A cluster expansion approach to exponential random graph models

Mei Yin

Department of Mathematics
University of Texas at Austin

December 20, 2011

Dynamical Gibbs-non-Gibbs Transitions Workshop, EURANDOM

The exponential family of random graphs is among the most widely-studied of network models. A host of analytical and numerical techniques have been developed in the past. We show that any exponential random graph model could be alternatively viewed as a lattice gas model with a finite Banach space norm. The system could then be treated by cluster expansion methods in statistical mechanics. In particular, we derive a convergent power series expansion for the limiting free energy in the case of small parameters. This hopefully would help with the application of renormalization group ideas to exponential random graph models.

Acknowledgments

I am grateful to Charles Radin, Persi Diaconis and Sourav Chatterjee for introducing me to the exciting subject of random graphs and for their many enlightening and encouraging comments.

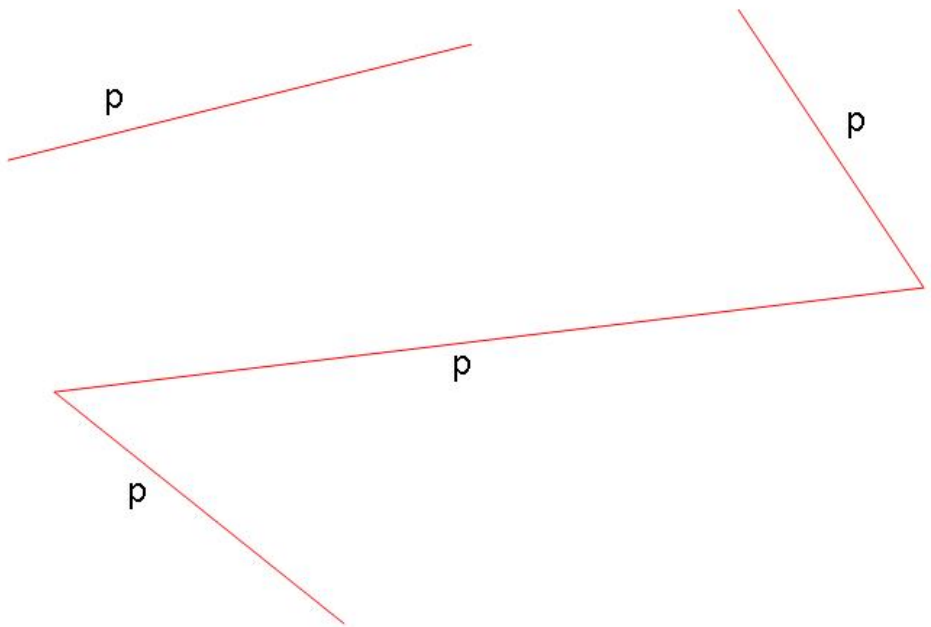
Outline

- Introduction and Background
- Framework and Notation
- Alternative View
- Cluster Expansion

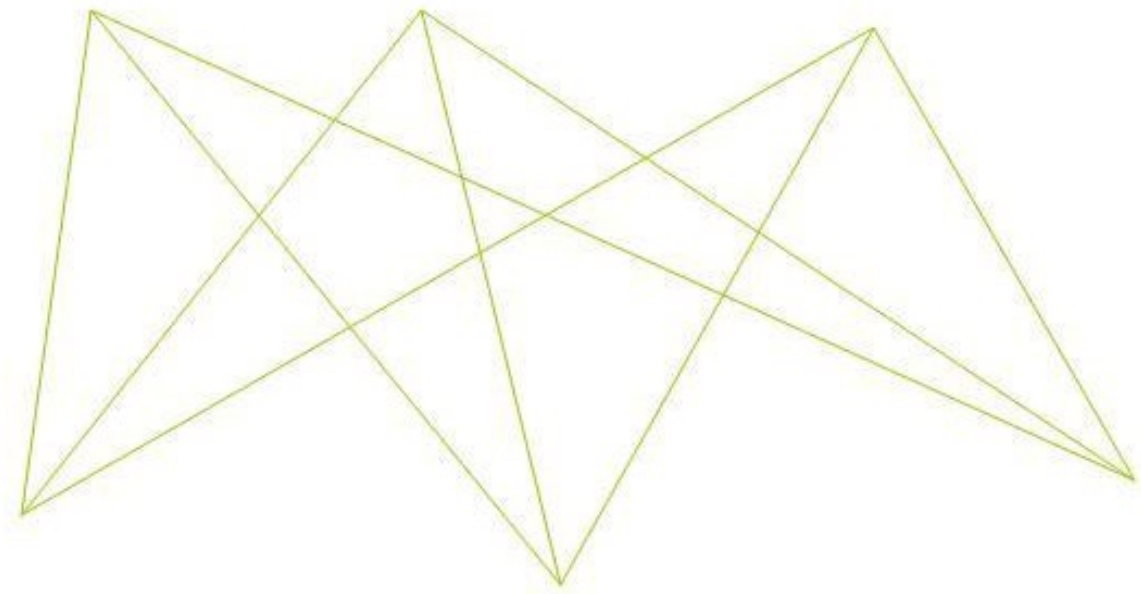
Pioneering work on the independent case: Erdős-Rényi graph $G(n, p)$,

$$\mathbb{P}_n^{\beta}(G) = e^{\beta E(G) - \psi_n} = p^{E(G)}(1-p)^{\binom{n}{2} - E(G)}.$$

Include edges independently with paramter $p = e^{\beta}/(1+e^{\beta})$.



Extremal Graph Theory (Turán)—Maximize number of edges without triangles. Unique solution: complete bipartite graph with equal parts.



More general statements...

Exponential random graph: Dependence between the random edges is defined through certain finite subgraphs, in imitation of the use of potential energy to provide dependence between particle states in a grand canonical ensemble of statistical physics. By varying the activity parameters, one could analyze the extent to which specific values of the subgraph densities interfere with one another. Estimation can be based on construction of a Markov chain that has the exponential random graph model as equilibrium distribution. Large deviation principle comes into play.

- Holland and Leinhardt studied the directed case.
- Frank and Strauss related random graph edges to Markov random field.
- Häggström and Jonasson examined phase transition in the random triangle model.
- More developments: Wasserman and Faust, Snijders et al., Rinaldo et al.
- Recent survey: Fienberg, Introduction to papers on the modeling and analysis of network data I & II. arxiv: 1010.3882 & 1011.1717.

Relevance to Gibbs measures:

- Ising model on complete graph: Curie-Weiss model. (Ellis and Newman, The statistics of Curie-Weiss models.)
- Ising model on sparse graph: No finite-dimensional structure. Distance between vertices. Phase transitions and coexistence phenomena are related to Gibbs measures on infinite trees. (Dembo and Montanari, Gibbs measures and phase transitions on sparse random graphs.)
- Ising model on lattice: Disordered limiting Gibbs state (with zero effective field) is pure up to the spin-glass critical temperature. (Bleher et al., On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice.)

- Graph homomorphism $\hom(H, G)$ is an edge-preserving map. Examples: H triangle, G triangle, $|\hom(H, G)| = 6$; H 2-star, G triangle, $|\hom(H, G)| = 12$.
- Homomorphism density $t(H, G) = \frac{|\hom(H,G)|}{|V(G)|^{|V(H)|}}$.
- $\beta_1, ..., \beta_k$ are k real parameters. $H_1, ..., H_k$ are finite simple graphs. Each H_i has m_i vertices $(2 \leq m_i \leq m)$ and p_i edges $(1 \leq p_i \leq p)$. In particular, H_1 is the complete graph on 2 vertices (i.e., a single edge).
- $\mathcal{G}_n$ is the set of simple graphs G on n vertices. Probability for $G \in \mathcal{G}_n$ is given by:

$$\mathbb{P}_n^{\{\beta_i\}}(G) = e^{n^2(\beta_1 t(H_1,G)+...+\beta_k t(H_k,G)-\psi_n)} := e^{n^2(T(G)-\psi_n)}.$$

- ψ_n is the normalization constant:

$$\psi_n = \frac{1}{n^2} \log \sum_{G \in \mathcal{G}_n} e^{n^2 T(G)}.$$

$\lim \psi_n$ is crucial for carrying out maximum likelihood and Bayesian inference.

- Park and Newman tried the technique of mean-field approximations.
- Monte Carlo schemes: Geyer and Thompson (MCMLE), Gelman and Meng (bridge sampling), Kou et al. (equi-energy sampler)
- More approaches: Besag, Comets and Janžura, Chatterjee, Snijders

Consider the space $\mathcal{W}$ of all symmetric measurable functions from $[0,1]^2$ into $[0,1]$. For $H \in \mathcal{G}_k$, let

$$t(H,h) = \int_{[0,1]^k} \prod_{(i,j)\in E(H)} h(x_i, x_j) dx_1 ... dx_k.$$

Lovász et al. developed graph limits (graphons): A sequence of graphs $\{G_n\}_{n\geq 1}$ is said to converge to h if for every finite simple graph H,

$$\lim t(H, G_n) = t(H, h).$$

Example: Erdős-Rényi graph $G(n,p)$, $h(x,y) = p$.

- Chatterjee and Diaconis gave the first rigorous proof of singular behavior in a specific exponential random graph model, the edge-triangle model. They also suggested that, quite generally, models with repulsion exhibit a transition qualitatively like the solid/fluid transition. (Estimating and understanding exponential random graph models. arXiv: 1102.2650.)
- Radin and Y derived the full phase diagram for a large family of 2-parameter exponential random graph models, each containing a first order transition curve ending in a second order critical point, qualitatively similar to the gas/liquid transition in equilibrium materials. (Phase transitions in exponential random graphs. arXiv: 1108.0649.)

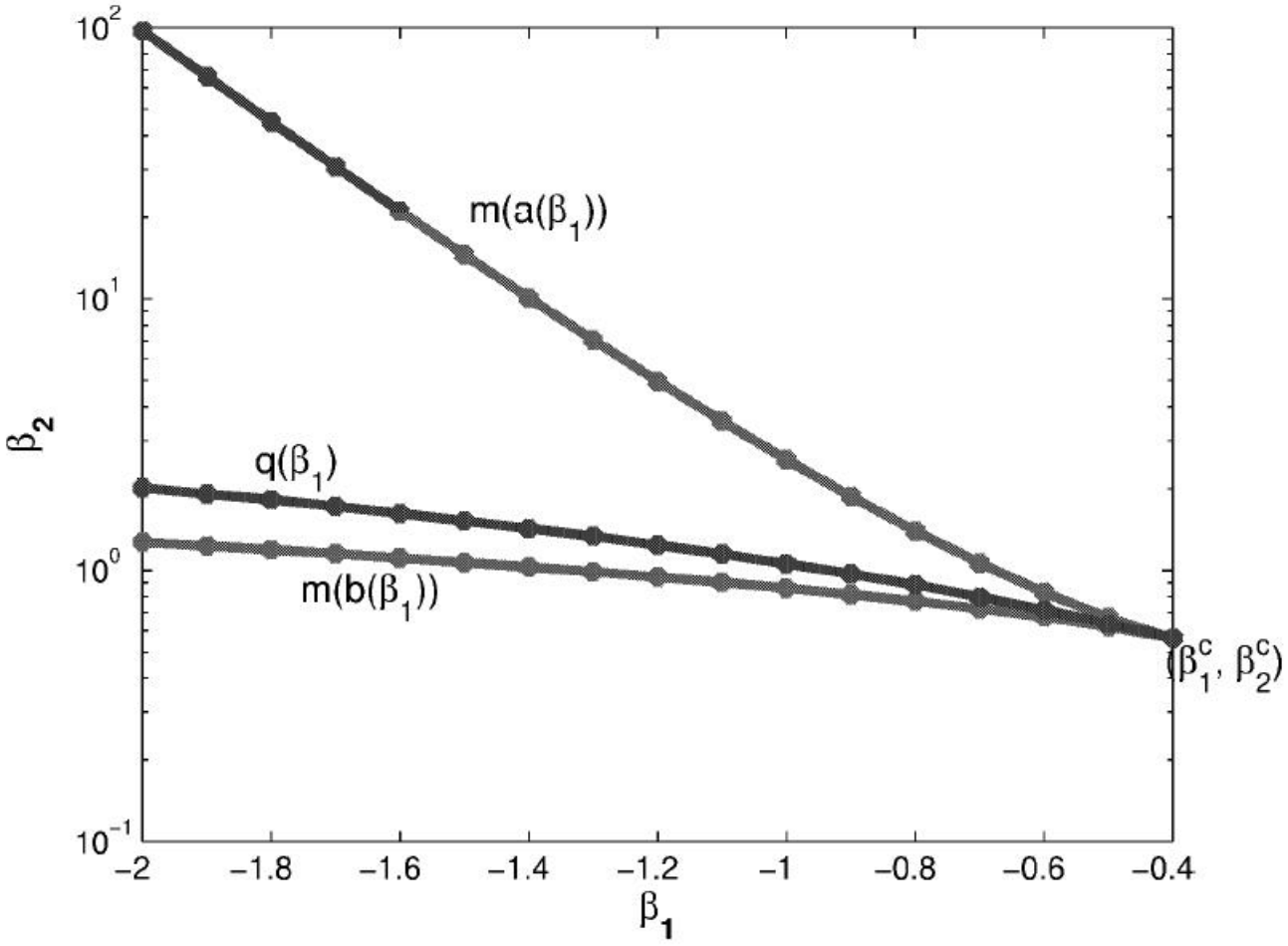
Optimization problem: Suppose $\beta_2, ..., \beta_k$ are nonnegative. Then

$$\lim \psi_n = \sup_{0 \le u \le 1} \left(\beta_1 u^{E(H_1)} + ... + \beta_k u^{E(H_k)} - \frac{1}{2} u \log u - \frac{1}{2}(1-u)\log(1-u) \right). \tag{2.1}$$

Behavior of $G \in \mathcal{G}_n$:

$$\min_{u \in U} \delta_{\square}(\tilde{G}_n, \tilde{u}) \to 0 \text{ in probability as } n \to \infty,$$

where U is the set of maximizers of (2.1).



The phase transition curve $\beta_2 = q(\beta_1)$ in the (β_1, β_2) plane. H_1 is a single edge and H_2 has 3 edges.

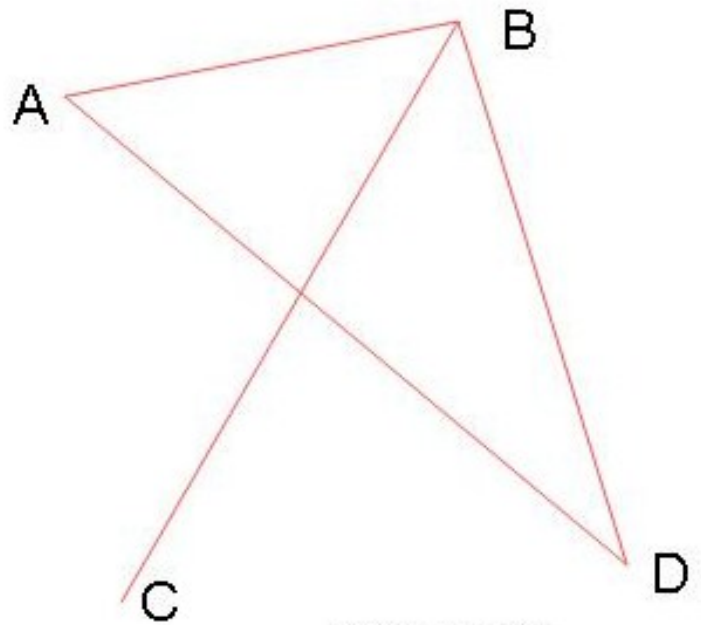
Fix $H \in \mathcal{G}_m$. Let $G \in \mathcal{G}_n$.
Proposition: The homomorphism density $t(H, G)$ has a lattice gas representation $\sum_X J(X)\sigma_X$.

- $\sigma_{ij} = \sigma_{ji}$ is an element of the adjacency matrix of G.
- X is any set of vertex pairs (i, j) of G.
- $\sigma_X = \prod_{(i,j)\in X} \sigma_{ij}$.

Proof: Number of possible (connected) image shapes of H in G under the graph homomorphism is finite. Consider such an image shape Y. Denote the corresponding homomorphism density by $t_Y(H, G)$. Define $J(X) = t_Y(H, G)$ for any X whose relative vertex positions are the same as in Y. This map becomes a homomorphism only when $\sigma_X = 1$, i.e., all corresponding edges between vertices in X exist.

Each edge, (A, B), (A, D), (B, C), (B, D), carries weight $2/4^3$.

Graph H



Graph G

Each 2-star, (A, B, C), (A, B, D), (C, B, D), (B, A, D), (A, D, B), carries weight $2/4^3$.

Proposition: Fix a vertex pair (i, j), denote by $t_{ij}(H, G)$ the part of the homomorphism density $t(H, G)$ that depends on σ_{ij}, we have

$$t_{ij}(H, G) = \sum_{X:(i,j)\in X} J(X) \leq \frac{m(m-1)}{n^2}.$$

Note: Sharp bound. Example: H and G are both a single edge.
Proof: The image of $V(H)$ in $V(G)$ consists of vertices i and j of G. To count these homomorphisms, we regard such a mapping as consisting of two steps. Step 1: We choose the vertices of G that the vertices of H are mapped onto. Step 2: We check whether these vertex-maps are valid homomorphisms (i.e., edge-preserving).

Graphs in the same exponential random graph family correspond to equilibrium ensembles.
Hamiltonian:

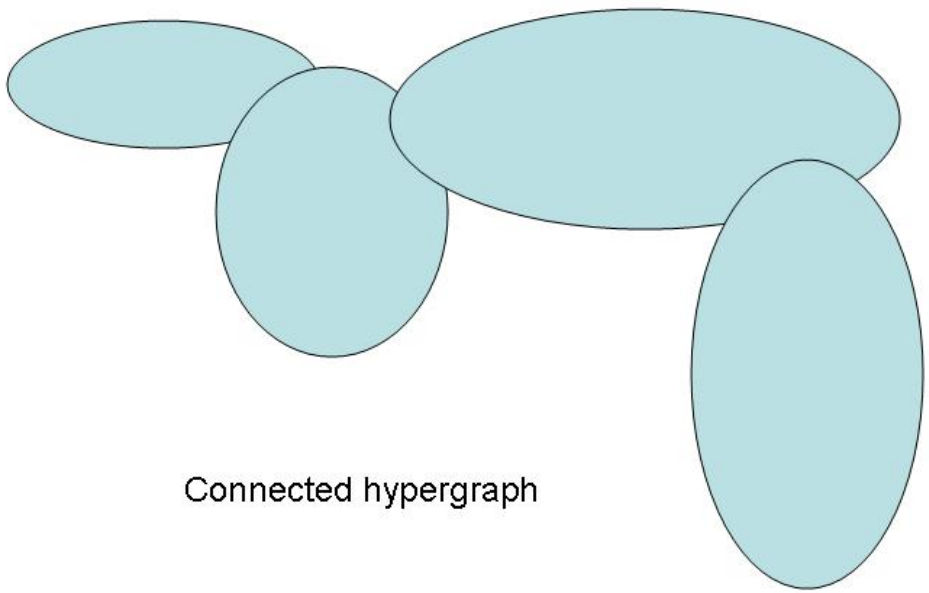
$$H(\sigma) = -n^2 \sum_{i=1}^{k} \beta_i J_i(X) \sigma_X = -\sum_X K(X) \sigma_X$$

Note: $K(X) = 0$ for $|X| > p$.
Banach space:

$$||K|| = \sup_{(i,j)} \sum_{X:(i,j)\in X} |K(X)| \leq m(m-1) \sum_{i=1}^{k} |\beta_i|.$$

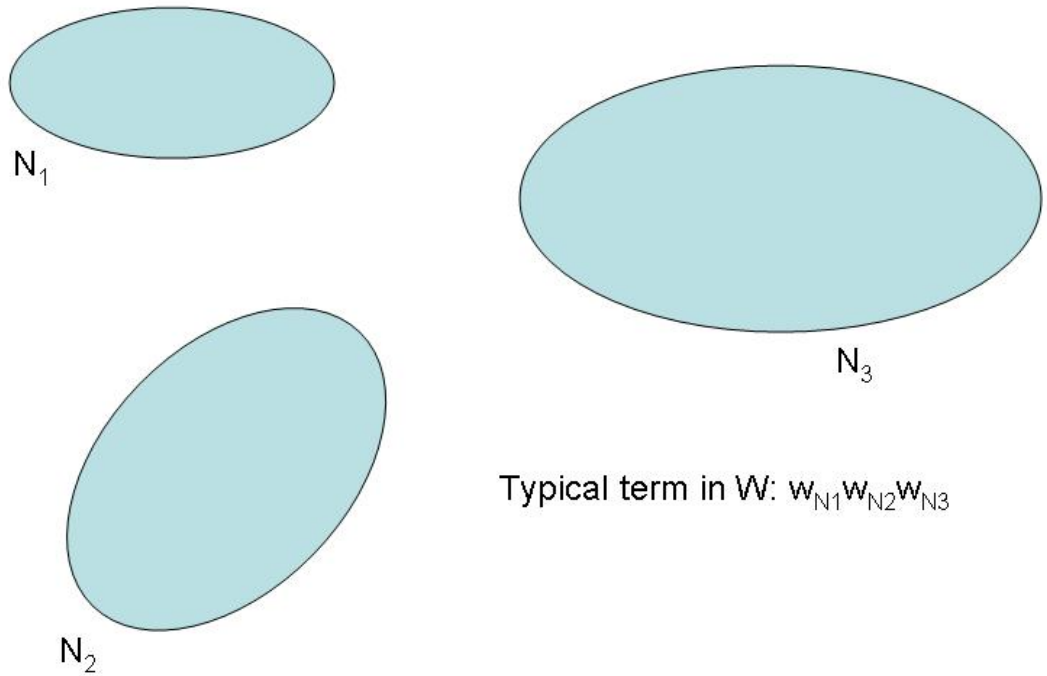
The limiting free energy (random graph model) ψ and the limiting free energy (lattice gas model) ϕ are related by $\psi = \frac{1}{2}(\log 2 + \phi)$.

- Hypergraph $\Gamma = (X, E)$.
- X is a set of sites, E (hyper-edge or link) is a set of nonempty subsets of X.
- Two links are connected if they overlap.
- Support of hypergraph: $\cup\Gamma$.
- Connected hypergraph Γ_c: $\cup\Gamma$ is nonempty and cannot be partitioned into nonempty sets with no connected links.


Connected hypergraph

$W = \sum_\sigma e^{-H(\sigma)} = \sum_\sigma e^{\sum_X K(X)\sigma_X}$. Let V be the set of all vertex pairs (i,j) of $G \in \mathcal{G}_n$.
Cluster representation for W: $W = \sum_\Delta \prod_{N\in\Delta} w_N$.

- Δ is a set of disjoint subsets N's of V.
- $w_N = \sum_{\cup\Gamma_c = N} \sum_\sigma \prod_{X\in\Gamma_c} (e^{K(X)\sigma_X} - 1)$.
- $|w_N| \le v_N = \sum_{\cup\Gamma_c = N} \prod_{X\in\Gamma_c} (e^{|K(X)|} - 1)$.



Typical term in W: $w_{N1}w_{N2}w_{N3}$

Proof: We rewrite $e^{\sum_X K(X)\sigma_X}$ as a perturbation around zero interaction:

$$W = \sum_{\sigma} \sum_{\Gamma} \prod_{X \in \Gamma} \left(e^{K(X)\sigma_X} - 1 \right).$$

The support of each hypergraph Γ on V would break up into connected parts Δ. We have

$$W = \sum_{\Delta} \prod_{N \in \Delta} \sum_{\cup \Gamma_c = N} \sum_{\sigma} \prod_{X \in \Gamma_c} \left(e^{K(X)\sigma_X} - 1 \right).$$

To apply standard results on cluster expansion, notice that

$$\sum_{\Delta} \prod_{N \in \Delta} w_N = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{N_1,\ldots,N_n} \prod_{(i,j)} (1 - c(N_i, N_j))\, w_{N_1} \cdots w_{N_n}$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{N_1,\ldots,N_n} \sum_{G} \prod_{(i,j) \in G} (-c(N_i, N_j))\, w_{N_1} \cdots w_{N_n},$$

where $G \in \mathcal{G}_n$ and

$$c(N_i, N_j) = \begin{cases} 1 & \text{if } N_i \text{ and } N_j \text{ overlap;} \\ 0 & \text{otherwise.} \end{cases}$$

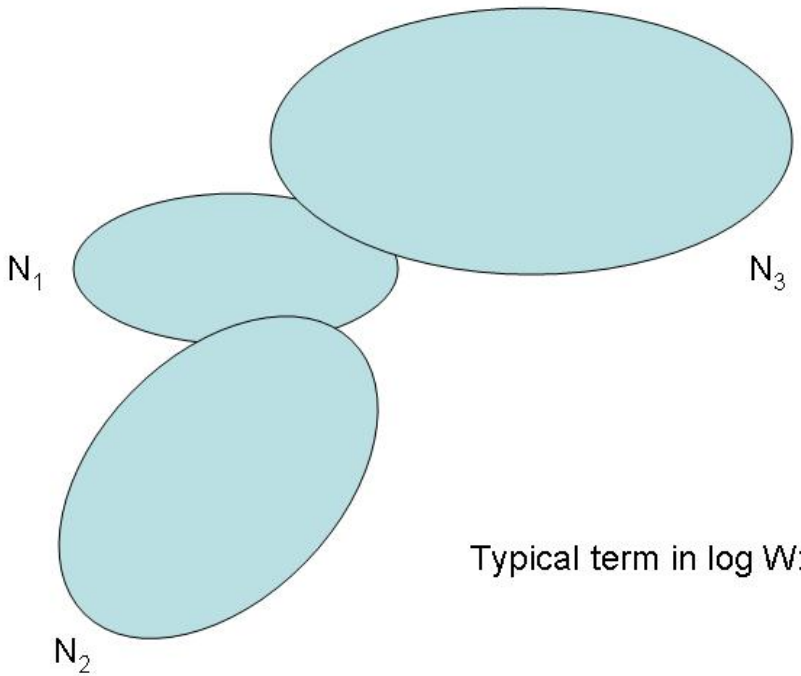
Cluster representation for $\log W$:

$$\log W = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{N_1,\ldots,N_n} C\left(N_1,\ldots,N_n\right) w_{N_1} \cdots w_{N_n},$$

where

$$C\left(N_1,\ldots,N_n\right) = \sum_{G_c} \prod_{(i,j)\in G_c} \left(-c(N_i,N_j)\right),$$

and $G_c \in \mathcal{G}_n$ is a connected graph.
Proof: The effect of taking the logarithm is that the sum over graphs is replaced by the sum over connected graphs.



Typical term in log W: $w_{N1}w_{N2}w_{N3}$

Kotecký-Preiss: Fix $M > 1$. Suppose that for each vertex pair (i, j), we have

$$\sum_{N:(i,j)\in N} v_N M^{|N|} \leq \log M. \tag{4.1}$$

Then the pinned free energy has a convergent power series expansion:

$$\sum_{n=1}^{\infty} \frac{1}{n!} \sum_{N_1,\ldots,N_n:\exists i N_i = N} |C(N_1, \ldots, N_n)| \, |w_{N_1}| \cdots |w_{N_n}| \leq v_N M^{|N|}.$$

How is K-P applicable here? Consider the coupling constants K with the Banach space norm $||K||$. Suppose $\sum_{i=1}^{k} |\beta_i|$ is small:

$$||K|| \leq m(m-1) \sum_{i=1}^{k} |\beta_i| \leq \frac{\log M(p-1)^p}{2(Mp)^p \left(1 + (p-1)\log M\right)}.$$

Then (4.1) holds for every vertex pair (i, j).
The maximal region of parameters $\{\beta_i\}$ is obtained by setting

$$\log M = \frac{-p + \sqrt{5p^2 - 4p}}{2p(p-1)}.$$

Main result: Fix $M > 1$. Consider the coupling constants K with the Banach space norm $||K||$. Suppose $\sum_{i=1}^{k} |\beta_i|$ is small. Then we have convergence of the cluster expansion for the limiting free energy $\phi = \lim_{V \to \infty} \frac{1}{|V|} \log W$.

$$
\begin{aligned}
|\log W| &\leq \sum_{N \subset V} \sum_{n=1}^{\infty} \sum_{N_1, \ldots, N_n : \exists i N_i = N} \frac{1}{n!} |C(N_1, \ldots, N_n)| \, |w_{N_1}| \cdots |w_{N_n}| \\
&\leq \sum_{N \subset V} v_N M^{|N|} \leq \sum_{(i,j) \in V} \sum_{N : (i,j) \in N} v_N M^{|N|} \leq |V| \log M.
\end{aligned}
$$

Proof: We notice that when $|K||$ is small (say $||K|| \leq \frac{1}{2}$), $e^{|K(X)|} - 1 \leq 2|K(X)|$ by the mean value theorem. For $\cup \Gamma_c = N$, $|N| \leq \sum |X|$ with X in Γ_c. We have

$$\begin{aligned} \sum_{N:(i,j)\in N} v_N M^{|N|} &\leq \sum_{N:(i,j)\in N} \sum_{\cup \Gamma_c = N} M^{|N|} \prod_{X\in\Gamma_c} 2|K(X)| \\ &\leq \sum_{\Gamma_c:(i,j)\in\cup\Gamma_c} \prod_{X\in\Gamma_c} 2|K(X)| M^{|X|}. \end{aligned}$$

A hypergraph Γ_c is rooted at the vertex pair (i,j) if $(i,j) \in \cup\Gamma_c$. Let $a_n(ij)$ be the contribution of all connected hypergraphs with n links that are rooted at (i,j),

$$a_n(ij) = \sum_{(i,j)\in\cup\Gamma_c : |\Gamma_c| = n} \prod_{X\in\Gamma_c} 2|K(X)|M^{|X|}. \tag{4.2}$$

Then

$$\sum_{N:(i,j)\in N} v_N M^{|N|} \leq \sum_{n=1}^{\infty} \sup_{(i,j)\in V} a_n(ij) := \sum_{n=1}^{\infty} a_n.$$

To be continued...

Lemma: Let a_n be the supremum over (i,j) of the contribution of connected hypergraphs with n links that are rooted at (i,j). Then a_n satisfies the recursive bound

$$a_n \leq 2||K||M^p \sum_{k=0}^{p} \binom{p}{k} \sum_{a_{n_1},\ldots,a_{n_k}:n_1+\cdots+n_k+1=n} a_{n_1} \cdots a_{n_k}$$

for $n \geq 1$, where $\binom{p}{k}$ is the binomial coefficient.

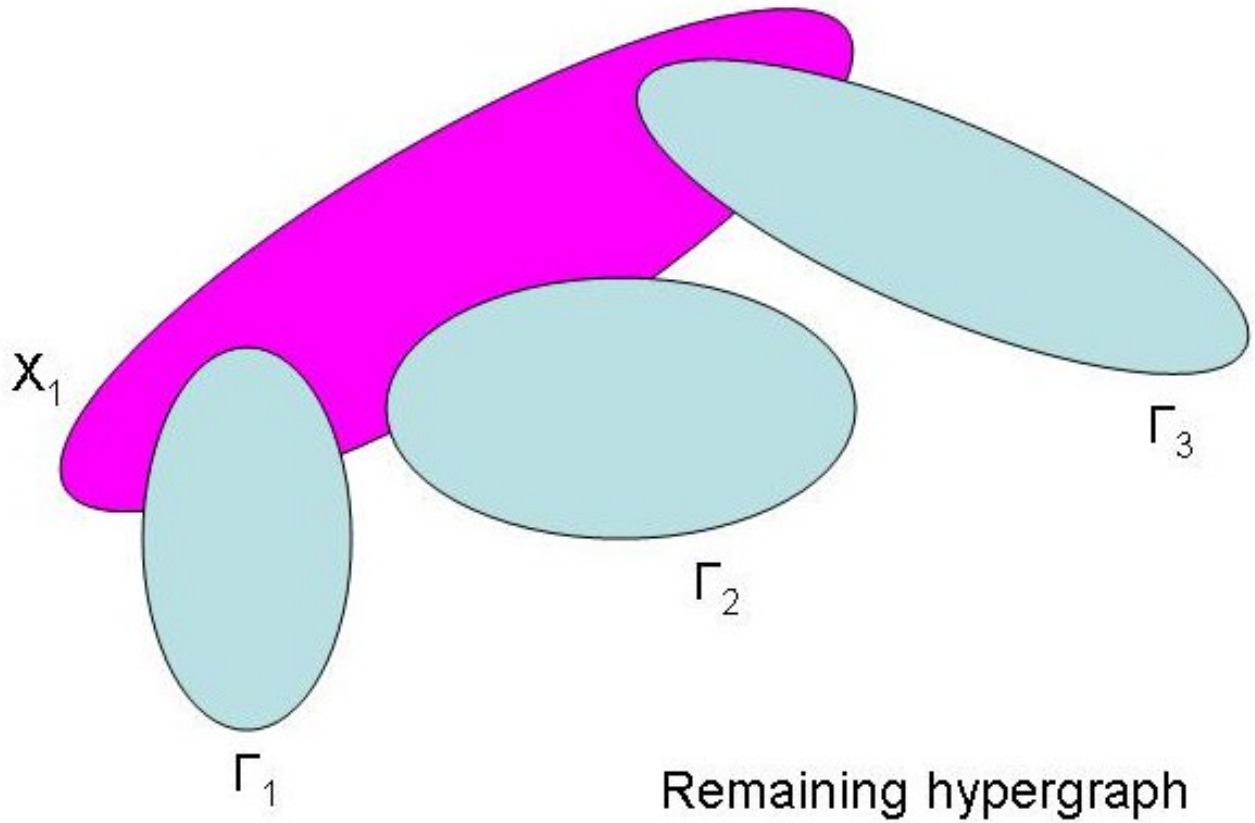
Proof: We linearly order the vertex pairs (i,j) in V and also linearly order the subsets X of V. For a fixed but arbitrarily chosen (i,j) in V, we examine (4.2). Write $\Gamma_c = \{X_1\} \cup \Gamma_c^1$, where X_1 is the least X in Γ_c with $(i,j) \in X_1$. Then

$$a_n(ij) \leq 2||K||M^p \sum_{\Gamma_c^1} \prod_{X \in \Gamma_c^1} 2|K(X)|M^{|X|}.$$

The remaining hypergraph Γ_c^1 has $n-1$ subsets and breaks into $k : k \leq p$ connected components $\Gamma_1, ..., \Gamma_k$ of sizes $n_1, ..., n_k$, with $n_1 + \cdots + n_k = n-1$.

$$a_n(ij) \leq 2||K||M^p \sum_{k=0}^{p} \binom{p}{k} \sum_{a_{n_1},...,a_{n_k} : n_1+\cdots+n_k+1=n} a_{n_1} \cdots a_{n_k}.$$

Finally, take the supremum over all (i,j) in V.


X_1
Γ_1
Γ_2
Γ_3
Remaining hypergraph

Clearly, $\sum_{N:(i,j)\in N} v_N M^{|N|}$ will be bounded above by $\sum_{n=1}^{\infty} \bar{a}_n$, if

$$\bar{a}_n = 2||K||M^p \sum_{k=0}^{p} \binom{p}{k} \sum_{\bar{a}_{n_1},\ldots,\bar{a}_{n_k}:n_1+\cdots+n_k+1=n} \bar{a}_{n_1} \cdots \bar{a}_{n_k} \qquad (4.3)$$

for $n \geq 1$.

Lemma: Consider the coefficients $\bar{a}_n$ that bound the contributions of connected and rooted hypergraphs with n links. Let $w = \sum_{n=1}^{\infty} \bar{a}_n z^n$ be the generating function of these coefficients. The recursion relation (4.3) for the coefficients is equivalent to the formal power series generating function identity

$$w = 2||K||M^p z(1+w)^p. \tag{4.4}$$

Proof: $(1+w)^p = \sum_{k=0}^{p} \binom{p}{k} w^k$, thus

$$w = 2||K||M^p z \sum_{k=0}^{p} \binom{p}{k} w^k.$$

Writing completely in terms of z,

$$\sum_{n=1}^{\infty} \bar{a}_n z^n = 2||K||M^p \sum_{k=0}^{p} \binom{p}{k} \sum_{\bar{a}_{n_1},\ldots,\bar{a}_{n_k} : n_1+\cdots+n_k+1=n} \bar{a}_{n_1} \cdots \bar{a}_{n_k} z^n.$$

Compare term-by-term.

Lemma: If w is given as a function of z as a formal power series by the generating function identity (4.4), then this power series has a nonzero radius of convergence $|z| \leq \frac{(p-1)^{p-1}}{2||K||(Mp)^p}$.

Proof: Without loss of generality, assume $z \geq 0$. Set $z_1 = 2||K||M^p z$. Solving (4.4) for z_1 gives

$$z_1 = \frac{w}{(1+w)^p}.$$

As z_1 goes from 0 to $(p-1)^{p-1}/p^p$, the w values range from 0 to $1/(p-1)$.

Going back...
We notice that in the above lemma, $w = \sum_{n=1}^{\infty} \bar{a}_n z^n = 1/(p-1)$ corresponds to $z_1 = 2||K||M^p z = (p-1)^{p-1}/p^p$, which implies that for each n,

$$\bar{a}_n \leq (2||K||(Mp)^p)^n (p-1)^{-(1+(p-1)n)}.$$

Gathering all the information we have obtained so far,

$$\begin{aligned} \sum_{N:(i,j)\in N} v_N M^{|N|} &\leq \sum_{n=1}^{\infty} (2||K||(Mp)^p)^n (p-1)^{-(1+(p-1)n)} \\ &= \frac{\frac{2||K||(Mp)^p}{(p-1)^p}}{1 - \frac{2||K||(Mp)^p}{(p-1)^{p-1}}} \leq \log M. \end{aligned}$$

Thank You!