Privacy, Anonymity, and Ambiguity in Social and Information Networks

Matthias Grossglauser, EPFL

CTW 2013

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AOL Search Log Release (2006)

```
4417749 care packages                               2006-03-02 09:19:32
4417749 movies for dogs                             2006-03-02 09:24:14
4417749 blue book                                   2006-03-03 11:48:52
4417749 best dog for older owner                    2006-03-06 11:48:24
4417749 best dog for older owner                    2006-03-06 11:48:24
4417749 rescue of older dogs                        2006-03-06 11:55:25
4417749 school supplies for the iraq children       2006-03-06 13:36:33
4417749 school supplies for the iraq children       2006-03-06 13:36:33
4417749 pine straw lilburn delivery                 2006-03-06 18:35:02
4417749 pine straw delivery in gwinnett county      2006-03-06 18:36:35
4417749 landscapers in lilburn ga.                  2006-03-06 18:37:26
4417749 pne straw in lilburn ga.                    2006-03-06 18:38:19
4417749 pine straw in lilburn ga.                   2006-03-06 18:38:27
4417749 gwinnett county yellow pages                2006-03-06 18:42:08
...
```

anonymized user ID

User 4417749 Uncovered by New York Times

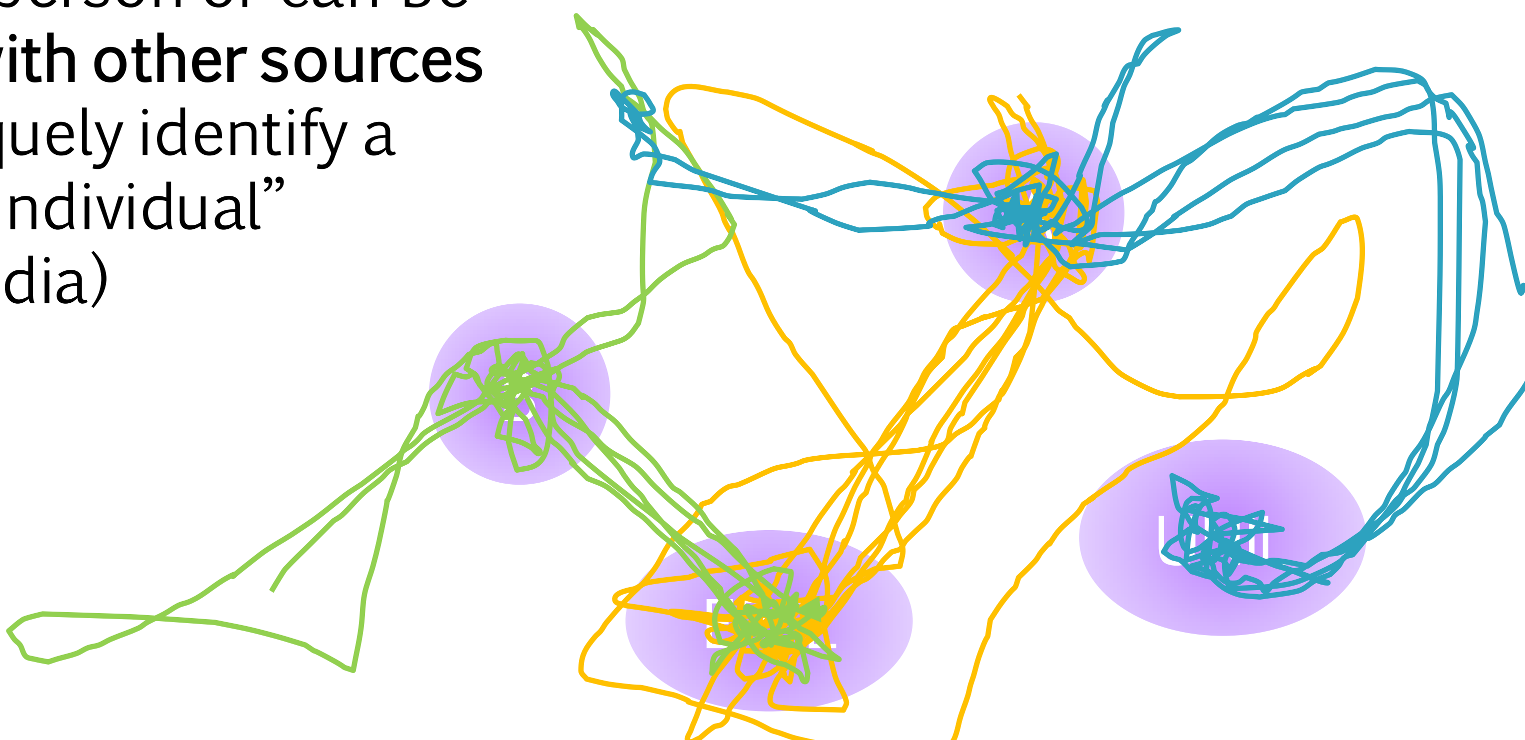
- Searches:
  - `“landscapers in Lilburn, Ga”`
  - `“homes sold in shadow lake subdivision gwinnett county georgia”`
  - `“jarrett t. arnold”, “jack t. arnold”`

- 4417749=Thelma Arnold
  - 62 years old widow and dog owner
  - home: Lilburn, GA
- AOL press release:
  - “There was no personally identifiable data provided by AOL with those records, but search queries themselves can sometimes include such information.”
- Heads had to roll…
  - AOL CTO Maureen Govern (+2 others) fired

Privacy: Hard to Define

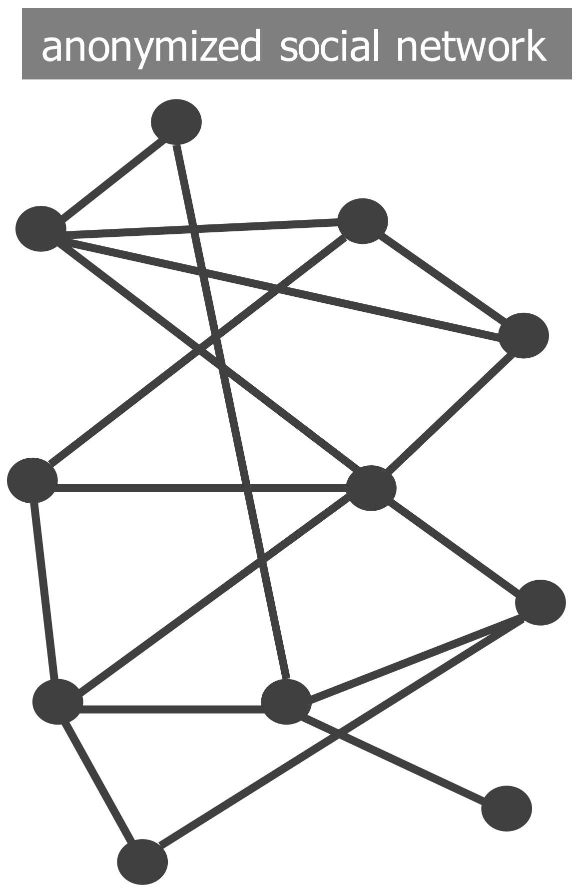
- **Personally identifiable information (PII):**
  - "information that can be used to uniquely identify, contact, or locate a single person or can be used **with other sources** to uniquely identify a single individual" (wikipedia)

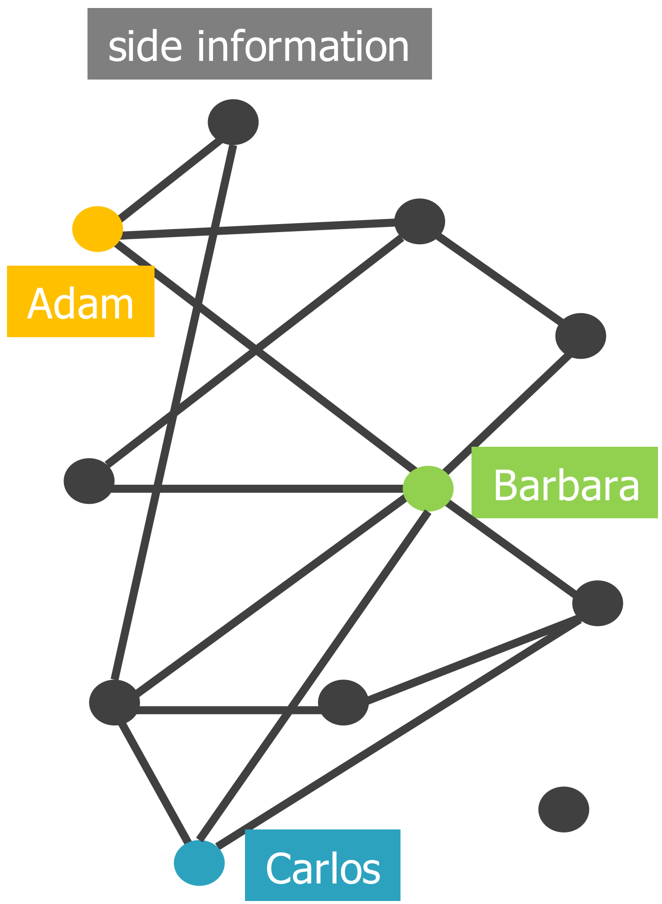
Name	Home	Work
Adam	A	EPFL
Barbara	B	EPFL
Carlos	A	UNIL



Privacy of Networks

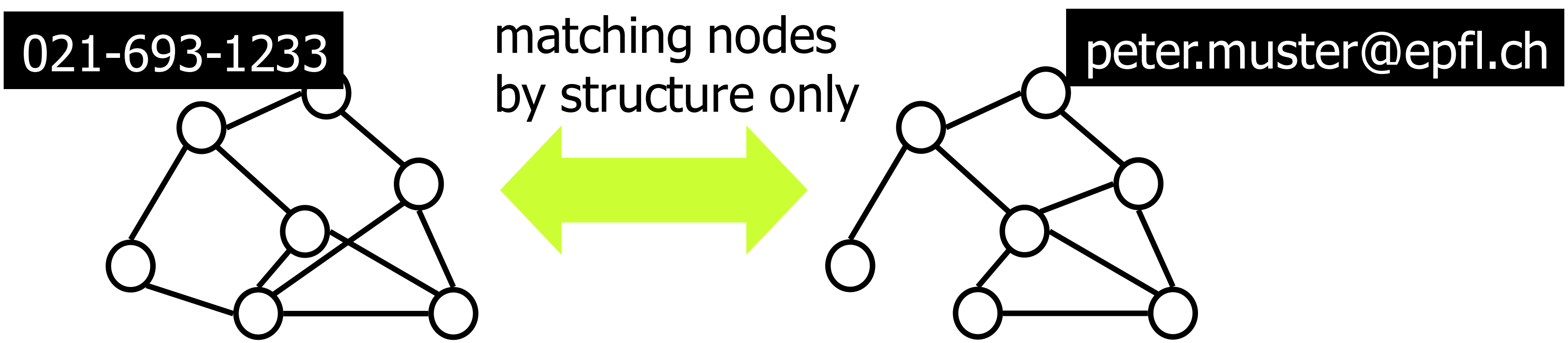
- **Adversary has:**
  - Anonymized network = unlabeled graph
  - Side information: subgraph; statistics on certain nodes; noisy version of whole network; …





Graph Matching

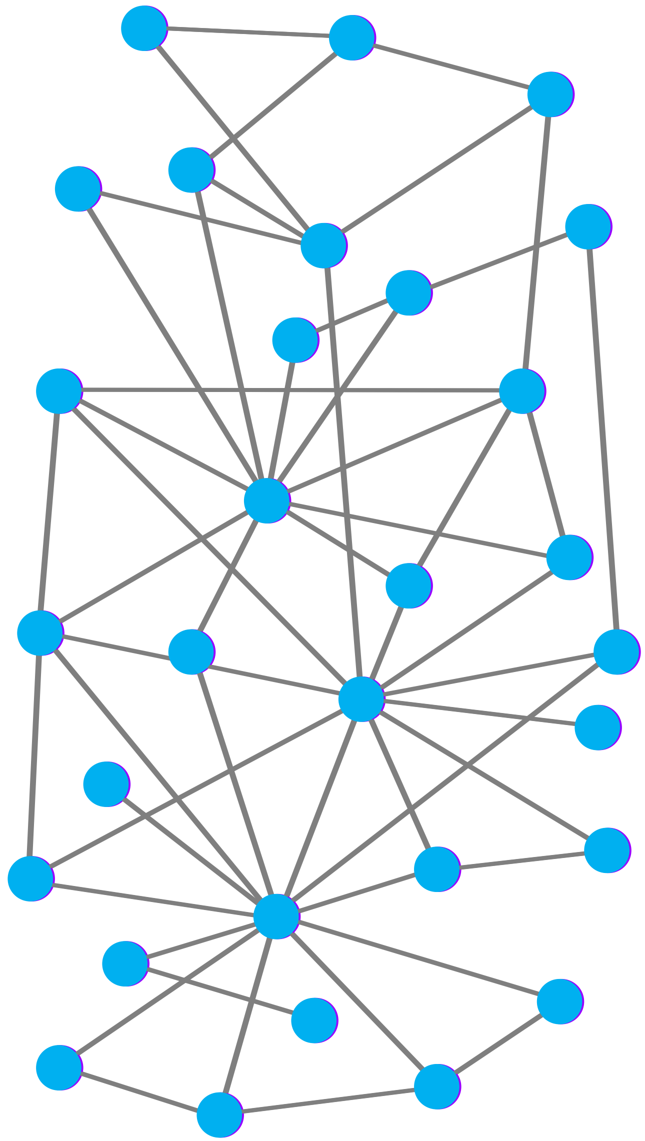
- **Other applications:**
  - Find overlap in networks:
    - Social networks from different domains & time slots
    - Identify viruses by function-call patterns
    - Computer vision: matching segment graphs for different viewing angles
    - …



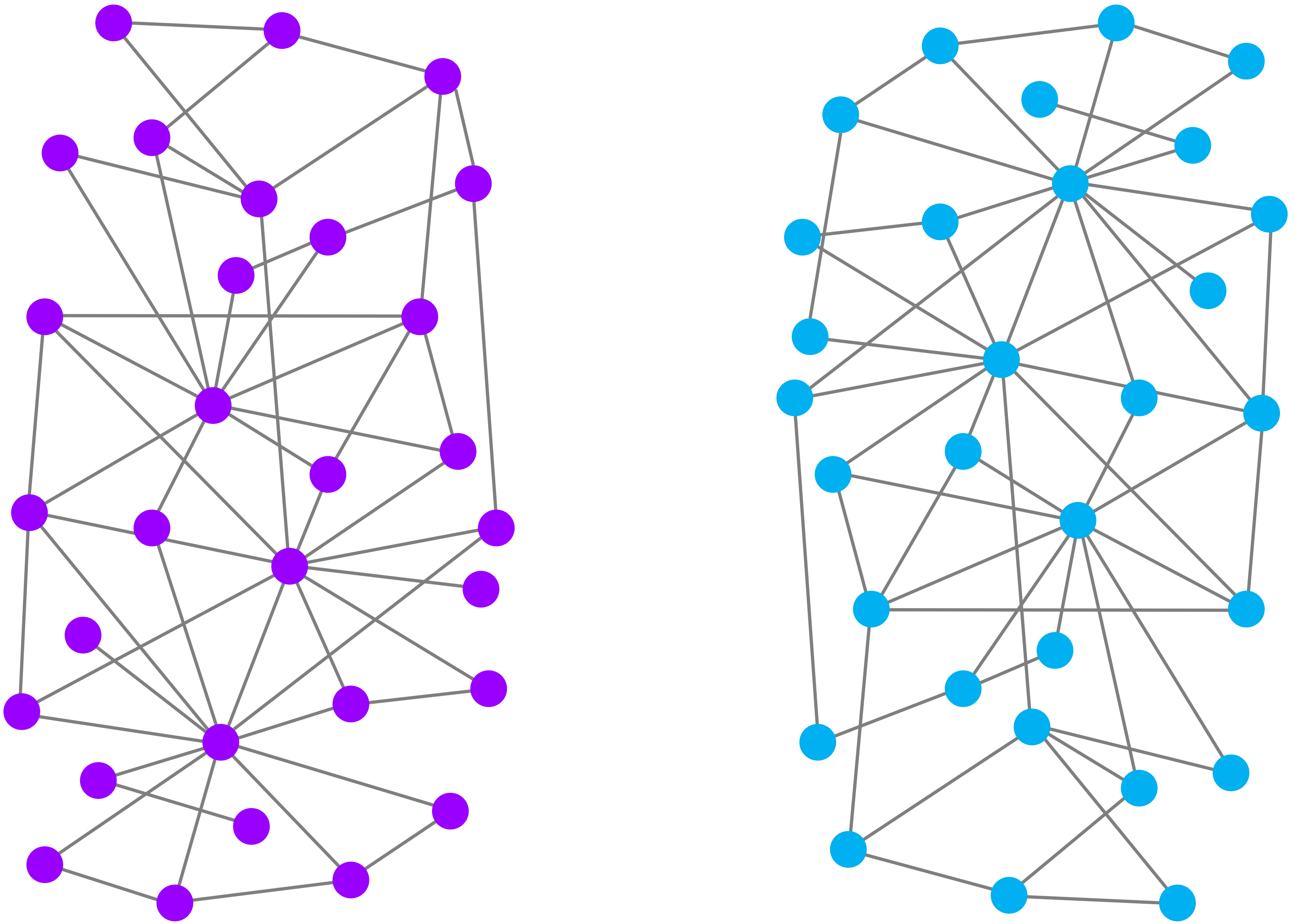
No limits

Fundamental feasibility w/o side information, but with ∞ time and memory

Graph Matching



Graph Matching with Noise

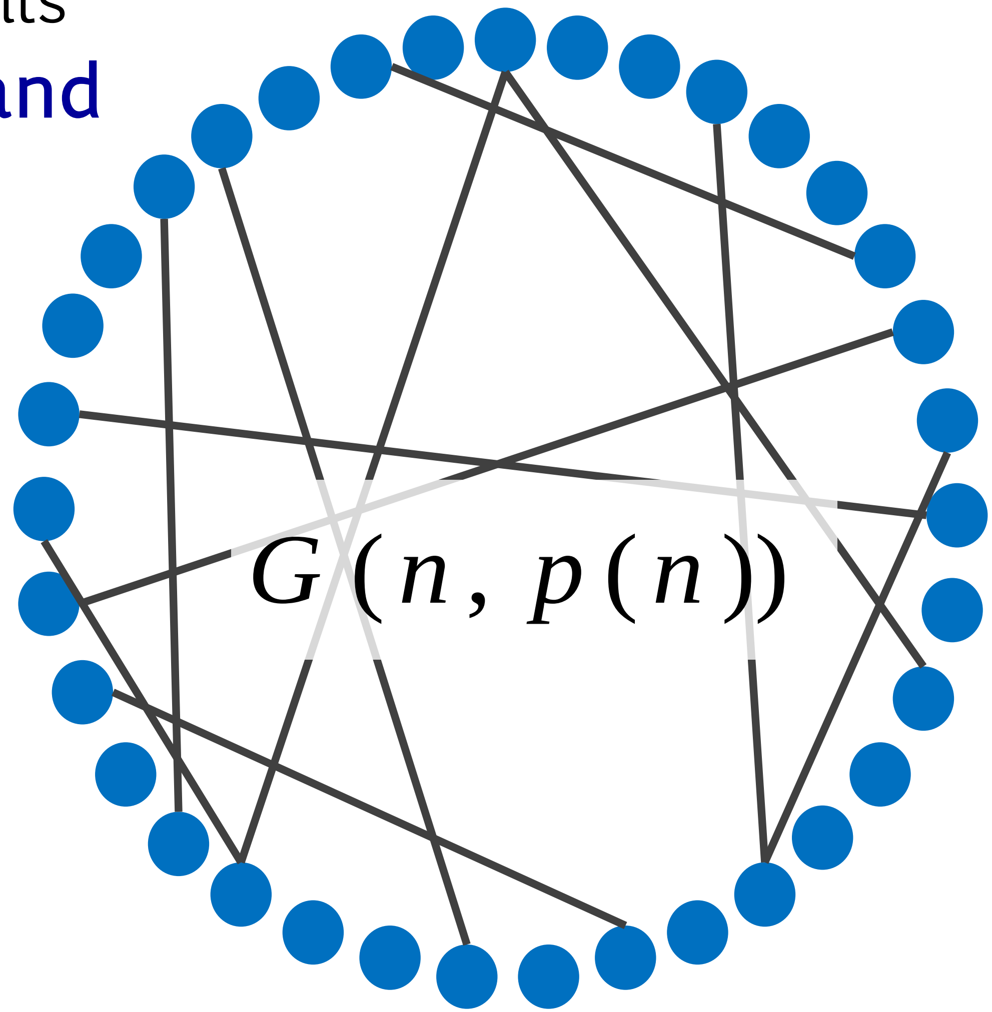


Question

- Is it fundamentally hard or easy to match similar graphs by structure?
- Fundamental =
  - Information-theoretic: ignore computational & memory cost
  - Hard: in addition to second graph, no other side information
  - Demanding: want to match every vertex

Random Graphs Instant Primer

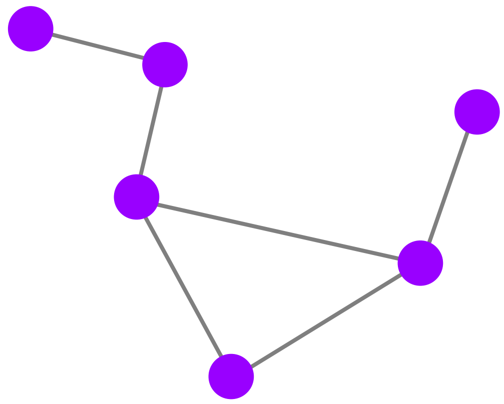
- First published 1959 by Erdös & Rényi
  - Focus on existence results
- Large n asymptotics and phase transitions
  - Connectivity
  - Existence of subgraphs
  - Giant component
  - Chromatic number
  - Automorphism group
  - …



Threshold for asymmetry:
$p = \log n / n$

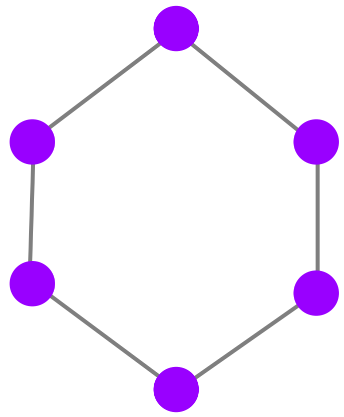
Automorphism: Asymmetric vs Symmetric

Asymmetric



AuG = 1

Symmetric



AuG = 12

AuG = size of automorphism group

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

Generator $G = G(n, p)$

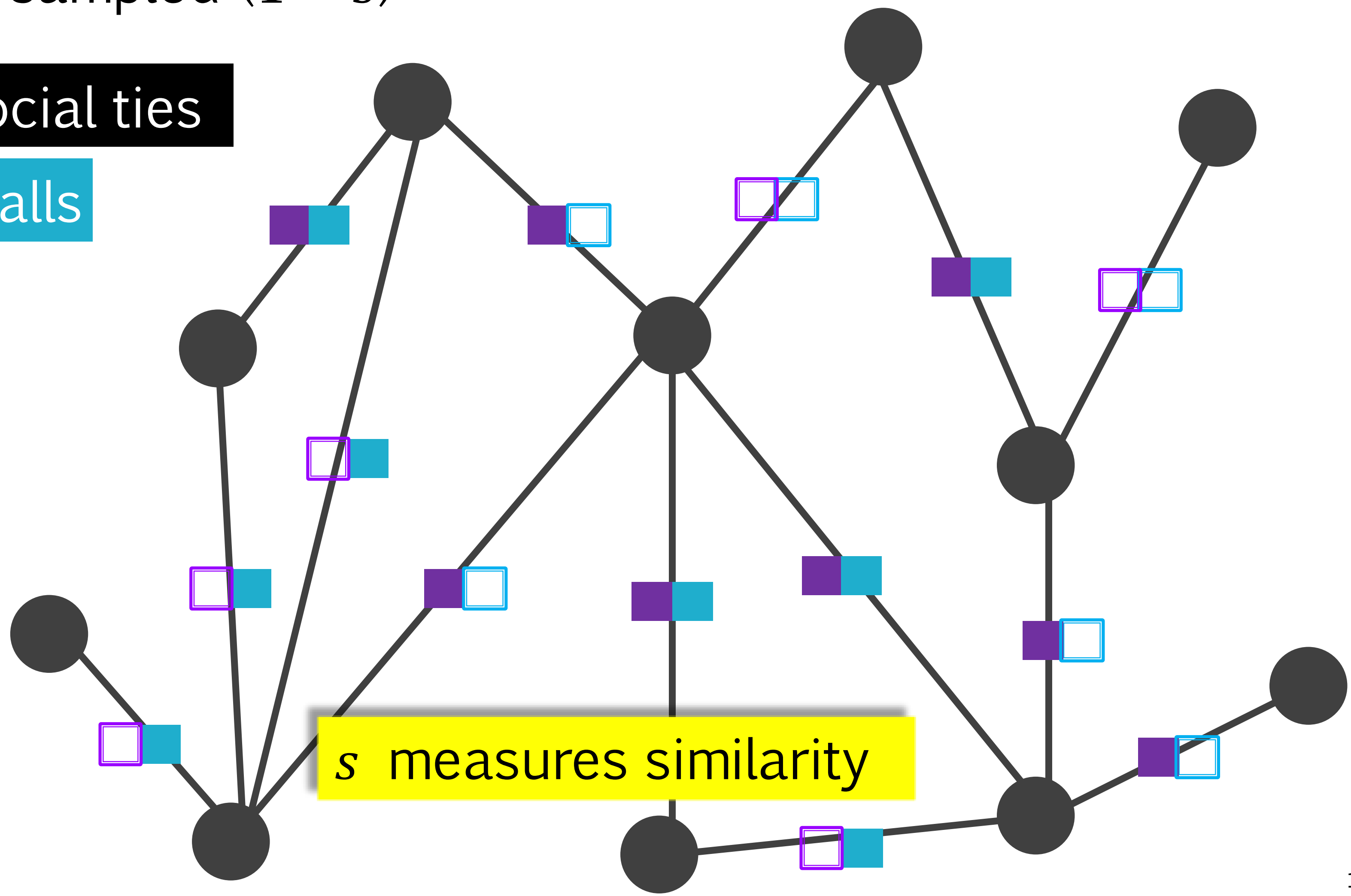
sampled (s)

not sampled ($1 - s$)

“real” social ties

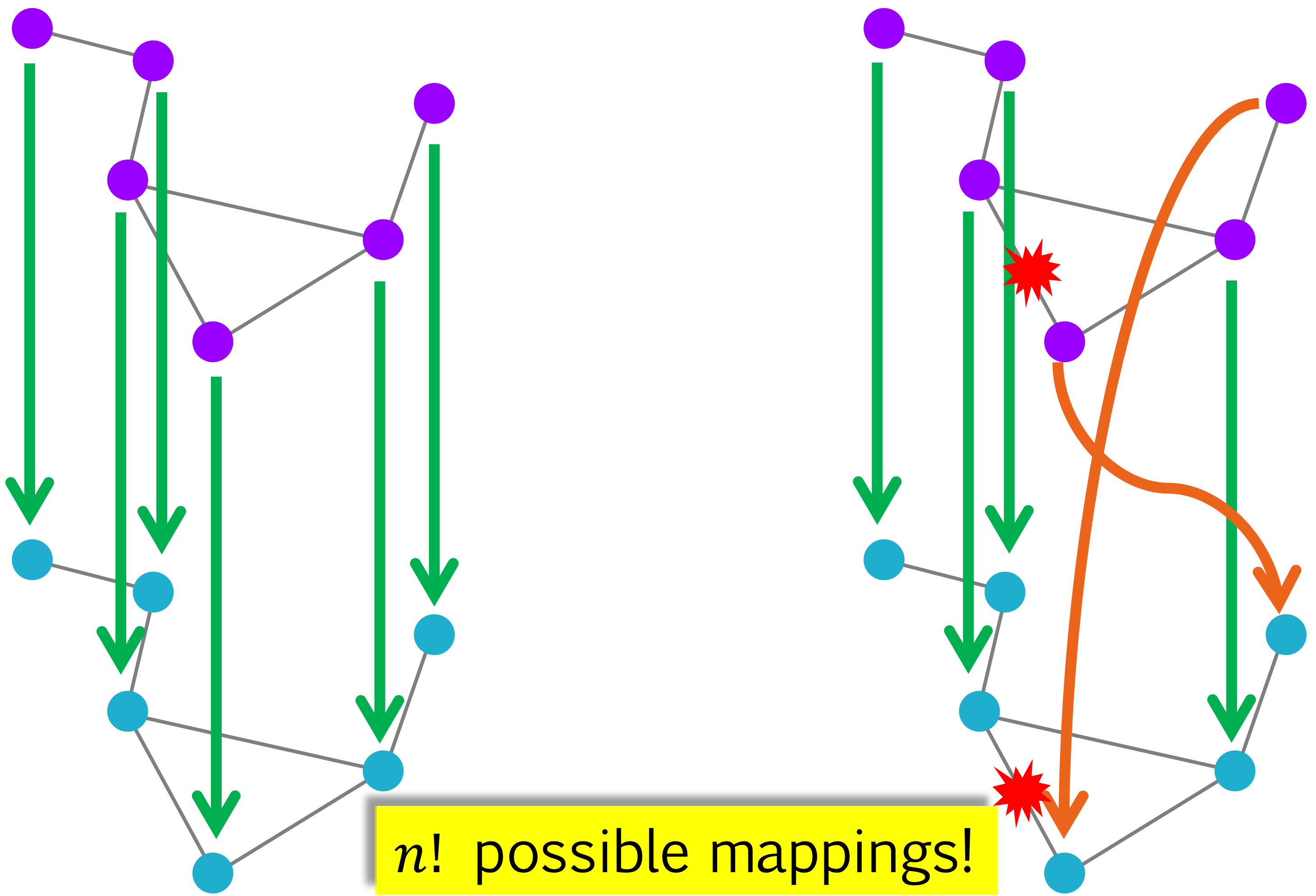
phone calls

emails



s measures similarity

Mappings and Edge Mismatch



$\Delta(\pi_0) = 0$ $\Delta(\pi) = 2$

Adversary Model

- **Assumption:**
  - Attacker has infinite computational power
  - Can try all possible mappings π and compute edge mismatch function Δ(π)
- **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: Difficult to Anonymize!

- The

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

threshold for aug(G)=1

  - For the G(n,p;s) matching problem, if

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

then the identity permutation minimizes Δ(.) a.a.s.

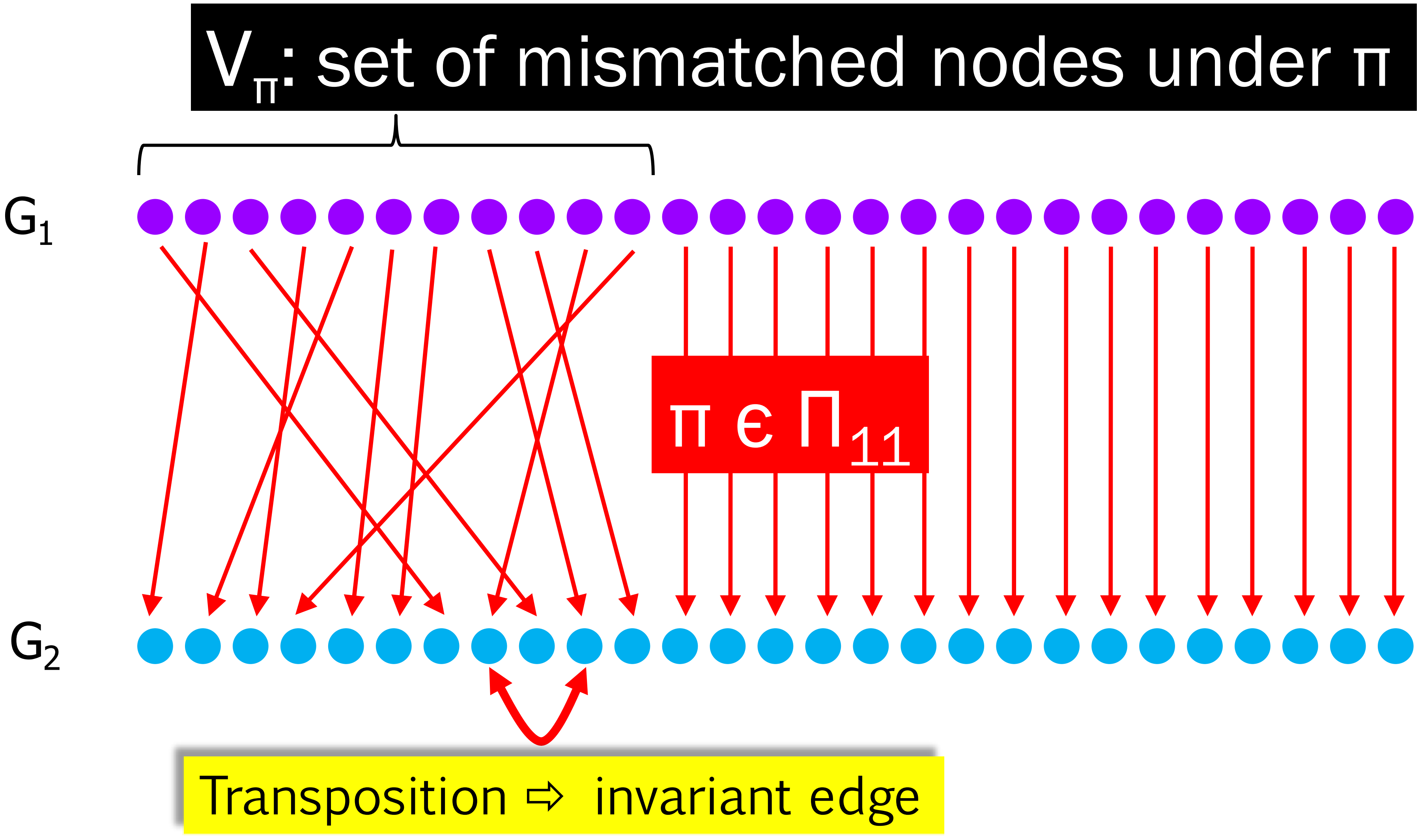
Penalty for difference G_1-G_2

"growing slowly"

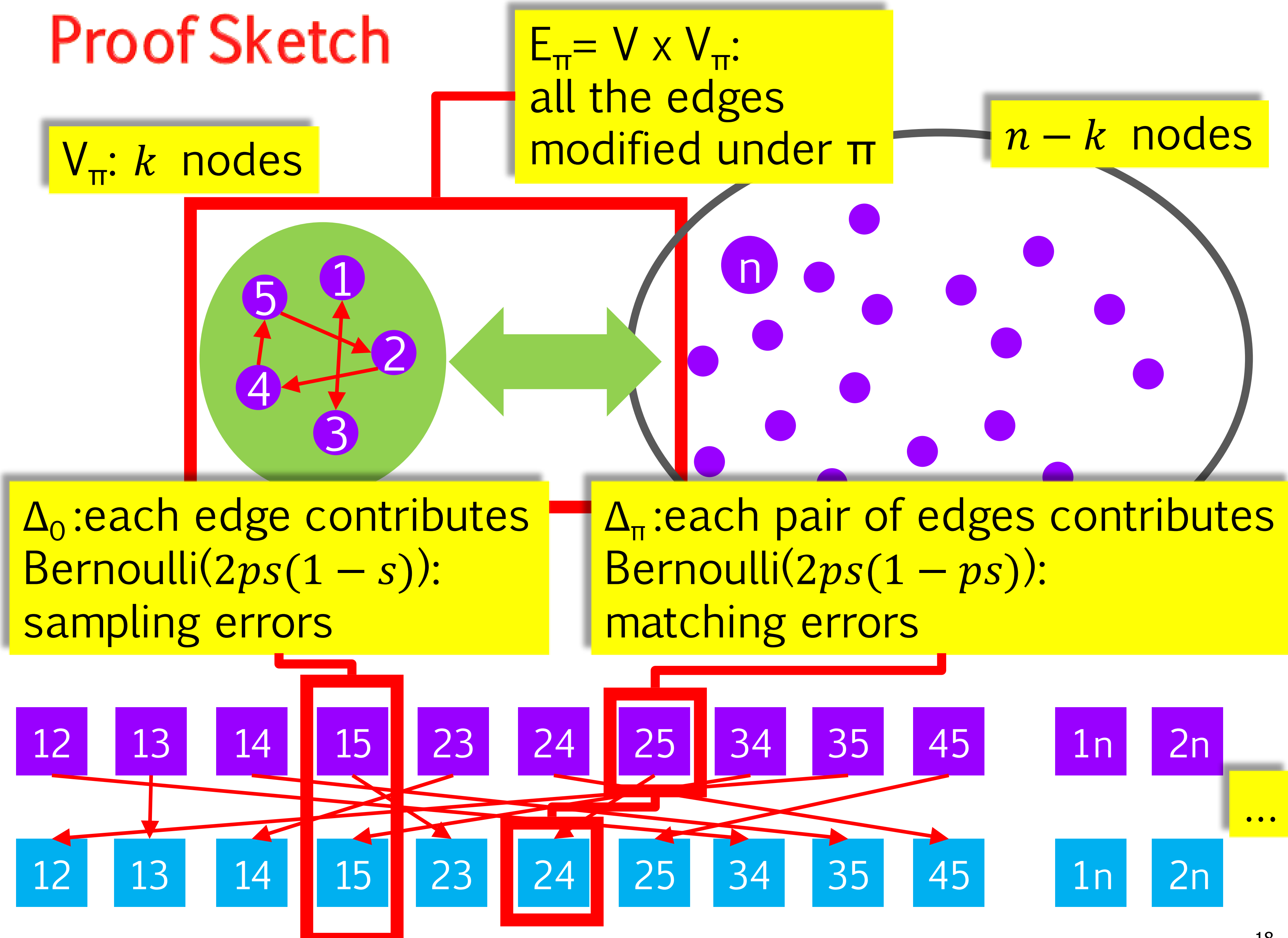
- Interpretation: two pieces of bad/good news
  - Surprisingly weak condition: degree growing faster than ~$\log n$ enough to break anonymity
  - Decrease with s only quadratic

Proof Sketch

- Fix a particular map **π**

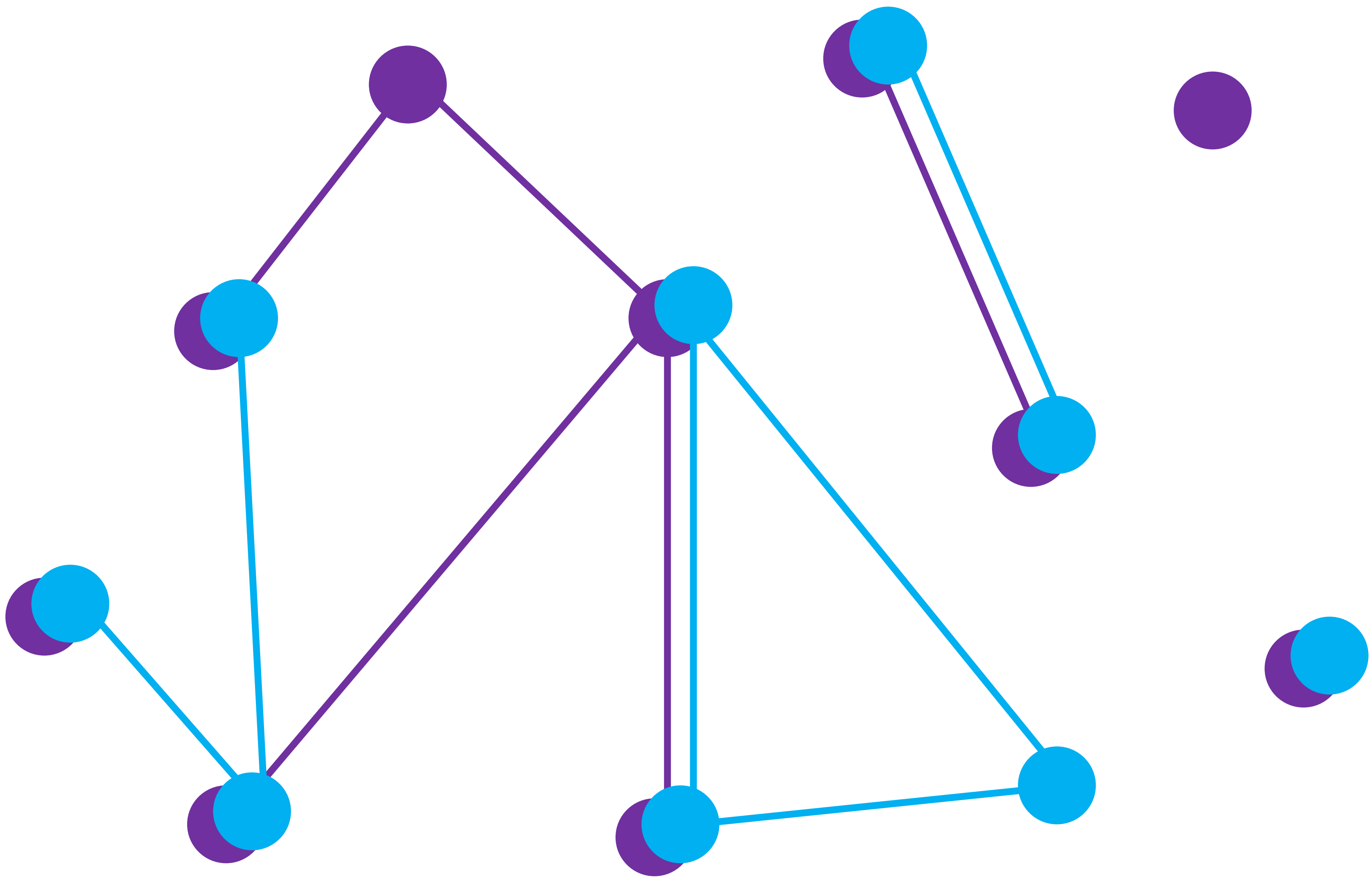


Proof Sketch



Extension: Node plus Edge Sampling

$G(n, p; s, t)$ matching problem



Extension: $G(n, p; s, t)$ Matching Problem

- **Result:**
  - Dependence on n still the same:

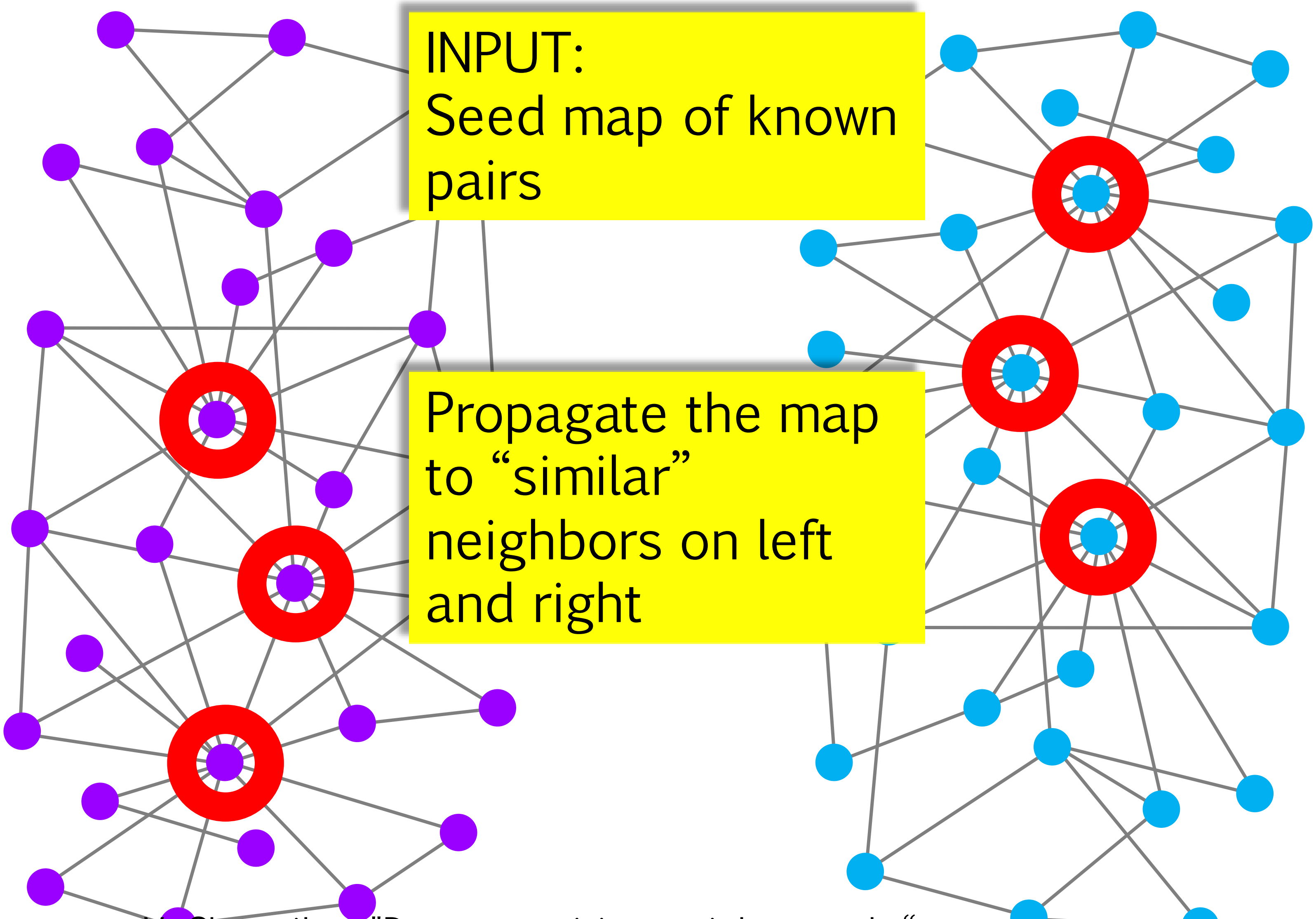
$$nps = c(s,t)\log n + \omega(1)$$

  - Dependence on s and t less intuitive
- **Interpretation:**
  - Node mismatch does not help/hurt too much either

Seeds = known matched pairs

Phase transition, and an efficient & tractable matching algorithm…

Map Propagation Heuristics



INPUT:
Seed map of known pairs

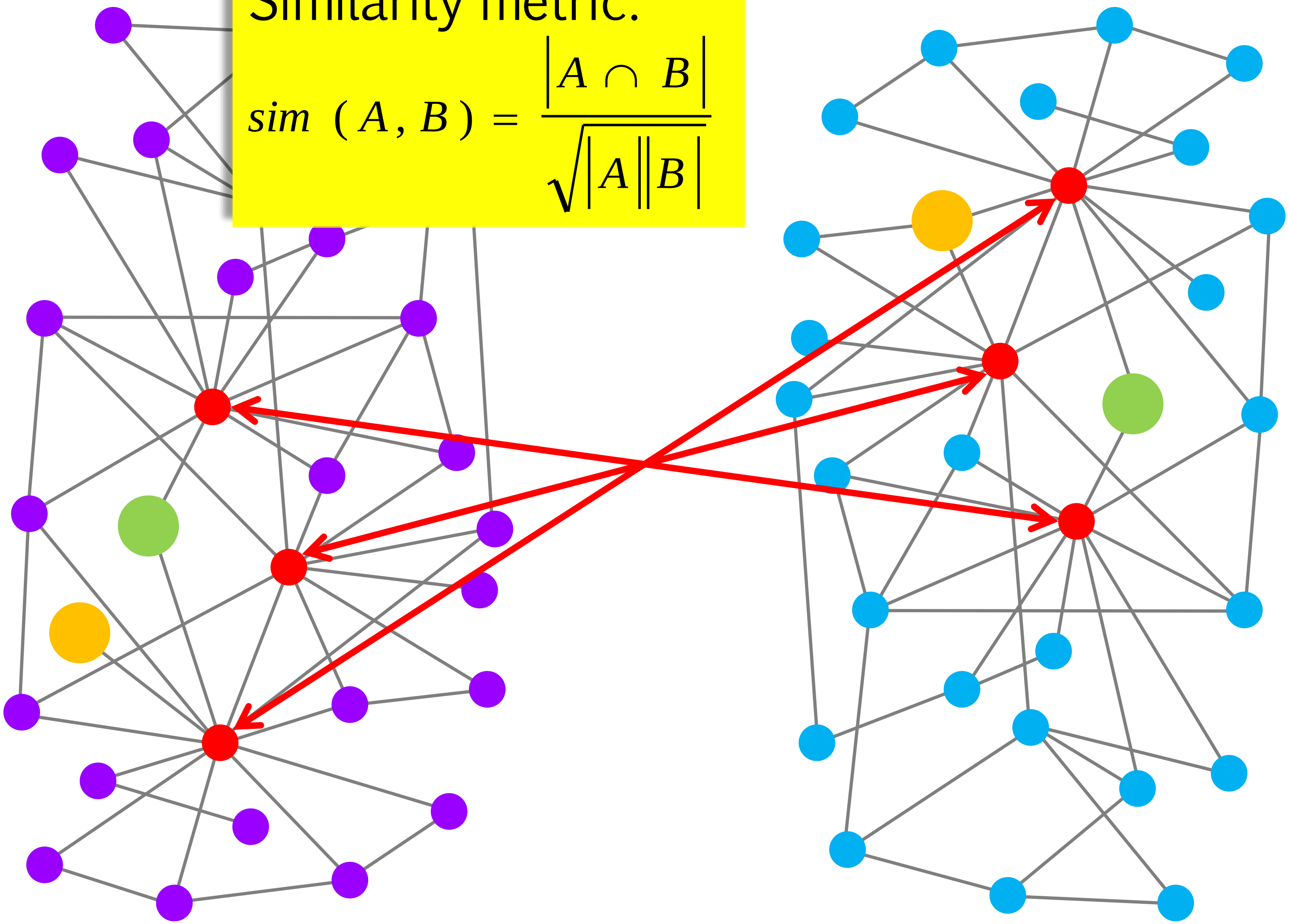
Propagate the map to “similar” neighbors on left and right

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

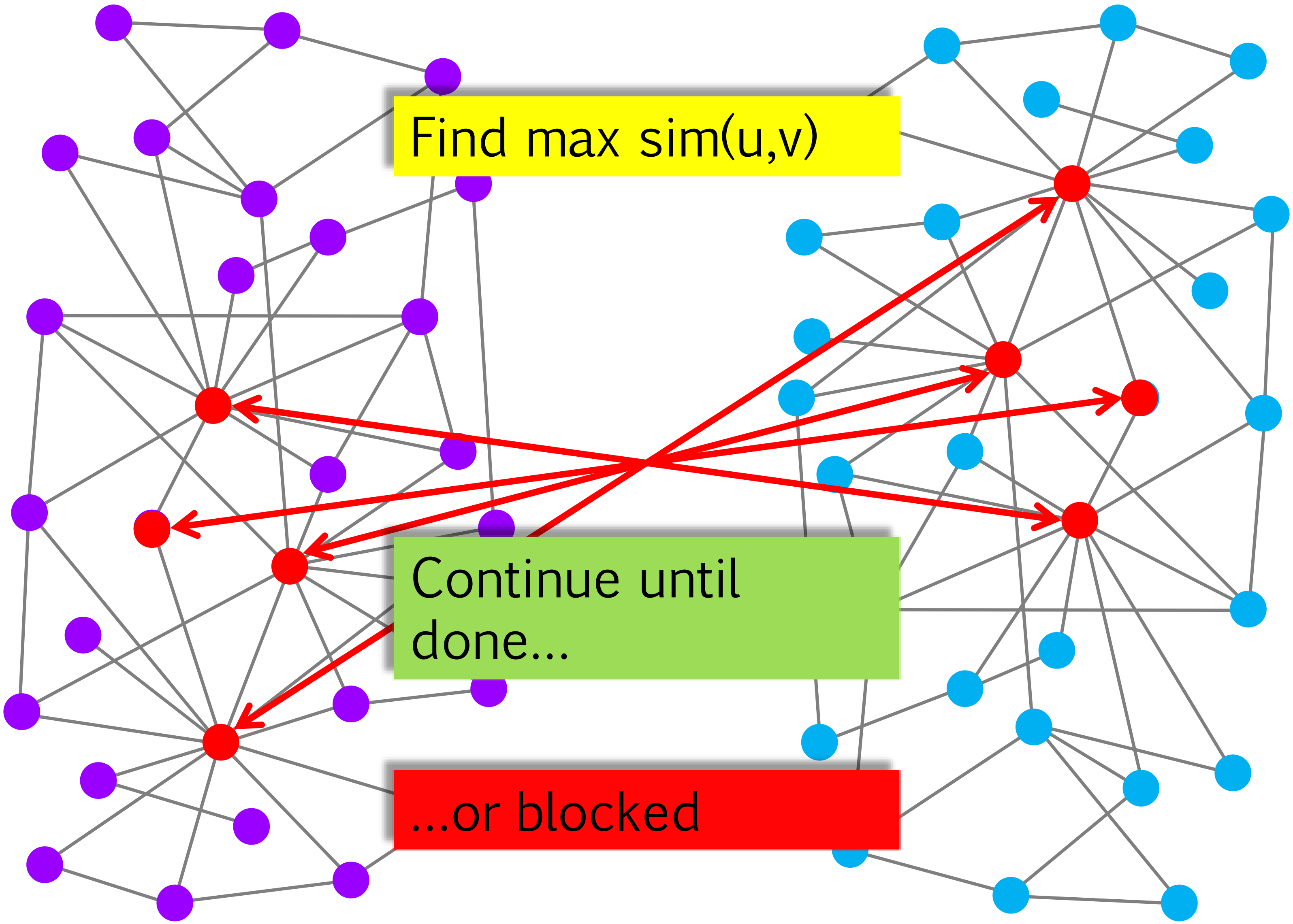
Neighborhood Overlap Metric

Similarity metric:

$$sim\ (A, B) = \frac{|A \cap B|}{\sqrt{|A||B|}}$$



Map Propagation



Questions

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

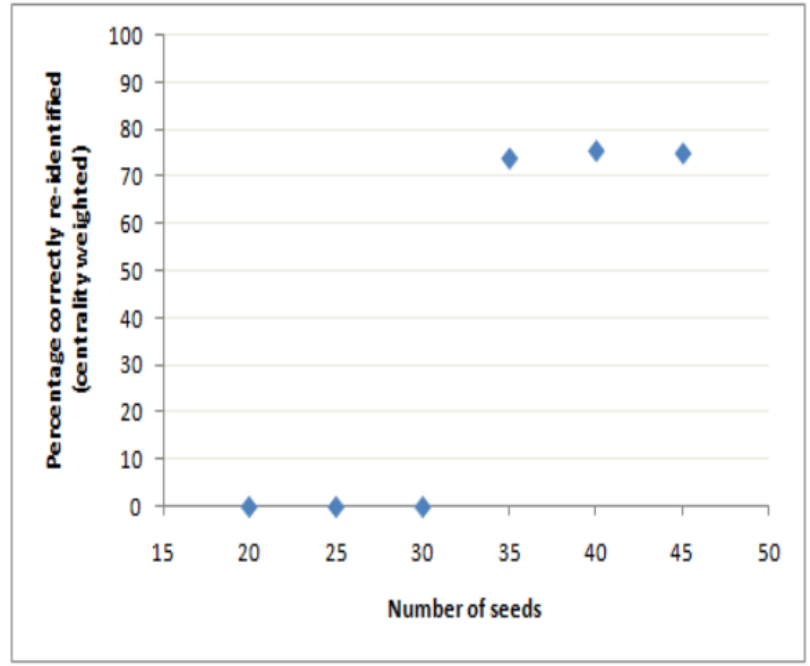
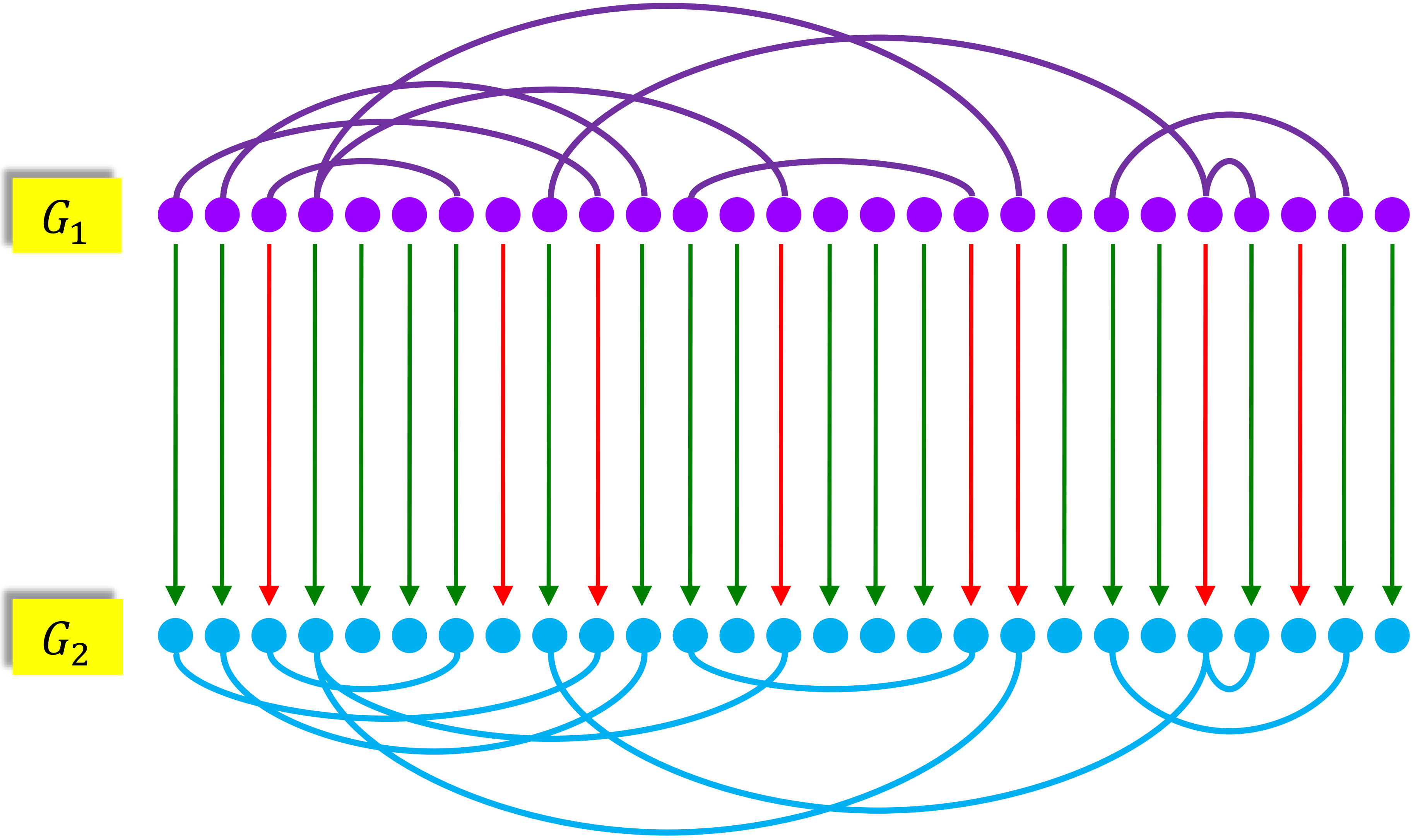


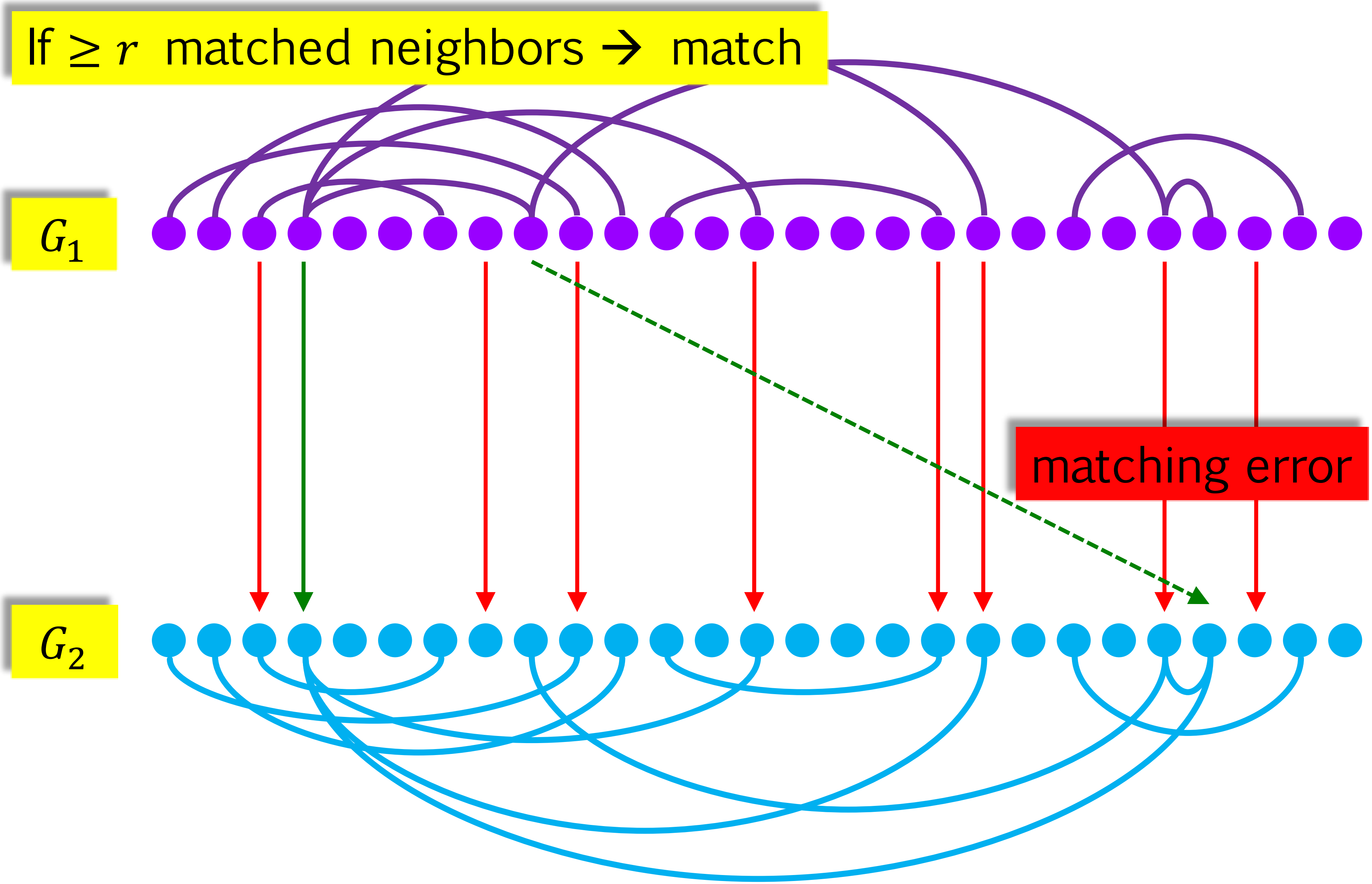
Figure 2. The fraction of nodes re-identified depends sharply on the number of seeds. Node overlap: 25%; Edge overlap: 50%

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

Model: Two Similar Graphs + Seeds

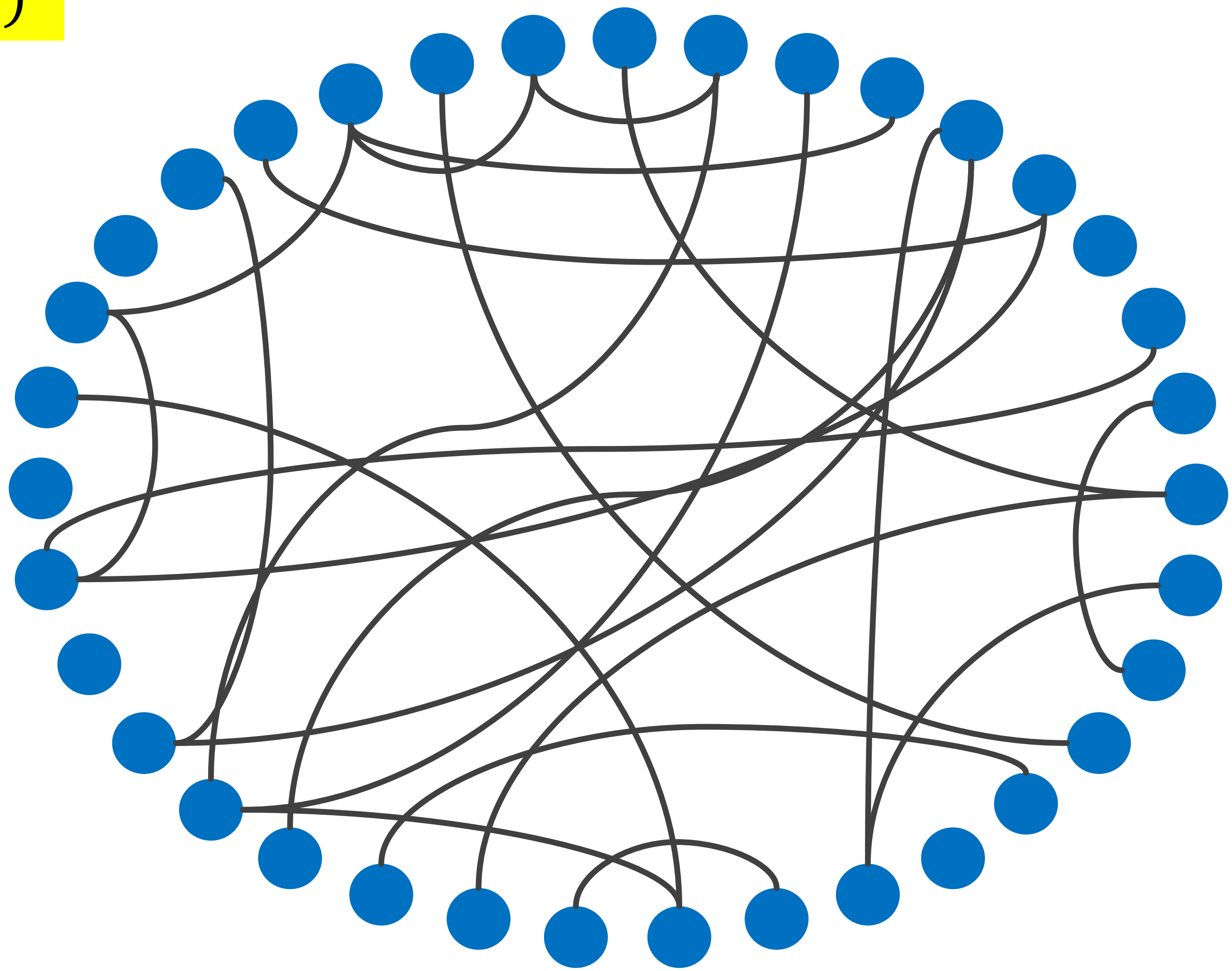


Percolation Graph Matching Algorithm

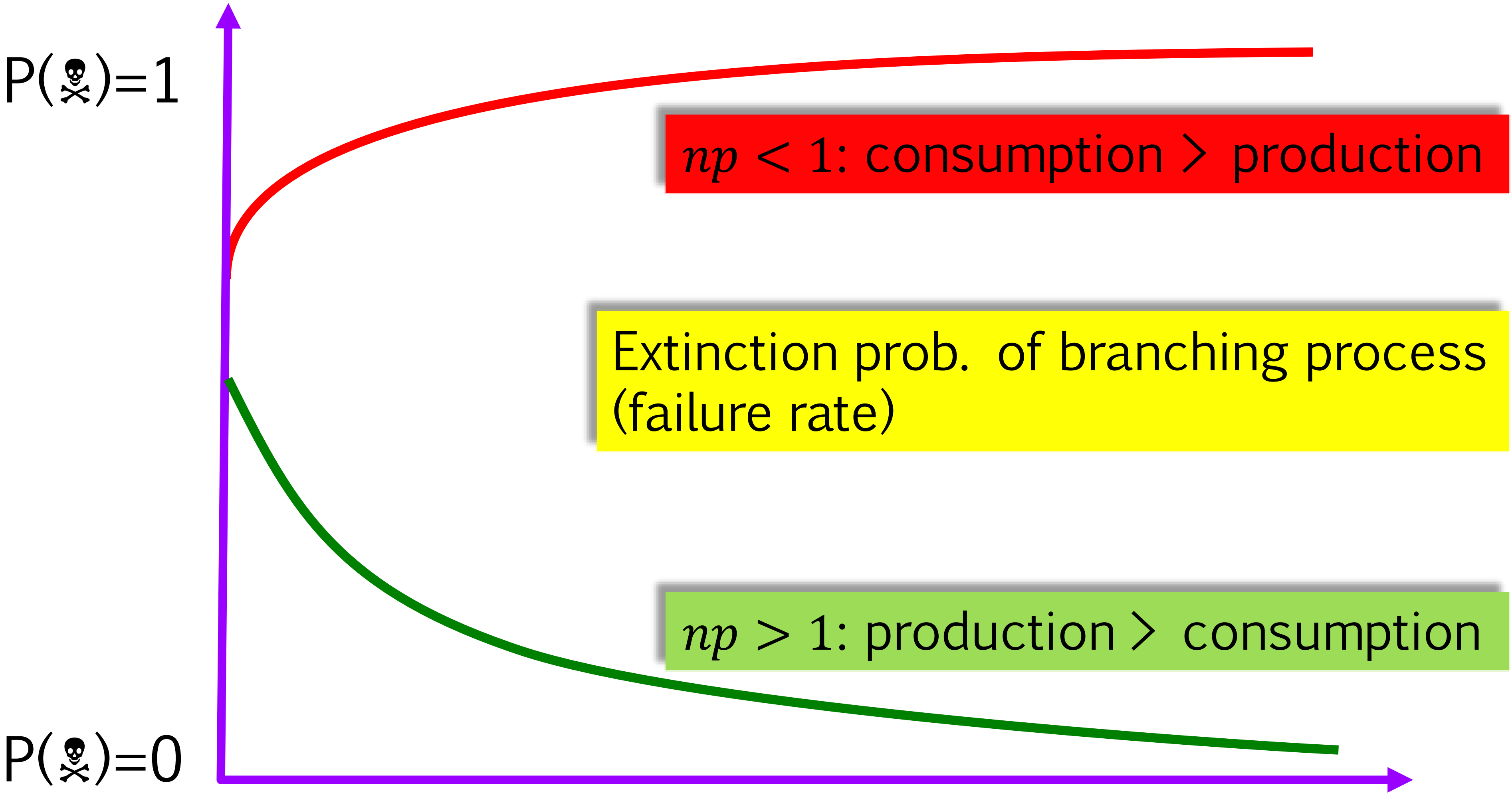


Giant Component for $G(n,p)$

$G(n,p)$

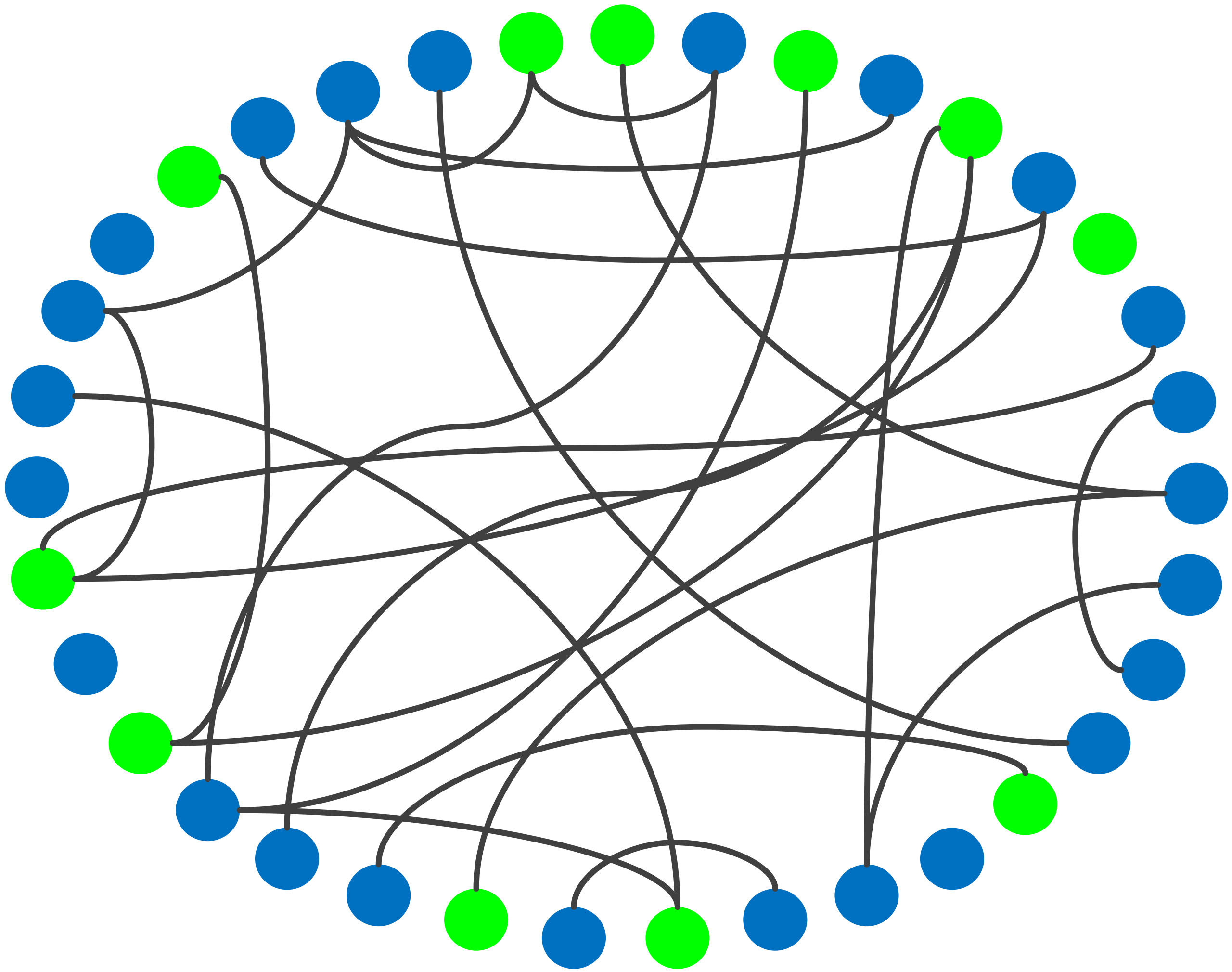


Giant Component: Branching Process



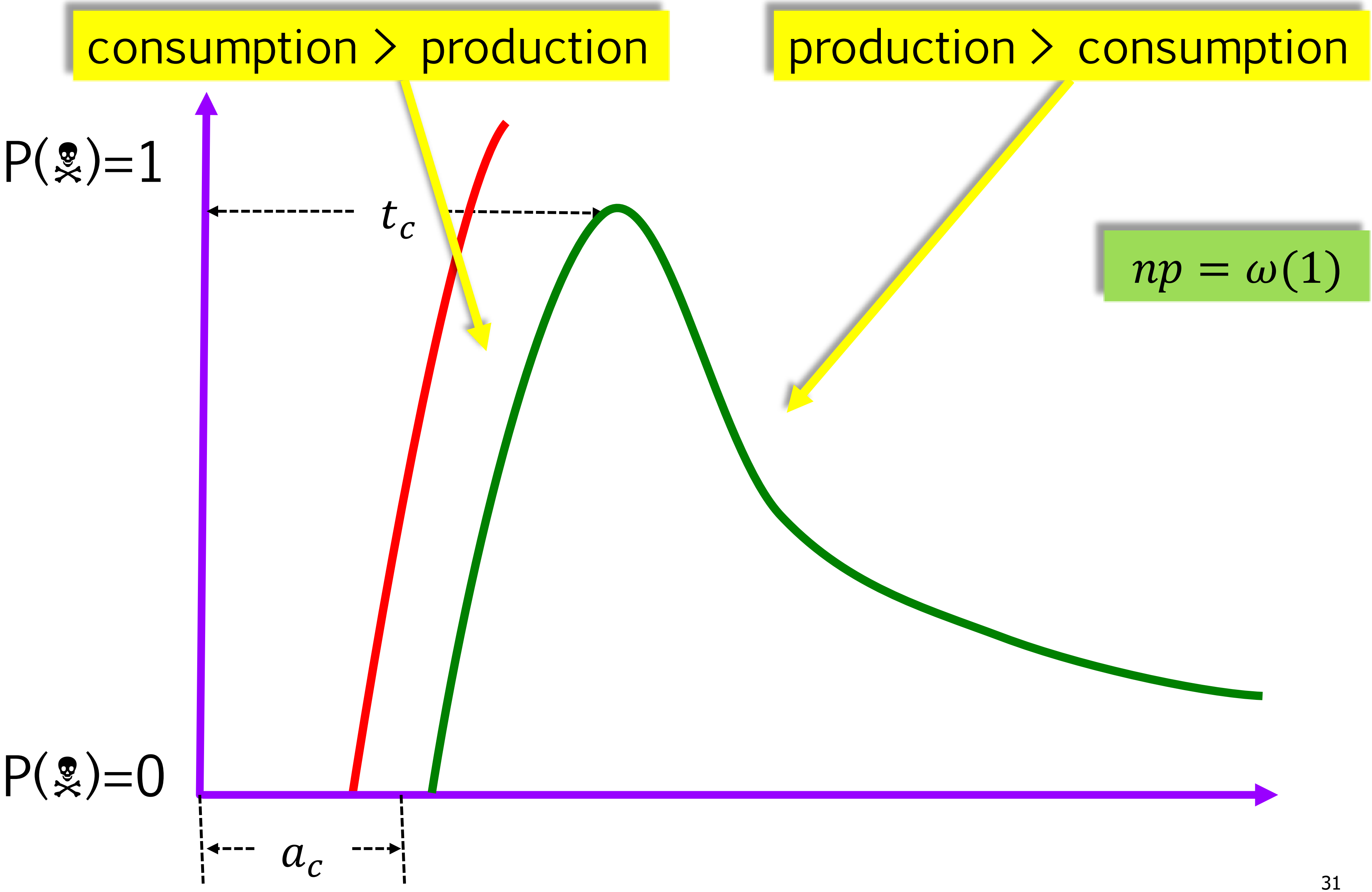
Bootstrap Percolation for $G(n, p)$

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]

Bottleneck in Bootstrap Percolation

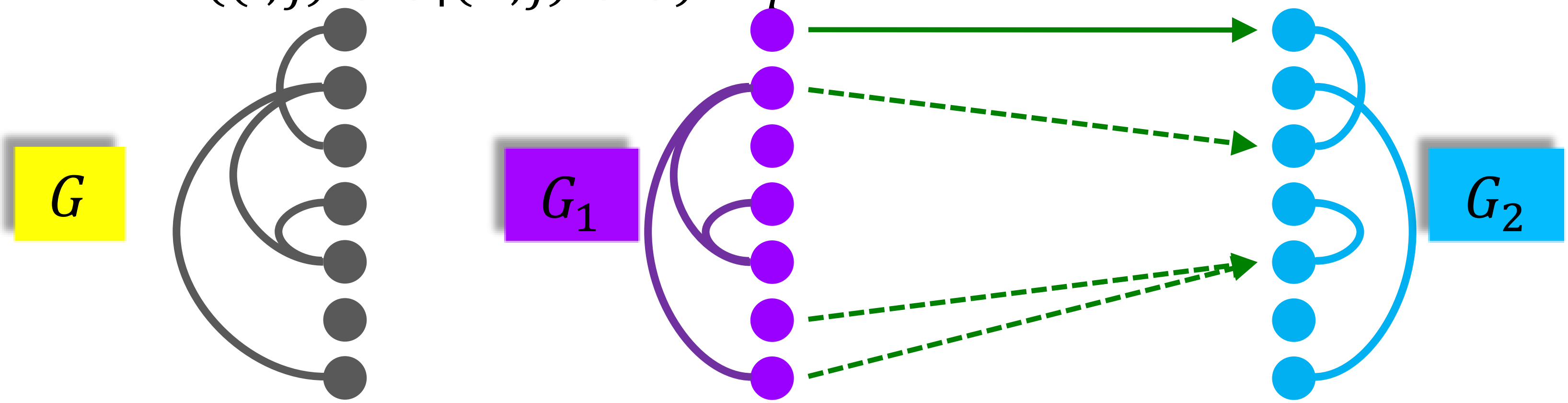


Results for Percolation Graph Matching

- **Theorem: phase transition in # seeds**
  - For $n^{-1} \ll ps \ll sn^{-\frac{1}{2}-\frac{3}{2r}}$:
  - 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:**
  - $a_c = (1 - r^{-1})t_c$
  - $t_c = \left(\frac{(r-1)!}{n(ps^2)^r}\right)^{1/(r-1)}$

Proof Sketch

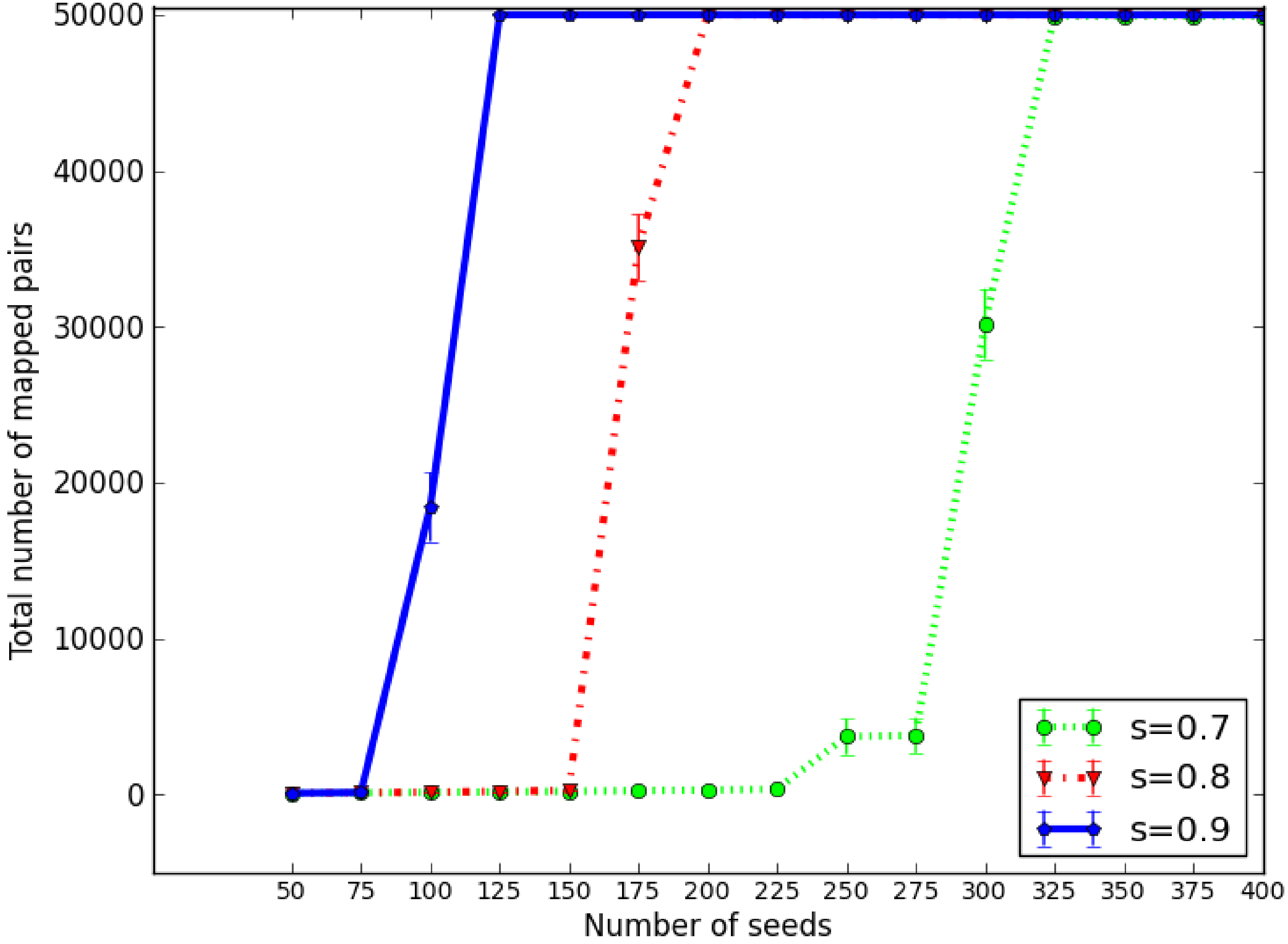
- **Bootstrap percolation in $G(n,p)$:**
  - \# credits of node i at time t: i.i.d. Binomials
- **Percolation graph matching in $G(n,p;s)$**
  - \# credits of pair (i,j) at time t: dependent, different Binomials
  - As long as no matching error so far, increments at t
  - Different: $(i,i) \sim Ber(ps^2), (i,j) \sim Ber((ps)^2)$
  - Dependent: for i, i', j all different:
    - $P\big((i,j)++\big) = (ps)^2$
    - $P((i,j)++|(i',j)++) = ps$



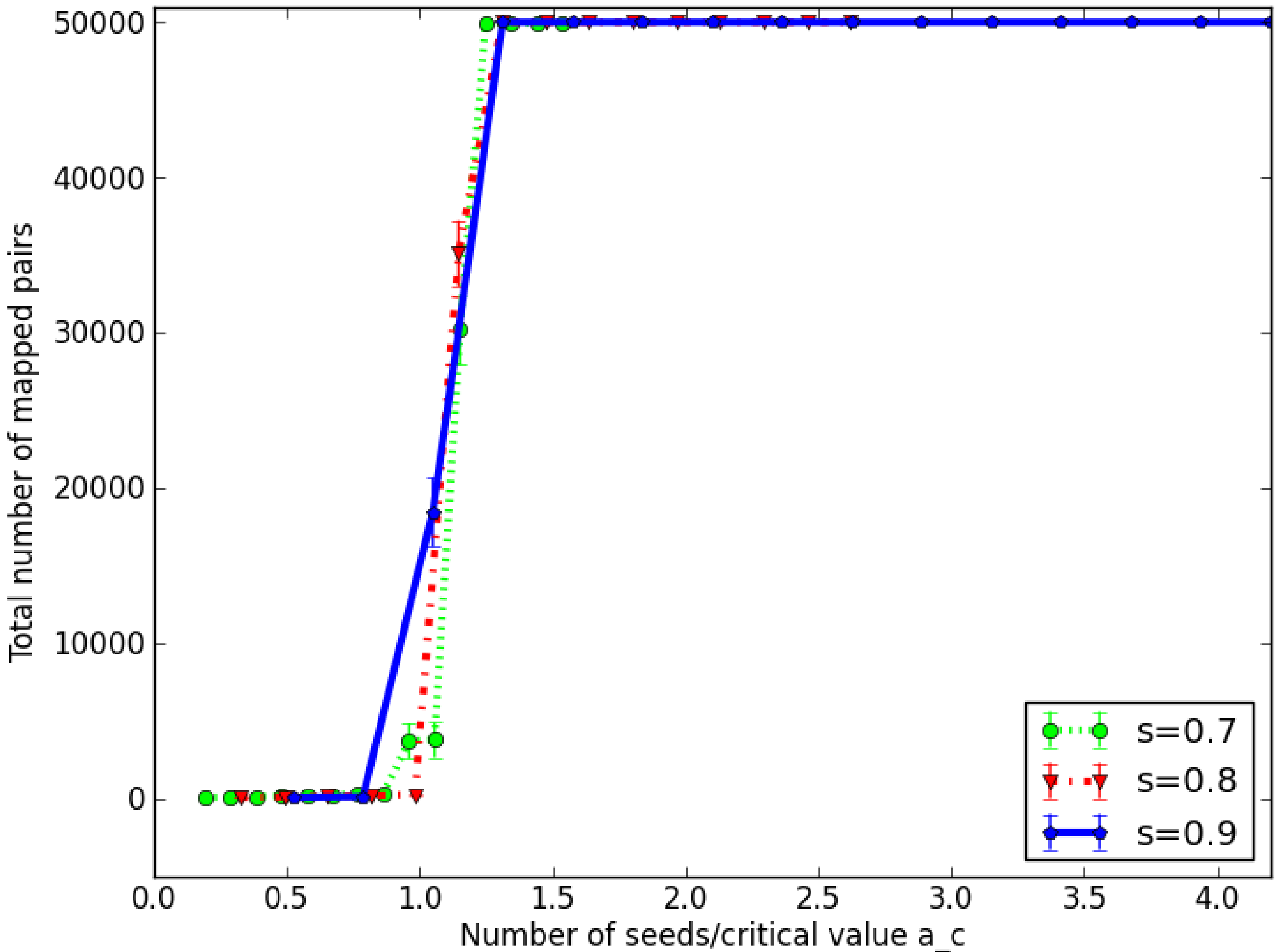
Proof Sketch

- Approach:
  - Focus on regime where X =no bad pair (i, j) get enough credits (r) to be potentially matched
  - True for $ps \ll n^{-\frac{1}{2}-\frac{3}{2r}}$
    - Need to choose r large enough (sparse graphs: $r \geq 4$, otherwise higher)
  - Conditional on X, only need to focus on good pairs (i, i)
  - Equivalence with bootstrap problem → does it percolate?
    - Need to have $n^{-1} \ll ps$
    - Need to have seed set size $a > a_c$ large enough

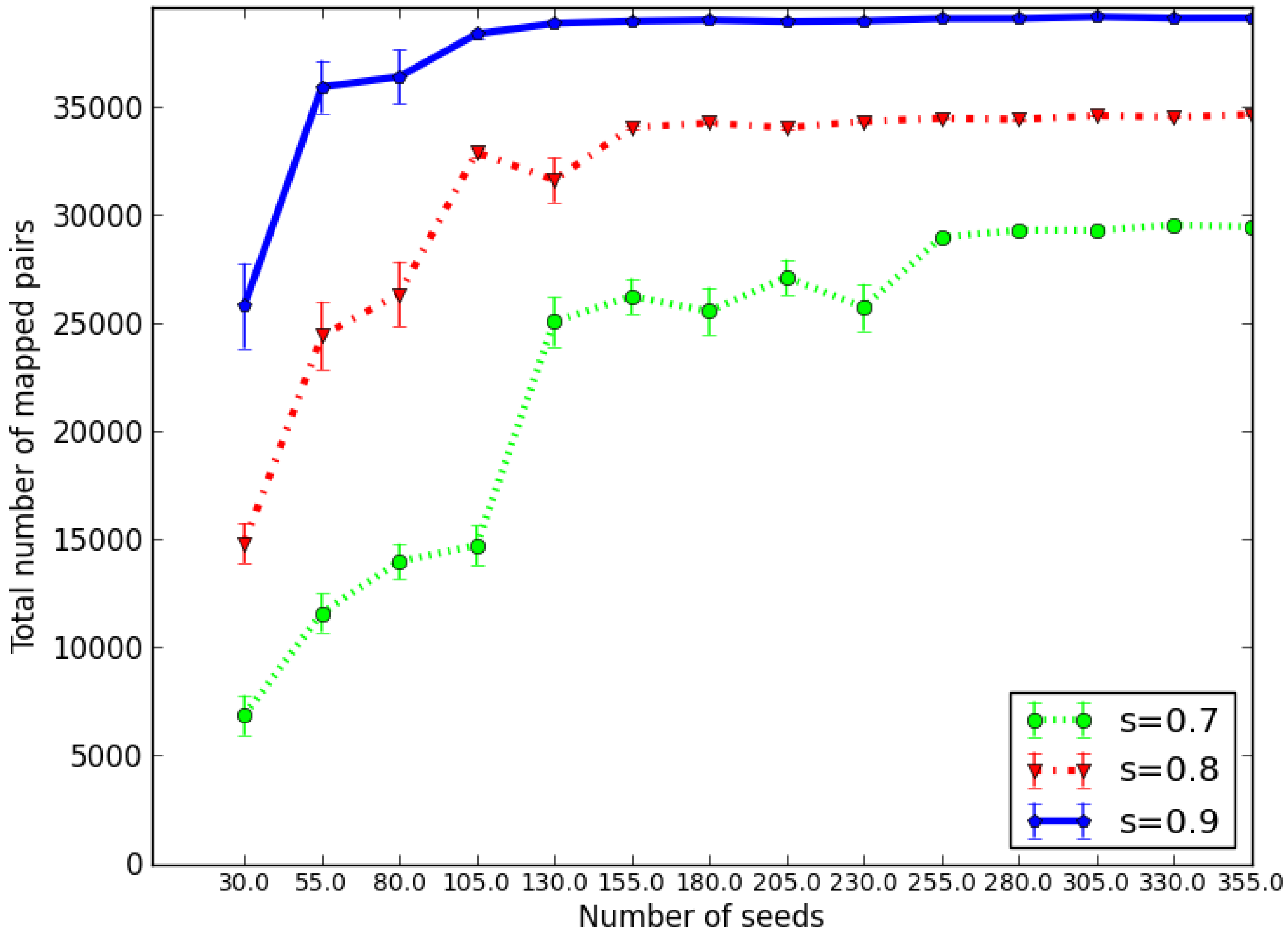
Simulation of PGM with $G(n, p; s)$ Network



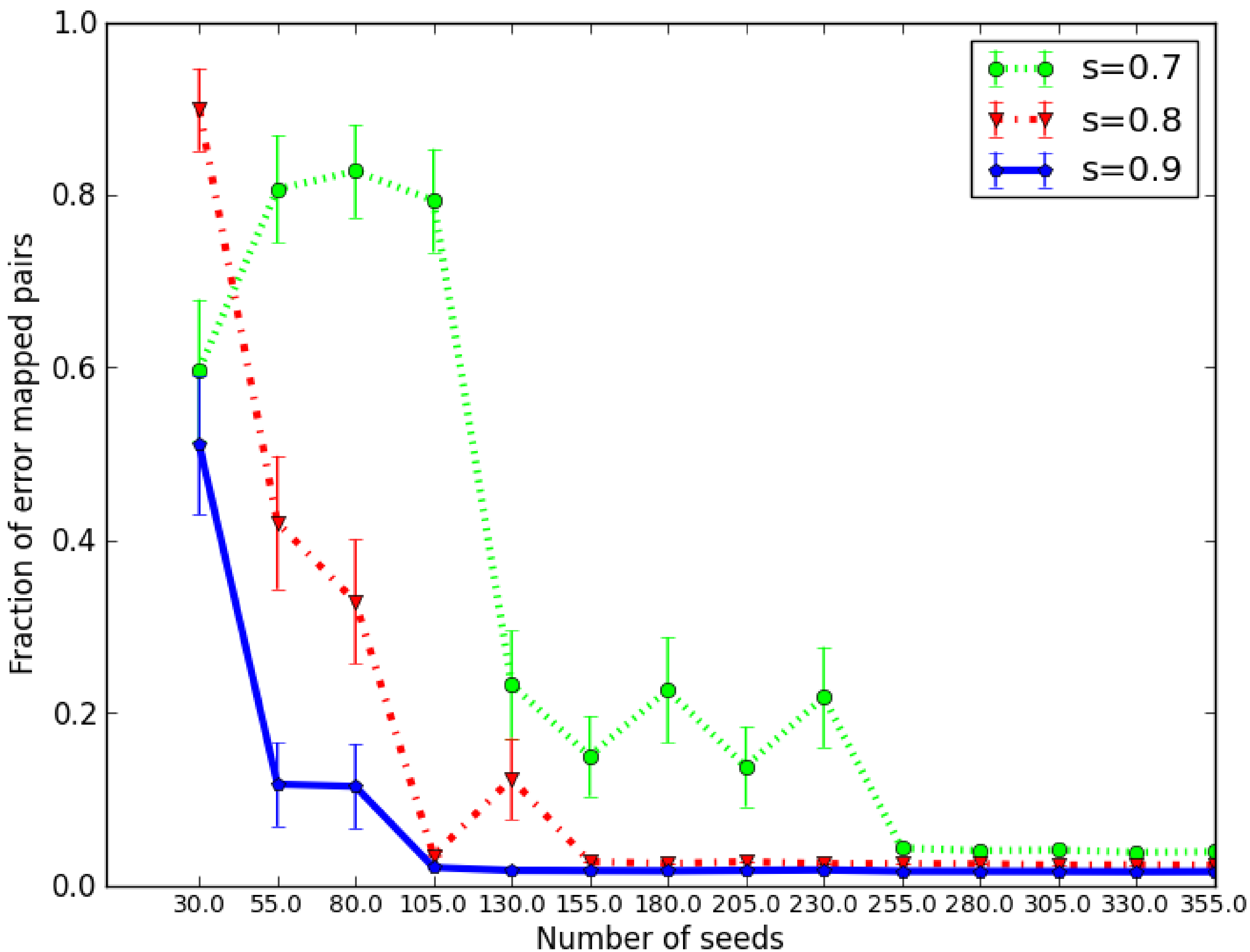
Simulation of PGM with $G(n, p; s)$ Network



Real Network: Slashdot Social Graph



Real Network: Slashdot Social Graph

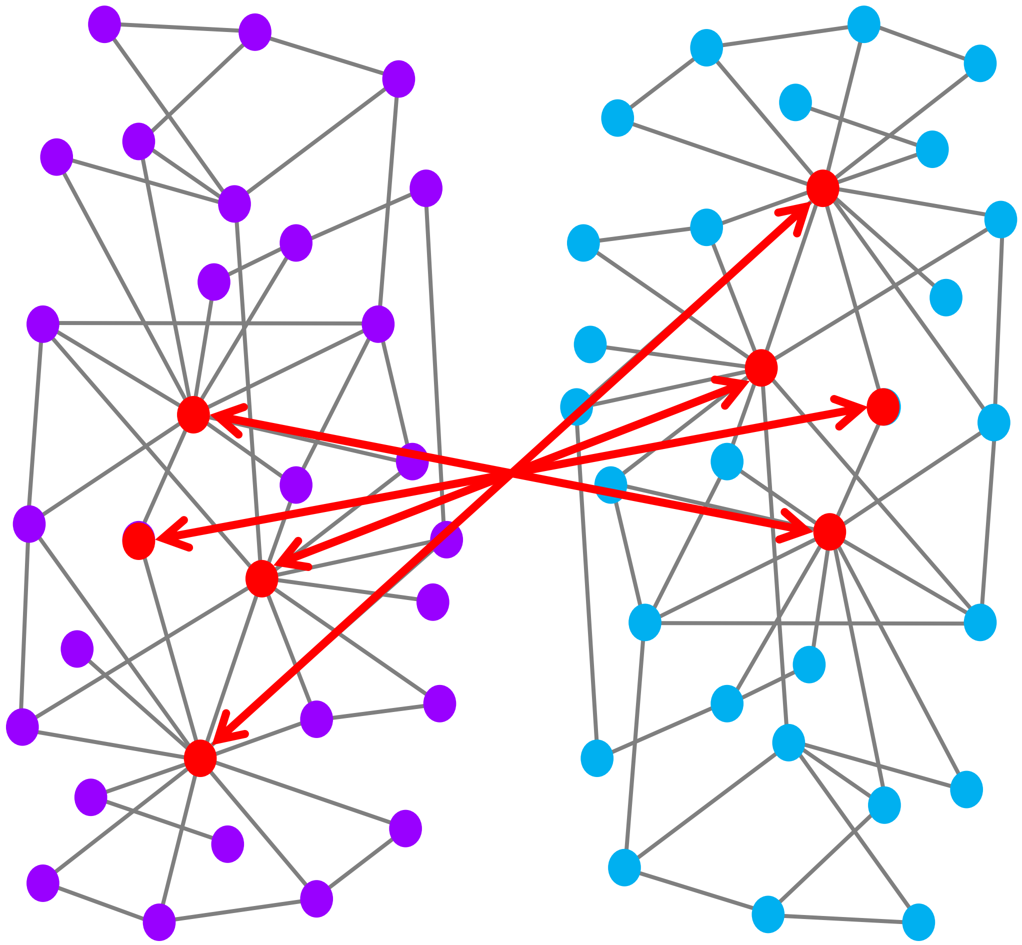


Real networks

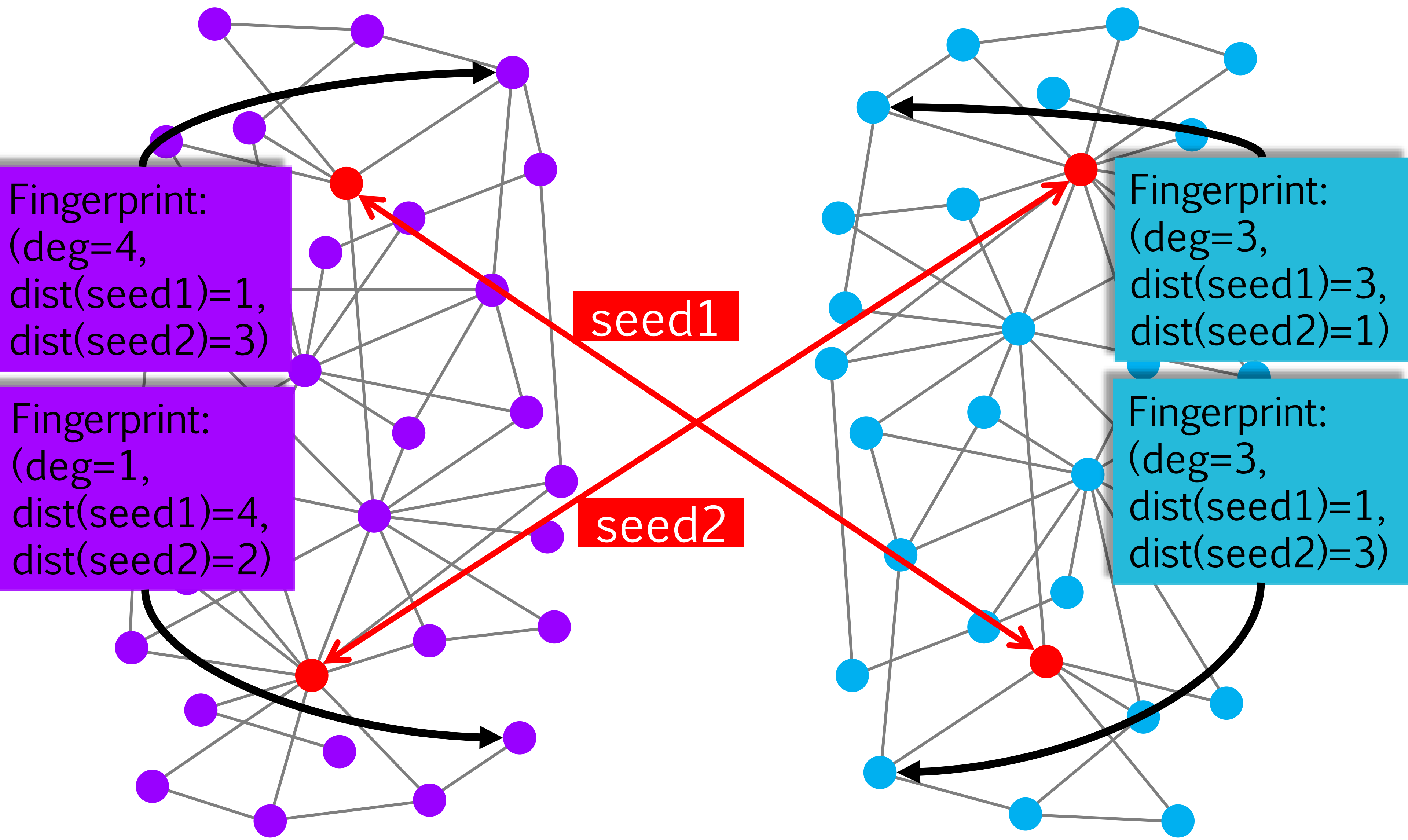
How to get started in practice

Matching Algorithm and Sampling Model

- **Question:**
  - Can similar idea inform algorithm design?
- **Wishlist:**
  - **Cold-start:** how to match without seeds?
  - **Sparse graphs:** how to avoid blocking?
  - **Error propagation:** how to correct mismatches?



Matching Algorithms: Bayesian Framework



Matching Algorithms: Bayesian Framework

Network sampling model:
P(fp1, fp2 | matched correctly),
P(fp1, fp2 | matched wrong)

Fingerprint:
(deg=4,
dist(seed1)=1,
dist(seed2)=3)

Fingerprint:
(deg=3,
dist(seed1)=3,
dist(seed2)=1)

Fingerprint:
(deg=1,
dist(seed1)=4,
dist(seed2)=2)

Single-pair posterior:
P(matched correctly | fp1, fp2)

Fingerprint:
(deg=3,
dist(seed1)=1,
dist(seed2)=3)

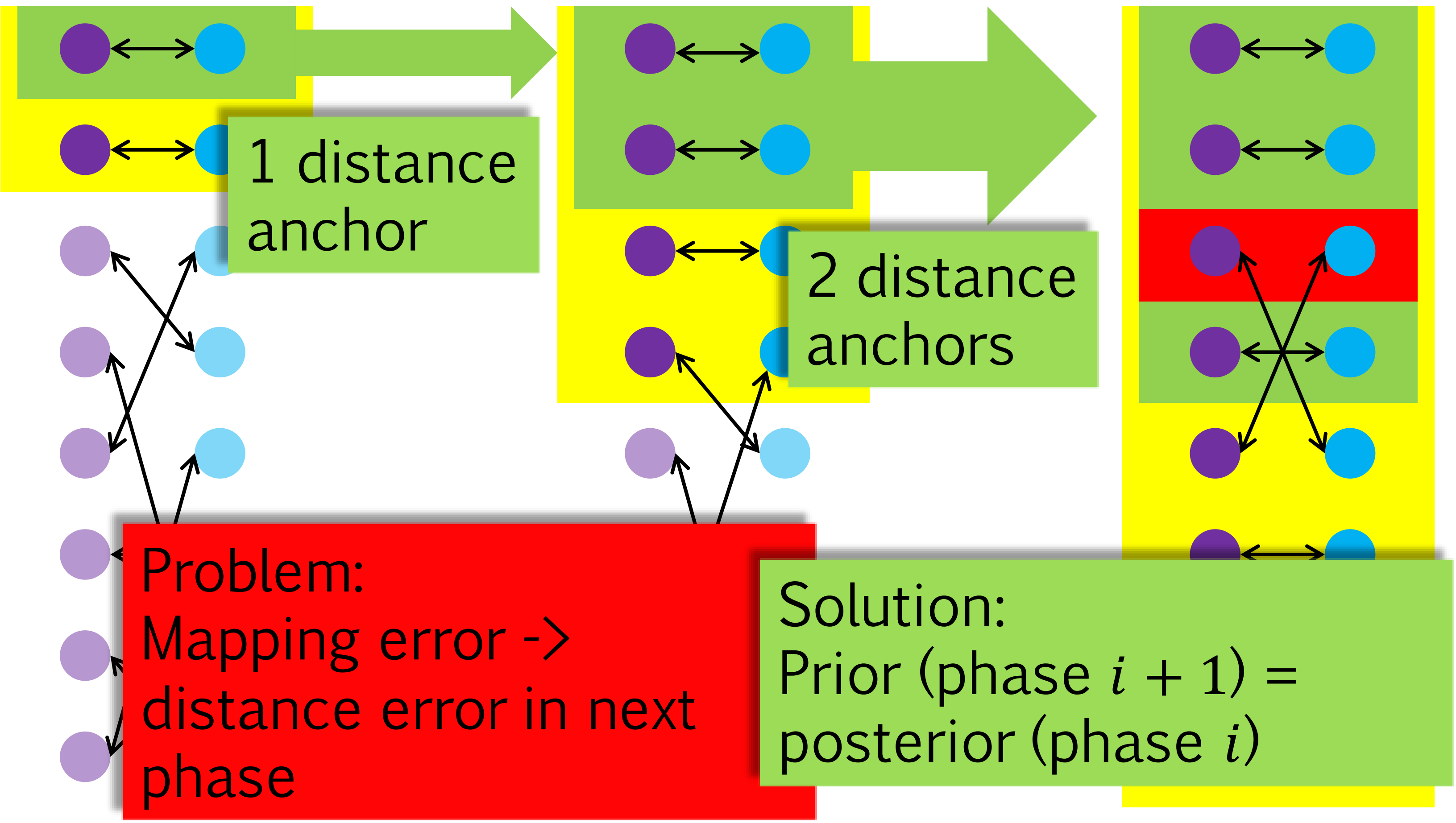
Jointly MAP matching:
Best bipartite matching π s.t. max
P(all matched correctly | all fingerprints)

Iterative Seedless Bayesian Matching

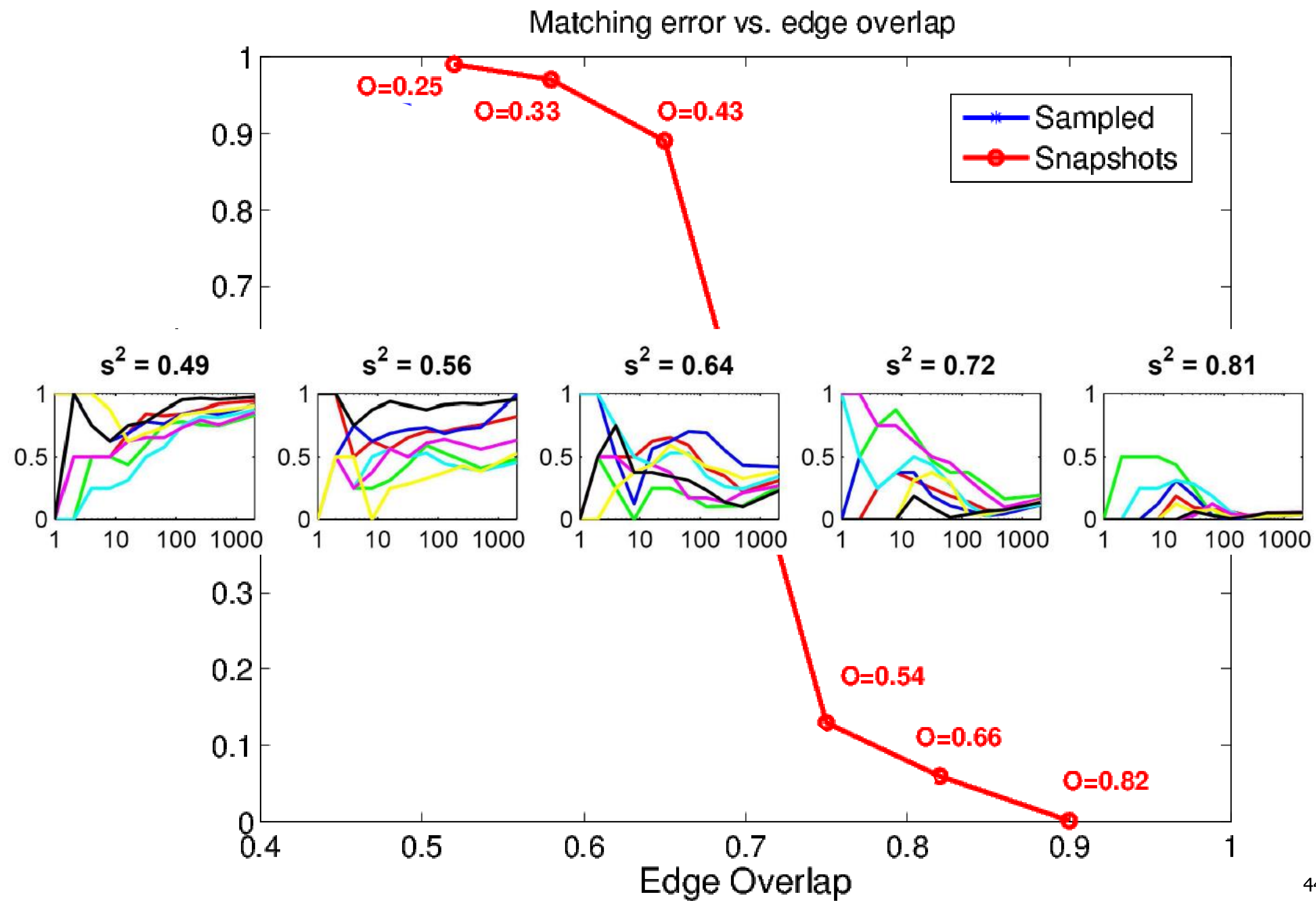
Phase 1:
2 candidates

Phase 2:
4 candidates

Phase 3:
8 candidates



Bayesian Seedless Matching: Performance



Conclusion

- **Graph Matching:**
  - Model as noisy graph isomorphism problem
  - How much information in network structure?
- **Information-theoretic:**
  - Matching is quite easy, benign growth of mean degree
  - $G(n, p; s)$ model: no a-priori structure
- **Percolation Graph Matching from seeds**
  - Phase transition in size of seed set → hard to control, tune, predict
  - Actually works very well in practice; parsimonious (r)
- **Finding seeds**
  - Bayesian framework & heuristics
  - Key idea: exploit known "couples" as references for new candidate pairs

Thank you!

CTW 2013

Collaborators:
Daniel R. Figueiredo,
Pedram Pedarsani,
Lyudmila Yartseva

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE