

Social and Information Networks 3: Processes

Internet Analytics (COM-308)

Prof. Matthias Grossglauser
School of Computer and Communication
Sciences

EPFL

Overview

- Computing measures over very large networks
 - Sampling
 - Random Walks on graphs
 - Handling degree bias: weighted estimator
 - Handling locality bias: mixing time of random walk
 - Mixing time depends on network structure
 - Conductance bound
- Epidemics
 - Models for spread of diseases, ideas, etc.
 - Individual evolution of infection: susceptible/infected/recovered
 - Network model: fully mixing vs structure
 - Phase transition: sharp jump from small to large epidemic

Measuring large networks

- Some networks are impossibly large:
 - Facebook: ~ 2.4bn members
 - Web: ~ 600m websites, 50bn pages
 - Constantly evolving/expanding
- Obstacles:
 - Storage
 - Computation
 - If you don't own it: cost of crawling
 - Network & server resources
 - Countermeasures implemented by service providers
 - Privacy, trade secrets, value of aggregated data

How to compute statistics?

- Assumption:
 - We cannot know the whole graph
 - But we can discover and explore a small fraction
- Examples:
 - What % of users are {female, male} on FB?
 - What fraction of web pages are in French?
- Approaches:
 - How to traverse the graph to obtain accurate statistics?

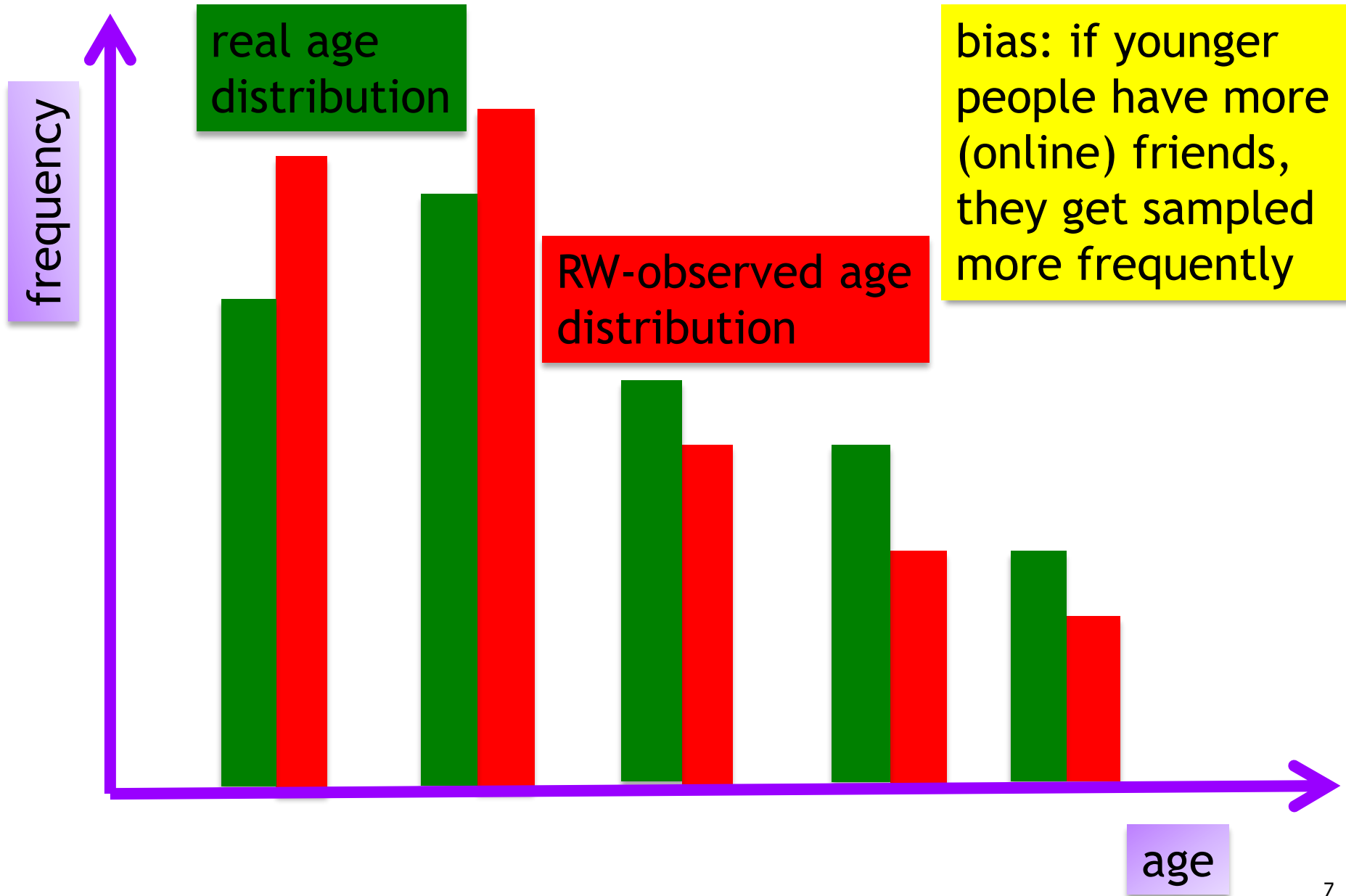
Node statistics in large networks

- Node sampling
 - Urn model: select every node with prob. $1/n$ indep. with replacement; compute average over many samples
 - Problem: usually not available, because we only have “neighbors of current node”!
- Breadth-First Search
 - Problem: “locality bias”
 - E.g.: starting node is a page in English → most nearby pages probably are as well
- Depth-First Search
 - Advantage: avoid locality bias
 - Problem: bias in ordering of links
 - E.g.: alphabetical → only visit people named “A*”; general links before specific links → favor generic

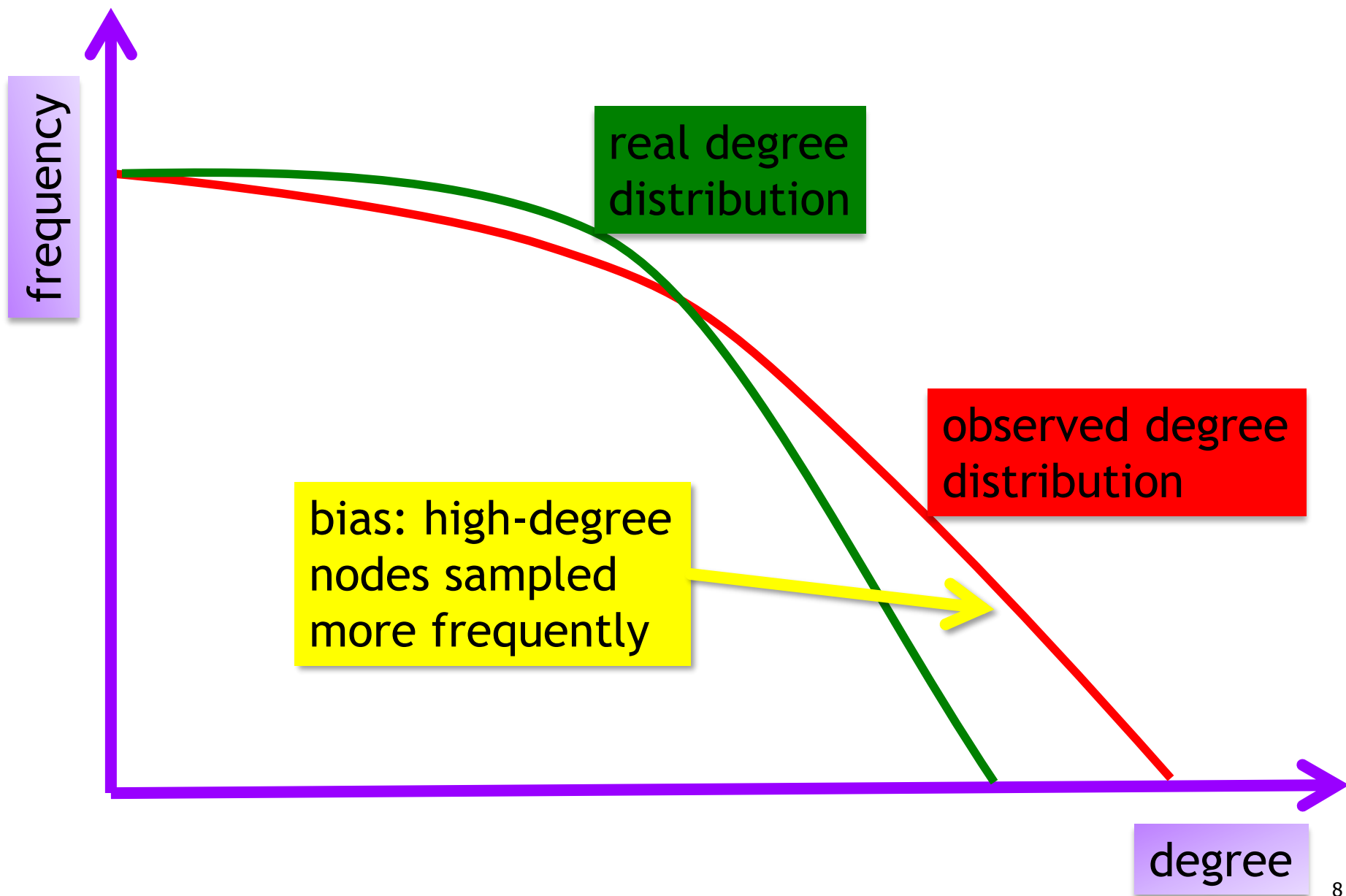
Random walks on graphs

- Random Walk
 - Advantage: no ordering bias (by def); no locality bias (under some conditions)
 - A bit like DFS with shuffled neighbors (but RW can return)
- Undirected graph $G(V, E)$
 - Assume connected (otherwise assume G is the GC for the actual network)
 - Random Walk:
 - Discrete time t
 - Node at time t : $X_t \in V$
 - At each time step, go to a neighbor of X_t uniformly at random $\rightarrow X_{t+1}$

Example of naïve RW sampling



Example of naïve RW sampling



Random walk as Markov chain

- Transition matrix P :

- $$P = \begin{cases} p_{ij} = 1/d_i & (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

- If $G(V, E)$ is undirected, connected and non-bipartite, then $\{X_t\}$ is an ergodic (irreducible, aperiodic) Markov chain
- Ergodicity:
 - Stationary distribution π
 - $p_{ij}(t) = \text{P}(\text{at } j \text{ after } t \text{ steps} \mid \text{starting at } i)$
 - $p_{ij}(t) \rightarrow \pi_j$ for all $i, j \in V$
 - RW “forgets” starting point i

Stationary distribution π

- Lemma:
 - $\pi \propto [d_1, d_2, \dots, d_n]$
- Proof:
 - Def of stationary distribution: $\pi = \pi P$
 - $[d_1, d_2, d_3, \dots, d_n]P = x$
 - $x_j = \sum_i d_i p_{ij} = \sum_i 1_{\{(i,j) \in E\}} = d_j$
 - $[d_1, \dots, d_n]$ is a left-eigenvector with eigenvalue=1 \rightarrow stationary distribution proportional to $[d_1, \dots, d_n]$
- Intuition:
 - Random walk “sees” uniformly random edges; the nodes visited by RW therefore biased by # of edges = degree
 - Similar to Friendship Paradox!

Obtaining unbiased estimator from RW

- Node statistic $f(i)$
- Would like to know $F = 1/n \sum_{v \in V} f(v)$
- Sampling:
 - Ideal: $P(X_t = v) = 1/n$
 - RW: $P(X_t = v) = \frac{d_v}{\|d\|_1} = \frac{d_v}{2m}$
- Compensate for degree bias:
 - Let RW run for T time steps
 - Compute $\hat{F} = \frac{2m \sum_t f(X_t)/d_{X_t}}{nT}$
- Stationary regime: unbiased $E[\hat{F}] = E[f(X_t)] = F$
 - But we cannot start in stationary regime - instead at a specific state \rightarrow how large does T have to be?

Unbiased estimator

- Because of ergodicity, time average = expectation

$$\frac{2m}{n} \cdot \frac{1}{T} \sum_t \frac{f_{X_t}}{d_{X_t}} \rightarrow \frac{2m}{n} \mathbb{E} \left[\frac{f_X}{d_X} \right], \text{ where } X \sim \pi$$

- $$\frac{2m}{n} \mathbb{E} \left[\frac{f_X}{d_X} \right] = \frac{2m}{n} \sum_i \pi_i \frac{f_i}{d_i} = \frac{2m}{n} \sum_i \frac{f_i}{\sum_j d_j} = F$$

Estimator without knowledge of n, m

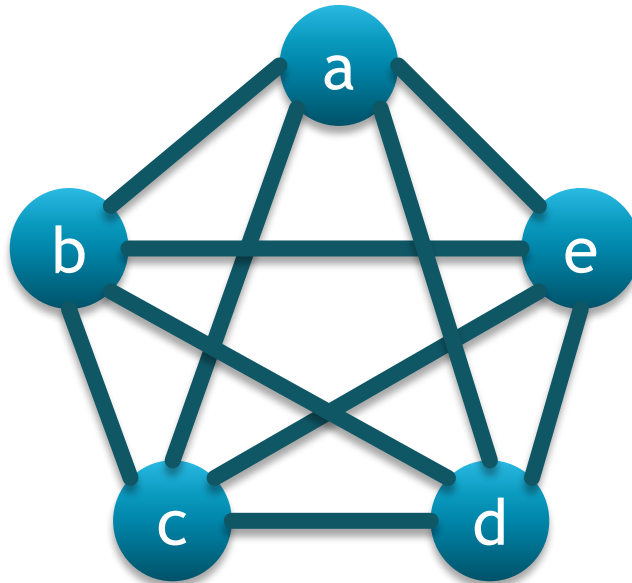
- In practice, we may not know n, m
- Eliminate from estimate:
 - Can estimate normalization constant from sample path

- $$\hat{F} = \frac{\sum_t f(X_t)/d_{X_t}}{\sum_t 1/d_{X_t}}$$

- Denominator: sum of all (random) weights

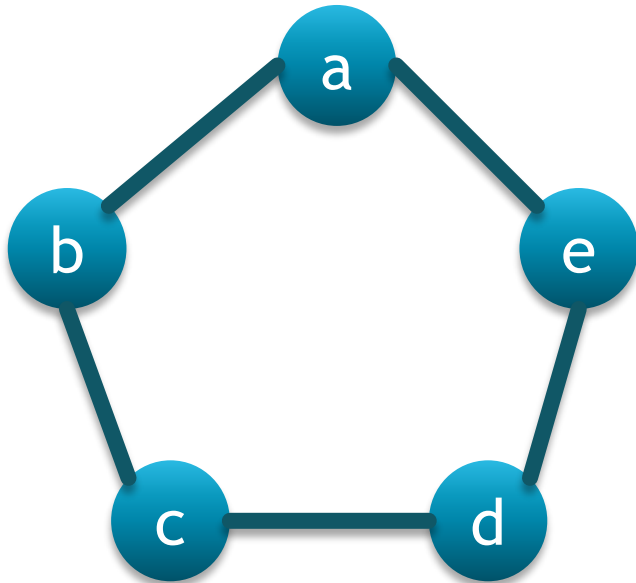
Transient analysis of $\{X_t\}$

- How quickly does RW converge?
- Depends on graph structure!
- Example 1: complete graph K_n :
 - $t = 1: p = (1,0,0,0,0)$
 - $t = 2: p = (0,1,1,1,1)/4$
 - Almost perfect mixing in 1 time step!



Transient analysis of $\{X_t\}$

- Example 2: cycle C_n :
 - $t = 1: p = (1,0,0,0,0)$
 - $t = 2: p = (0,1,0,0,1)/2$
 - $t = 3: p = (2,0,1,1,0)/4$
 - ...



- $n \rightarrow \infty$:
 - After t steps, standard deviation of distribution around start is $\propto \sqrt{t}$
 - Need $\omega(n^2)$ steps to mix
 - Would be better off doing exhaustive deterministic traversal!

RW sampling is worthwhile only if mixing time is $o(n)$

Mixing time of RW

- Theorem:

- $|p_{ij}(t) - \pi_j| \leq \sqrt{\frac{d_j}{d_i}} \lambda^t$

- λ : largest absolute value among $\lambda_2, \dots, \lambda_n$

- Exponentially fast convergence to stationary

- Note: for most graphs:

- $\lambda = \lambda_2$, i.e., second-largest eigenvalue of P

- How to determine λ_2 ?

Conductance bound

- Conductance with respect to a set:

- $$\Phi(S) = \frac{|\delta S|}{2m \pi(S)\pi(S')}$$

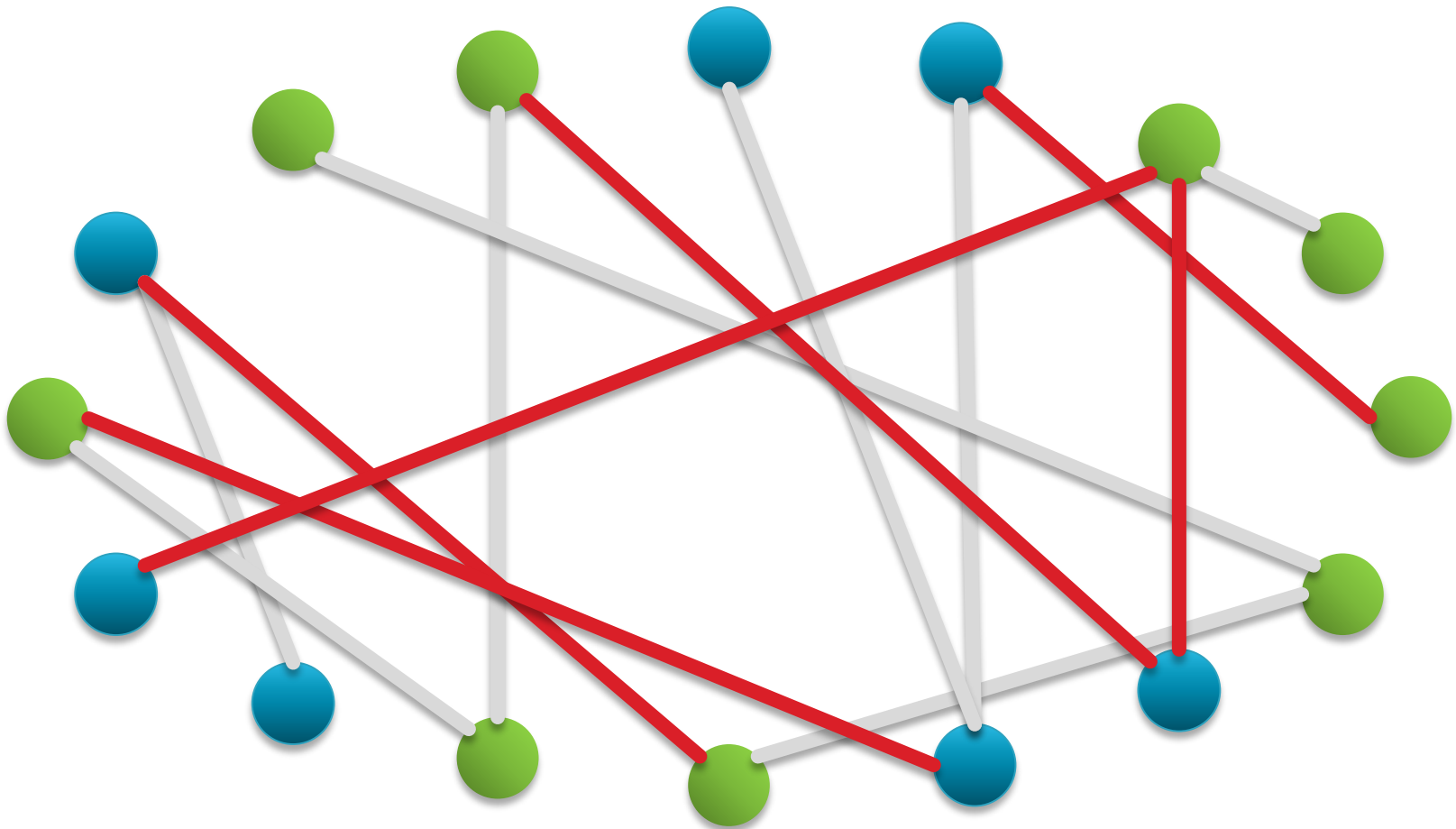
- where S is a nonempty set of nodes, $S' = V \setminus S$
- and $\pi(S) = \sum_{v \in S} \pi(v)$

- Conductance of the whole graph:

- $$\Phi = \min_{S \subset V} \Phi(S)$$

Conductance

- Red edges = δS



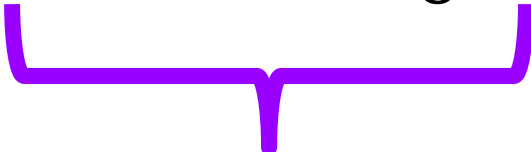
Interpretation of conductance

- Independent sampling (IS) process:
 - Sample each node u indep. with prob. π_u
 - Same state probabilities as RW, but memoryless
- Probability for IS to move $S \leftrightarrow S'$: $2\pi(S)\pi(S')$
- Probability for RW to move $S \leftrightarrow S'$: $|\delta S|/m$
 - Because prob. that RW traverses specific edge = $1/m$
- Conductance: $\Phi(S) = \frac{|\delta S|}{2m \pi(S)\pi(S')}$
 - Ratio of rate of switching of RW vs “ideal” memoryless IS
 - Worst case over all subsets of nodes
 - This graph-based quantity is often easier to compute/bound than λ_2

Conductance bound of mixing rate

- Theorem (Cheeger): bound on spectral gap:

- $1 - \Phi \leq \lambda_2 \leq 1 - \frac{\Phi^2}{8}$



Large conductance \rightarrow fast mixing

- Conductance “finds the bottleneck”
 - If any set S is poorly connected \rightarrow slow mixing
- Setting $T = c / \log \lambda_2$ ensures that RW has “almost” forgotten initial state

RW sampling: directed graph

- What if G is directed?
 - Could try to “undirect” G by adding (v, u) for every (u, v)
 - Not always easy: e.g., how to know incoming links to a web page?
- No straightforward way to determine π from local graph properties
- Sampling the web in an unbiased way is a challenge!

Epidemics

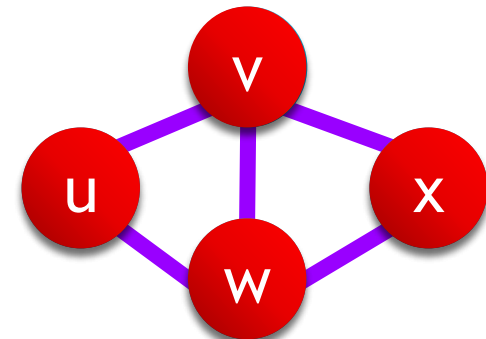
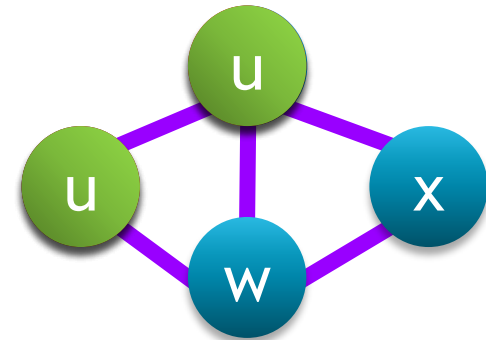
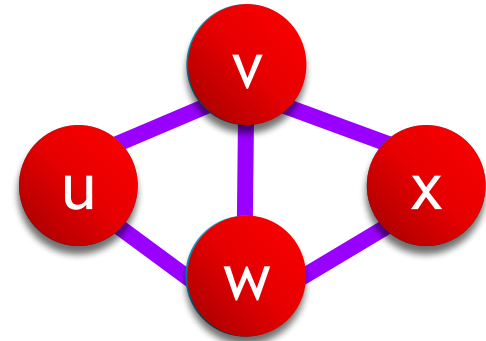
- Many processes in social and information networks have a notion of “infecting neighbors”
 - Infectious diseases: plague, cholera, HIV, corona,...
 - Ideas, preferences, believes - with applications in marketing and advertisement
 - Re-tweets, shares, chain letters, slash-dot,...
 - Computer viruses, internet worms,...
 - Non-infectious cultural/behavioral diseases: smoking, obesity, suicide,... (controversial)

Contact network

- Fully mixing:
 - Assumption that any two nodes (u, v) interact i.i.d.
- Restricted: context-dependent:
 - Airborne diseases (e.g., covid-19, SARS, influenza): physical proximity
 - Sexually transmitted diseases (e.g., HIV): sexual contact
 - Ideas, believes, behavioral patterns (e.g., music preference, smoking): close social relationship
 - Information epidemic (e.g., “gangnam style” video): online social network link
 - Computer security (e.g., virus, malware): online interaction through spam email, compromised web server or the like

Epidemics: models

- SI:
 - Susceptible \rightarrow Infected
 - Example: incurable infectious disease; knowledge
- SIR:
 - Susceptible \rightarrow Infected \rightarrow Recovered/Removed
 - Example: measles; computer virus (hopefully corona)
- SIS:
 - Susceptible \rightarrow Infected \rightarrow Susceptible \rightarrow Infected \rightarrow ...
 - Example: seasonal flu or cold
 - All-S is an absorbing state



Fully mixed SI model

- Each individual has contact rate β
 - Contact of $(S, I) \rightarrow (I, I)$ after contact
- $i(t)$: fraction infected at time t
- $s(t) = 1 - i(t)$: fraction susceptible at time t
- Large system ($n \rightarrow \infty$) - treat as continuous (“mean field approximation”):
 - $\frac{di}{dt} = \beta(1 - i)i$
 - Solution: logistic growth equation:

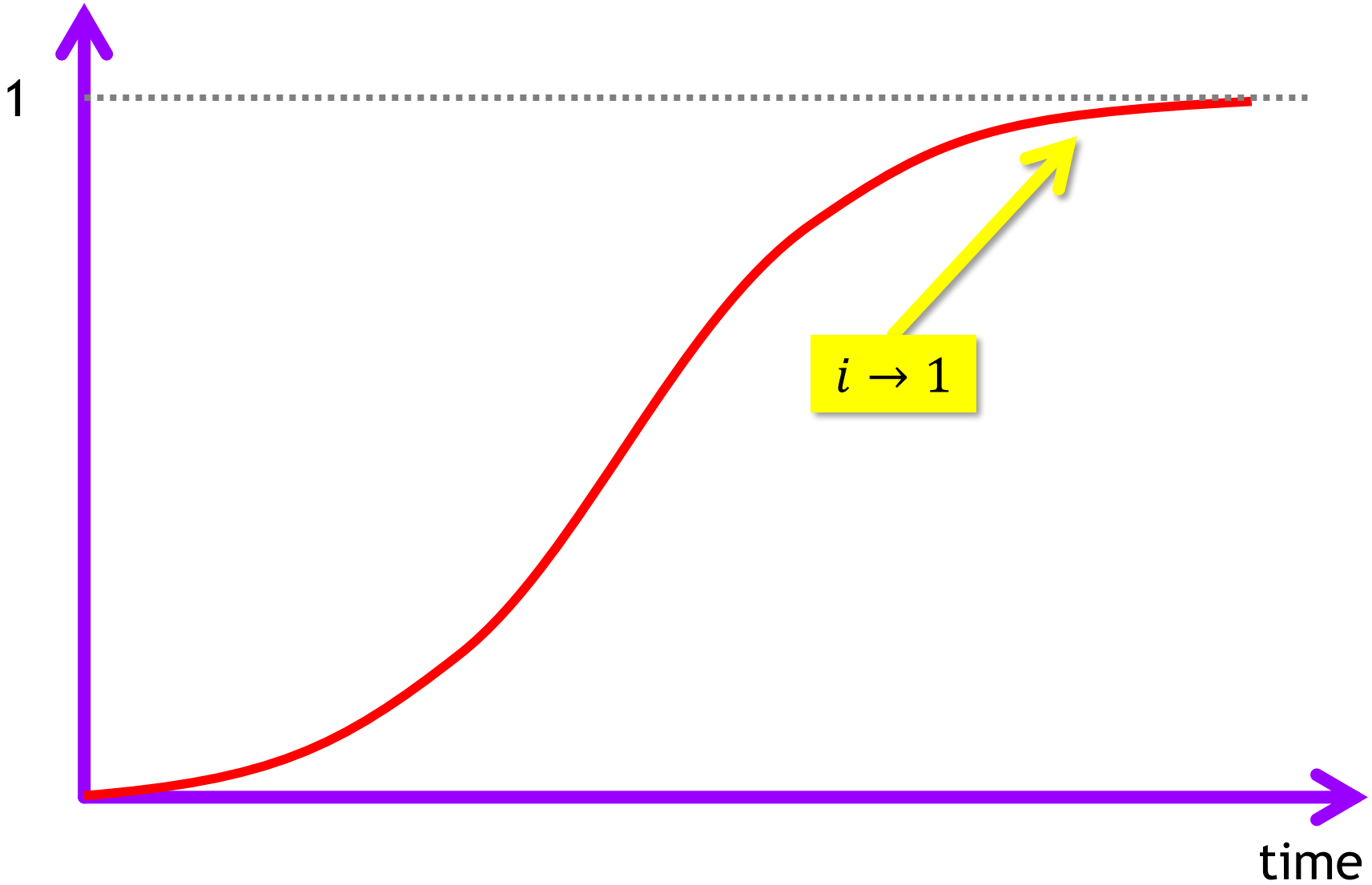
$$i(t) = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}}$$

Logistic growth equation

- $i(t) = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}}$
- $$\begin{aligned}\frac{di}{dt} &= \frac{\beta i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}} - \frac{\beta i_0 e^{\beta t} i_0 e^{\beta t}}{(1 - i_0 + i_0 e^{\beta t})^2} \\ &= \beta(i - i^2)\end{aligned}$$
- $i(0) = i_0$
- $i(\infty) = 1$

SI “s-curve”: logistic growth equation

size of infection

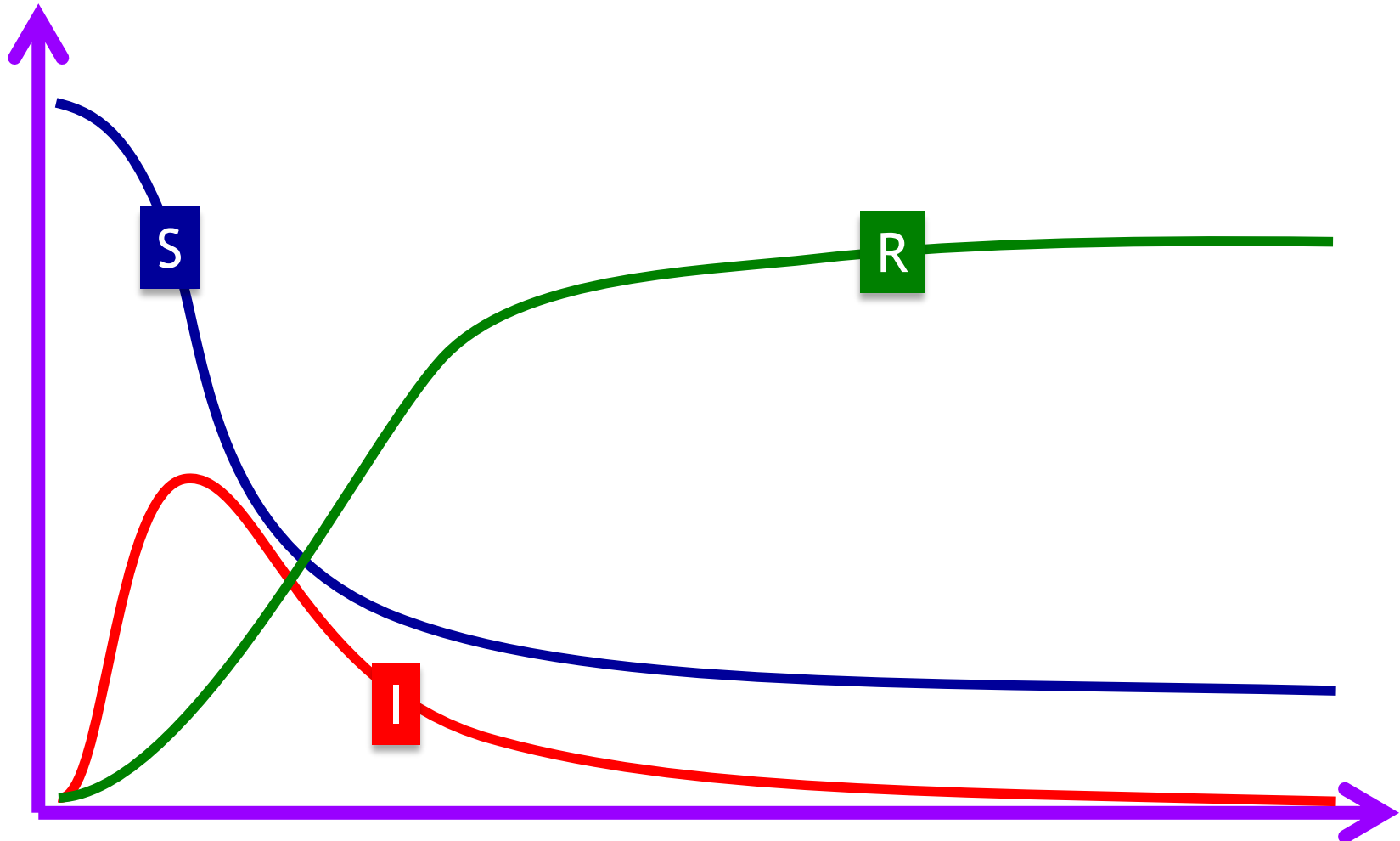


Fully mixing SIR model

- Similar to SI model, but:
 - Infected individual stays in I-state for time $\sim \exp(1/\gamma)$ (iid), then enters R-state forever
 - $\frac{ds}{dt} = -\beta si$; $\frac{di}{dt} = \beta si - \gamma i$; $\frac{dr}{dt} = \gamma i$
 - No closed form solution
- Basic reproductive number R_0 :
 - Informally: expected # of infections by patient zero
 - $R_0 = \beta/\gamma$

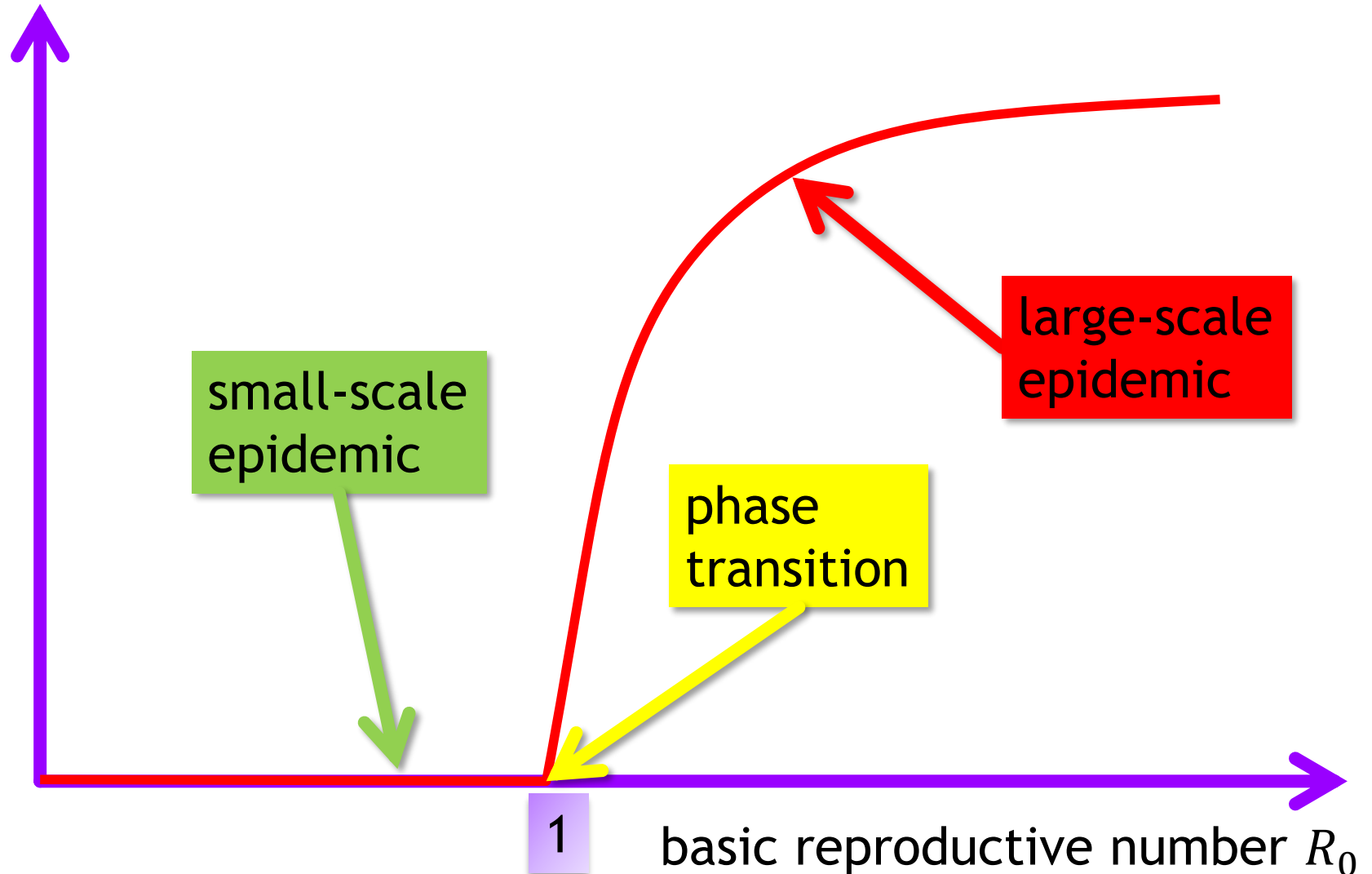
SIR: time evolution in finite system

- Typical evolution of 3 states (S,I,R)



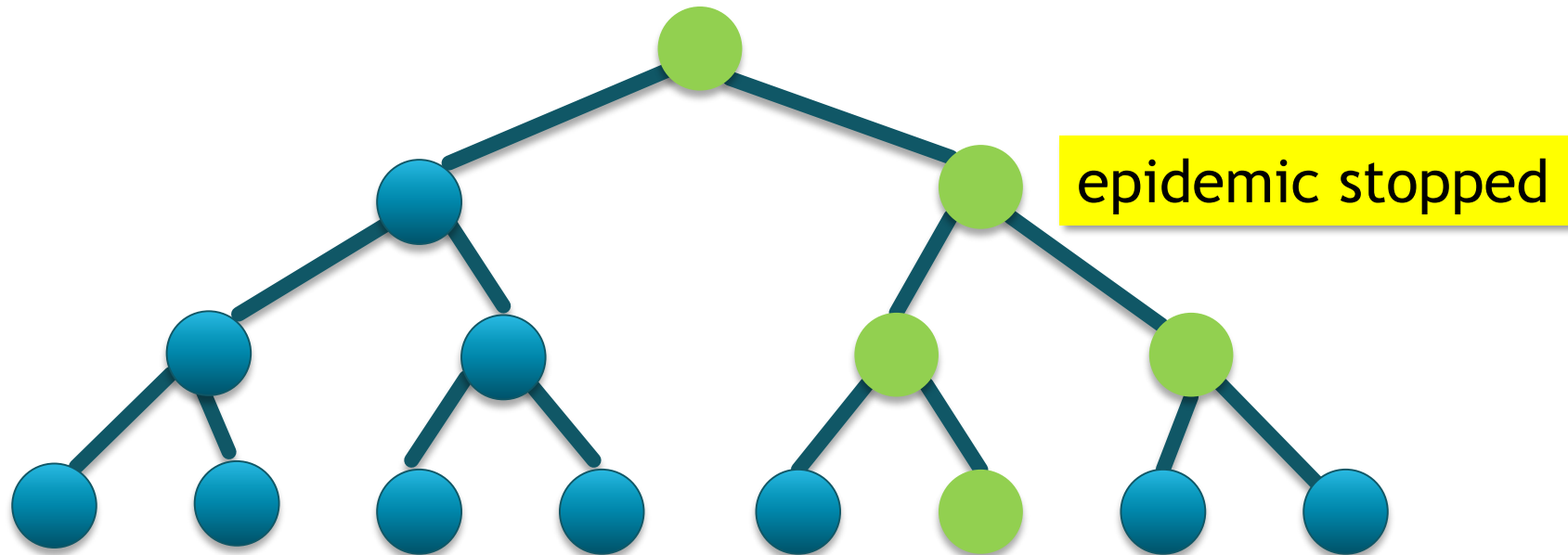
SIR: phase transition

final size of infection



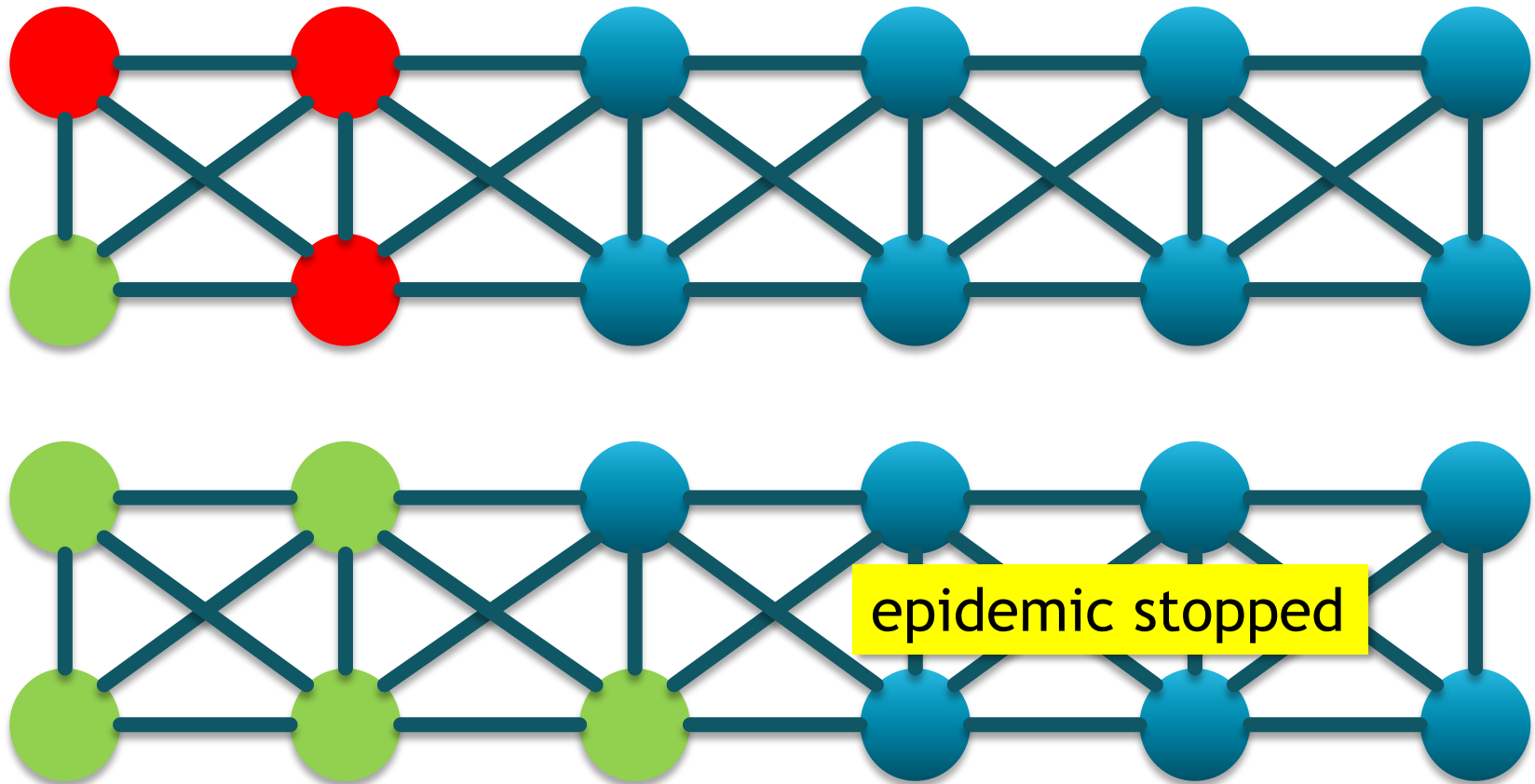
SIR in networks

- Epidemic is not guaranteed even when $R_0 > 1$
- Counterexample 1: tree network \rightarrow branching process
 - Even if $R_0 > 1$, extinction probability is > 0



SIR model in networks: bottleneck

- Counterexample 2: ladder network
 - Epidemic guaranteed to eventually stop, regardless of R_0



Summary & lessons

- Graph sampling:
 - Estimating node statistics without access to whole network
 - Beware the biases!
 - Degree bias: easy to compensate in undirected graphs
 - Locality bias: related to mixing time
- Epidemics
 - Models for many types of processes of local influence
 - SI/SIS/SIR
 - SIR has sharp phase transition
- Labs:
 - Unknown graph, try to crawl and compute statistic of interest
 - Stop an epidemic by removing some edges from network

References

- [L. Lovasz, Random Walks on Graphs: A Survey, Combinatorics, 1993]
- [M.E.J. Newman: Networks: An Introduction, Oxford, 2010 (chapter 17)]
- [D. Easley & J. Kleinberg: Networks, Crowds, and Markets, Cambridge, 2010 (chapter 21)]