On the Performance of Percolation Graph Matching

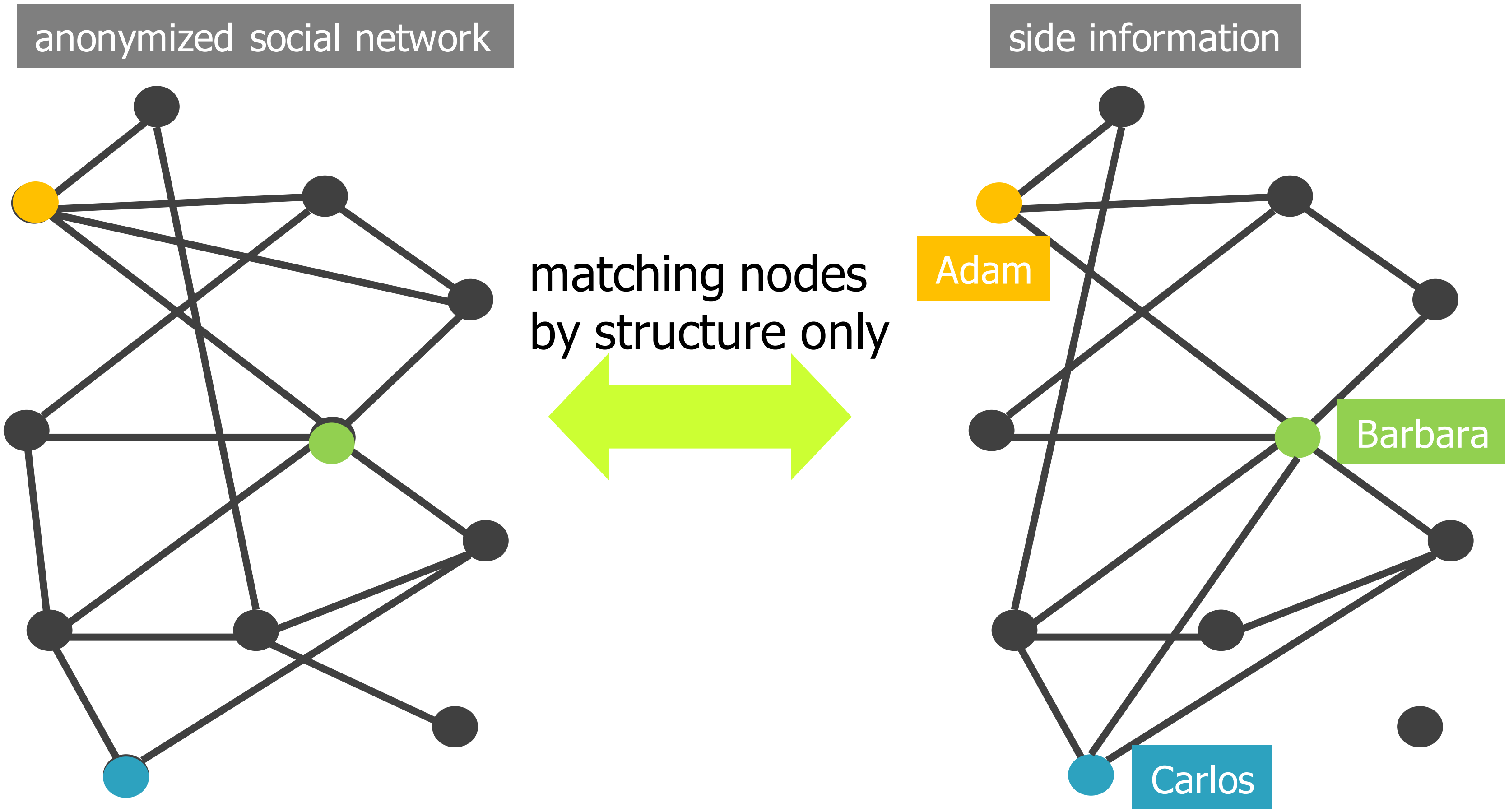
Lyudmila Yartseva
Matthias Grossglauser
EPFL

COSN
Oct 2013

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

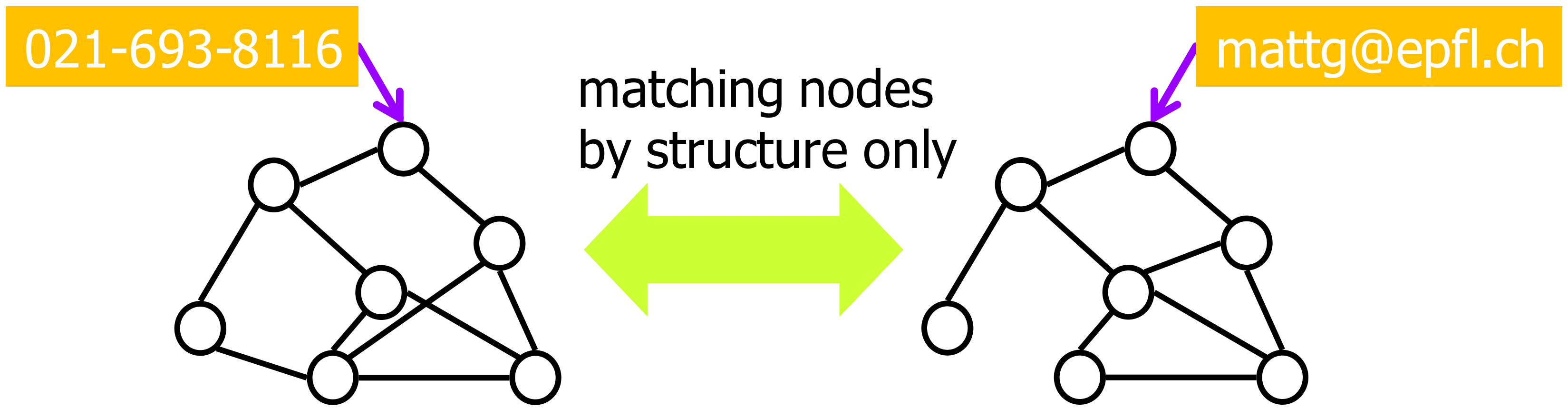
Privacy of Networks

- **Adversary has:**
  - Anonymized network: unlabeled graph
  - Side information: labeled graph – similar but not identical



Graph Matching Applications

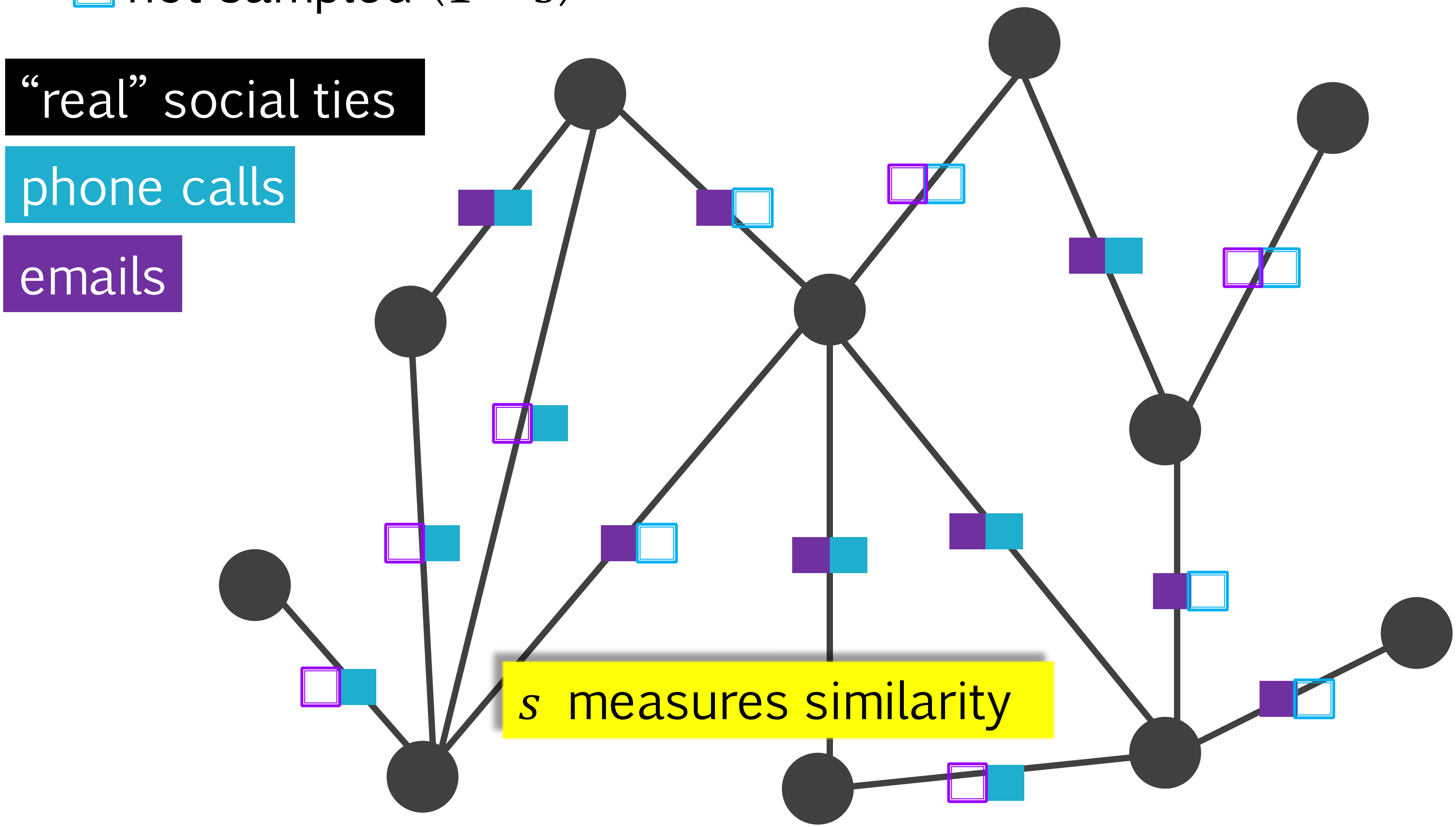
- Social networks:
  - Correlating different domains



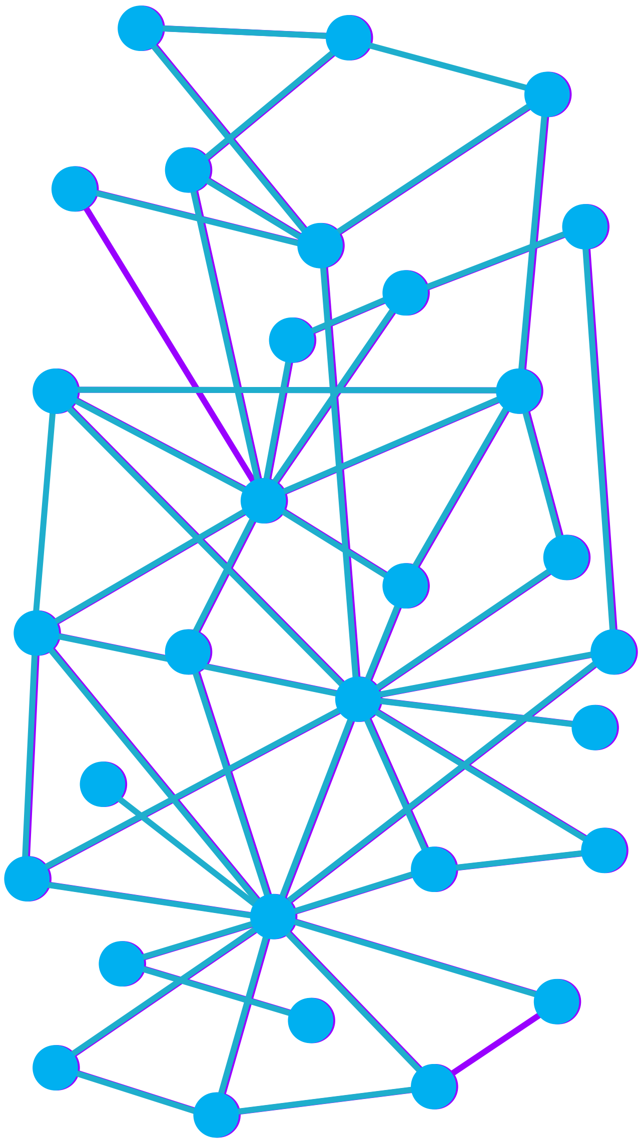
- Security:
  - Identifying computer viruses by function-call patterns
- Computer vision:
  - Segment adjacency graph to find similar images
- …

$G(n, p; s)$ Sampling Model

Generator $G = G(n, p)$
- ■ sampled (s)
- □ not sampled $(1 - s)$



$G(n, p; s)$: Two Correlated $G(n, ps)$'s



Result 1: GM is Easy with ∞ Resources

- **Theorem [PG11]:**
  - For the $G(n,p;s)$ matching problem, if

$$nps \frac{s^2}{2-s} = 8\log n + \omega(1)$$

threshold for $aug(G) = 1$

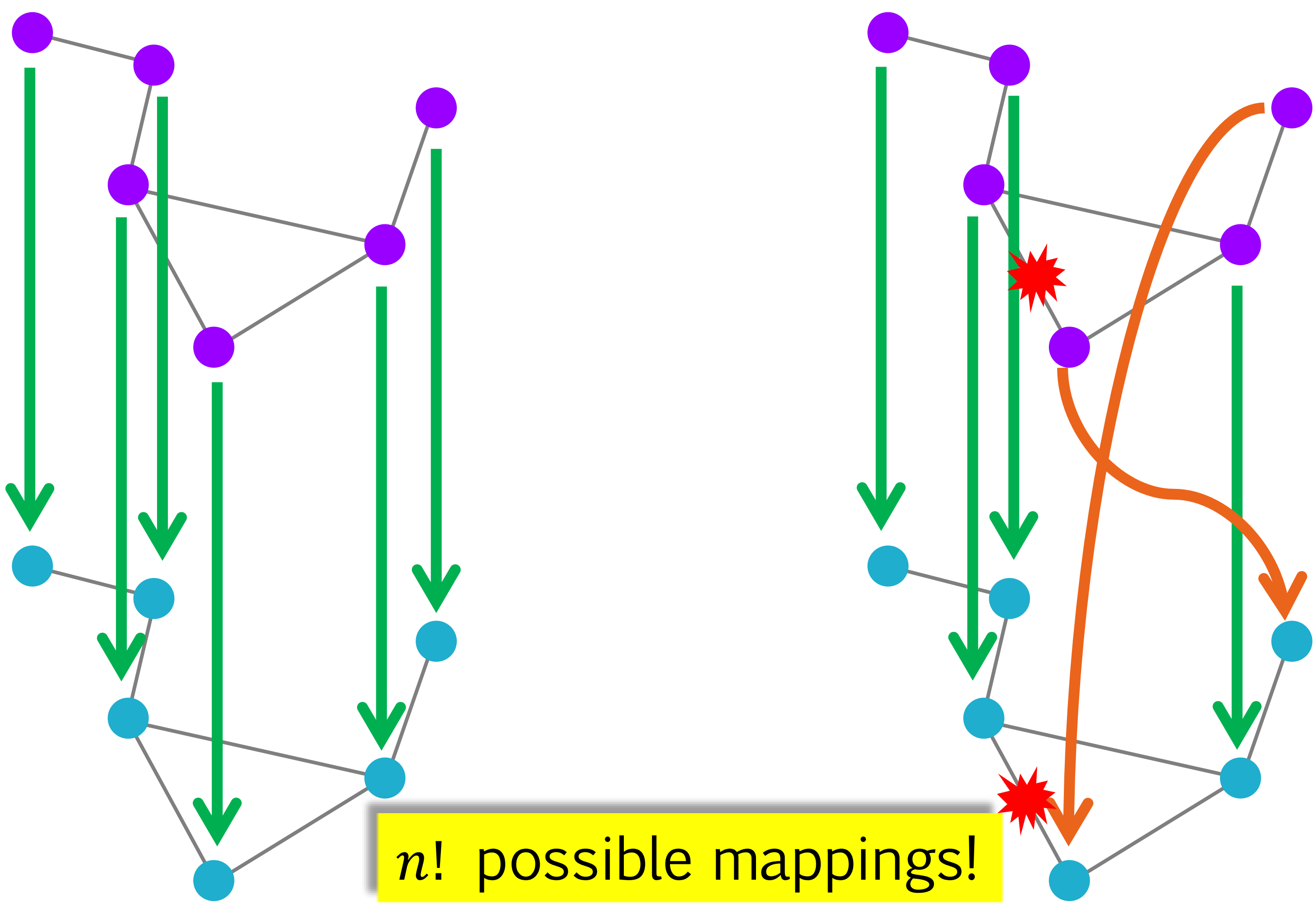
nps: $E[\text{degree}]$ of $G_{1,2}$

penalty for dissim of $G_{1,2}$

  then $G_{1,2}$ can be perfectly matched a.a.s.

- **Interpretation:**
  - Surprisingly weak condition: degree growing faster than $\sim\log n$ enough to break anonymity
  - Decrease with s only quadratic

Mappings and Edge Mismatch



$n!$ possible mappings!

$$\Delta(\pi_0) = 0$$

$$\Delta(\pi) = 2$$

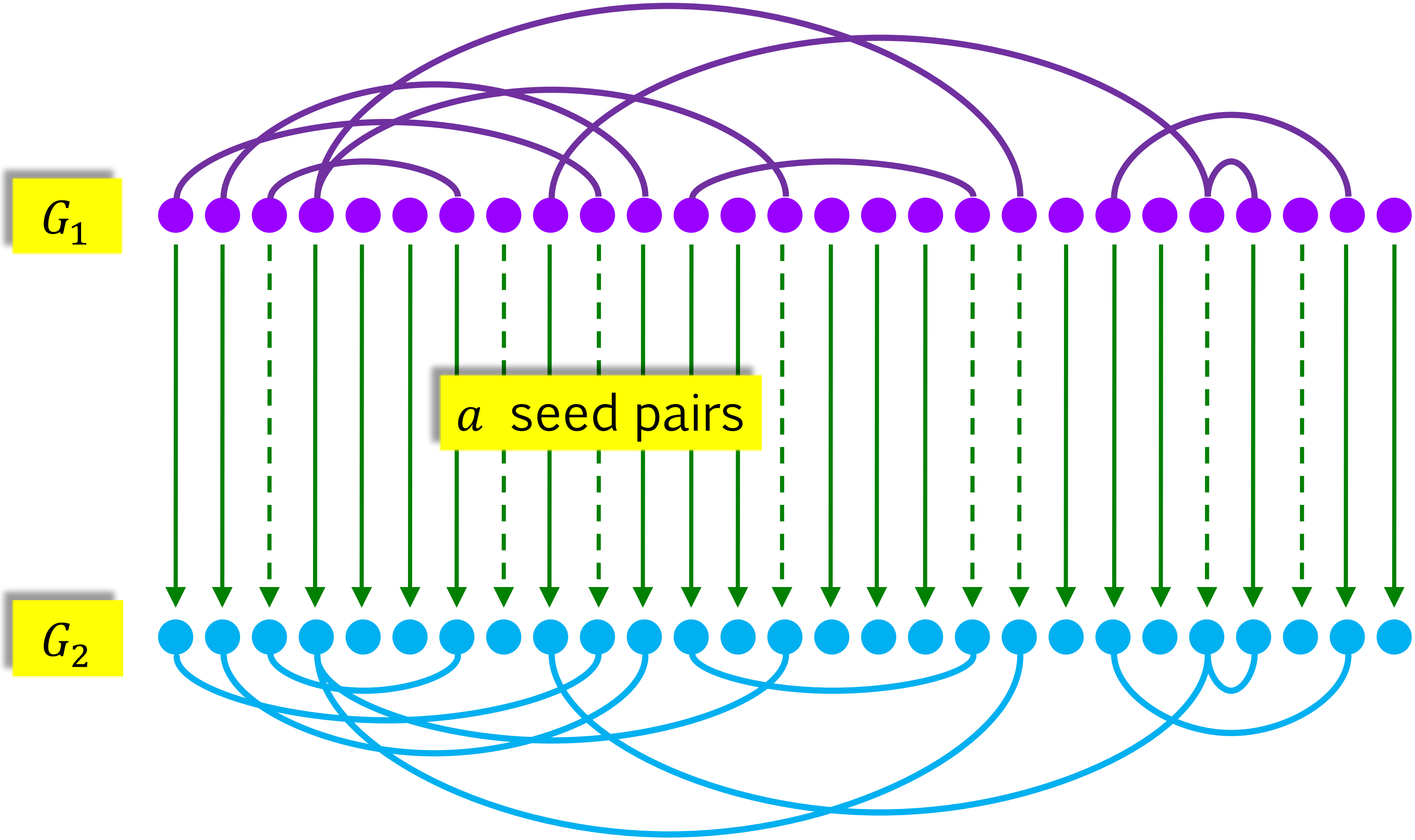
Approach

- **Assumption:**
  - Attacker has infinite computational power
  - Can try all possible mappings π and compute edge mismatch function $\Delta(\pi)$
- **Question:**
  - Are there conditions on p, s such that

$$P\{\pi_0 \text{ unique min of } \Delta(\pi)\} \to 1$$

  - If yes: adversary would be able to match vertex sets only through the structure of the two networks!
- **Note:**
  - $G(n, p; s)$ model: statistically uniform, low clustering, degree distribution not skewed -> conjecture: harder than real networks

Result 2: Graph Matching with Seeds



Questions

- How many seeds are needed?
- Is there a phase transition?
- How efficiently can we match?
- Tuning parameters?

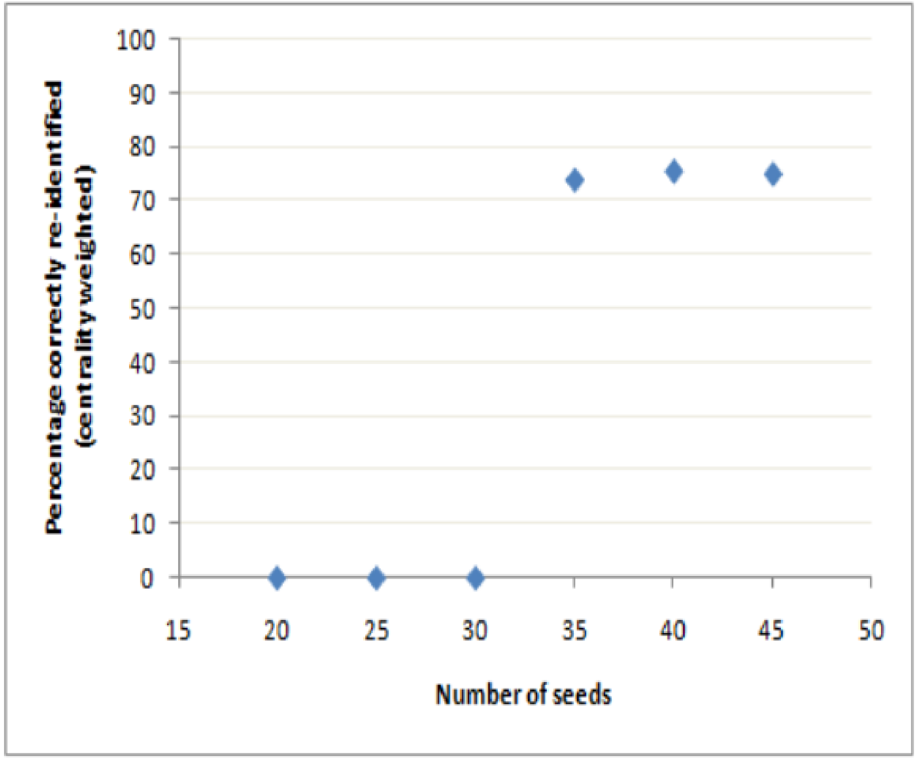
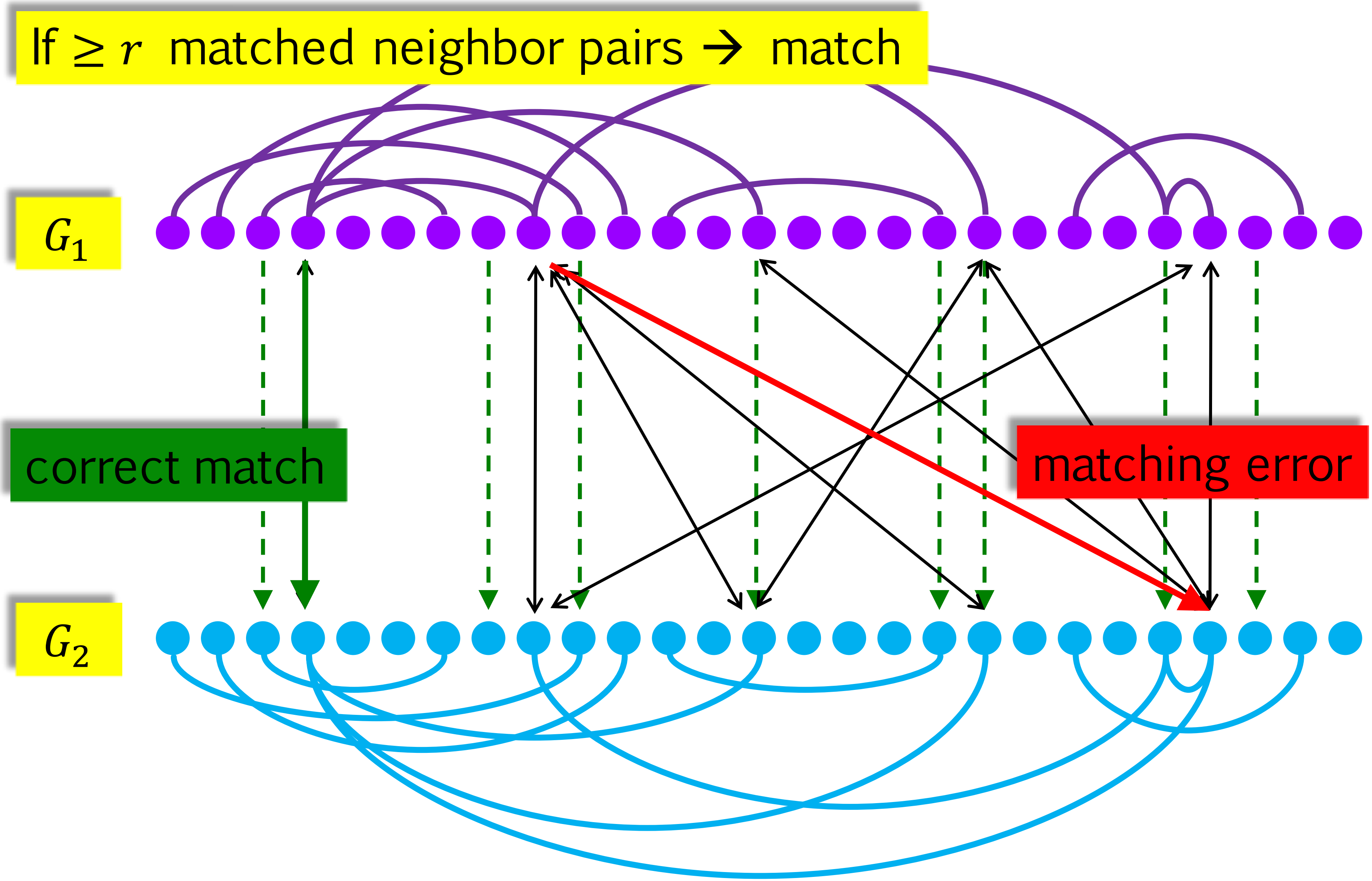


Figure 2. The fraction of nodes re-identified depen sharply on the number of seeds. Node overlap: 25 Edge overlap: 50%

From [A. Narayanan, V. Shmatikov, "De-anonymizing social networks", IEEE Symp. on Security and Privacy, 2009]

Percolation Graph Matching Algorithm

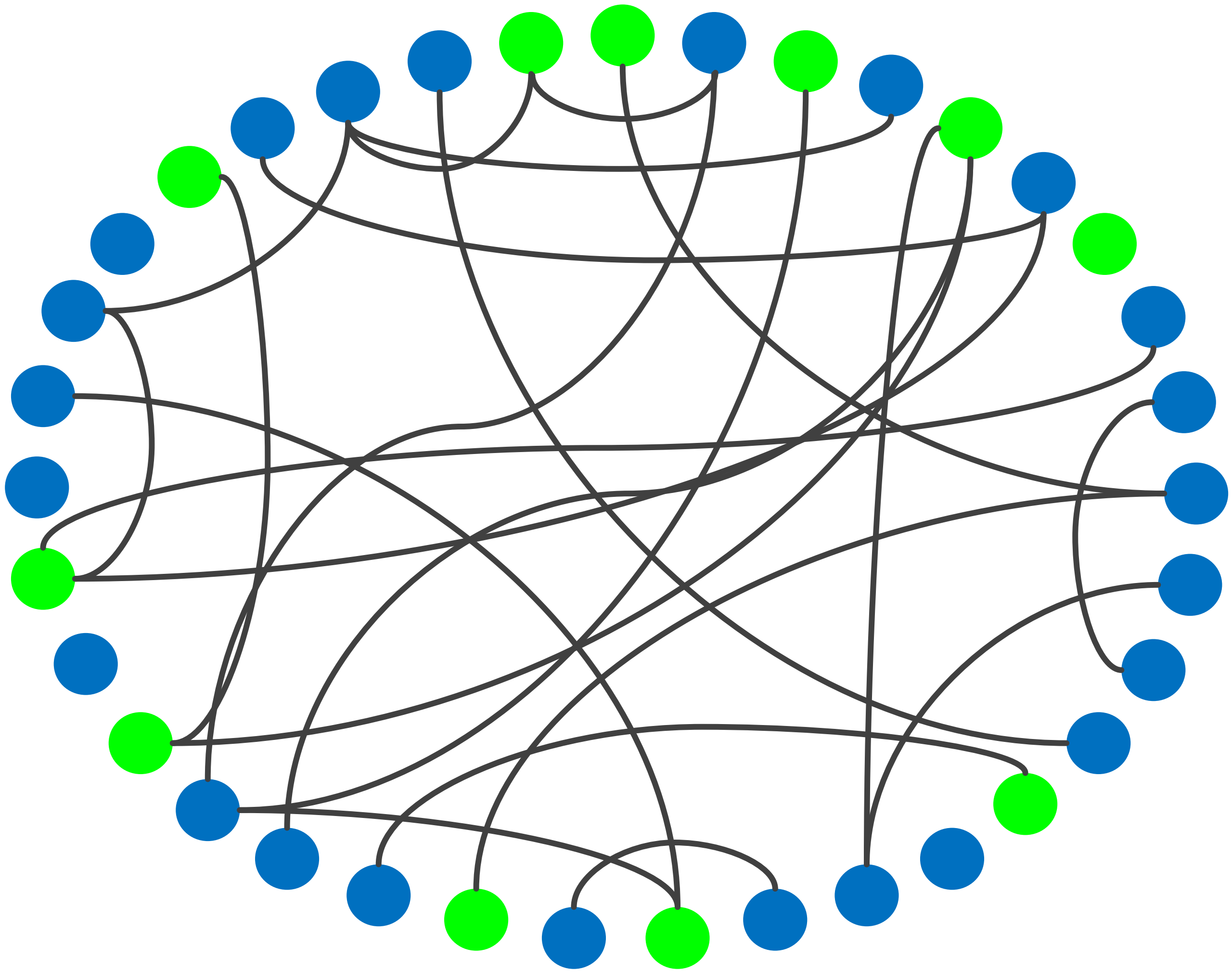


Result: Seed Set Size Threshold for $G(n, p; s)$

- Theorem 2: phase transition in # seeds a
  - For $n^{-1} \ll ps^2 \ll s^2 n^{-\frac{4}{r}}$:
    - If $\frac{a}{a_c} \to \alpha < 1$, final map is $o(n)$ w.h.p.
    - If $\frac{a}{a_c} > \alpha > 1$, final map is $n - o(n)$ w.h.p.
- Seed set size threshold:
  - $a_c = (1 - r^{-1}) \left(\frac{(r-1)!}{n(ps^2)^r} \right)^{1/(r-1)}$
  - Slowly densifying network: constant r
  - Growth of a_c: a bit less than linear
    - $p = \log n / n$, s fixed → $a_c \propto n (\log n)^{-r/(r-1)}$

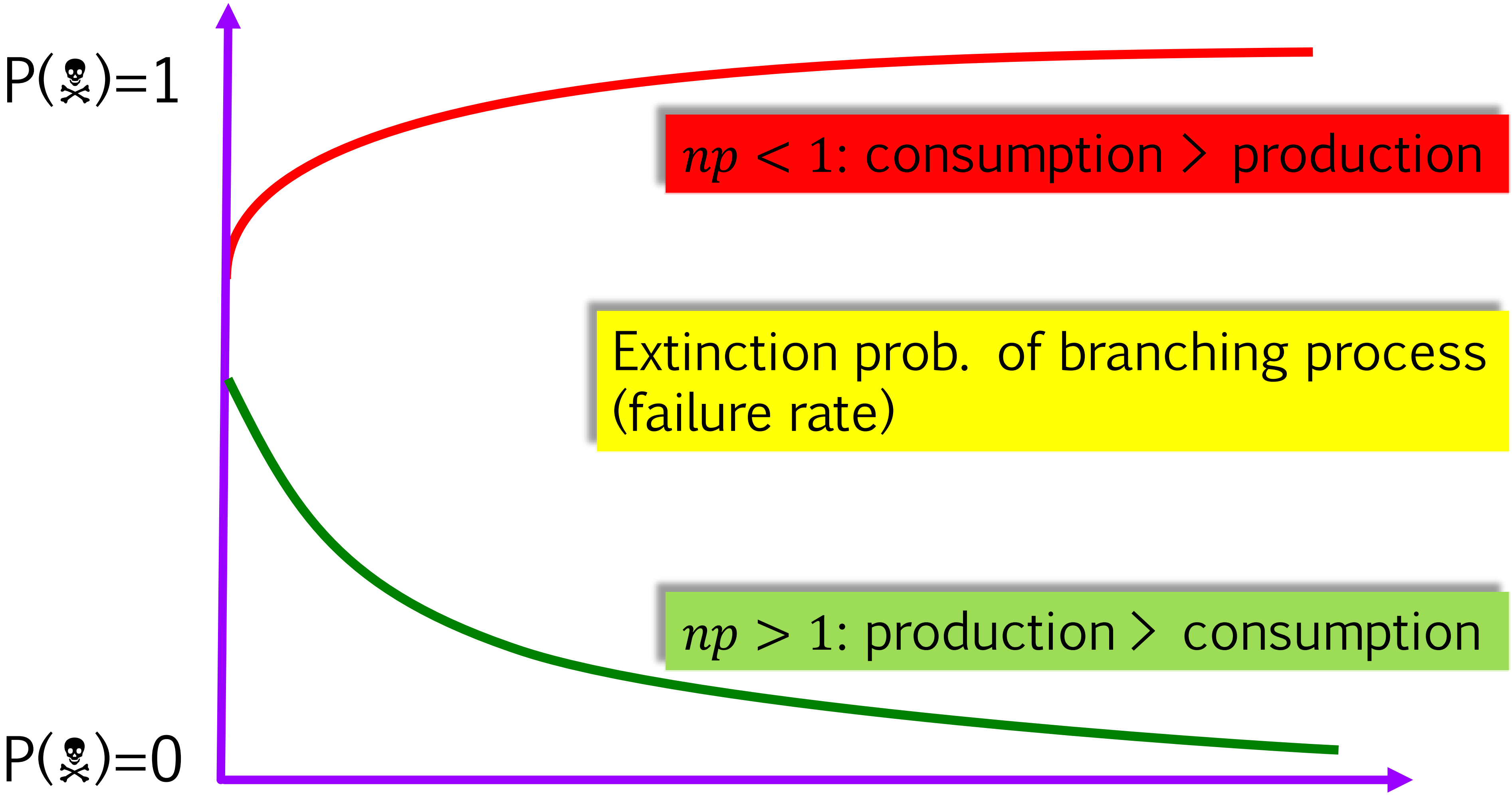
Bootstrap Percolation for $G(n,p)$

Activation from r neighbors

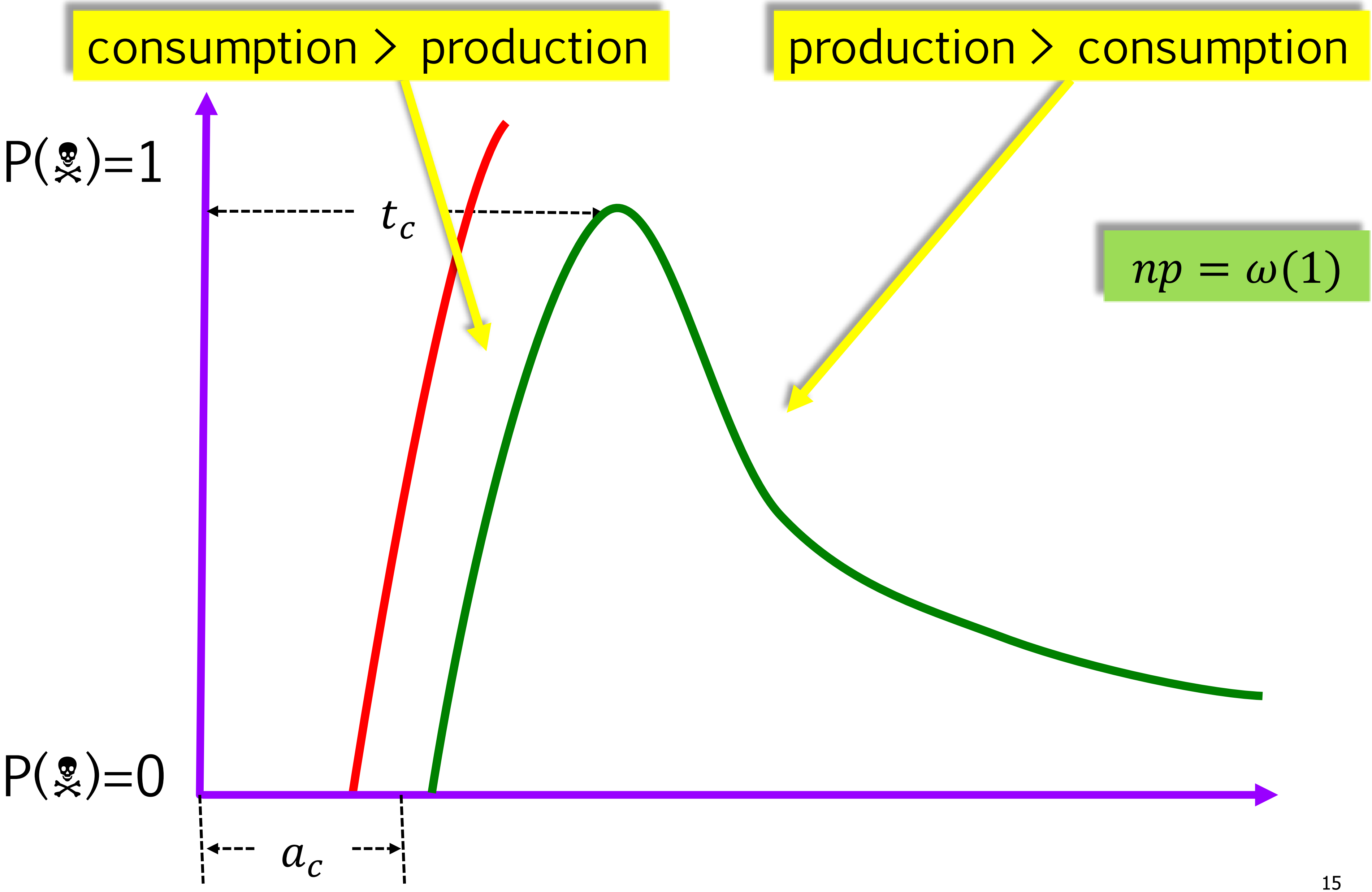


[S. Janson, T. Luczak, T. Turova, T. Vallier, Bootstrap Percolation on the Random Graph $G(n,p)$, Annals Applied Prob., 22(5), 2012]

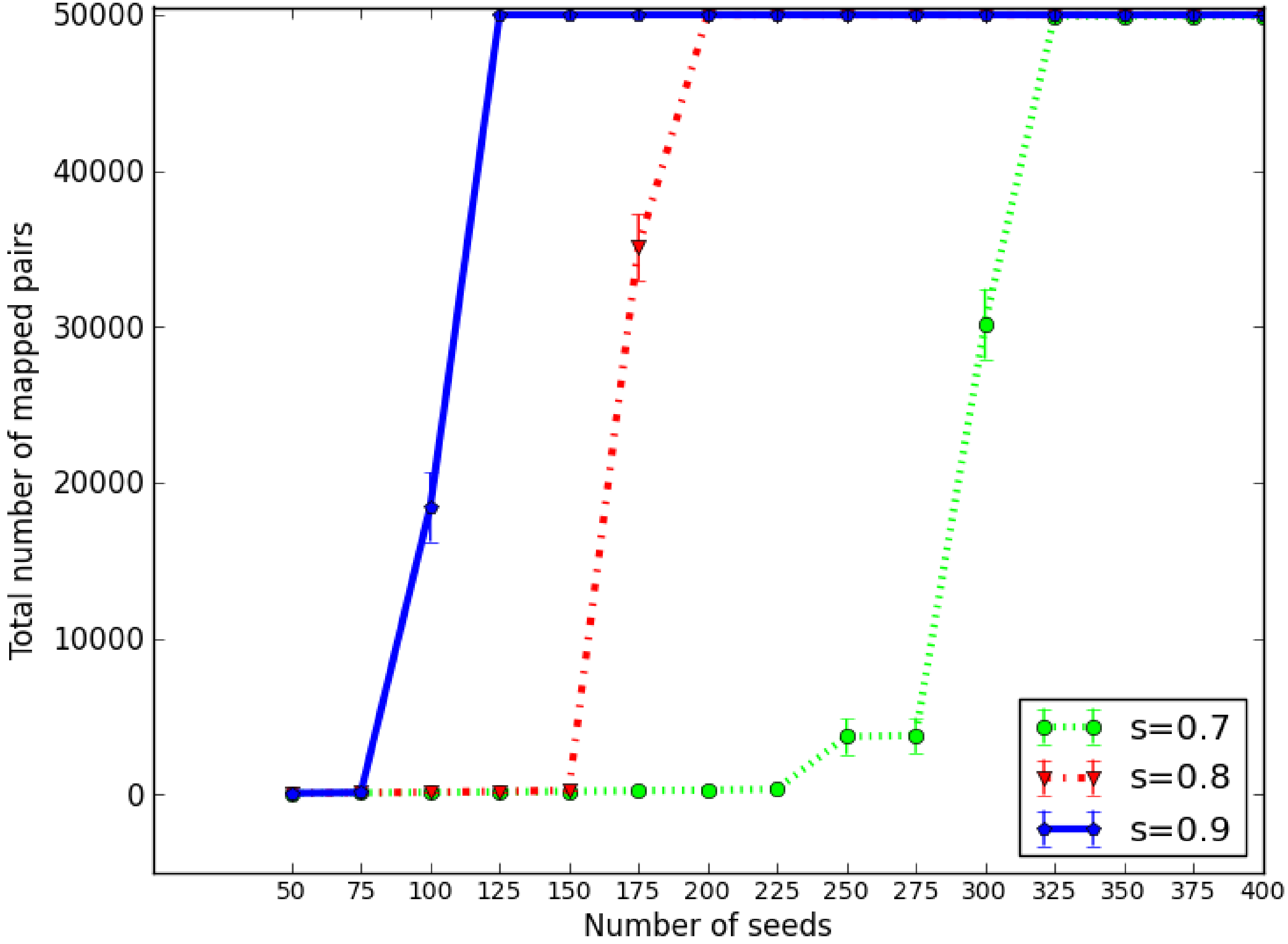
Giant Component: Branching Process



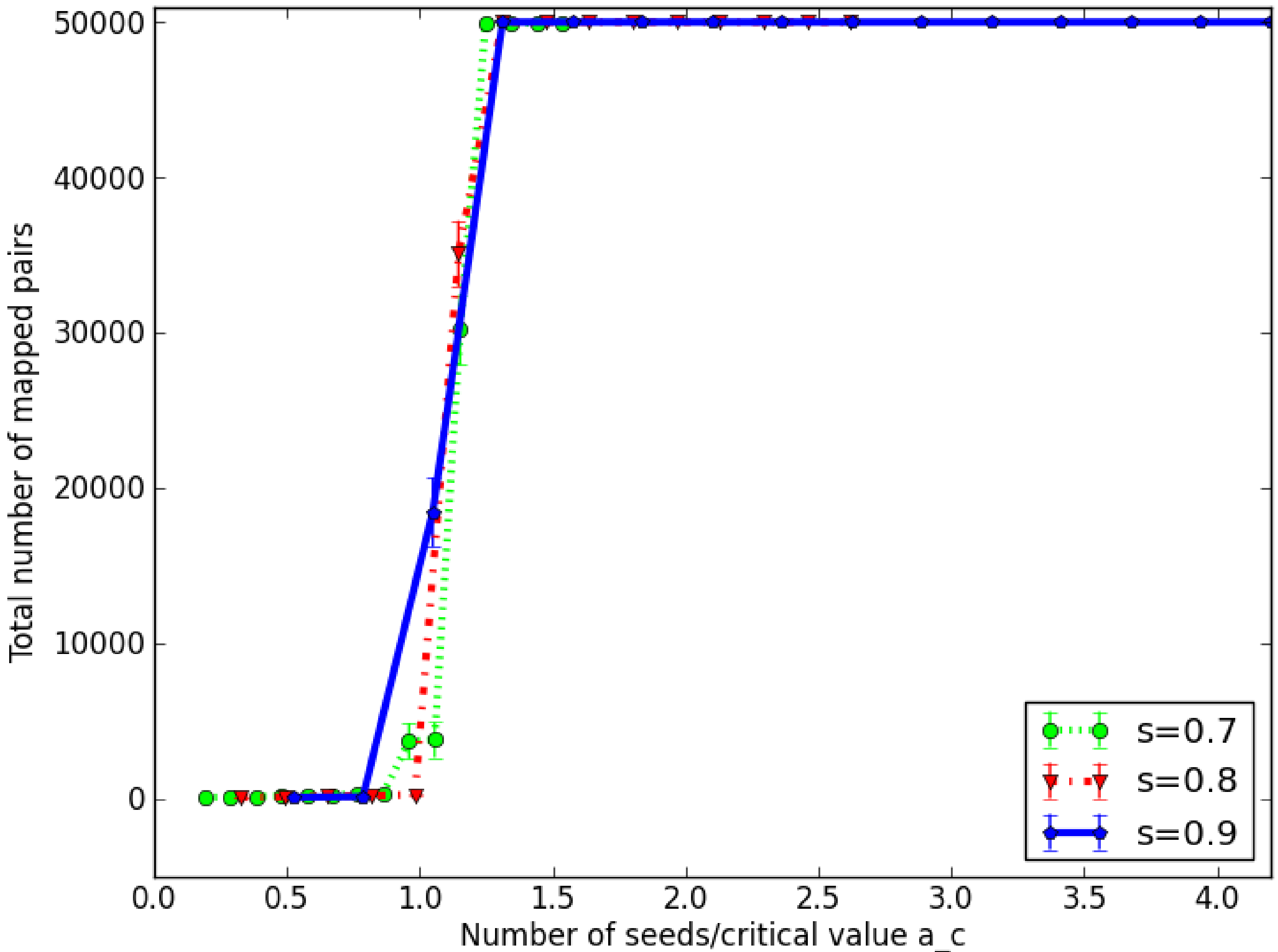
Bottleneck in Bootstrap Percolation



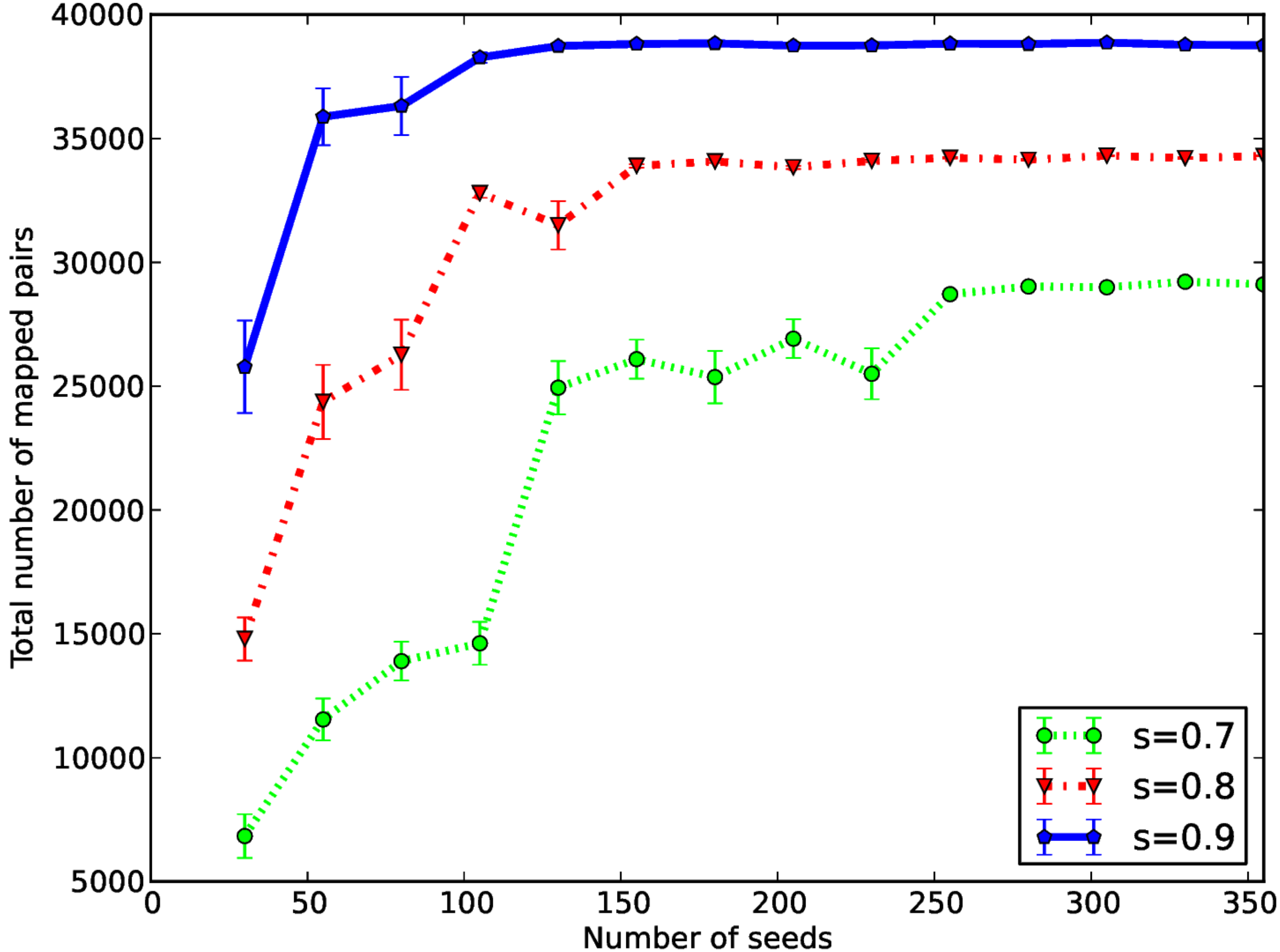
Simulation of PGM with $G(n, p; s)$ Network



Simulation of PGM with $G(n, p; s)$ Network

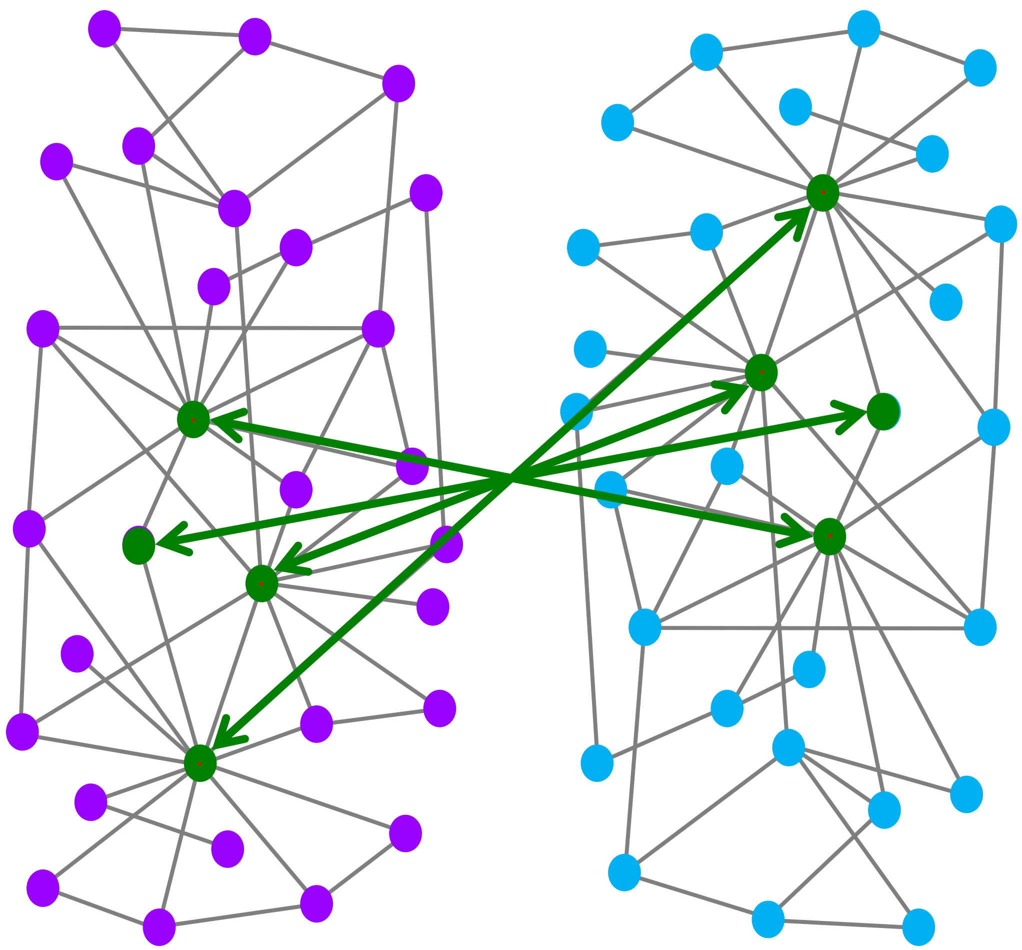


Slashdot Social Network

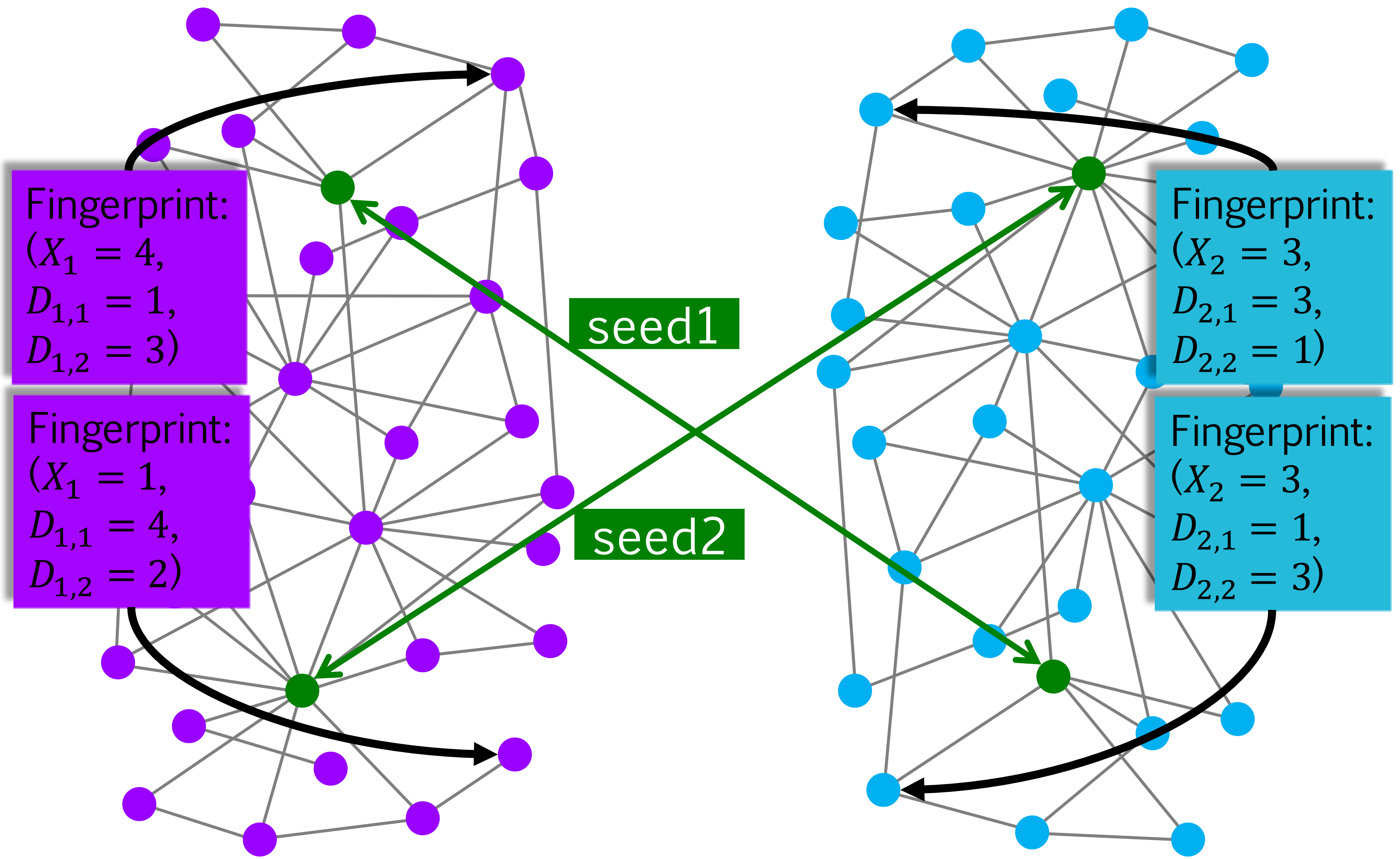


Result 3: Getting Started

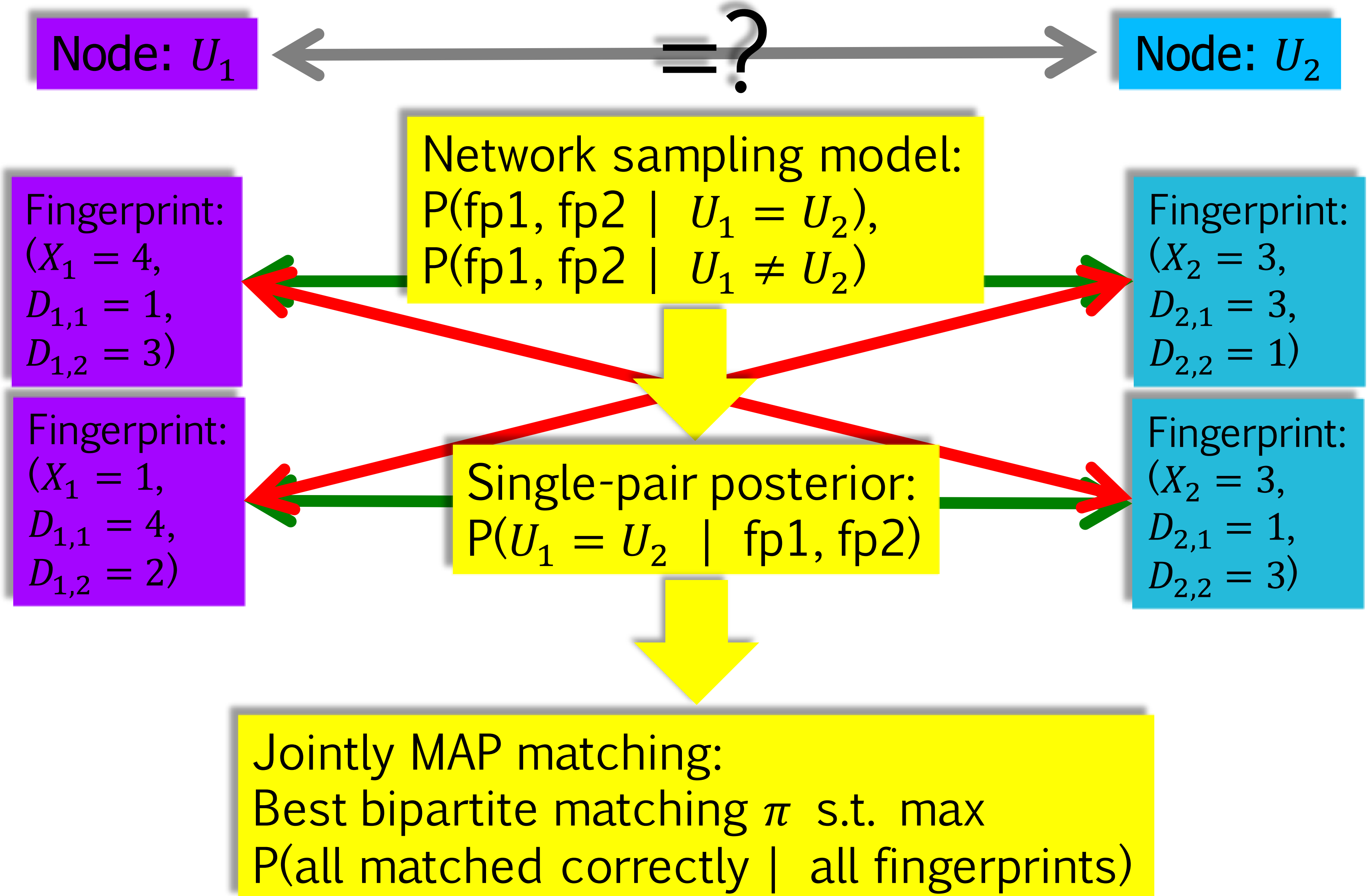
- **How to find seeds? [PFG13]**
  - Efficient (polynomial) algorithm to generate seed set
  - Does not work for $G(n, p)$
- **Real graphs:**
  - More heterogeneous than $G(n, p)$: degree skew, transitivity
  - Provides features for nodes



Finding Seeds: Bayesian Framework



Seed: Bayesian Framework



Conclusion

- **Graph matching problem:**
  - Social networks: privacy; merging
  - Model as noisy graph isomorphism problem
  - How much information in network structure?
- **$G(n, p; s)$ random graph model:**
  - Parsimonious: density (p), similarity (s)
  - Information-theoretic characterization of feasible region – condition is quite mild
- **Percolation Graph Matching algorithm:**
  - Simple algorithm, propagating evidence over node pairs
  - Actually works very well in practice; parsimonious (r)
- **Analysis:**
  - Sharp phase transition in seed set size (a), confirms empirical observation