Shortest Paths in Random Intersection Graphs

Gesine Reinert

Department of Statistics
University of Oxford
reinert@stats.ox.ac.uk

September 14$^{\text{th}}$, 2011

Outline

Bipartite graphs and random intersection graphs

The shortest distance - what to we know?

Finding the shortest path in random intersection graphs

What does the result tell us?

Joint work with Andrew Barbour (Zurich)

Examples of bipartite graphs

- Directors and companies (*Robins and Alexander*);
- Genes and gene properties (*Tanay et al.*);
- Persons and questions in an intelligence test (*Rasch*);
- Brazilian soccer players and clubs (*Onody and de Castro*);
- Plant-animal pollination, manufacturer-contractor interactions (*Saavedra et al.*).

We call the elements in two sets in the bipartite network *vertices* and *objects*.
In a bipartite network there are edges between vertices and objects, and between objects and vertices, but neither between vertices and vertices, nor between objects and objects.

Random intersection graphs

Random intersection graphs are derived from an underlying bipartite structure. In the simplest model, two vertices are connected if they have an edge to the same object. They were introduced by *Singer (1995)* and *Karonski et al. (1999)*.
They can be tuned to match degree distributions, and epidemics on random intersetion graphs have been studied by *Britton, Deijfen, Lageras and Lindholm (2008)*.
All edges are independent and undirected; and $p_{k,j}$ is the probability that there is an edge between vertex k and object j.
Here we think of K types of vertices and J types of objects; the probability of an edge between a vertex and an object depends only on their types.
We are interested in the shortest distance between two vertices in the random intersection graph.

The shortest distance in Bernoulli random graphs

The most standard random graph model is that of *Erdös and Renyi (1959)*; also called *Bernoulli random graphs*. The (finite) vertex set V is given, say $|V| = n$, and an edge between two nodes is present with probability p, independently of all other edges. *Erdös and Renyi (1960)*: If $p = p(n) = \frac{\log n}{n} + \frac{c}{n} + 0\left(\frac{1}{n}\right)$ then the probability that a Bernoulli graph, denoted by $\mathcal{G}(n, p)$ on n vertices with edge probability p is connected converges to $e^{-e^{-c}}$.

Here and in future log denotes the natural log.

If $p = \frac{\lambda}{n}$ and if $\lambda > 1$ does not depend on n, then if we pick two points A and B at random from the giant component, the shortest path $D(A, B)$ between these points satisfies (see *Durrett, 2006*) that, as $n \to \infty$,

$$\frac{D(A, B)}{\log n} \to \frac{1}{\log \lambda},$$

where the convergence is in probability.

In the Physics literature, the value $\frac{\log n}{\log(np)}$ is used for the average shortest path length in a Bernoulli random graph.

Newman (2001): If z_1 and z_2 are the average numbers of first and second neighbours of a vertex, and if n is the number of vertices, then in a general random graph with given degree sequence *(Molloy-Reed model)*, the average distance between pairs of vertices is approximately

$$\ell = \frac{\log(n/z_1)}{\log(z_2/z_1)} + 1.$$

Newman conjectured that many networks show this logarithmic behaviour.

The shortest distance in Watts-Strogatz small worlds

Ball et al. (1997), Watts and Strogatz (1998)
Newman, Moore, Watts (2000)
L vertices
k neighbours
ϕ probability of shortcut, per connection
$Lk\phi$ shortcuts on average

Results for Watts-Strogatz small worlds

Newman, Moore, Watts (2000)
based on continuous (great) circle model:
circle C of circumference L
Poisson $(L\rho/2)$ number of shortcuts added uniformly
neighbourhood collapsed by dividing distances by k
ρ corresponds to $2k\phi$
chords between points have length zero

Assume $L\rho > 1$
$\mathcal{D}$ shortest distance between two randomly chosen points
Barbour and R. (2001): uniformly in $|x| \leq \frac{1}{4}\log(L\rho)$,

$$\begin{aligned} & P\left(\mathcal{D} > \frac{1}{\rho}\left(\frac{1}{2}\log(L\rho) + x\right)\right) \\ = & \int_0^\infty \frac{e^{-y}}{1+e^{2x}y}\,dy + O\left((L\rho)^{-\frac{1}{5}}\log^2(L\rho)\right) \end{aligned}$$

(and exact expression for bound on the distance)

Idea of Proof

Pick point P at random from C
process walks from P at the same speed 2ρ in all possible directions, taking any shortcut that it can find
$R(t)$ the set of points that can be reached from P in time t
Taking shortcut means initially creating a new intervals on the circle will in due time meet some areas that it has covered before: dependence

Compare this process to a pure growth process $S(t)$:
start at P
growth rate $2p$
ignores overlap

For small times t: expect $R(t) \approx S(t)$

Pick another point P' at random from C
let an independent pure growth process run from that point
time at which the two independent pure growth processes will meet:
approximately $\frac{1}{2}\mathcal{D}$.

One way of realizing a random variable D with the above distribution is to realize W and W', and then to sample T from the conditional distribution

$$\mathbf{P}[T > x | W, W'] \;=\; e^{-\exp\{2x - G_1 - G_2\}},$$

where $G_1 := -\log W$ and $G_2 := -\log W'$ both have the Gumbel distribution. With this construction,

$$2T \;=_{\mathcal{D}}\; G_1 + G_2 - G_3,$$

where G_1, G_2 and G_3 are independent random variables with the Gumbel distribution (see Janson (1999) for an analogous result in a somewhat different context).

How do we proceed in bipartite networks?

Suppose that we have n_k vertices of type k, there are K vertex types in total; and there are m_j objects of type j; there are J object types in total. We use $n = n_1 + \cdots + n_k$ as the total number of vertices, and $m = m_1 + \cdots + m_J$ as the total number of objects. Write

$$P = (p_{k,j})_{k=1,\ldots,K;j=1,\ldots J}$$

for the edge probabilities, and define the diagonal matrices

$$N_X = diag(n_1, \ldots, n_k) \text{ and } N_Y = diag(m_1, \ldots, m_J).$$

The branching process approximation

Now we start a branching process in a general state $X(0)$ in the vertex space. Then the expected number of *offspring* of vertices of the different types is given by $X(0)M_X$, where

$$M_X = PN_Y P^T N_X$$

is the *mean matrix*. In general, the expected offspring vector in generation n is $X(0)M_X^n$.
We assume that the process is irreducible (i.e. could cover the whole space) and aperiodic.

The mean matrix is the key to branching processes. Let τ be the largest eigenvalue of M_X, and assume that $\tau > 1$ - then the probability that the branching process dies out is less than 1. Let μ and ν denote the left and right eigenvectors for τ, with $\mu_k \geq 0$ for all k, and $\sum \mu_k = 1$, and also $\mu^T \nu = 1$.

Fact: $W_i := \tau^{-i} X^T(i)\nu, i \geq 0$ converges (almost surely) to a limit W as $i \to \infty$.

Intuitively, this means that, if the branching process does not die out, its composition will stabilise, in the way described by the eigenvector for the largest eigenvalue.

Heuristic argument

We start two branching processes, one at a vertex A of type k_1, say, and the other one at vertex B of type k_2, say. The time when the offspring of the two branching processes meet is approximately half the shortest distance between A and B.
The time scale to look at it

$$i_0 \ := \ \left\lfloor \frac{\log n}{\log \tau} \right\rfloor ,$$

so that $\tau^{i_0} \leq n < \tau^{i_0+1}$, and we set

$$\tau^{-1} \ < \ \varphi(n) \ := \ n^{-1}\tau^{i_0} \ \leq \ 1.$$

To see this:

For X: no. individuals of type k in generation i is $\approx \tau^i W \mu_k$
No. descendants $X^A(i)$ of A at the i-th generation of X are $\approx \tau^i W^A \mu_k$, and those of B are $\approx \tau^i W^B \mu_k$
W^A and W^B are independent
Constructing the random intersection graph from the branching process:
indices are assigned to the vertices independently at random, with replacement.
Links between the A and B neighbourhoods occur whenever, for some $i \geq 1$ and some $1 \leq k \leq K$, one or more of the $X_k^B(i)$ are assigned the same index as one of the $X^A(i)$; other coincidences give rise to 'ghosts', and play no part in the intersection graph.

The mean number of links between the A and B neighbourhoods up to and including generation i is

$$\approx \sum_{s=1}^{i} \tau^{2s} W^A W^B \sum_{k=1}^{K} \frac{\mu_k^2}{n_k} \approx \kappa^X n^{-1} \tau^{2i} W^A W^B,$$

where

$$\kappa^X := \left(\frac{\tau^2}{\tau^2 - 1} \right) \sum_{k=1}^{K} \frac{\mu_k^2}{q_k^X},$$

and $q_k^X := n_k/n$. A similar formula hold for links occurring because of coincidence of indices at the object level; here, the expected number of links up to and including generation $i-1$ is $\approx \kappa^Y n^{-1} \tau^{2(i-1)} W^A W^B$, where $\kappa^Y = \tau \kappa^X$.

So overall mean number of links is $\approx \kappa n^{-1}\tau^{2i}W^A W^B$, where

$$\kappa \;=\; \frac{\tau}{\tau-1}\sum_{k=1}^{K}\frac{\mu_k^2}{n_k}.$$

Then, using Poisson approximation, the probability of there being no shared vertices in the i-neighbourhoods of A and B is

$$\approx \mathbf{E}_{k_1,k_2}\left\{e^{-\kappa n^{-1}\tau^{2i}W^A W^B}\right\},$$

this being the probability that the distance between A, of type $(k_1,1)$, and B, of type $(k_2,1)$, in the intersection graph exceeds $2i$.

While the branching process approximation is coarse, until the time when the two processes meet it turns out to be not a bad approximation, as in Watts-Strogatz small worlds.

Attention: The two processes could meet in vertex space, or in object space.

Our main result (2011)

Let W_A be the limiting variable for the process starting in A, and let W_B the limiting variable for the process starting in B. Let D denote the shortest distance between A and B.
If $W^A W^B$ has distribution function F_{k_1,k_2} on $\mathbf{R}_+$, and if

$$\begin{aligned}\mathbf{P}_{k_1,k_2}[U' \leq u] \\ := \quad \int_{(0,\infty)} \mathbf{P}[-(\log \tau)^{-1}(\Gamma + \log x + \log \kappa) \leq u] \, dF_{k_1,k_2}(x),\end{aligned}$$

where Γ denotes a standard Gumbel random variable, then

$$\mathbf{P}_{k_1,k_2}[U \leq u] \; = \; \mathbf{P}_{k_1,k_2}[U' \leq u + \log \varphi(n) / \log \tau].$$

Theorem

For $d = i_0 + u$, with $u \in \mathbb{Z}$ and $|u| < i_0/2$, we have

$$\begin{aligned}
&|\mathbf{P}_{k_1,k_2}[D \leq u + i_0] - \mathbf{P}_{k_1,k_2}[U' \leq u + \log \varphi(n)/\log \tau]| \\
&\quad \leq \quad \delta(\tau^u, m, n); \\
&|\mathbf{P}_{k_1,k_2}[D = \infty] - \mathbf{P}_{k_1,k_2}[U' = \infty]| \\
&\quad \leq \quad \delta(n^\alpha, m, n) + n^{-1} \\
&\qquad\quad + \mathbf{P}_{k_1,k_2}[0 < W^A W^B \leq \tau^2 \kappa^{-1} n^{-\alpha} \log n] + 2c_{26}\tau_1^{-i_0},
\end{aligned}$$

for any $0 < \alpha < (i_0 - 2)/2i_0 \approx 1/2$, where $\delta(y, m, n)$ is an explicit function, and $c_{26} > 0$ and $\tau_1 > 1$ are constants.

If $m \leq n$ then the first bound is of order $n^{-\frac{1}{4}}(\log n)^3$.

Remarks

- In general, the distribution of W_A and W_B is difficult to compute, but not so difficult to simulate from.
- Typical distances are of logarithmic order. This confirms *Newman*'s conjecture that essentially all average shortest paths are logarithmic in the number of vertices - to a crude level, as here the number of objects could dominate.
- The bound will be smaller, the larger τ is.
- The shortest paths will tend to be shorter, the larger τ is.

A Gumbel connection

We can write

$$\mathbf{P}[U = \infty] \;=\; \mathbf{P}[U' = \infty] \;=\; 1 - \mathbf{P}_{k_1}[W > 0]\mathbf{P}_{k_2}[W > 0],$$

where $\mathbf{P}_k$: start with a single individual of type $(k, 1)$;
Then $\mathcal{L}(U' \mid U' < \infty)$ is that of $\widetilde{U}$, realized as

$$\widetilde{U} \;=\; -\frac{1}{\log \tau}\{\Gamma + \log \widetilde{W}_A + \log \widetilde{W}_B + \log \kappa\},$$

where Γ, $\widetilde{W}_A$ and $\widetilde{W}_B$ are independent,

$$\begin{aligned} \mathbf{P}[\widetilde{W}_A \le w] &= \mathbf{P}_{k_1}[W \le w \mid W > 0] \quad \text{and} \\ \mathbf{P}[\widetilde{W}_B \le w] &= \mathbf{P}_{k_2}[W \le w \mid W > 0]. \end{aligned}$$

Also

$$F_{k_1,k_2}(0) \ = \ 1 - P_{k_1}[W > 0]P_{k_2}[W > 0]$$

approximates the probability that A and B are in different components of the graph, and are hence at infinite distance from one another.

Example: The Erdös-Rényi mixture model

Here $P = \alpha\beta^T$; the probability of a vertex of type i connecting to an object of type j is of product form; see (*Nowicky and Snijders*), and they have proved to be powerful for the analysis of sub-cellular networks, see (*Daudin et al*). In this case,

$$M_X(k,l) \quad = \quad \sum_{j=1}^{J} \alpha_k \beta_j m_j \beta_j \alpha_l n_l,$$

so that $M_X = C\alpha\alpha^T N_X$, with $C = \sum_{j=1}^J m_j \beta_j^2$, and has largest eigenvector $\tau = C\alpha^T N_X \alpha$.
Then τ is always at least as large as when a vertex makes no distinction as to which types of object it has links to.
For this model, *van den Esker, van der Hofstad, and Hooghstiemstra (2008)* obtained the same asymptotics, without bounds.

What does the result imply for random intersection graphs?

Spread of epidemics or rumours on bipartite graph: branching process approximation for initial spread
Rasch models, Erdös-Renyi mixture graphs: it is possible to estimate the edge probabilities (and the number of types/colours) using approximate methods, MCMC