Mining Spatiotemporal and Social Media Data

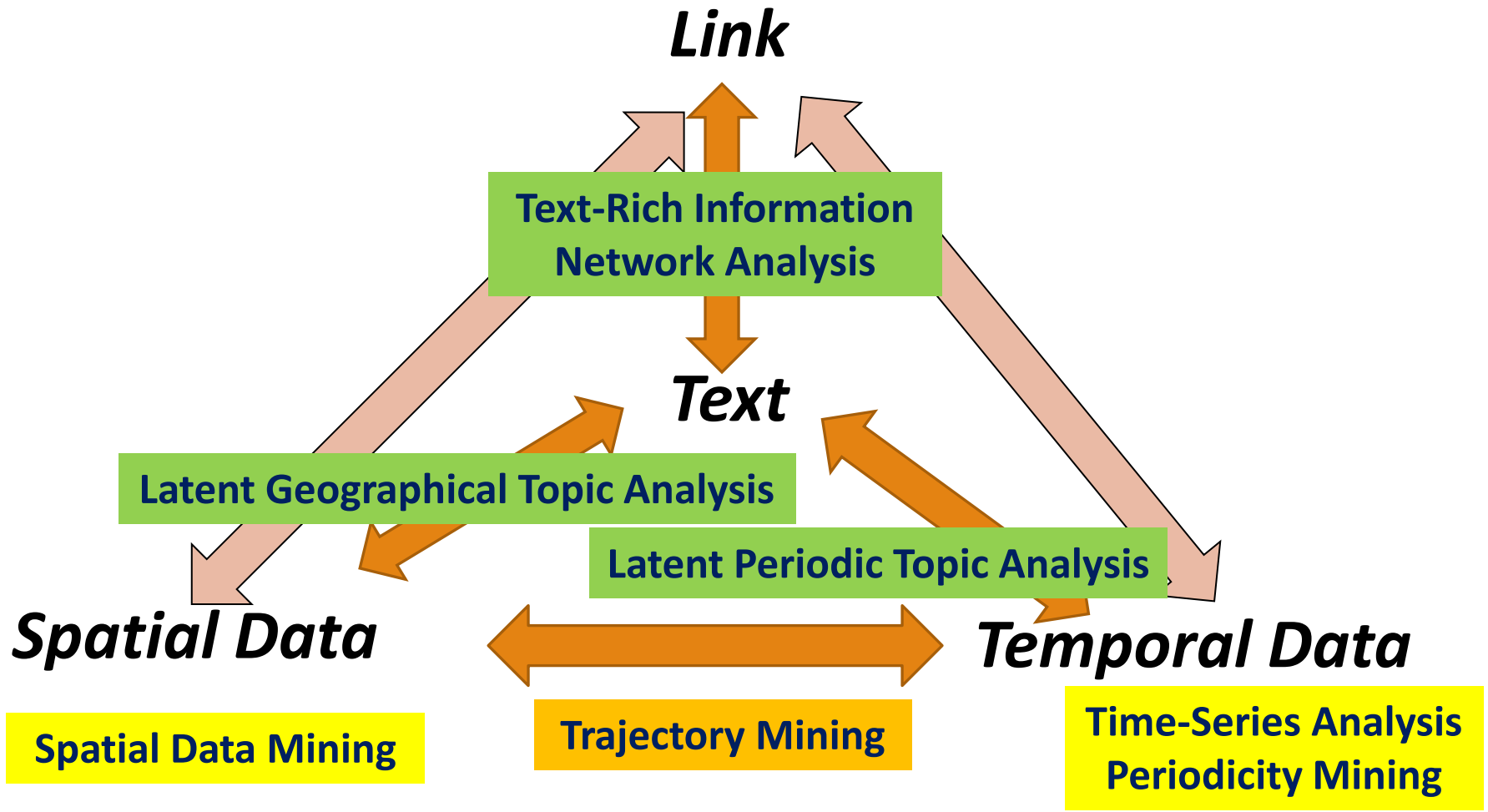
JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

APRIL 20, 2017

Outline

- Introduction: Integrated Mining of Spatio, Temporal and Text Data
- Mining Spatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary

Introduction: Integrated Mining of Spatial, Temporal and Text Data



Outline

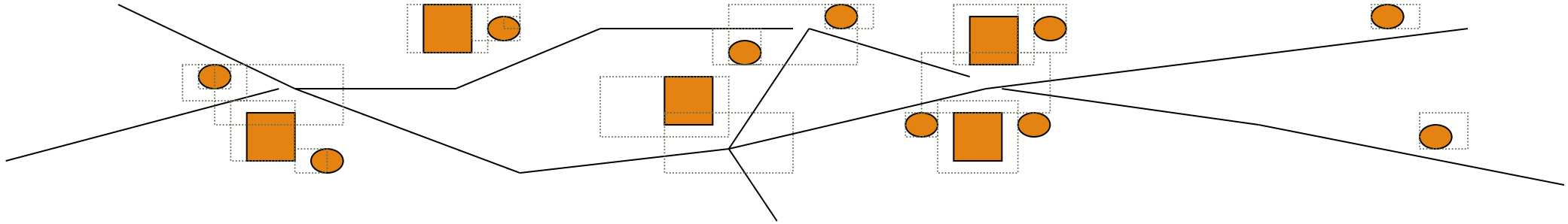
- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns 
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary

Spatial Patterns and Associations

- Spatial frequent patterns and association rule: $A \Rightarrow B$ [*s%, c%*]
  - A and B are sets of spatial or non-spatial predicates, e.g.,
    - Topological relations: *intersects, overlaps, disjoint,* etc.
    - Spatial orientations: *left_of, west_of, under,* etc.
    - Distance information: *close_to, within_distance,* etc.
  - Measures: *s%*: support, and *c%*: confidence of the rule
- Example: Rules likely to be found
  - *is_a(x, large_town) ^ intersect(x, highway) → adjacent_to(x, water) [7%, 85%]*
- Explore *spatial autocorrelation*: Spatial data tends to be highly self-correlated (*nearby things are more related than distant ones)*
  - E.g., neighborhood, temperature

Mining Spatial Associations: Progressive Refinement

- Hierarchy of spatial relationship:
  - *close_to* is a generation of *near_by*, *touch*, *intersect*, *contain*, ...
  - Progressive refinement: First search for rough relationship and then refine it
- Two-step mining of spatial association:
  - Step 1: Rough spatial computation (as a filter)
    - Using MBR (Minimum Bounding Rectangle) or R-tree for rough estimation
  - Step2: Detailed spatial algorithm (as refinement)
    - Apply only to those objects which have passed the rough spatial association test (no less than *min_support*)

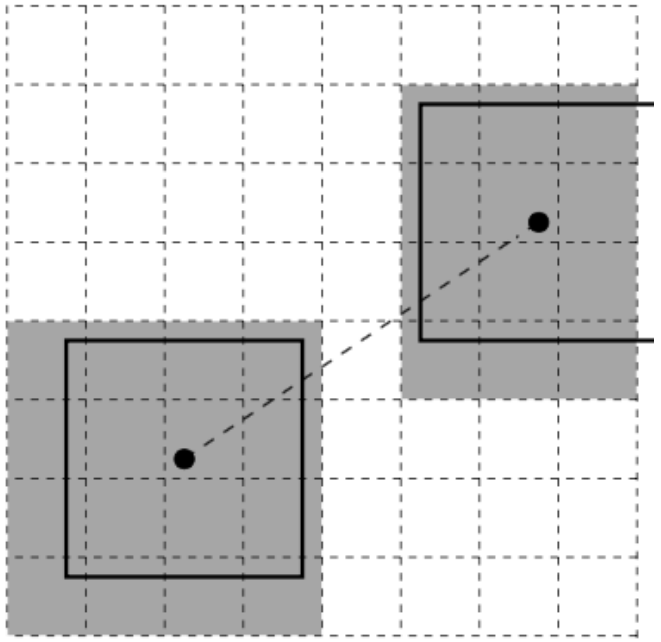


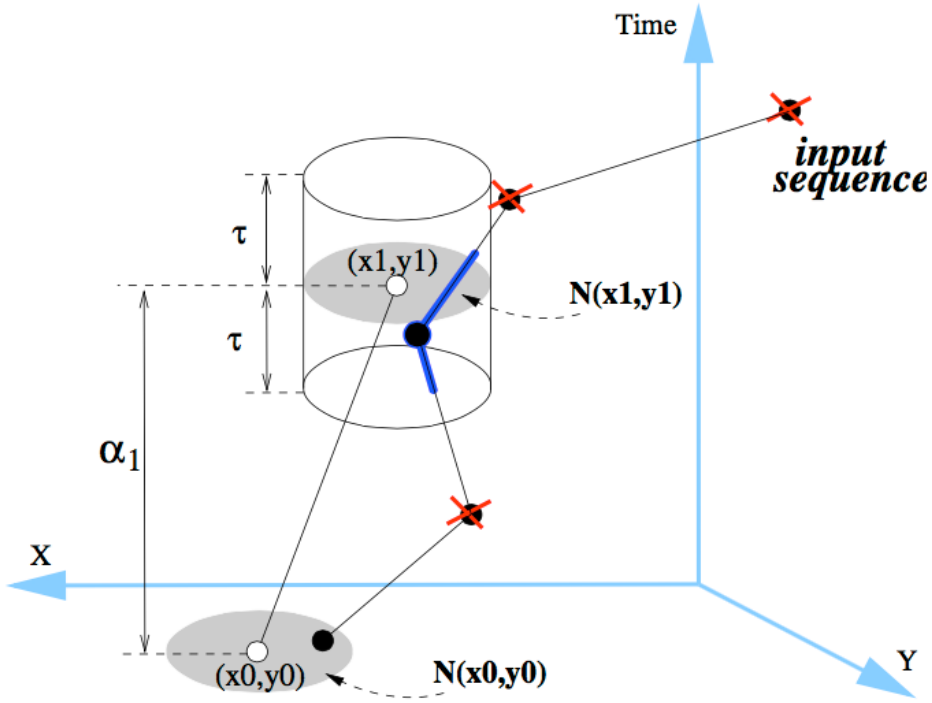
Outline

- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories 
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary

Partition-Based Trajectory Pattern Mining

- Partition-Based Trajectory Pattern Mining (e.g., Mining T-Patterns) [1]:
  - First partition the space into equal-width grids and obtain Regions-of-Interests (RoIs)
  - Then transform each input trajectory into a time-annotated symbolic sequence
  - Use constraint-based sequential pattern mining to find trajectory patterns

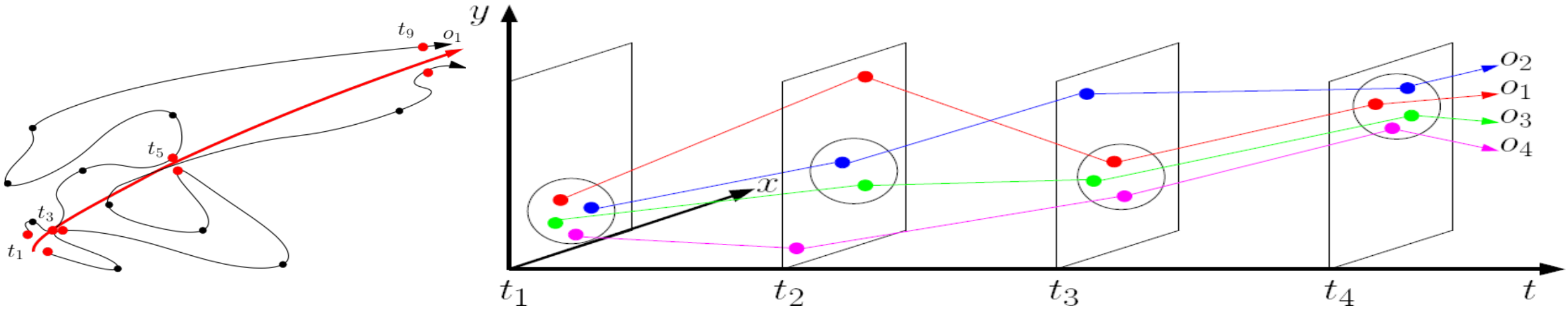




[1] F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory Pattern Mining, KDD'07

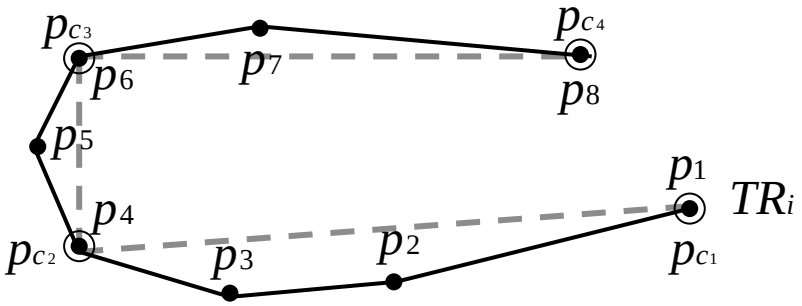
Detecting Moving Object Clusters

- **Flock and convoy**: Both require *k* consecutive time stamps
  - **Flock:** At least *m* entities are within a *circular* region of ***radius r*** and move in the same direction
  - **Convoy:** *Density-based clustering* at each timestamp; no need to be a rigid circle
- **Swarm:** Moving objects may not be close to each other for all the consecutive time stamps
  - Efficient pattern mining algorithms for uncovering such swarm patterns

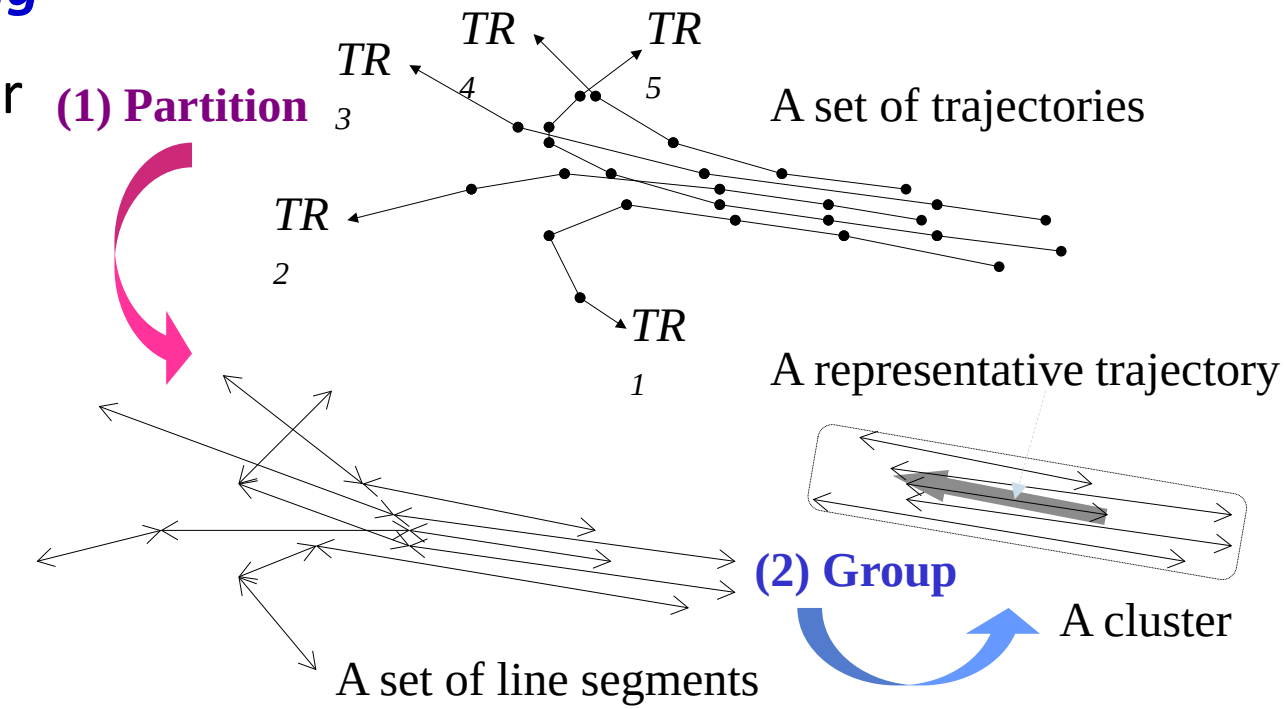


Trajectory Clustering: A Partition-and-Group Framework

- Grouping trajectories ***as a whole*** ⇒ cannot find ***similar portions*** of trajectories
- ***Solution:*** discovers common ***sub***-trajectories, e.g., *forecast hurricane landfall*
- Two phases: ***partitioning*** and ***grouping***
- Identify the points where the behavior of a trajectory changes rapidly ⇒ *characteristic points*
  - Based on the minimum description length (MDL) principle



◉ : characteristic point – – – : trajectory partition



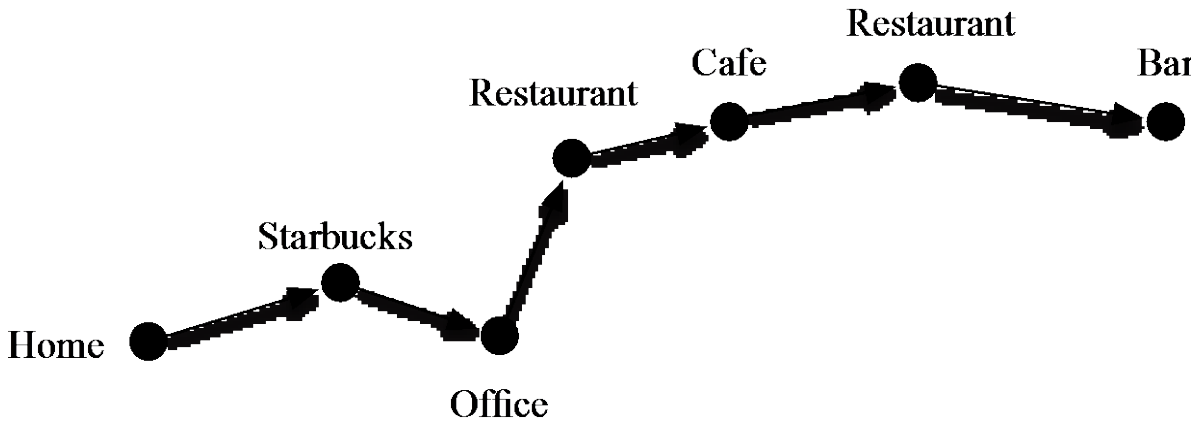
J.-G. Lee, et al., "Trajectory Clustering: A Partition-and-Group Framework", SIGMOD'07

Outline

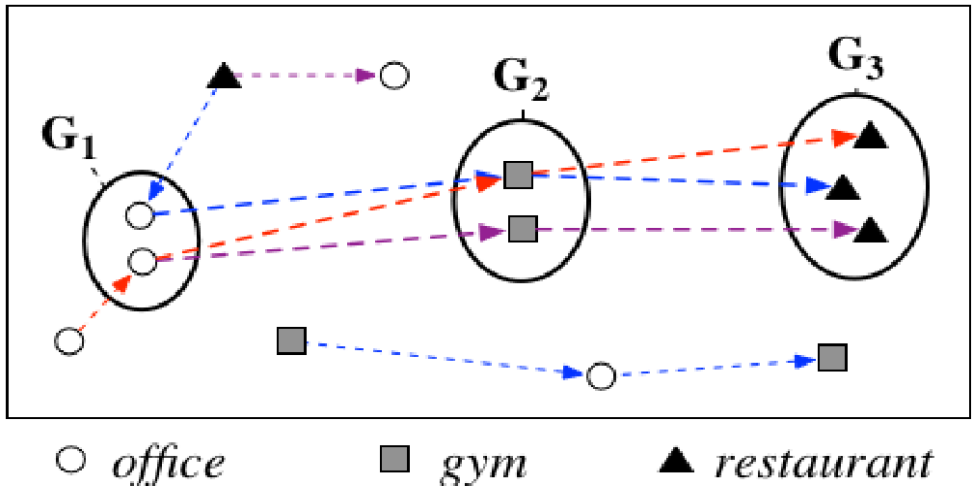
- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns 
- Mining Periodic Movement Patterns
- GeoTopic Discovery
- From Mining Social Relationships
- Summary

Mining Frequent Movement Patterns

- **Frequent Movement Pattern:** A movement sequence that frequently appears in the input trajectory database
- **Frequent Movement Pattern vs. Frequent Sequential Pattern**
  - Both aim at finding frequent subsequences from the input sequence database
  - For mining frequent movement patterns, similar places may need to be grouped to collectively form frequent subsequences



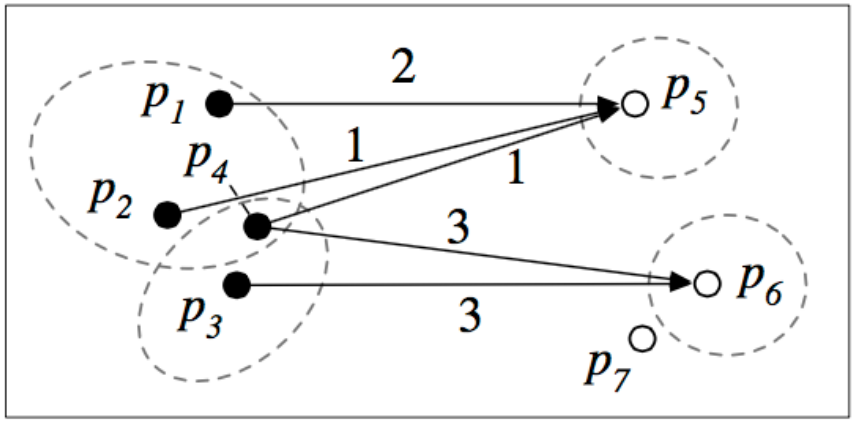
An example trajectory

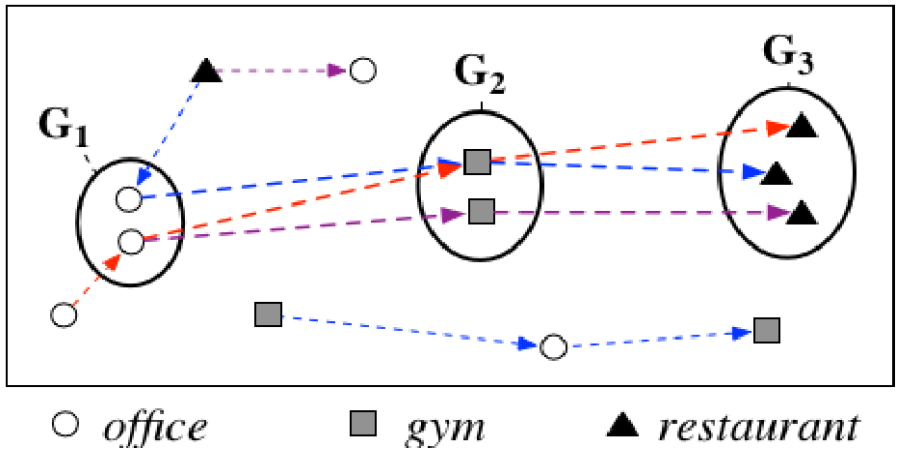


An example movement pattern

Mining Semantic-Rich Movement Patterns

- **Semantics-rich Movement Pattern**: In addition to knowing how people move from one region to another, we also want to understand the functions of the regions
- **A two-step top-down mining approach:**
  - Step 1: Find a set of coarse patterns that reflect people's semantics-level transitions (e.g., office → restaurant, home → gym)
  - Step 2: Split each coarse pattern into several fine-grained ones by grouping similar movement snippets





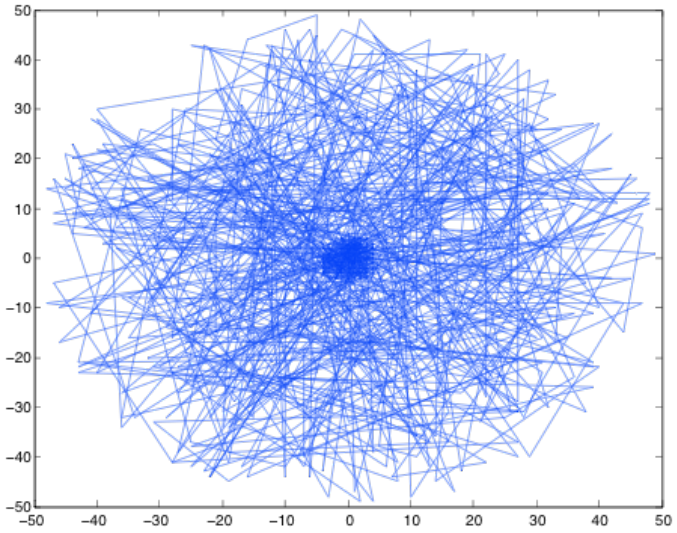
[1] C. Zhang et al., Splitter: Mining Fine-Grained Sequential Patterns in Semantic Trajectories, VLDB 2014

Outline

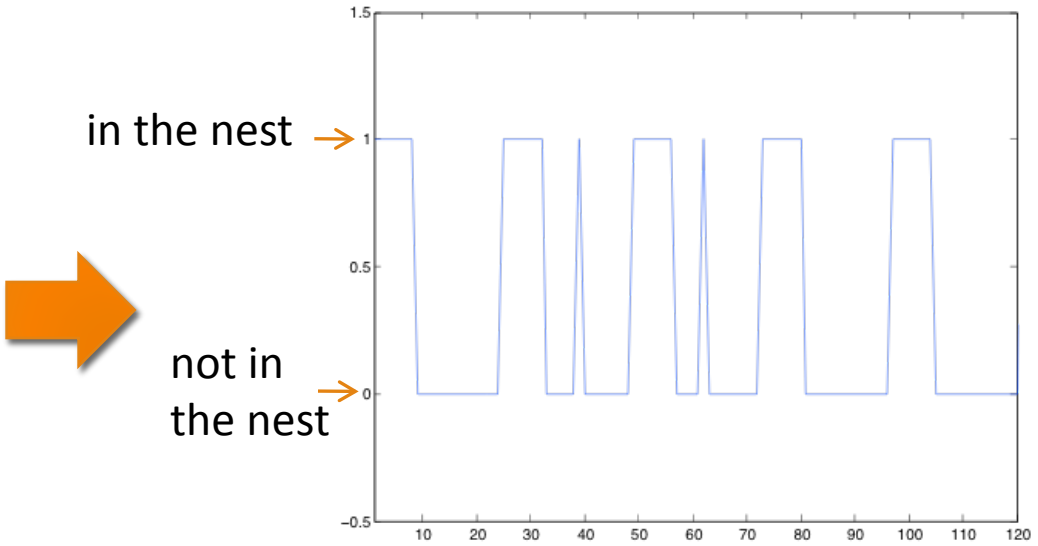
- Introduction: Integrated Mining of Spatial, Temporal and Text Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns 
- GeoTopic Discovery in Social Media
- From Mining Social Relationships
- Summary

Pattern Discovery in Sparse Movement Data: Finding Good Reference Points

- **Pattern discovery in sparse data**:
  - Periodicity shows up in some reference "spots" (or "cluster centers")
  - Reference spots can be detected using **density-based method**
  - Periods are detected for each reference spot using **Fourier Transform** and **auto-correlation**

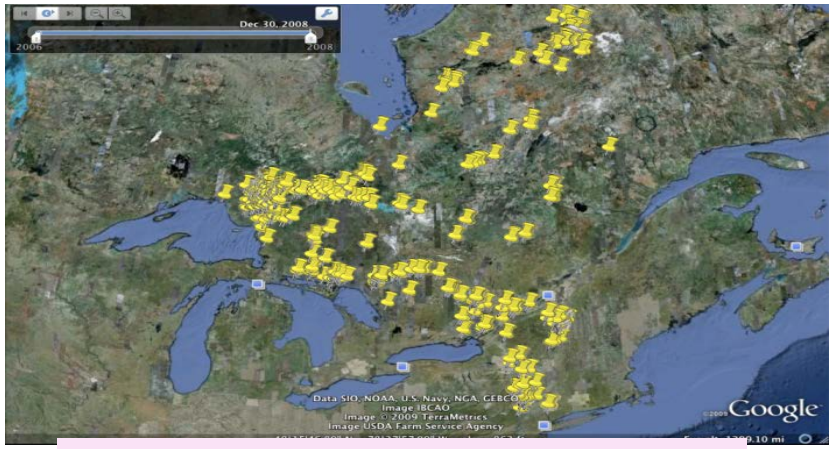


Finding being flying patterns? Bee hive is a good reference point

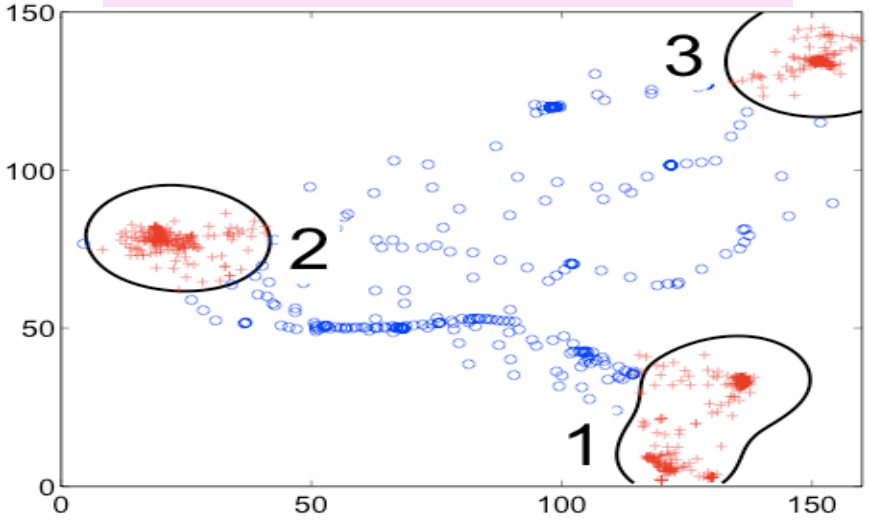


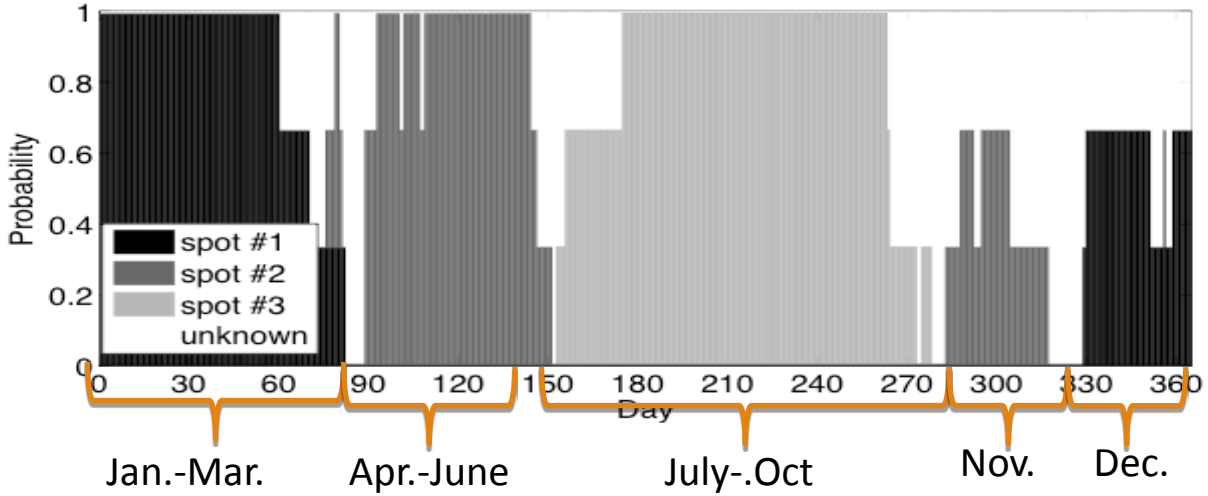
Period is more obvious in this binary sequence!

Example: Mining Periodic Patterns with Sparse Data



3-yr Bird migration data: very sparse

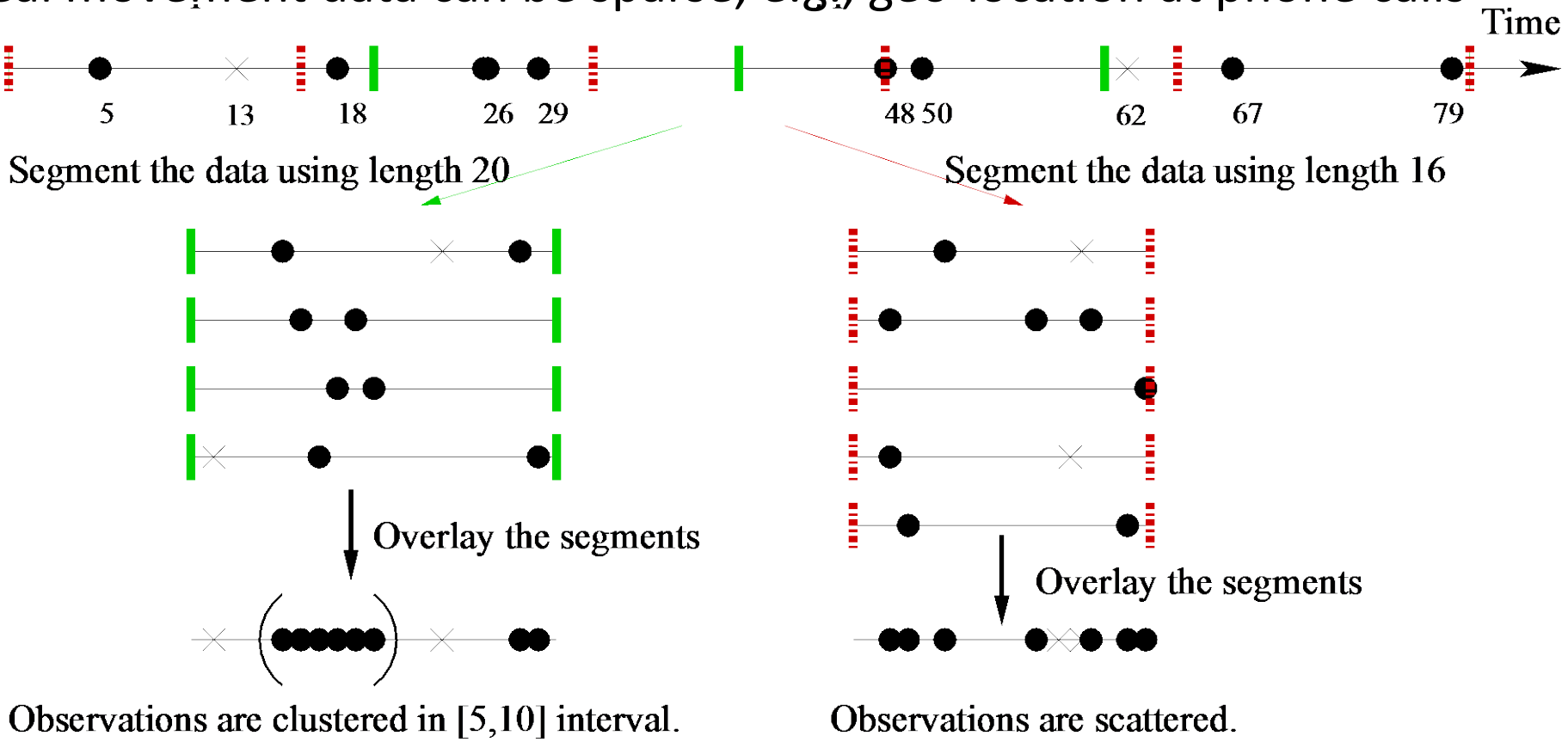




- **Detecting periods**: Cluster data to find reference "points" and then detect multiple interleaved periods by Fourier Transform and auto-correlation
- **Summarizing periodic patterns:** By clustering and pattern discovery

Periodicity Detection in Sparse Data

- Real movement data can be sparse, e.g., geo-location at phone calls



- Projecting on the true period, it shows a highly skewed (clustered) distribution
  - Effective method can be developed based on this observation (Li, et al., 2015)

Outline

- Integrated Mining of Spatial, Temporal and Social Media Data
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media
- From Mining Social Relationships
- Summary

Social Media Are Popular in Today's World

- Social media contains very rich spatial, temporal, text, photo, link, social network information
- Examples
  - Twitter: tweets from smart phones
    - Geo-tagged tweets
  - Flickr: geo-tagged photos
  - Advanced cameras with GPS receivers
  - Applications including Google Earth, Flickr, etc.
  - GPS functions in smart phones
- Social media data mining: A rich frontier

LGTA: 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
- 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.

GeoTopic: From Geo-Tagged Text to Topic Clusters

- Input: Text with spatial information

Geo-tagged photos related to *Food* in Flickr

ID	Image	Text	Latitude	Longitude	Location
1		dimsum breakfast dumplings …	22.377	114.185	
2		sushi sashimi rawfish …	35.669	139.762	
3		taco tacogrill crispybeef …	30.265	-97.680	
…	…	…	…	…	…

- Output:
  - Geographic topics: { p(w|z) }
    - $p(w|z_1)$, $p(w|z_2)$, $p(w|z_3)$
  - Topic distribution p(z|l)

Topic 1 (Chinese food)	Topic 2 (Japanese food)	Topic 3 (Mexican food)	…
noodles 0.067 dimsum 0.064 hotpot 0.039 rice 0.038 noodle 0.035 …	ramen 0.104 soba 0.066 noodle 0.065 sashimi 0.039 yakitori 0.030 …	tacos 0.069 taco 0.059 salsa 0.036 cajun 0.031 burrito 0.027 …	…

Location l = (40.70, 73.91)

topic z	p(z|l)
Topic 1 (Chinese food)	22%
Topic 2 (Japanese food)	14%
Topic 3 (Mexican food)	18%
…	…

Potential Solutions: Previous Work

- 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

LGTA: General Ideas (Location-Text Join Model)

- Geographical topic discovery
  - Topics are generated from regions instead of documents:
    - The geographic distribution of each region follows a Gaussian distribution
  - The words that are close in space likely belong to the same region and thus should be clustered into the same geographical topic
- To generate a geographical document d in a collection D:
  - Sample a region r from the discrete distribution of region importance α:
    - r ~ Discrete (α)
  - Sample location l_d from Gaussian distribution of μ_r and $\sum_r$
  - To generate each word in document d:
    - (a) sample a topic z from multinominal ϕ_r
    - (b) sample a word w from multinominal θ_z
  - Each topic can be related to several regions
  - Parameters can be estimated using an EM algorithm

$$p(l_d|\mu_r, \Sigma_r) = \frac{1}{2\pi\sqrt{|\Sigma_r|}} \exp\left(\frac{-(l_d-\mu_r)^T\Sigma_r^{-1}(l_d-\mu_r)}{2}\right)$$

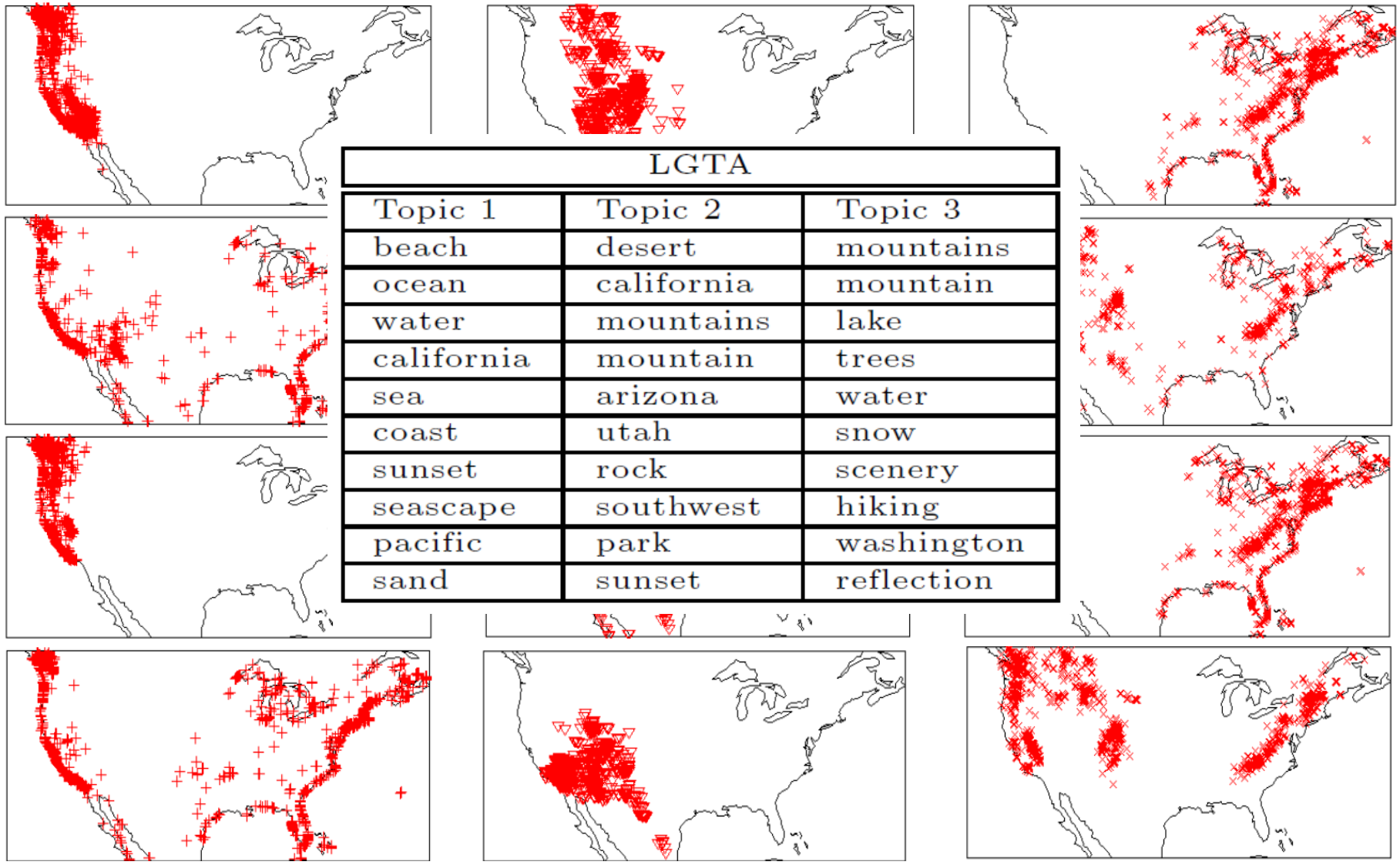
Performance Comparison: Geo-Tagged Photos Related to Landscape (coast vs. desert vs. mountain)

LDM

TDM

GeoFolk

LGTA

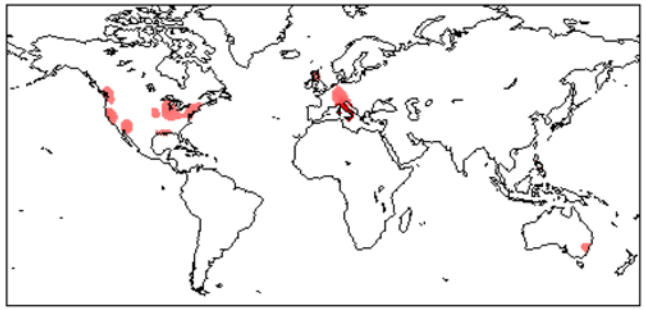


LGTA		
Topic 1	Topic 2	Topic 3
beach	desert	mountains
ocean	california	mountain
water	mountains	lake
california	mountain	trees
sea	arizona	water
coast	utah	snow
sunset	rock	scenery
seascape	southwest	hiking
pacific	park	washington
sand	sunset	reflection

LGTA: Geographical Food Topic Comparison

Chinese Food

Japanese Food



Italian Food

French Food

Spanish Food

Mexican Food

The larger p(topic|location) is, the darker the location is

Prior

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

French Food	Spanish Food	Mexican Food
french 0.564	spanish 0.488	mexican 0.484
bistro 0.070	tapas 0.269	tacos 0.069
patisserie 0.056	paella 0.076	taco 0.059
bakery 0.049	pescado 0.059	salsa 0.036
resto 0.044	olives 0.032	cajun 0.031
pastry 0.033	stickyrice 0.017	burrito 0.027
tarte 0.026	tortilla 0.013	crawfish 0.023
croissant 0.021	mediterranean 0.010	guacamole 0.022
baguette 0.019	mussels 0.008	margarita 0.020
mediterranean 0.018	octopus 0.008	cocktails 0.020

Outline

- Introduction
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media Data
- Latent Periodic Topic Discovery 
- Real-Time Local Event Detection from Geo-Tagged Social Media
- Summary

Latent Periodic Topic Analysis [ICDM'11]

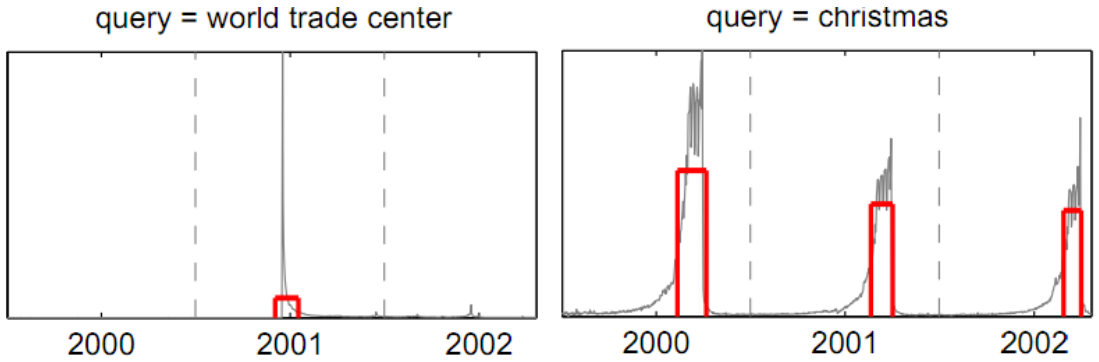
- Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "LPTA: A Probabilistic Model for Latent Periodic Topic Analysis", ICDM'11
- Periodic phenomena exist ubiquitously
  - Hurricanes
  - Music and film festivals
  - Product sales
  - TV program
  - Publicly traded company
- Most text articles have time associated with

- Ex. 1. News articles associated with pub. dates

- Ex. 2. 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 is insufficient 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 considered together
  - Synonyms and polysemy words due to the language diversity

Latent Periodic Topic Analysis (LPTA)

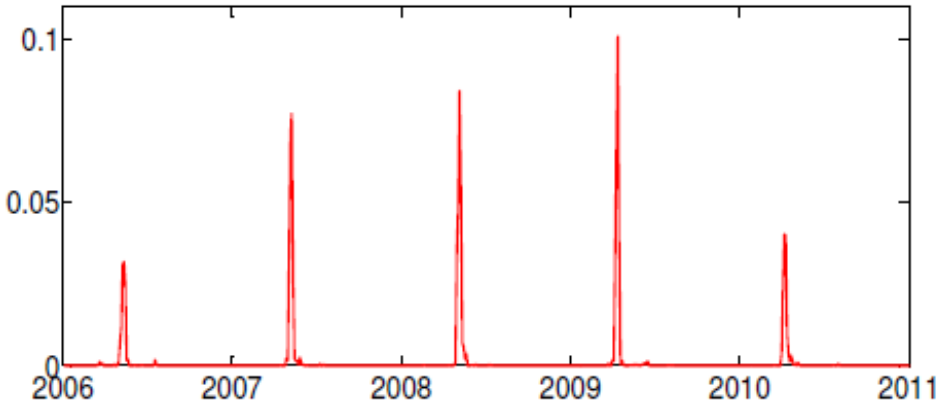
Input:
Time-stamped documents

ID	Text	Date
1	coachella, music, arts, festival, …	Apr 27 2008
2	sxsw, south by southwest, austin, …	Mar 14 2008
…	…	…

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

Periodic interval T, e.g., 1 year, etc.

Topic 1 (Coachella Festival)	…
coachella 0.1106	…
music 0.0915	…
indio 0.0719	…
california 0.0594	…
concert 0.0357	…
…	…



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 $\theta = \{\theta_z\}_{z \in Z}$

$$\theta_z = \{p(w \mid z)\}_{w \in V}$$

    - $p(w \mid z)$ is the probability of word w given topic z
  - The distribution of the timestamps for each topic

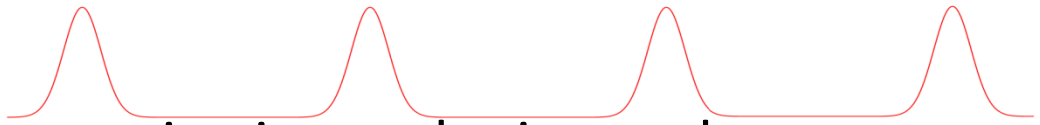
General Idea of LPTA

- 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

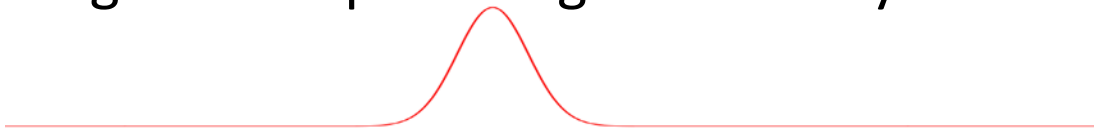
Temporal Patterns of Topics

- Periodic topics



  - A periodic topic is one repeating in regular intervals
  - 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

  - A background topic is one covered uniformly over the entire period
  - The timestamps of the background topics are generated by a uniform distribution
- Bursty topics



  - A bursty topic is a transient topic that is intensively covered only in a certain time period
  - 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:
  - Sample a topic z from multinomial ϕ_d i.e., $\{p(z|d)\}_{z \in Z}$
    - (a) If z is a background topic, sample time t from a uniform distribution [t_{start} , t_{end}], where t_{start} and t_{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
  - Sample a word w from multinomial θ_z i.e., $\{p(w|z)\}_{w \in V}$

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

$$L(\psi; D) = \log p(D \mid \psi) = \log \prod_{d \in D} p(w_d, t_d \mid \psi)$$

$$\log p(w_d, t_d \mid \psi) = \sum_d \sum_w n(d, w) \log \sum_z p(t_d \mid z)\, p(w \mid z)\, p(z \mid d)$$

- If topic z is a background topic, $p(t \mid z) = \dfrac{1}{t_{end} - t_{start}}$

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

- If topic z is a periodic topic, $p(t \mid z) = p(k) \dfrac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(t-\mu_z - kT)^2}{\sigma_z^2}}$

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

$$p(w|z) = \frac{\sum_d n(d,w)p(z|d,w)}{\sum_d \sum_{w'} n(d,w')p(z|d,w')} \qquad p(z|d) = \frac{\sum_w n(d,w)p(z|d,w)}{\sum_w \sum_{z'} n(d,w)p(z'|d,w)}$$

- For bursty topic z

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

- For periodic topic z

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

- Complexity: O(*iter* K|W|) where *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
  - Weekly seminar announcements for one semester from six research groups @CS, UIUC
  - 61 documents and 901 unique words
  - Set periodic interval T as 1 week
- DBLP (Computer Science Digital Bibliography)
  - The paper titles of several confs (WWW, SIGMOD, SIGIR, KDD, VLDB and NIPS) from 2003 to 2007
  - 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

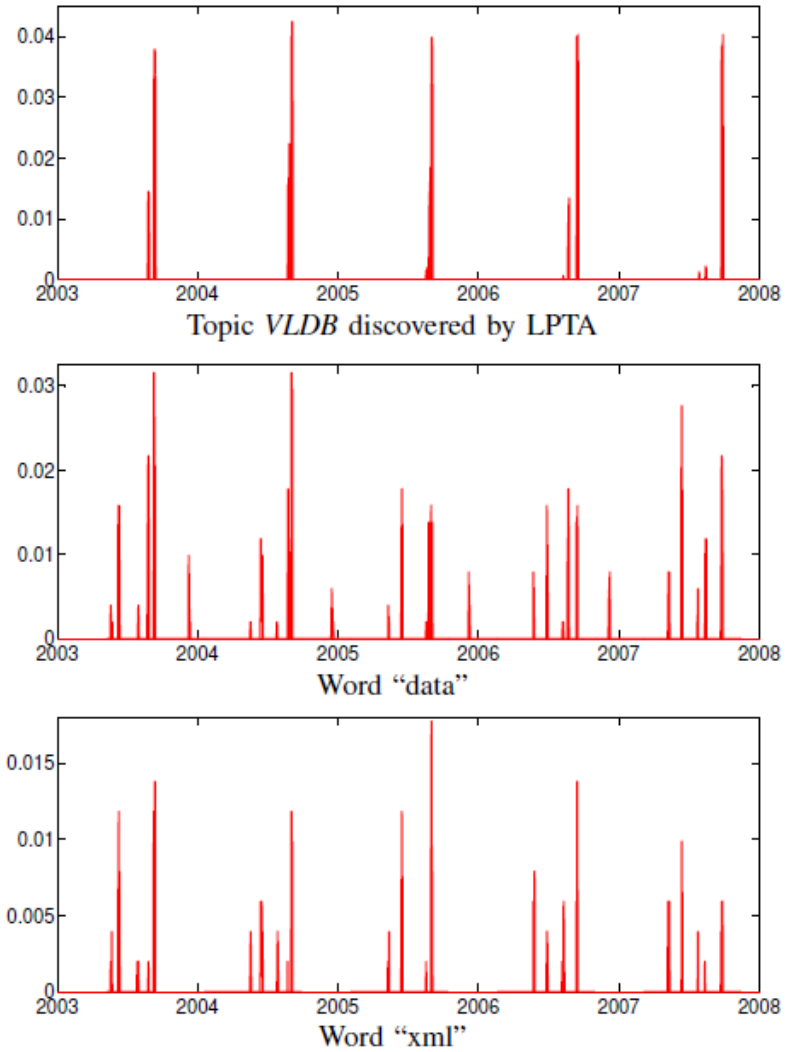
Topics Discovered by LPTA

- 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

Seminar		DBLP		Flickr	
Topic 1 (DAIS) Tue 16:00 (0h0m0s)	Topic 2 (AIIS) Fri 14:00 (0h0m0s)	Topic 1 (KDD) Aug 23 (10d3h11m)	Topic 2(SIGIR) Aug 3 (9d6h56m)	Topic 1 (ACL) Sep 29 (10d13h20m)	Topic 2 (Bonnaroo) Jun 16 (2d14h21m)
model 0.0166	computer 0.0168	mining 0.0353	retrieval 0.0495	acl 0.0945	bonnaroo 0.1066
based 0.0158	learning 0.0158	data 0.0289	based 0.0197	austin 0.0827	music 0.0870
mining 0.0151	machine 0.0138	search 0.0233	web 0.0189	music 0.0763	manchester 0.0587
text 0.0143	science 0.0128	clustering 0.0208	text 0.0171	austincitylim. 0.0442	tennessee 0.0518
network 0.0135	algorithms 0.0128	based 0.0195	query 0.0164	limits 0.0441	live 0.0327
web 0.0119	language 0.0118	web 0.0168	search 0.0162	city 0.0441	concert 0.0275
problem 0.0111	work 0.0108	learning 0.0159	document 0.0149	texas 0.0426	arts 0.0175
data 0.0111	problems 0.0108	networks 0.0114	language 0.0118	concert 0.0283	performance 0.0174
query 0.0111	models 0.0108	analysis 0.0105	relevance 0.0111	live 0.0212	backstagegall. 0.0113
latent 0.0095	prediction 0.0108	large 0.0104	evaluation 0.0111	zilker 0.0173	rock 0.0111

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



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

LPTA vs. Topic Models

Selected topics discovered for different datasets by using PLSA and LDA

Seminar				DBLP				Flickr			
PLSA		LDA		PLSA		LDA		PLSA		LDA	
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2
data	memory	problem	systems	web	search	web	system	sxsw	lollapaloo.	music	lollapaloo.
latent	computer	algorithm	computer	data	text	mining	database	austin	music	coachella	music
visualizati.	data	network	science	xml	databases	semantic	distributed	music	chicago	bonnaroo	chicago
intel	mining	graph	algorithms	queries	relational	detection	user	texas	concert	california	live
talk	parallel	time	time	mining	user	automatic	adaptive	southbyso.	acl	manchester	concert
analysis	science	networks	agent	semantic	analysis	services	content	live	grantpark	indio	grantpark
computer	pattern	influence	visualizati.	search	ranking	applicatic.	relevance	atx	live	tennessee	august
systems	programm.	online	data	streams	structure	graph	performan.	coachella	austincity.	arts	photos
machine	hardware	work	engineering	managem.	support	extraction	feedback	downtown	august	art	summer
visual	algorithms	question	function	adaptive	evaluation	patterns	image	livemusic	austin	palmsprin.	performan.

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.

- 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

SIGMOD vs. VLDB		SIGMOD vs. CVPR	
Topic 1 (SIGMOD) Jun 17 (7d11h6m)	Topic 2 (VLDB) Sep 11 (9d5h29m)	Topic 1 (SIGMOD) Jun 20 (7d15h42m)	Topic 2 (CVPR) Jun 21 (3d4h37m)
data	data	data	image
query	xml	query	based
xml	query	xml	tracking
database	queries	database	recognition
processing	efficient	processing	learning
efficient	database	efficient	object
databases	based	based	shape
queries	databases	system	segmentation
web	system	databases	detection
system	processing	queries	motion

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

- The words will fit into the corresponding periodic or bursty topics if they have periodic or bursty patterns

Bursty topics		Periodic topics	
Topic 1 (Lollapalooza) Aug 8 2009 (1d0h12m)	Topic 2 (Coachella) Apr 17 2009 (10d20h23m)	Topic 3 (SXSW) Mar 18 (6d8h33m)	Topic 4 (ACL) Sep 28 (14d7h22m)
lollapalooza	coachella	sxsw	acl
chicago	indio	austin	austin
concert	music	texas	music
music	california	music	austincityli.
grantpark	concert	southbysouth.	city
august	live	live	limits
live	desert	concert	texas
illinois	art	atx	concert
performance	musicfestival	downtown	live
lolla	livemusic	gig	zilker

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

K	Seminar						DBLP						Flickr					
	Accuracy(%)			NMI(%)			Accuracy(%)			NMI(%)			Accuracy(%)			NMI(%)		
	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA	PLSA	LDA	LPTA
2	31.1	31.8	**37.7**	11.7	12.3	**34.7**	24.2	25.4	**38.3**	1.9	2.8	**23.9**	45.7	48.9	**49.7**	22.4	28.3	**37.2**
3	37.0	38.0	**51.0**	19.0	19.9	**53.0**	26.8	26.8	**51.1**	3.6	3.8	**45.7**	57.7	59.9	**63.1**	35.9	42.1	**54.9**
4	39.4	41.3	**65.4**	23.6	24.0	**70.7**	26.5	27.7	**61.5**	3.8	4.5	**56.7**	63.7	70.6	**74.8**	42.2	53.8	**67.4**
5	40.1	42.1	**78.5**	25.7	26.6	**82.4**	27.1	28.7	**66.1**	4.5	5.6	**63.0**	69.2	74.8	**85.7**	48.6	59.9	**79.2**
6	43.0	41.9	**90.4**	30.6	28.9	**92.3**	26.6	27.8	**67.8**	4.7	5.7	**65.9**	67.6	78.5	**90.2**	47.9	60.2	**82.1**
7	40.8	39.5	**94.5**	30.5	29.7	**94.2**	24.0	26.2	**65.9**	4.3	5.8	**63.8**	67.2	71.5	**89.6**	46.5	54.3	**80.2**
8	39.0	40.0	**91.9**	30.4	31.0	**91.7**	22.3	23.9	**66.7**	4.4	5.6	**63.1**	66.0	69.8	**86.5**	45.7	53.1	**77.6**
9	35.3	36.9	**90.0**	30.5	30.8	**88.8**	20.8	22.3	**65.1**	4.4	5.6	**60.8**	64.2	64.5	**83.7**	44.3	50.6	**74.7**
10	34.9	33.9	**88.1**	31.7	30.2	**86.8**	19.6	20.6	**63.6**	4.5	5.5	**58.2**	63.1	67.7	**81.4**	43.5	51.4	**73.1**
Avg	37.9	38.4	**76.4**	26.0	26.0	**77.2**	24.2	25.5	**60.7**	4.0	5.0	**55.7**	62.7	67.3	**78.3**	41.9	50.4	**69.6**

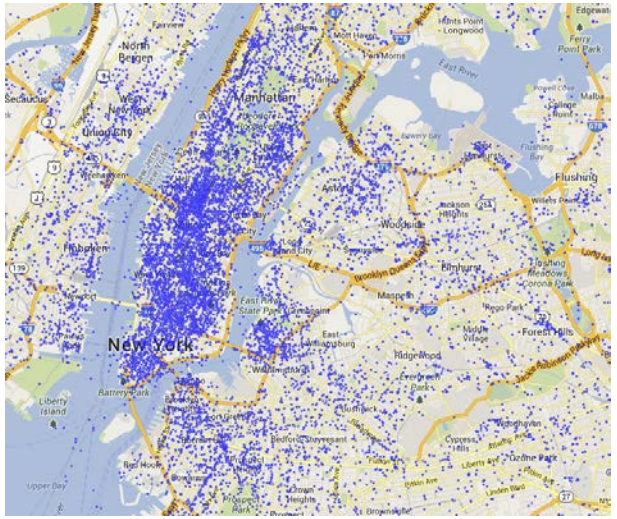
- Conclusion: The LPTA model discovers the latent periodic topics by combining the information from topical clusters and periodic patterns

Outline

- Introduction
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media Data
- Latent Periodic Topic Discovery
- Real-Time Local Event Detection from Geo-Tagged Social Media 
- Summary

GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams [SIGIR'16]

- C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, J. Han, "GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams", SIGIR'16
- Local Event: A local events is an ***unusual activity*** bursted within a ***local area*** and ***specific duration*** while engaging a considerable number of participants
  - E.g., parade, riot, sport game, concert, accident, disaster



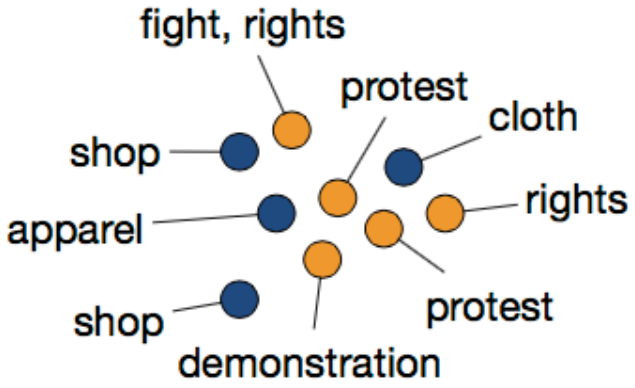
- The geo-tagged tweet stream brings new opportunities to this problem because of its (1) sheer size; (2) multi-dimensional information; and (3) real-time nature

Research Challenges

- Major challenges
  - Integrating multiple types of data: Location, time and text
  - Extracting interpretable events from tremendous noises (tweets are noisy and short)
  - On-line and real-time detection
- Previous work
  - Most existing event detection methods are designed for detecting ***global events***
    - Bursty in the entire stream; but local events are “bursty” in a small region and involve a relatively small number of tweets
  - Some local event detection methods
    - Not model the correlations between keywords; or are incapable of detecting local events in real time

Insight and Problem Definition

- A local event usually leads to many related tweets around the location **(a geo-topic cluster)**
- But **a geo-topic cluster is not necessarily a local event**
  - It may be a routine activity in that region (e.g., shopping)
  - It may be a global event rather than a local one (e.g., TV show)



We define a local event as a geo-topic cluster that shows clear spatiotemporal burstiness

- Our task: Given the geo-tagged tweet stream, we aim to
  - detect all local events in any query time window (**batch mode**)
  - update the result list in real time as the query window shifts continuously (**online mode**)

Overview of GeoBurst

- GeoBurst, a reference-based method for local event detection
  - It consists of three key components:
    - **A candidate generator** that finds geo-topic clusters in the query time frame, and regard them as *candidate events*
    - **A ranking module** that summarizes the routine activities in different regions to filter non-event candidates
    - **An updater** that updates local events in real time as the query window shifts

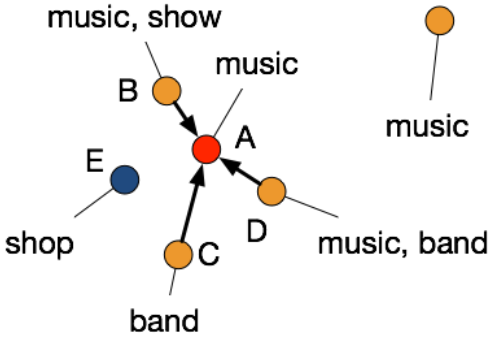
Candidate Event Generation (I) Find Geo-Topic Clusters

- Find geo-topic clusters in the query time frame as candidate events
  - Geo-topic cluster (a group of tweets): geographically close & semantically relevant
- Intuition: the spot where the event occurs is acting as a ***pivot*** that produces relevant tweets around it
- Our clustering algorithm is based on:
  - a geo-topic authority score for each tweet
  - an authority ascent process to find authority maxima as pivots
- Computing geo-topic authority
  - Geographical impact: calculated by a kernel function (*d* and *d'* are tweets)

$$G(d' \to d) = K(\|l_d - l_{d'}\|/h)$$

  - Semantic impact: calculated by random walk on a keyword co-occurrence graph

$$S(d' \to d) = \frac{1}{mn} \sum_{e \in E_d} \sum_{e' \in E_{d'}} r_{e' \to e}$$

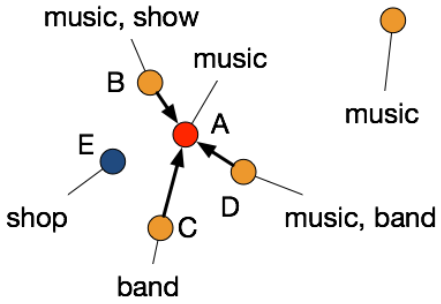


Candidate Event Generation (II) Pivot & Authority Ascent

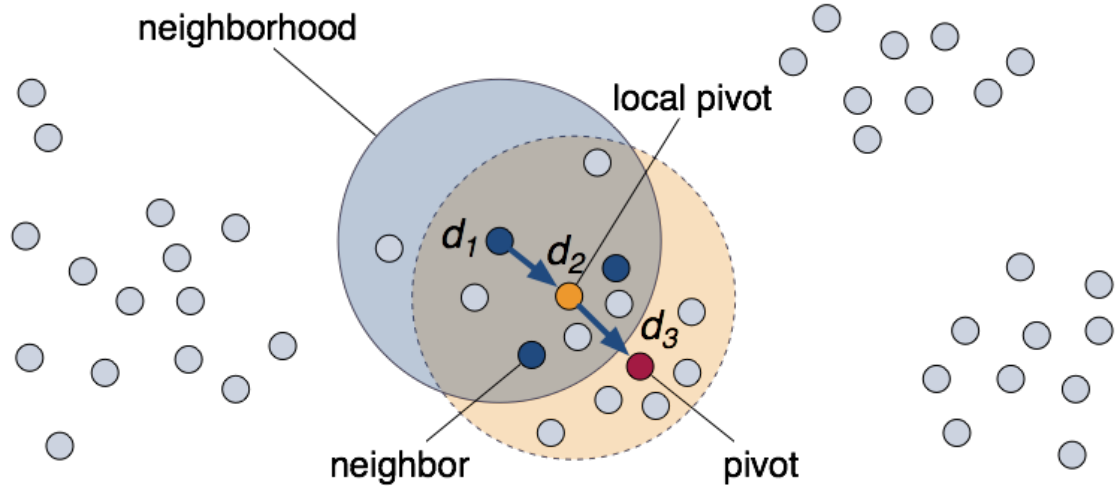
- Geo-topic authority: A tweet gets **an authority score** from neighbor tweets where
  - The geographical impact is captured by kernel function
  - The semantic impact is captured by random walk on the keyword co-occurrence graph
- A **pivot** is an authority maximum: a prominent tweet that is surrounded by many relevant tweets
- **A pivot attracts similar tweets** to form geo-topic clusters
- Find all the pivots in the geo-topic space by Authority Ascent

$$A(d) = \sum_{d' \in N(d)} G(d' \to d) \cdot S(d' \to d)$$

authority; geo-impact; semantic impact

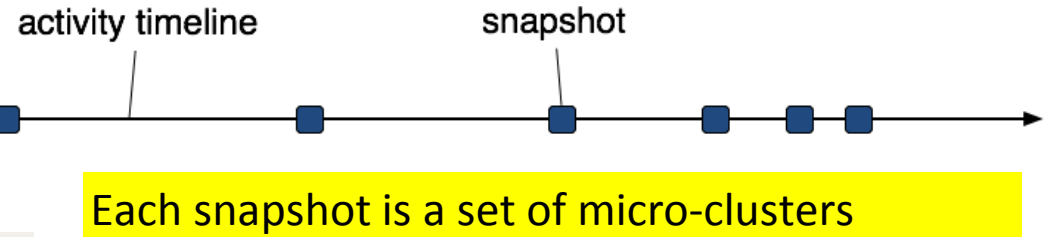


Authority can be interpreted as the total amount of energy received from the neighbors



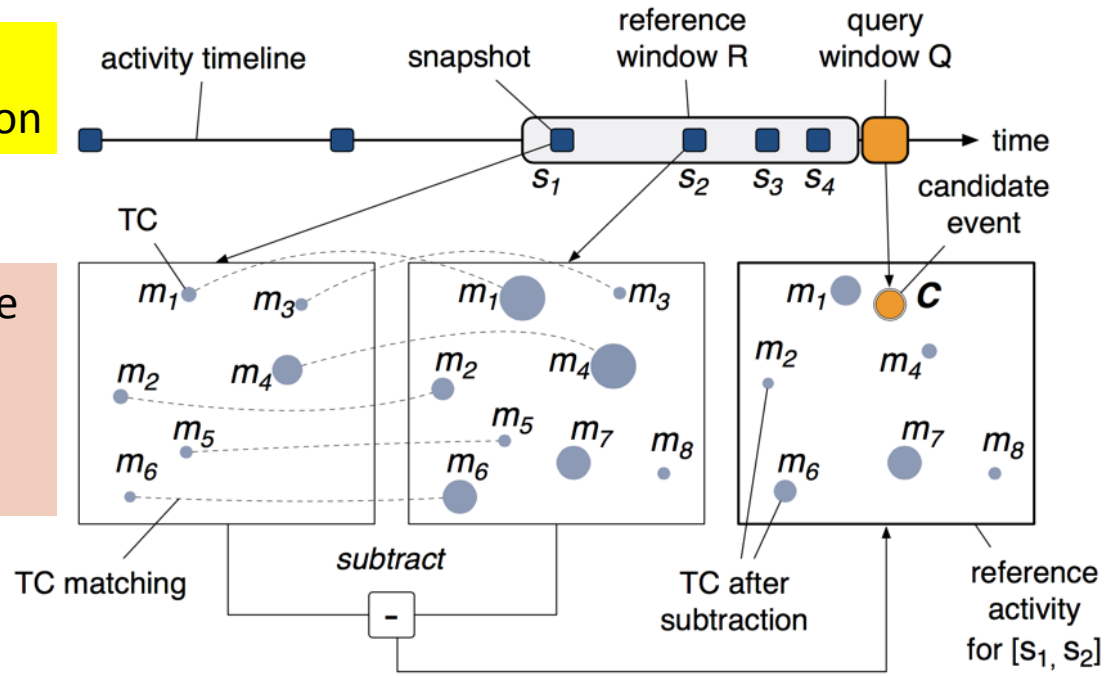
The Ranking Module

- We design the **activity timeline structure** to summarize the activities in different spatial regions and time periods
- The summaries in the activity timeline serve as background knowledge to quantify the spatiotemporal burstiness of candidates



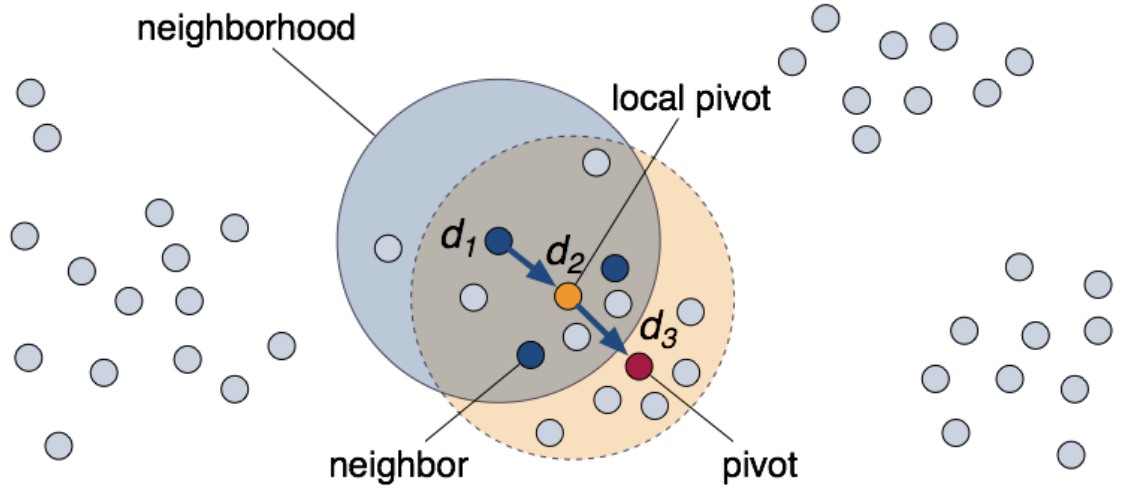
Each snapshot is a set of micro-clusters
Each cluster is an activity summary for a region

Retrieve the snapshots in a reference window as background knowledge
Compute z-score for each candidate as its ranking score



The Update Module

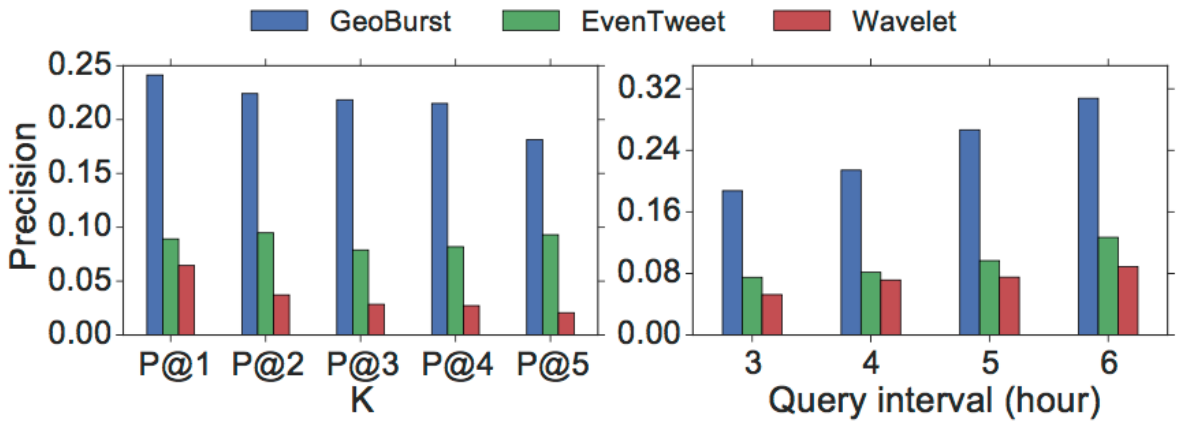
- In the entire process of GeoBurst, the most time-consuming step is pivot finding
- How to avoid finding pivots from scratch as the query window shifts?
  - The key is to maintain the local pivot for each tweet
- We design an updating strategy based on the additive property of authority score:
  - subtracting the contributions of outdated tweets
  - emphasizing the contributions of new tweets



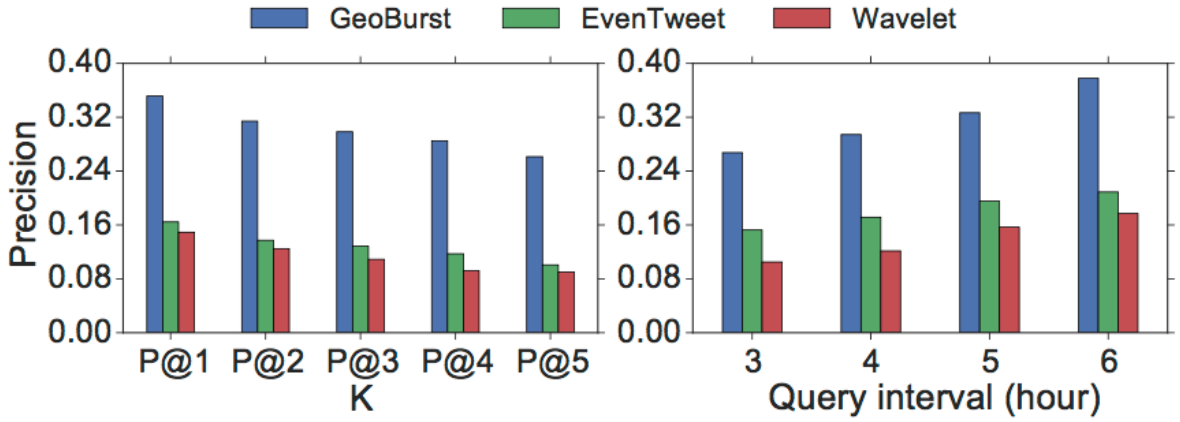
Experiments (Algorithm Comparison)

- Data:
  - NYC: 9M geo-tagged tweets in New York during 3 months
  - LA: 8M geo-tagged tweets in Los Angeles during 3 months
- Task: 80 queries with different durations (3h, 4h, 5h, 6h), find top-5 local events in each query window
- Compared Method: EvenTweet (PVLDB'13), Wavelet (CIKM'09)
- Evaluation: The crowdsourcing platform CrowdFlower
  - Ask the workers to judge whether the result is a local event

Precision Comparison



(a) Precision comparison (NY).



(b) Precision comparison (LA).

Experiments (Illustrative Cases)

	GeoBurst	Is event?
# 1	1. **Festival of Light**! **#nyfol** (@ **The Archway** under the **Manhattan Bridge** in **Brooklyn, NY**) 2. **#Lasers and beats** under the **Manhattan Bridge**! **#NewYorkFestivalofLight #NYFOL @ DUMBO** 3. **New York Festival of Lights #nyfol #dumbo @ DUMBO, Brooklyn**	**Yes**
# 2	1. **Knicks** vs. **Nets** at **Barclays Center**. @ Barclays Center http://t.co/PILk1xK3Tn 2. **Brooklyn** go hard @ **Barclays Center** http://t.co/iVUsJJ5TNG 3. Let's go **Knicks**! #NETS1107 (@ **Barclays Center** - @brooklynnets for @nyknicks vs @BrooklynNets)	**Yes**
# 3	1. **#Thai Restaurant** #spicythaifood (@ 104 **2nd Avenue** in **New York, NY**) 2. The **ASIAN DISHES** here are always my favorite. @ **Ugly Kitchen** 3. **Dinner time** with my family. Suuuuper Nice **Indian RESTAURANT**! @ Malai Marke **Indian Cuisine**.	No

	EvenTweet	Is event?
# 1	1. I practiced... Almost time for **Amy Schumer**. **Jennifer** (@ **Carnegie Hall**) https://t.co/HfqfTLmK2y 2. **2014 Gold Glove Awards Ceremony** with **Hall of Famers, All-Stars Jay Leno @ The Plaza Hotel** 3. My best attempt at a **selfie** with **Hugh Jackman** after **The River at CITS @ The River on Broadway**	No
# 2	1. **Knicks** vs. **Nets** at **Barclays Center**. @ **Barclays Center** http://t.co/PILk1xK3Tn 2. **Budweiser** brings everyone together #family #nonewfriends @ **Alchemy Tavern, Brooklyn** 3. **#Knicks** vs **#nets** with my best gal. @ **Barclays Center Brooklyn** http://t.co/eXXMUKxpIs	**Yes**
# 3	1. **#katespade** @ **Kate Spade / Jack Spade HQ** http://t.co/g6jiFwyc4M 2. Inspiring **keynote** by **Twitter CEO, Dick Costolo** @GirlsWhoCode Gala. http://t.co/yEGh803CuT 3. I wonder if **Jake** from **Statefarm** covers **Jumanji**?	No

Outline

- Introduction
- Mining Geospatial Patterns
- Mining and Aggregating Patterns over Multiple Trajectories
- Mining Semantic-Rich Movement Patterns
- Mining Periodic Movement Patterns
- GeoTopic Discovery in Social Media Data
- Latent Periodic Topic Discovery
- Real-Time Local Event Detection from Geo-Tagged Social Media
- Summary 

Summary

- Emerging: Integrated mining spatiotemporal and social media data
  - Mining Geospatial Patterns
  - Mining and Aggregating Patterns over Multiple Trajectories
  - Mining Semantic-Rich Movement Patterns
  - Mining Periodic Movement Patterns
  - GeoTopic Discovery in Social Media Data
  - Latent Periodic Topic Discovery
  - Real-Time Local Event Detection from Geo-Tagged Social Media
- Integrated data mining with spatiotemporal, social and trajectory data
  - Integrated mining with four dimensions: Spatial + Temporal + Text + Network

References (I)

- F. Giannotti, M. Nanni, F. Pinelli, D. Pedreschi, Trajectory Pattern Mining, KDD'07
- Y. Huang, S. Shekhar, H. Xiong, Discovering colocation patterns from spatial data sets: A general approach, IEEE Trans. on Knowledge & Data Eng., 16(12), 2004
- K. Koperski, J. Han, "Discovery of Spatial Association Rules in Geographic Information Databases", SSD'95
- J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory Clustering: A Partition-and-Group Framework", SIGMOD'07
- Z. Li, B. Ding, J. Han, R. Kays, "Swarm: Mining Relaxed Temporal Moving Object Clusters", VLDB'10
- Z. Li, B. Ding, J. Han, R. Kays, P. Nye, "Mining Periodic Behaviors for Moving Objects", KDD'10
- Z. Li, J. Wang and J. Han, "ePeriodicity: Mining Event Periodicity from Incomplete Observations", IEEE TKDE, 27(5): 1219-1232, 2015
- Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical Topic Discovery and Comparison", WWW'11
- Z. Yin, L. Cao, J. Han, J. Luo, and T. S. Huang, "Diversified Trajectory Pattern Ranking in Geo-tagged Social Media", SDM'11
- C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta, "Splitter: Mining Fine Grained Sequential Patterns in Semantic Trajectories", VLDB'14
- C. Zhang, G. Zhou, Q. Yuan, H. Zhuang, Y. Zheng, L. Kaplan, S. Wang, J. Han, "GeoBurst: Real-time Local Event Detection in Geo-Tagged Tweet Streams", SIGIR'16

Diversified Trajectory Pattern Ranking

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

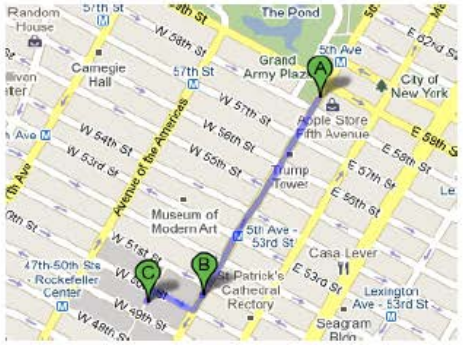
Input: A collection of geo-tagged photos (user, date time, GPS location)

- Our Framework
  - Extract trajectory patterns from the photo collection
  - Rank the trajectory patterns by estimating their importance according to user, location and trajectory pattern relations
  - Diversify the ranking result to identify the representative trajectory patterns from all the candidates

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

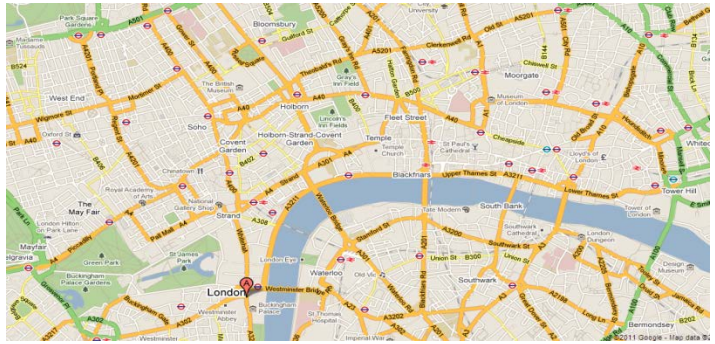
Output: Diversified trajectory pattern ranking result

Ex.: Top ranked trajectories in London, New York and Paris



Data Preprocessing and Pattern Discovery

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



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

- Form sequences
- PrefixSpan [*Pei* et al. TKDE 2004]
  - Ex. (min-support = 2)

ID	User	Date	Sequence
1	Alice	04/26/11	londoneye -> bigben -> downingstreet -> trafalgarsquare
2	Alice	04/27/11	londoneye -> tatemodern -> towerbridge
3	Bob	04/26/11	londoneye -> bigben -> tatemodern
…	…	…	…

- Three frequent sequential patterns:
  - londoneye → bigben
  - londoneye → bigben → trafalgarsquare
  - londoneye → tatemodern

ID	Travel sequence
1	londoneye → bigben → trafalgarsquare
2	londoneye → bigben → downingstreet → trafalgarsquare
3	londoneye → bigben → westminster
4	londoneye → tatemodern → towerbridge
5	londoneye → bigben → tatemodern

Rank Trajectory Patterns

- The *top frequent* trajectory patterns are short but not informative, e.g.,
- How to rank trajectory patterns?
  - A trajectory pattern is important if many important users take it and it contains important locations

$$P_T = M_{TU} \cdot P_U \quad P_T = M_{LT}^T \cdot P_L$$

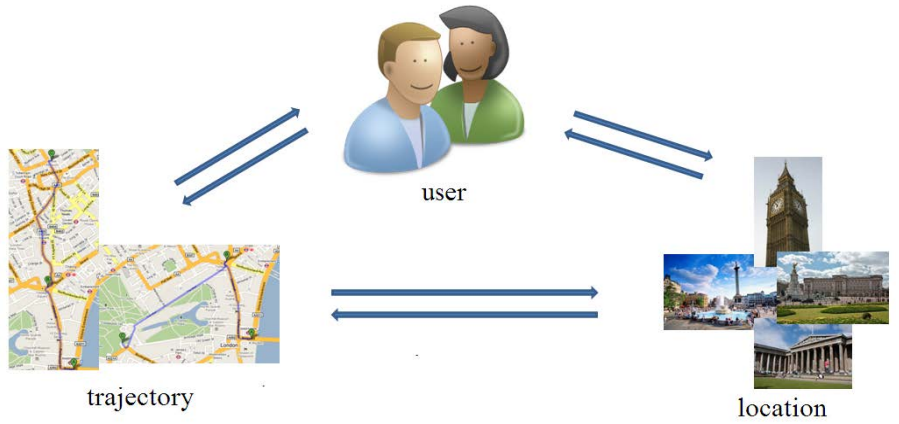
  - A user is important if the user takes photos at important locations and visits the important trajectory patterns

$$P_U = M_{UL} \cdot P_L \quad P_U = M_{TU}^T \cdot P_T$$

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

$$P_L = M_{LT} \cdot P_T \quad P_L = M_{UL}^T \cdot P_U$$

Trajectory pattern	Frequency
londoneye → bigben	21
bigben → londoneye	19
londoneye → tatemodern	18
londoneye → royalfestivalhall	15
londoneye → trafalgarsquare	14
londoneye → waterloobridge	12
towerbridge → cityhall	12
royalfestivalhall → londoneye	11
tatemodern → londoneye	11
bigben → parliamentsquare	10



Trajectory Pattern Ranking Algorithm

- P_T is the eigen-vector for M^TM 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^TM for the largest eigen-value if the intial P_T is not orthogonal to it
- Based on the algorithm, we can derive the top-trajectory in London

londoneye → bigben → downingstreet → horseguards → trafalgarsquare

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.

Top-Ranked Trajectories Are often Highly Biased to only a few Locations

- Top-Ranked Locations in London
  - P_L : the importance score for location L
  - # user: #users visited the location

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

- Top-Ranked Trajectories in London
  - highly biased to only a few locations
    - Trajectory 1 (londoneye → bigben → downingstreet → horseguards → trafalgarsquare)
    - Trajectory 5 (westminster → bigben → downingstreet → horseguards → trafalgarsquare)

Rank	Trajectory pattern
1	londoneye → bigben → downingstreet → horseguards → trafalgarsquare
2	londoneye → bigben → tatemodern
3	tatemodern → bigben → londoneye
4	londoneye → bigben → parliamentsquare → westminster
5	westminster → bigben → downingstreet → horseguards → trafalgarsquare
6	royalfestivalhall → londoneye → bigben
7	londoneye → royalfestivalhall → tatemodern
8	tatemodern → londoneye → royalfestivalhall
9	londoneye → tatemodern → towerbridge
10	londoneye → towerbridge → tatemodern

From Top-Ranked to Diversified Ranked Trajectories

- Diversified Ranked Trajectories in London
  - Trajectories 2, 4, & 5 are popular routes to explore street arts in London

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

- Location Recommendation in London
  - Rank the locations by the scores of trajectories (append current trajectory with next destination)

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
lodoneye→trafalgarsquare	buckinghampalace