On the number of circuits in random graphs

Guilhem Semerjian

[joint work with Enzo Marinari and Rémi Monasson]

[Europhys. Lett. 73, 8 (2006)]

and [cond-mat/0603657]

Orsay
13-04-2006

Outline of the talk

- Introduction
- A statistical mechanics approach
- The Bethe approximation / Belief Propagation
- An approximate algorithm
- Typical number of loops in random graph ensembles

Why counting loops (and why is it difficult ?)

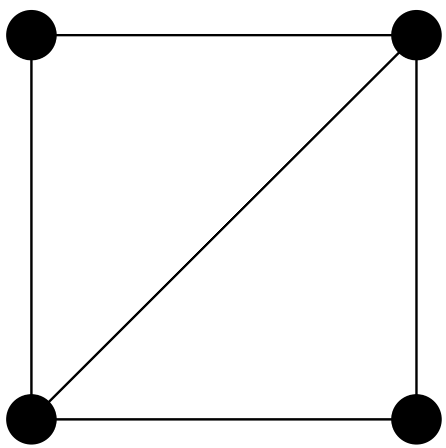
- Real-world networks motivation
  - Large amount of experimental datas on real-world networks (Internet for instance), characteristic features to (in)validate proposed models ?
  - Local properties (connectivity distribution, clustering coefficient, number of short loops,. . .) easy to measure.
  - Global properties (long loops or other large patterns) difficult :
    - exponential number of long loops $\Rightarrow$ tricky to enumerate
    - deciding if there is an Hamiltonian circuit is NP-complete
- Random graph motivation
  - Statistical properties of the number of loops (graph ensembles)

Definitions

Graphs G with N vertices and M edges

Loop (circuit, cycle) of length L : closed, non-intersecting path which visits L vertices (and L edges)

$\mathcal{N}_L$: number of distinct loops of length L



$N = 4$

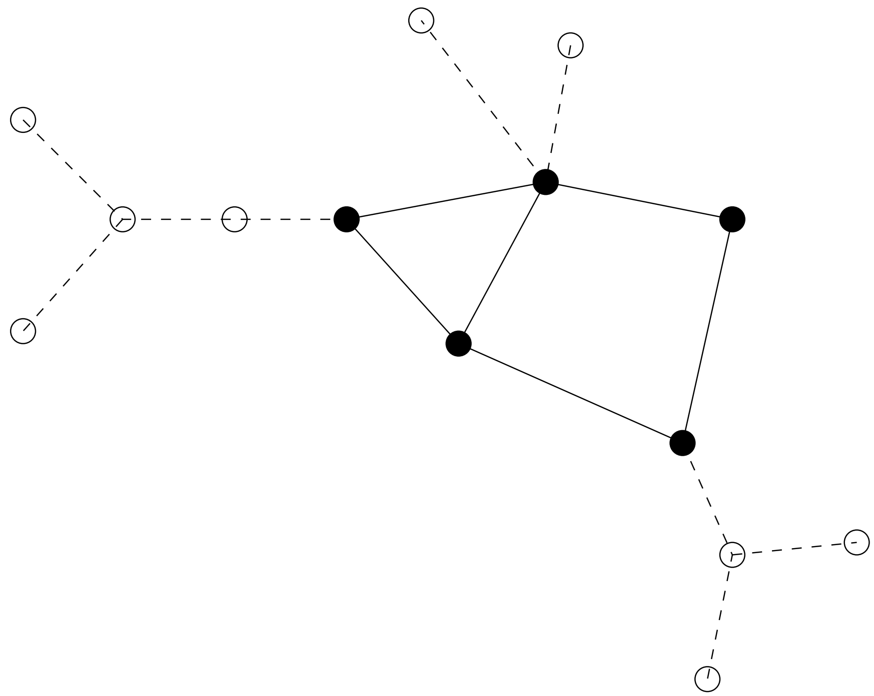
$M = 5$

$\mathcal{N}_3 = 2$

$\mathcal{N}_4 = 1$

2-core of a graph : largest of its subgraph in which all vertices have degree at least 2

Can be determined by leaf-removal



No loops out of the 2-core

A statistical mechanics approach

In general, computing the partition function
$\Leftrightarrow$ solving a difficult (microcanonical) enumeration problem :

$$\mathrm{Prob}[\mathcal{C}] = \frac{e^{-\beta E[\mathcal{C}]}}{Z(\beta)} \; , \qquad Z(\beta) = \sum_{\mathcal{C}} e^{-\beta E[\mathcal{C}]} = \sum_{E} \mathcal{N}_E \, e^{-\beta E}$$

For the enumeration of circuits, we need

$$Z(u) = \sum_{L} \mathcal{N}_L \, u^L$$

In the thermodynamic limit, $\ell = L/N$, $\mathcal{N}_L \sim e^{N\sigma(\ell)}$,

$$Z(u) = \sum_L \mathcal{N}_L \, u^L \quad \sim \quad \int d\ell \; e^{N[\sigma(\ell)+\ell \ln u]}$$

$$\sim \quad e^{N[\sigma(\ell(u))+\ell(u) \ln u]}$$

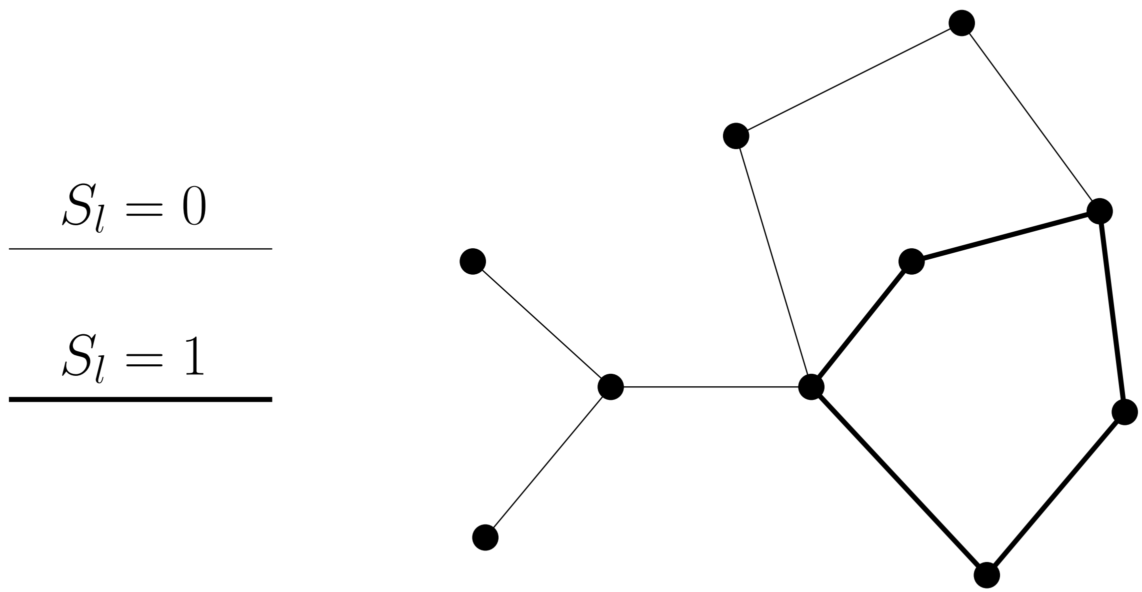
with $\ell(u)$ maximizing $[\ldots]$:

the « temperature » u selects loops of length $N\ell(u)$

What is the corresponding statistical mechanics model ?

degrees of freedom : $S_l \in \{0, 1\}$ on the M edges of the graph

$\mathcal{C} = \underline{S} = \{S_1, \ldots, S_M\}$ = subgraph of G (retain edges with $S_l = 1$)



$$\text{Prob}[\mathcal{C}] = \begin{cases} 0 & \text{if } \mathcal{C} \text{ is not a circuit} \\ \frac{u^L}{Z(u)} & \text{if } \mathcal{C} \text{ is a circuit of length } L \end{cases}$$

$$\mathrm{Prob}[\{S_1,\dots,S_M\}] = \frac{1}{Z(u)} \left(\prod_{l=1}^{M} u^{S_l} \right) \left(\prod_{i=1}^{N} w_i(\underline{S}_i) \right)$$

$$w_i = \begin{cases} 1 & \text{if there are 0 or } k=2 \text{ edges with } S_l = 1 \text{ around vertex } i \\ 0 & \text{otherwise} \end{cases}$$

[There could be several vertex disjoint circuits, should not be relevant in the thermodynamic limit]

Not simpler to solve exactly, but opens the way to

- Monte Carlo evaluation [Klemm and Stadler]
- analytical approximations

Related problems :

- $k = 1$ (matchings) [Zhou and Ou-Yang, Zdeborová and Mézard]
- $k \geq 3$ [Pretti and Weigt]

The Bethe approximation / Belief Propagation

Variational formulation of canonical computations (Gibbs free-energy) :

$$Z = \sum_{\mathcal{C}} w(\mathcal{C}) \ , \qquad \ln Z = - \min_{p_{\rm v}} \sum_{\mathcal{C}} p_{\rm v}(\mathcal{C}) \ln \left(\frac{p_{\rm v}(\mathcal{C})}{w(\mathcal{C})} \right)$$

Trial distributions $p_{\rm v}$:

- factorized $\longrightarrow$ mean-field, exact bound
- with neighboor correlations $\longrightarrow$ Bethe approximation, exact on trees

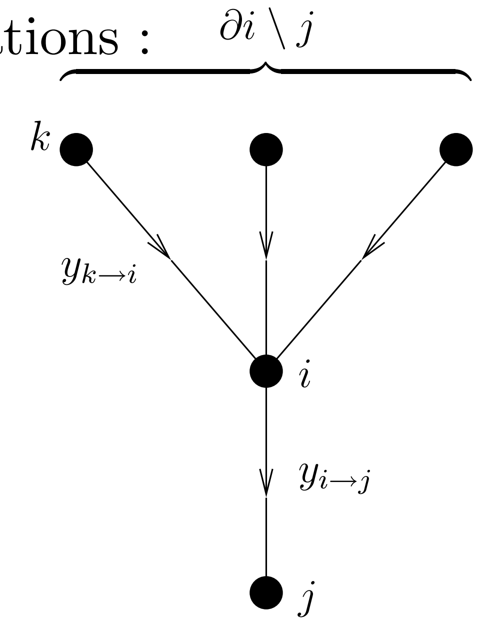
Minimizing Bethe free-energy $\Leftrightarrow$ finding fixed point of the (loopy)
Belief Propagation equations (message-passing algorithm)

[Yedidia et al.]

For the present model, directed messages $y_{i\to j}$, BP equations :

$$y_{i\to j} = \frac{u \sum_{k\in\partial i\setminus j} y_{k\to i}}{1+\frac{1}{2}u^2 \sum_{\substack{k,k'\in\partial i\setminus j\\ k\neq k'}} y_{k\to i}\ y_{k'\to i}}$$

$$p_l(1) = \frac{u y_{i\to j} y_{j\to i}}{1+u y_{i\to j} y_{j\to i}}$$



$p_l(1)$: fraction of circuits of length $N\ell(u)$ which go through $l = \langle ij \rangle$

$$\begin{cases} N\ell(u) = \sum_l p_l(1) \\ \sigma(\ell(u)) \text{ can also be computed from the messages } \{y_{i\to j}\} \end{cases}$$

An approximate enumeration algorithm on a given graph

- initialize messages $y_{i \to j}$ randomly
- iterate BP equations (for some u) until convergence
- compute $\ell(u)$ and $\sigma(\ell(u))$ from the fixed-point messages
- change u and do it again

$\Rightarrow$ parametric plot for the entropy of loops

$\Rightarrow$ yields also local informations (fraction of loops through one edge)

$\Rightarrow$ fast algorithm

Defects : convergence is not ensured, Bethe approximation is uncontrolled

Typical number of loops in random graph ensembles

Probability laws on the set of graphs with N vertices, examples :

- Erdös-Rényi (fixed number of edges/fixed probability of edges)
- fixed connectivity distribution (regular as particular case)
- growing networks

$\mathcal{N}_L = e^{N\sigma(\ell)}$ is a random variable, fluctuates from graph to graph in the ensemble

Statistical properties ?

$\mathcal{N}_L = e^{N\sigma(\ell)}$

« Annealed average » : $N\sigma_{\rm a}(\ell) = \ln \overline{\mathcal{N}_L}$

for arbitrary connectivity distribution : [Bianconi and Marsili]

« Quenched average » : $N\sigma_{\rm q}(\ell) = \overline{\ln \mathcal{N}_L}$

yields the « typical » value of σ :

$$\text{Prob}[e^{N[\sigma_{\rm q}(\ell)-\epsilon]} \le \mathcal{N}_L \le e^{N[\sigma_{\rm q}(\ell)+\epsilon]}] \to 1 \qquad \text{for } L/N \to \ell$$

In general, annealed averages are dominated by exponentially rare samples with exponentially more circuits ($\sigma_{\rm q} \le \sigma_{\rm a}$)

Annealed averages are very sensitive to « microscopic details » :
the two Erdös-Rényi ensembles have different annealed entropies

Simplest probabilistic proofs :

Compute the first two moments $\overline{\mathcal{N}_L}$ and $\overline{\mathcal{N}_L^2}$ (combinatorics)

If $\overline{\mathcal{N}_L^2} \sim \overline{\mathcal{N}_L}^2$, use Chebychev to prove $\sigma_{\mathrm{q}} = \sigma_{\mathrm{a}}$

But usually, $\overline{\mathcal{N}_L^2} \gg \overline{\mathcal{N}_L}^2$ ...

For the number of circuits :

$\overline{\mathcal{N}_L^2} \sim \overline{\mathcal{N}_L}^2$ only for regular random graphs,
in this case very detailed rigorous results

[Robinson and Wormald, Janson, Garmo]

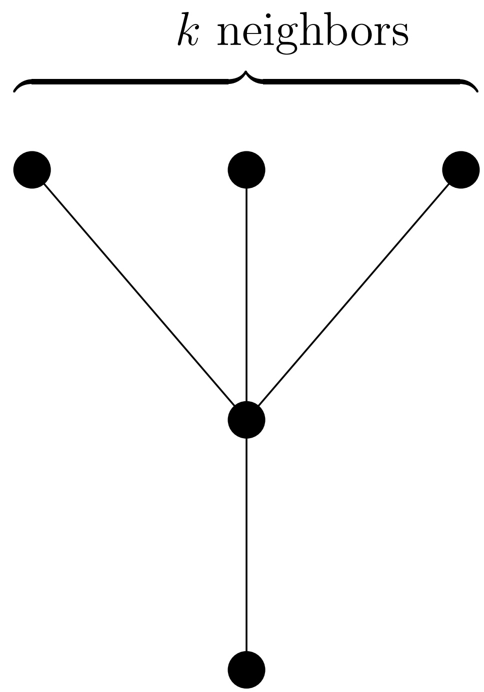
In all other ensembles, $\overline{\mathcal{N}_L^2} \gg \overline{\mathcal{N}_L}^2$, much less is known rigorously

The cavity computation for the quenched entropy

[Mézard and Parisi]

Connectivity distribution q_k (taking a site at random)

Mean degree $c = \sum_k k q_k$



« Offspring distribution » $\tilde{q}_k$
(taking an edge at random)

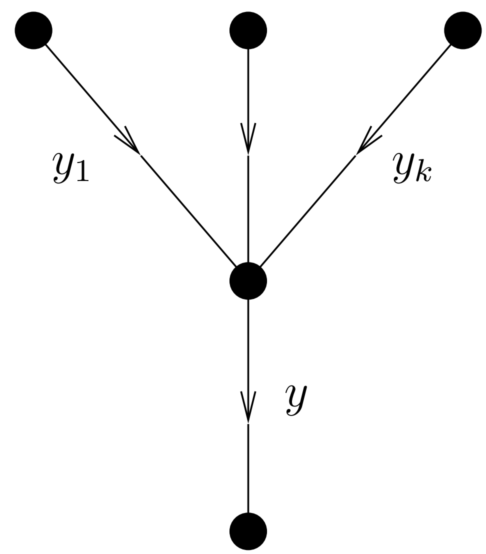
$$\tilde{q}_k = \frac{(k+1)q_{k+1}}{c}$$

BP messages $y_{i\to j}$ becomes random, with law $P(y)$

Self-consistent equation (with replica-symmetry assumption)

$$P(y) = \sum_{k=0}^{\infty} \tilde{q}_k \int dP(y_1)\dots dP(y_k)\ \delta(y - g_k(y_1,\dots,y_k))$$

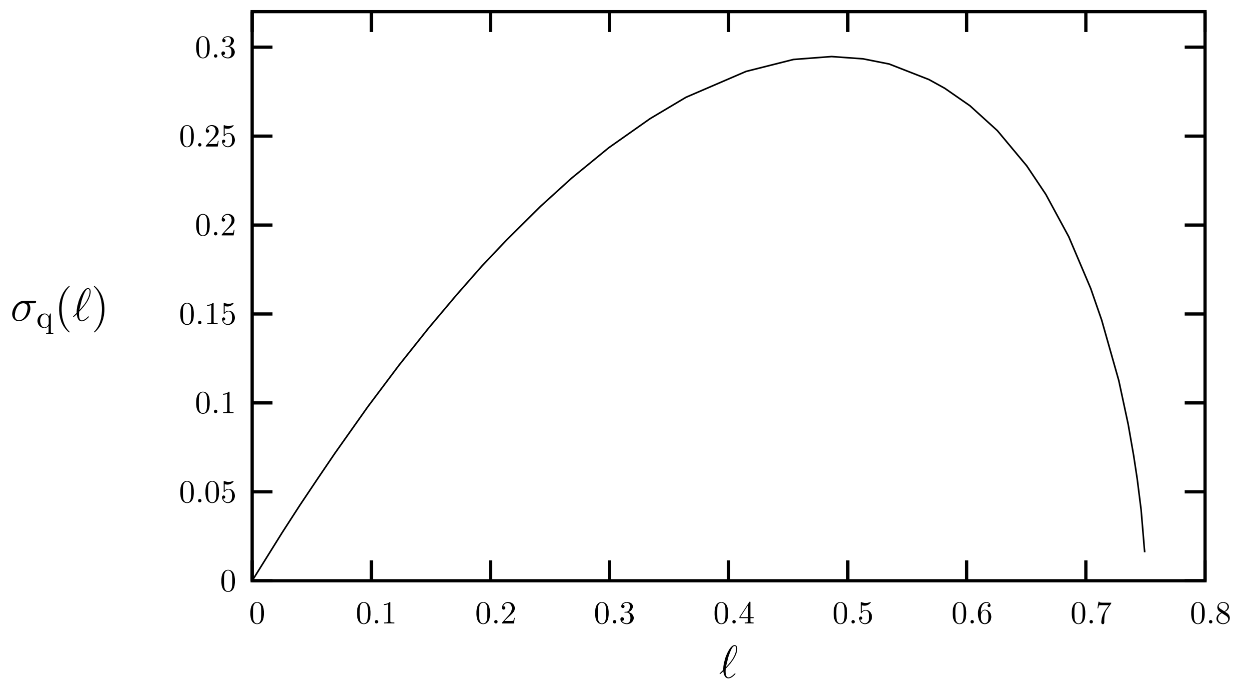
$$g_k(y_1,\dots,y_k) = \frac{u\ \sum_i y_i}{1 + u^2\ \sum_{i<j} y_i y_j}$$



From the solution $P(y)$, one finds $\ell(u)$ and $\sigma_{\mathrm{q}}(\ell(u))$

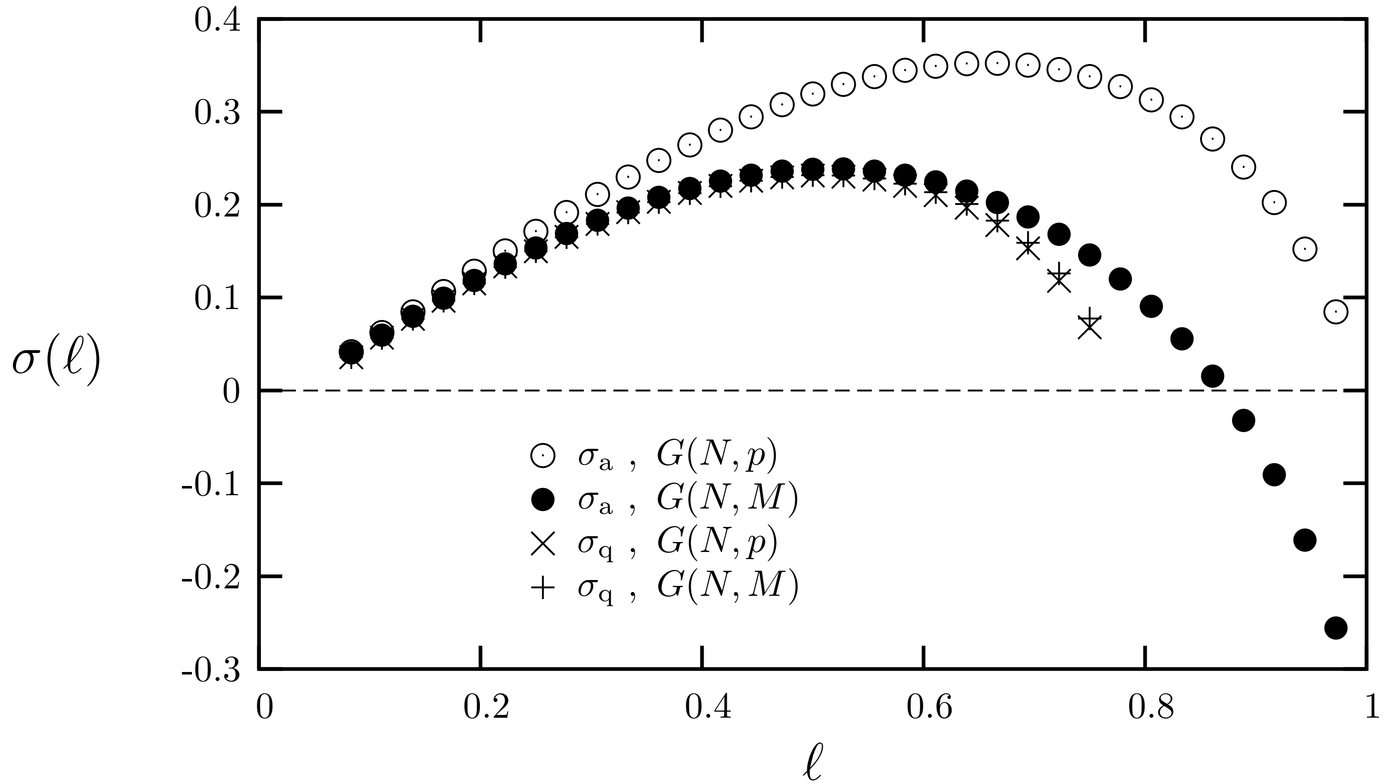
Distributional equation, easily solvable numerically by population dynamics algorithm

Example for Poisson random graphs with $c = 3$:

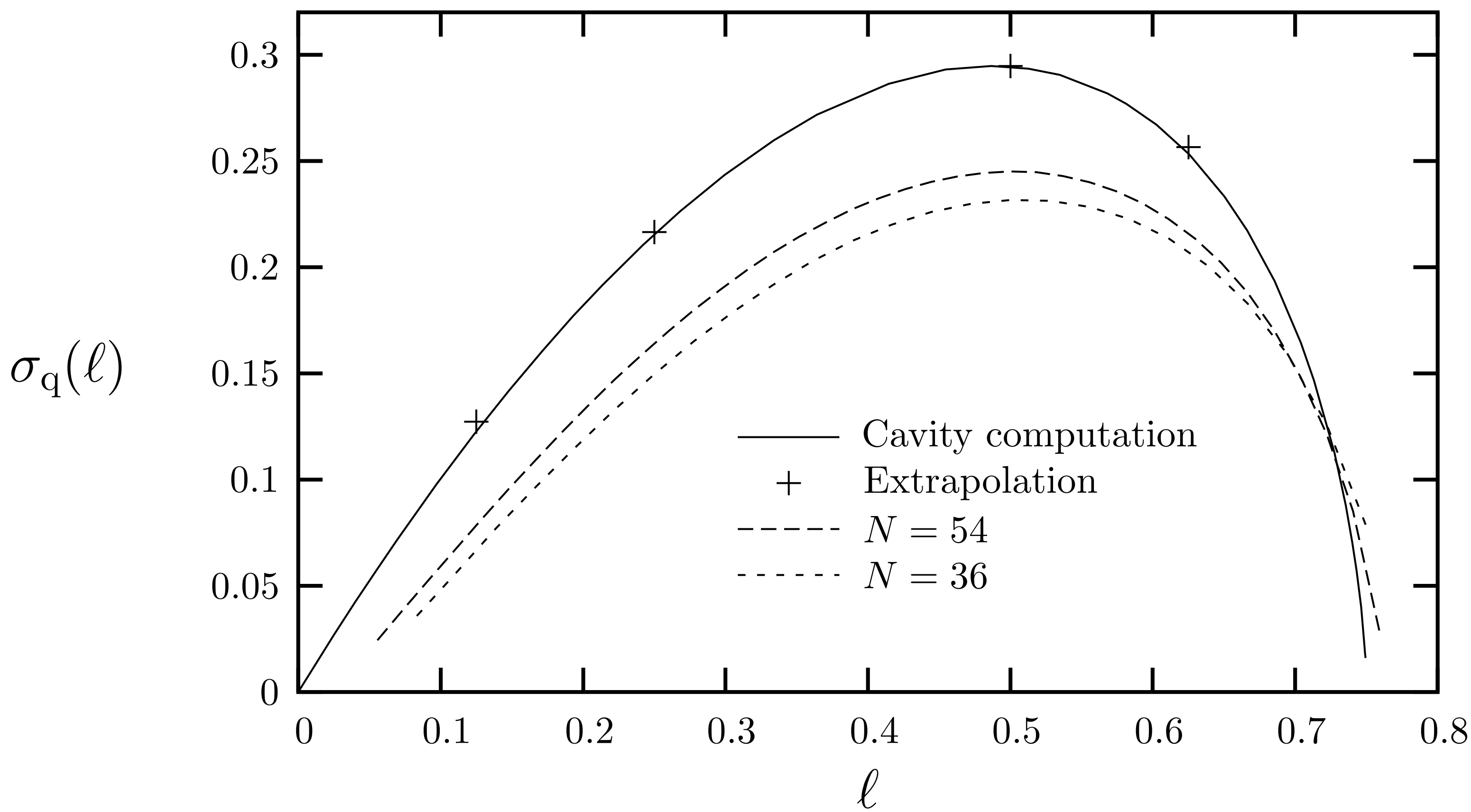


Comparison with exhaustive enumerations

(Erdös-Rényi ensembles)



$c = 3$, $N = 36$, quenched entropy estimated with the median



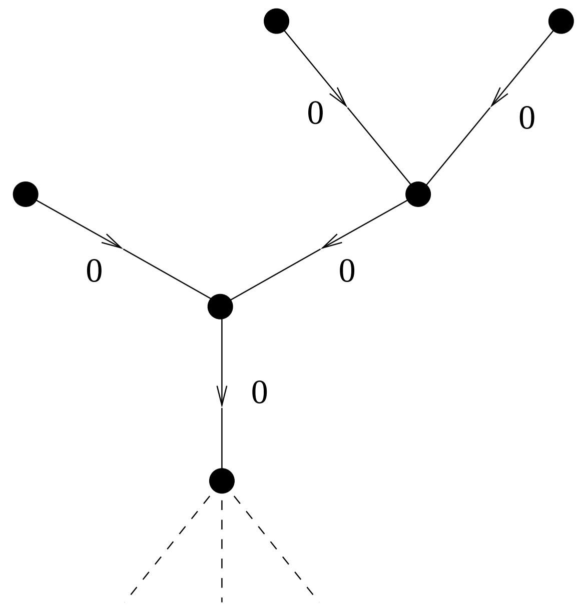
In $G(N, M)$, with $c = 3$

Some analytical predictions

In general $P(y)$ is known only numerically,

but some properties/limits can be investigated analytically

- Fraction of null messages yields the typical size of the 2-core for arbitrary connectivity distribution (confirmed by analysis of the leaf-removal algorithm)



- Short loops : expansion of $\sigma_{\rm q}(\ell)$ around $\ell = 0$

$$\underline{\text{First order :}}\ \sigma'_{\rm q}(0) = \ln\left(\frac{\sum_k k(k-1)q_k}{\sum_k kq_k}\right) = \sigma'_{\rm a}(0)$$

matches behaviour for $1 \ll L \ll \ln N$

$$\underline{\text{Second order :}}\ \sigma''_{\rm q}(0) \le \sigma''_{\rm a}(0)$$

with equality only in the regular case

- Longest loops (« zero-temperature » limit) :

  they have $L_{\max} = N\ell_{\max}$ edges (w.h.p.)

  – if minimal connectivity is 3 $(q_0 = q_1 = q_2 = 0)$

    the cavity computation yields $\boxed{\ell_{\max} = 1}$

    i.e. graphs in such ensemble are typically Hamiltonian

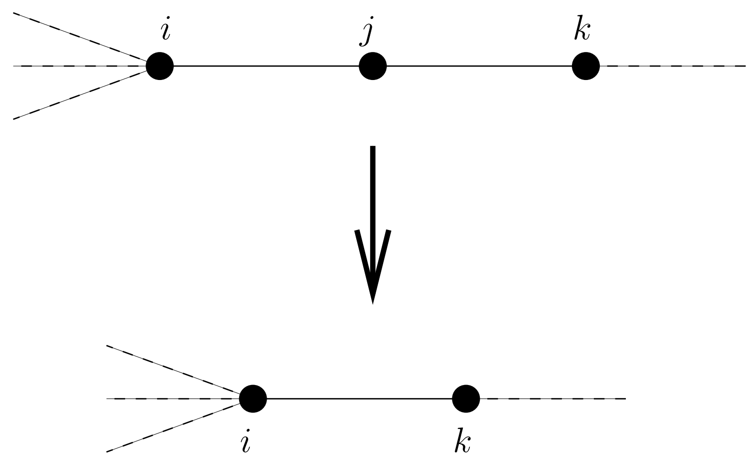
    was conjectured by Wormald

    statistical mechanics gives also $\sigma_{\mathrm{q}}(1)$

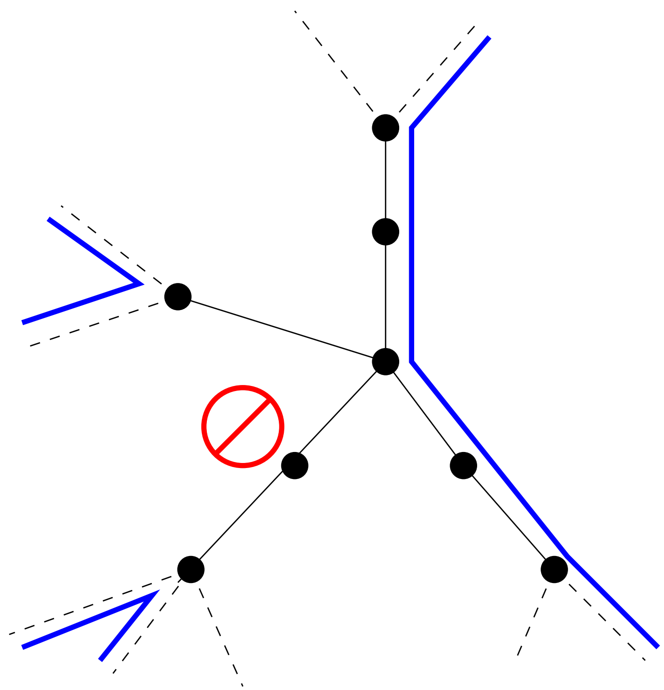
- Longest loops :
  - in the general case (no constraint on the minimal degree)

Bounds on $\ell_{\max}$:

- the 2-core contains $N\ell_{\text{core}}$ sites $\Rightarrow$ $\ell_{\max} \leq \ell_{\text{core}}$
- ℓ_{lb} of these sites have degree ≥ 3 $\Rightarrow$ $\ell_{\text{lb}} \leq \ell_{\max}$



Can the upperbound be saturated ?



No if there is a finite fraction of degree 2 sites in the 2-core

Conjecture : $\ell_{\max} = \ell_{\text{core}} - \sum_{k=3}^{\infty} q_k \binom{k}{3} \tilde{q}_1^3 + O(\tilde{q}_1^4)$

Confirmed by the small temperature expansion of the cavity results

Perspectives

- Rigorous proofs ? [Guerra et al, Aldous et al]
- Belief-inspired decimation algorithm to construct cycles
- Other random graph models, correlated networks, scale-free graphs (not Hamiltonian even for $k_{\min} = 3$) [Bianconi and Marsili]
- Large deviations from the typical case [Rivoire]
- Corrections to the Bethe approximation [Montanari and Rizzo, Parisi and Slanina, Chertkov and Chernyak]