Community Recovery in Graphs with Locality

Yuxin Chen[†], Govinda Kamath[†], Changho Suh[*], David Tse[†]

Stanford[†] KAIST[*]

Community recovery / graph clustering

Community structures are common in many social networks

Credit: The Future Buzz

Credit: S. Papadopoulos

Community recovery / graph clustering

Community structures are common in many social networks

Credit: The Future Buzz

Credit: S. Papadopoulos

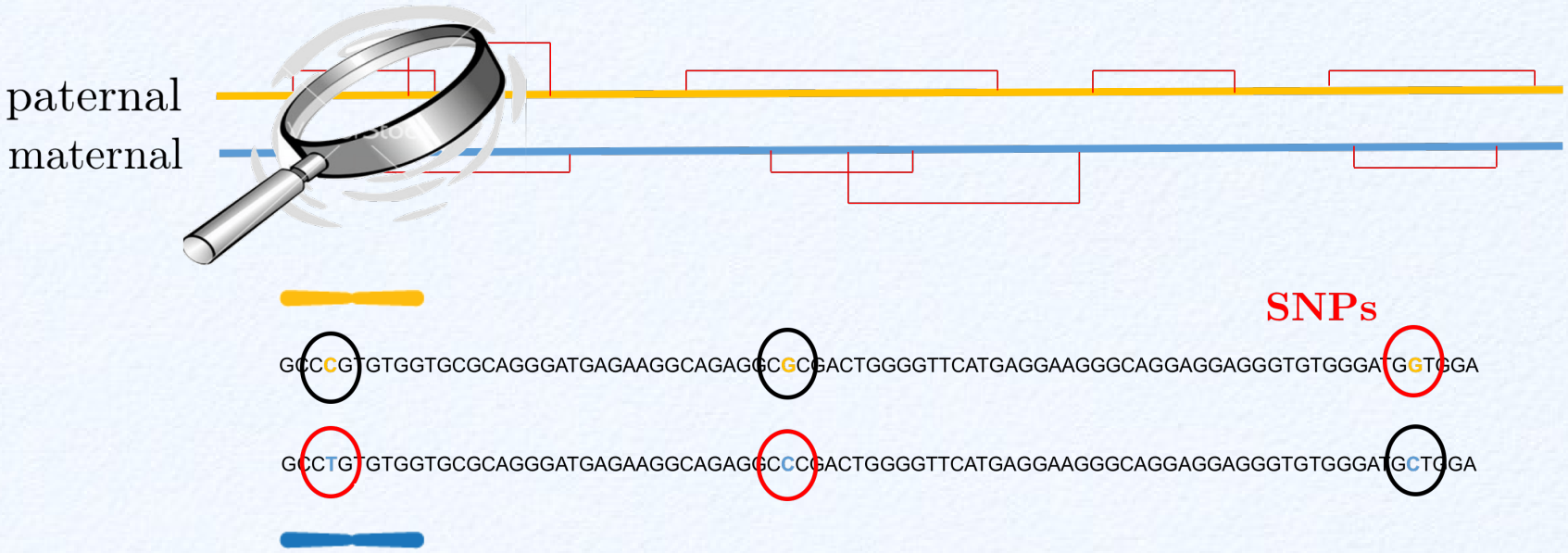
Community recovery: partition users into several clusters based on their friendships / similarities

Community recovery in computational biology

A genome phasing problem

Community recovery in computational biology

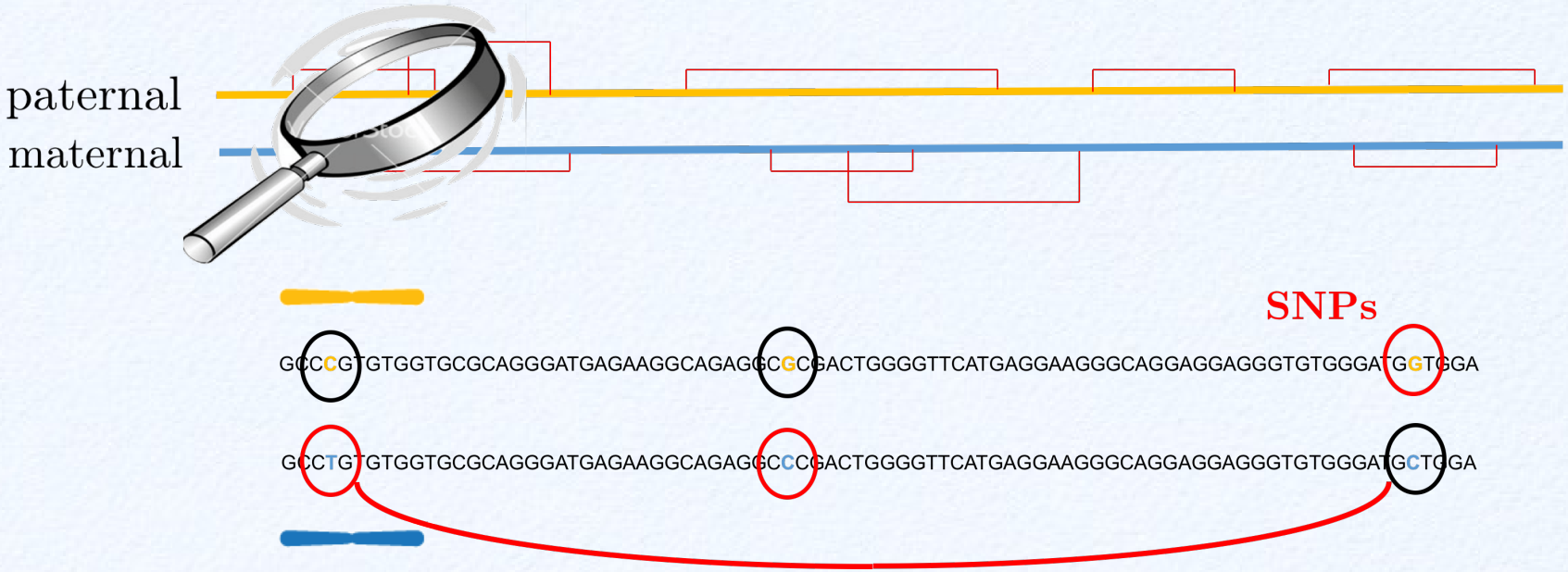
A genome phasing problem



phase info for each SNP: (1) maternally inherited (2) paternally inherited

Community recovery in computational biology

A genome phasing problem

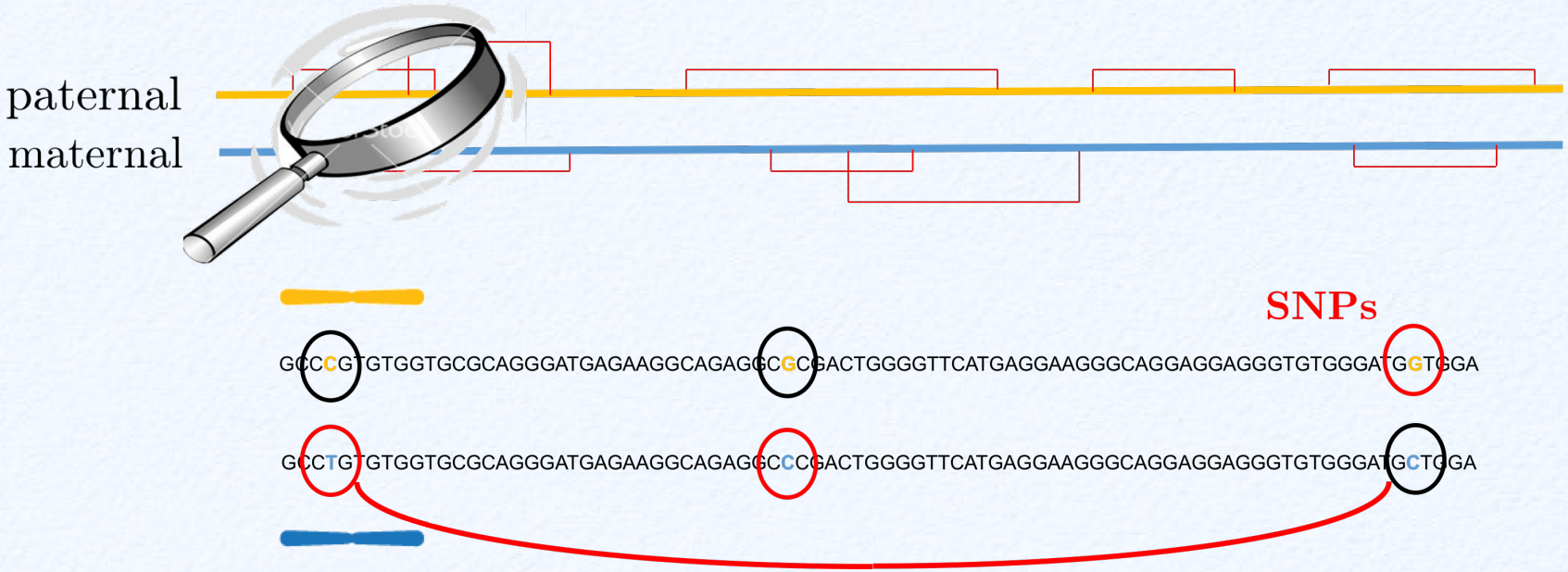


phase info for each SNP: (1) maternally inherited (2) paternally inherited

linking reads: relative phase relation of 2 (or more) SNPs

Community recovery in computational biology

A genome phasing problem



phase info for each SNP: (1) maternally inherited (2) paternally inherited

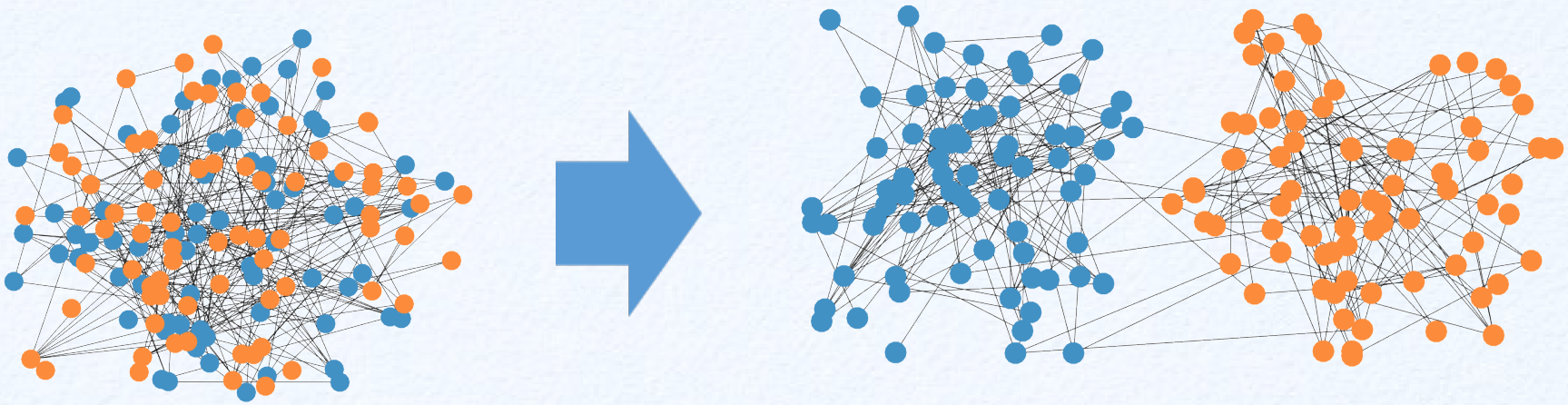
linking reads: relative phase relation of 2 (or more) SNPs

Haplotype phasing: retrieve phase info of all SNPs from linking reads

Stochastic block model / censored block model

Pairwise measurements for *any pair* (i, j) of nodes

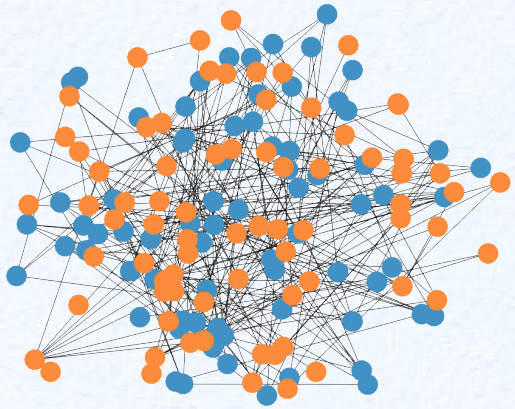
$$y_{i,j} \overset{\text{ind.}}{\sim} \begin{cases} P_0, & \text{if } i \text{ and } j \text{ are from same community} \\ P_1, & \text{else} \end{cases}$$



Problem: nodes often have locality

Most prior work: (almost) equally likely to sample between any pair of nodes

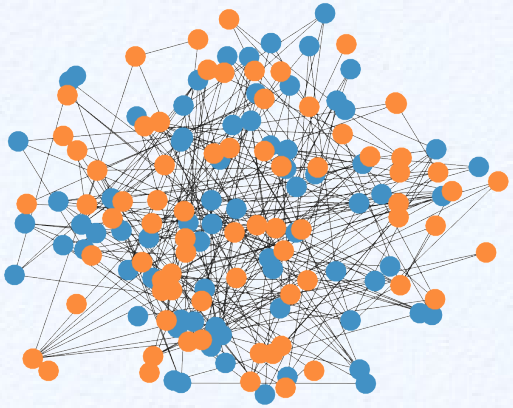
- *Condon et al., Jalali et al., Chen et al., Abbe et al., Mossel et al., Hajek et al., Chin et al...*



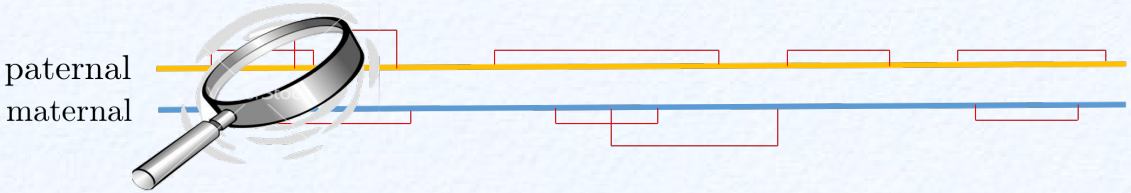
Problem: nodes often have locality

Most prior work: (almost) equally likely to sample between any pair of nodes

- *Condon et al., Jalali et al., Chen et al., Abbe et al., Mossel et al., Hajek et al., Chin et al...*



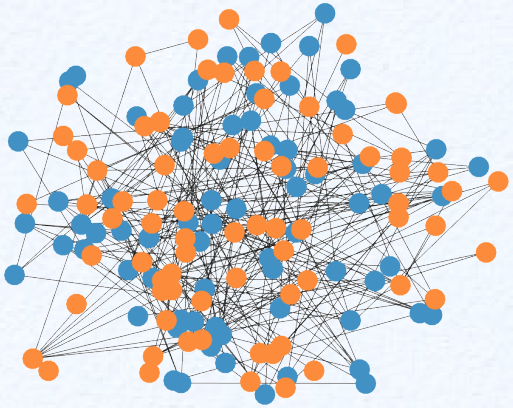
More realistically: samples come mainly (or exclusively) from nearby nodes



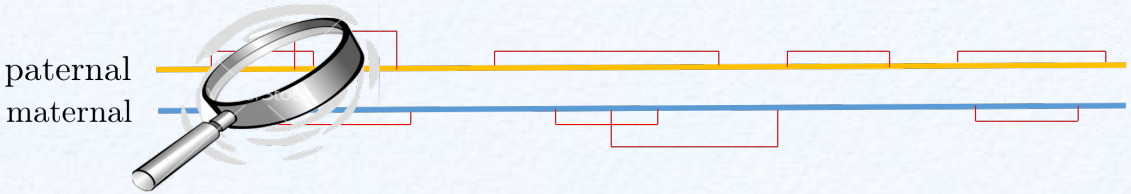
Problem: nodes often have locality

Most prior work: (almost) equally likely to sample between any pair of nodes

- *Condon et al., Jalali et al., Chen et al., Abbe et al., Mossel et al., Hajek et al., Chin et al...*



More realistically: samples come mainly (or exclusively) from nearby nodes

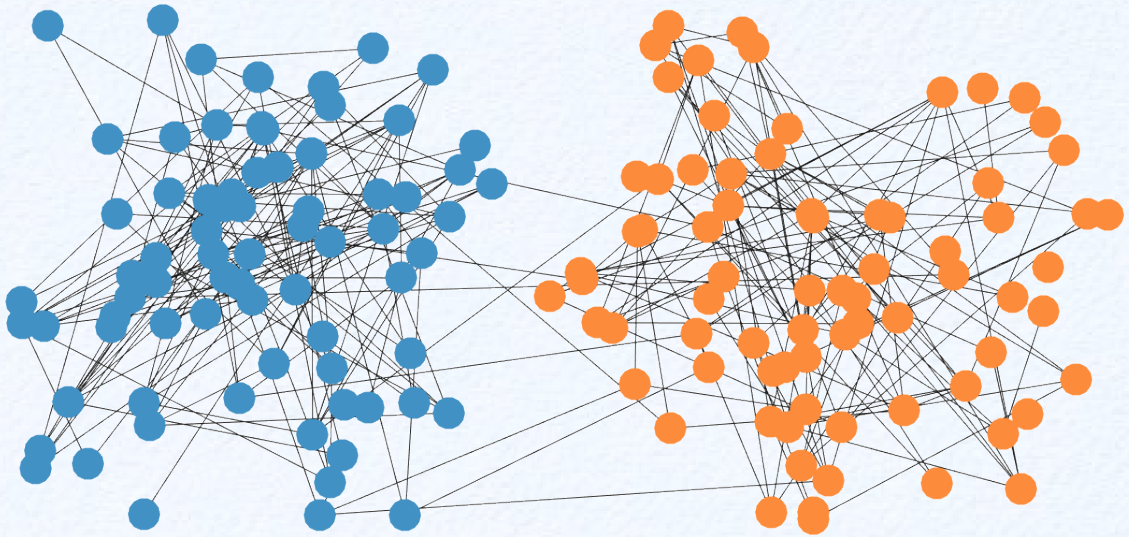


In new technologies like 10x-Genomics: (1) $n \sim 10^5$ SNPs; (2) linking range ~ 100 SNPs

This work: how to deal with measurement locality in community recovery?

A two-community model

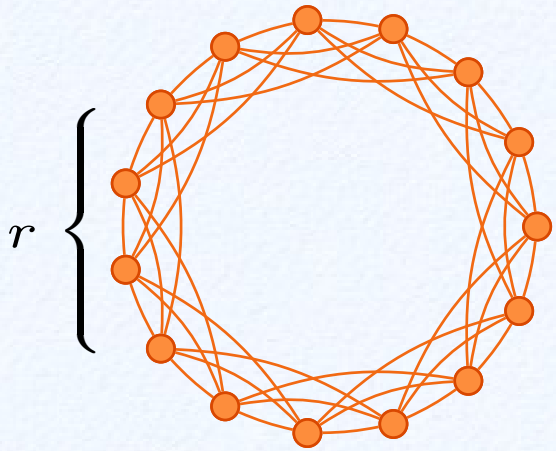
- n variables we seek: $x_1, \cdots, x_n \in \{0, 1\}$
  - encode community membership



$x_i = 0$ $\qquad$ $x_i = 1$

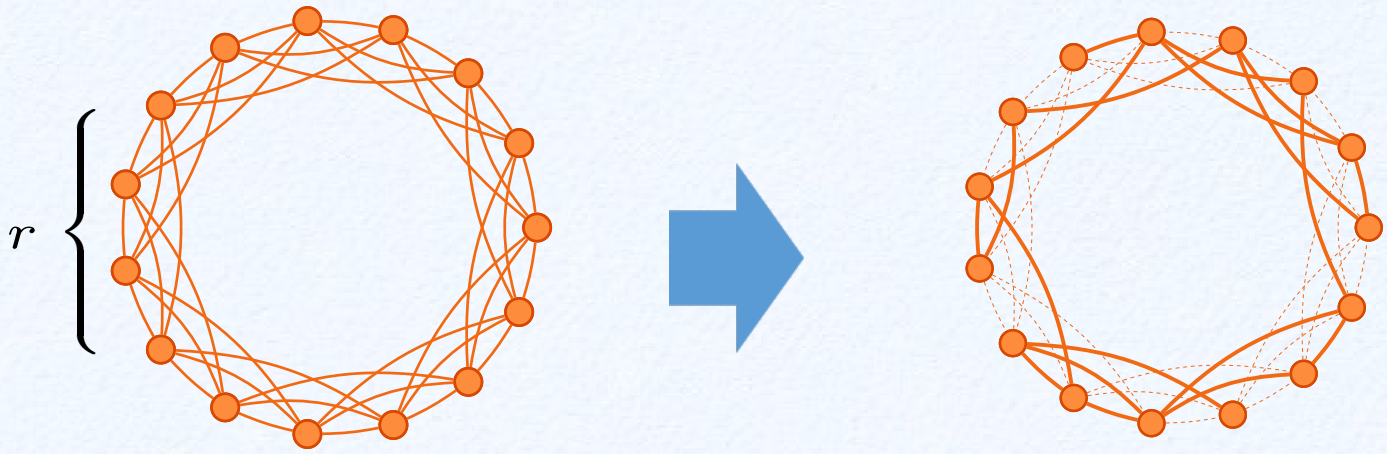
Measurement model: random sampling

- **Constraint graph $\mathcal{G}$**



Measurement model: random sampling

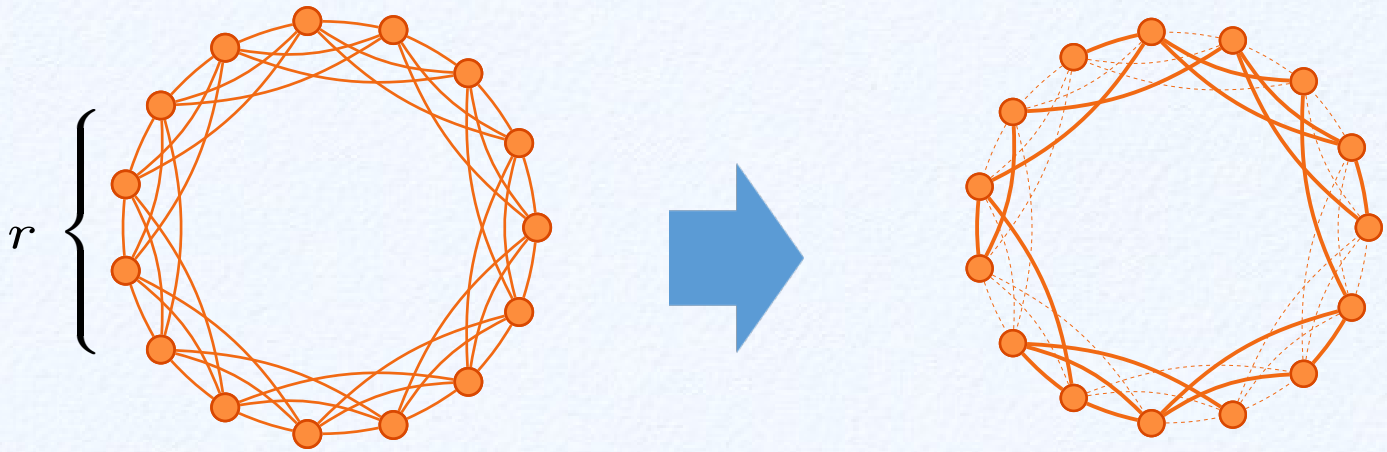
- **Constraint graph** $\mathcal{G}$



- **Random sampling:** pick m randomly chosen edges of $\mathcal{G}$

Measurement model: random sampling

- **Constraint graph** $\mathcal{G}$

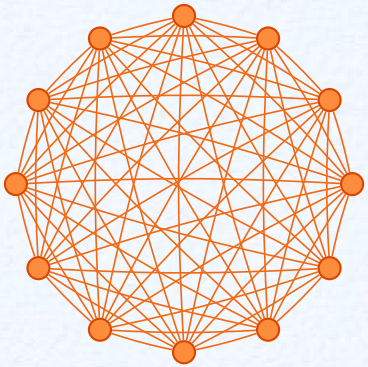


- **Random sampling:** pick m randomly chosen edges of $\mathcal{G}$
- **Noise model:** on each of these m edges (i, j), take an independent sample

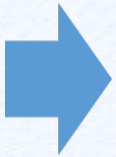
$$y_{i,j} \overset{\text{ind.}}{=} \begin{cases} x_i \oplus x_j, & \text{with prob. } 1 - \underbrace{\theta}_{\text{meas. error rate}} \\ x_i \oplus x_j \oplus 1, & \text{else} \end{cases}$$

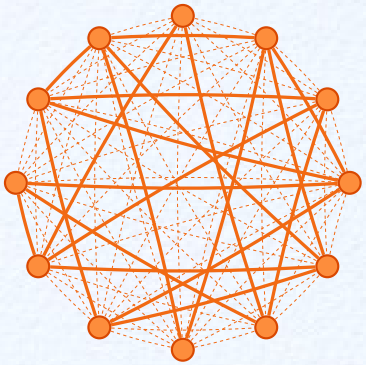
Modeling locality via constraint graph

Global / long-range measurements



constraint graph

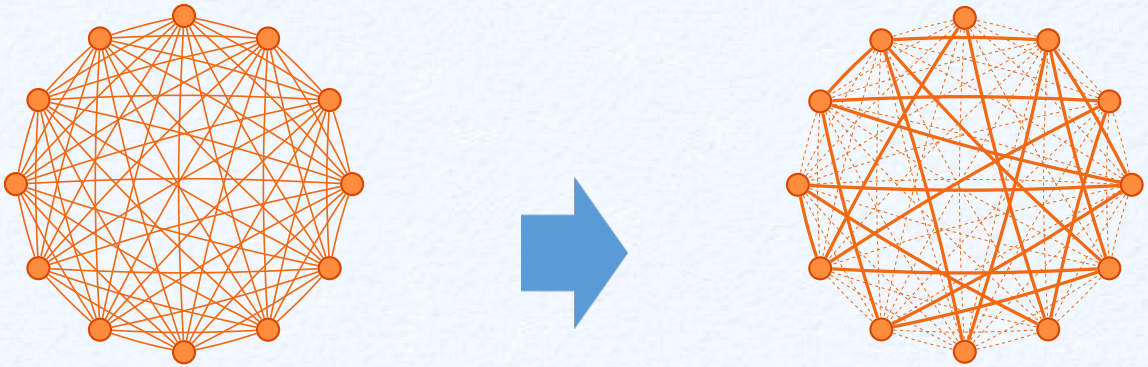




randomly picked edges

Modeling locality via constraint graph

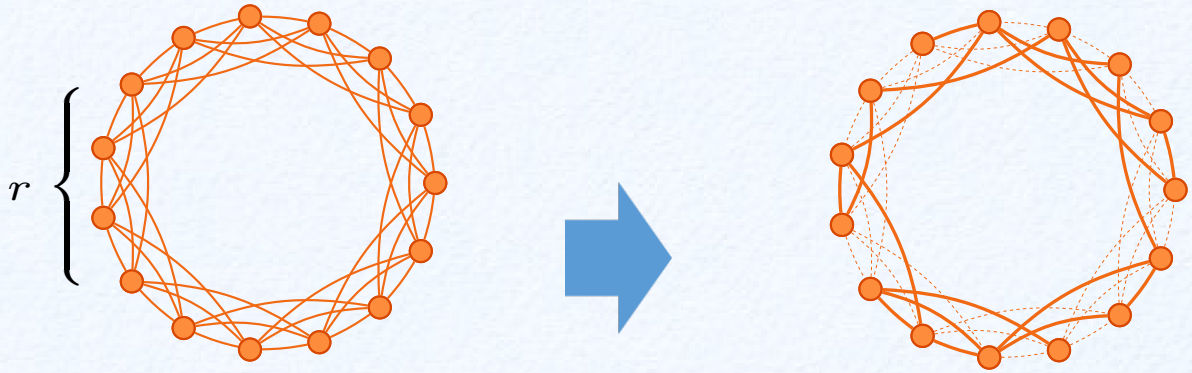
Global / long-range measurements



constraint graph

randomly picked edges

Local measurements



constraint graph
(e.g. $r \sim n^{0.4}$ for 10x)

randomly picked edges

Information and computation limits

1. How many samples are needed to recover $\{x_i\}$ reliably (up to global offset)?

Information and computation limits

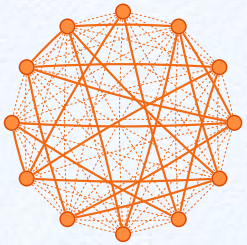
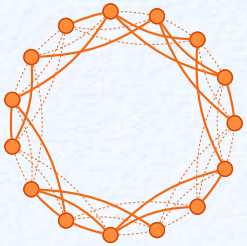
1. How many samples are needed to recover $\{x_i\}$ reliably (up to global offset)?

2. How to recover efficiently?

Information and computation limits

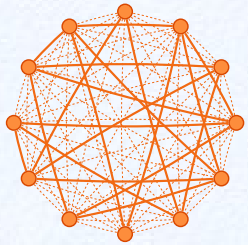
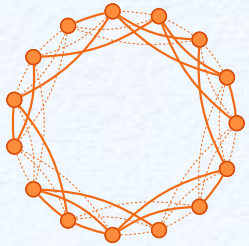
1. How many samples are needed to recover $\{x_i\}$ reliably (up to global offset)?

2. How to recover efficiently?

	Global samples	Local samples
		
prior works		

Information and computation limits

1. How many samples are needed to recover $\{x_i\}$ reliably (up to global offset)?
2. How to recover efficiently?

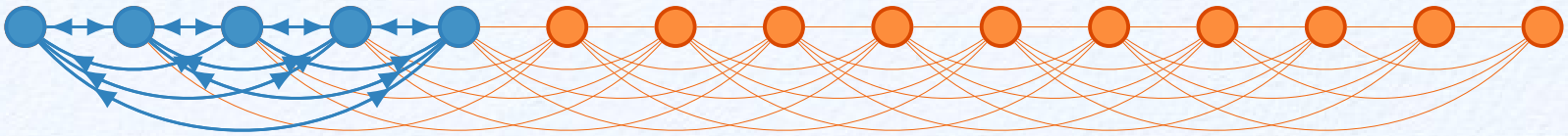
	Global samples	Local samples
		
prior works		

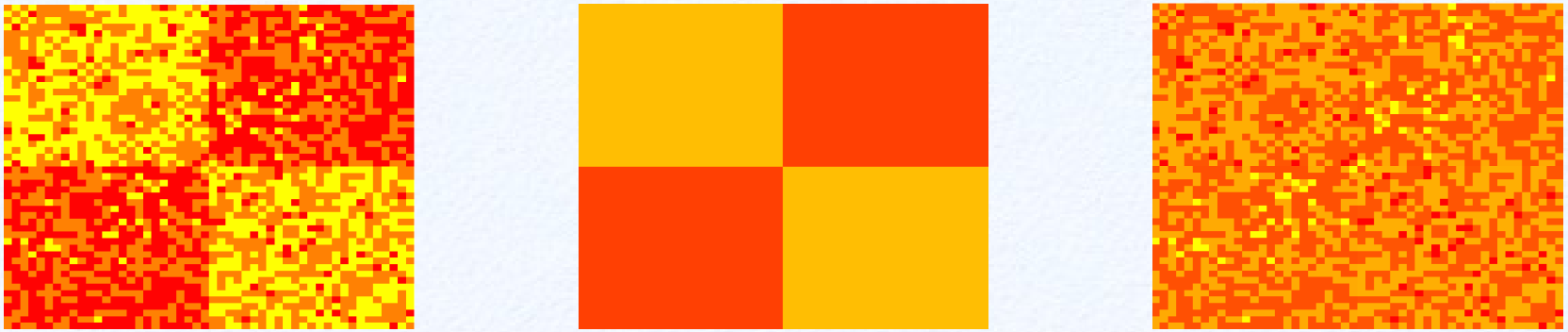
Encouraging news: one can obtain efficient recovery within linear time

Proposed algorithm: a 3-stage linear-time paradigm

Spectral-Stitching: Stage 1

Start by running spectral method on core complete subgraphs

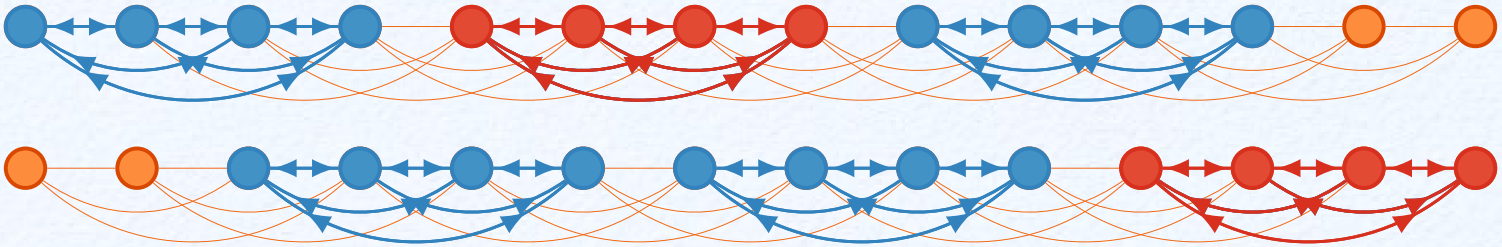




- Compute rank-1 approximation of $\boldsymbol{L}$ (*sample matrix restricted to the subgraph*)

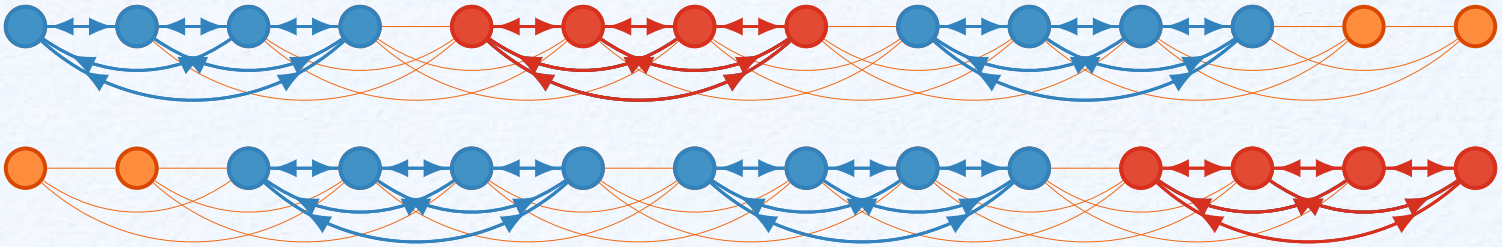
Spectral-Stitching: Stage 1

Split all nodes into overlapping subsets and run spectral methods separately



Spectral-Stitching: Stage 1

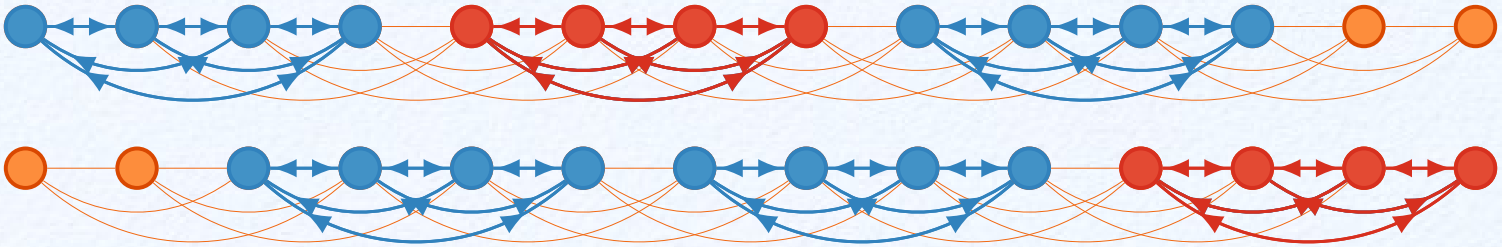
Split all nodes into overlapping subsets and run spectral methods separately



- Approximate solution within each subgraph
  - Key observation: approx. recovery needs only $O(1)$ samples per node

Spectral-Stitching: Stage 1

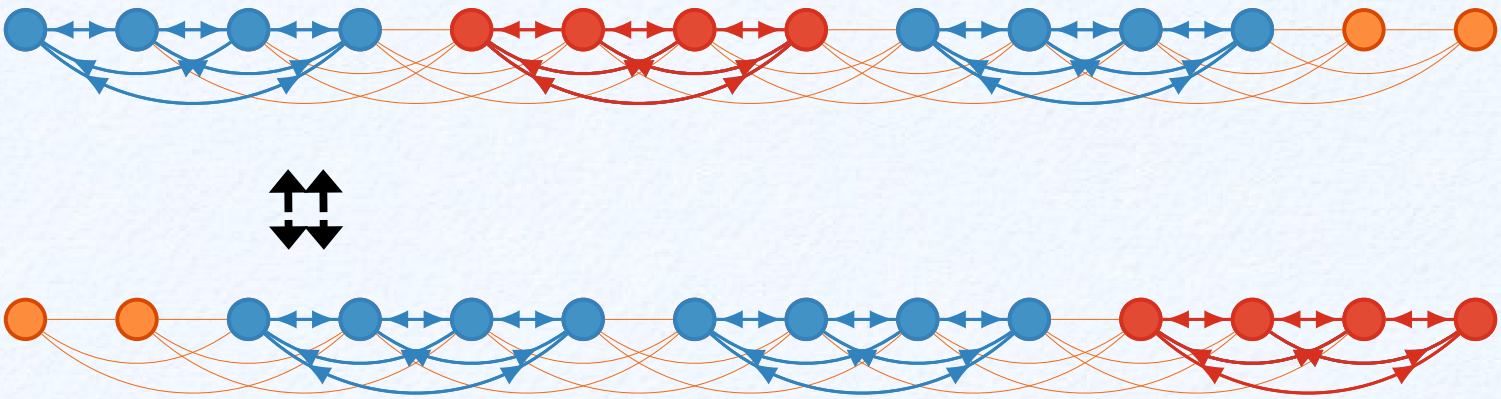
Split all nodes into overlapping subsets and run spectral methods separately



- Approximate solution within each subgraph
  - Key observation: approx. recovery needs only $O(1)$ samples per node
- Inconsistent global phases across subgraphs

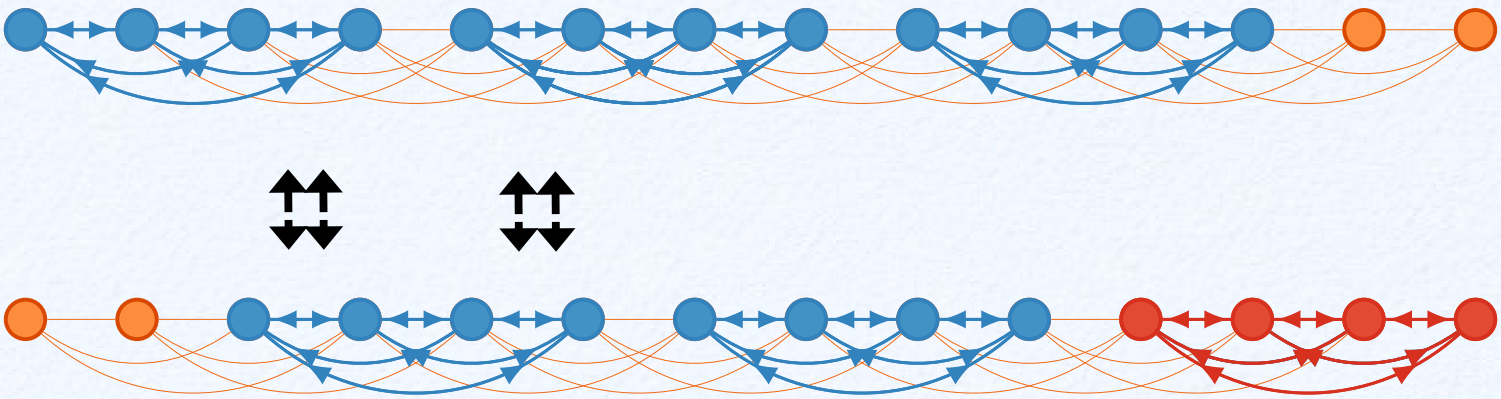
Spectral-Stitching: Stage 2

Calibrate phases across subgraphs by checking their correlations



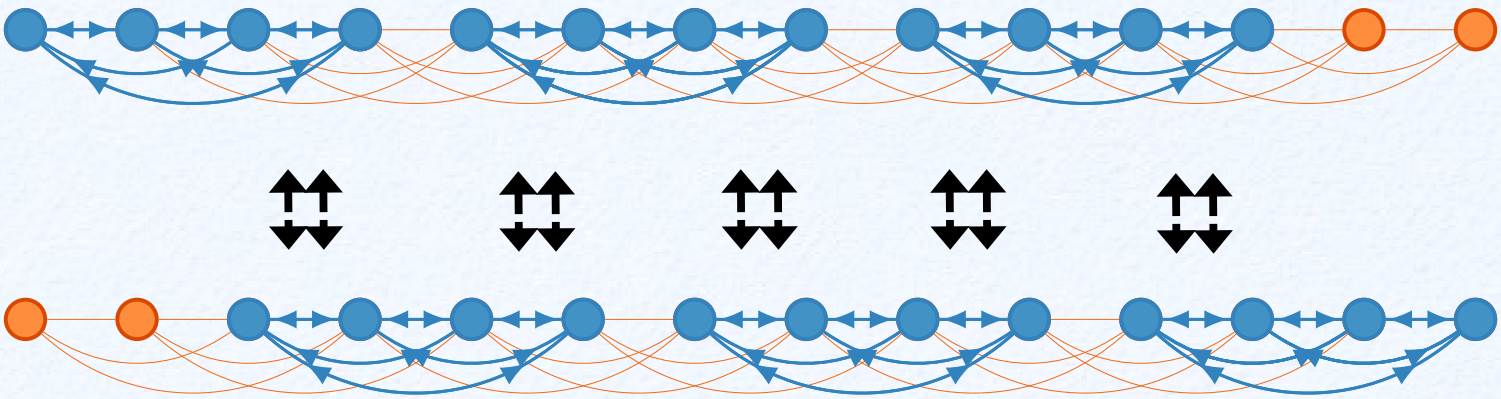
Spectral-Stitching: Stage 2

Calibrate phases across subgraphs by checking their correlations



Spectral-Stitching: Stage 2

Calibrate phases across subgraphs by checking their correlations

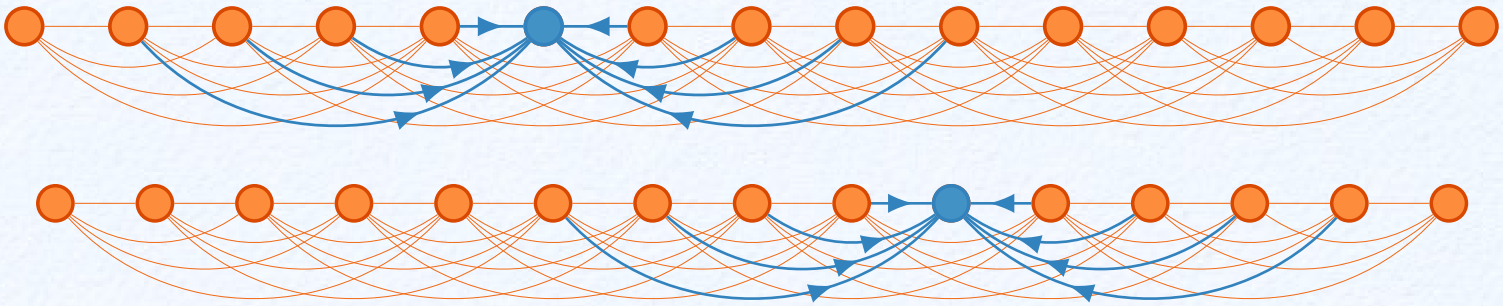


Purpose of Stages 1-2: obtain approximate solution of all nodes

Spectral-Stitching: Stage 3

Clean up all remaining errors by iterative refinement

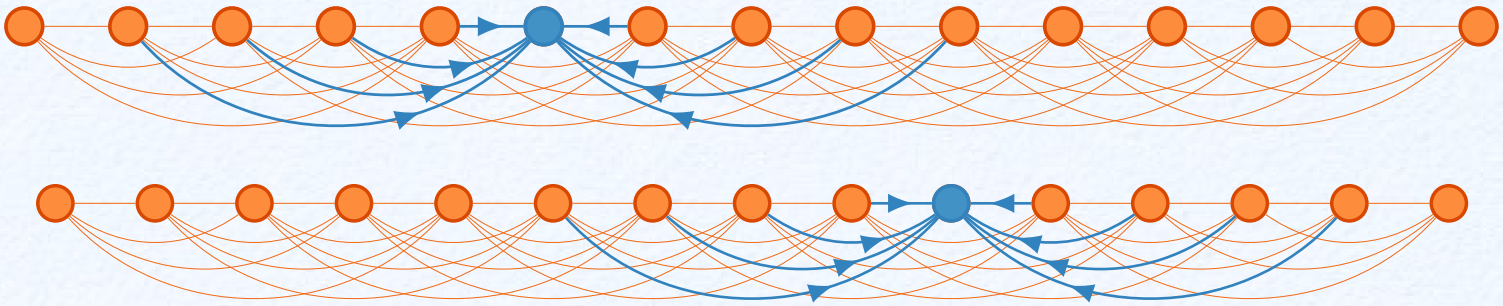
- local majority vote using *all* samples



Spectral-Stitching: Stage 3

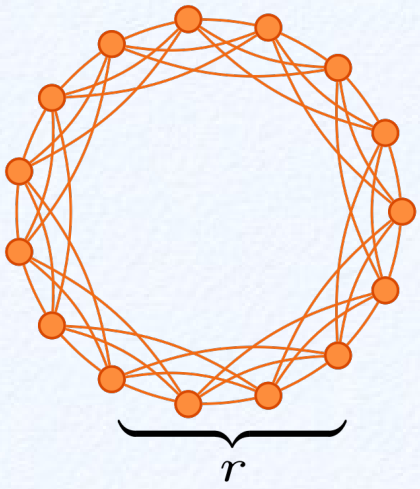
Clean up all remaining errors by iterative refinement

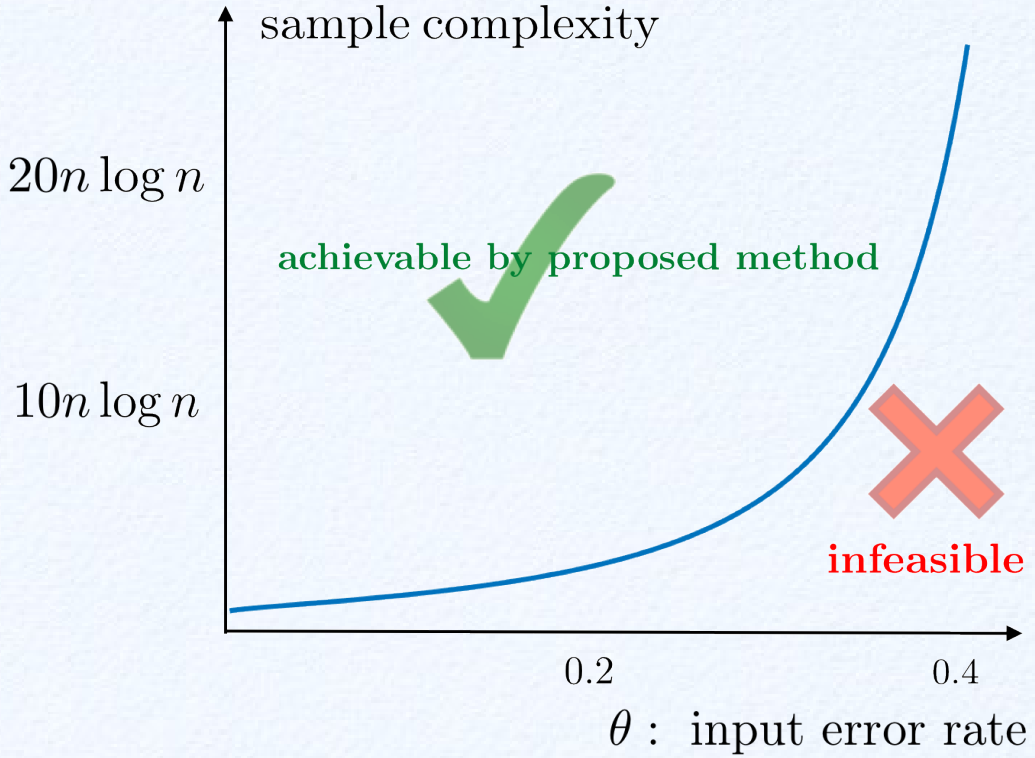
- local majority vote using *all* samples



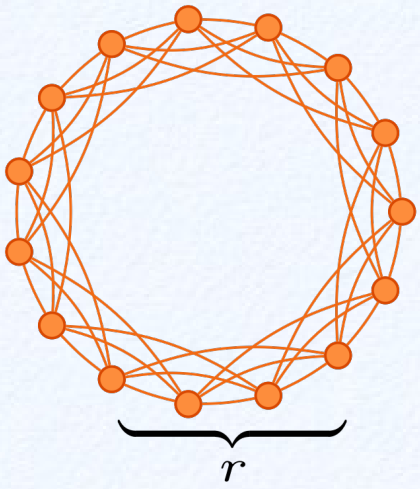
- Key observation: exact recovery needs at least $\Theta(\log n)$ samples per node

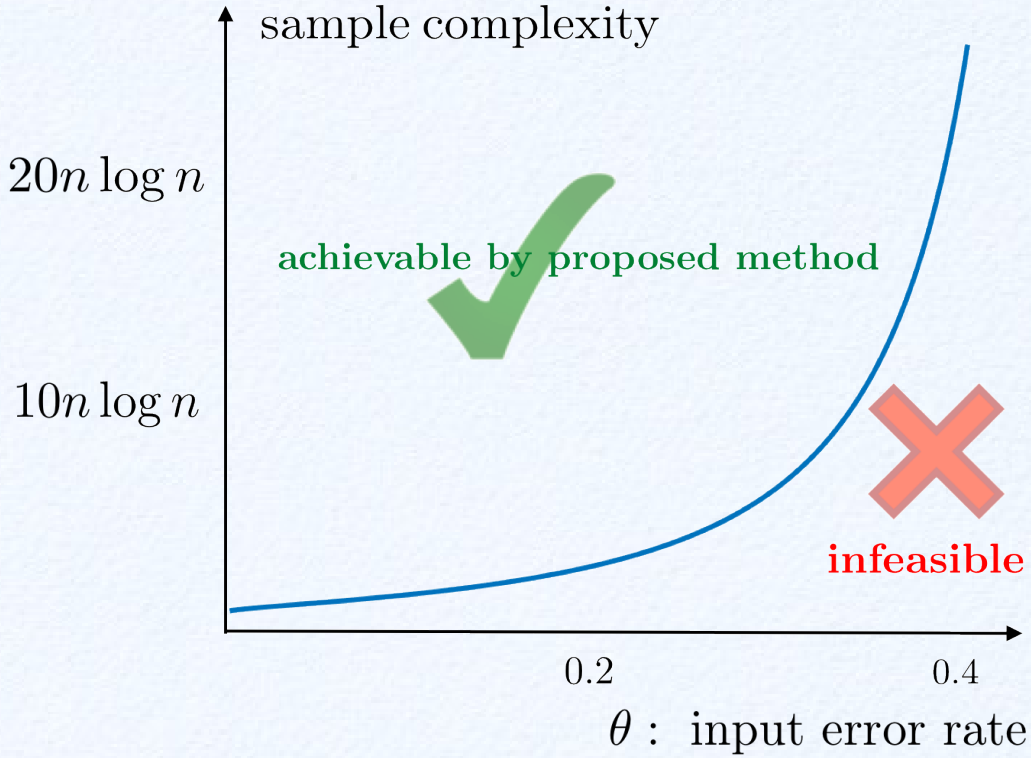
Main results: rings





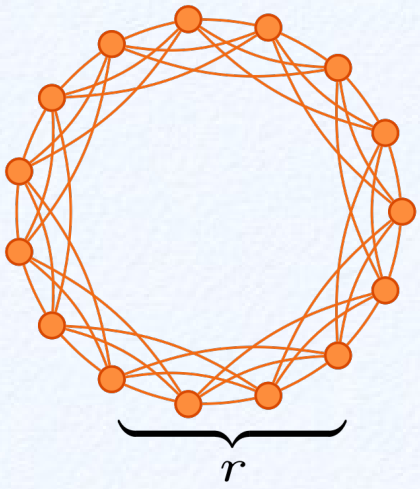
Main results: rings

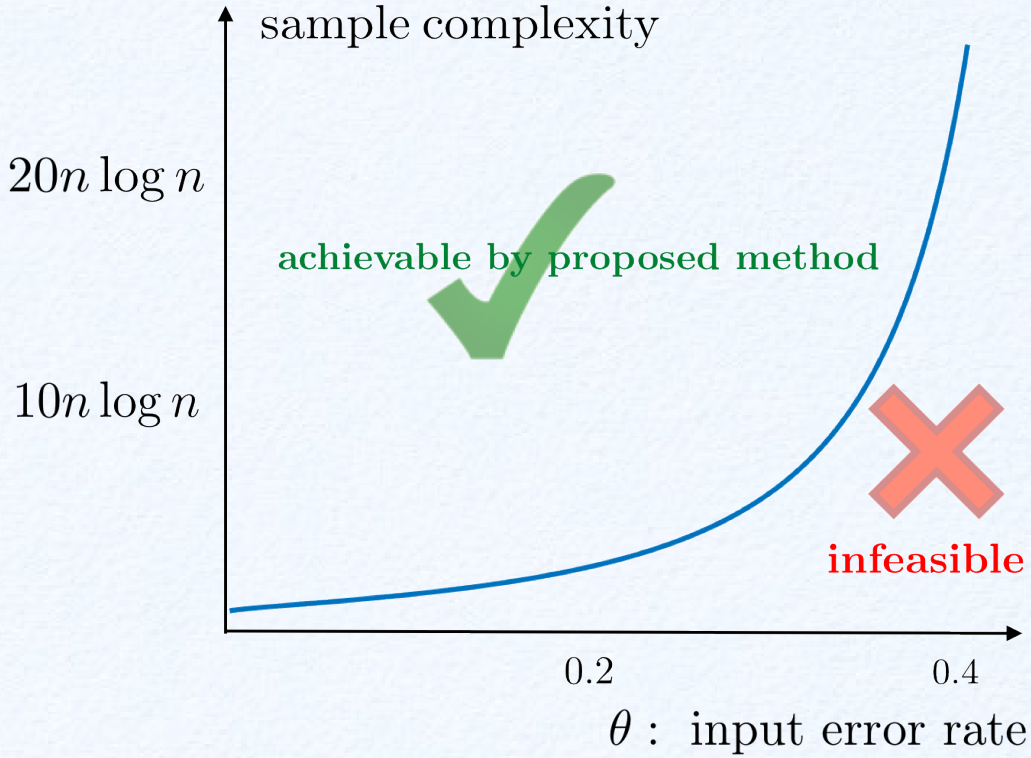




Theorem: minimum sample complexity $= \frac{0.5n \log n}{1-\exp\{-\mathsf{KL}(0.5 \,\|\, \theta\})}$

Main results: rings

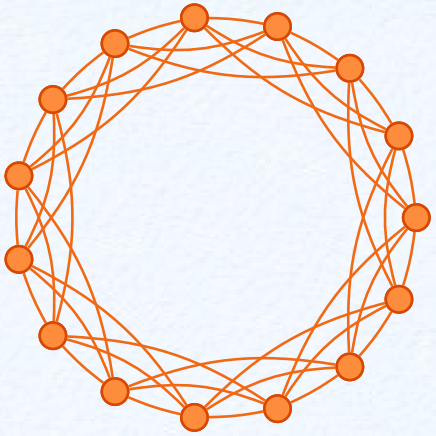




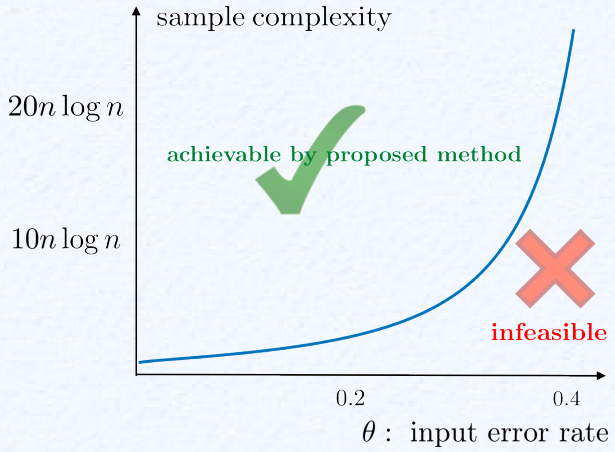
Theorem: minimum sample complexity $= \frac{0.5n \log n}{1-\exp\{-\mathsf{KL}(0.5 \,\|\, \theta\})}$

Info and comput. limits meet!

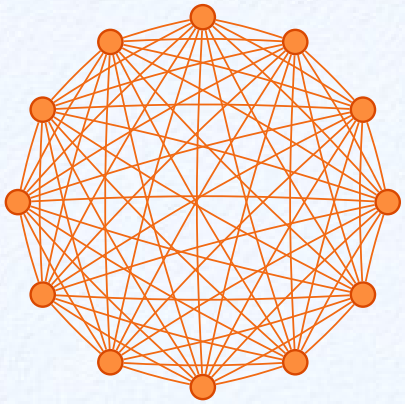
An insensitivity phenomenon



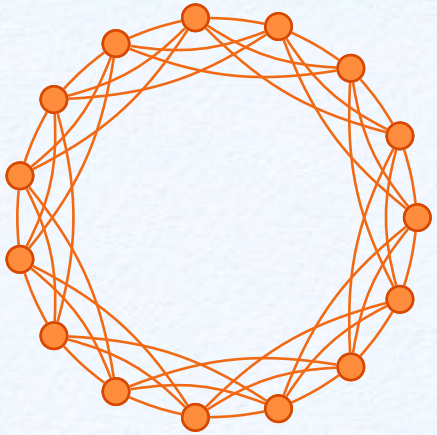
ring



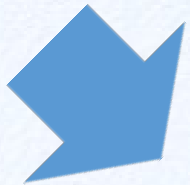
An insensitivity phenomenon

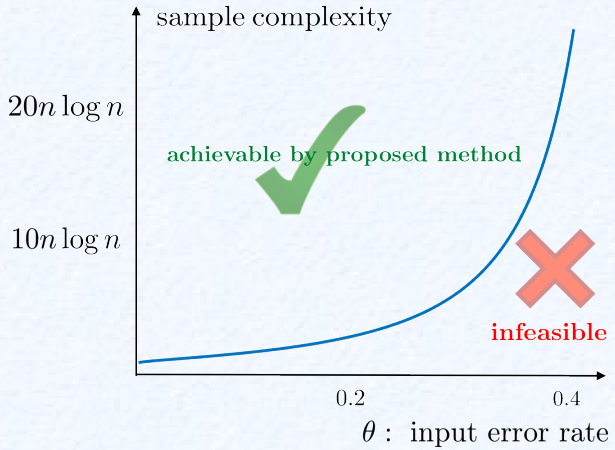


complete graph

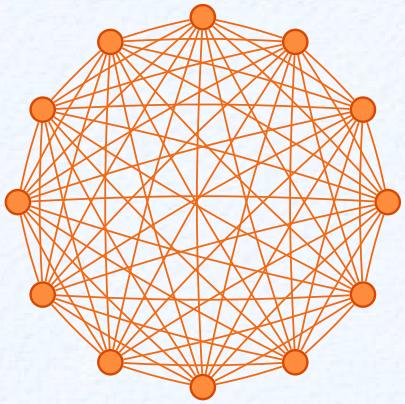


ring

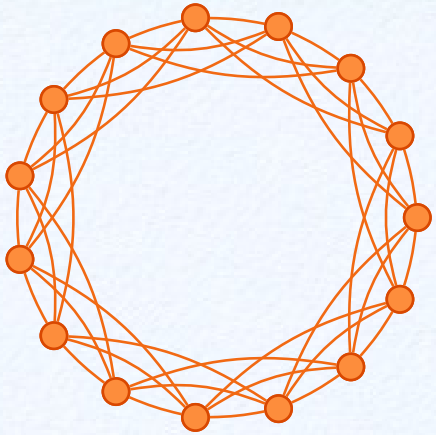




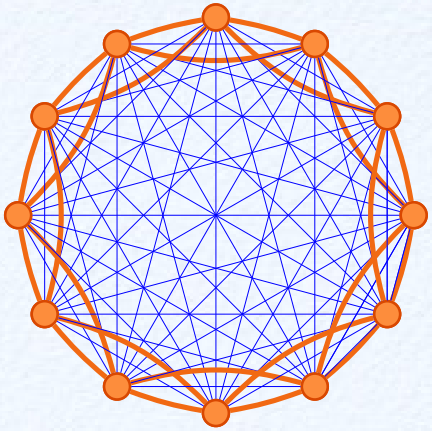
An insensitivity phenomenon



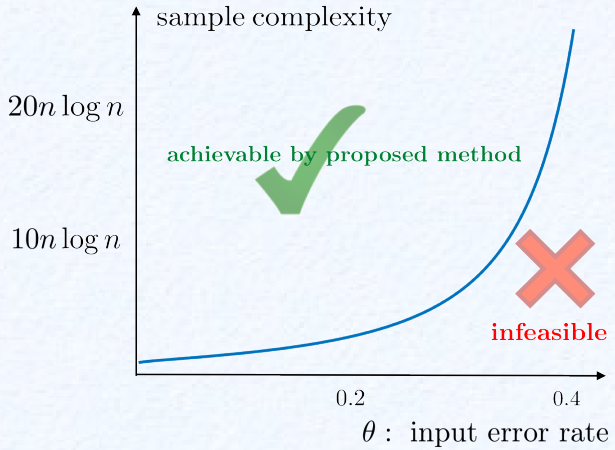
complete graph

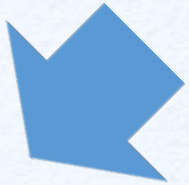


ring

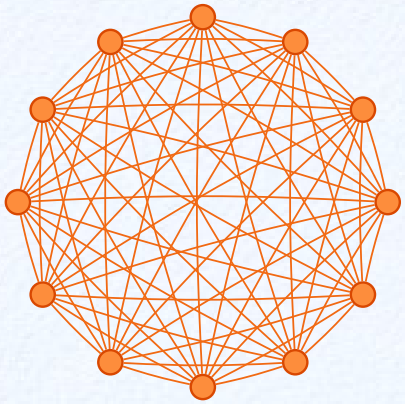


small-world

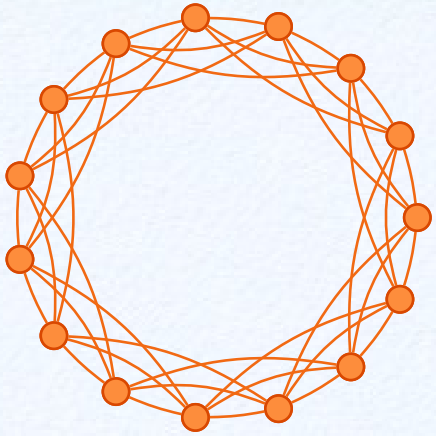




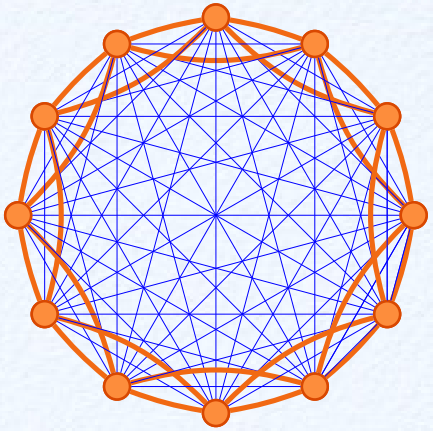
An insensitivity phenomenon



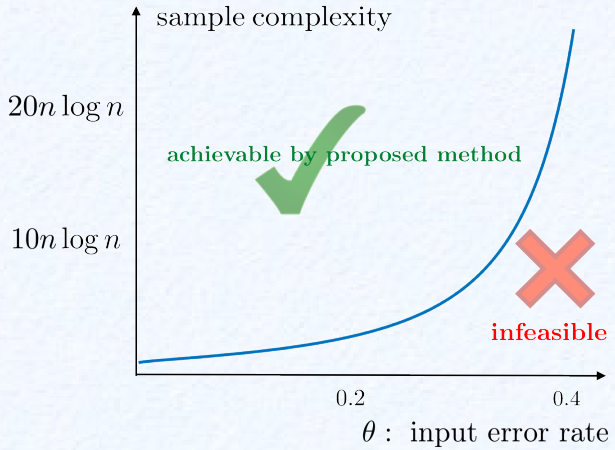
complete graph

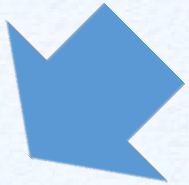


ring



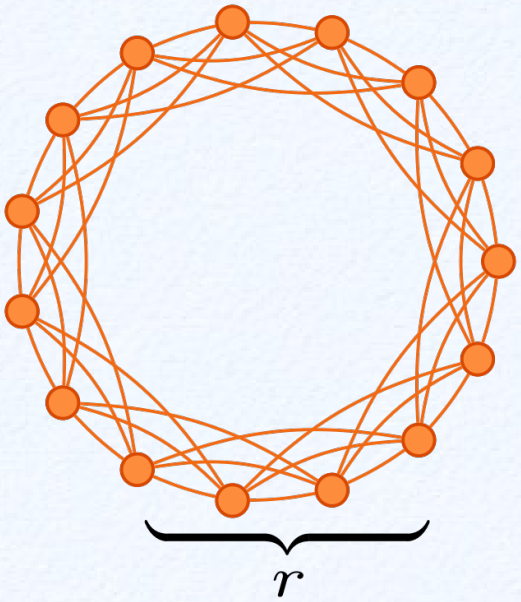
small-world

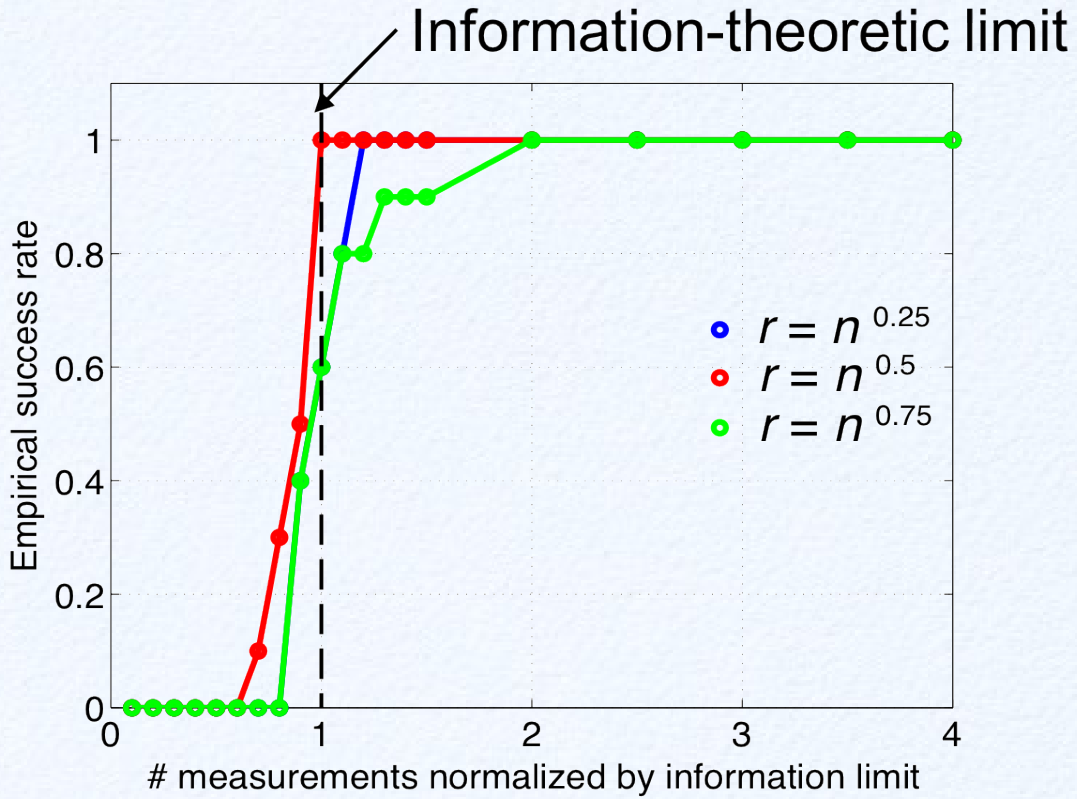




Info and comput. limits are identical for many spatially invariant graphs

Empirical success rate vs. sample size



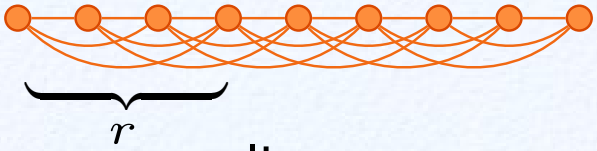


$n = 100,000$, input error rate $= 0.2$

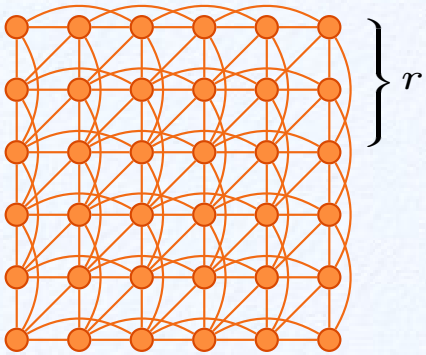
10 Monte Carlo runs to get each point

Each run takes $\sim$6.4 sec on a Mac Pro

Extension: beyond spatially invariant graphs

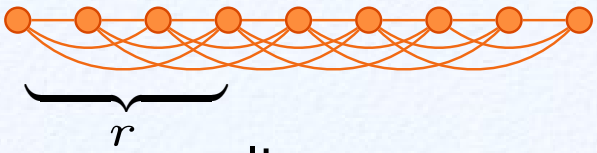


lines

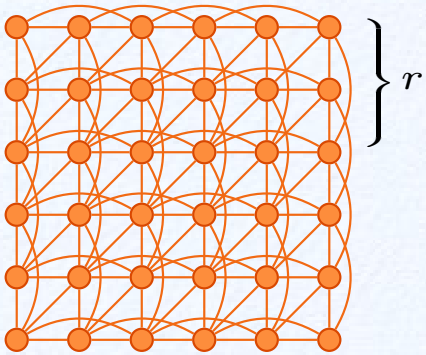


grids

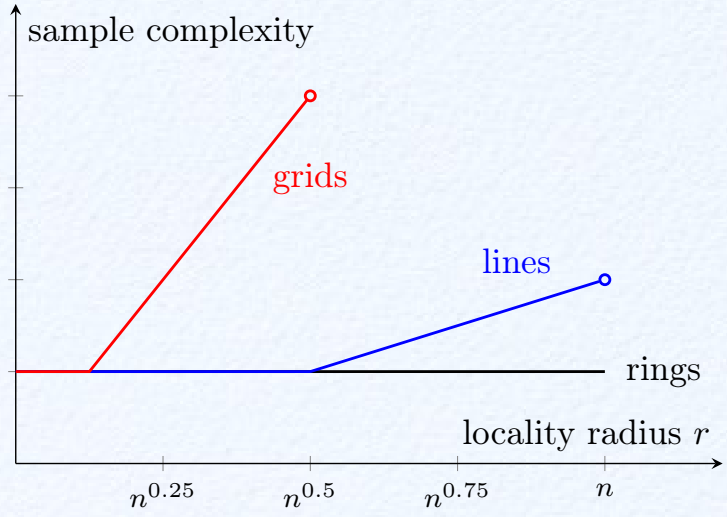
Extension: beyond spatially invariant graphs



lines



grids



Info limit vs. r

Infomation and comput. limits achievable by same algorithm

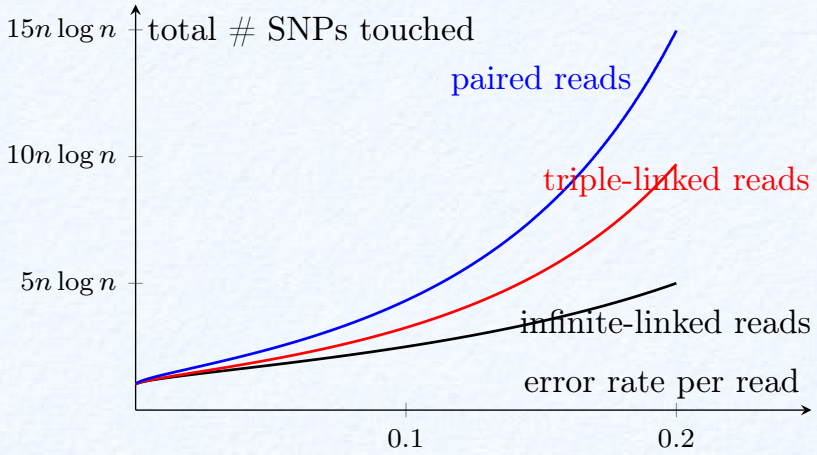
Extension: beyond pairwise measurements

New technologies (e.g. 10x) provide multi-linked reads from same chromosome, not just two

Extension: beyond pairwise measurements

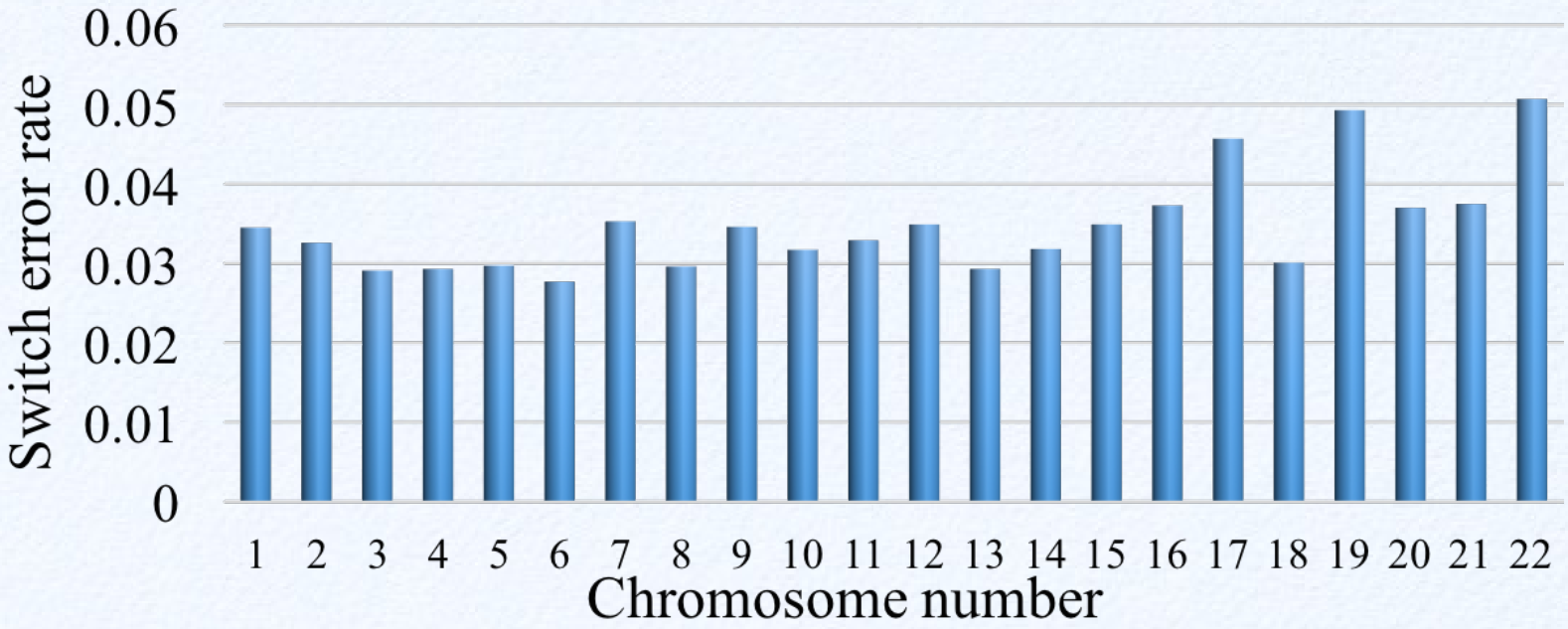
New technologies (e.g. 10x) provide multi-linked reads from same chromosome, not just two

Algorithm and theory can be easily extended to see performance gain



Initial results on real data (haplotype phasing)

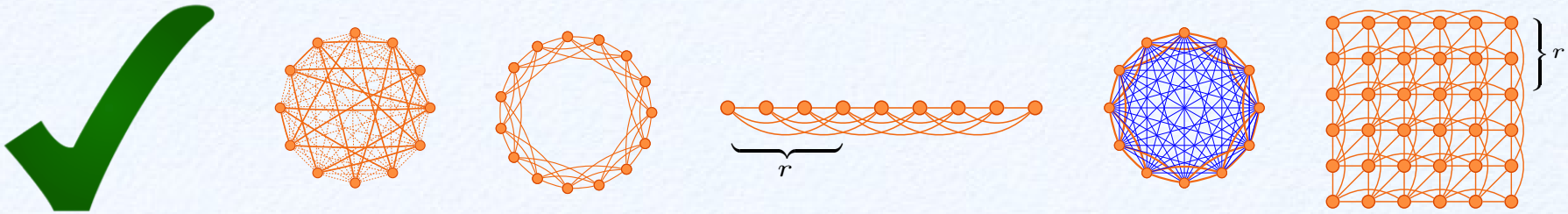
NA12878 dataset from 10x genomics



SNPs n : $34240 \sim 191829$, sample size m : $102633 \sim 574189$

Concluding remarks

- Studied community recovery when measurements are highly local
  - motivated by genome phasing and social networks
- Information limits can be achieved in linear time for a broad family of models



Full version of paper available at `http://arxiv.org/abs/1602.03828`