KDD2016

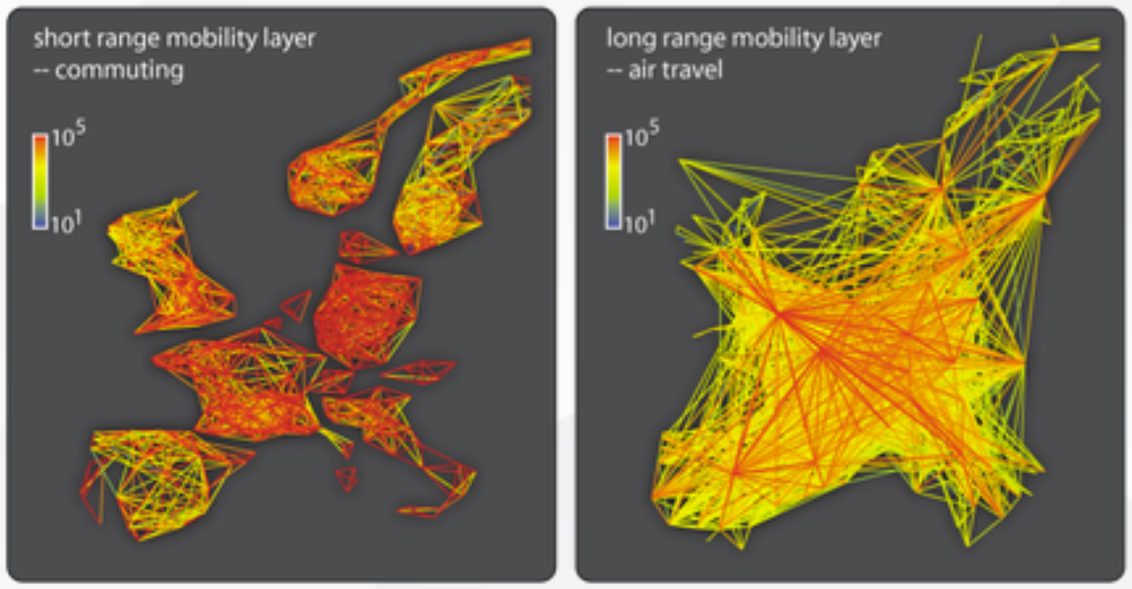
GMove: Group-Level Mobility Modeling Using Geo-Tagged Social Media

Chao Zhang, Keyang Zhang, Quan Yuan, Luming Zhang,
Tim Hanratty, and Jiawei Han

Presenter: Tongtong Liu

Background

- Mobility modeling aims at understanding human movement regularity.
- It is important to many applications:
  - Urban planning
  - Traffic scheduling
  - Location prediction
  - Activity recommendation
  - ……



Background

- Previous studies mostly use **GPS trace data** to model human mobility.
- The recent prevalence of geo-tagged social media (**GeoSM**) brings new opportunities to this task:
  - In addition to spatial and temporal information, each GeoSM record (e.g., tweet, Facebook post) also has text.
  - The GeoSM data has a much larger size and a much better coverage of the population than GPS trace data.

Our Goal

- We aim to unveil human movement regularity using large-scale GeoSM data.
- Specifically, we answer the following two questions:

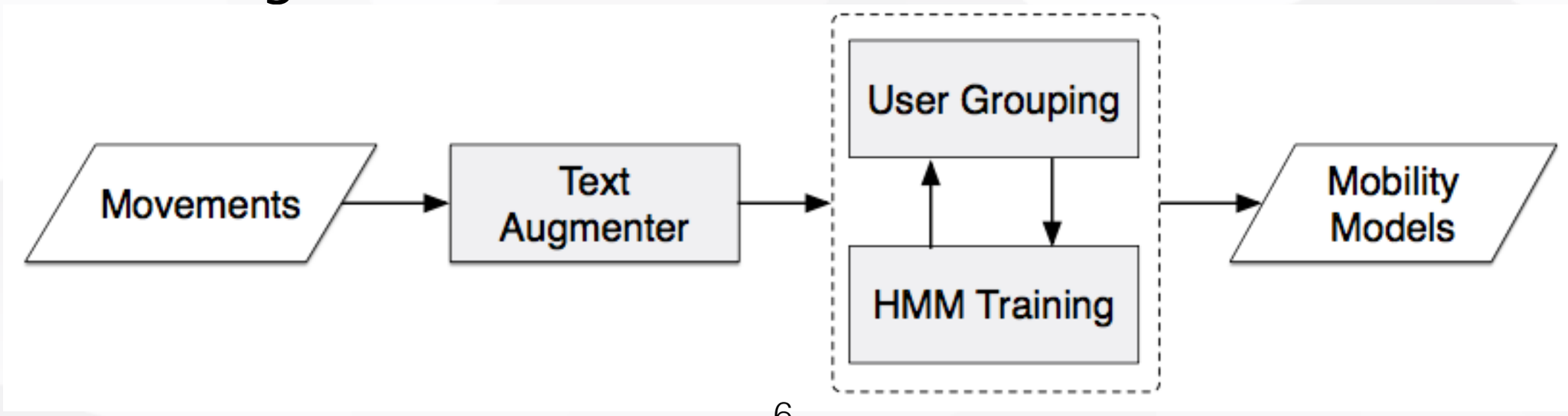
1. What are the intrinsic states underlying people's movements?
   - Here, a state should provide a 3W (where-what-when) view regarding the user's activity.
2. How do people move sequentially between those latent states?

Challenges

- GeoSM (e.g., tweets) have very short text, making it hard to model the semantics of human activities.
- Dilemma for mobility modeling using GeoSM data:
  - Each user typically has limited GeoSM records, learning a model for every user suffers from severe **data sparsity**.
  - Different users have totally different moving behaviors, learning one model for all the users suffers from **data inconsistency**.

Method Overview

- Relying on Hidden Markov Model, we propose an effective method named GMove
- Two key modules of GMove:
  - **Text augmenter:** reduces text sparsity using spatiotemporal signals
  - **HMM ensemble learner:** performs group-level HMM learning

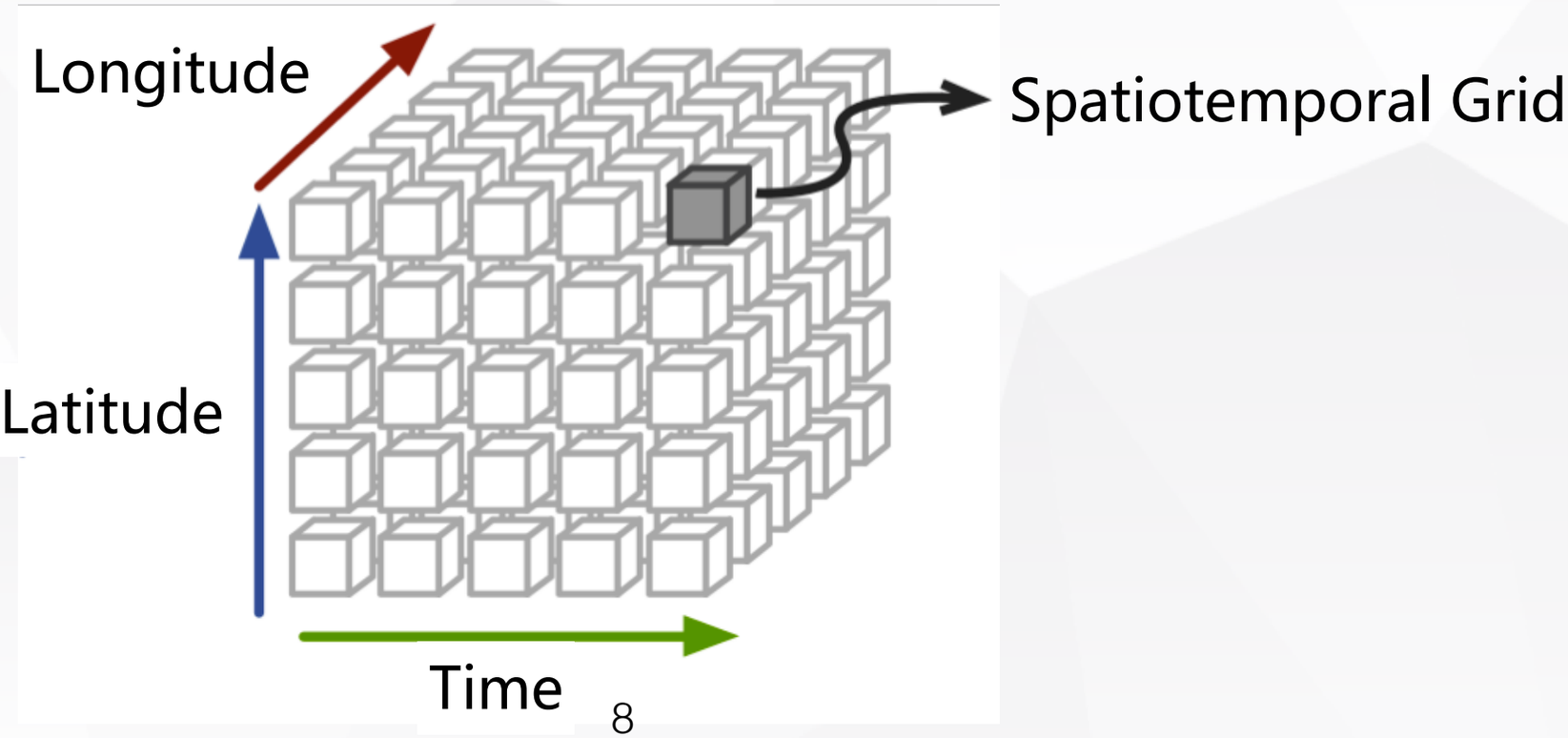


Module 1: Text Augmenter

- Why do we need the text augmenter?
  - The raw text messages are too short.
  - The spatiotemporal distributions of different words can unveil their semantical correlation.
- E.g., consider two users watching the Guoan' game at the Workers' Stadium.
  - They may post two tweets using two different keywords: "Guoan" and "Workers' Stadium".
  - Although those two keywords do not co-occur in the same tweet, they are spatially and temporally close, and thus correlated.

Text Augmentation

- We discretize the space D into equal-size grids.
  - For each keyword **k**, we use its spatiotemporal distribution over the 3-D cube to obtain a vector $\mathbf{V}_k$.
  - Given two keywords, we compute their correlation as the cosine distance between their vectors.



Grid Density & Signature

DEFINITION 2 (GRID DENSITY). *Given a keyword w, the density of w in grid $\langle n_x, n_y, n_t \rangle$ $(1 \leq n_x \leq N_x, 1 \leq n_y \leq N_y, 1 \leq n_t \leq N_t)$ is defined as w's frequency in that grid, namely*

$$d_w(n_x, n_y, n_t) = \frac{c_w(n_x, n_y, n_t)}{\sum_{n_x, n_y, n_t} c_w(n_x, n_y, n_t)},$$

where $c_w(n_x, n_y, n_t)$ is the number of GeoSM records that contain w and meanwhile fall in grid $\langle n_x, n_y, n_t \rangle$.

DEFINITION 3 (SIGNATURE). *Given a keyword w, its signature s_w is a $N_x N_y N_t$-dimensional vector, where $d_w(n_x, n_y, n_t)$ is the value for the $((n_t-1)N_x N_y + (n_y-1)N_x + n_x)$-th dimension.*

Augmentation Algorithm

- Once keyword correlations are computed, we perform weighted sampling to augment raw messages.

$$\mathcal{N}_w = \{v | v \in V \wedge corr(w, v) \geq \delta\}$$

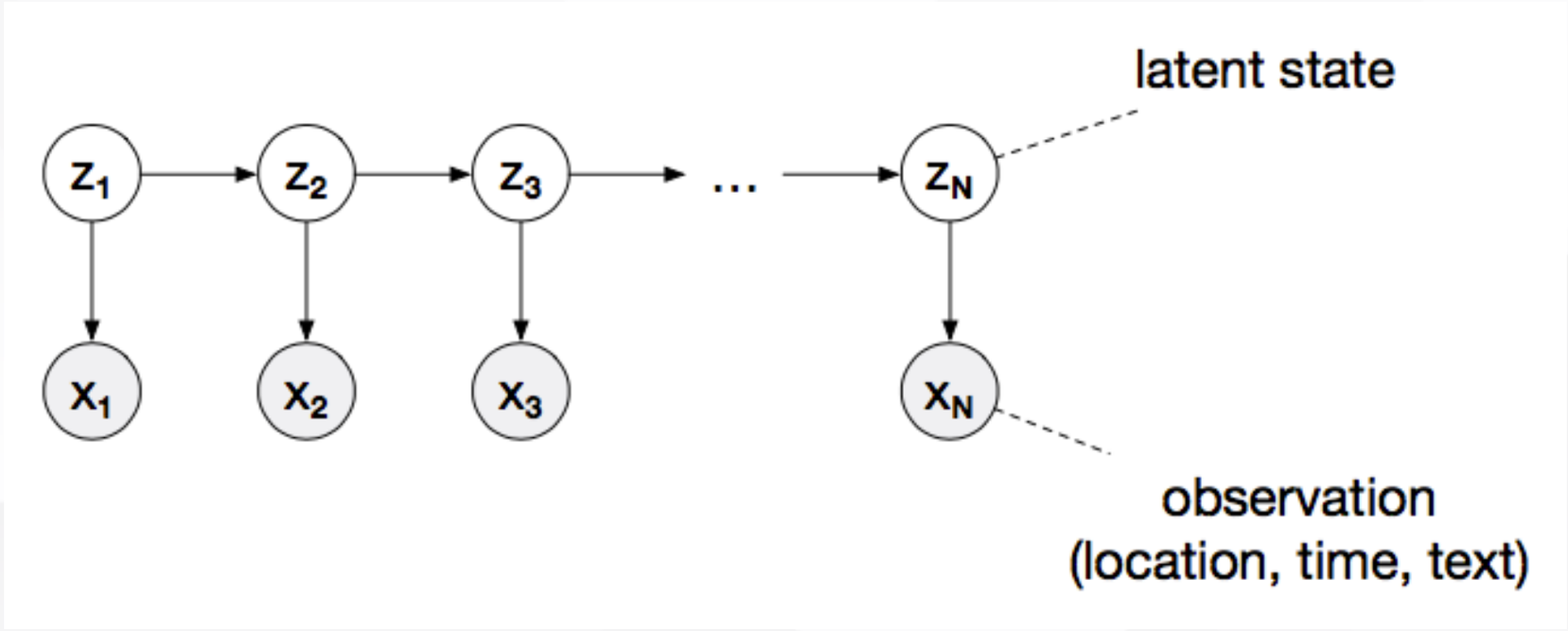
Algorithm 1: Text augmentation.

Input: A GSM record x, the target length L.
Output: The augmented text message of x.

$A_x \leftarrow$ The original text message e_x;
while $len(A_x) < L$ **do**
  Sample a word $w \in e_x$ with probability $\frac{TF-IDF(w)}{\sum_{v \in e_x} TF-IDF(v)}$;
  Sample a word $v \in \mathcal{N}_w$ with probability $\frac{corr(w,v)}{\sum_{u \in \mathcal{N}_w} corr(w,u)}$;
  Add v into A_x;
return A_x;

Module 2: HMM Ensemble Learner

- Hidden Markov Model (HMM) for mobility modeling:
  - It assumes multiple latent states (e.g., working at office) that govern a user’s movements.
  - The sequence of the latent states follows Markov process.



Note that, as the raw trajectory of each user is sparse, we impose a time constraint (e.g., three hours) to extract dense subsequences for model training.

Module 2: HMM Ensemble Learner

- We group like-behaved users (e.g., Stanford students) and train an HMM for each group:
  - Reduce data sparsity by aggregating the movements of multiple users.
  - Not compromising data consistency because the users in the same group share significant movement regularity.

	Data Sparsity	Data Consistency
Individual Level	**X**	O
Group Level	O	O
Global Level	O	**X**

Each user has limited records.

Different moving patterns are mixed.

How to Obtain Quality User Groups

- User grouping and mobility modeling mutually enhance each other:
  - Better user grouping leads to more consistent movement data within each group, which can improve the quality of the HMMs.
  - Better HMMs can better reveal movement regularities, which helps infer the group a user belongs to.

An Iterative Process

- GMove alternates between user grouping and HMM training.

- User Grouping
  - Soft group: membership vector
  - Assume the HMMs of different groups are already learnt
  - For user u, update the membership vector by computing the posterior probability that u belongs to group g
- HMM Training
  - Assume the group memberships of different users are fixed
  - Learn One HMM for each group g

Iterative Refinement Framework

1. **Initialization:** Let $U = \{u_1, u_2, \ldots, u_M\}$ be the user set, and $\mathcal{G} = \{1, 2, 3, \ldots, G\}$ be the G underlying user groups.

   (a) $\forall u \in U$, randomly generate an initial membership vector M_u *s.t.* $\sum_{g=1}^{G} M_u(g) = 1$, where $M_u(g)$ denotes the probability that user u belongs to group g.

   (b) $\forall g \in \mathcal{G}$, randomly initialize an HMM H_g. The ensemble of the HMMs are denoted as $\mathcal{H} = \{H_g | g \in \mathcal{G}\}$.

2. **HMM Training:** $\forall g \in \mathcal{G}$, use the membership vectors to reweigh all input trajectories, such that the weights of user u's trajectories are set to $M_u(g)$. Then refine every H_g to generate a new HMM ensemble, $\mathcal{H}^{\text{new}} = \{H_g^{\text{new}} | g \in \mathcal{G}\}$.

3. **User Grouping:** $\forall u \in U$, use $\mathcal{H}^{\text{new}}$ to update u's membership vector such that the g-th dimension is the posterior probability that user u belongs to group g, namely $M_u^{\text{new}}(g) = p(g|u; \mathcal{H}^{\text{new}})$.

4. **Iterating:** Check for convergence using the log-likelihood of the input trajectories. If the convergence criterion is not met, then let

$$\forall g, H_g \leftarrow H_g^{\text{new}}; \ \forall u, M_u \leftarrow M_u^{\text{new}};$$

   and return to Step 2.

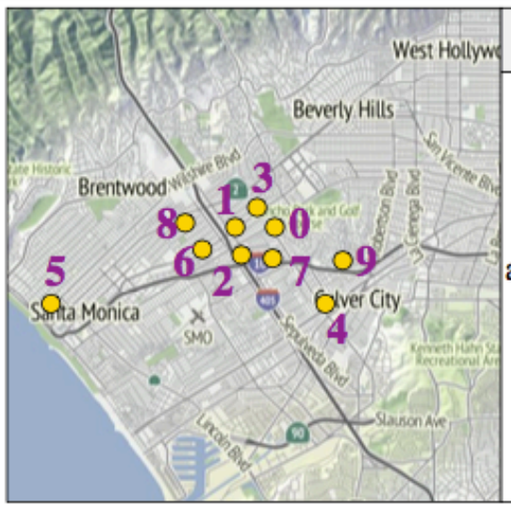
Experiments

- Data Sets
  - LA: ~0.6 million geo-tagged tweets published in Los Angeles.
  - NY: ~0.7 million geo-tagged tweets published in New York.

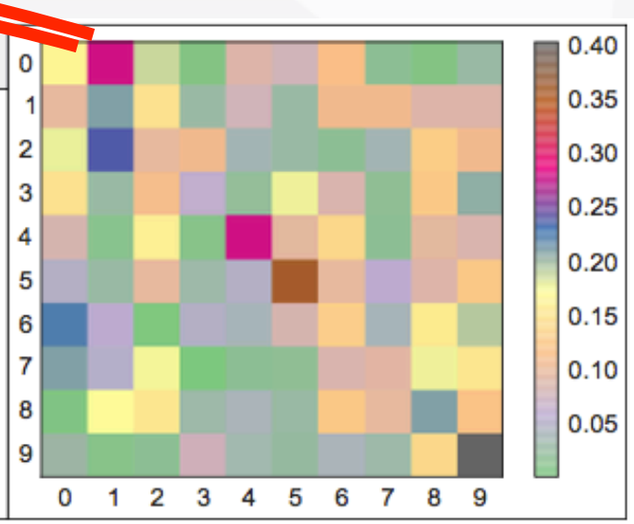
Case Study: Text Augmentation

Data	Raw tweet message	Augmented message
LA	Y'all just kobe fans not lakers. Let's go lakers!!	fans(11), game(7), kobe(19), jeremy(6), lakers(26), injury(8), staples(8), center(4), nba(9), bryant(12)
	Fun night! @ Universal Studios Hollywood http://t.co/wMibfyleTW	fun(4), universal(20), studio(16), hollywood(18), night(5), party(7), fame(6), people(13), play(11)
NY	Nothing better...fresh off the oven! #Italian #bakery #pizza	fresh(7), oven(21), italian(19), bakery(12), pizza(14), bread(6), cook(5), food(12), kitchen(4)
	My trip starts now! @ JFK Airport	jfk(24), international(5), trip(9), travel(6), john(13), kennedy(14), terminal(8), start(6), now(3), airport(12)

Group-Level Mobility Models

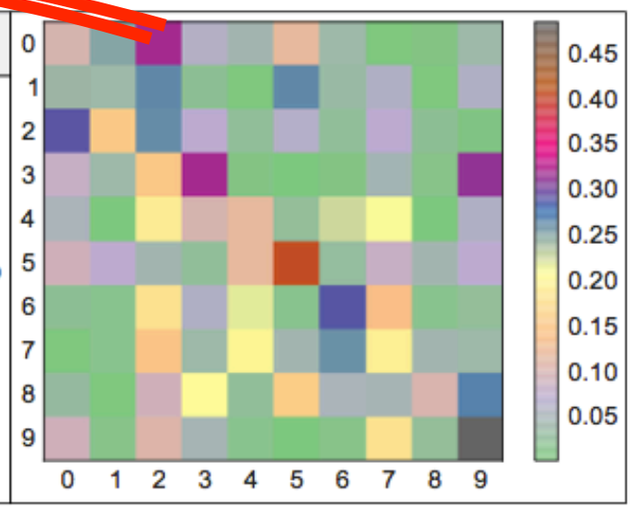


State 0	State 1	State 2	State 3	State 4	State 5	State 6	State 7	State 8	State 9
home	school	food	friend	buy	santa	wait	suck	home	night
morning	ucla	restaurant	hahaha	shop	monica	dinner	gym	rest	music
coffee	course	dinner	dream	mall	beach	food	football	sleep	tonight
dude	midterm	grill	song	beautiful	sunset	worst	team	live	bar
apartment	class	love	child	happy	dinner	grill	game	game	tbh
day	food	deli	birthday	mine	baby	restaurant	play	drink	nightlife
hall	drink	worst	drink	apparel	venice	delicious	buster	bed	evening
run	damn	come	kiss	bag	shore	amazing	let	show	movie
house	friend	waiter	bday	clothes	life	gather	beat	time	beer
sleep	stay	chipotle	snapchat	day	please	love	boy	miss	like



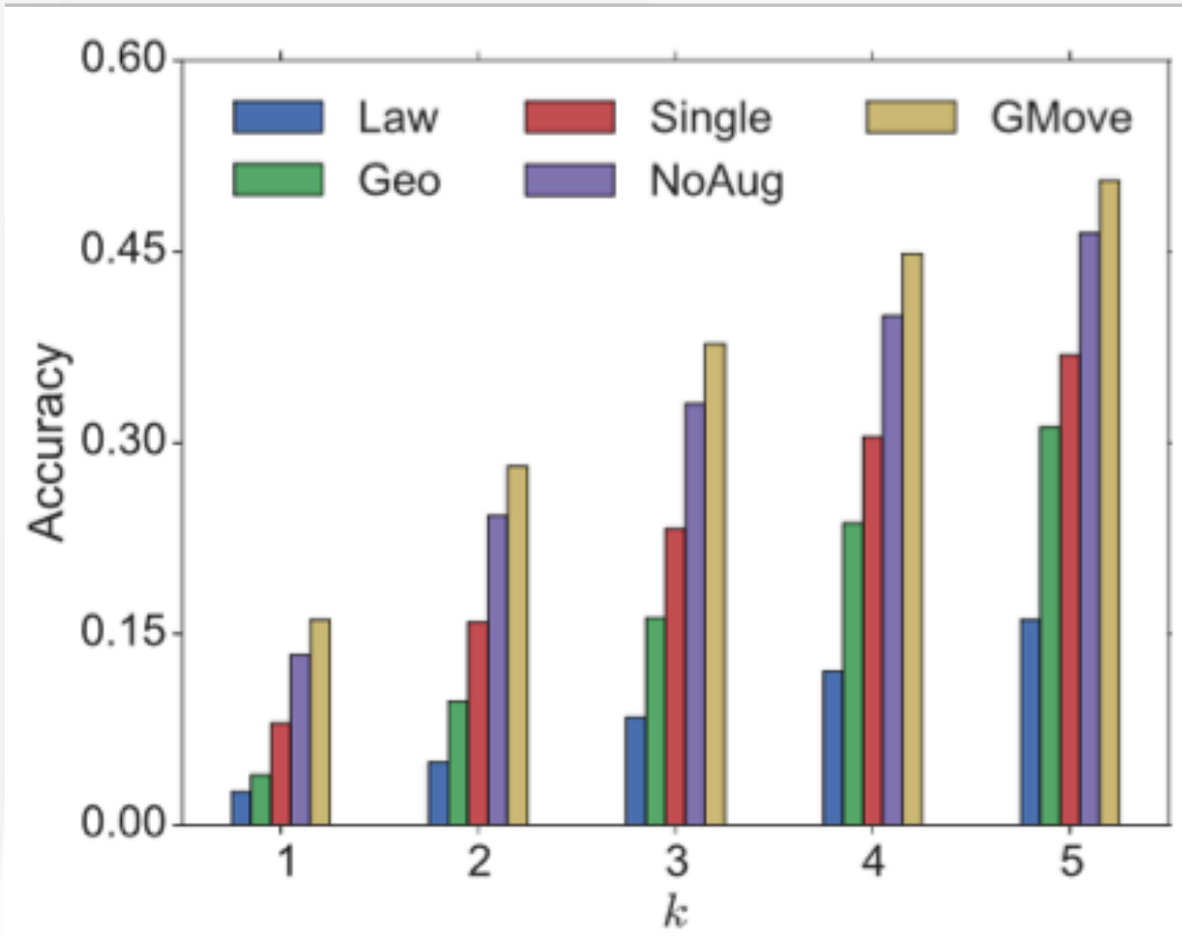
(a) The mobility model for the first user group (students).

State 0	State 1	State 2	State 3	State 4	State 5	State 6	State 7	State 8	State 9
hotel	restaurant	lax	universal	food	venice	shopping	night	food	club
sleep	food	travel	studio	restaurant	beach	buy	hotel	table	night
miss	amazing	airport	night	chinese	people	shop	live	taste	music
building	eat	international	party	mexican	life	clothes	lounge	salad	girl
center	lunch	come	hollywood	spicy	spa	cheap	place	quiet	nightclub
plaza	grill	flylax	buster	hometown	sunshine	sales	taxi	deli	money
flight	american	arrival	friend	taste	beautiful	nice	shower	hot	drink
home	deli	terminal	happy	china	santa	women	sleep	steak	sleep
city	together	train	haloween	town	monica	mall	queen	dining	dance
inn	pancake	depart	club	delicious	food	men	terminal	fame	rock

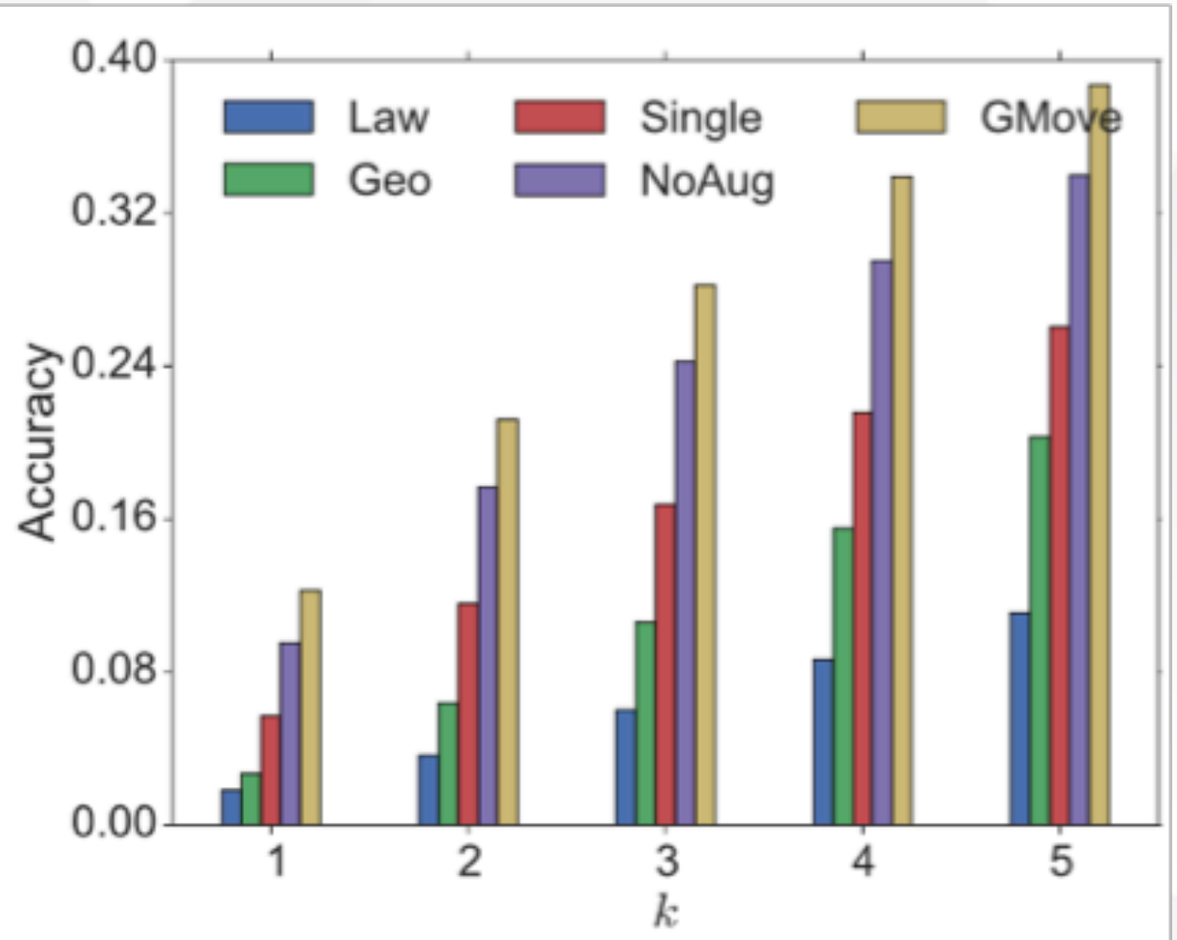


(b) The mobility model for the second user group (tourists).

Quantitative Evaluation: Location Prediction



(a) LA



(b) NY

Summary

- We study the problem of group-level mobility modeling using geo-tagged social media.
- We propose the GMove method:
  - It leverages keyword spatiotemporal correlations to reduce text sparsity.
  - It alternates between user grouping and HMM training to learn group-level models.
- Our experiments show that GMove can effectively retrieve group-level mobility models.