Week 9

10/19/2020

- Next HW due Monday 10/26, (extended from this Fri 10/23)
- Feedback from proposals → Inbox.
- This week: sampled networks.

Why sample?

① Access (e.g. private Twitter accts)

② Computation Time (algos are slow or memory runs out)

WWW ~ 10^{10} vertices

If we have $G = (V, E)$

a subsample or subgraph

$G' = (V', E')$

where $V' \subset V$ and

$E' \subseteq V' \times V' \subset E$

$V' \subset V$: is a subset of

$E' \subseteq V' \times V'$: subset with possible equality

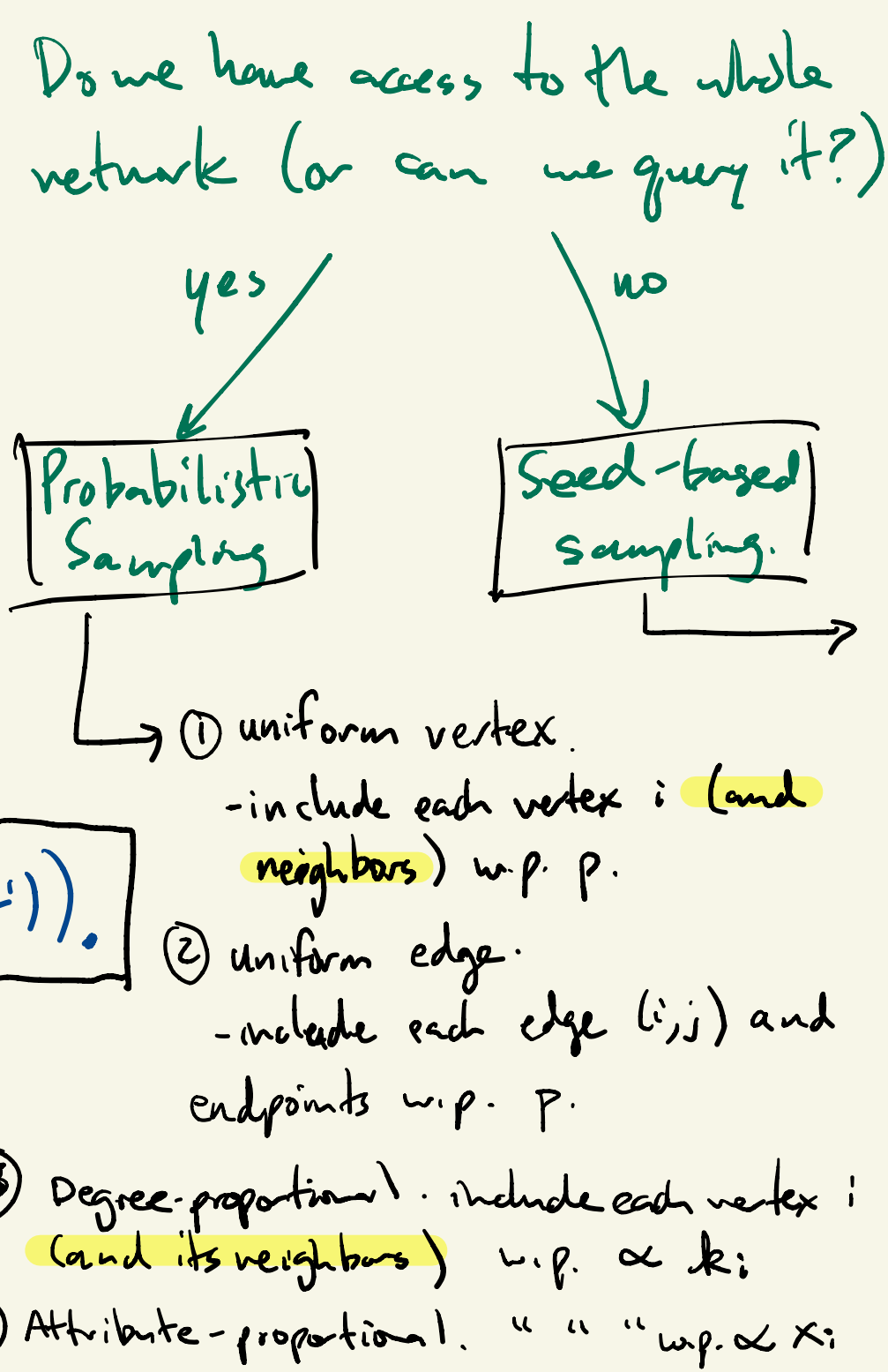
$E' \subset E$: If we don't sample the nodes, we don't get any of its edges.

Key:

Working with G', rather than G can be fine if the function f you are computing would yield equivalent results on G and G', i.e.

$f(G) = f(G')$, arrrr... more generally: if the distribution of outputs is the same $Pr(f(G)) = Pr(f(G'))$.

- Depends on f.
- Depends on G'.



Seed-Based Sampling.

$\ell \leftarrow \text{Nell}$

(Assumes access to one or more seed vertices)

① Snowball sampling.

For each seed vertex i, and distance ℓ, include all vertices and their neighbors, in an ℓ-step B.F.S. tree, rooted at i.

② B.F.S. edge sampling. For each seed vertex i, and distance ℓ, include all edges in an ℓ-step BFS, rooted at i.

③ Adaptive sampling: For each seed i, and int s, include all vertices (and nbrs) or include all edges, in an adaptively grow tree containing s vertices, rooted at i.

Analysis

Sampling induces Patterns.

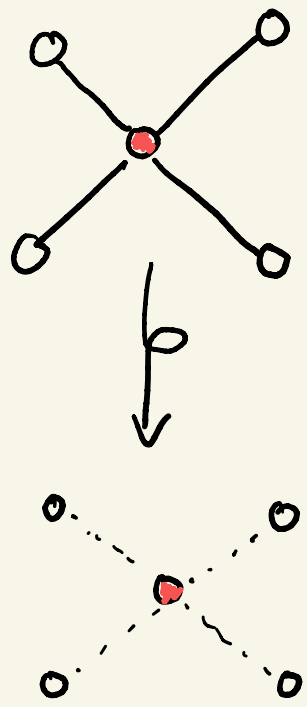
① Extreme sparsity (prob. sampling).

② Compact, but biased subgraph. (seed-based)

③ An overabundance of low-degree vertices in G'. (often $k=1$)

Why?

Network Immunization!



$n = 5$

$v = 1$

n nodes, m edges → adj A

v vaccines. 100% effective.

Whom to vaccinate?

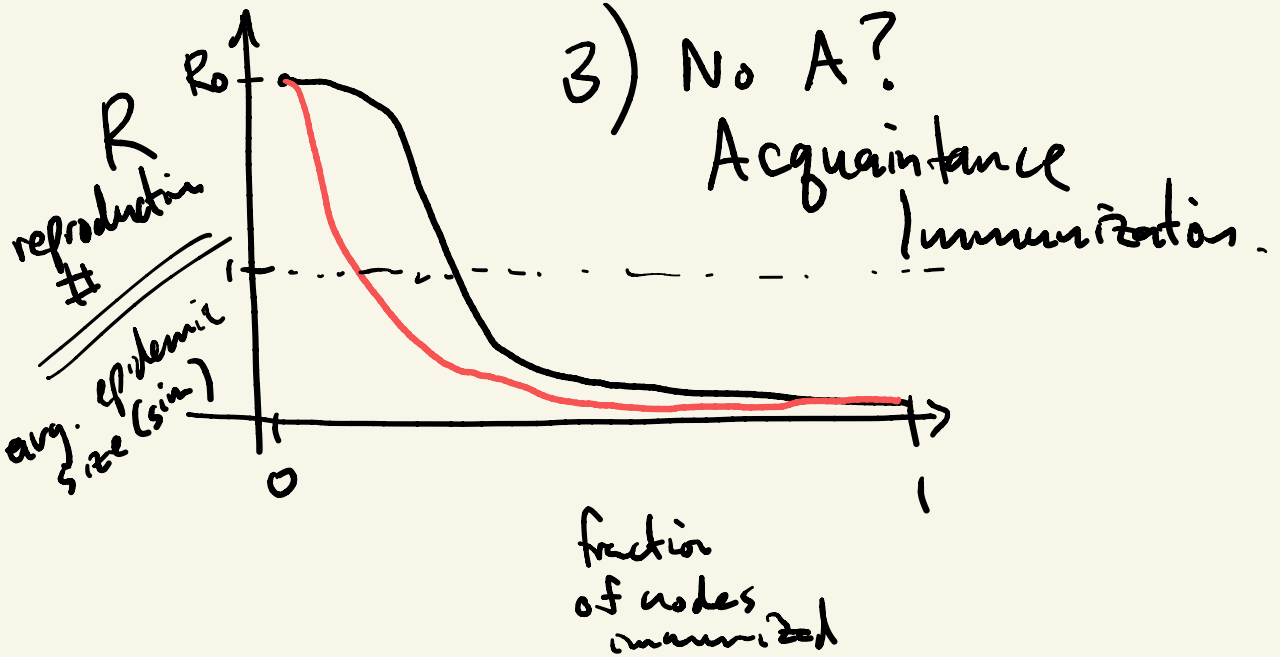
WHO: between countries: allocation.
within country: prioritization

Strategies?

1) High centrality
   - eigenvector
   - degree

2) Maximize # of disconnected components.

3) No A?
   Acquaintance Immunization.



10/21/20

Sampling distorts ...

- degrees – too many $k=1$ and low-degree nodes
- components. May have more components than in the original network!

Uniform Vertex Sampling

- target sample s vertices
- choose each vertex w.p. p. and, in so doing, also get its nbrs.

How should we choose p?

Each time we sample i, in expectation, we'll add $1+\langle k\rangle$ vertices to sample.

$\Rightarrow$ Choose each vertex w.p. $p = \frac{s}{(1+\langle k\rangle)n}$

we should get approx s vertices in the sample.

sub $\langle k\rangle = \frac{2m}{n}$

$$\Rightarrow p = \frac{s}{\left(1+\frac{2m}{n}\right)n} = \boxed{\frac{s}{n+2m} = p}$$

If network is sparse, $p \propto \frac{s}{n}$.

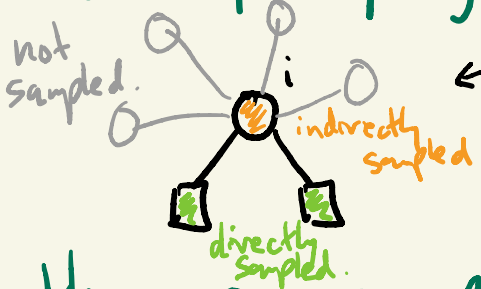
How will sampling change degree dist?

prime means sampled degree
no prime = original degree.

Note: • directly sampling a vertex $\Rightarrow$ we get k_i right, $k_i' = k_i$

• directly sampling 1 neighbor $\rightarrow k_i' = 1$

• $\leftarrow$ 2 neighbors $\rightarrow k_i' = 2$



How many sampled vertices have $k' = 1$?

Two ways to have $k' = 1$

① we sampled i, and it has $k = 1$

pn_1

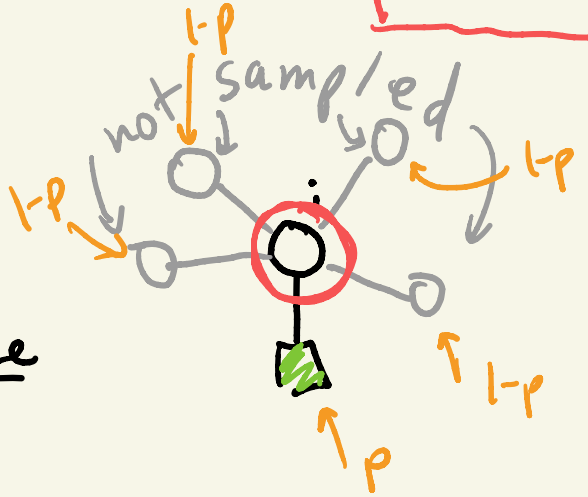
② we sampled another node, whose neighbor is i, but only sample one neighbor of i.

Let n_k be # of nodes in original graph w/ degree k

$$E[n_1'] = pn_1 + \sum_{k=2}^{n} p(1-p)^{k-1} n_k$$

$$E[n_{k'}'] = pn_k + p^{k'} \sum_{K=k'}^{n} (1-p)^{k-k'} n_{k'}$$

3 pts E.C.
email me, and tell me w/ short expl. which is correct?
where is P(did not sample i)?



$p(1-p)^{k-1}$

Uniform Edge

- choose to include each edge w.p. p.
- include neighbors

↓

sampling nodes $\propto k_i$

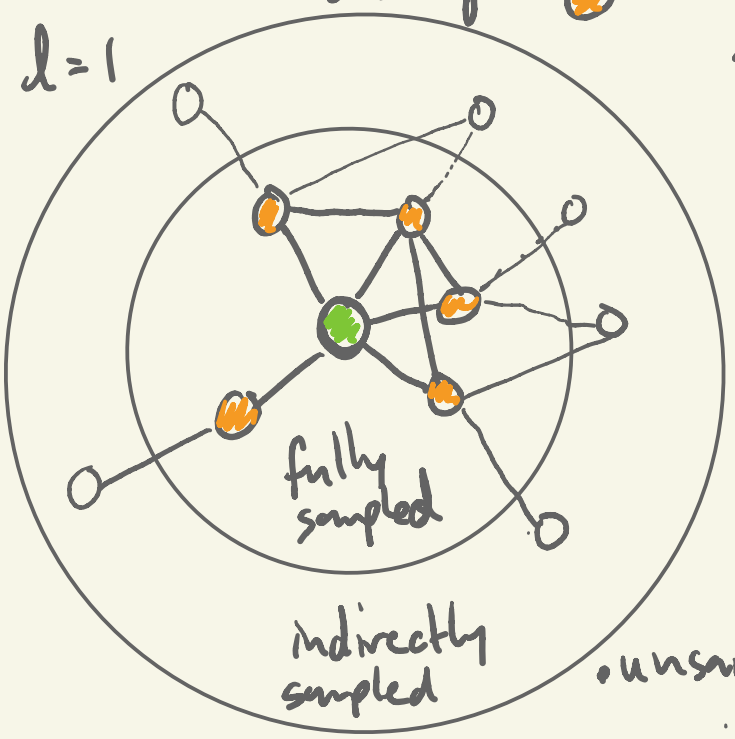
A node w/ degree k_i will have, in exp. degree $k_i' = p k_i$ in sampled graph.

Parallel: percolation.

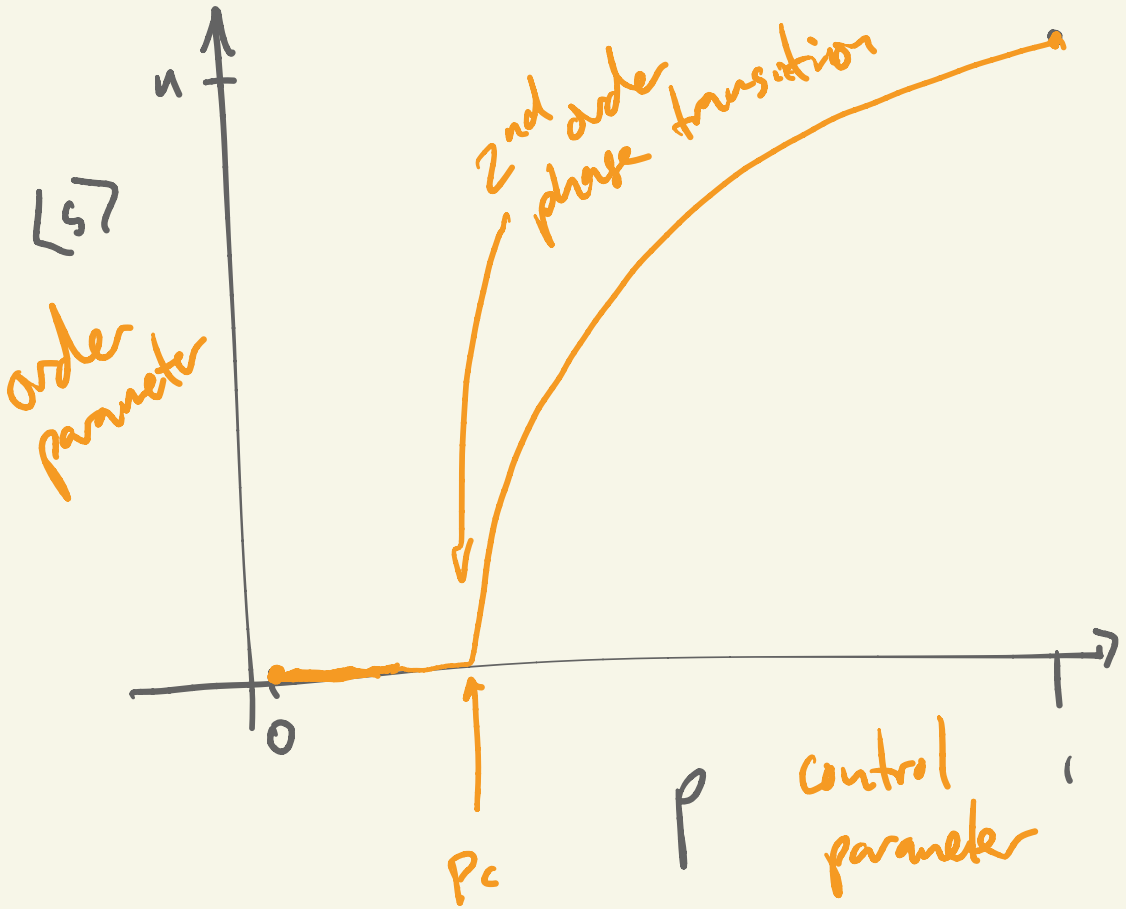
↳ if p too small, network will be disconnected, sparse. (if $\langle k \rangle$ falls below E.R. critical value $\langle k \rangle = 1$)

Snowball Sampling!

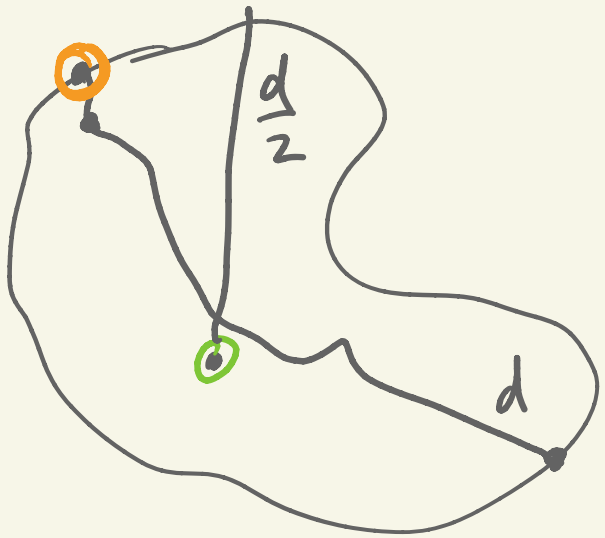
- start from a seed i
- include all vertices and nbrs up to a distance l. and edges



- fully sample
  ⋮
  and
  any j $d_{ij} \leq l$
- indirectly sampled
  j s.t. $d_{ij} = l+1$
- unsampled
  j s.t. $d_{ij} > l+1$



what would $\langle l \rangle$ be for $p = 1$?



How many steps to reach all nodes in network?

$$\frac{diam}{2} \leq \langle l \rangle \leq diam$$

$$\langle l \rangle = \frac{1}{n} \sum_i \text{longest shortest path from } i \text{ to any } j$$