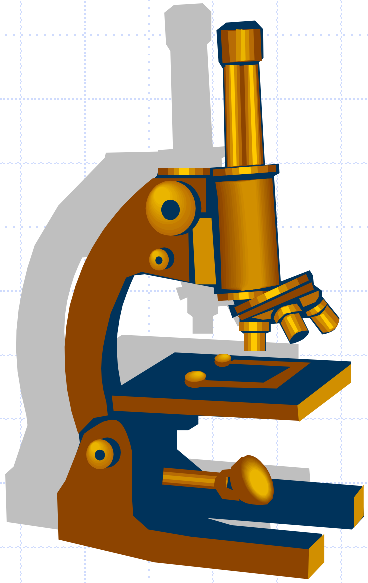
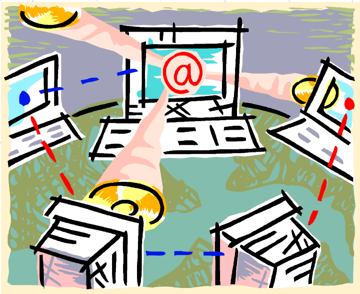
Improved Combinatorial Group Testing for Real-World Problem Sizes



Michael T. Goodrich
Univ. of California, Irvine

joint w/ David Eppstein and Dan Hirschberg

Group Testing



- *Input:* ***n*** items, numbered 0,1, …, n-1, at most ***d*** of which are **defective**.
- *Output:* the indices of all the defective items (or possibly an error condition indicating that more than d items are defective).
- Items can be grouped into arbitrary test subsets, which can be tested in whole to see they contain a defective item or not.

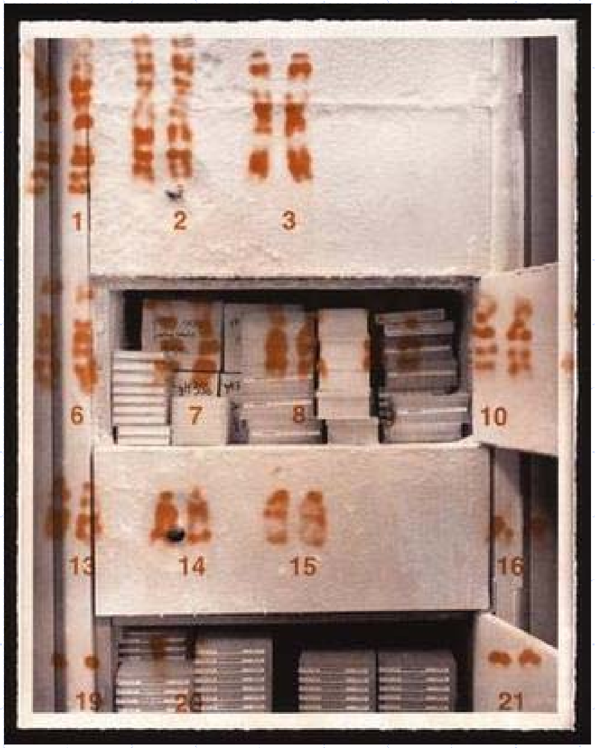
1st Motivation: Blood Testing

- Items are n **blood samples** (in the original problem, they were taken from WWII G.I.'s).
- Drops from different samples are **mixed together** and this mixture is tested for disease antigens.
- **Goal**: minimize the total number of tests

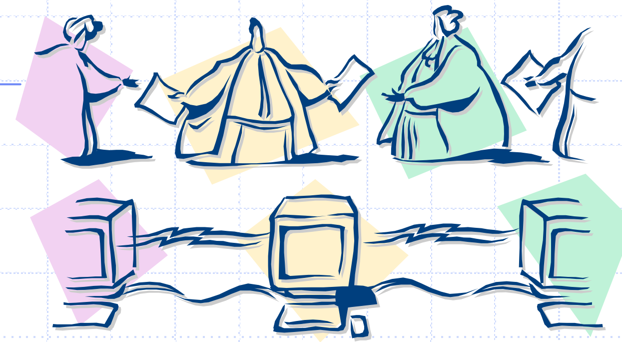
FIGURE 26.—U.S. Army cadets from Marquette University ready to give blood at the Milwaukee, Wis., donor center.

Modern Applications

- Screening vaccines for contamination
- Filtering clone libraries of DNA sequences (identifying which ones contain a certain DNA sequence)
- Computer security – for data forensics
- Computer fault diagnosis

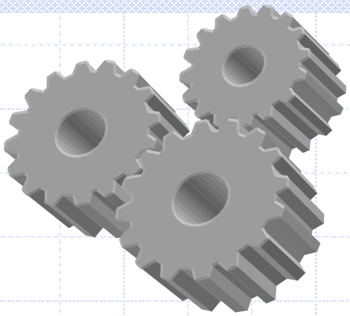


Testing Regimens



- **Non-adaptive**: All tests must be done in parallel
- **Partially adaptive**: Tests can be done in rounds (e.g., 2 rounds), with the tests in each round done in parallel
- **Fully adaptive**: Tests can be done sequentially

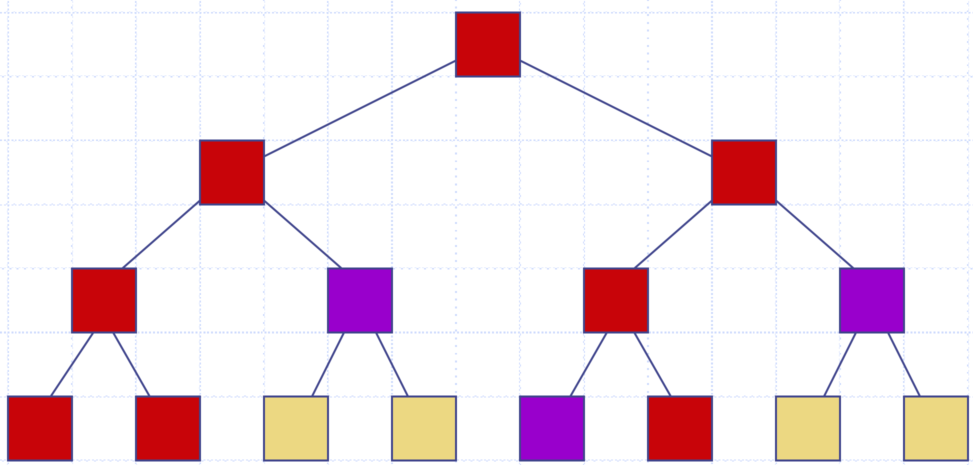
Efficiency Measures



- **t(n,d)** = number of tests to identify up to d defectives among n items.
  - t(n,d) must be $\Omega(\min\{n, d \log (n/d)\})$.
- **A(n,t)** = analysis time needed to determine which items are defective (after the tests are done).
  - time-optimal is A(n,t) is O(t).
- **S(n,d)** = sampling rate – the maximum number of tests in which any item may be included.
  - We would like S(n,d) to be O(t(n,d)/d)

A Simple Test Regimen

- For the fully-adaptive case:
  - Place a complete binary tree “on top of” the items
  - do a top-down search to defectives
  - will use O(d log (n/d)) tests



Another Simple Regimen

- For non-adaptive case when d=1:
  - Consider item numbers in binary
  - Test i is set of items w/ bit i = 1
  - Positive (defective) and negative (non-defective) tests identify the binary index of the defective item
  - t(n,d) is O(log n)
  - d=2 and d=3 cases are much harder…

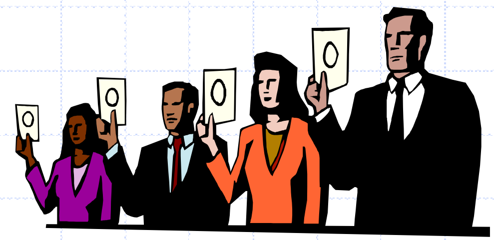
Previous Related Work

- [Du-Hwang, 00] achieve non-adaptive algorithm with t(n,d) being O(d^2log n).
- For two-stage case, [Debonis et al., 03] achieve t(n,d) < 7.5*(d log (n/d))
- For d=2, non-adaptive, [Kautz-Singleton, 64] achieve $t=3^{q+1}$ for $n=3^{2^q}$
- For d=2, non-adaptive, [Macula-Reuter, 98] achieve $t=(q^2+3q)/2$ for $n=2^q-1$
- For d=3, [Du-Hwang, 00] describe an approach that should achieve $t=18q^2-6q$ for $n=2^q-1$.

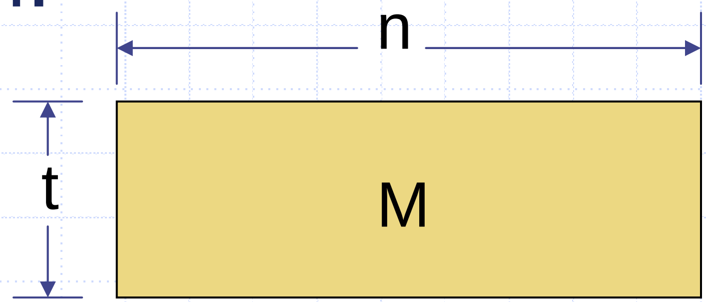
Our Results

- Chinese Remainder Sieve: An improved non-adaptive test regimen for general d and $n<10^{80}$
  - also an improvement for $n<10^{57}$ and small d values
- Rake-and-Winnow: A 2-stage algorithm with a better constant factor (4).
- Impoved (and time-optimal) algorithms for the d=2 and d=3 cases.

Matrix View of Testing

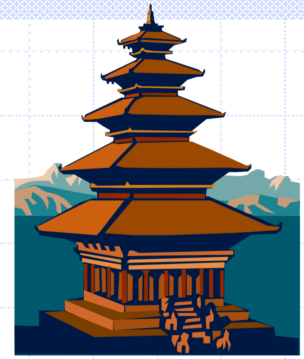


- A non-adaptive testing regimen can be viewed as a t x n binary matrix M:
  - M[i,j] = 1 if and only if test i includes item j



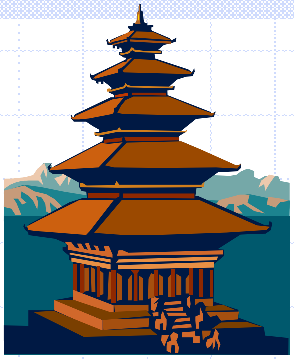
- M is **d-disjunct** if the Boolean sum of any d columns does not contain any other column.
  - An item is defective iff all its tests are positive
- M is **d-separable** if the Boolean sums of each set of at most d columns are distinct (harder analysis algorithm)

Chinese Remainder Sieve



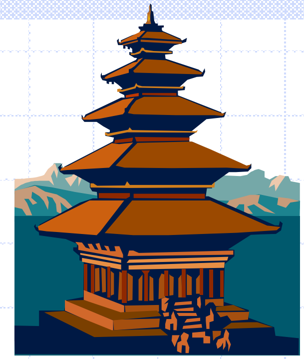
- Let $\{p_1^{e1}, p_2^{e2}, \ldots, p_k^{ek}\}$ be a set of prime powers multiplying to at least n^d.
- Construct a t x n matrix M as a concatenation of k submatrices, where M_j is t_j x n matrix where $t_j = p_j^{ej}$.
  - thus, $t = \Sigma\, p_j^{ej}$.
- Each row q of M_j has a 1 in column m if m mod t_j = q.
  - if q=2 and $t_j = 3^2 = 9$, then row q has 1's in columns 2, 11, 20, … .

Why it Works



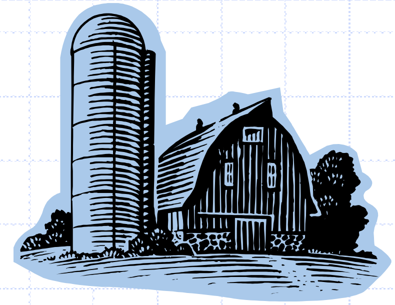
- If all tests are positive (defective) for column i, then i is defective.
  - For each (true) defective item h, let P_h be the product of moduli t_j associated with tests h has in common with i.
  - By a pigeon-hole argument, there is a (true) defective item h such that P_h is at least n.
  - By construction, i is congruent to the same values that h is contruent to, modulo each of the prime powers in P_h.
  - Thus, by Chinese Remainder Theorem, i is equal to h modulo a number that is at least n; hence, i=h.

Analysis



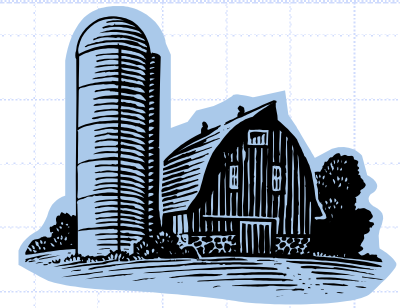
- The number of tests is the sum of the prime products (take each $e_j=1$ for simplicity)
- We need a bound on the sum of primes whose product is at least n^d.
- We show that this sum is at most $(1+o(1))*(2d \ln n)^2/2\ln (2d \ln n)$.
  - Uses a new bound on the sum of primes, which may be of independent interest.

Rake-and-Winnow



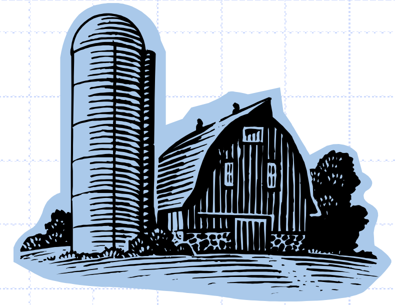
- Uses a randomized approach motivated by Bloom filtering.
- Also uses a matrix M, but in 2 rounds
- Given a set D of d columns in M and a column j, say j is **distinguishable** from D if there is a row i such that M[i,j]=1 but M[i,j']=0 for each j' in D.
- M is **(d,k)-resolvable** if, for any d-sized subset D, there are fewer than k columns that are not distinguishable from D.

The 2-Round Scheme

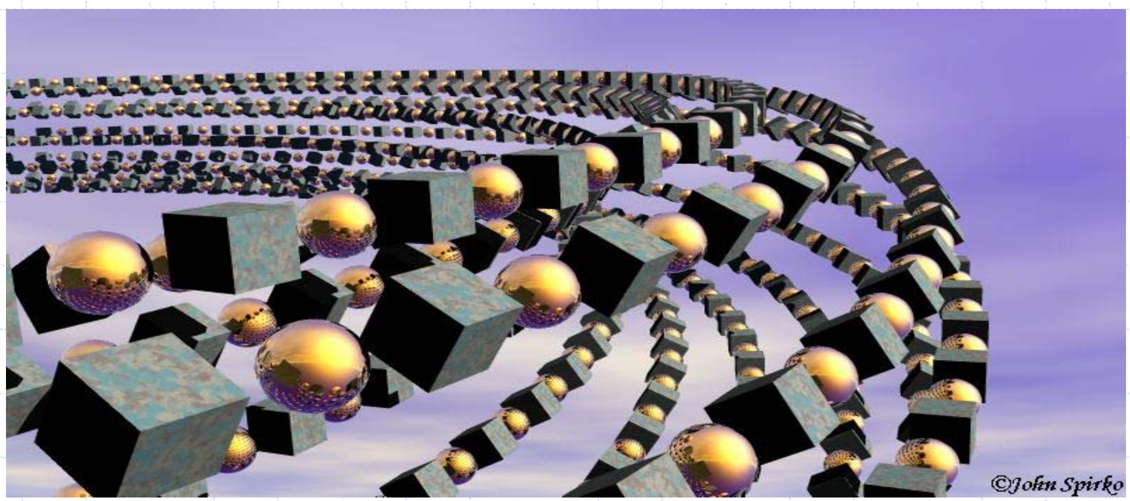


- Use a (d,k)-resolvable matrix M in the first round and make a test for each row.
- Discard all the items in negative (non-defective) tests.
- There are at most d+k remaining items.
- Test each remaining item individually.

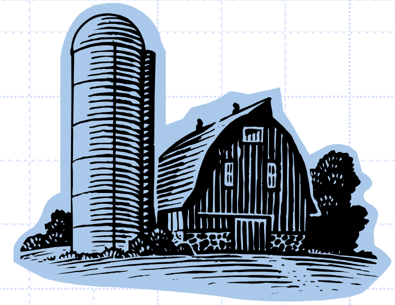
Constructing the Matrix



- Given t (set in the analysis), let M be a 2t x n matrix defined randomly:
  – For each column j, choose t/d rows of M at random and set these entries to 1.
  – that is, we "inject" j into those t/d tests



Analysis



- We show that M will be completely (d,1)-resolvable for any particular choice of D, with high probability, provided $t > 3.7183d \log n$.
  - that is, in practice, this will be a single-round scheme with t being $O(d \log n)$.
- There are a lot of possible D's however.
- Still, we show that if

$$t > 2d \log (en/d) + \log n,$$

  then M will be (d,d)-resolvable with high probability.

Conclusion and Questions

- We have presented improved algorithms for combinatorial group testing for real-world sizes.
- Open problem: design a single non-adaptive scheme that matches or improves our algorithms for small n, while being asymptotically optimal