Evolving a peer-to-peer network

Martin Dyer

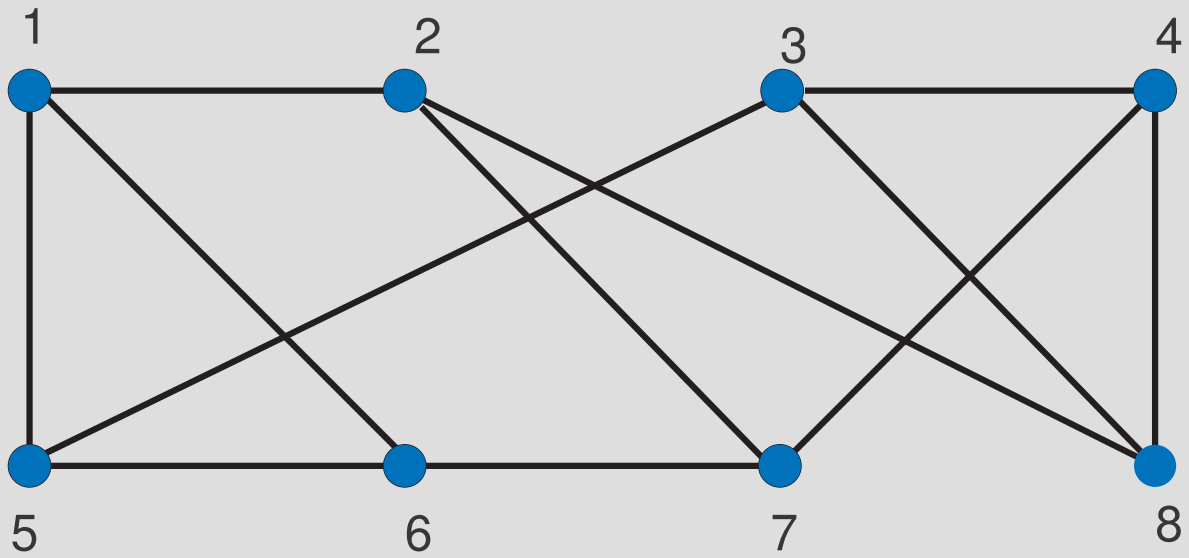
University of Leeds

Joint work with

Magnus Bordewich (Durham)
Colin Cooper (King's, London)
Catherine Greenhill (UNSW)

Sampling regular graphs

Let (d_n) be a sequence of positive integers such that $n \times d_n$ is even. Write $d = d_n$. For all n, let $\mathcal{S}_{n,d}$ be the set of all d-regular graphs on the vertex set $[n]$.

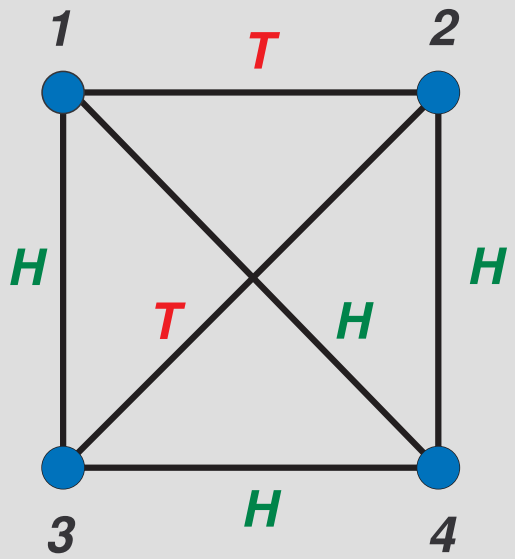


Question: How can we sample uniformly from $\mathcal{S}_{n,d}$?

It is easy to sample uniformly from the set of all graphs on vertex set $[n]$, using the Erdös-Rényi model.

For $1 \leq i < j \leq n$, independently flip a fair coin, and let

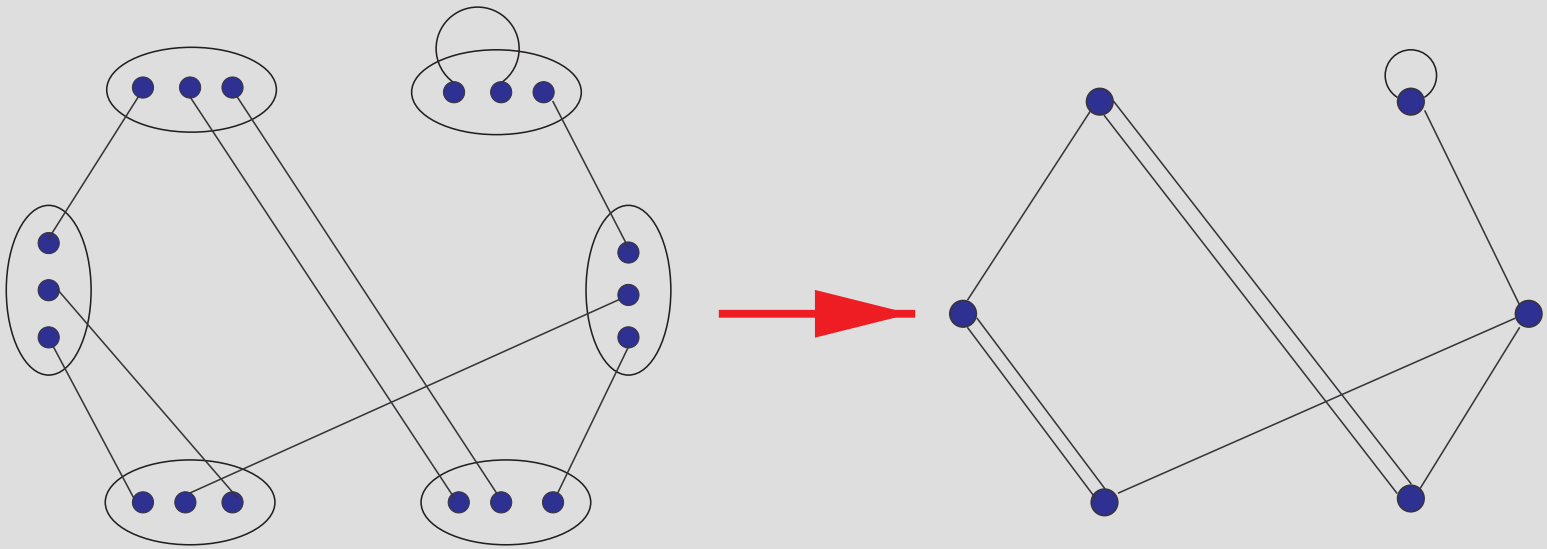
$$\begin{cases} \{i,j\} \in E(G) & \text{if heads,} \\ \{i,j\} \notin E(G) & \text{if tails.} \end{cases}$$



This sampling procedure takes $O(n^2)$ time.

Configuration model (Bollobás 1980, inter alia)

Take dn points in n classes of d points each. Choose a perfect matching of these points uniformly at random. Shrink the classes to vertices. This gives a d-regular multigraph on n vertices, which may have loops or multiple edges. If the multigraph is simple, accept it, otherwise try again.



This gives

$$|\mathcal{S}_{n,d}| = \frac{(dn)!}{(dn/2)!\,2^{dn/2}\,(d!)^n}\ \Pr(\text{simple}).$$

Moreover,

$$\Pr(\text{simple}) \sim \exp\left(\frac{1-d^2}{4}\right) \quad \text{as} \quad n \to \infty.$$

Conditioned on being simple, the output of the configuration model has uniform distribution on $\mathcal{S}_{n,d}$. Hence this gives an expected polynomial-time uniform sampling algorithm so long as $d = O(\sqrt{\log n})$.

Other uniform sampling algorithms

Wormald (1984): algorithms for $d = 3$ and $d = 4$.

Frieze (1988): expected polynomial time if $d = o(n^{1/5})$.

McKay and Wormald (1990): expected time $O(nd^3)$ if $d = O(n^{1/3})$. Algorithm uses switchings.

For larger d, or general degree sequences, this is a difficult problem, and some of the algorithms are quite complicated.

A simpler approach is to relax the uniformity condition and try to generate graphs which are only approximately uniform.

Approximately uniform sampling

Ideally we want an algorithm such that, given a constant $\varepsilon > 0$, the probability $\Pr(G)$ that a given graph $G \in \mathcal{S}_{n,d}$ is output satisfies

$$|\mathcal{S}_{n,d}|^{-1}(1+\varepsilon)^{-1} \leq \Pr(G) \leq |\mathcal{S}_{n,d}|^{-1}(1+\varepsilon).$$

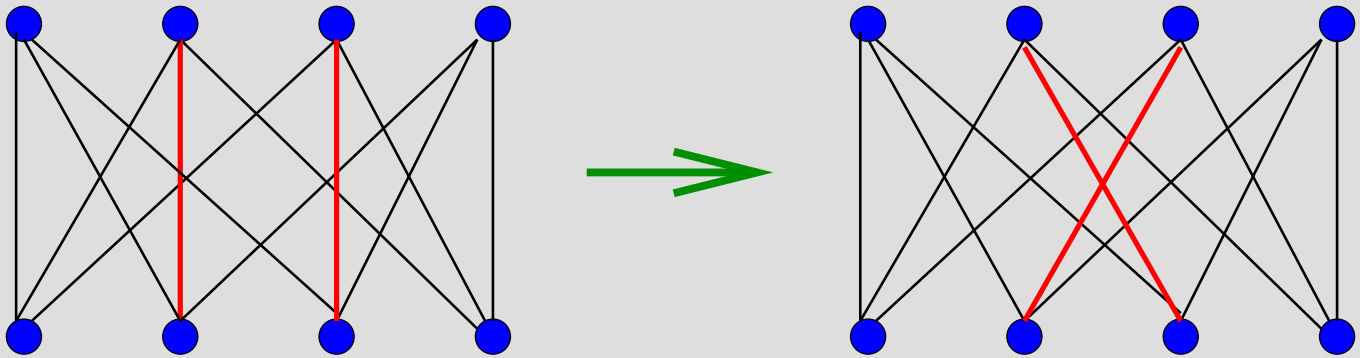
The algorithm is considered efficient if it runs in time polynomial in n, d and $\log(\varepsilon^{-1})$.

Tinhofer (1979): described an algorithm but no analysis of runtime or output distribution.

Jerrum and Sinclair (1990): approximate uniform sampler. Uses a Markov chain to sample matchings of a related graph. Polynomial time for "P-stable" degree sequences, or for all bipartite graphs (Jerrum, Sinclair and Vigoda, 2000).

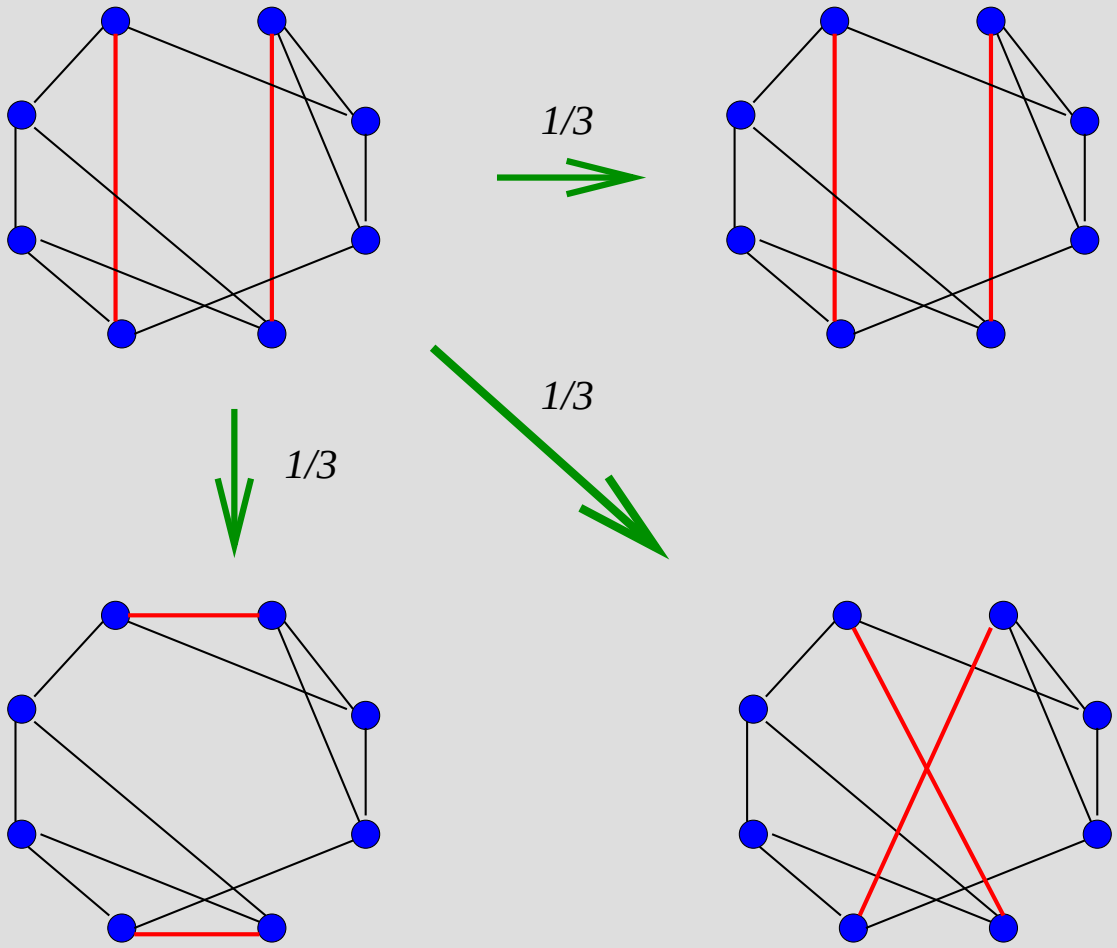
Steger-Wormald (1999): algorithm which produces graphs with asymptotically uniform distribution, when d is a small power of n. Can be thought of as a modification of the configuration model. Expected runtime $O(nd^2)$.

Kannan, Tetali and Vempala (1999): Markov chain for approximately uniform sampling of regular bipartite graphs. Polynomial time for regular bipartite graphs.



Analysis involves canonical paths. Arguments do not seem to work for general (non-bipartite) regular graphs.

Cooper, D and Greenhill (2005): analysed the switchings Markov chain for general regular graphs.

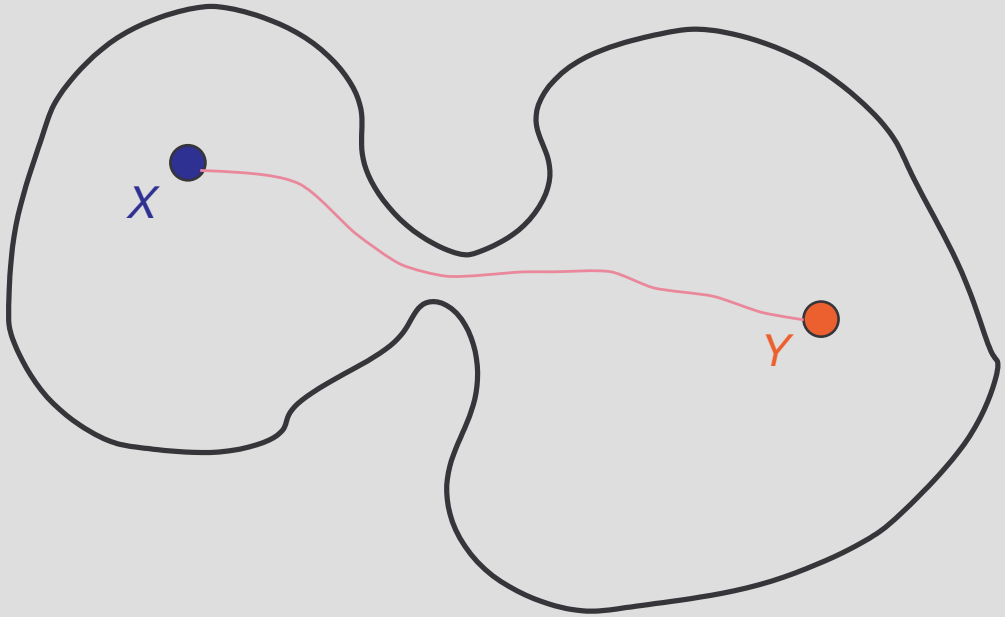


From $G \in \mathcal{S}_{n,d}$ do
  choose two nonadjacent distinct edges ij, $k\ell$, u.a.r.,
  choose a perfect matching M of $\{i, j, k, \ell\}$ u.a.r.,
  if $M \cap E(G) = \emptyset$ then
    delete the edges ij, $k\ell$ and add the edges of M,
  otherwise do nothing;
  endif;
end;

This Markov chain is irreducible and aperiodic on $\mathcal{S}_{n,d}$ with uniform stationary distribution.

The analysis uses canonical paths (Jerrum and Sinclair) and the multicommodity flow argument of Sinclair (also Diaconis and Stroock). Cooper, D and Greenhill use different canonical paths from those used by Kannan, Tetali and Vempala. (The canonical paths we use were suggested by Goldberg and Jerrum.)

Idea is to show that the Markov chain can move freely around the state space $\mathcal{S}_{n,d}$ so that it quickly approaches its uniform stationary distribution.

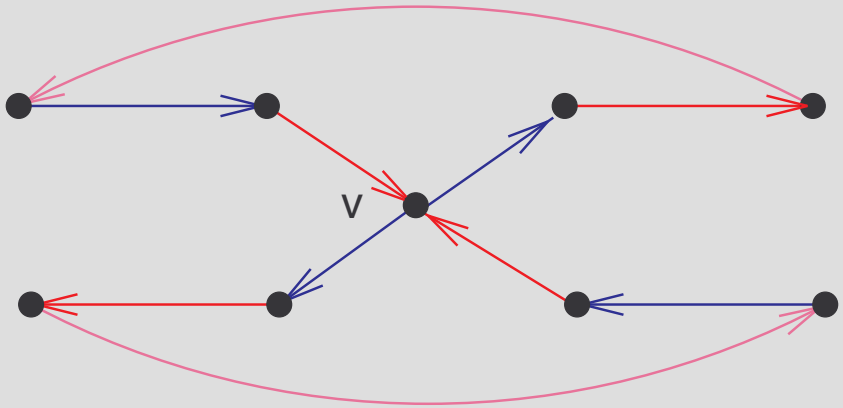


Given X, $Y \in \mathcal{S}_{n,d}$, define a path γ_{XY} from X to Y, where each step in the path corresponds to a move of the Markov chain.
If we can define the paths

$$\{\gamma_{XY} \mid X, Y \in \mathcal{S}_{n,d}\}$$

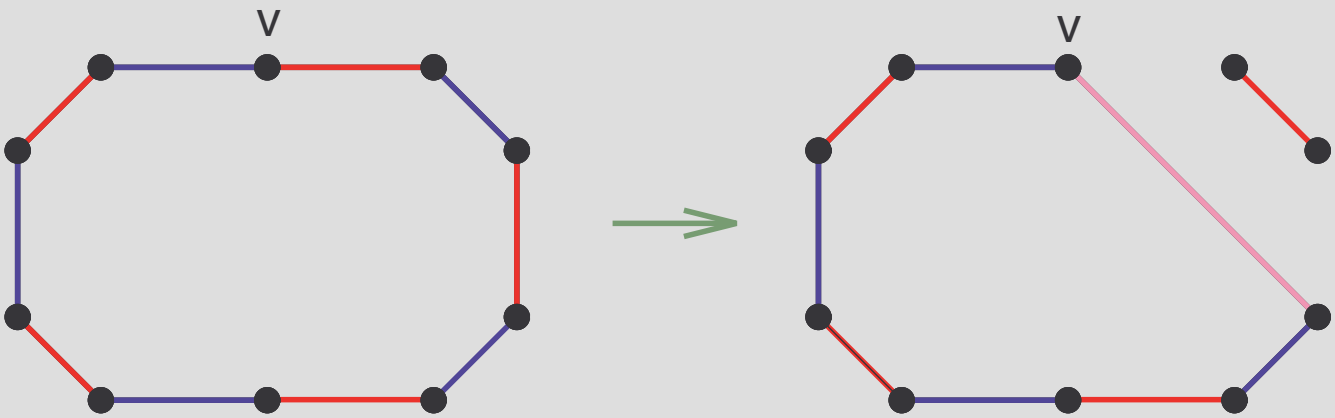
such that no edge is too badly overloaded, then this shows that there are no constrictions in the state space.

The paths transform graph X to graph Y by using a "canonical" procedure of flipping edges. First we construct "circuits" in the symmetric difference.

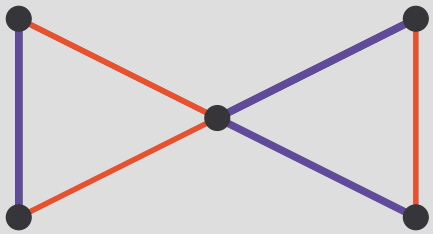


Canonical paths

We "unwind" cycles using a fixed vertex and a single non-edge.



Nonbipartiteness causes additional complications.



These can be overcome by using one more non-edge.

Multicommodity flow

The multicommodity flow approach generalises these ideas by using a weighted sum of canonical paths.

Putting all this together, we could show that the Markov chain will produce a sample with distribution within variation distance ε from uniform after at most

$$d^{17}\, n^{7}\, \log(dn\varepsilon^{-1})$$

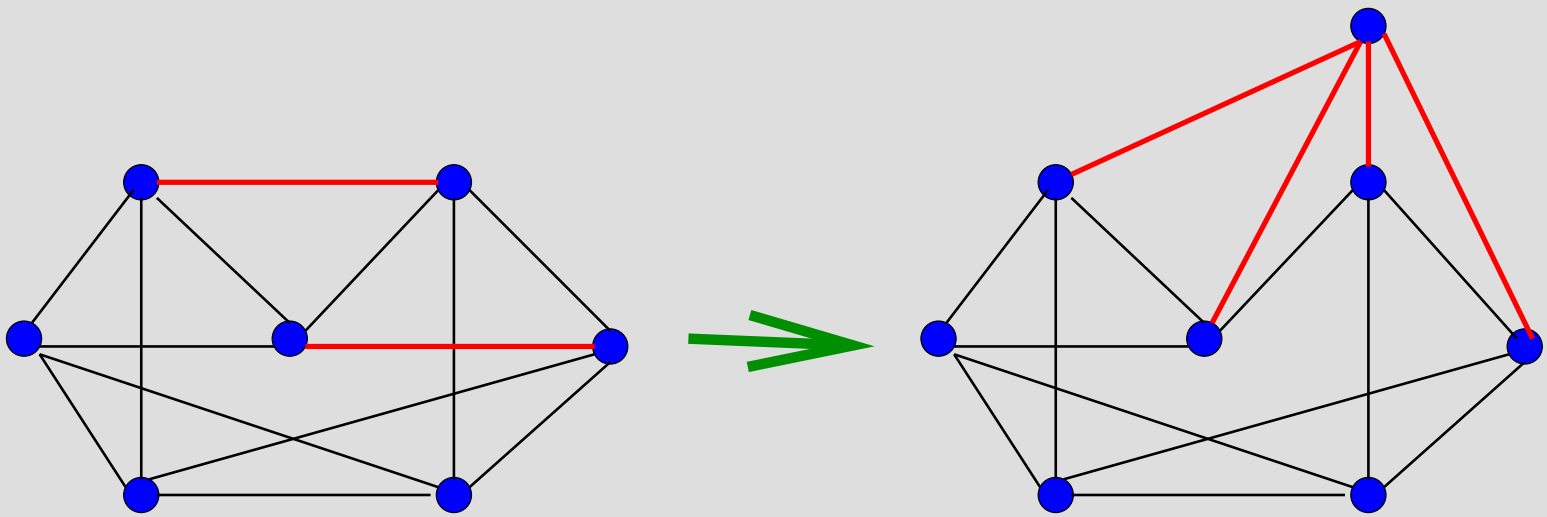
steps.

This bound is polynomial, but probably very much larger than necessary. Some improvements may be possible, but seem unlikely to give anything close to the truth.

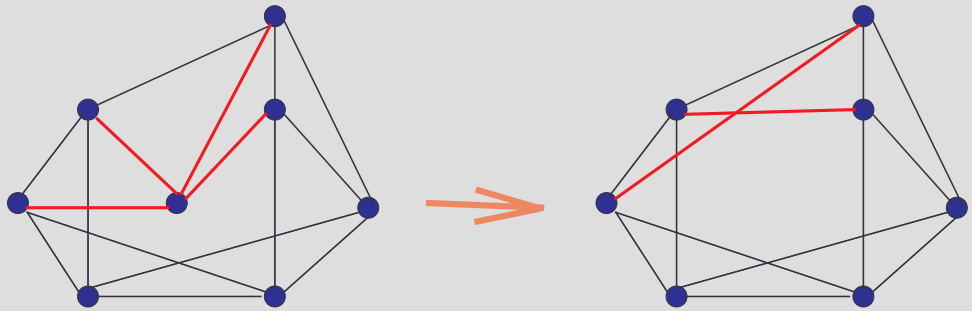
A peer-to-peer network

Bourassa and Holt (2003) proposed the following peer-to-peer network, called SWAN (Small-world Wide Area Networks).

Clients are arranged in a d-regular graph, where d is an even constant. When a new client arrives, $d/2$ independent (i.e. sharing no vertex) existing edges are chosen uniformly at random. This is done by performing random walks in the network. The new client then "pegs" these edges.



If a client wishes to depart, a perfect matching of its d neighbours is chosen and added to the graph.



Bourassa and Holt claimed that their network, started from an arbitrary small d-regular graph (e.g. the complete graph K_{d+1}), soon starts to behave like a random regular graph, and hence has attractive properties such as high connectivity and low diameter.

Connectivity and diameter

Random regular graphs have properties which are desirable for communication networks. In particular:

With high probability (i.e. probability $\to 1$ as $n \to \infty$),

1. They are d-connected, i.e. there are d vertex-disjoint paths between any pair of vertices. This provides protection against link or node failures and/or congestion in the network.
2. They have diameter $\sim \log_{d-1} n$. No d-regular graph on n vertices can have significantly smaller diameter. Low diameter is desirable since communications need to traverse fewer links.

For simplicity, let us consider $d = 4$ and write $\mathcal{S}_n = \mathcal{S}_{n,4}$. We model the process as a Markov chain with state space

$$\Omega = \bigcup_{n=5}^{\infty} \mathcal{S}_n.$$

Think of each set $\mathcal{S}_n$ as a level of the state space.

When there are n clients in the system, we assume that the interarrival time is exponentially distributed with mean ν_n, and residual service times for clients currently in the system are identically and independently exponentially distributed with mean μ_n. Then, at any event (arrival/departure), the probability that it is an arrival is

$$p_n = \frac{1/\nu_n}{1/\nu_n + n/\mu_n} = \frac{\mu_n}{\mu_n + n\nu_n}$$

and $q_n = 1 - p_n$ is the probability that it is a departure.

Arrival: When a new client (vertex) arrives, we choose two non-adjacent edges uniformly at random for the new vertex to peg.

Departure: If a client (vertex) wishes to leave, we choose a perfect matching of its neighbours uniformly at random and only allow the vertex to depart if none of these edges are present in the current graph.

Since this conceptual chain must always have vertex set $[n]$ for some n, it must do some "conceptual relabelling" which is not performed by the actual process.

From $G \in \mathcal{S}_n$ do
  with probability p_n do
    choose two distinct non-adjacent edges u.a.r.,
    add vertex $n+1$ on these edges,
    swap labels of vertex $n+1$ and a randomly chosen vertex,
  else (with probability q_n) do
    choose $i \in \{1, \ldots, n\}$ u.a.r.,
    choose a perfect matching M of the neighbours of i, u.a.r,
    if $M \cap E(G) = \emptyset$ then
      swap the labels of vertex i and vertex n,
      delete vertex n and add M to the edge set,
    else do nothing;
    endif;
  end;

We want the stationary distribution π of the Markov chain to be uniform when conditioned on a given level. If $\sigma_n = \pi(\mathcal{S}_n)$ then we must set

$$\pi(X) = \frac{\sigma_n}{|\mathcal{S}_n|}$$

for all $X \in \mathcal{S}_n$.

Suppose that $Y \in \mathcal{S}_{n+1}$ and that the Markov chain can move from X to Y in one step. If $\pi(X)P(X,Y) = \pi(Y)P(Y,X)$ for all such pairs then detailed balance ensures that π is the stationary distribution of the Markov chain. Use this to set the values of σ_n.

Now

$$P(X,Y) = \frac{p_n}{a_n(n+1)}, \quad P(Y,X) = \frac{q_{n+1}}{3(n+1)},$$

where $a_n = 2n^2 - 7n$ is the number of unordered pairs of non-adjacent edges in any 4-regular graph on n vertices. Detailed balance is equivalent to

$$\frac{\sigma_n\, p_n}{|\mathcal{S}_n|\, a_n} = \frac{\sigma_{n+1}\, q_{n+1}}{3\,|\mathcal{S}_{n+1}|}.$$

Since

$$\frac{|\mathcal{S}_{n+1}|}{|\mathcal{S}_n|} = \frac{a_n}{3}\,(1 + O(1/n))$$

we must set

$$\sigma_{n+1} = \frac{\sigma_n p_n}{q_{n+1}}\,(1 + O(1/n))\,.$$

Assumptions

The ratio between $\Pr(\text{departure})$ and $\Pr(\text{arrival})$ is $n\nu_n/\mu_n$. We assume that this is a strictly increasing function of n. Hence p_n is a strictly decreasing function of n: this puts the brakes on growth when the system gets large.

Also $p_0 = 1$. Write $p(n) = p_n$ and suppose that there exists $N > 0$ such that $p(N) = 1/2$. If no such N exists then the system grows indefinitely and there is no equilibrium behaviour.

Now write $p(n) = f(n/N)$ and assume that f is twice-differentiable on $x \geq 0$ and that $f''(x)$ is uniformly bounded by an absolute constant on $x \geq 0$. Then

$$f(0) = 1, \quad f(1) = 1/2, \quad f'(1) = -\alpha < 0$$

for some positive constant α.

Example

If arrivals are a Poisson process with rate $1/\nu$ and the departure of each client is Poisson with rate $1/\mu$, then

$$p_n = \frac{\mu}{\mu + n\nu}$$

when there are n clients in the system. Here $N = \mu/\nu$ and

$$p_n = \frac{1}{1 + n/N} = f(n/N)$$

where $f(x) = 1/(1 + x)$ satisfies our conditions.

This model was used by Pandurangan, Raghavan and Upfal (2003) in the analysis of a different architecture for peer-to-peer networks.

Lemma 1:

The equilibrium distribution of the size n of the system is approximately normal with mean N and variance $N/4\alpha$.

Define the centre of the state space to be

$$\Omega^* = \bigcup_{|N-n| \leq 2N^{3/4}} \mathcal{S}_n.$$

Lemma 2:

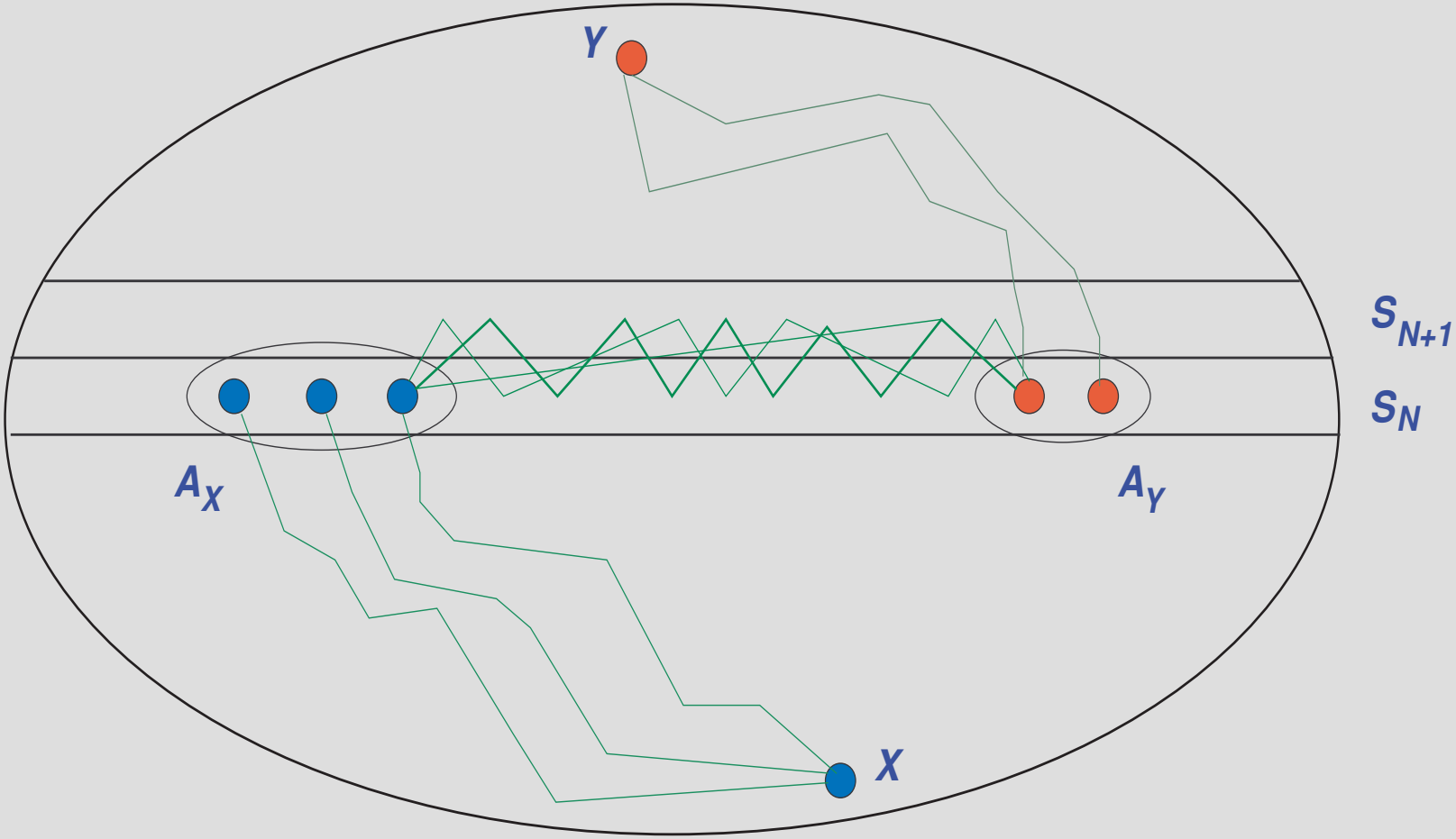
The following statements fail with probability exponentially small in N. If started from empty, the system reaches size $N - 2N^{3/4}$ in time $O(N^{5/4})$. Thereafter the system size does not leave the range $[N - 2N^{3/4}, N + 2N^{3/4}]$ in any time polynomial in N.

Analysis – multicommodity flow

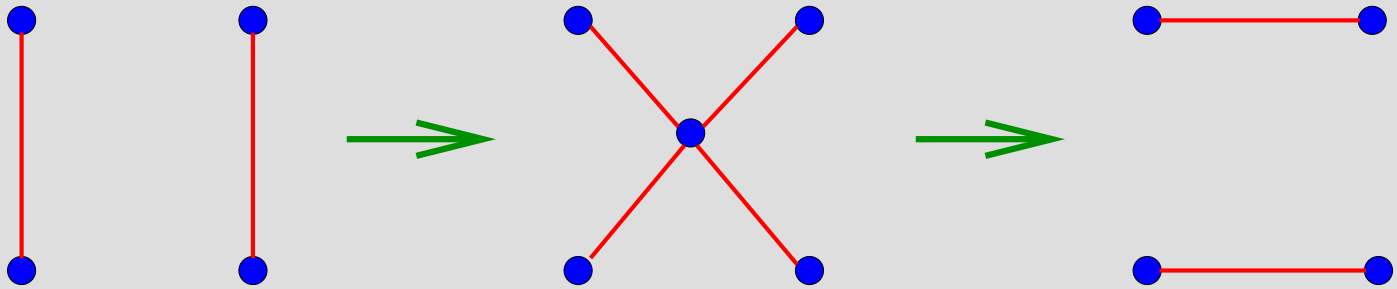
We condition on always remaining in the centre Ω^*.

For each central state $X \in \Omega^*$, we define a set of paths from X to a set of states $A_X \subseteq \mathcal{S}_N$ in a canonical way.

To define a set of paths between central states $X, Y \in \Omega^*$ we first route the flow from X to A_X, then use horizontal moves to route the flow from the set A_X to the set A_Y, and finally route the flow from A_Y to Y.


Y
S_{N+1}
S_N
A_X
A_Y
X

Horizontal moves are performed on the levels $\mathcal{S}_N \cup \mathcal{S}_{N+1}$. They simulate moves of the switchings chain described earlier, and implement the flow used to analyse that chain.



The edges with the greatest load are those between levels $\mathcal{S}_N$ and $\mathcal{S}_{N+1}$. We prove that the load on these edges is at most a factor $N^{3/4}$ higher than its load in the flow for the switchings chain. Hence, after a polynomial number of steps, the distribution of the Markov chain is almost uniform on each level.

Transient behaviour

Our analysis says that, after some polynomial number of steps, a snapshot of the SWAN process looks like a random regular graph. This gives some justification to Bourassa and Holt's claim that these networks have the good properties of random regular graphs (d-connected, logarithmic diameter).

But how do these desirable properties (high connectivity, low diameter) evolve over time? Do they hold (with high probability) at every step?

Can we model the behaviour of this network more closely?

Yes, provided we suitably limit the number of steps in the evolutionary phase.

Bordewich, D and Greenhill (2007) analyse the behaviour for the first $\kappa N \ln N$ steps, for a small enough constant κ.

During this period, departures are relatively less frequent than arrivals, and the network grows from constant size to $N - O(N^{1-\varepsilon})$ size, for some constant $\varepsilon > 0$.

This is close to the equilibrium size, and the analysis of equilibrium behaviour applies after this period.

The analysis uses a completely different approach from the equilibrium case. We compare the graphs produced by the process directly with random regular graphs, and show that their distributions remain "close" during the period.

The methods of analysis are based on random graph theory.

Relationship to random regular graphs

Our analysis is based on the following two lemmas.

Let ω_N be any sequence such that $\lim_{N\to\infty} \omega_N = \infty$.

Then, for $\omega_N \le t \le \kappa N \ln N$, with $n = n_t$ the size at time t,

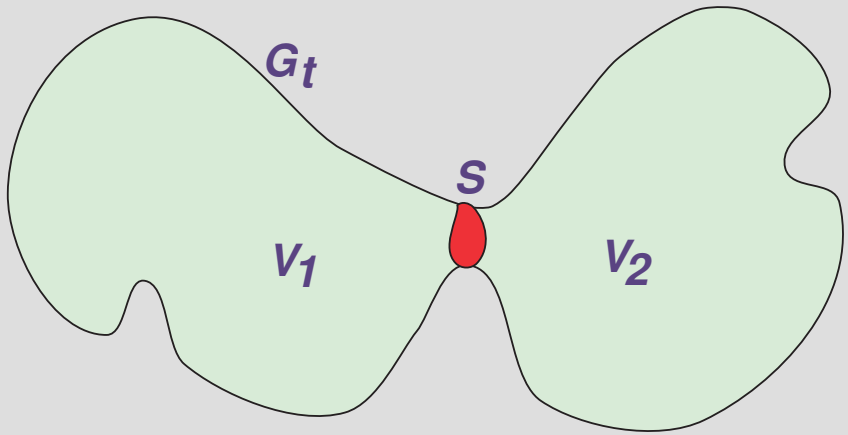
1. Comparison Lemma: Any event $\mathcal{E}$ which has probability $\Pr_{\mathrm{rg}}(\mathcal{E})$ in the regular random graph model, and probability $\Pr_{\mathrm{sw}}(\mathcal{E})$ in our network model satisfies
$$\Pr_{\mathrm{sw}}(\mathcal{E}) = O\big(n^{(d-1)^2/2+1/6}\big)\Pr_{\mathrm{rg}}(\mathcal{E}).$$
2. Small Subgraphs Lemma: For any constants k, e, let H be any graph with k vertices and e edges. Then
$$\Pr_{\mathrm{sw}}(H \subseteq G_t) = O\big(n^{k-e+1/6}\big).$$

Note that we cannot say anything about behaviour for constant t, since then every graph has a constant probability of occurring.

Connectivity

We can use these two lemmas to show that G_t is d-connected with high probability throughout $\omega_N \leq t \leq \kappa N \ln N$, as follows.

If G_t is not d-connected, there is some vertex subset S of size at most $d-1$ whose removal disconnects G_t into subgraphs with vertex sets V_1, V_2.

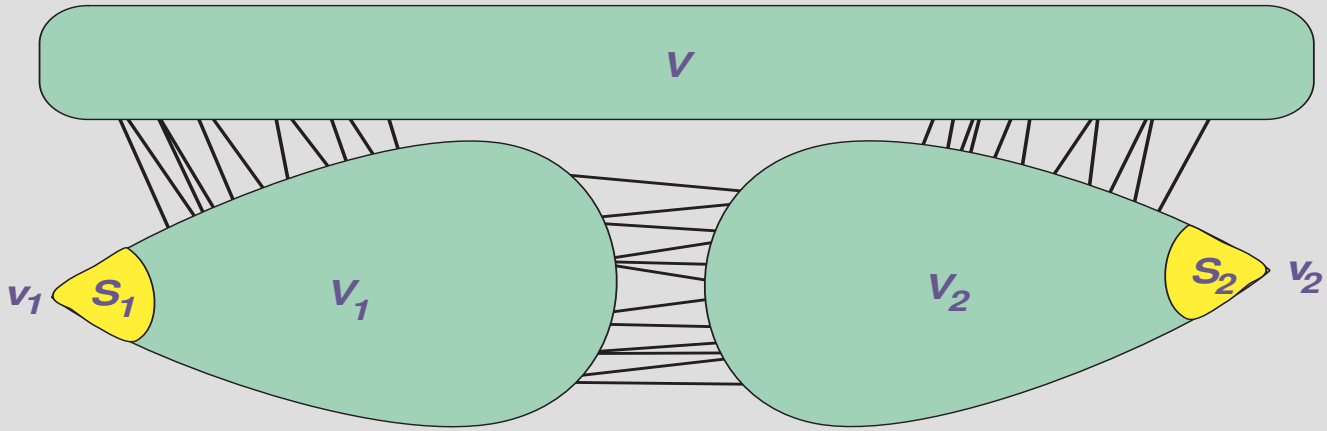


If both V_1 and V_2 are "large", the Comparison Lemma can be used, since there are too far few edges between V_1 and V_2, say.

If (say) V_1 is "small", the Small Subgraphs Lemma can be used, since there are too many edges in the V_1 subgraph.

Diameter (c.f. Bollobás & de la Vega, 1982)

We can also show that G_t has logarithmic diameter with high probability throughout $\omega_N \leq t \leq \kappa N \ln N$.



For any two vertices v_1 and v_2, we use the Small Subgraphs Lemma to show that G_t expands fast within a constant distance of v_1 and v_2, giving subgraphs S_1, S_2.

The Comparison Lemma is then used to show that G_t expands fast until we have subgraphs V_1, V_2 with $|V_1|, |V_2| \sim \sqrt{n \log n}$.

The Comparison Lemma is then used again to show that there must be an edge between V_1 and V_2.