REPETITIONS IN GRAPHS AND SEQUENCES

MARIA AXENOVICH

Abstract. The existence of unavoidable repeated substructures is a known phenomenon implied by the pigeonhole principle and its generalizations. A fundamental problem is to determine the largest size of a repeated substructure in any combinatorial structure from a given class. The strongest notion of repetition is a pair of isomorphic substructures, such as a pair of vertex-disjoint or edge-disjoint isomorphic subgraphs or a pair of disjoint identical subsequences of a sequence. A weaker notion of repetition is a pair of substructures that have the same value on a certain set of parameters. This includes vertex-disjoint induced subgraphs of the same order and size, disjoint vertex sets with the same multiset of pairwise distances, subgraphs with the same maximum degree. This paper surveys results on unavoidable repetitions, also referred to as twins, with a focus on three asymptotically tight results obtained over the past 5 years.

Keywords: sequence, subword, identical subwords, twins in sequences, twins in graphs, isomorphic subgraphs, repeated sequences, self-similarity, divisibility of graphs, twins in hypergraphs.

1. Introduction

There are many repetitions that one can observe in nature: identical twins, two leaves on a tree that look alike, repetitive motifs in a bird's song, segments of the DNA that are identical, etc.

Discrete mathematical structures possess repetitions as well, as is implied by the pigeonhole principle and its generalizations. A fundamental problem is to determine the largest size of a repeated substructure in any combinatorial structure from a given class. Here, we shall often refer to repeated structures as twins even if these twins are not identical.

The strongest notion of repetition deals with "identical" twins, i.e. with a pair of isomorphic substructures. A weaker notion of repetition is a pair of substructures with the same value on a certain set of parameters.

This survey focuses on three types of "identical" twins, which have corresponding size functions $f_v(n), f_e(n), f(n)$ and a few types of weaker twins, which have size functions such as $t(n), h(n)$.

Date: July 6, 2015.

Let $\mathcal{G}_n$ be the class of all n-vertex graphs, and $\mathcal{G}^m$ be the class of all graphs with m edges, let k be an integer. For all standard graph theoretic notions we refer the reader to the book of West, [57]. All graphs here are simple graphs with no repeated edges and no loops. We define the size functions:

$$
\begin{aligned}
f_v(n) &= \max\{\ k: && \text{any } G \in \mathcal{G}_n \text{ has two isomorphic vertex-disjoint induced subgraphs on} \\
& && k \text{ vertices each}\},\\
f_e(m) &= \max\{\ k: && \text{any } G \in \mathcal{G}^m \text{ has two isomorphic edge-disjoint subgraphs} \\
& && \text{on } k \text{ edges each}\},\\
f(n) &= \max\{\ k: && \text{any binary sequence with } n \text{ elements contains} \\
& && \text{two disjoint identical subsequences of } k \text{ elements each}\},\\
t(n) &= \max\{\ k: && \text{any } G \in \mathcal{G}_n \text{ has two vertex-disjoint induced subgraphs} \\
& && \text{on } k \text{ vertices and with the same number of edges}\},\\
h(n) &= \max\{\ k: && \text{any } G \in \mathcal{G}_n \text{ has two disjoint subsets of vertices} \\
& && \text{whose multisets of pairwise distances are identical}\}.
\end{aligned}
$$

For all of these functions, except for the last one, we now know exact asymptotic behavior. The following theorem is an easy consequence of Ramsey theorem [52, 24] and a property of random graphs, see Section 2.

Theorem 1.1.

$$f_v(n) = \Theta(\log n).$$

The next three theorems proved in 2012, 2014, and 2012 respectively involve more sophisticated proof techniques such as random methods on graphs, regular partitions of sequences, and balanced partitions of integers.

Theorem 1.2 (Lee, Loh, and Sudakov [44])**.** *There are absolute constants c and C for which*

$$c(m \log m)^{2/3} \leq f_e(m) \leq C(m \log m)^{2/3}.$$

Theorem 1.3 (Axenovich, Person, and Puzynina [8])**.** *There exists an absolute constant C such that*

$$\left(1 - C\left(\frac{\log n}{\log \log n}\right)^{-1/4}\right)\frac{n}{2} \leq f(n) \leq \frac{n}{2} - \frac{1}{2}\log n.$$

Theorem 1.4 (Bollobás, Kittipassorn, Narayanan, and Scott [15])**.** *For every $\epsilon > 0$, there is a natural number $N = N(\epsilon)$ such that for all $n > N$*

$$\frac{n}{2} - \epsilon n \leq t(n) \leq \frac{n}{2} - \log\log n.$$

Compared to the above theorems giving tight bounds, it is still not known what is the correct asymptotic behavior of the function $h(n)$. The gap between the upper and the lower bound is large. Here, the slight $\log\log n$ improvement of the upper bound is due to a result from [7].

Theorem 1.5 (Albertson, Pach, Young[1, 7])**.** *There are positive constants c, c', such that for any $n > 3$,*

$$\frac{c\log n}{\log\log n} < h(n) \leq \frac{n}{4} - c'\log\log n.$$

We address functions $f_v(n), f_e(m), f(n), t(n)$, and $h(n)$ in Sections 2, 3, 4, 5, and 6 respectively. We provide some insight into proof techniques, state generalizations and open problems. At the end of the survey, we mention some other weak twin problems.

2. Vertex-disjoint isomorphic subgraphs

Among all twin problems we consider here, the problem of finding the largest order of two vertex-disjoint isomorphic subgraphs is relatively easy and Theorem 1.1 answers it asymptotically. We prove it here for completeness.

Proof of Theorem 1.1. For the lower bound, consider an n-vertex graph G. By Ramsey theorem [52, 24] there is a complete subgraph or an independent set on $c\log n$ vertices. Thus G has two vertex disjoint induced subgraphs on $\frac{c}{2}\log n$ vertices both isomorphic to either a complete graph or an empty graph.

For the upper bound, we shall consider Erdős-Renyi random graph $G = G(n, 1/2)$ and follow the simple union bound approach of Lee, Loh, and Sudakov [44]. The fact that G has two vertex-disjoint isomorphic twins on t vertices each is equivalent to the existence of a $2t$-vertex subgraph H' that can be partitioned into a graph H and a vertex disjoint isomorphic copy of H. The expected number of such graphs H' is at most $\binom{n}{2t}\binom{2t}{t}t!2^{-\binom{t}{2}}$, where the first binomial coefficient gives the number of ways to choose $2t$ vertices in G, the second counts the number of ways to split $2t$ vertices in two equal parts, $t!$ is the number of ways to permute the vertices in one of the parts, and finally $2^{-\binom{t}{2}}$ is the probability that the second part with ordered vertices is isomorphic to the first part. If $t > 12\log n$, then $\binom{n}{2t}\binom{2t}{t}t!2^{-\binom{t}{2}} \leq (en/2t)^{2t}2^{2t}t^t2^{-t^2/4} \leq n^{2t}2^{-t^2/4} = 2^{t(2\log n - t/4)} \leq 2^{t(2\log n - 3\log n)} < 1$.

Thus there is a graph G with no two vertex disjoint isomorphic subgraphs on t vertices each. Note that this graph has no such subgraphs on $t' > t$ vertices each since otherwise one can delete the corresponding $t'-t$ vertices from each of these subgraphs and obtain isomorphic vertex-disjoint subgraphs on t vertices each. So, any two induced vertex-disjoint subgraphs on at least t vertices in G are not isomorphic. □

Algorithmic problems of finding vertex disjoint isomorphic subgraphs of a given graph and finding a largest induced subgraph common to two given graphs were studied extensively [9, 42, 46]. It was observed that one can map this twin problem to the problem of finding a largest clique in an auxiliary graph. Let G be a graph with $A, B \subseteq V(G)$, $A \cap B = \emptyset$. Let $G' = (V', E')$, where $V' = \{(u,v) : u, v \in V(G),\ u \neq v\}$, the edge set E' consists of pairs $((u,v),(u',v'))$, where u, v, u', v' are pairwise distinct and either $\{u,u'\}, \{v,v'\} \in E(G)$ or $\{u,u'\}, \{v,v'\} \notin E(G)$. Then it is clear that there is an isomorphism ϕ between $G[A]$ and $G[B]$, graphs induced by A and B in G, if and only if $\{(v,\phi(v)) : v \in A\}$ induces a clique in G'. Similarly, a clique in G' gives an isomorphism. Dzido and Krzywdziński [28] studied so-called local similarity of graphs. Specifically, for two parameters k, ℓ, they defined a function $g(k,\ell)$ that is the smallest n such that any ℓ graphs on n vertices have a common induced subgraph on k vertices. They proved that $g(k,3) = R(k,k)$, where $R(k,k)$ is the classical diagonal Ramsey number, and determined $g(3,\ell)$ and $g(4,\ell)$ for all $\ell > 1$.

3. Isomorphic edge-disjoint subgraphs

Let's consider the more complex situation with edge-disjoint twins in graphs. Let $f_e(G)$ be the largest k such that a graph or hypergraph G has two edge disjoint isomorphic subgraphs on k edges. It is clear that $f_e(G) \leq |E(G)|/2$, thus the optimization problem is set over graphs of the same size, m. In a general setting of hypergraphs, let

$$f_e(m,r) = \min\{f_e(G) :\ |E(G)| = m,\ G \text{ is an } r-\text{uniform hypergraph}\}.$$

The notion of $f_e(G)$ was introduced by Jacobson and Schónheim, where they called this parameter the self-similarity of G. This name may seem confusing, since self-similarity is most commonly used for the fractal structure of an object, i.e., for example, we would want a graph to be isomorphic to its subgraph.

Erdős, Pach, and Pyber [29] proved that there are positive absolute constants c and C, such that

$$cm^{2/(2r-1)} \leq f_e(m,r) \leq Cm^{2/(r+1)} \frac{\log m}{\log\log m}.$$

Lee, Loh, and Sudakov [44] improved the upper and the lower bounds to obtain a tight asymptotic result for graphs. Let $f_e(m) = f_e(m,2)$.

Theorem 3.1 ([44]). *There are absolute constants c and C for which*

$$c(m\log m)^{2/3} \leq f_e(m) \leq C(m\log m)^{2/3}.$$

The upper bound is based on a random graph with modified edge-probability. For the lower bound, the authors observe that a graph in which the vertex-disjoint induced twins are small behaves in a sense like a random graph.

When $r=3$, Gould and Rödl [32] proved that $f_e(m,3) \geq c_3 m^{1/2}$ for a constant c_3. Horn, Koubek, and Rödl [39] studied the hypergraph case further, proving the tightness, up to a logarithmic factor, of the upper bound in the result of Erdős, Pach, and Pyber [29] in case when $r=4,5,6$:

Theorem 3.2 ([39]). *There are constants c_4, c_5, c_6, such that $f_e(m,4) \geq c_4 m^{2/5}$, $f_e(m,5) \geq c_5 \frac{m^{1/3}}{\log m}$, and $f_e(m,6) \geq c_6 \frac{m^{2/7}}{\log^{35} m}$. For any $r \geq 7$,*

$$f_e(m,r) \geq \Omega(m^{\frac{2}{2r-\log_2 r}}).$$

Lee, Loh, and Sudakov [44] also generalized this problem to multiple twins. Let $f_e^s(m)$ be the largest k such that any graph on m edges has s pairwise edge-disjoint isomorphic subgraphs. Then, they showed in [44], that

$$f_e^s(m) = \Theta\left(m^{\frac{s}{2s-1}}(\log m)^{\frac{2s-2}{2s-1}}\right).$$

The earlier research was done on *perfect* edge-disjoint isomorphic twins. Here, we say that G has t perfect edge-disjoint isomorphic twins if the edge-set of G is partitioned into t isomorphic subgraphs. Graham, Harary, Robinson, Wallis, and Wormald considered this notion, also called t-divisibility of graphs, or isomorphic factorization of graphs.

Harary, Robinson, Wallis, and Wormald [36, 37], see also Guidotti [34], proved that K_n is t-divisible if t divides the total number of edges, $\binom{n}{2}$, in K_n. Alon, Caro, and Krasikov [3] showed that any m-edge tree could be made 2-divisible by deleting $O(m/\log\log m)$ edges. The notion of 2-divisibility was addressed also under the names *bisectable* and 2-*splittable*, see [35, 55]. There is an extensive research done on isomorphic factorizations, with a lot of papers devoted to factorizations of complete or complete multipartite graphs into specific trees or cycles. See also a book on decompositions by Bosák [16].

4. Repeated Subsequences

An *alphabet* Σ is a finite non-empty set of symbols called *letters*. Let $S = s_1 \dots s_n$ be a sequence $s_1, s_2, \dots, s_n$, where $s_i \in \Sigma$, $i=1,\dots,n$. Sometimes, one refers to S as a word over an alphabet Σ. A *subsequence* of S is a sequence $S' = s_{i_1}s_{i_2}\dots s_{i_s}$, where $i_1 < i_2 < \dots < i_s$. A *factor* of S is a subsequence obtained by taking consecutive elements from S, i.e., $s_is_{i+1}\dots s_j$, for some $1 \leq i \leq j \leq n$.

The study of repeated sequences is largely motivated by problems investigated in combinatorics on words and in the theory of formal languages with special attention given to counting subword occurrences, different complexity questions, and the word reconstruction problem (see, e.g., [27, 49, 54]). We say that two subsequences $s_{i_1}s_{i_2}\dots s_{i_s}$ and $s_{j_1}s_{j_2}\dots s_{j_t}$ of S are disjoint if the index sets are disjoint, i.e., $\{i_1,\dots,i_s\}\cap\{j_1,\dots,j_t\}=\emptyset$. For a sequence S, let $f(S)$ be the largest integer m such that there are two disjoint identical subsequences of S, each of length m. We call such subsequences *twins*. For example, if $S=s_1s_2s_3s_4s_5s_6s_7=0001010$, then $S_1=s_1s_2s_4$ and $S_2=s_3s_5s_6$ are both equal to 001, in addition $S_1'=s_1s_4s_5$ and $S_2'=s_2s_6s_7$ are both equal to 010. So, both pairs, S_1, S_2 and S_1', S_2' are pairs of twins. In this example $f(S)=3$.

The twin problem in sequences is closely related to the classical problem on common subsequences, see for example [25, 38, 58]. Indeed, if we consider two disjoint subsequences S_1 and S_2 of a given sequence S and find subsequences S_1' of S_1 and S_2' of S_2 such that $S_1'=S_2'$, then these subsequences correspond to twins in S. Optimizing over all such S_1 and S_2 gives largest twins. Let Σ^n be the set of sequences of length n over the alphabet Σ, let

$$f(n,\Sigma)=\min\{f(S):S\in\Sigma^n\}.$$

Observe first that $f(n,\{0,1\})\geq\lfloor(1/3)n\rfloor$. Indeed, consider any $S\in\Sigma^n$ and split it into consecutive triples. Each triple has either two zeros or two ones, so we can build a subsequence S_1 by choosing a repeated element from each triple, and similarly build a subsequence S_2 by choosing the second repeated element from each triple. For example, if $S=001\ 101\ 111\ 010$, then there are twins S_1, S_2, each equal to 0 1 1 0. If we mark one twin red and another bold, then $S=$ **0**01 **1**01 **1**11 **0**10.

We shall sometimes refer to two disjoint identical subsequences as *twins*, three disjoint identical subsequences as 3-*twins*, k disjoint identical subsequences as k-*twins*.

4.1. Two twins in a binary alphabet.

The main result of Axenovich, Person, and Puzynina [8] shows that any sufficiently large binary sequence can be split into two identical subsequences almost perfectly. This is one of the main theorems stated in the introduction.

Theorem 4.1 ([8]). *There exists an absolute constant C such that*

$$\left(1-C\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right)n\leq 2f(n,\{0,1\})\leq n-\log n.$$

The proof employs a classical density increment argument successfully applied in combinatorics and number theory, see e.g. the survey of Komlós and Simonovits [43] and some important applications [33] and [56]. One can partition any sequence S into factors that look as if they were random in a certain weak sense (call them

ε-regular). These ε-regular sequences can be partitioned (with the exception of ε proportion of letters) into two identical subsequences. By appending these together for every ε-regular sequence, we eventually obtain identical subsequences of S whose lengths satisfy the claimed inequality.

4.2. **More than two twins or larger alphabets.** For a given sequence S, let $f(S,k)$ be the largest m so that S contains k pairwise disjoint identical subsequences of length m each. Finally, let $f(n,k,\Sigma) = \min\{f(S,k) : \ S \in \Sigma^n\}$ and $f(n,\Sigma) = f(n,2,\Sigma)$.

Axenovich, Person, and Puzynina [8] showed that the situation differs significantly depending on whether the number of twins is larger or smaller than the size of the alphabet. If the size of the alphabet is at most the number of twins, the given sequences can be almost partitioned by these twins. Otherwise, in most cases the largest twins cover at most $(1-c)n$ elements of a given sequence of length n, for some constant $c > 0$, depending on the number of twins.

Theorem 4.2. [8] *For any integer $k \geq 2$, and alphabet Σ, $|\Sigma| \leq k$,*

$$\left(1 - C|\Sigma|\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right) n \leq k f(n,k,\Sigma).$$

Theorem 4.3. [8] *For any integer $k \geq 2$, and alphabet Σ, $|\Sigma| > k$,*

$$\left(\frac{k}{|\Sigma|} - C|\Sigma|\left(\frac{\log n}{\log\log n}\right)^{-1/4}\right) n \leq k f(n,k,\Sigma) \leq n - \max\{\alpha n, \log n\},$$

where $\alpha \in [0, 1/k]$ is the solution of the equation $\ell^{-(k-1)\alpha}\alpha^{-k\alpha}(1-k\alpha)^{k\alpha-1} = 1$, whenever such solution exists and 0 otherwise.

Bukh and Zhou [17] improved the above bounds in case when $k = 2$.

Theorem 4.4. [17] *Let Σ be an alphabet, $|\Sigma| \geq 3$. Then*

$$\max\left\{\frac{2.04}{|\Sigma|}n - o(n),\ 2 \cdot 3^{-4/3}|\Sigma|^{-2/3}n - 3^{-1/3}|\Sigma|^{1/3}\right\} \leq 2 \cdot f(n,2,\Sigma) \leq 2\alpha n,$$

where α is a real number in the interval $[1/|\Sigma|, 1/2]$ satisfying inequality

$$(1-2\alpha)\log\frac{1}{1-2\alpha} - \alpha\log(\alpha^2|\Sigma|) - 2\alpha\log\left(\frac{2}{1+\sqrt{1-1/|\Sigma|}}\right) + (1-2\alpha)\log(1-1/|\Sigma|) < 0.$$

The upper bound implies, in particular, that $2f(n,2,\Sigma) \leq 0.9864\, n$, for $|\Sigma| = 4$; $2f(n,2,\Sigma) \leq 0.96\, n$, for $|\Sigma| = 5$;

$$2f(n,2,\Sigma) \leq \left(\frac{2e}{\sqrt{|\Sigma|}} - (2e^2+1)/|\Sigma| + O(|\Sigma|^{-3/2})\right) n + o(n),$$

for all large $|\Sigma|$.

I.e., for alphabets on at least four letters, there are sequences that could not be almost split into two twins. In comparison, any sufficiently large binary sequence has such almost perfect splitting. It remains unclear what happens for alphabets on 3 letters. For the above theorems it follows that $\frac{2.04}{3}n - o(n) \leq 2 \cdot f(n, 2, \Sigma) \leq n - \log n$.

Open Problem 4.5. *Find $f(n, 2, \Sigma)$ for $|\Sigma| = 3$.*

4.3. **Tools.**

There were two main tools used in the twin problem in sequences so far: a version of the Regularity Lemma for sequences and common subsequences.

The first, introduced in [8], employs the Regularity Lemma variant also developed in that paper. The idea is to split the given sequence into a constant number of factors such that almost all of them are regular, i.e., have about the same density of each letter in sufficiently large subfactors, and then to find twins in each of these factors. To do so, say in case of the alphabet $\{0, 1\}$, delete at most an ϵ-proportion of elements, and spit a factor further into subfactors each containing exactly a zeros and exactly b ones, for some a and b. Build one twin by taking all a zeros from the 1^{st} subfactor, all b ones from the 2^{d} subfactor, all a zeros from the 3^{d}, all b ones from the 4^{th}, etc. The second twin is build by taking all a zeros from the 2^{d} subfactor, all b ones from the 3^{d}, all a zeros from the 4^{th}, etc.

The second tool, used for small cases in [8] and explored much more widely in [17], is based on common subsequences. For two sequences, a largest subsequence contained in both of these sequences is called a largest common subsequence. In order to find twins in a given sequence, one can split it into consecutive factors $S_1, \ldots, S_m$, split each S_i in two subsequences A_i, B_i, find a largest common subsequence C_i of A_i and B_i, and obtain twins each equal to $C_1C_2 \cdots C_m$. A refinement of this natural approach together with Regularity Lemma provided the best lower bounds on f in case of large alphabets, see [17], [18]. Next, we describe these approaches a bit more formally.

4.3.1. *Regularity Lemma for sequences.* First, we shall introduce some notations (for more detail, see for instance [22, 47]). Let $S = s_1 \ldots s_n$. We define a *support* of a subsequence to be its index set, i.e., for a subsequence $S' = s_{i_1}s_{i_2} \ldots s_{i_s}$ of S, $\{i_1, i_2, \ldots, i_s\}$ a support of S' in S. We write $\text{supp}(S')$ to denote support. The *length* of a sequence S is the number of elements in S, denoted $|S|$. For example $|(0010)| = 4$. If $S' = s_{i_1}s_{i_2} \ldots s_{i_s}$, we write $S' = S[I]$, for an index set $I = \{i_1, \ldots, i_s\}$. We use notation $S[i, i+m]$ for $s_is_{i+1} \ldots s_{i+m}$. If S is a sequence over alphabet Σ and $q \in \Sigma$, then we denote $|S|_q$ the number of elements of S equal to q. The *density* $d_q(S)$ is defined to be $|S|_q/|S|$. For two subsequences S' and S'' of S, we say that S' is contained in S'' if $\text{supp}(S') \subseteq \text{supp}(S'')$, we also denote by $S' \cap S''$ a subsequence of S, $S[\text{supp}(S') \cap \text{supp}(S'')]$. If $S = s_1 \ldots s_n$ and $S[1, i] = A$, $S[i+1, n] = B$, for some i, then we write $S = AB$ and call S a concatenation of A

and B. In all of the calculations below we omit floors and ceilings to avoid making the presentation too cluttered.

Next, we define *ε-regular sequence*. For a positive ε, $\varepsilon < 1/3$, call a sequence S of length n over an alphabet Σ *ε-regular* if for every i, $\varepsilon n + 1 \leq i \leq n - 2\varepsilon n + 1$ and every $q \in \Sigma$,

$$|d_q(S) - d_q(S[i, i + \varepsilon n - 1])| < \varepsilon.$$

Note that in the case of the binary alphabet $\Sigma = \{0, 1\}$, $d_0(S) = 1 - d_1(S)$ and thus $|d_0(S) - d_0(S[i, i + \varepsilon n - 1])| < \varepsilon \Longleftrightarrow |d_1(S) - d_1(S[i, i + \varepsilon n - 1])| < \varepsilon$. In this case, we shall denote $d(S) = d_1(S)$. For example, the sequence

$$S = 011101010101000101001100110101010100110101010101111000010100$$

of length $n = 60$ and density $1/2$ is ε-regular for $\varepsilon = 1/5$. This could be verified directly by considering all $60 - 12$ consecutive segments of length $\varepsilon n = (1/5) \cdot 60 = 12$ and comparing their densities with the density $1/2$ of S. Such a 12-letter sequence starting at first position is $S' = 011101010101$, $d(S') = 7/12$, $|7/12 - 1/2| < \epsilon = 1/5$. Such a 12-letter sequence starting at position 8 is $S'' = 101010001010$, $d(S'') = 5/12$, $|5/12 - 1/2| < 1/5$.

The notion of an ε-regular sequence resembles the notion of pseudorandom (quasirandom) sequence, see [23]. However, these two notions are quite different. A sequence that consists of alternating 0s and 1s is ε-regular but not pseudorandom. Also, unlike the case of stronger notions of pseudorandomness, one can check in a linear time, by observing at most n sequences $S[i, i + \varepsilon n - 1]$, whether a sequence is ε-regular, cf. [4] in the graph case.

Definition. We call $\mathcal{S} := (S_1, \ldots, S_t)$ a partition of S if $S = S_1 S_2 \ldots S_t$, (S is concatenation of consecutive S_is). A partition $\mathcal{S}$ is an **ε-regular partition** of a sequence $S \in \Sigma^n$ if

$$\sum_{\substack{i \in [t] \\ S_i \text{ is not } \varepsilon-\text{regular}}} |S_i| \leq \varepsilon n,$$

i.e., the total length of ε-irregular subsequences is at most εn.

Lemma 4.6 (Regularity Lemma for sequences, [8])**.** *For every ε, t_0 and n such that $0 < \varepsilon < 1/3$, $t_0 > 0$ and $n > n_0 = t_0 \varepsilon^{-\varepsilon^{-4}}$, any sequence $S \in \Sigma^n$ admits an ε-regular partition into t parts with $t_0 \leq t \leq T_0 = t_0 3^{1/\varepsilon^4}$.*

Using the Regularity Lemma, one can find large twins in sequences.

Outline of the proof of Theorem 4.1. We claim that if S is an ε-regular sequence, then $2f(S) \geq |S| - 5\varepsilon|S|$. Let $|S| = m$. We partition S into $t = 1/\varepsilon$ consecutive factors $S_1, \ldots, S_{1/\varepsilon}$, each of length εm. Since S is ε-regular, $|d(S_i) - d(S)| < \varepsilon$, for every $i \in \{2, \ldots, 1/\varepsilon - 1\}$. Thus each S_i has at least $(d(S) - \varepsilon)\varepsilon m$ occurrences of

1's and at least $(1-d(S)-\varepsilon)\varepsilon m$ occurrences of 0's. Let $S_i(1)$ be a subsequence of S_i consisting of exactly $(d(S)-\varepsilon)\varepsilon m$ letters 1 and $S_i(0)$ be a subsequence of S_i consisting of exactly $(1-d(S)-\varepsilon)\varepsilon m$ letters 0. When t is even, consider the following two disjoint subsequences of S: $A = S_2(1)S_3(0)S_4(1)\cdots S_{t-2}(1)$ and $B = S_3(1)S_4(0)S_5(1)\cdots S_{t-2}(0)S_{t-1}(1)$. When t is odd, A and B are constructed similarly. We see that A and B together have at least $m - 2\varepsilon^2 m(1/\varepsilon - 3) - 3\varepsilon m$ elements, where $2\varepsilon^2 m(1/\varepsilon - 3)$ is an upper bound on the number of 0's and 1's which we had to "throw away" to obtain *exactly* $(d(S)-\varepsilon)\varepsilon m$ letters 1 and $(1-d(S)-\varepsilon)\varepsilon m$ letters 0 in each S_i, $2\varepsilon m$ is the number of elements in S_1 and S_t, and εm is the upper bound on $|S_2(0)| + |S_{t-1}(1)|$. Thus, $2f(S) \geq m - 5\varepsilon m$. This concludes the proof of the claim. Now, consider an ϵ-regular partition of a given sequence of length n. Ignore the parts that are not ϵ-regular. Note, that there are at most ϵn letters in such parts. In each of the regular parts, find large twins using the claim. Finally, build large twins of the original sequence by concatenating the corresponding twins of each regular part. This gives the lower bound on $f(n)$. For the upper bound, consider a binary sequence $010001110000000001111111111\ldots$, where consecutive blocks have lengths 3^i. One can show that the union of any two twins in such a sequence omits at least one letter from each block. □

4.3.2. *Common Subsequences.* The length of a longest common subsequence of sequences W_1 and W_2 is denoted by $LCS(W_1, W_2)$. For example $LCS(001000, 110010) = 4$ with a common subsequence 1000. Here it is clear that there is no common subsequence on 5 letters because any 5-letter subsequence of the second sequence contains at least two 1's, and the first sequence has only one 1. The problem of finding common subsequences in sequences is of great importance in bioinformatics [58]. There is an extensive literature devoted to computational and algorithmic aspects of the problem, see for example [38]. Kiwi, Loebl, and Matoušek [41] considered random binary sequences. If $L'(n,k)$ is the length of the longest common subsequence of two n-element sequences chosen uniformly and independently at random over an alphabet of size k, then it was shown that

$$\lim_{k\to\infty} \sqrt{k} \lim_{n\to\infty} n^{-1}|\mathrm{Exp}(L'(n,k)| = 2.$$

Roughly speaking, the expected length of a common subsequence of two such random sequences is $\frac{2n}{\sqrt{k}}$, for sufficiently large n and k.

One of the very few extremal results regarding the common subsequences in general is done in case of permutations by Beane and Huynh-Ngoc [10], where it was shown that given three permutations W_1, W_2, W_3, the length of a longest common subsequence $LCS(W_i, W_j) \geq n^{1/3}$ for some $1 \leq i < j \leq 3$. An immediate corollary of this result implies that if W_1, W_2, W_3 are n-letter sequences over some alphabet such that each letter appears exactly the same number of times in each of W_i's, then for some $1 \leq i < j \leq 3$, $LCS(W_i, W_j) \geq n^{1/3}$. Bukh and Ma [18] built on this to provide asymptotically tight results concerning $L(n,k)$, the largest n' such that any two sequences over an alphabet of size k have a common subsequence of length n'. It was shown in [18], that for a fixed k, $L(n,k) = \frac{n}{k}(1+o(1))$.

The idea of Bukh and Zhou [17] was to apply this result, the Regularity Lemma and generalizations to the problem of twins in sequences. In order to prove the second lower bound in Theorem 4.4, an ϵ-regular sequence is split into consecutive factors $F_1,\dots,F_{n/k}$ of length k. In each of the factors F_i, three largest subsequences are chosen such that each letter of the alphabet is repeated the same number of times in each of these three subsequences. This could be done by taking, for a fixed letter, a largest subsequence consisting only of this letter and splitting it between the three equally with perhaps one or two elements leftover. Finally, the above results on common subsequences, applied to the three subsequences, gives twins of size roughly $k^{1/3}$ in F_i for each i. Concatenating the corresponding twins from all F_i's gives long twins from the original sequence.

Bukh and Zhou [17] and Bukh and Ma [18] built on the idea of common subsequences much more and provided many interesting generalizations.

4.4. **Small values of $f(n,k,\Sigma)$.** We will slightly abuse notation and denote by $f(n,k,\ell)$ the value of $f(n,k,\Sigma)$ with $|\Sigma| = \ell$. In the introductory section it was observed that $f(3,2,2) = 1$ yielding immediately a weak lower bound on $f(n,2,2)$ to be $\lfloor n/3 \rfloor$. In general, it is clear that

$$f(n,k,\ell) \geq \left\lfloor \tfrac{n}{m} \right\rfloor f(m,k,\ell).$$

For example, Theorem 4.3 gives a lower bound on $f(n,2,3)$ to be $\frac{1}{3}n - o(n)$, Theorem 4.4 provides the better lower bound of $\frac{1.02}{3}n - o(n)$, and gives another bound $f(n,2,3) \geq n/9 - 1$. There are no good lower bounds without $o(n)$ term, i.e., there is no known (reasonable) number t such that $f(t,2,3) \geq \frac{t}{3}$. If one could find such t this would immediately give another proof of $f(n,2,3) \geq \frac{1}{3}n - t$. However, the smallest value for such a possible t could be 21, which already presents a computationally challenging task. In Table 1 we summarize estimates on the values on $f(n,k,\ell)$, which were determined with the help of a computer. We see that for $n = 21$, $f(n,2,3) \leq 7 = n/3$, for $n = 22$, $f(n,2,3) \leq 7 < n/3$. Thus, the first "open" case which might improve lower bound $n/3$ on $f(n,2,3)$ is $f(22,2,3)$.

$\Sigma \setminus n$	6	7	8	9	10	11	12	13	14	15	16	17
$\{0,1\}$	2	2	2	3	3	4	4	5	5	5	6	6
$\{0,1,2\}$	1	1	2	2	2	3	3	3	4	4	4	4

Further we remark, that a more careful analysis below of the increment argument in the proof of Proposition 4.6 leads to the bound $T_0 \leq t_0 3^{(-2\log\varepsilon)/\varepsilon^3}$, which in

$\Sigma \setminus n$	18	19	20	21	22	23	24
$\{0,1\}$	7	7	8				
$\{0,1,2\}$	≤ 5	≤ 6	≤ 6	≤ 7	≤ 7	≤ 8	≤ 8

Table 1. Values for small t of $f(t,2,2)$ and $f(t,2,3)$.

turn improves the bounds in Theorems 4.1 and 4.2 to

$$\left(1 - C|\Sigma|\left(\frac{(\log\log n)^2}{\log n}\right)^{1/3}\right) n \leq kf(n,k,\Sigma).$$

5. Subgraphs with the same number of edges and vertices

The following twin problem in graphs was introduced by Caro and Yuster in [20]. Let $t(G)$ be the largest k such that there are two induced vertex-disjoint subgraphs of G on k vertices each and having the same number of edges. Note that the function $t(n)$ introduced in the beginning is equal to $\min\{t(G) : |V(G)| = n\}$.

Caro and Yuster [20] provided some bounds first, see Theorem 8. We include the proof of this theorem for completeness. The key idea is to use the chromatic number of the Kneser graph. For positive integers k, n ($k \leq n/2$), the *Kneser graph* $\mathcal{K} = KG(n,k)$ is the graph on the vertex set $\binom{[n]}{k}$ whose edge set consists of pairs of disjoint k-sets. Lovász [48] proved that the chromatic number of the Kneser graph $KG(n,k)$ is $n - 2k + 2$.

Theorem 5.1 ([20]). *There exists a positive constant c such that $\sqrt{n} \leq t(n) \leq n/2 - c\log\log n$.*

Proof. For a graph G with vertex set $[n]$, we define a vertex coloring of $\mathcal{K}$ by letting the number of edges induced by each k-subset of $V(G)$ be the color of the corresponding vertex in $\mathcal{K}$. The number of colors is at most $\binom{k}{2} + 1$. For $k = \sqrt{n}$, we have $\binom{k}{2} + 1 < n - 2k + 2$. Thus the number of colors is less than the chromatic number of $\mathcal{K}$ and there are two adjacent vertices in $\mathcal{K}$ of the same color. Therefore there are two disjoint k-sets of G that induce the same number of edges.

To verify the upper bound, consider a disjoint union of cliques of odd orders $a_1, \ldots, a_m$, where $a_1 = 1$, $a_j > 2(a_1^2 + a_2^2 + \cdots + a_{j-1}^2)$, and a_j is the smallest possible such number; so $m = c\log\log n$. One can show that any pair of twins in this graph either has at most $n - c\log\log n$ vertices together, or omits at least one vertex from each of the cliques. □

Ben-Eliezer and Krivelevich [12] proved that $t(G(n,p)) = \lfloor n/2 \rfloor$ with high probability, where $G(n,p)$ is the Erdős-Renyi random graph. In addition, Caro and Yuster proved that in a sparse graph there are twins of size almost $n/2$:

Theorem 5.2 (Caro and Yuster [20])**.** *For every fixed $\alpha > 0$ and for every $\epsilon > 0$ there exists $N = N(\alpha, \epsilon)$ so that for all $n > N$, if G is a graph on n vertices and at most $n^{2-\alpha}$ edges then $t(G) \geq (1-\epsilon)n/2$.*

Axenovich, Martin, and Ueckerdt [6] made further progress on a twin problem in terms of the number of edges in a graph. The main approach there was to split a large subset of vertices into sets A', B' of equal size, inducing graphs on almost the same number of edges, say with A' inducing smaller number of edges than B'. Then, using carefully chosen ahead of time extra vertices, that induce a matching and send no edges to A' and B', add a matching to A' and an independent set to B'.

Theorem 5.3 ([6])**.** *If G is a graph on n vertices and e edges then*

$$t(G) \geq \frac{n}{2}\left(1 - \frac{20\sqrt{e}\log n}{n}\right).$$

This theorem implies, in particular, that any n-vertex graph on $o\left(n^2/\log^2 n\right)$ edges has twins of size $n/2 - o(n)$ each and that planar graphs have twins of size at least $n/2 - c\sqrt{n}\log n$ each.

We say that G has *perfect twins* if the vertex set can be split in two sets of equal size inducing subgraphs with the same number of edges. Finding large twins is equivalent to finding a small subset of vertices whose deletion allows to find perfect twins in the remaining graph. Some criteria for the existence of perfect twins were provided in [6, 20]. The main observation here is that the graph has perfect twins if and only if the degree sequence could be split into two subsequences of the same length and having the same sum of the elements. We call this a *degree splitting property.* Indeed, sets A and B partitioning the vertex set of a graph G induce perfect twins iff $|A| = |B|$ and $|E(G[A])| = |E(G[B])|$. Since $2|E(A)| = \sum_{a\in A} deg(a) - |E(A,B)|$ and $2|E(B)| = \sum_{b\in B} deg(b) - |E(A,B)|$, where deg is the degree and $E(A,B)$ is the set of edges between A and B, the statement follows. Therefore the following theorem is basically number theoretic as it deals with degree sequence and not with the structure of the graph. Here Δ and δ are maximum and minimum degrees of a graph, respectively, V_i is the set of vertices of degree i.

Theorem 5.4. *Let G be a graph on n vertices, where n is even. If one of the following conditions holds, then $t(G) = n/2$.*

- *The degree sequence of G forms a set of consecutive integers.*
- *$|V_i|$ is even for each i.*
- *$n \geq 90$ and $|\{i : |V_i| \text{ is odd}\}| > n/2$.*
- *There are at least $\Delta(G) - \delta(G)$ disjoint consecutive pairs of vertices.*

In Theorem 5.4, the result of Karolyi [40] on partition of integers was used. We say that two integers are *almost equal* if they differ by $1, 0$, or -1.

Theorem 5.5 ([40]). *Let X be a set of m integers, each between 1 and $2m-2$. If $m \geq 89$, then one can partition X into two sets, X_1 and X_2 of almost equal sizes such that the sum of elements in X_1 is almost equal to the sum of elements in X_2.*

One of the few exact results in this area concerns trees and forests.

Theorem 5.6 ([6]). *If G is a forest then $t(G) \geq \lceil n/2 \rceil - 1$.*

If n is odd this is clearly best possible. For even n this is attained, for example, by a star. Such a graph has no perfect twins.

Finally, Bollobás, Kittipassorn, Narayanan, and Scott [15] proved the strongest result in the area showing that almost perfect twins exist in any sufficiently large graph. This is one of four main theorems we mention in the beginning of the survey.

Theorem 5.7 ([15]). *For every $\epsilon > 0$, there is a natural number $N = N(\epsilon)$ such that for any graph G on $n > N$ vertices, $t(G) \geq n/2 - \epsilon n$.*

The idea of the proof is to delete at most $2\epsilon n$ vertices of a given n-vertex graph such that the rest has perfect twins, or, as observed above, has the degree splitting property. This is done using probabilistic techniques. Specifically, the authors show that there is a probability p at most $\epsilon = \epsilon(G)$ such that removing vertices of G with probability p results in a graph that has degree splitting property with positive probability. In order to do so, many pairs of vertices with degree difference that is controlled are chosen and then the vertices of each pair are assigned to the first or the second part according to the following lemma.

Lemma 5.8. *Let H be a graph on an even number of vertices and $V(H)$ be partitioned into disjoint collections of pairs $\mathcal{P}_1, \dots, \mathcal{P}_k$ such that the degree difference for vertices in each pair in $\mathcal{P}_i$ is between a_i and b_i, $0 \leq a_1 \leq b_1$, $0 < a_i \leq b_i$, $2 \leq i \leq k$. Then H can be partitioned into two sets A and B of the same size such that the sum of the degrees of vertices from A differs from the sum of the degrees of vertices from B by at most b_k. In particular, if $b_k = 1$, then H has perfect twins.*

Bollobás, Kittipassorn, Narayanan, and Scott [15] note that their proof leads to a more precise lower bound $t(n) \geq n/2 - n/(\log\log n)^c$, for a positive constant c. They suspect that $n/2 - t(n) = o(n^\epsilon)$ for each positive ϵ and even state that $n/2 - t(n)$ could be $\Theta(\log n)$.

The twin problem in graphs could be interpreted as a problem in complete graphs edge-colored in 2 colors, say red and blue, where the twins are two vertex-disjoint complete subgraphs with the same number of red edges and the same number of blue edges. The authors of [15] define a natural generalization of the twin problem for multicolored complete graphs with multiple twins. Let G be an edge-colored complete graph on n vertices. Then $t(\ell, G)$ is the largest k such that there are ℓ pairwise disjoint sets of vertices, inducing the same number of edges of color c, for each color c. Finally, let

$$t(n,s,\ell) = \min\{t(\ell,G) : \ G \text{ is an } s-\text{edge-colored complete graph on } n \text{ vertices}\}.$$

So, $t(n,2,2) = t(n)$. The authors conjecture that $f(n,s,2) = n/2 - o(n)$ for any number of colors s.

6. Subgraphs with the same multiset of pairwise distances

In this section we consider one of the "weak" twins. For a vertex set $S \subseteq V(G)$ in a graph G, the *distance multiset*, $D(S)$, is the multiset of pairwise distances between vertices of S in G. Two vertex sets are called *homometric*, or homometric twins, if their distance multisets are identical. For a graph G, the largest integer h, such that there are two disjoint homometric sets of order h in G, is denoted by $h(G)$. This parameter was introduced by Albertson, Pach and Young [1].

Theorem 6.1 ([1]). $\frac{c\log n}{\log\log n} < h(n) \leq \frac{n}{4}$ *for* $n > 3$, *and a constant* c.

It is an easy observation that $h(G) = \lfloor |V(G)|/2 \rfloor$ when G is a path or G is a cycle. However, more is known. Note that the multisets of distances for a vertex subset of a path correspond to a multiset of pairwise differences between elements of a subset of positive integers. We shall say that two subsets of integers are homometric if their multisets of pairwise differences coincide. Piccard [51] claimed to prove that any two homometric sets of positive integers, whose multisets of distances contain no repeated elements are congruent to each other. Later, Bloom [13] found a counterexample to Piccard's claim. Recently, there has been more progress on this problem ([11]). Among others, Rosenblatt and Seymour [53] proved that two multisets A and B of integers are homometric if and only if there are two multisets U, V of integers such that $A = U + V$ and $B = U - V$, where $U + V$ and $U - V$ are multisets, $U+V = \{u+v : \ u \in U, \ v \in V\}$, $U-V = \{u-v : \ u \in U, \ v \in V\}$. Lemke, Skiena and Smith [45] showed that if G is a cycle of length $2n$, then any subset of $V(G)$ with n vertices and its complement are homometric sets. Suprisingly, when the class $\mathcal{G}$ of graphs under consideration is not a path or a cycle, the problem of finding $h(\mathcal{G})$ becomes nontrivial. For a class of graphs of diameter 2, the function h is equal to the function t introduced in the beginning, i.e., the largest integer k such that any graph on n vertices has two vertex-disjoint induced subgraphs on k vertices and with the same number of edges.

The homometric set problem we consider here has its origins in Euclidean geometry, with applications in X-ray crystallography introduced in the 1930's with later applications in restriction site mapping of DNA. In particular, the fundamental problem that was considered is whether one could identify a given set of points from its multiset of distances. There are several related directions of research in the area, for example the question of recognizing the multisets corresponding to a multiset of distances realized by a set of points in the Euclidean space of given dimension, see [50].

Theorem 6.2. *Let m_{ij}, $0 \leq i, j \leq n$ be nonnegative real numbers with $m_{ij} = m_{ji}$ for all i, j and $m_{ii} = 0$ for all i. There are points $P_0, \dots, P_n \in \mathbb{R}^n$ with $||P_i - P_j|| = m_{ij}$ for all i, j if and only if the matrix Q with entries $q_{ij} = \frac{1}{2}(m_{0i}^2 + m_{0j}^2 - m_{ij}^2)$ is positive semidefinite.*

Here, a matrix Q is positive semidefinite if it is symmetric and $\mathbf{x}^T Q \mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

Let us come back to the function $h(n)$. As we see from Theorem 6.1, the gap between the upper and the lower bound is significant. There were some improvements on the lower bound for graphs of given size or diameter.

Theorem 6.3 ([7])**.** *Let G be a graph on n vertices with e edges, diameter d, $n \geq 5$, $d \geq 2$. If for an integer k, $\binom{\binom{k}{2}+d-1}{d-1} < n - 2k + 2$, then*

$$h(G) \geq \max\{k,\ d/2,\ \sqrt{2e/n}\}.$$

In particular,

$$h(G) \geq \max\{0.5n^{1/(2d-2)},\ d/2,\ \sqrt{2e/n}\}$$

for $n \geq c(d) = d^{2d-2}$.

Since it is known that almost all graphs $G(n,p)$ have diameter 2, see [14], Theorem 5.7 implies that for any $\epsilon > 0$ and almost every graph $G = G(n,p)$, for $n > n(\epsilon)$, $h(G_{n,p}) = n/2 - \epsilon n$. As the problem of finding large homometric sets is quite difficult in general, there was an effort devoted to determining $h(G)$ for graphs from some specific classes, such as specific tree classes.

A *spider* is a tree that is a union of vertex-disjoint paths and a vertex that is adjacent to one of the endpoints of each path. A spider is k-legged, $k \geq 3$ if its maximum degree is k. A *caterpillar* is a tree whose vertex set consists of two parts, one inducing a path and another inducing an independent set of leaves adjacent to the path. A *haircomb* is a tree, that is a union of a path called the *spine* and a collection of vertex-disjoint paths, called legs, that have an endpoint on the spine. In the following theorems, ceilings and floors are omitted for a more clear presentation. Axenovich and Ozkahya [7] determined several bounds for these classes of trees.

Theorem 6.4 ([7])**.** *Let T be a tree on n-vertices.*

- $h(\mathcal{T}) \geq n^{1/3} - 1$.
- *If T is a caterpillar, then $h(T) \geq n/6$.*
- *If T is a haircomb, then $h(T) \geq \sqrt{n}/2$.*

- *If T is a k-legged spider, then $h(T) \geq \frac{5}{12}n$ for $k = 3$; $h(T) \geq \frac{1}{3}n$ for $k = 4$; $h(T) \geq \left(\frac{1}{4} + \frac{3}{8k-12}\right) n$ for $k \geq 5$; $h(T) = (n+2)/4$ for $k = n/2$.*

Fulek and Mitrović [31] improved the general result for trees and for haircombs:

Theorem 6.5 ([31])**.** *If T is an n-vertex tree, then $h(T) \geq \sqrt{n/2} - 1/2$. If T is an n-vertex haircomb tree, then $h(H) \geq c \cdot n^{2/3}$, for a constant c.*

Dubickas [26] generalized the problem of homometric sets to multiple copies. Let $h(n, r)$ be the largest integer h such that in any n-vertex graph there are r pairwise disjoint sets of vertices, each of size h which are pairwise homometric. In [26] it is proven that for any positive ϵ, there is n_ϵ such that for any $n > n_\epsilon$, $h(n, r) \geq (1 - \epsilon)\frac{1}{r}\frac{\log n}{\log \log n}$.

7. Subgraphs with the same maximum degree

One can set up a weak twin problem for any graph parameter or a set of parameters. When this parameter is the maximum degree, the twin problem has been addressed. Caro and Yuster [21] introduced a function $s_k(G)$ for a positive integer k and a graph G to be the largest integer m such that there are k induced pairwise vertex-disjoint subgraphs of G, each with m vertices, and with the same maximum degree.

Theorem 7.1 ([21])**.** *For any fixed k and for any graph G on n vertices $s_k(G) \geq n/k - o(n)$.*

The proof of this theorem implies that k could be taken to be n^ϵ with $o(n)$ replaced with $o(n^{1-\epsilon})$, for sufficiently small ϵ.

8. Acknowledgements

The author thanks collaborators with whom she worked on various twin problems: Ryan Martin, Lale Özkahya, Yury Person, Svetlana Puzynina, and Torsten Ueckerdt, as well as Boris Bukh, Radoslav Fulek, Alex Scott, and Benny Sudakov for interesting discussions. In addition, the author acknowledges generous support and effort of IMA during the thematic year of 2014-2015 and a partial support of NSF grant DMS-0901008 of the twin problem project.

References

[1] M. Albertson, J. Pach, M. Young, *Disjoint Homometric Sets in Graphs,* Ars Math. Contemp. 4 (2011), 1–4.

[2] N. Alon and B. Bollobás, *Graphs with a small number of distinct induced subgraphs*, Discrete Math. 75 (1989), 23-30. 1.5, 6, 6.1

[3] N. Alon, Y. Caro, and I. Krasikov, *Bisection of trees and sequences*, Combinatorics and algorithms (Jerusalem, 1988). Discrete Math. 114, no. 1–3 (1993), 3–7. 3

[4] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, and R. Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms 16, no. 1 (1994), 80–109. 4.3.1

[5] N. Alon, M. Krivelevich and B. Sudakov, *Large nearly regular induced subgraphs*, SIAM J. Discrete Math. 22, no. 4 (2008), 1325–1337.

[6] M. Axenovich, R. Martin, and T. Ueckerdt, *Twins in graphs*, European J. Combin. 39 (2014), 188–197. 5, 5.3, 5, 5.6

[7] M. Axenovich, L. Ozkahya, *On homometric sets in graphs*, Australasian Journal of Combinatorics, **55** (2013), 175–187. 1, 1.5, 6.3, 6, 6.4

[8] M. Axenovich, Y. Person, and S. Puzynina, *Regularity Lemma and twins in sequences*, Journal of Combinatorial Theory, Series A, **120** (2013), 733–743. 1.3, 4.1, 4.1, 4.2, 4.2, 4.3, 4.3, 4.6

[9] S. Bachl, F.-J. Brandenburg, D. Gmach, *Computing and Drawing Isomorphic Subgraphs*, Journal of Graph Algorithms and Applications, **8**, no. 2 (2004), 215–238. 2

[10] P. Beame, D.T. Huynh-Ngoc, *On the value of multiple read/write streams for approximating frequency moments,* Electronic Colloquium on Computational Complexity (ECCC), May 2008. Technical Report TR08-024. 4.3.2

[11] A. Bekir and S. W. Golomb, *There are no further counterexamples to S. Piccard's theorem*, Institute of Electrical and Electronics Engineers. Transactions on Information Theory, 53 (2007), 2864–2867. 6

[12] I. Ben-Eliezer, M. Krivelevich, *Perfectly balanced partitions of smoothed graphs,* Electronic Journal of Combinatorics, **16**, no. 1 (2009), note N14. 5

[13] G. S. Bloom, *A counterexample to a theorem of S. Piccard*, Journal of Combinatorial Theory. Series A, 22, no. 3 (1977), 378–379. 6

[14] B. Bollobás, *Modern graph theory*, Graduate Texts in Mathematics, 184, *Springer-Verlag, New York*, 1998, xiv+394 pp. 6

[15] B. Bollobás, T. Kittipassorn, B. Narayanan, A. Scott, *Disjoint induced subgraphs of the same order and size*, European Journal of Combinatorics volume 49 (2015), 153–166. 1.4, 5, 5.7, 5

[16] J. Bosák, *Decompositions of graphs,* Translated from the Slovak. With a preface by Štefan Znám. Mathematics and its Applications (East European Series), 47. Kluwer Academic Publishers Group, Dordrecht, 1990. xviii+248 pp. 3

[17] B. Bukh, L. Zhou, *Twins in words and long common subsequences in permutations,* arXiv:1307.0088. 4.2, 4.4, 4.3, 4.3.2

[18] B. Bukh, J. Ma, *Longest common subsequences in sets of words,* arXiv:1406.7017. 4.3, 4.3.2

[19] Y. Caro, A. Shapira, R. Yuster, *Forcing k-repetitions in degree sequences*, Electron. J. Combin. 21, no. 1 (2014), Paper 1.24, 9 pp.

[20] Y. Caro, R. Yuster, *Large disjoint subgraphs with the same order and size*, European Journal of Combinatorics, **30**, no. 4 (2009), 813–821. 5, 5.1, 5.2, 5

[21] Y. Caro, R. Yuster, *Large induced subgraphs with equated maximum degree*, Discrete Math. 310, no. 4 (2010), 742–747. 7, 7.1

[22] C. Choffrut, J. Karhumäki, *Combinatorics of words. In: Handbook of Formal Languages*, Springer, 1997. 4.3.1

[23] F. R. K. Chung, R. L. Graham, *Quasi-random subsets of* $\mathbb{Z}_n$, J. Combin. Theory Ser. A 61, no. 1 (1992), 64–86. 4.3.1

[24] D. Conlon, *A new upper bound for diagonal Ramsey numbers*, Ann. of Math., 170 (2009), 941–960. 1, 2

[25] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press and McGraw-Hill (2001), 350–355. 4

[26] A. Dubickas, *Equal distances between equal sets of vertices in graphs,* Šiauliai Math. Semin. 8(16) (2013), 37–47. 6

[27] M. Dudík, L. J. Schulman, *Reconstruction from subsequences*, J. Comb. Theory, Ser. A 103(2) (2003), 337–348. 4

[28] T. Dzido, K. Krzywdziński, *On a local similarity of graphs.* Discrete Math. 338, no. 6 (2015), 983–989. 2

[29] P. Erdős, J. Pach, and J. Pyber, *Isomorphic subgraphs in a graph*, Combinatorics (Eger, 1987), 553–556, Colloq. Math. Soc. Jnos Bolyai, 52, North-Holland, Amsterdam, 1988. 3, 3

[30] J. Fox and B. Sudakov, *Decompositions into subgraphs of small diameter*, Combinatorics, Probability and Computing, 19, no. 5-6 (2010), 753–774.

[31] R. Fulek and S. Mitrović, *Homometric sets in trees*, European J. Combin. 35 (2014), 256–263. 6, 6.5

[32] R. Gould, V. Rödl, *On isomorphic subgraphs,* Discrete Mathematics, 118, no. 1-3 (1993), 259–262. 3

[33] W. T. Gowers, *A new proof of Szemerédi's theorem*, Geom. Funct. Anal. 11, no. 3 (2001), 465–588. 4.1

[34] L. Guidotti, *Sulla divisibilita dei grafi completi*, Riv. Mat. Univ. Parma 1 (1972), 231–237. 3

[35] F. Harary, R. Robinson, *Isomorphic factorizations. VIII. Bisectable trees,* Combinatorica 4, no. 2-3 (1984), 169–179. 3

[36] F. Harary, R. W. Robinson, and N. C.Wormald, *Isomorphic factorisations. I: Complete graphs*, Trans. Amer. Math. Soc. 242 (1978), 243–260. 3

[37] F. Harary and W. D. Wallis, *Isomorphic factorizations II: Combinatorial designs*, Congr. Numer. 19 (1978) 13–28. 3

[38] D. S. Hirschberg, *A linear space algorithm for computing maximal common subsequences*, Communications of the ACM 18 (6) (1975), 341–343. 4, 4.3.2

[39] P. Horn, V. Koubek, V.Rödl, *Isomorphic edge disjoint subgraphs of hypergraphs*, preprint, 2012. 3, 3.2

[40] G. Károlyi, *Balanced subset sums in dense sets of integers*, Integers, **9** (2009), A45, 591–603. 5, 5.5

[41] M. Kiwi, M. Loebl, and J. Matoušek, *Expected length of the longest common subsequence for large alphabets*, Adv. Math. 197, no. 2 (2005), 480–498. 4.3.2

[42] I. Koch, *Enumerating all connected maximal common subgraphs in two graphs*, Theoretical Computer Science **250** (2001), 1–30. 2

[43] J. Komlós, M. Simonovits, *Szemerédi's regularity lemma and its applications in graph theory*, In: Combinatorics, Paul Erdős is Eighty, Vol. 2 (Keszthely, 1993), volume 2 of Bolyai Soc. Math. Stud., pp. 295–352. János Bolyai Math. Soc., Budapest, 1996. 4.1

[44] C. Lee, P. Loh, B. Sudakov, *Self-similarity of graphs*, SIAM J. of Discrete Math., 27(2) (2013), 959–972. 1.2, 2, 3, 3.1, 3

[45] P. Lemke, S. S. Skiena, and W. D. Smith, *Reconstructing sets from interpoint distances*, Discrete and Comput. Geom., Algorithms Combin., 25 (2003), 507–631. 6

[46] G. Levi, *A note on the derivation of maximal common subgraphs of two directed or undirected graphs,* Calcolo 9 (1972), 341–352 (1973). 2

[47] M. Lothaire, *Algebraic combinatorics on words*, Cambridge University Press, 2002. 4.3.1

[48] L. Lovász, *Kneser's conjecture, chromatic number, and homotopy*, Journal of Combinatorial Theory, Series A, **25**, no. 3 (1978), 319–324. 5

[49] A. Mateescu, A. Salomaa, and S. Yu. *Subword histories and parikh matrices,* J. Comput. Syst. Sci., 68(1) (2004), 1–21. 4

[50] J. Matoušek, *Thirty-three miniatures*, Student Mathematical Library, 53, Mathematical and algorithmic applications of linear algebra, American Mathematical Society, Providence, RI, 2010, x+182. 6

[51] S. Piccard, *Sur les ensembles de distances des ensembles de points d'un espace Euclidien*, Mém. Univ. Neuchâtel, vol. 13, Secrétariat de l'Université, Neuchâtel, (1939) 212. 6

[52] F. P. Ramsey, *On a Problem of Formal Logic*, Proc. London Math. Soc. S2-30 no. 1 (1930), 264. 1, 2

[53] J. Rosenblatt, P. D. Seymour, P. D., *The structure of homometric sets*, SIAM J. Algebraic Discrete Methods, Society for Industrial and Applied Mathematics. Journal on Algebraic and Discrete Methods, 3 (1982), 343–350. 6

[54] A. Salomaa, *Counting (scattered) subwords,* Bulletin of the EATCS, 81 (2003), 165–179. 4

[55] G. Stevens, *Some caterpillars are* 2*-splittable.* Proceedings of the Twenty-fifth Southeastern International Conference on Combinatorics, Graph Theory and Computing (Boca Raton, FL, 1994). Congr. Numer. 104 (1994), 181–186. 3

[56] E. Szemerédi, *Regular partitions of graphs,* (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), pp. 399–401, Colloq. Internat. CNRS, 260, CNRS, Paris, 1978. 4.1

[57] D. B. West, *Introduction to graph theory*, Prentice Hall Inc., Upper Saddle River, NJ, 1996. 1

[58] X. Xia, *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics,* New York: Springer, 2007. 4, 4.3.2

Karlsruher Institut für Technologie, Karlsruhe, Germany

E-mail address: maria.aksenovich@kit.edu