Identification of Test Collusion by the Methods of Information Theory and Combinatorial Optimization

Dmitry Belov

Law School Admission Council

Test collusion

Test collusion may be described as large-scale sharing of test materials or answers to test items prior to or during the exam.

An indication of test collusion:
Examinees involved in test collusion have unusual difference in performance between two subsets of items: items for which collusion took place and items for which collusion did not take place.

Large scale item preknowledge

At some test center (**affected test center**) a group of examinees (**aberrant examinees**) had access to answers of some items (**compromised items**) from an administered test prior to exam.
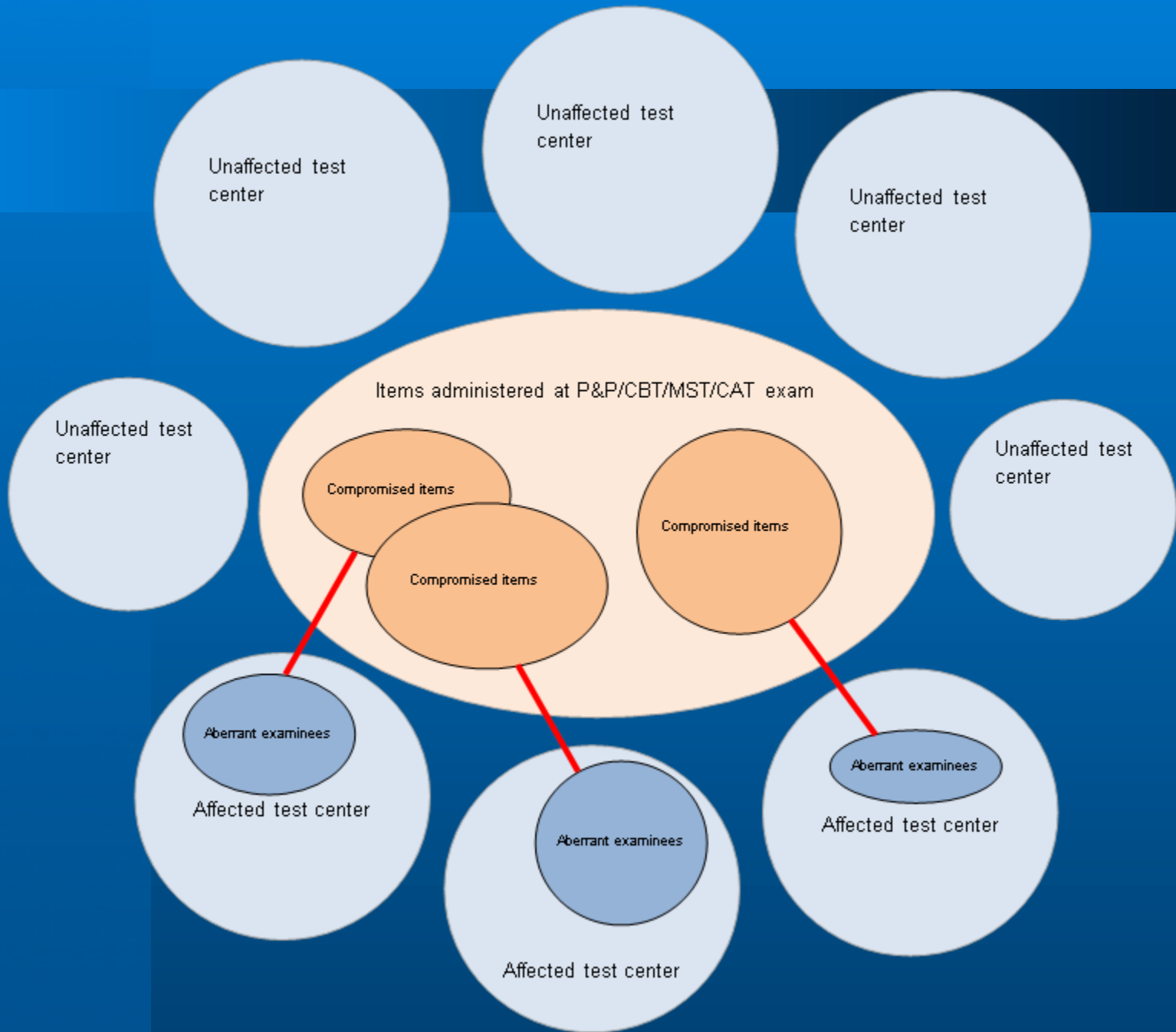
P&P, CBT, MST, CAT

An indication of large scale item preknowledge:
At each affected test center its aberrant examinees perform on compromised items unusually better than on uncompromised items.

Affected test centers (unknown)
Aberrant examinees (unknown)
Compromised items (unknown)

How to disentangle the issue with three unknowns (**Affected test centers**, **Aberrant examinees**, **Compromised items**)?

Current detectors (FLOR, KLD, …) assume that the compromised subset is known. They perform well when this assumption holds (e.g., compromised subset is known or compromised subset is covered by a known collection of subsets). In practice this assumption is violated which leads to a dramatic loss of power.

In general, item preknowledge is hard to detect within statistical framework only. At the same time, the problem is clearly combinatorial because of interactions between subsets of items, examinees, and test centers. Is it possible to disentangle the issue with three unknowns by applying the combinatorial optimization?

3D Algorithm
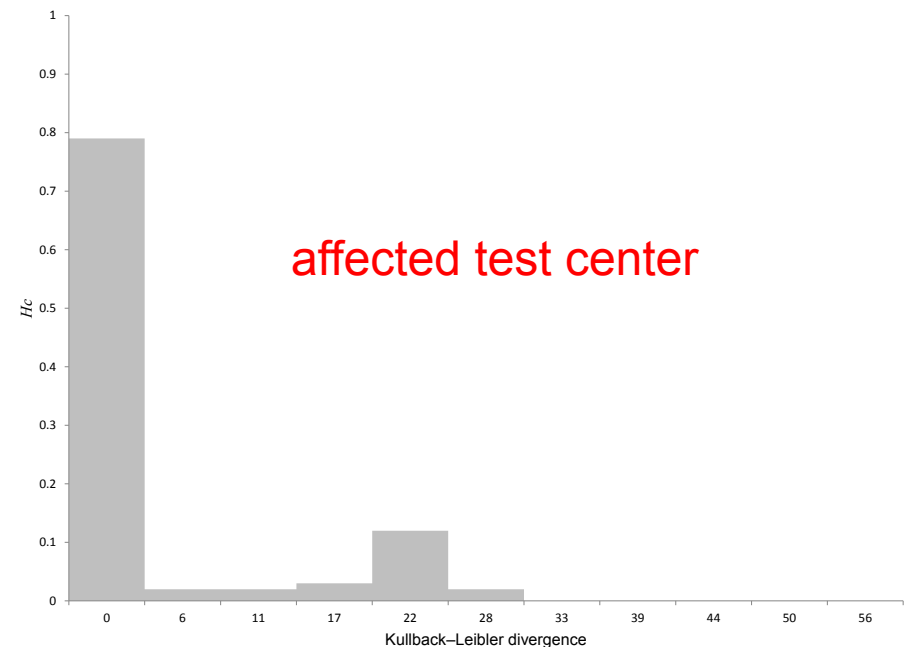
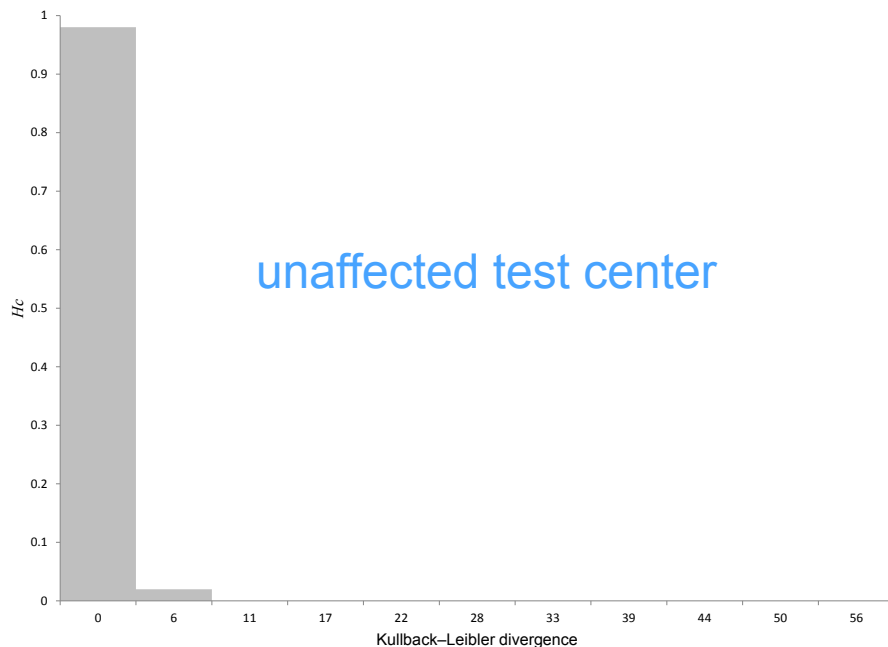**D**etect affected test centers

Detect compromised items for each affected test center using combinatorial search

Detect aberrant examinees for each affected test center and corresponding compromised items

Step 1: Detect affected test centers

Suppose compromised items are known.

If a test center is affected (i.e., has aberrant examinees) then the distribution of a person-fit statistic (FLOR, KLD, ...) computed at this test center should be unusual among distributions computed at unaffected test centers.

r is analyzed test center S is a random subset of items

X is a subset of unaffected test centers

$$g_{r,S} = \sum_{x \in X} \Big(\mathbf{D}(\mathbf{H}_{r,S} \parallel \mathbf{H}_{x,S}) + \mathbf{D}(\mathbf{H}_{x,S} \parallel \mathbf{H}_{r,S}) \Big)$$

Kullback–Leibler divergence:

$$\mathbf{D}(\mathbf{H}_1 \parallel \mathbf{H}_2) = \sum_{i=1}^{k} \mathbf{H}_1(z_i) \ln \frac{\mathbf{H}_1(z_i)}{\mathbf{H}_2(z_i)}$$

$$\mathbf{D}(\mathbf{H}_1 \parallel \mathbf{H}_2) \geq 0$$

$$\mathbf{D}(\mathbf{H}_1 \parallel \mathbf{H}_2) = 0 \Leftrightarrow \mathbf{H}_1 \equiv \mathbf{H}_2$$

$$\mathbf{D}(\mathbf{H}_1 \parallel \mathbf{H}_2) \neq \mathbf{D}(\mathbf{H}_2 \parallel \mathbf{H}_1)$$

$$\mathbf{G}_{r,S} = \frac{g_{r,S}}{\sum_{x \in X \cup \{r\}} g_{x,S}}$$

random subsets of items $S_1, S_2, \ldots, S_m$

$$c_r = \sum_{i=1}^{m} \mathbf{G}_{r,S_i}$$

Step 2: Detect compromised items by simulated annealing

Modifying S in order to maximize $\mathbf{G}_{r,S} = \dfrac{g_{r,S}}{\displaystyle\sum_{x \in X \cup \{r\}} g_{x,S}}$

Search space

S contains potentially compromised items

Adding
Removing
Swapping

$\gamma < \mathbf{P}(\mathbf{G}_{r,S} - \mathbf{G}_{r,S*}, t)$

Analysis of Type I Error

3D Algorithm is a sequence of two statistical tests:

1. Detect affected test centers
2. Detect aberrant examinees within each affected test center

If there are 100 test centers with 100 examinees in each test center (10000 total) then the number of falsely detected examinees can be approximated by $100*\alpha 1*100*\alpha 2$

Analysis of detection rate

CAT pool with 500 LSAT items

100 test centers, each with 100 examinees ~ N(0,1)

Simulate CAT with no aberrancy, compute item exposure, and form a **search space** with highly exposed items (items with a high risk to be compromised)

Simulate CAT with aberrancy:

- 10 random test centers are affected
- 10 random subsets of items from the **search space** are assigned to each affected test center as compromised subsets
- in each affected test center, 10 randomly chosen examinees give correct answers to corresponding compromised items
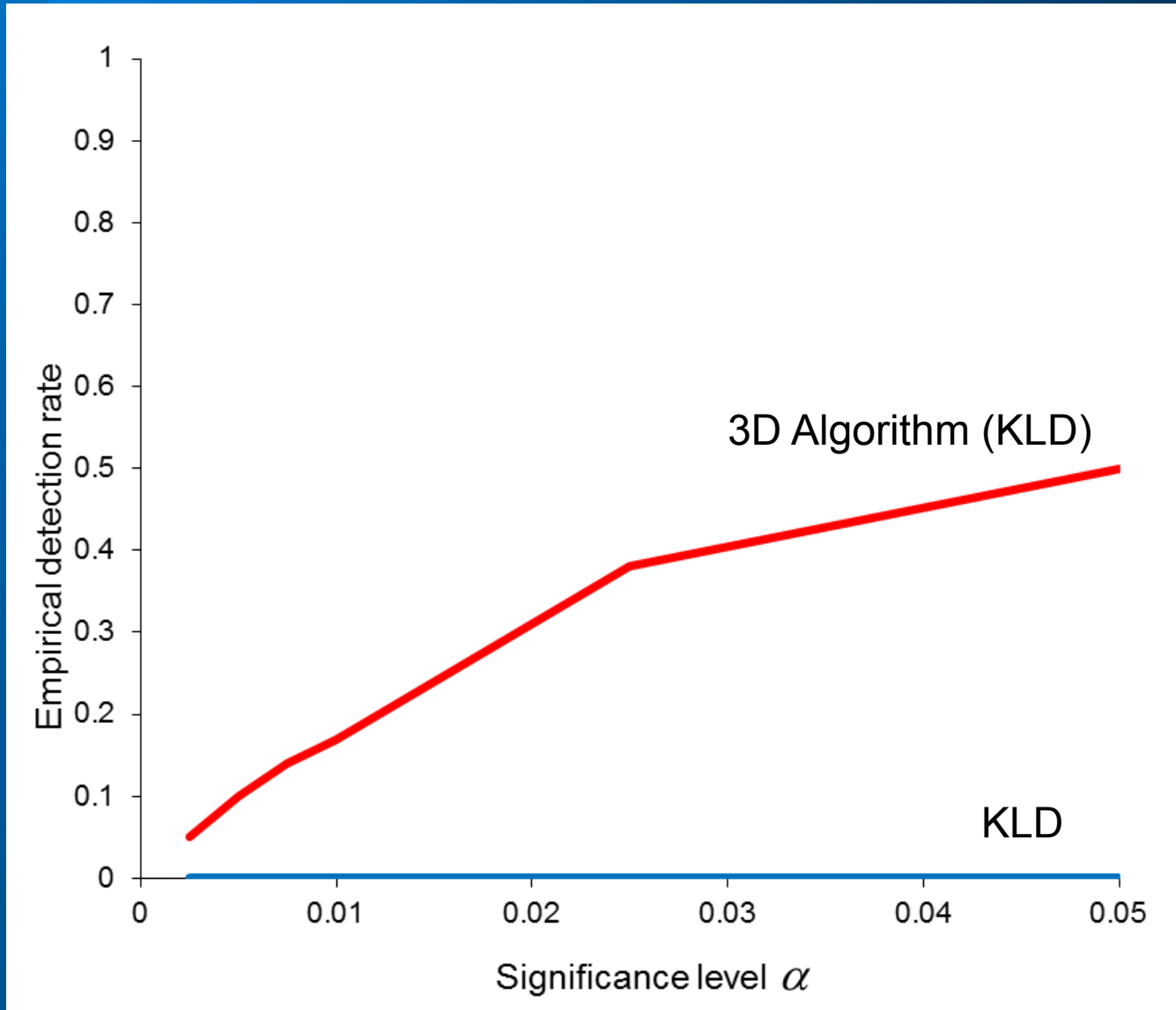
KLD between posteriors of ability = person fit statistic

Case 1: Search space has 51 items (>0.4)
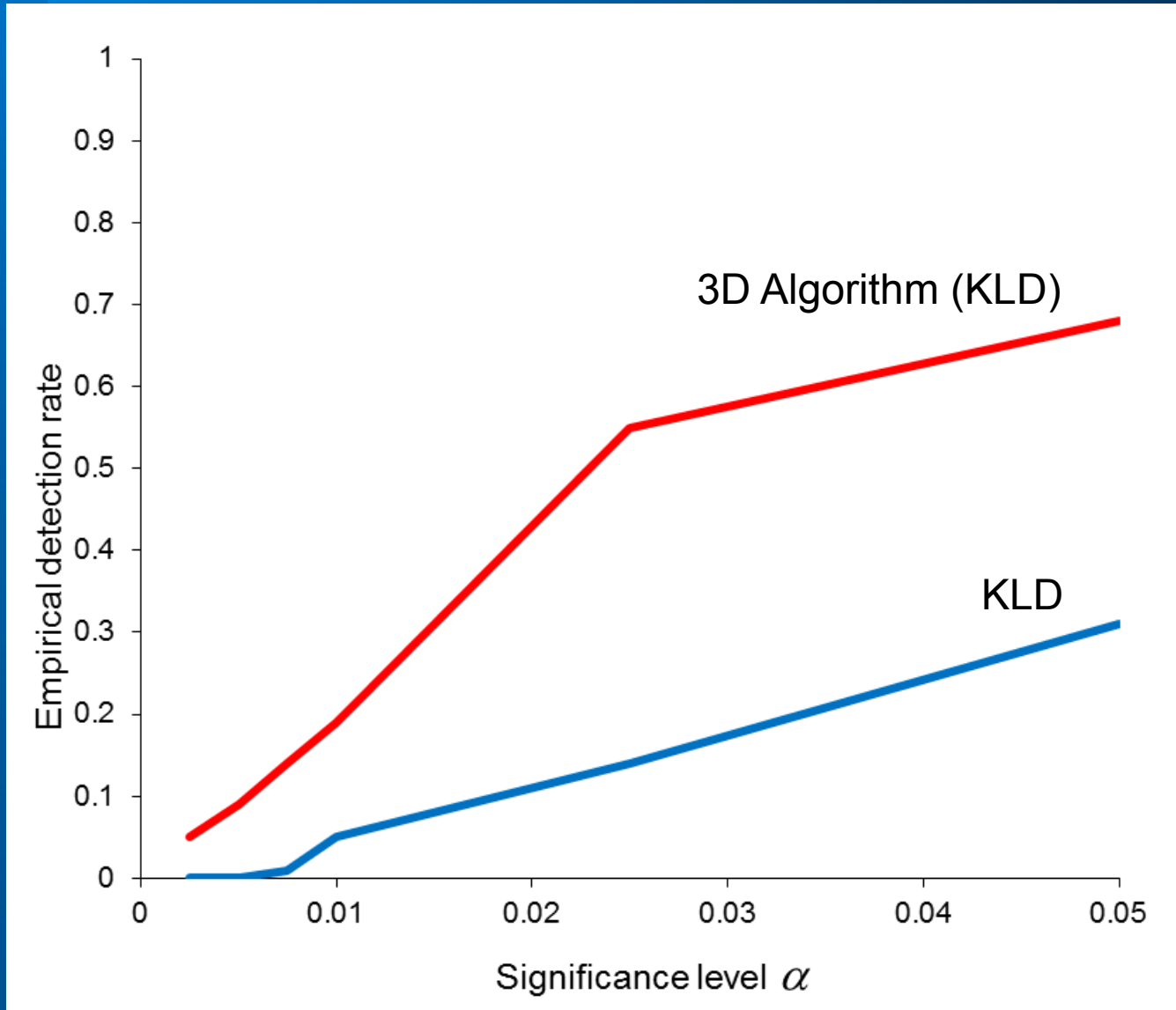
Similar results for
18 different
scenarios
simulating
large scale
item preknowledge

Case 2: Search space has 71 items (>0.3)

Similar results for
18 different
scenarios
simulating
large scale
item preknowledge

3D Algorithm is meta algorithm:

- Applicable to any testing program (P&P, CBT, MST, CAT)

- Various person-fit statistics can be plugged-in measuring performance differences in score and/or response time (FLOR, KLD[ability]+KLD[speed], etc.); thus, the detection of various types of test collusion can be supported

- Combinatorial search for compromised items can employ different approaches (simulated annealing, greedy algorithm, tabu search, genetic algorithm, and many other heuristics)

- Extendable to detect multiple groups of aberrant examinees within affected test center

- Definition of test center can be extended to support various relations between examinees

Test collusion happens at test centers. What is test center?

Common definition of test center is limited by the geographic location (room, class, college, etc.)

However, it can be extended to support other relations (went to same high school, went to same undergraduate college, went to same test-prep center, belong to the same group in a social network, etc.)

This extension allows the detection of groups of examinees involved in test collusion even if they take the actual exam at different geographic locations