Group Testing – a tool of protecting Network Security

Hung-Lin Fu 傅 恆 霖

Department of Applied Mathematics,
National Chiao Tung University, Hsin Chu,
Taiwan

Group testing (General Model)

- Consider a set *N* of *n* items consisting of at most *d* *positive* (used to be called defective) items with the others being *negative* (used to be called good) items.
- A group test, sometimes called a *pool*, can be applied to an arbitrary set S of items with two possible outcomes; *negative:* all items in S are negative*; positive:* at least one positive item in S, not knowing which one or how many.

Adaptive (Sequential) and Non-adaptive(Parallel) Algorithms

- *Adaptive Algorithm*: You ask the second question (query) after knowing the answer of the first one and continue …. That is, the previous knowledge will be used later.
- *Non-adaptive* Algorithm: All the queries are designed beforehand and then you ask all the questions (queries) *simultaneously.*

Algorithms

- Adaptive algorithm
- Non-adaptive algorithm
- k-stage algorithm

The most popular one is a 2-stage algorithm in which we use a non-adaptive algorithm first and then in the second stage we test the left "suspected" items one by one.

Save Money or Time

- An adaptive (sequential) algorithm conducts the tests *one by one* and the outcomes of all previous tests can be used to set up the later test. (*Save money!?*)
- A non-adaptive algorithm specifies a set of tests in advance so that they can be conducted *simultaneously*; thus forbidding using the information of previous tests. (*Save time!*)
- ***In general, an adaptive algorithm takes less queries (compare to a non-adaptive algorithm) to get the job done.***

Adaptive Algorithm

- By the nature of this algorithm, we normally find all the positives by way of a search algorithm step by step.
- The most popular one is the splitting algorithm which takes around $log_2 n$ steps to locate a positive items out of n items.
- In fact, this answer is best possible for one positive item by using the information lower bound.

Non-adaptive Algorithm

- A non-adaptive algorithm can be represented by a (0,1) – matrix $M = [m_{i,j}]$ such that columns are items, rows are tests and $m_{i,j}$ = 1 if and only if the jth item is "*involved*" in the ith test.
- The matrix represents a non-adaptive algorithm is also known as a "***pooling design***".

An example (12 items and 9 tests)

1			1			1			1		
1				1			1			1	
1					1			1			1
	1		1					1		1	
	1			1		1					1
	1				1		1		1		
		1	1				1				1
		1		1				1	1		
		1			1	1				1	

The blanks are zeros.

Set Notations

- Let $M = [m_{i,j}]$ be a txn matrix mentioned above. Then we can use n sets (ordered) S_i's to represent the matrix where

$S_k = \{i : m_{i,k} = 1, i = 1, 2, \ldots, t\}, k = 1, 2, \ldots, n.$

- The following sets represent the above (0,1)-matrix:

{1,2,3}, {4,5,6}, {7,8,9}, {1,4,7}, {2,5,8}, {3,6,9}, {1,5,9}, {2,6,7}, {3,4,8}, {1,6,8}, {2,4,9}, {3,5,7}.

Relation with Designs and Codes

- It is easier to apply combinatorial structures to construct pooling designs, therefore, we use sets (or codewords) for columns in non-adaptive algorithms.
- The set corresponding to a codeword (binary vector) is called the *support* of the codeword.

Various Models

- There are *inhibitors*.
- There is a *threshold*.
- Defective items are sets: *complex* model.
- *Competitive* model: the number of defectives is unknown.
- The defectives are *mutually obscuring*.
- *Size constraint group testing*.

Further Remarks

- We can approach this study via different topics such as *combinatorial design, algebraic combinatorics, coding theory and graph theory*.
- Group testing does play an important role in applications such as *computational molecular biology*, ***network security***, *image compression*, ..., etc.

Compressed Sensing

- In compressed sensing (CS), we are given an n-dimensional sparse signal with support size K.
- *Random (?)* projections of the sparse signal are obtained. (We notice the projection if it is not orthogonal to the signal, i.e. the inner product of the two vectors is not equal to zero.)
- *The goal is to identify the support set while minimizing the number of projections.*

Outcome Vectors

- The vector **y** is an *outcome vector* which is corresponding to an input **x**.
- If M is 1-1, then we can decode **x** as long as we know its outcome vector.
- In order to get the job done with lower decoding complexity, extra properties for m is needed. ***(This is the combinatorial part!)***

Signal: 010000010000

Outcome vector: 010111100

1			1			1			1		
1				1			1			1	
1					1			1			1
	1		1					1		1	
	1			1		1					1
	1				1		1		1		
		1	1				1				1
		1		1				1	1		
		1			1	1				1	

The blanks are zeros.

Network Security

- This is a topic which is too wide and large to be well studied.
- We can only pick up certain parts of network to see if we can do something about them.
- I am a combinatorial people and I know not much about network.
- We started to notice that group testing can play an important role on network security not long time ago, *I am learning!*

An Application of Size Constraint GT

- A size constraint GT is a pooling design such that both of the number of rows and the number of 1's in each row are bounded by fixed constants respectively. For example, a pooling design has *at most K rows* and each row is *of weight at most w*.
- This design is working for *Denial-of-Service (DoS) attack.*

DoS Attack

- The Denial-of-Service (DoS) attack is aiming at disrupting application service rather than depleting the network resource.
- Owing to its high similarity to legitimate traffic and much lower launching overhead than classic DoS attack, this new assault type cannot be efficiently detected or prevented by existing detection solutions.

Continued

- The key problem to identifying attackers lies in how to group clients together on each server so that if a server is under attack, we can quickly identify these attackers without examining each request.
- Since the number of available pools (virtual servers) and the number of clients/requests in a pool (server) has its limitation, an SCGT is required.

The Challenges

(1) How to construct a pooling design to enable prompt and accurate detection.

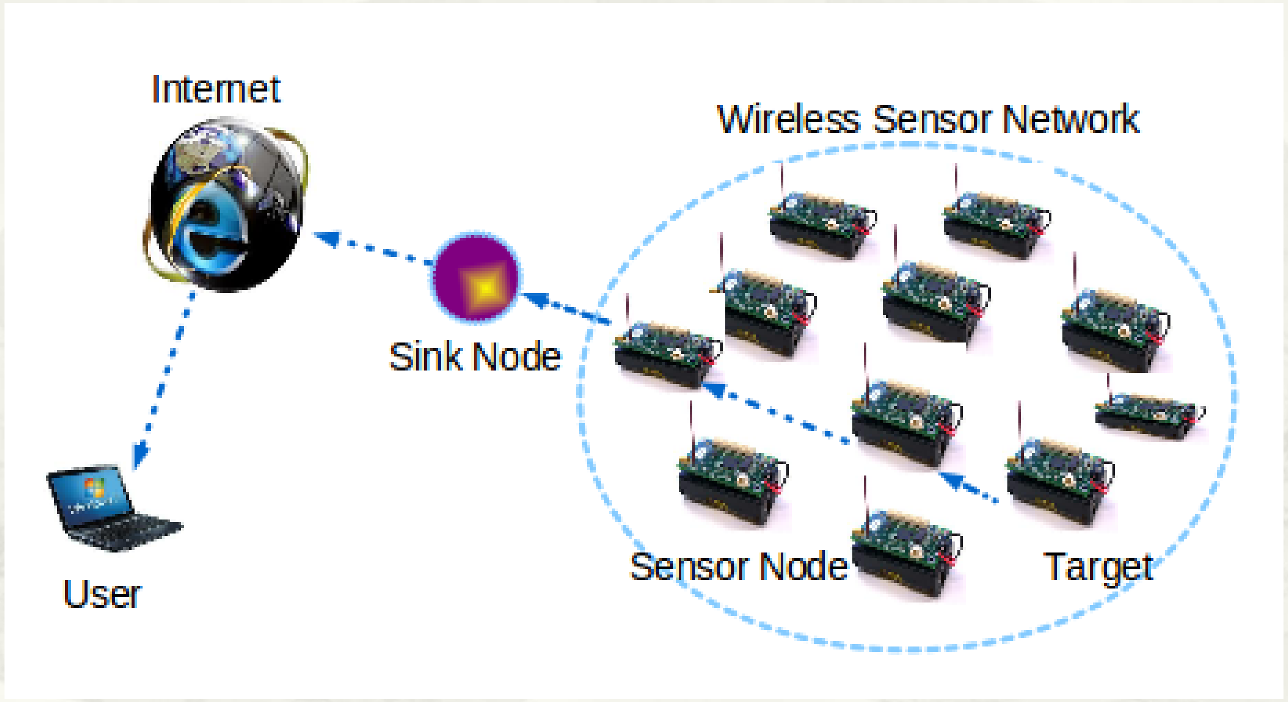
(2) How to regulate the service requests to match the design in a practical system.

(3) How to establish proper thresholds for server source usage indicator to generate accurate test outcomes.

Interference Free GT and Reactive Jamming Attacks

- Reactive Jamming attack has a great security threat to wireless sensor networks (WSN).
- There exist many studies against these attacks, however, these methods such as frequency hopping or channel surfing, require excessive computational capabilities on wireless devices.

Wireless Sensor Network



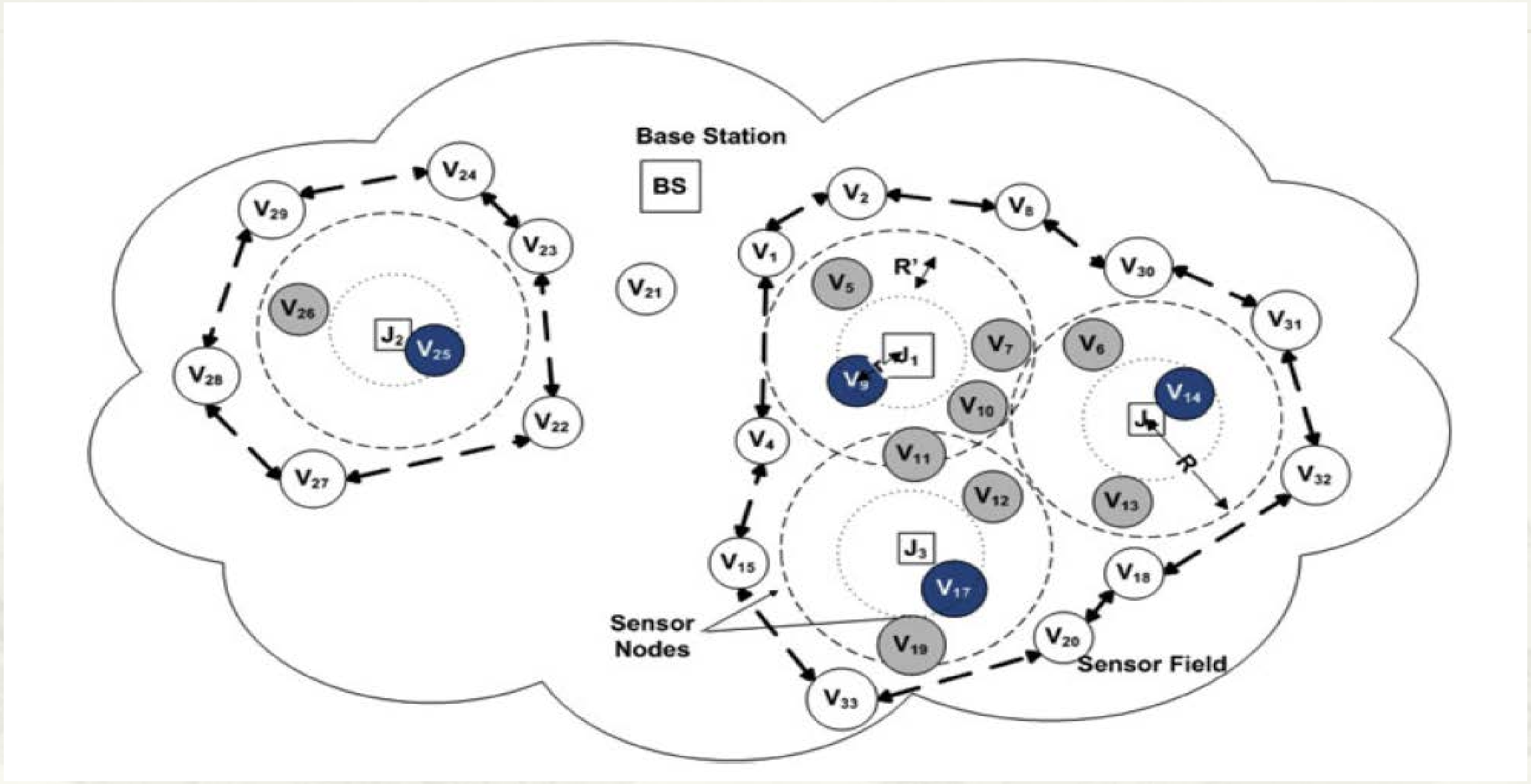
Key Idea

- We can apply an interference free group testing to identify the *trigger nodes*, whose transmissions activate any *reactive jammers*.
- The identification of these trigger nodes can help us to (1) design a better routing protocol by switching these nodes into only receivers to avoid activating jammers and (2) locate the jammers based on the trigger nodes, thus providing a mechanism against reactive jamming attacks.

Group Testing Works Here

- There are three types of nodes: (1) *Victim nodes* (VN): if a node v hears a jamming signal, then v is VN; (2) *Unaffected nodes* (UN): nodes are not being jammed, i.e., cannot hear the jamming signals; and (3) *Boundary nodes* (BN): vis said a BN if v is a UN who has a neighbor node as a victim.
- The purpose of BNs is to estimate the jamming range R where the VN set is to be tested to obtain the trigger nodes.

Wireless Sensor Network



Nodes in grey and blue are VNs around jammer nodes, where blue nodes are also trigger nodes, which invoke the jammer nodes. Nodes surrounding the jammed area are BNs, while the others are UNs.

For more details about the above two applications, please refer to the book "Group Testing Theory in Network Security" written by My T. Thai.

Multimedia Fingerprintings

- To prevent attacks from unauthorized users we *encrypt the copyrighted information*.
- To hinder attacks from authorized users we *fingerprint the copyrighted information*.
- Normally, fingerprints are embedded (through watermarking techniques) into multimedia contents.
- Information is represented by waveforms, which are viewed as ***vectors*** in signal space.

Fingerprinting Code: An Example

- Copyrighted sequence: 0101010101010101010100011
- Short sequences: 0001, 1010
- Fingerprinted sequence:

  0101110101010111010100011

Pirate sequences:

  0101010101010101010100011

  0101110100010111010100011

Copyrighted Information

Collusion Attacks

- The *averaging attack* (average multiple copies of the authorized content together) is the most feasible approach to perform collusion attacks.
- *This attack reduces the power of each contributing fingerprint (fairness between colluders) and makes the colluded signal have perceptual quality.*

Digital Fingerprinting Codes

- We try to find efficient *fingerprinting* codes which is effective for colluder-tracing.
- This is equivalent to require a good code which can be used by more users (large number of codewords) and also capable of tracing as many colluders as possible, moreover, the tracing algorithm had better with low complexity.

Apply Pooling Designs

- The idea of tracing colluders is similar to finding positive items in group testing.
- In group testing, in order to find d positive items, we use a *d-disjunct* matrix which has faster decoding algorithm or a *d-separable* matrix which has decoding algorithm with higher complexity.
- These matrices are (0,1) – matrices.
- How about [0, q-1] – matrices?
- *Find good anti-collusion codes!*

References for Fingerprinting Codes

1. M. Cheng, H.-L. Fu, J. Jiang, Y.-H. Lo and Y. Miao, Tracking codes: Another kind of codes with the identifiable parent property, submitted.
2. M. Cheng and Y. Miao, On anti-collusion codes and detection algorithms for multimedia fingerprinting, IEEE Trans. Inform. Theory 57 (2011), 4843 – 4851.
3. A. Barg, G.R. Blakley and G. Kabatiansky, Digital fingerprinting codes: Problem statements, constructions, identification of traitors, IEEE Trans. Inform. Theory 49 (2003), 852 – 865.

Keep Moving Forward!