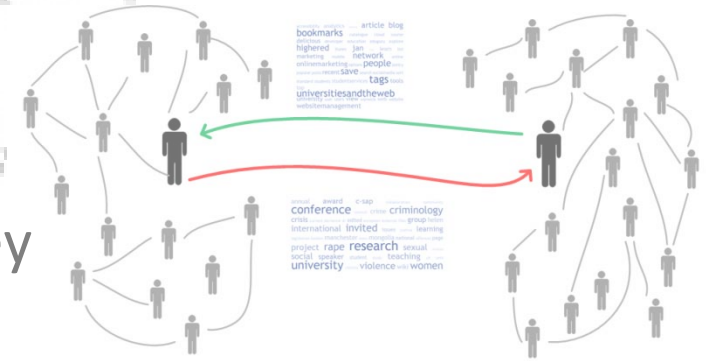
# Exploring the Social Media Landscape to Streamline Everyday Experiences

***Munmun De Choudhury***

School of Communication & Information

Rutgers, The State University of New Jersey



# Modern Social Interactional Modes



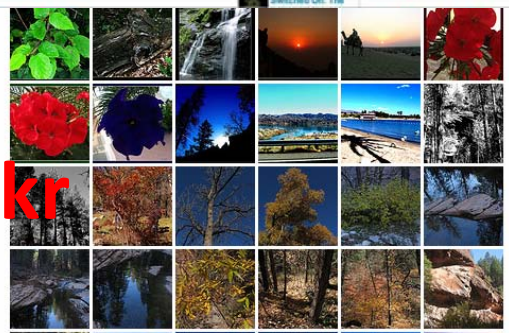











Twitter
Reddit
Blogger
MetaFilter
Orkut
MySpace
YouTube

Some of my most fav pics are here!

# 140 characters can cause revolutions


During the
elections in Iran


And during the earthquake in Haiti

# Sustainability of culture in the digital society

And to streamline our everyday life too!

**How do geo-temporal social breadcrumbs enable better decision making of our day-to-day actions?**

**How can social media inform us better about locally and globally distributed events?**

# Research Question 1

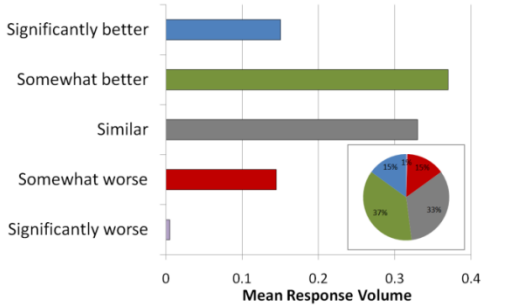
With colleagues at Yahoo! Research, New York

# Social geo-temporal breadcrumbs in travel itinerary planning

# Our Contributions

- ***Goal***
  - Construct intra-city travel itineraries automatically by tapping a latent source reflecting geo-temporal breadcrumbs left by millions of tourists – Flickr.
- ***Approach***
  - Extract photo streams of individual users.
  - aggregate all user photo streams into a POI graph; apply the orienteering algorithm to construct itineraries
- Our extensive survey-based user studies over about 450 workers on AMT indicate that high quality itineraries can be automatically constructed from Flickr data.

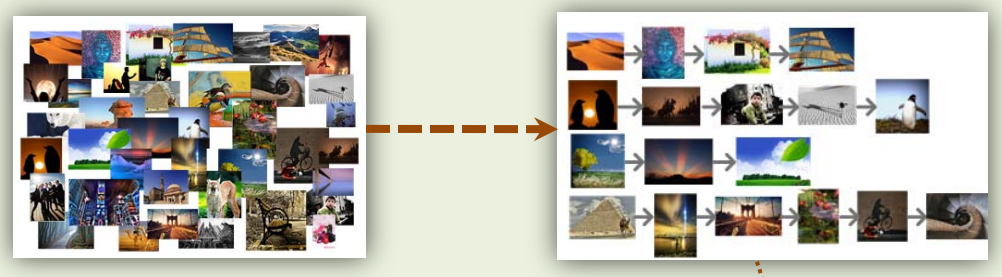
Time 09:00 : Start from ground zero
Time 09:00 : Spend 27 minutes at ground zero.
Time 09:27 : Transit to empire state building (estimated travel time: 52 minutes)
Time 10:19 : Spend 1 hour and 13 minutes at empire state building.
Time 11:32 : Transit to new york public library (estimated travel time: 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)



# Timed Paths

Step I

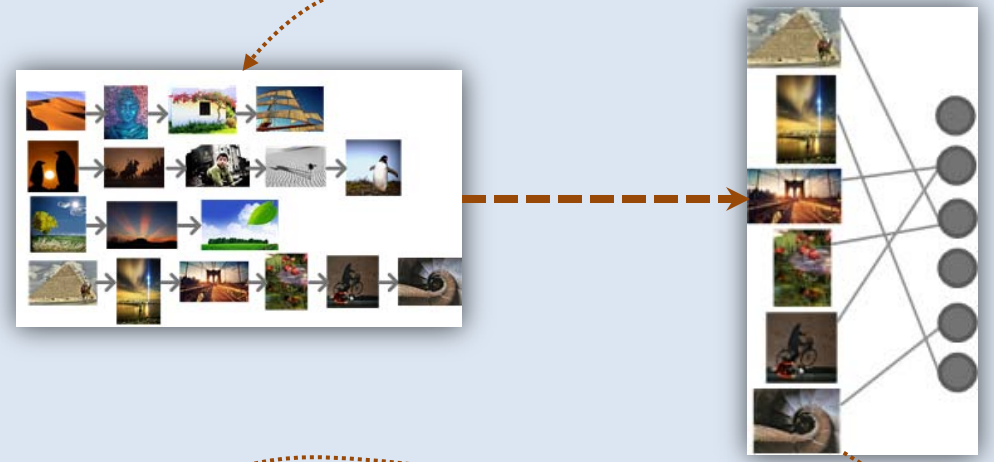
Photo Streams



- Identifying photos of a given city
- Filtering residents of a city
- Photo taken time verification

Step II

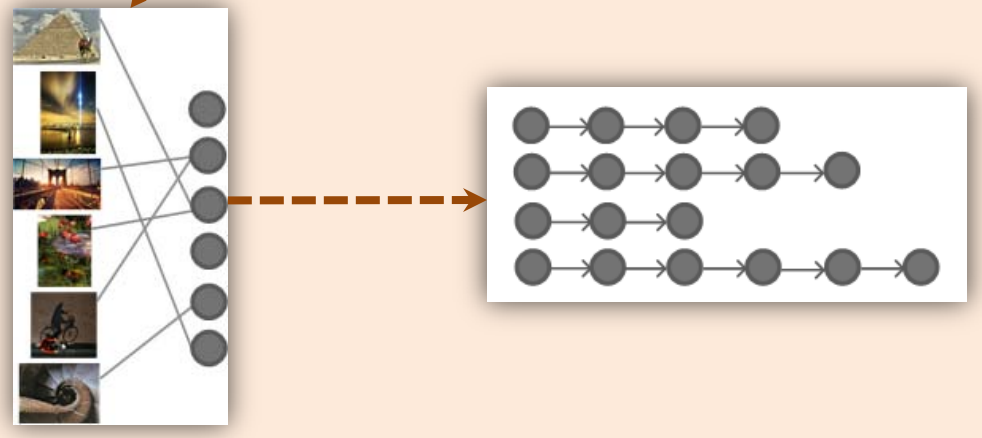
Photo-POI Mapping



- Extracting Candidate POIs
  - Lonely Planet to extract landmarks
  - Yahoo! Maps API to retrieve their geo-locations
- Tag-based / geo-based association

Step III

Timed Paths



- Segmentation of Photo Streams
  - Split the stream whenever the time difference between two successive photos of a user is greater than a threshold (8hrs)
- Construction of Timed Paths
  - “Timed visit” – the triplet of POI name, start time, end time

# Constructing Itineraries

# Constructing POI Graph

- Given the set of timed paths, our goal is to aggregate the actions of many individual travelers into coherent itineraries while taking into consideration POI popularity.
- Undirected POI graph, $G_C(V=L_C, E=L_C \times L_C)$ with the following predicates:
  - $T(\ell \in L_C)$, the visit time at each POI $\ell$. Longest visit time for each user; take the $75^{th}$ percentile among all users
  - $T(e \in E)$, the median transit time between two POIs
  - $V(\ell \in L_C)$, the prize or value that an itinerary gets from visiting each POI $\ell$ in $L_C$, and is a function of the popularity and visit duration of $\ell$

# Itineraries and Orienteering Problem

- An *itinerary* is a path in the graph $G_C$, where a node (POI) in the path may be visited more than once.
- **Problem Instance:**
  - Let I be an itinerary; its prize V(I) is defined as the sum of prizes of the unique set of POIs (i.e., a POI's prize is counted only once even if it is visited multiple times) along the path.
  - The time T(I) of the itinerary is the sum of visit times to the unique set of POIs along the path, plus the transit times along all edges on the path (including those that are traversed more than once).
- **Objective (solution using Orienteering Problem Approximation):**
  - Find an itinerary in $G_C$ from *s* to *t* of cost (=time) at most B maximizing total node prizes.
  - Note, B is typically whole days; *s* and *t* can be provided by the user

# Experimental Evaluation

# Data Preparation

- Five popular and geographically distributed cities were chosen: Barcelona, London, New York City (NYC), Paris, and San Francisco
- For each city, we obtained a list of POIs by pooling information from different sources (e.g., Lonely Planet, Wikipedia)

| 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 $\ell_1$ (most popular) through $\ell_4$.
  - 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 ($\ell_1$, $\ell_3$), ($\ell_1$, $\ell_4$), ($\ell_2$, $\ell_3$), ($\ell_2$, $\ell_4$), with a time budget of 12 hours.

## Single day itinerary

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)

## Two day itinerary

**DAY 1**

Time 09:00 : Start from Ground Zero
Time 09:00 : Transit to Metropolitan Museum of Art (estimated travel time: 1 hour and 43 minutes)
Time 10:43 : Spend 2 hours and 9 minutes at Metropolitan Museum of Art.
Time 12:52 : Transit to Empire State Building (estimated travel time: 1 hour and 30 minutes)
Time 14:22 : Spend 1 hour and 16 minutes at Empire State Building.
Time 15:38 : Transit to New York University (estimated travel time: 1 hour and 23 minutes)
Time 17:01 : Spend 18 minutes at New York University.
Time 17:19 : Transit to Staten Island Ferry (estimated travel time: 1 hour and 7 minutes)
Time 18:26 : Spend 1 hour and 10 minutes at Staten Island Ferry.
Time 19:36 : Transit to Ground Zero (estimated travel time: 56 minutes)
Time 20:32 : Reach Ground Zero

**DAY 2**

Time 09:00 : Start from Ground Zero
Time 09:00 : Transit to American Museum of Natural History (estimated travel time: 1 hour and 46 minutes)
Time 10:46 : Spend 2 hours and 25 minutes at American Museum of Natural History.
Time 13:11 : Transit to Wollman Skating Rink (estimated travel time: 1 hour and 2 minutes)
Time 14:13 : Spend 22 minutes at Wollman Skating Rink.
Time 14:35 : Transit to Rockefeller Center (estimated travel time: 1 hour and 2 minutes)
Time 15:37 : Spend 39 minutes at Rockefeller Center.
Time 16:16 : Transit to Radio City Music Hall (estimated travel time: 6 minutes)
Time 16:22 : Spend 30 minutes at Radio City Music Hall.
Time 16:52 : Transit to Chelsea Art Museum (estimated travel time: 34 minutes)
Time 17:26 : Spend 2 hours and 2 minutes at Chelsea Art Museum.
Time 19:28 : Transit to Grand Central Terminal (estimated travel time: 5 minutes)
Time 19:35 : Spend 17 minutes at Grand Central Terminal.
Time 19:52 : Transit to St Paul's Chapel (estimated travel time: 34 minutes)
Time 20:26 : Spend 26 minutes at St Paul's Chapel.
Time 20:52 : Transit to Ground Zero (estimated travel time: 4 minutes)
Time 20:56 : Reach Ground Zero

# Example Itineraries for the city NYC

# Ground Truth

- To compare our automatically constructed itineraries with baseline itineraries, we obtained itineraries provided by top tour bus companies for each city and considered them as ground truth itineraries
  - Visit or transit times do not come with typical bus tour itineraries; hence we derive these times using our system

| 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 |

# Experimental Methodology – AMT

- We design several user studies using the Amazon Mechanical Turk (AMT) based Human Intelligent Tasks (HITs).
- We seek feedback on various aspects of the itineraries constructed by our system from a large number of anonymous users
- **AMT:**
  - 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.

# Filtering Expert AMT Workers

- Qualification survey section in the user study:
  - Multiple choice questions on “less-known” POIs

**QUALIFICATION EVALUATION**

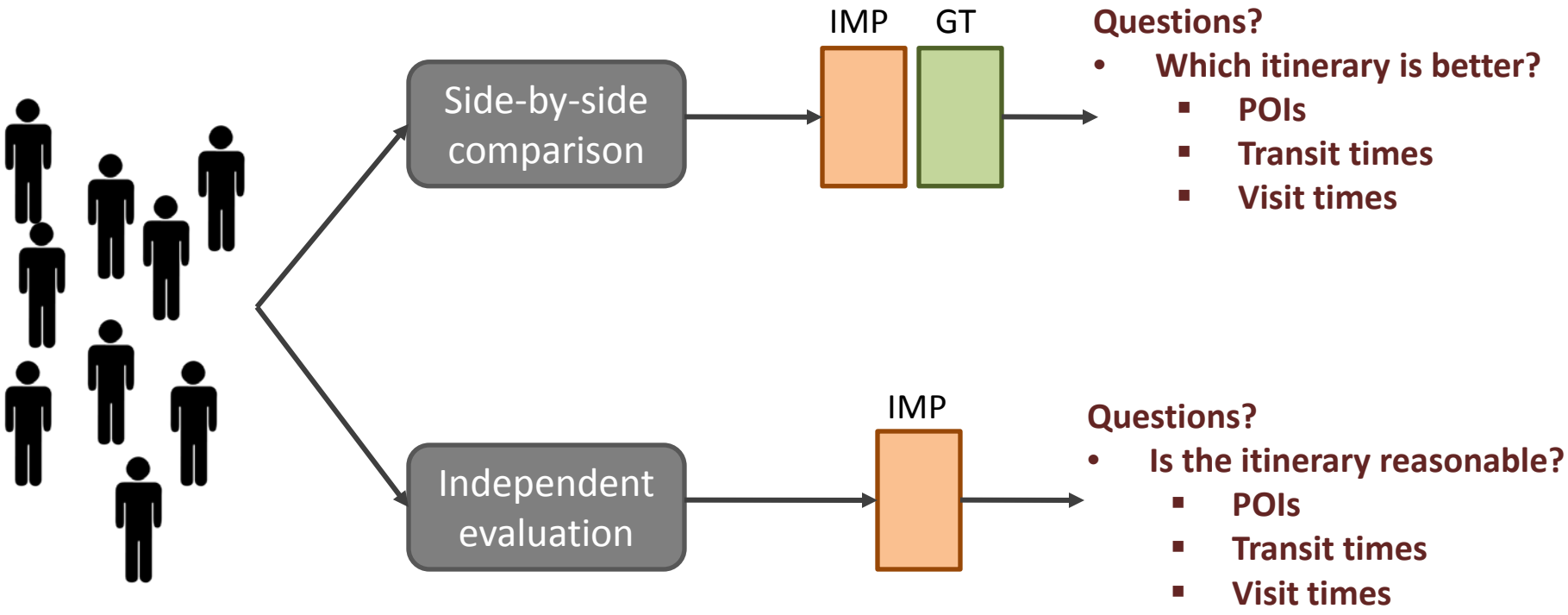
Please choose the most suitable name of the point of interest based on your experience. This would judge your fitness to take the travel itinerary evaluation task in the next section.



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

# 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?

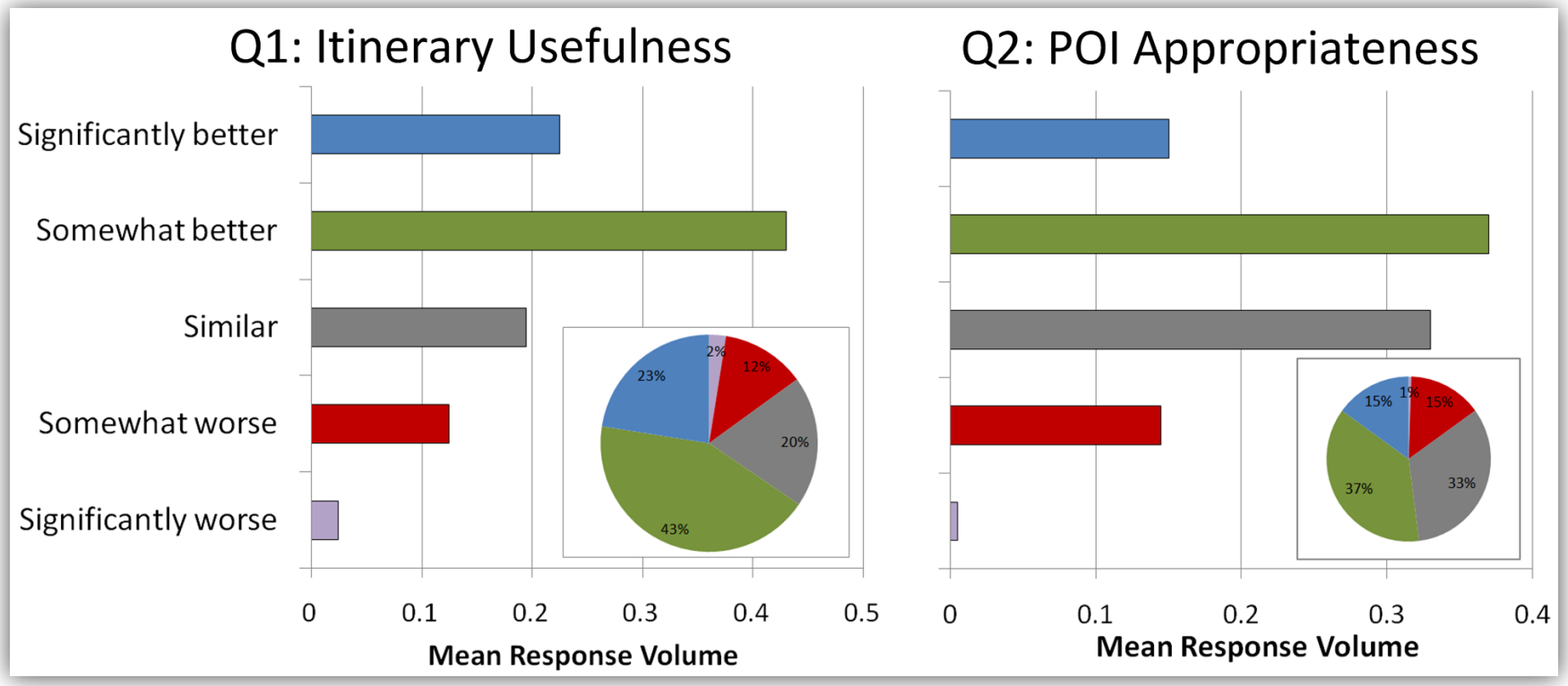
Overall itinerary quality comparison

Evaluation of the quality of suggested POIs

Transit time evaluation across consecutive POIs

# Results

- Evaluation Metric: estimate the usefulness of the itineraries from two aspects, such as the overall utility of the itineraries and appropriateness of POIs.
  - *Mean Response Volume* – it measures the number of worker responses received per option(in $Q_1$and $Q_2$) in the survey questionnaire.



# 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?

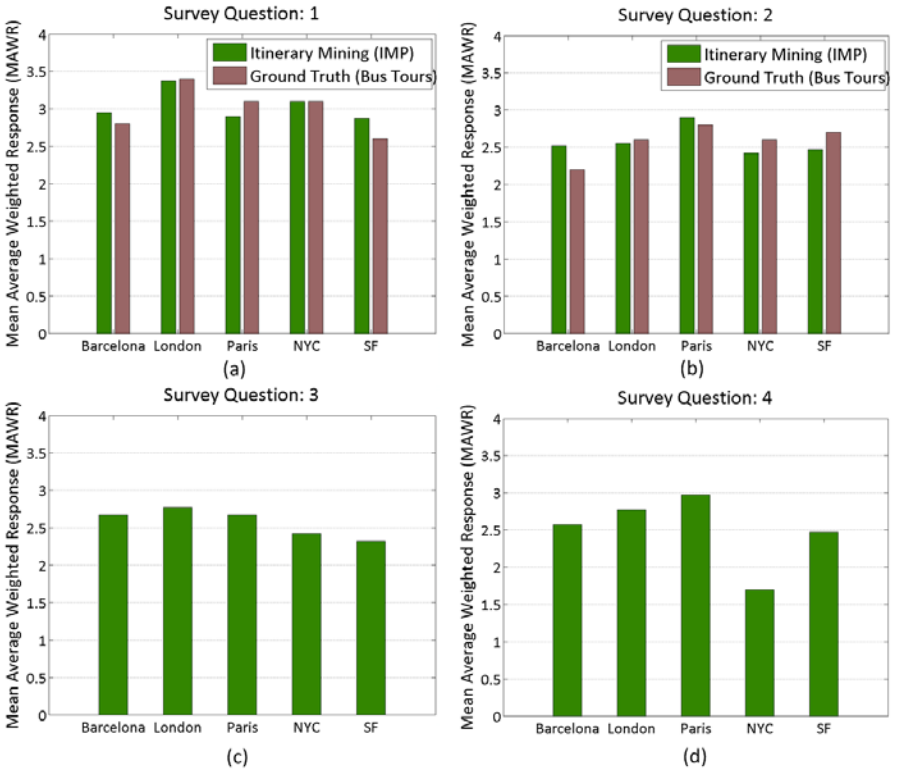
Q4: How would you rate the transit times between the landmarks, as proposed by the itinerary (from a tourist perspective)?
—Not accurate at all
—Somewhat accurate
—Mostly accurate
—Completely accurate
If you picked choices 3 or 4, did you find the transit times too short or too long?

# Results

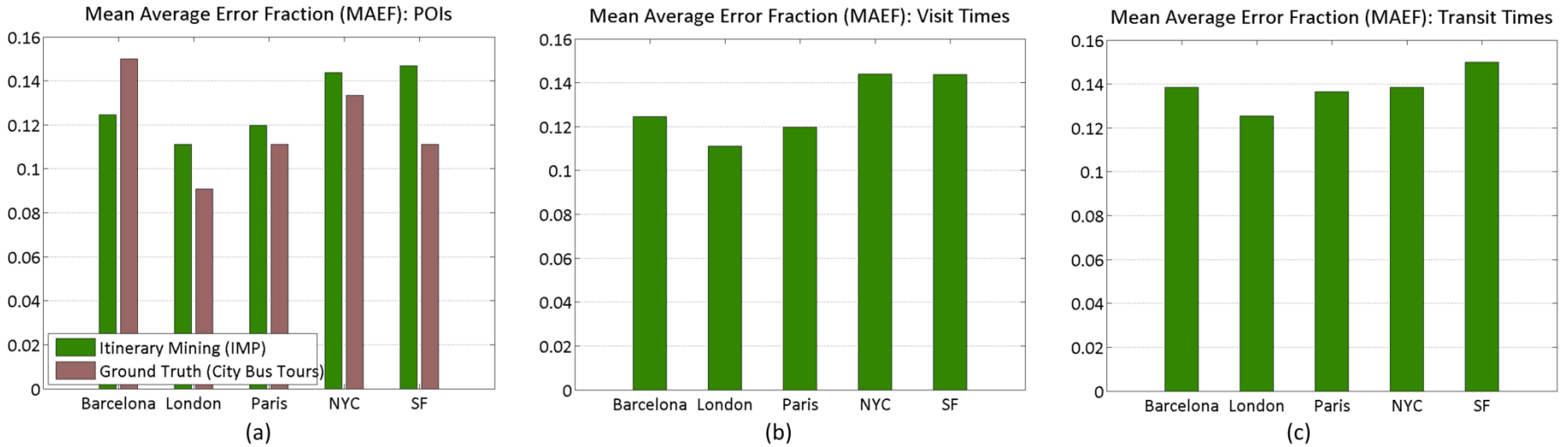
- Evaluation Metrics:
  - *Mean Weighted Response (MWR)* – aggregate the responses to each question from the workers in the same group, into a single number. Take mean across different itineraries generated by our method.
  - *Mean Average Error Fraction (MAEF)* – compute the percentage of the number of POIs, visit times, or transit times, that are considered bad or inaccurate by a particular worker, out of the total number of POIs

## 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 |



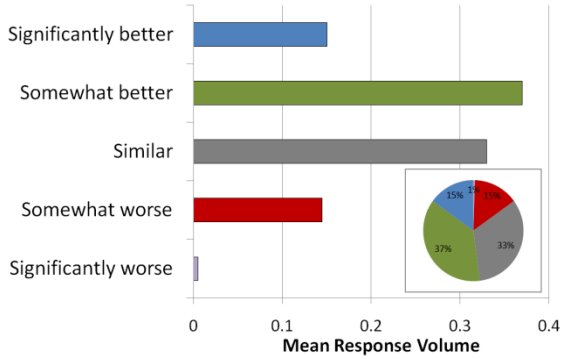

## The mean error fraction of (a) POIs, (b) Visit Times, and (c) Transit Times:



# Conclusions

- We addressed the question of automatic generation of travel itineraries for popular touristic cities from large-scale user contributed rich media repositories.
- Extensive survey-based user studies on AMT with promising results against bus tour companies' itineraries.
- ***Future extensions:***
  - Optimizing parameters
  - Incorporating traveler diversity
  - POI time constraints
  - Co-visitation patterns of users
  - City / POI coverage

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)



# Research Question 2

With colleagues at Rutgers University, NJ

# Geographically dispersed events


Key people?
Eyewitnesses?
Topical authorities?
Mass media?
Grassroot organizations?
Celebrities?

# Our Contributions

- Identifying user categories corresponding to events widely discussed on social media (Twitter)
  - Organizations
  - Journalists/Media bloggers
  - Ordinary individuals
  - Celebrities
- Develop a background training model
- Testing on a variety of events
- Exploring the user categories corresponding to these events

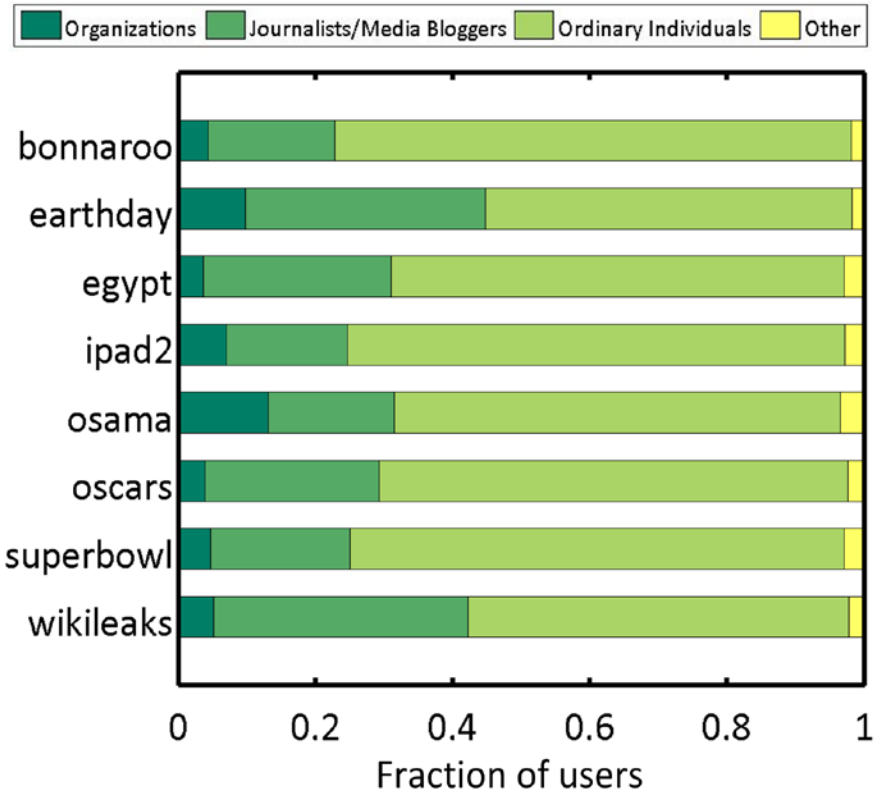
# Background Training Model

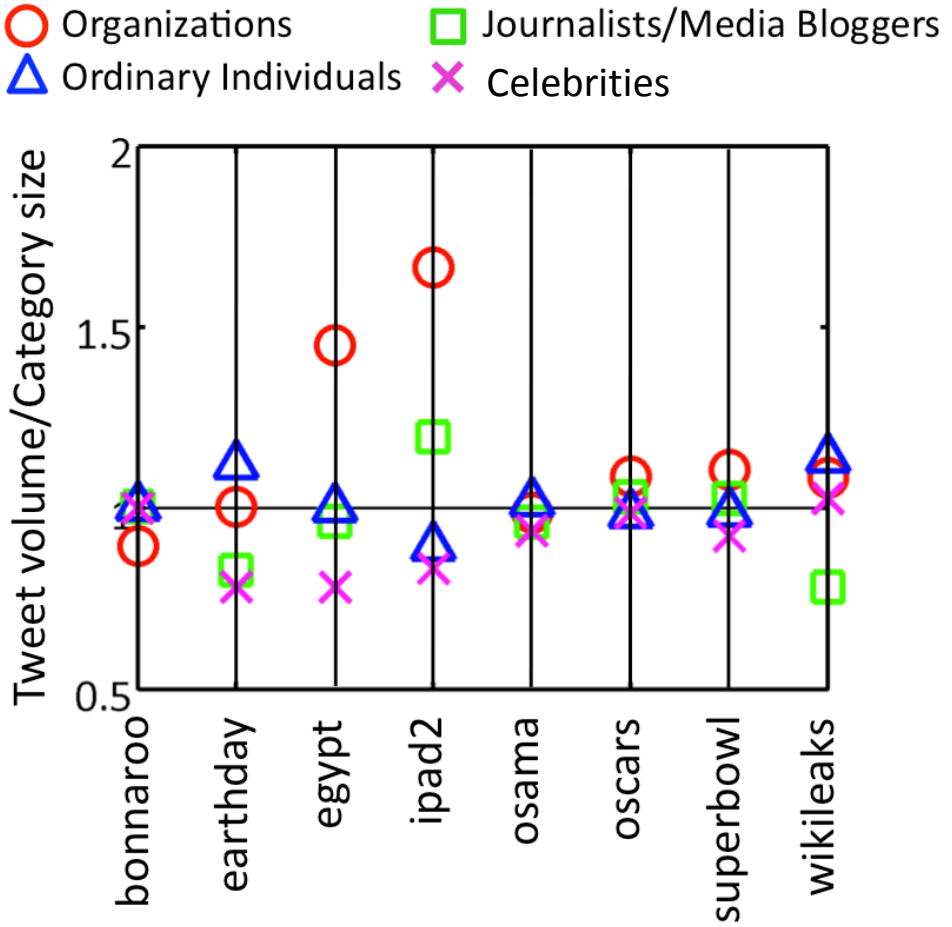
- ~5000 labeled Twitter accounts
  - Twellow, directory of Twitter users according to a list of broad classes
  - Muckrack, journalistic website with self-declared journalists and media bloggers
  - Twitter public timeline users, labeled using Amazon Mechanical Turk
- A number of features used to describe each user: network, behavioral and topical.
- A nearest neighbor classifier is then used to train the model spanning the users

# Examples of labeled users

| | |
|---|---|
| Organizations | @GLOBALHEALTHorg, @irnweb, @revradio, @PForus, @PowerWomenMag |
| Journalists/Media Bloggers | @DcTellAllMedia, @nieuwsmedia, @GrimaldiBiz, @AliceGomstyn, @foresmac |
| Ordinary Individuals | @KiidShyne, @eprinaexa, @katietip- ton95, @daphnebaks, @LaurenneK |
| Celebrities | @princessxtiana, @GugsSays, @willauvucastane, @NewYorkerLeaM |

# Exploring User Categories related to Events



# Conclusions

- Understanding user categories helps us understand the “character” of today’s geographically dispersed events
- Social media is exceedingly useful in understanding these temporal phenomena
- Enables easier journalistic inquiry and thereby eases “virtual transportation” planning


The End
NEXT EXIT

Social media are causing significant changes in our lives

# Questions?

Thanks!

For details: m.dechoudhury@rutgers.edu
Web: http://www.public.asu.edu/~mdechoud/
Twitter: @munmun10