

Random Simplicial Complexes and the Construction of Linear Error Correcting Codes

Adi Ben-Yoash

June 20, 2010

1 Abstract

It is a major open problem in coding theory to construct linear error correcting codes with a large rate and a large distance. Here we suggest a new approach to the construction of such codes that is based on the study of random two dimensional simplicial complexes. In this thesis we investigate some of the aspects of this main problem.

2 Introduction

2.1 Error Correcting Codes

A *binary error correcting code* of length n is a set $C \subseteq \{0, 1\}^n$. Members of C are called *code words*. If the code C has some desirable properties that we explain below, it can be used to communicate safely over noisy communication channels, where our transmissions include only code words from C . We have already defined the *length* of a code and now we turn to define the two main parameters of a code - its *rate* and its *distance*.

Since we only transmit code words from C it is obvious that we are utilizing the communication channel only partly. The bigger $|C|$ is, the higher our rate of channel utilization. Accordingly, we define the rate of the code as follows:

$$R(C) = \frac{1}{n} \log_2 |C|.$$

Finally, we come to error correction. By assumption we are transmitting over a noisy channel. Therefore, if we transmit the word $x \in C$, in the received message some of the bits of x may have changed. Since the receiver knows C as well, it can (if not too many errors have occurred) reconstruct x from its noisy version.

In order that we can correct errors using C it is necessary that different words in C not be too similar. If this is indeed the case, then the reconstruction of the transmitted code word that the receiver does (the *decoding*) works correctly. Specifically, we define the *Hamming Distance* of two words $x, y \in \{0, 1\}^n$:

$$d_H(x, y) = |\{i | x_i \neq y_i\}|.$$

This brings us to the definition of the *distance* of C :

$$d(C) = \min_{x, y \in C, x \neq y} (d_H(x, y)).$$

The main problem in the combinatorial theory of error correcting codes

is to construct codes C of length n for which both the rate $R(C)$ and the distance $d(C)$ are large.

There are many important aspects of the theory upon which we have not even touched. We have not discussed the statistical model of the noise generated by the communication channel and in particular did not even mention the important notion of channel *capacity*. We should also mention that there is another approach to this circle of problems which is more probabilistic (rather than combinatorial) in nature where, once a statistical model of the channel is at hand, seeks to minimize the overall *probability of error* in the transmission/reception process.

We have also ignored the *encoding* process by which the message we actually want to send is encoded in words from C . More importantly, we did not discuss the efficiency of the decoding algorithm (The reconstruction of x from its noisy version). These are fascinating and important issues, which must be kept outside the scope of this work.

2.1.1 A little about Linear codes

Linear codes constitute a very important family of error correcting codes. In order to define linear codes, we view $\{0, 1\}^n$ as the n -dimensional vector space over \mathbb{F}_2 , the field of order 2. In this case, we can describe C as the kernel of a binary matrix:

$$C = \{x \in \mathbb{F}_2^n \mid Ax = 0\}$$

where A is a binary matrix $A_{m \times n}$ ($m < n$). We usually assume that $\text{rank}(A) = m$, that is, the rows of A are linearly independent in the field \mathbb{F}_2 . The matrix A is usually called the *parity check matrix* of the code C . It is easy to see that in this case, $|C| = 2^{n-m}$, and therefore:

$$R(C) = 1 - \frac{m}{n}$$

The distance of C can be characterized as follows: It is the smallest number of columns of A that are linearly dependent. So, a major objective of coding theory is this:

Problem 1 Find $m \times n$ ($m < n$) binary matrices with no short dependencies between columns.

A single parameter version of this problem is:

Problem 2 Determine, for a given $0 < \delta < \frac{1}{2}$,

$$\limsup_{n \rightarrow \infty} \rho = \rho(\delta)$$

such that there exist arbitrarily large binary matrices of size $(1 - \rho)n \times n$ where every set of fewer than δn columns is linearly independent.

For a more detailed discussion concerning the known bounds on $\rho = \rho(\delta)$ see van Lint and Wilson [6]. We should mention that the general problem (for codes which are not necessarily linear) is widely open. A curious aspect of the theory is that the best known bounds for linear resp. linear codes are essentially identical. It is not clear if this reflects the actual truth or is just a failing of this research area.

A main purpose of this work is to suggest a new possible approach to the generation of such matrices.

We should mention that it is interesting to investigate these questions with the field \mathbb{F}_2 replaced by other finite fields \mathbb{F}_q . However, here we limit ourselves to the binary case.

2.2 A little about Random Graphs

The field of random graphs concerns graphs that are generated by various random processes. It is often useful to view a model of random graphs in two complementary ways:

1. In terms of the stochastic process that generates the graphs.
2. Investigate the distribution of the graphs thus generated.

The oldest model and most thoroughly explored model in this area is $G(n, p)$, The Erdős-Renyi model. Its description in terms of the underlying

stochastic process it is this: We start from a set of n vertices and $0 < p < 1$ is a given parameter. For every pair of vertices $\{x, y\}$ out of the n vertices we choose independently and with probability p to include the edge xy in the graph. In other words, Every two vertices are adjacent with probability p and are not adjacent with probability $1 - p$, independently of the other pairs.

Equivalently, we may think of $G(n, p)$ as a probability distribution over n -vertex graphs, where the probability of a given graph G with n vertices and $e(G)$ edges is:

$$\Pr(G) = p^{e(G)}(1 - p)^{\binom{n}{2} - e(G)}.$$

There are many more important models for random graphs. For instance:

1. Random regular graphs, where all vertices have the same degree.
2. More generally - Random graphs with a given degree sequence.
3. The $d_{out}(n, k)$ model: Here every vertex chooses a random set of k neighbours and connects with an edge to each of them. Note that this is a simple undirected graph. It contains the edge xy when either x chooses y or y chooses x or both.
4. In attempting to model the generation of The Internet, there is an interest in models of random graphs in which from time to time new vertices join an existing graph. A new vertex randomly chooses its set of neighbors with preference to vertices of higher degrees. Such models are called "Preferential attachment models" [3, 2].

Here we consider a model of random two dimensional simplicial complexes, which can be viewed as a two-dimensional analogue of the $d_{out}(n, 1)$ model.

A systematic study of random simplicial complexes was initiated by Linial and Meshulam [4]. Since then a number of papers were written in this area, but the whole subject is still in its infancy and much remains to be done.

2.3 A little about simplicial complexes

A *simplicial complex* is a finite family of sets F which is closed under inclusion. That is, if $A \in F$ and $B \subseteq A$ then $B \in F$. The elements of the underlying set are called *vertices* and the vertex set of F is denoted by $V(F)$. The sets in F are called *faces*, or *simplices*, and the *dimension* of the face $A \in F$ is defined as

$$\dim(A) := |A| - 1.$$

The *dimension of F* is defined as the highest dimension of a face in F ,

$$\dim(F) := \max_{A \in F} \dim A.$$

A *facet* in F is a face of dimension $\dim F$. Notice that a *graph is nothing but a one dimensional simplicial complex* whose facets are the edges. One more definition that we need is this: We say that a simplicial complex F is *k -full* if every subset of $V(F)$ of cardinality $\leq k + 1$ is a face of F .

This thesis can be viewed as part of a larger research effort which seeks to explore higher-dimensional analogues of various parts of graph theory which seem to be amenable to such extensions. To illustrate this point consider the following question

Question 3 *What is the high-dimensional analogue of a tree?*

This question has a number of meaningful answers. In this work we are interested in the notion of a *d -tree*, and more specifically, in 2-tree. Let us recall the recursive definition of a tree on n vertices. For $n = 2$, there is exactly one tree, that has two vertices and an edge connecting them. To obtain all the trees on vertex set $\{1, \dots, n+1\}$ we proceed as follows: Consider any tree on vertex set $\{1, \dots, n\}$ and add one more edge $(j, n+1)$ to some vertex $1 \leq j \leq n$.

Similarly, we define a 2-tree on the vertices 1, 2, 3 as the simplicial complex including all subsets of $\{1, 2, 3\}$. A 2-tree on vertex set $\{1, \dots, n+1\}$ is formed as follows: Start from any 2-tree T on vertex set $\{1, \dots, n\}$, and consider an edge (A one dimensional face) $(x, y) \in T$. Add the two dimensional face $(x, y, n+1)$, and of course the one dimensional faces $(x, n+1)$ and $(y, n+1)$.

2.4 The family of matrices

There are certain matrices that one usually associates with graphs. We recall one such matrix that plays a significant role here. The *incidence matrix* of a graph $G = (V, E)$ is a matrix A whose rows (resp. columns) are indexed by V (resp. E). For $v \in V$ and $e \in E$, we define $a_{v,e}$, the (v, e) entry of A as follows:

$$\text{If } v \in e \text{ then } a_{v,e} = 1 \text{ otherwise } a_{v,e} = 0.$$

We will be interested here in two-dimensional simplicial complexes F that have a full one-dimensional skeleton. Under these restrictions the natural analogue of a graph's incidence matrix is the *inclusion matrix* of one-vs.-two-dimensional faces of F . If $|V(F)| = n$, and since F is assumed to have a full one-dimensional skeleton the incidence matrix M has $\binom{n}{2}$ rows, one per each unordered pair of vertices. The columns of M are indexed by unordered triples for $[n]$ corresponding to F 's two-dimensional faces. As mentioned above, we are considering here a two-dimensional analogue of the random graph model $d_{out}(n, 1)$. Accordingly, the incidence matrix M is constructed as follows:

Every pair $\{x, y\} \subseteq [n]$ selects randomly, independently and uniformly an element $z \in [n] \setminus \{x, y\}$. We add a column to M corresponding to the triple $\{x, y, z\}$. Finally, M is the inclusion matrix of the selected triples vs. all pairs in $[n]$.

Here is a numerical example to illustrate this construction. with $n = 5$. For example, the pair (=the row) $(1, 4)$ can pick exactly one of the columns $(1, 2, 4)$, $(1, 3, 4)$ or $(1, 4, 5)$. This choice is made uniformly over the $n - 2$ possibilities. Note that a column may be chosen up to three times, but we ignore this aspect and include any triple that is selected any positive number of times. The matrix is binary, with every column having 3 entries of 1 and all other entries are 0. The 3 entries of 1 correspond to the rows that are indexed by a subset of the column at hand. For example, if $(1, 4, 7)$ is a column in M , its 3 one-entries reside in rows $(1, 4)$, $(1, 7)$ and $(4, 7)$.

Here is a concrete example matrix with $n = 5$. We first show an $\binom{n}{2} \times \binom{n}{2}$ matrix \tilde{M} which illustrates the process of selection and then the matrix M

where repeated columns are eliminated.

Table 1: Bold indicates the chosen facet

	1	1	1	1	1	2	2	3	1	1
	2	2	2	3	2	4	3	4	3	4
	4	3	4	5	3	5	5	5	5	5
(1,2)	1	1	1	0	1	0	0	0	0	0
(1,3)	0	1	0	1	1	0	0	0	1	0
(1,4)	1	0	1	0	0	0	0	0	0	1
(1,5)	0	0	0	1	0	0	0	0	1	1
(2,3)	0	1	0	0	1	0	1	0	0	0
(2,4)	1	0	1	0	0	1	0	0	0	0
(2,5)	0	0	0	0	0	1	1	0	0	0
(3,4)	0	0	0	0	0	0	0	1	0	0
(3,5)	0	0	0	1	0	0	1	1	1	0
(4,5)	0	0	0	0	0	1	0	1	0	1

As can be seen, this is a $\binom{n}{2} \times \binom{n}{2}$ the matrix is of size. The number of its 1 entries is $3 \cdot \binom{n}{2}$. However, since the same column can be chosen more than once (As is the case here with the columns (1, 2, 3), (1, 2, 4) and (1, 3, 5)), the matrix will usually have fewer columns (The distribution of this number is considered below). The elimination of such multiply-chosen columns clearly reduces as well the number of 1 entries.

Following the elimination of repeated columns we obtain the following matrix:

Table 2: Bold indicates the chosen facet

	1	1	1	2	2	3	1
	2	2	3	4	3	4	4
	4	3	5	5	5	5	5
(1,2)	1	1	0	0	0	0	0
(1,3)	0	1	1	0	0	0	0
(1,4)	1	0	0	0	0	0	1
(1,5)	0	0	1	0	0	0	1
(2,3)	0	1	0	0	1	0	0
(2,4)	1	0	0	1	0	0	0
(2,5)	0	0	0	1	1	0	0
(3,4)	0	0	0	0	0	1	0
(3,5)	0	0	1	0	1	1	0
(4,5)	0	0	0	1	0	1	1

We consider the matrix M thus constructed as the parity check matrix of a linear code. As constructed, M is inappropriate for the role of a parity check matrix, since it has fewer columns than rows. A possible way of overcoming this difficulty is for each row to choose randomly more than one column. This, however, is outside the scope of this thesis.

3 The main results

Our main focus in this work is the inclusion matrix M that was described above. We are specifically interested in understanding its left and right kernels. In order to develop some intuition for the problem, let us recall first what happens in an analogous one-dimensional situation, namely, incidence matrices of graphs. Let G be a graph and P its vertices vs. edges incidence matrix. A $0, 1$ vector is in the left kernel of P iff it is the indicator vector of the union of connected components in G . The right kernel of P is spanned by the indicator vectors of cycles in G and a general vector in the right kernel of P is the indicator vector of an Eulerian subgraph of G (a subgraph in which all vertices have even degrees). While the right kernel of M remains quite mysterious and would require additional investigation, we are able (Section 5) to determine (the leading term in) the typical dimension of the left kernel and to describe a corresponding number of independent vectors in the left kernel of M .

Theorem 4 *Consider a random 2-dimensional simplicial complex X on n vertices with a full 1-dimensional skeleton whose two-dimensional faces are chosen as follows: Every edge (one dimensional face) (x, y) of X selects uniformly and independently a vertex $z \neq x, y$ and the face (x, y, z) is included in X . Let M be the inclusion matrix of X , then the expected dimension of the left kernel of M is*

$$(1 + \epsilon)n + O(1)$$

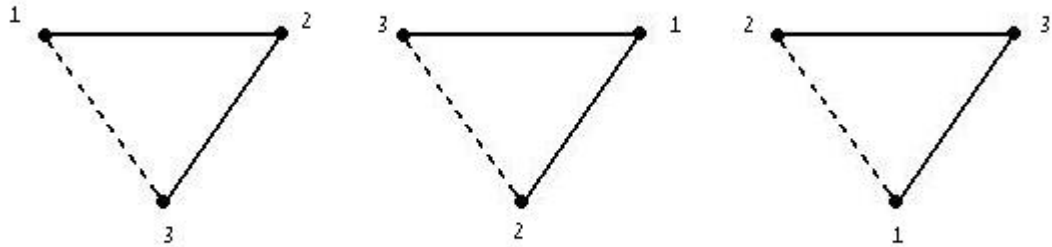
where $\epsilon = \sum_{j=0}^{n-3} \frac{1}{j!} (j+2)^{j-1} (\frac{2}{e^2})^j \approx 0.02$. Also, the expected dimension of X 's first homology is $\epsilon n + O(1)$ with the same ϵ .

Most of what follows is devoted to the proof of this theorem. We need to develop some background first.

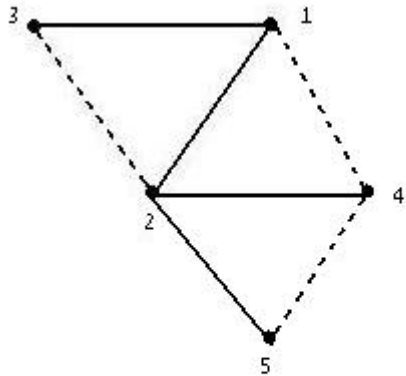
4 (1-2)-trees

A (1-2)-tree (V, T, τ) is a combinatorial structure that consists of a 2-tree T on vertex set V . An additional ingredient in the definition of a (1-2)-tree is a spanning tree τ on V such that every edge of τ is contained in some 2-dimensional face in T . We call τ the *skeleton* of the (1-2)-tree. (We note that this deviates from the usual meaning of skeleta in the context of simplicial complexes, but we hope that this creates no confusion).

To illustrate here are the three (1-2)-trees on vertex set $\{1, 2, 3\}$.



And here is a larger (1-2)-trees, on vertex set $\{1, 2, 3, 4, 5\}$



One of the basic facts in the study of trees within graph theory is **Cayley's Formula** which states that the number of labeled trees on n vertices is

n^{n-2} . One of the first challenges of the present work, is to develop a counting formula for (1-2)-trees. The answer is given in the following theorem.

Theorem 5 *The number of labeled (1-2)-trees on n vertices is $n(2n-2)^{n-3}$.*

Lemma 6 *Let τ be a tree with degree sequence $d_1 \dots d_n$. The number of labeled (1-2)-trees whose skeleton is τ is:*

$$\prod_{i=1}^n d_i^{d_i-2}$$

We prove this by constructing a bijection between the set of (1-2)-trees whose skeleton is τ and all ordered lists of n labeled spanning trees $(A_1 \dots A_n)$, where the tree A_i has d_i labeled vertices. By Cayley's formula, this will prove the Lemma. In our construction, the vertices of each tree A_i correspond to the edges incident with v_i , the i -th vertex in τ . The edge xy in A_i corresponds to the simplex $v_i xy$.

An *elementary cycle* in a two-dimensional simplicial complex is a set of simplices of the form $\{x, y_i, y_{i+1} | i = 1, \dots, t\} \cup \{x, y_1, y_t\}$.

Claim 7 *A 2-tree T contains no elementary cycles.*

Proof

By induction on the process by which T is constructed. Consider the first step at which an elementary cycle was created. This must be a step where one of the simplices $\sigma = x, y_i, y_{i+1}$ is added. But in this case, upon addition, σ shares at least two edges with already existing simplices, contrary to the definition of a 2-tree.

■

Given a (1-2)-tree T with skeleton τ , we now define a list of trees A_i as mentioned above. The tree A_i is defined in terms of the simplices in T that contain the i -vertex v_i . So consider a simplex $\sigma = v_i xy$ in T that contains v_i . Let $e = v_i x$ and $f = v_i y$ be the two corresponding edges in τ . Corresponding

to σ is an edge ef in A_i . Clearly A_i is a graph, but we need to show that it is in fact a tree.

By Claim 7, A_i is a cycle-free graph on d_i vertices and has, therefore, at most $d_i - 1$ edges. So, on one hand:

$$\sum_{i=1}^n e(A_i) \leq \sum_{i=1}^n (d_i - 1) = n - 2.$$

On the other hand, $\sum_{i=1}^n e(A_i)$ is the number of simplices in T , namely $n - 2$. It follows that all the inequalities above hold with equality. Consequently, for every i , the graph A_i is a tree, since it is cycle-free, has d_i vertices and $d_i - 1$ edges.

For the opposite direction, let τ be a tree on vertex set $\{v_1, \dots, v_n\}$ with corresponding vertex degrees d_1, \dots, d_n . Let E_i be the set of d_i edges incident with v_i . For $i = 1, \dots, n$, let B_i be a spanning tree with vertex set E_i . Let e, f be two adjacent vertices in B_i where e corresponds to the edge $v_i x$ in τ and f to $v_i y$. Then we associate with the edge ef in B_i the simplex $v_i xy$. We want to show that X the simplicial complex thus created is a (1-2)-tree that has τ as a skeleton.

Again we show this using the inductive definition of 2-trees. Consider two vertices v_i, v_j in τ whose distance in τ is the largest possible (i.e., their distance in τ is $\text{diam}(\tau)$). It is not hard to see that (i) v_i is a leaf in τ , and (ii) its (unique) neighbour, v_k , has at most one non-leaf neighbour, say v_l .

Now let us consider the tree B_k (on vertex set E_k). It has at least two leaves, so at least one of them differs from $v_k v_l$ (considered as a vertex of B_k), say it's $v_k v_m$. Let $v_k v_t$ be the unique neighbour of $v_k v_m$ in B_k . We claim that the simplex $\sigma = v_k v_m v_t$ is the only simplex containing v_m . This allows us to induct by (i) Removing σ from X , (ii) Removing the edge $v_t v_m$ from B_k (iii) Removing the vertex v_m from τ .

Now, we know from Cayley's formula that the number of labeled trees on d vertices is d^{d-2} , and since the choice of the trees is done independently, there are:

$$\prod_{i=1}^n d_i^{d_i-2}$$

possible labeled trees in our set.

Since we saw that there is a bijection between this set and the set of (1-2)-trees on the skeleton τ , our lemma is proved.

Lemma 8 *Let A be the set of trees on vertex set $\{v_1..v_n\}$, where v_i has degree d_i . The number of trees in A is:*

$$|A| = \binom{n-2}{d_1-1, d_2-1, \dots, d_n-1}$$

This is a well known lemma, and a proof can be found in many basic texts in discrete mathematics, e.g. [5].

Our goal is to show that there are $n(2n-2)^{n-3}$ labeled (1-2)-trees on n vertices. We already know that the number of trees τ with degree sequence d_1, \dots, d_n is $\binom{n-2}{d_1-1, d_2-1, \dots, d_n-1}$ and that there are $\prod_i d_i^{d_i-2}$ (1-2)-trees that have such a tree τ as their skeleton. Consequently, the number of (1-2)-trees on n vertices is:

$$\sum \binom{n-2}{d_1-1, d_2-1, \dots, d_n-1} \prod_{i=1}^n d_i^{d_i-2}$$

where the sum is over all sequences $d_1..d_n$ such that $\forall i d_i \geq 1$ and $\sum d_i = 2n-2$. We need to show now that this sum equals $n(2n-2)^{n-3}$. By moving all the terms that involve n to one side of the equation, this is the same as proving that:

$$\sum \prod_{i=1}^n \frac{d_i^{d_i-2}}{(d_i-1)!} = \frac{n(2n-2)^{n-3}}{(n-2)!}$$

Where the summation is as above. To this end we consider the exponential generating function of rooted trees [7]:

$$R(x) = \sum_{m \geq 1}^{\infty} \frac{m^{m-1}}{m!} x^m$$

or, written out:

$$R(x) = x + x^2 + \frac{3x^3}{2} + \frac{8x^4}{3} + \dots$$

Let us consider the formal power series R^n and evaluate the coefficient of x^{2n-2} in this sum. To find this coefficient we have to consider all the ways of expressing $2n - 2$ as an ordered sum $2n - 2 = d_1 + d_2 + \dots + d_n$ such that $d_i \geq 1$ for all i . It follows that this coefficient equals:

$$\sum_{d_1 \dots d_n, \forall i d_i \geq 1, \sum d_i = 2n-2} \prod_{i=1}^n \frac{d_i^{d_i-2}}{(d_i - 1)!}$$

(summing again, over the same range). This is precisely the expression we are after, so our proof will be completed if we can show that the coefficient of x^{2n-2} in $R^n(x)$ is indeed $\frac{n(2n-2)^{n-3}}{(n-2)!}$.

There is a very simple interpretation for the coefficient of $\frac{x^m}{m!}$ in $R^n(x)$. Namely, it is the number of rooted forests with n trees and m vertices, or, alternatively, the number of trees with $m + 1$ vertices and a given root with degree n whose incident edges are ordered ([6] pages 109-132).

In our case, this means the number of trees with $2n - 1$ vertices and a given root with degree n and ordered incident edges. Let us recall the following result of Clarke [1]: The total number of trees on r vertices in which the root has degree s is:

$$\binom{r-2}{s-1} (r-1)^{r-s-1}.$$

In our case, $r = 2n - 1$ and $s = n$. So, Clarke's formula becomes:

$$\begin{aligned} & \binom{2n-3}{n-1} (2n-2)^{n-2} = \\ & = \frac{(2n-3)!}{(n-1)!(n-2)!} (2n-2)^{n-2} \end{aligned}$$

Since the edges incident with the root are ordered, and there are n of them, there are $n!$ valid permutations for each tree, so, the coefficient of $\frac{x^{2m-2}}{(2m-m)!}$ in $R^n(x)$, which is the same as the number of trees with $2n - 1$ vertices and a given root with degree n and ordered edges, is:

$$\frac{(2n-3)!}{(n-1)!(n-2)!} (2n-2)^{n-2} (n!)$$

And therefore, the coefficient of x^{2n-2} is:

$$\begin{aligned} \frac{(2n-3)!}{(n-1)!(n-2)!(2n-2)!} (2n-2)^{n-2} (n!) &= \\ &= \frac{(2n-2)^{n-3} \cdot n}{(n-2)!} \end{aligned}$$

Which is what we wanted to prove.

■

5 Left Kernel

As it turns out, the leading term in the expected dimension of the left kernel of M consists of two ingredients, namely *stars* and *skeleta of (1-2)-trees*. It is convenient to associate an unordered pair with every coordinate of a vector v we consider here. Namely,

$$v = (v_{(1,2)}, v_{(1,3)} \dots v_{(1,n)}, v_{(2,3)} \dots v_{(2,n)} \dots v_{(n-1,n)}).$$

5.1 Stars

In the complete graph K_n there are n stars in which one vertex is the root and all other vertices are leaves. We consider the corresponding indicator vectors $v^1 \dots v^n$. They are $\binom{n}{2}$ -dimensional vectors indexed as above. The k -th star is defined via:

$$v_{(i,j)}^k = \begin{cases} 1 & \text{if } (i = k) \text{ or } (j = k) \\ 0 & \text{otherwise} \end{cases}$$

Every triple meets every star on either 0 or 2 edges. Consequently $v^k \cdot M = 0$ for every k . These n vectors span an $(n - 1)$ -dimensional subspace of the left kernel of M . (The sum of these vectors is zero, and this is easily seen to be the single linear dependency among these vectors).

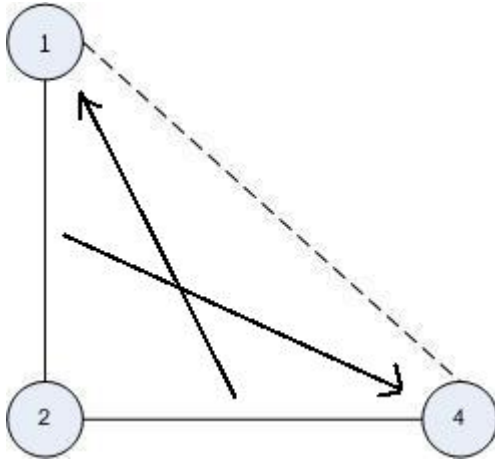
5.2 (1-2)-trees

The second type of vectors that have a statistically significant contribution to the left kernel are the characteristic vectors of the skeleton of an induced (1-2)-tree. To illustrate the idea here are two examples, a very small one and a slightly larger one

In our first (very small) example we let $n = 5$. Suppose that the pair (1, 2) chooses 4, and the pair (2, 4) chose 1. Let us also assume that the pairs (1, 2) and (2, 4) are not included in any other chosen simplex.

It follows that the rows of M that correspond to the pairs $(1,2)$ and $(2,4)$ are identical and therefore their mod 2 sum is zero (Both rows have a single 1 at the $(1, 2, 4)$ column and 0 otherwise).

In this case we associate the $(1-2)$ -tree in the following diagram with the linear dependency just described:

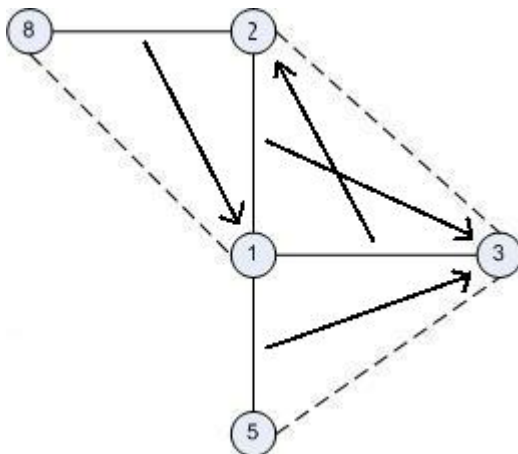


The full lines represent edges that participate in the linear dependency.

The dotted lines represent the additional pairs that complete the structure to a $(1-2)$ -tree.

The arrows represent the choice - An arrow between the edge xy and the vertex z indicates that the pair (x, y) chose the triple (x, y, z) .

We now turn to slightly larger example of a skeleton of a $(1-2)$ -tree:



In this example the following four rows of M sum to zero, thus yielding a vector in M 's left kernel.

	$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 8 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$
(1,2)	1	1	0
(2,8)	0	1	0
(1,3)	1	0	1
(1,5)	0	0	1

How many such linear dependencies do we expect to see in the random simplicial complex X we created?

Theorem 9 *For every $v \geq 3$, the expected number of (1-2)-trees with v vertices contained in the random simplicial complex X is*

$$(1 + o(1)) \binom{n}{v} \left(\frac{1}{n-2}\right)^{v-1} \left(\frac{1}{e^{2v-2}}\right) (v-2)v(2v-2)^{v-3}$$

The terms $\binom{n}{v}v(2v-2)^{v-3}$ account for choosing the v vertices of the (1-2)-tree and the actual (1-2)-tree.

Each of the $v-1$ tree edges must pick one vertex out of $n-2$ possible candidates to yield the $v-2$ simplices in the chosen (1-2)-tree. This explains the term $\left(\frac{1}{n-2}\right)^{v-1}$.

Each tree edge participates in no additional faces of the complex. In other words, for every edge (i, j) in the (1-2)-tree, for all other indices k , the edge (i, k) must not pick j and likewise (k, j) cannot pick i . The contribution of this condition, per each edge in T is $\left(1 - \frac{1}{n-2}\right)^{2(n-2)} = \frac{1+o(1)}{e^2}$. When we consider all $v-1$ edges, this accounts for the term $\frac{1+o(1)}{e^{2v-2}}$.

This takes care of the expected number of a (1-2)-tree shapes. In the diagrams above this accounts for the (1-2)-trees, but not for the arrows. Note that in every such diagram there is a unique face that is chosen twice (and hence contains two arrows in the diagram). We claim that once this face is chosen the whole system of edge to face selection is uniquely determined. Since there are $(v-2)$ faces in the (1-2)-tree, this explains the last unaccounted for term in the above formula.

Lemma 10 *A (1-2)-tree T on a skeleton τ formed in X is uniquely defined by the (single) face of T that is chosen twice.*

Proof

Let (x, y, z) be the twice chosen face, and let (x, y) and (y, z) be the two edges of this face that belong to τ . Consider next a face that contains one of these two, say (x, y, w) . We ask ourselves which edge of the face (x, y, w) chose it. Let's say that in addition to (x, y) , the other edge included in (x, y, w) that belongs to τ is (x, w) . Since (x, y) has already chosen the face (x, y, z) , it must be that (x, w) is the edge that chose (x, y, w) . We proceed this way from face to face and uniquely recover all choices that yielded the (1-2)-tree at hand.

■

It is not hard to see that no edge is in more than one such skeleton. Therefore in order to determine the typical dimension of the left kernel of M , it suffices to calculate the expected number of skeleton-type dependencies for a random matrix M . This expectation equals:

$$\begin{aligned}
& \sum_{v=3}^n \binom{n}{v} \left(\frac{1}{n-2}\right)^{v-1} \left(\frac{1}{e^{2v-2}}\right) (v-2) \cdot v(2v-2)^{v-3} \approx \\
& \approx n \sum_{v=3}^n \frac{1}{v!} (v-2)v \cdot 2^{v-3} (v-1)^{v-3} \frac{1}{e^{2v-2}} = \\
& = n \sum_{v=3}^n \frac{1}{(v-3)!} 2^{v-3} (v-1)^{v-4} \frac{1}{e^{2v-2}} = \\
& = n \sum_{j=0}^{n-3} \frac{1}{j!} 2^j (j+2)^{j-1} \frac{1}{e^{2j+4}} = \\
& = n \cdot e^{-4} \sum_{j=0}^{n-3} \frac{1}{j!} (j+2)^{j-1} \left(\frac{2}{e^2}\right)^j = \epsilon n
\end{aligned}$$

where it can be calculated that $\epsilon = e^{-4} \sum_{j=0}^{n-3} \frac{1}{j!} (j+2)^{j-1} \left(\frac{2}{e^2}\right)^j \approx 0.02$.

5.3 Proof of Theorem 4

It remains to prove that (1,2)-trees and stars exhaust the leading $\Theta(n)$ term in the expected dimension of the left kernel.¹ To this end consider a vector u in the left kernel of M . It is useful to view u as the indicator vector of a graph H . There is no loss of generality in assuming that H is connected, since otherwise the indicator vector of each connected component must be in the kernel as well, and the vectors corresponding to different connected components are linearly independent.

The same argument yields, in fact, more. We can assume that H is made connected by the simplices we select. Namely, it is possible to connect every two edges in H through a sequence of steps in each of which we can move from an edge $(x, y) \in E(H)$ to $(x, z) \in E(H)$ provided that the face (x, y, z) is in our random simplicial complex X . The reasoning is similar - We can divide $E(H)$ into parts which are “connected” in this sense. Each such connected part is in the kernel, and again the corresponding vectors are linearly independent.

If H has v vertices, then its number of edges and e can only be $v - 1$ or v . In the first case we are dealing with (1-2)-trees as above and in the second case H is a monocyclic graph. We already know (Theorem 9) the expected number of (1-2)-trees in our random complex. The number of monocyclic graphs H such that there is a collection of 2-dimensional faces that make H connected in the above sense does not exceed

$$(1 + o(1)) \binom{n}{v} \left(\frac{1}{n-2}\right)^v \left(\frac{1}{e^{2v}}\right) 2 \binom{v}{2} (2v-2)^{v-3}$$

It is easier to understand this expression vis-a-vis the one in Theorem 9. We view a monocyclic graph as a (1-2)-tree to which an arbitrary edge is added. This yields an over-count, since every edge in the single cycle could be the additional edge. This is, however, not a problem since we are only seeking an upper bound and not an exact expression. This choice of the additional edge accounts for the $\binom{v}{2}$ term. In a monocyclic graph there is no face that’s

¹It is conceivable that our methods are powerful enough to determine the next $\Theta(1)$ term in the asymptotic expansion of the expected dimension of the left kernel, but we have not pursued this possibility in full.

selected twice as above, the only freedom is in selecting an orientation of the single cycle, whence the factor of 2. The other two changes come from the fact that there are now v rather than $v - 1$ edges in H .

A direct calculation shows that this last expression does not exceed $(\frac{2}{e})^v$ so that when we sum over all $v \geq 3$ we get only $O(1)$.

6 Right Kernel

Unfortunately, at this stage, our understanding of M 's right kernel is still incomplete. This question is particularly interesting, since, as mentioned above, we'd like to consider M as the parity check matrix of a code. The code is just the right kernel of M , and in particular the code's distance is the least Hamming weight of a vector in M 's right kernel. We turn to study the vectors in M 's right kernel. One obvious type of such vectors are repetition vectors. We will also see vectors that correspond to triangulations of two-dimensional manifolds, but only a constant number of those.

6.1 Repetitions

Any two identical columns are obviously linearly dependent. Repeated columns occur when the same column is chosen by two distinct rows. We can exclude such columns from M entirely. If this is what we do, then we should determine the number of columns that remain in M . Let us denote by X the random variable that counts the number of repeated columns in M . Clearly,

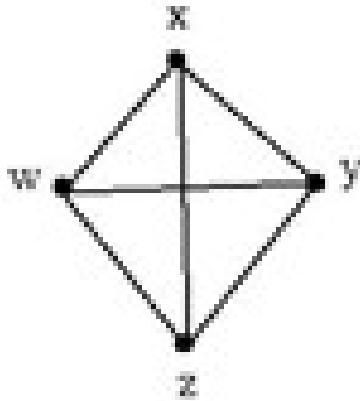
$$X = \sum_{|T|=3} Y_T + 2Z_T$$

where Y_T (resp. Z_T) is the indicator random variable of the event that the triple T was chosen exactly twice (resp. exactly three times).

$$\begin{aligned} \mathbb{E}(X) &= \binom{n}{3} \Pr(Y_T = 1) + 2 \cdot \binom{n}{3} \Pr(Z_T = 1) = \\ &= \binom{n}{3} \cdot 3 \cdot \frac{1}{(n-2)^2} \cdot \left(1 - \frac{1}{n-2}\right) + 2 \cdot \binom{n}{3} \cdot \left(\frac{1}{n-2}\right)^3 = \\ &= \frac{n}{2} + O(1). \end{aligned}$$

6.2 Triangulations

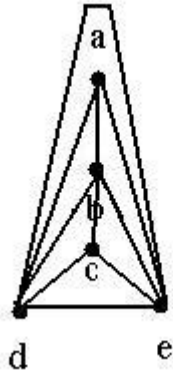
We can consider the collection of all faces in a triangulation of any two-dimensional manifold. For example, consider K_4 which we view as a triangulation of the plane or the 2-dimensional sphere. This complex has 4 vertices, and will look as follows:



This creates a dependency between the columns of M . Specifically, in this example, these columns would look like:

$$\begin{array}{rcccc}
 & \begin{pmatrix} w \\ x \\ y \end{pmatrix} & \begin{pmatrix} w \\ x \\ z \end{pmatrix} & \begin{pmatrix} w \\ y \\ z \end{pmatrix} & \begin{pmatrix} x \\ y \\ z \end{pmatrix} \\
 (w, x) & 1 & 1 & 0 & 0 \\
 (w, y) & 1 & 0 & 1 & 0 \\
 (w, z) & 0 & 1 & 1 & 0 \\
 (x, y) & 1 & 0 & 0 & 1 \\
 (x, z) & 0 & 1 & 0 & 1 \\
 (y, z) & 0 & 0 & 1 & 1
 \end{array}$$

A more complicated example is the following 6-vertex triangulation of the plane:



Here, the dependency between the columns in M would be:

$$\begin{array}{l}
 \begin{pmatrix} a \\ b \\ d \end{pmatrix} \quad \begin{pmatrix} a \\ b \\ e \end{pmatrix} \quad \begin{pmatrix} a \\ d \\ e \end{pmatrix} \quad \begin{pmatrix} b \\ c \\ d \end{pmatrix} \quad \begin{pmatrix} b \\ c \\ e \end{pmatrix} \quad \begin{pmatrix} c \\ d \\ e \end{pmatrix} \\
 (a, b) \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \\
 (a, d) \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \\
 (a, e) \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \\
 (b, c) \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \\
 (b, d) \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \\
 (b, e) \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \\
 (c, d) \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \\
 (c, e) \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \\
 (d, e) \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1
 \end{array}$$

How many of these dependencies will we see in M ? All the analysis below is carried out for oriented manifolds of genus g . The non-orientable case is similar and we omit the relevant (simple) modifications.

By the *Euler's formula* [6]:

$$v - e + f = 2 - 2g.$$

In a triangulation $3f = 2e$, and therefore:

$$v - \frac{f}{2} = 2 - 2g$$

or:

$$f = 2v - 4 + 4g$$

According to this, we can calculate how many v -vertex triangulations we'll see in M :

$$A_{v,g} \binom{n}{v} \left(\frac{1}{n-2}\right)^f = \Theta(n^{v-f}) = \Theta(n^{4-v-4g}).$$

Here $A_{v,g}$ is the number of v -vertex triangulations of an oriented 2-dimensional manifold of genus g . When $v = 4$ and $g = 0$, this number is constant, but for all other relevant values of v and g we get a $o(1)$ term.

Therefore, for large n we expect to see a constant number of K_4 's, and no additional dependencies of this kind.

Specifically, the expected number of K_4 's large ns is:

$$\begin{aligned} \lim_{n \rightarrow \infty} \binom{n}{4} \cdot \frac{1}{(n-2)^4} \cdot \binom{6}{4} &\approx \\ &\approx \frac{6!}{2! \cdot 4! \cdot 4!} = 0.625 \end{aligned}$$

6.3 Experiments

We have carried out a large number of numerical experiments. We observe additional vectors in the right kernel which we presently are unable to explain or analyze. The table below shows statistics about the number of repetitions, the number of K_4 's and the rank of M . This, in particular indicates those linear dependencies we which come neither from repetitions nor from K_4 's, for different values of n .

n	Runs	Reps (avg)	Reps (std)	K_4 (avg)	Other (avg)	Total (avg)	Total (std)
50	30	23.56	4.57	1	25.8	50.36	1.54
60	24	32.6	4.99	1.2	27.79	61.16	1.6
100	10	48.5	6.67	2.1	50.1	100.7	1.25
150	10	72.9	9.81	1.2	78.6	152.7	2.26
200	2	92	7.07	2	106.5	200.5	0.707

The total number of dependencies appears to be very close to n . This is in line with our analysis which shows that the left kernel has dimension $\geq 1.02n$. At any event, it seems very reasonable to expect, based both on our theoretical work on these experimental results, that the rank of M is $\binom{n}{2} - (1 + \gamma + o(1))n$ where $\gamma = 0.02\dots$ is the parameter computed above.

Also, it is evident that the number of repetitions is close to the theoretical calculation.

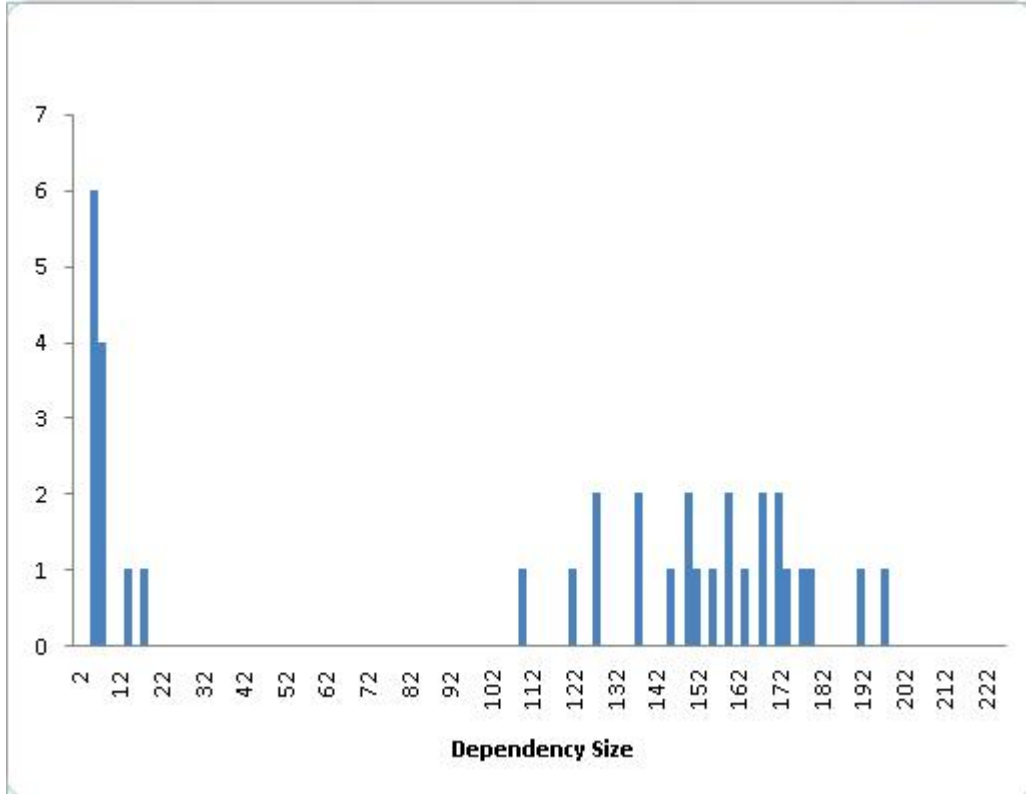
Another interesting parameter to look at is the shortest (in terms of Hamming weight) dependency found. This is interesting because for linear codes this determines the distance of the code.

The problem is that finding the shortest dependency is computationally hard [8]. If the dimension of the right kernel is k , we do not know of a much better way than going over 2^k vectors in order to find the one with the minimum weight.

On the other hand, for small n , we still see many dependencies that correspond to 6 and 8 vertex triangulations of the plane. Thus it is hard to draw any meaningful conclusions from experiments in which n is small.

As a compromise, we have conducted these experiments with $n = 40$. For such n , this computation is still feasible, and results can be interesting.

Here is a histogram of the shortest dependencies in matrices with $n = 40$, where the matrix has $\binom{40}{2} = 780$ rows.



It can be seen that most of the times, the shortest dependency is around 150. The short dependencies are all due to small (6, 8 or 10-vertex) triangulations of the plane. These should all disappear for higher values of n .

Left dependencies that correspond to stars can be considered "trivial", since they belong to the left kernel of the $\binom{n}{2} \times \binom{n}{3}$ inclusion matrix of pairs vs. triples. In this view it is interesting to experimentally examine the right kernel when we remove the $n - 1$ rows of the matrix M which correspond to all the pairs that include 0. (Any other spanning tree could be removed to this end). As expected and the next table shows, no significant changes were observed.

n	Runs	Reps (avg)	Reps (std)	K_4 (avg)	Other (avg)	Total (avg)	Total (std)
50	30	23.07	5	1.27	26.23	50.57	1.25
60	24	29.75	6.15	1.17	29.33	60.25	2.47
100	10	50.2	6.49	1.1	50.5	101.8	1.14
150	10	74.4	4.99	1	78.1	153.5	3.57
200	3	97.33	8.33	1	103	201.33	1.53

References

- [1] L. E. Clarke, On Cayley's Formula for Counting Trees. *J. London Math. Soc.* 33 1958 471-474.
- [2] Cooper, Colin and Frieze, Alan A general model of web graphs. *Random Structures Algorithms* 22 (2003), no. 3 311-335
- [3] H. Jeong, Z. Neda and A. L. Barabasi Measuring preferential attachment in evolving networks *Europhys. Lett.* 61 2003 567-572
- [4] N. Linial, and R. Meshulam Homological connectivity of random 2-dimensional complexes, *Combinatorica*, 26(2006) 475–487.
- [5] Nati Linial, and Michal Parnas, **Discrete Mathematics** (in Hebrew).
- [6] J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*. Cambridge University Press, 1992.
- [7] Moon, J. W. Counting labelled trees From lectures delivered to the Twelfth Biennial Seminar of the Canadian Mathematical Congress (Vancouver, 1969) *Canadian Mathematical Monographs*, No. 1
- [8] Vardy, Alexander The intractability of computing the minimum distance of a code *IEEE Trans. Inform. Theory* 43 (1997), no. 6 1757–1766