Learning Opinions in Social Networks

Vincent Conitzer   Debmalya Panigrahi   Hanrui Zhang

Duke University

Learning “opinions” in social networks

- a social media company (say Facebook) runs a poll
- ask users: “have you heard about the new product?”
- awareness of product <u>propagates</u> in social network
- observe: responses from <u>some random users</u>
- goal: <u>infer</u> opinions of users <u>who did not respond</u>

Learning “opinions” in social networks

more generally, “opinions” can be:

- awareness about a new product / political candidate / news item
- spread of a biological / computer virus

this talk:

- review propagation of opinions in social networks
- how to measure the complexity of a network for learning opinions?
- how to learn opinions with random propagation, when the randomness is unknown?

Related research topics

- learning propagation models: given outcome of propagation, infer propagation model

(Liben-Nowell & Kleinberg, 2007; Du et al., 2012; 2014; Narasimhan et al., 2015; etc)

- social network analysis & influence maximization: given fixed budget, try to maximize influence of some opinion

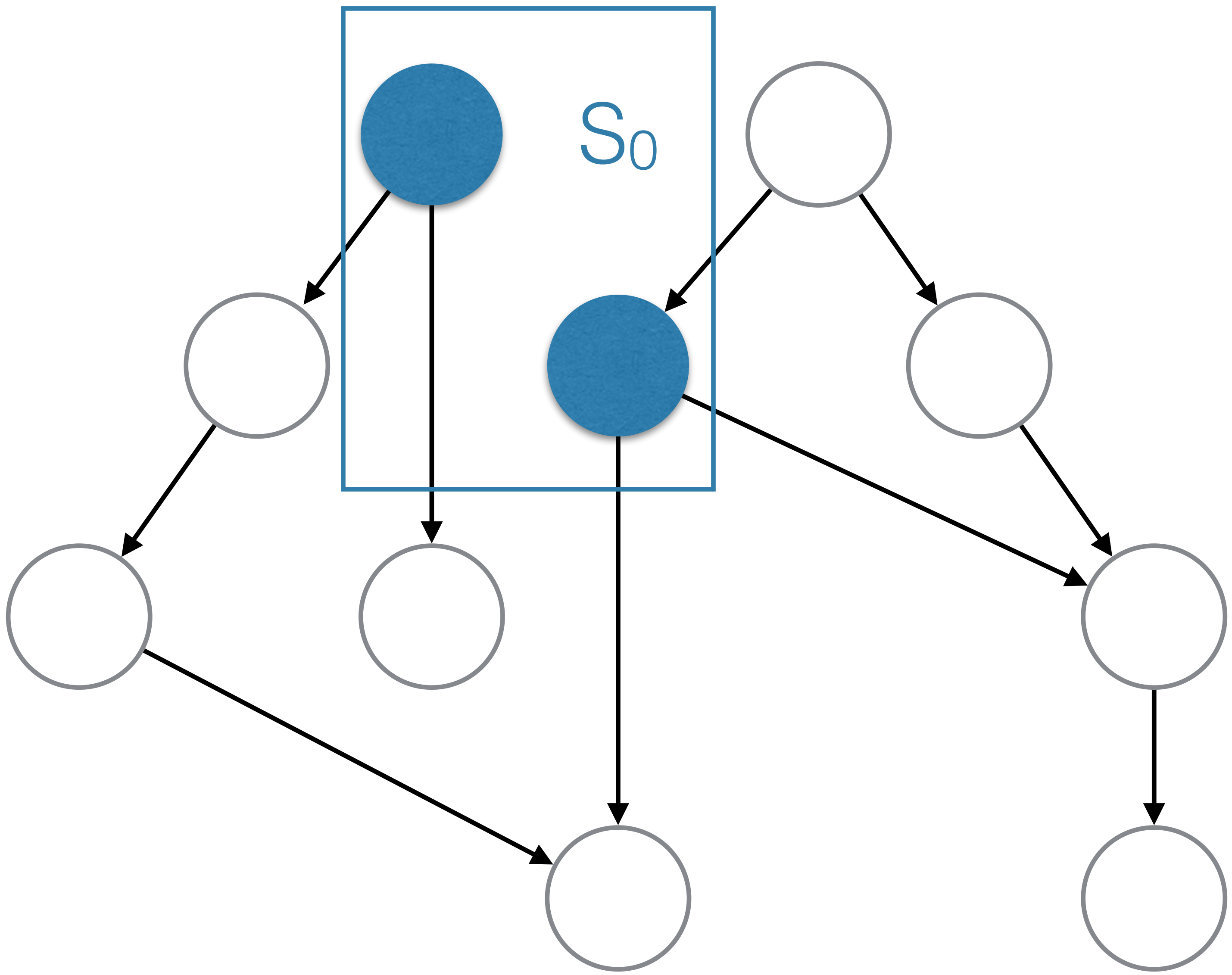
(Kempe et al., 2003; Faloutsos et al., 2004; Mossel & Roch, 2007; Chen et al., 2009; 2010; Tang et al., 2014; etc)

Information propagation in social networks

a simplistic model:

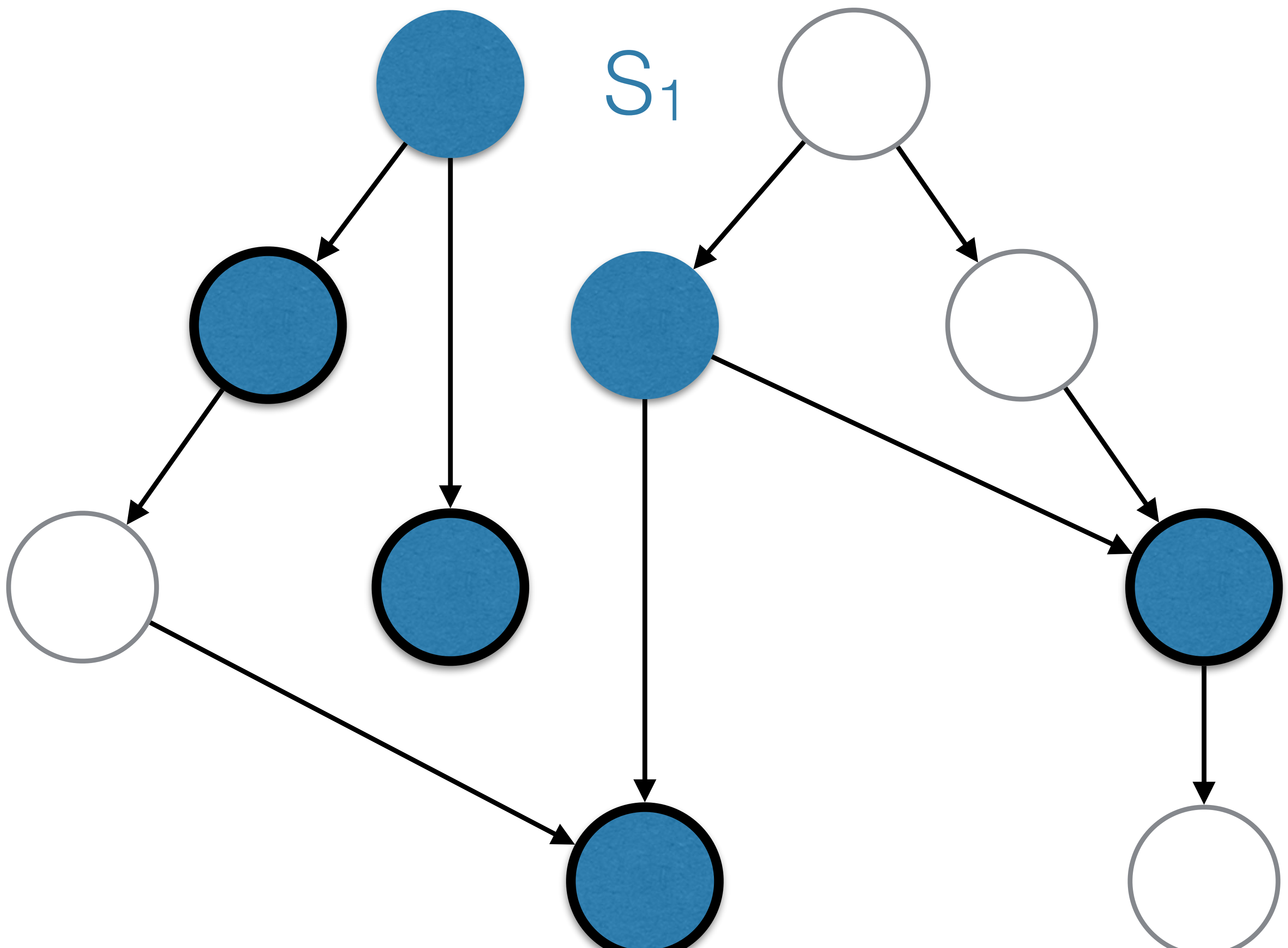
- network is a directed graph $G = (V, E)$
- a seed set S_0 of nodes which are initially informed (i.e., active)
- active nodes deterministically propagate the information through outgoing edges

Information propagation in social networks



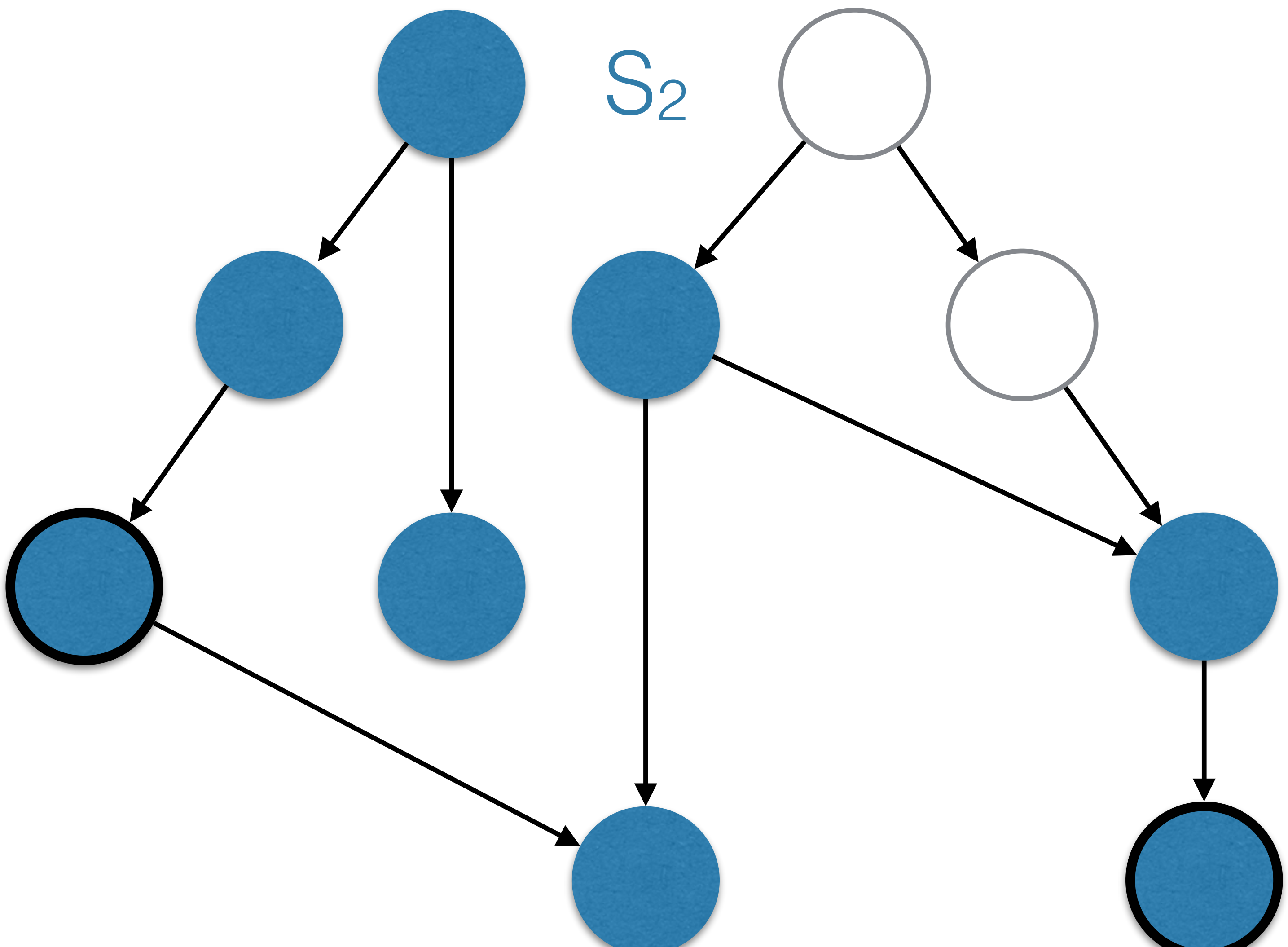
S_0: seed set that is initially active

Information propagation in social networks



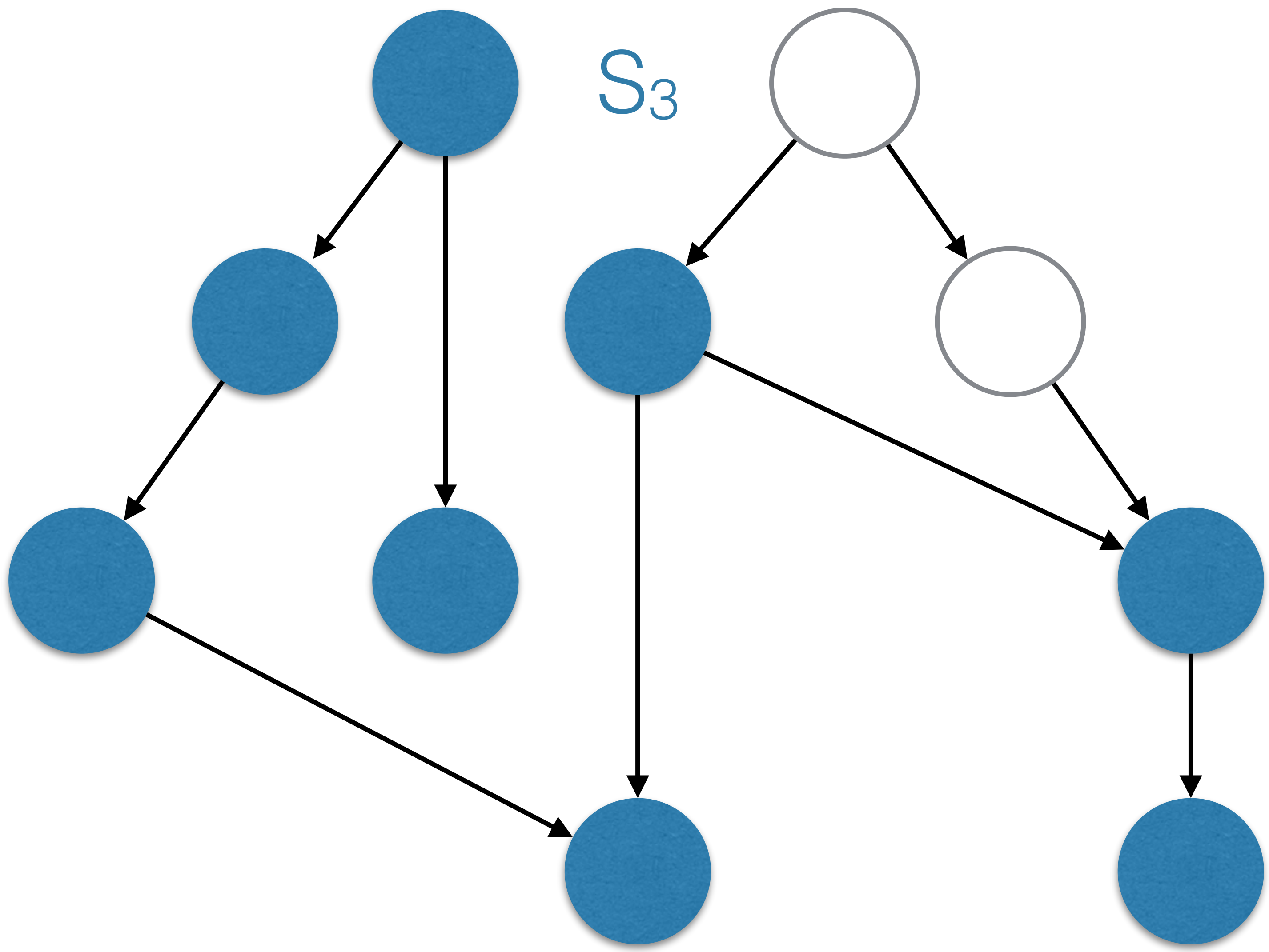
S_1: active nodes after 1 step of propagation

Information propagation in social networks



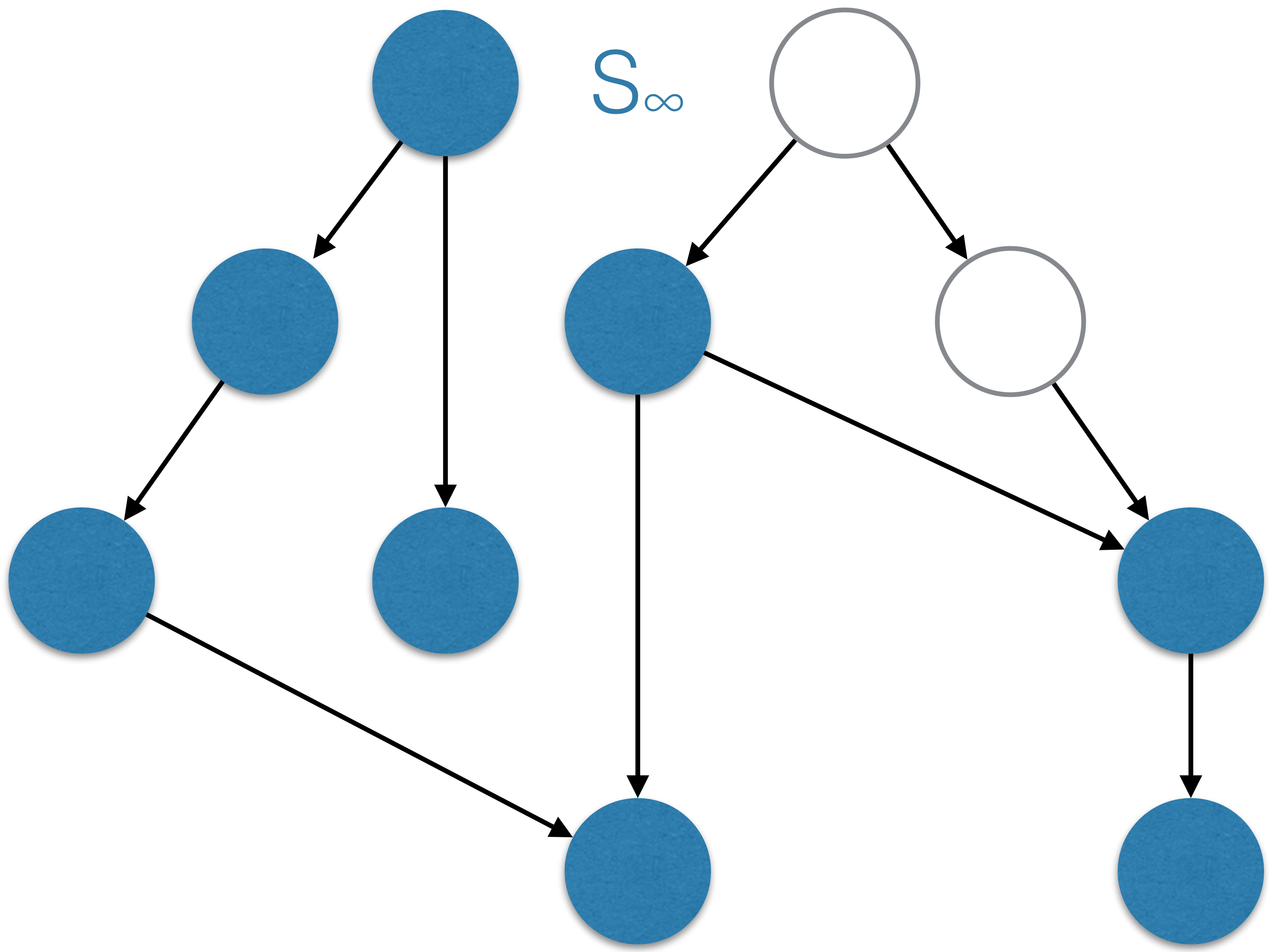
S_2: active nodes after 2 steps of propagation

Information propagation in social networks



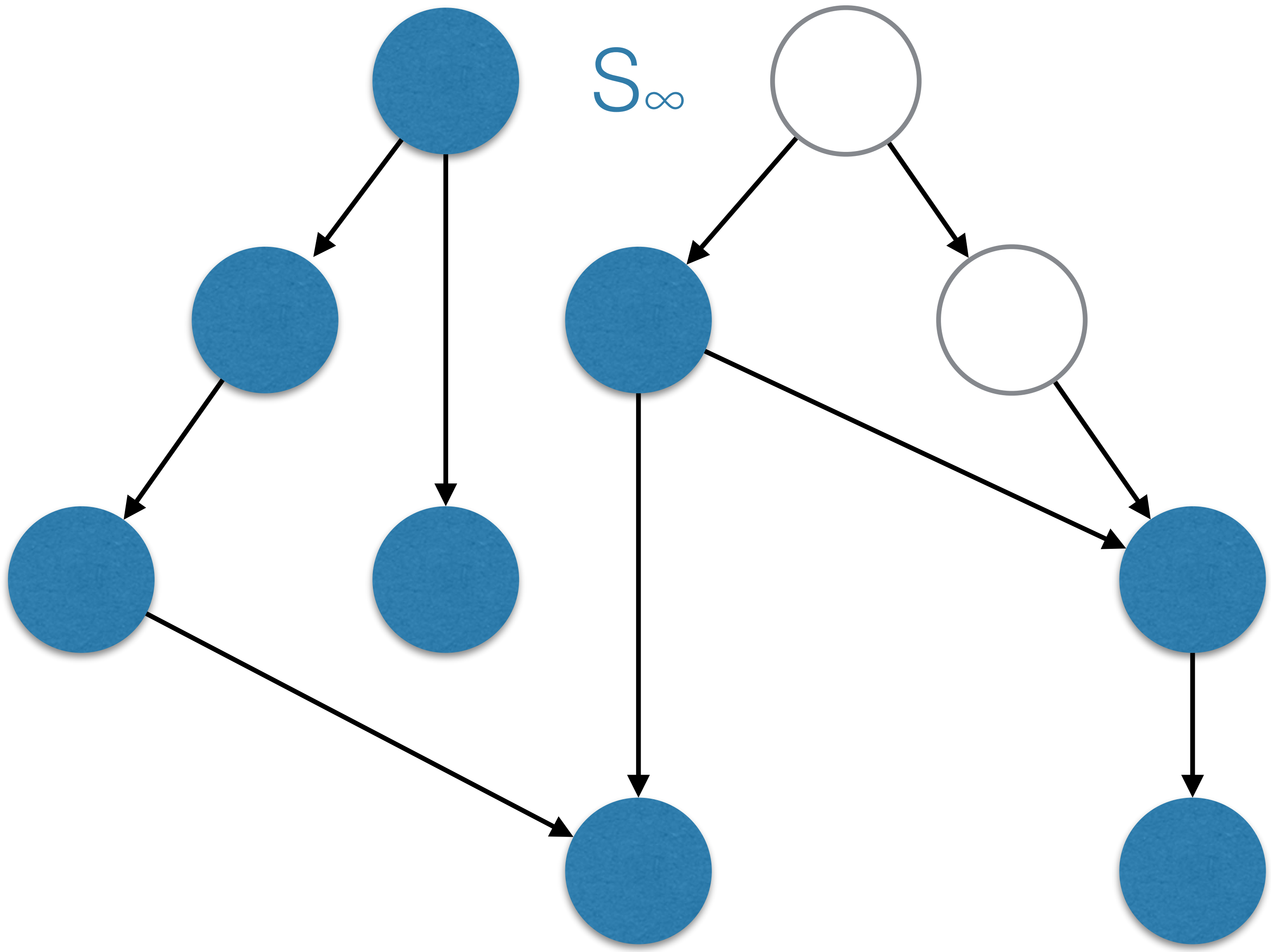
S_3: active nodes after 3 steps of propagation

Information propagation in social networks



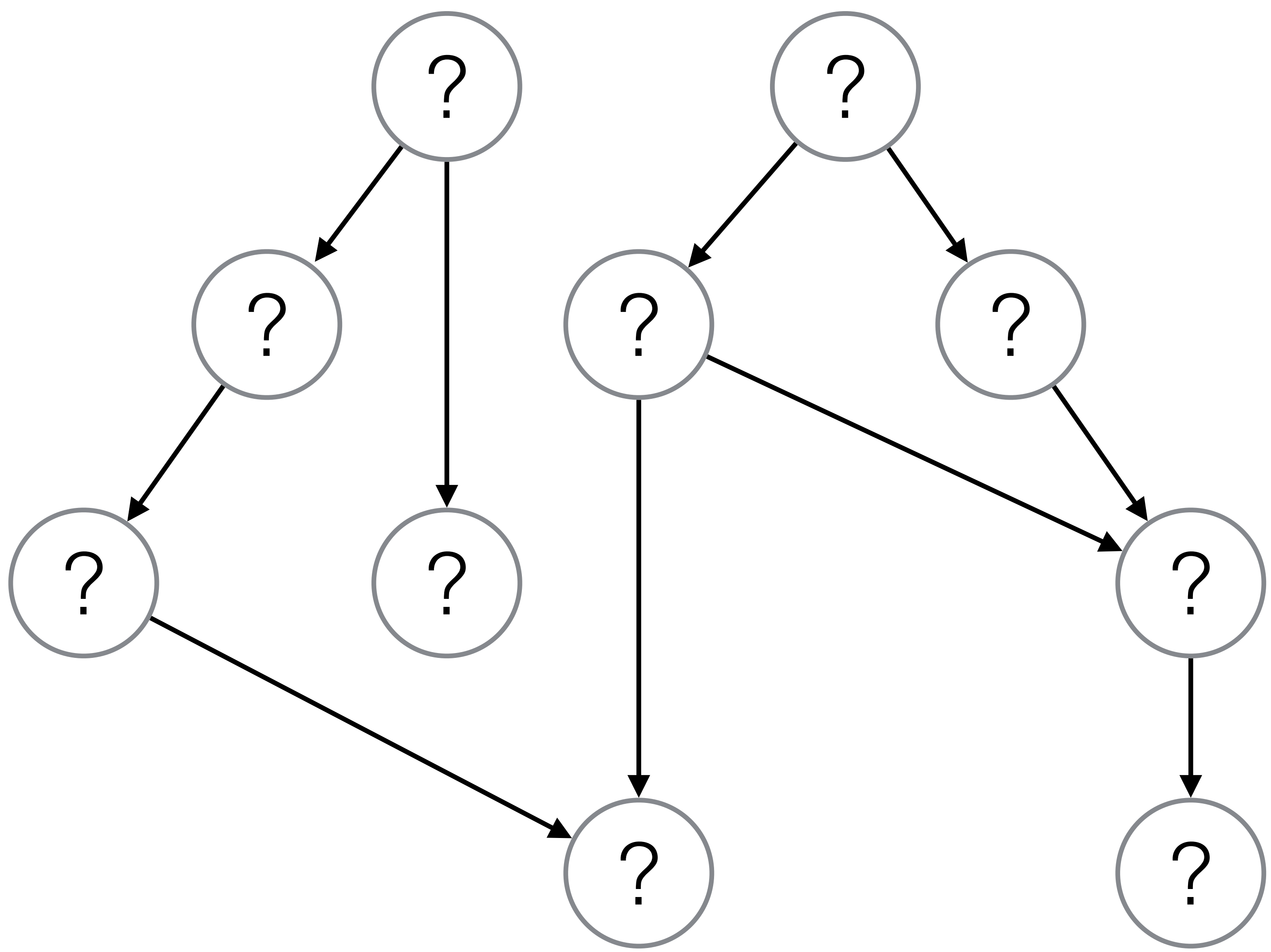
propagation stops after step 2
final active set $S_2 = S_3 = \ldots = S_\infty$

PAC learning opinions

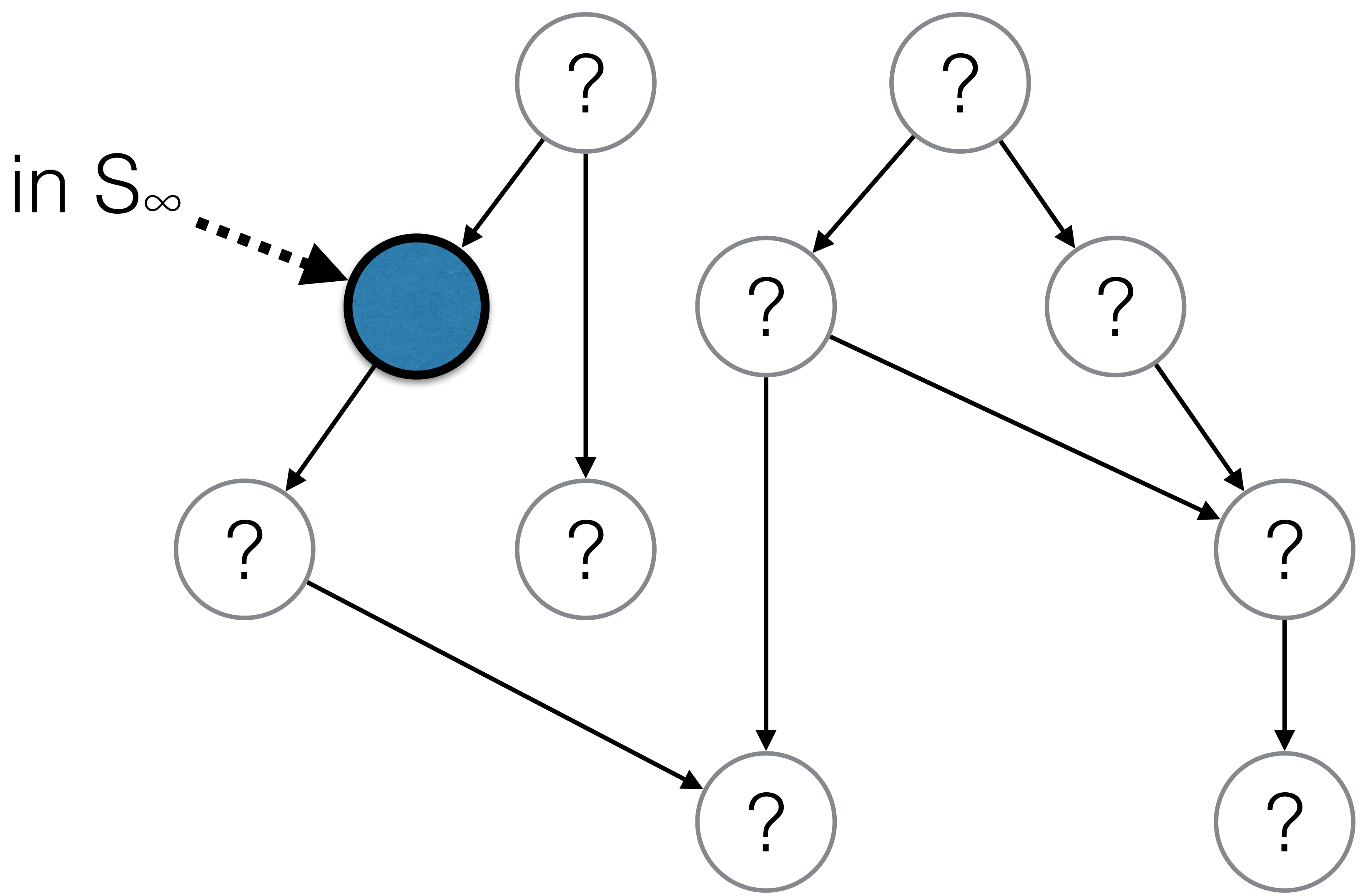


- fix G, unknown seed set S_0 and distribution $\mathscr{D}$ over V
- observe m iid labeled samples $\{(u_i, o_i)\}_i$, where for each i, $u_i \sim \mathscr{D}$, and $o_i = 1$ iff u_i in S_∞
- based on the sample set, predict if u in S_∞ for $u \sim \mathscr{D}$

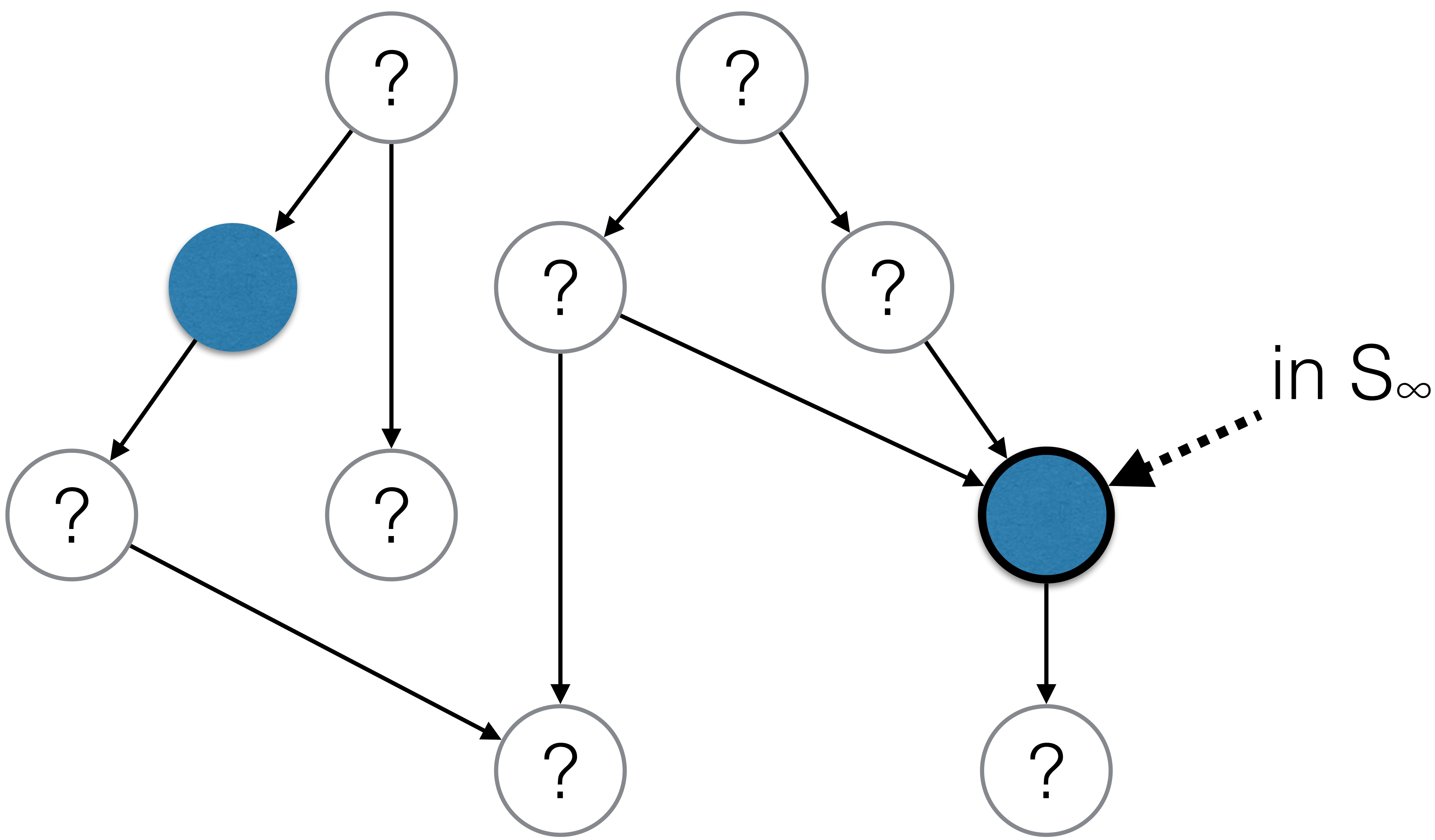
PAC learning opinions



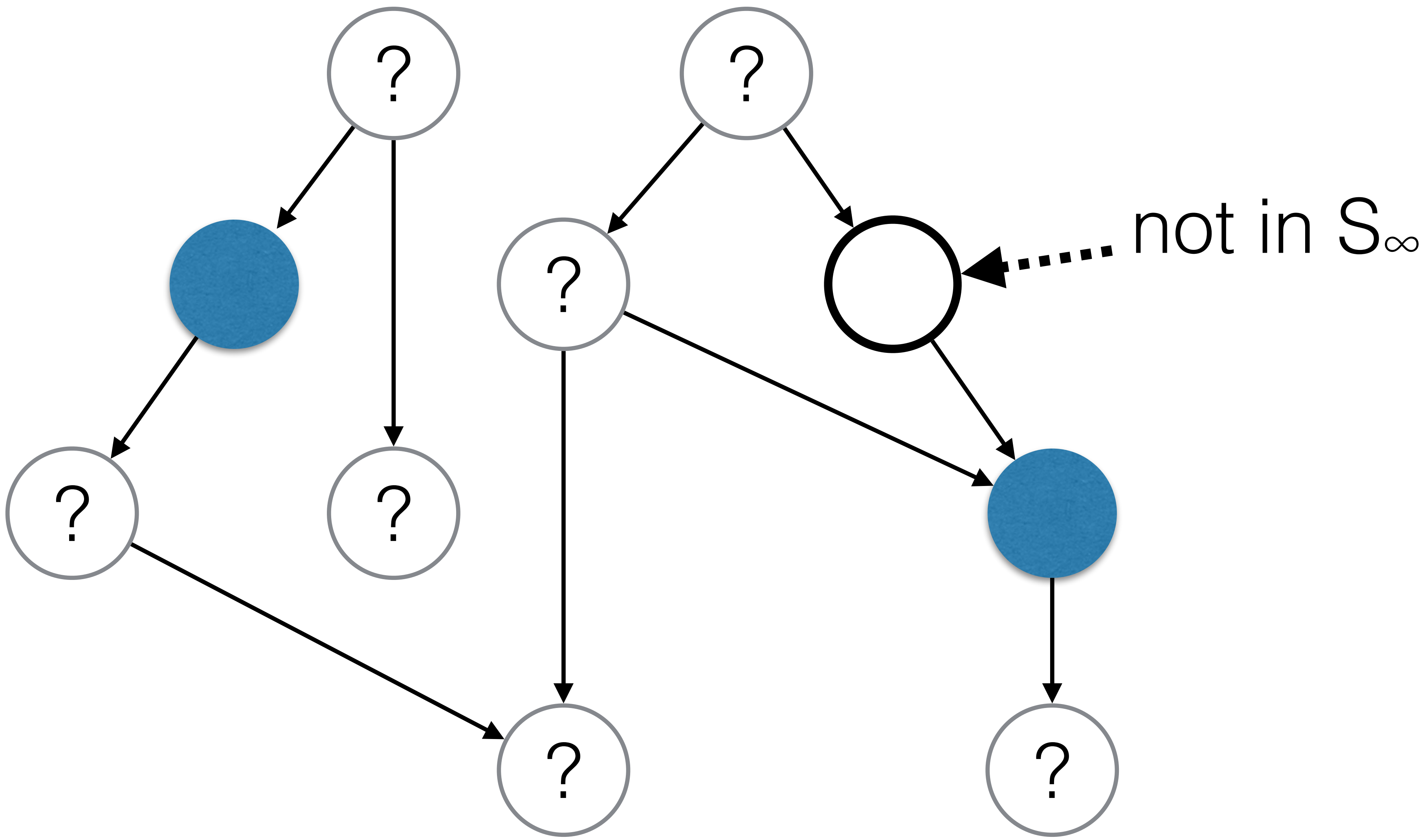
PAC learning opinions



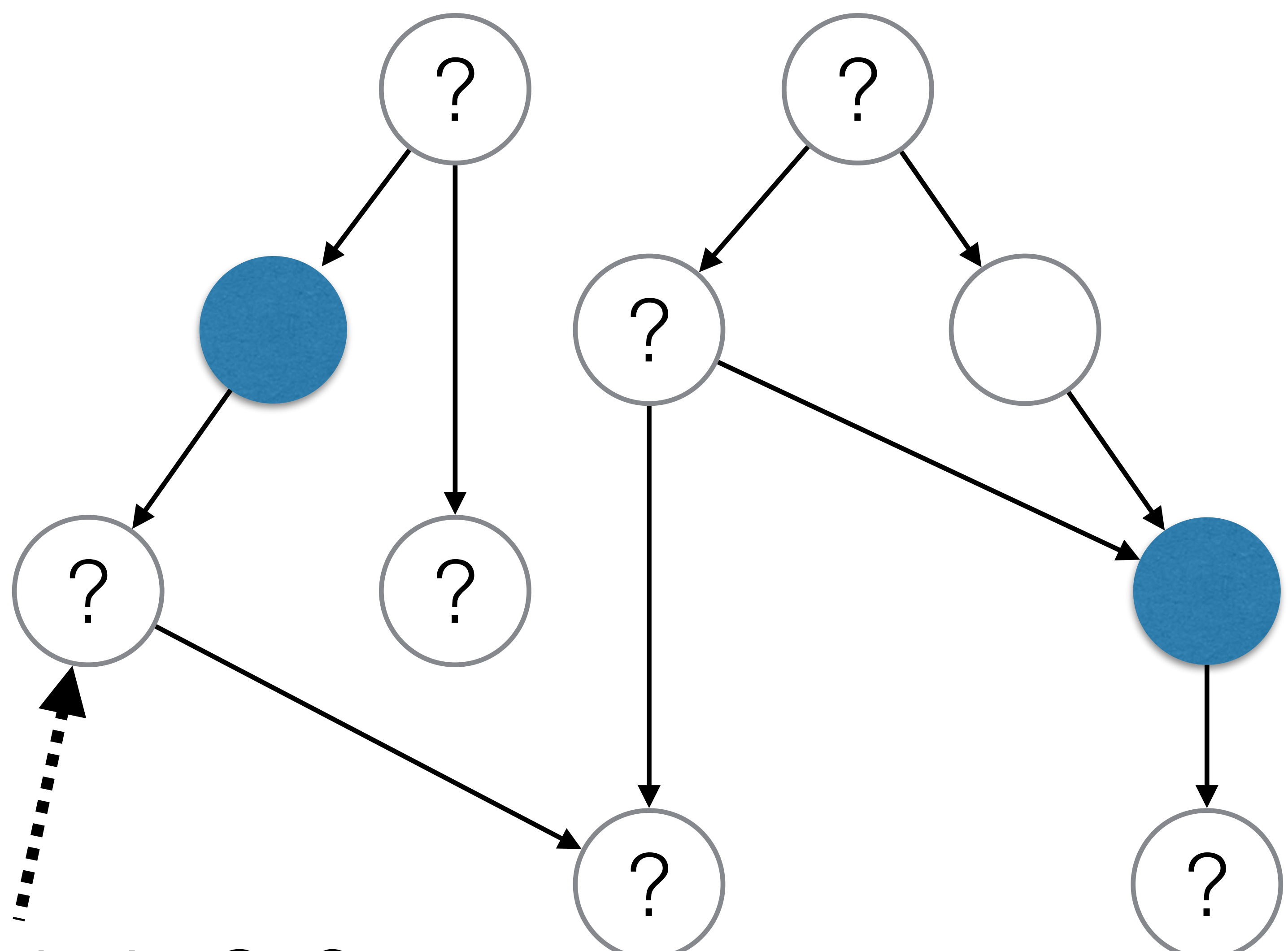
PAC learning opinions



PAC learning opinions



PAC learning opinions

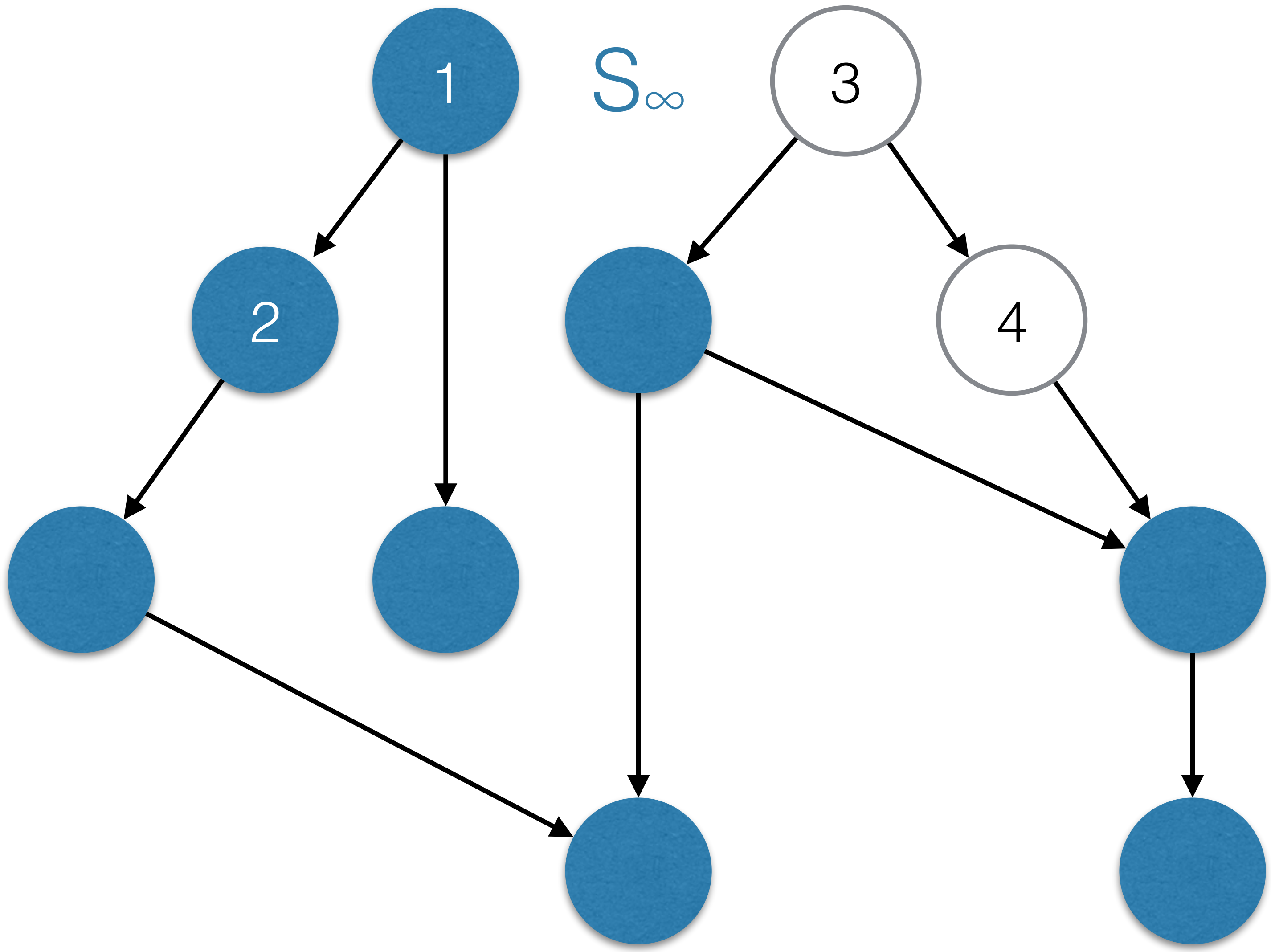


is this node in S_∞?

PAC learning opinions

- key challenge: how to generalize from observations to future nodes to make predictions for
- common sense: generalization is impossible without some prior knowledge
- so what prior knowledge do we have?
- answer: structure of the network

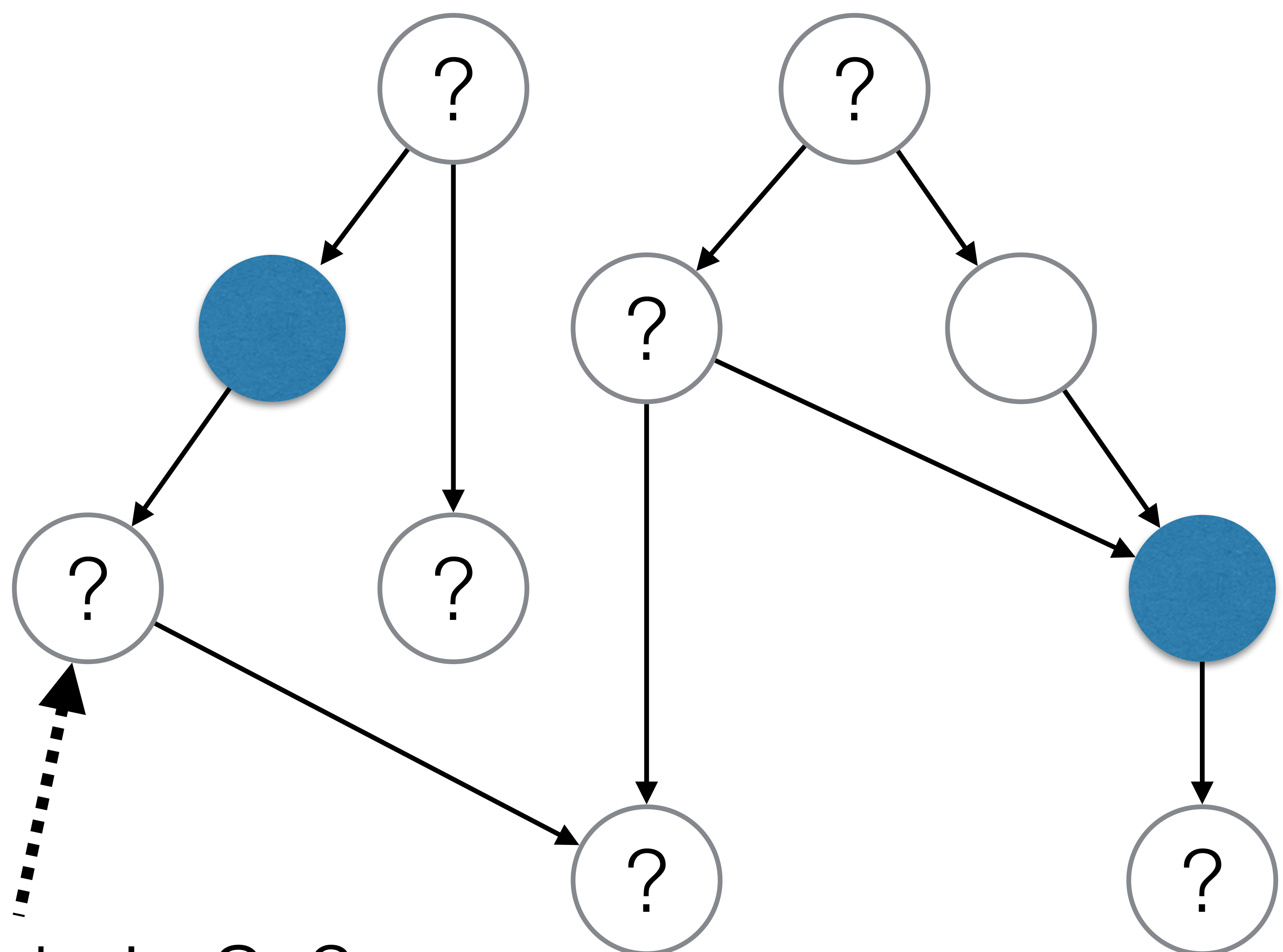
Implicit hypothesis class



for any pair of nodes u, v where u can reach v:

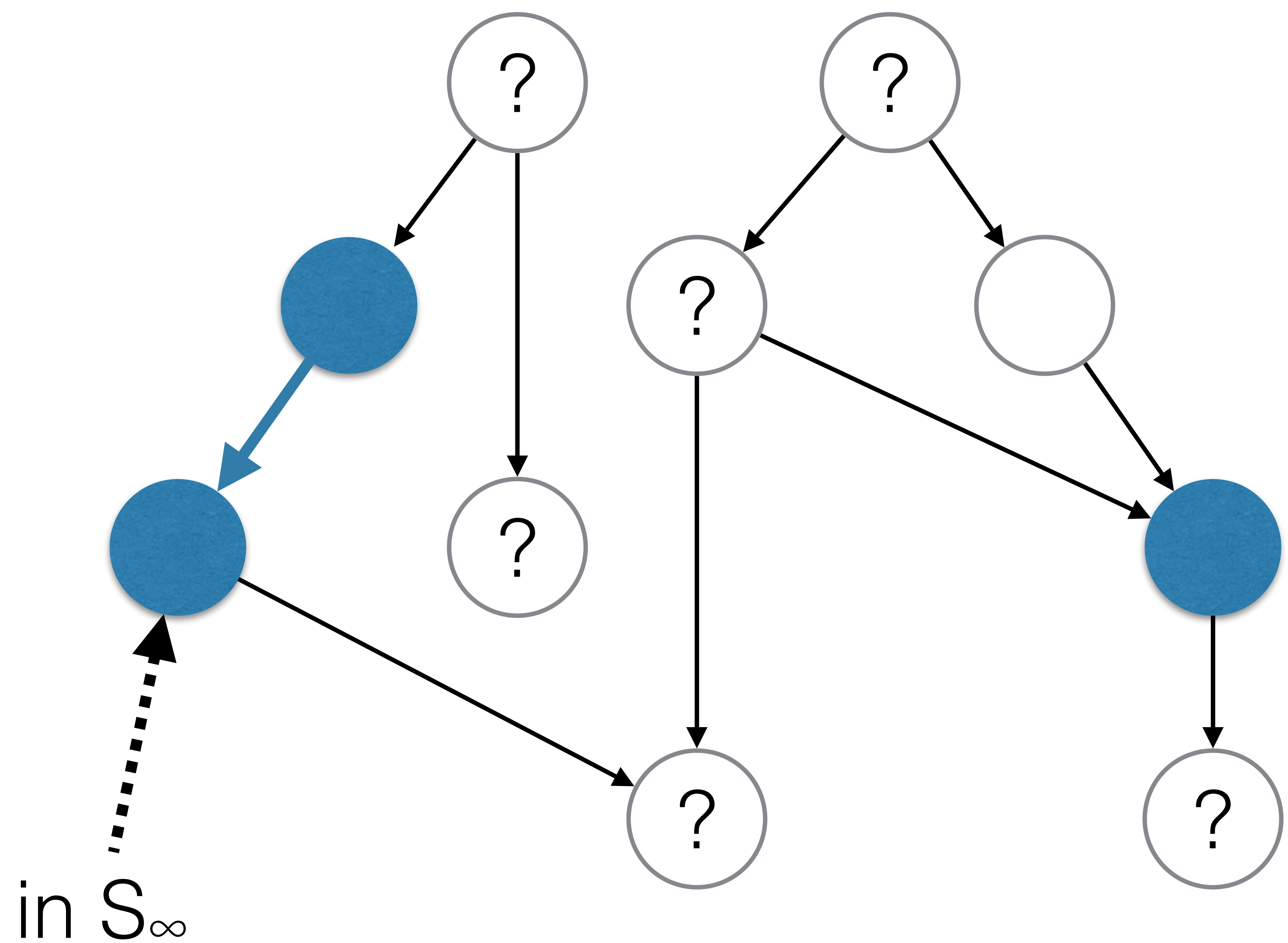
- if u is in S_∞, then v must be in S_∞ (e.g., u = 1, v = 2)
- equivalently, if v is not in S_∞, then u must not be in S_∞ (e.g., u = 3, v = 4)

PAC learning opinions



is this node in S_∞?

PAC learning opinions

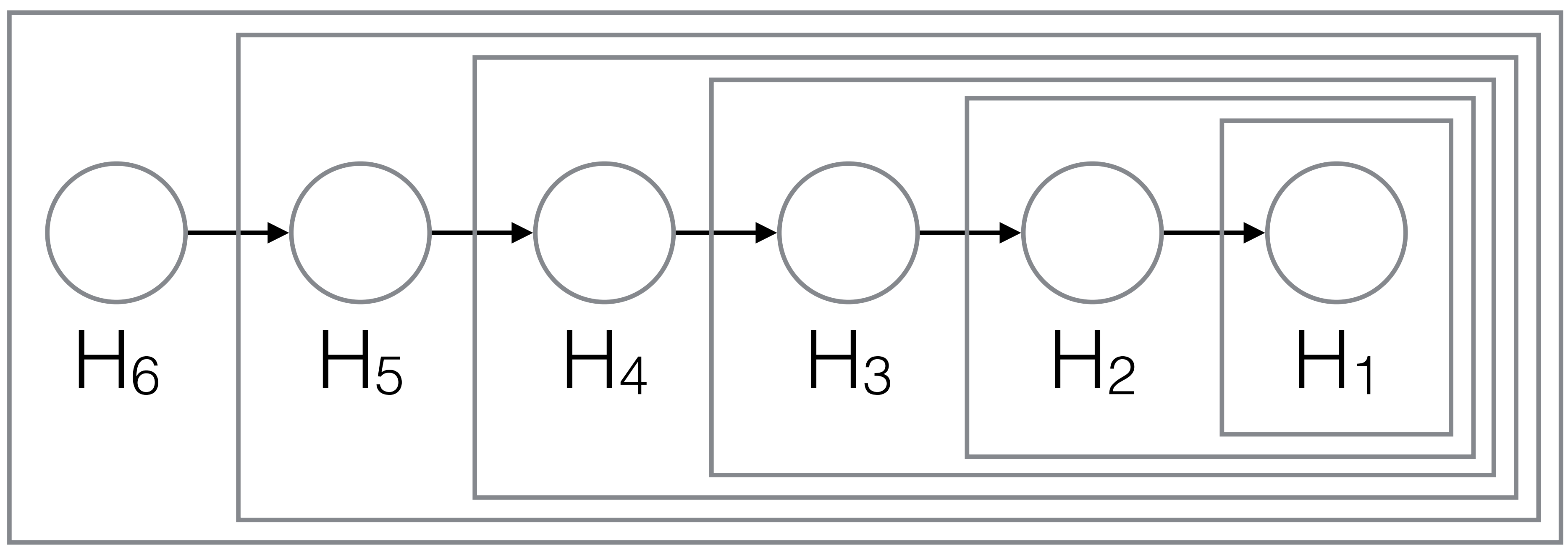


Implicit hypothesis class

for any pair of nodes u, v where u can reach v:

- if u is in S_∞, then v must be in S_∞ (e.g., u = 1, v = 2)
- equivalently, if v is not in S_∞, then u must not be in S_∞ (e.g., u = 3, v = 4)
- implicit hypothesis class associated with G = (V, E): family of all sets H of nodes consistent with the above (i.e., if u can reach v, then u in H implies v in H)
- implicit hypothesis class can be much smaller than 2^V

Implicit hypothesis class



implicit hypothesis class $\mathscr{H} = \{H_0, H_1, H_2, H_3, H_4, H_5, H_6\}$

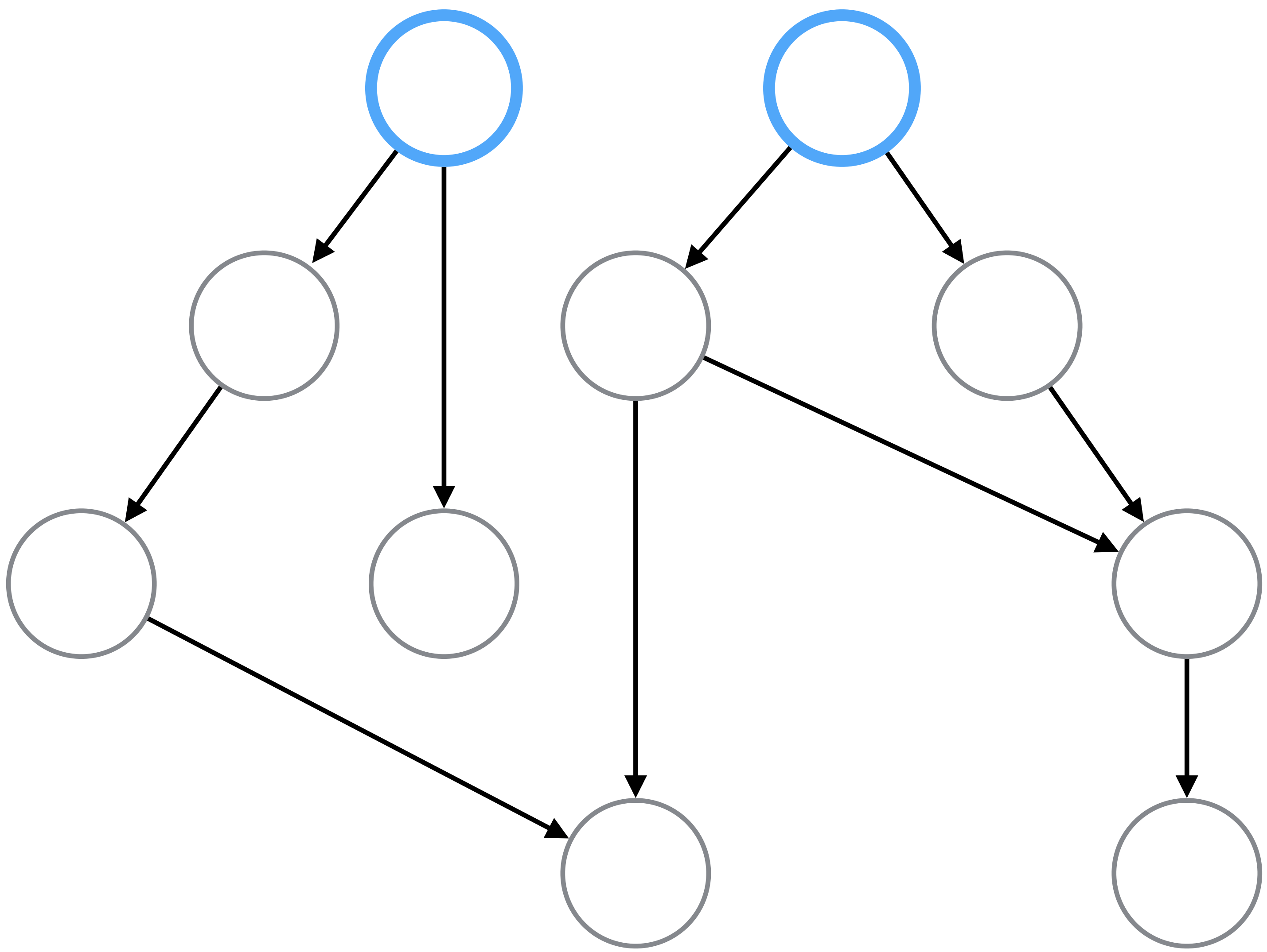
where $H_0 = \varnothing$ is the empty set

$|V| = 6, |2^V| = 64, |\mathscr{H}| = 7$

VC theory for deterministic networks

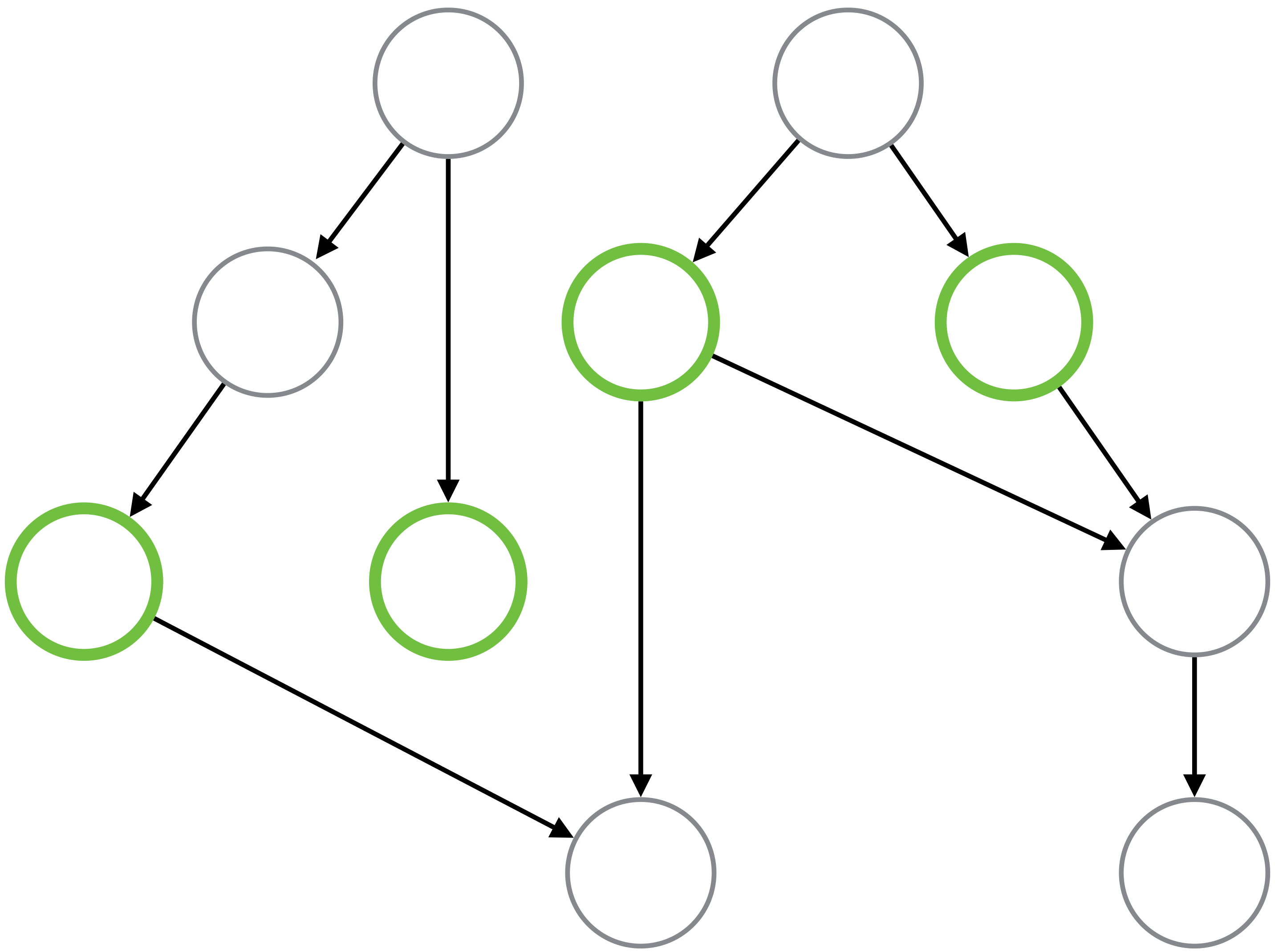
- VC(G): VC dimension of implicit hypothesis class associated with network G
- VC(G) = size of largest “independent” set (aka width), within which no node u can reach another node v

VC theory for deterministic networks



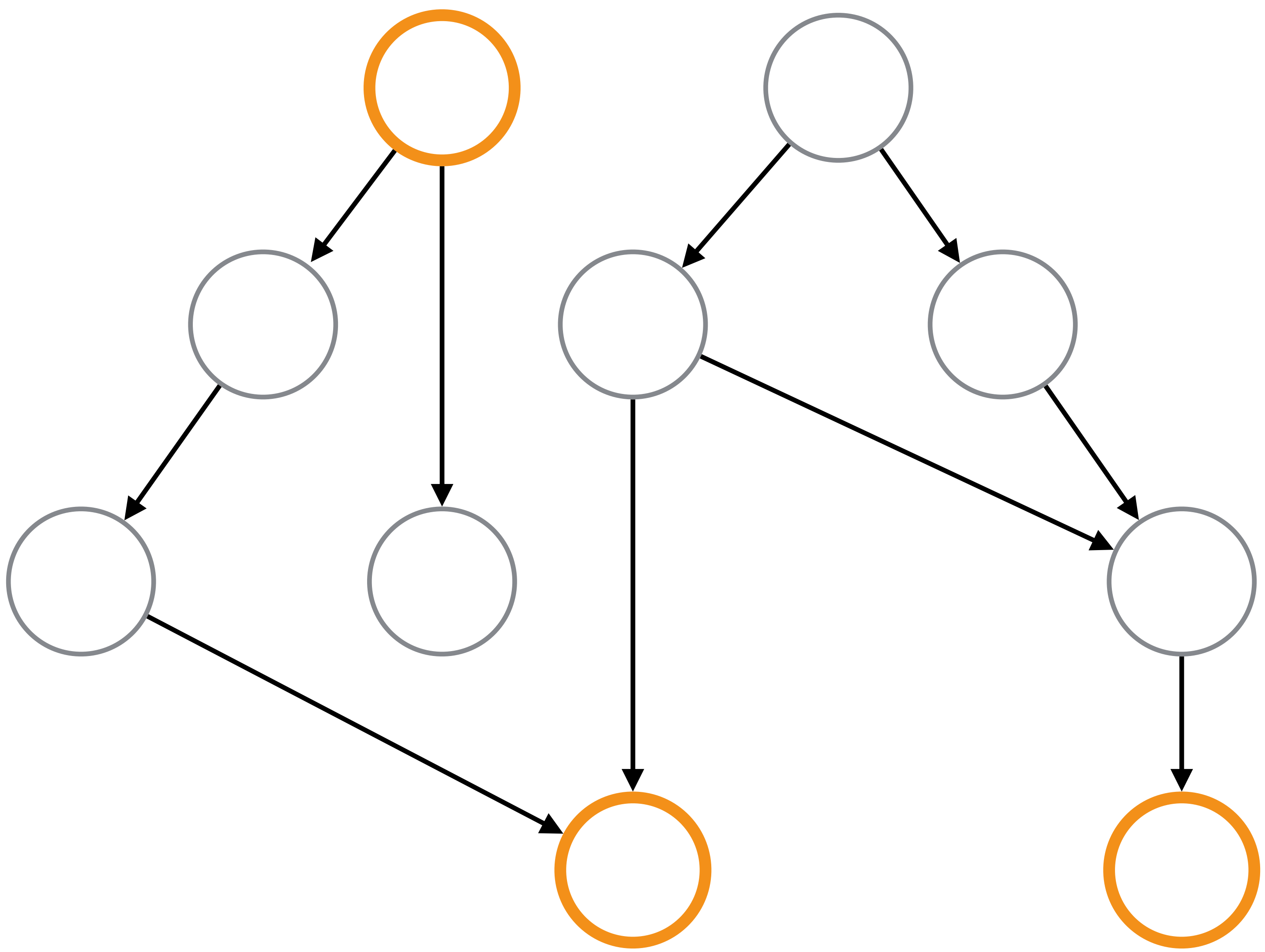
blue nodes: independent

VC theory for deterministic networks



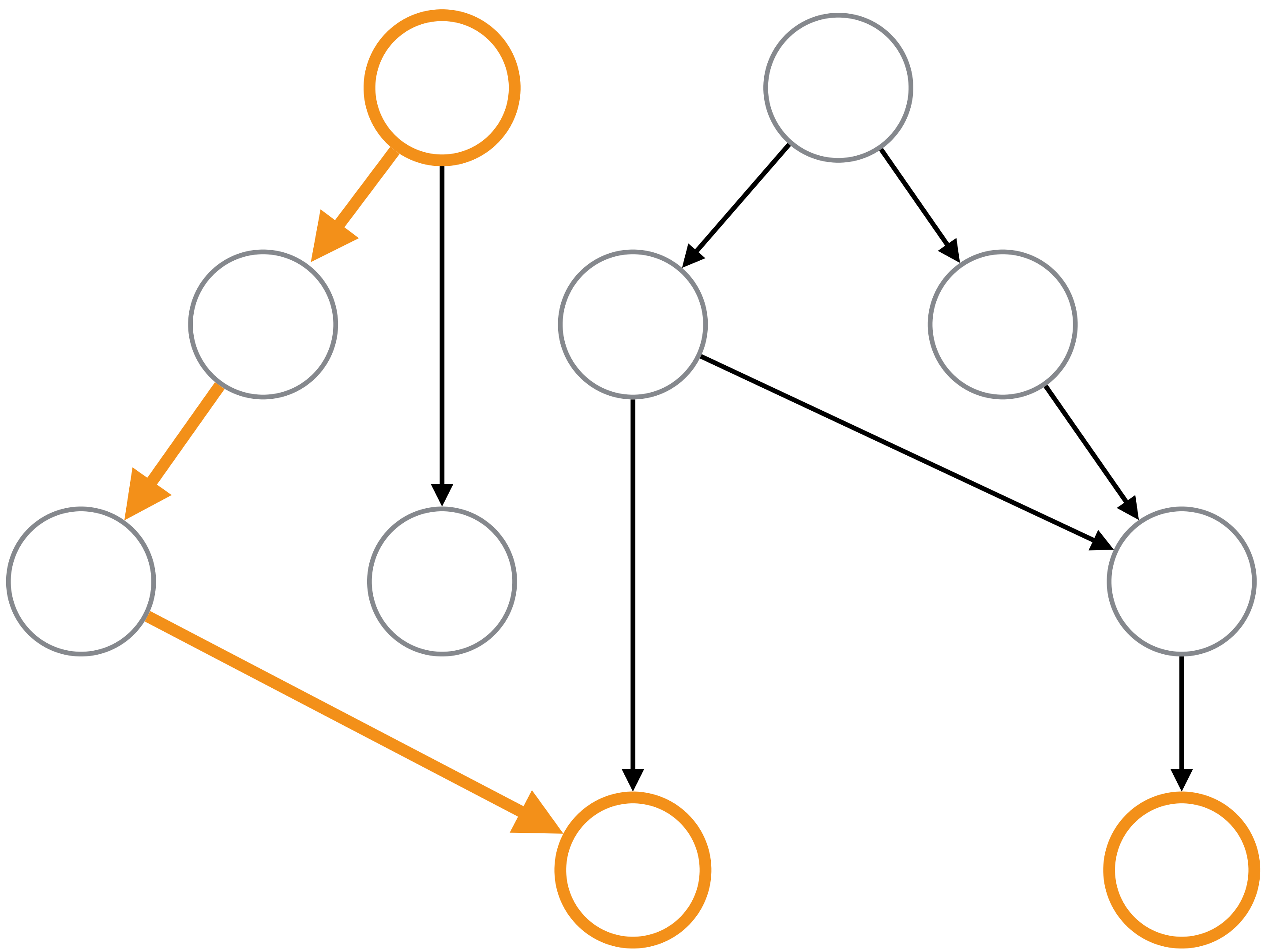
green nodes: independent

VC theory for deterministic networks



orange nodes: not independent

VC theory for deterministic networks



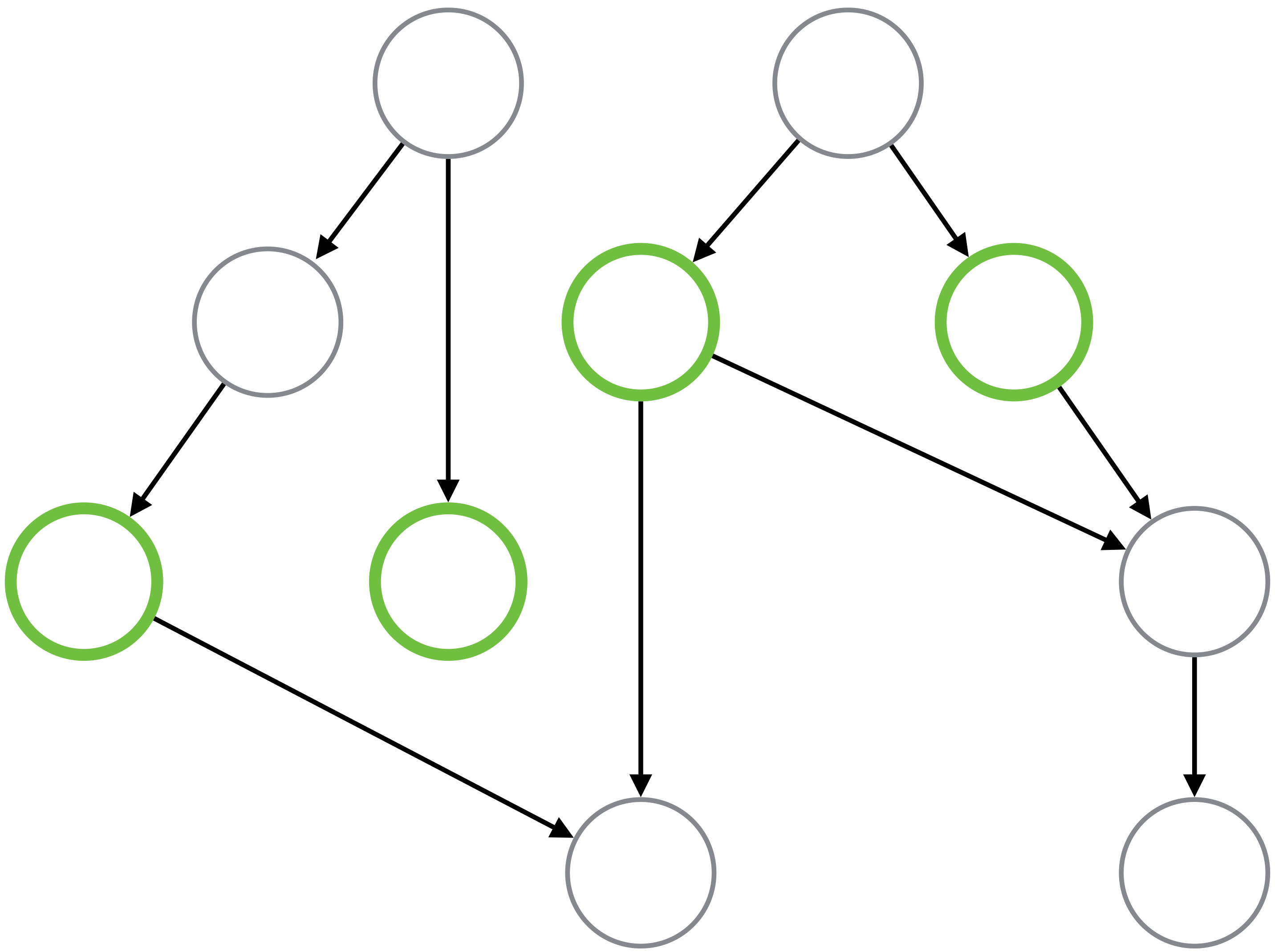
orange nodes: not independent

VC theory for deterministic networks

- VC(G): VC dimension of implicit hypothesis class associated with network G
- VC(G) = size of largest "independent" set (aka width), within which no node u can reach another node v
- VC(G) can be computed in polynomial time
- sample complexity of learning opinions:

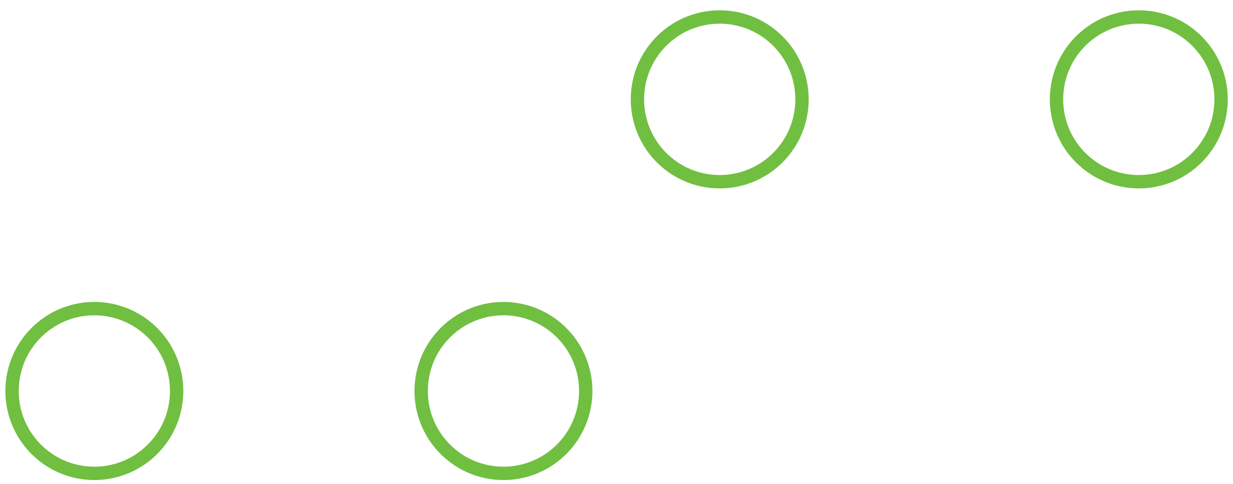
$$\tilde{O}(VC(G) / \varepsilon)$$

Why width?



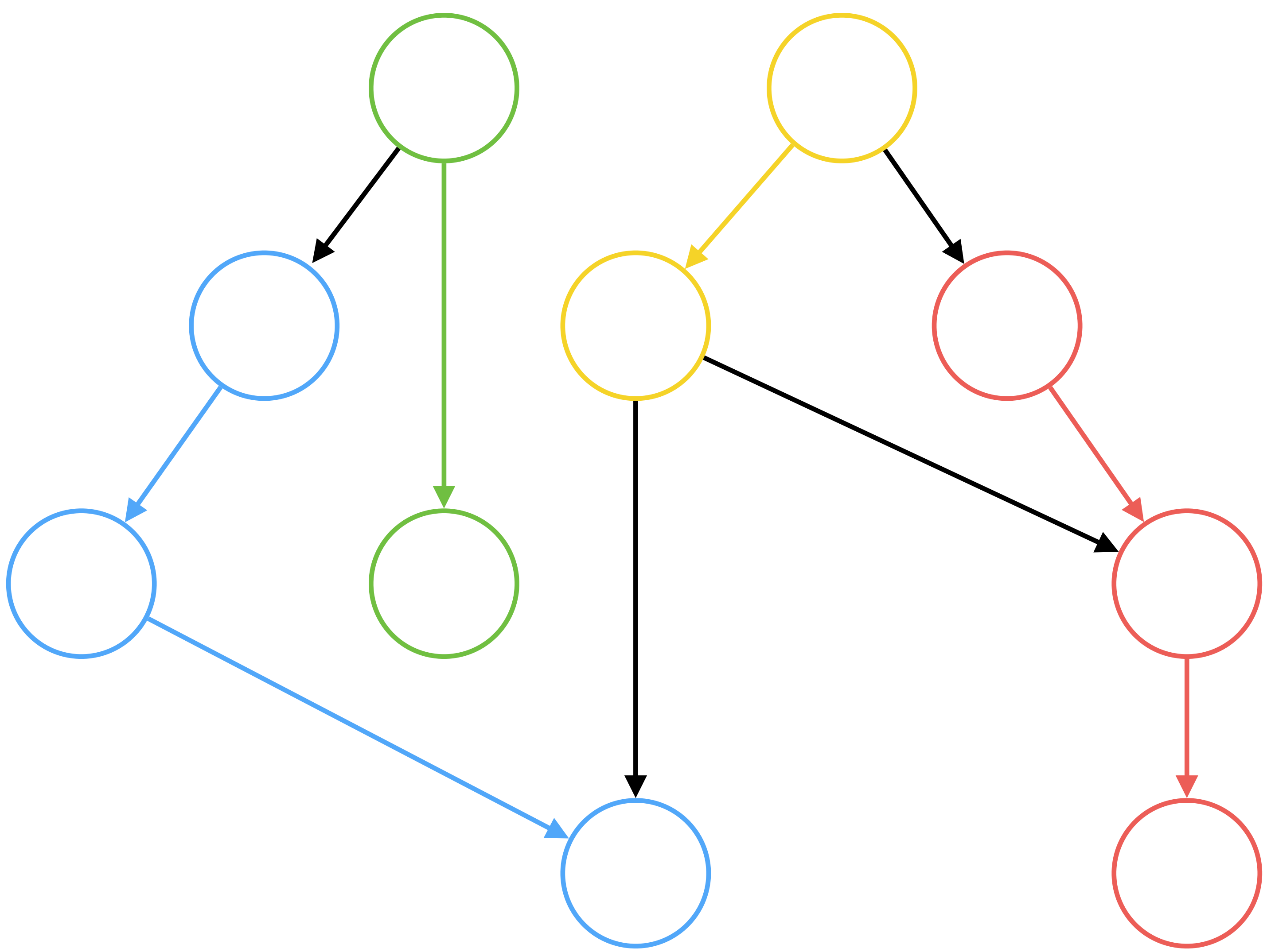
LB: $\mathscr{D}$ is uniform over a maximum independent set

Why width?



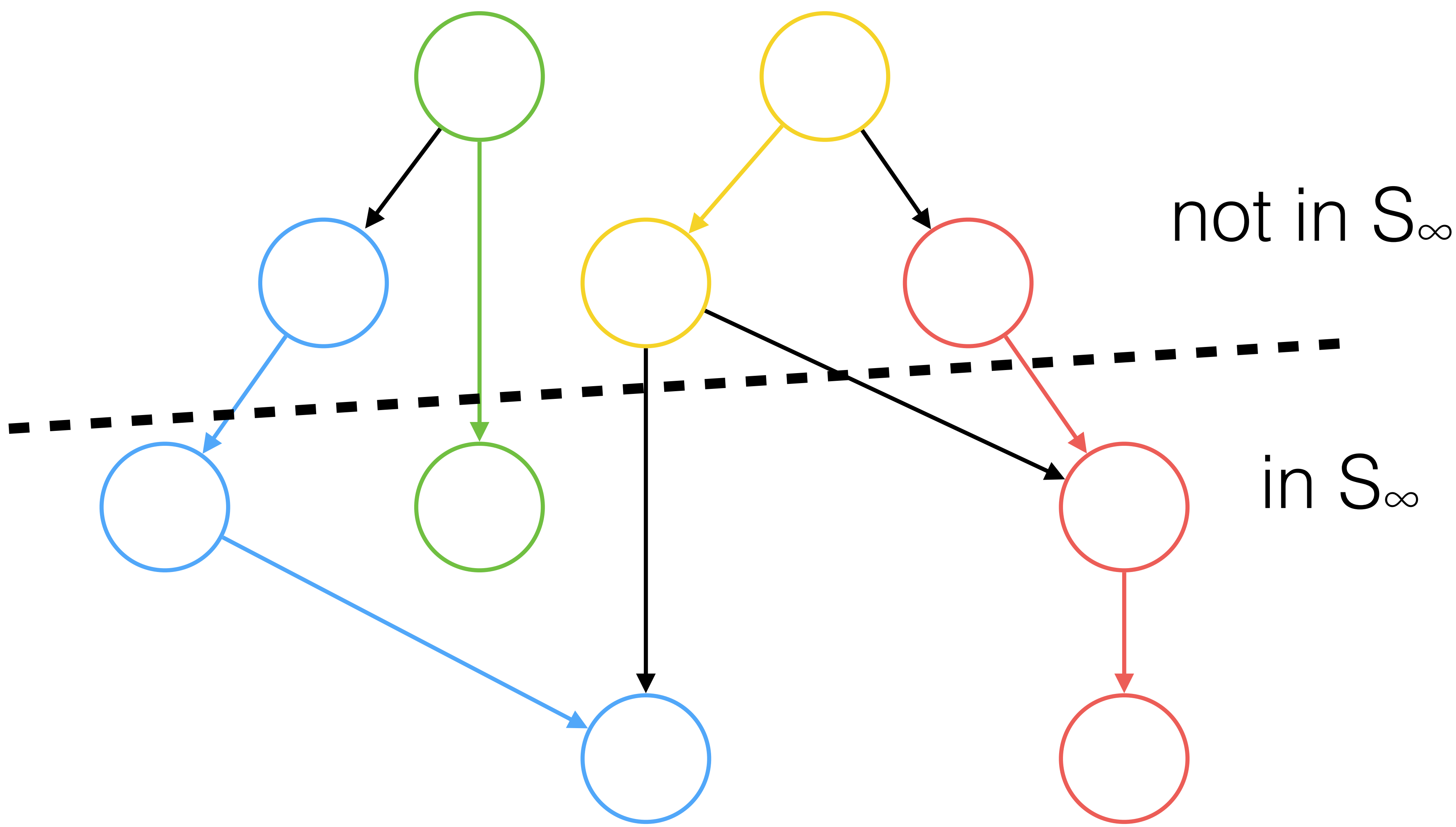
LB: $\mathscr{D}$ is uniform over a maximum independent set

Why width?



UB: number of chains to cover G = VC(G)
need to learn one threshold for each chain

Why width?



UB: number of chains to cover G = VC(G)
need to learn one threshold for each chain

- so far: VC theory for deterministic networks
- next: the case of random networks

Random social networks

- propagation of opinions is inherently random
- randomness in propagation = randomness in network
- random network $\mathscr{G}$: distribution over deterministic graphs
- propagation: draw G ~ $\mathscr{G}$, propagate from seed set S_0 in G

Random social networks

- random network $\mathscr{G}$: distribution over deterministic networks
- propagation: draw G ~ $\mathscr{G}$, propagate from seed set S_0 in G
- PAC learning opinions: fix $\mathscr{G}$, unknown S_0 and $\mathscr{D}$
- graph G ~ $\mathscr{G}$ realizes (unknown to algorithm), propagation happens from S_0 in G and results in S_∞
- algorithm observes m labeled samples, tries to predict S_∞
- “random” hypothesis class — VC theory no longer applies

Random social networks

- S_0: information to recover, G: noise
- learning is impossible when noise overwhelms information
- hard instance: nodes form a chain in a uniformly random order, S_0 = {node 1}
- learning the label of any other node requires $\Omega(n)$ samples

Random social networks

- S_0: information to recover, G: noise
- learning is impossible when noise overwhelms information
- when noise is reasonably small:

$\tilde{O}(\mathbb{E}[VC(G)] / \varepsilon)$ samples are enough to learn opinions

up to the intrinsic resolution of the network

Random social networks

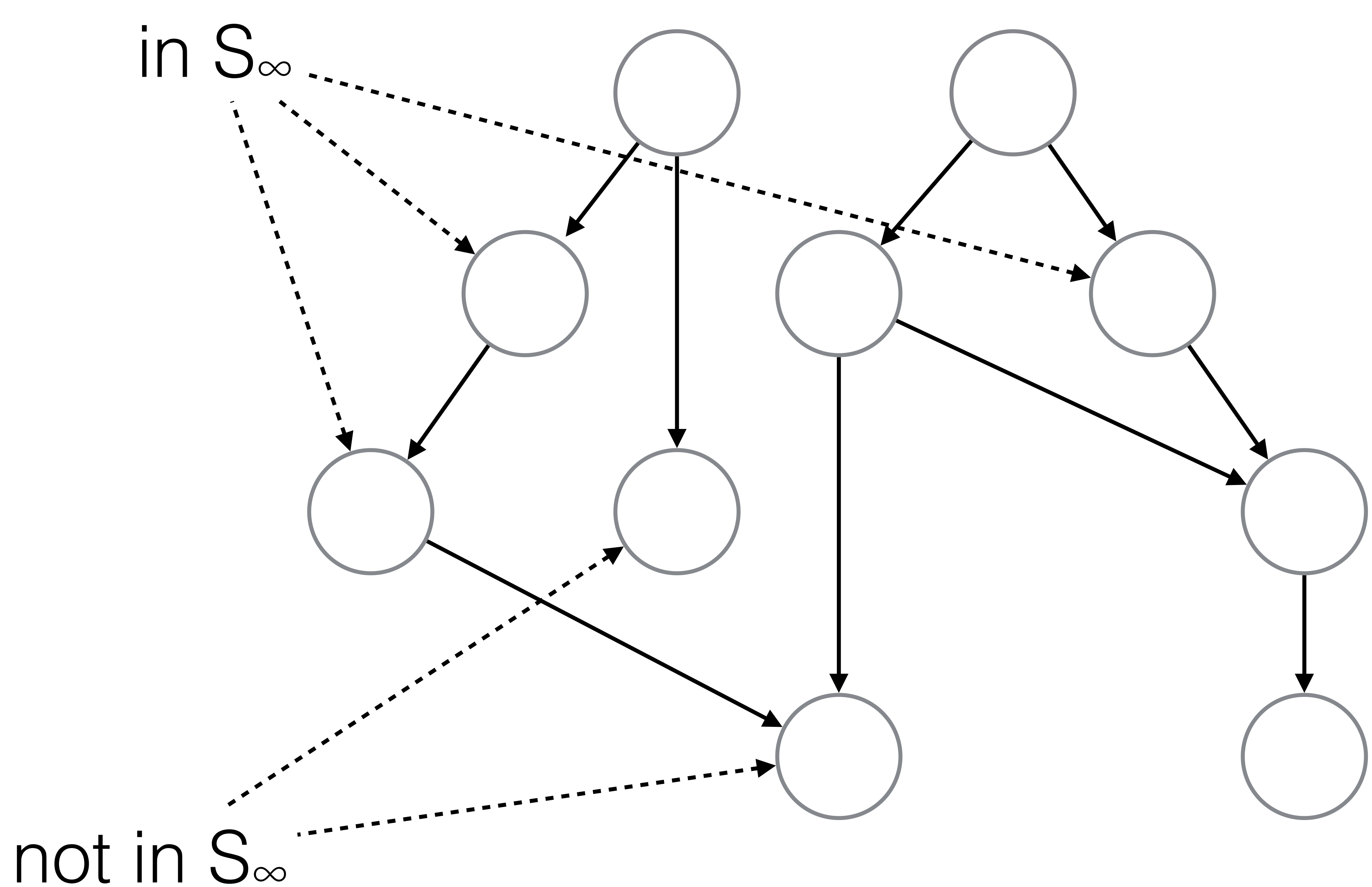
when noise is reasonably small:

$\tilde{O}(\mathbb{E}[VC(G)] / \varepsilon)$ samples are enough to learn opinions

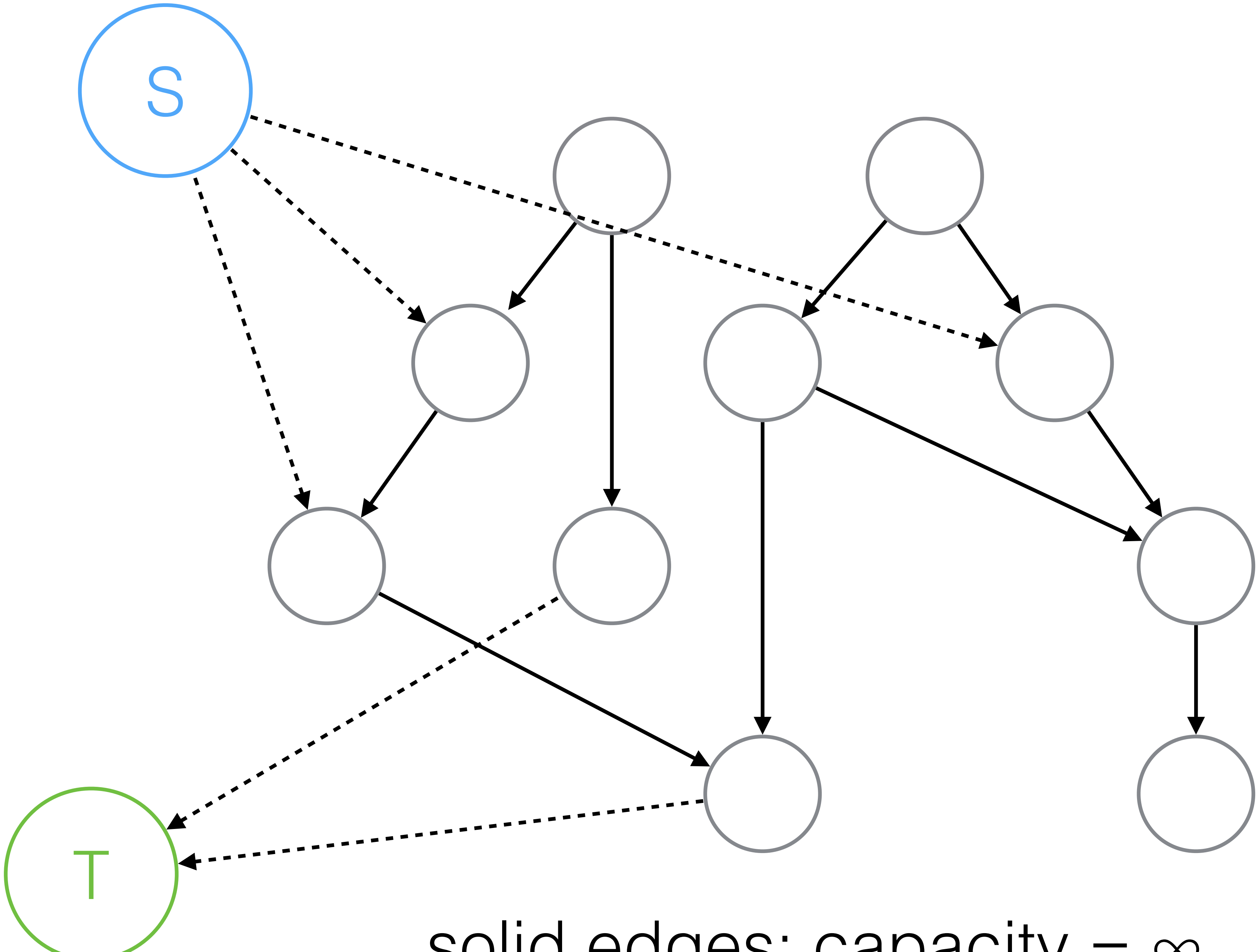
sketch of algorithm:

- draw iid sample realizations $G^j \sim \mathscr{G}$ of the network
- for each G^j, find the ERM H^j on G^j with the observed sample set $\{(u_i, o_i)\}$, by computing an s-t min-cut
- output H = node-wise majority vote by $\{H^j\}$, i.e., each node u is in H iff u is in at least half of $\{H^j\}$

Algorithm for ERM



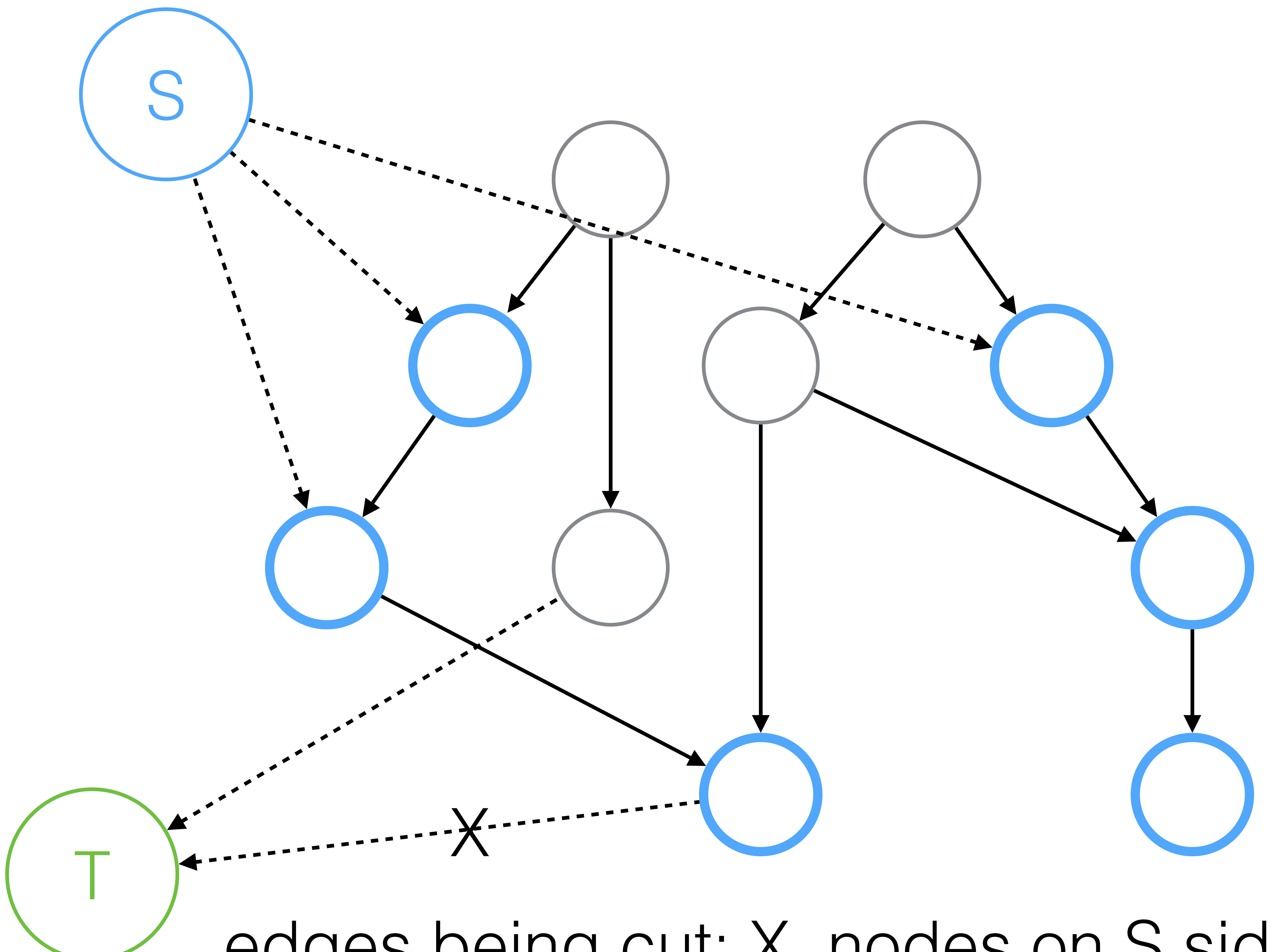
Algorithm for ERM



solid edges: capacity = ∞

dashed edges: capacity = 1

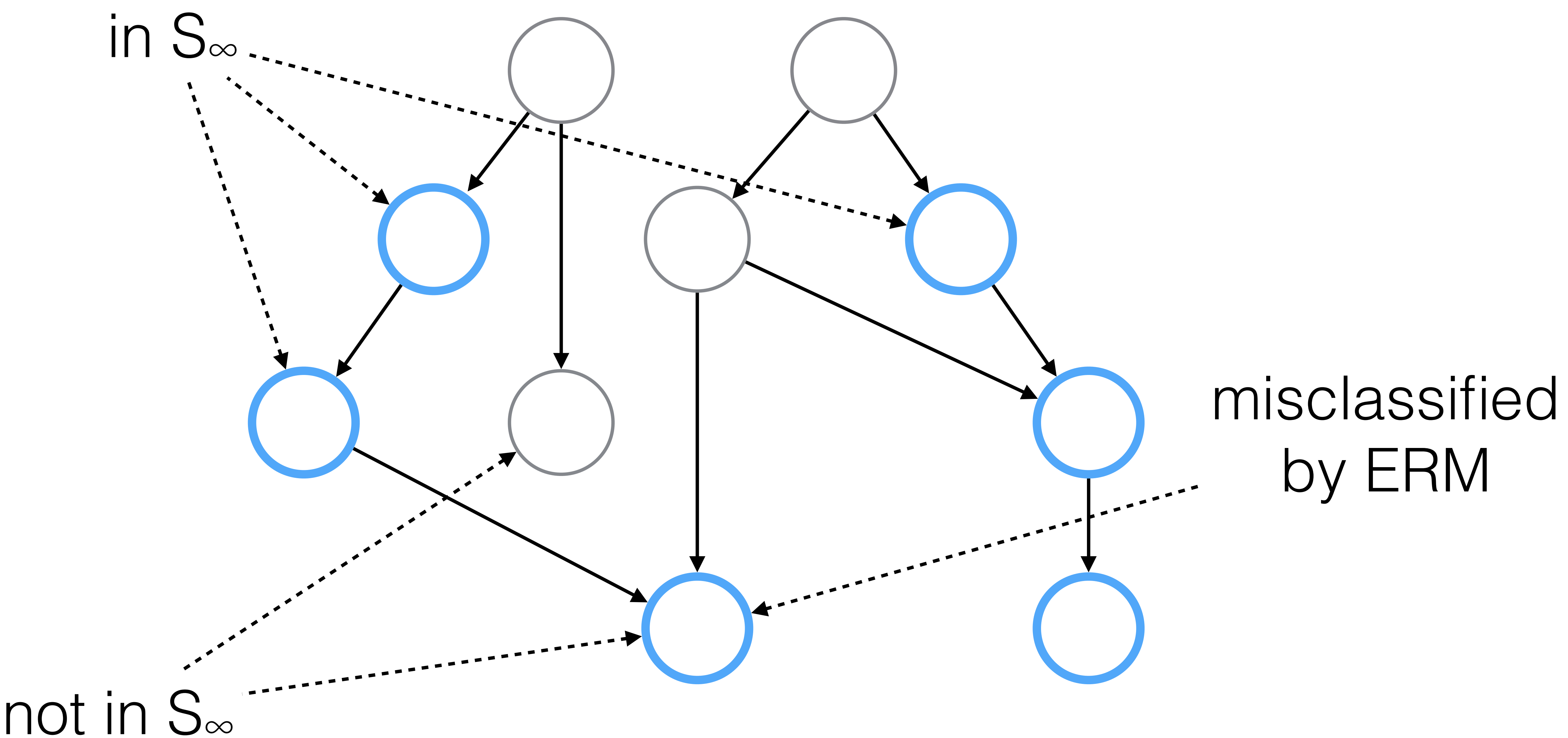
Algorithm for ERM



edges being cut: X, nodes on S side:

total capacity of S-T mincut = 1

Algorithm for ERM



Random social networks

- each ERM H^j has <u>expected</u> error $\boldsymbol{\varepsilon}$
- ... but probability of high error is still large
- use majority voting to boost probability of success

Future directions

- other propagation models
- non-binary / multiple opinions
- …

Thanks for your attention!

Questions?