Connectivity of a random directed graph model

Anne Benoit, Kamer Kaya, Bora Uçar

Connectivity of a random directed graph model

Anne Benoit, Kamer Kaya, Bora Uçar

Connectivity of a random directed graph model

Anne Benoit, Kamer Kaya, Bora Uçar

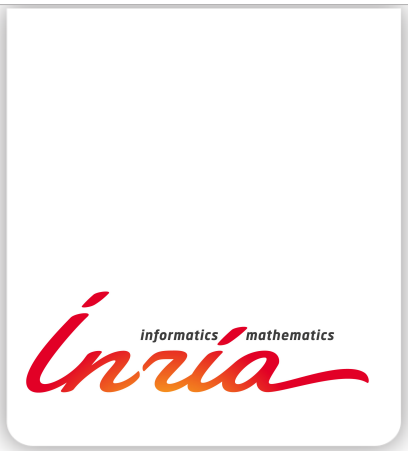
RESEARCH REPORT

N° 9540

January 2024

Project-Teams ROMA

ISSN 0249-6399 ISRN INRIA/RR--9540--FR+ENG



Connectivity of a random directed graph model

Anne Benoit*, Kamer Kaya†, Bora Uçar‡

Project-Teams ROMA

Research Report n° 9540 — version 2 — initial version January 2024
—revised version February 2024 — 12 pages

Abstract: We study the strong connectivity of directed graphs belonging to a random model with three parameters n, d, p. The parameter n defines the number of vertices. Each d-tuple of vertices is picked independently with probability p and if picked, $d-1$ directed edges from the vertex in the first position in the picked tuple to all others are added to the edge set. For $d = 2$, the model thus reduces down to the well-known Erdös–Renyi random directed graphs. The higher order case $d > 2$ arises in the spectral analysis of sparse tensors. We first investigate the threshold phenomenon for the strong connectivity of the directed random graphs from this model. Then, we conduct a series of experiments aimed at gaining a deeper understanding of these directed random graphs.

Key-words: Random graphs, strong connectivity, sparse tensor

* Univ Lyon, CNRS, ENS de Lyon, Inria, Université Claude Bernard Lyon 1, LIP UMR 5668, F-69007 LYON, France
† Sabancı University, Turkey
‡ Univ Lyon, CNRS, ENS de Lyon, Inria, Université Claude Bernard Lyon 1, LIP UMR 5668, F-69007 LYON, France

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Connectivité d'un modèle de graphe dirigé aléatoire

Résumé : Nous étudions la connectivité des graphes dirigés appartenant à un modèle aléatoire avec trois paramètres n, d, p. Le paramètre n définit le nombre de sommets. Chaque chaîne d de sommets, avec répétitions, est choisie indépendamment avec une probabilité p et, si elle est choisie, $d-1$ arêtes dirigées du sommet en première position dans la chaîne vers tous les autres sont ajoutées à l'ensemble des arêtes. Pour $d=2$, le modèle se réduit donc aux graphes aléatoires dirigés bien connus d'Erdös–Renyi. Le cas d'ordre supérieur $d>2$ se présente dans l'analyse des tenseurs creux. Nous étudions d'abord le phénomène de seuil pour la connectivité des graphes aléatoires dirigés de ce modèle. Ensuite, nous menons une série d'expériences visant à mieux comprendre ces graphes aléatoires dirigés.

Mots-clés : graphes aléatoires, connectivité, tenseurs creux

Contents

1 Introduction

One of the fundamental concepts in nonnegative matrix theory is the irreducibility of a matrix [2, Ch.2]. Once this property is verified for a matrix, one obtains a solid understanding of the spectral properties of the matrix by the use of the well-known Perron–Frobenius theorem and its extensions [2]. This analysis underlines many important applications, including PageRank [12]. An $n \times n$ matrix A is called irreducible if its directed graph $G_A = (V, E)$, where $|V| = n$ and for $v_i, v_j \in V$ the directed edge $v_i \to v_j \in E$ exists if and only if $a_{ij} \neq 0$, is strongly connected. Recall that a directed graph is strongly connected, if any vertex reaches to all other vertices by following a set of directed edges.

Extensions of the Perron–Frobenius theorem to tensors have been developed [5, 7, 17][15, Ch.3] and found various applications [9, 15]. By a tensor, we mean a multi-dimensional array, sometimes called a hypermatrix [13]. One of the extensions of the Perron–Frobenius theorem to tensors [5, 17] defines the irreducibility of a tensor T based on the strong connectivity of a directed graph associated with T. Let us now introduce the directed graph that is used in this context. Let T be a d-dimensional sparse tensor with n indices in each dimension, so that each nonzero is of the form $T(i_1, i_2, \ldots, i_d)$, with $i_1 \in \{1, 2, \ldots, n\}$. The directed graph $G_T = (V, E)$ associated with T has $|V| = n$ vertices, and each nonzero $T(i_1, i_2, \ldots, i_d)$ adds a total of $d - 1$ edges $v_{i_1} \to v_{i_j}$ for $j = 2, 3, \ldots, d$ to the directed graph G_T. There may be self-loops and parallel directed edges in G_T, which are of no importance in the context of connectivity. A tensor is called irreducible if and only if G_T is strongly connected. Note that when the definitions are applied to the case $d = 2$, they coincide with the relevant definitions in the matrix case.

In many case, matrices and tensors from a random family are used to evaluate algorithms [3, 4]. One common random matrix model is based on the Erdös–Renyi [8] or Gilbert [10] random graph model. This random graph model $G(n, p)$ has two parameters n and p. A graph from this family has n vertices, and includes each possible edge (in other words, all potential pair of vertices) with probability p. Then, the adjacency matrix of the graph is taken as a pattern of a random symmetric matrix for evaluating algorithms. The process can also be described in terms of matrices: each entry in the strictly upper triangular part of an $n \times n$ matrix is set to one with probability p, and the entries in the lower triangular part are set to be equivalent to the corresponding symmetric entries. In a similar fashion, random directed graph model $\overrightarrow{G}(n, p)$ can be used to create random unsymmetric matrices. In matrix terms, each of the n^2 entries of an $n \times n$ matrix is set to one with probability p. A similar model has been used for creating random tensors, with an additional parameter d specifying the number of dimensions [6][15, Ch. 4]. In this random tensor model, one builds a d-dimensional, 0-1 tensor T with n indices in each dimension, where $T(i_1, i_2, \ldots, i_d) = 1$ with probability p and zero otherwise.

The Erdös–Renyi graph models $G(n, p)$ and $\overrightarrow{G}(n, p)$ and hence the resulting matrices are well understood. In particular, the connectivity of undirected graphs from $G(n, p)$ and the strong connectivity of the directed graphs from $\overrightarrow{G}(n, p)$ are well established. Our aim in this paper is to understand the connectivity of the directed graphs associated with the random tensors created with parameters n, d, p and see when such directed graphs are strongly connected or when the random tensors are irreducible. We use $D(n, d, p)$ to refer to this family of directed graphs. To create a directed graph graph $G = (V, E)$ from $D(n, d, p)$, one builds a d-dimensional, 0-1 tensor T with n indices in each dimension, where $T(i_1, i_2, \ldots, i_d) = 1$ with probability p and zero otherwise. Then, $V = \{1, 2, \ldots, n\}$, and for each nonzero $T(i_1, i_2, \ldots, i_d)$ of the tensor T, E contains the $d-1$ directed edges from the vertex corresponding to i_1 to all vertices corresponding to $i_2, \ldots, i_d$. Here, we see that there are dependencies between edges in a graph from $D(n, d, p)$. The existence of a directed edge of the form $v_i \to v_j$ implies $d-2$ more directed edges emanating

from v_i. This dependency distinguishes $D(n,d,p)$ from $\overrightarrow{G}(n,p)$, and prevents us applying the results known for $\overrightarrow{G}(n,p)$ to our case.

Formally, this paper studies the connectivity properties of directed graphs belonging to a random model with three parameters $n \in \mathbb{Z}^+$, $d \in \mathbb{Z}^+$, and $p \in \mathbb{R}$ in the range $[0,1]$. Let $\mathbf{i} \in \{1,2,\ldots,n\}^d$, so that $\mathbf{i} = \langle i_1, i_2, \ldots, i_d \rangle$ is a d-tuple with $i_j \in \{1,2,\ldots,n\}$ for $j = 1,2,\ldots,d$. In a directed graph $G = (V,E)$ belonging to the family $D(n,d,p)$,

- there are n vertices, $|V| = n$;
- each d-tuple $\mathbf{i} = \langle i_1, i_2, \ldots, i_d \rangle$ for $i_j \in \{1,2,\ldots,n\}$ and $j = 1,2,\ldots,d$ is picked independently with probability p to generate edges. When a d-tuple $\mathbf{i} = \langle i_1, i_2, \ldots, i_d \rangle$ is picked, $d-1$ directed edges $i_1 \to i_j$ for $j = 2,3,\ldots,d$ are added to the edge set E.

The d-tuples are sampled independently, and the edges of G are created afterwards deterministically using the sampled edges. The duplicate edges and self-loops in G are of no importance for this study and can be discarded.

Our aim is to show a phase transition in the strong connectivity of $D(n,d,p)$ for a given n and d. That is, we want to specify a threshold $\hat{p}$ such that for all $p \gg \hat{p}$, a directed graph from the family $D(n,d,p)$ is strongly connected with a high probability; and a directed graph from the family $D(n,d,p')$ for $p' \ll \hat{p}$ is not strongly connected with a high probability.

Claim 1. *Let $\hat{p} = \frac{\log n + c}{n^{d-1}}$ for a real constant c. Then, for all $p \gg \hat{p}$, a directed graph from the family $D(n,d,p)$ is strongly connected with a high probability. In particular,*

$$\lim_{n\to\infty} \Pr(D(n,d,p) \text{ is strongly connected}) \geq e^{-2e^{-c}} .$$

For the *undirected* Erdös-Renyi (ER) graph model, the equivalent question is the threshold for an undirected ER graph model to be connected, which is $\frac{\log n + c}{n}$ for a $c \in \mathbb{R}$. Note that in this case, there are $n \log n + cn$ edges in expectation. The threshold for the directed graph model $D(n,2,p)$ is shown by Graham and Pike [11] to be $\frac{\log n + c}{n}$ for an arbitrary constant c; more precisely, Graham and Pike show that

$$\lim_{n\to\infty} \Pr(D(n,2,p) \text{ is strongly connected}) = e^{-2e^{-c}} .$$

As seen here, the threshold values for undirected ER graphs to be connected or directed ER graphs to be strongly connected are equivalent. The claimed results are equivalent to these results in that there are $n \log n + cn$ d-tuples to be considered for adding edges, and the limiting values for the directed graphs to be strongly connected are lower bounded by the same function.

2 Analysis

Let $\mathcal{N} = \{1,2,\ldots,n\}$. For a $D(n,d,p)$ not to be strongly connected, there should be a set of vertices $\mathcal{S} \subset \mathcal{N}$ from which no edge is directed to $\mathcal{N} \setminus \mathcal{S}$. We will bound the probability of this happening. A sample 3D tensor is shown in Figure 1 to aid discussion. In terms of tensors, there should be a subset of indices $\mathcal{S}$ in the first dimension such that all entries of the tensor with coordinates $\langle i, i_2, \ldots, i_d \rangle$ for $i \in \mathcal{S}$, $i_j \in \mathcal{N} \setminus \mathcal{S}$ and $j = 2,\ldots,d$ should be zero. In the figure, on the dimension I, there should be only zeros in the places shown with 0; that is $T(s,j,k) = 0$ for $s \in \mathcal{S}$ and $j,k \in \mathcal{N} \setminus \mathcal{S}$. The probability that this is happening for the example in Figure 1 is $(1-p)^{|\mathcal{S}|n^2 - |\mathcal{S}|^3}$. In order to compute this probability, we count the number of tensor entries of the form $T(s,j,k)$ for $s \in \mathcal{S}$ and $j,k \in \mathcal{N} \setminus \mathcal{S}$, and multiply the probabilities of each such

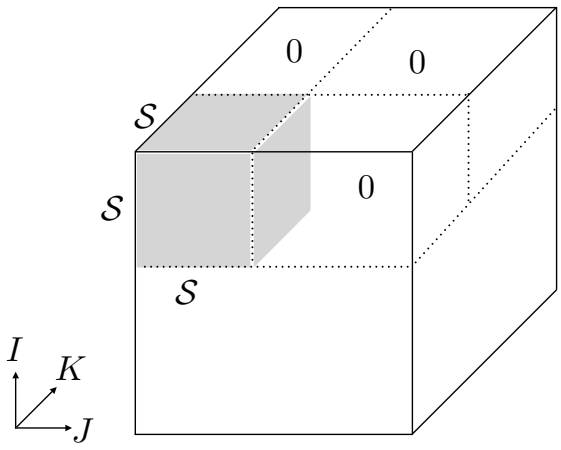


Figure 1: A sample 3D tensor T whose directed graph is not strongly connected.

entry to be zero. The number of entries in places shown with zero is obtained by subtracting the number of entries in the gray subtensor, which is $|\mathcal{S}|^3$, from the total number of entries of the form $T(s, j, k)$ for $s \in \mathcal{S}$ and $j, k \in \mathcal{N}$, which is $|\mathcal{S}|n^2$. In general, the probability that a given set $\mathcal{S}$ of vertices in $D(n, d, p)$ does not have edges outside of $\mathcal{S}$ can be computed as

$$(1-p)^{|\mathcal{S}|n^{d-1}-|\mathcal{S}|^d} . \tag{1}$$

In order for $D(n, d, p)$ not to be strongly connected, there must be at least one subset $\mathcal{S}$ of t vertices for $t = 1, \ldots, n-1$, which yields the form shown in Figure 1. The number of t-element subsets of $\mathcal{N}$ is $\binom{n}{t}$. By the union bound and (1), the probability that at least one of these events happens, in which case $D(n, d, p)$ is not strongly connected, is no larger than

$$\sum_{t=1}^{n-1} \binom{n}{t} (1-p)^{tn^{d-1}-t^d} .$$

Therefore,

$$\Pr(D(n,d,p) \text{ is strongly connected}) \geq 1 - \sum_{t=1}^{n-1} \binom{n}{t} (1-p)^{tn^{d-1}-t^d} . \tag{2}$$

Assuming that we did not lose too much in applying the union bound, the bound in (2) will be tight. We thus want to show that the right hand side of (2) is bounded from below by $e^{-2e^{-c}}$ to finalize the claimed result. We offer experimental results in the next section to demonstrate evidence that

$$\lim_{n\to\infty} \Pr(D(n,d,p) \text{ is strongly connected}) \geq e^{-2e^{-c}} .$$

holds for large c.

While one can develop a formula for the probability of an edge to be present in $D(n, d, p)$, there are dependencies among these events, as the presence of a directed edge $i \to j$ implies the presence of $d-2$ additional directed edges emanating from i. Therefore, a proper calculation of the probability that $D(n, d, p)$ is strongly connected cannot assume the independence of edges in the directed graph G; though the d-tuples are picked independently.

3 Experiments

We have implemented algorithms to create sparse tensors by first computing an expected number of nonzeros, $z = p \times n^d$, and then creating that many nonzeros using the standard `rand()` function from the standard `C` library. Assuming a fully random number generator, this procedure creates a sparse random tensor with z nonzeros among all $\binom{n^d}{z}$ such tensors, with uniform probability. A Matlab implementation of a similar random sparse tensor generator is available elsewhere (sptenrand.m from Tensor Toolbox [1]). In the language of the standard random graph models, this model is called binomial model, while the original one is called uniform graph model. The equivalence between the binomial and uniform graph and directed graph models in terms of thresholds are described by Łuczak [14] and Graham and Pike [11]. While we did not try to establish the equivalence between the binomial and uniform models analytically, we provide some experiments to argue thatthis being the case. Once the directed graph models are created, we check them for being strongly connected using Tarjan's algorithm [16]. Our codes are available at `https://gitlab.inria.fr/bora-ucar/ndp-graph-model`.

While the claimed results are asymptotic, we are interested in practical settings. We present experiments with $n \in \{100, 1000\}$, with $-2 \leq c \leq 4$, and with $d = 2, 4, 8, 16$. For each given triplet of the form (n, d, c), we create 100000 random directed graphs from the $D(n, d, p)$ family, where $p = \frac{\log n + c}{n^{d-1}}$. Counting the number of strongly connected graphs and dividing that number by 100000 gives the empirical probability. We then compare the empirical probabilities with the claimed bound $e^{-2e^{-c}}$. For a small value of $n = 50$, we also explicitly compute the right hand side of Equation (2) and plot its value against the empirical probability and the claimed lower bound for a small range.

3.1 Investigating the threshold

We investigate the empirical probability of $D(n, d, p)$ to be strongly connected with different parameter settings. This will experimentally verify the claimed bound and help us understand the $D(n, d, p)$ model.

Figure 2 plots the empirical probability of $D(n, d, p)$ being strongly connected for $n \in \{100, 1000\}$, $d = 4$, and c in the closed range $[-2, 4]$, defining the probability p. This figure also plots the curve of the lower bound $e^{-2e^{-c}}$. As seen in this figure, the empirical probabilities for $n = 100$ and $n = 1000$ are identical, and they are both larger than the bound $e^{-2e^{-c}}$, converging to it as c increases. The empirical probability at $c = 4$ is 0.98, for both $n = 100$ and $n = 1000$; the lower bound is 0.96. This figure thus suggests that the strong connectivty does not change with n, and the lower bound is tight for large values of c. In this figure, we also observe that for $c \leq -1.0$ and $c \geq 3$, the empirical values agree with the lower bound. These observations also hold for a smaller value of $n = 50$, with $c = [-2, 3]$.

Figure 3 plots the empirical probability of $D(n, d, p)$ being strongly connected for $n = 100$, $d \in \{2, 4, 8, 16\}$, and c in the closed range $[-2, 4]$, defining the probability p. This figure also plots the curve of the lower bound $e^{-2e^{-c}}$. As seen in this figure, the plot for $d = 2$ follows closely the curve $e^{-2e^{-c}}$. The others as a group are nearly identical among themselves and are always above that of $d = 2$, converging to the claimed lower bound $e^{-2e^{-c}}$ for $c = 4$. This figure suggests that the case $d > 2$ is different from the case $d = 2$ for moderate values c of interest such as $c = [1, 3]$, but the cases are similar for values of c outside of this range. These observations also hold for a smaller value of $n = 50$, with $c = [-2, 3]$.

We compare the right hand side of Equation (2), $1 - \sum_{t=1}^{n-1} \binom{n}{t} (1-p)^{tn^{d-1} - t^d}$, with the empirical probability in order to see whether the application of the union bound had introduced large factors. If that is the case, the right hand side of Equation (2) will not likely be as useful

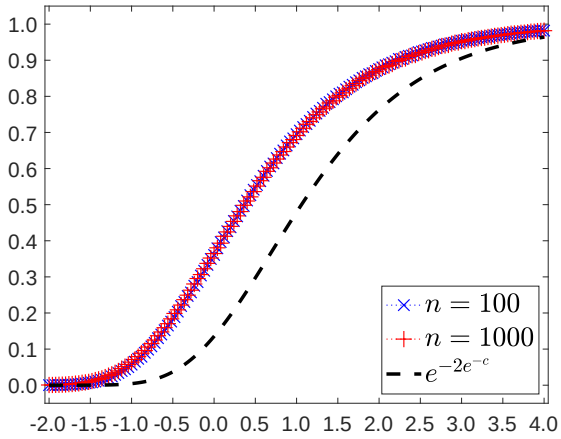


Figure 2: Empirical probability of $D(n, d, p)$ being strongly connected in the y-axis for $n = \{100, 1000\}$, $d = 4$, and $c = [-2, 4]$ in the x-axis. The values of c define the probability p of a nonzero to be chosen as $\frac{\log n + c}{n^{d-1}}$. The curve $e^{-2e^{-c}}$ is also plotted.

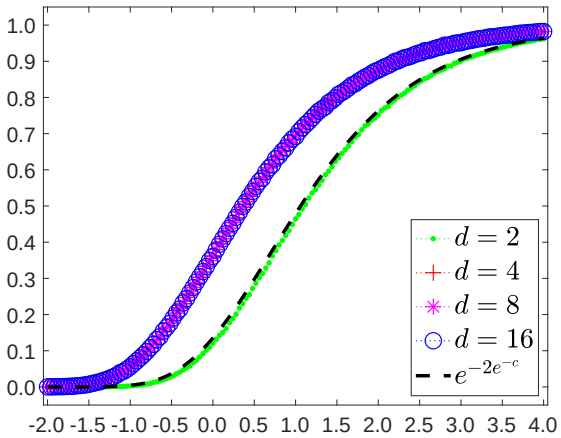


Figure 3: Empirical probability of $D(n, d, p)$ being strongly connected in the y-axis for $n = 100$, $d \in \{2, 4, 8, 16\}$, and $c = [-2, 4]$ in the x-axis. The values of c define the probability p of a nonzero to be chosen as $\frac{\log n + c}{n^{d-1}}$. The curve $e^{-2e^{-c}}$ is also plotted.

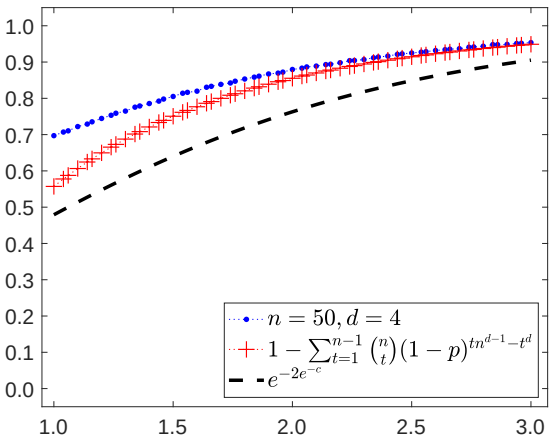


Figure 4: Empirical probability of $D(n, d, p)$ being strongly connected in the y-axis for $n = 50$, $d = 4$, and $c = [1, 3]$ in the x-axis. The values of c define the probability p of a nonzero to be chosen as $\frac{\log n + c}{n^{d-1}}$. The curve $e^{-2e^{-c}}$ and the right hand side of Equation (2) resulting from the union bound are also plotted.

	d4/d2	d8/d4	d16/d8
geomean	2.8183	2.0680	1.7177
slope	-0.0021	-0.0014	-0.0011

Table 1: The number of nonzeros at different $c = [-2, 4]$ with increments of 0.04 for $d = 2, 4, 8, 16$ and $n = 100$ are empirically observed (100000 instances at each value of c), and that of $d = 2k$ is divided to $d = k$ for $k = 2, 4, 8$. The geometric mean of those values are presented, and also the slope of the best fitting line.

lower bound as the claimed $e^{-2e^{-c}}$. Figure 4 compares these three quantities for $n = 50, d = 4$ and $c = [1, 3]$. In this figure, the formula resulting from the use of the union bound looks close enough to the empirical probability, especially for values of $c \geq 1.5$. In the shown range of c, that bound always dominates the claimed lower bound and hence the right hand side of Equation (2) is comparable to the claimed lower bound.

With a given pair of n and c and the claimed threshold probability, the expected number of nonzeros in the underlying d-dimensional tensor is always the same, which is $n \log n + cn$ and independent from d. Due to $d - 1$ edges being added for each selected nonzero, we expect the total number of edges in $D(n, d, p)$ to increase by the increasing d, for a given pair of n, c. This is shown in Figure 5 for $n = 100$. The curves flatten, and the difference between them reduces. We now substantiate this. For each different value of c that we used, we divide the average empirical number of nonzeros with $d = 2k$ to $d = k$, for $k = 2, 4$, and 8 to obtain the ratio of the number of edges. We then fit a line to those ratios, one for each k. Table 1 shows the geometric mean of those ratios and the slopes of each best fitting line to confirm the flattening of the curves and the difference between them.

3.2 Binomial vs uniform models

We present some results to compare binomial and uniform models. Strictly adhering to the uniform model requires investigating all n^d tuples for being a nonzero of the underlying tensor.

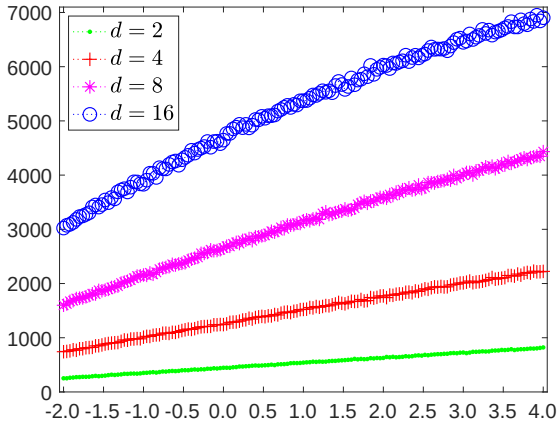


Figure 5: The number of edges in $D(n, d, p)$ in the y-axis for $n = 100$, $d \in \{2, 4, 8, 16\}$, and $c = [-2, 4]$ in the x-axis. The values of c define the probability p of a nonzero to be chosen as $\frac{\log n + c}{n^{d-1}}$.

As this will be very large unless n and d are small, we conduct only a limited set of experiments. We take $n = 50$, $d \in \{3, 4\}$, $c = [0, 3]$ and 10000 random different instances for the experiments. Figure 6 plots the empirical probability of $D(n, d, p)$ being strongly connected with binomial and uniform models. This figure also plots the empirical probability for $n = 100$ and $d \in \{3, 4\}$ with the binomial model (which corresponds to a sub-range of the plots in Figures 2 and Figure 3). This figures suggests that the binomial model results in the same directed graphs as the uniform model.

4 Conclusion

We investigated the strong connectivity of a family of random directed graph models having three parameters $D(n, d, p)$. These parameters are used to define a d-dimensional random tensor with n indices in each dimension. Each entry of the tensor is nonzero with probability p. Once the nonzeros are sampled, the associated directed graph is built on n vertices, with a tensor nonzero at position $\langle i_1, i_2, \ldots, i_d \rangle$ defining $d - 1$ edges from, i_1 to $i_2, i_3, \ldots, i_d$. We wanted to see when one can confidently say that a sampled graph from the family $D(n, d, p)$ is strongly connected. We observed that for $d > 2$ and $p = \frac{\log n + c}{n^{d-1}}$, there is a threshold around $c = -1$; after $c > -1$, the probability of having strongly connected directed graphs increases significantly. This is in accordance with the existing results on the Erdös–Renyi direct graph models. While we did not observe any difference for the cases $d = 4, 8, 16$, there is a noticeable difference between these as a group and $d = 2$.

One interesting observation for algorithm practitioners is that $z = n \log n + cn$ nonzeros are enough to obtain strongly connected directed graphs and investigate the connectivity properties, no matter how high dimensional the underlying tensors are. The binomial model allows generating zd coordinates to have those graphs without exponentially more work or exponentially smaller numbers (as probabilities). The binomial model thus allows creating d-dimensional tensors in time linearly proportional to d to define the graphs, not exponentially as in n^d. This can be useful in evaluating practical algorithms on high dimensional sparse tensors.

Note that, on one hand, a closed formula for the lower bound in Equation (2) would be useful. Indeed, the experiments indicate that the cases $d > 2$ and $d = 2$ are different, and the

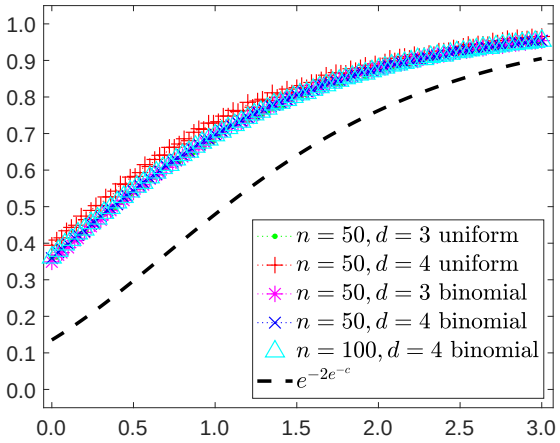
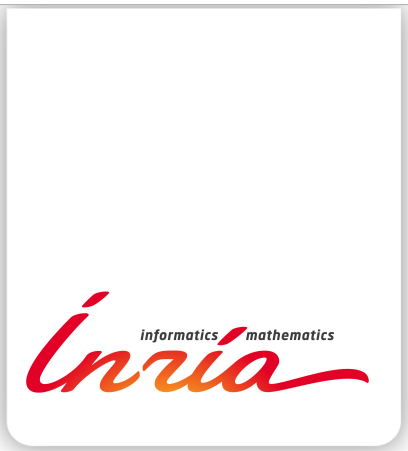


Figure 6: Comparing the empirical probability of a $D(n, d, p)$ being strongly connected for different n and d with binomial and uniform models, in the range of $c = [0, 3]$ in the x-axis. The curve $e^{-2e^{-c}}$ is also plotted in this range.

threshold for the $d > 2$ case is smaller (see Figure 3). On the other hand, the gap diminishes with increasing values of c, for example when $c \geq 3$. It is thus worthwhile to seek for a tighter lower bound for small c, e.g., for $c = [1, 3]$. This is also interesting from a practical perspective as $c = 1$ adds n more nonzeros in expectation, with an average of 1 per fixed index in one of the dimensions. On another note, the experiments suggest that the binomial and uniform models are equivalent (see Figure 6). A formal argument for this is welcome.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399