Algorithmic Crowdsourcing
and Applications in Big Data

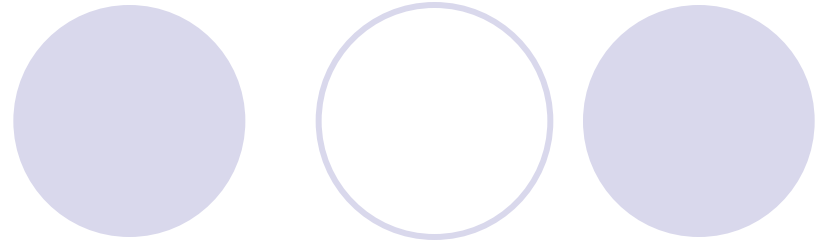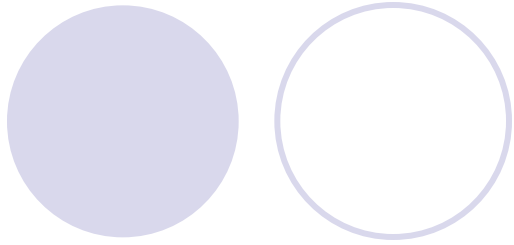Jie Wu

Dept. of Computer and Info. Sciences

Temple University

Road Map

- Introduction
- Mechanical Turk
- Applications
- Paradigms
- Challenges and Opportunities
- Social Crowdsourcing
- Conclusion

Transparent Computing and
Big Data

What

Why

Basic Components

Examples

INTRODUCTION

Big Data is Everywhere!

- Lots of data is being collected:

  Volume, Variety, Velocity

  - Web data, e-commerce
  - purchases
  - Bank/credit card transactions
  - Video and images
  - Social networks

How Much Data?

- Google processes 100 PB a day
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- WeChat has 600 M users and 20 B message per day
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)

640K ought to be enough for anybody.

Transparent Computing and Big Data

From Wikipedia

- "In information technology, big data consists of datasets that grow so large that they become awkward to work with using on-hand database management tools."

- Computers are not efficient in processing certain data (e.g., image processing)

What is Crowdsourcing?

- Coordinating a crowd (a large group of people online) to do micro-work (small jobs) that solves problems (that software or one user cannot easily do)

For Fun !

For Money !

Big Data

The Benefits of Crowdsourcing

- Performance
  - Inexpensive
  - Fast

- Human Processing Unit (HPU)
  - More effective than CPU (for some apps)
    - Image labeling
    - Language translation
    - Social network survey
    - ...

Basic Components

- ## Requester
  - People submit jobs
  - Human Intelligence Tasks (HITs)

- ## Worker
  - People work on jobs

- ## Platform
  - Job management
  - Amazon Mechanical Turk (MTurk)

Requester

Transparent Computing and Big Data

Worker Pool

Malaysia Airlines Flight 370

- ● **DigitalGlobel**
  - ○ Crowdsourcing volunteers comb satellite photos for Malaysia Airlines jet

- ● **March 11**

I just saw on our local Denver Fox news (KDVR.com) that a local company, Digital Globe, has reoriented their satellites to take high-res images in the area where the plane may have crashed. Crowdsourcing efforts are on to have people scan these images and find signs of debris. I was reminded of Jie Wu's talk earlier this month.

As you already know, Digital Globe has participated in the ISTeC IAC meetings (Jay Smith for example).

Example 1: Help Find Jim Gray
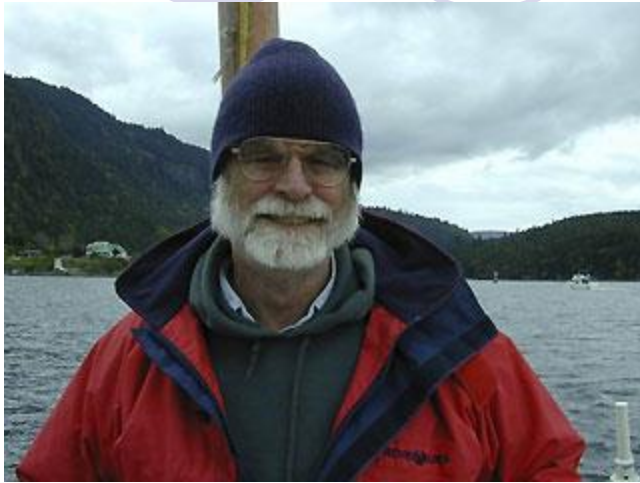
- Jim Gray, Turing Award winner, went missing with his sailboat outside San Francisco Bay in January 2007.

- Use satellite image to search for his sailboat.
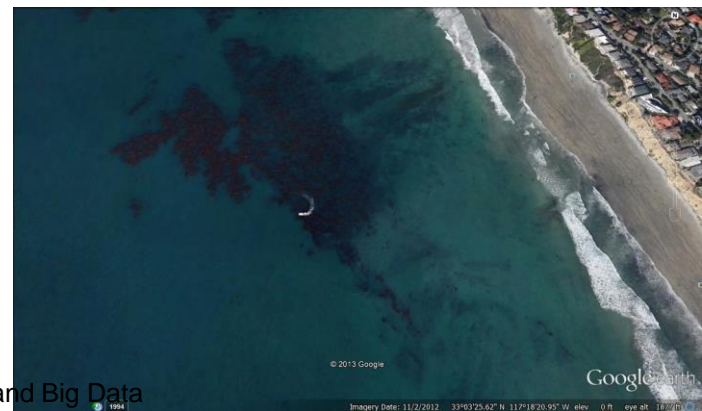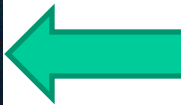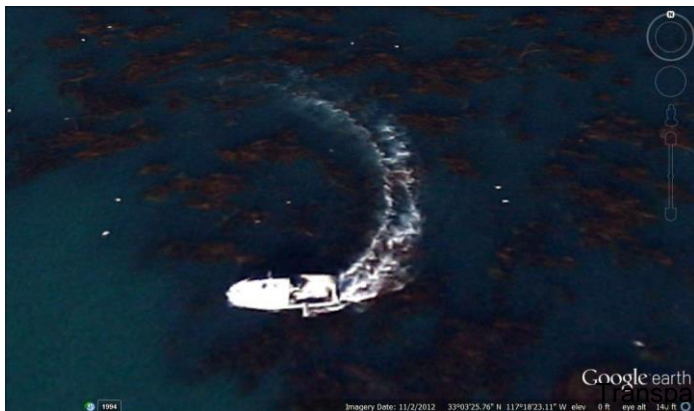
Example 2: DARPA Network Challenges

WE HAVE A WINNER!
MIT RED BALLOON CHALLENGE TEAM

- Problem (2009): $40,000 challenge award for the first team to find 10 balloons.

- MIT team won under 9 hours.

- Winning strategy
  - $2,000 per balloon to the first person to send the correct location
  - $1,000 to the person who invited the winner
  - $500 to whoever invited the inviter
  - ... (or to charity) ...

Example 3: Tag Challenges

- Problem (March 31, 2012): Find five suspects in Washington, D.C., New York, London, Stockholm, and Bratislava.

- Winner from UCSD CrowdScanner: located 3 of the 5 suspects.

- Winning strategy: same as MIT. Also, recruiters of the first 2,000 get $1.

Washington DC
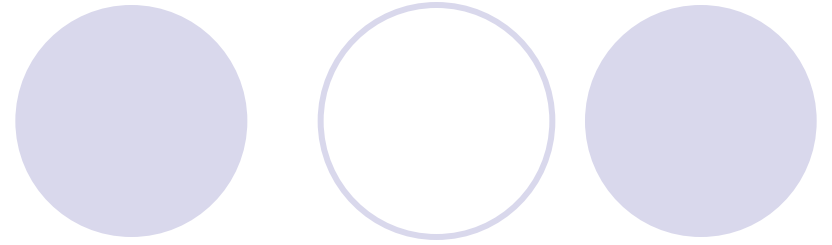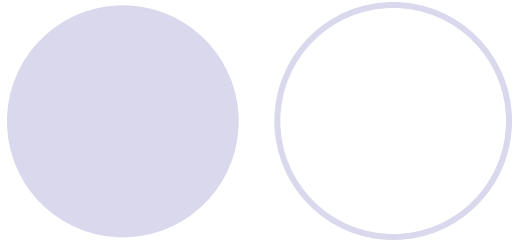
New York City

Bratislava

Transparent Computing and Big Data

Worker

HIT

Dashboard

Requester

MECHANICAL TURK

- As a worker, make an average of $0.03 per task

- As requester, set up simple tasks for workers to complete

- Paid directly to Amazon account

- Quality control is possible through MTurk services

Transparent Computing and Big Data

Worker: Contract for a HIT

- Select a HIT
  - By creation date, payment amount, time allotment

Transparent Computing and Big Data

Worker: Reviewing a HIT

Want to work on this HIT?
Accept HIT

Want to see other HITs?
Skip HIT

Total Earned: $4.72
Total HITs Submitted: 7

Copy Text from Business Card
Requester: Oscar Smith
Qualifications Required: None

Reward: $0.02 per HIT **HITs Available:** 39 **Duration:** 10 minutes

Please Copy Text from Business Card:

Your Current Quality Score is: **--**
If you have a high enough score, you will be ? considered for promotion to a Trusted Worker.

Name ?

Title   Company

Email   Website

Address: ?

Address Line 1

add line

City   State   Zip Code

Phone: click here if not a U.S. phone number ?

Work   Ext.

Mobile

Fax

add phone

Project Oscar
123 Arrowhead Ave.
Westborough, CO 12345
f - (213)233-1950   (213)574-4213
email - project.o...
Oscar Smith
Lead Research Manager

Please Accept HIT First

NOTE: This is a Sample

Please **select/crop** company logo or image from the business card above. Click + Drag to select the company logo.

- Review the HIT before accepting
  - Shown full task, allotted time (10 minutes), reward amount ($0.02)

Transparent Computing and Big Data

Worker: During a HIT

- Shows duration of time
- Gives worker the option to "Return" the HIT

Worker: Completing a HIT

Timer: 00:00:00 of 10 minutes   Want to work on this HIT?   Want to see other HITs?   Total Earned: $4.72
    [Accept HIT]   [Skip HIT]   Total HITs Submitted: 8

✅ **Your results have been submitted to Oscar Smith and will be approved or rejected shortly.**
You can work on this new HIT by clicking the "Accept HIT" button.

Copy Text from Business Card
Requester: Oscar Smith   **Reward:** $0.02 per HIT   **HITs Available:** 3   **Duration:** 10 minutes
Qualifications Required: None

- Confirmation message in green

- Automatically shows the next HIT submitted by the same requester

- Check Dashboard to see if HIT is accepted

Transparent Computing and Big Data

Worker: Sample Dashboard

Avoid Shady Requester

How Turkopticon works:

Turkopticon adds functionality to Amazon Mechanical Turk as you browse for HITs and review status of work you've done. As you browse HITs, Turkopticon places a button next to each requester and highlights requesters for whom there are reviews from other workers. Bad reviews let you avoid shady employers and good reviews help you find fair ones. You can view reports made against requesters with a quick click.

ment and vote on an article. Easy!

quester: ▼ Product Search   HIT Expiration Date:

communicativity: 1.00 / 5
generosity   :   2.57 / 5
fairness   :   2.86 / 5
promptness   :   2.00 / 5
What do these scores mean?

Scores based on 7 reviews
Report your experience with this requester »

Mr. Movie Quote   HIT Expiration Date:

As you review HITs you've completed, are there HITs you weren't fairly paid for? Turkopticon adds a button that lets you review requesters from your "Status Detail" page.
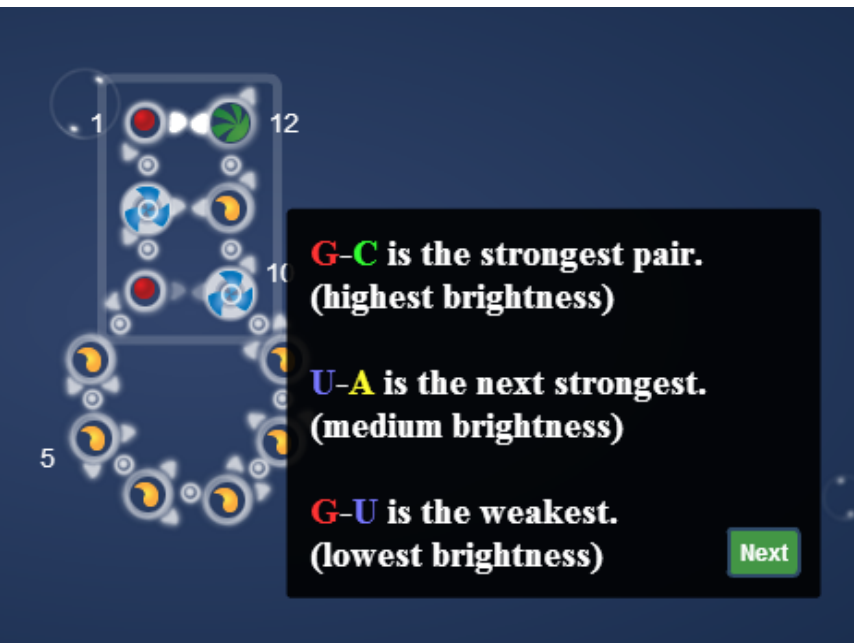
EteRNA

APPS: BIOLOGY

EteRNA: CMU, Stanford

- Aim: to gain mastery over the way RNA molecules folds.

EteRNA: CMU, Stanford

- By assigning different colors (RNA nucleotides), a RNA chain will fold into different structure

EteRNA: CMU, Stanford

EteRNA: CMU, Stanford

EteRNA: CMU, Stanford

EteRNA: CMU, Stanford

Galaxy Zoo

Fine-grained Recognition

APPS: IMAGE PROCESSING

GalaxyZoo: Zooniverse

GalaxyZoo: Zooniverse

GalaxyZoo: Zooniverse

Fine-Grained Recognition

Herring Gull (vs) / Common Tern (vs) Arctic Tern

Horned Puffin (vs) / Parakeet Auklet (vs) Least Auklet

J. Deng et al, "Fine-Grained Crowdsourcing for Fine-Grained Recognition," CVPR 2013
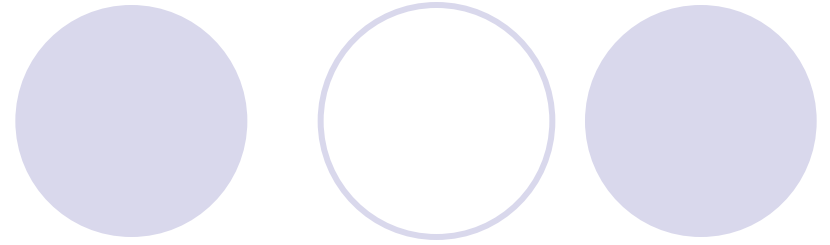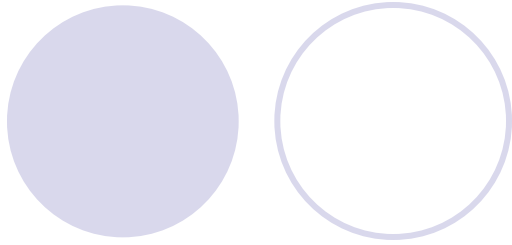
GWAP.com

reCAPTCHA

OnToGalaxy

APPS: COMMONSENSE KNOWLEDGE

GWAP.com: CMU
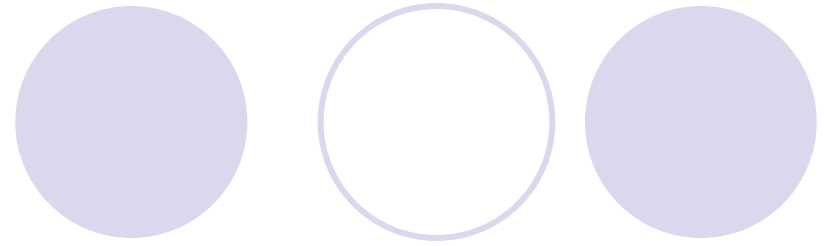
ESP Game
- Labeling images

Tag a Tune
- Labeling tunes

OnToGalaxy: University of Bremen

OnToGalaxy: University of Bremen

- Given a keyword
  - e.g., "tourism"

- Collect pods with words related to keyword
  - e.g., "voyage"
- Shoot down pods with unrelated words
  - e.g., "resist"

- An experimental game platform

Transparent Computing and Big Data

reCAPTCHA: CMU
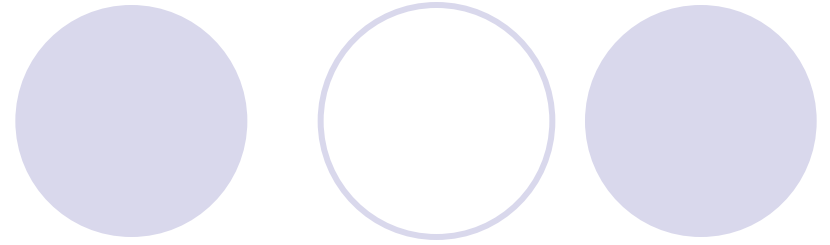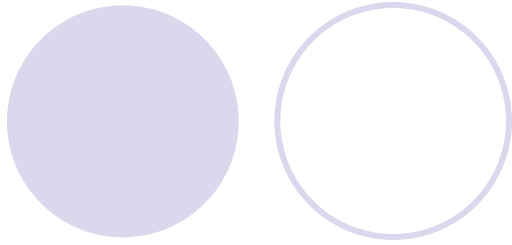
Sequential

Iterative and Parallel

Divide-and-Conquer and Aggregate

Map and Reduce: a Special Case

PARADIGMS

Sequential: Collaborative Workflow

- Lexical translation
  (weak bilinguals or machine)
- Assistive translation
  (strong bilinguals)
- Refine sentence
  (monolinguals)
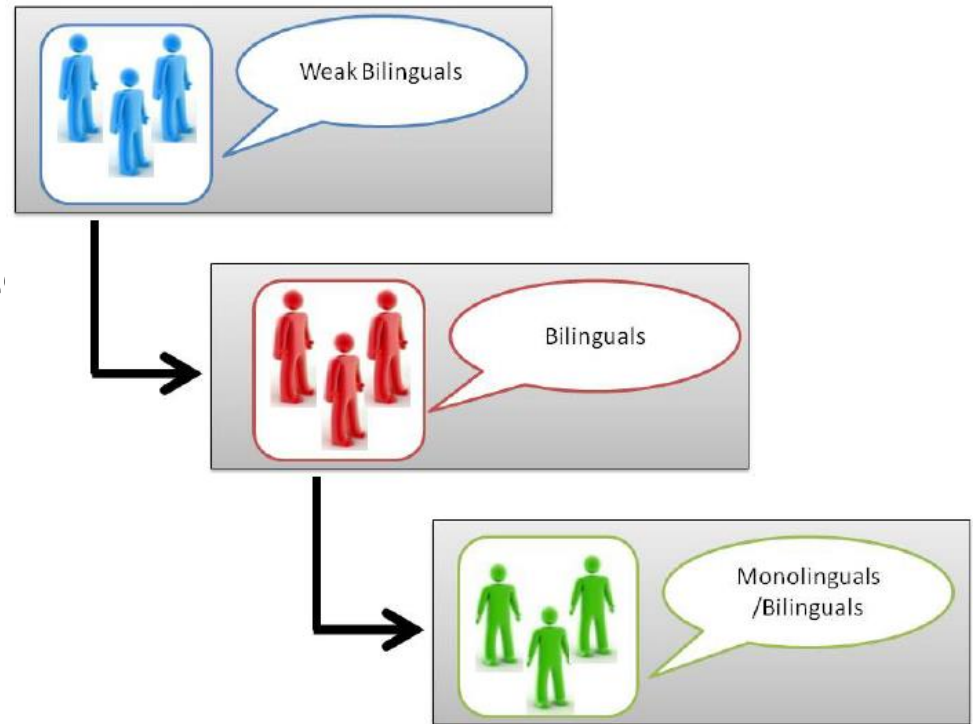
V. Ambati et al, "Collaborative Workflow for Crowdsourcing Translation," CSCW 2012

Iterative and Parallel

- Iterative improve and vote

G. Little et al, "Exploring Iterative and Parallel Human Computation Processes," HCOMP 2010

Transparent Computing and Big Data

Divide-and-Conquer and Aggregate

- **Divide-and-Conquer and Aggregate**
  - Decompose a problem statement and aggregate the results

Divide and Conquer
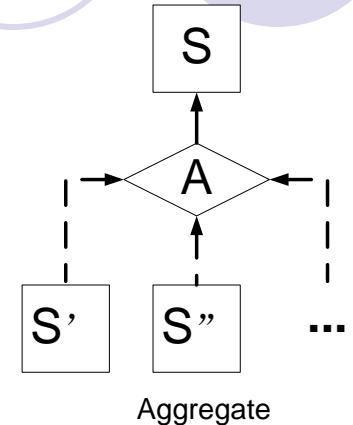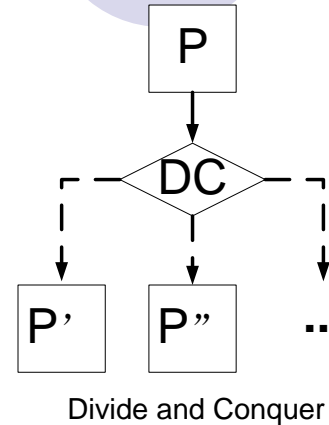
Aggregate

- **Two special aggregates**
  - Merge
  - Reduce

Merge
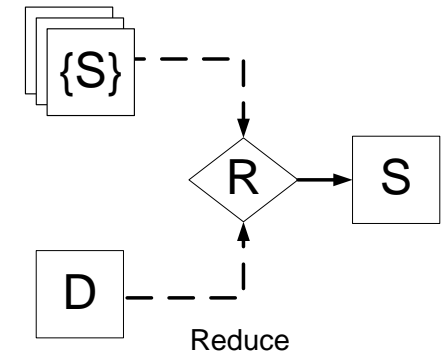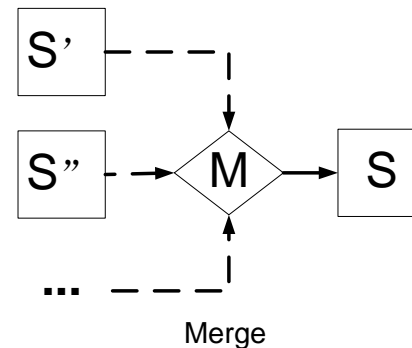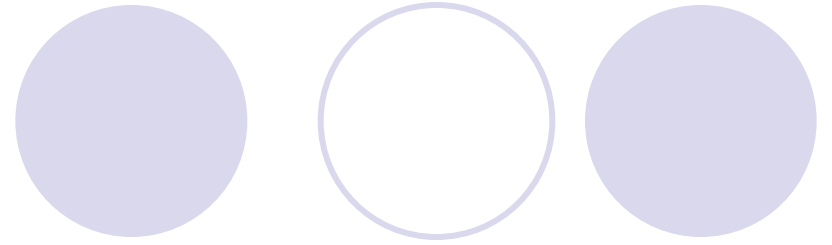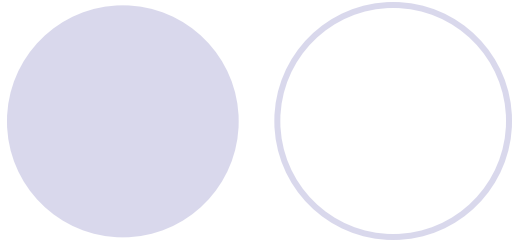
Reduce

P. Minder et al, "Crowdlang - First Steps Towards Programmable Human Computers for General Computation," AAAI 2011.

Map and Reduce: A Special Case
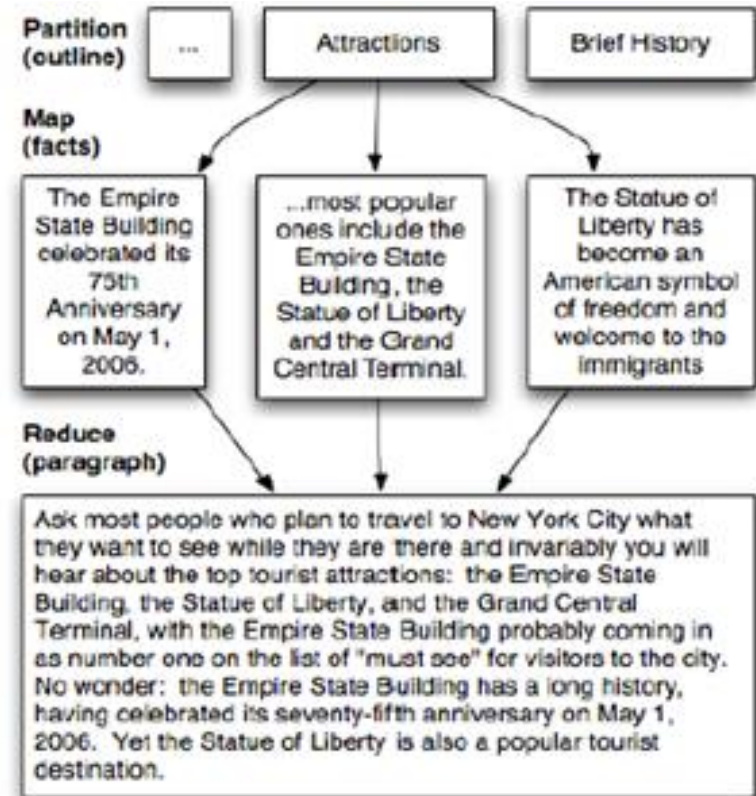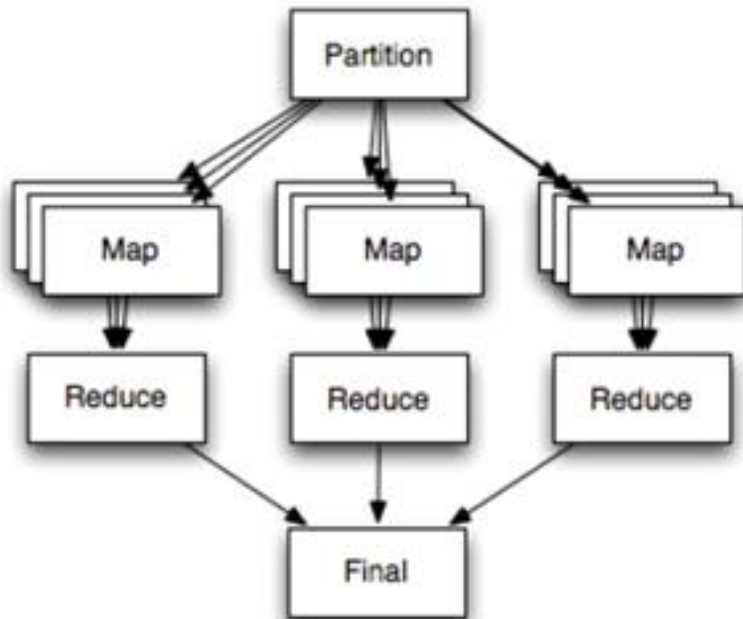
A. Kittur et al, "Crowdforge: Crowdsourcing complex work," UIST 2011

Transparent Computing and Big Data

Challenges

Opportunities

CHALLENGES AND OPPORTUNITIES

Transparent Computing and Big Data

Challenges

Each set has S/2 items

r workers | r

Each set has S/10 items

| r | r | r | r | r | r | r | r | r | r |

- **Trade-offs: time, cost, and quality**
  - Max algorithm with human error (with a probability)
  - Maximize quality (via redundancy) subject to cost and time

  P. Venetis et al, "Max Algorithms in Crowdsourcing Environments," WWW 2012

- **Incentive: money, glory, and love**
  - Platform-centric: a Stackelberg game
  - User-centric: auction-based incentive mechanism

  D. Yang et al, "Crowdsourcing to Smartphones: Incentive Mechanism Design for Mobile Phone Sensing," MobiCom 2012.

Challenges: HPU + CPU
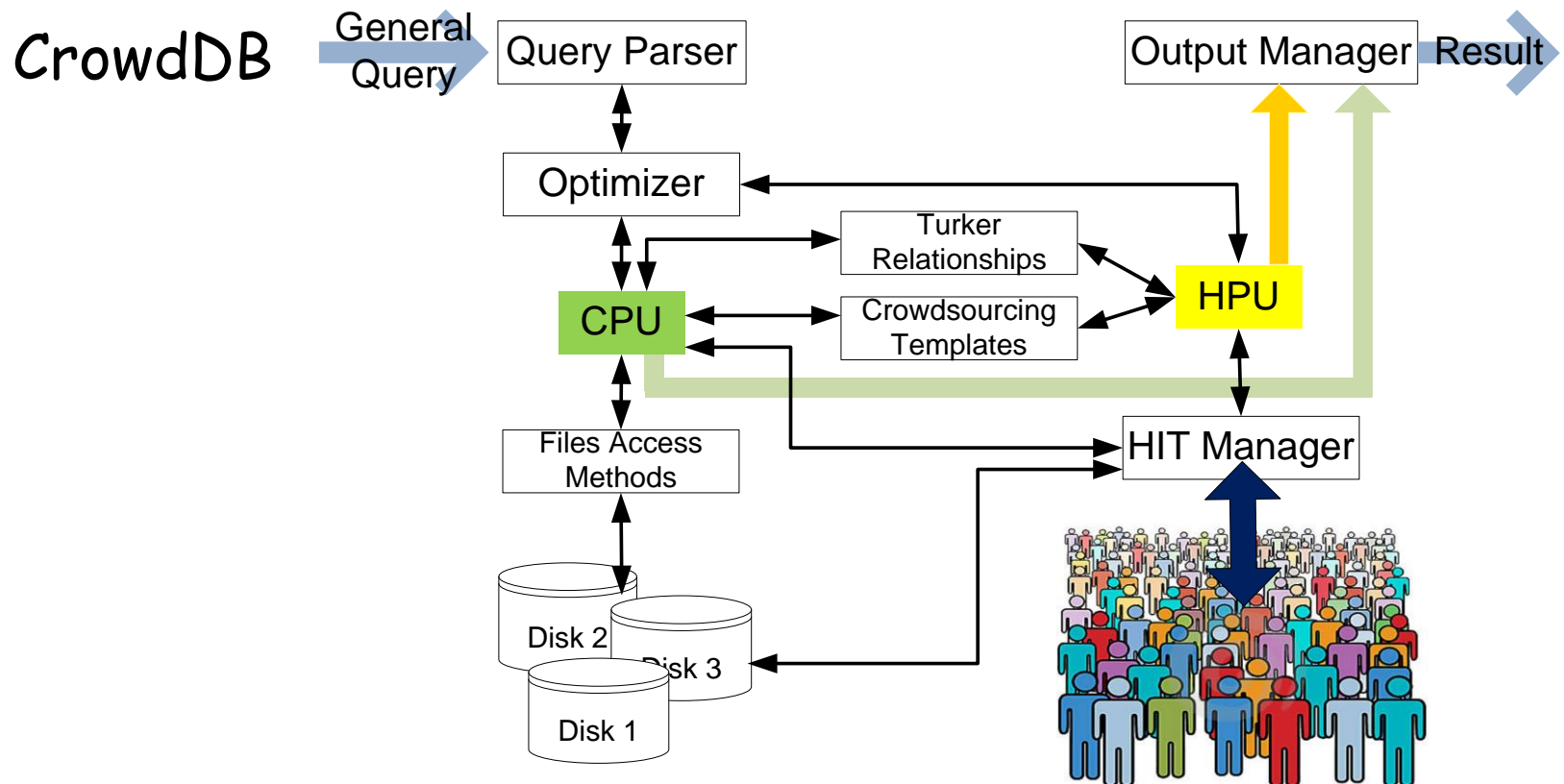
CrowdDB

M. Franklin et al, "CrowdDB: Answering Queries with Crowdsourcing," SIGMOD 2011

Transparent Computing and Big Data

Challenges: Collaborative Workflows

Turkomatic

- Complex works require careful and accurate design workflow
- Problems:
  - Loop subtasks
  - Task starvation
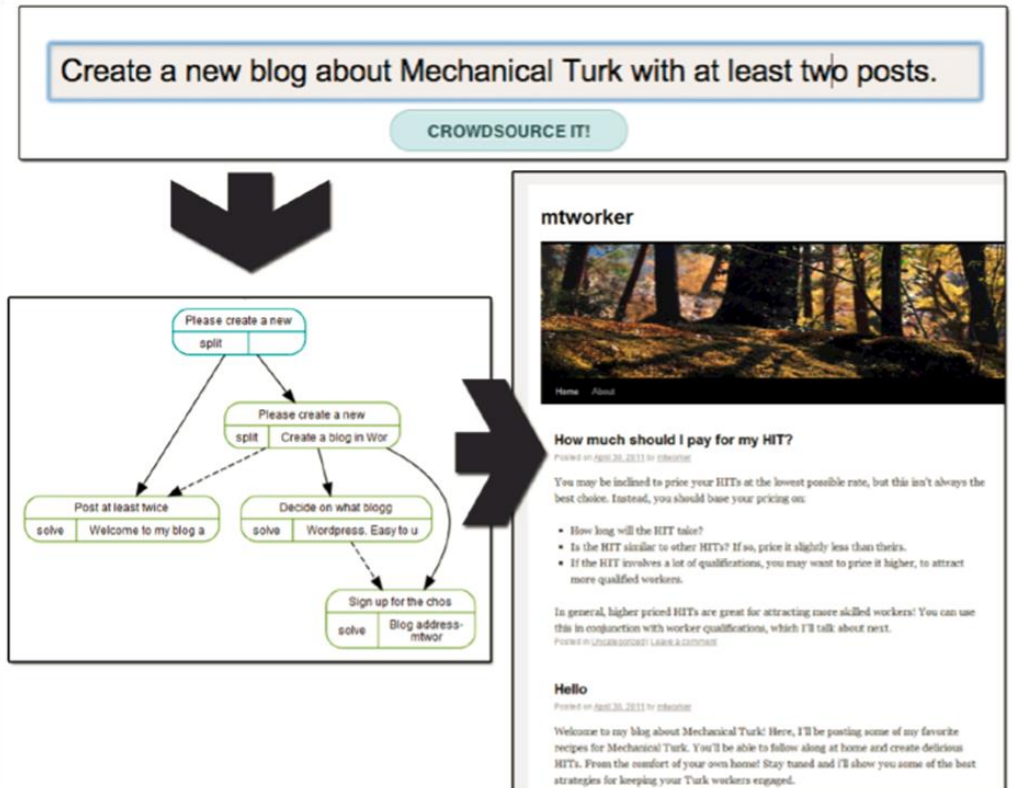
A. Kulkarni et al, "Collaboratively Crowdsourcing Workflows with Turkomatic," CSCW 2012

Transparent Computing and Big Data

Opportunities

- **Beyond simple workflows**
  - Graph search
  - Graph match

- **Beyond simple worker selection**
  - Dynamic procurement

- **Beyond independent workers**
  - Social networks

Beyond Simple Workflows

- **Graph search**
  - Human-assisted graph search
  - Best sequence of questions with simple Y/N answers

  A. Parameswaran et al, "Human-Assisted Graph Search: It's Okay to Ask Questions," VLDB 2010

- **Graph match**
  - People graph (who knows and/or communicates with whom)
  - Puzzle graph (ideas are compatible and can merge)
  - Natural dynamic for people to merge their compatible ideas

  C. Brummitt et al, "Jigsaw Percolation: What Social Networks Can Collaboratively Solve a Puzzle," 2012

Beyond Simple Worker Selection

Dynamic Procurement (multi-armed bandit)

- A gambler facing a row of slot machines

- Which one to play, how many times, and in which order

- Each machine having a random reward from a fixed distribution

- Objective: maximizing the sum of rewards earned through a sequence of lever pulls

Beyond independent workers

A. Badanidiyuru et al, "Bandits with Knapsacks: Dynamic Procurement for Crowdsourcing," 2013

Beyond Independent Workers

- Social network of workers
- Iterative recruitment of workers through social ties
- Challenges
  - Graph searching
  - Timeliness of responses
  - Stoppage condition

Computational Surplus Around

QQ Example

SOCIAL CROWDSOURCING

Computational Surplus Around

- ## Friends help friends
  - Fixed individual capability
  - Probabilistic friends' capability

- ## Makes dissemination decisions
  - Based on the estimations of the fixed and potential computational capacities
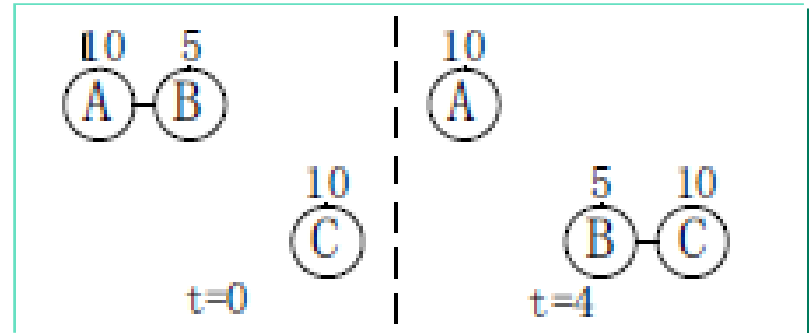
S. Zhang et al, "Minimum Makespan Workload Dissemination in DTNs: Making Full Utilization of Computational Surplus Around ," MobiHoc 2013

Water Filling Schedule

- Response delay
- Computation (by a friend)
- Reply delay

- d_i: response + reply
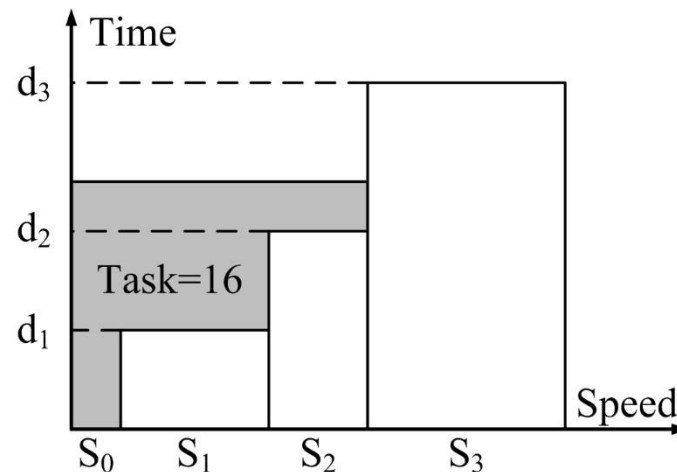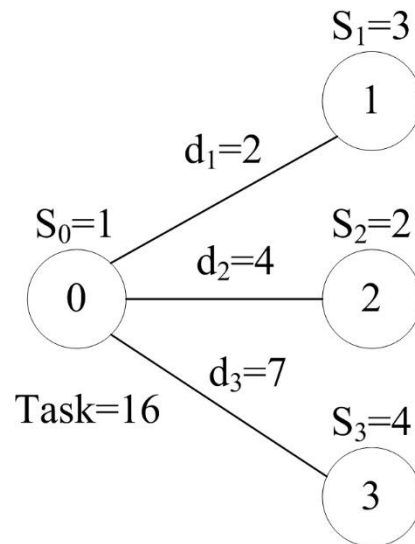
QQ Example

- Tencent QQ, or QQ
  - Instant messaging
- As of March 2013
  - 798.2 million active QQ accounts
  - Peak of 176.4 million simultaneous online users
- QQ experiment
  - Exploring social status of QQ users by responses

Iterative Request/Reply (reduce)

- Initial label is L = "2" (subtract L by 1 when forwarding this request to QQ friends)

- When L = 0, return the total number of QQ friends

- When L > 0, do the following:
  - Forward this request to all QQ friends
  - After receiving the first 10 replies, compute the average number of friends, and send them back to me

Iterative Request/Reply (merge)
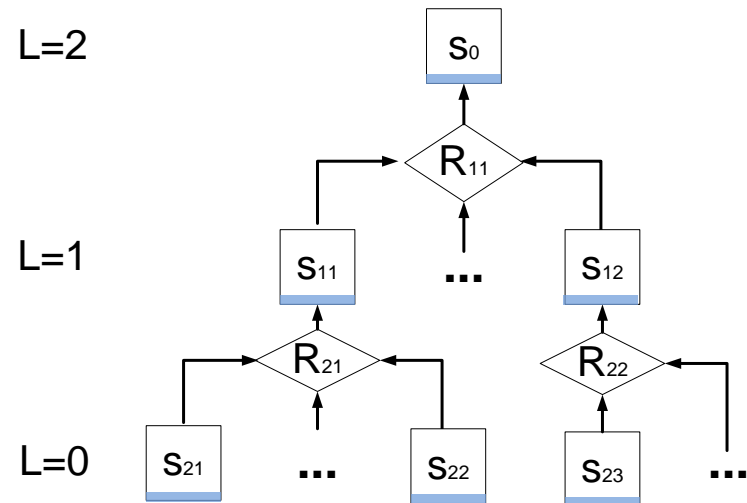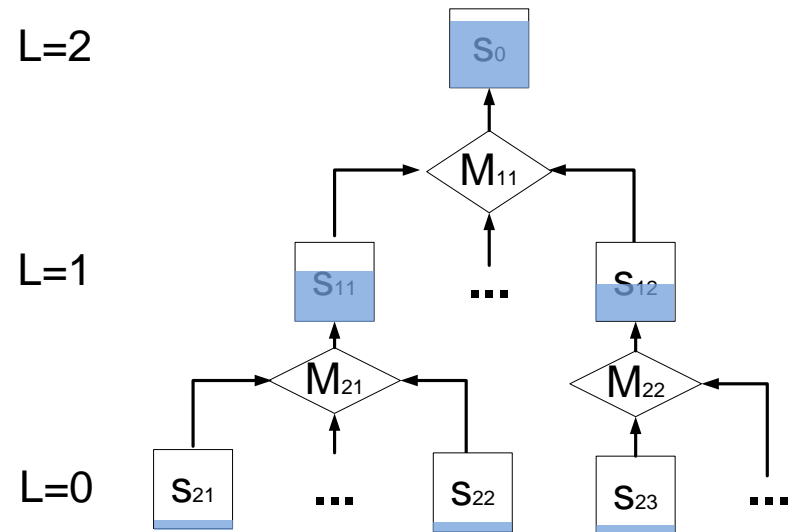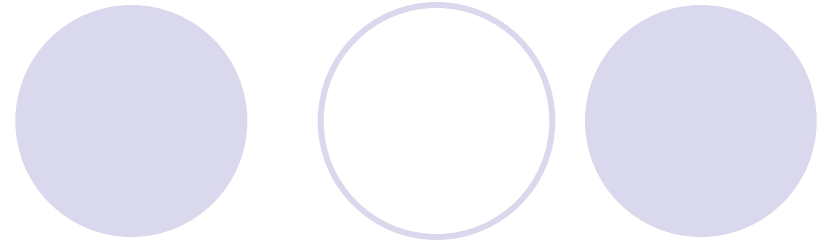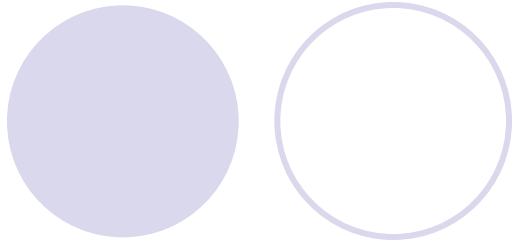
- Initial label is L = "2" (subtract L by 1 when forwarding this request to QQ friends)

- When L = 0, return the following:
  - Basic information (B)
  - Number of friends (N)
  - Timestamps (T)

- When L > 0, do the following:
  - Forward this request to all QQ friends
  - Pack the first 10 replies, together with your own information (B, N, T), and send them back to me

L=2

S_0

M_{11}

L=1 S_{11} ... S_{12}

M_{21} M_{22}

L=0 S_{21} ... S_{22} S_{23} ...

Summary

Acknowledgements

CONCLUSION

Summary

- HPU as a new paradigm to compliment the traditional CPU-based computing for big data

- Many unexplored algorithmic problems
  - Worker selection
  - Social connections of workers
  - Workflow design
  - Cost-time-quality trade-offs
  - Incentive mechanisms

Acknowledgements

○ Wei Chang

   Temple University

○ Grace Ju

   Carnegie Mellon University