Group Testing as Coding over the Binary Semifield via Residuation Theory

joint work with Cornelia Roessing

Marcus Greferath

Aalto University

Rijeka and Dublin, July 2021

Structure of the Presentation

The binary field $\mathbb{F}_2$

- Students of Sciences and Engineering are nowadays aware of the set $\mathbb{F}_2 = \{0, 1\}$ forming an algebraic structure known as *field*, provided we use the following operations:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

- Often $+$ runs under the name xor for *exclusive or* in contrast with the *inclusive or*.

- In fact, this field plays a dominant role in disciplines like Algebraic Coding Theory and Cryptography, to mention just a few.
- It gives rise to an entire powerful Linear Algebra relying on Gaussian Elimination and matrix inversion.
- Enriched with a distance function induced by the Hamming weight: $w_H : \mathbb{F}_2 \longrightarrow \mathbb{N}$ with

$$w_H(x) = \begin{cases} 1 & : \quad x \neq 0 \\ 0 & : \quad x = 0 \end{cases}$$

(and its additive extension) we enter Combinatorial Linear Algebra, another fancy term for Coding Theory.

The binary semifield $\mathbb{B}_2$

- This talk was motivated by the ongoing CoVID-19 pandemic and a mechanism that is called group testing.
- To get prepared, we will look into the *inclusive or* instead of the exclusive or xor.

$+$	0	1
0	0	1
1	1	1

$\cdot$	0	1
0	0	0
1	0	1

This semifield is at the same time the smallest non-trivial Boolean lattice, and it is natural to expect elements from order theory entering the game.

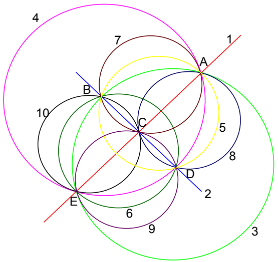
Group Testing

- Most pandemics initially spread slowly and show their true nature of growth (exponential) only at a later stage.
- At this stage the necessary counter measures have typically already become effectless.
- In the very early phase, particularly if a virus is novel, testing techniques might be complicated or costly.
- The goal is therefore to exploit testing resources efficiently.
- Around 1943, Dorfman [6] divised a technique, that is running under the name *group testing*.
- In the seventies of the previous century, a matrix-driven formalization of group testing took the lead.

Group Testing

- The underlying idea is that each test is used for a pool of specimen from different participants, while the specimen of every participant is spread over different pools.
- Dorfman observed, that it was possible to identify a small number of infected participants in a larger batch without having to medically check each individual.
- To organize group tests it is useful to have a table (a binary matrix) for tests and participants showing which participant's specimen is contained in which pool.
- We assume that we have only one round of testing, an approach known as *non-adaptive* group testing.

A small example



	1	2	3	4	5	6	7	8	9	10
A	1	0	1	1	1	0	1	1	0	0
B	0	1	1	0	1	1	1	0	0	1
C	1	1	0	0	0	0	1	1	1	1
D	0	1	0	1	1	1	0	1	1	0
E	1	0	1	1	0	1	0	0	1	1

- The table is the incidence matrix of the structure on the left side.
- This scheme allows to identify a single infected participant out of 10 by using only 5 tests.
- For more than one infected participant, the scheme will however fail.

- For example, assuming that exactly participant 3 is infected, we find that:

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

- This means, that test A, B, and E will be positive, while the remaining tests will show a negative result.
- The resulting vector $[1, 1, 0, 0, 1]^T$ uniquely determines participant 3 as infected.

Goals of Group Testing

- This small scheme shows at least the principle of identifying one out of 10 with just 5 tests.
- What to do, if we know that there are 2 infected participants in a group of 20 ?
- Dividing the 20 into two batches of 10 participants is possible, but suboptimal in principle!
- **Goals of Group Testing:**
  - For a group of size n that contains up to d infected individuals, divise a scheme of k tests that allow identification of all involved infected individuals.
  - Representing the scheme by the binary $k \times n$-matrix H solve the syndrome decoding problem $Hx = s \in \mathbb{B}_2^k$.

Mathematical Modelling of Group Testing

- **Definition:** Let n, k, and d be natural numbers with $d, k \leq n$. A group testing scheme is a $k \times n$-matrix H over the semifield $\mathbb{B}_2$ satisfying the following property:

  (S) The restriction of the mapping

  $$H : \mathbb{B}_2^n \longrightarrow \mathbb{B}_2^k, \; x \mapsto Hx$$

  to the disk of Hamming radius $d - 1$ centered in the origin is an injection.

- **Remark:** Group testing schemes are the check matrices of a Coding Theory over $\mathbb{B}_2$. The test result vector takes the role of the syndrome, and the all-0-word is the information transmitted, but then distorted by the infected cases.

- We suggest to refer to H as an $[n, k, d]$ group testing scheme.
- As a matter of fact, maximizing d and minimizing k are conflicting goals.
- For given n, d with $d \leq n$ an $[n, k, d]$ group testing scheme is called *optimal*, if for every $[n, k', d]$ group testing scheme there holds $k' \geq k$.
- **Further Goals:** Construct optimal group testing schemes, and develop and implement efficient syndrome decoders for them.

Residuated Mappings

Definition

Let $(A, \leq)$ and $(B, \leq)$ be two partially ordered sets. For mappings $f : A \longrightarrow B$ and $g : B \longrightarrow A$, the pair (f, g) is called a *residuated pair*, if there holds

$$f(x) \leq y \iff x \leq g(y), \quad \text{for all } x \in A \text{ and } y \in B.$$

A few points can be easily taken from [blyth].

1: $f : A \longrightarrow B$ may be called a *residuated mapping*, if there is $g : B \longrightarrow A$, such that (f, g) is a residuated pair. The mapping g is then uniquely determined by f. Dually, f is uniquely determined by g which is called the *residual* of f. It is usually denoted by f^+.

Residuated Mappings

2: f and f^+ are monotone mappings, and there holds $f^+ \cdot f \geq \mathrm{id}_A$ and $f \cdot f^+ \leq \mathrm{id}_B$. Conversely, if two monotone mappings f and g satisfy $g \cdot f \geq \mathrm{id}_A$ and $f \cdot g \leq \mathrm{id}_B$, then they will form a residuated pair.

3: $f \cdot f^+ \cdot f = f$ and $f^+ \cdot f \cdot f^+ = f^+$, and the sets $C := \{f^+ \cdot f(x) \mid x \in A\}$ (of closed elements in A) and $K := \{f \cdot f^+(y) \mid y \in B\}$ (of kernel elements in B) indeed form a closure/kernel system in their respective spaces A and B.

4: The according closure and kernel operators on A and B are induced by $h := f^+ \cdot f$ and $k := f \cdot f^+$, respectively. The mappings $f|C$ and $f^+|K$ are mutually inverse.

A Type of Inversion

5: If A and B are complete lattices, then f is residuated if and only if $f(\sum X) = \sum f(X)$ for all $X \subseteq A$. Accordingly g is a residual mapping iff $g(\prod Y) = \prod g(Y)$ for all $Y \subseteq B$.

6: Any residuated mapping on $\mathbb{B}_2^n \longrightarrow \mathbb{B}_2^k$ can be represented by a $k \times n$ matrix with entries in $\mathbb{B}_2$. The representation of its residual mapping is the subject of the following theorem.

Theorem

Let H be a $k \times n$-matrix describing a residuated mapping $\mathbb{B}_2^n \longrightarrow \mathbb{B}_2^k$. Let N_n and N_k denote the negation on $\mathbb{B}_2^n$ and $\mathbb{B}_2^k$, respectively. Then the residual mapping of H is $H^+ : \mathbb{B}_2^k \longrightarrow \mathbb{B}_2^n,\ y \mapsto N_n H^T N_k(y)$.

A Decision Scheme

- The above theorem practically yields a decision scheme $H^+ : \mathbb{B}_2^k \longrightarrow \mathbb{B}_2^n$.
- It works in the error-free syndromes case, because all occurring syndromes in $\mathbb{B}_2^k$ are then kernel elements, as described above.
- It will be correct, if the infection pattern in $\mathbb{B}_2^n$ is a closed element.
- If this is not the case, it will return the closure of the actual infection pattern, and hence in the worst case only false positive results.
- The natural interest is therefore in residuated mappings, where the all elements up to a given Hamming weight are closed elements.

Incidence Structures

- Recall that we used $k \times n$-matrices over $\mathbb{B}_2$ in order to set up a group testing scheme.
- In our initial motivating examples, they came from geometric structures.
- We will take this to a more rigid treatment.
- **Definition:** An *incidence structure* is a pair (P, B), where P is a set of points, and B, called the set of blocks, is a subset of 2^P.
- If a point $p \in P$ is contained in the block $C \in B$, then we say that p is *incident* with C.

Incidence Structures and Partial Linear Spaces

- If (P, B) is an incidence structure with $|P| = v$ and $|B| = b$. A binary matrix $M \in \mathbb{B}_2^{v \times b}$ is called an *incidence matrix* for (P, B), if its rows are labelled by the points in P, while its columns are labelled by the blocks in B, such that

$$M_{p,C} = \begin{cases} 1 & : \quad p \in C, \\ 0 & : \quad \text{otherwise.} \end{cases}$$

- An incidence structure (P, B) is called a *partial linear space* of order (s, t) if the following axioms hold:
  - every line is incident with $s + 1$ points, and every point is incident with $t + 1$ lines,
  - two different lines can intersect in at most one point, and two different points are connected by at most one line.

Designs

- For integers $0 \leq t \leq k \leq v$, a t*-design* is a set B of k-element subsets (blocks) of a v-element set P, such that every t-element subset of P is contained in the same number λ_t of blocks of B.
- If this is the case, then B will be referred to as a t-design with parameters (v, k, λ_t).
- For $t = 2$ these designs are known as a *balanced incomplete block design*.
- Every t-design is at the same time an s-design for all $0 \leq s \leq t$. The parameter λ_s can be computed from the other parameters of the design.

Main Technical Fact

Theorem (Identifiability of Blocks)

Let B be a $t-(v,k,1)$ Steiner System, and let $C_1,\ldots C_m$ denote a collection of m distinct blocks in B. If $k>(t-1)m$, then the following hold:

(a) $|C_1\cup\cdots\cup C_m| \;\geq\; mk-(t-1)\cdot\binom{m}{2}$.

(b) *If $C\in B$ is a block with $C\subseteq C_1\cup\cdots\cup C_m$ then $C=C_j$ for some $1\leq j\leq m$ which means C is determined.*

Remark: This statement can be proved by relatively simple counting and induction.

Corollary (Injectivity of group testing matrix)

Let H be the incidence matrix of a $t-(v,k,1)$ Steiner System with b blocks. Then the restriction of the mapping

$$H : \mathbb{B}_2^b \longrightarrow \mathbb{B}_2^v, \; x \mapsto Hx$$

to the disk of Hamming radius $d-1$ centered in the origin is injective, provided $k > (t-1)d$

Remark: The preferred incidence structures taken for group testing, are those t-designs with *small* t. Partial linear spaces form a huge class of tactical configurations ($t = 1$).

Examples

- A particularly well understood class of partial linear spaces is that of the *generalized quadrangles* introduced by J. Tits.

- **Definition:** A partial linear space (P, B) of order (s, t) is called a *generalized quadrangle*, denoted by $GQ(s, t)$, if it does not contain triangles.

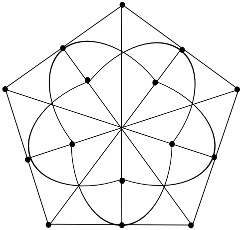


Figure: GQ(2,2)

Examples

- $GQ(s,t)$ has $(s+1)(st+1)$ points and $(t+1)(st+1)$ lines. This way, $GQ(2,4)$ has 27 points and 45 lines.
- This means (at least in theory), we obtain a group testing scheme that identifies up to 2 infected samples out of a batch of 45 using 27 tests.
- Here is the induced incidence matrix of $GQ(2,4)$.

$$\left[\begin{array}{ccccccccccccccccccccccccccccccccccccccccccccc}
1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&1&1&1&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&1&0&0&0&0&0&0&1&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&1&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&1&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&1&0&0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&1&0&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&1&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
1&0&0&1&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&1&1&1&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&1&0&1\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&1&0&1&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&1&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{array}\right]$$

False positives and false negatives

- What we have discussed so far was the error-free case.
- In the following we will consider two types of errors entering the syndrome.
- These are the so-called false positives and false negatives.
- We will of course model them by single bit flips, however note that they do not occur symmetrically.
- For the cheap antigen tests that you can purchase in the stores, it has been claimed that false negatives occur with probabilities up to 20%, while false positives are much rarer, making about 2%.

False positives and false negatives

- The above percentages describe the probabilities $P(\text{test pos} \mid \text{samp neg})$ as a false positive, and $P(\text{test neg} \mid \text{samp pos})$ as a false negative.
- For the applicant of the test, it is however much more interesting to obtain information about $P(\text{samp pos} \mid \text{test neg})$ and $P(\text{samp neg} \mid \text{test pos})$.
- These magnitudes can be related to each other which should remind us of the well-known channel forward and channel backward probabilities.
- It is known, that the distribution on the samples (here the prevalence σ as a probability) enters these relations.

False positives and false negatives

- If $P(\text{test pos} \mid \text{samp neg})$ and $P(\text{test neg} \mid \text{samp pos})$, are fixed, for example, the probability $P(\text{samp pos} \mid \text{test neg})$ will be higher under $\sigma = 90\%$ than when $\sigma = 1\%$.
- More-over, $P(\text{test pos} \mid \text{samp neg})$ assumes a possibly mixed sample, while $P(\text{samp pos} \mid \text{test neg})$ is mainly interested in the positiveness of the individual specimen.
- This means, that regarding the mentioned relationships, also the structure of the group testing scheme will play a role.

False positives and false negatives

- We have composed a little software tool in the *C* programming language in order to simulate the testing with various testing designs.
- We then fed the incidence matrices of the testing designs into the programme and ran a certain number of Monte-Carlo Simulations (typically 8192) in order to obtain an impression of the performance.
- It is clear, that sophisticated error correction mechanisms would be welcome to see the full quality of the test designs.
- In lack of such, we simply ran the *naked decoder* that I described in an earlier part of this presentation.

False positives and false negatives

- It turned out at the same time, that it makes definite sense to use as many tests as samples $k = n$, or even more of them $k \geq n$, should testing be very cheap but error-prone.
- For example, the incidence matrix of the $(7, 3, 1)$-Fano plane is of advantage when compared with the 7×7 identity matrix.
- This clearly is a (rather unexpected) extension of the original idea of group testing, where we try to minimize k under given n, because of the expensive nature of testing.
- It supports the idea of group testing just being a type of coding theory over $\mathbb{B}_2$.

Outlook

- This is work in progress. We have elaborated on the strong parallels between coding theory and group testing.
- Questions of further research may include the following:
  - Are there coding-type existence bounds for group testing? Singleten, Sphere-Packing, Elias, Varshamov, etc?
  - Optimality notions of group testing schemes depending on these existence bounds.
  - Construction of efficient decoding schemes for identification of infected individuals.
  - Are there Shannon-like theorems for the asymptotics?
  - Construction of infinite families of *asymptotically good* group testing schemes.

References I

V. Guruswami, A. Rudra, and M. Sudan
Essential Coding Theory
©V. Guruswami, A. Rudra, M. Sudan, 2018

The GAP Group.
GAP – Groups, Algorithms, and Programming,
Version 4.9.3, 2018,
https://www.gap-system.org

T. S. Blyth, and M. F. Janowitz
Residuation Theory.
1972,
Pergamon Press

References II

S. Mallapaty
The mathematical strategy that could transform coronavirus testing
Nature 583, 504-505 (2020)

G. Guglielmi
Fast coronavirus tests: what they can and cannot do
Nature 585, 496-498 (2020)

S. Ghosh et al.
Tapestry: A Single-Round Smart Pooling Technique for COVID-19 Testing
https://doi.org/10.1101/2020.04.23.20077727

References III

L. Mutesa et al.
A strategy for finding people infected with SARS-CoV-2: optimizing pooled testing at low prevalence
https://doi.org/10.1101/2020.05.02.20087924

R. Dorfman
The detection of defective members of large populations
J. Ann. Math. Statist. **14** (1943), pp. 436–440

Wikipedia
Semifields
https://en.wikipedia.org/wiki/Semifield

Thanks for your attention!