Collection of Social Networks Using Crowdsourcing

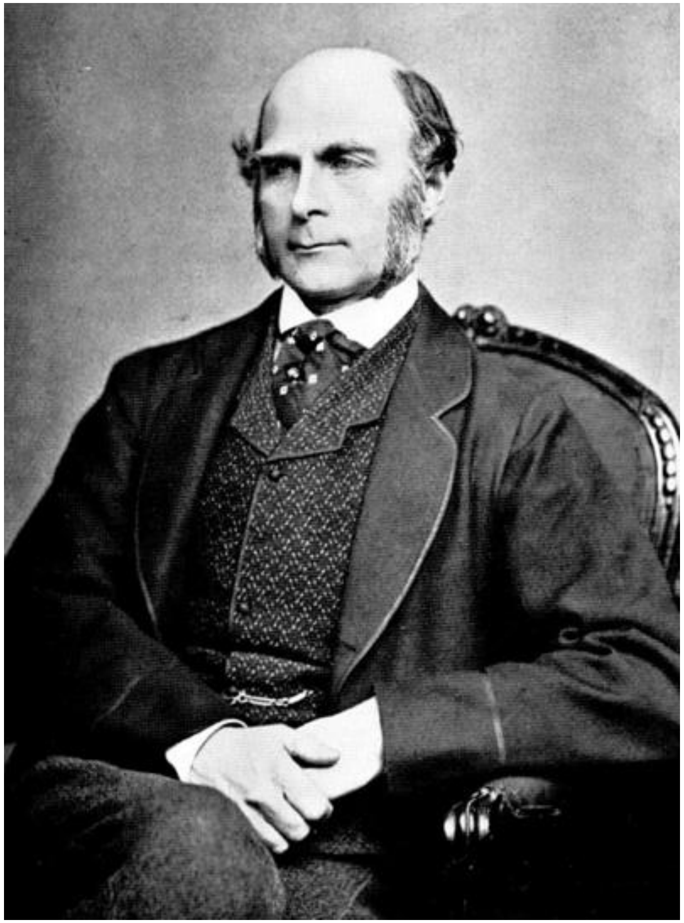
Younghoon Kim (nongaussian@gmail.com)

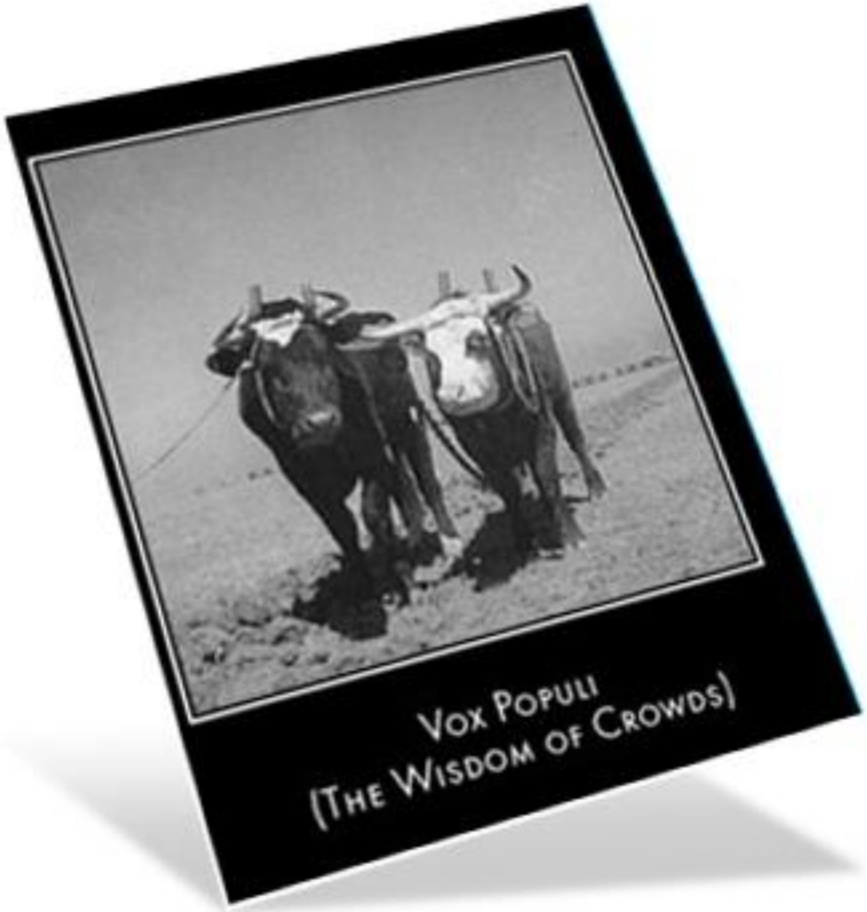
Division of Computer Science

Hanyang University ERICA, Ansan

Traditional Crowdsourcing

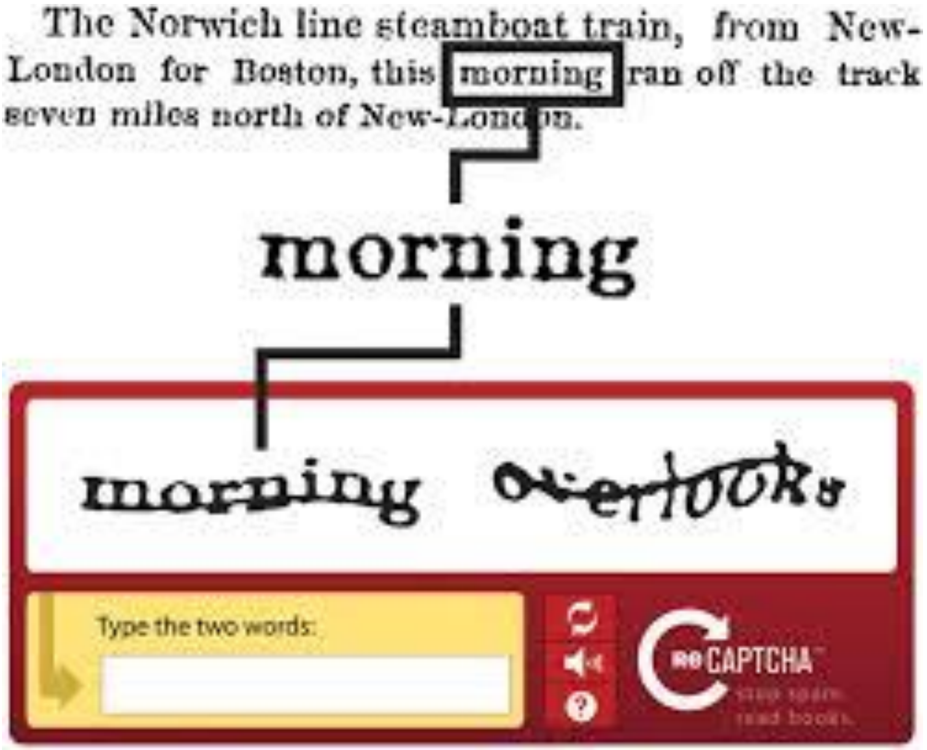
Weight-judging competition (Francis Galton, 1906):
1,197 (mean of 787 crowds) vs. 1,198 pounds (actual measurement)





[Slide by Dongwon Lee, PennState Univ.]

Eg, reCAPCHA

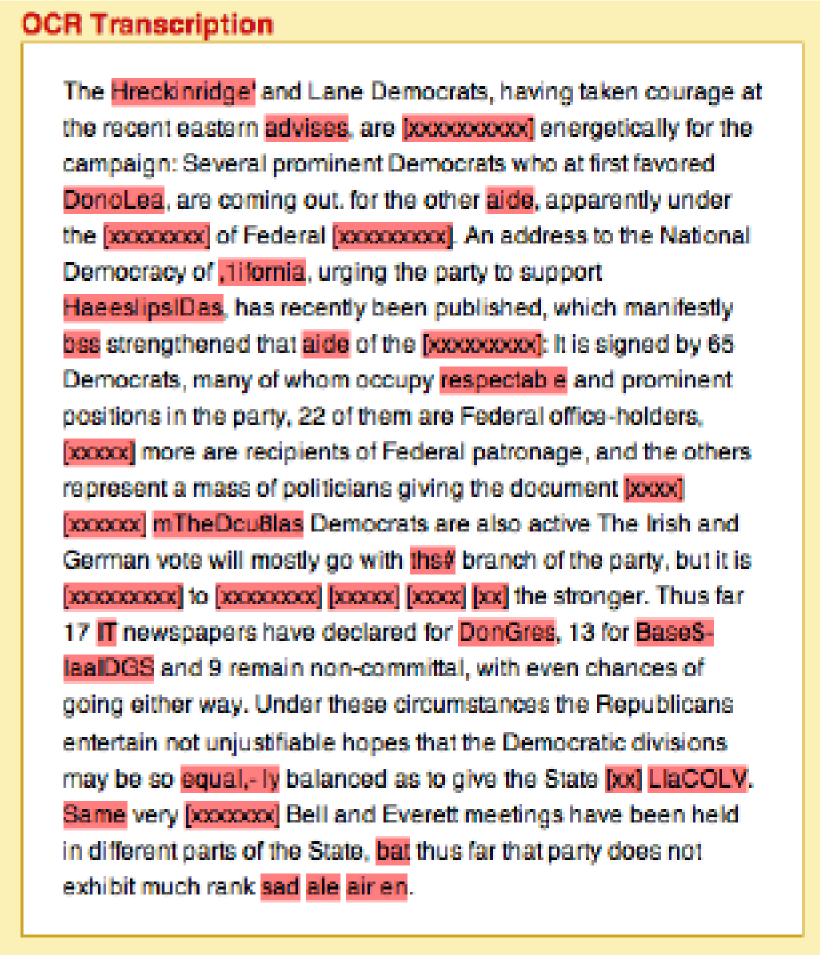

The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.
morning
morning overlooks
Type the two words:
reCAPTCHA
stop spam.
read books.

As of 2012

Captcha: 200M every day

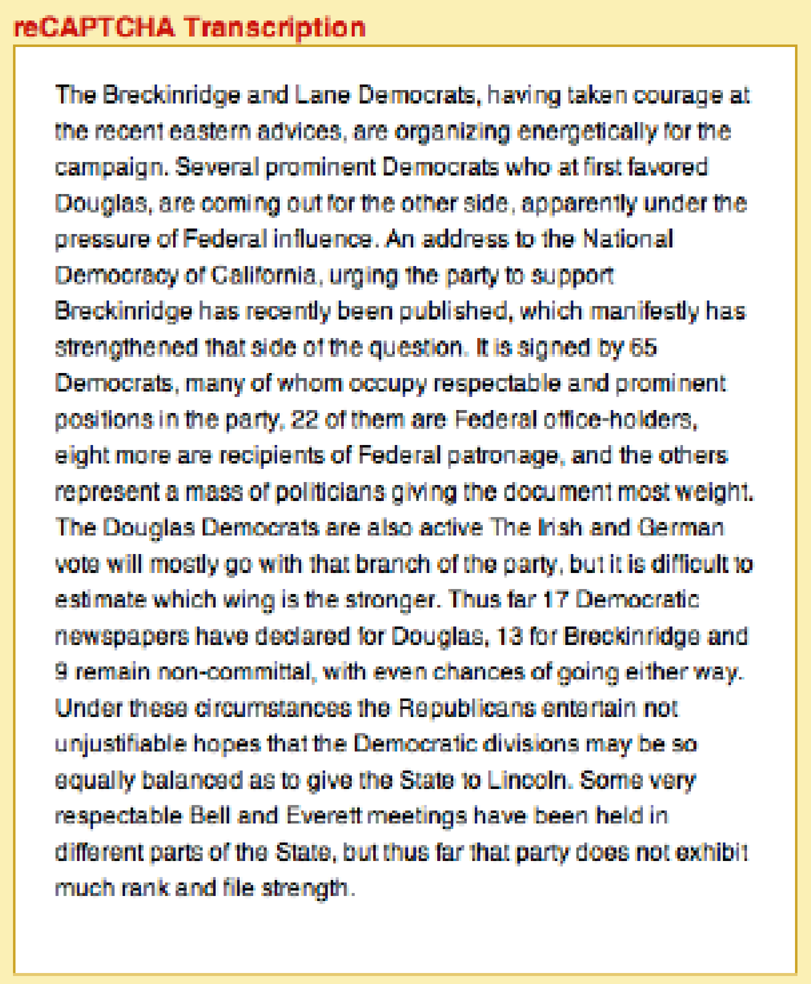
ReCaptcha: 750M to date

Eg, reCAPCHA



OCR Transcription

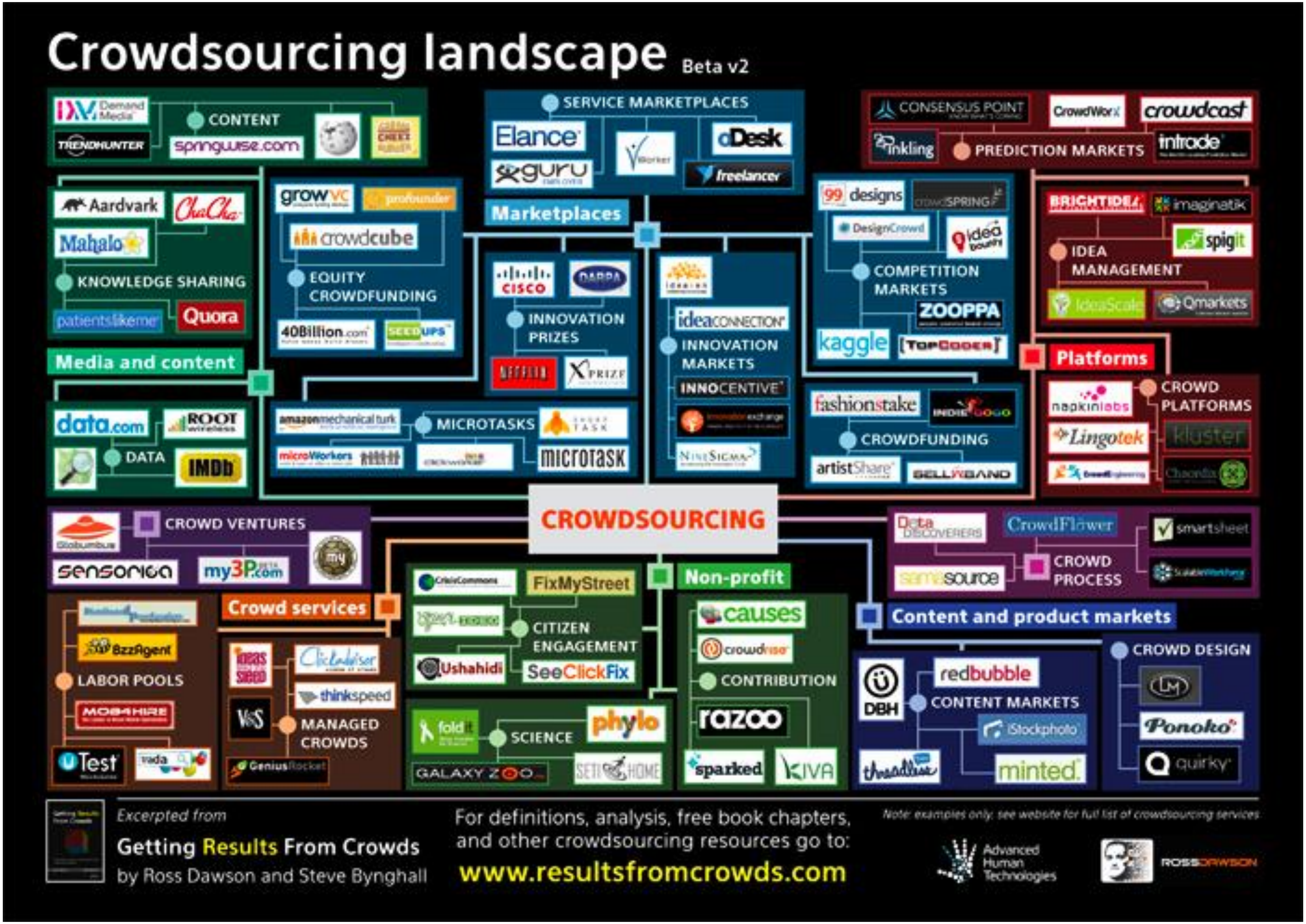
The Hreckinridge' and Lane Democrats, having taken courage at the recent eastern advises, are [xxxxxxxxxx] energetically for the campaign: Several prominent Democrats who at first favored DonoLea, are coming out. for the other aide, apparently under the [xxxxxxxx] of Federal [xxxxxxxxx]. An address to the National Democracy of ,1ifornia, urging the party to support HaeeslipsIDas, has recently been published, which manifestly bss strengthened that aide of the [xxxxxxxxx]: It is signed by 65 Democrats, many of whom occupy respectab e and prominent positions in the party, 22 of them are Federal office-holders, [xxxxx] more are recipients of Federal patronage, and the others represent a mass of politicians giving the document [xxxx] [xxxxxx] mTheDcuBlas Democrats are also active The Irish and German vote will mostly go with ths# branch of the party, but it is [xxxxxxxxx] to [xxxxxxxx] [xxxxx] [xxxx] [xx] the stronger. Thus far 17 IT newspapers have declared for DonGres, 13 for BaseS-IaaIDGS and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equal,- ly balanced as to give the State [xx] LIaCOLV. Same very [xxxxxxx] Bell and Everett meetings have been held in different parts of the State, bat thus far that party does not exhibit much rank sad ale air en.



reCAPTCHA Transcription

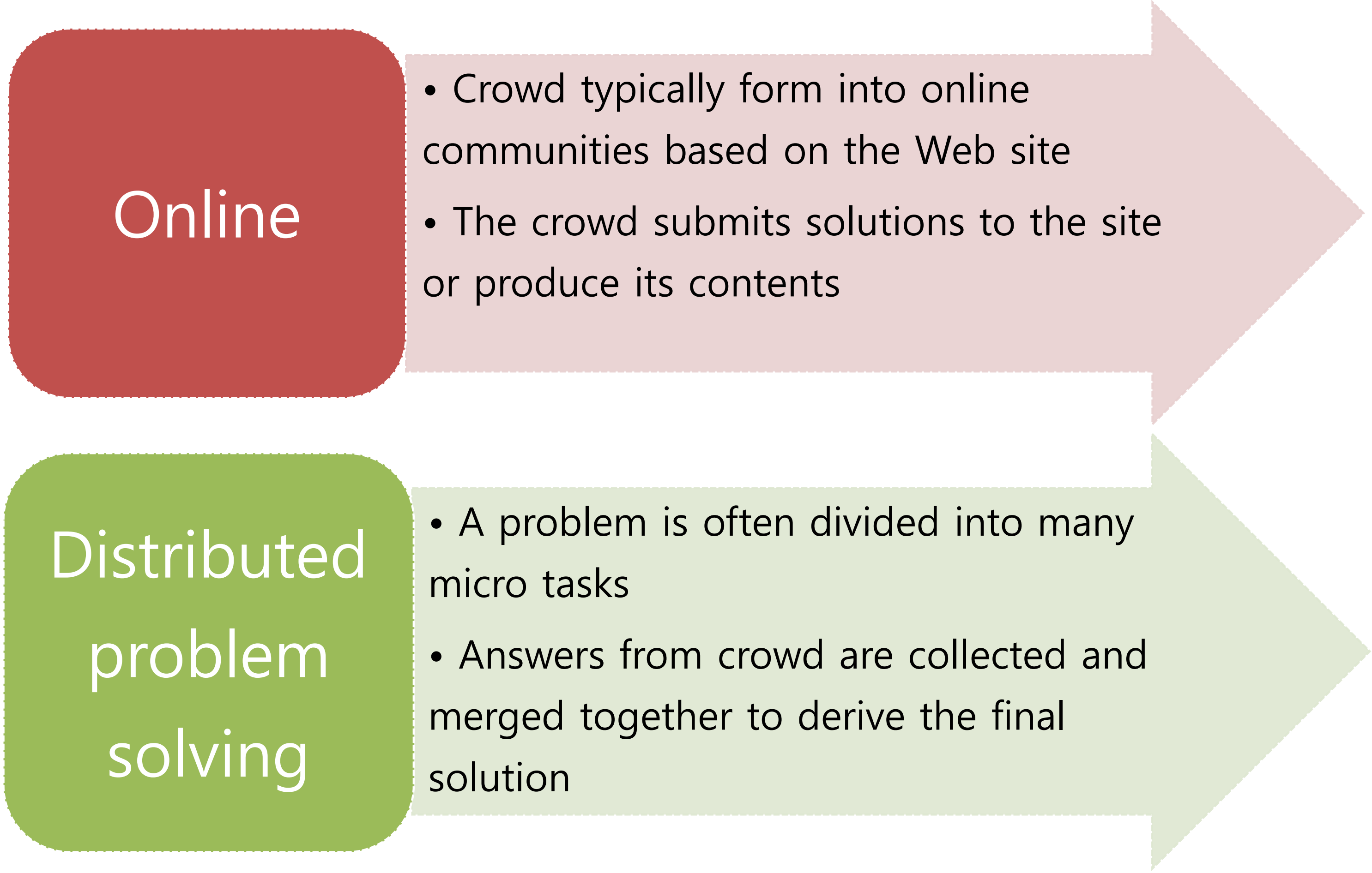
The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

http://www.google.com/recaptcha/digitizing



http://www.resultsfromcrowds.com/features/crowdsourcing-landscape/

Characteristics of Recent Crowdsourcing



Online

- Crowd typically form into online communities based on the Web site
- The crowd submits solutions to the site or produce its contents

Distributed problem solving

- A problem is often divided into many micro tasks
- Answers from crowd are collected and merged together to derive the final solution

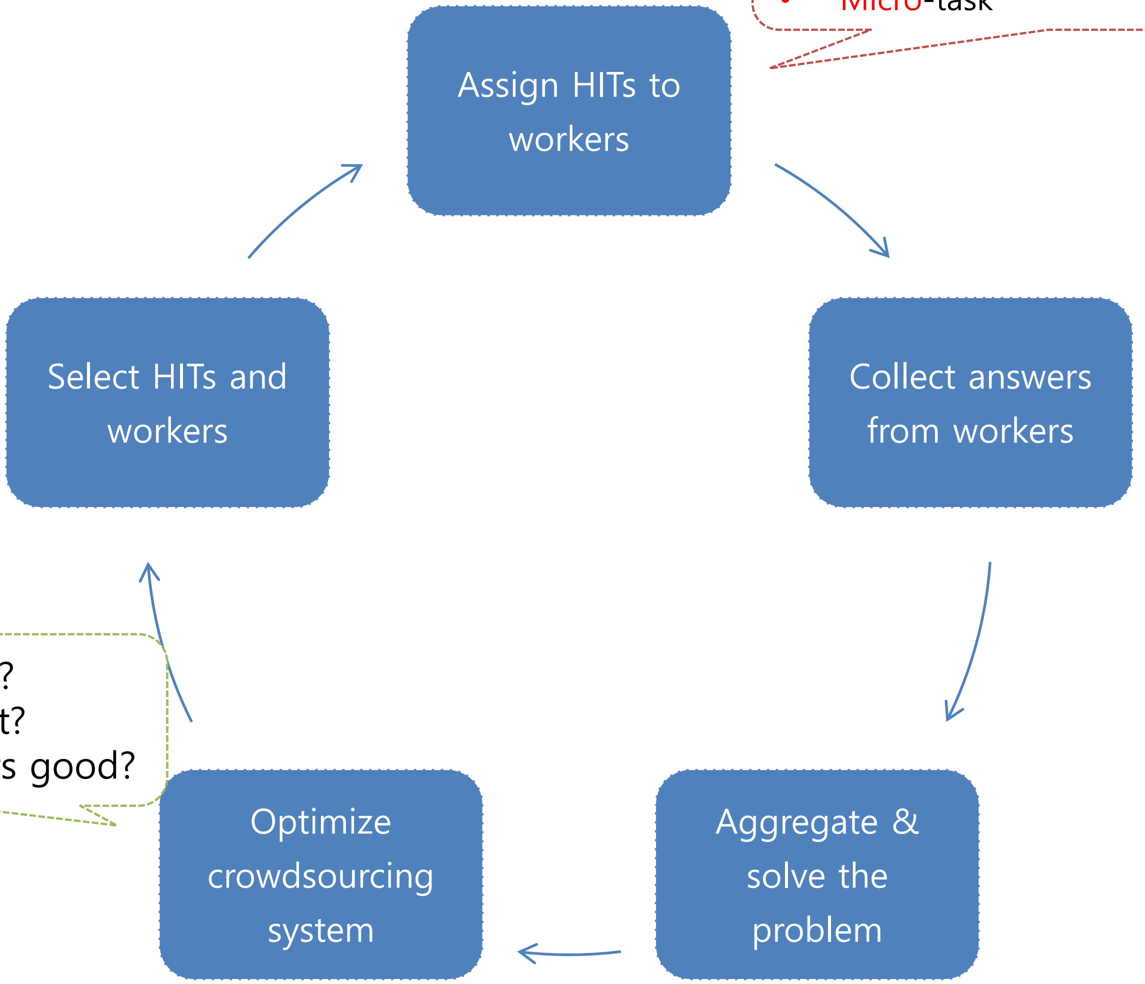
Computational Crowdsourcing

- Focus on computational aspect of crowdsourcing
  - Algorithmic aspect
  - Non-linear optimization problem
- When to use Computational Crowdsourcing?
  - Machine cannot do the task better than human
  - Large crowds can probably do it better than a small number of experts
  - Task can be split to many micro-tasks

Phases of Crowdsourcing

Tasks

- Called **HIT** (Human Intelligence Task)
- Micro-task



Assign HITs to workers

Collect answers from workers

Aggregate & solve the problem

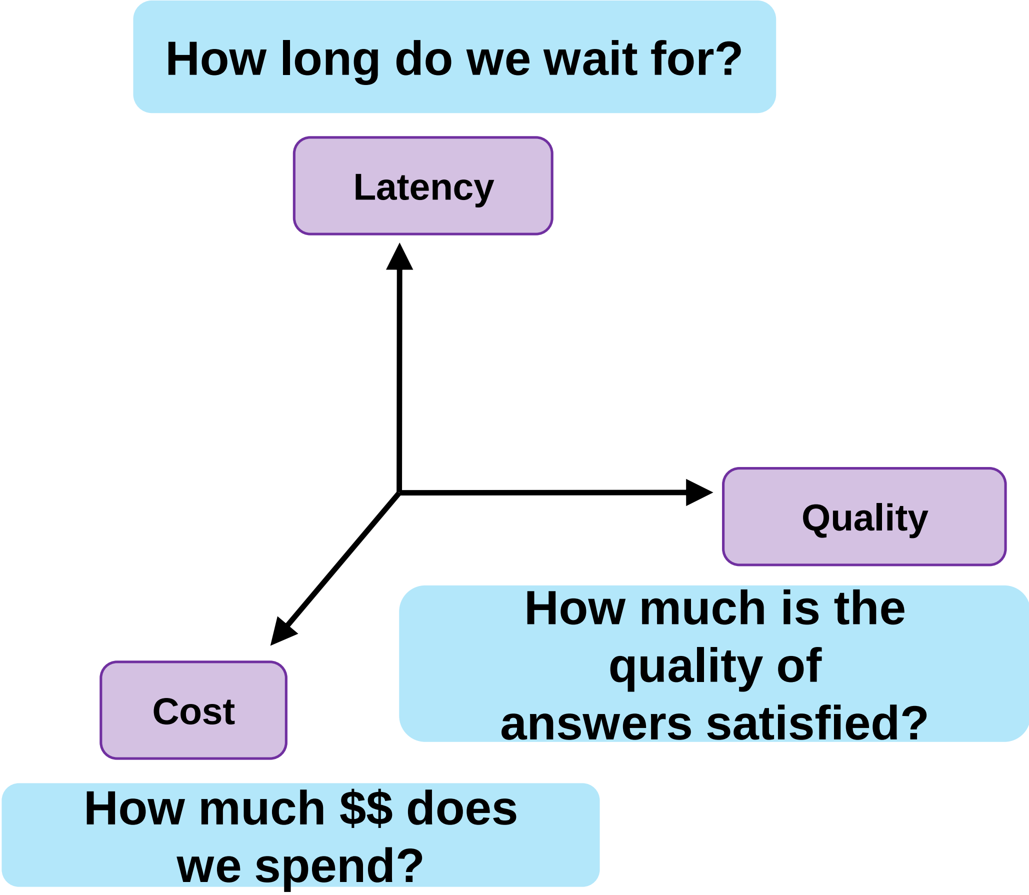
Optimize crowdsourcing system

Select HITs and workers

- How long time?
- How much cost?
- Are the answers good?

Three Computational Factors

- Latency (or execution time)
  - Worker pool size
  - Job attractiveness
- Monetary cost
  - Cost per question
  - # of questions (i.e., HITs)
  - # of workers
- Quality of answers
  - Worker maliciousness
  - Worker skills
  - Task difficulty



How long do we wait for?

How much is the quality of answers satisfied?

How much $$ does we spend?

[Slide by Dongwon Lee, PennState Univ.]

GRAPH INTEGRATION USING CROWDSOURCING

Graph Integration

- Graph integration by crowdsourcing
  - Given two graphs, find the identical pair of nodes by asking people
- Assumption
  - Graphs are written with different languages where inevitably we have to ask people for matching nodes
- Applications
  - Integration of different movie database such as IMDB and Dbpedia
  - User linkage between different social media such as Facebook and Twitter
  - Merge of literature databases like DBLP and CiteULike

Crowdsourcing for Graph Matching

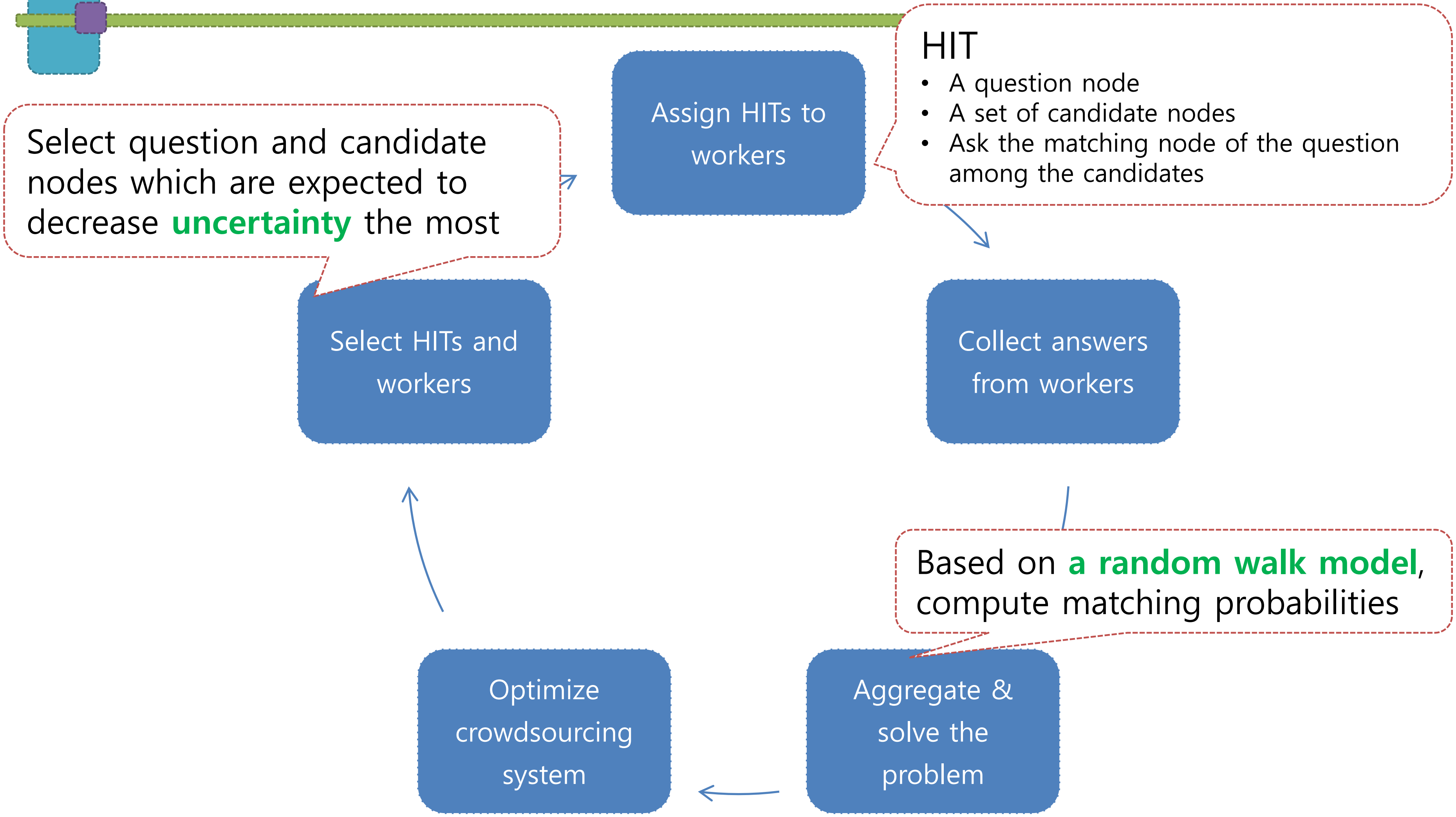
- Given
  - G_L: the left graph where we choose a quest node
  - G_R: the right graph to which workers refer to find its matching node
- Question
  - Which node is identical to this question node from G_L among the candidate nodes from G_R?

Match two graphs as exactly as possible based on the matching pairs collected **with a limited budge**

Problem Definition

- Similarity computation problem
  - Compute the most precise matches between the nodes of G_L and G_R based on the matching pairs collected so far
- Query selection problem
  - Select a query node from G_L and candidate nodes from G_R to ask annotators, which would provide the most useful information for matching two graphs in the next computation of matching nodes.

Phases of Crowdsourcing



Assign HITs to workers

HIT
- A question node
- A set of candidate nodes
- Ask the matching node of the question among the candidates

Collect answers from workers

Based on **a random walk model**, compute matching probabilities

Aggregate & solve the problem

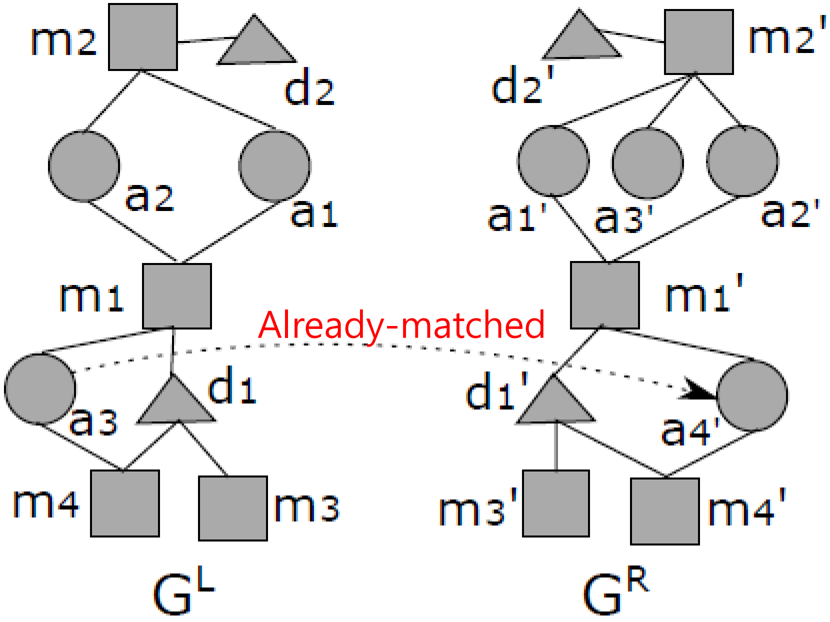
Optimize crowdsourcing system

Select HITs and workers

Select question and candidate nodes which are expected to decrease **uncertainty** the most

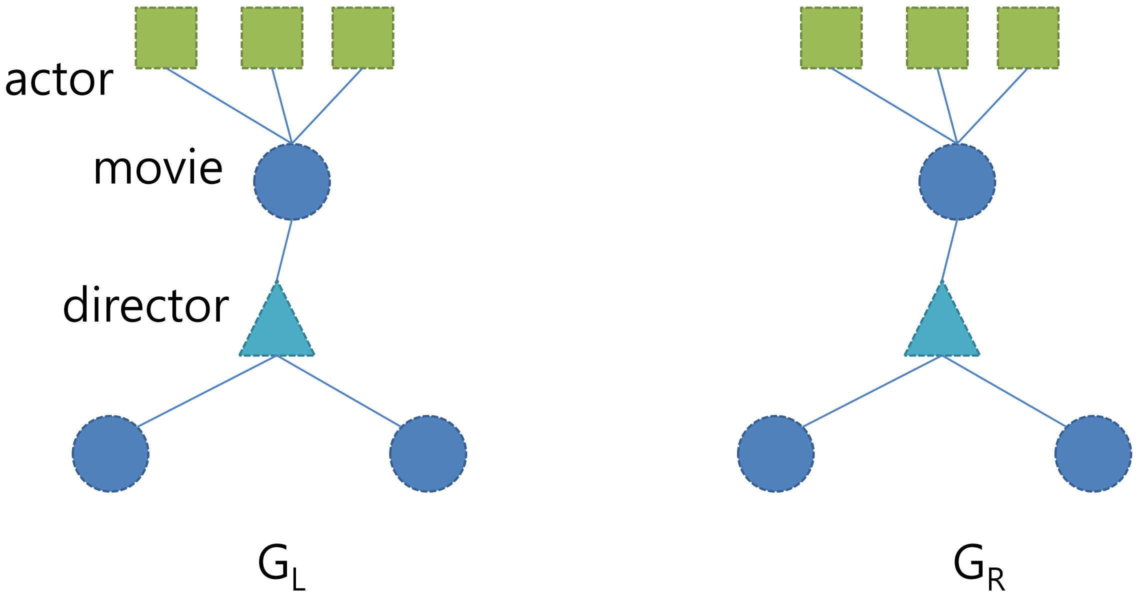
Key Idea

- Which node should be considered in the next question?
  - d_1 with d_1' and d_2'?
  - a_2 with a_1', a_2' and a_3'?



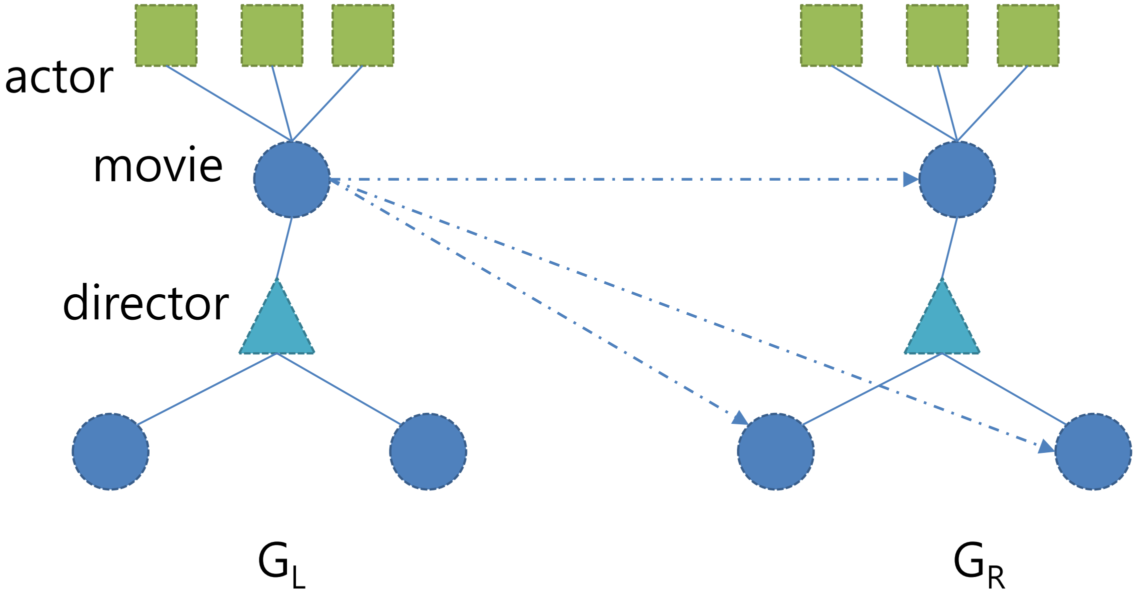
Data & Parameter Description

- Data
  - L: the query set of heterogeneous nodes with T types
  - L_t, $L_{t,i}$: the set of nodes of type t in L and the ith node in L_t respectively
  - R, R_t and $R_{t,i}$ is defined similarly
  - $N_{t,s}(i)$: the set of neighbor nodes of $L_{t,i}$ with type s
    - $n_{t,s}(i)$: the size of $N_{t,s}(i)$
  - M_t: the set of node indexes of type t which are labeled by annotators



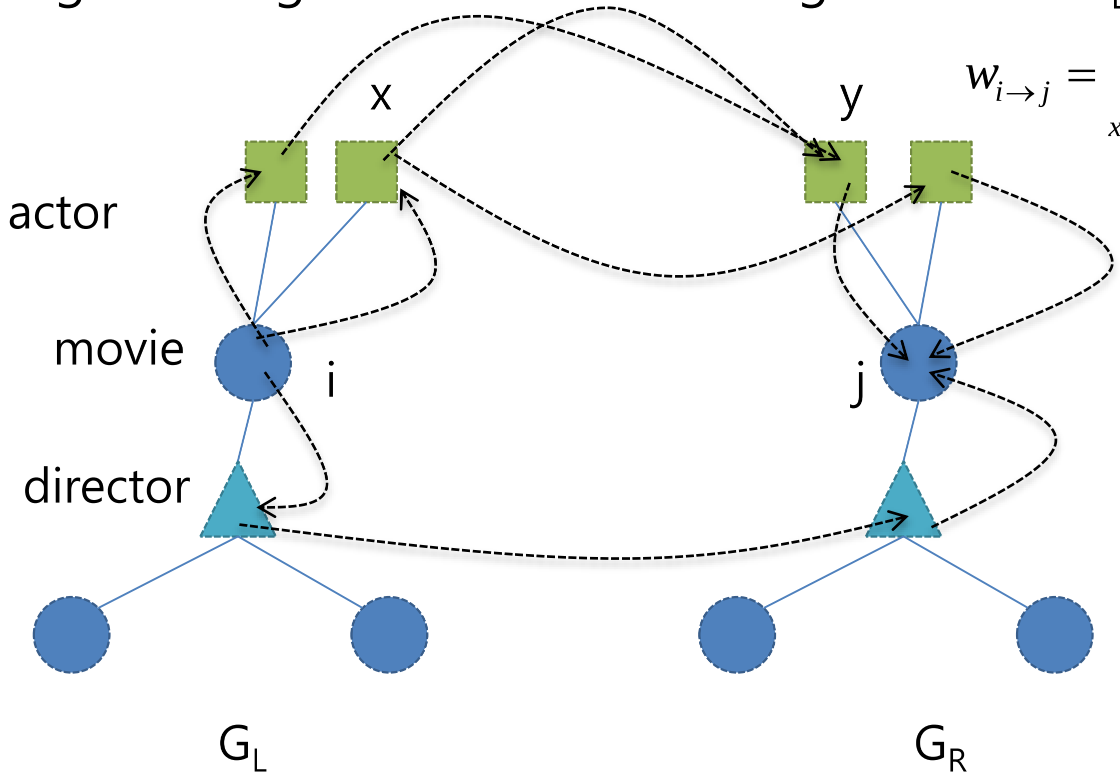
Data & Parameter Description

- Parameter
  - $w_{t,i\rightarrow j}$: for a query node i of type t (i.e., $L_{t,i}$), it is the probability that annotators answers $L_{t,i} = R_{t,j}$



Random Walk Model for Graph Matching

- Given
  - A query node i in G_L
- A worker finds the matching node from G_R for i by
  - Searching the neighbor nodes matching between G_L and G_R



$$w_{i\to j} = \sum_{x\in N(i)} \sum_{y\in N(j)} \frac{w_{x\to y}}{n(i)\cdot n(y)}$$

Inference of Matching Probabilities

- Minimize

Bhattacharyya coefficient B(P,Q):

$$B(P,Q)=-\log\sum_{i}\sqrt{P_i\cdot Q_i}$$

$$D=\sum_{i\in L}\left(-\log\sum_{j\in R}\sqrt{w_{i\to j}\sum_{x\in N(i)}\sum_{y\in N(j)}\frac{w_{x\to y}}{n(i)\cdot n(y)}}\right)+\sum_{i\in M}\left(-\log\sum_{j\in R}\sqrt{w_{i\to j}\overline{w}_{i\to j}}\right)$$

  - Subject to $\sum_{j\in R}w_{i\to j}=1,\ \forall i\in L$
- Where
  - $\overline{w}_{i\to j}$ represents the probability of $w_{i\to j}$ estimated with the matching pairs collected from workers

Inference by EM algorithm

Query Selection Problem

- Ask workers
  - To find *the matching node* in the right graph G_R for the given query node in the left graph G_L
- How to select the nodes for query?
  - Nodes with the largest entropy (called "*maximum expected model change*" [Burr Settles, 2010])
  - Ask an annotator with a query node which results in the largest change on the model

Expected Model Change

Model change = the distribution difference between before and after

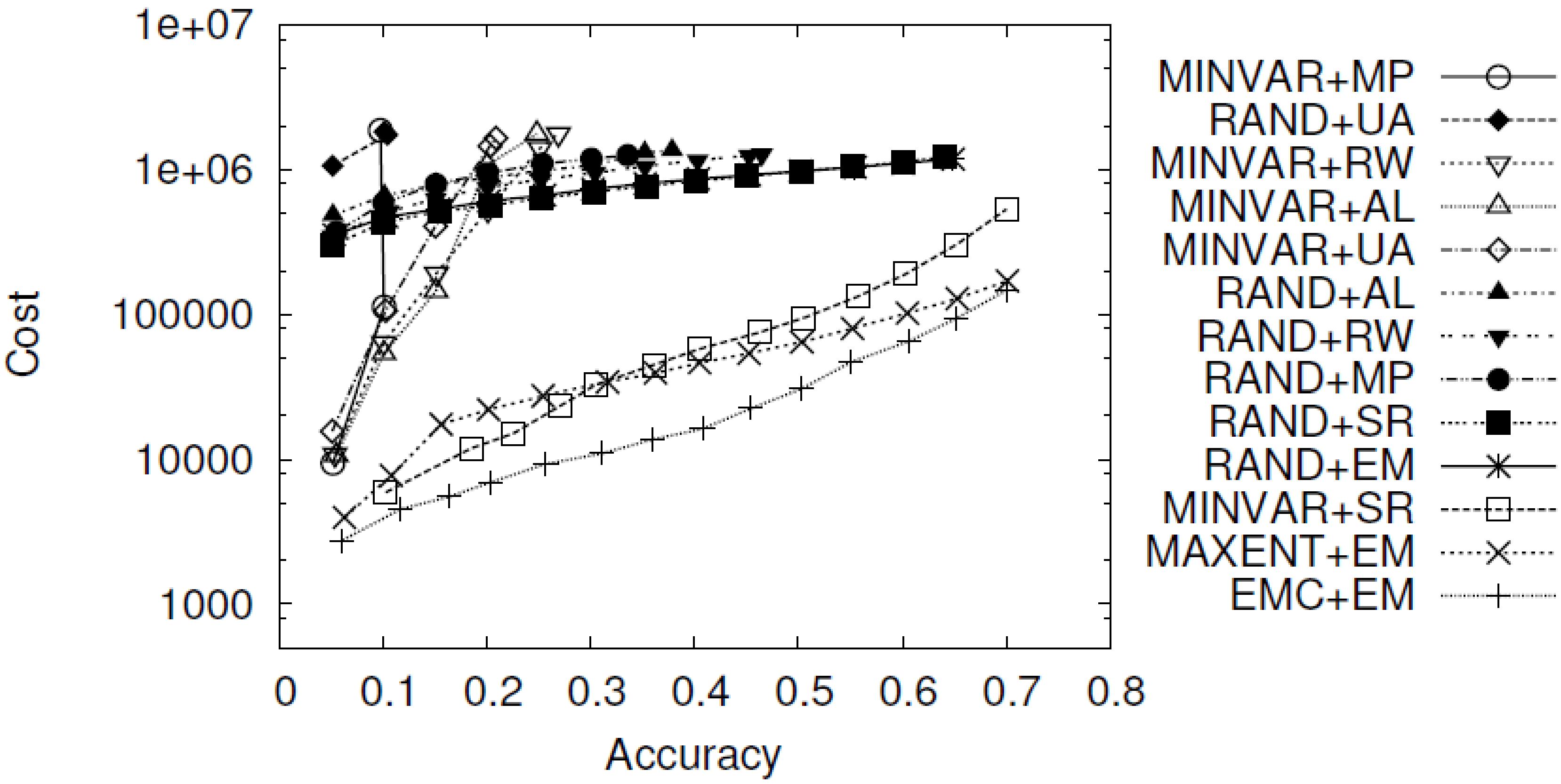
- Select a node i in G_L s.t.

$$\arg\max_{i \in L} \sum_{j \in R} w_{i \to j} \left[\sum_{x \in L} -\log \sum_{y \in R} \sqrt{w_{x \to y} w_{x \to y | i \to j}} \right]$$

  - where $w_{x \to y|i \to j}$ is the probability that x is labeled as y if an annotator labels that i → j
  - However, it is too expensive to find the optimal node maximize the expected model change → approximate calculation

Experiment

- Cost of query selection algorithms



GRAPH COLLECTION USING CROWDSOURCING

Graph Collection

- Graph collection by crowdsourcing
  - Facebook "Do you know XXX?"
  - Collecting edges by asking users
- Aims of the edge collection
  - Collect as much edges as possible with a limited budget?
  - A better(useful) social graph is the one which can spread information faster
    - *Small-world effect* [Watts and Strogatz, 1998 and Milgram, 1967]

~~llect~~
~~es~~
~~a ble~~
~~ted~~
~~dget~~

Increase connectivity between nodes as much as possible with a limited budget

Connectivity

- To measure the extent to which an information can be propagated in a directed graph G = (V,E)
  - *Compactness* [Botafogo et al., 1992]

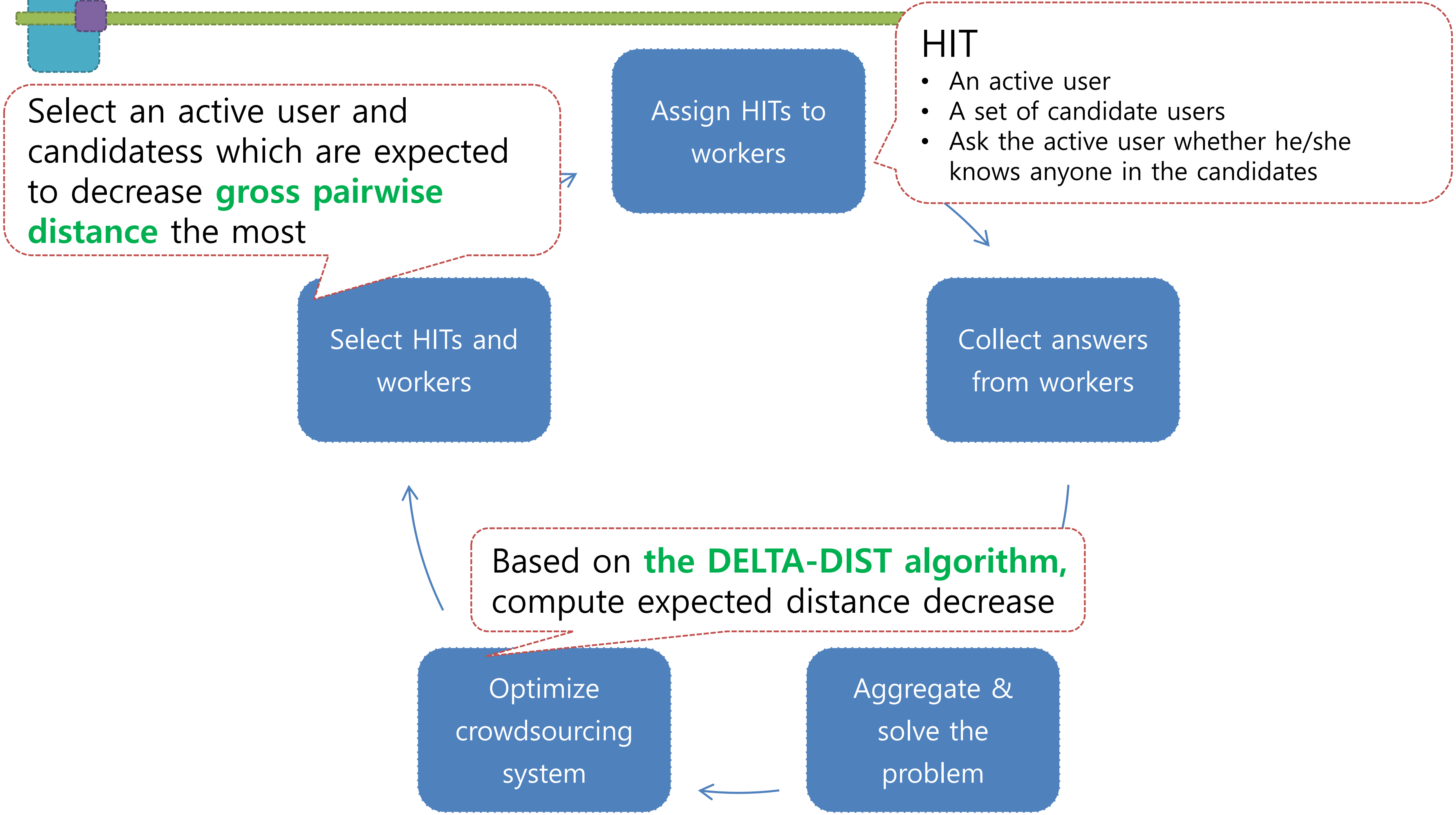
$$Compactness(G) = \frac{Max_{Compactness} - \sum_{u \in V} \sum_{v \in V} d(u,v)}{Max_{Compactness} - Min_{Compactness}}$$

  - which is maximized if the sum of distance of all vertex pairs (called *gross pairwise distance*) is the minimum

Problem Definition

- Given
  - A known directed graph G = (V, P, N)
    - V: vertexes
    - P: presenting edges
    - N: non-existing edges
- Distance changes computation
  - For every pair of vertexes, neither in P nor in N, compute the ***decrease in* the *gross pairwise distance*** if each vertex pair is connected (i.e., the edge of the vertex pair is added in P)
- Query selection problem
  - Discover an active vertex that we expect to bring the largest decrease in the gross pairwise distance if we ask his/her friend

Phases of Crowdsourcing



Assign HITs to workers

HIT
- An active user
- A set of candidate users
- Ask the active user whether he/she knows anyone in the candidates

Collect answers from workers

Aggregate & solve the problem

Optimize crowdsourcing system

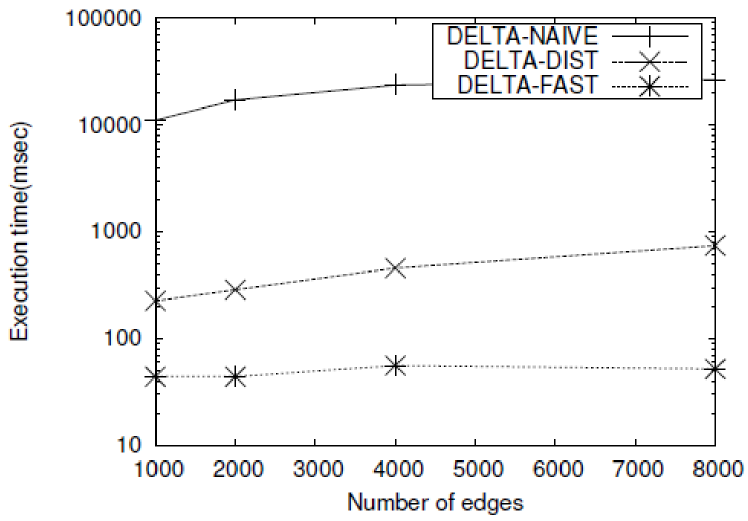
Based on **the DELTA-DIST algorithm,** compute expected distance decrease

Select HITs and workers

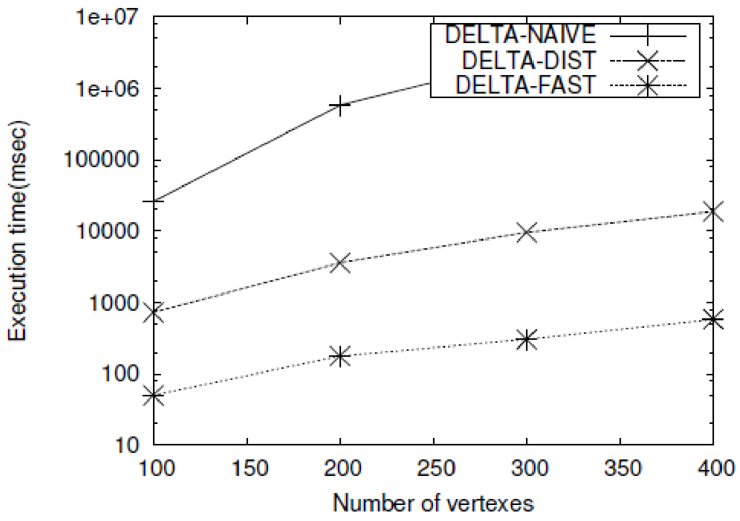
Select an active user and candidatess which are expected to decrease **gross pairwise distance** the most

Experiments

- Execution time of DELTA-DIST algorithm



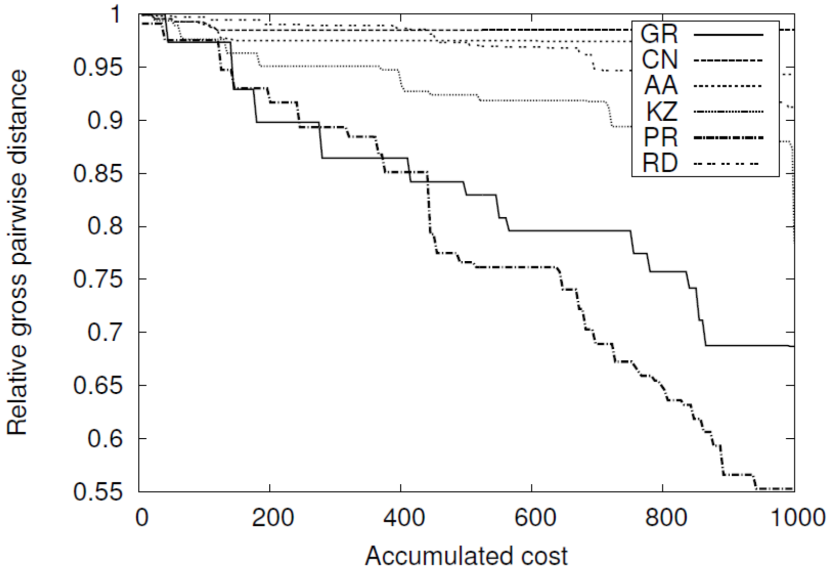
(a) Varying the number of edges with 100 vertexes



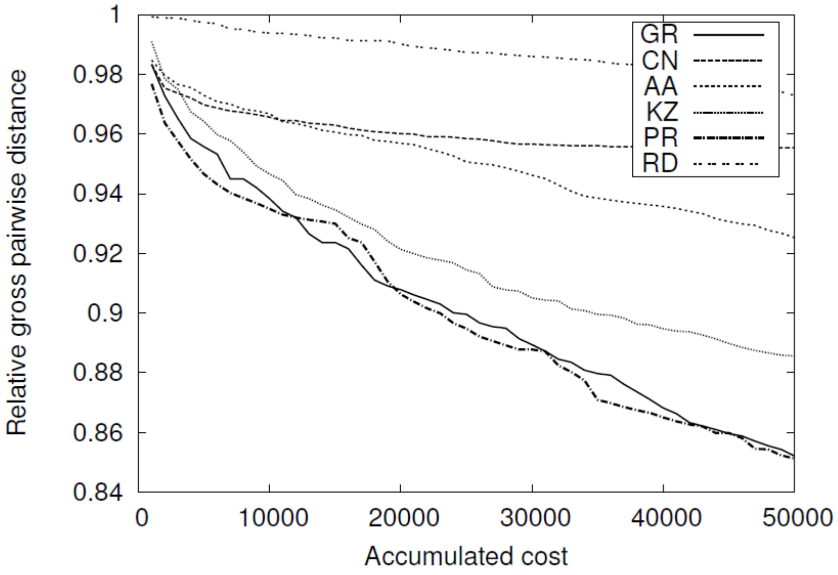
(b) Varying the number of vertexes with 8,000 edges

Experiments

- Cost efficiency for crowdsourcing



(a) Les Miserables data set



(b) OCLinks data set