Discrete Mathematics and Theoretical Computer Science DMTCS vol. **17:3**, 2016, 327–348

Connected Tropical Subgraphs in Vertex-Colored Graphs

Jean-Alexandre Anglès d'Auriac[1] Nathann Cohen[1]
Hakim El Maftouhi[1] Ararat Harutyunyan[2*]
Sylvain Legay[1] Yannis Manoussakis[1]

[1] *L.R.I., Université Paris-Sud, France.*
[2] *Mathematical Institute, University of Toulouse III (Paul Sabatier), France.*

received 2nd Feb. 2015, revised 14th July 2016, accepted 26th July 2016.

A subgraph of a vertex-colored graph is said to be tropical whenever it contains each color of the graph. In this work we study the problem of finding a minimal connected tropical subgraph. We first show that this problem is NP-Hard for trees, interval graphs and split graphs, but polynomial when the number of colors is logarithmic in terms of the order of the graph (i.e. FPT). We then provide upper bounds for the order of the minimal connected tropical subgraph under various conditions. We finally study the problem of finding a connected tropical subgraph in a randomly vertex-colored random graph.

Keywords: vertex-colored graph, connected subgraph, tropical subgraph, colorful subgraph, vertex-colored random graph.

1 Introduction

In this work, we deal with tropical substructures in vertex-colored graphs, first introduced in [AMK+]. Vertex-colored graphs are useful in various situations. For instance, the Web graph may be considered as a vertex-colored graph where the color of a vertex represents the content of the corresponding page (red for mathematics, yellow for physics, etc.) [BHKN13]. Applications can also be found in bioinformatics (Multiple Sequence Alignment Pipeline or for multiple protein-protein Interaction networks) [CPM10]. Given a vertex-colored graph, a *tropical subgraph* is defined to be a subgraph where each color of the initial graph appears at least once. Potentially, many graph invariants, such as the domination number and the vertex cover number, can be studied in their tropical version. This notion is close to the *colorful* concept used for paths in vertex-colored graphs [ALN11, Li01, Lin07] (with a colorful path being a

*Research supported by an FQRNT fellowship and a Digiteo postdoctoral scholarship, the latter when the author was at Université Paris-Sud.

tropical subgraph that is a path, and has no repeat color), though works on colored paths usually focus on finding colorings that fulfill specific criteria, one being admitting colorful path, while our work consider the coloring as an inherent property of the graph. It is also related to the concepts of *color patterns* used in bio-informatics [FFHV11, ZSLS11], which share our approach for the coloring of the graph. Here, we study minimum connected tropical subgraphs in vertex-colored graphs, focusing especially on the case where the number of colors used is large. The case where the number of colors is small is even more interesting in view of the aforementioned applications. Some related work can also be found in [BHKN13, BHK$^+$12, PA, ZSLS11], where the authors are looking for the minimum number of edges to delete in a graph such that all remaining connected components are colorful (i.e., do not contain two vertices of the same color). Some ongoing work on dominating tropical sets, tropical paths and tropical homomorphisms can be found in [AMK$^+$, FHH$^+$].

Throughout the paper, we let $G = (V, E)$ denote a simple undirected graph. Given a set of colors $\mathcal{C} = \{1, \ldots, c\}$, G^c denotes a vertex-colored graph whose vertices are each colored (not necessarily properly) by one of the colors in $\mathcal{C}$, and each color of $\mathcal{C}$ colors at least one vertex. For any subgraph H of G^c, we denote by $c(H)$ the set of colors of the vertices of H. A graph G^c is said to be *properly colored* when no adjacent vertices have the same color. The chromatic number of an uncolored graph G, denoted $\chi(G)$, is the smallest number of colors c such that there exists a graph G^c that is properly colored. A connected subgraph H of G^c is said to be *tropical* if $c(H) = \mathcal{C}$. The *connected tropical subgraph number* $\mathfrak{tc}(G^c)$ is the order of a smallest connected tropical subgraph of G^c. A *connected rainbow subgraph* of G^c is a connected subgraph in which each color is present at most once. A *connected colorful subgraph* of G^c is a connected rainbow subgraph which is tropical. The neighborhood $N(u)$ is the set containing all vertices adjacent to vertex u in G^c. The degree $d(u)$ is the number of vertices in $N(u)$. The closed neighborhood $N[u]$ is $N(u) \cup \{u\}$. We let $\delta(G^c)$ denote the minimum degree of G^c. When no confusion arises, we write $\mathfrak{tc}$ and δ instead of $\mathfrak{tc}(G^c)$ and $\delta(G^c)$. A *dominating set* S of a graph $G = (V, E)$ is a subset of V such that every vertex of V is either in S, or adjacent to a vertex in S. We denote by $\gamma(G)$ the minimum size of a dominating set of G. We call *blocks* of a graph its maximal 2-connected subgraphs, and say that a block is a *leaf* block when it contains exactly one cut-vertex. This paper is a study of the following problem:

MINIMUM CONNECTED TROPICAL SUBGRAPH PROBLEM (MCTS)

Input: A connected vertex-colored graph G^c, and an integer k.

Question: Is there a connected tropical subgraph of order k in G^c?

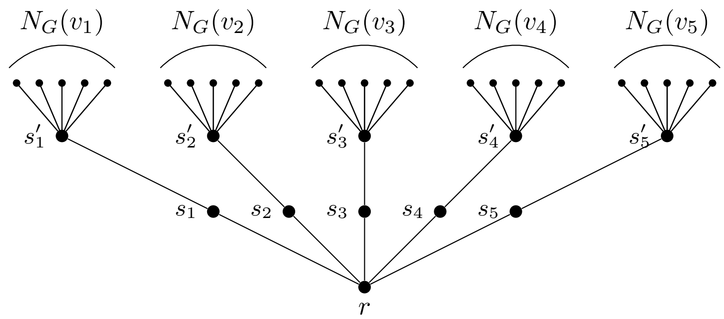
It is split into three parts. In Section 2 we prove that MCTS is NP-complete for general graphs, trees, interval and split graphs. We also give a dynamic programming FPT (Fixed Parameter Tractable) algorithm parametrized in the number of colors for general graphs. In Section 3 we give upper bounds for $\mathfrak{tc}$ related to various parameters (minimum degree, number of edges). In Section 4, we study how $\mathfrak{tc}$ behaves on random graphs.

2 NP-Hardness and FPT algorithms

Theorem 2.1. *MCTS is NP-Complete on trees.*

Proof: MCTS is in NP since testing whether a given set of vertices corresponds to a tropical connected subgraph can be done in polynomial time. The reduction is obtained from DOMINATING SET on general graphs. Consider an instance of DOMINATING SET on a graph G with vertices $v_1, v_2, \dots, v_n$. We define a colored tree T^c with $n+2$ colors $c_1, c_2, \dots, c_{n+2}$ in the following way. Let r be a vertex of color c_{n+2}, and for each $i \in \{1, \dots, n\}$:

- Let s_i be a vertex of color c_{n+1} adjacent to r.
- Let s'_i be a vertex of color c_i adjacent to s_i.
- For each vertex $v_j \in N(v_i)$, there is a vertex of color c_j adjacent to s'_i.



Given a dominating set S of G, along with a function f associating to each vertex $v \in G \setminus S$ an element of S that dominates it, we define a connected tropical subgraph H of T^c, containing the following vertices :

1. The vertex r.
2. The vertices s_i and s'_i of T^c for every $v_i \in S$.
3. For each vertex $v_i \in G \setminus S$, we take the vertex of color c_i which is adjacent to the vertex s'_j such that $v_j = f(v_i)$.

By construction, H is connected and tropical.

Reciprocally, given a minimal connected tropical subgraph H of T^c, we define the following set of vertices and f function: v_i belongs to S if and only if $s'_i \in H$, and for each $v_j \notin S$, $f(v_j) = v_i$ where i is such that s'_i is the neighbor of the only vertex of color c_j in H. Hence, there exists a bijection between a pair (S, f) associated with a dominant S of G and a minimal connected tropical subgraph H of T^c. Moreover, we have the following:

$$|H| = 1 + 2|S| + n - |S| = 1 + |S| + n.$$

Thus, by considering the minimum cardinality of S, we obtain,

$$\mathfrak{tc}(T^c) = 1 + \gamma(G) + n.$$

As a result, a minimum connected tropical subgraph of T^c corresponds to a minimum dominating set of G. Clearly, the reduction from the dominating set problem is in polynomial time, and the theorem is proved. $\square$

By the reduction, it follows that MCTS is NP-complete even when restricted to trees of height 3.

We recall that a graph G is called an *interval graph* if one can assign to each v in V an interval $I_v \subset \mathbb{R}$ such that $I_u \cap I_v$ is nonempty if and only if $uv \in E$.

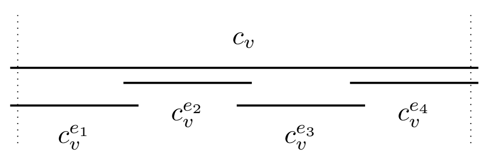
Theorem 2.2. *MCTS is NP-Hard for interval graphs, even when restricted to connected colorful subgraphs.*

Proof: As noted in the proof of Theorem 2.1, MCTS is in NP. We will show it is NP-hard by a reduction from the VERTEX COVER problem (VC). Consider an instance of VC on a graph G with n vertices and m edges and an integer k. To this instance we will associate a set of colored intervals. We introduce first the colors as follows. The colors are

- for each edge $e = (u, v) \in E$, two colors c_u^e and c_v^e,
- for each vertex $u \in V$, a color c_u, and
- colors c_{left} and c_{right}.

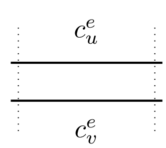
The set of intervals will be partitioned into subsets, called gadgets, and those gadgets will be ordered. The intervals which right extremity is the rightmost of gadget j will intersect with the intervals which left extremity is the leftmost of gadget $j + 1$. Apart from this rule, intervals will only intersect with other intervals from the same gadget. The very first gadget is going to contain only one interval of color c_{left}, and the very last gadget is going to contain only one interval of color c_{right}. In between, there will be a number of gadgets from the following three types, whose respective order do not matter for the proof.

Type 1: For each vertex v in V, the gadget g_v is defined as follows. Let $e_1, e_2, e_3, \ldots, e_{d(v)}$ be the edges adjacent to v. There are $d(v)$ intervals of color $c_v^{e_1}, c_v^{e_2}, \ldots, c_v^{e_{d(v)}}$, and one interval of color c_v. The interval of color $c_v^{e_i}$ intersects only the interval of color $c_v^{e_{i-1}}$, the interval of color $c_v^{e_{i+1}}$ (when those intervals exist), and the interval of color c^v. The intervals c_v and $c_v^{e_1}$ (respectively $c_v^{e_{d(v)}}$) have the same leftmost (respectively rightmost) extremity. For instance, if v is a vertex of degree four, the intervals are defined as follows.

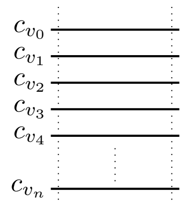


An interval will intersect with some intervals from the previous (respectively next) gadget if and only if it crosses the left (respectively right) dotted line.

Type 2: For each edge $e = uv$ in E, gadget g'_e is defined as follows. There are two intervals of color c_u^e and c_v^e, which share both their left and their right boundaries.



Type 3: The last type of gadget g'' uses n intervals with the same boundaries and colors $c_{v_1}, c_{v_2}, \ldots, c_{v_n}$, as follows.



So, there are n gadgets of type 1 (one for each vertex of G), m gadgets of type 2 (one for each edge of G), and we include $n-k$ gadgets of type 3. Figure 1 illustrates the reduction. Let us consider the interval graph implied by the obtained set of intervals. By coloring each vertex from this interval graph with the color of the corresponding interval, we obtain a vertex-colored interval graph I^c.

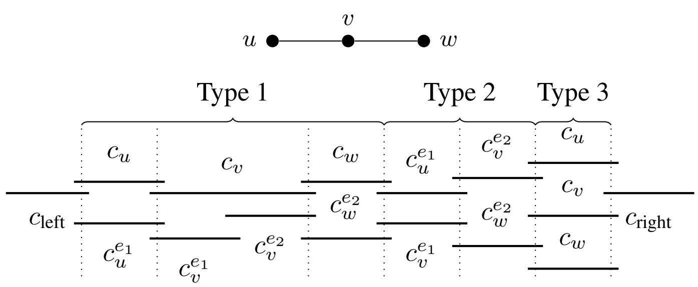


Fig. 1: A graph, and the set of intervals obtained by applying the reduction used in the proof of Theorem 2.2 when $k = 2$.

Let S be a vertex cover of G of size k. We will show a connected colorful subgraph of I^c can be build from S. Consider the set W of vertices from I^c associated to the following intervals:

- The two intervals of color c_{left} and c_{right},
- for each vertex $u \in S$, we take the intervals of color c_u from the type 1 gadget corresponding to u, along with all the intervals of color $c_u^{e_i}$ from the type 2 gadgets,
- for each vertex v_i among $\{v_1, v_2, \ldots, v_{n-k}\} = V \setminus S$, we take all the intervals of color $c_{v_i}^{e_j}$ from the type 1 gadget corresponding to v_i, along with the vertex of color c_{v_i} from the i-th gadget of type 3.

By construction, the obtained set of intervals will include exactly one interval of each color. Let us show that the union of all the intervals in the set is connected. To do so, let us show that for every gadget, we have intervals in W whose union is covering the whole gadget. This is the case directly for gadgets of type 1. This is the case for a gadget of type 2 because there is always at least one endpoint of the corresponding edge that belongs to S. This is also the case for gadgets of type 3 because there are exactly $n-k$ vertices in $G \setminus S$. Therefore, the set of vertices in I^c associated to this set of intervals is a connected colorful subgraph.

Let T be a set of intervals that correspond to a connected colorful subgraph of I^c. We will show how T implies a vertex cover of G of size k. We consider the set S of vertices of G such that a vertex u is in S if and only if there exists an edge e of G such that T contains the interval of color c_u^e from gadget $g'(e)$.

By construction of I^c, there is only one vertex of I^c with color c_{left}, and one with color c_{right}. Therefore, any connected subgraph of I^c must contain those two vertices. As the subgraph needs to be connected, it must also contain a path linking those two vertices. Therefore, for every gadget, the union of the intervals of T must include the whole gadget.

Let us consider the gadget $g'(e)$ for some edge e of G with $e = (u, v)$. The set T must contain either the interval of color c_u^e or the interval of color c_v^e. Therefore, at least one of the endpoints of e must be included in S. The set S is hence a vertex cover.

Consider now the gadgets of type 3. The set T must contain an interval from each of those $n - k$ gadgets, and each of those intervals must be of a different color. Hence there are $n-k$ vertices u_1, u_2,..., u_{n-k} such that T contains an interval of color c_{u_i} in a gadget of type 3.

Finally, consider the gadget $g(u)$ for some vertex u of G adjacent to edges e_1, e_2, e_3, ..., $e_{d(v)}$. The set T must contain either the interval of color c_u or all the intervals of color $c_v^{e_1}$, $c_v^{e_2}$, ..., $c_v^{e_{d(v)}}$. If S contains the vertex v, it means there exists some i such that T contains an interval of color $c_v^{e_i}$ from a gadget of type 2. As T contains exactly one interval of each color, it means that T cannot contain the interval of color $c_v^{e_i}$ in $g(u)$. As intervals of T must cover the whole of the gadget $g(u)$, T contains the interval of color c_u. This means that T does not contain another interval of color c_u from another gadget. As T already contains intervals of color c_{u_1}, c_{u_2}, ..., $c_{u_{n-k}}$, this means S can contain at most k vertices. □

A graph G is called a *split graph* if V can be partitioned into sets V_0 and V_1 such that the subgraphs induced by V_0, and V_1, are a clique and an independent set, respectively.

Theorem 2.3. *MCTS is NP-Hard for split graphs.*

Proof: We show that a polynomial algorithm for MCTS on split graphs can be used to solve VERTEX COVER on all graphs. To a given graph G with n vertices and m edges, we associate the vertex-colored split graph S_G^c defined as follows:

- For each vertex v) of G there is in S_G^c a vertex $S_G^c(v)$. All vertices $S_G^c(v)$ are pairwise adjacent, and are colored with color c_0.
- For each edge uv of G there is in S_G^c a vertex $S_G^c(uv)$ adjacent to $S_G^c(u)$ and $S_G^c(v)$. Each vertex $S_G^c(uv)$ is colored with a unique color.

S_G^c is a split graph colored with $m+1$ colors, and we can partition the set of vertices of S_G^c into sets V_0 and V_1 where V_0 is the set of vertices of color c_0 (which induce a clique) and V_1 is the set of the remaining vertices (which induce an independent set). We show that a bijection exists between the set of minimum connected tropical subgraphs of S_G^c and the set of optimal solutions to VERTEX COVER in G.

Let X be the vertices of a minimum connected tropical subgraph in S_G^c. As $V_1 \subseteq X$, let us write $V_2 = X \backslash V_1$. Observe that V_2 defines a vertex cover in G. Indeed, X is connected in S_G^c and so V_2 contains at least one of $S_G^c(u)$ or $S_G^c(v)$ for every $S_G^c(uv)$. On the other hand, every vertex cover of G of cardinality k defines in S_G^c a tropical set of cardinality $k + m$, which is necessarily connected as V_0 is a clique. Consequently, computing the minimum connected tropical subgraph in S_G^c determines the minimum Vertex Cover of G. □

Theorem 2.4. *MCTS can be solved in* $O(n^2 \times m \times 8^c)$ *time, where* n *and* m *are respectively the number of vertices and edges of* G^c.

Proof: We show that we can compute for each vertex $u \in G^c$ the function $f_u : \mathcal{P}(\mathcal{C}) \to \{1, \ldots, n\}$ which associates to a set of color S the order of the smallest connected subgraph containing u and at least one vertex of each color in S. The optimal value of MCTS is then the smallest value $f_v(\mathcal{C})$ reached for any vertex v. The algorithm is the following :

- Step 1: For $u \in V$, initialize $f_u(S) := 1$ when $S = \{c(u)\}$ and $f_u(S) := n$ otherwise.
- Step 2: While there exists an edge $e = uv$ in G^c and two sets of colors $S_u, S_v \in \mathcal{P}(\mathcal{C})$ such that $f_u(S_u) + f_v(S_v) < f_u(S_u \cup S_v)$, update f_u by setting $f_u(S_u \cup S_v) := f_u(S_u) + f_v(S_v)$.

Let us prove that the algorithm above is correct. Let us first show, by induction on the number of iterations, that at any iteration of Step 2, if $f_u(S) = k$, then there exists a connected subgraph of G^c of order k that contains at least one vertex of each color in S. After Step 1, $\{u\}$ is a suitable subgraph if $S = \{c(u)\}$, and G^c itself is suitable otherwise. Now suppose that the property is true before some iteration of Step two on edge uv of the algorithm. Let H_u (respectively, H_v), be the subgraph of order $f_u(S_u)$ (respectively, $f_v(S_v)$) containing u, (respectively, v), and at least one vertex of each color in S_u (respectively, S_v). By taking the union of H_u and H_v, we obtain a subgraph of order at most $f_u(S_u) + f_v(S_v)$ containing every color in $S_u \cup S_v$. This proves the claim. Hence, for every vertex u in G^c, $\mathfrak{tc}(G^c) \leq f_u(\mathcal{C})$.

We will show that the computed values of f_u correspond to the definition we gave of the function. Let us suppose, for contradiction, that at the end of the algorithm there is a vertex u and a connected subgraph H of order k containing u such that $k < f_u(c(H))$. Consider a spanning tree T of H rooted at u. For a vertex v in T, we denote by $T(v)$ the subtree of T rooted at v. Now, we claim that for every vertex v in T, $f_v(c(T(v))) \leqslant |T(v)|$. This is obvious if v is a leaf, as in this case $T(v) = \{v\}$, and $f_v(\{c(v)\}) = 1$ by Step 1 of the algorithm. Thus, we may suppose that the claim is true for all children $v_1, \ldots, v_r$ of a vertex v, and show that it holds for v. Since we can not apply Step 2 of the algorithm on the edge vv_1, with sets $c(v)$ and $c(T(v_1))$, it means that $f_v(c(v) \cup c(T(v_1))) \leqslant f_v(c(v)) + f_{v_1}(c(T(v_1)))$. We know that $c(v) \cup c(T(v_1)) = c(v \cup T(v_1))$, $f_v(c(v)) = 1$ (by Step 1 of the algorithm) and $f_{v_1}(c(T(v_1))) \leqslant |T(v_1)|$ (by induction), therefore $f_v(c(v \cup T(v_1))) \leqslant 1 + |T(v_1)|$. By the same reasoning, since we cannot apply Step 2 of the algorithm on vv_2, with sets $c(\{v\} \cup T(v_1))$ and $c(T(v_2))$, it means that $f_v(c(v \cup T(v_1) \cup T(v_2))) \leqslant 1 + |T(v_1)| + |T(v_2)|$. By repeating this argument for each child of v, we obtain that $f_v(c(T(v))) = f_v(c(v \cup T(v_1) \cup \cdots \cup T(v_j))) \leqslant 1 + |T(v_1)| + \cdots + |T(v_j)| = |T(v)|$. Hence for every vertex v in T, $f_v(c(T(v))) \leqslant |T(v)|$. Thus, $f_u(c(H)) = f_u(c(T(u))) \leqslant |T(u)| = k$, a contradiction. This proves the correctness of the algorithm.

Let us prove next the complexity of the algorithm. Setting up the initial values of every $f_u(S)$ in Step 1 can be done in $O(n \times 2^c)$. The identification of an edge uv and two sets S_u, S_v, suitable to apply Step 2 of the algorithm, takes at most $m \times 2^c \times 2^c$ operations. Applying Step 2 will strictly decrease the value of $f_u(S_u \cup S_v)$. There are only n functions on 2^c values, and each function can decrease at most n times on each value. Therefore, Step 2 is iterated at most $n \times 2^c \times n$ times. So the complexity is at most $O(n^2 \times m \times 8^c)$, as required. □

3 Sufficient Conditions

In this section, we give sufficient conditions for a vertex-colored graph to have connected colorful subgraphs of small order. Our first result relates $\mathfrak{tc}(G^c)$ to $\chi(G)$.

Proposition 3.1. *If G^c is a properly colored graph on $\chi(G)$ colors, then it contains a connected colorful subgraph.*

Proof: Let $V_1 \subseteq V$ be a color class of G^c. There must exist a vertex $v \in V_1$ whose neighborhood contains all the other colors, as otherwise all vertices of V_1 could be recolored with a color that does not appear in their neighborhood, yielding a proper coloring of G with $\chi(G) - 1$ colors. But now, $G^c[N[v]]$ contains a connected colorful subgraph. $\square$

Before we prove the next result, we need the following lemma.

Lemma 3.2. *Let G be a connected graph with n vertices and m edges. If G contains (at least) i cut vertices then*

$$m \leq \binom{n-i}{2} + i$$

Proof: We prove the result by induction on i, knowing that it holds when $i = 0$. We therefore assume that $i > 0$. Let v be a non-cut vertex from a leaf block of G. If v has degree 1 then $G\backslash v$ has at least $i - 1$ cut vertices, and by induction

$$m \leq |E(G\backslash v)| + 1 \leq \binom{(n-1)-(i-1)}{2} + (i-1) + 1 = \binom{n-i}{2} + i$$

Otherwise $G\backslash v$ has a set C of at least i cut vertices and v is adjacent with at most one of them. Note, however, that v cannot be adjacent to all of $V(G)\backslash C$, as every cut vertex splits $V(G)\backslash C$ into (at least) two non-empty connected components. Therefore, v has degree at most $(n - i - 2) + 1 = n - i - 1$ and by induction

$$m \leq |E(G\backslash v)| + n - i - 1 \leq \binom{n-i-1}{2} + (n-i-1) + i = \binom{n-i}{2} + i$$

$\square$

Theorem 3.3. *Let G^c be a connected vertex-colored graph with n vertices and m edges. For every non-negative integer $k \leq n - 4$, if $m \geq \binom{n-k-2}{2} + n - c + 2$, then $\mathfrak{tc}(G^c) \leq c + k$.*

Proof: By induction on n. If $n \leq c + k$, then G^c itself is a connected tropical subgraph of order at most $c+k$. We may therefore assume that $n \geq c+k+1$. Let $F \subseteq V$ be the set of vertices whose colors appear at least twice in the graph. Then $|F| \geq n - c + 1$ since at most $c - 1$ colors appear exactly once.

Let $i = n - c + 1$. We have assumed that $k \leq n - c - 1$, which implies that

$$m \geq \binom{n-k-2}{2} + n - c + 2 \geq \binom{n-i}{2} + i + 1$$

Using Lemma 3.2, we know that there is a vertex $v \in F$ such that v is not a cut vertex.

We assume first that $d(v) \geq n - k - 1$. Let $N[v]$ be the closed neighborhood of v. Then $G \setminus N[v]$ is of order at most k. Let p be the number of colors in $N[v]$. We will build a connected tropical subgraph of order at most $k + c$. First we take v and $p - 1$ of its neighbors colored with the $p - 1$ remaining colors.

For each missing color z, we add a connected component H of $G \setminus N[v]$ which contains z and one vertex from the neighborhood of v to keep H connected to v. In the worst case, we add every vertex in the graph $G \setminus N[v]$ and $c-p$ vertices in the neighborhood of v, which yields a connected tropical subgraph of order at most $c+k$.

Now, assume that $d(v) \leq n-k-2$. Let $G' = G \setminus \{v\}$ on n' vertices and m' edges. Then G' is connected (by definition of v), $n' = n-1$, G' is colored with c colors (as $v \in F$) and

$$\begin{aligned} m' &\geq m-(n-k-2) \\ &\geq \binom{n-k-2}{2} + n - c + 2 - (n-k-2) \\ &= \binom{n-k-2}{2} + n - c + 1 - (n-k-3) \\ &= \binom{n-k-3}{2} + n - c + 1 \\ &= \binom{n'-k-2}{2} + n' - c + 2. \end{aligned}$$

By induction, there exist a connected tropical subgraph of order $c+k$ in G'. It is also a tropical subgraph of order $c+k$ in G. This completes the argument and the proof. □

Note that the above proof leads to a polynomial time algorithm that finds a connected tropical subgraph of order $c+k$ under the hypothesis of the theorem. We now show that the bound given in the above theorem is tight. Fix two positive integers n and k. Consider now a rainbow complete graph K^c_{n-k-2} on $n-k-2$ vertices. Let x_1 be a vertex of color 1 in K^c_{n-k-2}. Add a path $x_1 v_1 v_2 \dots v_{k+2}$, $v_i \notin V(K^c_{n-k-2})$. Color v_{k+2} with color 0 and every other v_i with color 1. The resulting graph has exactly $\binom{n-k-2}{2} + n - c + 1$ edges, but has no connected tropical subgraph of order less than $c+k+1$.

Theorem 3.4. *Let G^c be a vertex-colored graph of minimum degree δ. If $\delta \geq \frac{n}{2}$ and $c \geq \frac{n}{2}$, then G^c has a connected colorful subgraph.*

Proof: Let S be the vertices of a largest connected rainbow subgraph of G^c. Assume $|S| < c$, otherwise the proof is done. As $|S| < c$, there exists a vertex $v \in G^c \setminus S$ which color does not appear in S. Also v has no neighbor in S, as S is maximal. Now, we distinguish between two cases depending upon the connectivity of S.

Suppose first S is 2-connected. For each vertex $u \in S$, $|N(u) \cap N(v)| \geq 2$, since u and v are not adjacent and $\delta \geq \frac{n}{2}$. If a vertex $w \in N(u) \cap N(v)$ is colored with a different color than u, then S can be extended to another connected rainbow subgraph, say S', by removing from S at most one vertex of the same color as w and adding to S vertices w and v. As S is 2-connected and only one vertex is removed from S, S' is connected. Furthermore, by definition, S' is rainbow, and $|S'| > |S|$, a contradiction. Thus, every vertex in $N(u) \cap N(v)$ has the same color as the vertex u. Since this is true for every u in S, $N(v)$ contains every color in S and there is a connected rainbow subgraph of G^c of order $|S|+1$ contained in $N[v]$, a contradiction to the maximality property of S.

Suppose next S is not 2-connected. Let U be a subset of S containing exactly one non-cut vertex from each leaf block of S. We define $T = \{w | w \in V \setminus S, \text{such that } w \text{ is neighbor of a vertex in S}\}$. Clearly, every color which appears in T also appears in S. This implies that $|V \setminus T| \geq c \geq n/2$ and, in turn, that $|T| \leq n/2$.

We now consider $e(U \cup \{v\}, T)$, i.e., the number of edges between $U \cup \{v\}$ and T. Note that a vertex $u \in U$ contained in a leaf block $B \subseteq S$ has at most $|B| - 1$ neighbors in S. It follows,

$$
\begin{aligned}
e(U \cup \{v\}, T) \geq & \overbrace{\Big(|U|\frac{n}{2} - |S| + 1\Big)}^{e(U,T)\geq} + \overbrace{\Big(\frac{n}{2} - \Big(n - |S| - |T|\Big) + 1\Big)}^{e(v,T)\geq} \\
=&(|U|+1)\frac{n}{2} - n + |T| + 2 \\
=&(|U|-1)\frac{n}{2} + |T| + 2 \\
\geq&(|U|-1)|T| + |T| + 2 \\
>&|U||T|.
\end{aligned}
$$

Thus, there is a vertex $u \in T$ adjacent to at least $|U| + 1$ vertices in $U \cup \{v\}$, i.e., to all of them. As u is connected to a non cut-vertex in every leaf block of S, $S \cup \{u\}$ is 2-connected. Let u' be the vertex of S colored with the same color as u. Then $(S \setminus \{u'\}) \cup \{u\} \cup \{v\}$ is a connected rainbow subgraph of order $|S| + 1$, a contradiction to the maximality property of S. $\square$

The proof of Theorem 3.4 immediately yields the following.

Corollary 3.5. *Let G^c be a vertex-colored graph of order n and of minimum degree δ. If $\delta \geq \frac{n}{2}$ and $c \geq \frac{n}{2}$, then a connected colorful subgraph can be found in polynomial time.*

We let $\delta_r(G^c)$ denote the minimum rainbow degree of G^c, i.e., the smallest number of colors a vertex in G^c has in its neighborhood.

Theorem 3.6. *The following holds:*

1. *For every $\epsilon, \epsilon' \in [0, 1)$, there exists a vertex-colored graph G^c such that $\delta(G^c) \geq \epsilon n$, $\delta_r(G^c) \geq \epsilon' c$ and G^c has no connected colorful subgraph.*
2. *For every positive integer p, there exists a vertex-colored graph G^c such that $\delta(G^c) \geq n - c + p$ and G^c has no connected colorful subgraph.*
3. *For every positive integer p and $\epsilon \in [0, 1)$, there exists a vertex-colored graph G^c such that $\delta(G^c) \geq \epsilon n$, G^c is p-connected and has no connected colorful subgraph.*

Proof: Let i, j, k be three positive integers, $k \geq 2$. We first define an uncolored graph $G(i, j, k)$ that will be used to prove all the three parts of the theorem. The graph $G(i, j, k)$ is defined as follows:

- $G(i, j, k)$ is composed of k vertex-disjoint cliques $H_1, H_2, \ldots, H_k$, each of order i, and k vertex-disjoint cliques $D_1, D_2, \ldots, D_k$, each of order j.
- For all $l \neq l'$, every vertex of H_l is adjacent to every vertex of $H_{l'}$ and to every vertex of $D_{l'}$.

$G(i, j, k)$ has the following properties :

- $n = |V(G(i, j, k))| = k(i + j)$.

- $\delta = (k-1)i + j - 1$.
- $G(i,j,k)$ is $i(k-1)$-connected.

By varying i, j, k and the coloring we can prove the three parts of the theorem as follows.

1. We consider $G(i,j,k)$ with each vertex in H_l colored with color 1, for $l = 1, \ldots, k$. Every D_l contains one vertex of colors $2, 3, \ldots j$ and one vertex of color $j + l$. The graph $G(i,j,k)$ colored this way is denoted by G^c. It satisfies the following properties :

   - $c = j + k$.
   - $\delta_r(G^c) = j$.

   Let $\epsilon, \epsilon' \in [0, 1)$. Let $k > \frac{1}{1-\epsilon}$ and $j > \frac{\epsilon' k}{1-\epsilon'}$. Then $\delta_r(G^c) = j = \epsilon' j + (1-\epsilon')j > \epsilon'(j+k) = \epsilon' c$ and $\frac{\delta}{n} \to \frac{k-1}{k} > \epsilon$ when $i \to \infty$. For i sufficiently large, we have a graph with $\delta > \epsilon n$, $\delta_r > \epsilon' c$. It has no connected colorful subgraph. Indeed, such a tropical subgraph would have to contain a vertex from each D_l. Thus, it would need to contain vertices in at least two H_l. This contradicts the fact that each color is present only once.

2. We consider $G(i,j,k)$ with each vertex in H_l colored with color 1, for $l = 1, \ldots, k$. Now, color all the jk vertices of the D_l with colors $\{2, 3, \ldots, jk+1\}$ so that each color appears on exactly one vertex. Let G^c denote the resulting colored graph. Then G^c is colored with $c = jk + 1$ colors.

   Given $p \in \mathbb{N}$, choose j such that $j \geq i + p$. Then $\delta(G^c) = (k-1)i + j - 1 \geq ki + p - 1 = n - c + p$. Graph G^c has no connected colorful subgraph. Indeed, such a tropical subgraph would need to contain every vertex in each D_l. Thus, it would need to contain vertices in at least two H_l to be connected. This contradicts the fact that each color is present only once.

3. We consider $G(i,j,k)$ colored the same way as in Case 2. Let G^c be the obtained graph. Given $p \in \mathbb{N}$, and $\epsilon \in [0, 1)$, we choose any $j \in \mathbb{N}$, and k such that $k > \frac{1}{1-\epsilon}$. If $i \geq \frac{p}{k-1}$, then G^c is p-connected. Also $\frac{\delta}{n} \to \frac{k-1}{k} > \epsilon$ when $i \to \infty$. For i sufficiently large, G^c is p-connected with $\delta(G^c) \geq \epsilon n$. By the argument in Case 2, G^c has no connected colorful subgraphs.

□

4 Random graphs

In this section we are interested in the problem of finding particular tropical subgraphs in a random graph such as cliques and connected components. Recall that the random graph $G(n,p)$ is the graph with vertex set $V = \{1, \ldots, n\}$ in which each of the possible $\binom{n}{2}$ edges appears with probability p, independently. In other words, if G is a graph with vertex set V and has m edges, then

$$\mathbb{P}[G(n,p) = G] = p^m (1-p)^{\binom{n}{2} - m}.$$

For more background on random graphs, we refer the reader to [Bol01] and [JLR00].

In our model, we will study a *randomly vertex-colored random graph*. Given a positive integer c, let $G(n,p,c)$ be the graph obtained from $G(n,p)$ by coloring each vertex with one of the colors $1, 2, \ldots, c$

uniformly and independently at random. The vertex coloring is independent of the existence of edges. Clearly, for any given vertex-colored graph G^c on V with m edges, we have

$$\mathbb{P}[G(n,p) = G^c] = \frac{p^m(1-p)^{\binom{n}{2}-m}}{c^n}.$$

We will say that $G(n,p,c)$ has a property $\mathcal{Q}$ *asymptotically almost surely* (abbreviated *a.a.s.*) if the probability it satisfies $\mathcal{Q}$ tends to 1 as $n \to \infty$.

We begin by recalling some notation and results that will be needed. Let X be a random variable. We denote by $\mathbb{E}(X)$ and $\mathbb{V}\mathrm{ar}(X)$ the *expectation* and the *variance* of X, respectively. For $r \geq 0$, $(n)_r = n(n-1)\dots(n-r+1)$ denotes the *falling factorial*. $\mathbb{E}(X)_r$ is called the *r-th factorial moment* of X. In particular, $\mathbb{E}(X)_0 = 1$ and $\mathbb{E}(X)_1 = \mathbb{E}(X)$. Let $X_1, X_2, \dots, X_n$ and X be integer-valued random variables. We say that X_n *converges in distribution* to X, as $n \to \infty$, and write $X_n \xrightarrow{d} X$, if $\mathbb{P}[X_n = k] \to \mathbb{P}[X = k]$ for every integer k.

The following useful bound is known as the *Chebyshev's inequality* which states that, for $t > 0$

$$\mathbb{P}\left[|X - \mathbb{E}(X)| \geq t\right] \leq \frac{\mathbb{V}\mathrm{ar}(X)}{t^2}.$$

In particular, if $\mathbb{E}(X) > 0$ and by setting $t = \mathbb{E}(X)$, we have

$$\mathbb{P}\left[X = 0\right] \leq \frac{\mathbb{V}\mathrm{ar}(X)}{\mathbb{E}^2(X)}.$$

The standard *second moment method* is based on Chebyshev's inequality. It consists in showing that, for a given sequence of non-negative, integer-valued random variables (X_n), $\mathbb{V}\mathrm{ar}(X_n)/\mathbb{E}^2(X_n)$ tends to 0 as $n \to \infty$, and thus concluding that $X_n > 0$ *a.a.s.*

Markov's inequality states that, if $X > 0$ and $t > 0$, then

$$\mathbb{P}\left[X \geq t\right] \leq \frac{\mathbb{E}(X)}{t}.$$

In particular, if $X_1, X_2, \dots, X_n$ are non-negative, integer-valued random variables, then $\mathbb{E}(X_n) \to 0$ as n $n \to \infty$ implies $\mathbb{P}[X_n = 0] \to 1$.

The following result is a variant of the so called *method of moments*. Let X be a random variable with a distribution that is determined by its moments (see [JLR00], p. 140, for a definition). If $X_1, X_2, \dots, X_n$ are random variables with finite moments ($\mathbb{E}(|X|^r) < \infty, r \geq 1$) such that $\mathbb{E}(X_n)_r \to \mathbb{E}(X)_r$ as $n \to \infty$ for every integer $r \geq 1$, then $X_n \xrightarrow{d} X$. An example of distribution which is determined by its moments is the *Poisson distribution*.

We will use the following asymptotic notation. Let $\{a_n\}$ and $\{b_n\}$ be two sequences of real numbers. For simplicity we assume that $a_n, b_n > 0$.

- $a_n = O(b_n)$ if there exist constants $n_0 \in \mathbb{N}$ and $C > 0$ such that $a_n \leq Cb_n$ for $n \geq n_0$.
- $a_n = o(b_n)$ and $a_n \ll b_n$ mean $a_n/b_n \to 0$ as $n \to \infty$.
- $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$.

4.1 Threshold for small tropical subgraphs

Let G be a fixed graph with v_G vertices and e_G edges. We denote by $a = |Aut(G)|$ the cardinality of the automorphism group of G. One of the first problems studied by Erdős and Rényi in [ER60] was that of the existence of at least one copy of G in $G(n, p)$. They determined the threshold function for that property in the special case in which G is a balanced graph (see below for the definition). Later in [Bol81] Bollobás extended this result to any arbitrary fixed graph. Formally, the threshold function for the property of containing a copy of G is $n^{-1/\rho(G)}$ where $\rho(G)$ is the ratio of the number of edges to the number of vertices in the densest subgraph of G, that is,

$$\rho(G) = \max\left\{\frac{e_H}{v_H} \;:\; H \subseteq G\;,\; v_H > 0\right\},$$

where v_H and e_H stand for the number of vertices and edges of H, respectively.

The next theorem follows from the result of Bollobás and shows that the above threshold also holds for the property of the existence of a tropical copy of a given graph in $G(n, p, c)$. In what follows, we will denote by $X_n = X_n(G)$ the number of tropical copies of G in $G(n, p, c)$. That is,

$$X_n = \sum_{G'} I_{G'},$$

where the sum is over all copies G' of G, and

$$I_{G'} = 1\left\{G(n, p, c) \supset G' \textit{ and } G' \textit{ is tropical}\right\}.$$

Theorem 4.1. *Let G be a fixed graph with at least one edge, $e_G > 0$. Let $c = v_G = |V(G)|$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left[G(n, p, c) \supset \text{tropical copy of } G\right] = \begin{cases} 0 & \text{if } p \ll n^{-1/\rho(G)} \\ 1 & \text{if } p \gg n^{-1/\rho(G)}. \end{cases}$$

Proof: The theorem clearly holds if $p \ll n^{-1/\rho(G)}$. Now we assume that $p \gg n^{-1/\rho(G)}$. We need to show that $\mathbb{V}\mathrm{ar}(X_n)/\mathbb{E}^2(X_n) \to 0$ as $n \to \infty$. The second moment of X_n is given by

$$\mathbb{E}(X_n^2) = \sum_{G',G''} \mathbb{E}(I_{G'} I_{G''}) = E_1 + E_2,$$

where

$$E_1 = \sum_{V(G')\cap V(G'')=\emptyset} \mathbb{E}(I_{G'} I_{G''}) \;\; \textit{and} \;\; E_2 = \sum_{V(G')\cap V(G'')\neq\emptyset} \mathbb{E}(I_{G'} I_{G''}).$$

As c is fixed, we have

$$E_1 = \binom{n}{c}\binom{n-c}{c}\left(\frac{c!p^{e_G}}{ac^c}(n)_c\right)^2 = (1 + o(1))E^2(X_n).$$

So, to complete the proof it suffices to show that $E_2/\mathbb{E}^2(X_n) = o(1)$. Since $\mathbb{P}[G', G'' \textit{ are tropical}\,] \leq c!/c^c$, it follows that

$$E_2 \leq \frac{c!}{c^c} \sum_{V(G')\cap V(G'')\neq\emptyset} \mathbb{P}[G(n, p) \supset G', G''].$$

Let Y_n denote the number of copies of G in $G(n,p)$. By splitting $\mathbb{E}(Y_n^2)$ into two parts in the same way as for $\mathbb{E}(X_n^2)$, we obtain

$$\sum_{V(G')\cap V(G'')\neq\emptyset} \mathbb{P}[G(n,p)\supset G',G''] = \mathbb{E}(Y_n^2) - \mathbb{E}^2(Y_n) + o(1)\mathbb{E}^2(Y_n).$$

Since $\mathbb{E}(X_n) = c!/c^c\mathbb{E}(Y_n)$, we get

$$\frac{E_2}{\mathbb{E}^2(X_n)} \leq \frac{c^c\mathbb{V}\mathrm{ar}(Y_n)}{c!\mathbb{E}^2(Y_n)} + o(1).$$

Thus, as c is fixed and $\mathbb{V}\mathrm{ar}(Y_n)/\mathbb{E}^2(Y_n) = o(1)$, it follows that $E_2/\mathbb{E}^2(X_n) = o(1)$, which completes the proof. □

In the next theorem we investigate the case in which $pn^{1/\rho(G)} \to \theta$ as $n \to \infty$, where θ is a positive constant. We are specially interested in a family of graphs called strictly balanced graphs defined as follows. A graph G is *balanced* if $\rho(G) = e_G/v_G$, that is, if $e_H/v_H \leq e_G/v_G$ for every $H \subset G$. G is *strictly balanced* if $e_H/v_H < e_G/v_G$ whenever $H \subsetneq G$, that is to say that every proper subgraph of G is strictly less dense than the graph itself. Trees, cycles and complete graphs are strictly balanced.

Theorem 4.2. *Let G be a fixed strictly balanced graph with v vertices and e edges. Denote by $a = |Aut(G)|$ the number of elements of the automorphism group of G. Let θ be a positive constant and set $p = \theta/n^{v/e}$. Let X_n denote the number of tropical copies of G in $G(n,p,c)$ with $c = v$. Then*

$$X_n \xrightarrow{d} \mathcal{P}(\lambda) \quad \textit{with} \quad \lambda = \frac{c!\theta^e}{ac^c},$$

where $\mathcal{P}(\lambda)$ is the Poisson distribution with mean λ.

Proof: The proof uses the method of moments described above. The expectation of X_n is easily estimated as follows.

$$\mathbb{E}(X_n) \sim \frac{c!\theta^e}{ac^c} = \lambda.$$

It is not hard to see that, for $r \geq 2$, the r-th factorial moment of X_n is equal to the expected number of ordered r-tuples $(G_1,\ldots,G_r)$ of tropical copies of G in $G(n,p,c)$, that is,

$$\mathbb{E}(X_n)_r = \sum_{G_1,\ldots,G_r} \mathbb{P}\left[I_{G_1}=1,\ldots,I_{G_r}=1\right].$$

We split $\mathbb{E}(X_n)_r$ into two parts

$$\mathbb{E}(X_n)_r = E'_r + E''_r.$$

E'_r is the expected number of ordered r-tuples of mutually vertex disjoint copies of G, while E''_r takes into consideration the other cases. We have

$$E'_r = \binom{n}{c}\binom{n-c}{c}\cdots\binom{n-(r-1)c}{c}\left(\frac{c!}{c^c}\right)^r\left(\frac{c!}{a}\right)^r p^{re} = (n)_{rc}\left(\frac{c!}{ac^c}\right)^r p^{re}.$$

Since c is fixed, it follows that

$$E'_r \sim \lambda^r = \mathbb{E}\,(X)_r\,,$$

where X is a random variable with Poisson distribution $\mathcal{P}(\lambda)$.

To complete the proof, it remains to show that $E''_r \to 0$ as $n \to \infty$. Clearly,

$$E''_r \le \sum_{G_1,\ldots,G_r} \mathbb{P}\left[G(n,p) \supset G_1, \ldots, G(n,p) \supset G_r\right].$$

It is shown in [JLR00], p. 67, that the right-hand side of the above inequality tends to 0 as $n \to \infty$, which completes the proof. □

4.2 Complete tropical subgraphs

One of the most interesting results in the study of random graphs was discovered by Matula [Mat76] who proved that the clique number $\mathrm{cl}(G(n,p))$ of $G(n,p)$ is asymptotically almost surely concentrated on two consecutive values. This result was also found independently by Bollobás and Erdős [BE76]. Let $0 < p < 1$ be fixed and set $b = 1/p$. Let the function $f(n)$ be defined by

$$f(n) = 2\log_b n - 2\log_b \log_b n + 1 + 2\log_b (e/2)\,.$$

Then, for any $\epsilon > 0$, the clique number of $G(n,p)$ satisfies

$$\mathbb{P}\Big[\lfloor f(n) - \epsilon \rfloor \le \mathrm{cl}\big(G(n,p)\big) \le \lfloor f(n) + \epsilon \rfloor\Big] \to 1 \;\; as \;\; n \to \infty.$$

This leads us to the natural question of what is the maximum number of colors $c = c(n)$ which *a.a.s.* guarantees the existence of a tropical clique of order r in $G(n,p,r)$, for every $r \le c(n)$. The answer to this question is given by the following theorem. In particular, it is shown that $c(n)$ differs from $f(n)$ by an additive constant (not depending on n).

Theorem 4.3. *Let $0 < p < 1$ be fixed. Let $c = c(n)$ be the function defined by*

$$c(n) = 2\log_b n - 2\log_b \log_b n - 2\log_b 2 + 1,$$

where $b = 1/p$. Then, for any $\epsilon > 0$, the following assertions hold.

(i) If $r > \lfloor c(n) + \epsilon \rfloor$, then a.a.s. there is no complete tropical subgraph of order r in $G(n,p,r)$.

(ii) If $r \le \lfloor c(n) - \epsilon \rfloor$, then a.a.s. $G(n,p,r)$ contains a complete tropical subgraph of order r.

Proof: The proof is based on the first and second moment methods. Let X_r be the random variable counting the number of tropical cliques of order r in $G(n,p,r)$. The expectation of X_r is given by

$$\mathbb{E}(X_r) = \binom{n}{r} p^{\binom{r}{2}} \cdot \frac{r!}{r^r} = \frac{(n)_r}{r^r} p^{r(r-1)/2}. \tag{1}$$

Using Stirling's formula, we get

$$\begin{aligned}\mathbb{E}(X_r) &= (1+o(1))\exp\left[\frac{r}{2}\left(2\log n + r\log p - \log p - 2\log r\right)\right]\\ &= (1+o(1))\exp\left[\frac{r\log b}{2}\Big(2\log_b n - r - 2\log_b r + 1\Big)\right],\end{aligned}$$

where $b = 1/p$. We observe that $\mathbb{E}(X_r)$ changes rapidly from $\omega(1)$ to $o(1)$ for values of r equivalent to $2\log_b n$. Indeed, let $\epsilon > 0$ be fixed, and set

$$c(n) = 2\log_b n - 2\log_b \log_b n - 2\log_b 2 + 1.$$

If $r > \lfloor c(n)+\epsilon\rfloor$, then, as r is an integer, we have $r \geq c(n)+\epsilon$. Using this lower bound, and replacing r by $(1-o(1))2\log_b n$ in $\log_b r$ of the above expression of $\mathbb{E}(X_r)$, we obtain

$$\mathbb{E}(X_r) \leq (1+o(1))\exp\left[\frac{r\log b}{2}\left(-\epsilon + o(1)\right)\right] = o(1).$$

Thus, by Markov's inequality, assertion *(i)* is proved.

Assume now that r is sufficiently large ($r\to\infty$) and $r \leq \lfloor c(n)-\epsilon\rfloor$. By a similar argument, we get

$$\mathbb{E}(X_r) \geq (1+o(1))\exp\left[\frac{r\log b}{2}\left(\epsilon + o(1)\right)\right] \to \infty \quad as \quad n\to\infty. \tag{2}$$

Thus, assertion *(ii)* will hold, if for every $r \leq \lfloor c(n)-\epsilon\rfloor$, $\mathbb{V}\mathrm{ar}(X_r)/\mathbb{E}^2(X_r) \to 0$ as $n\to\infty$. In what follows, we assume that $r = \lfloor c(n)-\epsilon\rfloor$. The case $r < \lfloor c(n)-\epsilon\rfloor$ can be done in the same way. First, we need to estimate $\mathbb{E}(X_r^2)$. Let S_1, S_2 be two subsets of vertices each of order r and having i vertices in common. Clearly,

$$\begin{aligned}\mathbb{P}\left[S_1 \text{ and } S_2 \text{ are tropical cliques}\right] &= \binom{r}{i}\frac{i!}{r^i}\left[\frac{(r-i)!}{r^{r-i}}\right]^2 p^{2\binom{r}{2}-\binom{i}{2}}\\ &= \frac{r!(r-i)!}{r^{2r-i}}p^{2\binom{r}{2}-\binom{i}{2}}.\end{aligned}$$

Thus,

$$\mathbb{E}(X_r^2) = \binom{n}{r}\sum_{i=0}^{r}\binom{r}{i}\binom{n-r}{r-i}\frac{r!(r-i)!}{r^{2r-i}}p^{2\binom{r}{2}-\binom{i}{2}}. \tag{3}$$

Relations (1) and (3) imply

$$\mathbb{E}(X_r^2) = \mathbb{E}^2\left(X_r\right)\Big[a_n + b_n\Big],$$

where

$$a_n = \binom{n}{r}^{-1}\left[\binom{n-r}{r} + r\binom{n-r}{r-1}\right]$$

and

$$b_n = \sum_{i=2}^{r} g(i),$$

with

$$g(i) = \binom{n}{r}^{-1}\binom{n-r}{r-i}\frac{r^i}{i!}b^{i(i-1)/2}.$$

We estimate a_n, for $r \sim 2\log_b(n)$, as follows.

$$a_n = 1 + O\left(\frac{(\log(n))^4}{n^2}\right).$$

To complete the proof, we need to show that, for $r = \lfloor c(n) - \epsilon \rfloor$, $b_n \to 0$. Let us consider the bottom term ($i = 2$) in the sum for b_n. We have

$$g(2) = b\binom{n}{r}^{-1}\binom{n-r}{r-2}\frac{r^2}{2} = O\left(\frac{(\log_b n)^2}{n^2}\right).$$

For $2 \le i \le r$,

$$\frac{g(i+1)}{g(i)} = \frac{b^i r(r-i)}{(i+1)(n-2r+i+1)} < \frac{r^2 b^i}{i(n-2r)}.$$

Let $t := \lfloor \alpha \log_b n \rfloor$, with $0 < \alpha < 1$. Then, for $2 \le i \le t-1$ and sufficiently large n, we have

$$\begin{aligned}
\frac{g(i+1)}{g(i)} &< \left(\frac{b^t}{2}\right)\left(\frac{r^2}{n-2r}\right) \\
&< \left(\frac{n^\alpha}{2}\right)\left(\frac{(2\log_b n)^2}{n-4\log_b n}\right) \\
&= \frac{4n^\alpha \log_b n}{2(n-4\log_b n)} \le 1.
\end{aligned}$$

It follows that, for sufficiently large n, the function $g(i)$ is decreasing. Thus

$$\sum_{i=2}^{t} g(i) < tg(2) = O\left(\frac{(\log_b n)^3}{n^2}\right).$$

Now we consider the second part of the sum for b_n. We have

$$\begin{aligned}
\sum_{i=t+1}^{r} g(i) &= \sum_{i=t+1}^{r}\binom{n}{r}^{-1}\binom{n-r}{r-i}\frac{r^i}{i!}b^{\binom{i}{2}} \\
&= \binom{n}{r}^{-1} b^{\binom{r}{2}}\frac{r^r}{r!}\sum_{i=t+1}^{r}\binom{n-r}{r-i}\frac{r^i}{i!}\frac{r!}{r^r}b^{\binom{i}{2}-\binom{r}{2}} \\
&= \mathbb{E}^{-1}(X_r)\sum_{i=t+1}^{r}\binom{n-r}{r-i}\frac{(r)_{r-i}}{r^{r-i}}b^{-\frac{(r-i)(r+i-1)}{2}}.
\end{aligned}$$

By setting $j = r - i$ and interchanging the order of summation, we get

$$\begin{aligned}\sum_{i=t+1}^{r} g(i) &= \mathbb{E}^{-1}(X_r) \sum_{j=0}^{r-t-1} \binom{n-r}{j} \frac{(r)_j}{r^j} b^{-\frac{j(2r-j-1)}{2}} \\ &< \mathbb{E}^{-1}(X_r) \sum_{j=0}^{r-t-1} \left((n-r) b^{-\frac{2r-j-1}{2}}\right)^j \\ &< \mathbb{E}^{-1}(X_r) \sum_{j=0}^{r-t-1} \left(n b^{\frac{-(r+t)}{2}}\right)^j.\end{aligned}$$

Since by assumption $r = \lfloor c(n) - \epsilon \rfloor \geq c(n) - 1 - \epsilon$, and as $t \geq \alpha \log_b n - 1$, we have, for n large enough,

$$n b^{\frac{-(r+t)}{2}} \leq \frac{2 b^{\frac{\epsilon+1}{2}} \log_b n}{n^{\alpha/2}} \leq 1.$$

Therefore

$$\sum_{i=t+1}^{r} g(i) \leq \frac{r-t}{\mathbb{E}(X_r)}.$$

Since, by (2), $\mathbb{E}(X_r) \geq (1 + o(1)) n^{\epsilon + o(1)}$, it follows that

$$\sum_{i=t+1}^{r} g(i) \leq O(1) \frac{\log n}{n^{\epsilon + o(1)}} \to 0 \quad \text{as} \quad n \to \infty.$$

The proof is complete. □

4.3 Tropical tree components

Let $G(n,p)$ be the random graph on n vertices with $p = \theta/n$, where θ is a positive constant. Erdős and Rényi discovered in their original work [ER60] that the structure of $G(n,p)$ undergoes sudden changes around $p = 1/n$. Roughly speaking, if $\theta < 1$ then $G(n,p)$ consists of small components, the largest of which is of order $O(\log n)$. While for $\theta > 1$ many of the small components join together to form a giant component of order $O(n)$. The remaining vertices are still in small components of order at most $O(\log n)$. This phenomenon is called *the double jump*, also known as the *phase transition phenomenon*. In the next theorem, we estimate the order of the largest tropical tree component in $G(n,p,c)$ at the subcritical phase ($\theta < 1$).

In what follows, we denote by T_k the number of components of $G(n,p,k)$ that are tropical trees of order k.

Theorem 4.4. *Let $p = \theta/n$, where $0 < \theta < 1$ is fixed. Let $\epsilon \in (0,1)$ be fixed. Set $k = k(n,\theta,\epsilon) = \left\lfloor (1-\epsilon) \dfrac{\log n}{\theta - \log \theta} \right\rfloor$. Then, asymptotically almost surely $G(n,p,k)$ has a tropical component of order k which is a tree.*

Proof: Using Cayley's formula for the number k^{k-2} of labeled trees of order k, we have

$$\begin{aligned}
\mathbb{E}(T_k) &= \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)+\binom{k}{2}-(k-1)} \frac{k!}{k^k} \\
&= \frac{n!}{(n-k)!} \frac{1}{k^2}\, p^{k-1}(1-p)^{kn-\frac{k^2}{2}-\frac{3k}{2}+1} \\
&= (1+o(1))\, n^k \frac{1}{k^2}\, p^{k-1}(1-p)^{kn-\frac{k^2}{2}-\frac{3k}{2}+1}.
\end{aligned}$$

Since $p = \theta/n = o(1)$ and $k = O(\log n)$, we have

$$\begin{aligned}
(1-p)^{kn-\frac{k^2}{2}-\frac{3k}{2}+1} &= e^{(kn-\frac{k^2}{2}-\frac{3k}{2}+1)(-p-\frac{p^2}{2}+o(p^2))} \\
&= e^{-knp+o(1)} \\
&= (1+o(1))\, e^{-knp}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}(T_k) &= (1+o(1))\, \frac{n^k p^{k-1} e^{-k\theta}}{k^2} \\
&= (1+o(1))\, \frac{n\theta^{k-1} e^{-k\theta}}{k^2} \\
&= (1+o(1))\, \frac{e^{\log n - k(\theta - \log\theta)}}{\theta k^2}.
\end{aligned}$$

Thus, for $k = (1-\epsilon)\dfrac{\log n}{\theta - \log\theta}$, we have $\mathbb{E}(T_k) = \dfrac{n^\epsilon}{\theta k^2} \to \infty$ as $n \to \infty$. To complete the proof, we need to compute the variance of T_k. Clearly,

$$\begin{aligned}
\mathbb{E}\left[T_k(T_k-1)\right] &= \binom{n}{k}\binom{n-k}{k} k^{2(k-2)} \left[\frac{k!}{k^k}\right]^2 p^{2(k-1)}(1-p)^{2(kn-\frac{k^2}{2}-\frac{3k}{2}+1)-k^2} \\
&= \mathbb{E}^2(T_k) \frac{\binom{n-k}{k}}{\binom{n}{k}} (1-p)^{-k^2} \\
&= (1+o(1))\mathbb{E}^2\left(T_k\right)\left(1-\frac{k}{n}\right)^k (1-p)^{-k^2} \\
&= (1+o(1))\mathbb{E}^2\left(T_k\right) e^{-\frac{k^2}{n}+k^2 p + o(1)} \\
&= (1+o(1))\mathbb{E}^2\left(T_k\right).
\end{aligned}$$

Consequently,

$$\frac{\mathbb{Var}(T_k)}{\mathbb{E}^2(T_k)} = \frac{1}{\mathbb{E}(T_k)} + o(1).$$

Since $\mathbb{E}(T_k) \to \infty$ as n tends to infinity, and by Chebyshev inequality, it follows that $\mathbb{P}[T_k = 0] \leq \mathbb{Var}(T_k)/\mathbb{E}^2(T_k) = o(1)$, which completes the proof. $\square$

In the next theorem, convergence in distribution of T_k is established for certain values of k. The proof is based on the following special case of the method of moments (see [Bol01], p. 25). Let $\lambda = \lambda(n)$ be a non-negative bounded function on $\mathbb{N}$. Let $X_1, \ldots, X_n$ be non-negative integer-valued random variables such that, for every $r = 1, 2, \ldots, \mathbb{E}(X)_r \to \lambda^r$ as $n \to \infty$. Then $X_n \xrightarrow{d} \mathcal{P}(\lambda)$.

Theorem 4.5. *Let $p = \theta/n$, where $0 < \theta < 1$ is fixed. Let*

$$k = \frac{1}{\theta - \log\theta}\Big[\log n - 2\log\log n - l\Big] \in \mathbb{N}, \qquad l = l(n) = O(1).$$

Denote by T_k the number of components of $G(n,p,k)$ that are tropical trees of order k. Then T_k has asymptotically Poisson distribution $\mathcal{P}(\lambda)$ with mean

$$\lambda = \frac{(\theta - \log\theta)^2 e^l}{\theta}.$$

Proof: Note first that by a judicious choice of l, k is an integer. From the proof of Theorem 4.4, we have

$$\mathbb{E}(T_k) \sim \frac{e^{\log n - k(\theta - \log\theta)}}{\theta k^2}.$$

It is easily checked that $\mathbb{E}(T_k)$ is asymptotically equivalent to λ. Since by assumption $l = O(1)$ and θ is fixed, λ is bounded. For every integer $r \geq 2$, the r-th factorial moment of T_k is estimated as follows.

$$\begin{aligned}
\mathbb{E}(T_k)_r &= \binom{n}{k}\binom{n-k}{k}\cdots\binom{n-(r-1)k}{k} \\
&\qquad \times \left[\frac{k!}{k^k}\right]^r \left(k^{k-2}\right)^r p^{r(k-1)} (1-p)^{rk(n-rk)+\binom{rk}{2}-r(k-1)} \\
&= (1+o(1))\,\frac{n^{rk}}{k^{2r}} p^{r(k-1)} (1-p)^{rk(n-rk)+\binom{rk}{2}-r(k-1)} \\
&= (1+o(1))\left[\frac{n^k}{k^2} p^{k-1} (1-p)^{k(n-k)}\right]^r (1-p)^{-r^2k^2+rk^2+\binom{rk}{2}-r(k-1)} \\
&= (1+o(1))\left[\mathbb{E}(T_k)\right]^r.
\end{aligned}$$

The result follows from the method of moments previously mentioned. □

References

[ALN11] S. Akbari, V. Liaghat, and A. Nikzad. Colorful paths in vertex-colorings of graphs. *Electronic Journal of Combinatorics*, 18:P17, 2011.

[AMK+] J.-A. Anglès d'Auriac, A. El Maftouhi, M. Karpinski, Y. Manoussakis, L. Montero, N. Narayanan, L. Rosaz, and J. Thapper. Tropical dominating sets in vertex-colored graphs. Unpublished results, submitted.

[BE76] B. Bollobás and P. Erdős. Cliques in random graphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 80, 1976.

[BHK+12] S. Bruckner, F. Hüffner, C. Komusiewicz, R. Niedermeier, S. Thiel, and J. Uhlmann. Partitioning into colorful components by minimum edge deletions. In *Combinatorial Pattern Matching*, pages 56–69, 2012.

[BHKN13] S. Bruckner, F. Hüffner, C. Komusiewicz, and R. Niedermeier. Evaluation of ilp-based approaches for partitioning into colorful components. In *Software Engineering and Applications*, pages 176–187, 2013.

[Bol81] B. Bollobás. Threshold functions for small subgraphs. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 90, 1981.

[Bol01] B. Bollobás. *Random Graphs - Second Edition*. Cambridge University Press, 2001.

[CPM10] Eduardo Corel, Florian Pitschi, and Burkhard Morgenstern. A min-cut algorithm for the consistency problem in multiple sequence alignment. *Bioinformatics*, 26:1015–1021, 2010.

[ER60] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[FFHV11] M. Fellows, G. Fertin, D. Hermelin, and S. Vialette. Upper and lower bounds for finding connected motifs in vertex-colored graphs. *J. Comput. Syst. Sci.*, 77(4):799–811, 2011.

[FHH+] F. Foucaud, A. Harutyunyan, P. Hell, S. Legay, Y. Manoussakis, and R. Naserasr. Tropical homomorphisms in vertex-coloured graphs. Unpublished results.

[JLR00] S. Janson, T. Luczak, and A. Ruciński. *Random Graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, 2000.

[Li01] A. Li. A generalization of the gallai-roy theorem. *Graphs andCombinatorics*, 17:681–685, 2001.

[Lin07] C. Lin. Simple proofs of results on paths representing all colors in proper vertex-colorings. *Graphs andCombinatorics*, 23:201–203, 2007.

[Mat76] D. W. Matula. The largest clique size in a random graph. *Technical Report, Dept. of Computer Science, Southern Methodist University Dallas*, 1976.

[PA] A. Popa and A. Adamaszek. Algorithmic and hardness results for the colorful components problems. In press, to appear in LATIN 2014.

[ZSLS11] C. Zheng, K. Swenson, E. Lyons, and D. Sankoff. Omg! orthologs in multiple genomes - competing graph-theoretical formulations. In *Workshop on Algorithms in Bioinformatics*, pages 364–375, 2011.

Email addresses: jagw40k@free.fr (J.-A. Anglès d'Auriac), nathann.cohen@gmail.com (N. Cohen), hakim.maftouhi@orange.fr (A. El Maftouhi), ararat.harutyunyan@math.univ-toulouse.fr (A. Harutyunyan), legay@lri.fr (S. Legay), yannis@lri.fr (Y. Manoussakis)