

Small world networks

A. D. Barbour

Department of Mathematics, University of Zürich

and

G. Reinert

Department of Statistics, University of Oxford

1 Small Worlds

It happens to most of us that we meet a stranger, and in conversation discover that we have a joint acquaintance. “It is a small world”, we might then say. The small world phenomenon has been studied in the 1960’s by Milgram [11], who sent a number of packets to people in Nebraska and Kansas with instructions to deliver these packets to one of two specific persons in Massachusetts as promptly as possible. The constraint was that the packets could be sent only to persons whom the sender knew on a first-name basis. Milgram determined a median of only about five intermediary recipients to be required to get such a packet to the final destination.

About thirty years later, Watts and Strogatz [17] suggested a mathematical model for social networks that was able to mimic this *small-world phenomenon*. The original model was soon modified, (see Newman, Moore and Watts [14]), to make it more amenable for mathematical analysis. In its simplest form, L vertices are put on a one-dimensional ring lattice. Each vertex is connected to its neighbours at distance at most k away. Distance here is lattice distance, each bond on the lattice has length one. To this deterministic graph, random shortcuts are added. With probability ϕ per connection in the deterministic graph, two points are connected by a shortcut. Thus there are $Lk\phi$ shortcuts on average. We are now faced with a random graph that is quite different to the Bernoulli random graphs introduced by Erdős and Rényi [9], in that small worlds display a higher degree of clustering for a given diameter; see also Bollobas [6].

To describe such a network, typically the following summary statistics are used (e.g. Dorogovtsev and Mendes [7], [8]). First, to measure the diameter, the average shortest path length ℓ is introduced. Pick two vertices, calculate their shortest path (using lattice distance, 1 unit per connection), and take the average over all these pairs. To measure clustering, the clustering coefficient C is introduced: let C_i be the fraction of existing connections between nearest neighbours of the node i , then C is the average over all C_i .

The small world phenomenon can intuitively be described as follows: if ℓ is approximately like that for a (Bernoulli) random graph, then C is much larger than for that random graph. As a Bernoulli random graph need not be connected, this formulation does not make rigorous sense, but an intuitive understanding can be obtained observing that, to a very crude order (see [7]),

$$\ell_{random} \approx \frac{\ln L}{\ln(\phi L)}, \quad C_{random} \approx \phi. \quad (1.1)$$

The argument for this is that, the average number of neighbours of a node is $z = L\phi$, so about z^ℓ nodes of the network are at distance ℓ or closer to it. With $L \sim z^{\mathbf{E}\ell}$, we obtain $\mathbf{E}\ell_{random} \approx \frac{\ln L}{\ln(\phi L)}$. The clustering coefficient is approximately $\frac{L\phi/2}{\binom{L}{2}} = \frac{L\phi}{L+1} \approx \phi$.

Examples where this small-world phenomenon has been assessed empirically include the neural network of **C. elegans**, the metabolic network for **E. coli**, and the power grid of the Western United States. The following table gives the summary statistic, and the comparison with random graphs in the sense of (1.1); see [7].

	ℓ_{actual}	ℓ_{random}	C_{actual}	C_{random}	L
C. elegans	2.65	2.65	0.26	0.05	282
E.coli	2.9	2.9	≈ 0.3	0.025	282
Power grid	18.7	18.7	0.08	0.0005	4941

Further examples include social networks, the world wide web, rumor propagation (see Zanette [18]), the spread of epidemics (see Ball *et al.* [2], scientific collaboration networks (such as described by Erdős-numbers), metabolic networks (Fell [10]; Ravasz *et al.* [15]), and many more. Indeed, for $k = 1$ the small-world model was first suggested in [2], where it is called the *great circle model*, to study the spread of disease. It can be shown that, to control a disease,

movement restrictions (elimination of shortcuts) slow down the spread considerably (see Brian Grenfell’s contribution to this conference). From a theoretical physics viewpoint, scaling and percolation are some of the features of interest (e.g. Newman and Watts [12]; Newman *et al.* [13]). In general, small-world networks serve as models for networks that do not appear to be “purely random”. More examples, more references as well as more details can be found in the recent books by Watts [16], Barabasi [3] and Dorogovtsev and Mendes [8] as well as in the survey papers by Albert and Barabasi [1] and by Dorogovtsev and Mendes [7].

For a small world network, ℓ and C will both be random quantities, and thus their distribution needs assessing. The presented work concerned approximating the distribution of ℓ , with a bound on the approximation error.

2 The distribution of the shortest path length

To study the distribution of ℓ , Newman, Moore and Watts [14] introduced the continuous circle model. Instead of the ring lattice, study a circle C of circumference L , to which a Poisson $(L\rho/2)$ number of shortcuts are added uniformly over the circle. With neighbourhood collapsed by dividing distances by k , ρ corresponds to $2k\phi$. In this continuous circle model, chords between points have length zero.

Using mean-field heuristics, [14] derive the approximate distribution of the shortest distance ℓ ; in particular they state that

$$\mathbf{E}\ell = \frac{L}{k}f(Lk\phi),$$

where

$$f(z) = \frac{1}{2\sqrt{z^2 + 2z}} \tanh^{-1} \sqrt{\frac{z}{z+2}}$$

$$\sim \begin{cases} \frac{1}{4} & \text{for } z \ll 1 \\ \frac{\log(2z)}{4z} & \text{for } z \gg 1. \end{cases}$$

Let us restrict attention to the case that $L\rho > 1$ – if the probability is high that there are no shortcuts, then the shortest distance between two points will mostly just be the distance on the deterministic graph.

Barbour and Reinert [5] show that, uniformly in $|x| \leq \frac{1}{4} \log(L\rho)$,

$$\mathbf{P} \left(\mathcal{D} > \frac{1}{\rho} \left(\frac{1}{2} \log(L\rho) + x \right) \right)$$

$$= \int_0^\infty \frac{e^{-y}}{1 + e^{2xy}} dy + O \left((L\rho)^{-\frac{1}{5}} \log^2(L\rho) \right). \quad (2.2)$$

Also an exact expression for the bound on the distance is given.

The corresponding approximating probability from [14] is

$$\mathbf{P}\left(\mathcal{D} > \frac{1}{\rho} \left(\frac{1}{2} \log(L\rho) + x\right)\right) \approx \frac{1}{1 + e^{2x}} \left(1 + O\left((L\rho)^{-\frac{1}{2}}\right)\right).$$

The difference to (2.2) can be explained by the mean-field approximation in [14] being of rather crude order. We will return to this point once the result (2.2) is explained in more detail.

To derive the limiting distribution, the following heuristic may be useful. For the rigorous argument, see [5]. Pick a point P at random from C , and denote by $R(t)$ the set of points that can be reached from P within time t . Here we assume that the process walks from P at the same speed 2ρ in all possible directions, taking any shortcut that it can find. Thus it will grow at rate 2ρ from P along the circle. Whenever it encounters a shortcut, it will take it, creating new intervals on the circle that are covered by the process. This process will in due time meet some areas that it has covered before. This introduces dependence in the intervals. We compare this process to a pure growth process $S(t)$ starting at P with growth rate 2ρ , which ignores overlap. For small times t , we expect that $R(t) \approx S(t)$.

Now pick another point P' at random from C , and let an independent pure growth process run from that point. The time at which the two independent pure growth processes will meet will be approximately $\frac{1}{2}\mathcal{D}$, where \mathcal{D} is the length of the shortest path between P and P' .

To make the above heuristic more precise, denote that for the pure growth process $S(t)$ started at P (which is a Yule process) by $M(t)$ the number of intervals at time t , and by $s(t)$ the total length of the circle covered at time t . Then

$$\mathbf{E}M(t) = e^{2\rho t}, \quad \mathbf{E}s(t) = \frac{1}{\rho} (e^{2\rho t} - 1).$$

Denote by $N(t)$ and $u(t)$ the corresponding quantities for the pure growth process started at the point P' . Running both pure growth processes from time 0, at time t there are approximately $e^{4\rho t}$ pairs of intervals, and each has approximately length $\frac{1}{\rho}$. If V_t denotes the number of intersecting pairs of intervals at time t , one from the process started at P , the other from the process started at P' , then

$$V_t \approx \frac{2}{L\rho} e^{4\rho t}.$$

The time scale at which the first encounter of the two processes will happen should be such that V_t is a small but visible number. Thus we put

$$\tau_x = \frac{1}{2\rho} \left\{ \frac{1}{2} \log(L\rho) + x \right\};$$

then

$$V_{\tau_x} \approx 2e^{2x}.$$

Indeed V_t is random, and a mixed Poisson approximation for V_t can be derived. Given that $M(\tau_x) = m$, with interval lengths s_1, \dots, s_m , and $N(\tau_x) = n$, with interval lengths u_1, \dots, u_n , we have that

$$V_{\tau_x} \approx \text{Poisson} \left(\frac{2}{L} \sum_{i=1}^m \sum_{j=1}^n \min(s_i, u_j) \right).$$

If \hat{V}_t denotes the number of intersections at time t in the original process $R(t)$, started from P and P' , then $\hat{V}_{\tau_x} \approx V_{\tau_x}$. Intuitively this stems from τ_x being rather a small time; for later times the process may differ considerably. As

$$\{V_{\tau_x} = 0\} \approx \{\hat{V}_{\tau_x} = 0\} = \{\mathcal{D} > 2\tau_x\},$$

we thus obtain

$$\begin{aligned} \mathbf{P}\{\mathcal{D} > 2\tau_x\} &\approx \mathbf{E} e^{-\frac{2}{L} \sum_{i=1}^{M(\tau_x)} \sum_{j=1}^{N(\tau_x)} \min(s_i, u_j)} \\ &= \mathbf{E} e^{-\frac{2}{L} \int_0^{\tau_x} M(v)N(v)dv}. \end{aligned}$$

To derive the final result, a martingale argument is employed. We know that

$$e^{-2\rho t} M(t) \rightarrow W \quad a.s.,$$

where W is exponentially distributed with parameter 1. As

$$\begin{aligned} e^{-\frac{2}{L} \int_0^{\tau_x} M(v)N(v)dv} &\approx e^{-\frac{2}{L} W W' \int_0^{\tau_x} e^{4\rho v} dv} \\ &\approx e^{-e^{2x} W W'}, \end{aligned}$$

where W and W' are independent, exponential random variables with parameter 1. Noting that

$$\mathbf{E} e^{-e^{2x} W W'} = \int_0^\infty \frac{e^{-y}}{1 + e^{2x} y} dy$$

yields the stated result (2.2). Moreover, bounds on these approximations are given.

It might be intuitive to compare the above result (2.2) to that obtained by [14]. From (2.2) we have that

$$\mathbf{P}\left(\rho\mathcal{D} > \left(\frac{1}{2}\log(L\rho) + x\right)\right) \approx \int_0^\infty \frac{e^{-y}}{1 + e^{2x}y} dy. \quad (2.3)$$

Note that

$$\begin{aligned} \mathbf{E}\left\{e^{-e^{2x}WW'} \mid W, W'\right\} &= e^{-e^{2x}WW'} \\ &= e^{-\exp\{2x + \log W + \log W'\}} \\ &= e^{-\exp\{2x - G_1 - G_2\}}, \end{aligned}$$

where $G_1 := -\log W$ and $G_2 := -\log W'$ both have the Gumbel distribution. Let T be a random variable such that $\mathbf{P}(T > x)$ is given by the right-hand side of (2.3), then with this construction,

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

whatever the values of W and W' , and hence of G_1 and G_2 , implying that, in distribution,

$$2T = G_1 + G_2 - G_3,$$

where G_1, G_2 and G_3 are independent random variables with the Gumbel distribution. In contrast, the limiting distribution in [14] can be written as the distribution of $G_1 - G_3$, thus ignoring some of the initial branching variation.

The generalization of the above to higher-dimensional lattices derived in [5] shows that the reduction in shortest distance as a result of introducing shortcuts decreases with increasing dimension.

In forthcoming work, Barbour and Reinert study discrete small worlds, first a discrete circle with continuous time evolution, secondly the discrete circle with discrete time. The latter covers the original small world model, where shortcuts have length one, and neighbourhood sizes come in explicitly.

3 Conclusion

The above work is but a start on studying the statistical properties of small world networks. Yet there are already more complicated models suggested that demand treatment. Possible extensions of the above work include the following.

Often real networks are found to display a hierarchical structure, such as many small, highly connected substructures linked by a larger structure, see,

for example, Ravasz *et al.* [15]. The above does not incorporate such hierarchical networks, but might be adaptable to do so.

Another summary statistic to describe networks is the degree $\langle k \rangle$, the average number of neighbours for a vertex (see for example [7]). In many real networks, it has been observed that there are some vertices that have a very large number of connections to other vertices; indeed, often a scaling law for the degree is postulated. This is modelled using so-called *scale-free networks*, where the probability of a shortcut is biased towards vertices that already have shortcuts. Due to the uniformity of the shortcut construction, small-world networks do not display this scale-free property. It would be interesting to explore this further.

References

- [1] ALBERT, R. AND BARABASI, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47–97.
- [2] BALL, F., MOLLISON, D. AND SCALIA-TOMBA, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89.
- [3] BARABASI, A.-L. (2002). *Linked: the new science of networks*. Perseus, Cambridge, Massachusetts.
- [4] BARBOUR, A.D., HOLST, L., AND JANSON, S. (1992). *Poisson Approximation*. Oxford Science Publications, Oxford.
- [5] BARBOUR, A.D. AND REINERT, G. (2001, 2003). Small Worlds. *Random Structures and Algorithms* **19**, 54–74. Correction: submitted to *Random Structures and Algorithms*, 2003.
- [6] BOLLOBAS, B. (1985) *Random Graphs*. Academic Press, London.
- [7] DOROGOVTSSEV, S.N. AND MENDES, J.F.F. (2002). Evolution of random networks. *Adv. Phys.* **51**, 1079–1187.
- [8] DOROGOVTSSEV, S.N. AND MENDES, J.F.F. (2003). *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford.
- [9] ERDÖS, P. AND RÉNYI, A. (1959). On random graphs. *I. Publicationes Mathematicae (Debrecen)* **6**, 290–297.

- [10] FELL, D.A. AND WAGNER, A. (2000). The small world of metabolism. *Nature Biotech.* **189**, 1121–1122.
- [11] MILGRAM, S. (1967). The small world problem. *Psychol. Today* **2**, 60–67.
- [12] NEWMAN, M.E.J. AND WATTS, D.J. (1999). Scaling and percolation in the small-world network model. *Phys. Rev. E* **60**, 7332–7344.
- [13] NEWMAN, M.E.J., JENSEN, I., AND ZIFF, R.M. (2002). Percolation and epidemics in a two-dimensional small world network. *Phys. Rev. E* **65**, 021904.
- [14] NEWMAN, M.E.J., MOORE, C. AND WATTS, D.J. (2000). Mean-field solution of the small-world network model. *Phys. Rev. Lett.* **84**, 3201–3204.
- [15] RAVASZ, E., SOMERA, A.L., MONGRU, D.A., OLTVAI, Z.N., AND BARABASI, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1554.
- [16] WATTS, D.J. (1999). *Small Worlds*. Princeton University Press, Princeton.
- [17] WATTS, D.J. AND STROGATZ, S.H. (1998). Collective dynamics of “small-world” networks. *Nature* **393**, 440–442.
- [18] ZANETTE, D.H. (2002). Dynamics of rumor propagation on small-world networks. *Phys. Rev. E* **65**, 041908.