Information theory via matroids and polymatroids

Thomas Westerbäck
Division of Mathematics and Physics, Mälardalen University

The 5th SNAG workshop in Algebra and Geometry,
28-29 March 2023

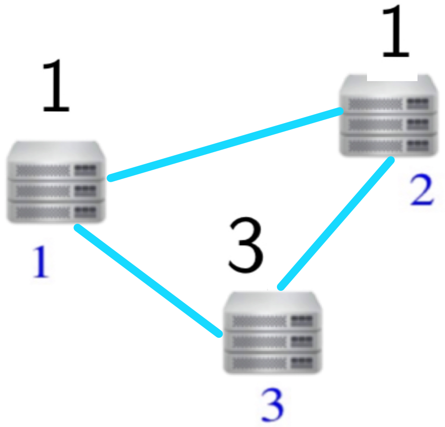
Entropy

Coded data

$$\begin{array}{ccc} 1 & 2 & 3 \\ \end{array}$$

$$\left\{\begin{array}{ccc} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 3 \\ 1 & 0 & 0 \\ 1 & 1 & 2 \end{array}\right\} \longrightarrow \begin{array}{ccc} 1 & 2 & 3 \\ (1 & 1 & 3) \end{array} \longrightarrow$$

Data network

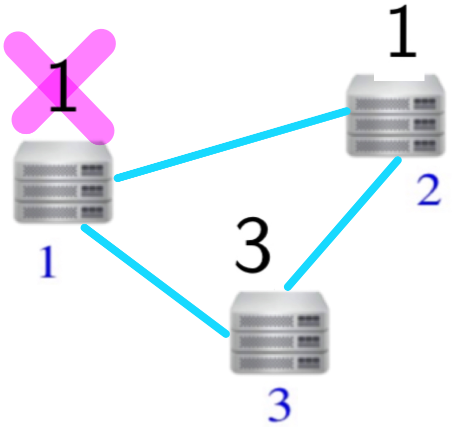


Entropy

Coded data

$$\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left\{ \begin{array}{ccc} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 3 \\ 1 & 0 & 0 \\ 1 & 1 & 2 \end{array} \right\} \end{array} \longrightarrow \begin{array}{ccc} 1 & 2 & 3 \\ (1 & 1 & 3) \end{array} \longrightarrow$$

Data network

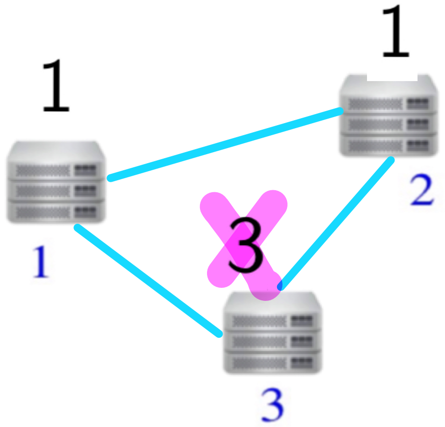


Entropy

Coded data

$$\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left\{ \begin{array}{ccc} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 3 \\ 1 & 0 & 0 \\ 1 & 1 & 2 \end{array} \right\} \end{array} \longrightarrow \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left(\begin{array}{ccc} 1 & 1 & 3 \end{array} \right) \end{array} \longrightarrow$$

Data network



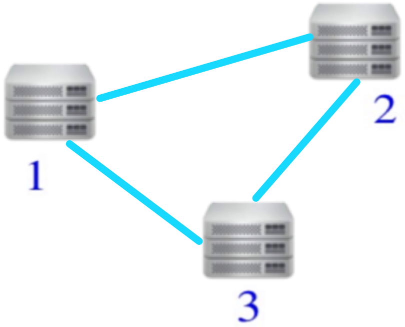
Entropy

- Entropy can be used to measure the amount of information in a data set.
- $\mathit{Entropy}(\{1\}) < \mathit{Entropy}(\{1,2\}) < \mathit{Entropy}(\{1,2,3\}) = \mathit{Entropy}(\{2,3\})$

Coded data

$$\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left\{ \begin{array}{ccc} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 3 \\ 1 & 0 & 0 \\ 1 & 1 & 2 \end{array} \right\} \end{array}$$

Data network



Entropy

Coded data

$$\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left\{ \begin{array}{ccc} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 1 & 1 & 3 \\ 1 & 0 & 0 \\ 1 & 1 & 2 \end{array} \right\} \end{array}$$

Probabilities

$$p(012) = 0.1,$$
$$p(001) = p(113) = p(100) = 0.2,$$
$$p(112) = 0.3$$

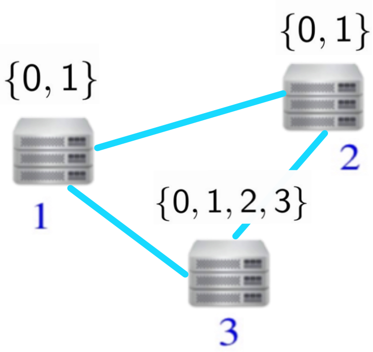
- Entropy of data set A: $H(A) = -\sum_{x \in A} p(x) \log(p(x))$.
- $H(\{1\}) = -(p(0)\log_2(p(0)) + p(1)\log_2(p(1)) = -(0.3\log_2(0.3) + 0.7\log_2(0.7) \approx 0.88$
- $H(\{1,2\}) \approx 1.76$
- $H(\{2,3\}) = H(\{1,2,3\}) \approx 2.25$

Entropy

- Capacity of a set of data nodes A in the data network: $\mathcal{C}(A) = \log(|A|)$

- $\mathcal{C}(\{1\}) = \log_2(|\{0,1\}|) = 1$
- $\mathcal{C}(\{1,2\}) = 2$
- $\mathcal{C}(\{2,3\}) = 3$
- $\mathcal{C}(\{1,2,3\}) = 4$

Data network



Entropy

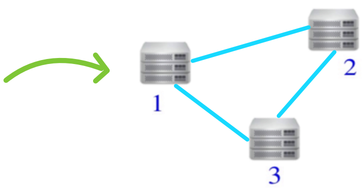
- Rate of a data set A: $\mathcal{R}(A) = \frac{Entropy(A)}{Capacity(A)} = \frac{H(A)}{\mathcal{C}(A)}$
- Rate gives a measure on the amount of redundancy in a data set, the smaller the rate, the greater the redundancy and vice versa.

$$
\begin{aligned}
\mathcal{R}(\{1\}) &= \frac{H(\{1\})}{\mathcal{C}(\{1\})} \approx \frac{0.88}{1} = 0.88 \\
\mathcal{R}(\{1,2\}) &= \frac{H(\{1,2\})}{\mathcal{C}(\{1,2\})} \approx \frac{1.76}{2} = 0.88 \\
\mathcal{R}(\{2,3\}) &= \frac{H(\{2,3\})}{\mathcal{C}(\{2,3\})} \approx \frac{2.25}{3} = 0.75 \\
\mathcal{R}(\{1,2,3\}) &= \frac{H(\{1,2,3\})}{\mathcal{C}(\{1,2,3\})} \approx \frac{2.25}{4} \approx 0.56
\end{aligned}
$$

Coded data

$$
\begin{array}{c}
\begin{array}{ccc} 1 & 2 & 3 \end{array} \\
\left\{ \begin{array}{ccc}
0 & 1 & 2 \\
0 & 0 & 1 \\
1 & 1 & 3 \\
1 & 0 & 0 \\
1 & 1 & 2
\end{array} \right\}
\end{array}
$$

Data network



Entropy

General block coded data:

- ambient space: $\mathbb{A}^E = \mathbb{A}_1 \times \ldots \times \mathbb{A}_n$ ($\mathbb{A}_i$ is finite)
- coded data : $C \subseteq \mathbb{A}^E$
- probability distribution on C: $p_C = \{p_{\mathbf{c}}\}_{\mathbf{c} \in C}$ ($p_{\mathbf{c}} > 0$ and $\sum_{\mathbf{c} \in C} p_{\mathbf{c}} = 1$)

Some properties

The following properties holds for any $A, B \subseteq E$ and $e \in E$,

$$
\begin{array}{ll}
\textit{(H1)} & 0 \leq H(A), \\
\textit{(H2)} & A \subseteq B \quad \Rightarrow \quad H(A) \leq H(B), \\
\textit{(H3)} & H(A) + H(B) \geq H(A \cap B) + H(A \cup B), \\
\textit{(C1)} & \mathcal{C}(e) > 0, \\
\textit{(C2)} & \mathcal{C}(e) \geq H(e), \\
\textit{(C3)} & \mathcal{C}(A) = \sum_{a \in A} \mathcal{C}(a).
\end{array}
$$

Entropy

Block linear coded data with uniform probability distribution:

- ambient space: $\mathbb{A}^E = \mathbb{F}_q^n$
- coded data : C is a subspace of $\mathbb{F}_q^n$
- probability distribution on C: $p_{\boldsymbol{c}} = \frac{1}{|C|}$ for all $\boldsymbol{c} \in C$

Some properties

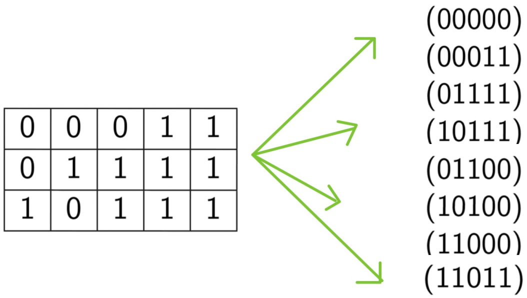
(R1)-(R3), (C1)-(C3) and the following properties holds for any $A \subseteq E$, using $\log_q$,

$$\begin{aligned} &(R4) \quad H(A) \in \mathbb{Z}, \\ &(C4) \quad \mathcal{C}(A) = |A|. \end{aligned}$$

Entropy

Block linear coded data C with uniform probability distribution can be represented by a matrix S.

- $C =$ rowspace(S)
- Entropy: $H(A) =$ rank(submatrix(A))



0	0	0	1	1
0	1	1	1	1
1	0	1	1	1

(00000)
(00011)
(01111)
(10111)
(01100)
(10100)
(11000)
(11011)

- $C =$ span((00011), (01111), (10111)) is a subspce of $\mathbb{F}_2^5$
- $H(\{1, 2, 3\}) =$ rank(submatrix({1,2,3}) $= 2$

Matroids and polymatroids

Definition

- A (finite) polymatroid $P = (\rho, E)$ is a pair consisting of
  - A finite set E.
  - A (rank) function $\rho : 2^E \to \mathbb{R}$ such that for all $X, Y \subseteq E$:

    (R1) $\rho(\emptyset) = 0$,

    (R2) $X \subseteq Y \Rightarrow \rho(X) \leq \rho(Y)$,

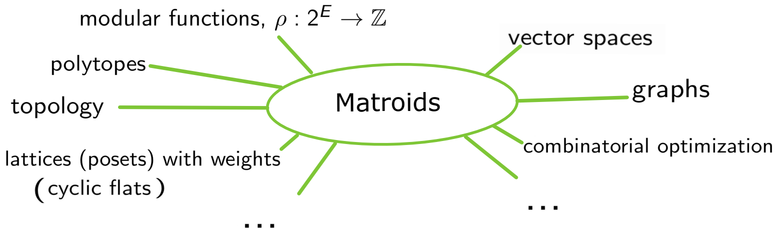
    (R3) $\rho(X) + \rho(Y) \geq \rho(X \cup Y) + \rho(X \cap Y)$.
- A matroid is a polymatroid which additionally satisfies the following two conditions for all $X \subseteq E$:

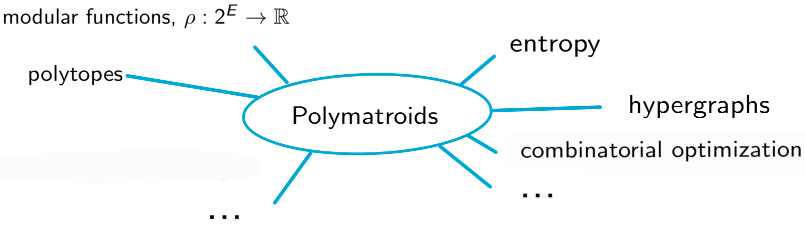
(R4) $\rho(X) \in \mathbb{Z}$,

(R5) $\rho(X) \leq |X|$.

Why using matroids and polymatroids?

- Axiomatic theories in algebraic combinatorics.
- Links to several different areas in mathematics, e.g. linear algebra, entropy, graph theory, hypergraph theory, combinatorial optimization, algebraic geometry, topology, ...
- Capture several properties of many mathematical objects, so called matrodial or polymatrodial properties.


modular functions, $\rho : 2^E \to \mathbb{Z}$
vector spaces
polytopes
Matroids
graphs
topology
combinatorial optimization
lattices (posets) with weights
(cyclic flats)
...
...


modular functions, $\rho : 2^E \to \mathbb{R}$
entropy
polytopes
Polymatroids
hypergraphs
combinatorial optimization
...
...

Matrodial and polymatrodial properties

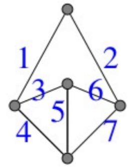
Matrodial

Matrices

1	2	3	4	5	6	7
1	0	0	0	0	1	1
0	1	0	0	0	1	1
0	0	1	0	1	1	0
0	0	0	1	1	0	1

$\rho(3,4,5) = 2$ (rank of submatrix)

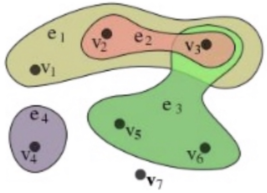
Graphs



$\rho(3,4,5,6,7) = 3$ ($|\{$largest non-cyclic subgraph$\}|$)

Polymatrodial

Hypergraphs



$\rho(e_2, e_3) = 4$ ($|\{$vertices in subhypergraph$\}|$)

Entropy

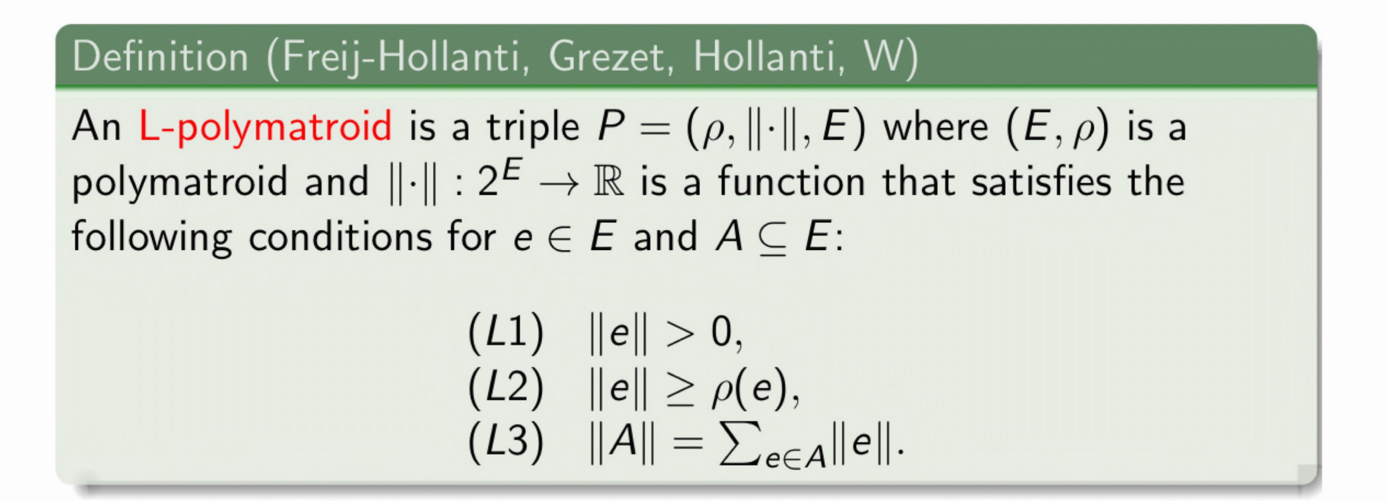
$$\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \end{array}$$
$$\left\{\begin{array}{ccccccc} 0 & 1 & 0 & 5 & 3 & 3 & 3 \\ 0 & 1 & 0 & 5 & 4 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 & 5 & 5 \\ 1 & 3 & 4 & 5 & 0 & 0 & 0 \\ 1 & 4 & 5 & 0 & 0 & 0 & 0 \end{array}\right\}$$

$\rho(1,2) = \frac{2}{5}\log_6(\frac{5}{2}) + \frac{3}{5} + \log_6(\frac{5}{3}) \approx 0.74.$
(joint entropy of subset)

Matroids and polymatroids

- Matroids were introduced independently by Hassler Whitney and Takeo Nakasawa in the 1930s.
- The interaction between matroid theory and other areas of mathematics has recently made great progress in several areas, e.g. Hodge theory (June Huh, ...), Stable polynomials (Petter Bränden, ...), tropical geometry (Erik Katz, ...), algebraic geometry, representation theory, topology, ...
- June Huh received the 2022 Fields Medal for having found striking connections between algebraic geometry and combinatorics, and among others this solved central problems in combinatorics that had been unsolved for decades, e.g. in matroid theory.
- Polymatroids were introduced by Jack Edmonds in 1970, and have especially proven to be useful in combinatorial optimization.

L-polymatroids



Definition (Freij-Hollanti, Grezet, Hollanti, W)

An L-polymatroid is a triple $P = (\rho, \|\cdot\|, E)$ where (E, ρ) is a polymatroid and $\|\cdot\| : 2^E \to \mathbb{R}$ is a function that satisfies the following conditions for $e \in E$ and $A \subseteq E$:

$$\begin{array}{ll} (L1) & \|e\| > 0, \\ (L2) & \|e\| \geq \rho(e), \\ (L3) & \|A\| = \sum_{e \in A} \|e\|. \end{array}$$

- L-polymatroids generalize concepts on matroids that polymatroids do not, e.g. duality, minors, cyclic flats.
- L-polymatroids capture concepts on mathematical objects which polymatroids don't, e.g. the capacity of a set of data nodes in a data network.

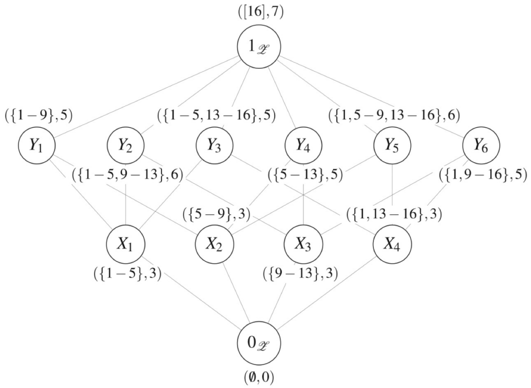
Cyclic flats

Definition (Freij-Hollanti, Grezet, Hollanti, W)

Let $P = (\rho, \|\cdot\|, E)$ be an L-polymatroid and $A \subseteq E$. Then

- A is a flat if $\rho(A) < \rho(A \cup e)$ for all $e \in E - A$.
- A is a cyclic set if $\rho(A) - \rho(A - a) < \|a\|$ for all $a \in A$.
- A is a cyclic flat if A is a flat and a cyclic set.
- The collection of cyclic flats is denoted by $\mathcal{Z}$.

1	2	3	4	5	6	7	8	9	10	11	12
1	1	0	0	0	0	0	0	0	0	1	3
0	2	1	0	0	1	0	0	0	0	1	3
0	3	0	1	0	3	0	0	0	0	1	3
0	0	0	0	1	2	0	0	0	0	1	3
0	0	0	0	0	0	1	0	1	0	1	3
0	0	0	0	0	0	0	1	2	0	1	3
0	0	0	0	0	0	0	0	3	1	1	1



Cyclic flats

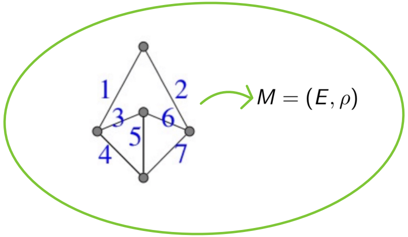
Results (Freij-Hollanti, Grezet, Hollanti, W)

Let $P = (\rho, \|\cdot\|, E)$ be an L-polymatroid and $A \subseteq E$, then

- $(\mathcal{Z}, \subseteq)$ *is a lattice.*
- $\rho(A) = \min\{\rho(Z) + \|A\| - \|A \cap Z\| : Z \in \mathcal{Z}\}$.
- E, $\{\|e\| : e \in E\}$ *and* $\{\rho(Z) : Z \in \mathcal{Z}\}$ *defines* P.
- $(\{Z \in: \rho(A) = \rho(Z) + \|A\| - \|A \cap Z\|\}, \subseteq)$ *is a sublattice of* $(\mathcal{Z}, \subseteq)$.

Characterization of classes of matroids

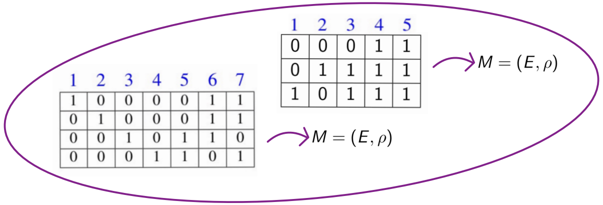
Graphical matroids



$M = (E, \rho)$

- Being able to characterize different classes of matroids that can be associated with different objects is generally very difficult.
- Almost all matroids are nonrepresentable.
- Rota's conjecture: All $\mathbb{F}_q$-representable matroids can be classified i.a. a finite list of prohibited minors. An outline of a proof for Rota's conjecture has been published, but not the entire proof.

$\mathbb{F}_2$-representable matroids



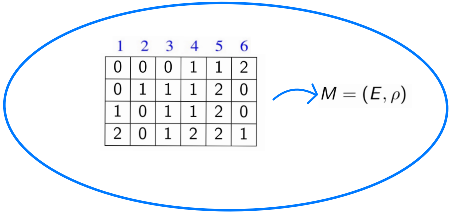
$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \hline 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \end{array}$$

$M = (E, \rho)$

$$\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{array}$$

$M = (E, \rho)$

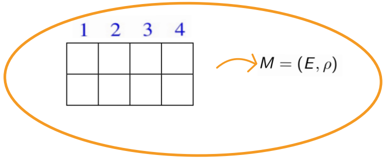
$\mathbb{F}_3$-reprerentable matroids



$$\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 1 & 2 & 0 \\ 1 & 0 & 1 & 1 & 2 & 0 \\ 2 & 0 & 1 & 2 & 2 & 1 \end{array}$$

$M = (E, \rho)$

$\mathbb{F}_q$-representable matroids



$$\begin{array}{|c|c|c|c|} 1 & 2 & 3 & 4 \\ \hline & & & \\ \hline & & & \\ \hline \end{array}$$

$M = (E, \rho)$

Characterization of classes of matroids

Results (Freij-Hollanti, Grezet, Hollanti, W)

- *A characterization of* $\mathbb{F}_2$*-representable matroids via cyclic flats.*
- *A first step in characterizing ”cyclic flats” of* $\mathbb{F}_q$*-representable matroids in general by finding some forbidden structures on the lattice of cyclic flats.*

Characterization of classes of matroids

Fundamentals:

- ambient space: $\mathbb{A}^E = \mathbb{A}_1 \times \ldots \times \mathbb{A}_n$ ($\mathbb{A}_i$ is finite)
- coded data : $C \subseteq \mathbb{A}^E$
- probability distribution on C: $p_C = \{p_{\mathbf{c}}\}_{\mathbf{c} \in C}$ ($p_{\mathbf{c}} > 0$ and $\sum_{\mathbf{c} \in C} p_{\mathbf{c}} = 1$)

Definition (Freij-Hollanti, Grezet, Hollanti, W)

The entropic L-polymatroid associated to C $\subseteq \mathbb{A}^E isP = (\rho, \|\cdot\|, E)$, where

$$\rho(A) = H_C(A) \text{ and } \|A\| = \log(|\mathbb{A}^A|).$$

- Informally, for entropic L-polymatroids, coded data A is a cyclic flat if adding a data node e increases the amount of information, and deleting a node a results in a possible loss of information which is less than the maximum amount of information $\|a\|$.

Different classes of coded data

- Ambient spaces, $\mathbb{A}^E = \mathbb{A}_1 \times \ldots \times \mathbb{A}_n$:
  - equicardinal alphabets ($|\mathbb{A}_i| = |\mathbb{A}_j|$ for all i, j),
  - non-equicardinal alphabets ($|\mathbb{A}_i| \neq |\mathbb{A}_j|$ for some i, j),
  - $\{\mathbb{A}^E\} \supseteq \{\mathbb{A}^E$ is a group$\} \supseteq \{\mathbb{A}^E$ is an Abelian group$\} \supseteq \{\mathbb{A}^E$ is a module$\} \supseteq \{\mathbb{A}^E = R^n$ for some finite ring $R\} \supseteq \{\mathbb{A}^E = R^n$, R is Frobenius$\} \supseteq \{\mathbb{A}^E = \mathbb{F}_q^n\} \supseteq \{\mathbb{A}^E = \mathbb{F}_2^n\}$.
- Codes, C:
  - $\{C \subseteq \mathbb{A}^E\} \supseteq \{$group codes$\} \supseteq \{$Abelian group codes$\} \supseteq \{$linear codes$\} \supseteq \{R$-vector-linear codes$\} \supseteq \{R$-linear codes$\}$
- Probability distribution, p_C:
  - $\{p_{\boldsymbol{c}} > 0, \sum_{\boldsymbol{c} \in C} p_{\boldsymbol{c}} = 1$ for $\boldsymbol{c} \in C\} \supseteq \{$uniform distribution: $p_{\boldsymbol{c}} = \frac{1}{|C|}\}$

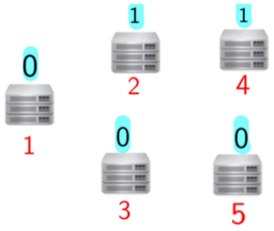
- A code C is Quasi-uniform if $H(A) = \log(|C(A)|)$ for all $A \subseteq E$.

  (Quasi-uniform codes can be considered as codes with maximal amount of information in comparison with the sizes of its code puncturing.)

Distributed storage

- Distributed storage is a technique for storing data on multiple storage devices that are interconnected in a network.
- The data is distributed over several units instead of one.
- Benefits of distributed storage are
  - ability to retrieve large amounts of data quickly and reliably.
  - high availability and scalability.
- Examples of applications: peer-to-peer networks, cloud-based storage services and data centers.

$$C = \left\{\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 0 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 2 & 2 & 1 \end{array}\right\} \longrightarrow$$

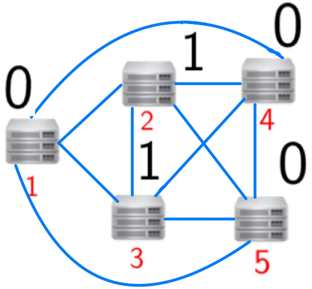


Distributed storage

- Desirable characteristics of a distributed storage system:
  - Efficient local repair of faulty storage nodes.
  - High global reliability.
  - Low storage excess.
  - High availability.
  - Effective hierarchy structure.
  - Small size of the coding alphabet.
- The properties given above can be measured with different parameters.
- There is a tradeoff between these parameters.

1	2	3	4	5
0	0	0	1	1
0	1	1	1	1
1	0	1	1	1

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ (0 & 1 & 1 & 0 & 0) \end{array}$$



Distributed storage

- Homogeneous linear distributed storage systems can be modeled by matrices over a finite fields.
- General heterogeneous distributed storage systems can be modeled by general coded data and entropy.
- Why general heterogeneous distributed storage systems may be preferable to homogeneous linear distributed storage systems:
  - The data storage devices can have different storage capacities.
  - Different data can have different probabilities to be stored.
  - Has the ability to achieve better values with respect to previously mentioned parameters in comparison to linear systems.
- Linear systems can be analyzed via matroids.
- General systems can be analyzed via entropic L-polymatroids and are generally more difficult to analyze and construct than linear ones.

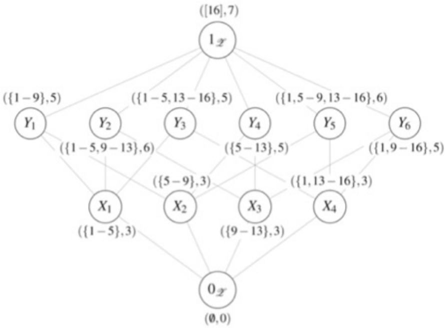
Distributed storage

Results (Freij-Hollanti, Grezet, Hollanti, W)

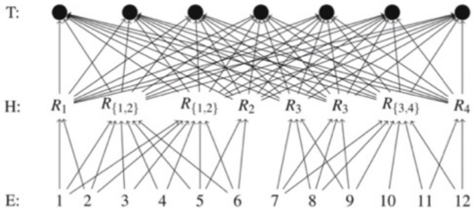
- *Introduced and developed the theory of L-polymatroids and entropic L-polymatroids to be able to use ”cylic flats” to analyze linear and general DSS.*
- *Limits for the various parameters given earlier*
- *Structures and constructions of good linear and general DSS.*

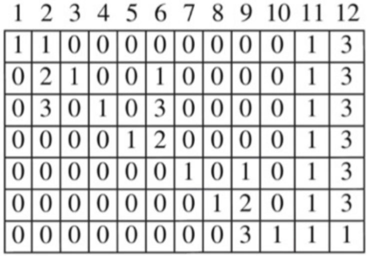
Distributed storage

cyclic flats



Gammoid





1	2	3	4	5	6	7	8	9	10	11	12
1	1	0	0	0	0	0	0	0	0	1	3
0	2	1	0	0	1	0	0	0	0	1	3
0	3	0	1	0	3	0	0	0	0	1	3
0	0	0	0	1	2	0	0	0	0	1	3
0	0	0	0	0	0	1	0	1	0	1	3
0	0	0	0	0	0	0	1	2	0	1	3
0	0	0	0	0	0	0	0	3	1	1	1

Data center

Distributed storage

Theorem (Freij-Hollanti, Grezet, Hollanti, W)

Let $P = (\rho, \|\cdot\|, E)$ be the associated entropic L-polymatroid of the coded data $C \subseteq \mathbb{A}^E$, (w.l.o.g. assume $\emptyset, E \in \mathcal{Z}$). Then

- *length:* $n = \|E\|$
- *rank:* $k = \rho(E)$
- *rate:* $\mathcal{R} = \frac{k}{n}$
- *failure tolerance:* $d = |E| - \max\{|Z| + \gamma(Z) : Z \in \mathcal{Z} - E\}$, *where* $\gamma(Z) = \max\{|A| : A \subseteq E - Z \text{ and } \rho(Z) + \|A\| < \rho(E)\}$.

Theorem (Freij-Hollanti, Grezet, Hollanti, W)

Let let $(\rho, \|\cdot, E\|)$ be an (n, k, d, r, δ, t)-L-polymatroid. Then

$$\begin{array}{ll} (i) & t(\delta - 1) + 1 \leq d \leq |E| - |\alpha|, \\ (ii) & \mathcal{R} = \frac{k}{n} \leq \frac{\beta}{\|E\|}, \end{array}$$

where α depends on $\|\cdot\|$ and (k, r, δ, t) and β on $\|\cdot\|$ and (d, r, δ, t).

Theorem (Freij-Hollanti, Grezet, Hollanti, W)

Let $P = (\rho, \|\cdot\|, E)$ be an (n, k, d, r, δ, t)-L-polymatroid that achieves the upper bound (i) or (ii) given above. Then,

$$\begin{array}{ll} (a) & R \text{ union of repair groups } \rho(R) < k \quad \Rightarrow \quad R \text{ is a cyclic flat}, \\ (b) & \text{if } Z \text{ is a cyclic flat, then } \frac{k - \rho(Z)}{\|E - Z\|} \text{ must satisfy (ii) on } E - Z. \end{array}$$