A random graph model with duplication

Tamás F. Móri
(joint work with Ágnes Backhausz)
Department of Probability Theory and Statistics
Eötvös Loránd University, Budapest, Hungary

Erdős Centennial, Budapest, July 3, 2013

Random graphs

Erdős–Rényi graphs (E–R 1959, 1960)

$G_{n,m}$ – n vertices; m edges, each with equal probability

Random graph process

in discrete time: $\left(G_{n,t},\ t = 0, 1, \ldots, \binom{n}{2}\right)$
probabilistic version: $\left(G_{n,t},\ t > 0\right)$ – edges come according to a homogeneous Poisson process

Random graphs

Erdős–Rényi graphs (E–R 1959, 1960)

$G_{n,m}$ – n vertices; m edges, each with equal probability

Random graph process

in discrete time: $\left(G_{n,t},\ t = 0, 1, \ldots, \binom{n}{2}\right)$
probabilistic version: $\left(G_{n,t},\ t > 0\right)$ – edges come according to a homogeneous Poisson process

Asymptotic results of the form *whp* (with high probability – i.e., with probability tending to 1 as $n \to \infty$)
E.g., threshold phenomenon for monotone graph properties

Random graphs

Erdős–Rényi graphs (E–R 1959, 1960)

$G_{n,m}$ – n vertices; m edges, each with equal probability

Random graph process

in discrete time: $\left(G_{n,t},\ t = 0, 1, \ldots, \binom{n}{2}\right)$
probabilistic version: $\left(G_{n,t},\ t > 0\right)$ – edges come according to a homogeneous Poisson process

Asymptotic results of the form *whp* (with high probability – i.e., with probability tending to 1 as $n \to \infty$)
E.g., threshold phenomenon for monotone graph properties

Very exciting mathematically, but real world networks often behave differently.

Degree distribution

The proportions of degree d vertices ($d = 0, 1, \dots$) decrease polynomially – scale free property.

Evolving random graphs

Degree distribution

The proportions of degree d vertices ($d = 0, 1, \dots$) decrease polynomially – scale free property.

Barabási and Albert (Science, 1999) – evolving graph process with preferential attachment dynamics.

At every step a new vertex and m new edges are added to the graph. The new edges connect the new vertex to old ones that are selected at random, with probabilities proportional to their current degrees.

Evolving random graphs

Degree distribution

The proportions of degree d vertices ($d = 0, 1, \dots$) decrease polynomially – scale free property.

Barabási and Albert (Science, 1999) – evolving graph process with preferential attachment dynamics.

At every step a new vertex and m new edges are added to the graph. The new edges connect the new vertex to old ones that are selected at random, with probabilities proportional to their current degrees.

$m = 1$ – a tree is built: plane oriented recursive tree (Yule 1925, Szymański 1987, etc.)

Evolving random graphs

Degree distribution

The proportions of degree d vertices ($d = 0, 1, \ldots$) decrease polynomially – scale free property.

Barabási and Albert (Science, 1999) – evolving graph process with preferential attachment dynamics.

At every step a new vertex and m new edges are added to the graph. The new edges connect the new vertex to old ones that are selected at random, with probabilities proportional to their current degrees.

$m = 1$ – a tree is built: plane oriented recursive tree (Yule 1925, Szymański 1987, etc.)

More general, more complicated models of web graphs (e.g., Cooper and Frieze, RSA 2003)

Evolving random graphs

An evolving random graph process means that the probability spaces describing the graph at subsequent stages of evolution are embedded one in another, so the whole process can be defined in the same probability space. This makes it possible to prove almost sure theorems. When it makes sense, it is stronger than the corresponding whp result, the two notions are related like *convergence in probability* vs *a.e. convergence*.

Evolving random graphs

An evolving random graph process means that the probability spaces describing the graph at subsequent stages of evolution are embedded one in another, so the whole process can be defined in the same probability space. This makes it possible to prove almost sure theorems. When it makes sense, it is stronger than the corresponding whp result, the two notions are related like *convergence in probability* vs *a.e. convergence*.

Asymptotic degree distribution: for every $d = 0, 1, \ldots$ the proportion of degree d vertices converges to a *constant* c_d with probability 1 as the size of the graph tends to infinity. The sequence (c_d) sums up to 1.
The asymptotic degree distribution is scale free, if $c_d \sim K\, d^{-\gamma}$ as $d \to \infty$, for some $\gamma > 0$ (characteristic exponent).

Evolving random graphs

An evolving random graph process means that the probability spaces describing the graph at subsequent stages of evolution are embedded one in another, so the whole process can be defined in the same probability space. This makes it possible to prove almost sure theorems. When it makes sense, it is stronger than the corresponding whp result, the two notions are related like *convergence in probability* vs *a.e. convergence*.

Asymptotic degree distribution: for every $d = 0, 1, \ldots$ the proportion of degree d vertices converges to a *constant* c_d with probability 1 as the size of the graph tends to infinity. The sequence (c_d) sums up to 1.
The asymptotic degree distribution is scale free, if $c_d \sim K\, d^{-\gamma}$ as $d \to \infty$, for some $\gamma > 0$ (characteristic exponent).

E.g., for the PORT $c_d = \frac{4}{d(d+1)(d+2)}$, $\gamma = 3$.

Heuristic reasoning

In the beginning of the new millenium we saw an explosion of (mostly heuristic) results, which were proved rigorously later. It is always striking when heuristics and approximate computations fail.

Duplication

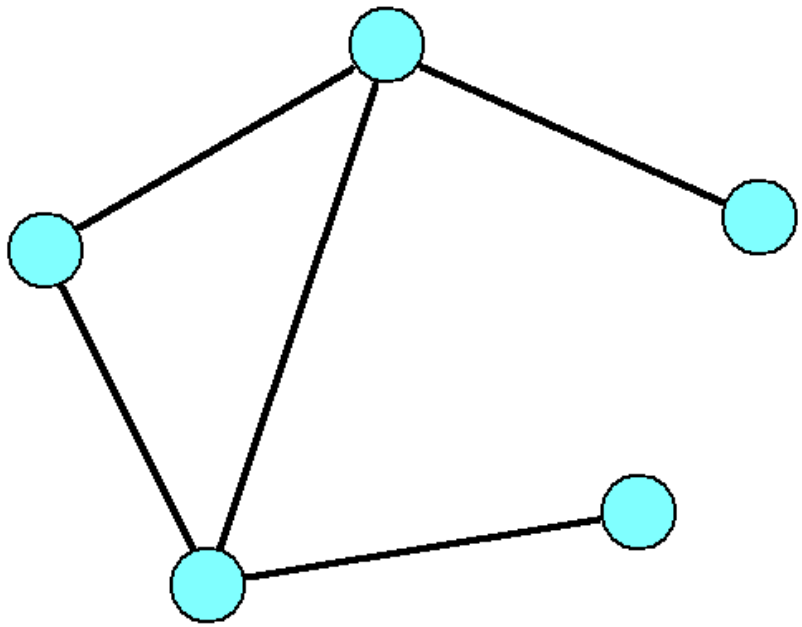
Biological networks within one cell (gene regulatory networks, protein-protein interaction networks, aka proteomes) are quite different from non-biological ones, e.g., $\gamma \in (1, 2)$. The central force of evolution is the duplication of the information in the genome.

Kim et al. (Phys. Rev. E, 2002)
Pastor-Satorras et al. (J. Theor. Biol., 2003)
Chung et al. (J. Comp. Biol., 2003)

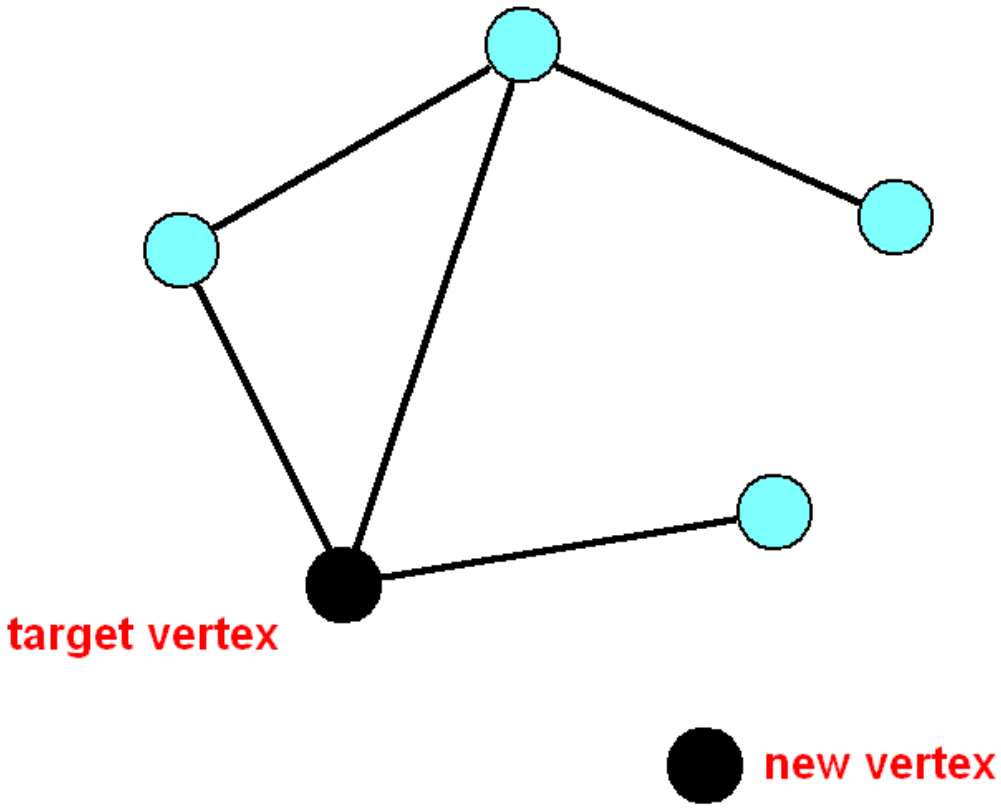
Start from a set of connected vertices and at each time step

- select one vertex of the graph at random and duplicate: add a new vertex and connect it to the neighbors of the selected vertex;
- delete the new edges independently, with probability δ;
- connect the new vertex to each old one independently, with probability α/n; merge multiple edges.

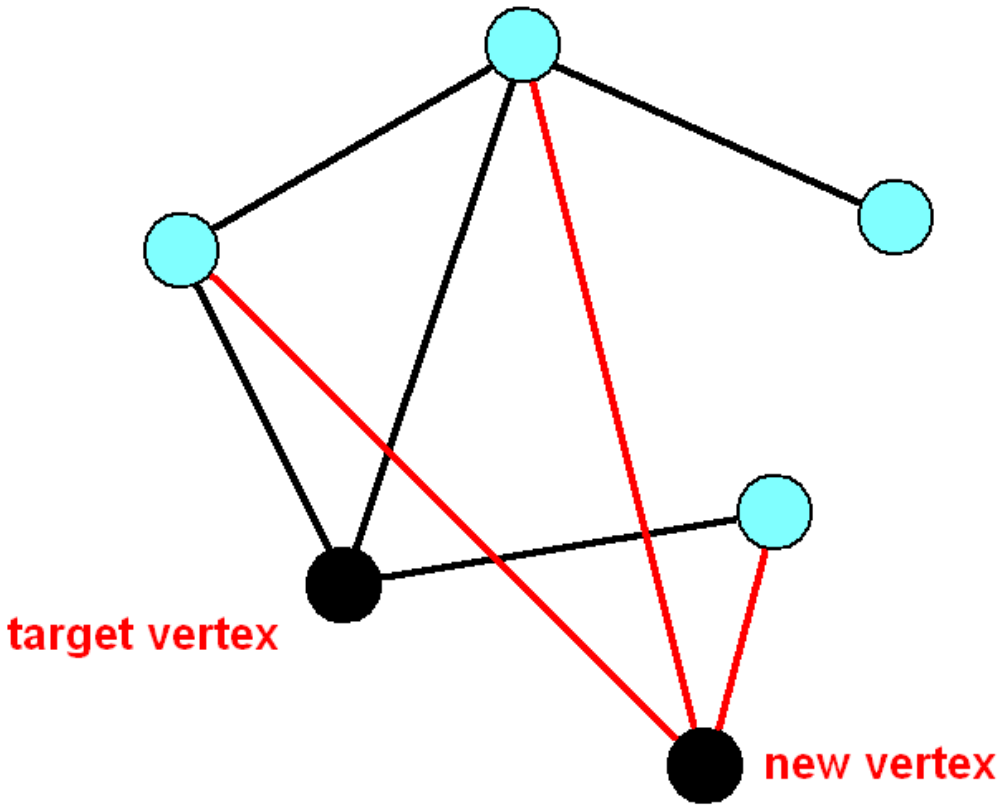
Pastor-Satorras et al. model



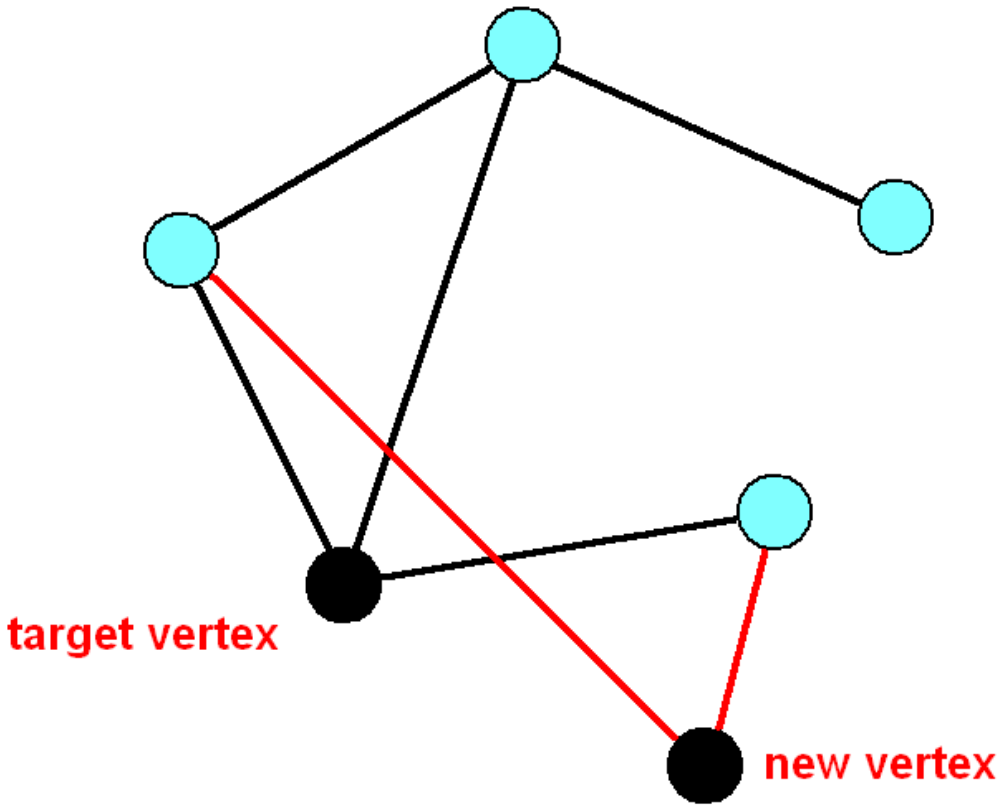
Pastor-Satorras et al. model



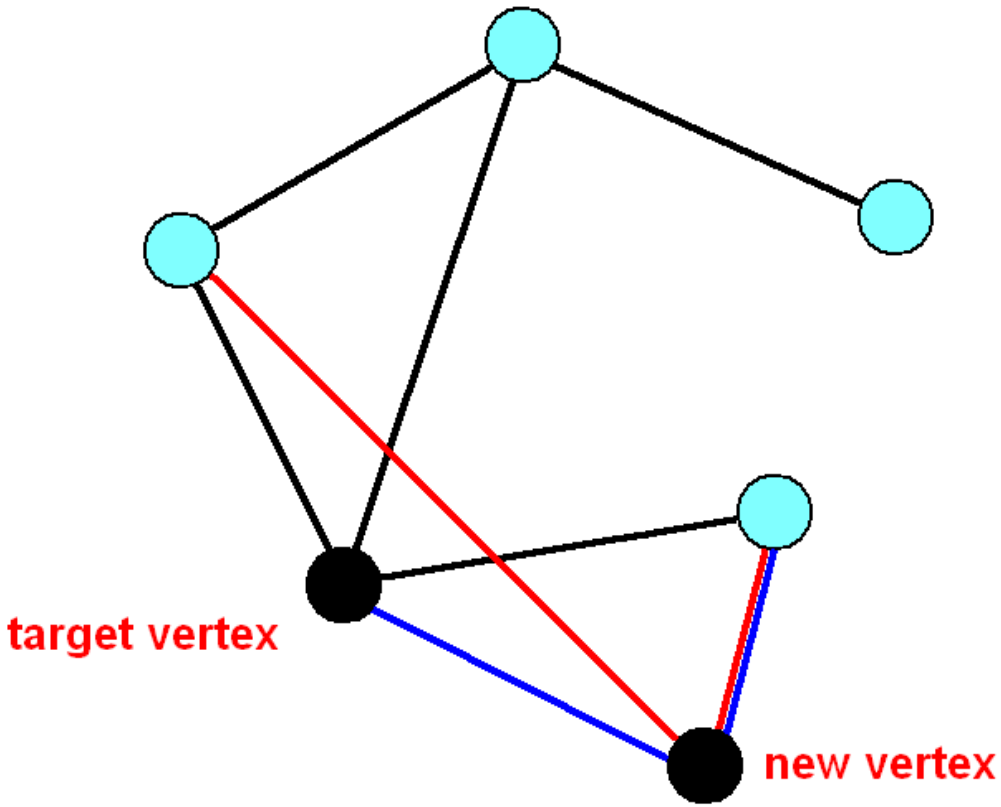
Pastor-Satorras et al. model



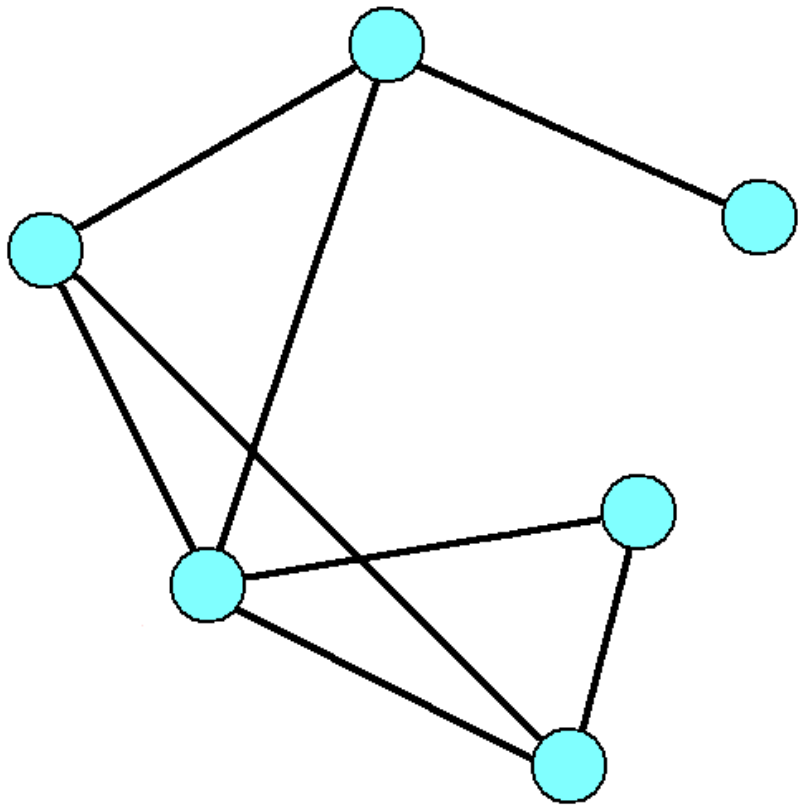
Pastor-Satorras et al. model



Pastor-Satorras et al. model



Pastor-Satorras et al. model



Pastor-Satorras et al. model

Results claimed

- The mean degree distribution is a power law with exponential cut-off, that is, $c_d \sim K\, d^{-\gamma} e^{-\lambda d}$ (Pastor-Satorras et al.)

FALSE

Pastor-Satorras et al. model

Results claimed

- The mean degree distribution is a power law with exponential cut-off, that is, $c_d \sim K\, d^{-\gamma} e^{-\lambda d}$ (Pastor-Satorras et al.)

FALSE

- The pure duplication model ($\delta = 0$) has an asymptotic degree distribution, which is a power law (Chung et al.)

FALSE

Bebek et al. (Theor. Comput. Sci., 2006)

Pastor-Satorras et al. model

Results claimed

- The mean degree distribution is a power law with exponential cut-off, that is, $c_d \sim K\, d^{-\gamma} e^{-\lambda d}$ (Pastor-Satorras et al.)

FALSE

- The pure duplication model ($\delta = 0$) has an asymptotic degree distribution, which is a power law (Chung et al.)

FALSE

Bebek et al. (Theor. Comput. Sci., 2006)

Modification: an extra step at the end of each turn. The new vertex is given a fixed number of additional edges, uniformly at random, so as to avoid singletons.

Then the expected number of degree d vertices, divided by n, converges, and the limits decrease polynomially (still a bit sketchy).

One more model with duplication and deletion

Inspired by these models we consider the following one.

Model 1

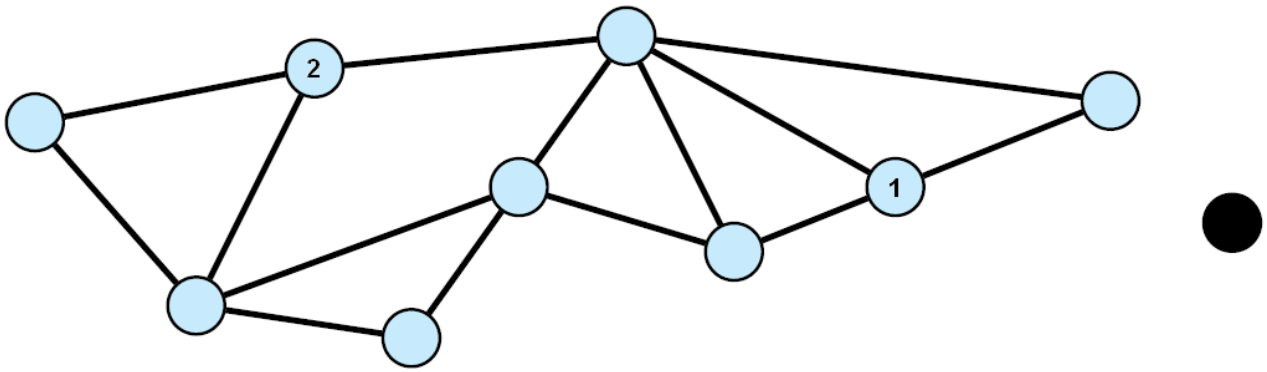
Start from a single vertex. The graph evolves in discrete time steps. At every step choose two (not necessarily different) vertices independently, uniformly at random.

- Duplication phase. Add a new vertex and connect it *to the first selected vertex* and to all of its neighbours.
- Deletion phase. Delete all *old* edges of the second selected vertex.

A kind of coagulation-fragmentation model. The presence of deletion causes intense fluctuation in the model's behavior, and makes it harder to access.

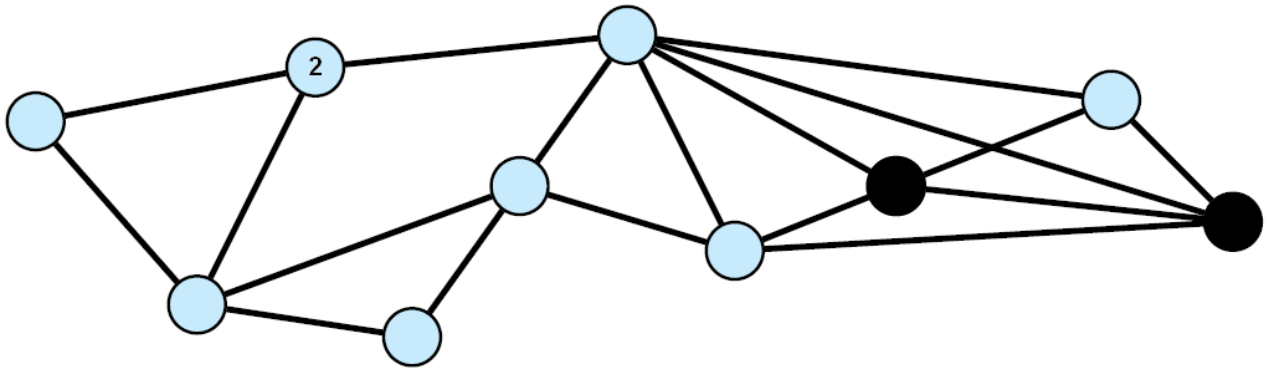
We do not claim that this model is biologically relevant.

One more model with duplication and deletion



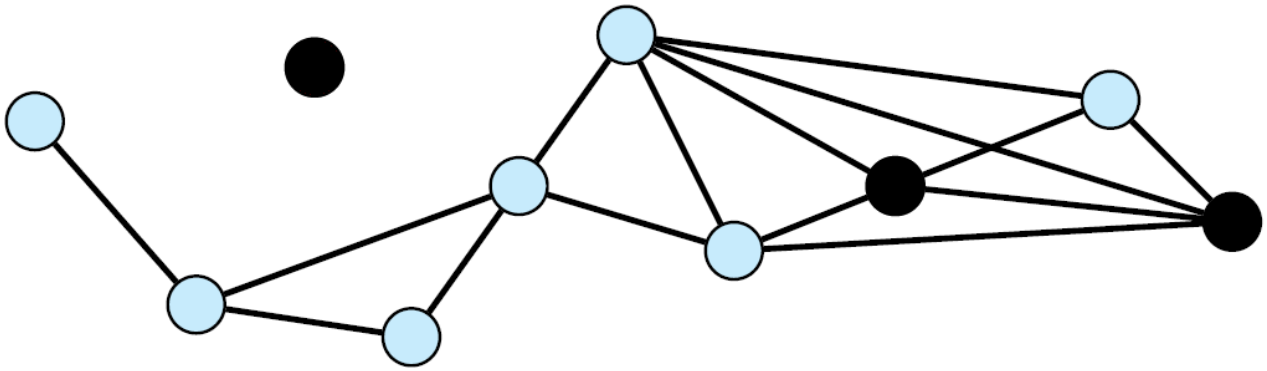
A new vertex is added (in black)

One more model with duplication and deletion



Duplication phase

One more model with duplication and deletion



Deletion phase

Properties

Linear growth

The number of vertices after step n is $n+1$, and the mean number of edges is equal to n — just like in a tree. However, in a recursive tree the individual degrees are increasing to infinity, while here they are regularly cut back to 0 due to deletion. In fact, individual degrees remain stochastically bounded, hence

- $\Rightarrow$ the asymptotic degree distribution (if exists) must have lighter tail;
- $\Rightarrow$ the maxdegree is of smaller order than any power of n;
- $\Rightarrow$ scale free property cannot be expected.

Asymptotic degree distribution

Theorem (Asymptotic degree distribution)

Let $X[n,d]$ denote the number of degree d vertices after step n. Then

$$\lim_{n\to\infty} \frac{X[n,d]}{n+1} = c_d, \quad d = 0, 1, \ldots,$$

with probability 1, where (c_d) is a sequence of positive numbers satisfying

$$c_0 = \frac{1+c_1}{3}; \quad c_d = \frac{(d+1)(c_{d-1}+c_{d+1})}{2d+3} \quad (d \geq 1);$$

$$c_0 + c_1 + \cdots = 1.$$

Not a recursion.

Method

A slight modification of the original model makes the analysis easier.

Model 2

The only point where Model 2 differs from Model 1 is that in the deletion phase the selected vertex loses *all of its edges* without any exceptions. That is, the new edges are protected in the erasure part of the same step in Model 1, but they might be deleted immediately in Model 2.

In Model 2, all connected components of the graph are cliques (including singletons). The number of d-cliques does not vary so vehemently as the number of degree d vertices: the fluctuation is bounded by 2.

Coupling

How to transfer the limit result from Model 2 to Model 1?

Coupling

We develop the two models simultaneously. The selected vertices are the same in both models. We color some of the edges and vertices of Model 1 red.

Edge coloring

- In the deletion phase, if a newborn edge is deleted in Model 2, color it red in Model 1.
- In the duplication phase, copies of red edges are also red in Model 1.
- All other edges are born to be black.

In this way, the edges in Model 1 are either black or red. The black subgraph is a realization of Model 2.

Coupling

Vertex coloring

- In the duplication phase, the new vertex becomes red if and only if the duplicated vertex is red.
- In the deletion phase, whenever an edge becomes red, color both endpoints red.
- If a red vertex is chosen to be deleted, and it loses all of its edges, color it back to black.

Black vertices can only have black edges, hence they have the same degree in both models. It is sufficient to show that the number of red vertices is $o(n)$. In fact, it is $o\left(n^{1/2+\varepsilon}\right)$.

A useful lemma

Lemma (simplified version)

Let $(\mathcal{F}_n)$ be a filtration, (ξ_n) a nonnegative adapted process. Suppose that

$$\mathbb{E}\big((\xi_n - \xi_{n-1})^2 \mid \mathcal{F}_{n-1}\big) = O\Big(n^{1-\delta}\Big)$$

holds with some $\delta > 0$. Let (u_n), (v_n) be nonnegative predictable processes such that $u_n < n$ for all $n \geq 1$.

(*a*) Suppose that $\mathbb{E}(\xi_n \mid \mathcal{F}_{n-1}) \leq \Big(1 - \dfrac{u_n}{n}\Big)\xi_{n-1} + v_n$, and $\lim_{n\to\infty} u_n = u$, $\limsup_{n\to\infty} v_n \leq v$ with some random variables $u > 0$, $v \geq 0$. Then

$$\limsup_{n\to\infty} \frac{\xi_n}{n} \leq \frac{v}{u+1} \quad \text{a.s.}$$

A useful lemma

Lemma (continued)

Let $(\mathcal{F}_n)$ be a filtration, (ξ_n) a nonnegative adapted process. Suppose that

$$\mathbb{E}\big((\xi_n - \xi_{n-1})^2 \mid \mathcal{F}_{n-1}\big) = O\Big(n^{1-\delta}\Big)$$

holds with some $\delta > 0$. Let (u_n), (v_n) be nonnegative predictable processes such that $u_n < n$ for all $n \geq 1$.

(*b*) Suppose that $\mathbb{E}(\xi_n \mid \mathcal{F}_{n-1}) \geq \Big(1 - \dfrac{u_n}{n}\Big)\xi_{n-1} + v_n$, and $\lim_{n\to\infty} u_n = u$, $\liminf_{n\to\infty} v_n \geq v$ with some random variables $u > 0$, $v \geq 0$. Then

$$\liminf_{n\to\infty} \frac{\xi_n}{n} \geq \frac{v}{u+1} \quad a.s.$$

Explicit form of the limits

Let $G(z)$ denote the generating function of the sequence (c_d), then

$$(1-z)^2 G'(z) = (3-2z)G(z) - 1, \quad G(0) = c_0.$$

This o.d.e can be solved.

$$G(z) = \frac{1}{(1-z)^2} \exp\Big(\frac{z}{1-z}\Big) \int_0^{1-z} \exp\Big(1 - \frac{1}{x}\Big)\, dx.$$

After some variable transformation we can expand it into Taylor series.

Theorem

$$c_d = (d+1) \int_0^{\infty} \frac{y^d e^{-y}}{(1+y)^{d+2}}\, dy, \quad d = 0, 1, \ldots .$$

Asymptotics

$$c_d = (d+1)\int_0^\infty \frac{y^d e^{-y}}{(1+y)^{d+2}}\, dy$$

In order to approximate the integral we first analyse the behavior of the integrand around the point where it attains its maximum. Second order Taylor approximation to the logarithm of the integrand turns the integral into a Gaussian one.

Theorem

$$c_d \sim (e\pi)^{1/2}\, d^{1/4}\, e^{-2\sqrt{d}}, \quad \text{as } d \to \infty.$$

Stretched exponential decay.

References

[1] Barabási, A-L., Albert, R., Emergence of scaling in random networks, *Science*, **286** (1999), 509–512.

[2] Bebek, G., Berenbrink, P., Cooper, C., Friedetzky, T., Nadeau, J., Sahinalp, S. C., The degree distribution of the generalized duplication model. *Theor. Comput. Sci.*, **369** (2006), 234–249.

[3] Chung, F., Lu, L., Dewey, T. G., Galas, D. J., Duplication models for biological networks, *J. Comput. Biol.*, **16** (2003), 677–687.

[4] Cooper C., Frieze, A., A general model of web graphs, *Random Structures Algorithms*, **22** (2003), 311–335.

[5] Erdős, P.; Rényi, A., On random graphs, I., *Publ. Math. Debrecen*, **6** (1959), 290–297.

References

[6] Erdős, P.; Rényi, A., On the evolution of random graphs, *MTA Mat. Kut. Int. Közl.*, **V(A1-2)** (1960), 17–61.

[7] Kim, J., Krapivsky, P. L., Kahng, B., Redner, S., Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev.*, **E66** (2002), 055101(R)

[8] Pastor-Satorras, R., Smith, E., Solé, R. V., Evolving protein interaction networks through gene duplication. *J. Theor. Biol.*, **222** (2003), 199–210.

[9] Szymański, J., On a nonuniform random recursive tree, *Ann. Discrete Math.* **33** (1987), 297–306.

[10] Yule, G. U., A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S., *Phil. Trans. R. Soc. B* **213** (1925), 21–87.