Finding short paths in small worlds: the idemetric property

Nen Huang[1] Andrew Lewis-Pye[2] Angsheng Li[3,1] Xuechen Li[4] Yicheng Pan[3]

March 27th, 2017

Abstract. We introduce the *idemetric* property, which formalises the idea that most nodes in a graph have similar distances between them, and which we suggest is likely to be satisfied by many small-world network models. As evidence for this claim, we show that the Watts-Strogatz model is idemetric for a wide range of parameters. For graphs with the idemetric property, we observe that the all-pairs shortest path problem can be easily reduced to the single-source shortest path problem, so long as one is prepared to accept solutions which are of stretch close to 2 with high probability. Applying Thorup's algorithm [MT99] to an undirected graph with m edges, for example, then provides a solution to the all-pairs problem for idemetric graphs, with gives paths of stretch close to 2 with high probability and runs in time $O(m)$.

[1]School of Computer Science, University of Chinese Academy of Sciences, Beijing, P. R. China.

[2]Department of Mathematics, London School of Economics, London, UK.

[3]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, 100190, P. R. China.

[4]Department of Computer Science, University of Toronto, Canada.

Lewis-Pye was partially supported by a Royal Society University Research Fellowship. This research was partially supported by the Grand Project "Network Algorithms and Digital Information" of the Institute of Software, Chinese Academy of Sciences, by an NSFC grant No. 61161130530, and by a China Basic Research Program (973) Grant No. 2014CB340302.

1 Introduction

The task of finding the shortest path between two nodes in a (possibly weighted and/or directed) graph is a fundamental problem in computer science, which has been extensively studied since the 1950s. Since often one would like to make queries for multiple pairs of nodes in the same graph, a number of variants of the problem also become significant:

The Single Source Shortest Paths (SSSP) Problem. The well known algorithm of Dijkstra [ED59] was originally formulated to give the shortest path between a single pair of nodes, but the version which is now better known solves the SSSP problem: a single node is fixed as the "source" node and the algorithm then finds shortest paths from the source to all other nodes in the graph. For a graph with n nodes and m edges, this algorithm for the SSSP problem terminates in time $O(n^2)$, but has been repeatedly improved upon. Thorup's algorithm[1] [MT99], for example, runs in time $O(m)$ for undirected weighted graphs with non-negative integer weights.

The All Pairs Shortest Paths (APSP) Problem. For the APSP problem one is required to output a data structure encoding all shortest paths between any pair of nodes. When a pair of nodes (a, b) is queried, the data structure should return a shortest path from a to b in time $O(\ell)$, where ℓ is the number of edges on this path. The classic Floyd-Warshall algorithm [RF62, SW62] solves this problem in time $O(n^3)$ for directed graphs. Running Thorup's solution to the SSSP for each node in an undirected graph gives an $O(nm)$ algorithm, which is much more efficient when the graph is sparse. For dense directed graphs, on the other hand, Williams' algorithm [RW14] is the best known, running in time $O(n^3/2^{\Theta(\log n)^{1/2}})$.

As pointed out by Thorup and Zwick [TZ05], however, there are many contexts in which the above solutions to the APSP problem are not satisfactory. Quite simply, the time and space complexity bounds provided by these algorithms are not practical for many real-world graphs of interest. One would like algorithms for which the preprocessing time – the time required to produce the data structure – is close to linear in n, or ideally, which can be run efficiently in an online and decentralised fashion, responding to changes in the graph structure as they occur. It therefore becomes natural to consider ways in which one can reasonably make the problem easier, and thereby obtain more efficient solutions. Of course, one option is to accept *approximate* solutions, i.e. algorithms which produce paths which are close to optimal. This can be made precise by requiring paths of small *stretch*, where a path from a to b is of stretch k if it is at most k times as long as the shortest path. Our concern here, will be in the production of very efficient algorithms which produce approximate solutions of small stretch ($k = 2$ or $k = 3$ say), and which are also probabilistic in the following sense (to be made precise later): we shall consider data structures which produce paths of small stretch with high probability, i.e. with probability tending to 1 as $n \to \infty$.

Small-world graphs. Another way in which to make the problem easier is to restrict attention to graphs with convenient properties. This will be a reasonable thing to do, so long as we consider properties which one can expect to be satisfied by graphs arising from real-world networks. While the algorithms described above are designed to run on arbitrary graphs, it may be that the situation changes dramatically when we restrict attention to *realistic* graphs. It is well known, for example, [WS98, AJB99] that graphs arising from real-world networks will very often have the *small world property*, meaning that the typical distance between nodes is $O(\log(n))$, while the clustering coefficient is not small – basically all graph models which are aimed at capturing common properties of real-world networks have this property that typical distances

[1] Here and throughout the paper we work under standard RAM model assumptions.

(and often the diameter) are $O(\log(n))$.

The question of finding short paths in small-world graphs was addressed by Kleinberg [JK00], for example. Since he considered decentralised algorithms in the absence of any preprocessing, however, his results were largely negative – the central point of that paper was that while short paths may exist, this does not mean that they can be easily found. Here we shall focus on random graph models which satisfy the small world property, but in contrast to Kleinberg we shall allow for preprocessing. Our results then stand in stark contrast to those of Kleinberg – precisely for the graphs which he is able to conclude that efficient decentralised algorithms do not exist in the absence of preprocessing, we shall show that very efficient decentralised algorithms do exist when reasonable levels of (even decentralised) preprocessing are permitted.

For the sake of simplicity, we shall focus for now on unweighted graphs. The expectation is, however, that all aspects of the analysis here can be extended to work for weighted graphs also.

1.1 Our results

If a and b are nodes in a (possibly directed) graph, then we let $d(a, b)$ denote the length of a shortest path from a to b (where all edges are traversed in the forward direction if the graph is directed, and with $d(a, b) = \infty$ if no such path exists). In the case of a weighted graph, the above definition still applies, if the length of a path is defined to be the sum of all edge weights in the path. Note that in a directed graph it may not be the case that $d(a, b) = d(b, a)$. In an undirected graph, however, d is a metric.

A very simple argument produces paths of stretch 3 with high probability, for any undirected graph and with preprocessing time $O(m\log(n))$. In order to see this, suppose that we pick k many *root* nodes uniformly at random, and then run Thorup's algorithm for the SSSP problem on each these k roots, in time $O(km)$. Given any pair of nodes a and b as a query, we then find the shortest path from a to b via any of the k roots and output this path. Now if a and b are chosen uniformly at random, consider the probability that the length of the path outputted by this algorithm is more than 3 times the length of the shortest path from a to b. Unless b is strictly closer to a than all of the k roots, which happens with probability $\leq 1/(k + 1)$, there will exist some root r such that $d(a, r) \leq d(a, b)$. Then we have that $d(r, b) \leq d(r, a) + d(a, b) \leq 2d(a, b)$, so that the length of the shortest path from a to b via r is at most 3 times $d(a, b)$. If we then set $k = \log(n)$ (or any function which tends to ∞ as n does) the probability of failure tends to 0 as $n \to \infty$. Note, though, that the query response time fails to be $O(d(a, b))$, since once has to search through each of the paths given by the $\log(n)$ many roots on each query.

It is also worth noting, that if we restrict to graphs of small diameter ($O(\text{polylog}(n))$ say), then one can run a decentralised version of the algorithm above, which may be seen as a distributed version of the Bellman-Ford algorithm, and which stills runs quite efficiently in an online fashion with the nodes computing in parallel. For the sake of simplicity, let's consider graphs of bounded degree. Rather than choosing $\log(n)$ many roots, it suffices to have each node choose to be a root with probability $\log(n)/n$. Then we can consider a stage by stage process, such that during each stage $s + 1$ and for each root r, each node u compares the shortest path $u_s(r)$ it has previously seen from u to r, with each of the paths given by taking the edge to a neighbour v and then following $v_s(r)$ (with information as to which nodes are roots disseminating simultaneously). If the graph is of diameter ℓ, then correct values are obtained after at most ℓ many of these update stages, or ℓ many stages after any changes to the graph. Each stage involves each node passing $O(\text{polylog}(n))$ many bits to each of its neighbours.

The complete bipartite graph suffices to show that one cannot improve upon the factor of stretch 3 in the observations above for general (undirected) graphs. By restricting to a certain natural class of graphs, however, we *can* remove the need for multiple roots (meaning that query response time is now $O(d(a,b))$), while simultaneously achieving stretch close to 2 for directed as well as undirected graphs.

The idemetric property. Recall that if X is a random variable and $\{X_i\}_{i\geq 0}$ is a sequence of random variables, then we say that X_n tends to X *in probability* if, for every $\epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

Definition 1.1. *Let $\mathbb{G}$ be a random network model, which produces for every $n \in \mathbb{N}$ (and perhaps taking other inputs, which we consider for now to be fixed) an ensemble of graphs with n nodes, i.e. a probability distribution over a certain set of graphs with n nodes. We say that $\mathbb{G}$ is **idemetric** if there exists a finite valued function f (i.e. $f(n) \neq \infty$), such that if a_n and b_n are nodes chosen uniformly at random from a graph generated according to $\mathbb{G}$ with n nodes, then $X_n = d(a_n, b_n)/f(n)$ tends to 1 in probability.*

For network models in which the graph produced may not be connected but does have a giant component, a weaker notion than idemetric is also useful.

Definition 1.2. *We say that a random network model $\mathbb{G}$ is **weakly idemetric** if both:*

1. *There exists $\kappa \in (0, 1]$ such that, with high probability (i.e. with probability tending to 1 as $n \to \infty$), the largest connected component of a graph generated according to $\mathbb{G}$ is of size at least κn.*
2. *There exists a function $f(n)$ such that if a_n and b_n are nodes chosen uniformly at random from the same largest connected component of a graph generated according $\mathbb{G}$ with n nodes, then $X_n = d(a_n, b_n)/f(n)$ tends to 1 in probability.*

So the idemetric property formalises the idea that most nodes in the graph have similar distances between them, while being weakly idemetric means that this holds for nodes in the giant component.

In fact it seems reasonable to expect *most* small-world network models to be idemetric. This becomes clear once one identifies the fact that most graph models which have the small-world property, have it for the same reason, which works roughly as follows. Suppose one picks a given node a at random, and considers the nodes at distance 1 from a, then the nodes at distance 2, and so on. If the graph model is small-world, then one expects that these successive generations can be accurately modelled as *something like* a branching process, with some reproductive rate $\mu > 1$ (giving of the order of μ^ℓ many nodes in the ℓ^{th} generation). This means that the number of nodes at a distance ℓ from a grows exponentially with ℓ until a reasonable fraction of the graph is attained, at which point the process ends quickly. As a result, most nodes will lie at a distance close to $\log_\mu(n)$ from a.

Classical results in graph theory (see [RD06] or [RH]) suffice to show, for example, that when the expected degree of each node is greater than 1, the Erdős-Rényi random graph $G(n, p)$ is weakly idemetric (although that term is not used in the literature). Similarly, without the term itself being applied, a range of inhomogeneous random graph models, such as the Chung-Lu model [CL02, CL03] and the Norros-Reittu model [NR06], have been shown to be weakly idemetric for a wide range of parameters [RH].

Now suppose that $\mathbb{G}$ is idemetric. For each n let G_n be a graph generated according to $\mathbb{G}$ with n nodes, and let r_n be a root node which is chosen uniformly at random in G_n. We can then apply any given solution to the SSSP problem with r_n as the source, and then again with edge directions reversed if the graph is directed, in order to find, for all $a \in G_n$, a shortest path $p(a, r_n)$ from a to r_n and a shortest path $p(r_n, a)$ from r_n to a. In network models with the small-world property, each node a can then store these short paths

$p(a, r_n)$ and $p(r_n, a)$, taking space at most $O(\text{polylog}(n))$ for each node. For a_n and b_n chosen uniformly at random from the nodes in G_n, let ℓ_n be the length of the path from a_n to b_n given by concatenating $p(a_n, r_n)$ and $p(r_n, b_n)$. Since $\mathbb{G}$ is idemetric, we then have that:

$$\frac{\ell_n}{d(a_n, b_n)} \to 2 \text{ in probability.}$$

So if $\mathbb{G}$ is idemetric then the all-pairs shortest path problem can be easily reduced to the single-source shortest path problem, so long as one is prepared to accept solutions which are of stretch close to 2 with high probability. If $\mathbb{G}$ is also small world then the SSSP problem can also be solved efficiently in a decentralised fashion, as described previously. Of course, if $\mathbb{G}$ is weakly idemetric, then similar results hold if we restrict to nodes in the giant component.

As evidence for the claim that most graph models aimed at capturing common properties of real-world networks will be idemetric, we prove this result for precisely those small-world graphs that Kleinberg gave negative results for in [JK00]. For the Watts-Strogatz model when $r < 2$, Kleinberg was able to make the crucial observation that, while short ($O(\log(n))$) paths will exist between all pairs of nodes with high probability, no efficient decentralised algorithms exist for finding short paths (in the absence of preprocessing)[2]. The notion of 'decentralised' which is used to get the negative results in that paper is somewhat more particular than the sense in which we have used the term here, since we allow each node to store some small amount of information about the graph structure after preprocessing – in fact that is the point of preprocessing. Before stating the result, let us define the model.

The Watts-Strogatz model. The model we consider is exactly the same as that in Kleinberg [JK00]. For any square number n, we begin with a set of nodes identified with the lattice points in a $\sqrt{n} \times \sqrt{n}$ square, $\{(x, y) : x \in \{1, 2, ..., \sqrt{n}\}, y \in \{1, 2, ..., \sqrt{n}\}\}$. The *lattice distance* between two nodes (x, y) and (r, s) is defined to be the number of lattice steps separating them when we fix periodic boundary conditions: $d_\ell((x, y), (s, t))$ is the minimum value of $|s' - x'| + |t' - y'|$, where the minimum is taken over all values $x' = x \pm \sqrt{n}, y' = y \pm \sqrt{n}, s' = s \pm \sqrt{n}, t' = t \pm \sqrt{n}$. For a universal constant $p \geq 1$, each node u has a directed edge to every other node within lattice distance p. For universal constants $q \geq 0$ and $r \geq 0$, we also construct directed edges from u to q other nodes (the long-range contacts) using independent random trials; the ith directed edge from u has endpoint v with probability proportional to $[d_\ell(u, v)]^{-r}$ (to obtain a probability distribution, we divide this quantity by the appropriate normalising constant[3]). This defines the network model $\mathbb{WS}(r, p, q)$. Note that the lattice distance between two nodes $d_\ell(a, b)$ may be quite different than $d(a, b)$. We shall prove the following:

Theorem 1.3. *For all $p, q \in \mathbb{N}$ with $p, q \geq 1$, and for all $r \in \mathbb{R}$ with $r \in [0, 2)$, the random network model $\mathbb{WS}(r, p, q)$ is idemetric.*

[2] Similar results also hold when $r > 2$ (but not for $r = 2$), but the graphs produced then are not small-world, and so are of less interest to us here.

[3] For the case $r = 0$, we fix the convention that $0^0 = 1$, so that the long-range outbound contact of a node u could be itself.

2 The proof of Theorem 1.3.

We consider first the case $p = q = 1$. Once we have dealt with this case, generalising to arbitrary $p, q \geq 1$ will then be straightforward. The proof for the case $p = q = 1$ is split into four sections A-D, and we then consider the general case $p, q \geq 1$ in Section E.

(A) The setup. Let a and b be nodes chosen uniformly at random from amongst the nodes in G, a graph with n nodes generated according to $\mathbb{SW}(r, p, q)$. We can consider the sets of nodes $A_1^*, A_2^*, \ldots$ and $B_1^*, B_2^* \ldots$, where:

$$A_i^* = \{u : d(a, u) = i - 1\}, \qquad B_i^* = \{u : d(u, b) = i - 1\}.$$

We are interested in finding the least i such that $b \in A_i^*$, and so are interested in the values $A_i = |A_i^*|$ (and later will also be interested in $B_i = |B_i^*|$). Clearly $A_1^* = \{a\}$, so $A_1 = 1$, and then A_2^* consists of the four lattice neighbours of a, together with the node which is the outbound long-range contact of a. As we continue to consider A_i^* and A_i for larger values of i, the process is complicated, however, by potential collisions: distinct nodes u and v may have outbound long-range contacts which are near to each other. The outbound long-range contact of a *could* be one of its four lattice neighbours, for example, or two of the lattice neighbours of a could have long-range contacts which are within distance one of each other. The latter possibility could then lead to double counting in A_4, unless one is careful. To give a useful upper bound for A_i, we therefore consider a simplified process in which, roughly speaking, every long-range contact appears on a new two dimensional lattice at an infinite lattice distance from the previous node. Formally, we can simply consider the sequence C_i, where $C_i = 0$ for $i < 0$ and, for $i \geq 0$:

$$C_0 = 1, \;\; C_{i+1} = C_i + \sum_{j \geq 1} C_{i-j} \cdot 4j = C_i + 4C_{i-1} + 8C_{i-2} + 12C_{i-3} \cdots \tag{2.0.1}$$

So, as depicted in Figure 1, $C_1 = 1, C_2 = 5, C_3 = 17, C_4 = 57$, and so on, and C_i is the value that A_i would take were it not for the possibility of collisions, as described above.

The enumeration of $\cup_i C_i^*$. It will also be useful to have a counterpart C_i^* to A_i^* and B_i^*, so that C_i^* is a set of nodes with $|C_i^*| = C_i$. To this end, we consider 3-dimensional points (x, y, z), which we shall call *3-nodes*, and specify that the lattice distance between (x, y, z) and (x', y', z') is infinite when $z \neq z'$, and is equal to the lattice distance between (x, y) and (x', y') otherwise. If $a = (x_a, y_a)$, we let $C_1^* = \{(x_a, y_a, 0)\}$. In order to enumerate C_{i+1}^*, we take each element $u = (x, y, z)$ of C_i^* in turn and proceed as follows.

1. Enumerate into C_{i+1}^* all 3-node lattice neighbours of u (i.e. those 3-nodes at lattice distance 1) which have not already been enumerated into $\cup_{j \leq i+1} C_j^*$.
2. Let $v = (x', y')$ be the outbound long-range contact of (x, y). Let z' be such that no nodes with third coordinate z' have been enumerated into $\cup_{j \leq i+1} C_j^*$ before, and enumerate (x', y', z') into C_{i+1}^* as the outbound long-range contact of u.

We say that the 3-node $u = (x, y, z) \in C_i^*$ *corresponds* to the node (x, y). Note that, due to collisions, it may be the case that a node (x, y) has several 3-nodes in C_i^* corresponding to it. The process above specifies an enumeration of $\cup_i C_i^*$, and we also consider it to specify an enumeration of $\cup_i A_i^*$ in the obvious way: nodes in this set are enumerated in the order of their first corresponding 3-nodes in $\cup_i C_i^*$. Distances between 3-nodes are defined in the standard way, in terms of the number of edges on the shortest directed path between

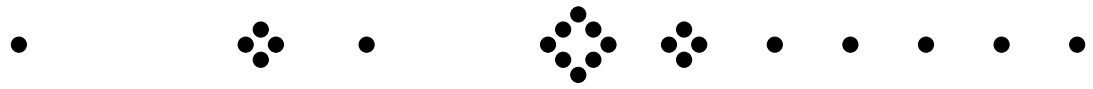

Figure 1: A schematic depiction of the growth of the sequence $\{C_i\}_{i\in\mathbb{N}}$.

two nodes: there are directed edges from each 3-node u to each of its lattice neighbours, i.e. those v such that $d_\ell(u, v) = 1$, and a directed edge from u to its outbound long-range contact. For 3-nodes u and v, we say that v is a descendant of u if u is a 3 node in C_i^*, v is a 3-node in C_j^* for some $j > i$ and $d(u, v) = j - i$. We will normally be interested in the descendants of u, only in the case that u is a long-range outbound contact.

It follows immediately from the definitions that for all i, $A_i \leq C_i$. Since $C_i = C_{i-1} + \sum_{j\geq 1} C_{i-j-1} \cdot 4j$ for $i > 0$, equation (2.0.1) can then be rewritten for $i > 0$:

$$C_{i+1} = 2C_i + 3C_{i-1} + 4\sum_{j\geq 2} C_{i-j}.$$

Given that, for $i > 2$, $C_{i-1} = 2C_{i-2} + 3C_{i-3} + 4\sum_{j\geq 2} C_{i-j-2}$, this in turn then gives, for $i > 2$:

$$C_{i+1} = 2C_i + 4C_{i-1} + 2C_{i-2} + C_{i-3}.$$

Standard techniques can then be applied in order to solve this linear recurrence relation. Since the characteristic polynomial

$$x^4 - 2x^3 - 4x^2 - 2x - 1 \tag{2.0.2}$$

has a largest root

$$\alpha \approx 3.38298, \tag{2.0.3}$$

it follows that for some constant $\rho > 0$:

$$\lim_{i\to\infty} \frac{C_i}{\rho \cdot \alpha^i} = 1. \tag{2.0.4}$$

Let us define:

$$\ell_n = \log_\alpha n.$$

It is immediate that $C_{i+1} \geq \sum_{j\leq i} C_j$. To within an additive constant, we then have that ℓ_n is the least k such that $\sum_{i=1}^{k} C_i \geq n$. Note that, for $i > 0$, $C_{i+1} \geq 2C_i$. For any $\epsilon > 0$, it follows that:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{(1-\epsilon)\ell_n} C_i}{n} = 0.$$

Since $\sum_{i=1}^{(1-\epsilon)\ell_n} A_i \leq \sum_{i=1}^{(1-\epsilon)\ell_n} C_i$, we conclude that, for any $\epsilon > 0$, with high probability:

$$d(a, b) > (1 - \epsilon)\ell_n.$$

To prove the theorem, it thus suffices to show that, for any $\epsilon > 0$, the following holds with high probability:

$$d(a, b) < (1 + \epsilon)\ell_n. \tag{2.0.5}$$

(B) Proving (2.0.5). First of all we establish a useful bound for the distribution on long-range contacts. Suppose that we are given arbitrary nodes u and v, but that we do not know the value of u' which is the outbound long-range contact of u. Let $\ln(n)$ denote the natural logarithm. Then we shall show that, irrespective of the given relative positions of u and v, the fact that $r \in [0, 2)$ implies:

$$\mathbb{P}(u' = v) \geq \frac{1}{4n\ln(n)}. \tag{2.0.6}$$

In order to establish (2.0.6), note that $\mathbb{P}(u' = v)$ is minimised by taking v at a maximum possible lattice distance from u, and by taking r as large as possible. So it suffices to prove the result for $r = 2$, and when v is at a maximum lattice distance from u. In that case:

$$\mathbb{P}(u' = v) = d_\ell(u, v)^{-2} / \sum_{v' \neq u} d_\ell(u, v')^{-2}. \tag{2.0.7}$$

Then we have:

$$\sum_{v' \neq u} d_\ell(u, v')^{-2} \leq \sum_{j=1}^{\sqrt{n}} (4j)(j^{-2}) = 4 \sum_{j=1}^{\sqrt{n}} j^{-1} \leq 4(1 + \ln(\sqrt{n})) \leq 4\ln(n), \tag{2.0.8}$$

where the second inequality follows from the general bound $\sum_{j=1}^{x} j^{-1} \leq (1 + \ln(x))$. Then (2.0.7) and (2.0.8) combine to give (2.0.6), as required.

Towards proving (2.0.5), fix ϵ with $0 < \epsilon < 1/6$, and for the remainder of the proof let $\ell^* = \frac{(1+\epsilon)\ell_n}{2}$ (note that we drop the subscript n for convenience). The basic idea, as depicted in Figure 2, is that so long as $A^*_{\ell^*}$ and $B^*_{\ell^*}$ are reasonably large, it is very likely the case that one of the nodes in $A^*_{\ell^*}$ has some node in $B^*_{\ell^*}$ as an outbound long-range contact. This gives a short path from a to b. More precisely, what we do is to show that for any constant $\mu > 1$, the following both hold with high probability:

$$A_{\ell^*} > \mu n^{1/2} \ln(n) \qquad \text{and} \qquad B_{\ell^*} > \mu n^{1/2} \ln(n). \tag{2.0.9}$$

If this holds for $\mu > 4$ then either $\bigcup_{i=1}^{\ell^*} A^*_i$ and $\bigcup_{i=1}^{\ell^*} B^*_i$ already have non-empty intersection, which gives the required short path from a to b, or else we can apply (2.0.6) to conclude that the probability every member of $A^*_{\ell^*}$ will fail to have a member of $B^*_{\ell^*}$ as a long-range contact is bounded above by:

$$\left(1 - \frac{4n^{1/2}\ln(n)}{4n\ln(n)}\right)^{\mu n^{1/2}} = \left(1 - \frac{1}{n^{1/2}}\right)^{\mu n^{1/2}} \to e^{-\mu} \text{ as } n \to \infty.$$

Given that μ was arbitrary, (2.0.9) therefore suffices to give (2.0.5). It remains to establish (2.0.9).

(C) Proving (2.0.9) for A_{ℓ^*}. We deal first with A_{ℓ^*} – achieving the lower bound for B_{ℓ^*} then uses most of the same ideas but is complicated by the fact that nodes do not have a fixed number of inbound long-range contacts. We deal with the case for B_{ℓ^*} in Section D of the proof. To achieve the lower bound for A_{ℓ^*}, we make some new definitions. Given the enumeration of $\cup_i C^*_i$ specified previously, we say $u \in C^*_{i+1}$ is *collision causing* if both:

(a) It is a long-range outbound contact of an element of C^*_i, and corresponds to a node in $\bigcup_{j=1}^{i+1} A^*_j$ which is within lattice distance $\log_2(n)$ of another previously enumerated element of $\bigcup_{j=1}^{i+1} A^*_j$, and;

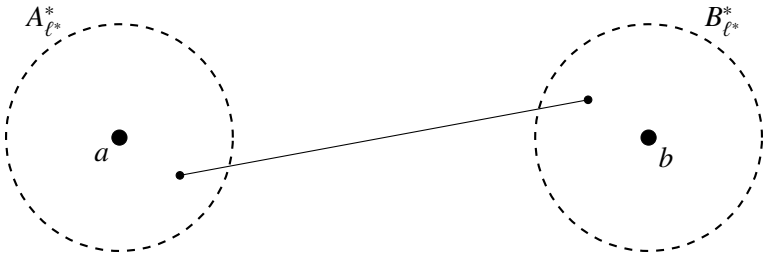

Figure 2: If A_{ℓ^*} and B_{ℓ^*} are much larger than $n^{1/2}\ln(n)$ there is likely to be an edge between them, giving a short path from a to b.

(b) It is not the descendant of a 3-node which is already collision causing.

Since we follow the process for less than $\log_2(n)/2$ many steps (i.e. $\ell^* < \log_2(n)/2$), long-range outbound contacts at a lattice distance $> \log_2(n)$ from all other nodes will not lead to collisions of the kind discussed previously. We then define the *discounted* 3-nodes, to be all those 3-nodes in $\bigcup_{i=1}^{\ell^*} C^*_i$ which are either collision causing, or else are descendants of a collision causing 3-node. The key point of these definitions is that any two 3-nodes in $\bigcup_{i=1}^{\ell^*} C^*_i$ which are not discounted correspond to distinct nodes in $\bigcup_{i=1}^{\ell^*} A^*_i$. For each i, let $\mathtt{nd}(C^*_i)$ be all those 3-nodes in C^*_i which are not discounted, and define $\mathtt{nd}(C_i) = |\mathtt{nd}(C^*_i)|$. The basic idea is now that we want to show, for all sufficiently large n and for all $i \leq \ell^*$, that $\mathtt{nd}(C_i)$ is unlikely to be very much smaller than C_i. Since distinct nodes in $\mathtt{nd}(C^*_i)$ correspond to distinct nodes in A^*_i, and since it follows from the definition of ℓ^* that $C_{\ell^*} = \Omega(n^{(1+\epsilon)/2})$, this will suffice to give the probabilistic lower bound for A_{ℓ^*} in (2.0.9).

In order to give the probabilistic lower bound for $\mathtt{nd}(C_i)$ just discussed, we will need to bound the probability that a given long-range contact is collision causing. So suppose that we are given an arbitrary node u and an arbitrary set of nodes C. While we know which node u is, and we know the elements of C, suppose that we do not know the value of u', which is the outbound long-range contact of u. We want to show that if $|C| \leq n^{2/3}$, then:

$$\mathbb{P}(u' \in C) = o(1). \tag{2.0.10}$$

The proof of (2.0.10) is given in the appendix. Using this bound, we can now provide the probabilistic lower bound for $\mathtt{nd}(C_i)$ discussed previously. Let α be defined as before, in (2.0.3). What we want to show is that, for every $\alpha_0 < \alpha$, the following holds with high probability:

$$\mathtt{nd}(C_{\ell^*}) > \alpha_0^{\ell^*}. \tag{2.0.11}$$

Given (2.0.11), let $\alpha_0 < \alpha$ be such that $\log_\alpha \alpha_0 > (1+\epsilon/2)/(1+\epsilon)$. Then with high probability:

$$A_{\ell^*} \geq \mathtt{nd}(C_{\ell^*}) > \alpha_0^{\ell^*} = (\alpha^{\log_\alpha \alpha_0})^{(1+\epsilon)\ell_n/2} > \alpha^{(1+\epsilon/2)\ell_n/2} = n^{\frac{1}{2}+\frac{\epsilon}{4}}.$$

For any constant $\mu > 0$, the last term $n^{\frac{1}{2}+\frac{\epsilon}{4}}$ is greater than $\mu n^{1/2}\ln(n)$ for all sufficiently large n. So this gives (2.0.9) for A_{ℓ^*}, as required.

The rough picture. To prove (2.0.9) for A_{ℓ^*}, it therefore remains to use (2.0.10) in order to establish (2.0.11). So suppose given α_0 such that $0 < \alpha_0 < \alpha$. The rough thinking as to why (2.0.11) should hold is as follows. We know from (2.0.4) that for large enough i, C_{i+1}/C_i is close to α. Now suppose that we have been given C^*_i (for $i < \ell^*$), and that as we enumerate each long-range outbound contact in C^*_{i+1}, we ask whether or not it is collision causing. From (2.0.10) it follows that for large enough n, each time that we ask this question,

the probability that the given node will be collision causing is less than $(\alpha - \alpha_0)/\alpha$. Suppose that there were m nodes in C_i, and for the sake of simplicity let's imagine that none of them were collision causing. Then (even forgetting about the fact that most nodes will not be enumerated into C_{i+1} as long range contacts and so *cannot* be collision causing) we can expect that the number of nodes we enumerate into C_{i+1} and which are not collision causing, is at least $(\alpha_0/\alpha)m\alpha = m\alpha_0$. So this approximate argument would seem to give exponential growth of rate at least α_0, for all sufficiently large n. While intuitively appealing, this argument quickly becomes complicated, however, when one tries to make it precise in a direct way. The description above did not properly address the relationships between nodes in C_i and C_{i+1}, and the number of nodes in C_{i+1} which are descendants of each node in C_i. If u is a long-range outbound contact in C_i, for example, then u will have 5 descendants in C_{i+1}. So if u is collision causing then this produces more than α many discounted nodes at the next level. In fact it is far more effective to formalise a variant of this same argument, which works with those modified forms of the characteristic polynomial (2.0.2) which result when we are only guaranteed a certain proportion of long range contacts.

The precise argument. Consider the version of the recurrence relation (2.0.1) which results when, in each generation, the nodes in C_i^* only have $(1-\beta)C_i$ many long-range outbound contacts (for $\beta < 1$). We'll use E_i to denote the new resulting sequence of values:

$$E_0 = 1/(1-\beta), \quad E_{i+1} = (1-\beta)E_i + \sum_{j \geq 1} E_{i-j} \cdot 4j(1-\beta). \tag{2.0.12}$$

Of course $(1-\beta)E_i$ may not be integer valued, but the sequence given by this recurrence relation is meaningful in giving a lower bound to the number of non-discounted 3-nodes there will be in each generation, in a context where we always have some integer number $\geq (1-\beta)\mathtt{nd}(C_i)$ of non-collision causing long-range outbound contacts for elements of $\mathtt{nd}(C_i^*)$. The same algebraic manipulations as before can be applied, in order to reduce (2.0.12) to:

$$E_{i+1} = (2-\beta)E_i + (4-3\beta)E_{i-1} + (2-3\beta)E_{i-2} + (1-\beta)E_{i-3}.$$

This is gives a characteristic equation which is a function of β:

$$f(x,\beta) = x^4 - (2-\beta)x^3 - (4-3\beta)x^2 - (2-3\beta)x - (1-\beta). \tag{2.0.13}$$

Now in a neighbourhood of $x = \alpha$, $\beta = 0$, the derivates of f with respect to x and β are both positive, meaning that the largest root of f is a continuous decreasing function of β in some neighbourhood of this point. For each $\alpha_0 < \alpha$ sufficiently close to α, it follows that there exists $\beta > 0$ for which α_0 is the largest root of $f(x,\beta)$. Similarly, for each $\alpha_0 > \alpha$ sufficiently close to α, there exists $\beta < 0$ for which α_0 is the largest root of $f(x,\beta)$.

So suppose given $\alpha_0 < \alpha$ and ϵ' with $0 < \epsilon' < 0.5$. To prove (2.0.11) it suffices to show, for all sufficiently large n, that $\mathtt{nd}(C_{\ell^*}) > \alpha_0^{\ell^*}$ with probability $> 1 - \epsilon'$. Choose α_1 and β, with $\alpha_0 < \alpha_1 < \alpha$, $0 < \beta < 0.5$, and such that α_1 is the largest root of $f(x,\beta)$. Suppose we are given $\mathtt{nd}(C_i^*)$ for some $i < \ell^*$. For all sufficiently large n, the fact that we chose $\epsilon < 1/6$ in the definition of ℓ^* means that:

$$\left|\cup_{j \leq \ell^*} C_j^*\right| < n^{2/3}. \tag{2.0.14}$$

Then, according to (2.0.10), for sufficiently large n the number of outbound long-range contacts of elements of $\mathtt{nd}(C_i^*)$ which are collision causing, is stochastically dominated by:

$$X \sim \mathrm{Bin}(\mathtt{nd}(C_i), \beta/2).$$

Applying the Chernoff bound to this stochastically dominating binomial, we conclude that for some $I_\beta > 0$, and for all sufficiently large n, the probability that more than a β proportion of the outbound long-range contacts of elements of $\mathtt{nd}(C_i^*)$ are collision causing is bounded above by:

$$e^{-I_\beta \mathtt{nd}(C_i)}.$$

Now $\mathtt{nd}(C_i)$ is guaranteed to grow at a certain rate, since $\mathtt{nd}(C_i) \geq 4(i-1)$ for $i \geq 2$. We can therefore choose i_0 such that $e^{-I_\beta 4(i_0-1)} < \epsilon'/2$, and this means that with probability 1,

$$e^{-I_\beta \mathtt{nd}(C_{i_0})} < \epsilon'/2.$$

We chose $\beta < 0.5$. So if at most a β proportion of the long-range contacts of elements of $\mathtt{nd}(C_{i_0}^*)$ are collision causing, we have $\mathtt{nd}(C_{i_0+1}) \geq 1.5\mathtt{nd}(C_{i_0})$. Since $\epsilon' < 0.5$ this gives:

$$e^{-I_\beta \mathtt{nd}(C_{i_0+1})} < (\epsilon'/2)^{1.5} = \epsilon' \sqrt{2\epsilon'}/4 < \epsilon'/4.$$

Then so long as at most a β proportion of the long-range contacts of elements of $\mathtt{nd}(C_{i_0+1}^*)$ are collision causing, we have $e^{-I_\beta \mathtt{nd}(C_{i_0+2})} < \epsilon'/8$, and so on. Define $\Pi_{\beta,\epsilon'}$ to be the following event: for all i with $i_0 \leq i < \ell^*$, at most a β proportion of the long-range contacts of elements of $\mathtt{nd}(C_i^*)$ are collision causing. The analysis above then gives, for sufficiently large n:

$$\mathbb{P}(\Pi_{\beta,\epsilon'}) > 1 - \epsilon'.$$

To complete the argument, consider all those 3-nodes at a lattice distance of $i_0 - 1$ from a. These are the 3-nodes we considered above, which are guaranteed to belong to $\mathtt{nd}(C_{i_0}^*)$. So long as $\Pi_{\beta,\epsilon'}$ holds, we can give a lower bound for the number of descendants of these nodes which belong to $\mathtt{nd}(C_{i_0+i-1}^*)$ for each $i > 0$ through the series:

$$E_0 = 4(i_0-1)/(1-\beta) \quad E_{i+1} = (1-\beta)E_i + \sum_{j \geq 1} E_{i-j} \cdot 4j(1-\beta). \tag{2.0.15}$$

There exists a constant $\rho > 0$ such that $\lim_{i\to\infty} E_i/(\rho\alpha_1^i) = 1$. Since $\alpha_0 < \alpha_1$, we can choose $i_1 > i_0$ such that $\mathtt{nd}(C_{\ell^*}^*) > \alpha_0^{\ell^*}$ whenever $\Pi_{\beta,\epsilon'}$ holds and $\ell^* > i_1$. This gives (2.0.11) as required.

(D) Proving (2.0.9) for B_{ℓ^*}. As remarked on previously, establishing (2.0.9) for B_{ℓ^*} is complicated by the fact that nodes do not have a fixed number of inbound long-range contacts. We begin by defining the sets D_i^*, which are to B_i^* as C_i^* is to A_i^*. As we enumerate the sets D_i^*, it is useful to monitor whether or not we have considered the long-range inbound contacts of each node before: all nodes start as *unseen* and will be labelled *seen* during the enumeration once we have considered their long-range inbound contacts. If $b = (x_b, y_b)$, we let $D_1^* = \{(x_b, y_b, 0)\}$. In order to enumerate D_{i+1}^*, we take each element $u = (x, y, z)$ of D_i^* in turn and proceed as follows.

1. Enumerate into D_{i+1}^* all 3-node lattice neighbours of u which have not already been enumerated into $\cup_{j \leq i+1} D_j^*$.
2. We divide into two cases. If (x, y) is not labelled as *seen* then proceed according to (a) below. Otherwise proceed according to (b).

(a) Let $v_1, \ldots, v_k$ be the inbound long-range contacts of (x, y), where $v_j = (x_j, y_j)$. For each $j \in [1, k]$ in turn, let z_j be such that no 3-nodes with third coordinate z_j have been enumerated into $\cup_{j \leq i+1} D^*_j$ before, and enumerate (x_j, y_j, z_j) into D^*_{i+1} as an inbound long-range contact of u. Label (x, y) as *seen*.

(b) Let k be sampled from a distribution $\mathrm{Bin}(n, 1/n)$ (independent from all other distributions considered). Let $z_1, ..., z_k$ be distinct and such that, for each $i \in [1, k]$, no node with third coordinate z_j has been enumerated into $\cup_{j \leq i+1} D^*_j$ before. For each $i \in [1, k]$ enumerate $(0, 0, z_i)$ into D^*_{i+1} as an inbound long-range contact of u.

We let $D_i = |D^*_i|$. The definitions we gave previously for collision causing nodes, discounted nodes and descendants are applied to the 3-nodes in $\bigcup_i D^*_i$ in the obvious way – replacing "outbound" everywhere with "inbound", replacing A^*_i with B^*_i, and replacing $d(u, v)$ in the definition of descendant with $d(v, u)$ (we are now interested in 3-nodes as elements of D^*_i rather than C^*_i, so these definitions *replace* rather than extend the previous ones).

In order to prove (2.0.9) for B_{ℓ^*}, it suffices to establish the following analogue of (2.0.11). For every $\alpha_0 < \alpha$, the following holds with high probability:

$$\mathtt{nd}(D_{\ell^*}) > \alpha_0^{\ell^*}. \tag{2.0.16}$$

The proof would proceed much as it did for (2.0.11), except that there are now *two* non-deterministic factors influencing $\mathtt{nd}(D_i)$: as well as the fact that 3-nodes may or may not be discounted, they may also have from 0 to n many inbound long-range contacts (0 to $n-1$ if $r \neq 0$). Whereas previously it followed directly from (2.0.4) that (2.0.14) held for all sufficiently large n, now we have to provide further proof that with high probability:

$$\left|\cup_{j \leq \ell^*} D^*_j\right| < n^{2/3}. \tag{2.0.17}$$

Only after (2.0.17) is established will we be in a position to conclude that, with high probability, the probabilistic bound (2.0.10) can be applied at all stages $i < \ell^*$. To prove (2.0.16) we therefore first have to prove that, for every $\alpha_1 > \alpha$, the following holds with high probability:

$$\left|\cup_{j \leq \ell^*} D^*_j\right| < \alpha_1^{\ell^*}. \tag{2.0.18}$$

In order to see that (2.0.18) gives (2.0.17), choose $\alpha_1 > \alpha$ such that $\log_\alpha(\alpha_1) < 4/(3(1+\epsilon))$. Then (2.0.18) gives that with high probability:

$$\left|\cup_{j \leq \ell^*} D^*_j\right| < \alpha_1^{\ell^*} = (\alpha^{\log_\alpha \alpha_1})^{\ell^*} < \alpha^{\frac{4(1+\epsilon)\ell_n}{6(1+\epsilon)}} = n^{2/3}.$$

We define $\ell^\dagger$ to be the minimum of ℓ^* and the first i such that $|\cup_{j \leq i+1} D^*_j| \geq n^{2/3}$. In order to establish that with high probability $\ell^\dagger = \ell^*$ and (2.0.18) holds, we can argue much as in the proof of (2.0.11). So suppose given $\alpha_1 > \alpha$ and ϵ' with $0 < \epsilon' < 0.5$. To prove (2.0.18) it suffices to show, for all sufficiently large n, that $\left|\cup_{j \leq \ell^*} D^*_j\right| < \alpha_1^{\ell^*}$ with probability $> 1 - \epsilon'$. Choose α_2 and β, with $\alpha < \alpha_2 < \alpha_1$, $0 < \beta < 0.5$, and such that α_2 is the largest root of $f(x, -\beta)$, where f is as defined in (2.0.13). Now suppose we are given $\cup_{j \leq i} D^*_j$ for some $i < \ell^\dagger$ — so for some $i < \ell^\dagger$ we are told precisely the elements of D^*_j for each $j \leq i$ (but we are not told the value of $\ell^\dagger$). Before we are told the long-range inbound contacts of a given node $u \in D^*_i$ during the enumeration of D^*_{i+1}, we do not know the outbound long range contacts of any nodes other than some of those we have enumerated into $\cup_{j \leq i+1} D^*_i$ already, and the outbound long-range contacts of all other

nodes remain uniformly and independently distributed amongst the unseen nodes. The number of inbound long-range contacts of elements of D_i^* is therefore stochastically dominated by X, which is the sum of D_i i.i.d. random variables, each with distribution $\mathrm{Bin}(n, 1/(n-n^{2/3}))$. At the same time, the number of inbound long-range contacts of elements of D_i^* stochastically dominates X', which is the sum of D_i i.i.d. random variables, each with distribution $\mathrm{Bin}(n-n^{2/3}, 1/n)$. Applying Chernoff bounds, we conclude that for some $I_\beta > 0$, and for all sufficiently large n, the probability that the number of inbound long-range contacts of elements of D_i^* is outside the interval $[(1-\beta)D_i, (1+\beta)D_i]$ is bounded above by[4]:

$$e^{-I_\beta D_i}.$$

We can then continue to argue much as in the proof of (2.0.11). Let i_0 be such that $e^{-I_\beta 4(i_0-1)} < \epsilon'/4$. We chose $\beta < 0.5$. So if the number of inbound long-range contacts of elements of D_i^* is at least $(1-\beta)D_i$, we have $D_{i_0+1} \geq 1.5D_{i_0}$, which means that $e^{-I_\beta D_{i_0+1}} < \epsilon'/8$, and so on. We can also choose N_0 which depends on ϵ' but is independent of n, and such that $\mathbb{P}(D_{i_0} > N_0) < \epsilon'/2$. Define $\Pi_{\beta,\epsilon'}$ to be the following event: $D_{i_0} \leq N_0$ and, for all i with $i_0 \leq i < \ell^\dagger$, the number of inbound long-range contacts of elements of D_i^* is in the interval $[D_i(1-\beta), D_i(1+\beta)]$. The analysis above then gives, for sufficiently large n:

$$\mathbb{P}(\Pi_{\beta,\epsilon'}) > 1-\epsilon'.$$

Consider the recurrence relation:

$$E_0 = N_0/(1-\beta) \quad E_{i+1} = (1-\beta)E_i + \sum_{j\geq 1} E_{i-j}\cdot 4j(1-\beta).$$

So long as $\Pi_{\beta,\epsilon'}$ holds, we have:

$$\forall i \in [i_0, \ell^\dagger], \quad D_i \leq E_{i-i_0+1}.$$

Since there exists $\rho > 0$ for which

$$\lim_{i\to\infty} \frac{E_{i-i_0+1}}{\rho\alpha_2^i} = 1,$$

and since $\alpha_2 < \alpha_1$, we can choose i_1 such that:

$$\forall i \in [i_1, \ell^\dagger], \quad \left|\cup_{j\leq i} D_j^*\right| < \alpha_1^i. \tag{2.0.19}$$

We also have that $E_{\ell^*} < n^{2/3}$ for all sufficiently large n, which means that so long as $\Pi_{\beta.\epsilon'}$ holds and n is sufficiently large, $\ell^\dagger = \ell^*$. Thus (2.0.19) gives (2.0.18), as required.

With (2.0.17) established, we can then prove (2.0.16) in almost exactly the same way that we proved (2.0.11). The proof is given in the appendix.

(E) The case $p, q \geq 1$. The generalised form of the recurrence relation (2.0.1) is given as follows. For $i < 0$, $C_i = 0$. Then $C_0 = 1/q$ and, for $i \geq 0$:

$$C_{i+1} = qC_i + \sum_{j\geq 1} \left(4p^2(i-1/2) + 2p\right) qC_{i-j}.$$

[4]Here I_β has a different but similar definition to its previous use in the proof of (2.0.11). Similarly $\Pi_{\beta,\epsilon}$ has a similar but different definition here in the proof of (2.0.18).

Using this expression for C_i, we get that, for $i > 0$:

$$C_{i+1} = (q+1)C_i + (2p^2 + 2p - 1)qC_{i-1} + \sum_{j \geq 2} 4p^2 qC_{i-j}.$$

For $i > 2$, this means that $C_{i-1} = (q+1)C_{i-2} + (2p^2 + 2p - 1)qC_{i-3} + \sum_{j \geq 2} 4p^2 qC_{i-j-2}$. For $i > 2$ we then have:

$$C_{i+1} = (q+1)C_i + \Big((2p^2 + 2p - 1)q + 1\Big) C_{i-1} + (4p^2 q - q - 1)C_{i-2} + (2p^2 - 2p + 1)qC_{i-3}.$$

The largest root α of the corresponding characteristic polynomial is the same as the largest eigenvalue of the matrix:

$$\boldsymbol{M} = \begin{bmatrix} q+1 & (2p^2 + 2p - 1)q + 1 & 4p^2 q - q - 1 & (2p^2 - 2p + 1)q \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

That the largest eigenvalue α of $\boldsymbol{M}$ is positive and real (and occurs with multiplicity 1) follows directly from the Perron-Frobenius Theorem for non-negative matrices. In a manner precisely analogous to the proof for $p = q = 1$, we conclude that, for some constant $\rho > 0$:

$$\lim_{i \to \infty} \frac{C_i}{\rho \cdot \alpha^i} = 1.$$

The remainder of the proof then goes through with only the obvious required modifications. The definitions of C_i^* and D_i^* have to be adjusted to incorporate the increased number of neighbours, for example, and the expression $\log_2(n)$ must be replaced throughout with $p\log_2(n)$. The expected number of inbound long-range contacts for each node is now q, and distributions on the number of inbound long-range contacts must be adjusted accordingly. The general form of (2.0.13) is:

$$x^4 - (q(1-\beta)+1)x^3 - \Big((2p^2 + 2p - 1)q(1-\beta) + 1\Big) x^2 - (4p^2 q(1-\beta) - q(1-\beta) - 1)x - (2p^2 - 2p + 1)q(1-\beta).$$

Again we have that if α is the largest root then in a neighbourhood of $x = \alpha, \beta = 0$, the derivates of f with respect to x and β are both positive.

References

[AJB99] Réka Albert, Hawoong Jeong & Albert-László Barabási. The diameter of the World Wide Web. *Nature*, 401, 130, 1999.

[CL02] Fan Chung & Linyuan Lu. The average distance in a random graph with given expected degrees. *Proceedings of the National Academy of Sciences USA*, 99 (25), 15879–15882, 2002.

[CL03] Fan Chung & Linyuan Lu. The average distance in a random graph with given expected degrees. *Internet Mathematics*, 1 (1), 91–113.

[ED59] Edsger Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1: 269–271, 1959.

[RD06] Rick Durrett. Random Graph Dynamics. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, New York, USA, 2006.

[RF62] Robert Floyd. Algorithm 97. *Communications of the Association for Computing Machinery*, volume 5, issue 6, page 345, 1962.

[RH16] Remco van der Hofstad. Random Graphs and Complex Networks. Volume 1. Cambridge University Press, 2016.

[RH] Remco van der Hofstad. Random Graphs and Complex Networks. Volume 2. Cambridge University Press, to appear.

[JK00] Jon Kleinberg. The small world phenomenon: an algorithmic perspective. *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 163–170, 2000.

[NR06] Ilkka Norros & Hannu Reittu. On a conditionally Poissonian graph process. *Advances in Applied Probability*, 38 (1), 59–75.

[MT99] Mikkel Thorup. Undirected single-source shortest paths with positive integer weights in linear time. *Journal of the Association for Computing Machinery*, 46 (3): 362–394, 1999.

[TZ05] Mikkel Thorup & Uri Zwick. Approximate distance oracles. *Journal of the Association for Computing Machinery*, 52 (1), 1–24, 2005.

[SW62] Stephen Warshall. A theorem on Boolean matrices. *Journal of the Association for Computing Machinery*, 9:11–12, 1962.

[WS98] Duncan Watts & Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 440–442, 1998.

[RW14] Ryan Williams. Faster all-pairs shortest paths via circuit complexity. *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 664–673, 2014.

3 Appendix

The proof of (2.0.10). Recall that we are given an arbitrary node u and an arbitrary set of nodes C. While we know which node u is, and we know the elements of C, we do not know the value of u', which is the outbound long-range contact of u. We want to show that if $|C| \leq n^{2/3}$, then $\mathbb{P}(u' \in C) = o(1)$.

In order to prove (2.0.10), we define the following sets of nodes:

$$H_i = \{x : \ 2^{i-1} \leq d_\ell(u, x) < 2^i\}, \qquad I_k = \{x : \ d_\ell(u, x) = k\}.$$

The number of nodes in $\bigcup_{j=1}^{i} H_j$ is $2^{i+1}(2^i - 1)$, so long as n is sufficiently large. We say that H_i is *complete* for n when the size of $\bigcup_{j=1}^{i} H_j$ takes this maximum value. Similarly, we say that I_k is complete for n when it is of size $4k$. In what follows we shall be concerned with $\mathbb{P}(u' \in H_i)$ and $\mathbb{P}(u' \in I_k)$ for various i and k. When we consider such values, the implicit assumption will always be that n is sufficiently large that H_i and I_k are complete for n.

To establish (2.0.10), our aim is to show that, for each $r \in [0, 2)$, there exists $\kappa_r > 0$ such that:

$$\frac{\mathbb{P}(u' \in H_{i+1})}{\mathbb{P}(u' \in H_i)} \geq 2^{\kappa_r} \text{ for all sufficiently large } i. \tag{3.0.1}$$

If H_i is complete for n then $|\cup_{j=1}^{i} H_j| = 2^{i+1}(2^i - 1)$. Note also that $\lim_{i\to\infty} 2^{i+1}(2^i-1)/(2\cdot 4^i) = 1$. It follows that for $\ell = \log_4(n)$ there exists a constant c such that, for all sufficiently large n:

$$\left|\cup_{j=1}^{\frac{2}{3}\ell+c} H_j\right| > n^{2/3}, \text{ while simultaneously } H_{\ell-c} \text{ is complete for } n.$$

Given (3.0.1), this means that (2.0.10) holds when C is chosen to be a set of $n^{2/3}$ many nodes which are as near as possible to u, and therefore holds for all C with $|C| \leq n^{2/3}$.

To prove (3.0.1), choose β with $0 < \beta < 2 - r$. For each $I_k \subseteq H_i$ consider $I_{2k} \cup I_{2k+1} \subseteq H_{i+1}$. For all sufficiently large k:

$$\frac{\mathbb{P}(u' \in I_{2k+1})}{\mathbb{P}(u' \in I_{2k})} > 2^{-\beta}. \tag{3.0.2}$$

Also:

$$\frac{\mathbb{P}(u' \in I_{2k})}{\mathbb{P}(u' \in I_k)} = 2^{1-r}. \tag{3.0.3}$$

Combining (3.0.2) and (3.0.3), we get that for all sufficiently large k:

$$\frac{\mathbb{P}(u' \in I_{2k+1}) + \mathbb{P}(u' \in I_{2k})}{\mathbb{P}(u' \in I_k)} > 2^{2-r-\beta},$$

so $\kappa_r = 2 - r - \beta$ suffices.

The proof of (2.0.16) given (2.0.18). The proof is very similar to the proof of (2.0.11). Suppose given $\alpha_0 < \alpha$ and ϵ' with $0 < \epsilon' < 0.5$. To prove (2.0.16) it suffices to show, for all sufficiently large n, that $\mathtt{nd}(D_{\ell^*}) > \alpha_0^{\ell^*}$ with probability $> 1 - \epsilon'$. Choose α_1 and β, with $\alpha_0 < \alpha_1 < \alpha$, $0 < \beta < 0.5$, and such that α_1 is the largest root of $f(x, \beta)$, where f is as in (2.0.13). Let $\ell^\dagger$ be the minimum of ℓ^* and the first i such that $|\cup_{j\leq i+1} D_j^*| \geq n^{2/3}$. By (2.0.18) it holds with high probability that $\ell^\dagger = \ell^*$. Suppose we are given $\mathtt{nd}(D_i^*)$ for some $i < \ell^\dagger$. Then, according to (2.0.10), for sufficiently large n the number of inbound long-range contacts of elements of $\mathtt{nd}(D_i^*)$ which are not collision causing, stochastically dominates X which is the sum of $\mathtt{nd}(D_i)$ many i.i.d random variables, each with mean $1 - \beta/2$. Applying the Chernoff bound to X, we conclude that for some $I_\beta > 0$, and for all sufficiently large n, the probability that we fail to have at least $(1-\beta)\mathtt{nd}(D_i)$ many long-range inbound contacts of elements of $\mathtt{nd}(D_i^*)$ which are not collision causing is bounded above by:

$$e^{-I_\beta \mathtt{nd}(D_i)}.$$

Now $\mathtt{nd}(D_i)$ is guaranteed to grow at a certain rate, since $\mathtt{nd}(D_i) \geq 4(i-1)$ for $i \geq 2$. We can therefore choose i_0 such that $e^{-I_\beta 4(i_0-1)} < \epsilon'/2$, and this means that with probability 1,

$$e^{-I_\beta \mathtt{nd}(D_{i_0})} < \epsilon'/2.$$

We chose $\beta < 0.5$. So if at most a β proportion of the long-range contacts of elements of $\mathtt{nd}(D_{i_0}^*)$ are collision causing, we have $\mathtt{nd}(D_{i_0+1}) \geq 1.5\mathtt{nd}(D_{i_0})$. Since $\epsilon' < 0.5$ this gives:

$$e^{-I_\beta \mathtt{nd}(D_{i_0+1})} < (\epsilon'/2)^{1.5} = \epsilon'\sqrt{2\epsilon'}/4 < \epsilon'/4.$$

Then so long as at least $(1-\beta)\mathtt{nd}(D_{i_0+1})$ many inbound long-range contacts of elements of $\mathtt{nd}(D^*_{i_0+1})$ are not collision causing, we have $e^{-I_\beta \mathtt{nd}(D_{i_0+2})} < \epsilon'/8$, and so on. Define $\Pi_{\beta,\epsilon'}$ to be the following event: for all i with $i_0 \leq i < \ell^\dagger$, at least $(1-\beta)\mathtt{nd}(D_i)$ many inbound long-range contacts of elements of $\mathtt{nd}(D^*_i)$ are not collision causing. The analysis above then gives, for sufficiently large n:

$$\mathbb{P}(\Pi_{\beta,\epsilon'}) > 1-\epsilon'.$$

To complete the argument, consider all those 3-nodes at a lattice distance of $i_0 - 1$ from b. These 3-nodes are guaranteed to belong to $\mathtt{nd}(D^*_{i_0})$. So long as $\Pi_{\beta,\epsilon'}$ holds, we can give a lower bound for the number of descendants of these nodes which belong to $\mathtt{nd}(D^*_{i_0+i-1})$ for each $i \leq \ell^\dagger - i_0 + 1$ through the series:

$$E_0 = 4(i_0-1)/(1-\beta) \quad E_{i+1} = (1-\beta)E_i + \sum_{j\geq 1} E_{i-j}\cdot 4j(1-\beta). \tag{3.0.4}$$

There exists a constant $\rho > 0$ such that $\lim_{i\to\infty} E_i/(\rho\alpha_1^i) = 1$. Since $\alpha_0 < \alpha_1$, we can choose $i_1 > i_0$ such that $\mathtt{nd}(D^*_{\ell^*}) > \alpha_0^{\ell^*}$ whenever $\Pi_{\beta,\epsilon'}$ holds, $\ell^\dagger = \ell^*$ and $\ell^* > i_1$. This gives (2.0.16) as required.