Coding for distributed storage &
the Birkhoff polytope

Sankeerth Rao (UC San Diego)

Joint with Daniel Kane and Shachar Lovett
An unexpected journey

Codes for distributed storage

Graph labeling

Birkhoff polytope graph

Structure vs Randomness

Image credit: www.maritzcx.com
Chapter 1: Codes for distributed storage
Codes for distributed storage

• Goal: store data reliably on multiple servers

• Model: erasure channel (node failures)

• Want to handle two types of node failures:
  • Local
  • Catastrophic

• Formalized as “Maximally Recoverable Codes” [Gopalan,Huang,Jenkins,Yekhanin 2014]
Simple grid topology

• Data is stored on $n^2$ nodes, viewed as $n \times n$ array
• Each entry is $k$ bits = element of $\mathbb{F}_2^k$

• Each row sums to zero: $\forall i, \sum_j x_{i,j} = 0$
• Each column sums to zero: $\forall j, \sum_i x_{i,j} = 0$

• One global linear constraint: $\sum_{i,j} y_{i,j} x_{i,j} = 0$
• Various extensions: multiple local/global constraints
Simple grid topology: local recovery

• Local failure recovery:
  Single node failure \((x_{i,j} \text{ erased})\)

• Recover from row (or from column)

• Read from \(n-1\) other nodes

• Quadratic speedup (codeword length = \# total nodes = \(n^2\))
Simple grid topology: catastrophic recovery

• **Catastrophic recovery:**
  Which erasure patterns can we hope to recover?

• Single erasure in row (row sum=0)
• Single erasure in column (column sum=0)
• Apply iteratively

• Global redundancy: handle cycles
Simple grid topology: catastrophic recovery

• Best understood if we view codeword as labeling of $K_{n,n}$, the complete bipartite graph

• Let $H \subset K_{n,n}$ be deleted nodes

• Can remove nodes of degree 1 (recover single erasure in row or column)

• Can recover remaining $H$ which is a simple cycle

• Forces requirement on the global redundancy:
  For any simple cycle $C$ in $K_{n,n}$, $\sum_{(i,j) \in C} \gamma(i,j) \neq 0$. 
Chapter 2:
Labeling the complete bipartite graph
Labeling the complete bipartite graph

• $\gamma: [n] \times [n] \to \mathbb{F}_2^k$

• Labeling edges of $K_{n,n}$

• Condition: for any simple cycle $C$
  \[ \sum_{(i,j) \in C} \gamma(i,j) \neq 0. \]

• When is this possible?

• How small can we choose the alphabet set (eg k)?
Labeling the complete bipartite graph: alphabet size

- $\gamma: [n] \times [n] \to \mathbb{F}_2^k$
- For any simple cycle $C$: $\sum_{(i,j) \in C} \gamma(i,j) \neq 0$.
- Upper bound: Random construction
  - $k = O(n \log n)$ suffices (as there are $n^{O(n)}$ simple cycles)
- Lower bounds:
  - $k \geq \Omega(\log n)$ (otherwise impossible even for cycles of length 4)
  - $k \geq \Omega(\log^2 n)$ [Gopalan,Hu,Kopparty,Saraf,Wang,Yekhanin 2017]
Labeling the complete bipartite graph: alphabet size

• \( \gamma: [n] \times [n] \to \mathbb{F}_2^k \)

• For any simple cycle \( C \): \( \sum_{(i,j) \in C} \gamma(i,j) \neq 0 \).

• **Main Theorem:** \( \frac{n}{2} \leq k \leq 3n \)

• Construction that beats the random construction
• Matching lower bound (up to constants)
Chapter 3: Explicit construction
Beats the random construction
Construction (beating random)

• Assume for simplicity: \( n \) is a power of 2
• Construct \( \gamma_n: [n] \times [n] \rightarrow \mathbb{F}_2^{4n} \) recursively

- First \( n \) bits: \( \gamma_n(i, j)_{1..n} = \begin{cases} 0 \ldots 010\ldots0, & j \leq n/2 \\ 00000000, & j > n/2 \end{cases} \)

- Next \( n \) bits: \( \gamma_n(i, j)_{n+1..2n} = \begin{cases} 0 \ldots 010\ldots0, & i \leq n/2 \\ 00000000, & i > n/2 \end{cases} \)

- Last \( 2n \) bits: \( \gamma_n(i, j)_{2n+1..4n} = \gamma_{n/2}(i \mod n/2, j \mod n/2) \)
Construction: correctness proof (1)

• fix simple cycle C
• Assume towards contradiction: $\sum_{(i,j) \in C} \gamma_n(i, j) = 0$.

• First n bits: $\gamma_n(i, j)_{1..n} = \begin{cases} 
0..010..0, & j \leq n/2 \\
00000000, & j > n/2 
\end{cases}$

• Claim: for every node $i$ on the left, its neighbours $j, j'$ are either both in the top $n/2$ or the both in the bottom $n/2$
• Proof: edges $(i, j), (i, j')$ only contribution to i-th coordinate of sum. As we assume sum=0, either $j, j' \leq n/2$ or else $j, j' > n/2$
Construction: correctness proof (2)

• fix simple cycle $C$
• Assume towards contradiction: $\sum_{(i,j) \in C} \gamma_n(i, j) = 0$.

• First $n$ bits of $\gamma_n$: for each node $i$ on the left, its neighbours are either both in the top $n/2$, or both in the bottom $n/2$

• Next $n$ bits of $\gamma_n$: for each node $j$ on the right, its neighbours are either both in the top $n/2$, or both in the bottom $n/2$

• Cycle $C$ is “trapped” in one of four $K_{n/2,n/2} : \{\text{top}, \text{bottom}\} n/2 \times \{\text{top}, \text{bottom}\} n/2$

• Last $2n$ bits: apply recursively $\gamma_{n/2}$ to each copy of $K_{n/2,n/2}$
Lesson

• Random is not always optimal
• (but it is always a useful benchmark)
Chapter 4:
Lower bound

This is where the Birkhoff polytope comes in...
Birkhoff polytope graph

• Birkhoff polytope: set of doubly stochastic $n \times n$ matrices
  = convex hull of $n \times n$ permutation matrices
• Its graph (nodes and edges) has an algebraic description

• $B_n = Cay(S_n, C_n)$ is a Cayley graph where:
  • Nodes = $S_n$ = symmetric group of permutations on $n$ elements
  • Generators = $C_n$ = cycles (permutations with exactly one nontrivial cycle)
  • Edges are $\{(\pi, \pi\sigma) : \pi \in S_n, \sigma \in C_n\}$

• Note: it is an undirected graph (because $C_n$ closed under inverse)
How is the Birkhoff polytope graph related?

Recall our problem:
• \( \gamma: \square \rightarrow \mathbb{F}^k \) such that for any simple cycle \( C \), \( \sum_{(i,j) \in C} \gamma(i, j) \neq 0 \).

• Let \( B_n = \text{Cay}(S_n, C_n) \) be the Birkhoff polytope graph.

• Define a coloring \( \chi: S_n \rightarrow \mathbb{F}_2^k \)

\[
\chi(\pi) = \sum_{i=1}^{n} \gamma(i, \pi(i))
\]

(i.e. sum of labels defined by over matching defined by \( \pi \))

• Lemma: \( \chi \) is a proper coloring of \( B_n \) (adjacent vertices get different values).
Proof of Lemma

- \( \gamma: [n] \times [n] \rightarrow \mathbb{F}_2^k \)
- \( B_n = \text{Cay}(S_n, C_n) \)
- \( \chi(\pi) = \sum_{i=1}^{n} \gamma(i, \pi(i)) \)

- **Lemma:** \( \chi \) is a proper coloring of \( B_n \)
- **Proof:**

Assume \( \chi(\pi) = \chi(\pi') \) where \( \pi, \pi' \in S_n \) are adjacent in \( B_n \) \iff \( \pi'\pi^{-1} \in C_n \)

Consider the matchings: \( M_\pi = \{(i, \pi(i)): i \in [n]\} \) and similarly \( M_{\pi'} \)

The symmetric difference \( M_\pi \oplus M_{\pi'} \) is a simple cycle \iff \( \pi'\pi^{-1} \in C_n \)

Its sum is \( \chi(\pi) + \chi(\pi') = 0 \). Contradiction to assumption on \( \gamma \).
Lower Bound Ideas:

• If $\gamma: [n] \times [n] \to \mathbb{F}_2^k$ satisfies that any simple cycle has nonzero sum, then:
  
  $B_n = Cay(S_n, C_n)$ can be colored with $2^k$ colors

• We prove a lower bound on chromatic number of $B_n$, by giving an upper bound on largest independent set.

• Let $A \subset S_n$ be an independent set in $B_n = Cay(S_n, C_n)$
• We prove: $|A| \leq \frac{n!}{c^n}$ for $c = \sqrt{2}$

• Our proof combines:
  • Representation theory of $S_n$
  • Structure-vs-pseudorandomness extension of the Hoffman bound
A standard approach, and why it fails here

• Standard approach: Hoffman bound
• Relies on extremal eigenvalues of adjacency matrix of $B_n$
• Eigenvalues $\leftrightarrow$ representations of $S_n$
• Extremal eigenvalues $\leftrightarrow$ low-dimensional representations
• These give poor quantitative bounds

• Our solution: if independent set is pseudo-random, then low-dimensional representations have negligible contribution
A pseudo-randomness condition

• $A \subset S_n$ independent set in $B_n = Cay(S_n, C_n)$

• Let $[n]_m = \{(i_1, \ldots, i_m) \text{ distinct elements of } [n]\}$

• $S_n$ acts on $[n]_m$: $\pi(i_1, \ldots, i_m) = (\pi(i_1), \ldots, \pi(i_m))$

• Definition: $A \subset S_n$ is $c$-pseudorandom is for any $m$, and any $I, J \in [n]_m$

$$\frac{|\{\pi \in A : \pi(I) = J\}|}{|A|} \leq \frac{c^m}{|[n]_m|}$$

(note that $S_n$ is 1-pseudorandom)
Usefulness of pseudo-randomness

• Pseudo-randomness condition implies that low-dimensional representations have negligible effect.

• Independent set is not Pseudo-random then we increment density to get a smaller similar problem.

• Structure vs randomness approach - We prove an upper bound on the independent set.
Epilogue
Summary of main results

• Motivation: understand grid topologies for codes for distributed storage

• Our journey took us to understand labeling of the complete bipartite graph, where the sum over any simple cycle is nonzero

• Reformulated the problem as understanding the chromatic number of the Birkhoff polytope graph

• Which in turn required a structure-vs-pseudorandomness approach to extend the standard Hoffman bound
If you want to see more details

• The paper is available online:

The independence number of the Birkhoff polytope graph, and applications to maximally recoverable codes

https://arxiv.org/abs/1702.05773

Thank you!
Lack of pseudo-randomness $\implies$ density increment

- $A \subset S_n$ independent set in $B_n = Cay(S_n, C_n)$
- Assume $A$ is not c-pseudorandom

- That is: For some $m$, and some $I, J \in [n]_m$, for $A' = \{ \pi \in A : \pi(I) = J \}$ we have

$$\frac{|A'|}{|A|} > \frac{c^m}{|[n]_m|}$$

- May assume WLOG that $I = J = \{n - m + 1, \ldots, n\}$ by multiplying all elements in $A$ from the left and right by appropriate permutations

- $A'$ embeds in $S_{n-m}$, is an independent set in $B_{n-m}$

- Induction: $|A'| \leq \frac{(n-m)!}{c^{n-m}}$

- Putting it together: $|A| < |A'| \frac{(n-m+1) \cdots n}{c^m} = \frac{n!}{c^n}$