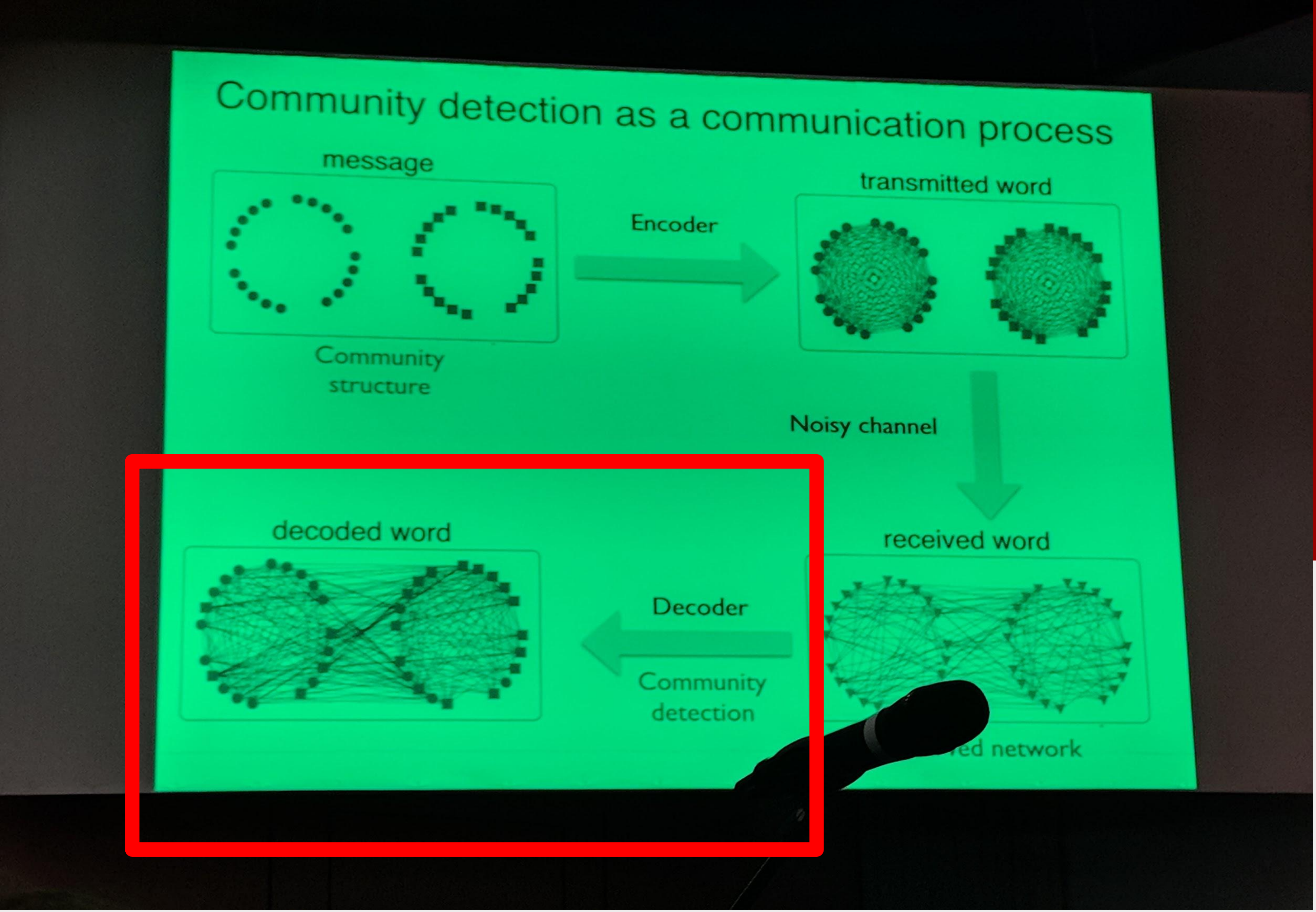
Error-Correcting Decoders for Communities in Networks

Krishna C. Bathina and Filippo Radicchi

Indiana University

CNetS

Center for Complex Networks and Systems Research

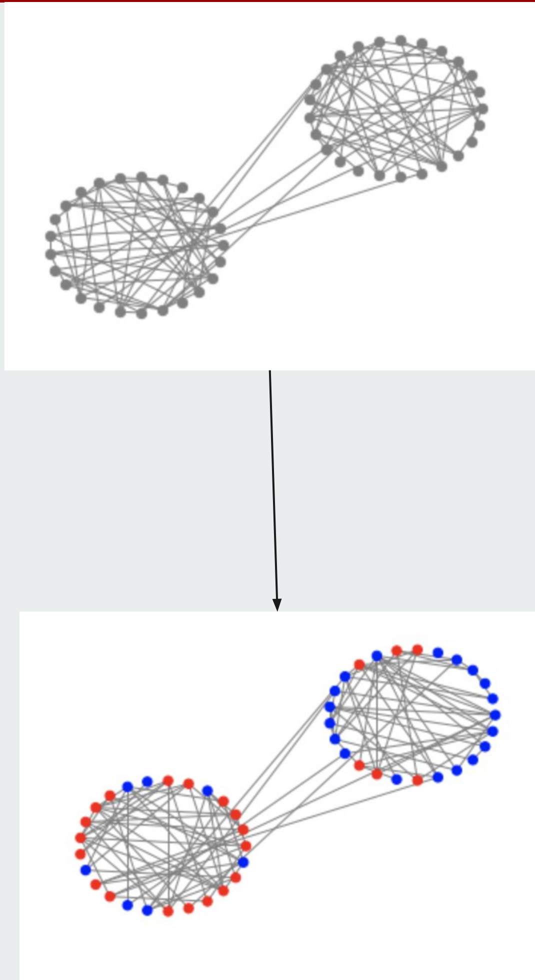


- Detectability threshold
- Noisy channel capacity
- Capacity achieving codes

Radicchi, Filippo. "Decoding communities in networks." *Physical Review E* 97.2 (2018): 022316.

Message becomes Decoded

A distorted message is passed through a decoding algorithm and the original message is returned - at least partially



0 if community of i = community of j
1 if community of i != community of j

Community membership of node i

$$\sigma_i + \sigma_j + \theta_{ij} \pmod{2} = 0$$

If the equation is solved for all n^2 pairs of nodes, then the distorted message has been perfectly decoded

Gallager Codes (LDPC codes)

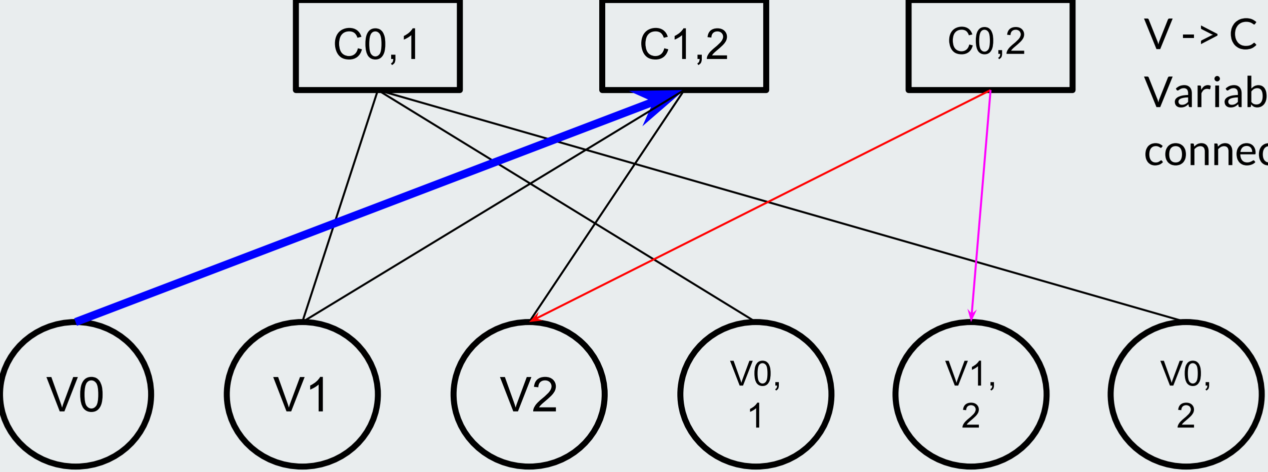
- Linear code based on a *Low-Density Parity Check* matrix H
- Check nodes -pairs of nodes in graph (3)
- Variable nodes - number of bits in codeword (6)

$$N + \frac{N(N-1)}{2}$$

$$\mathcal{H} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

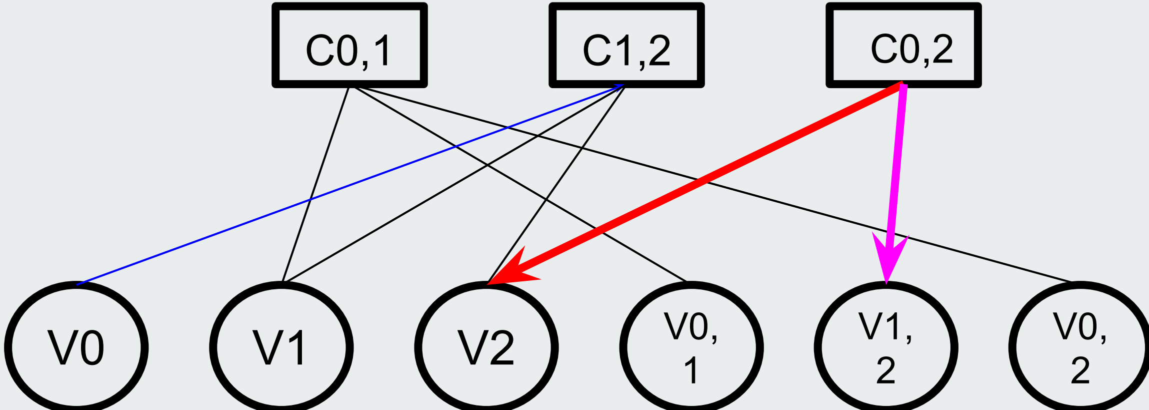
$$\frac{N(N-1)}{2}$$

Gallager, Robert. "Low-density parity-check codes." *IRE Transactions on information theory* 8.1 (1962): 21-28.



V -> C Message = probability of Variable's bit according the connected Check nodes

C -> V Message = probability of Variable's bit that would satisfy the other connected Variable nodes



Gallagher Decoder

a priori log likelihood ratio (LLR)

- Prior belief about the message given the information received
- Determines which steady state value the algorithm will converge to

Variable nodes i - Logarithm of the ratio of the community memberships given the received information bit

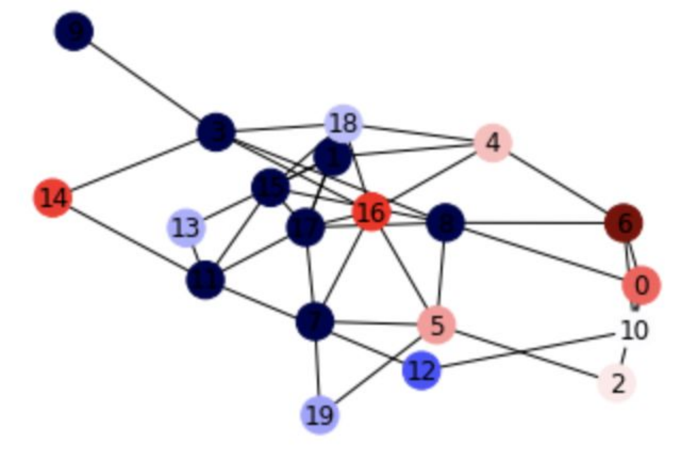
$$\ell_i = \log \frac{P(\sigma_i = 0|s_i)}{P(\sigma_i = 1|s_i)}$$

Variable nodes i,j - Logarithm of the ratio of the parity bits given the existence of an edge

$$\ell_{ij} = \log \frac{P(\theta_{ij} = 0|A_{ij})}{P(\theta_{ij} = 1|A_{ij})}$$

Random

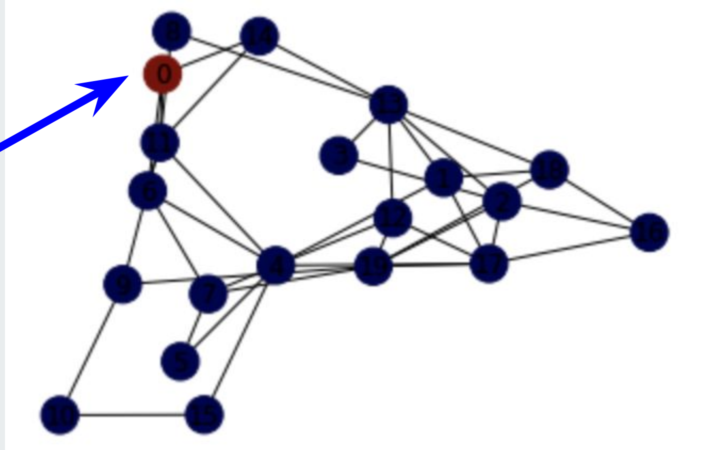
$$\ell_N \sim \mathcal{U}(-1, 1)$$



Regular

$$\ell_i = 1 \qquad \ell_{N \setminus i} = 0$$

Very confident about community membership



Variable nodes i

$A_{i,j}$	$\theta_{i,j}$	$P(A_{i,j}\|\theta_{i,j})$	$P(\theta_{i,j}\|A_{i,j})$
1	0	P_{in}	$\frac{P_{in}}{P_{in}+P_{out}}$
1	1	P_{out}	$\frac{P_{out}}{P_{in}+P_{out}}$
0	0	1 - P_{in}	$\frac{1-P_{in}}{2-(P_{in}+P_{out})}$
0	1	1 - P_{out}	$\frac{1-P_{out}}{2-(P_{in}+P_{out})}$

$$\ell_{ij} = \log \frac{P(\theta_{ij} = 0|A_{ij})}{P(\theta_{ij} = 1|A_{ij})}$$

$$= \begin{cases} \log(p_{in}) - \log(p_{out}) & \text{if} A_{ij} = 1 \\ \log(1 - p_{in}) - log(1 - p_{out}) & \text{if} A_{ij} = 0 \end{cases}$$

Stochastic block model

Variable nodes i,j

P_{in} is large enough to convey information about the community assignments

$$P_{in} - P_{out} \geq \frac{2\sqrt{k}}{N}$$

$$P_{in} + P_{out} = \frac{2k}{N}$$

2 communities

Tunable parameter

$$P_{in} = \alpha \frac{k + \sqrt{k}}{N}$$

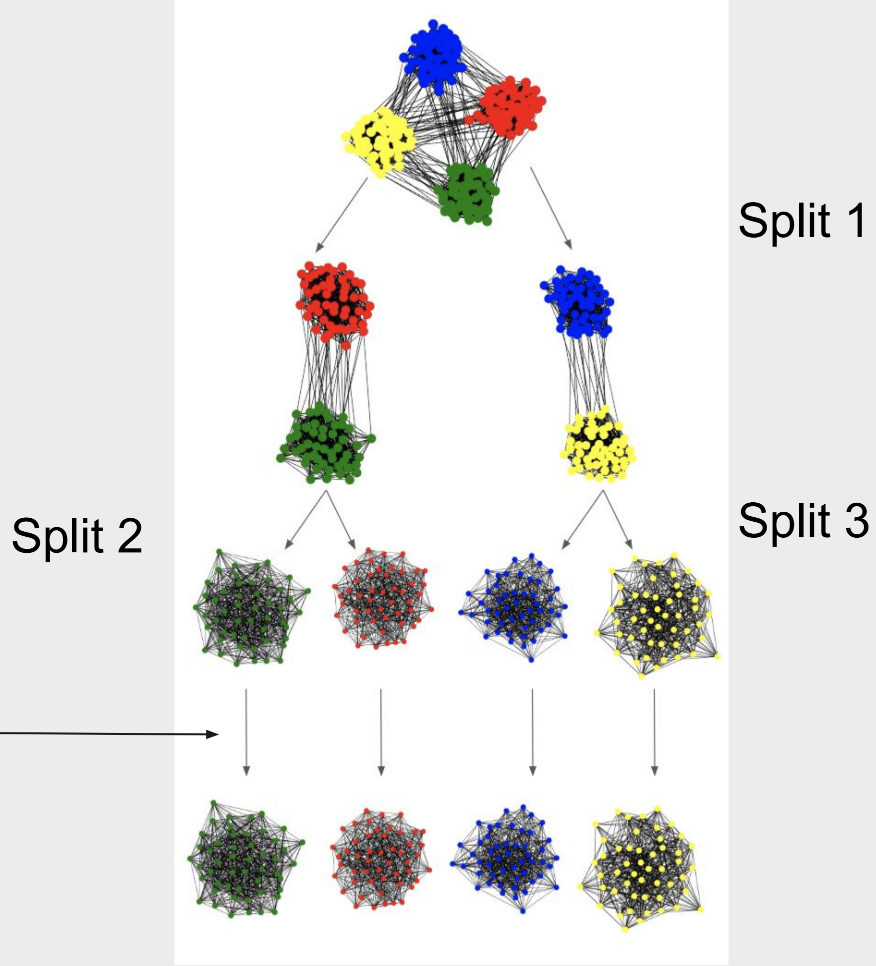
$$P_{out} = \max(0, \frac{\alpha k}{N} - P_{in})$$

Decelle, Aurelien, et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications." *Physical Review E*84.6 (2011): 066106.

Stochastic Block Model

Algorithm

1. Choose starting condition
2. Run on network
   a. Iterative decoding - new P_{in} and P_{out} for each iteration
   b. Return 2 sub-networks
3. Repeat on each subnetwork until no new splits are formed



Original Algorithm (Reformulated Gallagher)

$$F(a, x) = \log \frac{1 + a \tanh \frac{x}{2}}{1 - a \tanh \frac{x}{2}}$$

Best *a priori* initial estimate of message

$$\zeta_{i \to j}^{t=0} = \ell_i$$

Update to LLR based on all other nodes

$$\zeta_{i \to j}^{t} = \zeta_{i \to j}^{t=0} + \sum_{k \in N \setminus i,j} F[\tanh \frac{\ell_{i,j}}{2}, \zeta_{k \to i}^{t-1}]$$

Iterative Update of LLR

$$\ell_i^f = \ell_i + \sum_{k \in N \setminus i} F[\tanh \frac{\ell_{i,k}}{2}, \zeta_{k \to i}^{f-1}]$$

$$\ell_{i,j}^f = \ell_{i,j} + F[\tanh \frac{\ell_{i,j}}{2}, \zeta_{j \to i}^{f-1}]$$

Hard Decision

$$\begin{cases} \sigma_i = 0 \text{ if } \ell_i^f > 0 \\ \theta_{i,j} = 0 \text{ if } \ell_{i,j}^f > 0 \end{cases}$$

Best Estimate of LLR

Reduced Algorithm

- Original algorithm - messages are updated by all pairs of nodes
  - Even if $A_{ij} = 0$
- Assume messages passed between unconnected nodes are constant

$$\zeta^t_{i\to j} = \zeta^{t=0}_{i\to j} + \sum_{k\in N\setminus i,j} F[\tanh\frac{\ell_{i,j}}{2}, \zeta^{t-1}_{k\to i}]$$

LLR iterates on all node pairs, *even if* not connected

Constants

$A_{i,j}$	$\theta_{i,j}$	$P(A_{i,j}\|\theta_{i,j})$	$P(\theta_{i,j}\|A_{i,j})$
1	0	P_{in}	$\frac{P_{in}}{P_{in}+P_{out}}$
1	1	P_{out}	$\frac{P_{out}}{P_{in}+P_{out}}$
0	0	1 - P_{in}	$\frac{1-P_{in}}{2-(P_{in}+P_{out})}$
0	1	1 - P_{out}	$\frac{1-P_{out}}{2-(P_{in}+P_{out})}$

$$\ell_{con} = P(\theta = 0|A = 1) - P(\theta = 1|A = 1) = \frac{P_{in} - P_{out}}{P_{in} + P_{out}}$$

$$\ell_{non} = P(\theta = 0|A = 0) - P(\theta = 1|A = 0) = \frac{P_{out} - P_{in}}{2 - P_{in} - P_{out}}$$

Update to LLR if $A_{i,s} = 1$

$$\zeta_{i \to j}^{t=0} = \ell_i$$

Update to LLR for all unconnected nodes

Best *a priori* initial estimate

$$\zeta_{i \to j}^{t} = \zeta_{i \to j}^{t=0} + (N - k_i - 1) F\left[\frac{\ell_{non}}{2}, \mathcal{Z}^{t-1}\right] + \sum_{s \in \mathcal{N}_i \setminus j} F\left[\frac{\ell_{con}}{2}, \zeta_{s \to i}^{t-1}\right]$$

Number of unconnected nodes to node *i* excluding node j

Iterative Update - Updates to messages passed between edges

$$\mathcal{Z}^{t=0} = \frac{\sum_{i=1}^{N} (N - k_i - 1)\ell_i}{N(N-1) - 2M}$$

Sum of all initial messages sent to unconnected nodes

Total number of non-edges

Best *a priori* initial estimate

Average update to LLR between connected nodes

$$\mathcal{Z} = \mathcal{Z}^{t=0} + F\left[\frac{\ell_{non}}{2}, \mathcal{Z}^{t-1}\right] + \frac{\sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} F\left[\frac{\ell_{con}}{2}, \zeta_{i \to j}^{t-1}\right]}{N(N-1) - 2M)}$$

Iterative Update - Updates to messages passed between non-edges

Number of unconnected nodes to node *i*

$$\ell_i^f = \ell_i + \sum_{s \in \mathcal{N}_i} + F\left[\frac{\ell_{con}}{2}, \zeta_{s\to i}^{f-1}\right] + (N - k_i) F\left[\frac{\ell_{non}}{2}, \mathcal{Z}^{f-1}\right]$$

$$\ell_{i,j}^f = \log \frac{P_{in}}{P_{out}} + F[\tanh \frac{\zeta_{i\to j}^{f-1}}{2}, \zeta_{j\to i}^{f-1}]$$

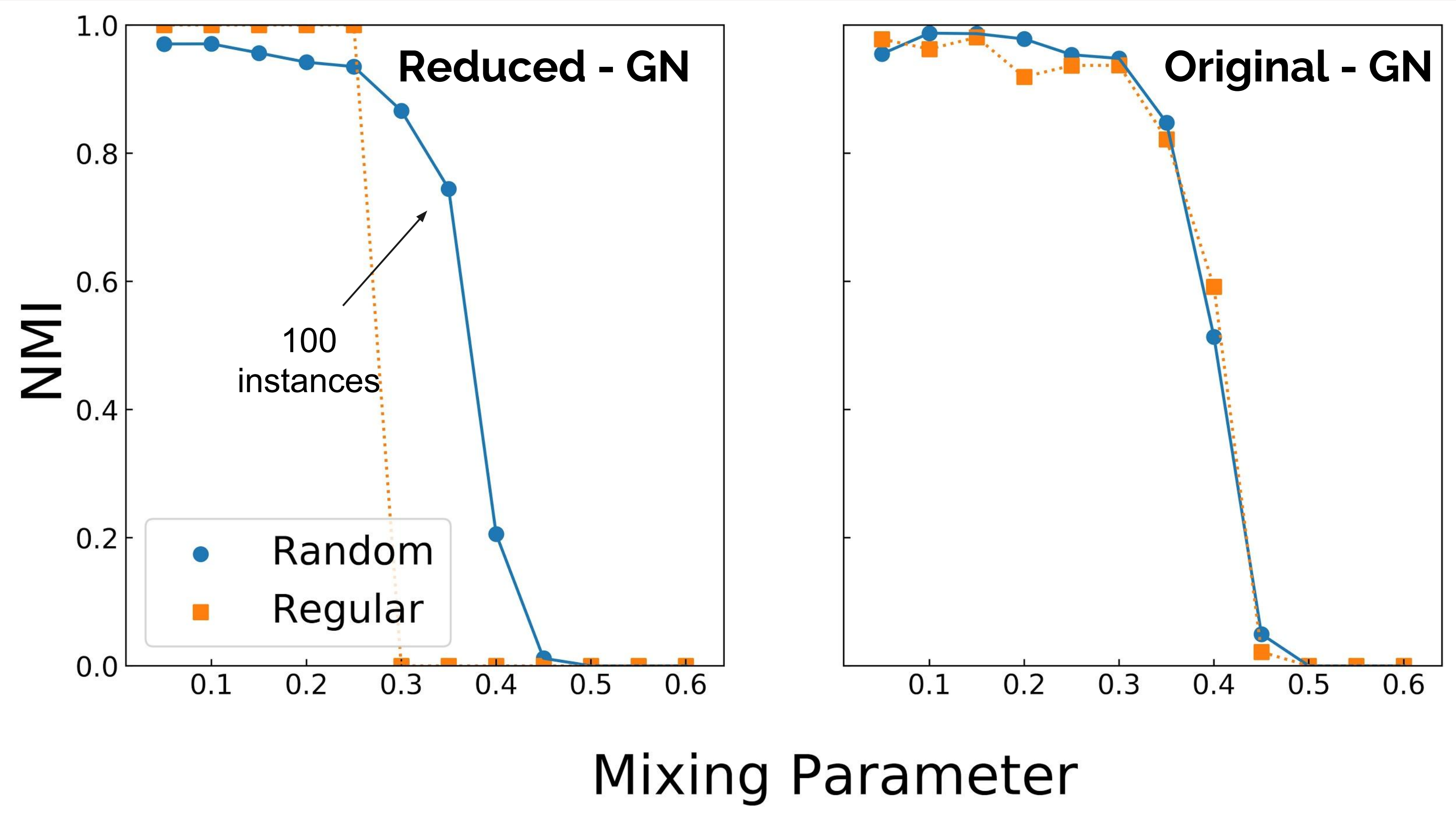
Hard Decision

$$\begin{cases} \sigma_i = 0 \text{ if } \ell_i^f > 0 \\ \theta_{i,j} = 0 \text{ if } \ell_{i,j}^f > 0 \end{cases}$$

Best Estimate of LLR

Girvan Newman

Mostly either completely perfect or completely imperfect recovery



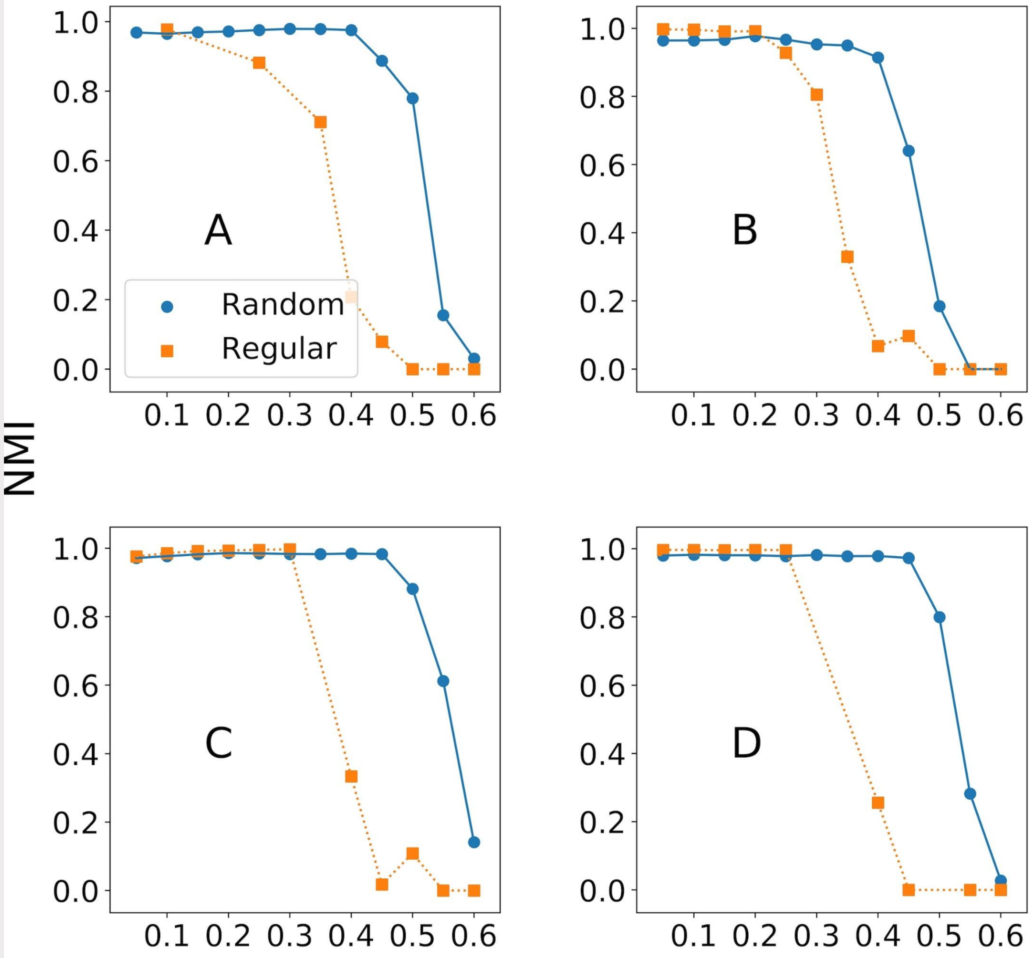
Reduced LFR

- Small - 10-50 nodes/community
- Big - 20-100 nodes/community

A. 1000 Small
B. 1000 Big
C. 5000 Small
D. 5000 Big

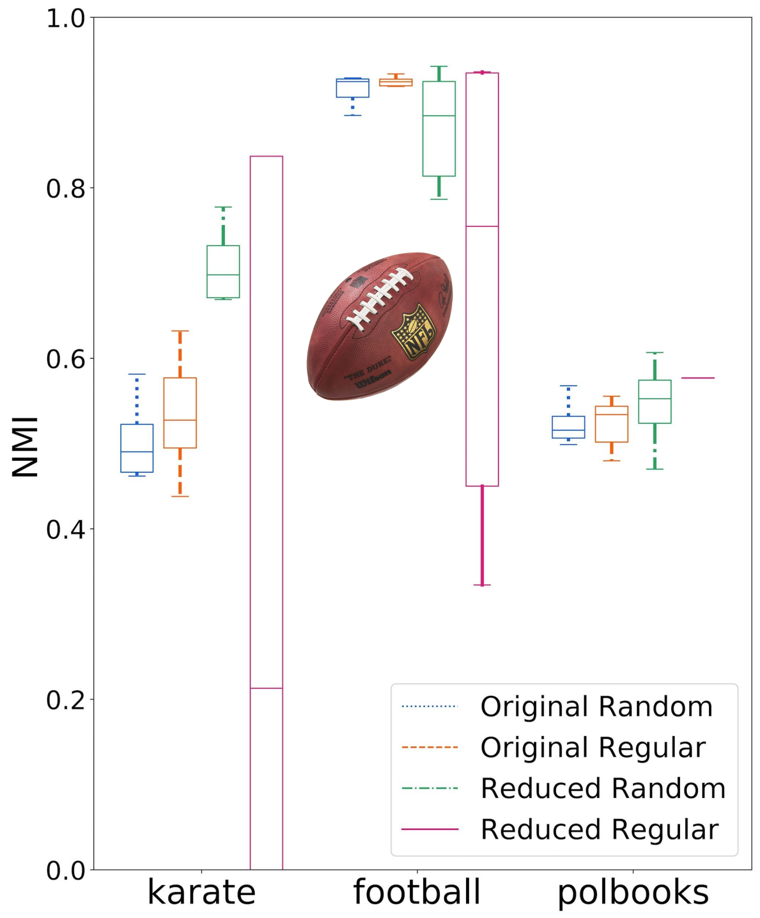
Mostly either completely perfect or completely imperfect recovery

Lancichinetti, Andrea, and Santo Fortunato. "Community detection algorithms: a comparative analysis." *Physical review E* 80.5 (2009): 056117.



Using Metadata

- Zachary Karate Club
- NCAA Football leagues
- US political books sold on Amazon during the 2004 election



http://www-personal.umich.edu/~mejn/netdata/

Thank you!