REMARKS ON A RAMSEY THEORY FOR TREES

JÁNOS PACH[*], JÓZSEF SOLYMOSI[†], GÁBOR TARDOS[‡]

Dedicated to Endre Szemerédi on the occasion of his 70th birthday.

Extending Furstenberg's ergodic theoretic proof for Szemerédi's theorem on arithmetic progressions, Furstenberg and Weiss (2003) proved the following qualitative result. For every d and k, there exists an integer N such that no matter how we color the vertices of a complete binary tree T_N of depth N with k colors, we can find a monochromatic replica of T_d in T_N such that (1) all vertices at the same level in T_d are mapped into vertices at the same level in T_N; (2) if a vertex $x \in V(T_d)$ is mapped into a vertex y in T_N, then the two children of x are mapped into descendants of the two children of y in T_N, respectively; and (3) the levels occupied by this replica form an arithmetic progression in $\{0,1,\ldots,N-1\}$. This result and its density versions imply van der Waerden's and Szemerédi's theorems, and laid the foundations of a new Ramsey theory for trees.

Using simple counting arguments and a randomized coloring algorithm called *random split*, we prove the following related result. Let $N = N(d,k)$ denote the smallest positive integer such that no matter how we color the vertices of a complete binary tree T_N of depth N with k colors, we can find a monochromatic replica of T_d in T_N which satisfies properties (1) and (2) above. Then we have $N(d,k) = \Theta(dk \log k)$. We also prove a density version of this result, which, combined with Szemerédi's theorem, provides a very short combinatorial proof of a quantitative version of the Furstenberg-Weiss theorem.

1. Introduction

Van der Waerden's celebrated theorem [12] states that for any positive integers d and k, there exists an integer $M = M(d,k)$ such that no matter how we color the elements of the set $\{1,2,\ldots,M\}$ with k colors, at least one of the color classes contains an arithmetic progression of length d.

Erdős and Turán [4] conjectured in 1936 and Szemerédi [10] proved in 1974 that this statement can be generalized as follows. For any positive integer d and real $\delta > 0$, there exists an integer $m = m(d,\delta)$ such that every subset of the set $\{1,2,\ldots,m\}$ of size at least δm contains an arithmetic progression of length d. Clearly, in van der Waerden's theorem, $M(d,k)$ can be chosen to be $m(d,1/k)$.

A second proof of Szemerédi's theorem was given by Furstenberg [5], using ergodic theory. Although qualitative in nature, this proof also has a quantitative version [11]. Furstenberg's proof represented a breakthrough, partly because of its flexibility. It led to a number of generalizations of Szemerédi's theorem that do not seem to follow by the original approach. These include the density Hales-Jewett theorem [6] and the polynomial Szemerédi theorem [2], [3].

In 2003, Furstenberg and Weiss [7] extended Furstenberg's proof to recurrence properties for Markov processes, which resulted in a series of new Ramsey-type theorems for trees. To formulate their results, we need to introduce some definitions.

For any positive integer d, let T_d denote the full binary tree of depth $d-1$. We will use the terms of root, leaf, child, descendant, and level in their usual meaning. In the standard implementation, for any $d > 0$, the vertex set $V(T_d)$ of T_d consists of the strings of length smaller than d over the binary alphabet $\{0,1\}$. The *level* of a vertex is the length of the string. The *root*, the only vertex at level 0, is the empty string. The *leaves* are the vertices at level $d-1$. The *children* of a non-leaf vertex x are $x0$ and $x1$. Finally, x is a *descendant* of y if y is an initial segment of x. Any vertex is considered a descendant of itself. The empty tree will be denoted by T_0.

We call a function $f\colon V(T_d) \to V(T_n)$ a *regular embedding* of T_d in T_n if the following two conditions are satisfied.

1. If y and z are the two children of x in T_d, then $f(y)$ and $f(z)$ are descendants of distinct children of $f(x)$ in T_n.
2. If x and y are vertices at the same level of T_d, then $f(x)$ and $f(y)$ are also at the same level in T_n.

For any subset $H \subseteq V(T_n)$, we say that H *contains a replica* of T_d if there is a regular embedding $f\colon V(T_d) \to H$. If, in addition, there exist suitable integers a and b such that every vertex at level i in T_d is mapped into a

vertex at level $a + ib$ in T_n, then H is said to contain an *arithmetic replica* of T_d. See Figure 1 for an example of a non-arithmetic replica of T_3 in T_5.
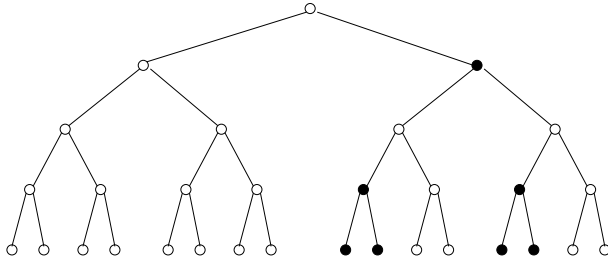
Figure 1. A non-arithmetic replica of T_3 in T_5

Furstenberg and Weiss [7] established the following theorem and various density versions of it.

Theorem A. [7] *For any positive integers d and k, there exists $N = N(d, k)$ such that for every coloring of the vertices of T_N with k colors, at least one of the color classes contains an arithmetic replica of T_d.*

Restricting this result to colorings of T_N, in which all vertices at the same level receive the same color, we obtain van der Waerden's theorem.

More than half of the vertices of T_N are leaves, yet the set of leaves of T_N contains no replica of T_2. Therefore, to formulate an analogue of Szemerédi's theorem for trees, we have to measure the "density" of a subset $H \subset V(T_N)$ differently.

Furstenberg and Weiss defined the *weight $w(x)$ of a vertex* $x \in V(T_n)$ to be $2^{-l(x)}$, where $l(x)$ denotes the level of x in T_n. The *weight of a set* $H \subseteq V(T_n)$ is

$$w(H) = \sum_{x \in H} w(x).$$

In other words, $w(H)$ is the expected size of the intersection of H with a uniformly selected random branch of T_n.

Theorem B. [7] *For any positive integer d and real $\delta > 0$, there exists $n = n(d, \delta)$ such that every subset of the vertex set of T_n with weight at least δn contains an arithmetic replica of T_d.*

Obviously, Theorem B generalizes both Theorem A and Szemerédi's theorem on arithmetic progressions.

The aim of this note is to offer a simple alternative approach to Theorems A and B. Using elementary combinatorial arguments and a randomized coloring algorithm, called *random split*, we prove the following results.

Theorem 1. *Let d, n be positive integers, and let H be a subset of the vertex set of T_n satisfying*

$$2^{w(H)} > \sum_{i=0}^{d-1} \binom{n}{i}.$$

Then H contains a replica of T_d.

Theorem 2. *Let $k, d, n \geq 2$ be integers.*

(i) *Suppose that $n > 5dk \log k$, where $\log$ denotes the logarithm of base 2. Then, for any coloring of the vertices of T_n with k colors, one can find in T_n a monochromatic replica of T_d.*

(ii) *If $n \leq (d-1)k \log(k/6)$, then there exists a coloring of T_n with k colors such that T_n contains no monochromatic replica of T_d.*

The first statement of Theorem 2 directly follows from Theorem 1. Indeed, for any k-coloring of $V(T_n)$, the weight of at least one of the color classes is at least $w(V(T_n))/k = n/k$. Thus, this color class contains a monochromatic replica of T_d, whenever we have

$$2^{n/k} > \sum_{i=0}^{d-1} \binom{n}{i}.$$

It follows by straightforward computation that this inequality holds for $n > 5dk \log k$.

By its nature, the original ergodic proof of the Furstenberg-Weiss theorem is purely existential. We finish this section by showing that Theorem 1 implies a *quantitative* version of Theorem B with

$$n(d, \delta) < 2^{2^{2^{2^{d+9}}}}.$$

Proof of Theorem B. Let $H \subset V(T_n)$ be a set of weight at least δn, and let l be a positive integer. We are going to prove that H contains an arithmetic replica of T_l, provided that n is sufficiently large.

Let $d = \varepsilon n$, for some $\varepsilon > 0$ to be specified later. It follows from Chernoff's bound [9] that

$$\sum_{i=0}^{d-1} \binom{n}{i} < 2^{h(\varepsilon)n},$$

where $h(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon) \log(1-\varepsilon)$ stands for the binary entropy of ε. Therefore, as long as $h(\varepsilon) \leq \delta$, we have

$$2^{w(H)} \geq 2^{\delta n} \geq 2^{h(\varepsilon)n} > \sum_{i=0}^{d-1} \binom{n}{i},$$

and the condition in Theorem 1 is satisfied. Setting $\varepsilon = h^{-1}(\delta)$, we obtain that T_n contains a replica of T_d with $d = \lfloor \varepsilon n \rfloor = \lfloor h^{-1}(\delta) n \rfloor$. Let $H_d \subset H$ denote the set of elements of such a replica.

The set of levels occupied by the elements of H_d in T_n is a subset of $\{0, 1, \ldots, n-1\}$ with density roughly $h^{-1}(\delta) > 0$. Thus, it follows from Szemerédi's theorem that this set contains an arithmetic progression of length l, provided that n is sufficiently large. This implies that there is a regular embedding $f \colon V(T_l) \to V(H_d)$ such that the levels of $f(V(T_l))$ in T_n form an arithmetic progression. In other words, T_n contains an arithmetic replica of T_l, as desired. If we plug in the best known quantitative version of Szemerédi's theorem, due to Gowers [8], we obtain the desired bound. ∎

2. Proof of Theorem 1

Let us start with a couple of definitions.

The *signature* of a regular embedding $f \colon V(T_d) \to V(T_n)$ is defined as the set of levels in T_n occupied by the images of the vertices $v \in V(T_d)$. Since all vertices at the same level of T_d are mapped by f into vertices at the same level of T_n, and vertices at different levels in $V(T_d)$ are mapped into vertices at different levels, we obtain that the signature of f is a d-element subset of $\{0, 1, \ldots, n-1\}$.

For a given subset $H \subset T_n$, We write $S(H)$ for the set of signatures of all regular embeddings of T_d in H, with $d \geq 0$. We have that $\emptyset \in S(H)$, which corresponds to the degenerate case when $d = 0$, and T_0 has no vertices.

Lemma 3. *Let $H \subseteq V(T_n)$. We have $|S(H)| \geq 2^{w(H)}$.*

Proof. The proof is by induction on n. If $n = 0$, we have $H = \emptyset$, $w(H) = 0$, and $S(H) = \{\emptyset\}$, so the statement is true.

Suppose now that $n \geq 1$ and that we have proved Lemma 3 for $n-1$. Let r denote the root of T_n, and let T' and T'' be the two subtrees isomorphic to T_{n-1} that $T_n - r$ splits into. We apply the lemma to these subtrees and to the sets $H' = H \cap V(T')$ and $H'' = H \cap V(T'')$. By the induction hypothesis, we have $|S(H')| \geq 2^{2w(H')}$. Note that the weight of H' inside the tree T' is $2w(H')$, because the levels are shifted by one. However, this shift does not affect the *size* of the set of signatures. Analogously, we have $|S(H'')| \geq 2^{2w(H'')}$.

We distinguish two cases. If $r \notin H$, then $w(H) = w(H') + w(H'')$ and $S(H) = S(H') \cup S(H'')$. The inequality claimed in the lemma follows:

$$|S(H)| \geq \max(|S(H')|, |S(H'')|) \geq 2^{\max(2w(H'), 2w(H''))} \geq 2^{w(H') + w(H'')}.$$

On the other hand, if $r \in H$, then we have $w(H) = w(H') + w(H'') + 1$, as $w(r) = 1$. In view of the fact that $S(H)$ is the disjoint union of the sets

$S(H') \cup S(H'')$ and $\{s \cup \{0\} \,|\, s \in S(H') \cap S(H'')\}$, we obtain that in this case $|S(H)| = |S(H')| + |S(H'')|$. Using the convexity of the function 2^w, we can conclude that in this case

$$|S(H)| = |S(H')| + |S(H'')| \geq 2^{2w(H')} + 2^{2w(H'')} \geq 2 \cdot 2^{w(H') + w(H'')} = 2^{w(H)},$$

as desired.   ∎

Proof of Theorem 1. By Lemma 3, the number of signatures determined by H is at least $2^{w(H)}$. By the assumption of Theorem 1, this quantity is larger than the number of signatures of size smaller than d. Therefore, $S(H)$ has an element of size at least d. In other words, there exists a regular embedding of T_d in H.   ∎

3. Random split and fit – Proof of Theorem 2(ii)

To prove the existence of a coloring which meets the requirements in Theorem 2(ii), we use a random coloring algorithm which will be called *random split*.

   We color the vertices of T_n by the positive integers in order of increasing level (breadth first). While performing the coloring procedure, we maintain a list of "forbidden colors" for each vertex of T_n not yet colored. These lists are empty at the beginning of the procedure. When we reach a vertex x, we assign to x the smallest permitted color: the smallest positive integer c that does not appear on its list of forbidden colors. If x is not a leaf, we update the lists associated to its descendants as follows. Let y and z be the two children of x, and let D_y and D_z denote their sets of descendants. For each level l larger than the level of x, we make an independent uniform random choice and either add c to the list of forbidden colors of every element of D_y on level l or we add c to the list of every vertex in D_z on level l.

Lemma 4. *The random split coloring of T_n admits no monochromatic regular embedding (replica) of T_2.*

Proof. Consider any regular embedding f of T_2 in T_n. Denote by x the image of the root of T_2 and let y and z be the images of the leaves. By definition, y and z are on the same level l and they are descendants of distinct children of x. In the random split coloring, x receives some color c and at the same time the c is added to the list of forbidden colors to all descendants of one of its children on level l. In particular, c will be forbidden either for y or for z. Thus, f cannot be monochromatic with respect to this coloring.   ∎

Lemma 5. *Restricted to any one root-leaf branch of T_n, the random split coloring is equivalent to the following "random fit" procedure: We color the vertices one by one, starting at the root. For each vertex, we consider the positive integers in increasing order until one is accepted and give the vertex the accepted color. When considering the integer c, we accept it with probability 2^{-m}, where m is the number of vertices (along this branch) that have earlier been colored with the color c. In particular, we do accept c the first time it is considered.*

Proof. Restricting our attention to a single branch simplifies the procedure of updating the lists of forbidden colors in random split: once a vertex is colored, its color is added to the list of each uncolored vertex independently with probability $1/2$. Equivalently, in the random fit procedure, if a color c appears m times along the branch up to a certain point, then for every remaining vertex y, the color c appears on the list of y with probability $1 - 2^{-m}$. These events are independent for different pairs (c, y), so deciding them can be postponed until the particular vertex y is colored, as done in random fit. ∎

The key to the proof of Theorem 2(ii) is the following statement.

Lemma 6. *Let $n \geq 8$ and $k = 2\lfloor 3n/\log n \rfloor$. For any branch of T_n of length n, the probability that the random fit algorithm uses a color higher than k is smaller than 2^{1-n}.*

Before proving Theorem 2(ii) in its full generality, we show that Lemma 6 implies the result for $d = 2$. Indeed, let us apply random split coloring to T_n. This coloring admits no regular embedding of T_2, by Lemma 4. By Lemmas 5 and 6, the probability that it uses a color higher than k on any given branch is less than 2^{1-n}. So, by the union bound, it uses at most k colors on the entire tree T_n, with positive probability.

Proof of Lemma 6. Fix a branch of length n of T_n, and consider one by one the sequence of all choices made by the random fit algorithm. The maximum number of choices is $N := \binom{n+1}{2}$, and after each choice we either accept or reject a color. Let p_j denote the probability with which we accept the color at the j'th choice.

Set $X_0 = 0$, and for any $j > 0$, define the random variable X_j as follows. Let $X_j = X_{j-1} + p_j$ if random fit rejects the corresponding color considered at the j'th choice, and let $X_j = X_{j-1} + p_j - 1$ if it accepts. If random fit makes fewer than j individual choices, we simply set $X_j = X_{j-1}$. Obviously, the random variables X_j ($j = 1, 2, 3, ...$) define a martingale with differences bounded by 1, and X_j stabilizes for $j \geq N$.

There are exactly n choices at which a color is accepted, and the corresponding -1 terms contribute $-n$ to X_N. If a color larger than k was ever

used, then every color up to $k+1$ must have been used at least once. For simplicity, we set $l=k/2+1$ and use the fact that each color $i\leq l$ must have been considered at least l times, and every time it was considered, it gave a positive contribution to X_n of at least 2^{-m_i}, where m_i is the total number of vertices along this branch that were assigned color i. Thus, we have

$$X_n \geq \sum_{i=1}^{l} \frac{l}{2^{m_i}} - n \geq \frac{l^2}{n^{1/3}} - n \geq \frac{9n^{5/3}}{\log^2 n} - n,$$

where the middle inequality comes from the fact the $\sum_i m_i \leq n$ and 2^{-m} is a convex function.

Azuma's inequality [1] bounds the probability that $X_N = X_N - X_0 > T$ by $e^{-\frac{T^2}{2N}}$. Substituting $T = \frac{9n^{5/3}}{\log^2 n} - n$, we obtain the desired bound for the probability that a color larger than k is assigned to some vertex. ∎

Proof of Theorem 2(ii). We have already seen that for $d=2$ the statement directly follows from Lemma 6. This means that there is a k-coloring χ' of $T_{n'}$ with $n' = \lfloor k\log(k/6)\rfloor$, which does not admit a monochromatic regular embedding of T_2.

To tackle the case $d>2$, let $n=(d-1)n'$ and split T_n into subtrees isomorphic to $T_{n'}$, in the usual way: the levels 0 to $n'-1$ form one subtree, the levels n' to $2n'-1$ form $2^{n'}$ subtrees, etc. Coloring each of these subtrees separately according to χ', we obtain a coloring that admits no monochromatic regular embedding of T_d. ∎

4. Concluding remarks

Furstenberg and Weiss generalized Theorem B in two directions. First of all, instead of binary trees, one can consider ternary trees or, in general, trees in which every non-leaf vertex has s children, for some $s\geq 2$.

Obviously, our approach also applies to this case. Let $T_{n,s}$ denote a full tree of depth n with this property. The only difference in our argument is that now the *weight* of a vertex $x \in V(T_{n,s})$ at level l has to be defined as $w(x) = s^{-l}$. The weight of a subset of $V(T_{n,s})$ is the sum of the weights of its elements. We can define the regular embeddings of $T_{d,s}$ in $T_{n,s}$ as in the binary case, but now the s children of a vertex $v \in V(T_{d,s})$ have to be mapped to descendants of distinct children of the image of v. Instead of Lemma 3, now we have

Lemma 3'. *Let* $H \subseteq V(T_{n,s})$. *Then the number of signatures of all regular embeddings of* $T_{d,s}$ *in* H *satisfies*

$$|S(H)| \geq \sum_{\sigma \in S(H)} (s-1)^{-|\sigma|} \geq \left(\frac{s}{s-1} \right)^{w(H)}.$$

Lemma 3' readily implies the following version of Theorem 1:

Theorem 1'. *Let* d, n, *and* s *be positive integers, and let* H *be a subset of the vertex set of* $T_{n,s}$ *satisfying*

$$\left(\frac{s}{s-1} \right)^{w(H)} > \sum_{i=0}^{d-1} \frac{\binom{n}{i}}{(s-1)^i}.$$

Then H *contains a replica of* $T_{d,s}$.

Using Theorem 1' and Szemerédi's theorem, one can easily deduce the corresponding version of Theorem B: Any subset H of the vertex set of $T_{n,s}$ with weight $w(H) \geq \delta n$ contains an arithmetic replica of $T_{d,s}$ provided $n > n_0(d,s,\delta)$.

As another variant of their result, Furstenberg and Weiss considered arithmetic embeddings in not necessarily full trees. Nevertheless, in what follows, we use the same definition of regular embeddings of T_d in T as for embeddings in T_n.

Theorem B'. [7] *Let* α *be a positive real, let* $s > 1$ *be an integer, and* T *a rooted tree satisfying the following three conditions:*

(a) *every vertex has at most* s *children;*
(b) *every leaf is at level* n; *and*
(c) *the number of leaves is at least* $2^{\alpha n}$.

Then there is an arithmetic replica of T_d *in* T *provided* $n > n'_0(s,d,\alpha)$.

Proof. Let us define the map $g \colon V(T_n) \to V(T)$ as follows. For the root r of T_n, let $g(r)$ be the root of T. For any non-leaf vertex $v \in V(T_d)$, let g map the two children of v to the two distinct children of $g(v)$ which have the largest number of descendants that are leaves, unless $g(v)$ has only one child. In the latter case, we map both children of v to the only child of $g(v)$. Let

$$H = \{v \in V(T_d) \mid g(v) \text{ has 0 or at least 2 children}\}.$$

Since g preserves levels, it maps every replica of T_d in H into a replica of T_d in T, and every arithmetic replica is mapped into an arithmetic replica.

Note that T has at most $s^{w(H)-1}$ leaves. Thus, by our assumption, $w(H) > \alpha n / \log s$. By Theorem B, if $n > n'_0(s,d,\alpha)$, then H contains an arithmetic replica X of T_d. Consequently, $g(X)$ is an arithmetic replica of T_d in T. ∎

References

[1] N. ALON and J. SPENCER: *The Probabilistic Method, 3rd ed.*, Wiley, New York, 2008.

[2] V. BERGELSON and A. LEIBMAN: Polynomial extensions of van der Waerden's and Szemerédi's theorems, *J. Amer. Math. Soc.* **9** (1996), 725–753.

[3] V. BERGELSON and A. LEIBMAN: Set polynomials and a polynomial extension of the Hales-Jewett theorem, *Ann. of Math. (2)* **150** (1999), 33–75.

[4] P. ERDŐS, P. TURÁN: On some sequences of integers, *J. London Math. Soc.* **11** (1936), 261–264.

[5] H. FURSTENBERG: Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions, *J. Analyse Math.* **31** (1977), 204–256.

[6] H. FURSTENBERG and Y. KATZNELSON: A density version of the Hales-Jewett theorem, *J. d'Analyse Math.* **57** (1991), 64–119.

[7] H. FURSTENBERG and B. WEISS: Markov processes and Ramsey theory for trees, *Combin. Probab. Comp.* **12** (2003), 547–563.

[8] W. T. GOWERS: A new proof of Szemerédi's theorem, *Geometric And Functional Analysis* **11** (2001), 465–588.

[9] A. RÉNYI: *Foundations of Probability*, Holden-Day, San Francisco, 1970.

[10] E. SZEMERÉDI: On sets of integers containing no k elements in arithmetic progression, *Acta Arith.* **27** (1975), 299–345.

[11] T. TAO: A quantitative ergodic theory proof of Szemerédi's theorem, *Electronic J. Comb.* **13** (2006), # R99.

[12] B. L. VAN DER WAERDEN: Beweis einer Baudetschen Vermutung, *Nieuw. Arch. Wisk.* **15** (1927), 212–216.

János Pach

EPFL, Lausanne
Switzerland
       and
Rényi Institute
Budapest, Hungary
pach@cims.nyu.edu

József Solymosi

University of British Columbia
Vancouver
Canada
solymosi@math.ubc.ca

Gábor Tardos

Simon Fraser University
Vancouver
Canada
       and
Rényi Institute
Budapest, Hungary
tardos@renyi.hu