, E. Ziv and C. H. Wiggins. *Proc. Natl. Acad. Sci. USA* 102, 3192 (2005).
- Robustness Can Evolve Gradually in Complex Regulatory Gene Networks with Varying Topology. S. Ciliberti, O. C. Martin and A. Wagner. *PLoS Comp. Bio.* 3, e15 (2007).
- Simple rules yield complex food webs. R. J. Williams and N. D. Martinez. *Nature* 404, 180 (2000).
- A network analysis of committees in the U.S. House of Representatives. M. A. Porter, P. J. Mucha, M. E. J. Newman and C. M. Warmbrand. *Proc. Natl. Acad. Sci. USA* 102, 7057 (2005).
- On the Robustness of Centrality Measures under Conditions of Imperfect Data. S. P. Borgatti, K. M. Carley and D. Krackhardt. *Social Networks* **28**, 124 (2006).