Discovering Geographical Topics in Twitter

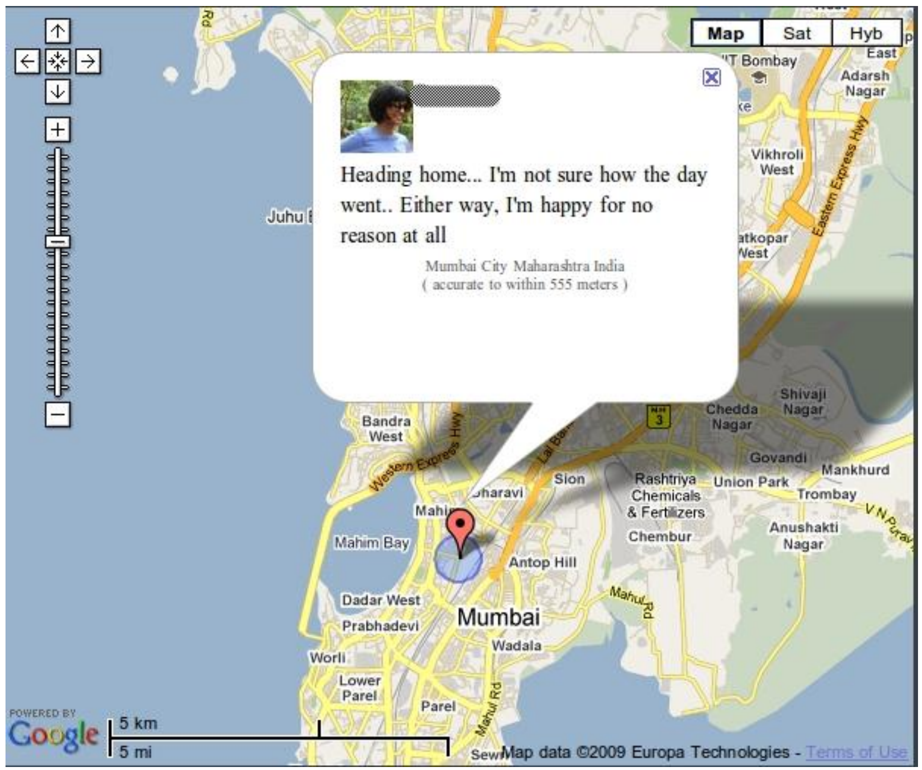
Liangjie Hong, Lehigh University
Amr Ahmed, Yahoo! Research
Alexander J. Smola, Yahoo! Research
Siva Gurumurthy, Twitter
Kostas Tsioutsiouliklis, Twitter

Overview

- Motivations
- Our Proposed Model
- Experiments
- Conclusions

Motivations

Twitter messages + Locations



Motivations

We want to know…

- How is information created and shared in different geographic locations? What is the inherent geographic variability of content?
- What are the spatial and linguistic characteristics of people? How does this vary across regions?
- Can we discover patterns in users' usage of micro-blogging services?

- Can we predict user location from tweets?

Motivations

Applications

Behavioral targeting and user modeling

Better local information filtering

Motivations

Challenges

- Tweets
  - noisy and short (140 characters)
- Only 1% of tweets geo-tagged
  - Can we predict locations for non-tagged tweets?
- Many intuitions to be combined
  - Background, regional language models, topics
  - Personal preferences, regional preferences…

…

Can we really infer locations for a tweet?

Yes via tweet decomposition

Just landed after a long flight. It is raining here at Lyon though!

What is the user's location?

Just landed after a long flight. It is raining here at Lyon though!

background

just
after
It
be
the
can
cant
will

Just landed after a long flight. It is raining here at Lyon though!

Travel	background
landed	just
flight	after
delay	It
TSE	be
Gate	the
terminal	can
	cant
	will

Just landed after a long flight. It is raining here at Lyon though!

Travel/airport	background	SE airport area
landed	just	Lyon
flight	after	Saint
delay	It	Exupery
TSE	be	convention
Gate	the	center
terminal	can	raining
	cant	
	will	

Semantic Topic

Travel/airport

landed
flight
delay
TSE
Gate
terminal

Background Language Model

background

just
after
It
be
the
can
cant
will

Regional Language Model

SE airport area

Lyon
Saint
Exupery
convention
center
raining

Delayed again at the TSE check point and might miss my flight. way to go SF!

Travel/airport	background	SFO
landed	just	SF
flight	after	SFO
delay	It	San
TSE	be	Fransisco
Gate	the	airport
terminal	can	
	cant	
	will	

Can we always do that?

Life is good! Feeling great today!

Life is good! Feeling great today!

Daily life

life
feeling
good
today
morning

background

just
after
It
be
the
can
cant
will

?

Life is good! Feeling great today!

If we know something extra about the context and **user location preferences**, perhaps we can do better than random guessing!

Motivations

Previous work

- Simple regional language models
  - No factorization
- No personal preferences
- Complicated inference algorithms
  - Usually two step process
  - Fails to learn coherent regions

Overview

- Motivations
- Our Proposed Model
- Experiments
- Conclusions

Our Proposed Model

- A novel probabilistic model considers
  - Regional language models
  - Global topics
  - Personal preferences
- Sparse modeling + Bayesian treatment
- An efficient inference algorithm

The Model

- Basic Intuition
  - Regions
  - Topics
  - Users
  - Tweets
- The generative process
  - Intuition
  - Glory details

Basic Intuition: Region

- Must be coherent
  - There is enough traffic in it
  - Affects the way we write tweets
    - Has preference over what topic discussed
    - Specific keywords
  - Area over the map
  - Example
    - An airport
    - A park
    - A mall
    - A city

Basic Intuition: Topic

- Classify the content of the tweet
- Might not tell us the location
- Puts a distribution over words
- Examples
  - Sports
  - Politics
  - Travel
  - Daily life, etc

Basic Intuition: User

- Has preferences over locations
  - Where he usually spends his/her time
- Has preference over topics
  - What he tweets about

Basic Intuition: Tweet

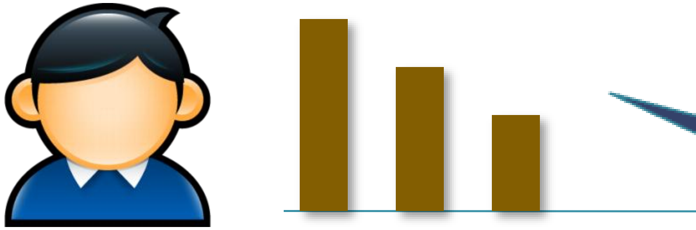
- Written by a given user
- At a specific location (region)
  - Depends on the user
- About a specific topic
  - Depends on
    - What the user talks about
    - What is being discussed at this location
- Composed of a bag of words from
  - Topic + location + background language models

The Model

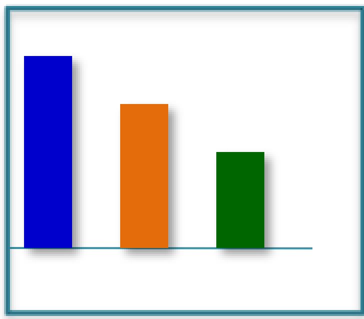
- Basic Intuition
  - Regions
  - Topics
  - Users
  - Tweets
- The generative process
  - Intuitive explanation
  - Glory details

How a tweet is being generated?

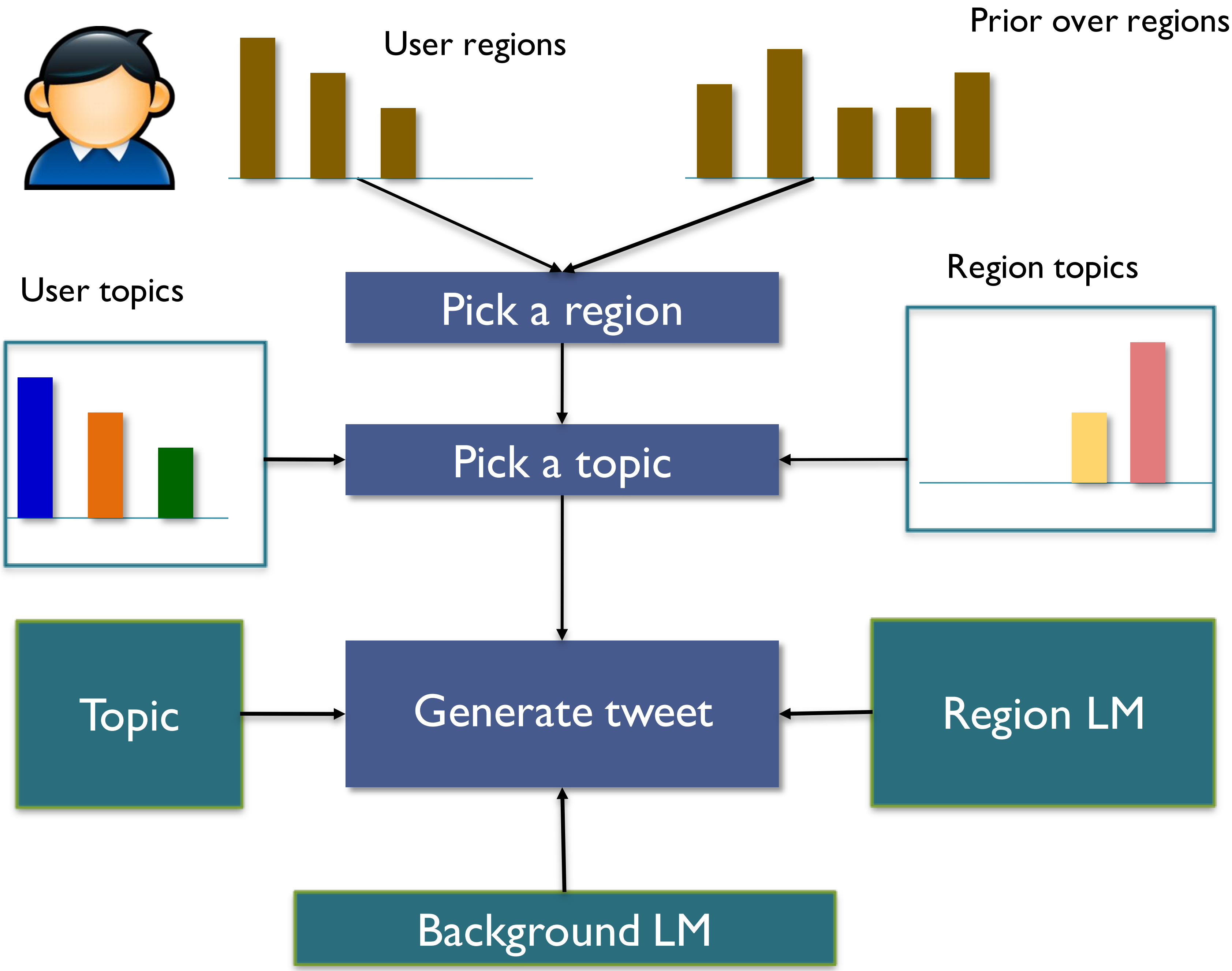
- Pick a location
- Pick a topic
- Generate the words

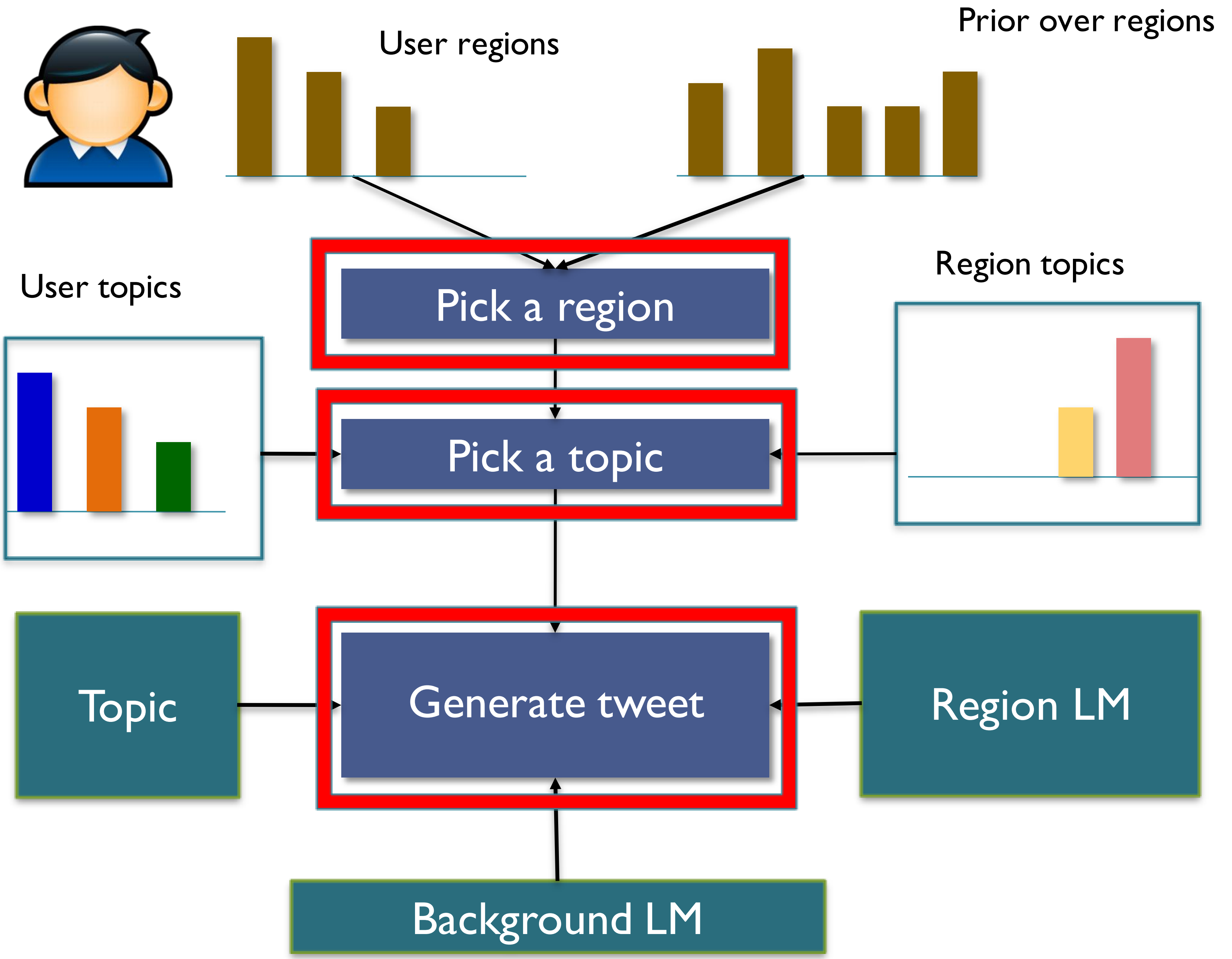


Preferences over regions
Regions are unsupervised
Just an area over the map



Preference over topics:
What he likes to talk about


User regions
Prior over regions
User topics
Pick a region
Region topics
Pick a topic
Topic
Generate tweet
Region LM
Background LM


User regions
Prior over regions
User topics
Pick a region
Region topics
Pick a topic
Topic
Generate tweet
Region LM
Background LM

Discrete Additive Models

- Switch-based models
  - Normalized distributions
  - Pick one distribution
  - Sample from it
- SAGE (Eisenstein, Ahmed, Xing, 2011)
  - Un-normalized distribution
    - Log frequencies
  - Add them all together
  - Exponentiate and sample

SAGE

An Additive model for discrete distributions

- Discrete distribution via natural parameters
  Example:

$$p(v|\phi) = \exp\left(\phi_v - g(\phi)\right) \text{ where } g(\phi) = \log \sum_v \exp\left(\phi_v\right)$$

- Log-frequency differences
- Addition of multiple models
  Example:

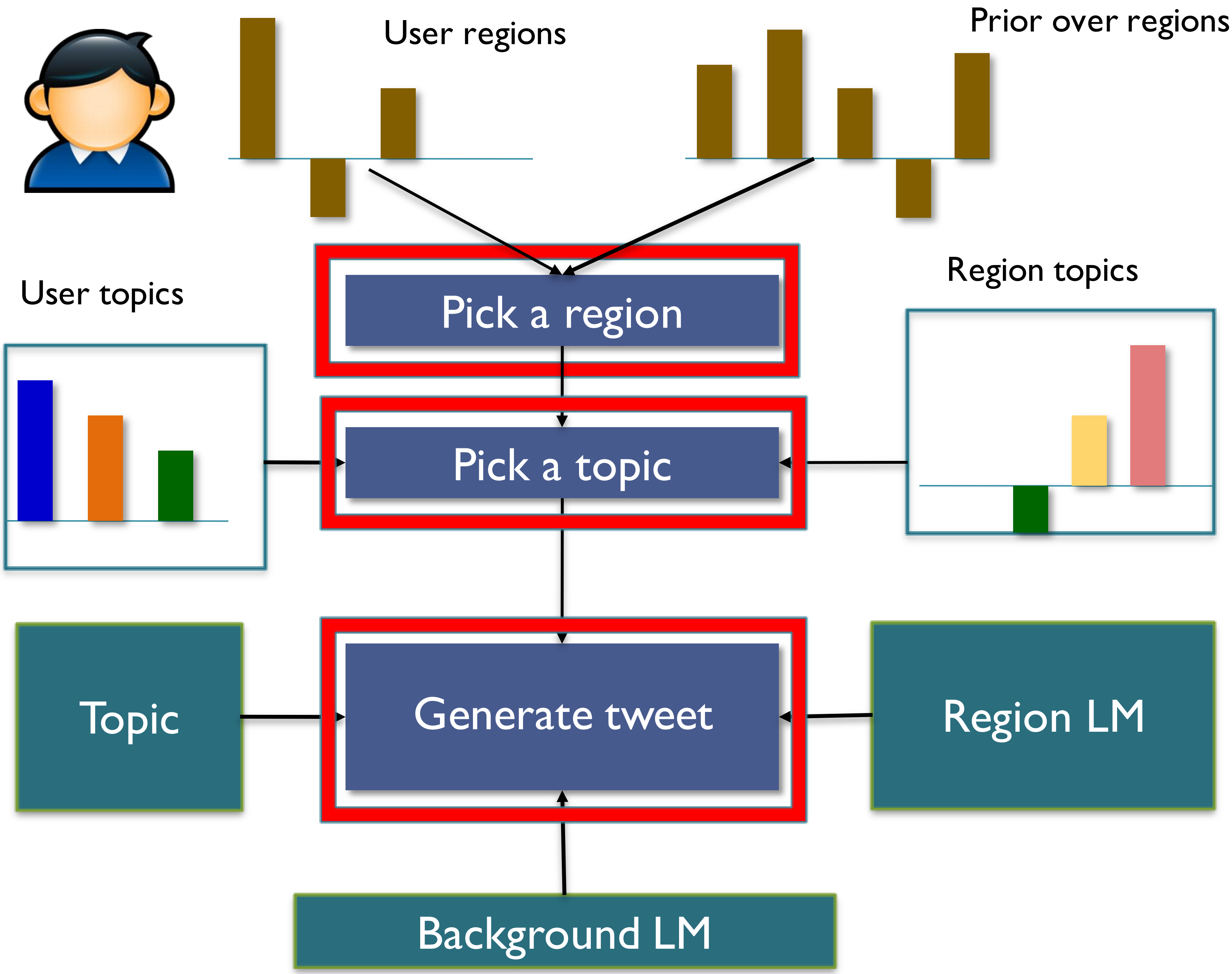
$$P(v|\phi_0, \phi_u, \phi_g) := p(v|\phi_0 + \phi_u + \phi_g)$$

SAGE

Use `SAGE` to replace "switch" variables to enable us incorporate multiple sources in different levels of our model easily

- Language models
  Example: background, regional, global…
- User preferences
  Example: global, regional, personal…

…


User regions
Prior over regions
User topics
Pick a region
Region topics
Pick a topic
Topic
Generate tweet
Region LM
Background LM

The Model

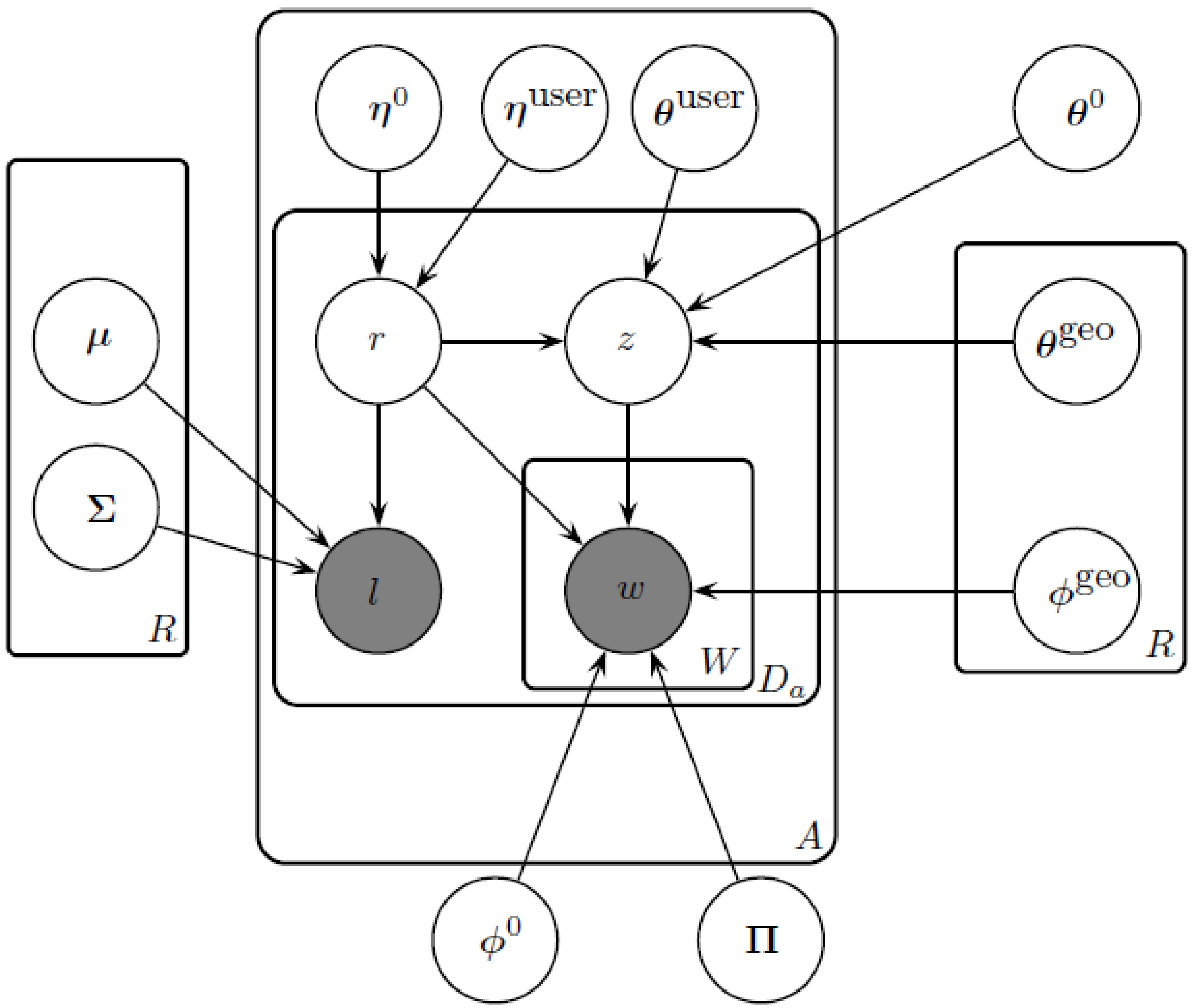
- Basic Intuition
  - Regions
  - Topics
  - Users
  - Tweets
- The generative process
  - Intuitive explanation
  - Glory details

Generative Process

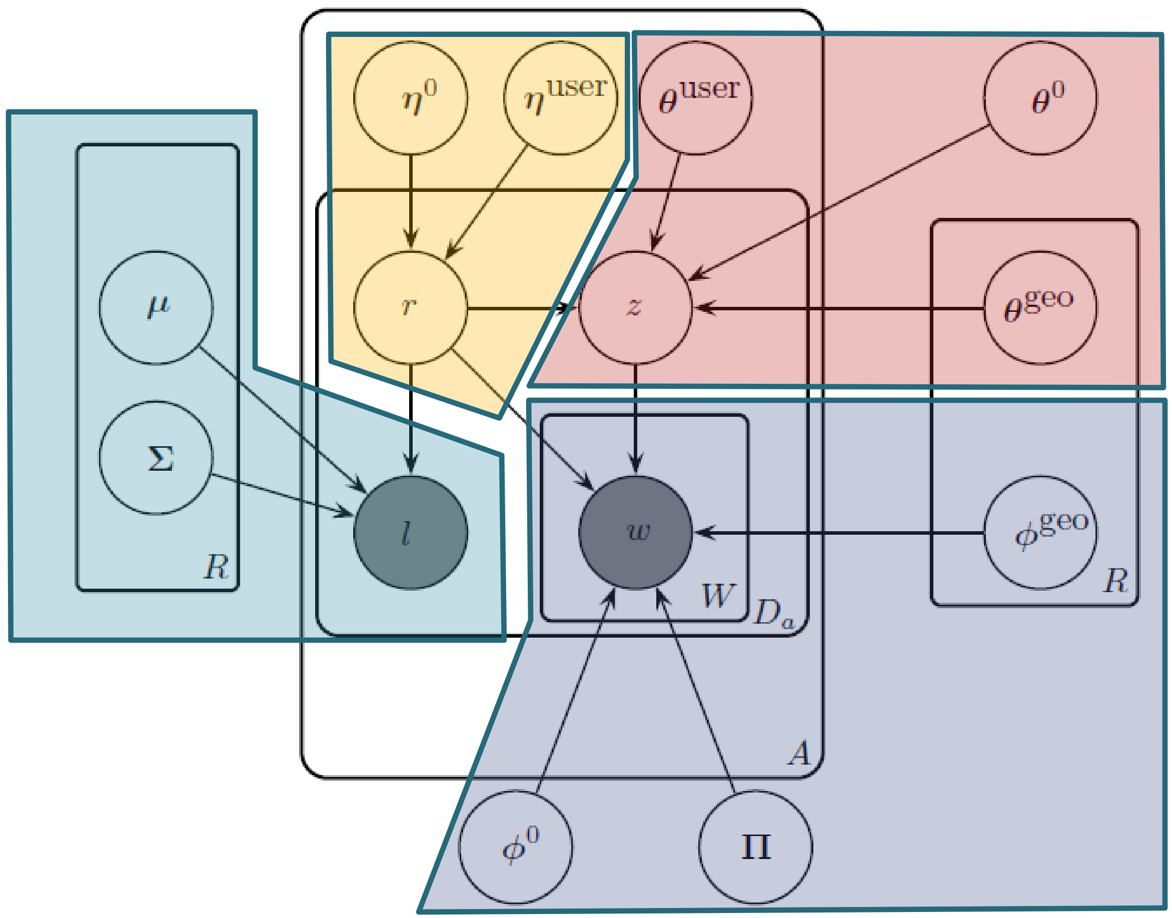
Notations

Symbol	Size	Usage
$\boldsymbol{\eta}^0$	$1 \times \mathbb{R}$	global region distribution
$\boldsymbol{\eta}^{\text{user}}$	$\mathbb{U} \times \mathbb{R}$	user-dependent region distribution
$\boldsymbol{\theta}^0$	$1 \times \mathbb{K}$	global topic distribution
$\boldsymbol{\theta}^{\text{geo}}$	$\mathbb{R} \times \mathbb{K}$	region-dependent topic distribution
$\boldsymbol{\theta}^{\text{user}}$	$\mathbb{U} \times \mathbb{K}$	user-dependent topic distribution
$\boldsymbol{\phi}^0$	$1 \times \mathbb{V}$	global term distribution
$\boldsymbol{\phi}^{\text{geo}}$	$\mathbb{R} \times \mathbb{V}$	region-dependent term distribution
$\boldsymbol{\Pi}$	$\mathbb{K} \times \mathbb{V}$	a global topic matrix
$\boldsymbol{\mu}$	$\mathbb{R}^2$	mean location of a latent region
$\boldsymbol{\Sigma}$	$\mathbb{R}^{2\times 2}$	covariance matrix of a latent region

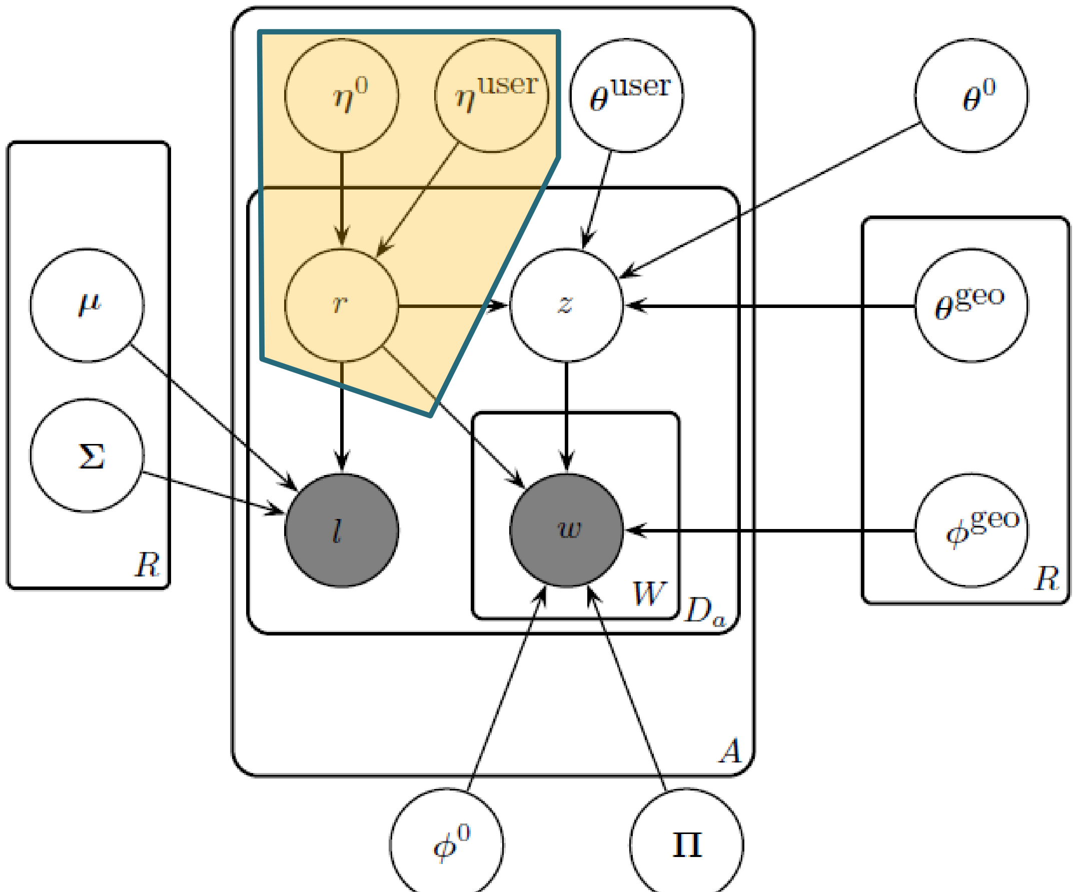
The Graphical Model



The Graphical Model



Region Selection



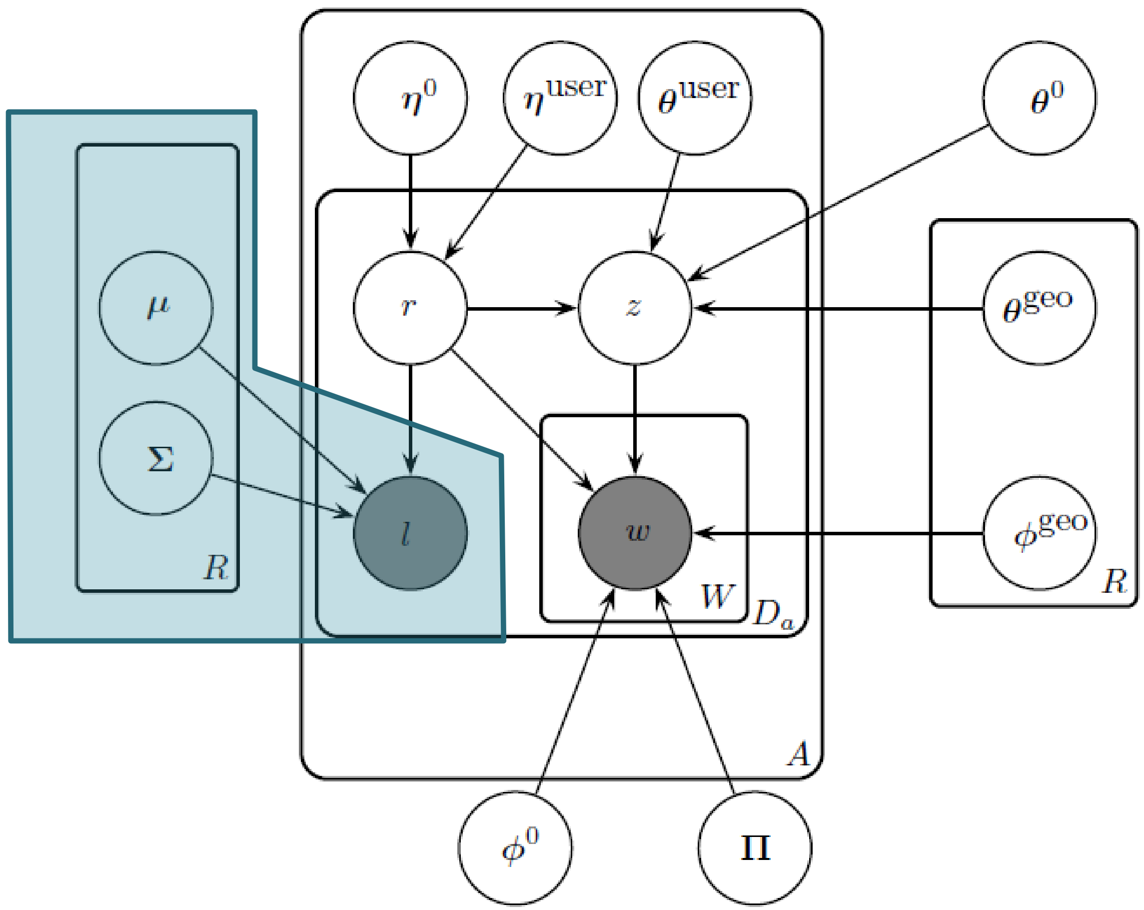
Region Selection

Step-by-Step

- Users tend to appear in a handful geographical locations.

$$P\left(r|\boldsymbol{\eta}^{0}, \boldsymbol{\eta}_{u}^{\text{user}}\right) = p\left(r|\boldsymbol{\eta}^{0} + \boldsymbol{\eta}_{u}^{\text{user}}\right)$$

Location Generation

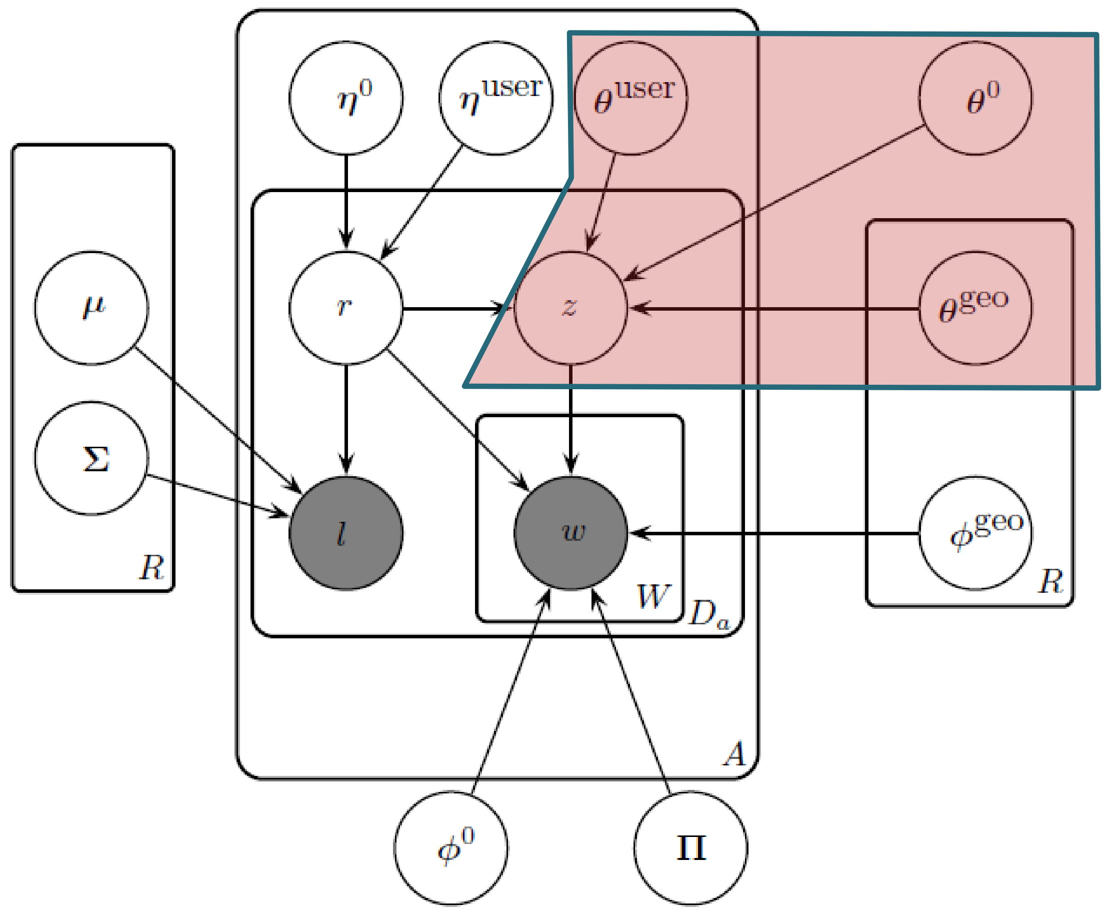


Location Generation

- Once a region is selected, locations can be generated.

$$\mathbf{l}_d \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r).$$

Topic Selection

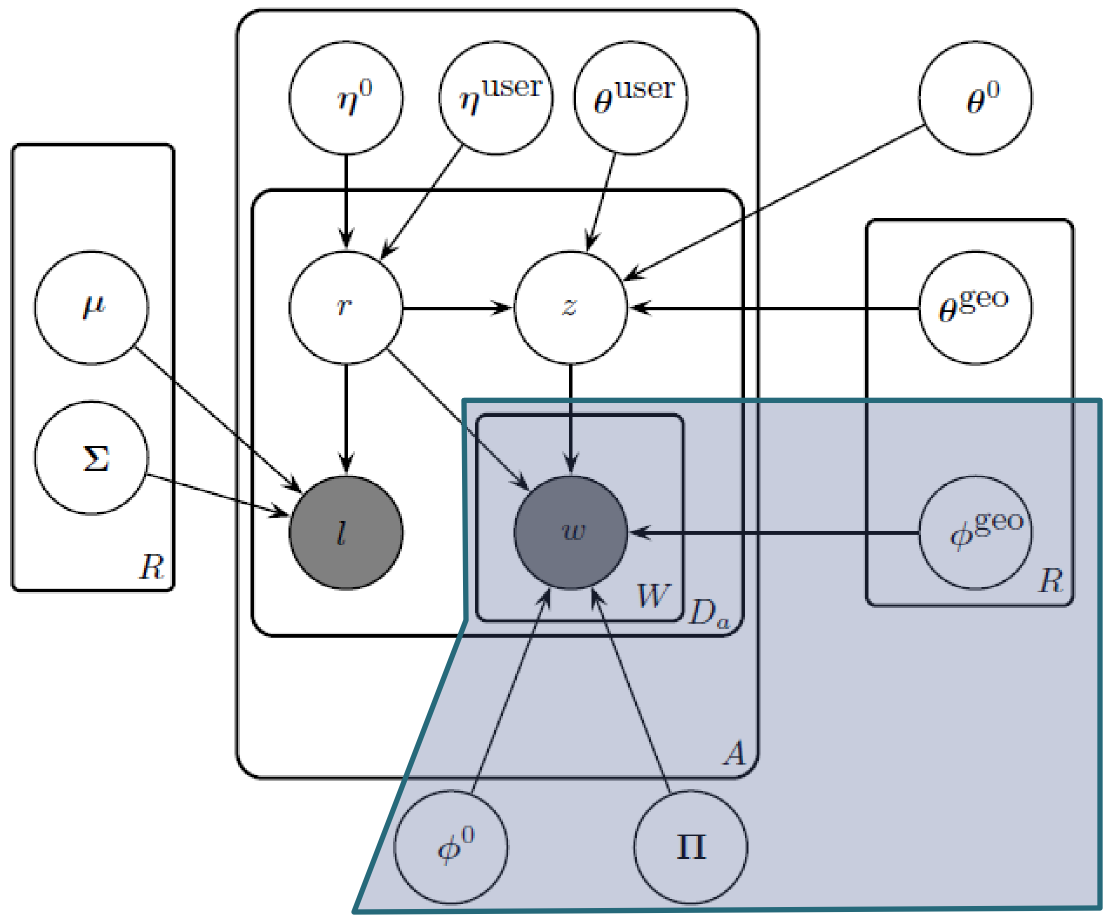


Topic Selection

- Topics have different chances to be discussed in different regions by different users

$$P\left(z|\theta^0, \theta_u^{\mathrm{user}}, \theta_r^{\mathrm{geo}}\right) = p\left(z|\theta_j^0 + \theta_{u,j}^{\mathrm{user}} + \theta_{r,j}^{\mathrm{geo}}\right)$$

Word Generation



Word Generation

- Words used in a tweet depend on both the location and topic of the tweet.

$$P\left(w|z,\phi^{0},\phi_{r}^{\mathrm{geo}},\mathbf{\Pi}_{z}\right)=p\left(w|\phi^{0}+\phi_{r}^{\mathrm{geo}}+\mathbf{\Pi}_{z_{d}}\right)$$

Sparse Modeling

- Laplace Priors

$$\eta_r^0 \sim \mathcal{L}(0, \omega^0) \quad \eta_{u,r}^{\text{user}} \sim \mathcal{L}(0, \omega_u)$$

$$\theta_z^{\text{geo}} \sim \mathcal{L}(0, \lambda_l) \quad \theta_{u,z}^{\text{user}} \sim \mathcal{L}(0, \lambda_u) \quad \theta_{r,z}^{\text{geo}} \sim \mathcal{L}(0, \lambda_r)$$

$$\phi_v^0 \sim \mathcal{L}(0, \psi^0) \quad \phi_{r,v}^{\text{geo}} \sim \mathcal{L}(0, \psi_l)$$

$$\mathbf{\Pi}_{z,v} \sim \mathcal{L}(0, \psi_t)$$

Sparsity results in predictive models

Bayesian treatment

- Prior distributions over mean and covariance matrix
- Jeffery prior

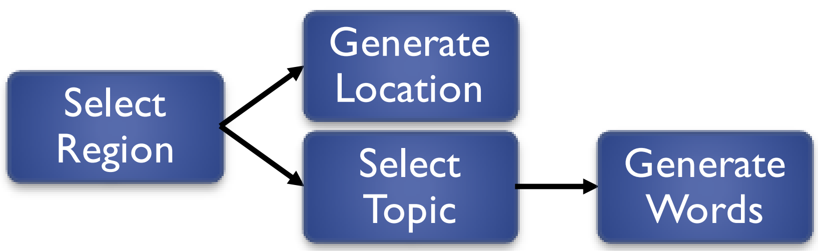
$$\mu \sim \text{Unif.}$$

$$P(\Sigma) \propto |\Sigma|^{-(3/2)}.$$

- Penalize large regions
  - We want region to be predictive as much as the data supports

Recap

- Generative Process



- Sparse Modeling
  - L_1 regularization (Laplace priors)
- Geographical Modeling
  - Bayesian treatment

Inference Algorithm

- A variant of Monte Carlo EM
  - "E-Step": Sample latent discrete variables
  - "M-step": Update all model parameters

- Sparse update of gradients
- L_1 regularization: `ISTA` algorithm
- Initialize regions with `K-means` algorithm

Overview

- Motivations
- Our Proposed Model
- Experiments
- Conclusions

Experiments

Dataset

- Twitter data
  - Randomly sample 1,000 users
  - All tweets from Jan 2011 to May 2011
  - 573,203 distinct tweets
- Twitter geographical data
  - Locations + Twitter Places

Experiments

Location Prediction

- Metric
  - average error distance
  - Kilometers

Experiments

Location Prediction

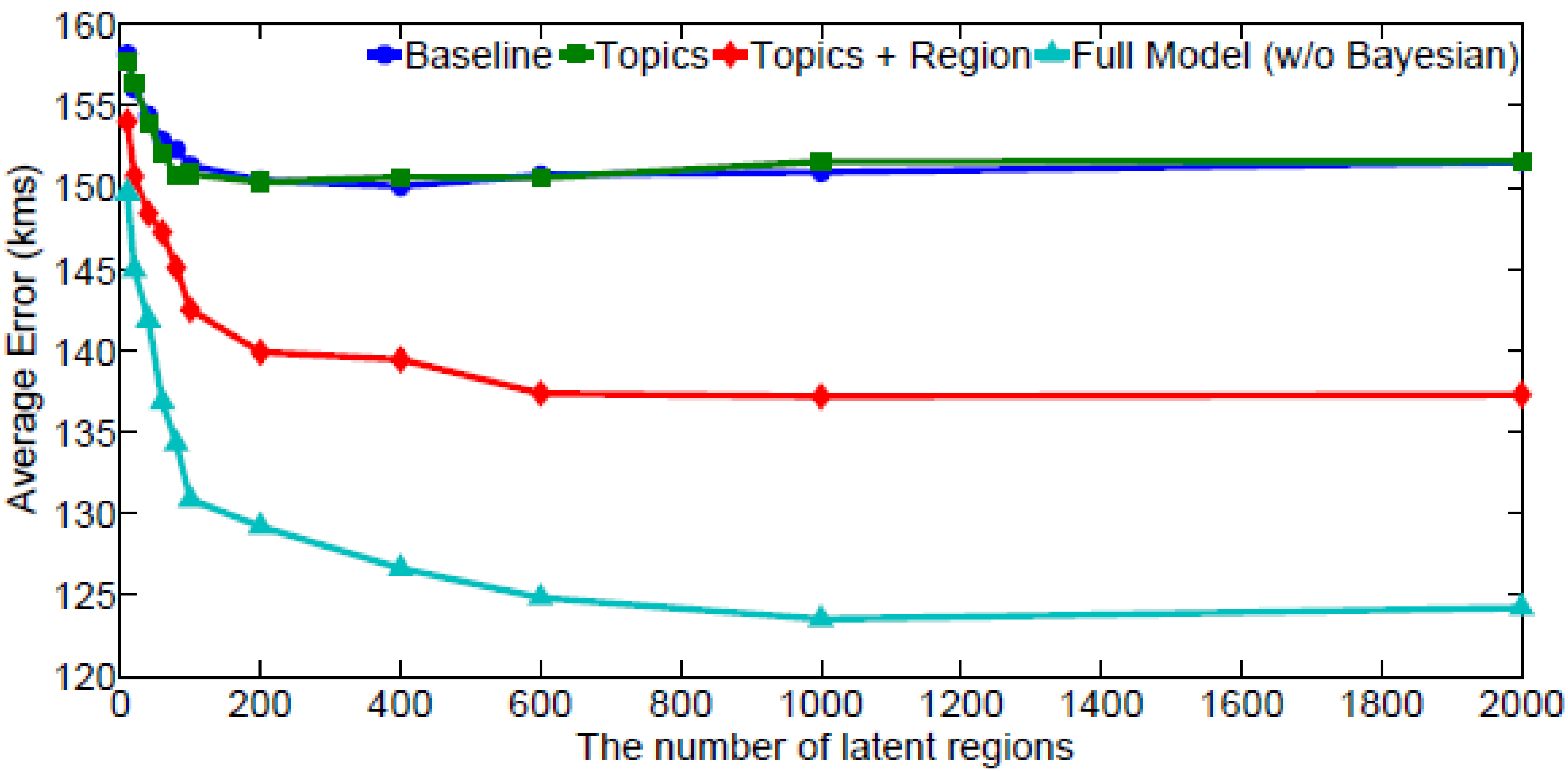
- Baselines
  - Yin et al. WWW 2011 paper
    - PLSA formalism
    - No personalization
  - Our model without ϕ^{geo}, η^{user} and θ^{user}
    - Similar to Yin et al.'s formalism but SAGE model
  - Our model without η^{user} and θ^{user}

Experiments

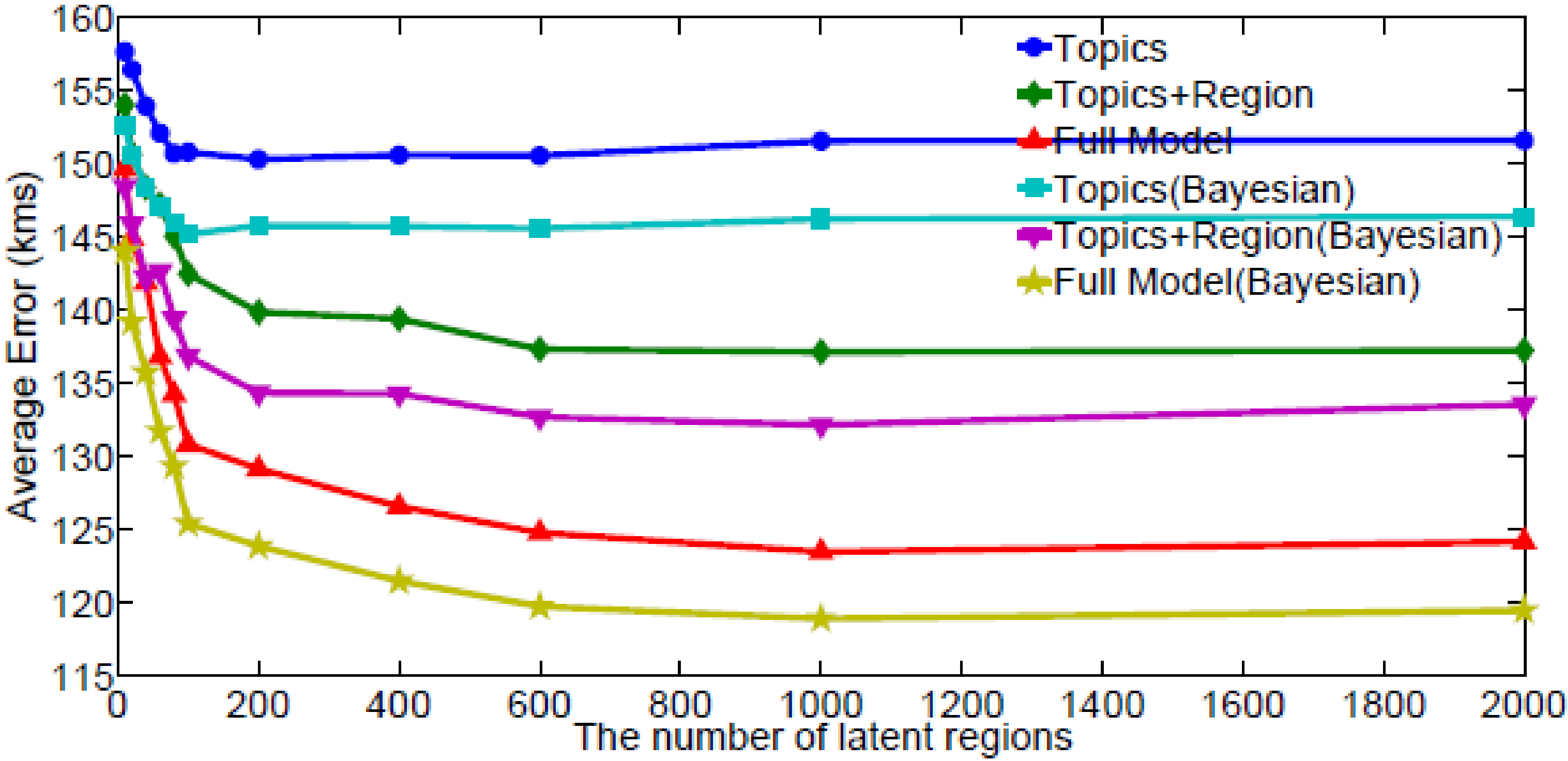
Location Prediction

- Baselines
  - Yin et al. WWW 2011 paper
    - PLSA formalism
    - No personalization
  - Our model without ϕ^{geo}, η^{user} and θ^{user}
    - Similar to Yin et al.'s formalism but SAGE model
  - Our model without η^{user} and θ^{user}

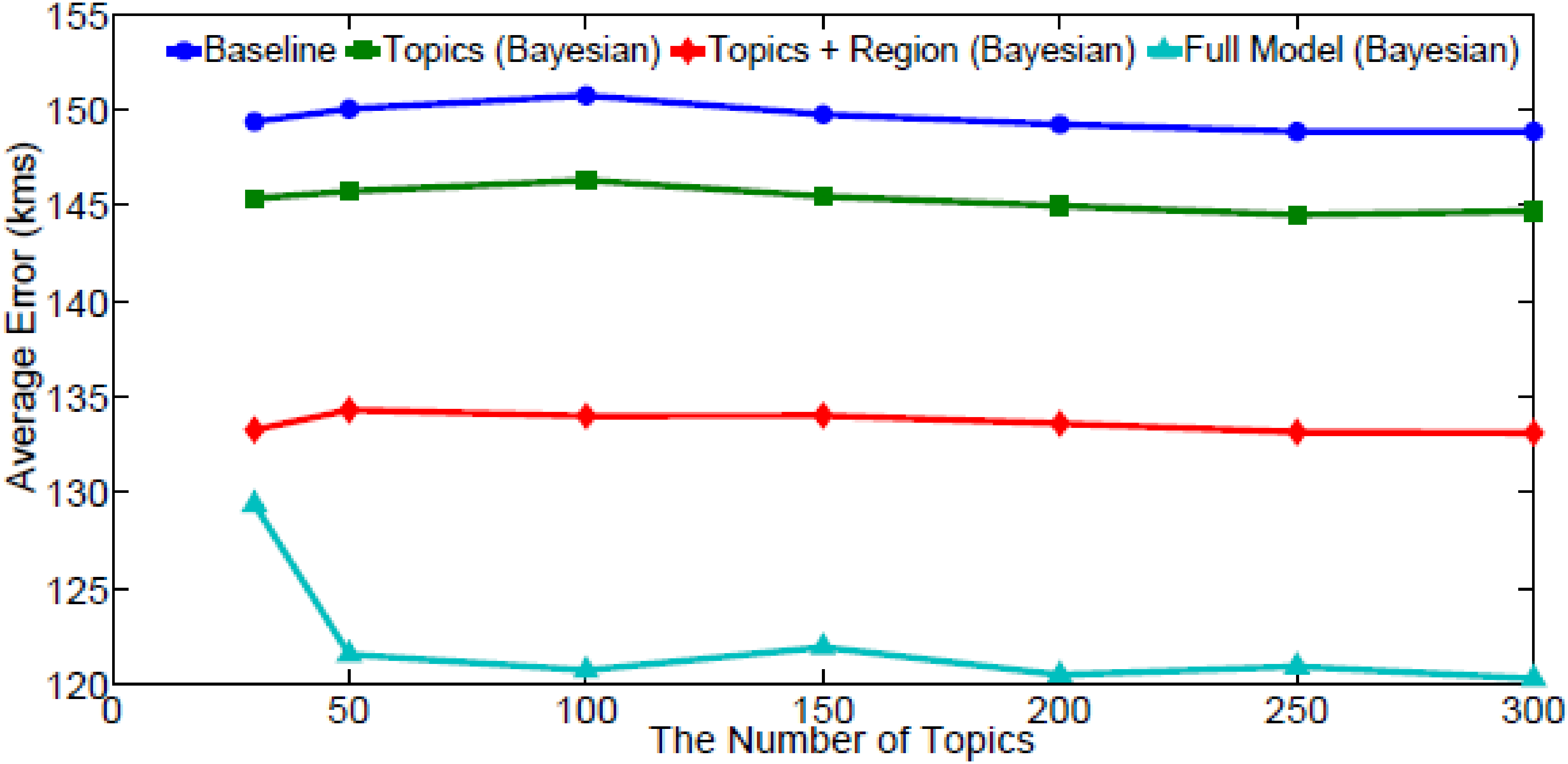
Location Prediction



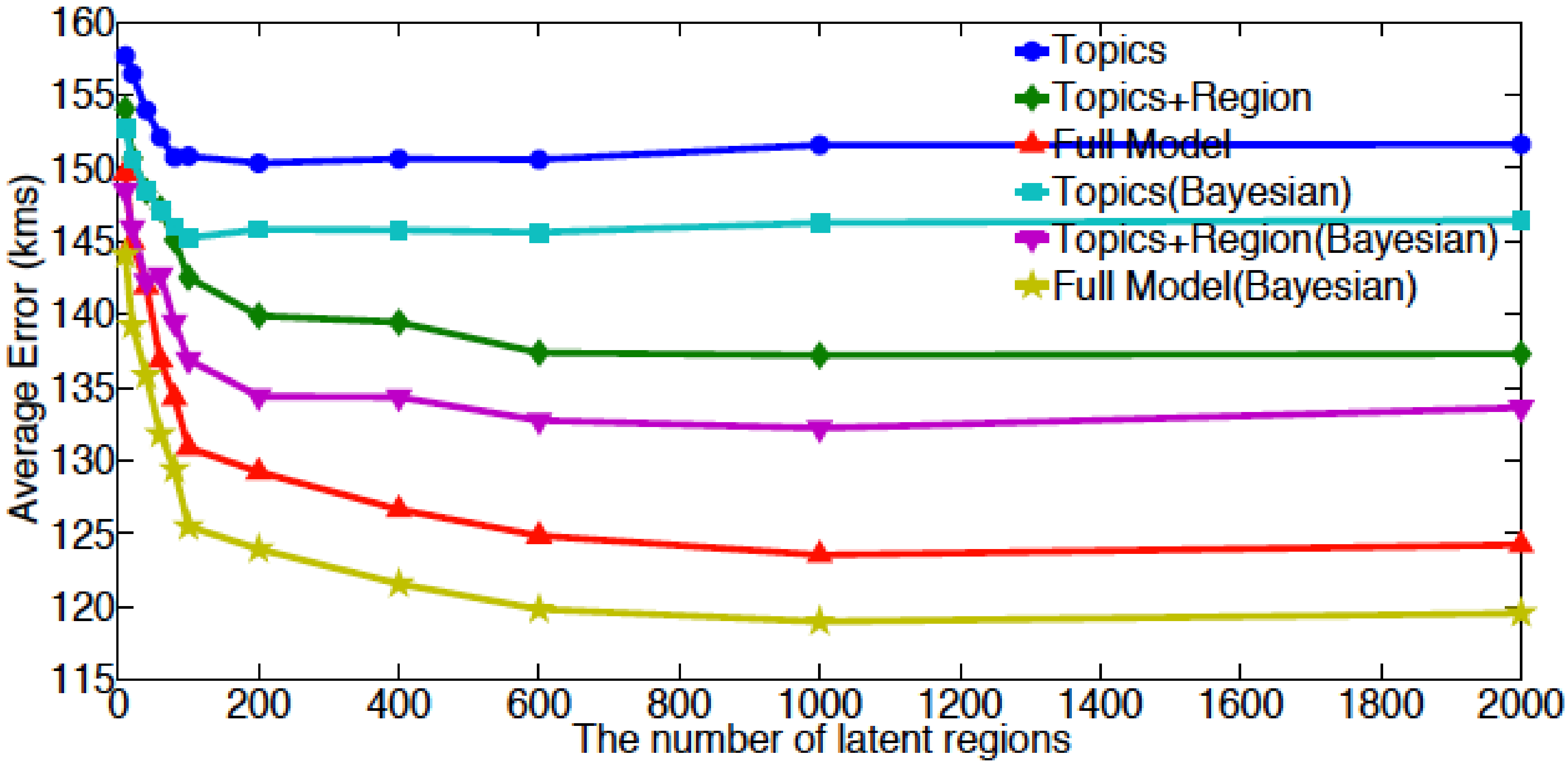
Bayesian Treatment



Number of Topics



Number of Regions

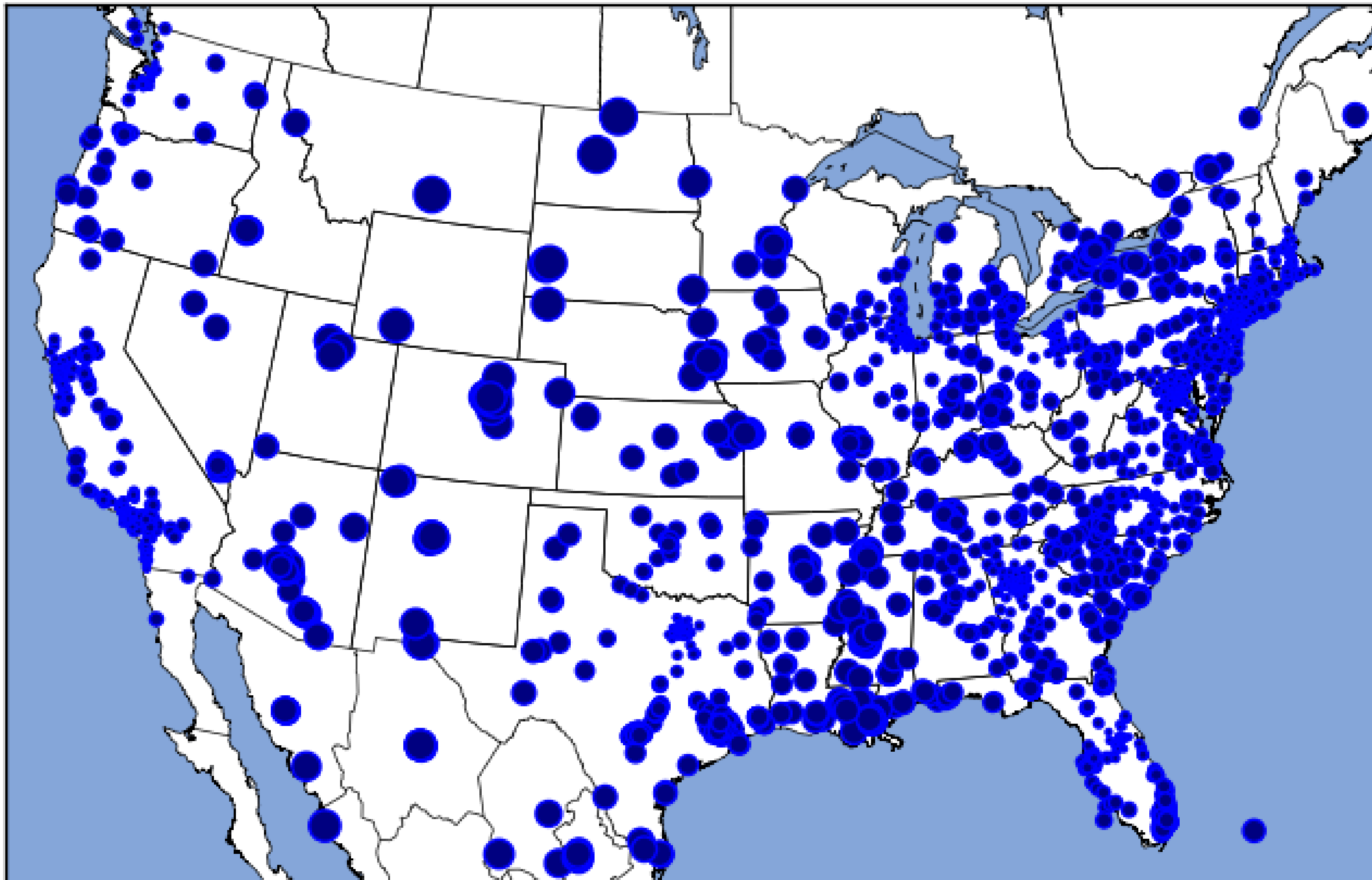


Experiments (Public Data)

# of regions	[3]	[2]	[1]	Topics	Topics + Region	Full Model
10	494	479	501	540.60	481.58	449.45
20	494	479	501	522.18	446.03	420.83
40	494	479	501	513.06	414.95	395.13
60	494	479	501	507.37	410.09	380.04
80	494	479	501	499.42	408.38	374.01
100	494	479	501	498.94	407.78	**372.99**

[1] Eisenstein et al. EMNLP 2010.
[2] Wing and J. Baldridge. ACL 2011.
[3] Eisenstein, Ahmed, Xing ICML 2011.

Error Analysis



Global and local topics

Entertainments
lady bieber album music beats artist video listen itunes apple produced movies #bieber lol new songs
Sports
yankees match nba football giants wow win winner game weekend horse #nba
Politics
obama election middle east china uprising egypt russian tunisia #egypt afghanistan people eu

Location with Top Ranked Terms
United States->New York->Brooklyn
brooklyn ave flatbush avenue mta prospect 5th #brooklyn spotlight carroll bushwick museum broadway madison vanderbilt coney slope eastern subway new york pkwy #viernesnayobon #mets otsego greenwich starbucks
United States->California->San Francisco
sfo francisco san airport international millbrae terminal flight burlingame bart mateo boarding bayshore telecommute landed heading bay airlines united bound flying #sfo camino groupon caltrain moon tsa baggage california engineer valley
United States->Pennsylvania->Philadelphia
philadelphia #philadelphia phl #jobs market others #job street philly walnut septa chestnut the cherry sansom arch spruce citizens locust btw temple pennsylvania rittenhouse passyunk bitlyetq7a6 bookrenters pike international
United Kingdom->England->London
winds lhr hounslow terminal the cloudy mph ickenham bath heathrow temperature airport car only airways uxbridge sun splendid fair london british lounge tothers harmondsworth speedbird whens for stars day flight dominos navigation brunel
Australia->New South Wales->Sydney
sydney #sydney bondi george street mascot domestic syd surry station cnr platforms harbour darlinghurst qantas hoteloxford eddy haymarket terminal wales australia chalmers uts pitt #marketing junction darling centre #citijobs citigroup druitt

Conclusions

- Probabilistic model for geographical information
  - Regional variations
  - Personal preferences
- Effective inference algorithm
- Best location prediction
- Discriminatively learned language models
- Future work
  - Hierarchical model
  - Hash tags
  - Temporal location model