Criticality and Phase Transitions in Diversity Colouring

http://www.ncrg.aston.ac.uk/

K.Y.M Wong and D. Saad

HKUST and Aston University

17 July 2007

Statistical mechanics of distributed information systems

Outline

- What is distributed storage?
- Coding and segment distribution
- Message passing for segment distribution
- Segment distribution - analysis
- Numerical results - transition point
- Future research directions

Distributed storage

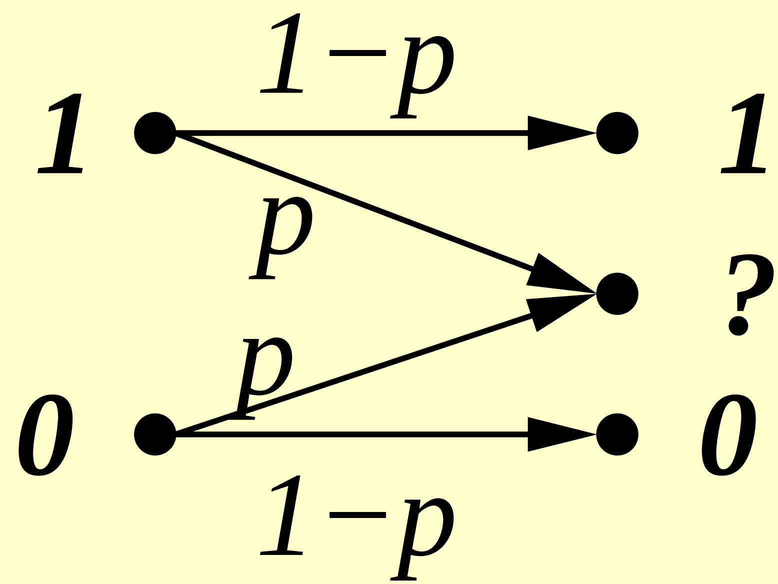
- A system of nodes (computers?) on a graph
- Main problems -
  - Limited storage and computing capability
  - Bottlenecks, speed of download
- Distributing file (segments), retrieval from neighbours
- Available methods
  - Replication
  - Cashing
  - Coding

Distributed storage - advantages

- Decentralised
- Computationally efficient
- Minimises latency
- Fault-tolerant; high availability of data even with some component failure
- Load balancing (both storage and traffic)
- Increased data security against eavesdropping over individual links

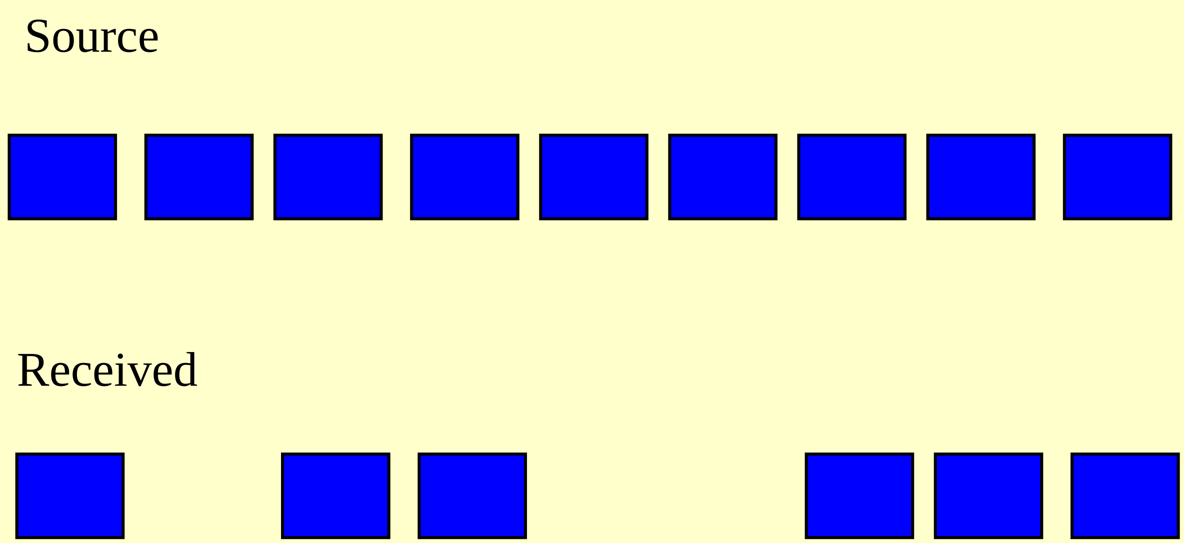
Coding - Binary Erasure Channel

- The message is a binary vector
- Bits get lost while being transmitted via a Binary Erasure Channel
- Received bits are correct
- Coding techniques for retrieving lost information, e.g., Low Density Parity Check Codes



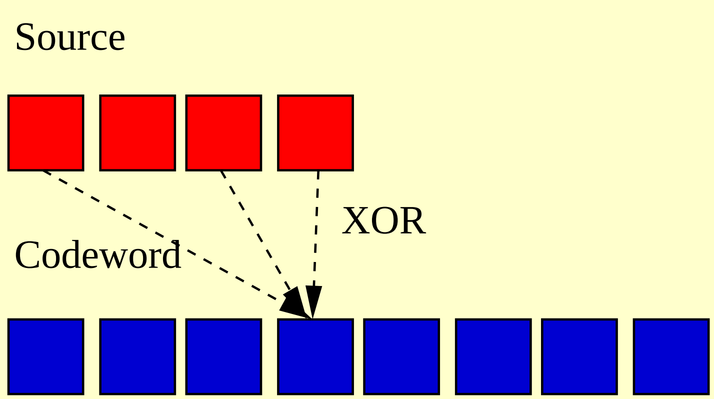
Coding - file segments

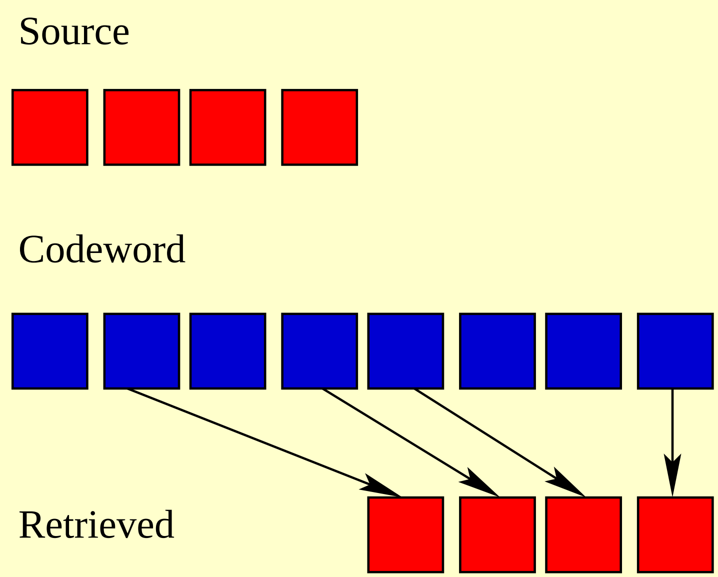
- Distributed storage can be viewed as file reconstruction from partial information
- Only part of the original segments are available



Fountain codes

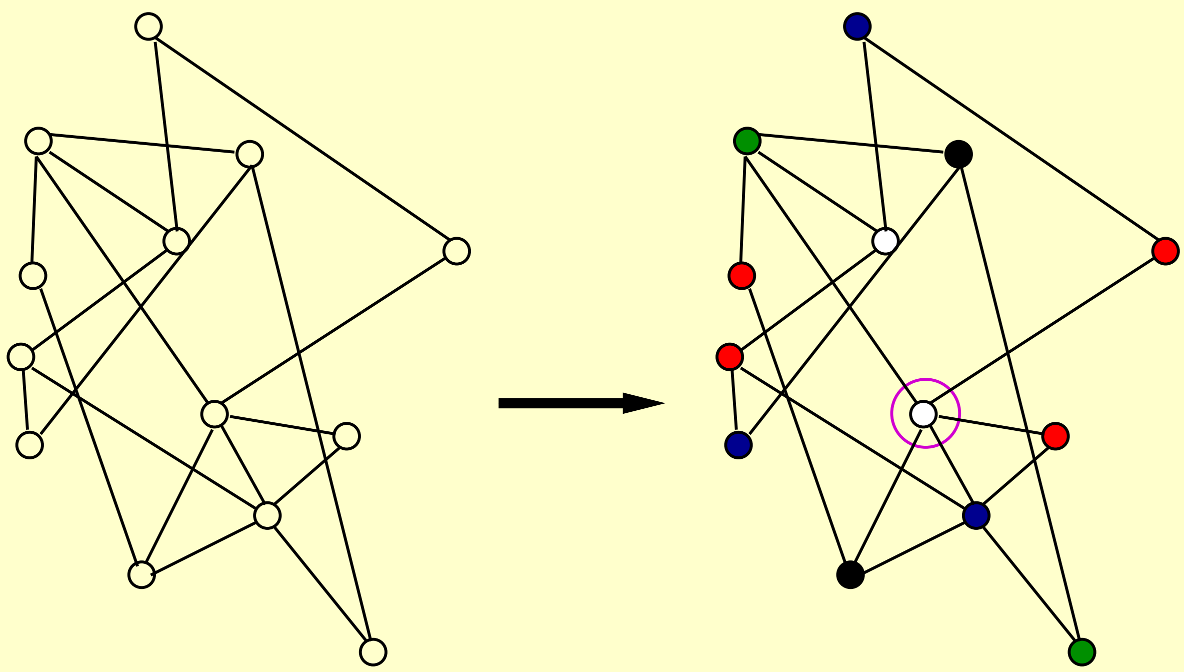
- Best suited for BEC, highly robust
- Generating a fountain of codewords from data packets, by XOR of d packets taken from some distribution of $\rho(d)$
- Decoding by message passing/deterministic





Segment distribution

- Different segments can be viewed as colours
- Main problem - how to maximise the number of different colours in a neighborhood - Diversity Colouring
- File retrieval - when a sufficiently large subset of segments is collected



Performance measures

- Number of identical segments (colours) in neighbourhood $\mathcal{L}_i \equiv \{q_i\} \cup \{q_j | j \in N_i$

$$E = \sum_i \phi\left(\mathcal{L}_i\right) = \sum_i \sum_{q=1}^{Q} \left[\delta\left(q, q_i\right) + \sum_{j \in N_i} \delta\left(q, q_j\right)\right]^2$$

- Incomplete fraction $f_{\rm incom}$ - average fraction of nodes with an incomplete colour set

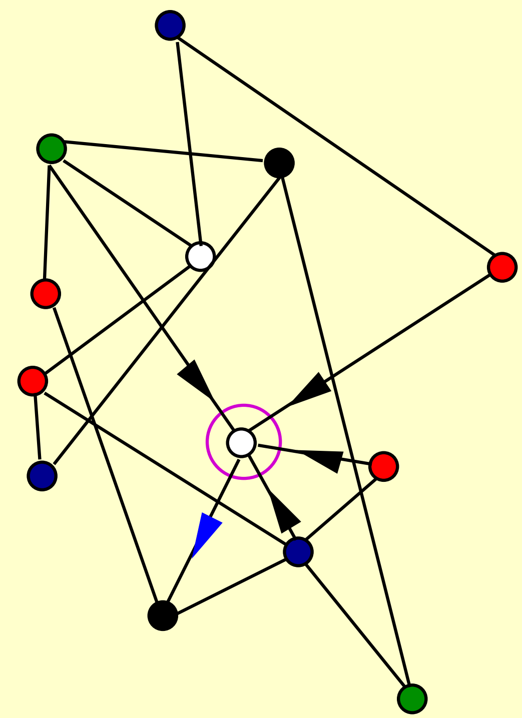
$$f_{\textsf{incom}} = \left\langle \Theta\left[Q - \sum_{q=1}^{Q} \Theta\left(\delta(q, q_i) + \sum_{j \in N_i} \delta(q, q_j)\right)\right]\right\rangle$$

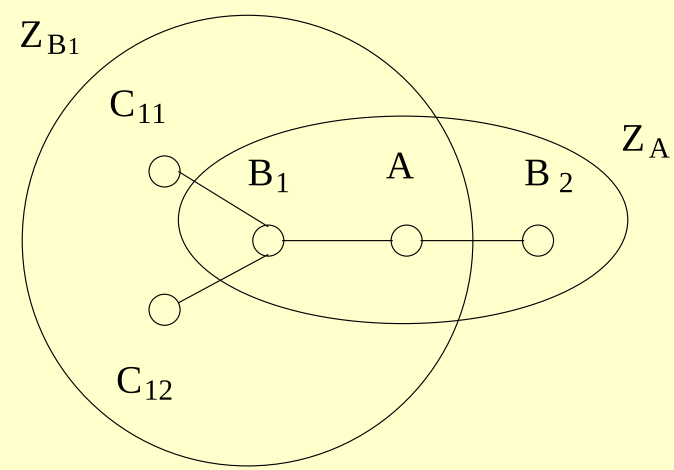
- Unsatisfied fraction $f_{\textsf{unsat}}$ - average fraction of unavailable colours at a neighbourhood ($Q \leq$ number of nearest neighbors plus 1)

$$f_{\textsf{unsat}} = \left\langle \left[1 - \frac{1}{Q} \sum_{q=1}^{Q} \Theta\left(\delta(q, q_i) + \sum_{j \in N_i} \delta(q, q_j)\right)\right]\right\rangle$$

Message passing

- Previous work [a] used message passing technique for assigning segments to nodes (factoring approximation)
- Vector messages represent probabilities of having certain colours
- Requires messages from 2nd order neighbours
- Pseudoposteriors are averaged over a time window





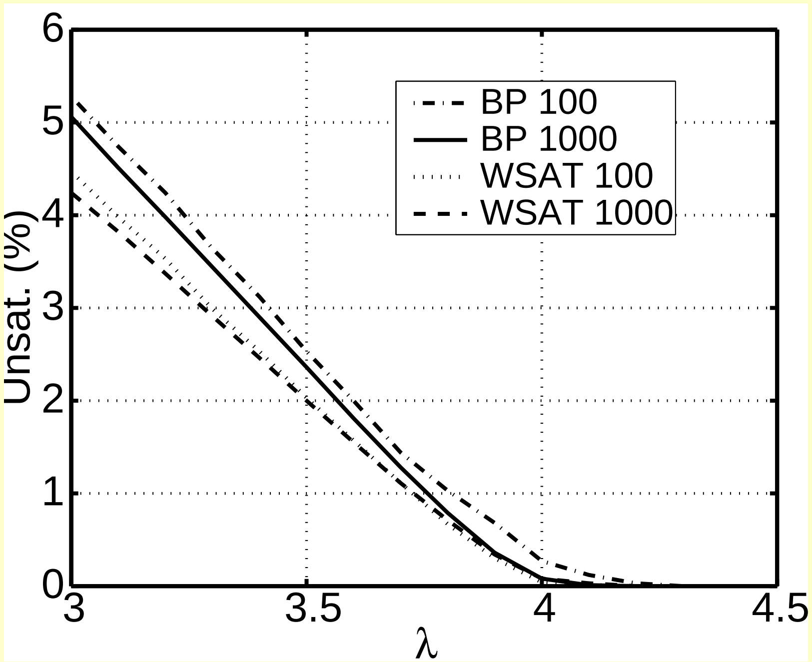
[a] Bounkong, van Mourik, Saad, PRE **74** 057101 2006

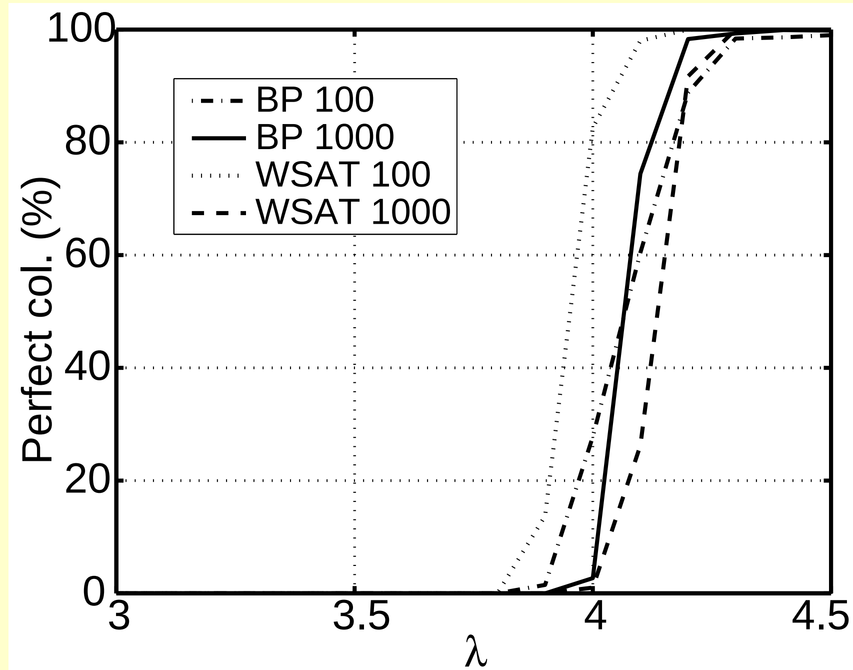
Experiments

- Graph structure
  - Random graph - Poissonian distribution of a given mean (limited)
  - Linear fixed connectivity - nodes have one of two consecutive fixed connectivity value (i.e. connectivity 3.5 would mean that node have connectivities 3 and 4 with equal probability)
- Network size - 100, 1000 nodes
- Number of segments - 4
- Averaged over - 1000 graphs
- Comparison with WalkSAT -
  - GSAT/random with probability 1/2
  - Very high limit for the number of search steps

Results - BP vs. WSAT

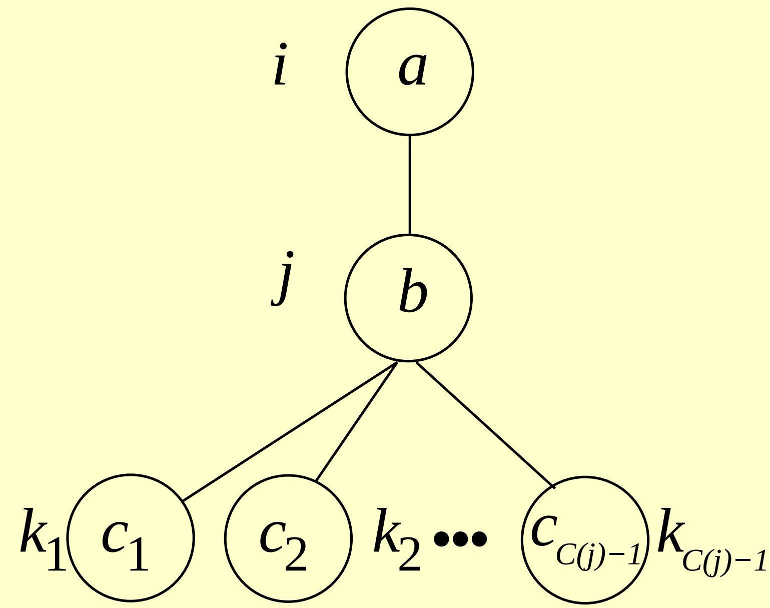
For the case of linear connectivity we get





Tree approximation

- Analysis can be carried out using the replica method and the tree approximation.
- $F = -T \ln Z$ where $Z = \mathrm{Tr}_{\{q_i\}} \exp\left[-\beta \sum_i \phi(\mathcal{L}_i)\right]$
- Vertex free energy - the key property to be evaluated recursively



$$F_{ij}(a,b) = -T \ln \mathrm{Tr}_{\{c_k|k\in N_j\backslash\{i\}\}} \exp\left[-\beta \sum_{k\in N_j\backslash\{i\}} F_{jk}(b,c_k) - \beta\phi\left(b, \{a\} \cup \{c_k|k \in N_j\backslash\{i\}\}\right)\right] - F_{\mathrm{av}}$$

Expected values

- $F_{ij}(a,b)$ values are obtained recursively

- Expectation values can be calculated

$$\langle \mathcal{A} \rangle = \left\langle \frac{\text{Tr}_{\{\mathcal{L}_i\}} \exp\left[-\beta \sum_{j\in N_i} F_{ij}(q_i,q_j) - \beta\phi\left(\mathcal{L}_i\right)\right] \mathcal{A}(\mathcal{L}_i)}{\text{Tr}_{\{\mathcal{L}_i\}} \exp\left[-\beta \sum_{j\in N_i} F_{ij}(q_i,q_j) - \beta\phi\left(\mathcal{L}_i\right)\right]} \right\rangle_{\text{net}}$$

for instance $\langle E \rangle = N\langle \phi \rangle$

- We calculate f_{incom}, f_{unsat} energy, entropy

Obtaining solutions

- In the zero temperature limit

$$F_{ij}(a,b) = \min_{\{q_k|k\in N_j\backslash\{i\}\}} \Bigg[\sum_{k\in N_j\backslash\{i\}} F_{jk}(b,q_k)$$

$$+\phi\left(b, \{a\} \cup \{q_k | k \in N_j\backslash\{i\}\}\right)\Bigg] - F_{\mathsf{av}}$$

- Due to possible degeneracy expectation values should be weighted by the vertex entropy

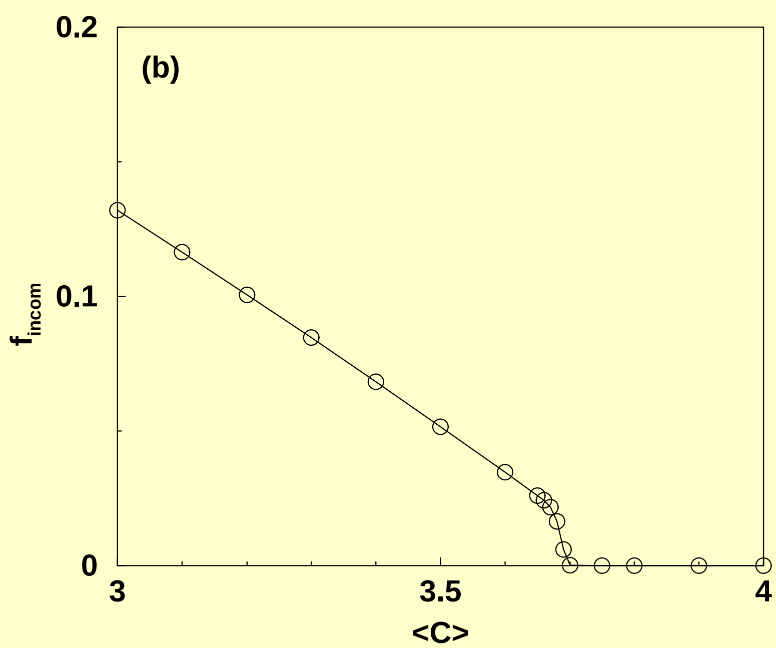
$$\langle \mathcal{A} \rangle = \left\langle \frac{\mathrm{Tr}_{\cup_{j\in N_i}\mathcal{X}_{ij}} \exp\left[\sum_{j\in N_i} S_{ij}(q_i,q_j)\right] \mathcal{A}(\mathcal{L}_i)}{\mathrm{Tr}_{\cup_{j\in N_i}\mathcal{X}_{ij}} \exp\left[\sum_{j\in N_i} S_{ij}(q_i,q_j)\right]} \right\rangle_{\mathrm{net}}.$$

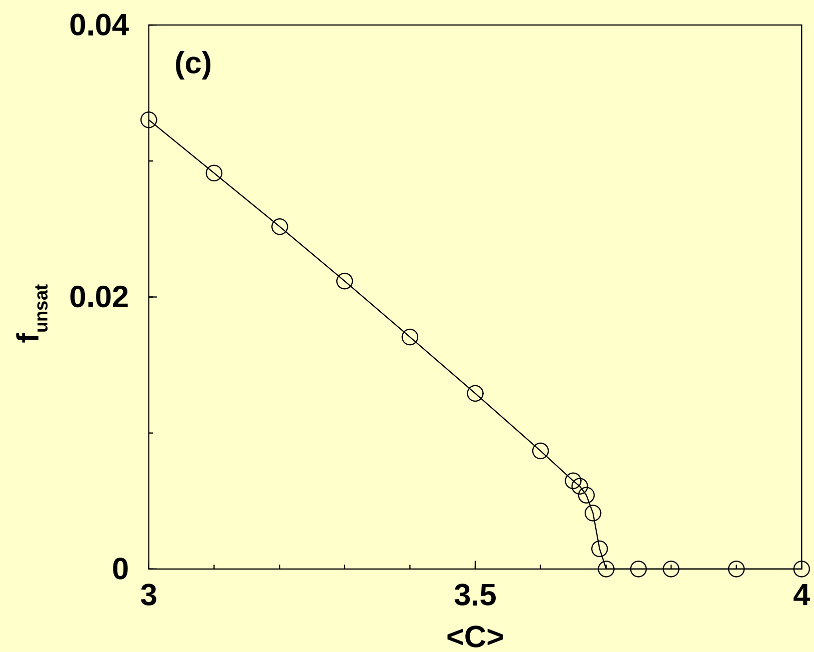
Experiments

- Graph structure - Linear fixed connectivity:
  for average connectivity $\langle C\rangle$, C_j selected as $\lfloor\langle C\rangle\rfloor + 1$ / $\lfloor\langle C\rangle\rfloor$ with probabilities $\langle C\rangle - \lfloor\langle C\rangle\rfloor$, and $1 - \langle C\rangle + \lfloor\langle C\rangle\rfloor$
- Network size - 1000 nodes at each generation
- Number of segments - $Q = 4$
- Critical transition - $\langle C\rangle = 3.68$

Results - $f_{incom/unsat}$

For the case of linear connectivity and $Q = 4$ we obtain the following f_{incom}, f_{unsat} values

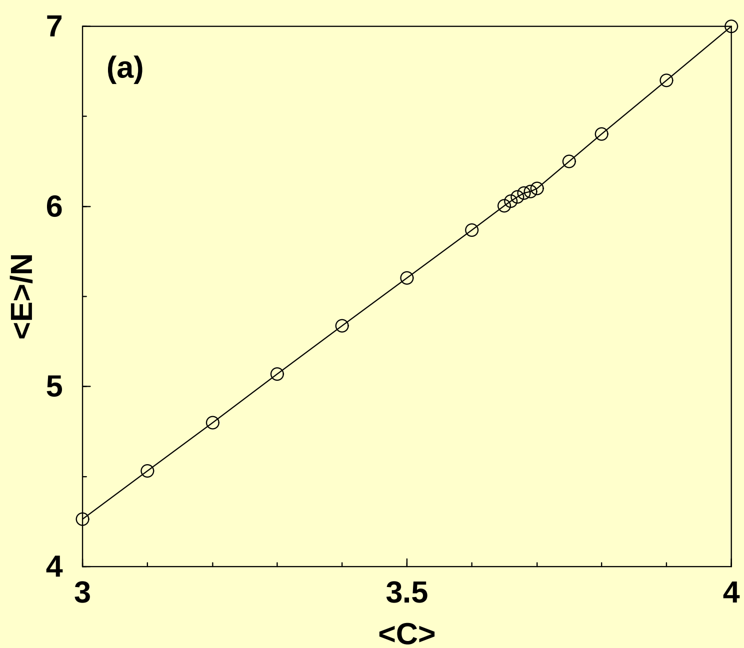


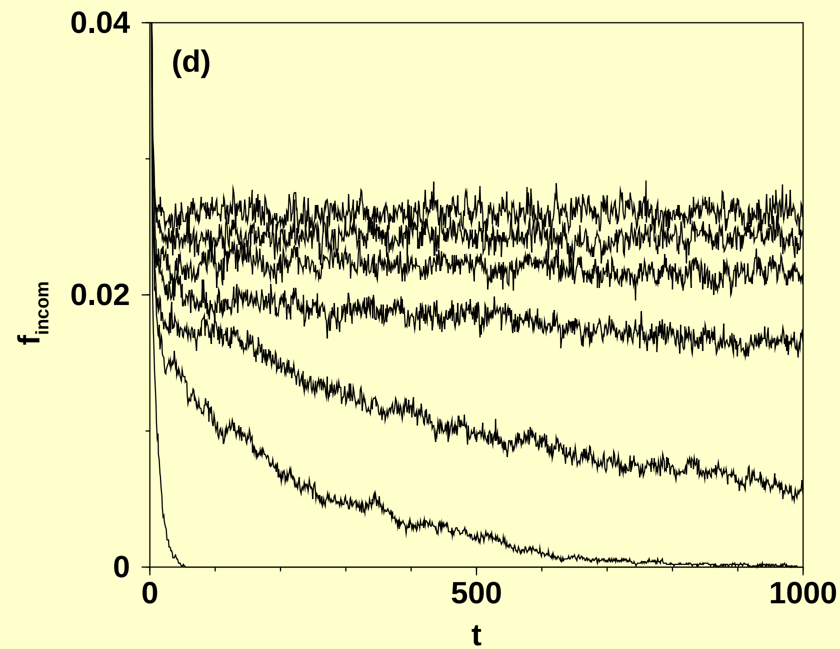


Results - E

- The energy E is growing linearly with connectivity

- Convergence times slow considerably as $\langle C \rangle$ approaches 3.68

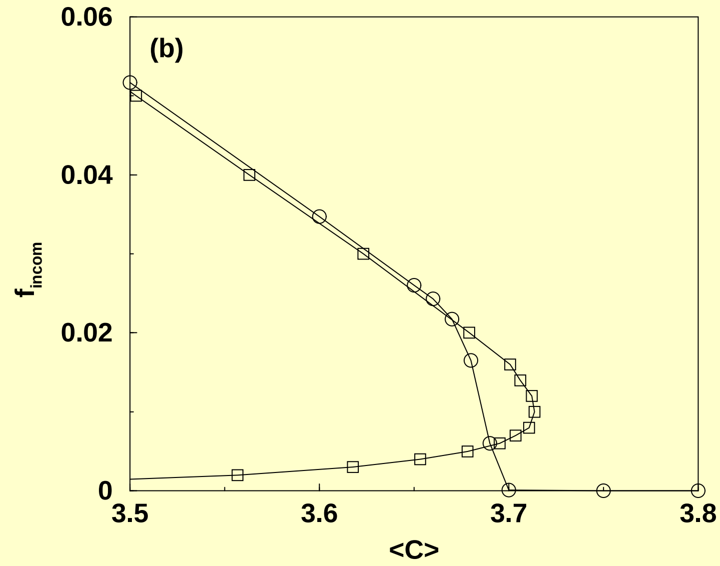




Results - fixed performance measures

- For fixed f, update vertex free energies for $\langle C\rangle$.

- Compute $f(\lfloor\langle C\rangle\rfloor)$ and $f(\lfloor\langle C\rangle\rfloor+1)$ (two samples)



- Set trial connectivity $\langle C\rangle$ for the next time step to satisfy $f=f(\lfloor\langle C\rangle\rfloor)(\lfloor\langle C\rangle\rfloor+1-\langle C\rangle)+f(\lfloor\langle C\rangle\rfloor+1)(\langle C\rangle-\lfloor\langle C\rangle\rfloor)$

- Repeat until convergence

Validity of the results

- To determine the range in which the results are valid one should look for an indication of inconsistency (RSB)

- Negative entropy

- Stability

Summary & future research

- We use a tree based analysis to study the properties of diversity colouring on sparse graphs
- Transitions points for various performance measures calculated
- Future Research:
  - Improved algorithms
  - Multiple segments per node
  - Multiple hops
  - Dynamical allocation of segments
  - RSB