# An introduction to group testing

Oliver Johnson
O.Johnson@bristol.ac.uk
bristoliver.substack.com

Institute for Statistical Science, University of Bristol, UK

LMS Summer School Sheffield, 24th–25th July 2023

# Outline of the course

# Section 1:
# Introduction to group testing

# Toxic talk treat teaser

- Professor J is talking to students at Shedfield University.
- 7 plates of delicious post-lecture snacks.
- Professor J's evil nemesis, Dr X, has poisoned one of them.
- Whoever eats that snack will fall asleep for 24 hours.
- How to find the poisoned food, as efficiently as possible?
- Can pay any PhD student £10 to eat what we tell them.

# How to solve the mystery?

- One idea: pay 7 PhD students to eat one snack each ('individual testing')
- One will fall asleep.
- Will cost us £70.
- Better idea: use the following strategy (only costs £30).

| | Olives | Nuts | Bread sticks | Crisps | Dip | Cheese straws | Jelly |
|---|---|---|---|---|---|---|---|
| PhD 1 | ✓ | × | ✓ | × | ✓ | × | ✓ |
| PhD 2 | ✓ | ✓ | × | × | ✓ | ✓ | × |
| PhD 3 | ✓ | ✓ | ✓ | ✓ | × | × | × |

# Outcome

| | Olives | Nuts | Bread sticks | Crisps | Dip | Cheese straws | Jelly | |
|---|---|---|---|---|---|---|---|---|
| PhD 1 | ✓ | × | ✓ | × | ✓ | × | ✓ | 😴 |
| PhD 2 | ✓ | ✓ | × | × | ✓ | ✓ | × | 😮 |
| PhD 3 | ✓ | ✓ | ✓ | ✓ | × | × | × | 😴 |

- Strategy based on binary representation?
- Solution: breadsticks were poisoned.
- This strategy would always work.
- But what if more than one snack poisoned?

# More interesting problems

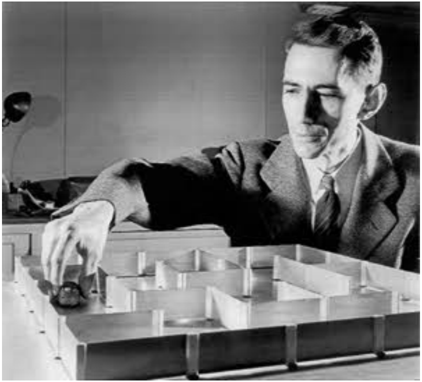
- What if you had 500 snacks, with 10 poisoned?
- How many students would we need?
- What should we get them to eat?
- How should we find the poisoned snacks?
- What if some students are immune to poison . . . or fall asleep anyway?
- This is **group testing**.

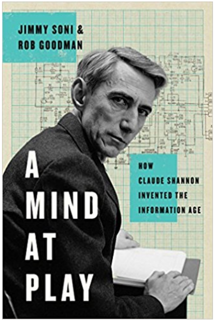
# What is group testing?

- It's a way of efficiently testing a large population for a rare disease.
- It's a toy model that throws up some interesting combinatorial problems.
- But also gets used in a variety of applied fields.
- Sometimes referred to as 'pooled testing'.
- Will give an introduction without too many pre-requisites.
- If you want more detail, see our survey: *Foundations and Trends in Communications and Information Theory* Vol. 15, No. 3–4, pp 196–392, 2019 (Joint with Matt Aldridge, Jon Scarlett)
- Free version at `arxiv:1902.06002`.

# Section 2:
# Shannon and information theory

# Claude Shannon (1916–2001)

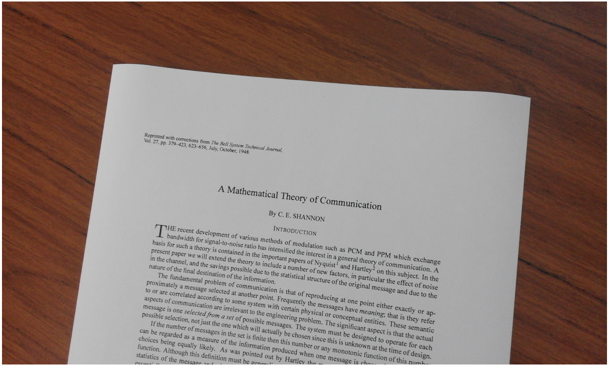





*It's rare to talk about maths and science as opportunities to revel in discovery. We speak, instead about their practical benefits – to society, the economy, our prospects for employment. STEM courses are the means to job security, not joy. Studying them becomes the academic equivalent of eating your vegetables – something valuable, and state sanctioned, but vaguely distasteful.*

– from *A Mind at Play* (2017) by Jimmy Soni and Rob Goodman

# Shannon's 1948 paper



Reprinted with corrections from *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the ...

- One of the most influential scientific papers ever.
- 110k citations and counting.
- Impact comparable with e.g. Einstein's work on relativity?
- Recommended movie *The Bit Player* (on YouTube).

# What did Shannon do?

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.

from *A Mathematical Theory of Communication (1948)*

- Boole had introduced binary arithmetic (0s and 1s).
- Shannon realised any information can be represented as series of these 'bits'.
- Understood we can compress information down to a limit (entropy).
- **Think about** "amount of stuff" (uncertainty) a message contains.
- Remove redundancy.
- **Key Idea: Predictable messages are compressible.**

# We all live in Shannon's world

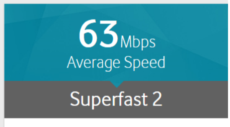
Phones

iPhone XS Deals on Contract - Get 10GB Extra Data | Vodafone
https://www.vodafone.co.uk › Pay monthly phones ▾
Buy now. Best iPhone **XS** deals on pay monthly in Gold, Space Grey or Silver. Specs of 64GB, 256GB & 512GB. Trade-in & save up to £240. Free 10GB data.

Hard Drives


5TB Toshiba HDWE150EZSTA X300, 3.5" HDD, SATA III - 6Gb/s, 7200rpm, 128MB Cache, 4.17ms, Shock Sensor, NCQ, Retail
**LN82058**

Broadband



# Information

- Shannon realised the key quantity is *information*.

## Definition

- Consider an event $A$ which happens with probability $p(A)$.
- If it happens, amount of information we gain from that is:

$$I(A) = -\log p(A).$$

- Take log to base 2, then unit is bits.
- An unlikely event occurring brings more information.
- Logs play nicely with independence: given independent events $A$, $B$:

$$p(A \cap B) = p(A)p(B).$$

- Means that information adds up from independent events:

$$I(A \cap B) = I(A) + I(B).$$

# Entropy definition

**Definition (Shannon 1948)**

Suppose we have outcomes $1, \ldots, n$ which occur with probability $\mathbf{p} = (p_1, \ldots, p_n)$ respectively. Then the *entropy*

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i .$$

- Amount of information we expect to gain from source of randomness.

# Coin toss example

- **Fair coin:**

$$\begin{aligned} \text{Heads with probability } 0.5 &\implies \text{See Heads, gain 1 bit,} \\ \text{Tails with probability } 0.5 &\implies \text{See Tails, gain 1 bit.} \end{aligned}$$

  Overall expect to gain $H = 0.5 \times 1 + 0.5 \times 1 = 1$ bit.

- **Biased coin:**

$$\begin{aligned} \text{Heads with probability } 0.11 &\implies \text{See Heads, gain 3.184 bit,} \\ \text{Tails with probability } 0.89 &\implies \text{See Tails, gain 0.168 bit.} \end{aligned}$$

  Overall expect to gain $H = 0.11 \times 3.184 + 0.89 \times 0.168 = 0.5$ bit.

# Maximising entropy

- **More unpredictable means bigger entropy**.
- Can show that if $n$ outcomes possible:
  1. in general: $0 \leq H(\mathbf{p}) \leq \log n$.
  2. entropy is maximised iff $p_i \equiv 1/n$ (uniform). In this case $H(\mathbf{p}) = \log n$.
  3. entropy is maximised iff $p_j \equiv 1$ for some $j$, $p_i = 0$ else (deterministic). In this case $H(\mathbf{p}) = 0$.
- Challenge: can you prove this?

# Channels and noise

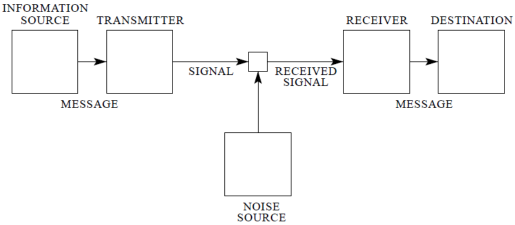



Fig. 1—Schematic diagram of a general communication system.

- Shannon also modelled noisy (imperfect) 'communications channels'.
- Imagine we send message $X = (0\ 0\ 1\ 0\ 1\ 0\ 0\ 1 \ldots)$
- Measurements inaccurate, environment has interference etc.
- Through noisy channel receive $Y = (0\ 1\ 1\ 0\ 1\ 0\ 0\ 0 \ldots) \neq X$
- Could model this as $Y = X \oplus Z$, where $Z$ is noise (randomly flip bits).
- If $Z$ IID Bernoulli($p$), this is binary symmetric channel.

# Coping with noise

- Shannon realised noise needn't be a problem.
- Can make messages longer (deliberately introduce redundancy).
- e.g. Naively just repeat message 3 times – call this 'rate $1/3$'.
- Even with some errors know what was (probably) sent.
- In fact, better strategies available.
- In general rate is 'proportion of bits that are message'.
- How big can rate be?

# Shannon capacity

- Shannon introduced 'capacity' $C$ of a channel, gave general formula.
- **Think about** "width of a pipe".
- Roughly speaking:
  1. **[Achievability]** For any rate $< C$, there is a strategy so the message gets through.
  2. **[Converse]** For any rate $> C$, no such strategy exists.
- **Key Idea: Some problems can't be solved.**

# Example: Binary symmetric channel

- In some cases can explicitly calculate the capacity.
- 'Memoryless' case relatively straightforward.
- Recall binary symmetric channel (bits independently flipped with probability $p$)
- Capacity

$$C = 1 - h(p) = 1 - \left(-p \log_2 p - (1-p) \log_2 (1-p)\right).$$

- What's the worst case?
- How does this relate to one-time pad?

# Section 3:
# Adaptive group testing algorithms

# Early history of the group testing problem

- 1942/3, US wanted to test all men joining army for syphilis.
- Tests not cheap, so overall potentially very expensive.
- Condition rare, so test outcomes known with high probability.
- Idea: pool blood from a group of people, **test it together**.
- If syphilis present in any blood sample, test outcome is positive.
- If no syphilis present, test outcome is negative.
- Obviously an idealization.
- Call this standard noiseless group testing.

# Variations on the model

- Can make assumptions more realistic:
- Binary vs non-binary outcomes
- Guaranteed ('zero error') vs high probability ('small error') recovery
- Partial vs full recovery
- Noiseless vs noisy
- Number of defectives known vs unknown?
- Fixed number of defectives vs defective with fixed probability (or e.g. Markov structure)

# Notation

- Refer to population as 'items' (say $N$ of them).
- Refer to infected people as 'defective' (say $K \ll N$ of them).
- Sometimes refer to defective set $\mathcal{K}$ (collection of infected people).
- Often write $\widehat{\mathcal{K}}$ for estimate of defective set produced by particular algorithm.
- Define error probability $\mathbb{P}(\mathrm{err}) = \mathbb{P}(\widehat{\mathcal{K}} \neq \mathcal{K})$.
- Do $T$ tests, hope to have low $\mathbb{P}(\mathrm{err})$ (high $\mathbb{P}(\mathrm{suc})$).
- Think about $N \to \infty$:
  1. **[Very sparse regime:]** $K$ is constant (or bounded as $K = O(1)$)
  2. **[Sparse regime:]** $K$ scales sublinearly as $K \sim N^\alpha$ for some *sparsity parameter* $\alpha \in [0, 1)$.
  3. **[Linear regime:]** $K$ scales linearly as $K \sim \beta N$ for $\beta \in (0, 1)$.
- Sparse regime turns out to be mathematically the most interesting.
- Linear regime perhaps easiest to motivate, but mathematically not interesting (individual testing generally wins).

# Dorfman testing

- Simplest strategies are adaptive (choice of tests depends on results of previous tests).

## Algorithm (Dorfman testing)

*Dorfman suggested two stage strategy:*

1. *Split large group of people into pools, test each pool together.*
2. *If negative, know everyone is disease-free.*
3. *If positive, suggested retesting individually.*

- Basic strategy for COVID-19 pooled testing.
- At low prevalence can save many tests.

# Run-time of Dorfman testing (see Feller Section IX.9)

- Suppose we split $N$ items into $\ell$ pools of size $N/\ell$.
- Procedure requires at most

$$T = \ell + K\frac{N}{\ell} \quad \text{tests.}$$

- Differentiating w.r.t. $\ell$, optimal pool size satisfies $1 - KN/\ell^2 = 0$.
- That is, the best choice $\ell = \sqrt{KN}$.
- Hence Dorfman testing needs at most $2\sqrt{KN}$ tests.
- Can we do better though?

# Better adaptive testing

- With large initial pools can be inefficient to retest individually.
- Binary search strategy is better.
- Hwang's algorithm (JASA 1972) essentially optimal in terms of number of tests.

# Binary splitting algorithm (Sobel and Groll)

## Algorithm (Binary splitting)

*Given a set $A$:*

1. *Initialize the algorithm with set $A$. Perform a single test containing every item in $A$.*
2. *If the preceding test is negative, $A$ contains no defective items, and we halt. If the test is positive, continue.*
3. *If $A$ consists of a single item, then that item is defective, and we halt. Otherwise, pick half of the items in $A$, and call this set $B$. Perform a single test of the pool $B$.*
4. *If the test is positive, set $A := B$. If the test is negative, set $A := A \setminus B$. Return to Step 3.*

This is guaranteed to find one defective item in $\lceil \log_2 |A| \rceil$ tests, if there is one.

# Binary splitting performance

## Theorem

*We can find all $K$ defectives in a set of $N$ items by repeated rounds of binary splitting, using a total of $K \log_2 N + O(K)$ adaptive tests, even when $K$ is unknown.*

## Proof.

Each round finds 1 defective. At the end of each round we can test the remaining items to see if there is another defective left. □

# Hwang's algorithm

## Algorithm

*Divide the $N$ items into $K$ subsets of size $N/K$ (rounding if necessary), and apply binary splitting to each subset in turn.*

This will find the defective set with certainty using $K \log_2(N/K) + O(K)$ adaptive tests.

# Magic number

- This performance is close to the 'magic number' of $T = \log_2 \binom{N}{K}$.
- For $K \ll N$:

$$\log_2 \binom{N}{K} \sim K \log_2(N/K).$$

- Argument: if we want an error probability of *exactly* zero, a simple counting argument based on the pigeonhole principle reveals that we require $T \geq \log_2 \binom{N}{K}$.
- There are only $2^T$ combinations of test results, but there are $\binom{N}{K}$ possible defective sets that each must give a different set of results.

**Theorem (BJA, *ISIT13*)**

*Choose defective set uniformly from $\binom{N}{K}$ sets of size $K$. For adaptive or non-adaptive standard noiseless group testing, any test design and any algorithm:*

$$\mathbb{P}(\mathrm{suc}) \leq \frac{2^T}{\binom{N}{K}}.$$

# Drawback of adaptive testing

- However, binary search can take many rounds.
- For COVID this could be close to generation time.
- Trade-off between tests saved and value of each one.
- PCR testing is parallel: need to declare whole test strategy in advance.
- How well can we do non-adaptively?

# Section 4:
# Applications of group testing

# Applications of group testing

- Group testing is used in a variety of applied settings.
- Specifically (see Section 1.7 of survey): biology, communications, information technology and data science.
- Early applications in industrial process monitoring (quality control).
- e.g. Test string of Christmas tree lights

# Applications in biology

- Already described applications to finding infection
- Works similarly for e.g. screening for cancer, finding rare DNA sequences.
- Sometimes can't or don't want to identify infected individuals, only count.
- e.g. mosquitos with malaria, HIV prevalence (privacy questions)
- Versions of algorithms specialised to counting defectives (see Section 5.3 of survey)
- Key result: if $K$ defectives, each item in test with probability $p$, test is positive with probability $1-(1-p)^K$.
- Hence using Bernoulli tests with $p(\ell) = 1 - 2^{-1/\ell}$, if $K \gg \ell$ then many more than half the tests positive.
- Similarly if $K \ll \ell$ then many fewer than half positive.

# Applications in communications

- *Multiple access channel*: several users communicating with single receiver.
- Group testing style protocols (see Wolf)
- People broadcast at times when they see fit (think of test matrix)
- Three outcomes: idle, success, collision (zero, one, multiple users sending signal at same time)
- *Cognitive radio*: similar, but frequencies not times
- Pick a frequency band to broadcast in, hope it's not occupied.
- *Network tomography*: infer structure of network by time taken for packets to go between particular pairs of computers.

# Applications in information technology

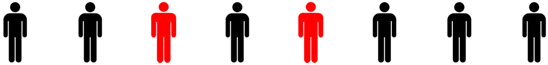
- *File comparison problem*: $N$ files, suppose $K$ of them have changed, want to know which.
- In advance can store result of hashes (pseudorandom function) of subsets of files.
- If file changes, so do corresponding hashes. Just like group testing.
- *Bloom filters*: Test whether a given item lies in a special 'interesting' set.
- *Database systems*: Want to work out which items in database are 'hot' (in high demand).
- Similarly, spotting heavy hitters (high-traffic flows) in Internet traffic.

# Applications in data science

- Group testing is example of search problem
- Similarities to ‘find a counterfeit coin’ problem
- In general area of sparse inference (c.f. compressed sensing)
- Also classification problems
- Crops up in various theoretical computer science problems (pattern matching, $K$-junta problem – if function depends on at most $K$ inputs).
- Quantum analogues of group testing exist.

# Section 5:
# Non-adaptive group testing algorithms

# Standard non-adaptive noiseless group testing



| | | | | | | | | Outcome **Y** |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | Positive |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | Positive |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | Negative |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Positive |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | Positive |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | Positive |

- Represent pooling strategy via binary test matrix X.
- Rows are tests, columns are people or 'items'.
- Put a 1 if item is in test.
- Red denotes being defective
- Can describe test outcomes as (here $\bigvee$ represents max or *OR*):

$$y_t = \bigvee_{i \in \mathcal{K}} x_{ti}.$$

# In practice

| ? | ? | ? | ? | ? | ? | ? | ? | Outcome **Y** |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | Positive |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | Positive |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | Negative |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | Positive |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | Positive |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | Positive |

- Want to infer list of defective items ('defective set' $\mathcal{K}$)
- Want as few tests as possible.
- Can separate choice of algorithm and design of matrix.

# Group Testing Algorithms

- Will discuss some algorithms that find defective set
- Mostly focus on computationally feasible algorithms ...
- ... with provable performance guarantees.
- Infer defective items from test matrix X and outcomes **Y** only.
- Negative test ... all items in it are non-defective.
- Positive test with one item in ... item is defective.

# Capacity and rate

## Definition

Channel coding analogy applies:

- In general represent defective set by binary random vector $\mathbf{U}$
- Magic number = entropy $H(\mathbf{U})$.
- Do $T$ tests.
- Rate of algorithm $= H(\mathbf{U})/T$ (bits learned per test).
- Hence (for binary outcomes) rate $\leq 1$.
- Write $\mathbb{P}(\text{suc})$ for success probability.
- Constant $C$ is (weak) group testing capacity if for any $\epsilon > 0$:
  1. **[Achievability]** there exist algorithms with rate $\geq C - \epsilon$ with $\mathbb{P}(\text{suc}) \to 1$,
  2. **[Weak converse]** all algorithms with rate $\geq C + \epsilon$ have $\mathbb{P}(\text{suc})$ bounded away from 1.

# COMP algorithm (Chan, Che, Jaggi, Saligrama 2011)

## Algorithm (COMP)

- *Algorithm has two stages:*
  1. *All items in a negative test are non-defective ...*
  2. *Mark all remaining items (Possible Defectives) as defective.*

# COMP example

| ? | ? | ? | ? | ? | ? | ? | |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 | 0 | Negative |
| 1 | 1 | 0 | 1 | 0 | 0 | 1 | Positive |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | Negative |
| 0 | 1 | 1 | 0 | 1 | 1 | 0 | Positive |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | Positive |

- First, look at negative tests.
- Test 1 is negative, so items 1,3,6 are non-defective.
- Test 3 is negative, so items 1,5 are non-defective.
- Hence items 2,4,7 are possible defectives (PDs).
- COMP algorithm: declare 2,4,7 to be defective.

# $\mathrm{COMP}$ performance

- $\mathrm{COMP}$ makes no errors in the first stage.
- Item labelled as non-defective must indeed be so.
- $\mathrm{COMP}$ estimate $\widehat{\mathcal{K}}$ contains no false negatives.
- Can take a long time for all non-defectives to appear in a negative test.
- Works fine, but not really optimal.

# DD algorithm (Aldridge, Baldassini, Johnson 2014)

## Algorithm (DD)

- *Algorithm has three stages (first stage is same as COMP):*
  1. *All items in a negative test are non-defective . . . leaving smaller set of possible defectives (PDs).*
  2. *Look for positive tests with exactly one PD item in . . . that item must be defective.*
  3. *Deal with others arbitrarily e.g. mark as non-defective.*

# DD example: Stage 2

| | | | |
|---|---|---|---|
| 0 | 0 | 0 | |
| 1 | 1 | 1 | Positive |
| 0 | 0 | 0 | |
| 1 | 0 | 0 | Positive |
| 0 | 1 | 0 | Positive |

- Restrict to submatrix corresponding to the PD set
- Test 4 is positive with one PD item in, so item 2 is defective.
- Test 5 is positive with one PD item in, so item 4 is defective.
- Up to this point, inference is definitely correct.

# DD example: Stage 3


| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0 | | 0 | | | 0 | |
| | 1 | | 1 | | | 1 | Positive |
| | 0 | | 0 | | | 0 | |
| | 1 | | 0 | | | 0 | Positive |
| | 0 | | 1 | | | 0 | Positive |

- Don't know about item 7.
- Arbitrarily, make it non-defective (sparsity grounds).
- Item 7 **masked** by defective items (important later).
- This is probably the obvious algorithm.
- However possible to prove performance bounds.

# SCOMP algorithm

- SCOMP is a greedy algorithm which builds on DD.

## Algorithm (SCOMP)

1. *Initialize $\hat{\mathcal{K}}$ as the estimate $\hat{\mathcal{K}}_{\mathrm{DD}}$ produced by DD algorithm, and declare any definitely nondefective items (items appearing in a negative test) to be nondefective. The other possible defectives are not yet declared either way.*
2. *Any positive test is called unexplained if it does not contain any items from $\hat{\mathcal{K}}$.*
3. *Add to $\hat{\mathcal{K}}$ the possible defective not in $\hat{\mathcal{K}}$ that is in the most unexplained tests, and mark the corresponding tests as no longer unexplained. (Ties may be broken arbitrarily.)*
4. *Repeat step 2 until no tests remain unexplained. The estimate of the defective set is $\hat{\mathcal{K}}$.*

# SSS algorithm

- Another (idealized) algorithm is $\mathrm{SSS}$.

**Algorithm (SSS)**

*Look for Smallest Set $\mathcal{L}$ that 'Satisfies' all the tests, i.e.*

1. *Each positive test has at least one member of $\mathcal{L}$,*
2. *No negative test has any member of $\mathcal{L}$.*

- Not practically feasible?
- (Problem relaxed to reals soluble via linear programming)
- But if $\mathrm{SSS}$ doesn't work, why would any other algorithm?
- Intuition formalised by Aldridge (*IEEE Trans. Inform Theory*, 2019)

# SSS example

| ? | ? | ? | ? | ? | ? | ? | |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |

- Since we have only $n = 7$ items, it is, in this small case, practical to check all $2^7 = 128$ subsets.
- It is not difficult to check that the sets $\{2, 4\}$ and $\{2, 4, 7\}$ are the only satisfying sets.
- Of these, $\{2, 4\}$ is the smallest satisfying set.
- Can use $\mathrm{COMP}$ and $\mathrm{DD}$ as pre-processing steps.

# Section 6:
# Group testing and COVID

# Group testing in Wuhan

## Coronavirus: China's plan to test everyone in Wuhan

8 June 2020

The other way they sped up the process was to use a batch testing method, which groups individual test samples together.

Reports suggest they used batches of between five and 10 samples in Wuhan, only carrying out individual tests if a batch proved positive.

# Group Testing and COVID

- Early in pandemic, tests v. scarce, so group testing natural strategy.
- Has been applied in practice in various countries.
- Much of it has been Dorfman type pooling (e.g. 11 million in 5 days in Wuhan in May 2020).
- As much a regulatory problem as a maths one?
- Very extensive literature developed in last few years.
- See survey by Aldridge and Ellis.
- Need finite blocklength results, not just asymptotic rate.
- Maybe isn't (binary) group testing at all?

# Proof of concept: March 2020

*Clinical Infectious Diseases*

MAJOR ARTICLE

IDSA Infectious Diseases Society of America hivma hiv medicine association OXFORD

## Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools

**Idan Yelin,[1,a] Noga Aharony,[1,a] Einat Shaer Tamar,[1,a] Amir Argoetti,[1,a] Esther Messer,[2] Dina Berenbaum,[1] Einat Shafran,[3] Areen Kuzli,[3] Nagham Gandali,[3] Omer Shkedi,[4] Tamar Hashimshony,[1] Yael Mandel-Gutfreund,[1,5] Michael Halberthal,[4,6,b] Yuval Geffen,[7,b] Moran Szwarcwort-Cohen,[3,b] and Roy Kishony[1,5,b]**

[1]Faculty of Biology, Technion - Israel Institute of Technology, Haifa, Israel, [2]Safety Unit, Technion - Israel Institute of Technology, Haifa, Israel, [3]Virology laboratory, Rambam Health Care Campus, Haifa, Israel, [4]Faculty of Medicine, Technion - Israel Institute of Technology, Haifa, Israel, [5]Computer Science Department, Technion - Israel Institute of Technology, Haifa, Israel, [6]Rambam Health Care Campus, Haifa, Israel, and [7]Bacteriology laboratory, Rambam Health Care Campus, Haifa, Israel

***Background.*** The recent emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) led to a current pandemic of unprecedented scale. Although diagnostic tests are fundamental to the ability to detect and respond, overwhelmed healthcare systems are already experiencing shortages of reagents associated with this test, calling for a lean immediately applicable protocol.

***Methods.*** RNA extracts of positive samples were tested for the presence of SARS-CoV-2 using reverse transcription quantitative polymerase chain reaction, alone or in pools of different sizes (2-, 4-, 8-, 16-, 32-, and 64-sample pools) with negative samples. Transport media of additional 3 positive samples were also tested when mixed with transport media of negative samples in pools of 8.

***Results.*** A single positive sample can be detected in pools of up to 32 samples, using the standard kits and protocols, with an estimated false negative rate of 10%. Detection of positive samples diluted in even up to 64 samples may also be attainable, although this may require additional amplification cycles. Single positive samples can be detected when pooling either after or prior to RNA extraction.

- ‘A single positive sample can be detected in pools of up to 32 samples’

# Rwanda work: Nature paper October 2020

Article

## A pooled testing strategy for identifying SARS-CoV-2 at low prevalence

https://doi.org/10.1038/s41586-020-2885-5

Received: 13 May 2020

Accepted: 12 October 2020

Published online: 21 October 2020

Check for updates

Leon Mutesa[1,2], Pacifique Ndishimye[2,3], Yvan Butera[1,2], Jacob Souopgui[1,2,4], Annette Uwineza[1,2], Robert Rutayisire[1,2], Ella Larissa Ndoricimpaye[2], Emile Musoni[2], Nadine Rujeni[2], Thierry Nyatanyi[2], Edouard Ntagwabira[2], Muhammed Semakula[2], Clarisse Musanabaganwa[2], Daniel Nyamwasa[2], Maurice Ndashimye[2,3], Eva Ujeneza[3], Ivan Emile Mwikarago[2], Claude Mambo Muvunyi[2], Jean Baptiste Mazarati[2], Sabin Nsanzimana[2], Neil Turok[3,5,6]✉ & Wilfred Ndifon[3]✉

- Hypercube algorithm, practical trials

# Tapestry pooling

medRχiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

CSH Cold Spring Harbor Laboratory BMJ Yale

HOME | ABOU

Search

Comment on this paper

**Tapestry: A Single-Round Smart Pooling Technique for COVID-19 Testing**

Sabyasachi Ghosh, Ajit Rajwade, Srikar Krishna, Nikhil Gopalkrishnan, Thomas E. Schaus, Anirudh Chakravarthy, Sriram Varahan, Vidhya Appu, Raunak Ramakrishnan, Shashank Ch, Mohit Jindal, Vadhir Bhupathi, Aditya Gupta, Abhinav Jain, Rishi Agarwal, Shreya Pathak, Mohammed Ali Rehan, Sarthak Consul, Yash Gupta, Nimay Gupta, Pratyush Agarwal, Ritika Goyal, Vinay Sagar, Uma Ramakrishnan, Sandeep Krishna, Peng Yin, Dasaradhi Palakodeti, Manoj Gopalkrishnan

**doi:** https://doi.org/10.1101/2020.04.23.20077727

- Non adaptive, each sample tested in 3 different pools (Reed–Solomon design)
- Outperforms binary converse bounds due to varying Ct levels – cf compressed sensing

# Section 7:
# Algorithm performance

# Matrix designs

- First question is how to choose test strategy (design the text matrix)
- Algorithms like $\mathrm{COMP}$ and $\mathrm{DD}$ should work for any matrix design given enough tests.
- Hope a sensible random matrix performs well on average.
- cf random coding, RIP in compressed sensing.
- Common strategy in combinatorics – see 'probabilistic method'.
- Though COVID has shown good performance from structured matrices (finite blocklength problem?)

# Simple matrix designs

- Most basic design is Bernoulli($p$) – every item in every test independently with probability $p$.
- This design is relatively easy to analyse.
- Number of defective items in each test is $\sim \mathrm{Bin}(K, p) \simeq \mathrm{Po}(pK)$.
- So taking $p = 1/K$ implies about 36% of tests are negative.
- Hence entropy of outcomes is quite high, tests are informative.
- Some items don't get tested very often which can be a problem.
- So constant column-weight matrix can perform better (though harder to analyse).

# COMP performance

- Given a nondefective item, the probability that it appears in a particular test, and that such a test is negative, is $p(1-p)^K$.
- Hence, probability this given nondefective appears in no negative tests is $(1 - p(1-p)^K)^T$.
- COMP succeeds precisely when every nondefective item appears in a negative test, so the union bound gives

$$\begin{aligned}
\mathbb{P}(\mathrm{err}) &= \mathbb{P}\left( \bigcup_{i \in \mathcal{K}^c} \{\text{item } i \text{ does not appear in a negative test}\} \right) \\
&\leq |\mathcal{K}^c| \left( 1 - p(1-p)^K \right)^T \\
&\leq N \exp(-Tp(1-p)^K).
\end{aligned}$$

# COMP performance (cont.)

- Taking $p = 1/K$ meaning that $(1 - 1/K)^K \to e^{-1}$.
- Taking $T = (1+\delta)eK \ln N$ means that $Tp(1-p)^K \sim (1+\delta)\ln N$.
- Deduce that

$$\mathbb{P}(\mathrm{err}) \leq N \exp(-Tp(1-p)^K) \sim N^{-\delta}.$$

- Hence if $K \sim N^{\alpha}$ the COMP rate is

$$\begin{aligned} \frac{H(U)}{T} &= \frac{\log_2 \binom{N}{K}}{T} = \frac{(1-\alpha)K \ln N}{T \ln 2} \\ &= \frac{(1-\alpha)}{(1+\delta)e \ln 2} \sim 0.531(1-\alpha). \end{aligned}$$

# DD rate

- With $K$ defectives, Bernoulli$(1/K)$ matrix does well:

**Theorem (ABJ, *IEEE Trans. Inform Theory*, 2014)**

*Using a Bernoulli$(1/K)$ matrix in the regime $K = N^\alpha$, the* DD *algorithm has $\mathbb{P}(\mathrm{suc}) \to 1$ when*

$$\textit{rate} \; < \frac{1}{e \ln 2} \min\left(1, \frac{1-\alpha}{\alpha}\right) \simeq 0.53 \min\left(1, \frac{1-\alpha}{\alpha}\right).$$

- Constant column weight even better (see Johnson, Aldridge, Scarlett, *IEEE Trans. Inform Theory*, 2019).
- Improves the 0.53 to 0.69.
- i.e. rate $0.69 \min\left(1, \frac{1-\alpha}{\alpha}\right)$ achievable.

# Intuition for DD rate

- Assuming that $p = 1/K$ and that $T = CeK \ln N$, expected number of negative tests is

$$\mathbb{E}M_0 = T(1-p)^K \sim Te^{-1} = CK \ln N.$$

- Write $G$ for the (random) number of non-defectives wrongly surviving $\mathrm{COMP}$ step.
- Can argue $G | M_0 = m_0 \sim \mathrm{Bin}(N-K, (1-p)^{m_0})$.
- Binomial concentration tells us that $M_0$ is close to its mean with high probability, so $G$ has

$$\mathbb{E}G \sim N(1-p)^{\mathbb{E}M_0} \sim N \exp(-p\mathbb{E}M_0) \sim N^{1-C}.$$

- If $C = \max\{\alpha, 1-\alpha\} + \epsilon \geq 1 - \alpha + \epsilon$, then deduce $\mathbb{E}G \leq N^{\alpha-\epsilon} \ll K$ (since $K \sim N^\alpha$).
- I.e. number of possible defectives is close to number of true defectives, so true defectives should not get 'drowned out'.

# SCOMP performance

- SCOMP is harder to analyse
- Greedy, sequential algorithms like this mean you need to keep track of all the previous stages.
- Not many rigorous results.
- Believe it should increase success probability.
- But Coja-Oghlan *et al.* show it doesn't improve the rate.

# Converse results: Bernoulli matrices?

- One defective can get masked by other $K-1$ defectives.
- If so, $\mathrm{SSS}$ (and hence any algorithm) will fail.
- For Bernoulli matrices, inclusion-exclusion gives probability of this.

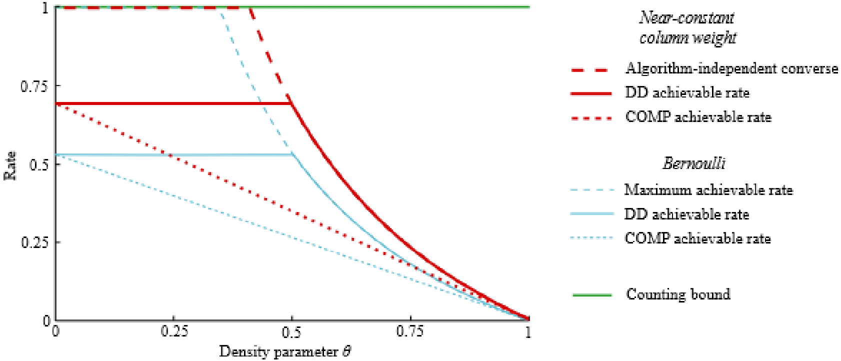
**Theorem (ABJ, *IEEE Trans. Inform Theory*, 2014)**

*Using any Bernoulli matrix in the regime $K = N^\alpha$, the $\mathrm{SSS}$ algorithm has $\mathbb{P}(\mathrm{suc})$ bounded away from 1 when*

$$\textit{rate} > \frac{1}{e \ln 2} \left( \frac{1-\alpha}{\alpha} \right) \simeq 0.53 \left( \frac{1-\alpha}{\alpha} \right).$$

- Hence, although simple, $\mathrm{DD}$ is essentially optimal for $\alpha > 1/2$

# Rate bounds



1
0.75
0.5
0.25
0
Rate
0
0.25
0.5
0.75
1
Density parameter θ
Near-constant column weight
Algorithm-independent converse
DD achievable rate
COMP achievable rate
Bernoulli
Maximum achievable rate
DD achievable rate
COMP achievable rate
Counting bound

# Converse results in general

- Can provide a similar converse for constant column weight – again, 0.53 improves to 0.69.
- In fact this is optimal for **all** matrix designs.
- See Coja-Oghlan, Gebhard, Hahn-Klimroth, Loick *COLT* 2020.
- Also give practical algorithm (under tweaked matrix design) that matches SSS performance.
- Hence they found the capacity of noiseless group testing.

# Section 8:
# Testing with noise

# Effect of noise

- Unrealistic to imagine noiseless model holds in practice.
- Tests will not be perfect, and give measurement errors.
- Work in progress to develop corresponding theory for noisy measurements.
- Key is model for noise: different model, different result.
- Simplest model: send test outputs through BSC.
- Can extend this to independent noise channel with $\mathbb{P}(0 \to 1) \neq \mathbb{P}(1 \to 0)$.
- In practice: dilution/threshold effects?

# Noisy testing notation

- Recall that noiseless group testing has outputs

$$y_t = \bigvee_{i \in \mathcal{K}} x_{ti}$$

- Can consider models where output of test $t$ is (random) function of $\bigvee_{i \in \mathcal{K}} x_{ti}$.

**Example (Binary symmetric noise)**

In the binary symmetric noise model, the $t$-th test outcome is given by

$$Y_t = \begin{cases} \bigvee_{i \in \mathcal{K}} x_{ti} & \text{with probability } 1 - \rho \\ 1 - \bigvee_{i \in \mathcal{K}} x_{ti} & \text{with probability } \rho. \end{cases}$$

This is, each test is flipped independently with probability $\rho$.

# Noisy testing notation (cont.)

## Definition

We define the probability transition function $p(\cdot \mid m, \ell)$ such that for a test containing $m$ items, $\ell$ of which are defective, for each outcome $y \in \mathcal{Y}$ we have

$$\mathbb{P}\left( Y_t = y \;\middle|\; \sum_{i=1}^{N} X_{ti} = m, \sum_{i \in \mathcal{K}} X_{ti} = \ell \right) = p(y \mid m, \ell),$$

independently of all other tests.

## Example (Standard noiseless group testing)

$$\begin{aligned} p(1 \mid m, \ell) &= 1 \quad \text{if } \ell \geq 1, & p(0 \mid m, \ell) &= 0 \quad \text{if } \ell \geq 1, \\ p(1 \mid m, \ell) &= 0 \quad \text{if } \ell = 0, & p(0 \mid m, \ell) &= 1 \quad \text{if } \ell = 0, \end{aligned}$$

independent of $m$.

# Only defects matter

- Many noise models have the following property:

**Definition**

A noise model satisfies the *only defects matter* property if the probability transition function is of the form

$$p(y \mid m, \ell) = p(y \mid \ell).$$

# Noisy model examples – models not equivalent?

## Example (Addition (reverse $Z$-channel) noise - false positives only)

In the addition noise model, the probability transition function is given by

$$\begin{aligned} p(1 \mid m, \ell) &= 1 \quad \text{if } \ell \geq 1, & p(0 \mid m, \ell) &= 0 \quad \text{if } \ell \geq 1, \\ p(1 \mid m, \ell) &= \varphi \quad \text{if } \ell = 0, & p(0 \mid m, \ell) &= 1 - \varphi \quad \text{if } \ell = 0, \end{aligned}$$

where $\varphi \in (0, 1)$ is a noise parameter.

## Example (Z channel noise - false negatives only)

In the Z channel noise model, the probability transition function is given by

$$\begin{aligned} p(1 \mid m, \ell) &= 1 - \vartheta \quad \text{if } \ell \geq 1, & p(0 \mid m, \ell) &= \vartheta \quad \text{if } \ell \geq 1, \\ p(1 \mid m, \ell) &= 0 \quad \text{if } \ell = 0, & p(0 \mid m, \ell) &= 1 \quad \text{if } \ell = 0, \end{aligned}$$

where $\vartheta \in (0, 1)$ is a noise parameter.

# Noisy model examples

## Example (Dilution noise)

In the dilution noise model, the probability transition function is given by

$$p(1 \mid m, \ell) = 1 - \vartheta^\ell, \qquad p(0 \mid m, \ell) = \vartheta^\ell, \quad \text{for all } \ell \geq 0,$$

where $\vartheta \in (0, 1)$ is a noise parameter.

## Example (Threshold group testing)

In the probabilistic threshold group testing noise model, the probability transition function is given by

$$\begin{aligned} p(1 \mid m, \ell) &= 1 && \text{if } \tfrac{\ell}{m} \geq \overline{\theta}, & p(0 \mid m, \ell) &= 0 && \text{if } \tfrac{\ell}{m} \geq \overline{\theta}, \\ p(1 \mid m, \ell) &= 0 && \text{if } \tfrac{\ell}{m} \leq \underline{\theta}, & p(0 \mid m, \ell) &= 1 && \text{if } \tfrac{\ell}{m} \leq \underline{\theta}, \\ p(1 \mid m, \ell) &= \tfrac{1}{2} && \text{if } \underline{\theta} < \tfrac{\ell}{m} < \overline{\theta}, & p(0 \mid m, \ell) &= \tfrac{1}{2} && \text{if } \underline{\theta} < \tfrac{\ell}{m} < \overline{\theta}\ , \end{aligned}$$

where $\underline{\theta} \leq \overline{\theta}$ are thresholds.

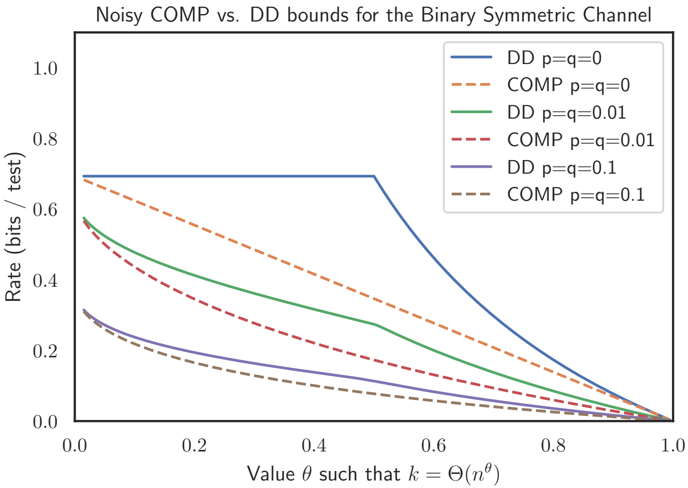
# Noisy results

- Fair to say that there isn't a general theory for noisy channels yet.
- Algorithms need to be developed on slightly ad hoc basis.
- Performance bounds, tuning parameters similarly ad hoc.
- Some ideas pass over though.

## Algorithm (Noisy DD)

1. *Declare every individual that appears in $\alpha\Delta$ or more displayed negative tests as healthy and remove such individual from every assigned test.*
2. *Declare every yet unclassified individual who is now the only unclassified individual in $\beta\Delta$ or more displayed positive tests as infected.*
3. *Declare all remaining individuals as healthy.*

# Can give performance bounds




from Gebhard, Johnson, Loick and Rolvien, *IEEE Transactions IT*, 2022

# Some theoretical results

- Strengthened form of only defects matter property:

## Definition (Noisy defective channel)

If we can express

$$p(y \mid m, \ell) = p(y \mid \mathbb{I}\{\ell \geq 1\}),$$

where $p(y \mid \mathbb{I}\{\ell \geq 1\})$ is the transition probability function of a noisy binary communication channel, then we say that the *noisy defective channel property* holds.

## Theorem

*If the noisy defective channel property holds then the group testing capacity $C$ is bounded above by $C_{\mathrm{chan}}$, the Shannon capacity of the corresponding noisy communication channel $p(y \mid \mathbb{I}\{\ell \geq 1\})$:*

$$C \leq C_{\mathrm{chan}}.$$

# Section 9:
# Group testing extensions

# Different defectivity models

- Have focused on model where $\mathcal{K}$ is chosen uniformly among all sets of size $K$ (combinatorial prior).
- Alternative models exist: e.g. every item defective independently with same probability $p$ (probabilistic prior).
- To make more realistic, could remove independence assumption.
- e.g. simulate model with local interactions, infections grow in spatial clusters.
- In theory, makes things easier (magic number lower than if infections independent).
- But some practical questions: group testing within families might have multiple positives (would prefer 0 or 1).

# Group testing with constraints

- Have assumed we can use any test strategy that we like.
- In practice, this may not be possible.
- For example, medical testing, can perhaps only split a sample into $\gamma$ parts, or have at most $\rho$ samples in one test.
- Here $\gamma$, $\rho$ are bounded, whereas e.g. Bernoulli designs would want to let these quantities tend to infinity (see Gandikota et al.).
- See also model where items correspond to nodes on a graph (e.g. testing a computer network)
- May only allow test pools that correspond to physically allowable paths in a network (see work of Chergaghchi *et al.*).

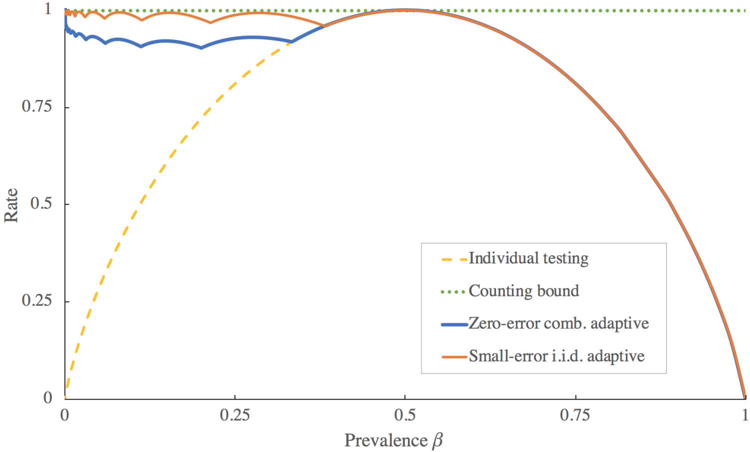
# Finite blocklength

- A lot of the theory is in asymptotic regimes: e.g. $N \to \infty$, $K \sim N^\alpha$ or $K \sim N\beta$, how can $T$ behave?
- But in many practical situations we have a finite size problem.
- e.g. (PCR testing) $T = 96$, $N = 500$, how many defectives can we find?
- Which regime is this (linear or sublinear)?
- It shouldn't really matter – would like *universal* performance bounds.
- Fairly open question . . .

# Linear regime

- Mostly focused on the sparse ($K = N^{\alpha}$) regime
- Some interesting questions in the linear ($K = N\beta$) regime.
- Turns out you can't beat individual testing for non-adaptive problem (see Aldridge).
- Adaptive testing more interesting.
- Binary splitting algorithms work (though get into issues about rounding set sizes to powers of 2).
- For small error probability, results also due to Aldridge.
- For zero error probability, individual testing is suboptimal for $\beta < 1/3$, conjectured to be optimal for $\beta \geq 1/3$, known to be optimal for $\beta > 0.369$.

# Linear regime results



Thank you for listening!