On Graphs and Algebraic Graphs that do not Contain Cycles of Length 4

Noga Alon* H. Tracy Hall† Christian Knauer‡ Rom Pinchasi§
Raphael Yuster¶

August 28, 2011

Abstract

We consider extremal problems for *algebraic graphs*, that is, graphs whose vertices correspond to vectors in $\mathbb{R}^d$, where two vectors are connected by an edge according to an algebraic condition. We also derive a lower bound on the rank of the adjacency matrix of a general abstract graph using the number of 4-cycles and a parameter which measures how close the graph is to being regular. From this we derive a rank bound for the adjacency matrix A of any simple graph with n vertices and E edges which does not contain a copy of $K_{2,r}$: $\mathrm{rank}(A) \geq \frac{E-2n(r+1)}{r^2\sqrt{n}}$.

1 Introduction

This paper is devoted to the study of matrices in general and adjacency matrices of graphs in particular. It is the goal of this paper to study the relation between the rank of a matrix, the number of nonzero entries of it, and the structure of the matrix. Specifically, we will consider n by n matrices of a given rank d that do not contain a certain pattern of nonzero entries. We will try to bound the number of nonzero entries in those matrices in terms of n and d. Such questions can be considered as part of the study of sign patterns of matrices. The results in [CPR00, P02] provide a good example for such a study and they are particularly

*Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. nogaa@tau.ac.il. Research supported in part by the Israel Science Foundation, by a USA-Israel BSF grant, and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

†Department of Mathematics, Brigham Young University, Provo Utah 84602, U.S.A. h.tracy@gmail.com. This research has been conducted while the author was visiting the Technion – Israeli Institute of Technology.

‡Institut für Informatik Freie Universität Berlin. christian.knauer@inf.fu-berlin.de.

§Mathematics Dept., Technion—Israel Institute of Technology, Haifa 32000, Israel. room@math.technion.ac.il. Research supported by the Israel Science Foundation and G.I.F., the German-Israeli Foundation for Scientific Research and Development.

¶Mathematics Dept., University of Haifa, Haifa 31905, Israel. raphy@math.haifa.ac.il.

closely related to our paper. At the time of writing this paper we were not aware of that part of our results can be shown to follow from the results in [CPR00, P02]. We will remark on these results of [CPR00, P02] later.

Our paper also shades some light on the relation between extremal problems for graphs, and the rank of the adjacency matrices of these graphs. Let G be a graph on n vertices that does not contain a copy of $K_{q,r}$ as a subgraph where $q \leq r$. It is a well known theorem of Kövari, Sós, and Turán ([KST54]) that such a graph G can have at most $c(r)n^{2-\frac{1}{q}}$ edges.

In Section 2 we restrict our attention to a family of bipartite graphs whose vertices correspond to two finite sets of vectors in $\mathbb{R}^d$. Let $U = \{u_1, \ldots, u_m\}$ and $V = \{v_1, \ldots, v_n\}$ be two sets of vectors in $\mathbb{R}^d$. We will identify U and V with the vertex classes of the bipartite graph G where v_i is connected to u_j iff the scalar product (u_j, v_i) is positive. Such a bipartite graph G will be called an *algebraic graph* of dimension d. We would like to get an improvement over the Kövari-Sós-Turán Theorem for the family of algebraic graphs in a given dimension. More specifically, we consider algebraic graphs that do not contain a copy of $K_{2,r}$ as a subgraph and show that if $d \leq 4$, these graphs have only linearly many edges in terms of the number of their vertices.

In Section 3 we take a closer look at the case $q = 2$ in the Kövari-Sós-Turán Theorem. Let A be the adjacency matrix of the graph G, that is, $A = \{a_{ij}\}$ is the n-by-n matrix with $a_{ij} = 1$ when $\{i, j\}$ is an edge of G, and $a_{ij} = 0$ otherwise. We consider the rank of the matrix A over $\mathbb{R}$ and show the following relation between the number of edges E in G and $\text{rank}(A)$, the rank of A.

Theorem 1. *Let G be a simple graph not containing $K_{2,r}$ as a subgraph. If G has n vertices and E edges, then the rank of A satisfies*

$$\text{rank}(A) \geq \frac{E - 2n(r+1)}{r^2\sqrt{n}}.$$

At the time of writing this paper we were not aware of how closely related the result in Theorem 1 is to the results in [CPR00, P02]. It is not hard to show that the case $r = 2$ in Theorem 1 follows from the results in [CPR00, P02]. In fact also the ideas used in the proofs are very similar. We will elaborate on this before proving Theorem 1 in Section 3.

Observe in particular that Theorem 1 implies that if G has roughly the maximum possible number of edges determined by the Kövari-Sós-Turán theorem, then the rank of A must be high (linear in n).

The proof of Theorem 1 will follow from a more general theorem, developed in Section 3, that bounds the rank of an adjacency matrix using the number of 4-cycles in the graph and a parameter that measures how close the graph is to being regular.

2 Extremal Problems for Algebraic Graphs

In this section we restrict our attention to the family of algebraic graphs, which are formally defined as follows.

Definition 1. Let $U = \{u_1, \ldots, u_m\}$ and $V = \{v_1, \ldots, v_n\}$ be two finite sets of vectors in $\mathbb{R}^d$. The corresponding *algebraic graph of dimension* d is the bipartite graph with color classes U and V and with edges connecting u_i to v_j iff the scalar product (u_i, v_j) is positive.

We would like to examine extremal problems in graph theory restricted to the family of graphs G that are algebraic of dimension d. We begin with a geometric result.

Definition 2. Let $P \subset \mathbb{R}^d$ be a set of n points. A subset $A \subset P$ is called a k-set of P if $|A| = k$ and A is the intersection of P with some open affine half-space.

Theorem 2. *Let $P \subset \mathbb{R}^3$ be a set of n points, and let $r \geq 2$ be a fixed integer. Let $A_1, \ldots, A_m$ be (not necessarily distinct) subsets of P. Assume that for every i, A_i is a k_i-set of P for some k_i, and that $|A_{i_1} \cap \ldots \cap A_{i_r}| \leq 1$ for every $1 \leq i_1 < \ldots < i_r \leq m$. Then $\sum_{i=1}^m k_i \leq m + 4(r-1)n$.*

Proof. Define a graph H whose vertices are the points of P. The edges of H are all the edges of the polytope $\mathrm{conv}(P)$. In addition, for every $x \in P$ that is not a vertex of $\mathrm{conv}(P)$ connect x to 4 points of P $y_1, \ldots, y_4$ such that $x \in \mathrm{conv}(\{y_1, \ldots, y_4\})$. H has at most $4n$ edges. This is because the edges of $\mathrm{conv}(P)$ form a planar graph on the vertices of $\mathrm{conv}(P)$.

We claim that if A is a k-set of P, then there are at least $k-1$ edges of H induced by the vertices in A. Indeed, this is because the graph H has the property that no matter which orientation of $\mathbb{R}^3$ we choose, every point in P is connected in H to a higher point, except the highest point with respect to the given orientation. This is true for the vertices of $\mathrm{conv}(P)$ and for the points of P that are not vertices of $\mathrm{conv}(P)$ as well, and implies that the induced subgraph of H on the vertices of any k-set is connected.

Considering $A_1, \ldots, A_m$, it follows from the condition that $|A_{i_1} \cap \ldots \cap A_{i_r}| \leq 1$ for every $1 \leq i_1 < \ldots < i_r \leq m$ that each edge in H may be induced by at most $r-1$ sets A_i. Therefore $\sum_{i=1}^m (k_i - 1) \leq 4(r-1)n$. From here it easily follows that

$$\sum_{i=1}^m k_i \leq 4(r-1)n + m.$$

∎

We are now ready to prove our first theorem concerning algebraic bipartite graphs that do not contain $K_{2,r}$ as a subgraph.

Theorem 3. *Let G be an algebraic bipartite graph of dimension 4, where m and n are the cardinalities of the two color classes of G. If G does not contain a copy of $K_{2,r}$, then the number of edges of G is $O(m + rn)$.*

Proof. Let $U = \{u_1, \ldots, u_m\}$ and $V = \{v_1, \ldots, v_n\}$ be the two sets of vectors corresponding to the vertices of G. By a suitable rotation of $\mathbb{R}^4$ one may assume that the last coordinate of each v_i and u_j is nonzero. Since the definition of the graph G depends only on the sign of the scalar products of vectors from U and vectors from V, by multiplying each vector in $U \cup V$ by a positive number, we may assume that the last coordinate of each v_i and u_j is either $+1$ or -1. Let U^+ denote the set of all vectors in U the last coordinate of which is $+1$, and let $U^- = U \setminus U^+$. Similarly define V^+ as the set of all vectors in V the last coordinate of which is $+1$ and let $V^- = V \setminus V^+$.

Without loss of generality we assume that at least $1/4$ of the edges in G are between vectors in U^+ and vectors in V^+. We will restrict our attention to the subgraph G' of G induced by $U^+ \cup V^+$. Without loss of generality we denote $U^+ = \{u_1, \ldots, u_{m'}\}$ and $V^+ = \{v_1, \ldots, v_{n'}\}$.

Let $T : \mathbb{R}^4 \to \mathbb{R}^3$ be the projection defined by $T(x_1, x_2, x_3, x_4) = (x_1, x_2, x_3)$. For every $1 \leq i \leq m'$ and $1 \leq j \leq n'$ let $\tilde{u_i} = T(u_i)$ and $\tilde{v_j} = T(v_j)$. For every $1 \leq i \leq m'$ let H_i be the affine hyperplane in $\mathbb{R}^3$, $H_i = \{\tilde{v} \in \mathbb{R}^3 | (\tilde{u_i}, \tilde{v}) = -1\}$. Let $P \subset \mathbb{R}^3$ be the set of vectors $P = \{\tilde{v_1}, \ldots, \tilde{v_{n'}}\}$. For every $1 \leq i \leq m'$ let A_i denote the set of points $\tilde{v_j}$ of P satisfying $(\tilde{u_i}, \tilde{v_j}) > -1$. A_i is a k_i-set of P for some k_i. The crucial observation is that u_i and v_j are connected by an edge in G' exactly when $\tilde{v_j} \in A_i$. Therefore, the number of edges in G' is precisely $\sum_{i=1}^{m'} |A_i| = \sum_{i=1}^{m'} k_i$. Observe that because G' does not contain a copy of $K_{2,r}$, for every $1 \leq i_1 < \ldots < i_r \leq m'$ we have $|A_{i_1} \cap \ldots \cap A_{i_r}| \leq 1$. Hence, by Theorem 2, $\sum_{i=1}^{m'} k_i \leq m' + 4(r-1)n' = O(m + (r-1)n)$. ∎

It is tempting to conjecture that Theorem 3 is true for algebraic bipartite graphs of any dimension, perhaps with the bound depending linearly on a function of d. This is however false at least when the dimension d is greater or equal to 5.

Recall that the bound $O(n^{4/3})$ of Szemerédi and Trotter ([ST83]) on the number of incidences between n point and n lines in the plane is best possible. This was shown in a construction by Erdős (see [E87]). We consider such a construction of n lines $\ell_1, \ldots, \ell_n$ and n points $p_1, \ldots, p_n$ with at least $cn^{4/3}$ incidences. For every $i = 1, \ldots, n$ let a_i and b_i be the parameters such that the line ℓ_i is represented by $\ell_i = \{(x, y) \in \mathbb{R}^2 \mid a_i x + b_i y = 1\}$. Let (A_i, B_i) denote the coordinates of the point p_i for $i = 1, \ldots, n$. Therefore, there are $cn^{4/3}$ pairs of a line ℓ_i and a point p_j such that $p_j \in \ell_i$, that is, $a_i A_j + b_i B_j = 1$. For every i and j consider the number $s_{i,j} = -(a_i A_j + b_i B_j - 1)^2$. This number is always negative except when $p_i \in \ell_j$ where it is equal to zero. Let $\epsilon = \min_{s_{i,j} \neq 0} |s_{i,j}|$. Then the number $-(a_i A_j + b_i B_j - 1)^2 + \epsilon/2$ is positive exactly when $p_j = (A_j, B_j)$ belongs to $\ell_i = \{(x, y) \in \mathbb{R}^2 \mid a_i x + b_i y = 1\}$.

We now present the expression $-(a_i A_j + b_i B_j - 1)^2 + \epsilon$ as a scalar product of two vectors, one depending only on p_j and the other depending only on ℓ_i. This is easy to do because

$$-(a_i A_j + b_i B_j - 1)^2 + \epsilon/2 = -a_i^2 A_j^2 - b_i^2 B_j^2 - 2a_i b_i A_j B_j + 2a_i A_j + 2b_i B_j - 1 + \epsilon/2.$$

Hence, $-(a_i A_j + b_i B_j - 1)^2 + \epsilon = (u_i, v_j)$, where $u_i = (-a_i^2, -b_i^2, -2a_i b_i, 2a_i, 2b_i, 1 - \epsilon/2)$ and $v_j = (A_j^2, B_j^2, A_j B_j, A_j, B_j, 1)$. Note that the vectors u_i and v_j are in $\mathbb{R}^6$.

Define $U = \{u_i \mid 1 \leq i \leq n\}$ and $V = \{v_j \mid 1 \leq j \leq n\}$ and consider the algebraic graph G with vertex set $U \cup V$. G is a graph of dimension $d = 6$ and has $cn^{4/3}$ edges and $2n$ vertices. On the other hand G does not contain a copy of $K_{2,2}$ as a subgraph, for no two distinct points among $p_1, \ldots, p_n$ belong to two distinct lines among $\ell_1, \ldots, \ell_n$.

In the next theorem we consider a weaker notion of an algebraic graph over any field and in any dimension.

Theorem 4. *Let $U = \{u_1, \ldots, u_m\}$ and $V = \{v_1, \ldots, v_n\}$ be two sets of vectors in $\mathbb{F}^d$, where $\mathbb{F}$ is any given field. Let G be the bipartite graph whose vertices are the disjoint union $U \cup V$ where u_i is connected by an edge to v_j iff the scalar product (u_i, v_j) is nonzero. If G does not contain a copy of $K_{2,r}$, for some $r \geq 2$, then the number of edges in G is at most $(r-1)d(m+n)$.*

Proof. Let $A = \{a_{i,j}\}$ be the m by n matrix with entries in $\mathbb{F}$ defined by $a_{i,j} = (u_i, v_j)$. Clearly, the number of nonzero entries in A equals the number of edges in G. We will show that there are at most $(r-1)d(m+n)$ nonzero entries in A.

Since the rank of A over $\mathbb{F}$ is at most d, there are d rows of A which span the entire row space of A. Without loss of generality assume that every row in A is a linear combination of the first d rows of A.

We claim that in each of the last $m-d$ rows of A there are at most $(r-1)d$ nonzero entries. Indeed, consider the i'th row $(a_{i,1} \ldots a_{i,n})$ of A where $d+1 \leq i \leq m$. Since this row is a linear combination of the first d rows of A, for every $1 \leq j \leq n$ for which $a_{i,j} \neq 0$ there exists $1 \leq t_j \leq d$ such that $a_{t_j,j} \neq 0$. Because G does not contain a copy of $K_{2,r}$, no two rows of A share more than $r-1$ nonzero coordinates. Therefore, there are at most $r-1$ indices j such that t_j is the same number between 1 and d. It follows that for at most $(r-1)d$ indices $1 \leq j \leq n$ $a_{i,j} \neq 0$. Consequently, the nonzero entries in A consist of at most dn possible entries in the first d rows of A, together with at most $(r-1)d(m-d)$ nonzero entries in the last $m-d$ rows of A. Adding the two bounds together we get an upper bound of (less than) $(r-1)d(m+n)$ for the number of nonzero entries in A and therefore also for the number of edges of G. ∎

Remark. Observe that the matrix A in the proof of Theorem 4 has rank d and can be considered as a representation of the adjacency matrix of a bipartite graph G. Theorem 4 studies the matrix A under the additional condition that A does not contain a 2 by r sub-matrix with nonzero entries. Such matrices were considered in [CPR00, P02] from a slightly different points of view. Specifically, in [P02] and [CPR00] A is a square matrix that has nonzero values on its diagonal. It is also assumed that A does not contain a 2 by 2 submatrix with nonzero entries. It is shown in [P02] that such a matrix has a very high rank, namely, $d \geq n/2$. For further extensions of this study we refer the reader to [P02].

Considering the proof of Theorem 4 in the case $r = 2$, one obtains the following result: If A is an m by n matrix over a field $\mathbb{F}$ and A does not contain a 2 by 2 submatrix all of whose entries are nonzero, then the rank d of A is at least the number of nonzero entries in A divided by $(m+n)$. In the next theorem we get a small improvement over this estimate when the number of nonzero entries in A is small.

Theorem 5. *Let A be an m-by-n matrix over an arbitrary field $\mathbb{F}$. Assume that A does not have any 2×2 submatrix all of whose entries are nonzero. Then the rank d of A over $\mathbb{F}$ satisfies: $d \geq \Omega((\#NZ(A) - m - n)^{1/3})$, where $\#NZ(A)$ is the number of nonzero entries in A.*

Proof. Let d denote the rank of A over $\mathbb{F}$. Then A can be written as a product $A = BC$ where B is an m-by-d matrix and C is a d-by-n matrix, both of rank d. We regard the columns of C as points in $\mathbb{F}^d$ that we denote by $P = \{p_1, \ldots, p_n\}$, where the i'th column of C represents the coordinates of p_i in $\mathbb{F}^d$. We regard the rows of B as vectors in $\mathbb{F}^d$ perpendicular to hyperplanes $H_1, \ldots, H_n$. That is

$$H_i = \{(x_1, \ldots, x_d) \in \mathbb{F}^d | B_{i,1}x_1 + \ldots + B_{i,d}x_d = 0\},$$

where $(B_{i,1} \ldots B_{i,d})$ is the i'th row of B. Observe that $\cap_{i=1}^m H_i$ is trivially $\{0\}$ for otherwise the rank of B is smaller than d.

Since $\cap_{i=1}^m H_i = \{0\}$ there are d hyperplanes among $H_1, \ldots, H_m$ whose intersection is trivial. Without loss of generality we will assume that $H_1 \cap \ldots \cap H_d = \{0\}$.

The condition on the matrix A is equivalent to the fact that for every $1 \leq i < j \leq m$ $H_i \cup H_j$ includes all points of P except maybe one. Let P' be the set of all points of P not included in at least one of the sets $H_i \cup H_j$, where $1 \leq i < j \leq d$. Then $|P'| \leq \binom{d}{2}$.

For each $1 \leq i \leq d$ let $\overline{H_i} = \cap_{j \neq i, 1 \leq j \leq d} H_j$. Observe that for every $1 \leq i \leq d$, $\overline{H_i}$ is a linear space of dimension 1.

We claim that each point in $P \setminus P'$ must lie in the union $\overline{H_1} \cup \ldots \cup \overline{H_d}$. Indeed, let $p \in P \setminus P'$. If $p = 0$ there is nothing to prove. For $p \neq 0$, there is $1 \leq i \leq d$ such that $p \notin H_i$. But then $p \in \overline{H_i}$ for otherwise there exists $j \neq i$ such that $1 \leq j \leq d$ and $p \notin H_j$. This is a contradiction because as $p \in P \setminus P'$ we have $p \in H_i \cup H_j$ for every $1 \leq i < j \leq d$.

Suppose next that $s \neq t$ and $p_s, p_t \in P$ are two nonzero points on $\overline{H_i}$. Then p_s and p_t must belong to every H_j for any $j \neq i$ between 1 and m. This is because otherwise both p_s and p_t do not belong to $H_i \cup H_j$ contradicting our assumption. In other words if p_s and p_t both belong to the same $\overline{H_i}$ then the corresponding columns of A (columns number s and t) have only one nonzero entry each, unless one of them is the zero column.

It follows that besides the columns of A that correspond to the points of P' and possibly an additional d columns corresponding to points on $\cup_{i=1}^d \overline{H_i}$, all other columns of A have at most one nonzero entry each.

The argument above is symmetric for n and m and therefore, we conclude that apart from at most d^2 rows of A and d^2 columns of A, all other rows and columns have at most one nonzero entry.

By the theorem of Kövari, Sós, and Turán, a d^2-by-d^2 submatrix of A has $O(d^3)$ nonzero entries. Therefore $\#NZ(A) \leq O(d^3) + m + n$, and the assertion of Theorem 5 follows. ∎

Remark. Following the remark of one of the referees, the result in Theorem 5 can be improved over the field of reals as follows:

Without loss of generality assume that $m \leq n$ and extend A to an n by n matrix with the least $n - m$ rows equal to zero. Let G be the bipartite graph on $2n$ vertices $v_i, \ldots, v_n$ and $u_1, \ldots, u_n$ whose adjacency matrix corresponds to the matrix A in the sense that v_i is connected to u_j iff $A_{ij} \neq 0$. The condition on the matrix A implies that G does not contain a copy of $K_{2,2}$. Let k be the largest size of a matching in G and without loss of generality assume that $v_1, \ldots, v_k$ are matched to $u_1, \ldots, u_k$, respectively. Therefore, every edge in G is incident to one of the vertices $v_1, \ldots, v_k, u_1, \ldots, u_k$. It follows that A is a matrix whose support consists of k rows that correspond to the the vertices $v_1, \ldots, v_k$ and k columns that corresponds to the vertices $u_1, \ldots, u_k$. Because G does not contain $K_{2,2}$ as a subgraph it follows from the bipartite version of the Kövari-Sós-Turán Theorem that A has at most $c(k\sqrt{n}+n)$ nonzero entries, for some absolute constant $c > 0$. In other words $\#NZ(A) \leq c(k\sqrt{n}+n)$. On the other hand by Corollary 13 in [CPR00] (see also [P02]), the rank of the k by k submatrix of A which correspond to the subgraph of G induced by $\{v_1, \ldots, v_k\} \cup \{u_1, \ldots, u_k\}$ is at least $k/2$. Therefore $d = rank(A) \geq k/2$. We can now deduce that $d \geq \frac{\frac{1}{c}\#NZ(A)-n}{\sqrt{n}}$. Observe that this result implies also the case $r = 2$ in Theorem 1, up to constant multipliers.

From each of theorems 4 and 5 the following observation follows. Let A be an n-by-n matrix with entries in a field $\mathbb{F}$. Assume that A does not have a 2-by-2 submatrix all of whose entries are nonzero, and that the number of nonzero entries in A is roughly the maximum possible value determined by the Kövari-Sós-Turán theorem, that is $\Omega(n^{3/2})$. Then the rank of A is $\Omega(\sqrt{n})$. In Section 3 we will significantly improve this lower bound on the rank of A in the case where A is a $\{0, 1\}$ matrix and the field $\mathbb{F}$ is $\mathbb{R}$.

3 Regularity, Rank, and Counting Squares

In this section we take a closer look at graphs that do not contain $K_{2,r}$ as a subgraph.

Definition 3. The *adjacency matrix* of a graph G is the matrix $A = \{a_{ij}\}$ whose rows and columns are indexed by the vertices of G, with a_{ij} equal to the number of edges from vertex i to vertex j.

All graphs considered will be undirected simple graphs, thus for our purposes an adjacency matrix is always a symmetric $\{0, 1\}$-matrix with zeros on the diagonal. If G has E edges, then $2E$ entries of the adjacency matrix of G are equal to 1.

We will be particularly interested in the relation between the rank and the number of 1 entries in adjacency matrices of graphs not containing $K_{2,r}$ as a subgraph. It follows from Theorem 4 that the rank d, over any field, of the adjacency matrix of a $K_{2,r}$-free graph with E edges and n vertices satisfies $2E \leq 2(r-1)dn$. One can interpret this inequality as saying that the rank of the adjacency matrix of a $K_{2,r}$-free graph cannot be very low if the number of edges in this graph is high. In particular when the number of edges in a $K_{2,r}$-free graph G is the maximum possible value determined by the Kövari-Sós-Turán theorem, then the rank of the adjacency matrix of G is at least of the order of $\sqrt{n}$. We will show that this bound can be significantly improved when the field in question is $\mathbb{R}$.

We begin with a definition relating graphs to $\{0, 1\}$-matrices, and two definitions which can be applied either to graphs or to matrices.

Definition 4. Given any $\{0,1\}$-matrix $A = \{a_{ij}\}$ of size m-by-n, the *bipartite graph defined by A* is the graph whose vertices are numbered $(1, \ldots, m+n)$, with an edge $\{i, m+j\}$ iff $a_{ij} = 1$.

The bipartite graph defined by A has as many edges as the sum of the entries of A.

The next definition comes in stages. To start, given a multiset of numbers there is a natural measure of how close they are to being equal: the ratio between the square of the average value of an element of the multiset and the average value of a square of an element.

Definition 5. Let $\mathcal{L} = (\ell_1, \ldots, \ell_n)$ be a finite list of real numbers, not all zero. Then the *regularity* of $\mathcal{L}$, denoted $\rho(\mathcal{L})$, is defined as follows:

$$\rho(\mathcal{L}) = \frac{\left(\sum \ell_i/n\right)^2}{\sum \ell_i^2/n} = \frac{\left(\sum \ell_i\right)^2}{n \sum \ell_i^2}.$$

We observe the following properties of $\rho(\mathcal{L})$:

- $\rho(\mathcal{L}) \in [0,1]$
- $\rho(\mathcal{L}) = 1$ if and only if $\ell_i = \ell_j$ for all i, j.
- If $\ell_i = 0$ for $i > k$, then $\rho(\mathcal{L}) = \frac{k}{n}\rho(\ell_1, \ldots, \ell_k)$, and in particular $\rho(\mathcal{L}) \leq \frac{k}{n}$.

By an abuse of notation we now extend the notion of regularity to graphs and to symmetric matrices:

Definition 5.1. In case G is a graph on n vertices with E edges, $E \geq 1$, the regularity of G, denoted $\rho(G)$, is the regularity of the degree sequence $(d_1, \ldots, d_n)$ of G.

This gives us

$$\rho(G) = \frac{\left(\sum d_i\right)^2}{n \sum d_i^2} = \frac{4E^2}{n \sum d_i^2}.$$

A regular graph is a graph G with regularity $\rho(G) = 1$.

Definition 5.2. In case A is a symmetric n-by-n matrix with at least one nonzero entry, the regularity of A, denoted $\rho(A)$, is the regularity of the spectrum $(\lambda_1, \ldots, \lambda_n)$ of A.

This can also be expressed as follows:

$$\rho(A) = \frac{\left(\sum \lambda_i\right)^2}{n \sum \lambda_i^2} = \frac{\mathrm{tr}(A)^2}{n\ \mathrm{tr}(A^2)}.$$

Since A is symmetric, $\mathrm{rank}(A)$ is the number of non-zero eigenvalues, and we have

$$\frac{\mathrm{tr}(A)^2}{\mathrm{tr}(A^2)} = \rho(A)\, n \leq \mathrm{rank}(A).$$

The inequality between the right-hand-side and the left-hand-side in the above formula is well known and has several applications, see, for example, [CPR00], [A03]. The only matrices A with regularity $\rho(A) = 1$ are multiples of the identity matrix.

One last graph parameter counts the number of 4-cycles in a graph:

Definition 6. For a graph G the *square count* of G, denoted $\Box(G)$, is the number of 4-cycles in G.

In case G is the bipartite graph defined by a $\{0,1\}$-matrix A, $\Box(G)$ is the number of submatrices of A equal to $\left[\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right]$.

We can now state a theorem giving a rank bound on $\{0,1\}$-matrices.

Theorem 6. *Let $A = \{a_{ij}\}$ be an m-by-n matrix with entries in $\{0,1\}$, not all of them zero. Let E be the number of nonzero entries of A, and let G be the bipartite graph defined by A, which has $(m+n)$ vertices and E edges. Then the rank of A over $\mathbb{R}$ is bounded by*

$$\operatorname{rank}(A) \geq \rho\big(AA^T\big)\, m$$

and furthermore

$$\rho\big(AA^T\big)\, m = \frac{E^2}{4E^2/\left(\rho(G)\,(m+n)\right) - E + 4\,\Box(G)}.$$

Proof. Let A and G be as in the statement of the theorem, and let $r_1, \ldots, r_m$ and $c_1, \ldots, c_n$ denote the row sums and column sums of A respectively, so that the vertex degrees $(d_1, \ldots, d_{m+n})$ of G are $(r_1, \ldots, r_m, c_1, \ldots, c_n)$, and $E = \sum r_i = \sum c_i = \frac{1}{2} \sum d_i$.

Although A is not necessarily a square matrix, we have

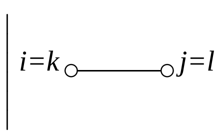
$$\operatorname{rank}(A) = \operatorname{rank}(AA^T)$$

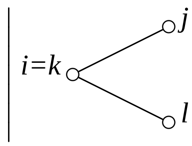
where AA^T is square of size m-by-m, and

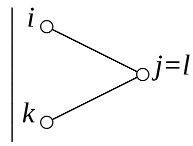
$$\operatorname{rank}(AA^T) \geq \rho\big(AA^T\big)\, m = \frac{\operatorname{tr}(AA^T)^2}{\operatorname{tr}(AA^T AA^T)},$$

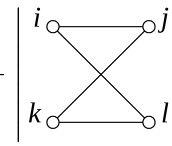
by a previous observation. We also have $\operatorname{tr}(AA^T) = \sum_{i,j} a_{ij}^2 = E$. To bound $\operatorname{rank}(A)$ in terms of the graph G, it remains to give a combinatorial description of $\operatorname{tr}(AA^T AA^T)$. This we do as follows:

$$\operatorname{tr}(AA^TAA^T) = \sum_{i,k}^{m}\sum_{j,l}^{n} a_{ij}a_{kj}a_{kl}a_{il}$$

$=$ 

$+$ 

$+$ 

$+$ 

$$= E + \sum_{i=1}^{n} c_i(c_i-1) + \sum_{j=1}^{m} r_j(r_j-1) + 4\,\square(G)$$

$$= E + \sum_{i=1}^{m+n} d_i^2 - 2E + 4\,\square(G)$$

$$= \frac{4E^2}{\rho(G)\,(m+n)} - E + 4\,\square(G)\,.$$

We thus have

$$\operatorname{rank}(A) \geq \rho\big(AA^T\big)\, m = \frac{\operatorname{tr}\left(AA^T\right)^2}{\operatorname{tr}\left(AA^TAA^T\right)} = \frac{E^2}{4E^2/\left(\rho(G)\,(m+n)\right) - E + 4\,\square(G)}$$

as claimed. ∎

Before giving the proof of Theorem 1, we remark again that our ideas are very similar to those in [CPR00, P02], and we were unaware of this at the time of writing this paper. In fact the case $r = 2$ in Theorem 1 can be easily deduced, up to constant multipliers, from the results in [CPR00, P02]. This was shown in the remark after the proof of Theorem 5.

Theorem 1 is a little more general because it deals also with the case $r > 2$. We shall now bring its proof.

Proof of Theorem 1

We claim that less than $2\sqrt{n}$ vertices of G have degree greater than $2r\sqrt{n}$. Indeed, let $k = \lceil 2\sqrt{n}\rceil$ and let $B = \{b_1, \ldots, b_k\}$ be a set of vertices of G with degree greater than $2r\sqrt{n}$. Each vertex in B has more than $2(r-1)\sqrt{n}$ neighbors in $V(G)\backslash B$. Since for every $i \neq j$, b_i and b_j have at most $r-1$ common neighbors, it follows that the number of vertices in $V(G)\backslash B$ that are neighbors of at least one of $b_1, \ldots, b_k$ is at least $2(r-1)\sqrt{n} + (2(r-1)\sqrt{n} - (r-1)) + \cdots + 1 > n$, a contradiction.

Let B denote be the set of vertices in G with degree greater than $2r\sqrt{n}$. By the bipartite version of the Kövari-Sós-Turán Theorem ([KST54]), denoting by d_v the number of neighbors in B for a vertex $v \in V(G) \setminus B$ we know that

$$\sum_{v \in V(G)\setminus B} \binom{d_v}{2} \leq (r-1)\binom{|B|}{2}\,.$$

Since $|B| < 2\sqrt{n}$ we have that the number of edges connecting the vertices in B to vertices not in B, which is at most $\sum_{v \in V(G)\setminus B} d_v$ is, in turn, at most $2nr$.

We therefore consider the subgraph G' of G obtained from G by deleting from it the vertices in B and all the edges incident to them. The number of vertices in G' is $n' = n - |B| \geq n - 2\sqrt{n}$ and the number of edges of G' is $E' \geq E - 2n(r+1)$. The maximum degree in G' is at most $2r\sqrt{n}$.

Let A denote the adjacency matrix of G'. Let H be the bipartite graph defined by A, so that H has $2n'$ vertices and $2E'$ edges.

We have $\sum_{i=1}^{n'} c_i = 2E'$. Moreover, for every $1 \leq i \leq n'$ we have $c_i \leq 2r\sqrt{n}$. Therefore:

$$\sum_{i=1}^{n'} c_i^2 \leq \frac{2E'}{2r\sqrt{n}}(2r\sqrt{n})^2 = 4rE'\sqrt{n}.$$

We deduce:

$$\rho(c_1, \ldots, c_{n'}) = \frac{(2E')^2}{n' \sum c_i^2} \geq \frac{E'}{rn'\sqrt{n}}.$$

We can also bound the quantity $\square(H)$. Observe that because G' does not contain a copy of $K_{2,r}$, we have $\square(H) \leq \frac{\sum_{i=1}^{n'} \binom{c_i}{2}}{r-1}\binom{r-1}{2}$. Hence,

$$\square(H) \leq \frac{r-2}{4}(\sum_{i=1}^{n'} c_i^2 - \sum_{i=1}^{n'} c_i) \leq \frac{r-2}{4}(4rE'\sqrt{n} - 2E') \leq (r-2)rE'\sqrt{n}.$$

Since A is symmetric, we have $\rho(H) = \rho(c_1, \ldots, c_{n'})$, and by Theorem 6 we have:

$$\begin{aligned} \operatorname{rank}(A) &\geq \frac{(2E')^2}{4\,(2E')^2 / (\rho(H)\,2n') - 2E' + 4\square(H)} \\ &\geq \frac{4E'^2}{8E'r\sqrt{n} + 4(r-2)rE'\sqrt{n}} \\ &\geq \frac{E'}{r^2\sqrt{n}} \geq \frac{E - 2n(r+1)}{r^2\sqrt{n}} \ . \end{aligned}$$

Since the adjacency matrix of G' is a submatrix of the adjacency matrix of G, then the rank of the adjacency matrix of G is at least $\frac{E-2n(r+1)}{r^2\sqrt{n}}$, as claimed.

∎

Remark. The result in Theorem 1 is asymptotically best possible. To see this recall that there exists a graph G with n vertices and $cn^{3/2}$ edges that does not contain $K_{2,2}$ as a subgraph, where $c > 0$ is an absolute constant. Let A be the adjacency of G. Let E be an integer smaller than $cn^{3/2}$. A has $k = \frac{E}{c\sqrt{n}}$ rows with a total of at least E 1-entries. Let A' be the $n \times n$ matrix which consists only of these k rows of A and zeros otherwise. Consider the $2n \times 2n$ matrix

$$B = \left[\begin{smallmatrix} 0 & A' \\ A'^T & 0 \end{smallmatrix}\right]$$

.

B is a symmetric matrix which is the adjacency matrix of a graph on $2n$ vertices that does not contain $K_{2,2}$ as a subgraph. B has at least $2E$ 1-entries while $rank(B) \leq 2k = \frac{2E}{c\sqrt{n}}$.

References

[A03] N. Alon, Problems and results in extremal combinatorics, I, Discrete Math. 273 (2003), 31–53.

[CPR00] B. Codenotti, P. Pudlák, and G. Resta, Some structural properties of low-rank matrices related to computational complexity, Theoret. Comput. Sci. 235 (2000), 89–107.

[E87] H. Edelsbrunner, Algorithms in combinatorial geometry. EATCS Monographs on Theoretical Computer Science, 10. Springer-Verlag, Berlin, 1987.

[KST54] T. Kövari, V. Sós, and P. Turán, On a problem of K. Zarankiewicz. *Colloquium Math.* **3**, (1954). 50–57.

[P02] P. Pudlák, Cycles of nonzero elements in low rank matrices. Special issue: Paul Erdős and his mathematics. Combinatorica **22** (2002), no. 2, 321–334.

[ST83] E. Szemerédi and W. Trotter, jr., Extremal problems in discrete geometry, *Combinatorica*, **3** (1983), 381–392.