PCR, Tropical Arithmetic, and Group Testing

Hsin-Po Wang
with Ryan Gabrys and Alexander Vardy

Department of Electrical and Computer Engineering, University of California San Diego

2022/05/23 NTU GICE Seminar

To be in ISIT 2022. Slides at http://h-p.wang/ntu. Preprint at abs/2201.05440

動機

大目標是篩檢 Covid

抗原、抗體 = 便宜但不準

PCR (Polymerase Chain Reaction 聚合酶連鎖反應) = 準但貴
而且可以追蹤新變種 (alpha/delta/omicron/...)

Group Testing (GT) 可不可以「疊」在現有的篩檢方法上以加快其速度？

大綱

PCR 的原理

其他人的 Group Testing 方法（如 Quantitative GT 跟 Semi-Quantitative GT）

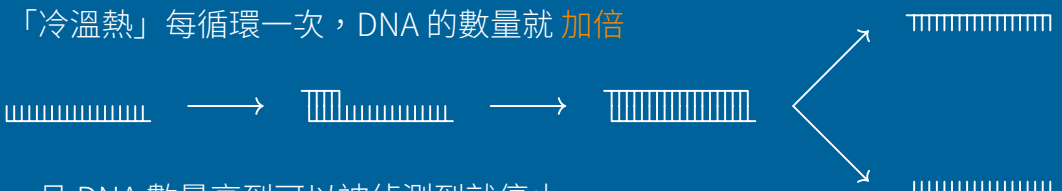
我們提出的 Group Testing 方法（叫做 Tropical GT）

PCR 的原理

PCR 機器基本上是一個三溫暖機，有冷、溫、熱三種溫度

冷 = Annealing：引子（primer）跟聚合酶（polymerase）會黏到單股 DNA 上
溫 = Elongation：聚合酶會從引子開始，把單股 DNA 補完成雙股 DNA
熱 = Denaturation：一條雙股 DNA 分裂成兩條單股 DNA

「冷溫熱」每循環一次，DNA 的數量就 加倍



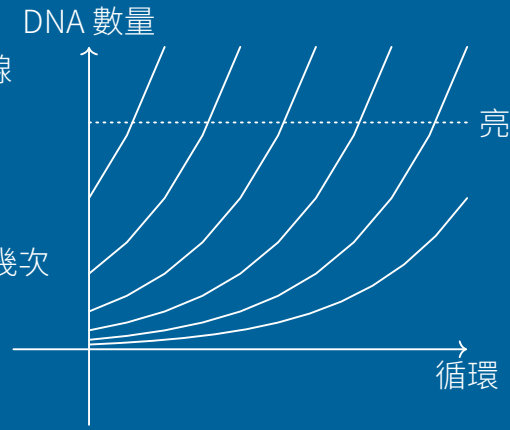
一旦 DNA 數量高到可以被偵測到就停止

偵測 DNA 的方法與 Ct 值的算法

在試管裡放入「喜歡黏在 DNA 上，且黏上去之後照紫外線會發出螢光」的分子

所以每次「溫」跟「熱」之間就照一下紫外線
如果偵測到螢光就代表試管裡有病毒的 DNA

Ct 值就是偵測到螢光之前「冷溫熱」循環了幾次



問題

如何結合 GT 跟 Ct？

複習：Binary Group Testing (GT)

在回答問題之前，先複習一下最傳統的 Binary GT 概念
想像篩檢的結果只有 陰性 或 陽性

把五個人的唾液混在一起測一次
若陰性，則五個人皆陰性
若陽性，則至少有一個人是陽性，但你不知道是誰；故而這五個人要各自測一次

相較於每個人測一次
五個人測出來陰性的話節省了四次篩檢，陽性的話多花費了一次篩檢
如果陰性機率比陽性機率高很多，就有「賺到」

起源：[Dorfman 1943]。書：[Du–Hwang 1993]。上課講義：[Ngo–Rudra 2011]。
Survey paper：[Aldridge–Johnson–Scarlett 2019]

複習：Threshold GT

篩檢的結果同樣分成陰性或陽性或 不確定
病人的數量小於 L 時結果為陰性，病人的數量大於 U 時結果為陽性

[Damaschke 2006] [Dyachkov 2013] [Cheraghchi 2013]

複習：Quantitative GT

你有一個 精密的電子秤 跟十袋硬幣，每袋有數千個硬幣
真幣每個重 5 g，假幣每個重 4.999 g
最多有一袋是假幣（是的話，則整袋都是假幣）

一個例子：
從第一袋硬幣裡抓一個，從第二袋硬幣裡抓兩個，依此類推
把這 55 個硬幣拿去電子秤上秤
秤出來是 275 g 就沒有假幣
比 275 g 少的話，少多少 mg 就代表第幾堆是假幣

另一個名字是 Coin-Weighing Problem
[Hwang 1987] [Guy–Nowakowski 1995] [Bshouty 2009]

複習：Semi-Quantitative GT

你有一個 生鏽的彈簧秤，精確到 0.1 g

In general 這就是「刻度比較粗」的 Quantitative GT

[Emad–Milenkovic 2014] [Cheraghchi–Gabrys–Milenkovic 2021]

複習：Compressed Sensing

意思跟 Semi-Quantitative GT 差不多，看你從哪個科目來
跟看你用「零壹矩陣」還是正常的矩陣
跟矩陣乘法要不要換成「邏輯矩陣乘法」 $\bigvee_j (A_{ij} \wedge B_{jk})$

補充一點是 Compressed Sensing 人可能會喜歡 minimize $|A\mathbf{x} - \mathbf{y}|_2^2 + \lambda |\mathbf{x}|_1$.

[Ghosh et al. 2021] [Shental et al. 2020] [Mutesa et al. 2021]。
Survey paper：[Aldridge–Ellis 2022]。
台： [10.1109/TNSE.2021.3121709] [10.1155/2021/6636396] [10.1016/j.dam.2020.11.022]

下面講為什麼這些 approaches 不見得適合 PCR

PCR 的精度問題

因為 DNA 可以無限複製，PCR 非常的靈敏，最低可以檢測到 100 copies/ml

但同時，PCR 的相對精度爛到，如果病毒數量差不到 兩倍，Ct 值就有可能一樣！

對數尺度的奧妙

疊加 50dB 跟 30dB 的白噪音，可得 50.043dB 的白噪音

混合等體積的 pH 1 跟 pH 3 的鹽酸溶液，等價於 pH 0.9957 稀釋兩倍

規模 9 的地震跟規模 8 的地震同時同地發生，與規模 9.009 相當

視星等 1 的天體跟視星等 6 的天體靠太近，變一顆視星等 0.9892

真正的問題

如何理解對數尺度裡的「相加」？

用 Tropical Arithmetics 熱帶算術！

Tropical Arithmetic 的規則如下：
domain 是實數跟正無限大： $\mathbb{R} \cup \{\infty\}$
熱帶加法： $x \oplus y := \min(x, y)$
熱帶乘法： $x \odot y := x + y$

$x \oplus \infty = x$
$x \odot \infty = \infty$

叫熱帶是因為巴西的數學家先開始研究的

提示：想成對數
$2^{-x} + 2^{-y} \approx 2^{-\min(x,y)}$
$2^{-x} \cdot 2^{-y} \approx 2^{-(x+y)}$

熱帶算術 + 矩陣乘法 = 熱帶矩陣乘法

假設 $X_1, \ldots, X_\ell, Y_1, \ldots, Y_m, Z_1, \ldots, Z_n$ 是一些地點
令 A_{ij} 是從 X_i 到 Y_j 的距離
令 B_{jk} 是從 Y_j 到 Z_k 的距離

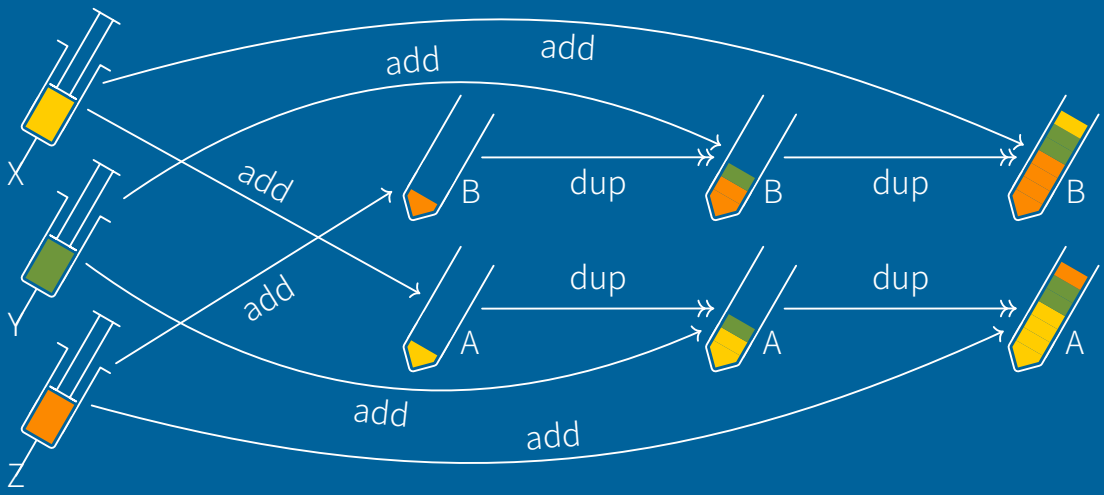
熱帶矩陣乘法：$A \odot B$ 是一個矩陣，它的 (i,k) 那格定成
$\bigoplus_j (A_{ij} \odot B_{jk}) = \min_j (A_{ij} + B_{jk})$
這是從 X_i 經過某個 $Y_?$ 再到 Z_k 的最短距離

公理化「PCR + 池化 Pooling」

假設有 Ct 值分別為 $x_1, x_2, \ldots, x_n$ 的 n 個樣本
分別在第 $\delta_1, \delta_2, \ldots, \delta_n$ 個循環之後加入同一個試管裡
此試管的 Ct 值應為 $-\log_2\left(\sum_j 2^{-\delta_j - x_j}\right)$ (cf. softmin)

我們 假裝 該試管的 Ct 值為 $\boldsymbol{\delta} \odot \mathbf{x} = \bigoplus_j (\delta_j \odot x_j) = \min_j(\delta_j + x_j)$
粗體 $\boldsymbol{\delta}$ 是 δ_j 組成的 row vector，粗體 $\mathbf{x}$ 是 x_j 組成的 column vector

令 S 是 $T \times N$ 矩陣，代表有 T 個試管跟 N 個人， 我們叫它 Schedule 時刻表
$S \odot \mathbf{x}$ 是很多個 Ct 值的 column vector
下一頁是最一開始我寫信給 advisor 時用的例子

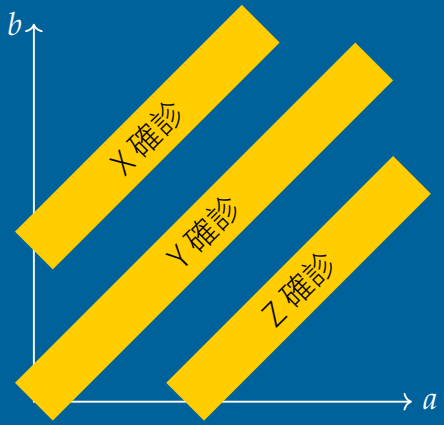


$$\begin{bmatrix} a \\ b \end{bmatrix} \coloneqq \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} \odot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \min(0+x,\, 1+y,\, 2+z) \\ \min(2+x,\, 1+y,\, 0+z) \end{bmatrix}$$

Decoding 上一頁那個 Schedule

假設只有一個人有病毒

X 確診 iff $a - b = -2$
Y 確診 iff $a - b = 0$
Z 確診 iff $a - b = 2$



小結

這個 work 有兩個觀念上的創新

其一是我們用 $x \oplus y := \min(x, y)$ 來刻畫 Ct 值 x 混合 Ct 值 y 的結果

其二是我們引入 $\delta \odot x := \delta + x$（蓄意拖延），來加強 GT

其三是熱帶矩陣乘法剛好同時囊括了這兩個觀念
讓「nonadaptive 熱帶 GT」變得很像「熱帶 Compressed Sensing」

從更多人裡找出一位患者

$\begin{bmatrix} a \\ b \end{bmatrix} \coloneqq \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} \odot \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ 用 兩管 PCR 找到 三人 中的 一位 患者，且拖延不超過 兩圈

這樣叫做 (2管, 3人, 1病)-tropical code within maximum delay 2 圈

有更多人來醫院篩檢的話，$\begin{bmatrix} a \\ b \end{bmatrix} \coloneqq \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 & 0 \end{bmatrix} \odot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix}$

這樣叫做 (2管, 7人, 1病)-tropical code within maximum delay 6 圈

兩管一病的 General Case，但是盡量不要拖

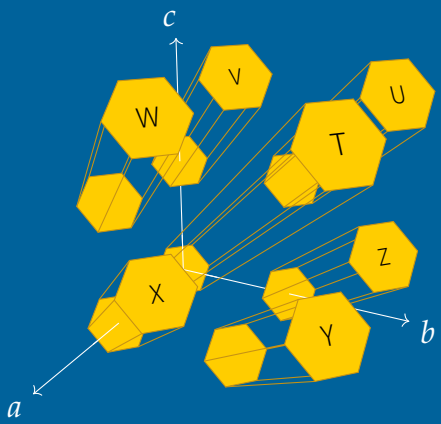
$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$ 是 (2管, 7人, 1病)-tropical code within maximum delay 3 圈

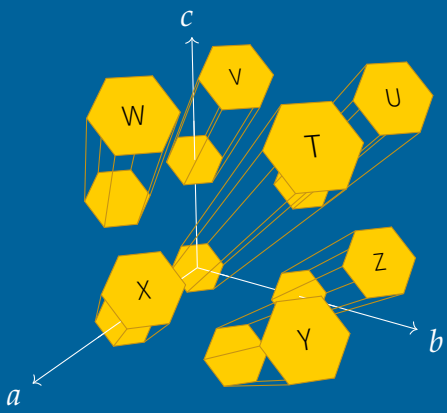
$\begin{bmatrix} 0 & 0 & 1 & 1 & 2 & 2 & 3 \\ 3 & 2 & 2 & 1 & 1 & 0 & 0 \end{bmatrix}$ 亦是 (2管, 7人, 1病)-tropical code within maximum delay 3

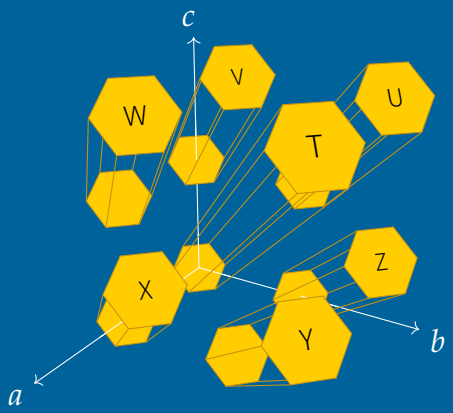
一般式：$(2\text{管}, 2\ell + 1\text{人}, 1\text{病})$-tropical code within maximum delay ℓ 圈

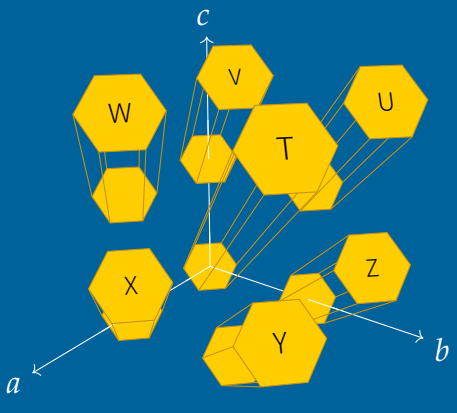
延伸問題

測三次，但是少一點拖延？


c
V
W
U
T
Z
X
Y
b
a


c
V
W
U
T
Z
X
Y
b
a


c
V
W
U
T
X
Z
Y
a
b


c
V
W
U
T
X
Z
Y
a
b

更重要的問題

要是有兩個或更多患者呢？

要 Minimize 哪個 Metric？

Minimize 把樣本滴來滴去的 pipetting work

Maximize 平均每管 PCR 可以篩檢的人數

Minimize 拖延的時間

Minimize 滴管使用量

小推理：若 Xavier 只參與一個試管，那 Yvonne 就不能參與該管
（否則 Yvonne 比較多病毒的話就會「蓋掉」Xavier）

如果不想要退化成 individual testing 的話，每個人都至少參與兩個試管

先統一設定為兩個
可以用 graph theory 裡的 graph 來描述誰參與了哪兩個試管
試管 = vertex，人 = edge

每個人要決定要拖延多久才把樣本放進試管裡
delay 多久其實不重要，重要的是兩個試管的 delays 差多少

發明一個單字：diff-lay
每個人把自己的 diff-lay 寫在 graph 的 edge 上

A
B
C
D
E
4
2
1
1
9
5

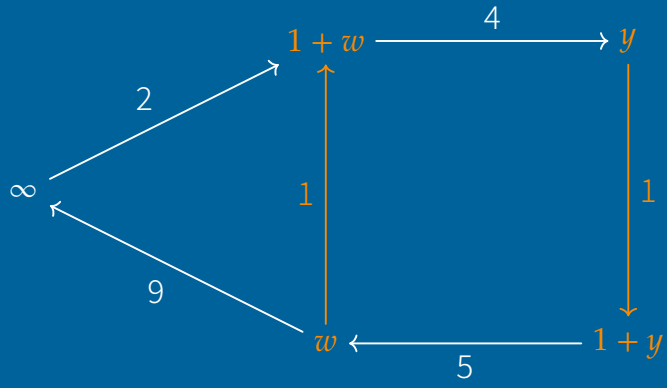
11
4
16
2
8
1
1
9
10
8
5

11
4
∞
2
30
1
1
9
10
∞
5

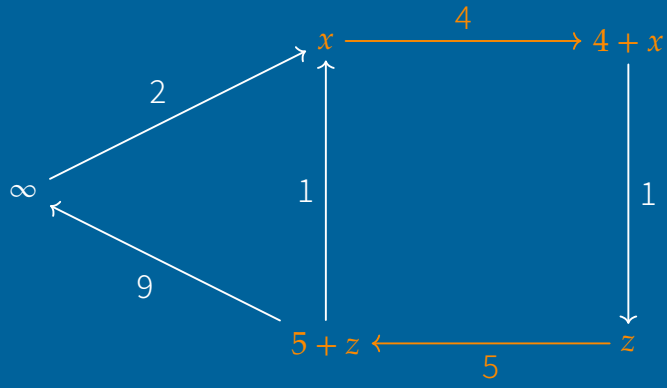
從剛剛的例子我們學到

三角形不好

四邊形有條件好，條件是：
沿著四個 edges 把 diff-lays 加起來之後要非零（逆向要加負號）



$$a - b + c - d = (1 + w) - y + (1 + y) - w = 2$$



$$a - b + c - d = x - (4 + x) + z - (5 + y) = -9$$

每個人參與兩個試管的 Design Principle

不能有三角形，但是四邊形可以，而且 edges 要盡量多 ➡ 二部圖 bipartite graph

每個人要決定自己的 diff-lay，使得每個四邊形的「順向和」 $\neq 0$
➡ 用 mod-p 乘法表 當作 bi-adjacency matrix，p 是質數
➡ 從「左邊第 i 個試管」連到「右邊第 j 個試管」這條邊的 diff-lay 是 $ij \bmod p$
p 是質數

➡ $(2p$管$, p^2$人$, 2$病$)$-tropical code within maximum delay $(p-1)/2$ 圈

要 Minimize 哪個 Metric？

Minimize 把樣本滴來滴去的 pipetting work ⬅ 這個講完了

Maximize 平均每管 PCR 可以篩檢的人數 ⬅ 換這個

Minimize 拖延的時間

Maximize 平均每管 PCR 可以篩檢的人數

令 X 為 the subset of tubes Xavier is in
令 Y 為 the subset of tubes Yvonne is in

若 $|X \setminus Y| = 0$，代表 Yvonne 病得很重的話會「蓋住」Xavier
若 $|X \setminus Y| = 1$ 且 Yvonne 病重，我們只剩一個試管可以「看到」Xavier

基於上開理由，希望至少要 $|X \setminus Y| \geq 2$

每個人都比另一個人「多兩管」

我們需要 a family of subsets of $\{1, \ldots, T\}$ 滿足這個條件：
$|X \setminus Y| \geq 2$ 對任兩個 family 裡的 subsets X, Y

取所有大小是 $\lfloor T/2 \rfloor$ 的 subsets，但是限制 $\sum_{a \in X} a \equiv 0 \pmod{T}$.

令 $x \in \{1, \ldots, N\}$ 是 Xavier 的編號
X 是 Xavier 參與的試管的 subset，再令 $a \in X$ 是一個試管編號

「Xavier 把他樣本放進試管 a 裡」after 拖延 $ax \bmod N$ 圈，N 是質數

要 Minimize 哪個 Metric？

Minimize 把樣本滴來滴去的 pipetting work ⬅ 這個講完了，bipartite graph

Maximize 平均每管 PCR 可以篩檢的人數 ⬅ 這也個講完了

Minimize 拖延的時間 ⬅ 換這個

用現成的構造「張量積」出新構造

令 S 是一個 (t管, n人, 2病)-tropical code

考慮 $\begin{bmatrix} S \otimes \text{全一矩陣}_{1\times n} \\ \text{全一矩陣}_{1\times n} \otimes S \end{bmatrix}$ ⬅ 把連續 n 個人當作一組，用 S 測這 n 組
⬅ 把模 n 同餘的人當作一組，用 S 測這 n 組

假設 $y_{10}n + y_1$ 跟 $z_{10}n + z_1$ 染疫、Ct 值分別是 y_* 跟 z_*
那我們會知道 $\{(y_{10}, y_*), (z_{10}, z_*)\}$ 跟 $\{(y_1, y_*), (z_1, z_*)\}$

用 Systematic Reed–Solomon 生成 Checksum

我們知道 $\{(y_{10}, y_*), (z_{10}, z_*)\}$
如果兩個病人 Ct 值一樣會出包
➡ 用 systemic Reed–Solomon code 生成 checksum 做最後檢查：

$$\begin{bmatrix} S \otimes \text{全一矩陣}_{1\times n} \\ \text{全一矩陣}_{1\times n} \otimes S \\ \text{one row of checksums} \end{bmatrix}$$

第 $in + j$ column 的 checksum k 滿足： (i, j, k) is a codeword of $[3, 2, 2]$-RS

結論（倒數第二張，快講完了）

這個 work 有兩個觀念上的創新

數值模擬顯示，假裝 $x \oplus y := \min(x, y)$ 可以大幅簡化 decoding

蓄意拖延，意即 $\delta \odot x := \delta + x$，製造了很多有趣的組合設計問題
我們用到二部圖、區塊設計、有限體乘法、張量積、里德–索羅門碼等工具

展望（可以問問題了）

把 decoding 包裝成 convex optimization（化為 LASSO？）

病人數 ≥ 3 知道得很少

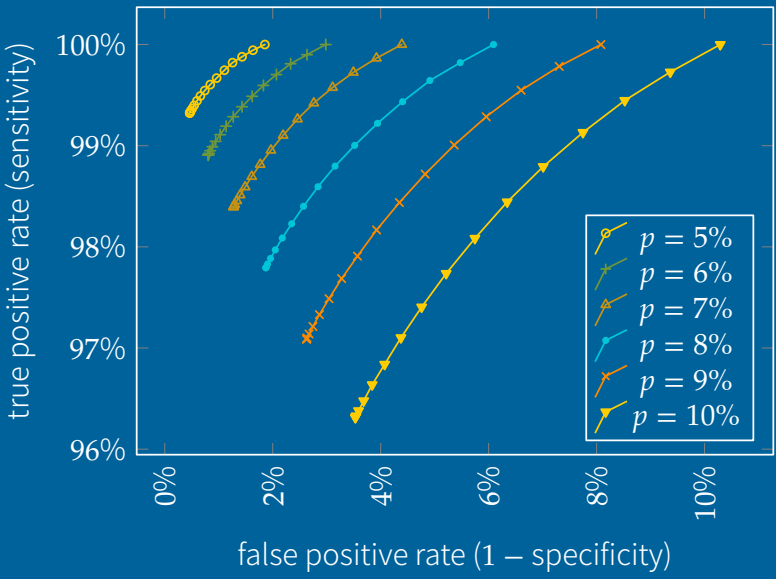
Asymptotic behavior 都猶未了解

Noisy measurement

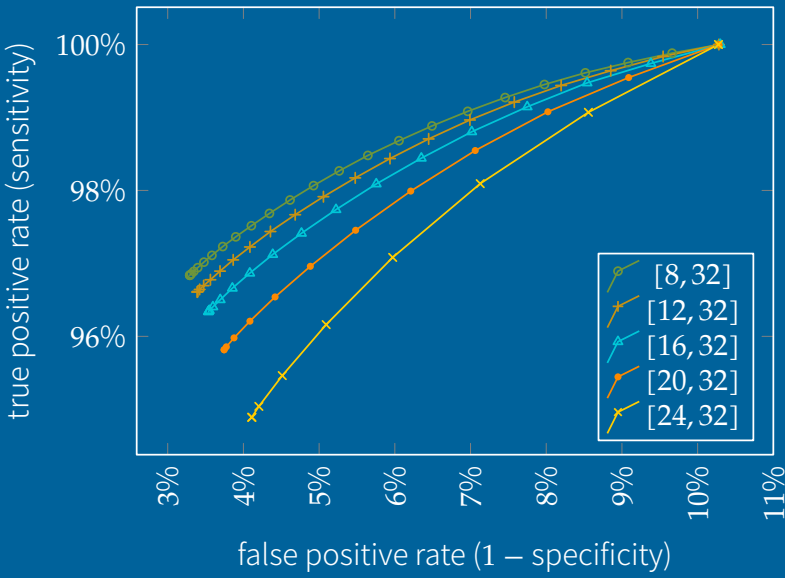
To be in ISIT 2022. Slides at http://h-p.wang/ntu. Preprint at abs/2201.05440

闌尾

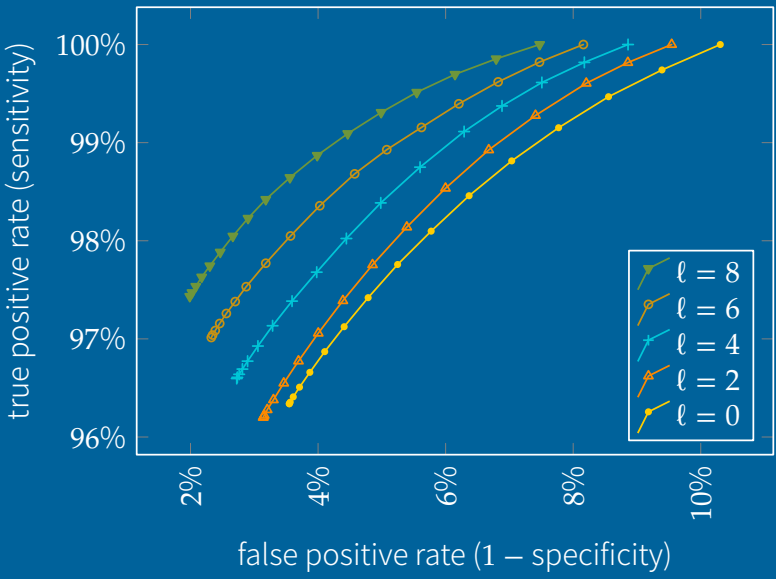
Appendix



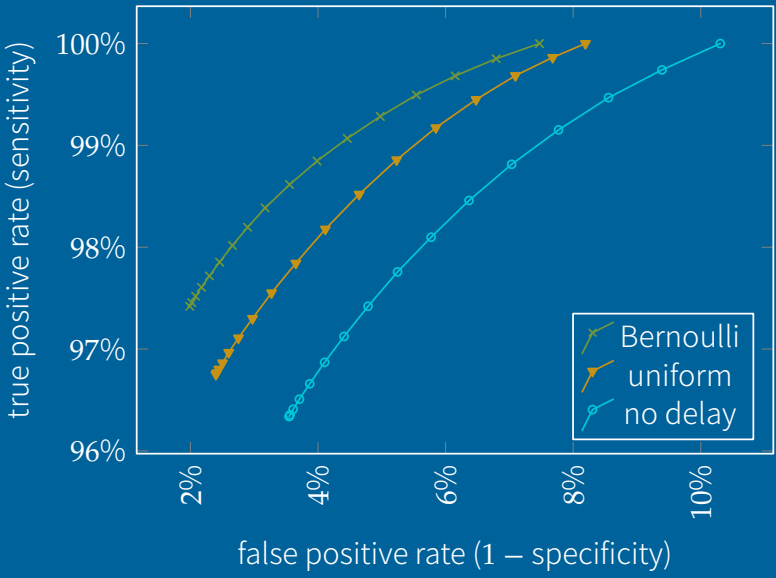
Assume uniform Ct values on the interval $[16, 32]$, 15×35 Kirkman triple system, and no delay ($\ell = 0$). We vary the prevalence rate p and plot the ROC curves.



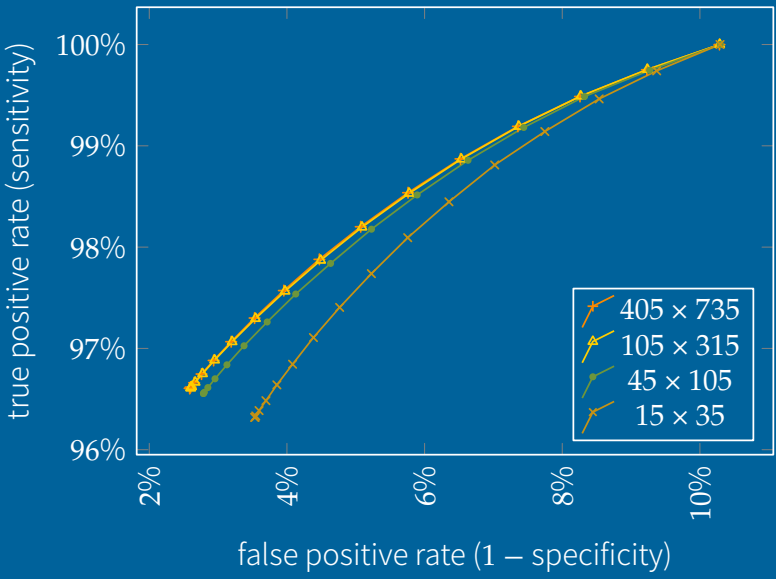
Assume prevalence rate $p = 10\%$, uniform Ct values, 15×35 Kirkman triple system, and no delay ($\ell = 0$). We vary the range of the Ct values and plot the ROC curves. Surprisingly, larger interval (consequently larger variance) is easier to decode.



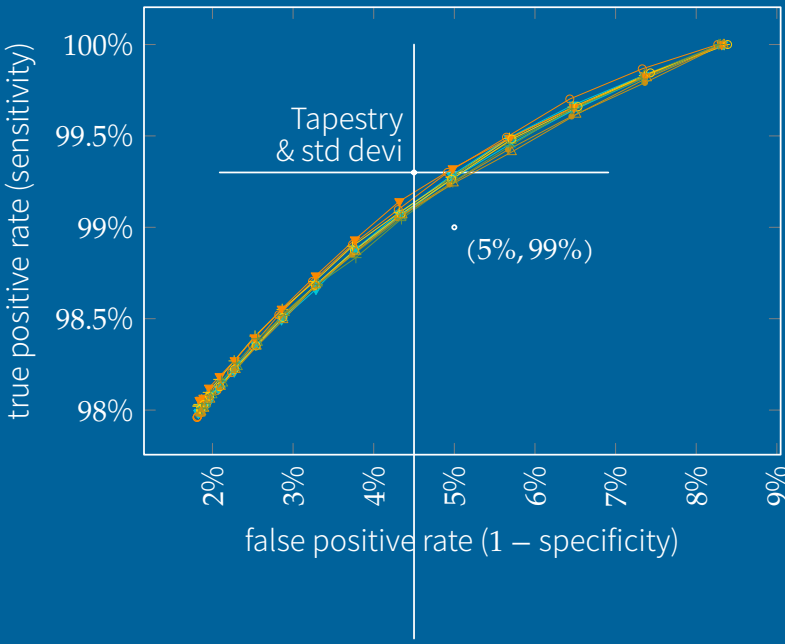
Assume prevalence rate $p = 10\%$, uniform Ct values on the interval $[16, 32]$, 15×35 Kirkman triple system, and $\ell \cdot \text{Bernoulli}(1/2)$ delay. We vary the limit of delay ℓ and plot the ROC curves.



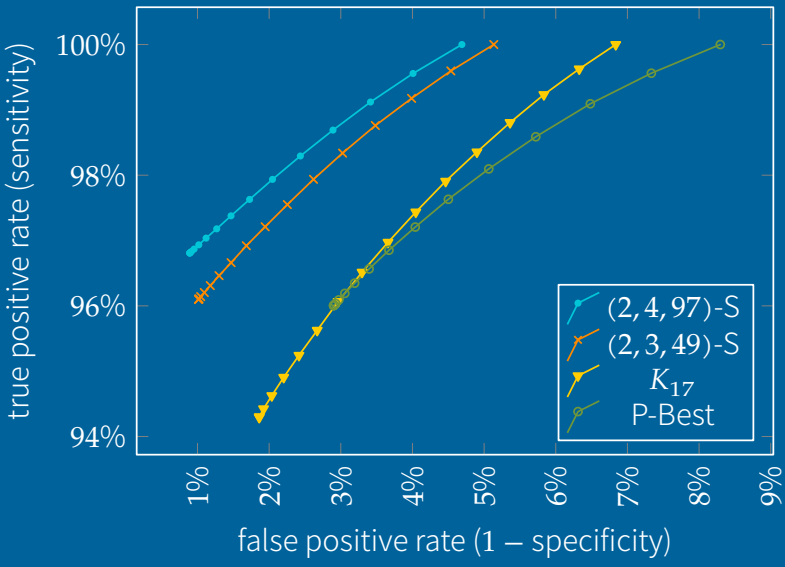
Assume prevalence rate $p = 10\%$, uniform Ct values on the interval $[16, 32]$, 15×35 Kirkman triple system, and $\ell = 8$. We vary the distribution of the random delay δ and plot the ROC curves.



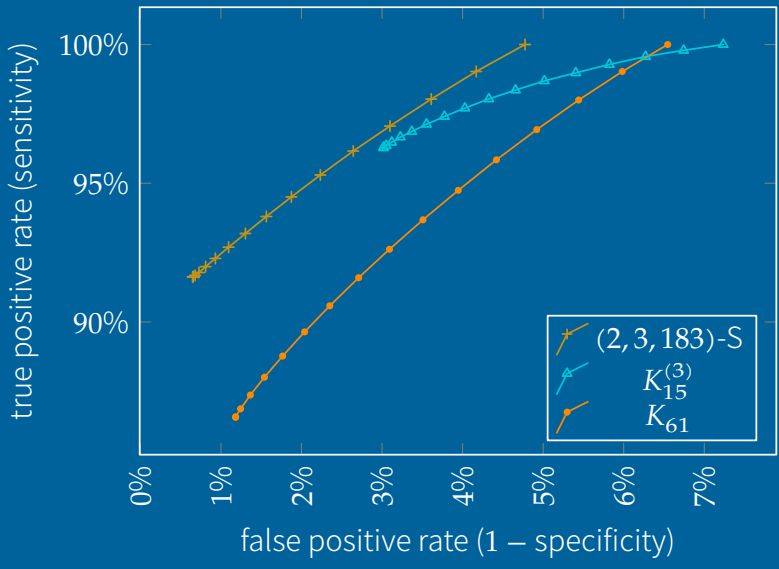
Assume prevalence rate $p = 10\%$, uniform Ct values on $[16, 32]$, and no delay ($\ell = 0$). We consider Kirkman triple systems of different size (after truncation so that the code rate $N/T = 7/3$ is fixed) and plot the ROC curves.



Assume $D = 10$ patients within $N = 105$ persons (infection rate 9.52%), uniform Ct values on $[16, 32]$, 45×105 Kirkman triple system (truncation of a 45×330 Kirkman triple system), and no delay ($\ell = 0$). We plot 10 ROC curves. Each curve is 10,000 encoding–decodings, i.e., 450,000 tubes, 100,000 patients, and 1,050,000 test takers. Compare this to Tapestry's data point and its standard deviations ($4.50\% \pm 2.41\%, 99.30\% \pm 2.55\%$) (Table S.XII of the preprint version [Ghosh et al. 2020]).



Assume prevalence rate $p = 2\%$, uniform Ct values on $[16, 32]$, and no delay ($\ell = 0$). We consider $(2, 4, 97)$-Steiner system (aka 2-$(97, 4, 1)$ design), $(2, 3, 49)$-Steiner system (aka 2-$(49, 3, 1)$ design), complete graph on 17 vertices, and P-BEST [Shental]. They all have code rate $N/T = 8$. We plot their ROC curves.



Assume prevalence rate $p = 0.5\%$, uniform Ct values on $[16, 32]$, and no delay ($\ell = 0$). We consider Kirkman triple system on 183 vertices, complete 3-uniform hypergraph on 15 vertices, and complete graph on 61 vertices. The first two have code rate $N/T = 30 + 1/3$; the last one has code rate $N/T = 30$. We plot their ROC curves.