Data Mining with Social and Trajectory Data: Urban Computing in the Big Data Age

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

Major collaborators: Hector Gonzalez, Jae-Gil Lee, Zhenhui Li, Lu An Tang, Jingjing Wang, Zhijun Yin

Acknowledgement of funding: NSF, ARL, NASA, DHS, MURI (AFOSR), Boeing, IBM, and MSR
Outline

- Why Mining Social and Trajectory Data?
- Periodica and MoveMine
- Social-Related Data Mining: GeoTopic Discovery
- From Mining Social Relationships to Cyberphysical Data Analysis
- Looking forward: Future Research Topics
Urban Computing in the Big Data Age

- Urban computing: Integrated computing using urban-related data (location/map, time, people/crowd, social network) to improve the quality of urban life
- Big data age for urban computing
  - Big data: Data has been collected from all the aspects of urban life, from sensors, GPS, cell-phones,
  - What should we do? Turning massive urban data into information and knowledge
- Mining data in four dimensions
  - Location
  - Time
  - Text
  - Links (social and information networks)
- Integrated multi-dimensional mining of social and trajectory data
Booming Age of Social and Trajectory Data

Advanced satellite, sensors, RFID, and wireless technologies:

- Prevalence of mobile devices such as smart phones
- GPS embedded in cars
- Sensors attached on animals

A trajectory: A sequence of timestamps and locations

<table>
<thead>
<tr>
<th>ID</th>
<th>Timestamp</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Peter&quot;</td>
<td>2010-04-02 13:12</td>
<td>37.5, -122.5</td>
</tr>
<tr>
<td>&quot;Peter&quot;</td>
<td>2010-04-02 15:22</td>
<td>37.2, -123.5</td>
</tr>
</tbody>
</table>
Importance on Mining Social and Trajectory Data

Trajectory data mining has many important, real-world applications driven by the real need

Human movement studies

Biological studies

Weather forecasting

GPS wildlife tracking is more common

Science
• help biologists understand the large-scale data

Disease control
• animal-borne communicable diseases such as the H5N1 strain of avian influenza

Environmental Study
• abnormal movements could be signs of environmental changes

Suspicious relationships
Complexity of the Moving Object Data

- **Uncertainty**
  - Sampling rate could be inconstant: From every few seconds transmitting a signal to every few days transmitting one
  - Data can be sparse: A recorded location every 3 days
- **Noise**
  - Erroneous points (e.g., a point in the ocean)
- **Background**
  - Cars follow underlying road network
  - Animals movements relate to mountains, lakes, ...
- **Movement interactions**
  - Affected by nearby moving objects
Research Impacts

- Moving object and trajectory data mining has many important, real-world applications driven by the real need
  - Homeland security (*e.g.*, border monitoring)
  - Law enforcement (*e.g.*, video surveillance)
  - Ecological analysis (*e.g.*, animal scientists)
  - Weather forecast
  - Traffic control
  - Location-based service
  - ...
Outline

- Why Mining Social and Trajectory Data?
- Periodica and MoveMine
- Social-Related Data Mining: GeoTopic Discovery
- From Mining Social Relationships to Cyberphysical Data Analysis
- Conclusions
Periodicity: A Fundamental Regularity

“Human trajectories show a high degree of temporal and spatial regularity.” —Nature’08

“93% potential predictability in human movement based on regularity.” —Science’10

Mining periodic behaviors is useful for:
• Summarize long historical movement
• Predict future movement
• Detect abnormal event

None of previous work ever studied how to automatically detect periods in movements
Challenges in Detecting Periods

Data complexity challenge:
• Raw movement is not periodic
• Movements are affected by many factors
No Periodicity in Raw Trajectory

Raw movement sequence:
\[(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\]

Transform points to complex numbers:
\[x_1 + y_1i, x_2 + y_2i, x_3 + y_3i, \ldots, x_n + y_ni\]

Apply Fourier Transform on complex numbers
\[X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi \frac{k}{N} n}\]

Bee example:
8 hours in hive
16 hours fly nearby

Fail to detect the period of one day
Key Insight: Use a Reference Spot [KDD’10]

Bee example: Observe the in-and-out movements from the hive region.

Reference spot: hive

Easy to see the periodicity

Two-Dimensional Movement

One-Dimensional Binary Sequence

Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, “Mining Periodic Behaviors for Moving Objects”, KDD’10, July 2010
Benefits of Using Reference Spots

- Reveal hidden periods
- Discover multiple periods
- Detect periods at different spatial granularities

Reference spots are **data-dependent and should be automatically detected** instead of given.
Find Reference Spots

- **Reference spot:**
  - Frequently visited spatial regions
  - Higher density than other locations

\[ f(c) = \frac{1}{n \gamma^2} \sum_{i=1}^{n} \frac{1}{2\pi} \exp\left(-\frac{|c - loc_i|^2}{2\gamma^2}\right) \]

(1) Use kernel-based method to calculate the densities

(2) Reference spots: contours of high density places
Detect Periods in Binary Sequence

Fourier transform (Periodogram) gives a set of candidates

\[ X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i kn}{N}} \quad k = 0, \ldots, N - 1 \]

\[ F_k = ||X_k||^2 \]

Autocorrelation further locates the exact periods

\[ R(\tau) = \sum_{i=1}^{n} b_\tau b_{i+\tau} \]
A Bald Eagle Case: **Step 1. Detect Periods**

- Tracking time: Jan, 2006 to Dec., 2008
- Number of locations: 2204

### Density Map

- Reference spots: 1, 2, 3

### Periods detected using Fourier transform and autocorrelation:

<table>
<thead>
<tr>
<th>ref. spot</th>
<th>period</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>363</td>
</tr>
<tr>
<td>2</td>
<td>363</td>
</tr>
<tr>
<td>3</td>
<td>364</td>
</tr>
</tbody>
</table>
A Bald Eagle Case: **Step 2. Summarize Behaviors**

“This bald eagle stays in New York area (i.e., reference spot #1) from December to March. In March, it flies to Great Lakes area (i.e., reference spot #2) and stays there until the end of May. It flies to Quebec area (i.e., reference spot #3) in the summer and stays there until late September. Then it flies back to Great Lake again staying there from mid October to mid November and goes back to New York in December.”
Human Movement:  Step 1. Detect Periods

- Data from Nokia Competition 2012
- Tracking time: 11/30/2009-4/26/2010
- GPS device actively tracks locations every 20 minutes

<table>
<thead>
<tr>
<th>ref. spot</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period (hours)</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
</tbody>
</table>
Human Movement: Step 2. Summarize Behaviors

Spot #1: Office
Spot #2: Commuting city
Spot #3: Home
Spot #4: Vacation place
Periodicity Detection in **Sparse** Data (KDD’12)

- Human movements:
  - Nokia Data: GPS devices actively report locations every 20 minutes
  - Majority of existing data are collected by cell-phones: only report locations when making calls
- Animal movements:
  - Birds data: 2~3 locations in 3~5 days
  - Can we detect daily periodicity?

```
2009-05-02 01:03 in
2009-05-03 11:30 out
2009-05-05 03:12 in
2009-05-09 12:03 in
2009-05-10 11:14 out
2009-05-11 02:15 in
...
```

Sparse in-and-out observations w.r.t a reference spot

Only focus on period detection on binary observations
Finding Periodicity in Sparse Data? It Is a Challenge

Sparse Raw Data

2009-05-02 01:03 in
2009-05-03 11:30 out
2009-05-05 03:12 in
2009-05-09 12:03 in
2009-05-10 11:14 out
2009-05-11 02:15 in
...

Any periodicity in the above sequence?
Mining Periodicity in Sparse Data

- Event has a period of 20
- Occurrences of the event happen between $20k+5$ to $20k+10$

Event has a period of 20. Occurrences of the event happen between $20k+5$ to $20k+10$.

Segment the data using length 20

Segment the data using length 16

Overlay the segments

Overlay the segments

Observations are clustered in [5,10] interval. Observations are scattered.
A Probabilistic Model for Periodic Event

Ratio of positive (in) and negative (out) observations:

\[ \mu^+(I, T) = \frac{|S_I^+|}{|S^+|}, \quad \mu^-(I, T) = \frac{|S_I^-|}{|S^-|} \]

where \( I \) is a set of timestamps, \( T \) is a potential period.
Periodicity Measure

Discrepancy score \[ \Delta_{\chi}(I, T) = \mu^+_{\chi}(I, T) - \mu^-_{\chi}(I, T) \]

- If \( T \) is the correct period, the discrepancy score should be **large** for certain set of timestamps
- If \( T \) is the wrong period, the discrepancy scores are likely to be **zero** for any set of timestamps

Periodicity measure \[ \gamma_{\chi}(T) = \max_{I \in \mathcal{I}_T} \Delta(I, T) \]

**Theorem:** Suppose \( d = \{d_t\}_{t=0}^{n-1} \) are i.i.d. random variables in \([0, 1]\) with nonzero mean, and a sequence \( \chi \) is generated according to any \((p^{T_0}, d)\) for some \( T_0 \), then

\[ \lim_{n \to \infty} \gamma_{\chi}(T) \leq \lim_{n \to \infty} \gamma_{\chi}(T_0), \quad \forall T \in \mathbb{Z}. \]
Performance Comparisons

(Ratio of observed points in the complete sequence)
Experiment on Real Human Data

- Nokia dataset: Pick one person and one of his frequently visited locations

Sampling rate: 20min

Sampling rate: 1hour
Moving Object Clustering

- A moving cluster is a set of objects that move close to each other for a long time interval
  - **Note**: Moving clusters and flock patterns are essentially the same

- Formal Definition [Kalnis et al., SSTD’05]:
  - A moving cluster is a sequence of (snapshot) clusters \( c_1, c_2, ..., c_k \) such that for each timestamp \( i \) (1 \( \leq \) \( i \) \( < \) \( k \)), \( \frac{|c_i \cap c_{i+1}|}{|c_i \cup c_{i+1}|} \geq \vartheta \) (0 \( < \) \( \vartheta \) \( \leq \) 1)
Four Kinds of Relative Motion Patterns (Laube et al. 04, Gudmundsson et al. 07)

- **Flock** (Parameters: $m > 1$ and $r > 0$) At least $m$ entities are within a circular region of radius $r$ and they move in the same direction

- **Leadership** (Parameters: $m > 1$, $r > 0$, and $s > 0$) At least $m$ entities are within a circular region of radius $r$, they move in the same direction, and at least one of the entities was already heading in this direction for at least $s$ time steps

- **Convergence** (Parameters: $m > 1$ and $r > 0$) At least $m$ entities will pass through the same circular region of radius $r$ (assuming they keep their direction)

- **Encounter** (Parameters: $m > 1$ and $r > 0$) At least $m$ entities will be simultaneously inside the same circular region of radius $r$ (assuming they keep their speed and direction)
An example of a *flock* pattern for $p_1$, $p_2$, and $p_3$ at 8th time step; also a *leadership* pattern with $p_2$ as the leader

A *convergence* pattern if $m = 4$ for $p_2$, $p_3$, $p_4$, and $p_5$
A new definition considers *multiple* time steps, whereas the previous definition *only one* time step.

**Flock**: A *flock* in a time interval $I$, where the duration of $I$ is at least $k$, consists of at least $m$ entities such that for every point in time within $I$ there is a disk of radius $r$ that contains all the $m$ entities.

- e.g.)

![A flock through 3 time steps](image-url)
An Extension of Flock Pattern – Convoy
(Jeung et al. ICDE’08 & VLDB’08)

Flock pattern has rigid definition with a circle.
Convoy use density-based clustering at each timestamp.
From Flocks and Convoys to Swarms

- Flock and convoy all require \( k \) consecutive time stamps (still very rigid definition)
- Moving objects may not be close to each other for consecutive time stamps (need to relax time constraint)
Discovery of Swarm Patterns of Moving Objects

- A system that mines moving object patterns: Z. Li, et al., “MoveMine: Mining Moving Object Databases", SIGMOD’10 (system demo)
- Z. Li, B. Ding, J. Han, and R. Kays, “Swarm: Mining Relaxed Temporal Moving Object Clusters”, in submission.

\[\text{Swarm} \text{ discovers more patterns} \rightarrow \]

\[\leftarrow \text{Convoy} \text{ discovers only restricted patterns} \]

Trajectory Data Mining: A “Partitioning + X” Approach

- Task: Mining a large collections of trajectory data

- Philosophy: A long trajectory as a whole should be partitioned in mining

- A partitioning-and-regrouping methodology

- Trajectory clustering

- Trajectory classification

- Trajectory outlier detection
Trajectory Clustering [SIGMOD’07]

- Two phases: *partitioning* and *grouping*

(1) Partition

(2) Group

Note: A representative trajectory is a common sub-trajectory
Sample Clustering Results

570 Hurricanes (1950~2004)
7 Clusters from Hurricane Data

A red line: a representative trajectory

2 Clusters from Deer Data
Trajectory Classification [VLDB’08]

- Parts of trajectories near the container port and near the refinery enable us to distinguish between container ships and tankers even if they share common long paths.
- Those in the fishery enable us to recognize fishing boats even if they have no common path there.
Region-Based Clustering

Trajectory-Based Clustering

Features

Trajectory Partitions

Region-Based Clustering

Trajectory-Based Clustering

Trajectory-Based Clustering

Region-Based Clustering

(1) A B C

(2) A B C

(3) 1 2 3 4 5 6

(4) 7 8 9 10
Extracted Features

Features:
- 10 Region-Based Clusters
- 37 Trajectory-Based Clusters

Data (Three Classes)

Accuracy = 83.3%
Partition-and-Detect Framework (Lee et al. ICDE’08)

- Existing algorithms compare trajectories as a whole. They might not be able to detect outlying portions of trajectories.
  - e.g., TR$_3$ is not detected as an outlier since its overall behavior is similar to those of neighboring trajectories.

- The partition-and-detect framework is proposed to detect outlying sub-trajectories.

[Diagram showing multiple trajectories with TR$_3$ highlighted as an outlying sub-trajectory]
Overall Procedure

- Two phases: *partitioning* and *detection*

```
\begin{itemize}
  \item \textit{TR}_5, \textit{TR}_4, \textit{TR}_3, \textit{TR}_2, \textit{TR}_1
  \item A set of trajectories
  \item A set of trajectory partitions
  \item \textit{TR}_3
  \item An outlier
  \item Outlying trajectory partitions
\end{itemize}
```
Trajectory Outlier Detection: Example

3 Outliers from Elk Data
MoveMine: Mining Moving Object Databases

A system that mines moving object patterns: Z. Li, et al., “MoveMine: Mining Moving Object Databases”, SIGMOD’10 (system demo)
Effectiveness Testing on Real Data

Number of individuals: 165

Time span:
11/14/2000 6:49:00 AM
to
11/21/2006 6:15:00 AM

Raw buffalo data
165 buffalo from
Years 2000 to 2006
Outline

- Why Mining Social and Trajectory Data?
- Periodica and MoveMine
- Social-Related Data Mining: GeoTopic Discovery
- From Mining Social Relationships to Cyberphysical Data Analysis
- Conclusions
Spatial Data are Popular on the Web

- Advanced cameras with GPS receivers
- Applications including Google Earth, Flickr, etc.
- GPS functions in smart phones
Spatial Text Documents

- **Flickr**: geo-tagged photos

  *Sunset over the Brooklyn Bridge and Lower Manhattan, New York City*

  *Sunset from the Brooklyn side of the Manhattan Bridge (there’s the view at night). Clouds to the left blocked any sunbeam below, but there was a brief moment where the sky really lit up as the sun peaked through the cloud cover and sat right between two towers (I think it’s the short/long 180°). I transitioned exposures to capture more of the dynamic range of the scene. Exposure Standard with Photomatix. I used LiveView to focus, which often makes things a bit clearer as you need to focus on the image.*

  *Shooted exposures (EV 1.7): 1/180° - 24mm (35mm eq).*
GeoTopic: Mining Spatial Text Documents

- **Applications**
  - Analyze the cultural differences around the world
  - Explore the hot topics or events in different places
  - Compare the popularity of specific products in different regions

- **Discover** different topics of interests those are coherent in geographical regions

- **Compare** several topics across different geographical locations

- **GeoTopic:** Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang, “Geographical Topic Discovery and Comparison”, Proc. of 2011 Int. World Wide Web Conf. (WWW'11), Hyderabad, India, Mar. 2011.
Input:
Text with spatial information

Output:
1. Geographic topics: \{ p(w|z) \}  
2. Topic distribution \( p(z|l) \)

\[
p(w|z_1) \quad p(w|z_2) \quad p(w|z_3)
\]

<table>
<thead>
<tr>
<th>Location</th>
<th>= (40.70, 73.91)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Topic 1 (Chinese food)</th>
<th>Topic 2 (Japanese food)</th>
<th>Topic 3 (Mexican food)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>noodles 0.067</td>
<td>ramen 0.104</td>
<td>tacos 0.069</td>
<td>...</td>
</tr>
<tr>
<td>dimsum 0.064</td>
<td>soba 0.066</td>
<td>taco 0.059</td>
<td>...</td>
</tr>
<tr>
<td>hotpot 0.039</td>
<td>noodle 0.065</td>
<td>salsa 0.036</td>
<td>...</td>
</tr>
<tr>
<td>rice 0.038</td>
<td>sashimi 0.039</td>
<td>cajun 0.031</td>
<td>...</td>
</tr>
<tr>
<td>noodle 0.035</td>
<td>yakitori 0.030</td>
<td>burrito 0.027</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

| topic z | \( p(z|l) \) |
|---------|-------------|
| Topic 1 (Chinese food) | 22%         |
| Topic 2 (Japanese food) | 14%         |
| Topic 3 (Mexican food)  | 18%         |
| ...               | ...         |
Potential Solutions

- LDM: Location-driven model
  - Clustering based on document locations
  - One location cluster is a topic

- TDM: Text-driven model [Mei et al. WWW’08]
  - Topic modeling with network regularization
  - Documents that are close in space should have similar topic distributions

- GeoFolk [Sizov WSDM’10]
  - A topic modeling that uses both text and spatial information
  - The geographical distribution of each topic is Gaussian
**An Illustrating Dataset**

Geo-tagged photos related to landscape (coast vs. desert vs. mountain)

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>beach</td>
<td>desert</td>
<td>mountains</td>
</tr>
<tr>
<td>ocean</td>
<td>california</td>
<td>mountain</td>
</tr>
<tr>
<td>water</td>
<td>mountains</td>
<td>lake</td>
</tr>
<tr>
<td>california</td>
<td>mountain</td>
<td>trees</td>
</tr>
<tr>
<td>sea</td>
<td>arizona</td>
<td>water</td>
</tr>
<tr>
<td>coast</td>
<td>utah</td>
<td>snow</td>
</tr>
<tr>
<td>sunset</td>
<td>rock</td>
<td>scenery</td>
</tr>
<tr>
<td>seascape</td>
<td>southwest</td>
<td>hiking</td>
</tr>
<tr>
<td>pacific</td>
<td>park</td>
<td>washington</td>
</tr>
<tr>
<td>sand</td>
<td>sunset</td>
<td>reflection</td>
</tr>
</tbody>
</table>
The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance.

\[
\text{perplexity}_{\text{test}}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d \in D_{\text{test}}} \log p(w_d)}{\sum_{d \in D_{\text{test}}} N_d}\right\}
\]
Based on the perplexity of both the location and text, the advantage of our model is much larger.
### Geographical Topic Comparison

<table>
<thead>
<tr>
<th>Chinese Food</th>
<th>Japanese Food</th>
<th>Italian Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>chinese 0.552</td>
<td>japanese 0.519</td>
<td>italian 0.848</td>
</tr>
<tr>
<td>noodles 0.067</td>
<td>ramen 0.104</td>
<td>cappuccino 0.067</td>
</tr>
<tr>
<td>dimsum 0.064</td>
<td>soba 0.066</td>
<td>latte 0.048</td>
</tr>
<tr>
<td>hotpot 0.039</td>
<td>noodle 0.065</td>
<td>gelato 0.030</td>
</tr>
<tr>
<td>rice 0.038</td>
<td>sashimi 0.039</td>
<td>pizza 0.002</td>
</tr>
<tr>
<td>noodle 0.035</td>
<td>yakitori 0.030</td>
<td>pizzeria 0.002</td>
</tr>
<tr>
<td>tofu 0.020</td>
<td>okonomiyaki 0.026</td>
<td>mozzarella 0.001</td>
</tr>
<tr>
<td>dumpling 0.018</td>
<td>udon 0.026</td>
<td>pasta 0.001</td>
</tr>
<tr>
<td>duck 0.018</td>
<td>tempura 0.020</td>
<td>ravioli 0.000</td>
</tr>
<tr>
<td>prawn 0.017</td>
<td>curry 0.016</td>
<td>pesto 0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>French Food</th>
<th>Spanish Food</th>
<th>Mexican Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>french 0.564</td>
<td>spanish 0.488</td>
<td>mexican 0.484</td>
</tr>
<tr>
<td>bistro 0.070</td>
<td>tapas 0.269</td>
<td>tacos 0.069</td>
</tr>
<tr>
<td>patisserie 0.056</td>
<td>paella 0.076</td>
<td>taco 0.059</td>
</tr>
<tr>
<td>bakery 0.049</td>
<td>pescado 0.059</td>
<td>salsa 0.036</td>
</tr>
<tr>
<td>resto 0.044</td>
<td>olives 0.032</td>
<td>cajun 0.031</td>
</tr>
<tr>
<td>pastry 0.033</td>
<td>stickyrice 0.017</td>
<td>burrito 0.027</td>
</tr>
<tr>
<td>tarte 0.026</td>
<td>tortilla 0.013</td>
<td>crawfish 0.023</td>
</tr>
<tr>
<td>croissant 0.021</td>
<td>mediterranean 0.010</td>
<td>guacamole 0.022</td>
</tr>
<tr>
<td>baguette 0.019</td>
<td>mussels 0.008</td>
<td>margarita 0.020</td>
</tr>
<tr>
<td>mediterranean 0.018</td>
<td>octopus 0.008</td>
<td>cocktails 0.020</td>
</tr>
</tbody>
</table>
The larger $p(\text{topic|location})$ is, the darker the location is.
Latent Periodic Topic Analysis [ICDM’11]

- Periodic Phenomena Exist Ubiquitously
  - Hurricanes
  - Music and film festivals
  - Product sales
  - TV program
  - Publicly traded company

- Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang, "LPTA: A Probabilistic Model for Latent Periodic Topic Analysis", Proc. 2011 IEEE Int. Conf. on Data Mining (ICDM'11), Dec. 2011.
Text Data with Temporal Information

News articles associated with their publishing dates

Obama sets campaign theme: Middle class at stake

OSAWATOMIE, Kan. (AP) — Declaring the American middle class in jeopardy, President Barack Obama on Tuesday outlined a populist economic vision that will drive his re-election bid, insisting the United States must reclaim its standing as a country in which everyone can prosper if provided “a fair shot and a fair share.”

While never making an overt plea for a second term, Obama’s offered his most comprehensive lines of attack against the candidates seeking to take his job, only a month before Republican voters begin choosing a presidential nominee. He also sought to inject some of the long-overshadowed hope that energized his 2008 campaign, saying: “I believe America is on its way up.”

Tweets published with their upload times in Twitter

Tagged photos annotated with dates in Flickr
Apply Periodicity Analysis on Text Data

- Periodicity detection for time series database [Elfeky et al. TKDE 2005]
- Some studies follow the similar strategies to analyze the time distribution of a single tag or query to detect periodic patterns [Vlachos et al. SIGMOD 2004]
Challenges

- A single word is not enough to describe a topic and more words are needed to summarize a topic comprehensively.

- Analyzing the periodicity of single terms only is not sufficient to discover periodic topics.
  - E.g., “music”, “festival” and “chicago” may not have periodic patterns if considered separately, but there may be periodic topics if they are considered together.

- Synonyms and polysemy words due to the language diversity.
Latent Periodic Topic Analysis (LPTA)

Input: Time-stamped documents

<table>
<thead>
<tr>
<th>ID</th>
<th>Text</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>coachella, music, arts, festival, ...</td>
<td>Apr 27 2008</td>
</tr>
<tr>
<td>2</td>
<td>sxsw, south by southwest, austin, ...</td>
<td>Mar 14 2008</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Output:
1. Periodic topics: \( \{ p(w|z) \} \)
2. Time distribution of topics

Topic 1 (Coachella Festival)

- coachella 0.1106
- music 0.0915
- indio 0.0719
- california 0.0594
- concert 0.0357
- ...

Periodic interval \( T \), e.g., 1 year, etc.

The distribution of the timestamps for the topic related to Coachella festival
Latent Periodic Topic Analysis: Problem Formulation

- **Input:**
  - A collection of time-stamped documents $D$
  - The number of topics $K$
  - Periodic interval $T$

- **Output:**
  - $K$ periodic topics
    - $\theta = \{\theta_z\}_{z \in \mathbb{Z}}$
      - $\theta_z = \{p(w \mid z)\}_{w \in V}$
    - $p(w \mid z)$ is the probability of word $w$ given topic $z$
  - The distribution of the timestamps for each topic
General Idea of LPTA

- Related work
  - Periodicity Analysis in time-series DB [Elfeky et al., 2005]
  - Topic models: PLSA [Hofmann SIGIR 1999] and LDA [Blei et al. JMLR 2003]
  - Topic Over Time [Wang et al. KDD 2006], etc.
- LPTA (Latent Periodic Topic Analysis): General Ideas
  - Term co-occurrence
    - If two words co-occur often in the same documents, they are more likely to belong to the same topic
  - Temporal structure
    - We assume that there are many consecutive periods across the time line. The words occurring around the same time in each period are likely to be clustered
Temporal Patterns of Topics

- Periodic topics
  - A periodic topic is one repeating in regular intervals
- Background topics
  - A background topic is one covered uniformly over the entire period
- Bursty topics
  - A bursty topic is a transient topic that is intensively covered only in a certain time period
Temporal Patterns of Topics (Cont.)

- Periodic topics
  - The distribution of timestamps for each periodic topic as a mixture of Gaussian distributions where the interval between the consecutive components is $T$

- Background topics
  - The timestamps of the background topics are generated by a uniform distribution

- Bursty topics
  - The timestamps of the bursty topics are generated from a Gaussian distribution

- The document collection is modeled as a mixture of background topics, bursty topics and periodic topics
Generative Process of LPTA

- For each word in document \( d \) from collection \( D \):
  - (1) Sample a topic \( z \) from multinomial \( \phi_d \), i.e., \( \{p(z \mid d)\}_{z \in Z} \)
    - (a) If \( z \) is a background topic, sample time \( t \) from a uniform distribution \([t_{\text{start}}, t_{\text{end}}]\), where \( t_{\text{start}} \) and \( t_{\text{end}} \) are the start time and end time of the document collection
    - (b) If \( z \) is a bursty topic, sample time \( t \) from \( N(\mu_z, \sigma_z^2) \)
    - (c) If \( z \) is a periodic topic, sample period \( k \) of document \( d \) from a uniform distribution. Sample time \( t \) from \( N(\mu_z + kT, \sigma_z^2) \) where \( T \) is periodic interval
  - (2) Sample a word \( w \) from multinomial \( \theta_z \), i.e., \( \{p(w \mid z)\}_{w \in V} \)
Log-likelihood of Document Collection

Given the data collection \( \{(w_d, t_d)\}_{d \in D} \) where \( w_d \) is the word set in document \( d \) and \( t_d \) is the timestamp of document \( d \), the log-likelihood of the collection given \( \psi = \{\theta, \phi, \mu, \sigma\} \) is as follows

\[
L(\psi; D) = \log p(D | \psi) = \log \prod_{d \in D} p(w_d, t_d | \psi)
\]

\[
\log p(w_d, t_d | \psi) = \sum_d \sum_w n(d, w) \log \sum_z p(t_d | z) p(w | z) p(z | d)
\]

- If topic \( z \) is a background topic,
  \[
p(t | z) = \frac{1}{t_{\text{end}} - t_{\text{start}}}
\]

- If topic \( z \) is a bursty topic,
  \[
p(t | z) = \frac{1}{\sqrt{2\pi \sigma_z^2}} e^{-\frac{(t - \mu_z)^2}{\sigma_z^2}}
\]

- If topic \( z \) is a periodic topic,
  \[
p(t | z) = p(k) \frac{1}{\sqrt{2\pi \sigma_z^2}} e^{-\frac{(t - \mu_z - kT)^2}{\sigma_z^2}}
\]
Parameter Estimation

- EM (Expectation Maximization) algorithm
  - E-step
    \[ p(z|d,w) = \frac{p(t_d|z)p(w|z)p(z|d)}{\sum_z p(t_d|z')p(w|z')p(z'|d)} \]
  - M-step
    \[ p(w|z) = \frac{\sum_d n(d,w)p(z|d,w)}{\sum_d \sum_{w'} n(d,w')p(z|d,w')} \]
    \[ p(z|d) = \frac{\sum_w n(d,w)p(z|d,w)}{\sum_w \sum_{z'} n(d,w)p(z'|d,w)} \]

For bursty topic \( z \)

\[ \mu_z = \frac{\sum_d \sum_w n(d,w)p(z|d,w)t_d}{\sum_d \sum_w n(d,w)p(z|d,w)} \]
\[ \sigma_z = \left( \frac{\sum_d \sum_w n(d,w)p(z|d,w)(t_d - \mu_z)^2}{\sum_d \sum_w n(d,w)p(z|d,w)} \right)^{1/2} \]

For periodic topic \( z \)

\[ \mu_z = \frac{\sum_d \sum_w n(d,w)p(z|d,w)(t_d - I_d T)}{\sum_d \sum_w n(d,w)p(z|d,w)} \]
\[ \sigma_z = \left( \frac{\sum_d \sum_w n(d,w)p(z|d,w)(t_d - \mu_z - I_d T)^2}{\sum_d \sum_w n(d,w)p(z|d,w)} \right)^{1/2} \]

Complexity: \( O(\text{iter} K|W|) \) where \( \text{iter} \) is the number of the iterations in EM, \( K \) is the number of topics, \( |W| \) is the total count of the words in all the documents.
Experimental Datasets

- **Seminar**
  - The weekly seminar announcements for one semester from six research groups in computer science department at University of Illinois at Urbana-Champaign
  - 61 documents and 901 unique words
  - Set periodic interval T as 1 week

- **DBLP (Computer Science Digital Bibliography)**
  - The paper titles of several different conferences from 2003 to 2007. The conferences include WWW, SIGMOD, SIGIR, KDD, VLDB and NIPS
  - The timestamps of the documents are determined w.r.t. the conference programs
  - 4070 documents and 2132 unique words
  - Set periodic interval T as 1 year

- **Flickr**
  - The photos for several music festivals from 2006 to 2010 including SXSW (South by Southwest), Coachella, Bonnaroo, Lollapalooza and ACL (Austin City Limits)
  - The tags of a photo are considered as document text, while the time when the photo was taken is considered as document timestamp
  - 84244 documents and 7524 unique words. Set periodic interval T as 1 year
Topics Discovered by LPTA

Selected periodic topics discovered by LPTA. The date and the duration in the parentheses are the mean and standard deviation of the timestamps for the corresponding periodic topic.

<table>
<thead>
<tr>
<th>Seminar</th>
<th>DBLP</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 0.0166</td>
<td>Computing 0.0168</td>
<td>Acl 0.0945</td>
</tr>
<tr>
<td>Based 0.0158</td>
<td>Learning 0.0158</td>
<td>Austin 0.0827</td>
</tr>
<tr>
<td>Mining 0.0151</td>
<td>Machine 0.0138</td>
<td>Music 0.0763</td>
</tr>
<tr>
<td>Text 0.0143</td>
<td>Science 0.0128</td>
<td>Manchester 0.0587</td>
</tr>
<tr>
<td>Network 0.0135</td>
<td>Algorithms 0.0128</td>
<td>Austin City 0.0441</td>
</tr>
<tr>
<td>Web 0.0119</td>
<td>Language 0.0118</td>
<td>Tennessee 0.0518</td>
</tr>
<tr>
<td>Problem 0.0111</td>
<td>Work 0.0108</td>
<td>Limits 0.0441</td>
</tr>
<tr>
<td>Data 0.0111</td>
<td>Problems 0.0108</td>
<td>City 0.0441</td>
</tr>
<tr>
<td>Query 0.0111</td>
<td>Models 0.0108</td>
<td>Concert 0.0275</td>
</tr>
<tr>
<td>Latent 0.0095</td>
<td>Prediction 0.0108</td>
<td>Texas 0.0426</td>
</tr>
<tr>
<td></td>
<td>Analysis 0.0105</td>
<td>Concert 0.0283</td>
</tr>
<tr>
<td></td>
<td>Large 0.0104</td>
<td>Performance 0.0174</td>
</tr>
<tr>
<td></td>
<td>Evaluation 0.0111</td>
<td>Zilker 0.0173</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rock 0.0111</td>
</tr>
</tbody>
</table>
LPTA vs. Periodicity Detection

- AUTOPERIOD [Vlachos et al. SDM 2005], a two-tier approach by considering the information in both the autocorrelation and the periodogram, fails to detect meaningful periodic words because the time series are sparse and few words have apparent periodic patterns.

- Compared with single word representation, LPTA uses multiple words to describe a topic
  - In DBLP, topic “VLDB”: data 0.0530, xml 0.0208, query 0.0196, queries 0.0176, efficient 0.0151, mining 0.0142, database 0.0136, streams 0.0112, databases 0.0111
LPTA vs. Periodicity Detection (Cont.)

Time distribution of topic VLDB discovered by LPTA and time distributions of the words in the topic.

- Word "xml"
# LPTA vs. Topic Models

Selected topics discovered for different datasets by using PLSA and LDA

<table>
<thead>
<tr>
<th>Seminar</th>
<th>PLSA</th>
<th>LDA</th>
<th>DBLP</th>
<th>PLSA</th>
<th>LDA</th>
<th>Flickr</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td>Topic 2</td>
<td>Topic 1</td>
<td>Topic 2</td>
<td>Topic 1</td>
<td>Topic 2</td>
<td>Topic 1</td>
<td>Topic 2</td>
</tr>
<tr>
<td>data</td>
<td>memory</td>
<td>problem</td>
<td>systems</td>
<td>web</td>
<td>search</td>
<td>web</td>
<td>system</td>
</tr>
<tr>
<td>latent</td>
<td>computer</td>
<td>algorithm</td>
<td>computer</td>
<td>data</td>
<td>text</td>
<td>mining</td>
<td>database</td>
</tr>
<tr>
<td>visualizati.</td>
<td>data</td>
<td>network</td>
<td>science</td>
<td>xml</td>
<td>databases</td>
<td>semantic</td>
<td>distributed</td>
</tr>
<tr>
<td>intel</td>
<td>mining</td>
<td>graph</td>
<td>algorithms</td>
<td>queries</td>
<td>relational</td>
<td>detection</td>
<td>user</td>
</tr>
<tr>
<td>talk</td>
<td>parallel</td>
<td>time</td>
<td>time</td>
<td>mining</td>
<td>user</td>
<td>automatic</td>
<td>adaptive</td>
</tr>
<tr>
<td>analysis</td>
<td>science</td>
<td>networks</td>
<td>agent</td>
<td>semantic</td>
<td>analysis</td>
<td>services</td>
<td>content</td>
</tr>
<tr>
<td>computer</td>
<td>pattern</td>
<td>influence</td>
<td>visualizati.</td>
<td>search</td>
<td>ranking</td>
<td>applicatic.</td>
<td>relevance</td>
</tr>
<tr>
<td>systems</td>
<td>programm.</td>
<td>online</td>
<td>data</td>
<td>streams</td>
<td>structure</td>
<td>graph</td>
<td>performan.</td>
</tr>
<tr>
<td>machine</td>
<td>hardware</td>
<td>work</td>
<td>engineering</td>
<td>managem.</td>
<td>support</td>
<td>extraction</td>
<td>feedback</td>
</tr>
<tr>
<td>visual</td>
<td>algorithms</td>
<td>question</td>
<td>function</td>
<td>adaptive</td>
<td>evaluation</td>
<td>patterns</td>
<td>image</td>
</tr>
</tbody>
</table>
Integration of Text and Time

Periodic topics for SIGMOD vs. VLDB and SIGMOD vs. CVPR datasets by using LPTA. The date and the duration are the mean and standard deviation of the timestamps.

<table>
<thead>
<tr>
<th>SIGMOD vs. VLDB</th>
<th>SIGMOD vs. CVPR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic 1</strong></td>
<td><strong>Topic 1</strong></td>
</tr>
<tr>
<td>SIGMOD (SIGMOD)</td>
<td>SIGMOD (SIGMOD)</td>
</tr>
<tr>
<td>Jun 17 (7d11h6m)</td>
<td>Jun 20 (7d15h42m)</td>
</tr>
<tr>
<td><strong>Topic 2</strong></td>
<td><strong>Topic 2</strong></td>
</tr>
<tr>
<td>VLDB (VLDB)</td>
<td>CVPR (CVPR)</td>
</tr>
<tr>
<td>Sep 11 (9d5h29m)</td>
<td>Jun 21 (3d4h37m)</td>
</tr>
</tbody>
</table>

- data, data
- query, xml
- xml, query
- database, queries
- processing, efficient
- efficient, database
- databases, based
- queries, databases
- web, system
- system, processing

SIGMOD and VLDB are two reputed conferences in database area, and it is difficult to differentiate these two conferences based on text only.

SIGMOD and CVPR are held in June, so it is difficult to differentiate these two if we rely on time information only.
Periodic vs. Bursty Topics

Instead of pooling the photos related to music festivals all together, we keep the photos related to SXSW and ACL festivals from 2006 to 2010 and those related to Coachella and Lollapalooza in 2009 only.

<table>
<thead>
<tr>
<th>Bursty topics</th>
<th>Periodic topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1 (Lollapalooza) Aug 8 2009 (1d0h12m)</td>
<td>Topic 2 (Coachella) Apr 17 2009 (10d20h23m)</td>
</tr>
<tr>
<td>lollapalooza</td>
<td>coachella</td>
</tr>
<tr>
<td>chicago</td>
<td>indio</td>
</tr>
<tr>
<td>concert</td>
<td>music</td>
</tr>
<tr>
<td>music</td>
<td>california</td>
</tr>
<tr>
<td>grantpark</td>
<td>concert</td>
</tr>
<tr>
<td>august</td>
<td>live</td>
</tr>
<tr>
<td>live</td>
<td>desert</td>
</tr>
<tr>
<td>illinois</td>
<td>art</td>
</tr>
<tr>
<td>performance</td>
<td>musicfestival</td>
</tr>
<tr>
<td>lolla</td>
<td>livemusic</td>
</tr>
<tr>
<td></td>
<td>sxsw</td>
</tr>
<tr>
<td></td>
<td>austin</td>
</tr>
<tr>
<td></td>
<td>texas</td>
</tr>
<tr>
<td></td>
<td>music</td>
</tr>
<tr>
<td></td>
<td>austincityli.</td>
</tr>
<tr>
<td></td>
<td>southbysouth.</td>
</tr>
<tr>
<td></td>
<td>city</td>
</tr>
<tr>
<td></td>
<td>live</td>
</tr>
<tr>
<td></td>
<td>limits</td>
</tr>
<tr>
<td></td>
<td>concert</td>
</tr>
<tr>
<td></td>
<td>texas</td>
</tr>
<tr>
<td></td>
<td>concert</td>
</tr>
<tr>
<td></td>
<td>gig</td>
</tr>
<tr>
<td></td>
<td>zilker</td>
</tr>
</tbody>
</table>

The words will fit into the corresponding periodic or bursty topics if they have periodic or bursty patterns.
Quantitative Evaluation

- The latent topics discovered by the topic modeling approaches can be regarded as clusters.
- Accuracy and normalized mutual information (NMI) can be used to measure the clustering performance.

<table>
<thead>
<tr>
<th>K</th>
<th>Accuracy(%)</th>
<th>NMI(%)</th>
<th>Accuracy(%)</th>
<th>NMI(%)</th>
<th>Accuracy(%)</th>
<th>NMI(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLSA</td>
<td>LDA</td>
<td>LPTA</td>
<td>PLSA</td>
<td>LDA</td>
<td>LPTA</td>
</tr>
<tr>
<td>2</td>
<td>31.1</td>
<td>31.8</td>
<td>37.7</td>
<td>11.7</td>
<td>12.3</td>
<td>34.7</td>
</tr>
<tr>
<td>3</td>
<td>37.0</td>
<td>38.0</td>
<td>51.0</td>
<td>19.0</td>
<td>19.9</td>
<td>53.0</td>
</tr>
<tr>
<td>4</td>
<td>39.4</td>
<td>41.3</td>
<td>65.4</td>
<td>23.6</td>
<td>24.0</td>
<td>70.7</td>
</tr>
<tr>
<td>5</td>
<td>40.1</td>
<td>42.1</td>
<td>78.5</td>
<td>25.7</td>
<td>26.6</td>
<td>82.4</td>
</tr>
<tr>
<td>6</td>
<td>43.0</td>
<td>41.9</td>
<td>90.4</td>
<td>30.6</td>
<td>28.9</td>
<td>92.3</td>
</tr>
<tr>
<td>7</td>
<td>40.8</td>
<td>39.5</td>
<td>94.5</td>
<td>30.5</td>
<td>29.7</td>
<td>94.2</td>
</tr>
<tr>
<td>8</td>
<td>39.0</td>
<td>40.0</td>
<td>91.9</td>
<td>30.4</td>
<td>31.0</td>
<td>91.7</td>
</tr>
<tr>
<td>9</td>
<td>35.3</td>
<td>36.9</td>
<td>90.0</td>
<td>30.5</td>
<td>30.8</td>
<td>88.8</td>
</tr>
<tr>
<td>10</td>
<td>34.9</td>
<td>33.9</td>
<td>88.1</td>
<td>31.7</td>
<td>30.2</td>
<td>86.8</td>
</tr>
<tr>
<td>Avg</td>
<td>37.9</td>
<td>38.4</td>
<td>76.4</td>
<td>26.0</td>
<td>26.0</td>
<td>77.2</td>
</tr>
</tbody>
</table>

Conclusion: The LPTA model discovers the latent periodic topics by combining the information from topical clusters and periodic patterns.
Spatial-Temporal Data in Social Media

Share your photos. Watch the world.

4.2 million things geotagged this month
Motivation: Diversified Trajectory Pattern Ranking

- Explore the common wisdom from spatial-temporal data in geo-tagged social media
- Discover trajectory patterns interesting to two kinds of users
  - Some users are interested in the most important trajectory patterns ➔ Ranking
  - Some users are interested in exploring a new place in a diverse way ➔ Diversification
Diversified Trajectory Pattern Ranking

Input: A collection of geo-tagged photos (user, date time, GPS location)

(1) Extract trajectory patterns
(2) Rank trajectory patterns
(3) Diversify ranked patterns

Output: Diversified trajectory pattern ranking result
Data Preprocessing

(1) Cluster locations: mean-shift algorithm (27974 photos in London)

londoneye, trafalgarsquare, britishmuseum, bigben, towerbridge, piccaillycircus, buckinghampalace, tatemodern, …

(2) Form sequences

<table>
<thead>
<tr>
<th>ID</th>
<th>User</th>
<th>Date</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice</td>
<td>04/26/11</td>
<td>londoneye -&gt; bigben -&gt; downingstreet -&gt; trafalgarsquare</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>04/27/11</td>
<td>londoneye -&gt; tatemodern -&gt; towerbridge</td>
</tr>
<tr>
<td>3</td>
<td>Bob</td>
<td>04/26/11</td>
<td>londoneye -&gt; bigben -&gt; tatemodern</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Extract Trajectory Patterns

- PrefixSpan [Pei et al. TKDE 2004]
  - Example (minimum support = 2)

<table>
<thead>
<tr>
<th>ID</th>
<th>Travel sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>londoneye → bigben → trafalgarsquare</td>
</tr>
<tr>
<td>2</td>
<td>londoneye → bigben → downingstreet → trafalgarsquare</td>
</tr>
<tr>
<td>3</td>
<td>londoneye → bigben → westminster</td>
</tr>
<tr>
<td>4</td>
<td>londoneye → tatemodern → towerbridge</td>
</tr>
<tr>
<td>5</td>
<td>londoneye → bigben → tatemodern</td>
</tr>
</tbody>
</table>

- Three frequent sequential patterns:
  - londoneye -> bigben
  - londoneye -> bigben -> trafalgarsquare
  - londoneye -> tatemodern
### Top Frequent Trajectories in London

<table>
<thead>
<tr>
<th>Trajectory pattern</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>londoneye → bigben</td>
<td>21</td>
</tr>
<tr>
<td>bigben → londoneye</td>
<td>19</td>
</tr>
<tr>
<td>londoneye → tatemodern</td>
<td>18</td>
</tr>
<tr>
<td>londoneye → royalfestivalhall</td>
<td>15</td>
</tr>
<tr>
<td>londoneye → trafalgarsquare</td>
<td>14</td>
</tr>
<tr>
<td>londoneye → waterloobridge</td>
<td>12</td>
</tr>
<tr>
<td>towerbridge → cityhall</td>
<td>12</td>
</tr>
<tr>
<td>royalfestivalhall → londoneye</td>
<td>11</td>
</tr>
<tr>
<td>tatemodern → londoneye</td>
<td>11</td>
</tr>
<tr>
<td>bigben → parliamentsquare</td>
<td>10</td>
</tr>
</tbody>
</table>

The top frequent trajectory patterns are short and not informative.
A trajectory pattern is important if many important users take it and it contains important locations.

A user is important if the user takes photos at important locations and visits the important trajectory patterns.

An location is important if it occurs in one or more important trajectory patterns and many important users take photos at the location.

\[ P_T = M_{TU} \cdot P_U \]
\[ P_T = M_{LT}^T \cdot P_L \]
\[ P_U = M_{UL} \cdot P_L \]
\[ P_U = M_{TU}^T \cdot P_T \]
\[ P_L = M_{LT} \cdot P_T \]
\[ P_L = M_{UL}^T \cdot P_U \]
Ranking Algorithm

**Algorithm** Trajectory pattern ranking

**Input:** $M_{TU}$, $M_{UL}$, $M_{LT}$

**Output:** A ranked list of trajectory patterns

1. Initialize $P_T^{(0)}$

2. Iterate

   \[
   P_L = M_{LT} \cdot P_T^{(t)} \quad P_U = M_{UL} \cdot P_L \\
   P_T = M_{TU} \cdot P_U \quad P_U = M_{TU}^T \cdot P_T \\
   P_L = M_{UL}^T \cdot P_U \quad P_T^{(t+1)} = M_{LT}^T \cdot P_L \\
   P_T^{(t+1)} = P_T^{(t+1)} / \|P_T^{(t+1)}\|_1
   \]

   until convergence.

3. Output the ranked list of trajectory patterns in the decreasing order of $P_T^*$, i.e., the converged $P_T$.

$P_T$ is the eigen vector for $M^T M$ for the largest eigen value, $M = M_{TU} M_{UL} M_{LT}$. The algorithm is a normalized power iteration method to detect the eigen vector of $M^T M$ for the largest eigen value if the initial $P_T$ is not orthogonal to it.
Top Trajectory in London

londoneye -> bigben -> downingstreet -> horseguards -> trafalgarsquare
### Top Ranked Locations in London

<table>
<thead>
<tr>
<th>Location</th>
<th>$P_L$</th>
<th># User</th>
<th>Location</th>
<th>$P_L$</th>
<th># User</th>
</tr>
</thead>
<tbody>
<tr>
<td>londoneye</td>
<td>0.0157</td>
<td>528</td>
<td>southwark</td>
<td>0.0062</td>
<td>57</td>
</tr>
<tr>
<td>trafalgarsquare</td>
<td>0.0125</td>
<td>456</td>
<td>stpaulscathedral</td>
<td>0.0058</td>
<td>77</td>
</tr>
<tr>
<td>bigben</td>
<td>0.0121</td>
<td>205</td>
<td>downingstreet</td>
<td>0.0053</td>
<td>52</td>
</tr>
<tr>
<td>tatemodern</td>
<td>0.0119</td>
<td>491</td>
<td>horseguards</td>
<td>0.0051</td>
<td>25</td>
</tr>
<tr>
<td>royalfestivalhall</td>
<td>0.0093</td>
<td>175</td>
<td>londonbridge</td>
<td>0.0049</td>
<td>37</td>
</tr>
<tr>
<td>towerbridge</td>
<td>0.0089</td>
<td>185</td>
<td>embankment</td>
<td>0.0047</td>
<td>23</td>
</tr>
<tr>
<td>cityhall</td>
<td>0.0077</td>
<td>141</td>
<td>harrods</td>
<td>0.0047</td>
<td>39</td>
</tr>
<tr>
<td>waterloobridge</td>
<td>0.0076</td>
<td>198</td>
<td>toweroflondon</td>
<td>0.0046</td>
<td>91</td>
</tr>
<tr>
<td>parliamentsquare</td>
<td>0.0075</td>
<td>150</td>
<td>naturalhistorymuseum</td>
<td>0.0046</td>
<td>97</td>
</tr>
<tr>
<td>piccadillycircus</td>
<td>0.0074</td>
<td>182</td>
<td>monument</td>
<td>0.0046</td>
<td>59</td>
</tr>
<tr>
<td>britishmuseum</td>
<td>0.0074</td>
<td>230</td>
<td>victoriaandalbertmuseum</td>
<td>0.0045</td>
<td>64</td>
</tr>
<tr>
<td>gherkin</td>
<td>0.0073</td>
<td>75</td>
<td>bank</td>
<td>0.0044</td>
<td>63</td>
</tr>
<tr>
<td>lloyds</td>
<td>0.0070</td>
<td>121</td>
<td>royalacademy</td>
<td>0.0040</td>
<td>34</td>
</tr>
<tr>
<td>coventgarden</td>
<td>0.0070</td>
<td>169</td>
<td>oxfordstreet</td>
<td>0.0040</td>
<td>51</td>
</tr>
<tr>
<td>buckinghampalace</td>
<td>0.0064</td>
<td>107</td>
<td>bloomsbury</td>
<td>0.0038</td>
<td>27</td>
</tr>
</tbody>
</table>

$P_L$ is the importance score for location L

# user is the number of users visiting the location
Top Ranked Trajectories in London

<table>
<thead>
<tr>
<th>Rank</th>
<th>Trajectory pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>londoneye → bigben → downingstreet → horseguards → trafalgarsquare</td>
</tr>
<tr>
<td>2</td>
<td>londoneye → bigben → tatemodern</td>
</tr>
<tr>
<td>3</td>
<td>tatemodern → bigben → londoneye</td>
</tr>
<tr>
<td>4</td>
<td>londoneye → bigben → parliamentsquare → westminster</td>
</tr>
<tr>
<td>5</td>
<td>westminster → bigben → downingstreet → horseguards → trafalgarsquare</td>
</tr>
<tr>
<td>6</td>
<td>roylalfestivalhall → londoneye → bigben</td>
</tr>
<tr>
<td>7</td>
<td>londoneye → roylalfestivalhall → tatemodern</td>
</tr>
<tr>
<td>8</td>
<td>tatemodern → londoneye → roylalfestivalhall</td>
</tr>
<tr>
<td>9</td>
<td>londoneye → tatemodern → towerbridge</td>
</tr>
<tr>
<td>10</td>
<td>londoneye → towerbridge → tatemodern</td>
</tr>
</tbody>
</table>

The top trajectories are highly biased in only a few regions
Trajectory 1 (londoneye -> bigben -> downingstreet -> horseguards -> trafalgarsquare)
Trajectory 5 (westminster -> bigben -> downingstreet -> horseguards -> trafalgarsquare)
Diversified Ranked Trajectories in London

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tourist route pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bigben → downingstreet → horseguards → trafalgarsquare</td>
</tr>
<tr>
<td>2</td>
<td>spitalfields → shoreditch(1) → shoreditch(2) → shoreditch(3) → shoreditch(4)</td>
</tr>
<tr>
<td>3</td>
<td>charingcross → londoneye</td>
</tr>
<tr>
<td>4</td>
<td>bricklane(1) → bricklane(2)</td>
</tr>
<tr>
<td>5</td>
<td>londoneye → royalfestivalhall → tatemodern</td>
</tr>
<tr>
<td>6</td>
<td>oldstreet(1) → oldstreet(2)</td>
</tr>
<tr>
<td>7</td>
<td>piccadillycircus → soho → oldcomptonstreet</td>
</tr>
<tr>
<td>8</td>
<td>londonbridge → cityhall → towerbridge</td>
</tr>
<tr>
<td>9</td>
<td>gherkin → lloyds → londonbridge → southwark</td>
</tr>
<tr>
<td>10</td>
<td>leicestersquare → chinatown</td>
</tr>
</tbody>
</table>

Trajectories 2, 4, & 5 are popular routes to explore street arts in London
### Location Recommendation in London

#### Current trajectory | Recommended next destination
---|---
londoneye | bigben, tatemodern, trafalgarsquare, southbank, parliamentsquare, towerbridge, piccadillycircus, buckinghampalace
londoneye → bigben | downingstreet, horseguards, trafalgarsquare, parliamentsquare
londoneye → bigben → downingstreet | horseguards, trafalgarsquare
londoneye → tatemodern | southbank, towerbridge, piccadillycircus
londoneye → trafalgarsquare | buckinghampalace

Rank the locations by the scores of trajectories (append current trajectory with next destination)
Trajectory Ranking (SDM’2011)

- Problem Formulation
  - Given a collection of geo-tagged photos along with users, locations and timestamps, how to rank the mined trajectory patterns with diversification into consideration?

- Our Framework
  1. extract trajectory patterns from the photo collection
  2. rank the trajectory patterns by estimating their importance according to user, location and trajectory pattern relations
  3. diversify the ranking result to identify the representative trajectory patterns from all the candidates

- Example: Top trajectories in London, New York City and Paris
Outline

- Why Mining Social and Trajectory Data?
- Periodica and MoveMine
- Social-Related Data Mining: GeoTopic Discovery
- From Mining Social Relationships to Cyberphysical Data Analysis
- Conclusions
Hang out at night for a dinner or on the weekend for outdoor activities

Being together during daytime on weekdays

- Finding a *particular relationship* among moving objects, such as *friends*
  - Two objects having such relationship could meet in certain times
- It is hard to give a concrete definition for each relationship.
  - Both friends and random people could meet on weekends
- Solution: relationship detection in a supervised way:
  - Given some object pairs having/not having such relationship
  - Automatically learn the temporal characteristics
Clustering Cannot Uncover Every Relationship

- We are interested in **friend** relationship, but friends may not be together for LONG time
- Example:
  
  ![Graph showing meeting minutes in a week for a friend pair and a non-friend pair](image)

  - For a specific relationship, there might be some time intervals playing a discriminative role
Relationship Detection Framework

- **T-Motif**: a time interval $[S,T]$, that
  - **many positive** pairs meet at that time
  - **few negative** pairs meet at that time

- Ex.: MIT Reality mining dataset:
  - 94 people tracked for 10 months
  - Use only spatiotemporal info

- Algs. for efficient mining of T-motifs and effective classification

![Graph showing meeting frequency over days of the week]

### Table 1. Top-10 T-Motifs for friend relationship

| T-Motif $[S,T]$ | best split point | $\frac{|D_{\geq T} \cap D^+|}{|D^+|}$ | $\frac{|D_{\geq T} \cap D^-|}{|D^-|}$ |
|-----------------|------------------|--------------------------|--------------------------|
| [21:56 Wed., 23:08 Wed.] | 12              | 0.372881                 | 0.031746                 |
| [22:45 Tue., 23:39 Tue.] | 55              | 0.305085                 | 0.0181406                |
| [19:07 Sat., 7:07 Sun.] | 249             | 0.220339                 | 0.00453515               |
| [20:56 Tue., 22:44 Tue.] | 1               | 0.508475                 | 0.113379                 |
| [23:55 Tue., 1:42 Wed.] | 10              | 0.355932                 | 0.0453515                |
| [23:22 Wed., 3:43 Thurs.] | 53              | 0.220339                 | 0.00680272               |
| [7:08 Sun., 16:49 Sun.] | 53              | 0.40678                  | 0.0770975                |
| [1:20 Fri., 5:12 Fri.] | 12              | 0.20339                  | 0.00680272               |
| [21:52 Mon., 9:00 Tue.] | 11              | 0.644068                 | 0.240363                 |
| [18:12 Sun., 20:01 Sun.] | 3               | 0.389831                 | 0.0793651                |

104
Classification: Using T-Motifs to Detect Relationship

- Given some labeled pairs, discover T-Motifs (discriminative time intervals) for such interesting relationship
- Efficient algorithm for finding T-Motifs and effective algorithm for T-motif-based classification
- Experiments on human mobility data show T-Motifs can better capture the characteristics of relationship
- See: Zhenhui Li, Cindy Xide Lin, Bolin Ding, and Jiawei Han, “Mining Significant Time Intervals for Relationship Detection”, Proc. of 2011 Int. Symp. on Spatial and Temporal Databases (SSTD'11), Minneapolis, MN, Aug. 2011
Geo-Friends Recommendation

- Social network with data collected from sensors is usually referred as Cyber-Physical Social Network
- Problem to be solved: Friend recommendation in GPS-based cyber-physical social networks, by combining GPS data with social network information
- Our method discovers real life friends on web-based social network
  - Geo-Friends: Potential real life friends, who have both social similarities and geographical correlation
**Popularity of Mobile Devices**

- Mobile devices: Very popular, a major media of communication
- Data from mobile devices (like real time GPS location, moving trajectories): Reflect users’ daily activities and real life social interactions
- Social network services: Allow users to store and share locations and trajectories collected from their mobile devices

<table>
<thead>
<tr>
<th>A List of Major Location-Based Social Network Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foursquare</td>
</tr>
<tr>
<td>Google Latitude</td>
</tr>
<tr>
<td>Yelp Check-in</td>
</tr>
</tbody>
</table>

......
A Geo-Friend Finding Example

- **Real life** friends play an important role in off-line social events while most virtual on-line friends can fulfill such social function.

- **Bob** is a college friend who lives in another country now.

- **Carlos** is a co-worker but no social network similarity with Alex.

- **Alex needs geo-friends** join him in a local charity event.

- **David** is more likely to be Alex’s geo-friend, but we cannot get this information by only analyzing social network or GPS data.

- **David** shares common friends and goes to same gym, same game store with Alex.
Data Model

- **GPS Trajectory**: Sequentially connecting GPS records of a particular user, following the ascending order of timestamps.
- **GPS-Based Cyber Physical Social Network**: 

  \[ G(S, V, E) : \]
  - \( V \): Set of people in the network
  - \( E \): Set of edges, represents all the links between people
  - \( S \): Set of GPS trajectories associated with people
Pattern-Based Social Network

- Build a pattern-based heterogeneous information network by combining GPS patterns and social network structures.
- Given $G(S, V, E)$, first discard raw GPS trajectory set $S$.
- Then for each GPS pattern, create an additional node $p$, and link corresponding person node $v$ with $p$ if this GPS pattern exists in person $v$’s GPS trajectory history.
Pattern-Based Social Network (2)

- Create a new edge \(<v, p>\), and add it to \(E'\). Set \(E'\) in contains three types of edges: edges between people, edges from person nodes to pattern nodes, and edges from pattern nodes to person nodes.
Pattern Refinement

- Adding a large number of GPS patterns without selection may decrease the performance badly
  - Common locations contains no social similarity, e.g., bus stop, and hospital
- Instead of manually refining patterns, we employ an entropy-based thresholding measure* to refine and select discriminative GPS patterns
  - This method filter out patterns with high frequency and low length

Datasets

- We generate 4 synthetic datasets with different sizes, attributes and distributions in order to cover different scenarios and thoroughly test our framework.
- Also, apply our method on MIT Reality Mining dataset.
Competitor Methods

- Random: random selection
- Same Edge: choose friends based on number of same friends
- GPS Similarity: choose friends by measuring GPS location and trajectory similarity
- Random Walk without GPS Patterns: Recommend friends by applying random walk with restart on the original social network
- Bluetooth (only MIT dataset): Recommend friends by returning people who share high meeting frequency
Performance Study

gpsnet120 precision

MIT dataset precision

gpsnet120 recall

MIT dataset recall
Outline

- Why Mining Social and Trajectory Data?
- Periodica and MoveMine
- Social-Related Data Mining: GeoTopic Discovery
- From Mining Social Relationships to Cyberphysical Data Analysis
- Looking forward: Future Research Topics
Urban Computing with Social & Trajectory Data

Spatial Data  Text  Temporal Data

- Link
- Text-Rich Information Network Analysis
- Latent Geographical Topic Analysis
- Latent Periodic Topic Analysis
- Periodicity Mining
- Trajectory Pattern Mining
- Trajectory Ranking
Future Urban Computing in the Big Data Age

- What have we done for urban computing?
  - Spatiotemporal and trajectory analysis
  - Space + text analysis (e.g., GeoTopic modeling)
  - Text + time analysis (e.g., Periodic topic analysis: LPTA)
  - Social and information network analysis
  - Integrated analysis across multiple dimensions

- Future directions:
  - Integrated data mining with social and trajectory data
  - Integrated computing with four dimensions
    - Spatial + Temporal + Text + Network
References (I): MoveMine & Social Relationship Mining

- Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, “Mining Periodic Behaviors for Moving Objects”, Proc. 2010 ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'10), July 2010
- Z. Li, J. Wang, and J. Han, "Mining Periodicity for Sparse and Incomplete Event Data", Proc. of 2012 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12), Aug. 2012
- Z. Li, B. Ding, J. Han, and R. Kays, “Swarm: Mining Relaxed Temporal Moving Object Clusters", Proc. 2010 Int. Conf. Very Large Data Bases (VLDB'10), Sept. 2010
- Jingjing Wang and Bhaskar Prabhala, "Periodicity Based Next Place Prediction", Proc. of Workshop on Mobile Data Challenge by Nokia, June 2012
- Zhenhui Li, Cindy Xide Lin, Bolin Ding, and Jiawei Han, “Mining Significant Time Intervals for Relationship Detection”, Proc. of 2011 Int. Symp. on Spatial and Temporal Databases (SSTD'11), Aug. 2011
References (II): Trajectory Mining

- J.-G. Lee, J. Han, and X. Li, "Trajectory Outlier Detection: A Partition-and-Detect Framework", ICDE 2008
- Jae-Gil Lee, Jiawei Han, Xiaolei Li, and Hector Gonzalez, “TraClass: Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering”, Proc. 2008 Int. Conf. on Very Large Data Base (VLDB'08), Auckland, New Zealand, Aug. 2008.
- Xiao Yu, Ang Pan, Lu-An Tang, Zhenhui Li, and Jiawei Han, "Geo-Friends Recommendation in GPS-based Cyber-Physical Social Network", Proc. of 2011 Int. Conf. on Advances in Social Network Analysis and Mining (ASONAM'11), Kaohsiung, Taiwan, July 2011
References (III): GeoTopic Discovery

- Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang, "Geographical Topic Discovery and Comparison", Proc. of 2011 Int. World Wide Web Conf. (WWW'11), Hyderabad, India, Mar. 2011.

- Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas Huang, "Diversified Trajectory Pattern Ranking in Geo-tagged Social Media", Proc. of 2011 SIAM Conf. on Data Mining (SDM'11), Phoenix, AZ, Apr. 2011.

- Zhijun Yin, Liangliang Cao, Quanquan Gu and Jiawei Han, "Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling", ACM TIST, 2012

- Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai and Thomas S. Huang, “LPTA: A Probabilistic Model for Latent Periodic Topic Analysis”, ICDM’2011

- Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas S. Huang, "Diversified Trajectory Pattern Ranking in Geo-tagged Social Media", SDM'11

- Liangliang Cao, Xin Jin, Zhijun Yin, Andrew D. Del Pozo, Jiebo Luo, Jiawei Han, Thomas S. Huang, “RankCompete: Simultaneous Ranking and Clustering of Information Networks “, Neurocomputing
FREEDOM IS A LIGHT
FOR WHICH MANY MEN HAVE DIED IN DARKNESS

IN UNMARKED GRAVES WITHIN
THIS SQUARE LIE THOUSANDS
OF UNKNOWN SOLDIERS OF
WASHINGTON'S ARMY WHO DIED
OF WOUNDS AND SICKNESS DURING
THE REVOLUTIONARY WAR

THE INDEPENDENCE AND LIBERTY
YOU POSSESS ARE THE WORK OF
JOINT COUNCILS AND JOINT
EFFORTS OF COMMON DANGERS,
SUFFERINGS AND SUCCESS.
WASHINGTON'S FAREWELL ADDRESS SEPT. 17, 1796