Top-*k* Processing for Search and Information Discovery in Social Applications

Lecture 4: User Studies

Sihem Amer-Yahia

Julia Stoyanovich

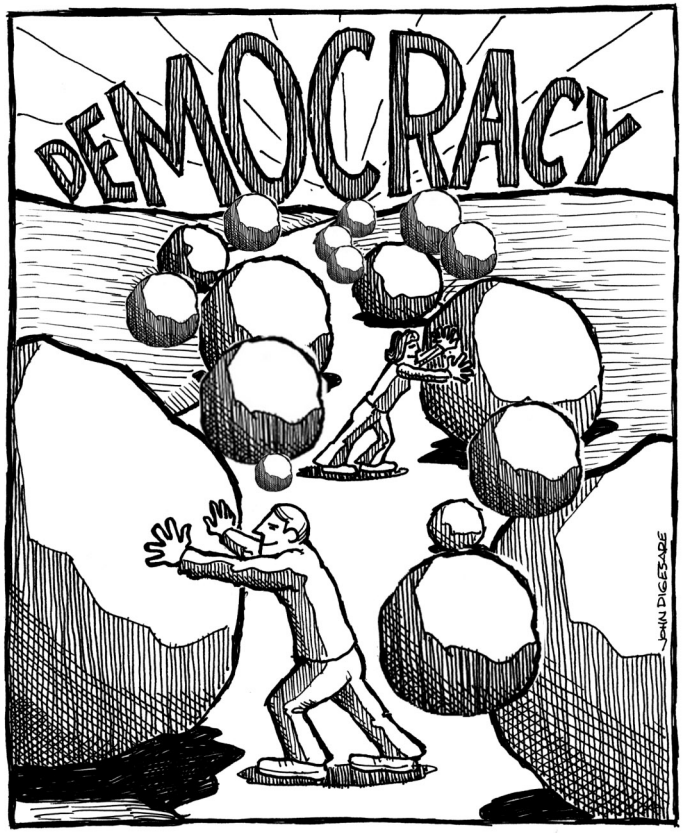
Social top-*k* @ *Joint RuSSIR/EDBT Summer School 2011*

Summary of last lectures

- **Fundamental algorithms**
  - Use the inverted list indexing structure
  - Have an access strategy and a stopping condition
  - TA – instance-optimal over the class of *reasonable* algorithms
  - NRA – useful when random access is expensive or impossible
- **Network-aware search**
  - Ubiquitous on the Social Web
  - Careful modeling of inverted lists enables top-*k* applicability
  - Space/time tradeoff exploration for scalable network-aware search (Cluster-Seekers and Cluster-Taggers)
- **Group recommendation**
  - Top-k algorithms for ad-hoc groups rely on pre-computed pair-wise disagreement lists
  - Space-saving strategies for disagreement lists

Quote of the day

Democracy is not a spectator sport. ~Lotte Scharfman



How do we evaluate accuracy?

- **Reflects how satisfied users are with ranked results**
- **Typical approaches:**
  - Online deployment
  - A/B testing/split testing (slide 12 in network-aware search class)
  - User studies: offline controlled experiments
- **2 applications**
  - Group recommendation
  - Itinerary recommendation

Outline

- ✓ Intro
- **Amazon Mechanical Turk**
- **User study for group recommendation**
- **User study for itinerary extraction**

Experimental methodology – AMT

- **Amazon Mechanical Turk (AMT) is based on Human Intelligent Tasks (HITs)**
  - The concept of AMT is to provide a crowd-sourcing marketplace where *requesters* (i.e., individuals or institutions who have tasks to be completed) and *workers* (i.e., individuals who can perform the tasks in exchange for monetary reward) can come together.
  - AMT provides a platform where the tasks (i.e. HITs) are hosted and executed, money is transferred securely, and the reputation of workers and requesters is tracked.
- **HITs allow seeking feedback from a large number of participants**

Lifecycle of a HIT (ICWSM'11 tutorial)

- **Requester builds a HIT**
  - Internal HITs are hosted by Amazon
  - External HITs are hosted by the requester
  - HITs can be tested on {requester, worker}sandbox.mturk.com
- **Requester posts HIT on mturk.com**
  - Can post as many HITs as account can cover
- **Workers do HIT and submit work**
- **Requester approves/rejects work**
  - Payment is rendered
  - Amazon charges requesters 10%
- **HIT completes when it expires or all assignments are completed**

Main API functions (ICWSM'11 tutorial)

- **CreateHIT** (Requirements, Pay rate, Description) **– returns HIT Id and HIT Type Id**

– SubmitAssignment (AssignmentId) – notifies Amazon that this assignment has been completed

– ApproveAssignment (AssignmentID) – Requester accepts assignment, money is transferred, also RejectAssignment

– GrantBonus (WorkerID, Amount, Message) – Give the worker the specified bonus and sends message, should have a failsafe

– NotifyWorkers (list of WorkerIds, Message) – e-mails message to the workers.

Command-line tools (ICWSM'11 tutorial)

- **Configuration files**
  - mturk.properties – for interacting with MTurk API
  - [*task name*].input – variable name & values by row
  - [*task name*].properties – HIT parameters
  - [*task name*].question – XML file
- **Shell scripts**
  - run.sh – post HIT to Mechanical Turk (creates .success file)
  - getResults.sh – download results (using .success file)
  - reviewResults.sh – approve or reject assignments
  - approveAndDeleteResults.sh – approve & delete all unreviewed HITs
- **Output files**
  - [*task name*].success – created HIT ID & Assignment IDs
  - [*task name*].results – tab-delimited output from workers

Outline

- ✓ Intro
- ✓ Amazon Mechanical Turk
- **User study for group recommendation**
- **User study for itinerary extraction**

GroupRecs experiments on AMT

- **Dataset**
  - *MovieLens* data set
  - 71,567 users, 10,681 movies, 10,000,054 ratings
- **User Studies**
  - Compare effectiveness of proposed Group Recommendation algorithms with existing approaches
  - Small and large groups of similar, dissimilar and random users are formed.
  - Algorithms Average Relevance Only (AR), Least Misery Only (LM), Consensus with Pair-wise Disagreements (RP), Consensus with Disagreement Variance (RV) are compared

User study

- **Four group recommendation mechanisms**
  - Average Rating (AR)
  - Least-Misery Only (MO)
  - Consensus with Pairwise Disagreement (RP)
  - Consensus with DisagreementVariance (RV)
- **User collection phase**
  - Recruit users
  - Obtain their movie preferences
  - Group formation
  - Group size and group cohesiveness

User study

- **Group judgment phase**
  - obtain ground truth judgments on movies by users in a group setting.
- **Result interpretation*:***
  - user similarity in a group as well as group size should be accounted in modeling disagreement in the consensus function
- **Effectiveness of group ratings**
  - proposed group recommendation strategies are highly rated

User study results

Table 3 Dissimilar user group—overall model ratings

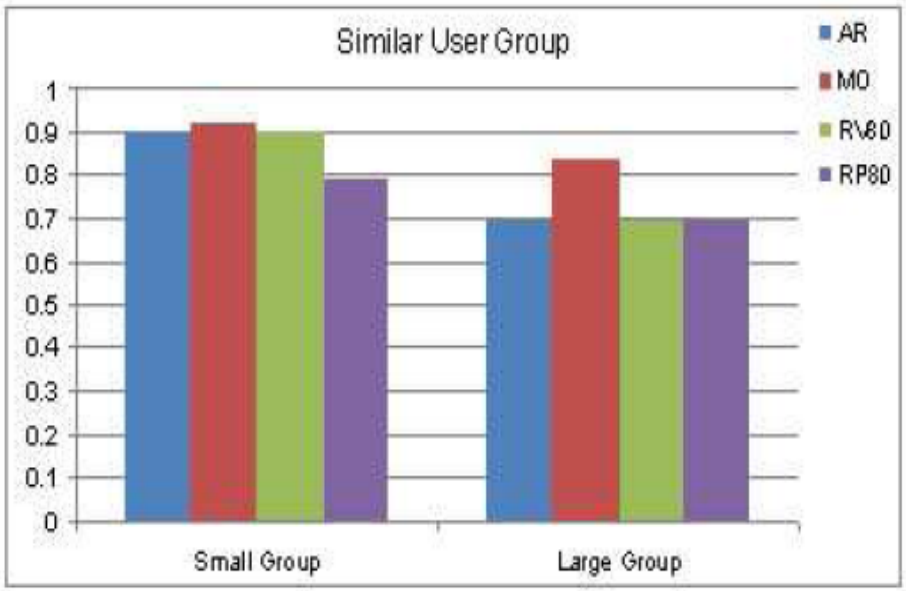
Rating	RP80		RV80	
	Small (%)	Large (%)	Small (%)	Large (%)
1	0	0	0	0
2	5	0	8	3
3	31	20	28	17
4	42	44	60	36
5	22	36	4	44

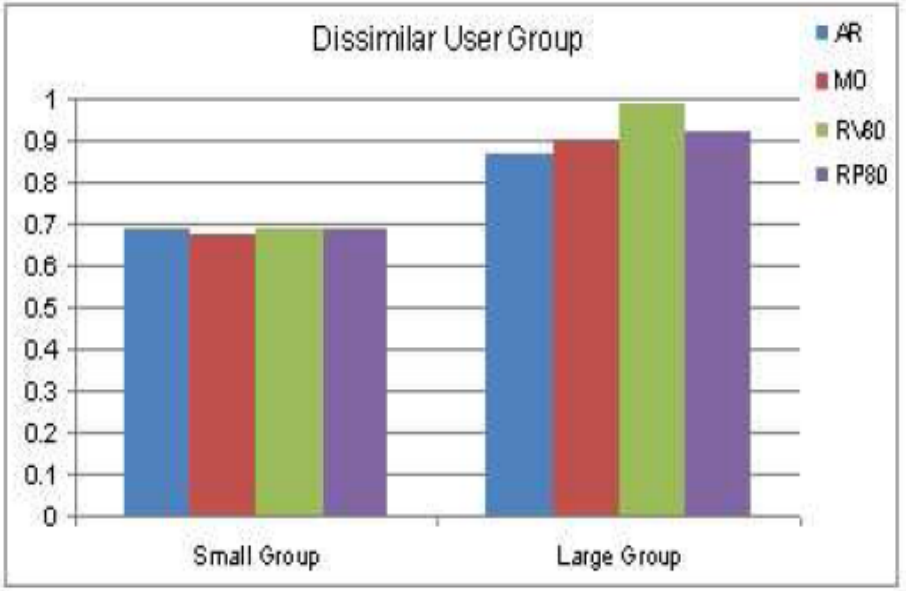
User study results

Table 4 Similar user group—overall model ratings

Rating	RP80		RV80	
	Small (%)	Large (%)	Small (%)	Large (%)
1	3	0	5	0
2	14	0	8	0
3	14	14	20	11
4	52	30	40	41
5	17	56	27	48

Disagreement is important for Dissimilar User Groups





- **Misery Only (MO) is the best model for similar user group.**
- **Disagreement is important for dissimilar users. Consensus with Disagreement Variance (RV80) is the best model.**

Outline

- ✓ Intro
- ✓ Amazon Mechanical Turk
- ✓ User study for group recommendation
- • **User study for itinerary extraction**

Extracting travel itineraries from Flickr

Goal: extract the itinerary of each traveler by mapping photos into Points Of Interest (POIs) and aggregate actions of many travelers into coherent queryable itineraries.

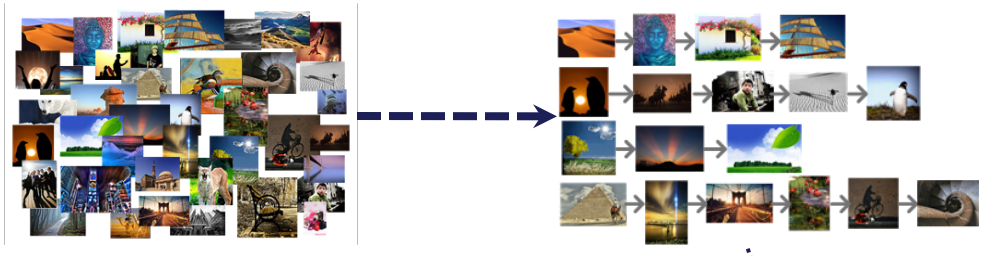
- **Feedback on various aspects of the itineraries constructed by our system from a large number of anonymous users**

Problem definition

- **Definitions**
  - Each itinerary is a timed path
  - The set of timed paths implies a *weighted graph* G over POIs
  - An *itinerary* is a path in the graph G
  - The *value* of an itinerary is the sum of popularities of its POIs
  - The *time* of an itinerary is the sum of POI visit and transit times

- **Problem Instance ("Orienteering")**
  - Find an itinerary in G from a *source* POI to a *target* POI of budget (=time) at most B maximizing total value
  - The time budget B is typically whole days
  - *source* and *target POIs* provided by user (e.g. hotel)

Photo Streams

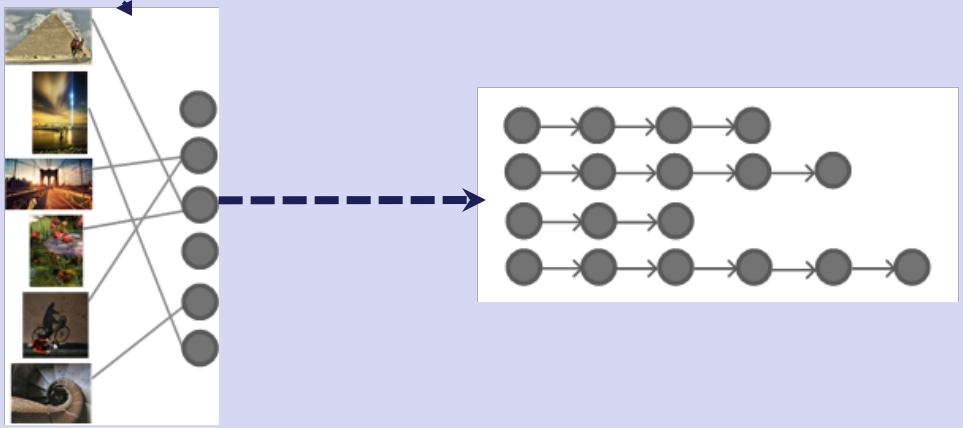


- *Identify photos of a given city*
- *Filter out residents of a city*
- *Validate photo timestamps*

Photo-POI Mapping

- *Extract Candidate POIs*
  - *Lonely Planet/Y! Travel to extract landmarks*
  - *Yahoo! Maps API to retrieve their geo-locations*
- *Tag & geo-based POI association*

Timed Paths



- *Photo Streams Segmentation*
  - *Split the stream whenever the time difference between two successive photos is "large"*
- *Distillation of Timed Visits*
  - *<POI, start time, end time>*
- *Construction of Timed Paths*
  - *A sequence of Timed Visits*

Data preparation

- **Five popular and geographically distributed cities were chosen: Barcelona, London, New York City (NYC), Paris, and San Francisco**
- **For each city, we generate four itineraries using our system**

City	#POIs	#Timed Paths	Sample POIs
Barcelona	74	6,087	Museu Picasso, Plaza Reial
London	163	19,052	Buckingham Palace, Churchill Museum, Tower Bridge
New York City	100	3,991	Brooklyn Bridge, Ellis Island
Paris	114	10,651	Tour Eiffel, Musee du Louvre
San Francisco	80	12,308	Aquarium of the Bay, Golden Gate Bridge, Lombard Street

Itinerary generation

- **For each city, we generate four itineraries using our system.**
- **We first select the city's four most popular POIs and designate them as ℓ_1 (most popular) through ℓ_4.**
  - The popularity of a POI is determined by the number of distinct users who have provided a photo associated with the POI.
- **The four itineraries for each city are then constructed by setting the starting point and ending point as (ℓ_1, ℓ_3), (ℓ_1, ℓ_4), (ℓ_2, ℓ_3), (ℓ_2, ℓ_4), with a time budget of 12 hours.**

Example itinerary for NYC (single-day)

Time **09:00** : Start from **ground zero**
Time **09:00** : Spend 27 minutes at **ground zero**.
Time **09:27** : Transit to **empire state building** (estimated travel time: 52 minutes)
Time **10:19** : Spend 1 hour and 13 minutes at **empire state building**.
Time **11:32** : Transit to **new york public library** (estimated travel time: 15 minutes)
Time **11:47** : Spend 29 minutes at **new york public library**.
Time **12:16** : Transit to **radio city music hall** (estimated travel time: 24 minutes)
Time **12:43** : Spend 51 minutes at **radio city music hall**.
Time **13:34** : Transit to **central park** (estimated travel time: 23 minutes)
Time **13:57** : Spend 40 minutes at **central park**.
Time **14:37** : Transit to **rockefeller center** (estimated travel time: 33 minutes)
Time **15:10** : Spend 37 minutes at **rockefeller center**.
Time **15:47** : Transit to **grand central terminal** (estimated travel time: 22 minutes)
Time **16:09** : Spend 27 minutes at **grand central terminal**.
Time **16:36** : Transit to **chrysler building** (estimated travel time: 6 minutes)
Time **16:42** : Spend 31 minutes at **chrysler building**.
Time **17:13** : Transit to **brooklyn bridge** (estimated travel time: 32 minutes)
Time **17:45** : Spend 36 minutes at **brooklyn bridge**.
Time **18:21** : Transit to **statue of liberty** (estimated travel time: 21 minutes)
Time **18:42** : Spend 42 minutes at **statue of liberty**.
Time **19:24** : Transit to **little korea** (estimated travel time: 26 minutes)
Time **19:50** : Spend 31 minutes at **little korea**.
Time **20:21** : Transit to **ground zero** (estimated travel time: 38 minutes)

Goal of user study

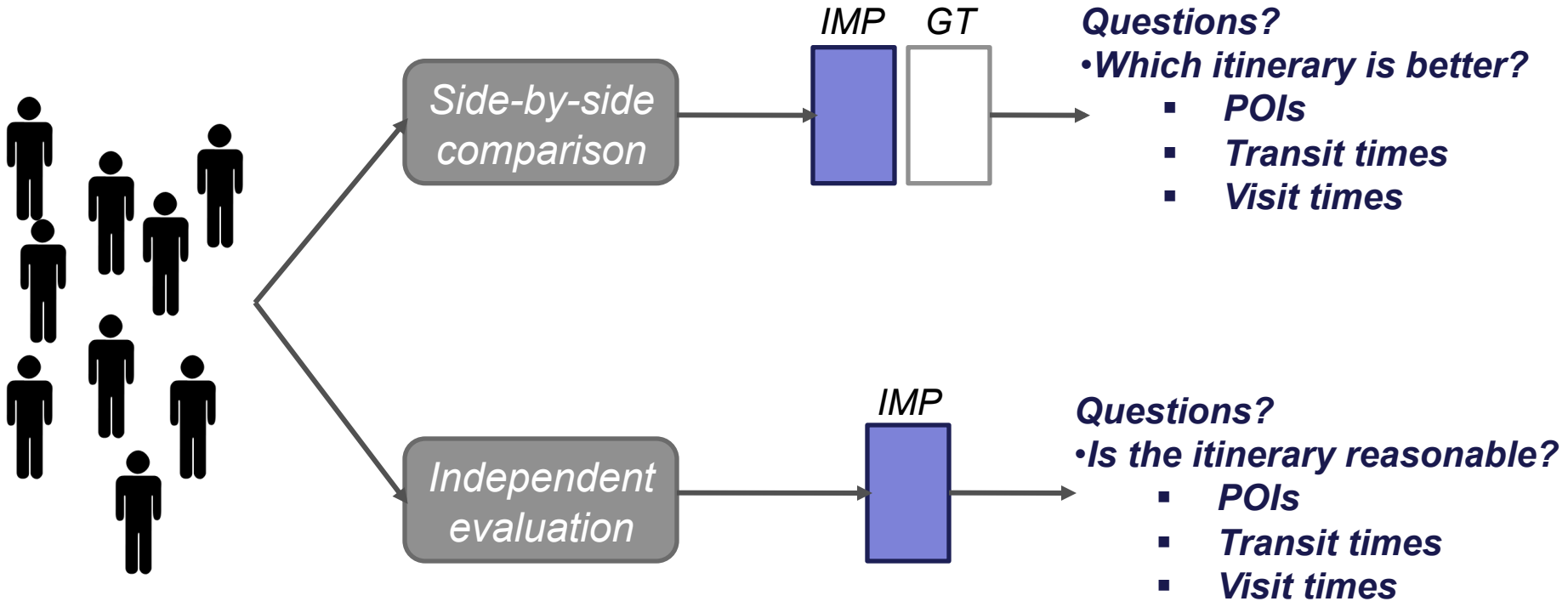
- **Estimate the usefulness of the itineraries from two aspects:**
  - overall utility of the itineraries
  - appropriateness of POIs
- **Challenge**
  - design a set of questions to AMT users and collect and interpret feedback
  - what is our ground truth?

Ground truth

City	Ground Truth Sources
Barcelona	www.barcelona-tourist-guide.com
London	www.theoriginaltour.com
New York City	www.newyorksightseeing.com
Paris	www.carsrouges.com
San Francisco	www.allsanfranciscotours.com

User study design summary

- **Side-by-side evaluation comparing our itineraries to ground-truths**
- **Independent evaluation examining our itineraries in detail**



Comparative evaluation

Evaluation Questions:

I. Overall, which one of the above two proposed itineraries you would rate higher?
- Itinerary 1 is significantly more useful than Itinerary 2.
- Itinerary 1 is somewhat more useful than Itinerary 2.
- Both are similar.
- Itinerary 2 is somewhat more useful than Itinerary 1.
- Itinerary 2 is significantly more useful than Itinerary 1.

II. How would you rate the set of points of interest included in the two itineraries?
- Itinerary 1 has significantly more appropriate points of interest than Itinerary 2.
- Itinerary 1 has somewhat more appropriate points of interest than Itinerary 2.
- Both are comparatively similar.
- Itinerary 2 has somewhat more appropriate points of interest than Itinerary 1.
- Itinerary 2 has significantly more appropriate points of interest than Itinerary 1.

III. How would you rate the transit times at the points of interest in the two itineraries (from a tourist perspective)?
- Itinerary 1 has significantly more accurate transit times than Itinerary 2.
- Itinerary 1 has somewhat more accurate transit times than Itinerary 2.
- Both are comparatively similar.
- Itinerary 2 has somewhat more accurate transit times than Itinerary 1.
- Itinerary 2 has significantly more accurate transit times than Itinerary 1.

IV. Any additional comments?

Overall itinerary quality comparison

Evaluation of the quality of suggested POIs

Transit time evaluation across consecutive POIs

Independent evaluation

Q1: Overall, would you rate the proposed itinerary as:
—Not at all useful to a tourist
—Not so useful to a tourist
—Somewhat useful to a tourist
—Very useful to a tourist

Q2: How would you rate the set of points of interest included in the itinerary?
—Make no sense
—Mostly inappropriate
—Somewhat appropriate
—Mostly appropriate

Q3: How would you rate the visit times at the landmarks, as proposed by the itinerary (from a tourist perspective)?
—Not accurate at all
—Somewhat accurate
—Mostly accurate
—Completely accurate
If you picked choices 3 or 4, did you find the visit times too short or too long?

Q4: How would you rate the transit times between the landmarks, as proposed by the itinerary (from a tourist perspective)?
—Not accurate at all
—Somewhat accurate
—Mostly accurate
—Completely accurate
If you picked choices 3 or 4, did you find the transit times too short or too long?

Evaluation measures

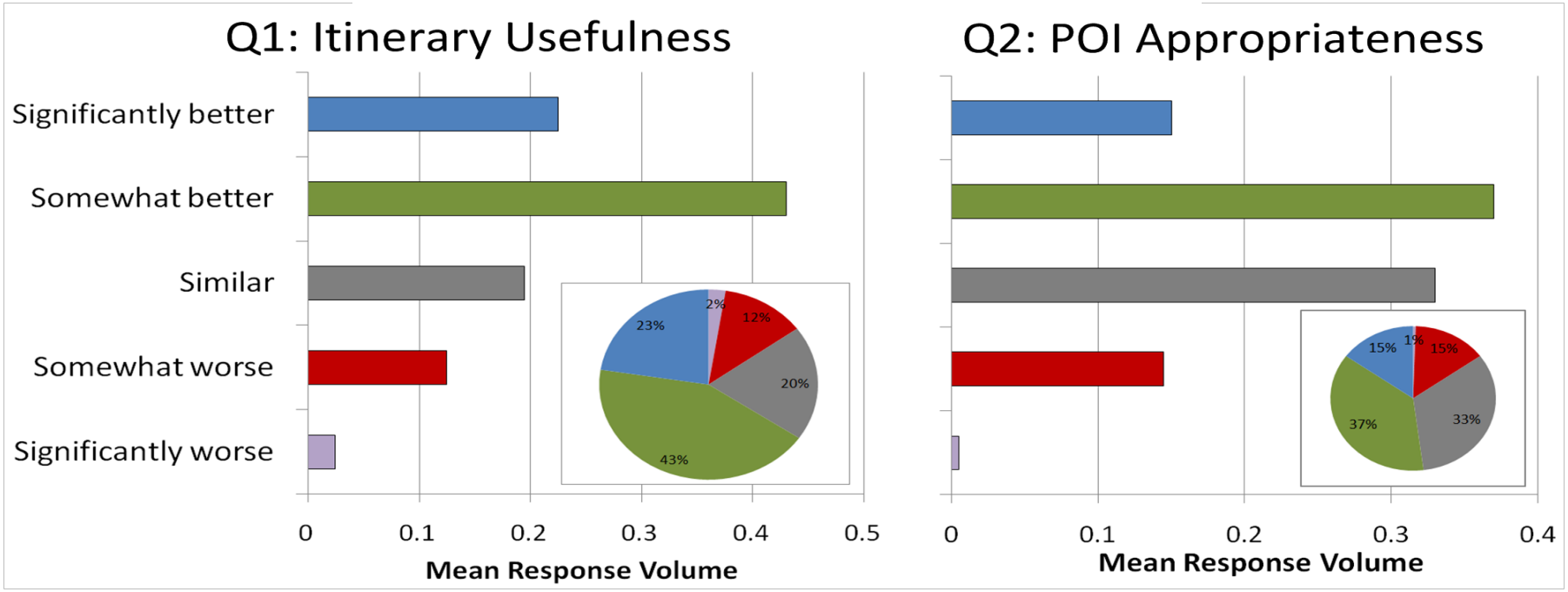
- *Mean Weighted Response (MWR)* – aggregate the responses to each question from the workers in the same group, into a single number. Take mean across different itineraries generated by our method.

- *Mean Average Error Fraction (MAEF)* – compute the percentage of the number of POIs, visit times, or transit times, that are considered bad or inaccurate by a particular worker, out of the total number of POIs

Results for side-by-side comparison

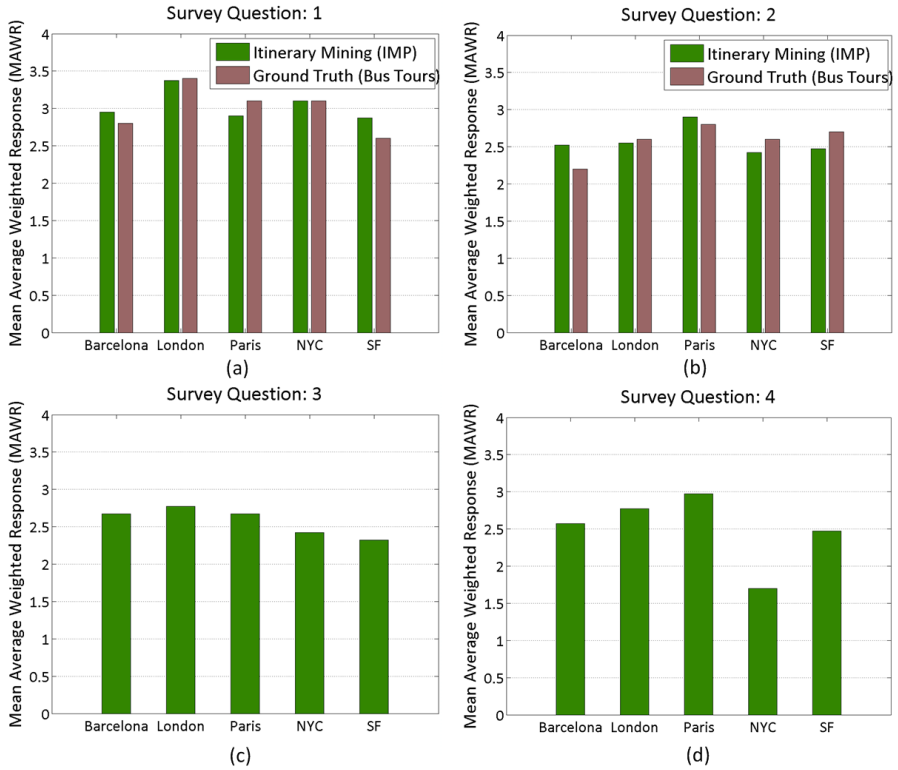
$$\mathrm{MRV}(opt, q) = \frac{1}{n_q(opt)} \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} \sum_{\mathcal{I}} n_q^{\mathcal{I},C}(opt), \qquad (1)$$

where $n_q^{\mathcal{I},C}(opt)$ is the number of workers who chose the option *opt* in question q for the HIT involving our system-generated itinerary I and city C; and $n_q(opt)$ is the total number of workers who responded to option *opt* for question q across all HITs.

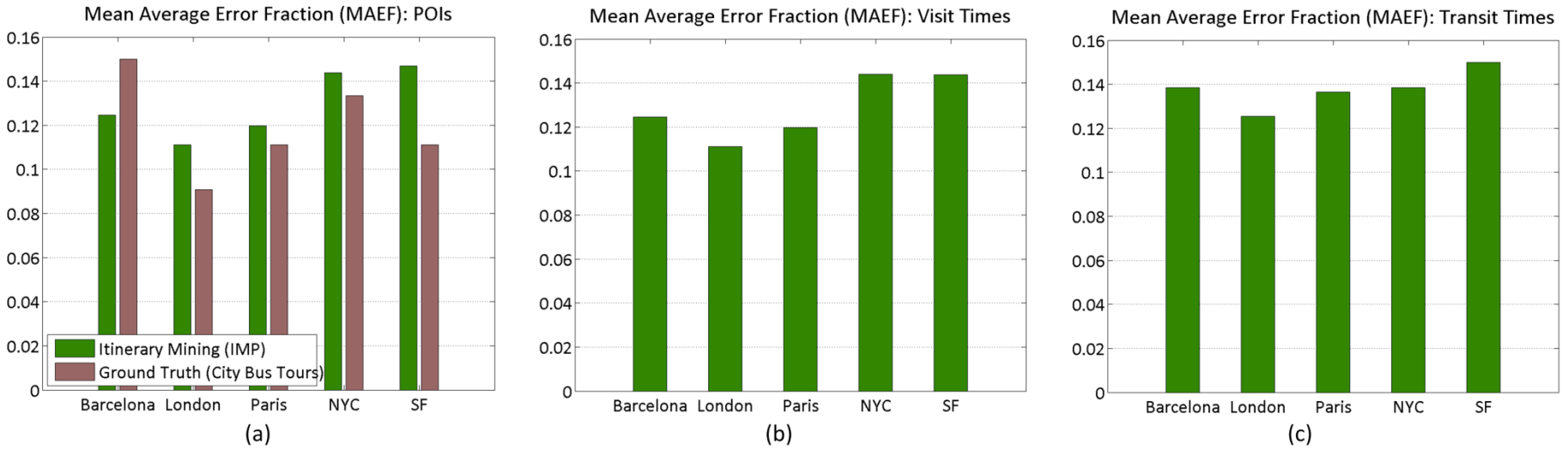


MWR for London Itineraries

London Itineraries	Q1	Q2	Q3	Q4
IMP It. 1	3.1	2.9	2.7	2.8
IMP It. 2	3.5	2.1	2.7	2.1
IMP It. 3	3.4	2.5	2.8	2.7
IMP It. 4	3.5	2.7	2.9	3.1
Ground Truth	3.4	2.6	2.6	2.6



The mean error fraction of (a) POIs, (b) Visit Times, and (c) Transit Times:



Summary and challenges

- **AMT enables scaling up user studies to hundreds, thousands of users**
- **AMT is just a hiring platform**
- **Experiment designer must "track" users and enforce consistency**
  - in group recommendations, have users really seen the movies they are asked to rate to build their profile?
  - in itinerary planning, do hired users really know about a city?

Filtering expert AMT workers

- **Multiple-choice questions on "less-known" POIs**

QUALIFICATION EVALUATION

Please choose the most suitable name of the point of interest based on your experience. This would judge your fitness to take the travel itinerary evaluation task in the next section.



○ Empire State Building	○ Flatiron Building	○ Herald Square
○ Rockefeller Center	○ Saint Patrick's Cathedral	○ Washington Sq Park
○ Chrysler Building	○ Trinity Church	○ Lincoln Center

More challenges (ICWSM'11 tutorial)

- What are the conditions in which workers perform differently than the laboratory setting?
- How often does one person violate Amazon's terms of service by controlling more than one worker account?
- Although the demographics of the workers on Mechanical Turk is clearly not a sample of either the U.S. or world populations, could one devise a statistical weighting scheme to achieve this?
- Since workers are tied to a unique identifier (their Worker ID) one could conduct long term, longitudinal studies about how their behavior changes over time.

[*task name*].results (ICWSM'11 tutorial)

hitid	Assignment id	Worker id	accepted	submitted	feed back	reject	Answer. bonus
14SBGD GM5ZHZ FE3OU26 DJESC20 DXKY	1BPE1URVWQKM6DSG40M WDVKIAJ93B4	A2IB92P5729K3Q	Sat Oct 02 16:03:49 EDT 2010	Sat Oct 02 16:43:55 EDT 2010			1.39
14SBGD GM5ZHZ FE3OU26 DJESC20 DXKY	1GMFLPGSL0NMWZJSTFXN J1FS74J6KW	A2LKKOAIMEF1PT	Sat Oct 02 16:10:23 EDT 2010	Sat Oct 02 16:44:33 EDT 2010			1.54
14SBGD GM5ZHZ FE3OU26 DJESC20 DXKY	1VQ5ID82X6TJXBU4EKXYI SVF8C4BWJ	A15T1WFW5B2OPR	Sat Oct 02 16:13:22 EDT 2010	Sat Oct 02 16:44:56 EDT 2010			1.49
14SBGD GM5ZHZ FE3OU26 DJESC20 DXKY	16XXR2KPFCB31UOCMBG7 8KLMAD4HND	A16ME0W2U4THE0	Sat Oct 02 16:00:21 EDT 2010	Sat Oct 02 16:45:08 EDT 2010			1.67

References and further reading

1. *Automatic construction of travel itineraries using social breadcrumbs*.
   Munmun De Choudhury, Moran Feldman, Sihem Amer-Yahia, Nadav Golbandi, Ronny Lempel, Cong Yu. HyperText 2010.

2. *Space Efficiency in Group Recommendation.*
   Senjuti Roy, Sihem Amer-Yahia, Ashish Chawla, Gautam Das, Cong Yu. VLDB J. 2010.

3. *How to use Mechanical Turk for Behavioral Research?*
   Winter Mason and Siddharth Suri. CWSM 2011 (panel).

Questions?