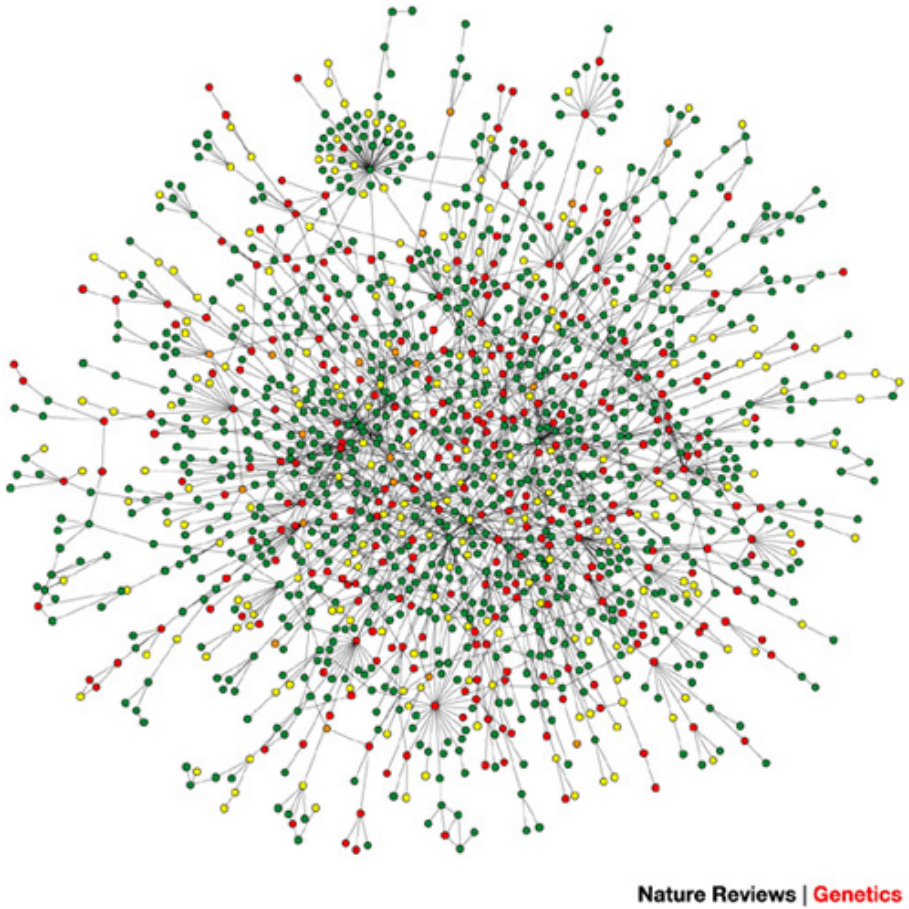
The partial duplication random graph

Peter Pfaffelhuber

with Felix Hermann

Munich, July 2014

Yeast protein-protein interaction network



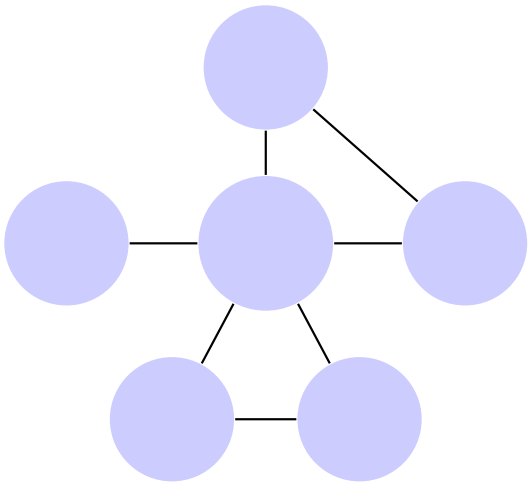
from Barabsi & Oltvai, *Network biology: understanding the cell's functional organization*, 2004

The model from Pastor-Sartorras, Chung, Bebek,...

- **Protein-protein interactions:** key for biological functions of cells
- **Evolution:** proteins are copied, their function partially retained, new interactions can emerge
- **Model:** copy vertex with all edges, keep each edge with probability p, insert additional r edges at random
- *With the right selection of parameters q and r, the general duplication model well approximates the degree distribution of the yeast proteome network.* (Bebek et al 2006)

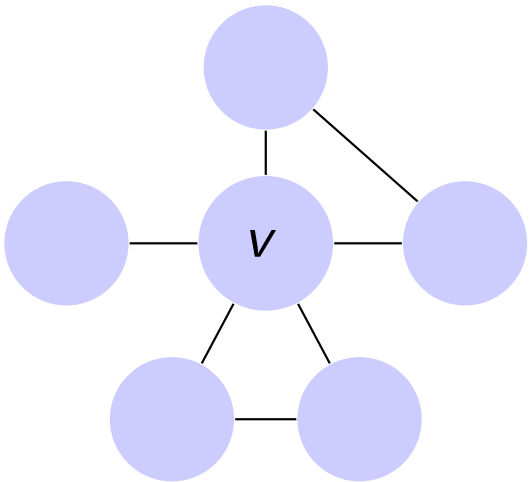
- In the following, we analyse the model for $q = 0$

The partial duplication random graph



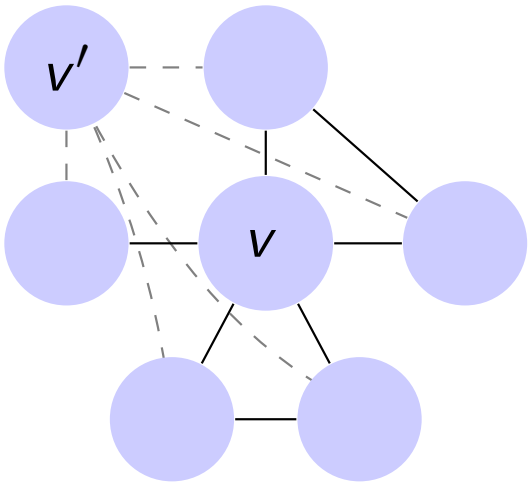
Start with some graph

The partial duplication random graph



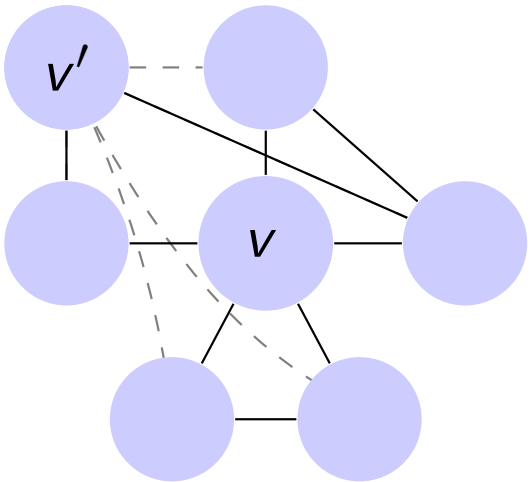
Pick a vertex purely at random

The partial duplication random graph



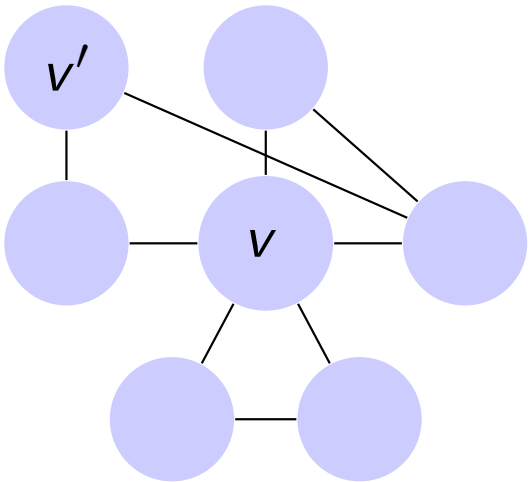
Copy the vertex together with all edges

The partial duplication random graph



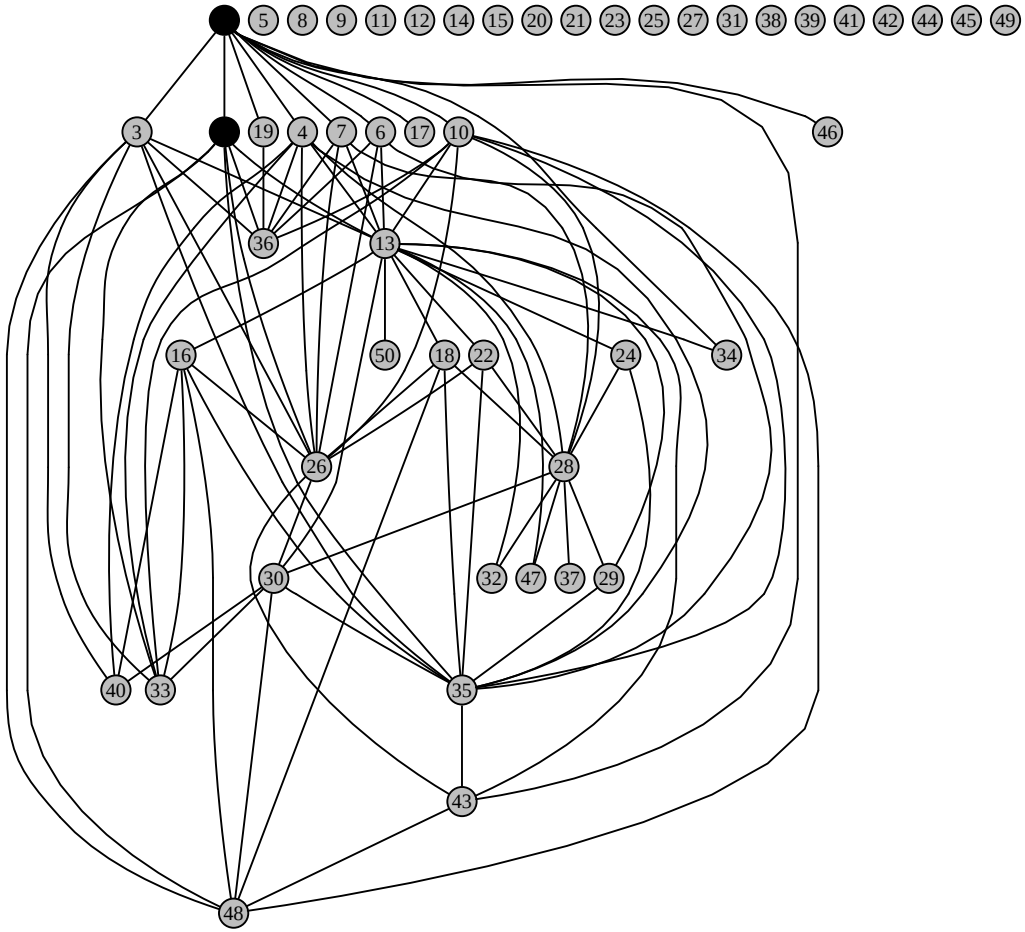
Keep every edge with probability p

The partial duplication random graph



This was one step in the partial duplication random graph

Example with 50 vertices and $p = 0.6$



Theorem 1
Let $E^\circ(n)$ be the average degree when there are n vertices. Then,

$$E^\circ(n) \xrightarrow{n\to\infty} \begin{cases} \infty, & p > 1/2, \\ 0, & p < 1/2. \end{cases}$$

Theorem 2
Let $F_0^\circ(n)$ be the frequency of singletons and $p^* + \log p^* = 0$. (Note $p^* \approx 0.56$.) Then,

$$F_0^\circ(n) \xrightarrow{n\to\infty} 1 \qquad \text{iff} \qquad p \le p^*.$$

The average degree

- Let $E(n)$ be the number of edges when there are n vertices

$$\mathbf{E}[E(n+1)|\mathcal{F}_n] = E(n)\Big(1+\frac{2p}{n}\Big),$$

$$\mathbf{E}\Big[E(n+1)\underbrace{\frac{n_0\cdots n}{(n_0+2p)\cdots(n+2p)}}_{\xrightarrow{n\to\infty}\frac{\Gamma(n_0+2p)}{\Gamma(n_0)}n^{-2p}}\Big|\mathcal{F}_n\Big] = \underbrace{E(n)\frac{n_0\cdots(n-1)}{(n_0+2p)\cdots(n-1+2p)}}_{\text{non-negative martingale}}$$

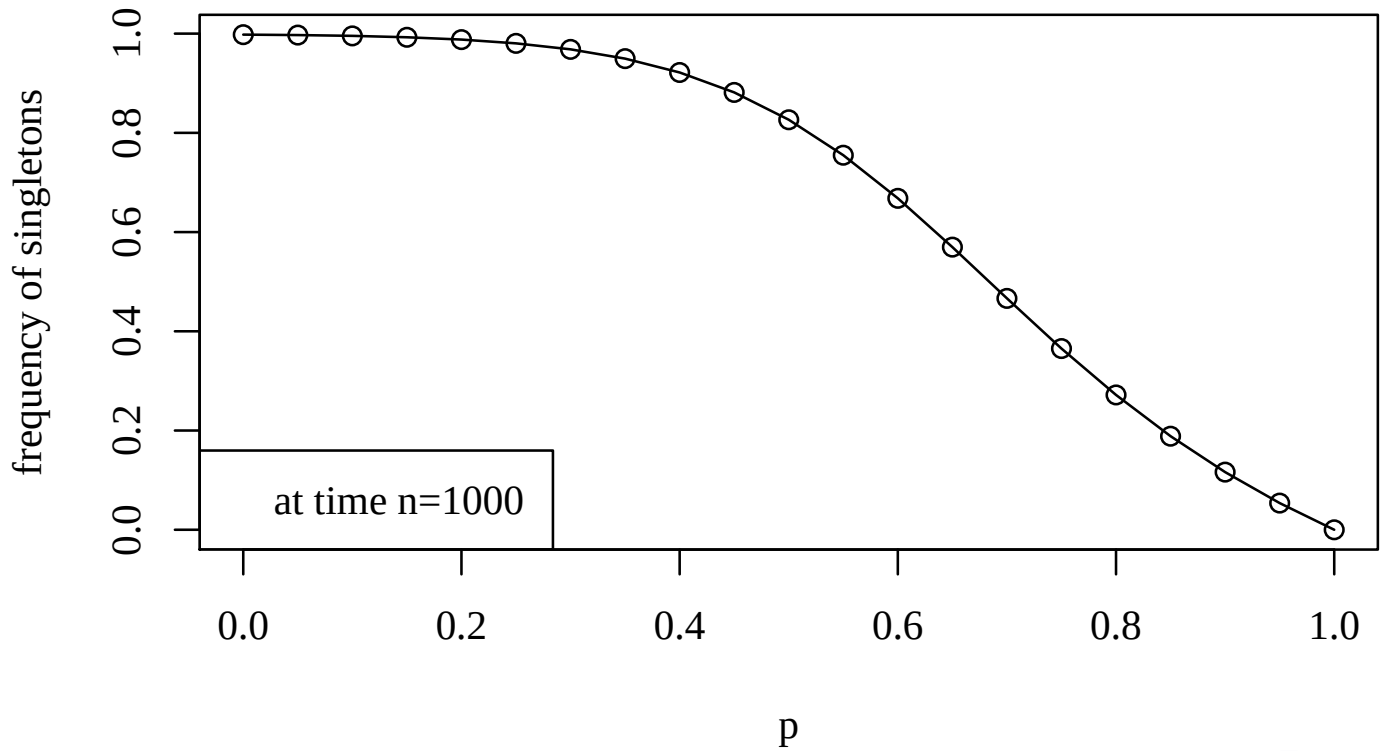
- This implies $n^{-2p}E(n) \xrightarrow{n\to\infty}_{a.s.} E(\infty)$.
- $E^\circ(n) := \frac{E(n)}{n}$ is the average degree of a vertex. So,

$$E^\circ(n) \xrightarrow{n\to\infty} \begin{cases} \infty, & p > 1/2, \\ 0, & p < 1/2. \end{cases}$$

The frequency of singletons

- Let $F_0^\circ(n)$ be the frequency of singletons

$$p < 1/2: \qquad E^\circ(n) \xrightarrow{n\to\infty} 0 \quad \Longrightarrow \quad F_0^\circ(n) \xrightarrow{n\to\infty} 0$$



The frequency of singletons

- $F_k(n)$: number of vertices of degree k, $F_k^\circ(n) = \frac{F_k(n)}{n}$,

$$H_x^\circ(n) := \sum_{k=0}^{\infty} F_k^\circ(n)(1-x)^k \quad \Longrightarrow \quad F_0^\circ(n) = H_1^\circ(n).$$

- Using recursions,

$$\mathbf{E}[H_x^\circ(n+1) - H_x^\circ(n)|\mathcal{F}_n] = \frac{1}{n+1}\left(px(1-x)\frac{d}{dx}H_x^\circ(n) + H_{px}^\circ(n) - H_x^\circ(n)\right)$$

The frequency of singletons

- Time continuous partial duplication graph $\mathcal{G}$: Every vertex splits at rate $1 + \frac{1}{|G_s|}$. Then,

$$\frac{d}{ds}\mathbf{E}[H_x^\circ(s)|\mathcal{F}_s] = px(1-x)\frac{d}{dx}H_x^\circ(s) + H_{px}^\circ(s) - H_x^\circ(s)$$

- Let $\mathcal{X}$ be a Markov process which jumps from x to px at rate 1, and increases at rate $x(1-x)$ between jumps

$$\Rightarrow \frac{d}{ds}\mathbf{E}[f(X_s)|\mathcal{F}_s] = pX_s(1-X_s)f'(X_s) + f(pX_s) - f(X_s)$$

- Let $\mathcal{G}$, $\mathcal{X}$ be independent. Then,

$$\frac{d}{ds}\mathbf{E}[H_{X_{t-s}}^\circ(s)] = 0$$

The frequency of singletons

$$\frac{d}{ds}\mathbf{E}[H^{\circ}_{X_{t-s}}(s)] = 0 \qquad \Longrightarrow \qquad \mathbf{E}[H^{\circ}_{X_0}(t)] = \mathbf{E}[H^{\circ}_{X_t}(0)]$$

- The process $\mathcal{X}$ satisfies, for $p^* + \log p^* = 0$

$$X_t \xrightarrow{t\to\infty} \begin{cases} 0, & p \le p^* \\ X_\infty \neq 1, & p > p^* \end{cases}$$

- Hence,

$$\lim_{n\to\infty} \mathbf{E}[F^{\circ}_0(n)] = \lim_{t\to\infty} \mathbf{E}[F^{\circ}_0(t)]$$

$$= \lim_{t\to\infty} \mathbf{E}[H^{\circ}_1(t)] = \mathbf{E}[H^{\circ}_{X_\infty}(0)] \begin{cases} = 1, & p \le p^* \\ < 1, & p > p^* \end{cases}$$

What next?

- Study connected component (power law?)
- Include additional r edges as in the original model.
- Parameter estimation using protein-protein-interaction network.