On the average path length of deterministic and stochastics recursive networks

Philippe Giabbanelli, Dorian Mazauric, Jean-Claude Bermond

On the average path length of deterministic and stochastics recursive networks*

Philippe J. Giabbanelli, Dorian Mazauric, and Jean-Claude Bermond

Mascotte, INRIA, I3S(CNRS,UNS), Sophia Antipolis, France,
{Philippe.Giabbanelli,Dorian.Mazauric,Jean-Claude.Bermond}@sophia.inria.fr

Abstract. The average shortest path distance ℓ between all pairs of nodes in real-world networks tends to be small compared to the number of nodes. Providing a closed-form formula for ℓ remains challenging in several network models, as shown by recent papers dedicated to this sole topic. For example, Zhang *et al.* proposed the deterministic model ZRG and studied an upper bound on ℓ. In this paper, we use graph-theoretic techniques to establish a closed-form formula for ℓ in ZRG. Our proof is of particular interests for other network models relying on similar recursive structures, as found in fractal models. We extend our approach to a stochastic version of ZRG in which layers of triangles are added with probability p. We find a first-order phase transition at the critical probability $p_c = 0.5$, from which the expected number of nodes becomes infinite whereas expected distances remain finite. We show that if triangles are added independently instead of being constrained in a layer, the first-order phase transition holds for the very same critical probability. Thus, we provide an insight showing that models can be equivalent, regardless of whether edges are added with grouping constraints. Our detailed computations also provide thorough practical cases for readers unfamiliar with graph-theoretic and probabilistic techniques.

1 Introduction

The last decade has witnessed the emergence of a new research field coined as "Network Science". Amongst well-known contributions of this field, it was found that the average distance ℓ in a myriad of real-world networks was small compared to the number of nodes (*e.g.*, in the order of the logarithm of the number of nodes). Numerous models were proposed for networks with small average distance [1, 2] such as the static Watts-Strogatz model, in which a small percentage of edges is changed in a low-dimensional lattice [3], or dynamic models in which ℓ becomes small as nodes are added to the network [4]. However, proving a closed form formula for ℓ can be a challenging task in a model, and thus this remains a current research problem with papers devoted to this sole task [5]. In this paper, we prove a closed form formula for a recently proposed model, in which the authors showed an upper bound on ℓ [6]. While the model presented

* Research funded by the EULER project and *région PACA*.

in [6] is deterministic, a stochastic version was also studied for which the authors approximated an upper bound on ℓ [8]. Thus, we present two stochastic versions and we rigorously characterize their behaviour using both upper and lower bounds on ℓ, and studying the ratio with the number of nodes.

The paper starts by establishing the notation, then each of the three Sections focusses on a model. Firstly, we consider the model as defined in [6]: we prove a closed-form formula for the average distance ℓ, and characterize the ratio between the number of nodes and ℓ. Secondly, we propose a version of the model in which edges and nodes are randomly added but in specific groups. In this version, we establish bounds on the expected value of ℓ and we provide a closed-form formula for the expected number of nodes. While the former is always finite, the latter becomes infinite from a critical probability $p_c = 0.5$, thus the ratio between ℓ and the number of nodes can be arbitrarily large. However, the infinite value of the expected number of nodes results from a few very large instances, and thus does not represent well the trend expressed by most instances for $p \geq p_c$. Consequently, we also study the ratio between the number of nodes and ℓ by considering all instances but very large ones. Thirdly, we study the number of nodes and ℓ in a stochastic version that does not impose specific groups, similarly to [8]. We show that this version also has a finite expected value for ℓ, and an infinite expected number of nodes from $p = p_c$.

2 Notation

We denote by ZRG_t the undirected graph defined by Zhang, Rong and Guo, obtained at step t [6]. It starts with ZRG_0 being a cycle of three nodes, and "ZRG_t is obtained by ZRG_{t-1} by adding for each edge created at step $t-1$ a new node and attaching it to both end nodes of the edge" [6]. The process is illustrated by Figure 1(a). We propose two probabilistic versions of ZRG. In the first one, each of the three original edges constitutes a *branch*. At each time step, a node is added *for all* active edges of a branch with independent and identical (*iid*) probability p. If a branch does not grow at a given time step, then it will not grow anymore. We denote this model by $BZRG_p$, for the probabilistic *branch* version of ZRG with probability p. Note that while the probability p is applied at each time step, the resulting graph is not limited by a number of time steps as in ZRG_t: instead, the graph grows as long as there is at least a branch for which the outcome of the stochastic process is to grow, thus there exist arbitrarily large instances. The process is illustrated in Figure 1(b). Finally, the second stochastic version differs from $BZRG_p$ by adding a node with probability p *for each* active edge. In other words, this version does not impose to grow all the 'layer' at once, but applies the probability edge by edge. We denote the last version by $EZRG_p$ for the probabilistic *edge* version of ZRG with probability p.

In this paper, we are primarily interested in the average distance. For a connected graph G having a set of nodes $V(G)$, its average distance is defined by $\ell(G) = \frac{\sum_{u \in V(G)} \sum_{v \in V(G)} d(u,v)}{|V(G)|*(|V(G)|-1)}$, where $d(u,v)$ is the length of a shortest path between u and v. In a graph with N nodes, ℓ is said to be *small* when proportional to $\ln(N)$, and *ultrasmall* when proportional to $\ln(\ln(N))$ [9].

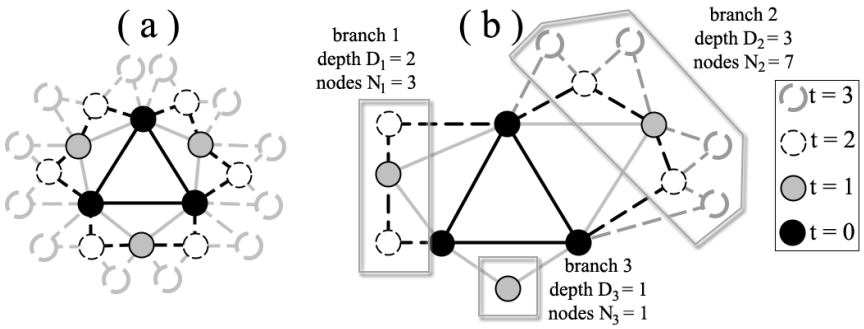

Fig. 1. The graph ZRG_0 is a triangle, or cycle of 3 nodes. At each time step, to each edge added at the previous time step, we add a node connected to the endpoints of the edge. This can be seen as adding a triangle to each outer edge. The process is depicted step by step up to ZRG_3 (a). A possible instance of $BZRG_p$ illustrates the depth and number of nodes in each of the three probabilistic branches (b). The graph grew for 3 time steps, at the end of which the outcome of the last active branch was not to grow.

3 Deterministic version

In this Section, we consider the version introduced by [6] and defined in the previous Section. We denote by V_t the vertices of ZRG_t, and A_t the vertices added at step t. We established in [7] that $|A_t| = 3 * 2^{t-1}$ for $t \geq 1$.

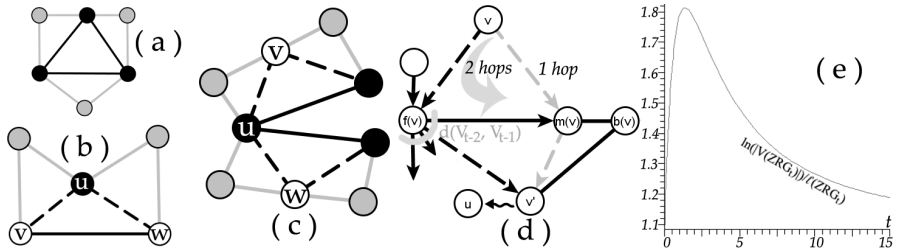

Fig. 2. In ZRG_1, each of the black initial nodes is connected to two grey added nodes (a). We assume that $u \in A_{t-1}$, thus it stems from an edge (v, w). As the edges (u, v) and (u, w) are active, u will also be connected to two nodes in A_t (b). We assume that $u \in V_{t-2}$: thus, it is connected to two (children) nodes $v, w \in A_{t-1}$. The edges (u, v) and (u, w) being active, u will also be connected to the two nodes they generate at the next step, belonging to A_t (c). Shortest paths to compute $d(A_t, V_{t-1})$ (d). The average distance in ZRG_t is close to the logarithm of the graph's size (e).

By construction, a node $u \in A_t$ is connected to the two endpoints of a formerly active edge. One endpoint was created at the step immediately before (*i.e.*, $t-1$), and we call it the *mother* $m(u) \in A_{t-1}$, while the other endpoint was created at an earlier time, and we call it the *father* $f(u) \in V_{t-2}$. A node having

same mother as u is called its uterine brother and denoted as $b(u)$. Furthermore, we observe that each node $v \in V_{t-1}$ is connected to two nodes of A_t. This is proved by induction: it holds for $t = 1$ (see Figure 2(a)), we assume it holds up to $t-1$ and we show in Figure 2(b-c) that it follows for t. Since each node in A_t is connected to two nodes in V_{t-1}, and each node in V_{t-1} is connected to two nodes in A_t, the graph has a bipartite structure used in our proof.

We now turn to the computation of $\ell(ZRG_t)$. We denote by $d(X,Y) = \sum_{u \in X} \sum_{v \in Y} d(u,v)$ the sum of distances from all nodes in X to all nodes in Y. Theorem 1 establishes the value of $g(t) = d(V_t, V_t)$, from which we will establish the average distance using $\ell(ZRG_t) = \frac{g(t)}{|V(ZRG_t)|*(|V(ZRG_t)|-1)}$. We gave the sketch of a different proof in [7], thus the interested reader can compare it with the full proof given here to better illustrate graph-theoretic techniques.

Theorem 1. $g(t) = 4^t(6t+3) + 2 * 3^t$

Proof. By definition, $V_t = V_{t-1} \cup A_t$. Thus, $d(V_t, V_t) = d(V_{t-1}, V_{t-1}) + d(A_t, V_{t-1}) + d(V_{t-1}, A_t) + d(A_t, A_t)$. Since the underlying graph is undirected, $d(A_t, V_{t-1}) = d(V_{t-1}, A_t)$ hence

$$g(t) = g(t-1) + 2d(A_t, V_{t-1}) + d(A_t, A_t), t \geq 2 \tag{1}$$

In the following, we consider that a shortest path from $v \in A_t$ always goes through $f(v)$, unless the target is the brother $b(v)$ or the mother $m(v)$ in which case correction factors are applied. Suppose that we instead go through $m(v)$ to reach some node u: since $m(v)$ is only connected to $b(v)$, $f(v)$ and some node v' (see Figure 2(d)) then the route continues through v'. However, the path $v, m(v), v'$ can be replaced by $v, f(v), v'$ without changing the length.

We compute $d(A_t, V_{t-1})$ for $t \geq 2$. Since we always go through $f(v)$, we use a path of length 2 in order to go from v to $m(v)$ whereas it takes 1 using the direct link. Thus, we have to decrease the distance by 1 for each $v \in A_t$, hence a correcting factor $-|A_t|$. We observe that each node in V_{t-2} is the father of two nodes in A_t, hence routing through the father costs $2d(V_{t-2}, V_{t-1})$ to which we add the number of times we use the edge from v to the father. As each $v \in A_t$ goes to each $w \in V_{t-1}$, the total cost is

$$d(A_t, V_{t-1}) = 2d(V_{t-2}, V_{t-1}) + |A_t||V_{t-1}| - |A_t| \tag{2}$$

We have that $2d(V_{t-2}, V_{t-1}) = 2d(V_{t-2}, V_{t-2}) + 2d(V_{t-2}, A_{t-1}) = 2g(t-2) + 2d(V_{t-2}, A_{t-1})$. Furthermore, using Eq. 1 we obtain $g(t-1) = g(t-2) + 2d(A_{t-1}, V_{t-2}) + d(A_{t-1}, A_{t-1}) \Leftrightarrow 2d(A_{t-1}, V_{t-2}) = g(t-1) - g(t-2) - d(A_{t-1}, A_{t-1})$. Substituting these equations with Eq. 2, it follows that

$$d(A_t, V_{t-1}) = g(t-1) + g(t-2) - d(A_{t-1}, A_{t-1}) + |A_t||V_{t-1}| - |A_t| \tag{3}$$

We compute $d(A_t, A_t)$, for $t \geq 2$. In order to go from v to its uterine brother $b(v) \in A_t$, it takes 2 hops through their shared mother, whereas it takes 3 hops through the father. Thus, we have a correction of 1 for $|A_t|$ nodes. The path from a v to a w is used four times, since $f(v)$ has two children in A_t and so does

$f(w)$. Finally, we add 2 for the cost of going from a node to its father at both ends of the path, and we have $|A_t|(|A_t|-1)$ such paths. Thus, the overall cost is

$$d(A_t, A_t) = 4g(t-2) + 2|A_t|(|A_t|-1) - |A_t| \quad (4)$$

We combine. Given that $|A_t| = |V_{t-1}|$, we substitute Eq. 3 into Eq. 1 hence

$$g(t) = 3g(t-1) + 2g(t-2) + d(A_t, A_t) - 2d(A_{t-1}, A_{t-1}) + 2|A_t|^2 - 2|A_t| \quad (5)$$

From Eq. 4, for $t \geq 3$ we obtain $d(A_{t-1}, A_{t-1}) = 4g(t-3)+2|A_{t-1}|^2-3|A_{t-1}|$. Given that $|A_t| = 2|A_{t-1}|$, we substitute $d(A_{t-1}, A_{t-1})$ and Eq. 4 in Eq. 5:

$$g(t) = 3g(t-1) + 6g(t-2) - 8g(t-3) + 3|A_t|^2 - 2|A_t|, t \geq 3 \quad (6)$$

We manually count that $f(0) = 6$, $f(1) = 42$ and $f(2) = 252$. Thus the equation can be solved into $g(t) = 4^t(6t+3) + 3 * 2^t$ using standard software.

Corollary 1. *Since* $|V(ZRG_t)| = 3 * 2^t$ *[6], it follows from the Theorem that* $\ell(ZRG_t) = \frac{4^t(6t+3)+3*2^t}{3*2^t(3*2^t-1)} = \frac{t2^{t+1}+2^t+1}{3*2^t-1}$.

Using this corollary, we obtain $\lim_{t\to\infty} \frac{\ln(|V(ZRG_t)|)}{\ell(ZRG_t)} = \frac{3*\ln(2)}{2} \approx 1.03$ and $\lim_{t\to\infty} \frac{\ln(\ln(|V(ZRG_t)|))}{\ell(ZRG_t)} \approx 0$. Thus, the average size is almost exactly $\ln(|V(G)|)$ for large t. Since the size of the graph is exponential in t, it is important that the graphs obtained for small values of t have a similar ratio, which is confirmed by the behaviour illustrated in Figure 2(e).

4 Stochastic branch version

As in the previous Section, we are interested in the ratio between the number of nodes and the average path length. In this Section, our approach is in three steps. Firstly, we establish bounds on the *depth* of branches, defined as the number of times that a branch has grown. Secondly, we study the number of nodes. We find that the number of nodes undergoes a first-order phase transition at the critical probability $p_c = 0.5$: for $p < 0.5$, there is a finite number of nodes, whereas for $p \geq 0.5$ this number becomes infinite. Since in the latter the expected depth of branches is bounded by finite numbers, the expected graphs have an arbitrarily small average distance compared to the number of nodes. However, the expected number of nodes only provides a mean-field behaviour that can may lack representativeness due to a few very large instances. Thus, we conclude by investigating the behavior of instances of bounded depth.

4.1 Depth of branches

To fully characterize the depth of branches, we are interested in their expected depth for the standard case as well as the two extremal cases consisting of the *deepest* and *shallowest* branches. In other words, we study the mean-field behaviour and we provide a lower and an upper bound. We start by introducing our notation for the depth in Definition 1. We start by establishing the expected depth of a branch in Theorem 2, then we turn to the expected shallowest depth, and we conclude in Theorem 4 showing that the expected deepest depth of a branch is finite.

Definition 1. *We denote D_1, D_2, D_3 the depth of the three branches. The depth of the deepest branch is $D_{max} = max(D_1, D_2, D_3)$ and the depth of the shallowest branch is $D_{min} = min(D_1, D_2, D_3)$.*

Theorem 2. *The expected depth of a branch is $\mathbb{E}(D_i) = \frac{p}{1-p}$, $i \in \{1, 2, 3\}$.*

Proof. The probability $P(D_i = k)$ that a branch grows to depth k is the probability p^k of successily growing k times, and the probability not to grow once (*i.e.*, to stop at depth $k+1$). Thus, $P(D_i = k) = \underbrace{p \cdot \cdots \cdot p}_{k} \cdot (1-p) = p^k(1-p)$. Since the expected value of a discrete random variable D_i is given by $\mathbb{E}(D_i) = \sum_{k=0}^{\infty}(kP(D_i = k))$, it follows that $\mathbb{E}(D_i) = \sum_{k=0}^{\infty}(kp^k(1-p)) = p(1-p)\sum_{k=0}^{\infty}(kp^{k-1})$. Since $\frac{dp^k}{dp} = kp^{k-1}$, we further simplify into $\mathbb{E}(D_i) = p(1-p)\sum_{k=0}^{\infty}(\frac{dp^k}{dp})$. As the sum of a derivative is equal to the derivative of the sum, it follows that $\mathbb{E}(D_i) = p(1-p)\frac{d\sum_{k=0}^{\infty} p^k}{dp}$. We note that $\sum_{k=0}^{\infty} p^k$ is an infinite geometric sum, hence

$$\mathbb{E}(D_i) = p(1-p)\frac{d}{dp}\frac{1}{1-p} = \frac{p(1-p)}{(1-p)^2} = \frac{p}{1-p}$$

Theorem 3. $\mathbb{E}(D_{min}) = -\frac{p^3}{p^3-1}$

Proof. The probability of the shallowest depth to be at least k knowing that the probability p applies iid to each branch is $P(D_{min} \geq k) = P(D_1 \geq k)P(D_2 \geq k)P(D_3 \geq k) = p^{3k}$. By definition, $P(D_{min} = k) = P(D_{min} \geq k) - P(D_{min} \geq k+1^{)}$, thus $P(D_{min} = k) = p^{3k} - p^{3(k+1)}$. This probability is plugged into the definition of the expected value as in Theorem 2 hence

$$\mathbb{E}(D_{min}) = \sum_{k=0}^{\infty}(kP(D_{min} = k)) = \sum_{k=0}^{\infty}(k(p^{3k} - p^{3(k+1)})) = -\frac{p^3}{p^3-1}$$

Theorem 4. $\mathbb{E}(D_{max}) = -\frac{p(p^3+4p^2+3p+3)}{(p-1)(p^2+p+1)(p+1)}$.

Proof. By construction, the deepest branch does not exceed k iff none of the branches has a depth exceeding k. Since the probability p applies iid to all three branches, we have $P(D_{max} \leq k) = P(D_1 \leq k)P(D_2 \leq k)P(D_3 \leq k)$. Furthermore, a branch is strictly deeper than k if it successfully reaches depth $k+1$. Thus, $P(D_i > k) = \underbrace{p \cdot \cdots \cdot p}_{k+1} = p^{k+1}$, $i \in \{1, 2, 3\}$. By algebraic simplification, we have $P(D_{max} \leq k) = (1 - P(D_1 > k))(1 - P(D_2 > k))(1 - P(D_3 > k)) = (1 - p^{k+1})^3$. By definition, $P(D_{max} = k) = P(D \leq k) - P(D \leq k-1) = (1 - p^{k+1})^3 - (1 - p^k)^3$. Given that $\mathbb{E}(D_{max}) = \sum_{k=0}^{\infty}(kP(D_{max} = k))$, we replace the expression of $P(D_{max} = k)$ to obtain $\mathbb{E}(D_{max}) = \sum_{k=0}^{\infty}(k((1 - p^{k+1})^3 - (1 - p^k)^3))$. The final expression results from algebraic simplification using standard software.

4.2 Average number of nodes

We introduce our notation in Definition 2, and Theorem 5 provides a closed-form of the expected number of nodes.

Definition 2. *We denote by N_1, N_2, and N_3 the number of nodes in the three branches. Since we start from a cycle with three nodes, the total number of nodes is $N = N_1 + N_2 + N_3 + 3$.*

Theorem 5. *For $p < \frac{1}{2}$, the expected number of nodes is $\mathbb{E}(N) = \frac{3(1-p)}{1-2p}$.*

Proof. First, we focus on the expected number of nodes in a branch. As the probability p applies iid to all three branches, we select the first branch without loss of generality. By construction, the total number of nodes N_1 in the branch 1 at depth $D_1 = k \geq 0$ is $N_1 = 2^k - 1 = \sum_{i=1}^{k} 2^{i-1}$. Thus, the expected value of the random variable N_1 is given by $\mathbb{E}(N_1) = \sum_{k=0}^{\infty}((2^k - 1)P(D_1 = k))$. As shown in Theorem 2, $P(D_1 = k) = p^k(1-p)$. We replace it in the equation to obtain $\mathbb{E}(N_1) = \sum_{k=0}^{\infty}((2^k - 1)p^k(1-p))$. After expanding the equation, we have $\mathbb{E}(N_1) = \sum_{k=0}^{\infty}(2^k p^k(1-p) - p^k(1-p)) = (1-p)\sum_{k=0}^{\infty}((2p)^k) - (1-p)\sum_{k=0}^{\infty}(p^k)$. As in Theorem 2, we have $\sum_{k=0}^{\infty}(p^k) = \frac{1}{1-p}$ thus the second term simplifies and yields $\mathbb{E}(N_1) = (1-p)\sum_{k=0}^{\infty}((2p)^k) - 1$. It is well known that a serie of the type $\sum_{k=0}^{\infty}(x^k)$ diverges to infinity for $x \geq 1$. Thus, our serie diverges for $2p \geq 1 \Leftrightarrow p \geq \frac{1}{2}$. In other words, this results only holds for $0 \leq p < \frac{1}{2}$.
The infinite geometric sum $\sum_{k=0}^{\infty}((2p)^k)$ can be simplified in $\frac{1}{1-2p}$ hence $\mathbb{E}(N_1) = \frac{1-p}{1-2p} - 1 = \frac{1-p-1+2p}{1-2p} = \frac{p}{1-2p}$. The probability p applies iid to all three branches hence $\mathbb{E}(N_1) = \mathbb{E}(N_2) = \mathbb{E}(N_3)$. Thus, by Definition 2, the expected number of nodes in the overall graph is given by $\mathbb{E}(N) = 3 + 3\mathbb{E}(N_1) = 3 + \frac{3p}{1-2p} = \frac{3(1-p)}{1-2p}$.

Theorem 5 proved that the average number of nodes diverges to infinity at the critical probability $p_c = 0.5$. This may appear to be a discrepancy with Theorem 4 stating that the expected depth of the deepest branch is finitely bounded. For the sake of clarity, we provide an intuition and an example on this point. First, we note that the *expected* deepest depth and the *expected* number of nodes have different growth rates. Indeed, even if graphs with very deep depth scarcely occur at $p = 0.5$, their impact on the expected number of nodes is tremendous since the number of nodes grows *exponentially* with the depth. On the other hand, the impact of such graphs on the expected deepest depth is only linear. To illustrate different growth rates with a known network, consider the complete graph K_n, in which each of the $n \geq 1$ nodes is connected to all others. In K_n, the number of nodes grows linearly whereas the distance is constant. Thus, the distance between two nodes is 1 even with an infinite number of nodes.

In a nutshell, the expected number of nodes for $p \geq 0.5$ may not be representative of most instances due to the large impact of very deep graphs. Thus, it remains of interest to investigate the number of nodes for graphs with bounded depths. This is established in Theorem 6, in which we consider the $q\%$ possible instances with smallest depth. By lowering the impact of the very deep graphs, this theorem consistutes a lower bound that better describes practical cases.

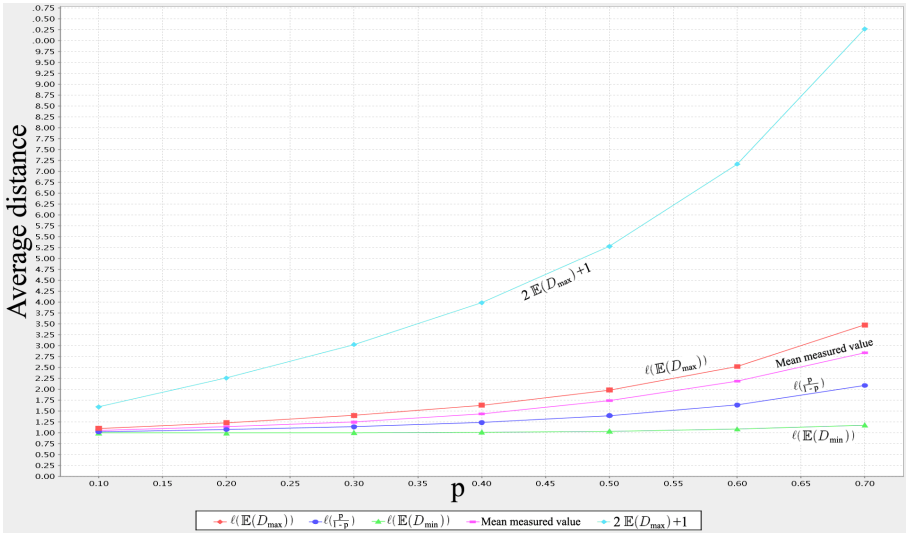

Fig. 3. The measured average distance compared with three bounds and an estimate, proved or conjectures. The simulations valide the conjectures. *Color online.*

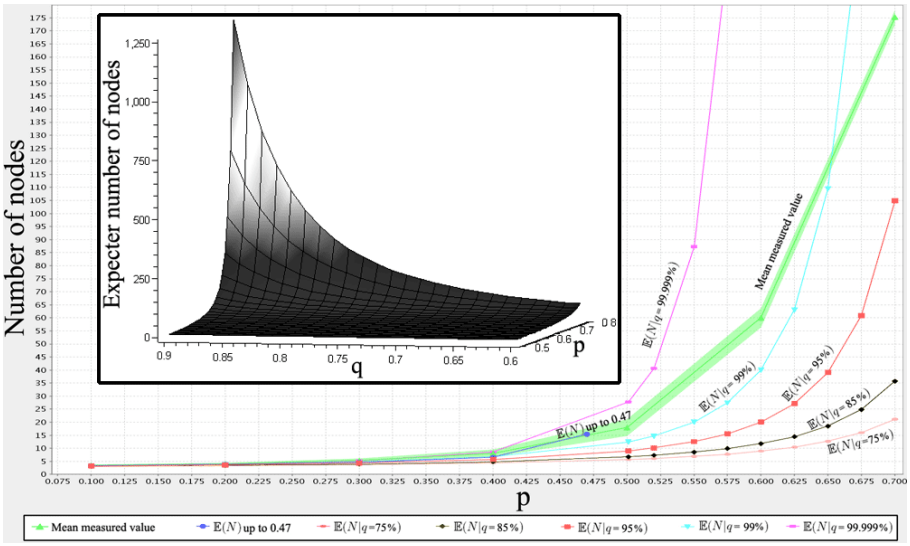

Fig. 4. Up to $p = 0.5$ (excluded), the expected number of nodes is finite. From $p = 0.5$, the expected number of nodes is infinite due to very large instances, thus we provide a finite estimate by considering the $q\%$ smallest instances, from $q = 75\%$ to $q = 100 - 10^{-3}\%$. Simulations were limited to the graphs that fitted in memory thus the mean measured value represents only an estimate based on small graphs for values beyond $p = 0.5$. *Color online.*

Lemma 1. $\mathbb{E}(N|D_{max} \leq K) = -\frac{(3p-6p^{K+2}2^K+6p^{K+1}2^K-3)}{(2p-1)(p^{K+1}-1}$

Proof. The expected number of nodes is adapted from Theorem 5 by the following substitutions: $\mathbb{E}\underbrace{(N_1)}_{(N_1|D_{max}\leq K)} = \sum_{k=0}^{\infty}((2^k-1)\underbrace{P(D_1=k)}_{P(D_1=k|D_{max}\leq K)})$. We simplify:

$$P(D_1 = k|D_{max} \leq K) \overset{\text{formula}}{=} \frac{P(D_1=k\cap D_{max}\leq K)}{P(D_{max}\leq K)}$$

$$\overset{\text{expend } D_{max}}{=} \frac{P(D_1=k\cap D_1\leq K\cap D_2\leq K\cap D_3\leq K)}{P(D_1\leq K\cap D_2\leq K\cap D_3\leq K)}$$

$$\overset{\text{independence}}{=} \frac{P(D_1=k)P(D_1\leq K)P(D_2\leq K)P(D_3\leq K)}{P(D_1\leq K)P(D_2\leq K)P(D_3\leq K)} \overset{\text{simplifying}}{=} \frac{P(D_1=k)}{P(D_1\leq K)}$$

Thus $\mathbb{E}(N_1|D_{max} \leq K) = \sum_{k=0}^{\infty}((2^k-1)\frac{P(D_1=k)}{P(D_1\leq K)})$. We showed in the proof of Theorem 4 that $P(D_1 \leq K) = 1 - p^{K+1}$, and we showed in the proof of Theorem 2 that $P(D_1 = k) = p^k(1-p)$. By substituting these results, and using from the previous Theorem that $\mathbb{E}(N) = 3 + 3\mathbb{E}(N_1)$, it follows that

$$\mathbb{E}(N|D_{max} \leq K) = 3 + 3\sum_{k=0}^{K}(\frac{(2^k-1)p^k(1-p)}{1-p^{K+1}})$$

The closed form formula follows by algebraic simplification.

Theorem 6. *By abuse of notation, we denote by $\mathbb{E}(N|q)$ the expected number of nodes for the q% of instances of $BZRG_p$ with smallest depth. We have*

$$\mathbb{E}(N|q) = -\frac{3(-1+(1-\sqrt[3]{q})^{\frac{\ln(2)+\ln(p)}{\ln(p)}})(p-1)}{-\sqrt[3]{q}(2p-1)}$$

Proof. Theorem 4 proved that the expected deepest depth of a branch was at most K with probability $(1-p^{K+1})^3$. Thus, if we want this probability to be q, we have to consider branches whose depth K is at most:

$$(1-p^{K+1})^3 = q \Leftrightarrow K = log_p(1-\sqrt[3]{q}) - 1$$

The Theorem follows by replacing K in Lemma 1 with the value above.

The effect of a few percentages of difference in q is shown in Figure 4 together with the results from Theorem 5 and 6. In the inset of Figure 4, we show that the number of nodes grows sharply with q.

4.3 Average path length

We conducted experiments in order to ensure the veracity of the results presented in the two previous Sections, and to compare them with devised bounds. For values of p from 0.1 to 0.7 by steps of 0.1, we measured the average distance of the resulting graphs, obtained as the average over 1000 instances. In Figure 3, we plot it against four bounds and an estimated mean:

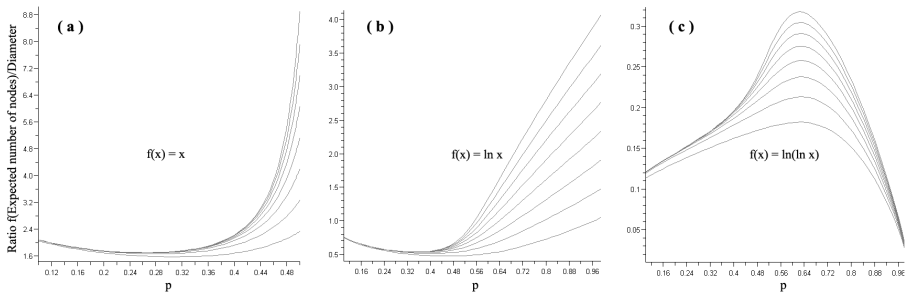

Fig. 5. We measured the ratio between $\mathbb{E}(N|100-10^x)$ (a), $\ln(\mathbb{E}(N|100-10^x))$ (b), $\ln(\ln(\mathbb{E}(N|100-10^x)))$ (c) and the diameter $2\mathbb{E}(D_{max})+1$ for x going from 0 (bottom curve) to 7 (top curve). We determined a critical probability $p_c = 0.5$ at which the regime changes, and this is confirmed by these measures showing that the average distance goes from linear in the number of nodes (a) for $p < p_c$ to small in the number of nodes (b) for $p \geq p_c$.

- *Proven bound.* Theorem 4 established the expected deepest depth. At any time, the graph has three branches, and we can only go from one branch to another through the basic cycle. Thus, the expected maximum distance between any two points in the graph consists of going from the most remote node of two branches to the cycle, and going through the cycle. As a node is at most at distance $\mathbb{E}(D_{max})$ from the cycle and we can go from any node of the cycle to another in one hop, the expected maximum distance is $2\mathbb{E}(D_{max}) + 1$. Since this is the *maximum* distance, we use it as a proven upper bound on the *average* distance.
- *Conjecture bounds.* Our intuition is that since $\ell(t)$, proven in Theorem 1, provides the average distance for a graph in which *all* branches have depth t, then a lower and upper bound can be obtained by considering the graphs with shallowest (Theorem 3) and deepest (Theorem 4) depths respectively. This is confirmed by the simulations.
- *Conjectured mean.* Similarly to the conjecture bounds, we have proven the expected depth $\mathbb{E}(D) = \frac{p}{1-p}$ of a branch in Theorem 2, and the simulation confirms that $\ell(\frac{p}{1-p})$ constitute an estimate of the average distance.

As we previously did for the deterministic version, we now investigate whether the average distance $\ell(BZRG_p)$ can be deemed small compared to the number of nodes $|V|$. As explained in the previous Section, we proved a (first-order) phase transition at the critical probability $p_c = 0.5$. The behaviour of the graph can be characterized using the ratios displayed in Figure 5: for $p \ll p_c$, we observe an average distance proportional to the number of nodes, and for $p > p_c - \epsilon$ the average distance is proportional to the logarithm of the number of nodes which is deemed small. The ratio in Figure 5(c) is too low, and tends to 0 for a large probability p, thus the graph cannot be considered ultra-small. The separation at $p_c - \epsilon$ can also be understood from a theoretical perspective. For $p \geq p_c$, we proved that $\ell(BZRG_p)$ can be arbitrary small compared to $|V|$ since $\ell(BZRG_p)$

is finite whereas $|V|$ is infinite. When $p = 0.5-\varepsilon$, the average distance is bounded by a finite number: by Theorem 2 we have that the expected depth of a branch is $\mathbb{E}(D_i) < 1$ and, using the aforementioned argument regarding the maximum distance, this entails $\ell(BZRG_p) < 2 * 1 + 1 = 3$. Furthermore, as stated in the proof of Theorem 6, the expected number of nodes in a branch is $\mathbb{E}(N_i) = \frac{0.5-\varepsilon}{\varepsilon}$ which can thus be arbitrarily large. Thus, the behaviour for $p \geq p_c$ is also expected to hold in a small neighborhood of the critical probability.

5 Stochastic edge version

In order to show a broad variety of approaches, we prove the number of nodes and the average path length of $EZRG_p$ using different tools from the previous Section. Theorem 7 establishes the number of nodes in the graph.

Theorem 7. *For $p < \frac{1}{2}$, the expected number of nodes is $\mathbb{E}(N) = 3 + \frac{3p}{1-2p}$.*

Proof. We consider the dual of the graph, which we define using a visual example in Figure 6. The dual is a binary tree: for an edge, a triangle is added (root of the tree) with probability p, to which two triangles are added iid at the next step (left and right branches of the tree) with probability p, and so on. Since one node is added to the tree when a node is added to the original graph, studying the number of nodes in $EZRG_p$ is equivalent to studying the number of nodes in the **t**ree. We denote the latter by $t(p)$. The number of nodes starting from any edge is the same, and we denote it by N. Thus, N corresponds to the probability of starting a tree (*i.e.*, adding a first node that will be the root) multipled by the number of nodes in the tree: $N = pt(p)$. Once the tree has been started, the number of nodes corresponds to the sum of the root and the number of nodes in the two branches, hence $t(p) = 2pt(p) + 1$. Note that there is a solution if and only if $p < \frac{1}{2}$, and otherwise the number of nodes is infinite. By arithmetic simplification, we obtain $t(p)(1-2p) = 1 \Leftrightarrow t(p) = \frac{1}{1-2p}$. Thus, $N = \frac{p}{1-2p}$. Since the graph starts as a cycle of length three, the number of counts corresponds to the three original nodes, to which we add the number of nodes starting in each of the three trees, thus $\mathbb{E}(N) = 3 + \frac{3p}{1-2p} = \frac{3(1-p)}{1-2p}$.

A proof similar to Theorem 7 could be applied to the number of nodes in $BZRG_p$. However, the tree associated with $BZRG_p$ grows by a complete level with probability p, instead of one node at each time. Thus, the current depth of the tree has to be known by the function in order to add the corresponding number of nodes, hence $N = pt(p, 0)$ and $t(p, k) = pt(p, k + 1) * 2^k$.

In the previous model, we showed that the expected average distance had a constant growth whereas the expected number of nodes had an exponential growth. Thus, the gap between the two could be arbitrarily large. Using simulations reported in Figure 7, we show that the same effect holds in this model.

6 Conclusion and future work

We proved a close-form formula for the average distance in ZRG_t. We proposed two stochastic versions, and showed that they had a first-order phase transition

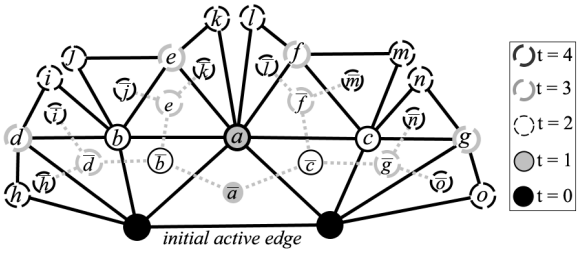

Fig. 6. Nodes linked by black edges correspond to four successive growths from an initial active edge. When a node x is added, it creates a triangle, to which we associate a node $\bar{x}$. If two triangles share an edge, their nodes $\bar{x_1}$ and $\bar{x_2}$ are connected by a grey dashed edge. The graph formed by the nodes associated to triangles is called *dual*.

at the same critical probability. In the recent years, we have witnessed many complex network models in which nodes are added at each time step. The graph-theoretic and probabilistic techniques illustrated in our paper can thus be used to rigorously prove the behaviour of models.

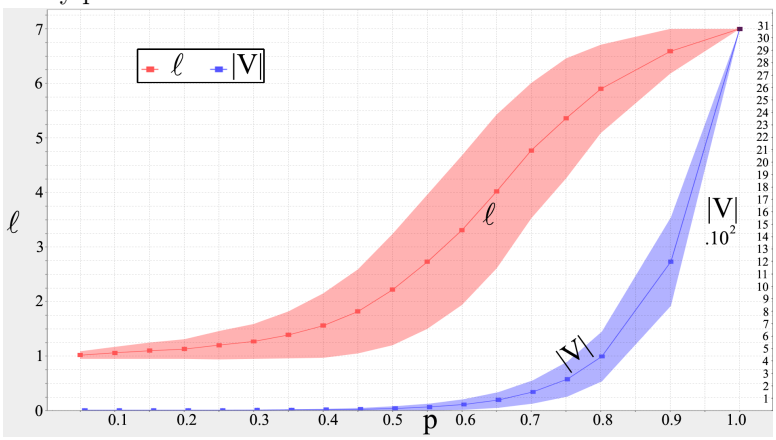

Fig. 7. The average path length has a slow growth compared to the number of nodes, as in $BZRG_p$. We show the values averaged accross simulations and the standard deviation. *Color online.*

References

1. S. Schnettler, Social Networks **31**, 165 (July 2009), ISSN 03788733
2. P. Giabbanelli and J. Peters, Technique et Science Informatiques (to appear)(2010)
3. D. J. Watts, *Small worlds: the dynamics of networks between order and randomness* (Princeton University Press, Princeton, NJ, 1999)
4. J. Davidsen, H. Ebel, and S. Bornholdt, Phys. Rev. Lett. **88** (2002)
5. Z. Zhang, L. Chen, S. Zhou, L. Fang, J. Guan, and T. Zou, Phys. Rev. E **77** (2008)
6. Z. Zhang, L. Rong, and C. Guo, Physica A **363**, 567 (2006)
7. P. Giabbanelli, D. Mazauric, and S. Perennes, in *Proc. of the 12th AlgoTel*, 2010
8. Z. Zhang, L. Rong, and F. Comellas, J. of Physics A **39**, 3253 (2006)
9. R. Cohen and S. Havlin, Phys. Rev. Lett. **90** (2003)