Group testing meets traitor tracing

Peter Meerwald and Teddy Furon

INRIA Rennes Bretagne Atlantique, Rennes, France
Email: {peter.meerwald,teddy.furon}@inria.fr

Abstract

1. Link Group Testing (GT) and Traitor Tracing (TT)
2. Apply our Traitor Tracing decoding algorithm to Group Testing

What is Group Testing?

Epidemiology: Identify a small set of virally-infected people in a large population.
It is too expensive to test all the blood samples.

Setup N: population size, K: number of infected, T: number of pools of blood samples

Pooling Design a contact matrix $\mathbf{M} \in \mathbb{B}^{T\times N}$: $M_{ij} = 1$ if test i uses blood of person j.

Testing Realize T tests in parallel: results $\mathbf{y} \in \mathbb{B}^{T\times 1}$ which depend on $\{\mathbf{M}_j | j \in \mathcal{K}\}$
If the testing procedure is perfect: $\mathbf{y} = \mathbf{M} \otimes \mathbf{x}$ (where $\mathbf{x}$ the indicator vector).
In practice:

q: False positive probability	Test is positive whereas no infected triggers it.
u: Dilution factor.	One infected triggers the test with probability $(1-u)$.

Decoding Identify the infected persons: binary vector $\hat{\mathbf{x}} \in \mathbb{B}^{T\times 1}$.

Goal Minimize the number of tests T.

What is Traitor Tracing?

Content Security: Identify a small set of dishonest users illegally distributing their copies of Video-on-Demand movies. Embed the user's codeword in his content copy (versioning by watermarking).

Setup N: number of a VoD portal users, K: number of colluders, T: bits in codeword

Coding Design a binary code matrix $\mathbf{M} \in \mathbb{B}^{T\times N}$.

Collusion The colluders mix their copies to forge a pirated copy. The watermark decoder retrieves the pirated sequence $\mathbf{y} \in \mathbb{B}^{T\times 1}$. Marking assumption: $y_i \in \{M_{ij_1}, \cdots, M_{ij_K}\}$.

Decoding Identify the colluders: binary vector $\hat{\mathbf{x}} \in \mathbb{B}^{N\times 1}$.

Goal Minimize the number of bits T to be embedded in the content.

Differences

Requirements What does matter is ...

GT: Probability of false negative $\rightarrow$ Missing at least one infected patient.
TT: Probability of false positive $\rightarrow$ Avoid accusing at least one innocent user.

Nuisance parameters What do we know?
GT: K is unknown, but (u, q) are accurately measured (depends on biological test).
TT: Collusion strategy is unknown, but, $y_i = x$ if $M_{ij_1} = \cdots = M_{ij_K} = x$.

TT is a harder problem than GT: $T = O(K^2 \log N)$ versus $T = O(K \log N)$

Similarities

$$\begin{array}{ccccccc} \mathbf{M}_{j_1} & = & M_{1j_1} & M_{2j_1} & \cdots & M_{Tj_1} \\ \mathbf{M}_{j_2} & = & M_{1j_2} & M_{2j_2} & \cdots & M_{Tj_2} \\ \vdots & = & \vdots & \vdots & \vdots & \vdots \\ \mathbf{M}_{j_K} & = & M_{1j_K} & M_{2j_K} & \cdots & M_{Tj_K} \\ \hline \mathbf{y} & = & y_1 & y_2 & \cdots & y_T \end{array}$$

Mathematical Model How is $\mathbf{y}$ related to the codewords $\{\mathbf{M}_j | j \in \mathcal{K}\}$?
$\Longrightarrow$ Think of $\mathbf{y}$ as a random vector.

TT: Collusion strategy $\boldsymbol{\theta}$ s.t. $\theta_k = \mathbb{P}[Y_i = 1 | \sum_{j\in\mathcal{K}} M_{ij} = k]$
GT: The same model holds. $\theta_k = \mathbb{P}[Y_i = 1 | \sum_{j\in\mathcal{K}} M_{ij} = k] = 1 - (1-q)u^k$.

Application of TT methods to GT

Generation of Matrix $\mathbf{M}$

In TT, the Tardos Code [1] is the optimum code construction: matrix $\mathbf{M}$ is randomly drawn!

1. Randomly draw T variables $p_i \overset{\text{i.i.d}}{\sim} f(p)$ with $f(p) : (0,1) \rightarrow \mathbb{R}^+$
2. Randomly draw M_{ij} s.t. $\mathbb{P}(M_{ij} = 1) = p_i$

Probabilities

Thanks to the probabilistic construction of $\mathbf{M}$ and the mathematical model based on $\boldsymbol{\theta}$:

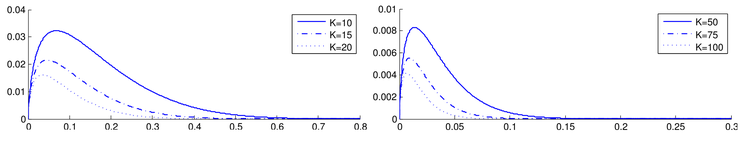
$$\begin{aligned} \mathbb{P}(Y = 1 | p, K) &= \sum_{k=0}^{K} \mathbb{P}(Y = 1 | k \text{ infected }) \cdot \mathbb{P}(k \text{ infected } | p, K) \\ &= \sum_{k=0}^{K} \theta(k) \binom{K}{k} p^k (1-p)^{(K-k)} = 1 - (1-q)(1-p+up)^K. \end{aligned}$$

There are similar expressions for the following cases:

We know the identity of one infected:	$\mathbb{P}(Y_i = 1 \| M_{ij}, p_i, K)$
We know the identities of ℓ infected:	$\mathbb{P}(Y_i = 1 \| \Sigma_i, p_i, K)$ with $\Sigma_i = (M_{ij_1}, \cdots, M_{ij_\ell})$

Mutual Information

This allows us to compute $I(Y; X|p, K)$ and to find $p^\star(K) = \arg\max I(Y; X|p)$.
But we do not know K. Assume that $K \in [\underline{K}, \overline{K}]$, and choose $f = \mathbb{U}_{[p^\star(\underline{K}), p^\star(\overline{K})]}$.



$I(Y; X|p)$ in nats as a function of p. (left) $(q, u) = (0, 0.2)$, (right) $(q, u) = (0.01, 0.05)$.

Funded by French national project MEDIEVALS ANR-07-AM-005.

Decoding

Estimation of K If $(u, q) \neq (1, 1)$, then identifiable: $\hat{K} = \arg\max \sum_{i=1}^{T} \log \mathbb{P}(Y = y_i | p_i, K)$

Single decoder For each user, test the following hypothesis:

$\mathcal{H}_0$	Patient i is not infected:	$\mathbb{P}(\mathbf{Y}, \mathbf{M}_j \| \mathbf{P}, K) = \mathbb{P}(\mathbf{Y} \| \mathbf{P}, K) \cdot \mathbb{P}(\mathbf{M}_j \| \mathbf{P})$
$\mathcal{H}_1$	Patient i is infected:	$\mathbb{P}(\mathbf{Y}, \mathbf{M}_j \| \mathbf{P}, K) = \mathbb{P}(\mathbf{Y} \| \mathbf{M}_j, \mathbf{P}, K) \cdot \mathbb{P}(\mathbf{M}_j \| \mathbf{P})$

Score based on Log-Likelihood Ratio: $s_j = \sum_{i=1}^{T} \log \frac{\mathbb{P}(y_i | M_{ij}, p_i, \hat{K})}{\mathbb{P}(y_i | p_i, \hat{K})}$
Patients with the highest scores are more likely to be infected.

Joint decoder Compute scores for subsets of ℓ patients.
Inf. Theory tells scores more discriminative, but never done before because of complexity $O(N^\ell)$.

$$s_k = \sum_{i=1}^{T} \log \frac{\mathbb{P}(y_i | \Sigma_{ik}, p, \hat{K})}{\mathbb{P}(y_i | p_i, \hat{K})} \text{ with } \Sigma_{ik} = (M_{ij_1}, \cdots, M_{ij_\ell})$$

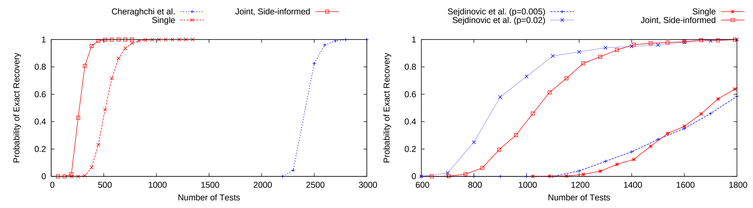
1. Single decoder over population, and isolate $\lceil \sqrt{2!N} \rceil$ persons with highest scores in $\mathcal{S}^{(2)}$
2. Pair decoder over $\mathcal{S}^{(2)}$ and isolate $\lceil \sqrt[3]{3!N} \rceil$ persons with highest scores in $\mathcal{S}^{(3)}$
3. Triple decoder over $\mathcal{S}^{(3)}$...

This idea is to gradually discard the less likely infected while maintaining a list of suspects short enough to allow joint decoding with bigger subsets.

Side-Informed decoders Deem as infected the most likely individuals and include them in the side-information set $\mathcal{SI}$. Denote $\Xi_i = \{M_{ij} | j \in \mathcal{SI}\}$.

$$s_k = \sum_{i=1}^{T} \log \frac{\mathbb{P}(y_i | \Sigma_{ik} \bigcup \Xi_i, p, \hat{K})}{\mathbb{P}(y_i | \Xi_i, p_i, \hat{K})} \text{ with } \Sigma_{ik} = (M_{ij_1}, \cdots, M_{ij_\ell})$$

Experiments Comparison with prior art [2, 3, 4].



(left) $N = 10^5$, $K = 10$, $(q, u) = (0, 0.2)$ [2]; (right) $N = 5000$, $K = 50$, $(q, u) = (0.01, 0.05)$ [3, 4]

References

[1] G. Tardos, "Optimal probabilistic fingerprint codes," in *Proc. 35th ACM Symposium on Theory of Computing*, San Diego, CA, USA, 2003, pp. 116–125.

[2] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli, "Group testing with probabilistic tests: Theory, design and application," in *Proc. 47th Allerton Conf. on Commun., Control and Computing*, Monticello, IL, USA, Sept. 2009, ext. version on arXiv:1009.3186v1.

[3] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *submitted to IEEE Trans. Inf. Theory*, 2009, arXiv:0907.1061v3.

[4] D. Sejdinovic and O. Johnson, "Note on noisy group testing: asymptotic bounds and belief propagation reconstruction," in *Proc. 48th Allerton Conf. on Commun., Control and Computing*, Monticello, IL, USA, Oct. 2010, arXiv:1010.2441v1.