TEL AVIV UNIVERSITY אוניברסיטת תל-אביב * UC SANTA CRUZ ^ IBM +

Asking the Right Questions in Crowd Data Sourcing

Rubi Boim * Ohad Greenshpan * Tova Milo * Slava Novgorodov *
Neoklis Polyzotis ^ Wang-Chiew Tan ^+

Goal

Crowd-based data sourcing engages Web users to collectively contribute information.

AskIt! determines which questions should be directed to which users to reduce data uncertainty

What To Take Into Consideration

- **Not all users are equal**
- **Not all questions are equal**
- **Constraints on whom to ask**
  - **Bounded budget**
  - **User availability**
  - **Relevance**

Simplified Example

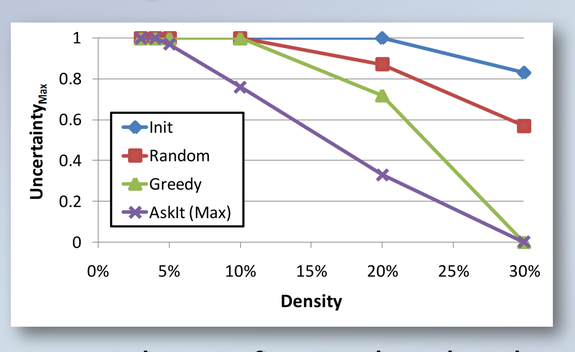
Seller A	Seller B
Good	*Good*
Good	*Good*
Good	*Bad*
Bad	*Bad*
?	*?*

Who would you ask and why?

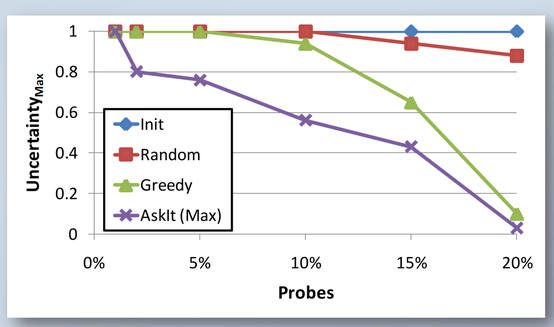
What impact would each have on the data?

- Asking about Seller B cannot shift the overall distribution
- Asking about Seller A can make a large difference (either 4/1 split or a 3/2 split)

Experimental Results



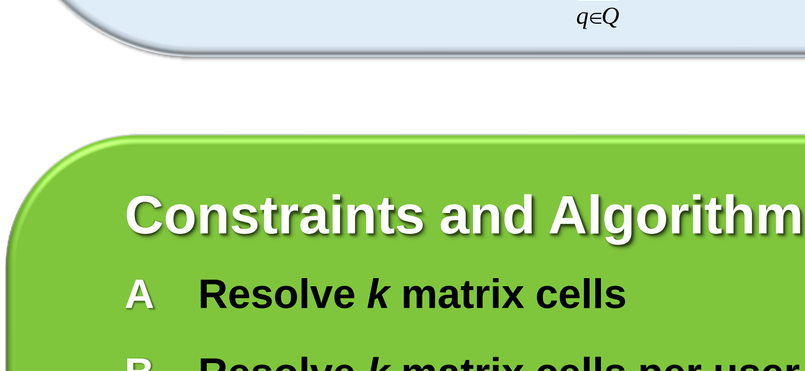
uncertainty$_{max}$ for varying density



uncertainty$_{max}$ for varying probes

Problem Statement

We are given a matrix *M* of user answers to questions



$$entropy(q) = -\sum p_i \log p_i$$ (Probability distribution over the current answers of *q)*

Minimizing the uncertainty in the entropy of *q* due to the unknown cells

$$uncertainty(q) = \max Ent(q) - \min Ent(q)$$

max and min values?

$$uncertainty_{\max}(M) = \max_{q \in Q} uncertainty(q)$$

$$uncertainty_{sum}(M) = \sum_{q \in Q} uncertainty(q)$$

Constraints and Algorithms

A **Resolve *k* matrix cells**

B **Resolve *k* matrix cells per user**

C **Resolve *k* matrix cells per question**

D **Combination of B and C**

$uncertainty_{max}$ PTIME solution for all cases

$uncertainty_{sum}$ PTIME solution for **A** and **C**

B and **D** are NP-complete

(employ greedy heuristics)