Increasing paths in the temporal stochastic block model
Discrete Mathematics Days 2024

Sofiya Burova*[1,2], Gábor Lugosi†[2,3,5], and Guillem Perarnau‡[1,4]

[1]IMTECH and Departament de Matematiques, Universitat Politecnica de Catalunya, Spain
[2]ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain
[3]Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain
[4]Centre de Recerca Matematica, Barcelona, Spain
[5]Barcelona School of Economics, Barcelona, Spain

Abstract

We study random temporal graphs, in the context of the Stochastic Block Model. Temporal graphs naturally model time-dependent propagation processes, for instance social interactions or infection processes. In these graphs, every edge has a unique timestamp, chosen uniformly at random, and the notion of connectivity is limited to sequences of edges with timestamps that increase over time. Our goal is to understand the asymptotic behavior of the temporal diameter of these graphs, especially in the subcritical regime where the average number of connections per node is of order $c \log n$, with $c < 1$. We analyze the first-order asymptotics for various aspects, including the length of the longest increasing paths starting at a typical vertex, as well as the size of the set of vertices reachable from a typical vertex via increasing paths.

1 Introduction

The Stochastic Block Model (SBM) is a foundational random graph model used to understand and analyze the structural properties of networks. It models networks by dividing vertices into groups (blocks or communities) and specifying different connection probabilities between these groups. They are versatile tools used for clustering, community detection, anomaly detection, and link prediction. A Temporal Stochastic Block Model extends the traditional SBM by incorporating time into the edge formation process. Each edge in the network not only connects two nodes but also has an associated timestamp, indicating when the interaction occurred. The model can be extended to study dynamic processes over time, such as infection or diffusion processes. TSBMs can model the spread of diseases, where nodes represent individuals and edges represent interactions that could lead to transmission. Temporal information allows for the analysis of how infections spread over time. This perspective emphasizes the importance of *increasing paths* as a key object of interest.

In this work, we consider the stochastic block model of parameters $n \in \mathbb{N}$, $a, p_1, p_2, q \in [0, 1]$, denoted by $\mathbb{SBM}(n, a, p_1, p_2, q)$, comprised of two independent Erdős-Rényi *communities*, $G_1 \sim \mathbb{G}(\lfloor an \rfloor, p_1)$ and

*Email: sofiburova@gmail.com. Research of S.B. supported by the Spanish Agencia Estatal de Investigación under project PID2022-138268NB-100

†Email: gabor.lugosi@gmail.com. Research of G.L. supported by Ayudas Fundación BBVA a Proyectos de Investigación Científica 2021 and the Spanish Ministry of Economy and Competitiveness, Grant PID2022-138268NB-I00, financed by MCIN/AEI/10.13039/501100011033, FSE+MTM2015- 67304-P, and FEDER, EU

‡Email: guillem.perarnau@upc.edu. Research of G. P. supported by the Spanish Agencia Estatal de Investigación under projects PID2020-113082GB-I00 and the Severo Ochoa and Maria de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M)

$G_2 \sim \mathbb{G}(\lceil(1-a)n\rceil, p_2)$, where for all $u \in V(G_1), v \in V(G_2)$, the edge uv appears with probability q independently from all other edges. Additionally, to each edge e we assign a label $L(e)$, a uniform random variable on $[0,1]$ independent from all the others. Note that this is equivalent to taking a uniform permutation of $\{1,\ldots,m\}$ where m is the total number of edges since we are only interested in the order of the edges with respect to one another. Let $G = (V,E,L) \sim \mathbb{SBM}(n,a,p_1,p_2,q)$ be a temporal stochastic block model where V is the vertex set, E is the set of edges and $(L(e))_{e\in E}$ is a family of iid[1] uniform random variables on $[0,1]$. We say a path $(w_1,\ldots,w_k)$ is *increasing* if the sequence of labels $(L(v_iv_{i+1}))_{i\in[k-1]}$ is increasing and determine its length to be the number of edges in it. For $u, v \in V(G)$, we say u is reachable from v if there exists an increasing path from v to u. Denote by $B_\ell(v)$ where $\ell \in \mathbb{N}$, the set of reachable vertices from v via increasing paths of length ℓ. Let $B_n(v) = \cup_\ell B_\ell(v)$ be the reachable set of v in G.

The topic of random simple temporal graphs has been widely discussed in previous works ([2, 3, 4, 5]). It is known that the temporal Erdős-Rényi random graph $\mathbb{G}(n,p)$ undergoes a phase transition around $p = c\log n/n$. In particular, Casteigts et al. showed in [5] that the thresholds for the properties that a typical pair of vertices is connected, a typical vertex can reach all other vertices, and any pair of vertices is connected are, respectively, $\log n/n, 2\log n/n$ and $3\log n/n$. Broutin, Kamčev and Lugosi, [4], recently extended these results providing tight bounds for the longest and shortest increasing paths in each regime, thus completing the study of the temporal diameter of $\mathbb{G}(n,p)$.

In this work, we generalize their approach to the stochastic block model as we suspect it exhibits similar behaviour. Set $p_1 = \lambda_1 \log n/n$, $p_2 = \lambda_2 \log n/n$ and $q = c\log n/n$. We focus on this regime as it is where the phase transition occurs in each community. We naturally assume that edges are more common within communities than between communities. In our setting, it suffices to assume that $c \le \sqrt{\lambda_1\lambda_2}$. The parameter of interest is

$$\theta = \theta(a,\lambda_1,\lambda_2,c) \coloneqq \frac{1}{2}(a\lambda_1 + (1-a)\lambda_2) + \frac{1}{2}\sqrt{(a\lambda_1 - (1-a)\lambda_2)^2 + 4a(1-a)c^2}. \tag{1}$$

Notice that, the case $a = 1-a = 1/2$ and $\lambda_1 = \lambda_2 = c$ yields the $\mathbb{G}(n, c\log n/n)$ setting from [4]. We say that a sequence of events $(E_n)_{n\in\mathbb{N}}$ occurs *with high probability* (whp for brevity) if $\lim_{n\to+\infty}\mathbb{P}_n[E_n] = 1$. We are now ready to state our main results.

Theorem 1 (Reachability from a vertex). *Let $w \in V(G)$. If $0 < c \le \sqrt{\lambda_1\lambda_2}$ and $\theta < 1$, then whp*

$$\frac{\log|B_n(w)|}{\log n} \to \theta.$$

We present a constructive algorithm for building the reachable set from a typical vertex via increasing paths. In particular, we show that one can embed into the temporal $\mathbb{SBM}$ a *uniform random recursive tree* (URRT) of size roughly n^θ, which is a well-understood random object, obtained by repeatedly attaching a leaf to a random vertex of the existing tree, thus providing deeper understanding of the structure of the graph. The main challenge is that the labels of the edges of the tree in the temporal stochastic block model depend on the number of vertices it contains from each community. We couple the evolution of these quantities with a Pólya urn process to keep track and use a result by Janson, [8], to understand their asymptotic behaviour. Furthermore, since the height of a random recursive tree of size m is known to be asymptotically $e\log m$ ([6, 9]), we can derive from this result a lower bound on the size of the longest path with a fixed starting point.

Theorem 2 (Longest increasing paths). *If $0 < c \le \sqrt{\lambda_1\lambda_2}$ and $\theta < 1$, then whp*

(i) there are no increasing paths between two fixed vertices;

[1] independent and identically distributed

(ii) the size of the longest increasing path starting at a fixed vertex $w \in V(G)$, denoted by $\gamma_{\max}(w)$,

$$\frac{|\gamma_{\max}(w)|}{\log n} \to e\theta.$$

The upper bounds in both Theorem 1 and Theorem 2 are obtained by combinatoric computations and standard probabilistic methods, involving some cumbersome calculations, see Section 2. In Section 3, we prove the lower bounds through well-placed couplings, the most important of which is described in our key Lemma 10. To fit our setting, we extend a concentration inequality by Janson ([7]) on sums of independent exponential random variables of deterministic parameters to randomized ones, under some additional assumptions.

Remark 3. *Observe that $\theta = \theta(c)$ is increasing as a function of c, hence*

$$\theta(0) = \max\{a\lambda_1, (1-a)\lambda_2\} \leq \theta \leq a\lambda_1 + (1-a)\lambda_2 = \theta(\sqrt{\lambda_1\lambda_2})$$

since $0 \leq c \leq \sqrt{\lambda_1\lambda_2}$. It follows from the first inequality that the regime $\theta < 1$ renders both G_1 and G_2 subcritical.

Remark 4. *The threshold for the emergence of a giant component of linear size in the classic $\mathbb{SBM}$ is in the regime where $p_1 = \lambda_1/n, p_2 = \lambda_2/n, q = c/n$ and is known to be $\theta = 1$ ([1]). We expect the critical regime in the temporal setting to be delayed by a factor of $\log n$ and preserve the threshold $\theta = 1$ as is observed for $\mathbb{G}(n,p)$.*

2 Upper bounds: First moment method

2.1 Combinatorics of paths

Fix $u, v \in V(G)$ distinct. Denote by $\Lambda^k = \Lambda^k(u,v)$ the set of all self-avoiding paths on G of length $k \in \mathbb{N}$ from u to v. We partition

$$\Lambda_k = \bigcup_{\ell \in [k], h \in [k-\ell]} \Lambda^k_{\ell,h}$$

where ℓ is the number of inter-communal edges and h is the number of inner-communal edges of the community of origin used to construct the paths. By convention, for $i \in \{1,2\}$, we write $\bar{i} = 3 - i$.

Proposition 5. *Suppose $u \in V(G_i)$ and $v \in V(G_j)$, $i, j \in \{1,2\}$. Let $\ell \in [k], h \in [k-\ell]$ and set $s_1 = h + \lfloor \ell/2 \rfloor - \mathbb{1}_{j=i}$ and $s_2 = k - h - \lfloor \ell/2 \rfloor - \mathbb{1}_{j=\bar{i}}$. Then, the size of the set $\Lambda^k_{\ell,h}$ is given by*

$$|\Lambda^k_{\ell,h}| = |\Lambda^k_{\ell,h}(i,j)| = \begin{cases} \frac{n_i!}{(n_i-k-1)!} & \text{if } \ell = 0,\ h = k,\ i = j; \\ \frac{n_i!}{(n_i-s_1)!}\frac{n_{\bar{i}}!}{(n_{\bar{i}}-s_2)!}\binom{h+\lceil\frac{\ell+1}{2}\rceil-1}{h}\binom{k-h-\lceil\frac{\ell+1}{2}\rceil}{k-\ell-h} & \text{if } \ell \in 2\mathbb{N}^* - |i-j|; \\ 0 & \text{otherwise.} \end{cases}$$

2.2 First moment method

Let $u \in V(G_i), v \in V(G_j)$ and ℓ, h be such that $\Lambda^k_{\ell,h} \neq \emptyset$. Then, for all $\gamma \in \Lambda^k_{\ell,h}$, define the event $A(\gamma) = \{\gamma$ is increasing and all its edges are present in $G\}$. It is straightforward that

$$\mathbb{P}[A(\gamma)] = \frac{p_i^h q^\ell p_{\bar{i}}^{k-\ell-h}}{k!} = \frac{(\log n)^k}{n^k k!}\lambda_i^h c^\ell \lambda_{\bar{i}}^{k-\ell-h}$$

where the $1/k!$ factor ensures that the path is increasing. Denote by X_k the number of increasing paths of length k in G. Denote by $Y_k(i)$ the number of increasing paths of length k that start at a fixed vertex in G_i and let $Y_k = Y_k(1) + Y_k(2)$. Similarly, denote by $Z_k(i)$ the number of increasing paths of length k whose two endpoints are fixed vertices from the same community G_i and by Z^*_k the number of increasing paths of length k whose endpoints are fixed vertices from different communities. Let $Z_k = Z_k(1) + Z_k(2) + Z^*_k$.

Lemma 6. *Let $k \in \mathbb{N}$. There exists a finite natural number $d \in \mathbb{N}$, not depending on k, such that*

$$\mathbb{E}[X_k] \leq O\left(\frac{nk^d}{k!}(\theta \log n)^k\right), \quad \mathbb{E}[Y_k] = O\left(\frac{k^d}{k!}(\theta \log n)^k\right), \quad \mathbb{E}[Z_k] = O\left(\frac{k^d}{nk!}(\theta \log n)^k\right).$$

Proof of the upper bound in Theorem 1 and Theorem 2. First, we prove (i). Fix two vertices $w, z \in V(G)$. Then, the probability that there exists an incresing path from w to z in G is given by

$$\mathbb{P}\left[\bigcup_{k\in\mathbb{N}} \bigcup_{\gamma\in\Lambda^k(w,z)} A(\gamma)\right] \leq \sum_{k\geq 1} \mathbb{P}[Z_k \geq 1] \leq C(\theta \log n)^d n^{\theta-1} = o(1)$$

since $\theta < 1$. For (ii), we proceed in a similar manner. Notice that, the existence of an increasing path starting at w of length k implies there are increasing paths starting at w of length ℓ for any $1 \leq \ell \leq k$. Thus, Markov's inequality implies

$$\mathbb{P}[|\gamma_{\max}(w)| \geq k] \leq \mathbb{P}[Y_k \geq 1] \leq \mathbb{E}[Y_k] \leq Ck^d\left(\frac{e\theta \log n}{k}\right)^k$$

where the last inequality follows by Lemma 6 and Stirling's approximation. Setting k to be $(1+\epsilon)e\theta \log n$ yields the desired result. For the reachable set of w, first notice that $|B_n(w)| \leq \sum_{k\geq 1} Y_k$. Hence, by Markov's inequality and Lemma 6,

$$\mathbb{P}[|B_n(w)| \geq n^{\alpha(1+\epsilon)}] \leq \frac{\sum_{k\geq 1}\mathbb{E}[Y_k]}{n^{\alpha(1+\epsilon)}} \leq C(\theta \log n)^d n^{-\alpha\epsilon} = o(1)$$

which concludes the proof. □

3 Lower bounds: Embedding a random recursive tree

We will couple $G = (V, E, L)$ with $G' = (V, E', L')$ defined as follows. To each pair of distinct vertices u, v, associate an exponential random variable W_{uv} independent of the others, of parameter p_{uv} where

$$p_{uv} := \begin{cases} q & \text{if } u \in G_1, v \in G_2 \text{ or } u \in G_2, v \in G_1, \\ p_i & \text{if } u, v \in G_i \text{ for } i \in \{1, 2\}. \end{cases}$$

An edge $e = uv$ has label $L'(e) = W_{uv}$ and $e \in E'$ if and only if $W_{u,v} \leq 1$. Notice that

$$\mathbb{P}[e \in G'] = \mathbb{P}[W_{uv} \leq 1] = 1 - e^{-p_{uv}} \leq p_{uv} = \mathbb{P}[e \in G]$$

so we can couple G and G' in such a way that $E' \subseteq E$.

Lemma 7. *The total variation distance between L and L' restricted to the edges that appear in G'*

$$d_{TV}(L_{|E'}, L'_{|E'}) \to 0 \qquad \text{as } n \to \infty.$$

Corollary 8. *For n large enough, there is a coupling between G and G' such that for any path γ,*

$$\mathbb{P}_{G'}[A(\gamma)] \leq \mathbb{P}_G[A(\gamma)],$$

i.e. if γ is open and increasing in G', then it is open and increasing in G as well.

Janson proved the following concentration inequality for sums of independent exponential random variables of deterministic parameters (Theorem 5.1, [7]). Here, we extend the result by considering randomized parameters under mild additional assumptions. The last ingredient that we need to prove our key lemma is the following concentration inequality.

Lemma 9. *Let $(\xi_i)_{i\in\mathbb{N}}$ be a sequence of real random variables such that*

- *there exist $c_1, c_2 \in \mathbb{R}_+$ such that $c_1 i \leq \xi_i \leq c_2 i$ deterministically for all $i \in \mathbb{N}$;*
- *ξ_i/i converges almost surely to a constant θ as $i \to \infty$.*

Let $X_i \sim Exp(\xi_i \log n)$ be exponential random variables such that, conditional on $(\xi_i)_{i\in\mathbb{N}}$, $(X_i)_i$ are mutually independent. Then, for all $\epsilon \in (0,1)$, if $r \leq n^{(1-\epsilon)\theta}$, with high probability $\sum_{i=1}^{r} X_i < 1$.

Lemma 10 (Key Lemma)**.** *Fix a vertex $w \in V(G)$. Let $r = \lfloor n^{(1-\epsilon)\theta} \rfloor$ and let T_r be a random recursive tree on r vertices. Then, whp, there is a coupling between G' and the random recursive tree T_r such that there is a tree $\tilde{T}_r$ rooted at vertex w which consists of increasing paths from w and is isomorphic to T_r.*

Proof of the lower bound in Theorem 1 and Theorem 2. Fix $\epsilon > 0$. From Corollary 8 and Lemma 10, it follows that we can embed into G a tree T of size $r = n^{(1-\epsilon)\theta}$ that is isomorphic to a URRT T_r. The lower bound in Theorem 1 follows immediately. It is known that the height of T_r is whp $e \log r$ ([6, 9]). Hence, whp $|\gamma_{\max}| \geq e \log r = (1-\epsilon)e\theta \log n$. □

References

[1] E. Abbe, Community Detection and Stochastic Block Models, *Foundations and Trends® in Communications and Information Theory* **14(1-2)** (2018), 1-162.

[2] O. Angel, A. Ferber, B. Sudakov and V. Tassion, Long monotone trails in random edge-labellings of random graphs, *Combinatorics, Probability and Computing* **29(1)** (2020), 22–30.

[3] R. Becker, A. Casteigts, P. Crescenzi, B. Kodric, M. Renken, M. Raskin and V. Zamaraev, Giant components in random temporal graphs, arXiv preprint, 2022, `arXiv:2205.14888`.

[4] N. Broutin, N. Kamčev and G. Lugosi, Increasing paths in random temporal graphs, arXiv preprint, 2023, `arXiv:2306.11401`.

[5] A. Casteigts, M. Raskin, M. Renken and V. Zamaraev, Sharp thresholds in random simple temporal graphs, *In 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (2022), 319–326.

[6] L. Devroye, Branching processes in the analysis of the heights of trees, *Acta Informatica* **24** (1987), 277–298.

[7] S. Janson, Tail bounds for sums of geometric and exponential variables, *Statist. Probab. Lett.* **135** (2018), 1–6.

[8] S. Janson, Functional limit theorems for multitype branching processes and generalized Pólya urns, *Stochastic Processes and their Applications* **110(2)** (2004), 177–245.

[9] B. Pittel, Note on the height of random recursive trees and m-ary search trees, *Random Structures and Algorithms* **5** (1994), 337–347.