

# Social and Information Networks 2: Evolution

Internet Analytics (COM-308)

Prof. Matthias Grossglauser  
School of Computer and Communication  
Sciences



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

# Overview

- Herding and “watching thy neighbor”
  - Information cascades: why imitating your friends makes sense - and how it can lead to surprising group behavior
  - Heavy-tailed degree distributions: “the rich get richer” applied to networks
- Observing network properties
  - The importance of the observer
  - Example: your friends are more popular than you!

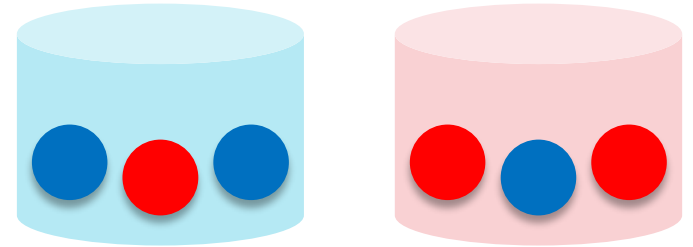
# Watching thy Neighbor?

- Human decision-making:
  - Primary private information...
  - Heavily influenced by what decisions taken by others
- Reason:
  - Primary information: often too voluminous, noisy, not trustworthy,...
  - By imitating others, piggyback on their effort to interpret primary information
- Question:
  - Macro behavior of such systems?



# Herding: how it can go wrong

- Urn with 3 balls
  - A priori distribution (blue/red majority) = (0.5,0.5)
  - majority blue: 2 blue + 1 red
  - majority red: 2 red + 1 blue
- A group of people take turns:
  - Draw a ball from the urn at random
  - Check the color of the ball privately, put it back in urn
  - Announce their **guess** (blue/red majority) to everybody
- Assumption:
  - ❶ Each individual is altruistic: do what allows others to make best guess
  - ❷ Each individual is selfish = tries to make best guess for himself



# Urn model: altruistic (1)

- Every person:
  - Selects a ball at random (with replacement)
  - Announces the color of the ball to everybody
- As  $n \rightarrow \infty$ , majority color of urn is equal to color most frequently observed
  - Consequence of law of large numbers
- After a few “sacrifices”, everybody could produce best guess
  - Sacrifice in the sense that first few individuals might be forced to say **red** (color of their ball) even if previous information suggests **blue** majority

# Urn model: selfish (2)

- Sequential decision-making
  - Public guess - goal: correct guess for many/most people
  - Observed color remains private
- First individual:
  - Blue ball: announce  $\text{guess}(1) = \text{blue}$
  - Red ball: announce  $\text{guess}(1) = \text{red}$
  - Public guess of first fully reveals private information
- Second individual:
  - If  $\text{color}(2) = \text{guess}(1)$ : announce this color
  - If  $\text{color}(2) \neq \text{guess}(1)$ : does not matter (assume  $\text{color}(2)$ )
  - Public guess of second fully reveals private information

# Urn model: selfish (2)

- Third individual:
  - If  $\text{guess}(1) \neq \text{guess}(2)$ : announce  $\text{guess}(3) = \text{color}(3)$
  - If  $\text{guess}(1) = \text{guess}(2)$ :
    - Announce  $\text{guess}(3) = \text{guess}(2) = \text{guess}(1)$ , regardless of  $\text{color}(3)$
  - Why is this?
    - Person 3 knows that guesses 1+2 reveal perfect information
    - Therefore, regardless of  $\text{color}(3)$ ,  $\text{guess}(1) = \text{guess}(2)$  dominates guess
- Fourth,..., $\infty$ th individual:
  - If  $\text{guess}(1) = \text{guess}(2)$ :
    - Announce  $\text{guess}(i) = \text{guess}(2) = \text{guess}(1)$ , regardless of  $\text{color}(i)$

# Urn model: (2) leads to cascade

- If  $\text{guess}(1) = \text{guess}(2)$  were both wrong, then all future guesses are wrong!
- This happens with prob.  $1/9$
- Even though each individual is using available information in the best way to make a guess
- Can show that in this model, this is sure to happen eventually (even if not for 3<sup>rd</sup> individual)



# Information cascade: suboptimal decision

- Cascade: sequential decisions
- Individual:
  - Efficiency gain by observing others' decisions
- Global behavior:
  - Primary information can “wash out”
  - Suboptimal or random decisions
- Might these be cascades:
  - Stock market gyrations, “flash crash”
  - Inexplicable shifts in popularity of {restaurants, clubs, celebrities,...}
  - Fashion, style, celebrity,...
  - ...



# Herding in networks

- Observation:
  - Degree distributions in networks often resemble power laws
- Power law:
  - $P(D > d) \propto d^{-\gamma}$
- Most distributions have “light tails”:
  - $P(D > d) \propto e^{-\alpha d}$  (or lighter/bounded)
  - Exponential, Geometric, Gaussian, Poisson, ...

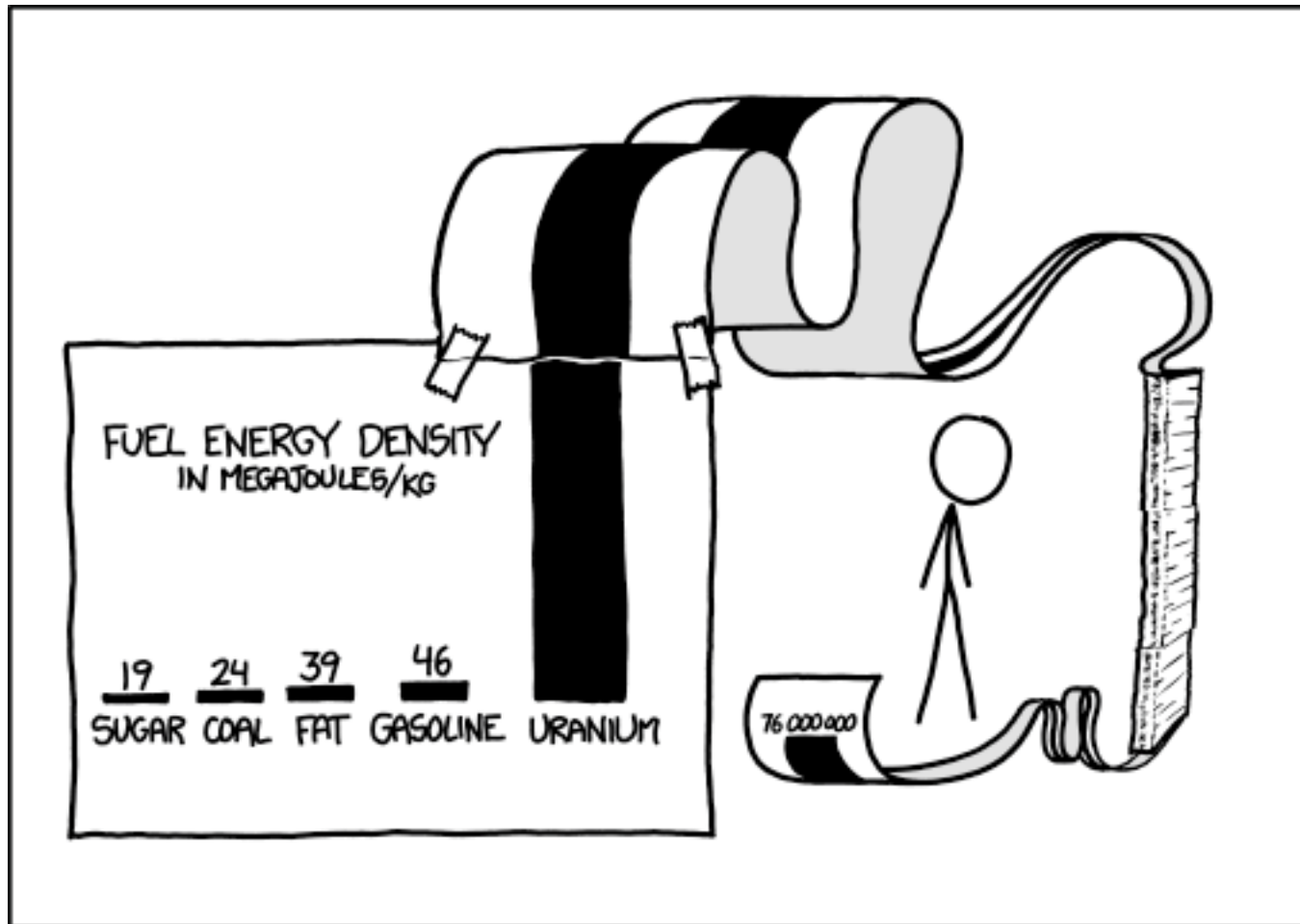
# Pareto distribution

- Support:  $d \in [\beta, \infty)$
- CCDF (Complementary Cumulative Distribution Function):
  - $P(D > d) = 1 - F_D(d) = \begin{cases} \left(\frac{d}{\beta}\right)^{-\gamma}, & d \geq \beta \\ 1 & \text{otherwise} \end{cases}$
  - $\gamma$ : exponent, also called “Pareto index”
- Moments:
  - $E[D^k] = \begin{cases} \frac{\beta^k \gamma}{\gamma - k}, & k < \gamma \\ \infty & \text{otherwise} \end{cases}$

# Numerical comparison exp/power

- Distribution of human height:
  - Mean = 178 cm
  - Stddev = 8 cm
- Compare tails: how tall are extremely tall people?
  - What is  $d^*$  such that  $P(D > d^*) = 10^{-9}$
- Normal:
  - $d^* = 226$  cm
- Pareto:
  - $\gamma \cong 23, \beta \cong 170\text{cm}$
  - $d^* = 420$  cm !!
- Assumption very important for extremal values!

# Log-log plot

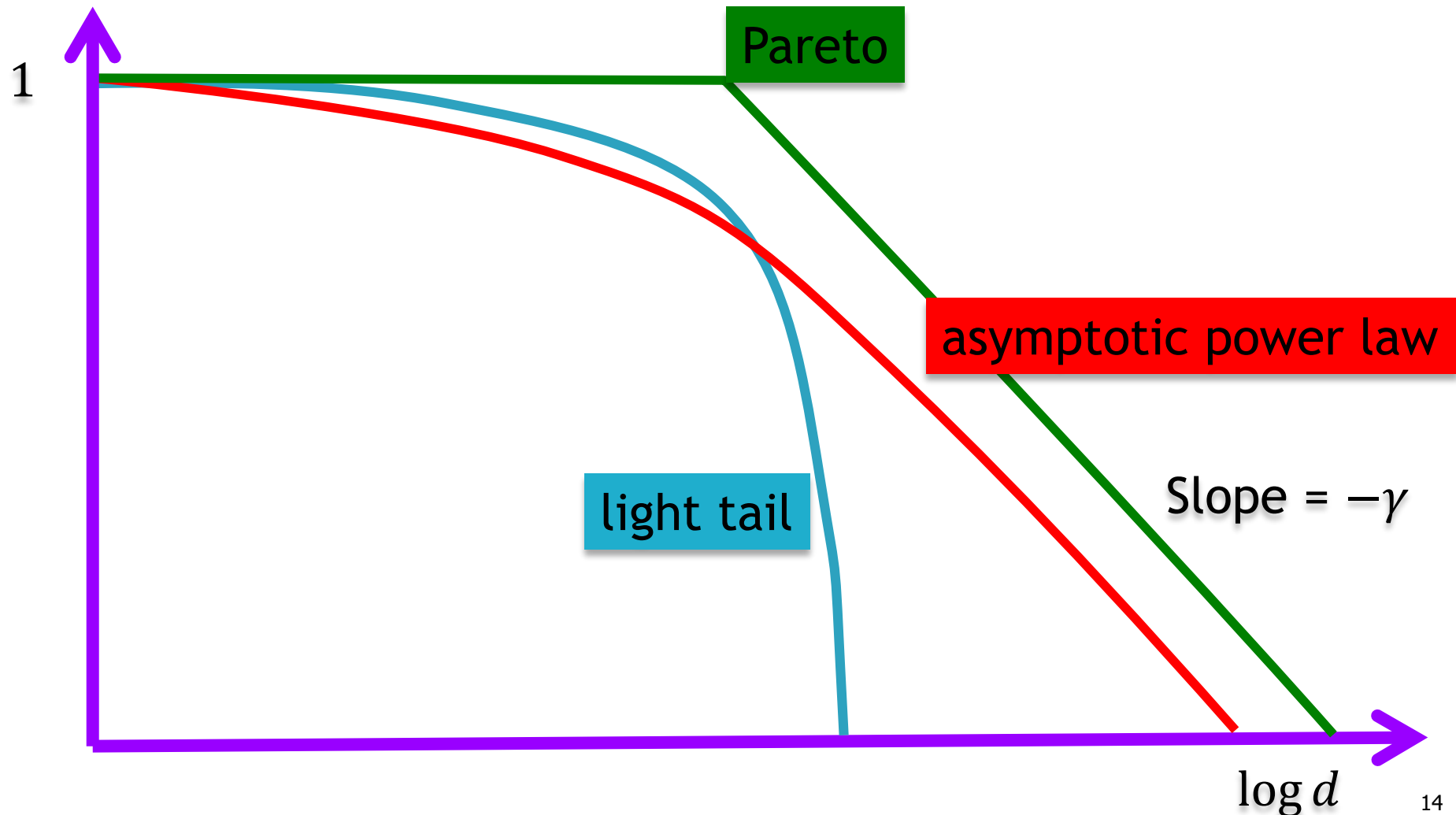


SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T  
FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY*.

Source: xkcd #1162

# Log-log distribution plot

- $\log P(D > d)$



# Examples of observed power laws

- File sizes on a computer
- Stock market crashes
- Sizes of cities
- Phone call length
- Wealth & income distribution
- Sizes of floods
- Popularity of web pages
- Word frequencies in prose
- Degree distribution in social networks
- ...

# Why worry about the tail?

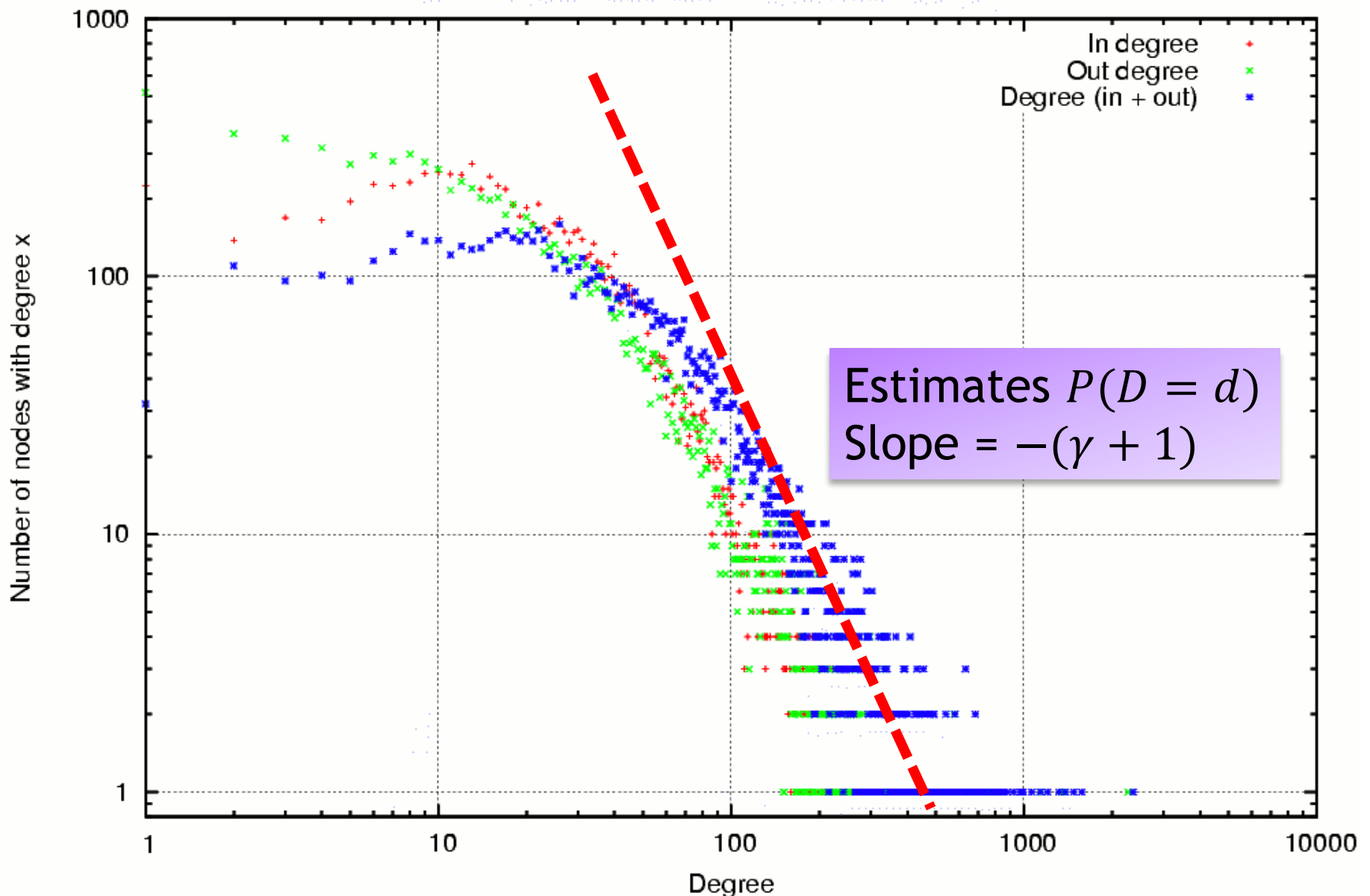
- Would you like to sit on a plane engineered under a Gaussian assumption for turbulence? ;-)





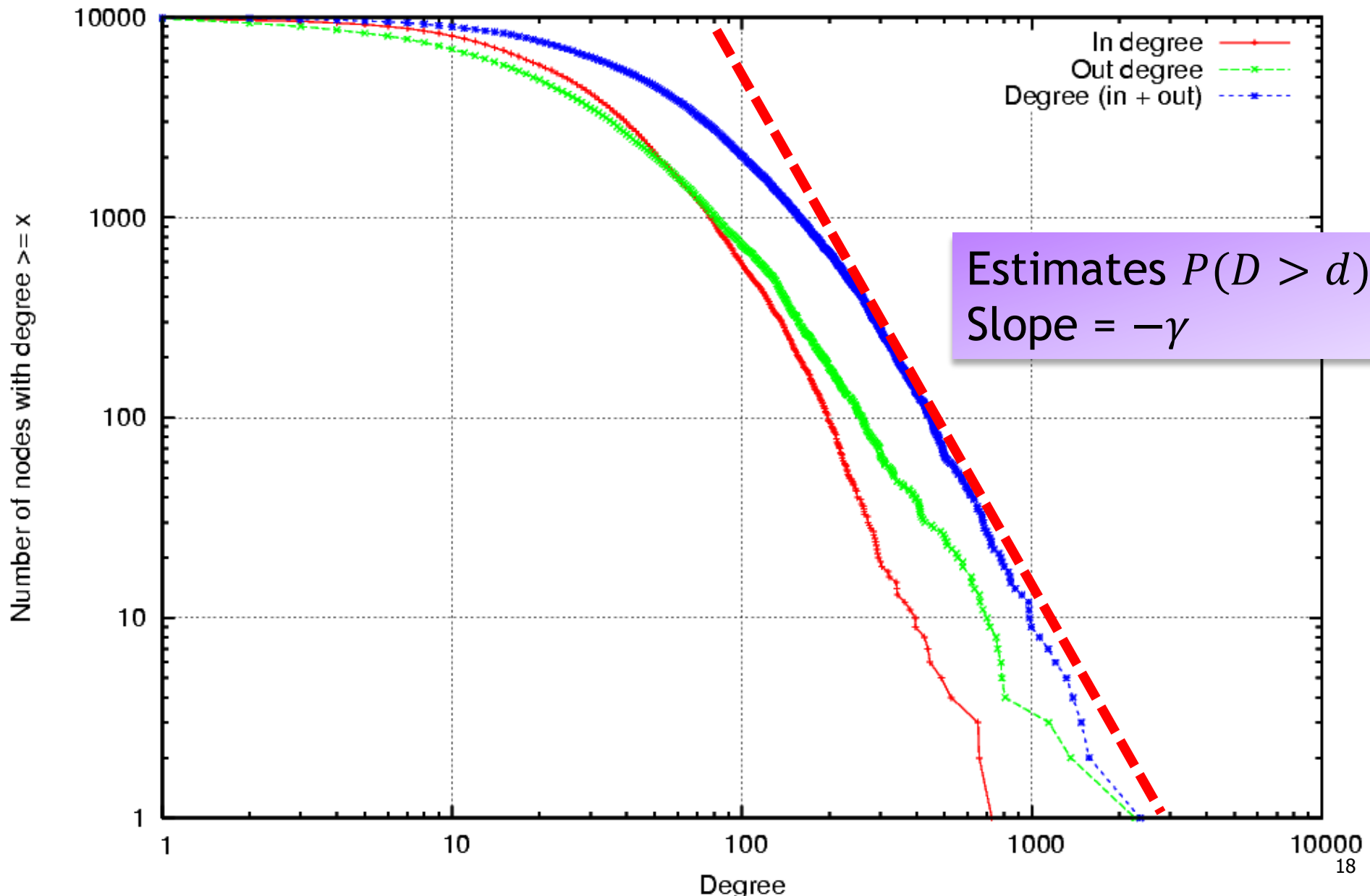
# Log-log distribution plot

26 weeks data, 9909 nodes, 355954 directed edges



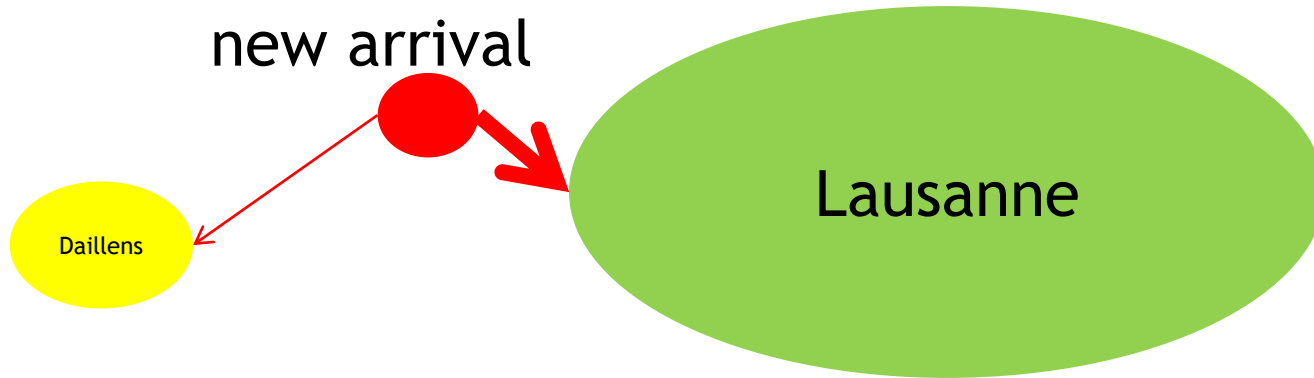
# Log-log cumulative plot

26 weeks data, 9909 nodes, 355954 directed edges



# One explanation: the rich get richer

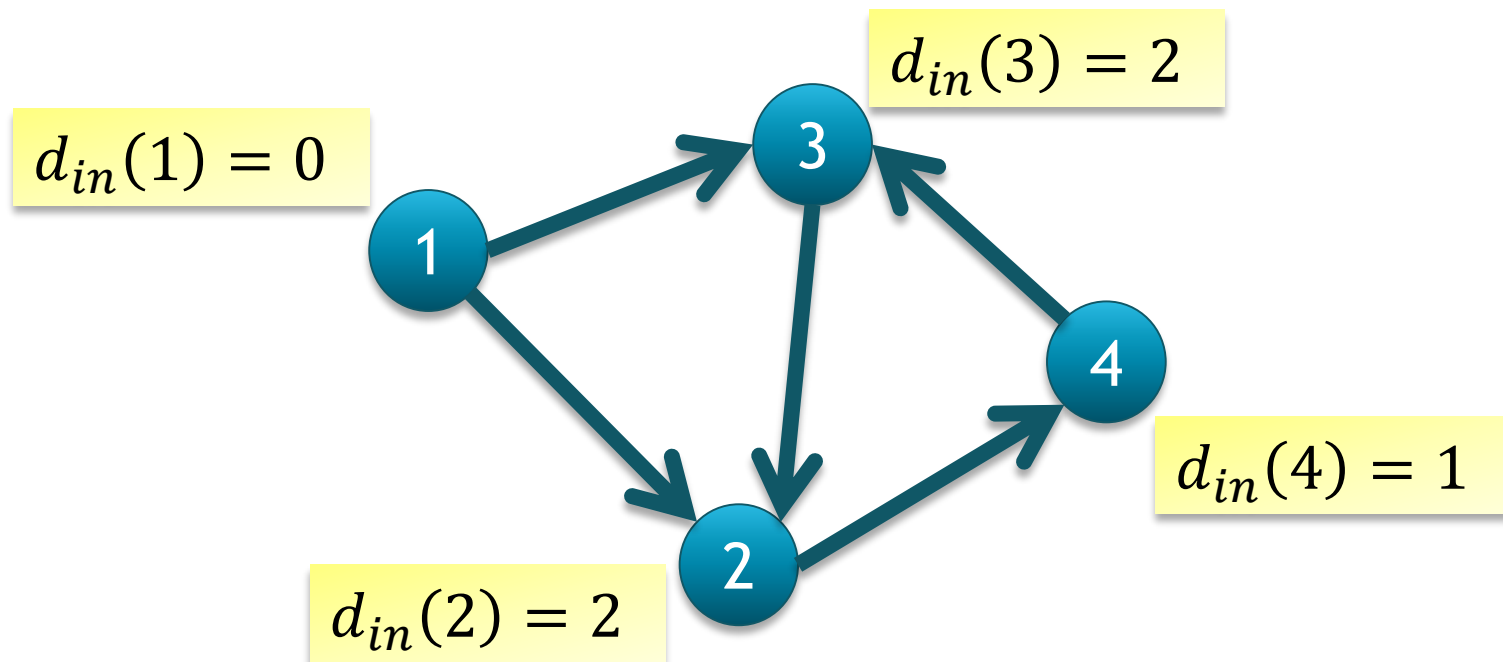
- New arrival in our region: move to Daillens or Lausanne?
  - More likely Lausanne, because more people already there
- City size distribution after many arrivals?



- Also: “the first million is the hardest” ;-)

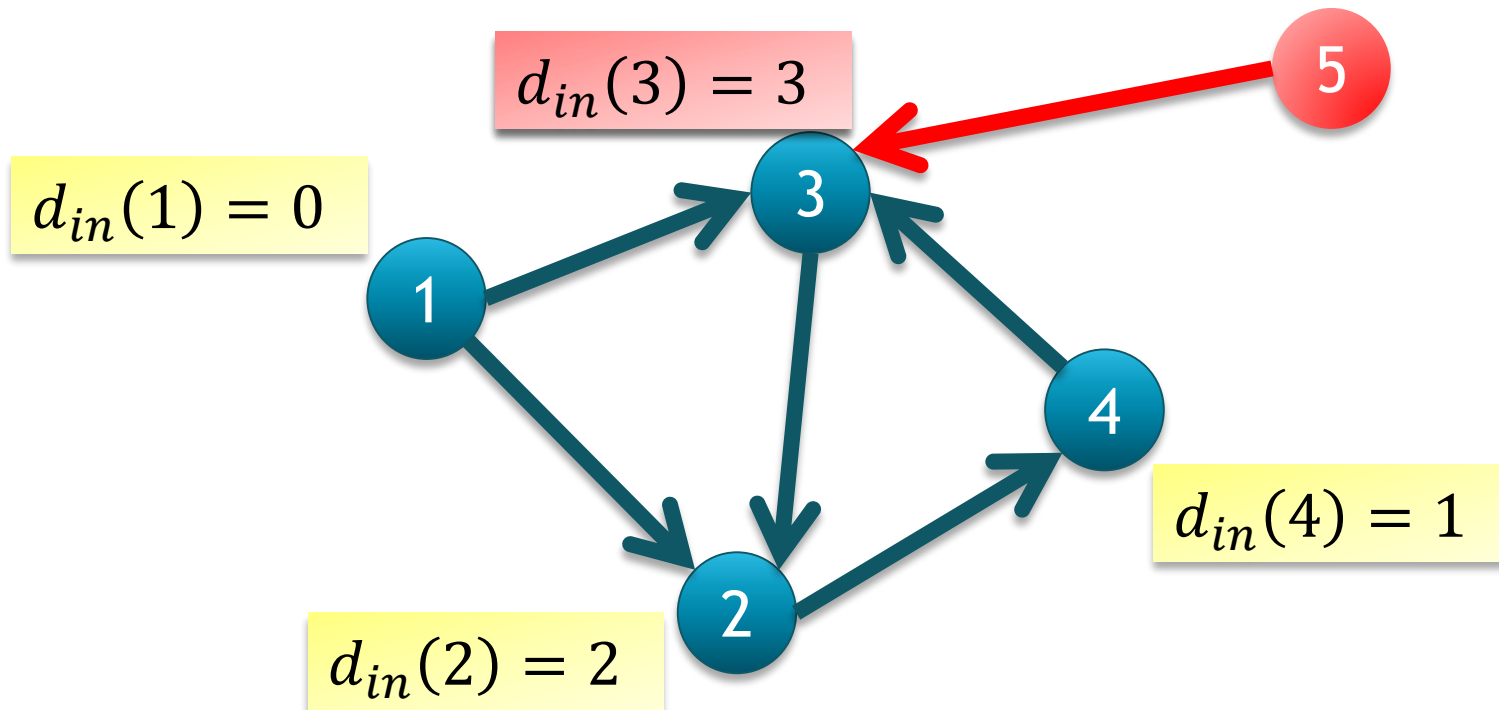
# Preferential attachment in growing nets

- Growth model: nodes arrive one by one and join the existing network
  - Directed graph
  - In-degree  $d_{in}(v)$  measures “popularity” and “attractiveness” of a node



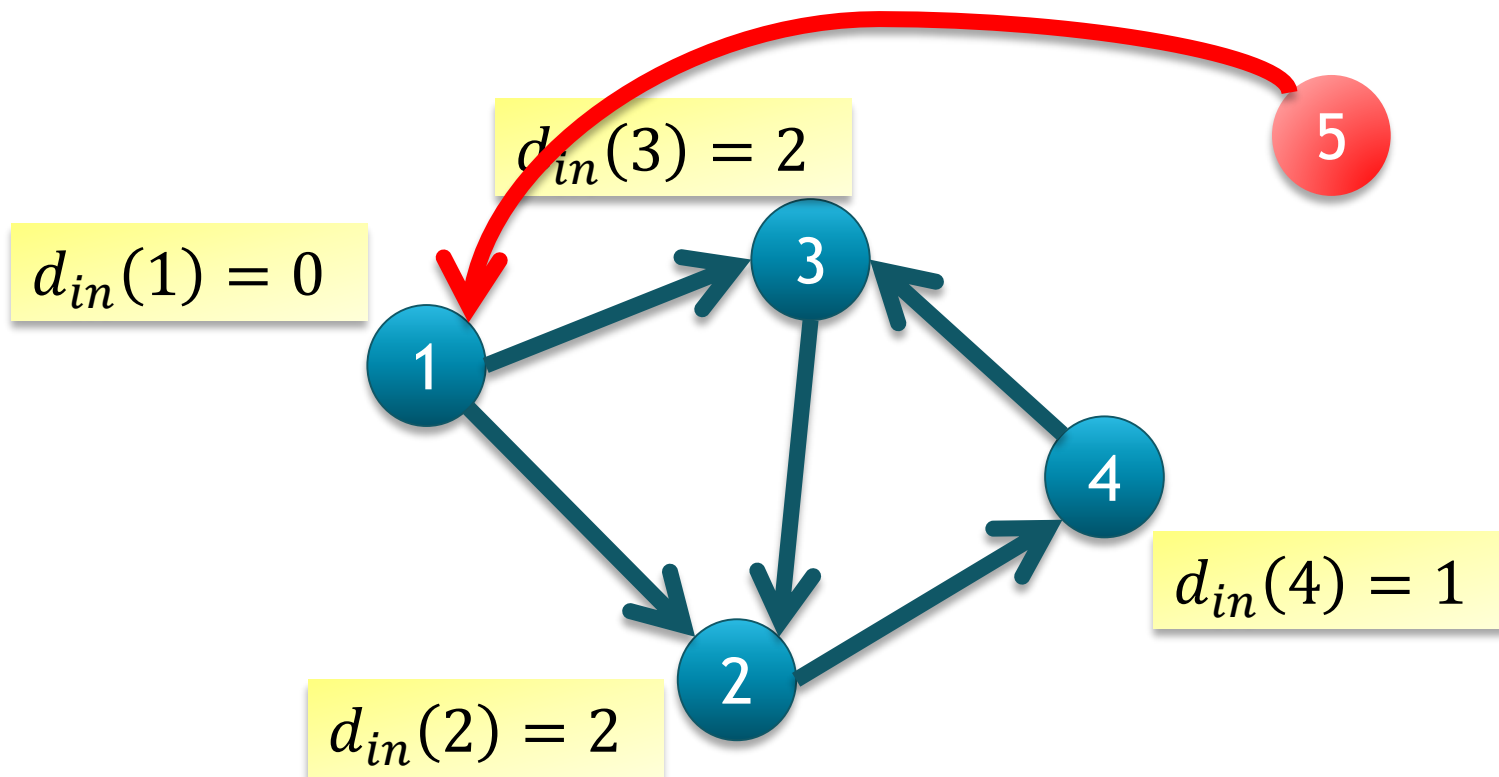
# Preferential attachment

- Preferential attachment: new node creates one edge
- Prob. of connecting to  $v$  is  $\propto d_{in}(v)$ 
  - Intuition: high-degree easier to meet; more popular; more useful;...



# Preferential attachment

- Node with in-degree 0 never gets “started”
- Need another assumption:
  - With prob.  $\alpha$ , new node connects uniformly at random
  - With prob.  $(1 - \alpha)$ , preferential attachment



# Pref attachment: analysis

- Evolution of this system:
  - Graph structure only matters through in-degrees
- Markov chain  $\{X_j(t)\}$ : # nodes with in-degree =  $j$  at time  $t$ 
  - Total # of nodes and edges at time  $t = t$
  - Notation:  $X_j := X_j(t)$

- Drift:

- $$P(X_j(t+1)=X_j(t)+1) = \underbrace{\alpha \frac{X_{j-1}}{t}}_{\text{prob. of selecting a node uniformly}} + (1-\alpha) \underbrace{(j-1) \frac{X_{j-1}}{t}}_{\text{prob. of selecting a node } \propto \text{degree}}$$

- $$P(X_j(t+1)=X_j(t)-1) = \alpha \frac{X_j}{t} + (1-\alpha)j \frac{X_j}{t}$$

# Pref attachment: analysis

- Combined drift (pretend  $X_j \in \mathbb{R}$ )

- $$\frac{dX_j}{dt} = \frac{\alpha(X_{j-1} - X_j) + (1-\alpha)((j-1)X_{j-1} - jX_j)}{t}$$

- Assume as  $t \rightarrow \infty$ , degree sequence converges  $\frac{X_j}{t} \rightarrow c_j$ , then solve for  $c_j$ :

- $c_j$ : fraction of nodes with degree  $j$

- $$\frac{c_j}{c_{j-1}} = \frac{\alpha + (1-\alpha)(j-1)}{1 + \alpha + (1-\alpha)j} = 1 - \frac{2-\alpha}{1 + \alpha + (1-\alpha)j}$$

- for large  $j$ , this is  $\cong 1 - \frac{2-\alpha}{1-\alpha}j^{-1}$



# Pref attachment: analysis

- Note that  $\left(\frac{j}{j-1}\right)^{-(\gamma+1)} = \left(1 - \frac{1}{j}\right)^{\gamma+1} \sim 1 - \frac{\gamma+1}{j}$
- So  $\gamma = \frac{2-\alpha}{1-\alpha} - 1 = \frac{1}{1-\alpha}$
- Putting together:
  - $\frac{c_j}{c_{j-1}} = \left(\frac{j}{j-1}\right)^{-(\gamma+1)}$ , hence
- $c_j \propto j^{-(\gamma+1)}$ : power law
- The stronger the preferential attachment ( $\alpha$  small), the “heavier” the tail of the degree distribution
- Arguments can be made rigorous

# Network effects and “winner-takes-all”

- Example:
  - Facebook vs {friendster, sixdegrees, xing,...}
  - Technology standards: BluRay
- Metcalfe's Law:
  - The value of a network is proportional to  $n^2$
  - Because the value to an individual is proportional to  $n$
- Lock-in
  - Being early is very important

# Observer: Friendship Paradox

- “Your friends have more friends than you”
- Experiment:
  - Get on facebook and compute the average # friends of your friends
  - How does this compare to your own # friends?

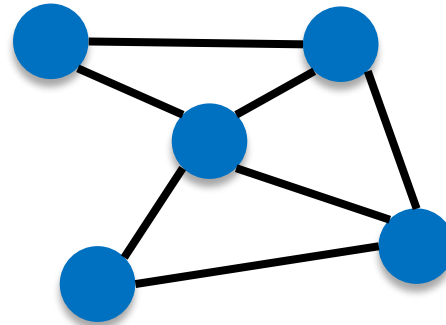
# Friendship Paradox

- Formally:
  - Social network =  $G(V, E)$
  - $d_v$ : degree of node  $v$
  - $n = |V|$ : number of nodes,  $m = |E|$ : number of edges
- Average number of friends:  $\mu = \frac{\sum d_v}{n}$
- How to talk about average number of friends' friends?

# Sampling nodes vs sampling edges

- Average degree over nodes:

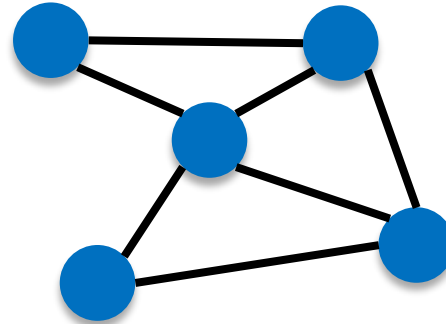
- $$\mu = \frac{\sum d_v}{n} = 2 \frac{m}{n}$$



“look at each person”

- Average degree over edges:

- $$\frac{\sum_{(u,v) \in E} d_v}{2m}$$



“look at each person’s list of friends”

# Friendship Paradox

- Lemma:

- $$\frac{\sum_{(u,v) \in E} d_v}{2m} = \mu \left( 1 + \frac{\sigma^2}{\mu^2} \right)$$

- Degree variance:

$$\sigma^2 = \frac{1}{n} \sum_{v \in V} d_v^2 - \left( \frac{1}{n} \sum_{v \in V} d_v \right)^2$$

# Friendship Paradox

- Proof:

- $$\frac{\sum_{(u,v) \in E} d_v}{2m} =$$
- $$= \frac{\sum_{v \in V} d_v^2}{2m} = \text{(because } v \text{ appears } d_v \text{ times in sum over } E)$$
- $$= \frac{\sum_{v \in V} d_v^2}{\mu n} = \text{(because avg degree is } \mu = 2m/n)$$
- $$= \frac{\sigma^2 + \left(\frac{1}{n} \sum_v d_v\right)^2}{\mu}$$

# Why is it important?

- Epidemiology:
  - Best protection for a population with a given budget?
  - Assume social network is not knowable globally
- Two strategies:
  - (a) immunize a random set of people
  - (b) immunize random friends of a random set of people
- Friendship Paradox:
  - (b) better than (a)!
  - Bias towards “higher-degree friends”
- Other applications:
  - Finding good monitors, trend-setters, etc.



# The observer matters

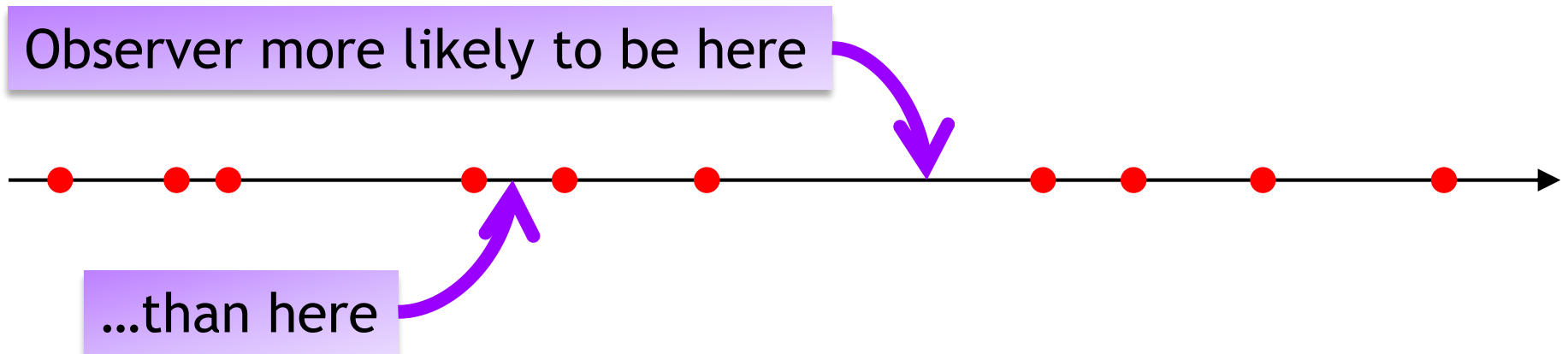
- Other examples:
  - Occupancy distribution:
    - Suppose a train is full 50% of the time, and empty 50% of the time
    - Observer: train is full 100% of the time



The observer  
only sees this

# The observer matters

- Waiting time:
  - Suppose buses arrive a Poisson( $\lambda$ ) point process
  - Average interarrival interval:  $1/\lambda$
  - Observer point of view:
    - Residual time (until next bus): mean =  $1/\lambda$
    - Since last bus: mean =  $1/\lambda$
    - Mean observed interval length:  $2/\lambda$  !



# Summary and lessons

- Herding
  - Following others' decisions: natural social mechanism, can lead to suboptimal global behavior
  - Information cascades: watching others can wash out primary information
  - Rich-get-richer: huge differences in {wealth/degree/influence/membership/...}, winner-takes-all markets
- Observing
  - Choice of observer - sampling bias
  - Paradox: average friend is more popular than average individual
- Next week:
  - Processes on networks: epidemics, sampling

# References

- [D. Easley and J. Kleinberg, Networks, Crowds, and Markets (chapter 16), 2010]
- [Grossglauser & Thiran, Models and Methods for Random Networks (class notes)]