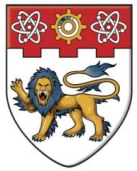


NANYANG TECHNOLOGICAL UNIVERSITY

# Who, Where, When and What: Discover Spatio-Temporal Topics for Twitter Users

Presenter: Quan Yuan (SCE)

Supervisor: Asst. Prof. Gao Cong

Prof. Nadia Magnenat- Thalmann

Nanyang Technological University

KDD 2013

# Motivation

- Micro-blogging services (e.g., Twitter) and location-based social networks (LBSNs, e.g., Foursquare) have generated a great number of geotagged short text messages.


- A short message contains a user id, a text message, posting time and a venue.
- Such short messages offers a good opportunity to study the behaviors of individuals (**who**) from geographical location (**where**), time (**when**) and activity (**what**).

NANYANG TECHNOLOGICAL UNIVERSITY

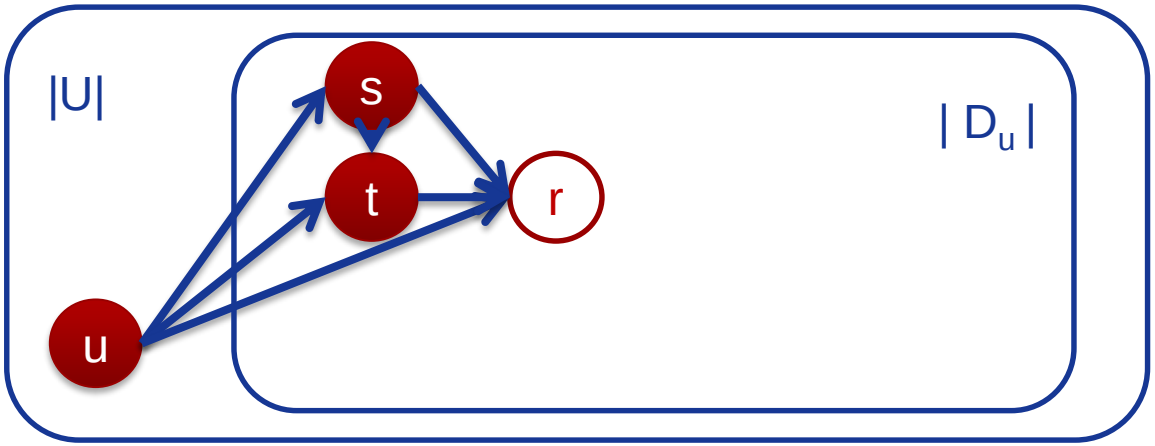
# Related Work

- Previous studies focused on at most three factors of **Who, Where, When**, and **What**
  - **Where What**: geographical topic modeling
    - *Yin et al.* Geographical topic discovery and comparison. *In WWW 2011*
    - Chinatown: dinning, Raffles: sightseeing, etc..
  - **Where When What**: geographical event detection
    - *Sakaki et al.* Earthquake shakes twitter users: real-time event detection by social sensors. *In WWW 2010*
    - China @ Jan 28st: lunar new year
  - **Who Where When**: modeling spatio-temporal mobility behaviors of individual users
    - *Cho et al.* Friendship and mobility: user movement in location-based social networks. *In KDD 2011*
    - Tony: office@2:00 pm, home@9:00pm
  - **Who Where What**: geographical topic profiling of users
    - *Hong et al.* Discovering geographical topics in the twitter stream. *In WWW 2012*
    - Tony @ Jurong Point: shopping, dinning

# Overview: Region

- A tweet *d* is modeled as a five-tuple $\{u_d, l_d, \boldsymbol{w_d}, t_d, s_d\}$
  - *u*: user, *l*={*id, coordinate*}: venue id and geographical coordinate
  - ***w***: words, *t*={*hh:mm:ss*}: time in a day, *s*: workday/weekend
- Intuitions:
  - An individual *u*'s mobility centres at different personal geographical regions *r* (e.g., home region, work region, shopping region, etc).
  - The region where a user stays is influenced by time (time in a day *t* & day of a week *s*).
  - Eg: Tony: 2:00 pm, weekday – NTU; 9:00 pm, weekend – Home
  - Region: Gaussian distribution over latitude/longitude.

**Graphical Model**



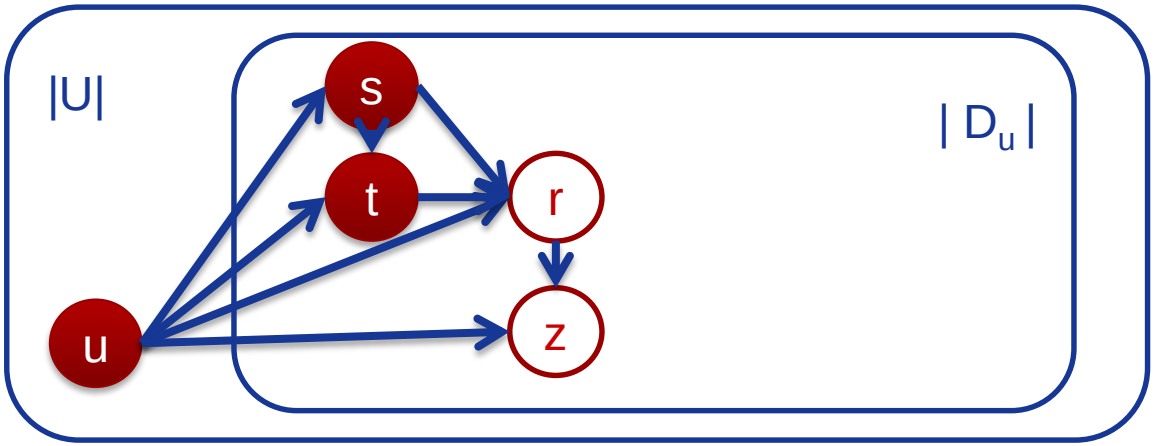

NANYANG TECHNOLOGICAL UNIVERSITY

# Overview: Topic

- A tweet is modeled as a five-tuple $d = \{u_d, l_d, \boldsymbol{w_d}, t_d, s_d\}$
  - $u$: user, $l$={*id, coordinate*}: venue id and geographical coordinate
  - $\boldsymbol{w}$: words, $t$={*hh:mm:ss*}: time in a day, $s$: workday/weekend
- Intuitions:
  - Different users $u$ and regions $r$ has different preferences over topics $z$, {$P(z|u)$}, {$P(z|r)$}
  - The topics $z$ of a user $u$ at a place are influenced by region $r$ and $u$'s topic preference.
  - Eg.: Home: cooking, music. Tony: cooking, reading

    Tony @ Home: cooking

**Graphical Model**



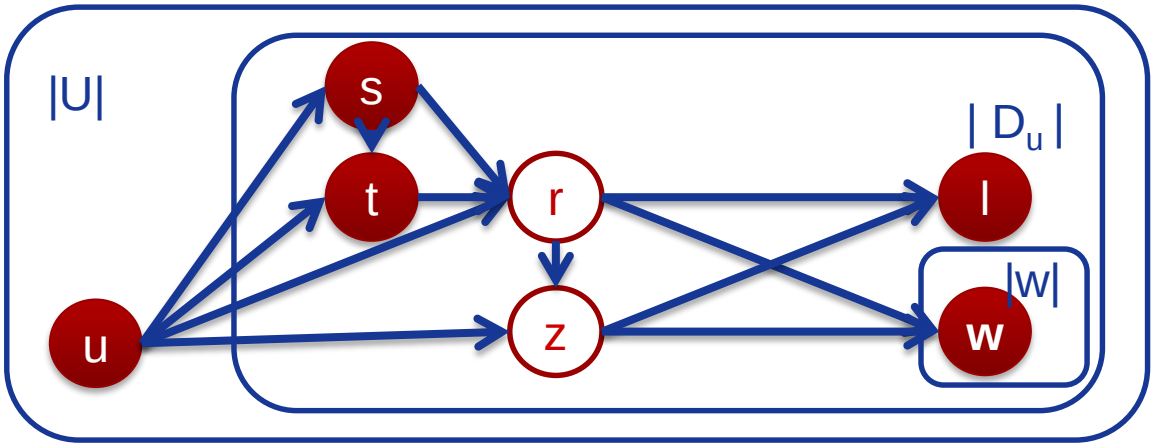

NANYANG TECHNOLOGICAL UNIVERSITY

# Overview: Venue and Words

- A tweet is modeled as a five-tuple $d = \{u_d, l_d, \boldsymbol{w_d}, t_d, s_d\}$
  - *u*: user, *l*={*id, coordinate*}: venue id and geographical coordinate
  - ***w***: words, *t*={*hh:mm:ss*}: time in a day, *s*: workday/weekend
- Intuitions:
  - When choosing a venue *l* to visit, the user considers the topic requirement and the region. $P(l|r,z) = \alpha * P(id|z) + (1-\alpha) * P(coor|r)$
  - E.g.: Tony, shopping: fairprice
  - The selection of words are also influenced by topic and region.

    $P(w|r,z) = \beta * P(w|z) + (1-\beta) * P(w|r)$

**Graphical Model**



# Parameter Estimation

- The likelihood:

$$\sum_d log \sum_z \sum_r p(u_d, r, z, s_d, t_d, \ell_d, \boldsymbol{w}_d) = \sum_d log \sum_z \sum_r p(u_d) p(s_d|u_d) p(t_d|u_d, s_d) p(r|u_d, s_d, t_d)$$

$$[\prod_{w \in \boldsymbol{w}_d} (\lambda p(w|z) + (1-\lambda) p(w|r))^{c(w, \boldsymbol{w}_d)}][\kappa p(\ell_d|z) + (1-\kappa) p(\ell_d|r)] p(\boldsymbol{w}_d|r, z)$$

- We use Expectation-Maximization (EM) to estimate parameters
- E-step: $p(r, z|d) = \dfrac{p(d, r, z)}{p(d)} = \dfrac{p(d, r, z)}{\sum_r \sum_z p(d, r, z)}$
- M-step:

$$p(r|u, s) = \frac{\sum_{d \in D_{u,s}} \sum_z p(r, z|d)}{\sum_{d \in D_{u,s}} \sum_z \sum_{r'} p(r', z|d)},$$

$$p(w|r) = \frac{\sum_{d \in D_w} \sum_z c(w, \boldsymbol{w}_d) p(r, z|d)}{\sum_{w'} \sum_{d \in D'_w} \sum_z c(w', \boldsymbol{w}_d) p(r, z|d)},$$

$$p(w|z) = \frac{\sum_{d \in D_w} \sum_r c(w, \boldsymbol{w}_d) p(r, z|d)}{\sum_{w'} \sum_{d \in D'_w} \sum_r c(w', \boldsymbol{w}_d) p(r, z|d)},$$

$$p(z|u, r) = \frac{\sum_{d \in D_u} p(r, z|d)}{\sum_{d \in D_u} \sum_{z'} p(r, z'|d)},$$

$$\boldsymbol{\mu}_{u,s,r} = \frac{\sum_{d \in D_{u,s}} \sum_z p(r, z|d) \cdot \boldsymbol{cd}_{\ell_d}}{\sum_{d \in D_{u,s}} \sum_z p(r, z|d)},$$

$$\nu_{u,s,r} = \frac{\sum_{d \in D_{u,s}} \sum_z p(r, z|d) \cdot t_d}{\sum_{d \in D_{u,s}} \sum_z p(r, z|d)},$$

$$\boldsymbol{\Sigma}_{u,s,r} = \frac{\sum_{d \in D_{u,s}} \sum_z p(r, z|d) \cdot (\boldsymbol{cd}_{\ell_d} - \boldsymbol{\mu}_{u,s,r})^T (\boldsymbol{cd}_{\ell_d} - \boldsymbol{\mu}_{u,s,r})}{\sum_{d \in D_{u,s}} \sum_z p(r, z|d)},$$

$$\sigma^2_{u,s,r} = \frac{\sum_{d \in D_{u,s}} \sum_z p(r, z|d) \cdot td^2(t_d, \nu_{u,s,r})}{\sum_{d \in D_{u,s}} \sum_z p(r, z|d)},$$

$$p(\ell|z) = \frac{\sum_{d \in D_\ell} \sum_r p(r, z|d)}{\sum_{d \in D_\ell} \sum_{z'} \sum_r p(r, z'|d)}.$$

# Datasets and Evaluation Tasks

- Datasets:
  - **WW**: 89,007 world-wide tweets, 3,883 users, 60,962 venues
  - **USA**: 171,768 microblogs in USA, 4,122 users, 35,989 venues
- Venue prediction for tweets
  - Given text contents, user ids and posting time, predict the most likely venue at which the tweet is posted.
  - Rank candidate venues *l* by $p(l \mid u, s, t)$
- User prediction:
  - Predict the likelihood of a user visiting a venue at a given time.
  - Rank candidate users *u* by $p(u \mid l, s, t)$
- Venue prediction for user:
  - Predict the place where a user stays at a given time.
  - Rank candidate venues *l* by $p(l \mid u, s, t, \mathbf{w})$

# Venue prediction for tweets

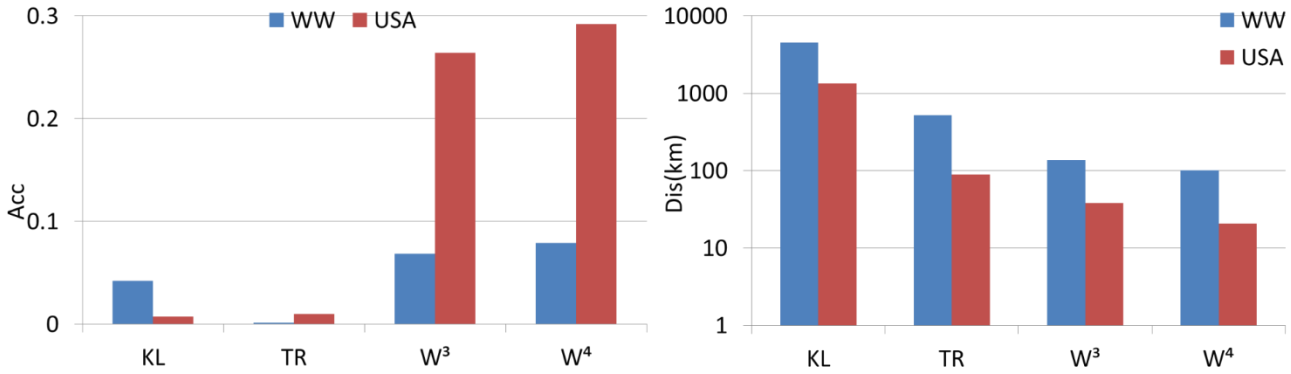
- Metric:
  - Prediction accuracy (Acc): percentage of tweets whose predicted venues are the true venues.
  - Average error distance (Dis): the average geographical distance between the predicted and true venues for all tweets.
  - larger Acc and smaller Dis indicate better performance.
- Baselines:
  - KL-divergence based method (KL) ***W. Li et al. CIKM 2011***
  - Topic+Region (TR) ***L. Hong et al. WWW 2012***
  - Who+Where+What (W3)
  - Who+Where+When+What (W4)

| Factors in modeling | KL | TR | $W^3$ | $W^4$ |
|---|---|---|---|---|
| Who (User) | × | √ | √ | √ |
| Where (Geo) | × | GlbR | PsnR | PsnR |
| When (Time) | × | × | × | √ |
| What (Words) | √ | √ | √ | √ |

NANYANG TECHNOLOGICAL UNIVERSITY

# Venue prediction for tweets

- Prediction results :




- W4 outperforms KL and TR by more than 80% in terms of both metrics.
- W3 utilizes the same information as does TR, but gains better results
  - Regions are personal in W3, but global in TR
- W4 achieves best results
  - Time factor is important

NANYANG TECHNOLOGICAL UNIVERSITY

# User Prediction and Venue Prediction

- Baselines: PMM [*Cho et al. In KDD 2011*]
- Metric: Accuracy (Acc)
- User Prediction:
  - Predict the likelihood of a user visiting a venue at a given time

| Acc | WW | USA |
|---|---|---|
| PMM | 0.4163 | 0.4021 |
| W4 | 0.5063 | 0.5863 |

- Venue Prediction for User
  - Predict the place where a user stays at a given time.

| Acc | WW | USA |
|---|---|---|
| PMM | 0.0423 | 0.1102 |
| W4 | 0.0776 | 0.2953 |

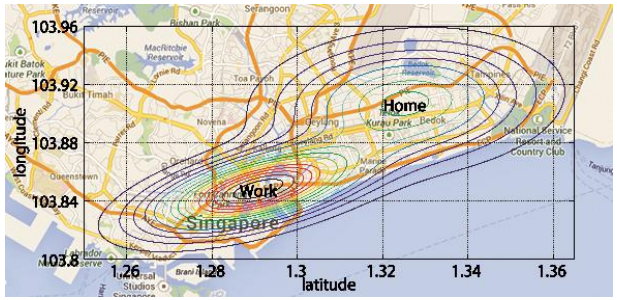
- W4 outperforms PMM by more than 20% and 45% on the two datasets, respectively

# Results of Other Tasks

- Activity prediction:
  - Predict word of a user at a given time

| time | words |
|---|---|
| 14:30 weekday | break work coffee resting gym international |
| 10:00 weekend | good morning home breakfast shopping eat |

- User mobility




Personal regions

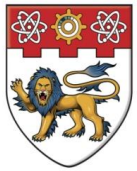
- Representative words for topics

| Topic | Representative words |
|---|---|
| Home | family fun offroad rental home love |
| Dining | lunch dinner birthday breakfast drinks eat |
| Nightlife | night happy singing playing dance football |
| Work | working tonight coffee tired money Friday |
| Holiday | Christmas friends holiday merry celebrating choir |

NANYANG TECHNOLOGICAL UNIVERSITY

# Conclusion

- The large availability of geo-tagged tweets enables us to study individuals' mobility behaviors user, geolocation, time, and activity factors.
- We propose W4 (**Who Where When What**) to model the interactions of all the four factors in a unified way to better understand individuals' behaviors.
- Experimental results on two real-world datasets show that the proposed method outperforms baselines on various applications.



NANYANG
TECHNOLOGICAL
UNIVERSITY

# Q & A?

# Thank You !

# Publications


- Q. Yuan, G. Cong, Z. Ma, A. Sun, N. M. Thalmann, **Who, Where, When and What: Discover Spatio-Temporal Topics for Twitter Users**, *In Proc. of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD), 605-613, 2013.*
- Q. Yuan, G. Cong, Z. Ma, A. Sun, N. M. Thalmann, **Time-aware Point-of-interest Recommendation**, *In Proc. of ACM SIGIR Conference (SIGIR), 363-372, 2013.*
- B. Liu, Q. Yuan, G. Cong, D. Xu, **User Profile Enhanced Geolocation Suggestion for Social Images**, *Journal of the American Society for Information Science and Technology (JASIST), 2013.*
- Q. Yuan, G. Cong, A. Sun, C. Lin, N. M. Thalmann, **Category Hierarchy Maintenance: A Data-Driven Approach**, *In Proc. of ACM SIGIR Conference (SIGIR), 791-800, 2012.*
- Q. Yuan, G. Cong, N. M. Thalmann, **Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification**, *In Proc. of WWW, Poster track, 645-646, 2012.*
- Z. Ma, A. Sun, Q. Yuan, G. Cong, **Topic-Driven Reader Comments Summarization**, *In Proc. of ACM Conference on Information and Knowledge Management (CIKM), 265-274, 2012.*
- X. Cao, G. Cong, B. Cui, C. S. Jensen, Q. Yuan, **Approaches to Exploring Category Information for Question Retrieval in Community Question-Answer Archives**, *ACM Transactions on Information Systems (TOIS), Volume 30 Issue 2, Article No. 7, May 2012.*