Optimal Testing of Structured Knowledge

M. Munie and Y. Shoham, AAAI 2008

Presented by Jon Napolitano

Testing Students

- What do they know?
  (Structured Knowledge)

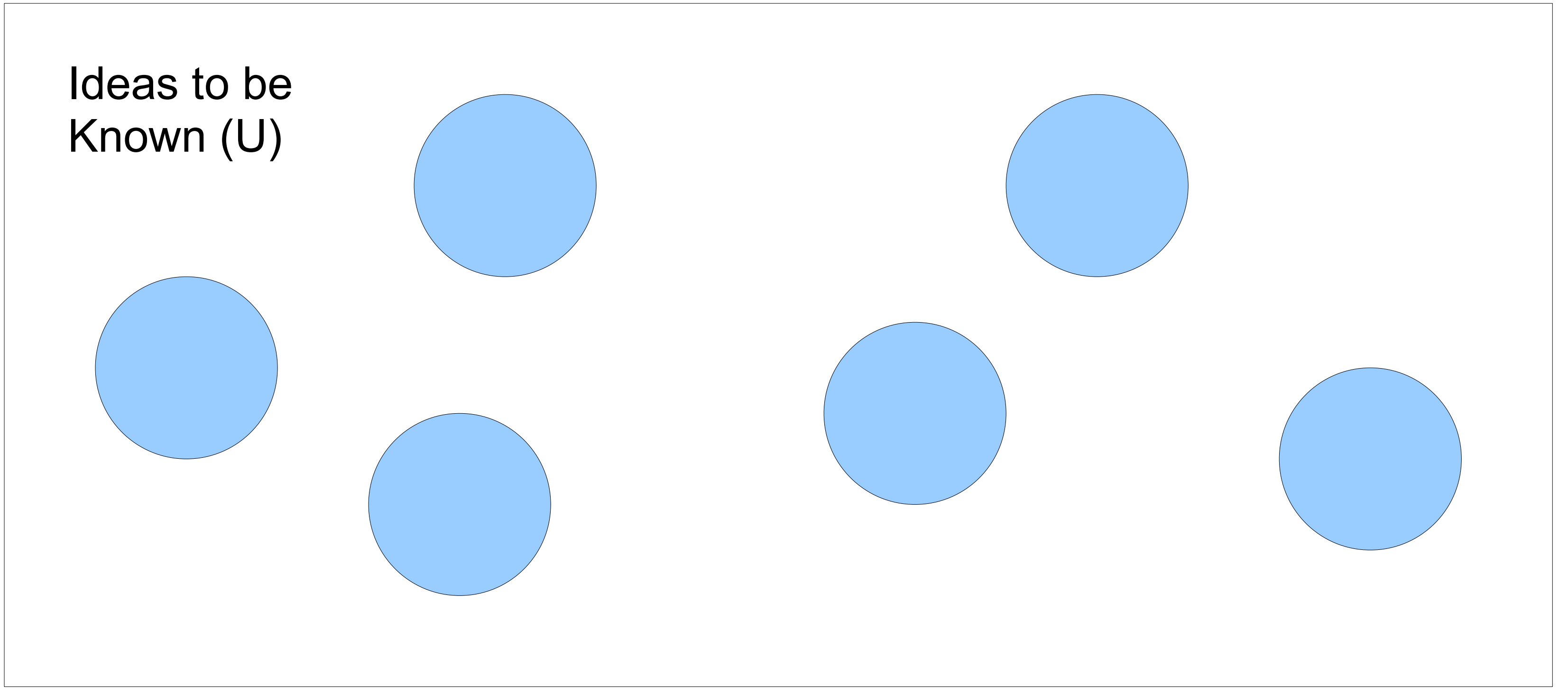
Testing Students

- What do they know?
(Structured Knowledge)
- Paper test or oral exam?
(Parallel or sequential tests)

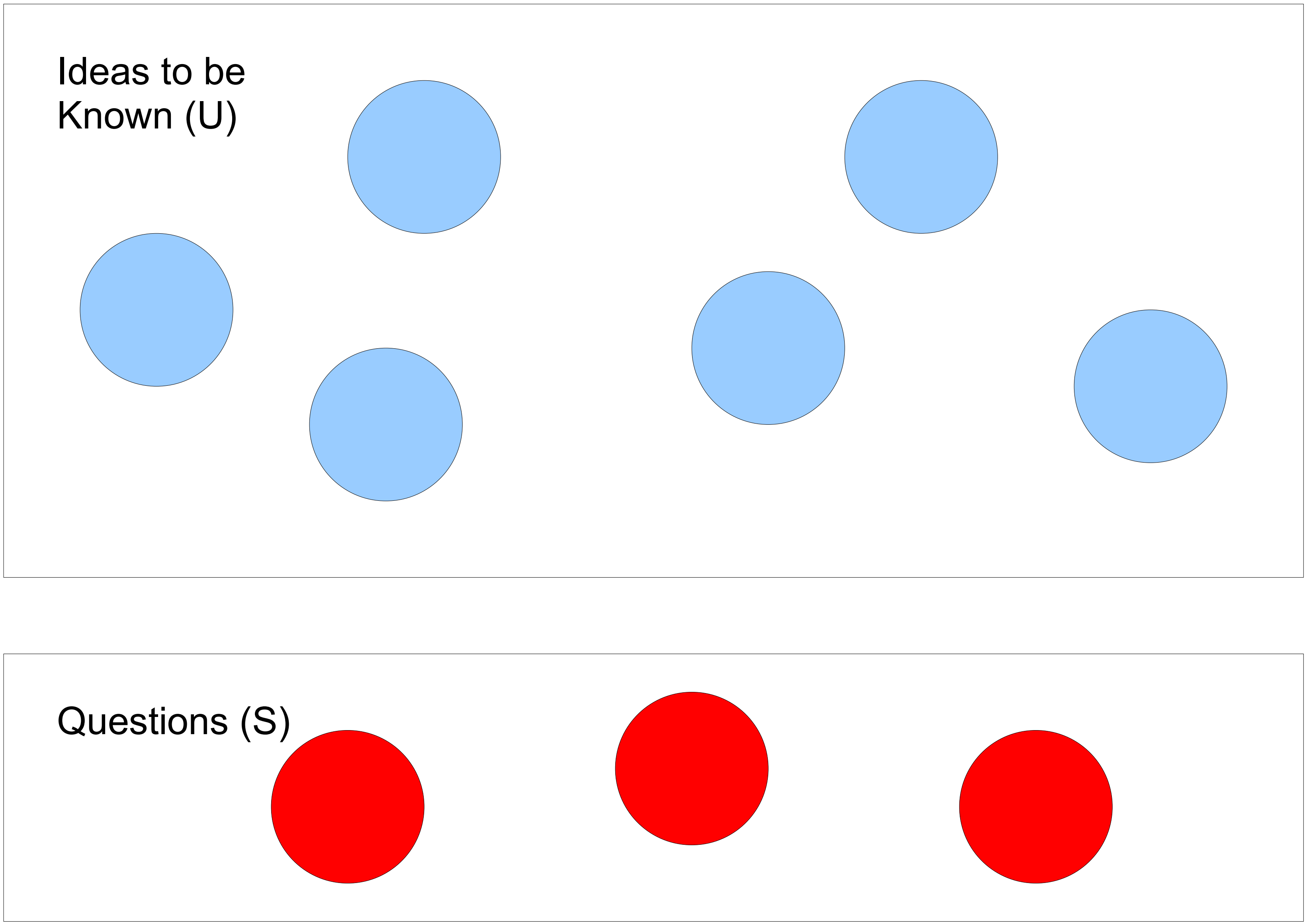
Testing Students

- What do they know?
  (Structured Knowledge)
- Paper test or oral exam?
  (Parallel or sequential tests)
- One Professor, or multiple?
  (Single vs. Multi-agent learning)

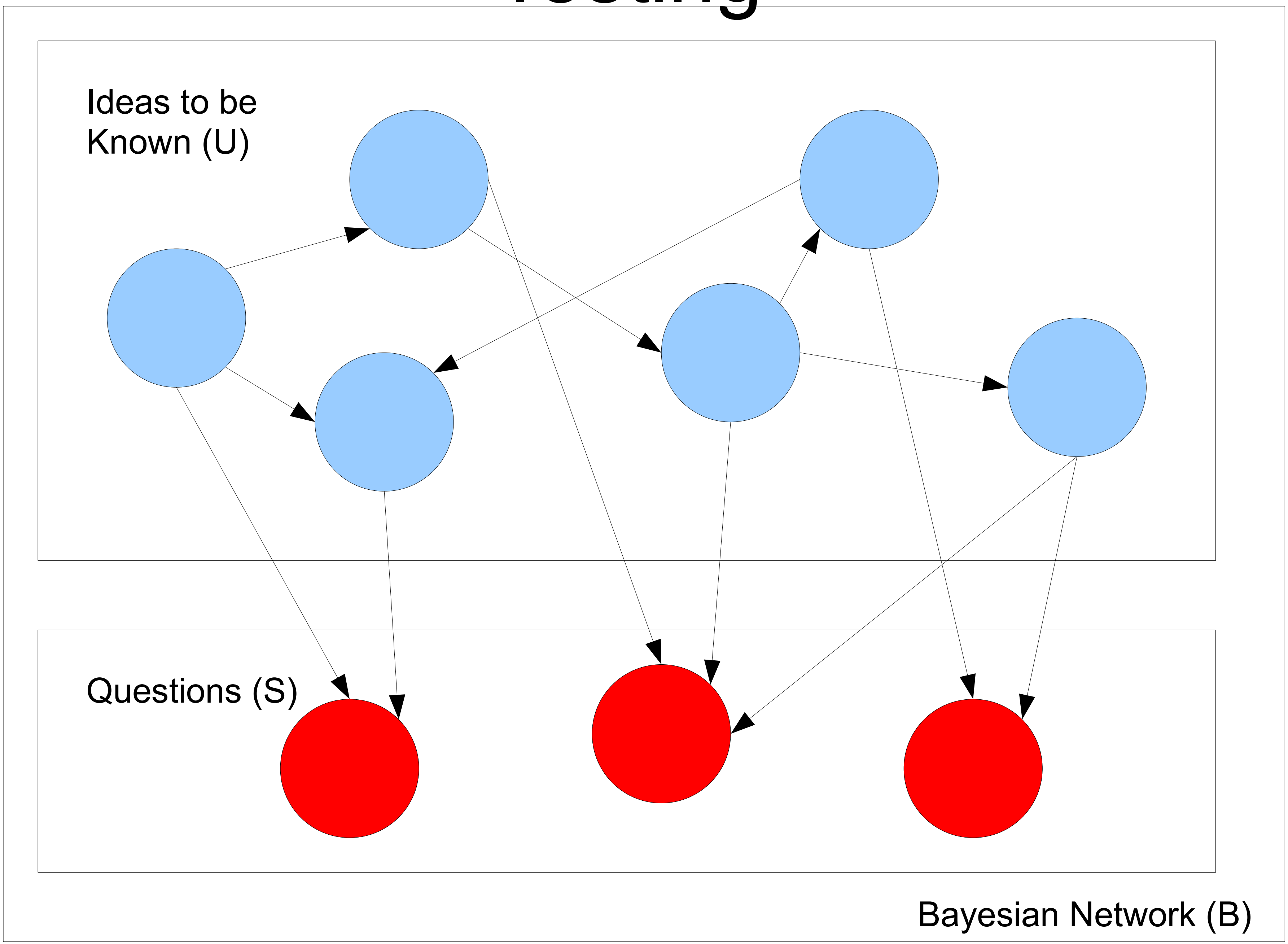
Structured Knowledge



Testing



Testing



Other Formalisms

- b in **R+**
  (Budget for questions)

Other Formalisms

- b in **R**+
  (Budget for questions)
- c : S **->** **R**+
  (Cost function for questions)

Other Formalisms

- b in **R**+
  (Budget for questions)
- c : S **->** **R**+
  (Cost function for questions)
- D = {0,1}
  (Decisions, i.e. Pass/Fail)

Other Formalisms

- b in **R**+
  (Budget for questions)
- c : S -> **R**+
  (Cost function for questions)
- D = {0,1}
  (Decisions, i.e. Pass/Fail)
- A, the “design space”
  (Possible sets of questions within budget)

Utility Function

$$u = \begin{cases} 1 & \text{if } d=1 \text{ and } |\{x \text{ in } u: x=1\}| \geq z|U| \\ 1 & \text{if } d=0 \text{ and } |\{x \text{ in } u: x=1\}| \leq z|U| \\ 0 & \text{otherwise} \end{cases}$$

How do we decide?

Q: Once we have made our observations, what is the optimal decision?

How do we decide?

Q: Once we have made our observations, what is the optimal decision?

A: For this utility function, if
$P(\text{student knows his stuff} \mid \text{observations}) \geq .5$
then pass him.

How Hard is the Problem?

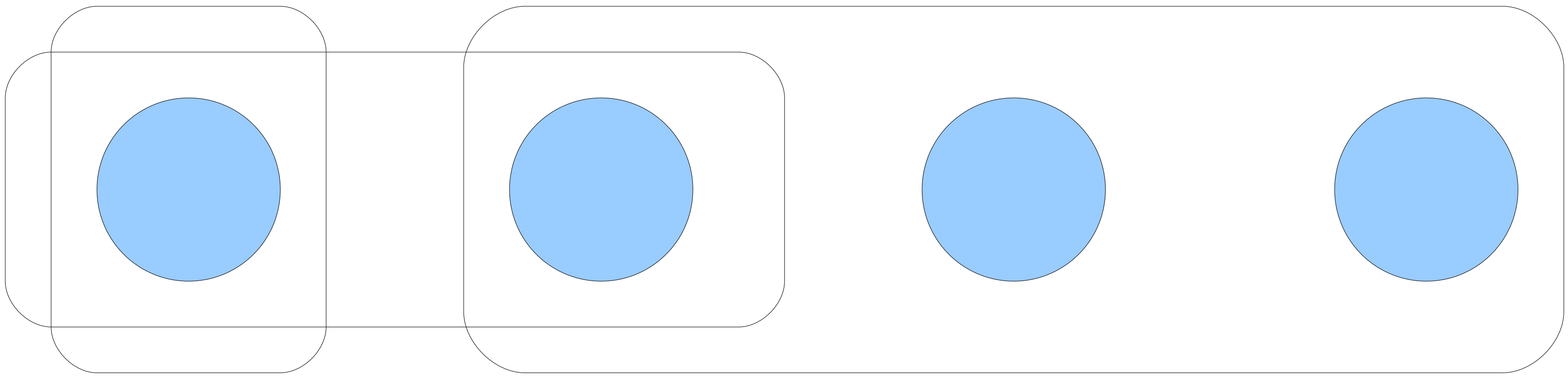
Q: Is it tractable to attempt to compute the optimal question choice?

How Hard is the Problem?

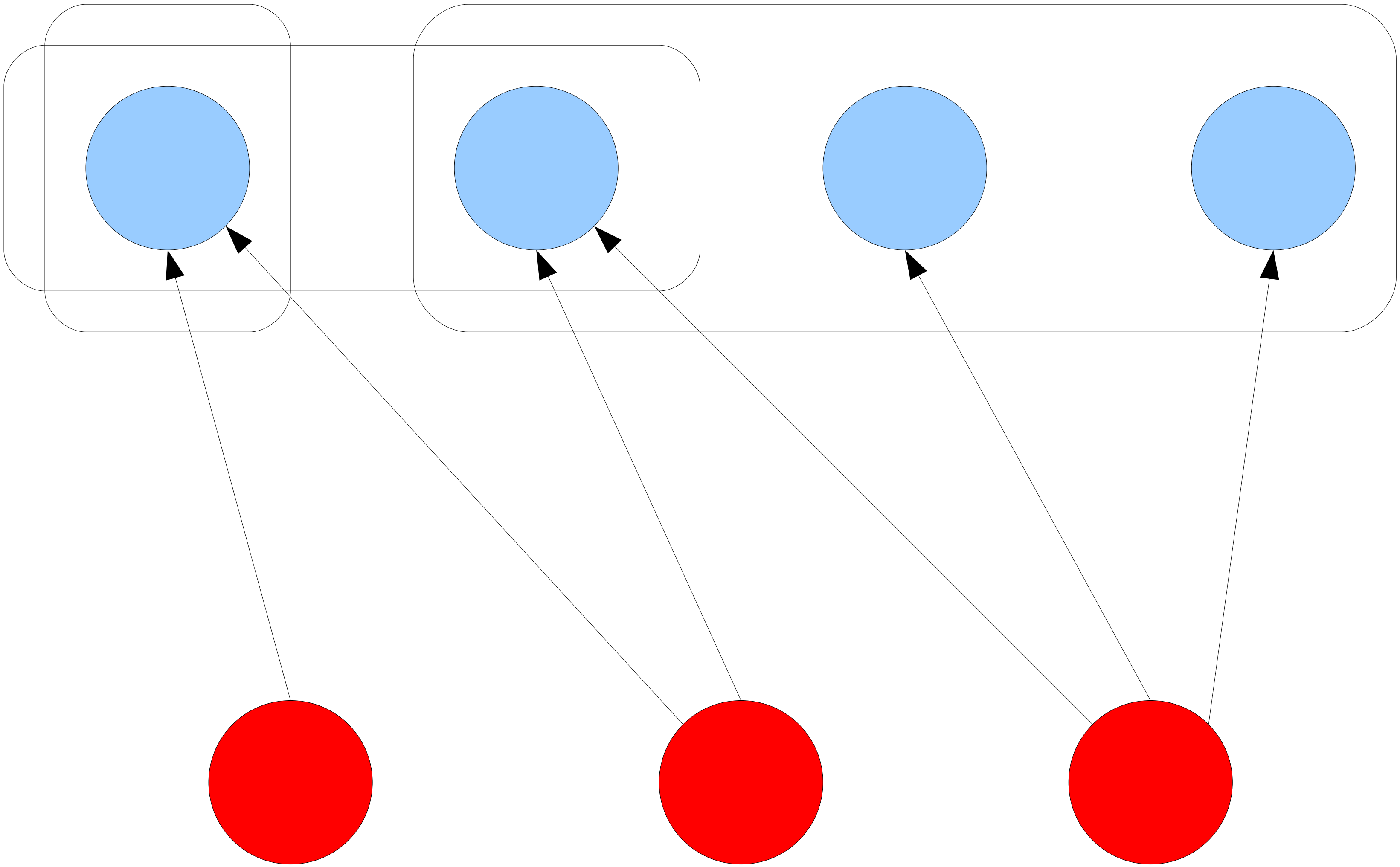
Q: Is it tractable to attempt to compute the optimal question choice?

A: No.

Proof: Reduction from Set Cover



Proof: Reduction from Set Cover



Can We Approximate?

- In the general case, no, the best approximation is simply deciding entirely based on the prior probability of meeting the threshold.

Can We Approximate?

- In the general case, no, the best approximation is simply deciding entirely based on the prior probability of meeting the threshold.
- But we can if we restrict the graph a little bit!

A Restricted Approximation

- None of the nodes in U can be connected to each other

A Restricted Approximation

- None of the nodes in U can be connected to each other
- Every node in S is connected to exactly one node in U

A Restricted Approximation

- None of the nodes in U can be connected to each other
- Every node in S is connected to exactly one node in U
- Each parent in U is deterministically equal to its child in S.

Greedy Algorithm

- Every step, add the node x where
  x = argmax P(u=0)

Greedy Algorithm

- Every step, add the node x where
  x = argmax P(u=0)
- Repeat until budget is used up.

Greedy Algorithm

- Every step, add the node x where
  x = argmax P(u=0)
- Repeat until budget is used up.
- This algorithm is optimal!

Oral Exams

- Obviously at least as good as paper tests.

Oral Exams

- Obviously at least as good as paper tests.
- Given a budget b, improvement can be at least .5-(1/2^b).

Oral Exams

- Obviously at least as good as paper tests.
- Given a budget b, improvement can be at least .5-(1/2^b).
- Still just as NP-hard.

Does Smooth Utility Help?

Define $f(d,u)$ such that
$f(Pass,u) \geq f(Fail,u)$ when $|\{x \text{ in } u: x=1\}| \geq z|U|$
and vice versa.

Does Smooth Utility Help?

Define $f(d,u)$ such that
$f(Pass,u) \geq f(Fail,u)$ when $|\{x \text{ in } u: x=1\}| \geq z|U|$
and vice versa.

No, it's still NP-hard by the same reduction,
and still relatively hard to approximate.

Multiple Professors

- Every professor gets a copy of S to observe separately, and a budget of size b/|P|

Multiple Professors

- Every professor gets a copy of S to observe separately, and a budget of size b/|P|
- Grand committees view question-groups as a whole, and so can perceive the correctness better.

Multiple Professors

- Every professor gets a copy of S to observe separately, and a budget of size $b/|P|$
- Grand committees view question-groups as a whole, and so can perceive the correctness better.
- Singleton sessions allow for more questioning.

Multiple Professors

- Every professor gets a copy of S to observe separately, and a budget of size $b/|P|$
- Grand committees view question-groups as a whole, and so can perceive the correctness better.
- Singleton sessions allow for more questions.
- So which is better?

Grand Committees vs. Singleton Sessions

- For some networks, GCs do better than SSs with expected utilities of 1 vs. less than .5.

Grand Committees vs. Singleton Sessions

- For some networks, GCs do better than SSs with expected utilities of 1 vs. less than .5.
- For other networks, SSs do better than GCs with an expected utility of 1 vs. a utility equivalent to choosing based on the prior.

Grand Committees vs. Singleton Sessions

- For some networks, GCs do better than SSs with expected utilities of 1 vs. less than .5.
- For other networks, SSs do better than GCs with an expected utility of 1 vs. a utility equivalent to choosing based on the prior.
- Deciding which is the case is NP-hard!

Multiple Students

Fairness criterion:

Multiple Students

Fairness criterion:

For every student, the value
H(D|A) (the entropy of the decision given the solution) should be identical.

Does Fairness Reduce Effectiveness?

Consider the case where every student of n students has a 50% chance of knowing that 1+1=2

Does Fairness Reduce Effectiveness?

Consider the case where every student of n students has a 50% chance of knowing that 1+1=2

given a budget b = n-1

Does Fairness Reduce Effectiveness?

Consider the case where every student of n students has a 50% chance of knowing that 1+1=2

given a budget b = n-1

We can't ask every student a question, so the fair solution is to decide based on the prior, with expected utility of n/2

A Better Strategy

But if we choose instead to ask the question “what is 1+1?” of n-1 of the students, our expected utility becomes n-1/2, which is almost twice as good.

In Short

It's awfully hard to test well, and when you do it may not be fair.