Avoiding patterns in matrices via a small number of changes

Maria Axenovich* Ryan Martin†
Department of Mathematics
Iowa State University
Ames, IA 50011

May 4, 2005

Abstract

Let $\mathcal{A} = \{A_1, \ldots, A_r\}$ be a partition of a set $\{1, \ldots, m\} \times \{1, \ldots, n\}$ into r nonempty subsets, and $A = (a_{ij})$ be an $m \times n$ matrix. We say that A has a pattern $\mathcal{A}$ provided that $a_{ij} = a_{i'j'}$ if and only if $(i, j), (i', j') \in A_t$ for some $t \in \{1, \ldots, r\}$. In this note we study the following function f defined on the set of all $m \times n$ matrices M with s distinct entries: $f(M; \mathcal{A})$ is the smallest number of positions where the entries of M need to be changed such that the resulting matrix does not have any submatrix with pattern $\mathcal{A}$. We give an asymptotically tight value for

$$f(m, n; s, \mathcal{A}) = \max\{f(M; \mathcal{A}) : M \text{ is an } m \times n \text{ matrix with at most } s \text{ distinct entries}\}.$$

1 Introduction

The problem of studying the properties of matrices that avoid certain submatrices or patterns is a classical and well studied problem in combinatorics. It is investigated from a matrix point of view as well as in an equivalent formulation of forbidden subgraphs of bipartite graphs; see [1], [7], [4], [12], et al. Most of the previous research is devoted to extremal and structural problems of matrices with no forbidden submatrices. There are only a few results studying efficient modifications of matrices or graphs such that the resulting structure satisfies certain properties, for example, [5] and [6]. In this paper, we apply powerful graph theoretic techniques to study the distance properties between certain classes of matrices. Our main goal is to investigate the number of positions where the entry-changes need to be performed on a given matrix such that the resulting matrix does not have a fixed subpattern. Although this problem is of independent theoretical interest, it has multiple applications in computational biology such as in the compatibility of evolutionary trees and in studying metabolic networks, see [3], [13].

For positive integers m, n, s, with $s \leq mn$, let $\mathcal{M}(m, n; s)$ denote the set of all $m \times n$ matrices with a fixed number, s, of distinct entries. Let $\mathcal{A} = \{A_1, \ldots, A_r\}$ be a partition of pairs from $\{1, \ldots, m\} \times \{1, \ldots, n\}$ into r nonempty classes. An $m \times n$ matrix $A = (a_{ij})$ is said to have a *pattern* $\mathcal{A}$ provided that $a_{ij} = a_{i'j'}$ if and only if $(i, j), (i', j') \in A_t$ for some $t \in \{1, \ldots, r\}$. It follows, in particular, that two $m \times n$ matrices A and B with sets of distinct entries $S(A)$ and $S(B)$, respectively, have the same pattern if there is a bijection $g : S(A) \to S(B)$ such that $B(i, j) = g(A(i, j))$ for all $1 \leq i \leq m$ and all $1 \leq j \leq n$.

Example 1 *Matrices A and B have the same pattern with a corresponding bijection g; matrices A and B' have different patterns:*

$$A = \begin{pmatrix} 1 & 4 & 3 \\ 1 & 1 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 5 & 1 & 2 \\ 5 & 5 & 1 \end{pmatrix}, \qquad B' = \begin{pmatrix} 5 & 1 & 2 \\ 0 & 5 & 1 \end{pmatrix}.$$

*axenovic@math.iastate.edu
†rymartin@iastate.edu

In this case, $g(1) = 5$, $g(4) = 1$, $g(3) = 2$.

A $k \times \ell$ matrix B is a *submatrix* of an $m \times n$ matrix A if there are nonempty subsets $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_\ell\}$ of distinct indices with $\{i_1, \ldots, i_k\} \subseteq [m]$, $\{j_1, \ldots, j_\ell\} \subseteq [n]$ such that $B(\alpha, \beta) = A(i_\alpha, j_\beta)$, $1 \leq \alpha \leq k$, $1 \leq \beta \leq l$. If, for a matrix M', there is a submatrix M with pattern $\mathcal{A}$ then we say that M' has a *subpattern* $\mathcal{A}$.

Definition 1 *For a pattern $\mathcal{A}$ and positive integers m, n, s, we define* $\mathrm{Forb}(m, n; s, \mathcal{A})$ *to be the set of all $m \times n$ matrices with at most s distinct entries and not containing subpattern $\mathcal{A}$.*

Example 2 *Let $\mathcal{A} = \{\{(1,1),(1,2),(2,1)\},\{(2,2)\}\}$. The set* $\mathrm{Forb}(m, n; 2, \mathcal{A})$ *consists of all $m \times n$ matrices which have at most 2 distinct entries and contain no submatrix of the form* $\begin{pmatrix} x & x \\ x & y \end{pmatrix}$, $\begin{pmatrix} y & x \\ x & x \end{pmatrix}$, $\begin{pmatrix} x & x \\ y & x \end{pmatrix}$, $\begin{pmatrix} x & y \\ x & x \end{pmatrix}$, $x \neq y$. *In particular,* $\mathrm{Forb}(m, n; 2, \mathcal{A})$ *consists of $m \times n$ matrices with all entries equal and all $m \times n$ matrices with two distinct entries such that each row has all equal entries.*

Next we define the distance between two matrices and between classes of matrices. For two matrices A and B of the same dimensions, we say that $\mathrm{Dist}(A, B)$ is the number of positions in which A and B differ; i.e., it is the matrix Hamming distance. For a class of matrices $\mathcal{F}$ and a matrix A, all of the same dimensions, we denote $\mathrm{Dist}(A, \mathcal{F}) = \min\{\mathrm{Dist}(A, F) : F \in \mathcal{F}\}$. Finally,

$$f(m, n; s, \mathcal{A}) = \max\{\mathrm{Dist}(A, \mathcal{F}) : A \in \mathcal{M}(m, n; s), \mathcal{F} = \mathrm{Forb}(m, n; s, \mathcal{A})\}.$$

This function corresponds to the minimum number of positions on which the entries need to be changed in any $m \times n$ matrix with at most s distinct entries in order to eliminate all subpatterns $\mathcal{A}$. This problem is also called an *editing distance problem*, since we consider the minimum number of editing operations on a matrix, where each editing operation is a change of an entry in some position.

Note that $\mathrm{Forb}(m, n; s, \mathcal{A})$ might be an empty set of matrices for some patterns $\mathcal{A}$. For example, let s be fixed, and $\mathcal{A}$ be a pattern having exactly one set; i.e., a pattern corresponding to matrices with all entries being equal. We call such a pattern a *trivial pattern*. If m and n are large, then there is no $m \times n$ matrix with fixed number of distinct entries avoiding pattern $\mathcal{A}$. This follows from the finiteness of the bipartite Ramsey number, see [8]. On the other hand, when a pattern $\mathcal{A}$ has at least two distinct entries, then the class $\mathrm{Forb}(m, n; s, \mathcal{A})$ is nonempty since it contains all $m \times n$ matrices with a trivial pattern. Our main result is the following:

Theorem 1.1 *Let s, r be positive integers, $s \geq r$. Let b_1, b_2 be positive constants such that $b_1 \leq m/n \leq b_2$. Let $\mathcal{A}$ be a non-trivial pattern with r distinct entries, then*

$$f(m, n; s, \mathcal{A}) = (1 + o(1)) \left(\frac{s - r + 1}{s} \right) mn.$$

We shall prove these results using graph-theoretic formulations. A graph $H = (V, E)$ is bipartite if its vertex set can be partitioned such that $V = X \cup Y$, $X \cap Y = \emptyset$, and its edge set E is a subset of $X \times Y$. If $m = |X|$, $n = |Y|$ and $E = X \times Y$, then this graph is denoted $K_{m,n}$ and called a complete bipartite graph. Now, we can introduce a pattern on the edges of a complete bipartite graph as a partition of the edges in exactly the same manner as above. Let $\mathcal{A} = \{A_1, \ldots, A_r\}$ such that $E = A_1 \cup \cdots \cup A_r$ and A_i's are nonempty and pairwise disjoint. Then $\mathcal{A}$ is called a pattern on E. Now, let c be a coloring of edges of $K_{m,n}$. We say that c has a *pattern* $\mathcal{A}$ if it satisfies the property that $c(e) = c(e')$ if and only if $e, e' \in A_i$ for some $i = 1, \ldots, r$. If c is an edge-coloring of a graph G, we say that a coloring c' of a graph G' occurs in G under coloring c if there is a subgraph H of G isomorphic to G' such that the coloring c restricted to H coincides with the coloring c' of G'. Similar to the case with matrices, for a color pattern $\mathcal{A}$ defined on the edges of a graph G', we say that G has a subpattern $\mathcal{A}$ if there is an occurrence of a subgraph H in G such that H is isomorphic to G' and the coloring c restricted to H has a pattern $\mathcal{A}$.

For two edge-colorings c and c' of a graph G, we say that the *edit distance* between c and c' on G is the smallest number of edge-recolorings in G colored under c needed to obtain c'. For a given pattern $\mathcal{A}$ on edges of a complete bipartite graph, and an edge-colored $K_{m,n}$ with coloring c, let $F(m,n;c,\mathcal{A})$ be the smallest number of edge-recolorings of $K_{m,n}$ colored by c such that the resulting coloring does not contain a subpattern $\mathcal{A}$. Define

$$F(m,n;s,\mathcal{A}) := \max\{F(m,n;c,\mathcal{A}) : c \text{ uses } s \text{ colors}\}.$$

Observation. There is a bijection g between all $m \times n$ matrices with s distinct entries and all edge-colorings of $K_{m,n}$ using s colors. Indeed, this bijection can be defined as $g(M(i,j)) = c(\{i,j\})$, $i \in \{1,\ldots,m\}$, $j \in \{1,\ldots,n\}$; where $c(\{i,j\})$ is the color of an edge $\{i,j\}$ and $M(i,j)$ is the (i,j)th entry of the matrix. Moreover, a matrix M does not have subpattern $\mathcal{A}$ if and only if a coloring $g(M)$ does not have a subpattern $\mathcal{A}$.

For all other graph-theoretic terminology, we refer the reader to [14]. Our main theorem is proven in terms of graph colorings.

Theorem 1.2 *Let ϵ, $0 < \epsilon < 1$ be fixed, and let m', n', s, r; $s \geq r$, be fixed as well. Let $m+n$ be sufficiently large and let $\mathcal{A}$ be a pattern on $K_{m',n'}$ with r colors. Then,*

$$\left(1 - \epsilon\left(5s + 2 + (s+1)\left(\frac{m}{n} + \frac{n}{m}\right)\right)\right)\left(\frac{s-r+1}{s}\right)mn \leq F(m,n;s,\mathcal{A}) \leq \left(\frac{s-r+1}{s}\right)mn.$$

Observe that now Theorem 1.1 is an immediate corollary of Theorem 1.2 which we prove in Section 3. Section 2 describes the techniques that we use in the proof.

2 Main tools

For two disjoint sets of vertices X and Y, we shall refer to a pair (X,Y) as a complete bipartite graph with partite sets X and Y, and we denote its edges by $E(X,Y)$. Let $c : E(X,Y) \to \{1,\ldots,s\}$ be an edge-coloring of a pair (X,Y). For each color $\nu \in \{1,\ldots,s\}$, and any two subsets $X' \subseteq X$, $Y' \subseteq Y$, we denote by $E_\nu(X',Y')$ the set of edges of color ν in a pair (X',Y'). Then $d_\nu(X',Y')$ is the *density of a color ν* in the subgraph induced by X' and Y', defined as follows:

$$d_\nu(X',Y') = \frac{|E_\nu(X',Y')|}{|X'||Y'|}.$$

For $x \in X \cup Y$, we define $N_\nu(x)$ to be the set of all vertices joined to x by edges of color ν. We say that a pair (X,Y) is *ϵ-regular in color ν* if for every $X' \subseteq X$ and $Y' \subseteq Y$ with sizes $|X'| \geq \epsilon|X|$, $|Y'| \geq \epsilon|Y|$, we have

$$|d_\nu(X,Y) - d_\nu(X',Y')| < \epsilon. \tag{1}$$

Lemma 2.1 is based on the so-called many-color regularity lemma of Szemerédi, see [10], and is an implication of the refinement argument, i.e., Theorem 8.4 in [11].

Lemma 2.1 (Bipartite Many-Color Regularity Lemma [11]) *For any $\epsilon > 0$ and integers s, m' there exists M, a positive integer, such that if the edges of a bipartite graph $G = (X,Y;X \times Y)$ are colored with $1,\ldots,s$, then the vertex set $V(G)$ can be partitioned into sets $V_0, V_1, \ldots, V_k$, for some k, $m' \leq k \leq M$, so that $|V_0| < \epsilon(|X| + |Y|)$, and $|V_i| = |V_j|$, for $i,j \in \{1,\ldots,k\}$, and all but at most ϵk^2 pairs (V_i,V_j) are ϵ-regular in color ν, for each $\nu = 1,\ldots,s$, and either $V_i \subseteq X$ or $V_i \subseteq Y$, for $i = 1,\ldots,k$.*

In addition, we need to prove a multicolor version of the so-called intersection property, which is stated in [11] and revised in [2].

Fact 2.2 (Many-Color Intersection Property) *Let $\epsilon > 0$ and $\delta > 0$ be fixed and r and ℓ be positive integers. Let (A, B) be a pair with edges colored such that color ν is ϵ-regular with density d_ν, $d_\nu \geq \delta$ for $\nu = 1, \ldots, r$. Let $Y \subset B$. Assume that $(\delta - \epsilon)^{\ell-1}|Y| > \epsilon|B|$. Let k_ν for $\nu = 1, \ldots, r$ be a positive integer such that $\sum_{\nu=1}^{r} k_\nu = \ell$ and let any vector $\mathbf{a} \in A^\ell$ be indexed such that*

$$\mathbf{a} = \left(a_{[1,1]}, \ldots, a_{[1,k_1]}, a_{[2,1]}, \ldots, a_{[r-1,k_{r-1}]}, a_{[r,1]}, \ldots, a_{[r,k_r]}\right).$$

Then,

$$\#\left\{\mathbf{a} \in A^\ell : \left|Y \cap \bigcap_{\nu=1}^{r} \bigcap_{i=1}^{k_\nu} N_\nu(a_{[\nu,i]})\right| < \prod_{\nu=1}^{r} (d_\nu - \epsilon)^{k_\nu} |Y|\right\} \leq \ell\epsilon|A|^\ell. \tag{2}$$

The proof of Fact 2.2 is a standard argument which follows by induction on ℓ.

Corollary 2.3 *Let $\epsilon > 0$ and $\delta > 0$ be fixed and r and ℓ be positive integers. Let c be an edge-coloring of a pair (A, B) with at least r colors from $\{1, \ldots, r, \ldots\}$ such that color ν is ϵ-regular with density d_ν, $d_\nu \geq \delta$, for $\nu = 1, \ldots, r$. Let us be given that $(\delta - \epsilon)^{\ell-1} > \epsilon$ and $2r^\ell\ell\epsilon < 1$ and $(\delta - \epsilon)^\ell|B| \geq \ell$. Then any edge-coloring of $K_{\ell,\ell}$ with colors from $\{1, \ldots, r\}$ will occur as a subcoloring of c.*

3 Proof of Theorem 1.2

3.1 Upper bound

We shall show that for any s-edge-coloring of a complete bipartite graph with partite sets of sizes m and n, there are at most $\left(\frac{s-r+1}{s}\right)mn$ editing operations sufficient to destroy a fixed color pattern with r colors.

Let $\mathcal{A}$ be a color pattern with r sets defined on a complete bipartite graph G and let c be an edge-coloring of $K_{m,n}$ with s colors. Without loss of generality, let 1 be the color of the largest color class in c. We shall recolor the $s - r + 1$ smallest color classes of c so that their new color is 1. The resulting coloring will use only $r - 1$ colors and thus will not contain a forbidden pattern. The $s - r + 1$ smallest color classes account for at most $(1 - (r-1)/s)\,mn$ edges. Thus,

$$F(n, m; s, \mathcal{A}) \leq \left(\frac{s-r+1}{s}\right)mn.$$

3.2 Lower bound

To establish the lower bound, we show that there is a coloring of the given complete bipartite graph requiring many edit-operations to destroy a forbidden pattern. We begin with a claim that gives us a coloring which is highly regular.

Claim 1 *Let s be a positive integer, and $0 < \epsilon < 1/2$. There is an integer M such that if $|X| \geq M$ and $|Y| \geq M$ then there is an edge-coloring c of a complete bipartite graph $G = X \times Y$, with colors $1, 2, \ldots, s$, satisfying the following property: If $X' \subseteq X$ and $Y' \subseteq Y$, such that $|X'|, |Y'| > (|X| + |Y|)(1 - \epsilon)/M$, then $d_\nu(X', Y') \in (1/s - \epsilon, 1/s + \epsilon)$, $\nu = 1, \ldots, s$.*

Claim 1 is proven by choosing a coloring at random and by showing that, with high probability, this coloring has the desired properties, see [9].

Fix $\epsilon > 0$, let c' be a coloring of $G = (X \cup Y, X \times Y)$, $|X| = m, |Y| = n$, of minimum edit distance from c with the property that c' contains no subpattern $\mathcal{A}$. Apply Lemma 2.1 with parameter ϵ to the coloring c'. Let M be the constant given by Lemma 2.1 and the partition having all the non-leftover sets being enumerated as $X_1, \ldots, X_p, Y_1, \ldots, Y_q$ with $|X_i| = |Y_j| = Q$ and $X_i \subseteq X$, $Y_j \subseteq Y$ for $1 \leq i \leq p$, $1 \leq j \leq q$. We call a pair (X_i, Y_j), a *good* pair if it is ϵ-regular in each color $\nu \in \{1, 2, \ldots, s\}$ in coloring c'. We have that there are at most $s\epsilon(p + q)^2$ pairs which are not good. Moreover, for each good pair (X_i, Y_j) there are at most $r - 1$ colors such that the denstity of those classes in coloring c' is at least $\delta = 2\epsilon$. Otherwise, Corollary 2.3 would imply that pattern $\mathcal{A}$ appears in c', a contradiction. Therefore, for a good pair (X_i, Y_j),

there are at least $(s-r+1)\left(\frac{1}{s}-3\epsilon\right)Q^2$ editing operations needed to obtain coloring c' from the coloring c. The regularity lemma gives that $m \geq pQ \geq m-\epsilon(m+n)$ and $n \geq qQ \geq n-\epsilon(m+n)$. Therefore, the total number of recolored edges is at least

$$\begin{aligned}
&(s-r+1)\left(\frac{1}{s}-3\epsilon\right)Q^2\left(pq-s\epsilon(p+q)^2\right)\\
&\geq \left(\frac{s-r+1}{s}\right)(1-3s\epsilon)\left(pQqQ-s\epsilon(pQ+qQ)^2\right)\\
&\geq \left(\frac{s-r+1}{s}\right)(1-3s\epsilon)\left((m-\epsilon(m+n))(n-\epsilon(m+n))-s\epsilon(m+n)^2\right)\\
&\geq \left(\frac{s-r+1}{s}\right)mn\left(1-\epsilon\left(5s+2+(s+1)\left(\frac{m}{n}+\frac{n}{m}\right)\right)\right)\\
&\geq \left(\frac{s-r+1}{s}\right)mn(1-C\epsilon),
\end{aligned}$$

where $C = 5s+2+(s+1)(b_1+b_2)$. ■

Remark. It should be noted that, although we prove theorems for submatrices, our results easily follow for other patterns. Suppose we wish to forbid patterns of the form $\begin{pmatrix} 1 & 2 \\ 1 & * \end{pmatrix}$, where the $*$ represents any entry, either a repeated 1 or 2 or a new entry 3. Our result depends only on the number of distinct entries in the pattern, so the (asymptotic) number of changes necessary and sufficient to forbid this pattern is the same as the number of changes needed to forbid $\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$ (that is, $(1+o(1))\left(\frac{s-1}{s}\right)mn$) but fewer than to forbid $\begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$ (that is, $(1+o(1))\left(\frac{s-2}{s}\right)mn$).

Acknowledgement. We are indebted to anonymous referees whose careful reading and friendly suggestions helped to significantly improve the presentation of the results.

References

[1] R. Anstee, *General forbidden configuration theorems,* J. Combin. Theory Ser. A **40** (1985), no. 1, 108–124.

[2] M. Axenovich, A. Kézdy and R. Martin, *On editing distance in graphs*, submitted.

[3] D. Chen, O. Eulenstein, D. Fernández-Baca and M. Sanderson, *Flipping: A supertree construction method* in Bioconsensus, AMS **61** DIMACS Series in Discrete Mathematics and Theoretical Computer Science (2003), 135–160.

[4] V. Deĭneko, R. Rudolf and G. Woeginger, *A general approach to avoiding two by two submatrices.* Computing 52 (1994), no. 4, 371–388.

[5] P. Erdős, A. Gyárfás and M. Ruszinkó, *How to decrease the diameter of triangle-free graphs.* Combinatorica **18** (1998), no. 4, 493–501.

[6] P. Erdős, E. Győri and M. Simonovits, *How many edges should be deleted to make a triangle-free graph bipartite?* Sets, graphs and numbers (Budapest, 1991), 239–263, Colloq. Math. Soc. János Bolyai, **60**, North-Holland, Amsterdam, 1992.

[7] Z. Füredi, *Turán type problems.* Surveys in combinatorics, 253–300, London Math. Soc. Lecture Note Ser., **166**, Cambridge Univ. Press, Cambridge, 1991.

[8] R. Graham, B. Rothschild and J. Spencer. *Ramsey theory.* Second edition. Wiley-Interscience Series in Discrete Mathematics and Optimization. A Wiley-Interscience Publication. John Wiley and Sons, Inc., New York, 1990.

[9] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs.* Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[10] J. Komlós, A. Shokoufandeh, M. Simonovits and E. Szemerédi, *The regularity lemma and its applications in graph theory.* Theoretical aspects of computer science (Tehran, 2000), 84–112, Lecture Notes in Comput. Sci., **2292**, Springer, Berlin, 2002.

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

[12] H. Prömel and A. Steger, *Excluding induced subgraphs. II. Extremal graphs.* Discrete Appl. Math. **44** (1993), no. 1-3, 283–294.

[13] G. Stephanopoulos, A. Aristidou and J. Nielsen, *Metabolic engineering: principles and methodologies.* Academic Press, San Diego, 1998.

[14] D. West, *Introduction to Graph Theory*, second edition, Prentice Hall, 2001.