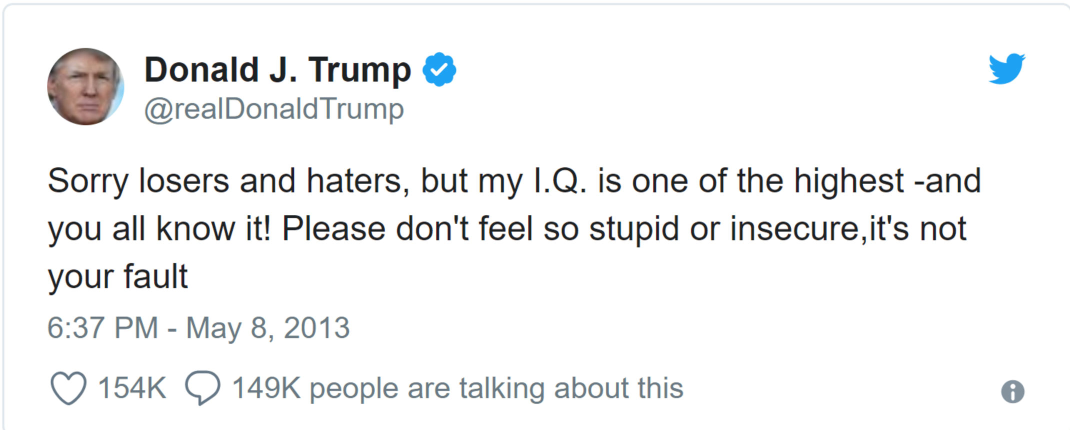
Dynamic Topic Modeling: Spatiotemporal Analysis of Los Angeles Twitter Data

D. J. Arnold[1], J. Du[2], K. Flood[1], M. Ghavamizadeh[3] B. Kim[1], C. Parkinson[1], M. Plack[4], S. Tan[5], H. Yao[1], A. L. Bertozzi[1], P. J. Brantingham[1]

[1]University of California, Los Angeles, [2]Princeton University, [3]University of California, Berkeley, [4]University of Siegen, [5]Harvey Mudd College

Problem

As a widely used social media and news service, Twitter is a valuable source of data for understanding social trends. However, the enormous amount of data Twitter contains requires an efficient method for discovering hidden semantic structure and common themes among tweets. We create a topic model using a dynamic, word-embedded non-negative matrix factorization (Semantic NMF), and apply this model to a dataset of Los Angeles tweets. We analyze the spatiotemporal patterns of topics and explore the task of location inference.



Example Tweet

Definitions

- $X \in \mathbb{R}^{n\times m}$ is the **term-document** matrix containing occurrences of the n unique terms in the m documents.
- $W \in \mathbb{R}^{n\times k}$ is the latent **term-topic** matrix, describing which words define the k latent topics.
- $H \in \mathbb{R}^{k\times m}$ is the **topic-document** matrix showing topic assignment for each document.
- The goal of **NMF** is to find the solution of

$$\operatorname*{arg\,min}_{W,H} \|X - WH\|_F^2 \quad \text{s.t. } W \geq 0, H \geq 0,$$

thus splitting the body of text into k latent topics.

Word embeddings

Word embeddings capture the semantic meaning of words. The **NMF** is modified to compute distance in the word embedding space as

$$\operatorname*{arg\,min}_{W,H} \|V(X - WH)\|_F^2 \quad \text{s.t. } W \geq 0, H \geq 0.$$

V can come from any word embedding model. We use *word2vec* trained on a Google News dataset.[1]

- The objective function can be minimized using a modified **Hierarchical Alternating Least Squares** algorithm.
- Initializing W with standard NMF improves convergence.

Comparing the results with standard NMF it can be seen that

- topics are more diverse and
- topic assignment vectors are more sparse.

A more thorough evaluation will follow.

[1] https://code.google.com/archive/p/word2vec/

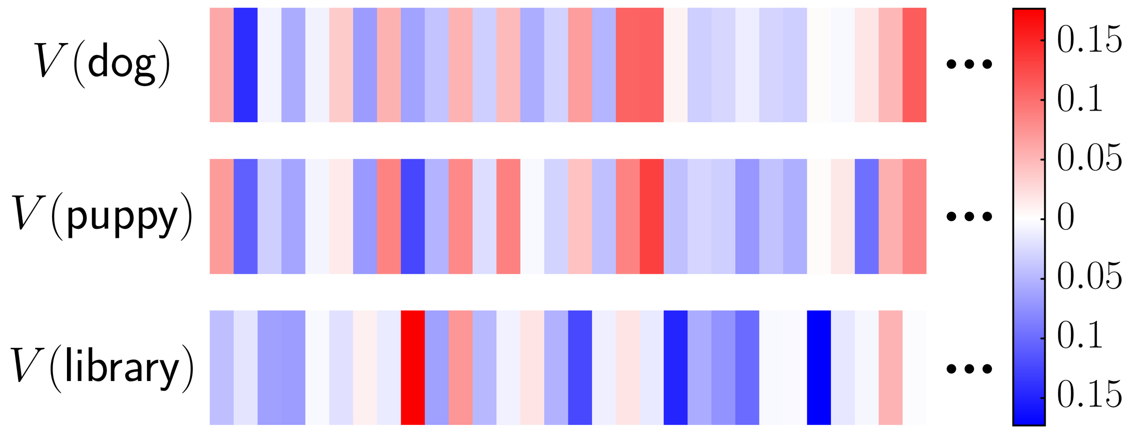


Figure: Word embeddings figure

Dynamic Topic Modeling

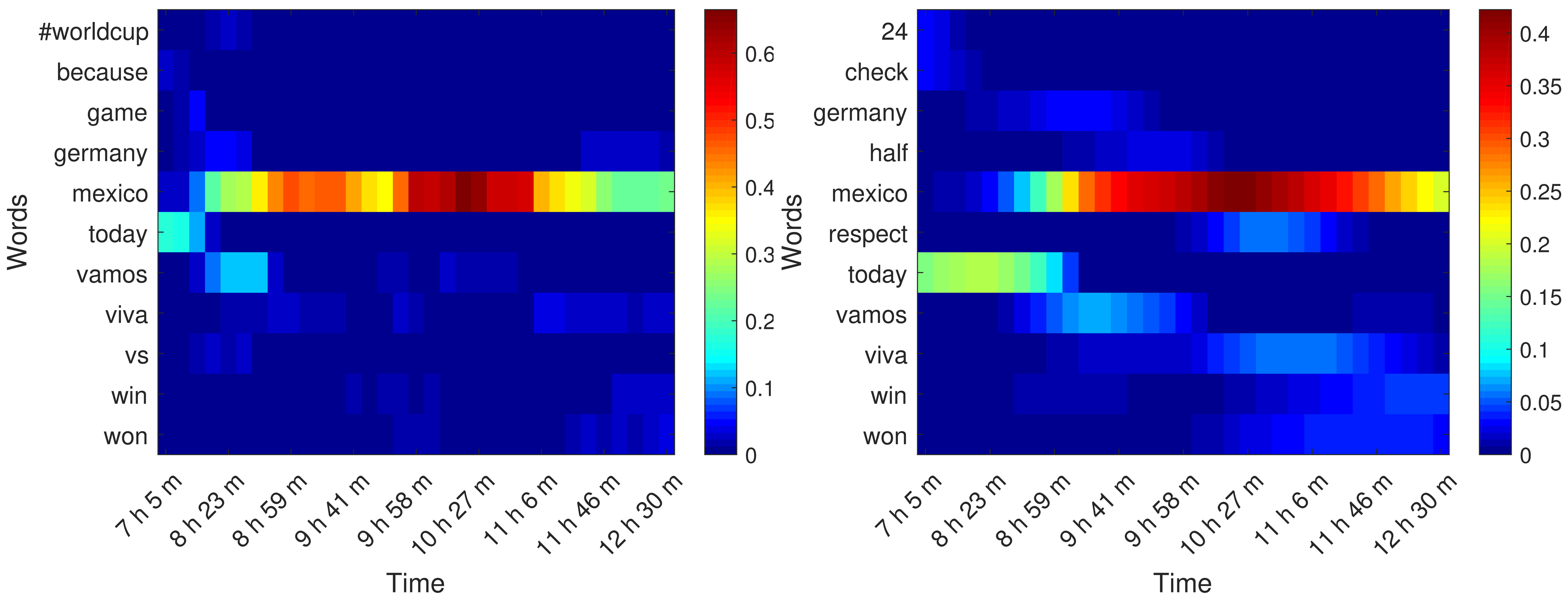
We use a **sliding time window** and run NMF on each epoch t. In order to achieve consistency of topics we add a **temporal regularization** of W similar to [3] as

$$\operatorname*{arg\,min}_{W^t,H^t} \|X^t - W^tH^t\|_F^2 + \lambda\|W^t - W^{t-1}\|_F^2$$

$$\text{s.t.} \quad W^t \geq 0, H^t \geq 0,$$

where W^{t-1} denotes the term-topic matrix from the previous epoch. This regularization ensures that topics do not change drastically from one time window to the next. To solve this, we initialize $W^t = W^{t-1}$ and $H^t = [H^{t-1} \ \hat{H}]$ where $\hat{H} = (W^t)^+ X^t$. Using these initial guesses, the minimization problem requires fewer iterations to converge and the computation remains feasible.

Figure: Example Topic from June 17th, 2018



(a) Without regularization (b) With regularization

Current Events

The aim is the identification of topics that are related to a specific event occurring at a specific place in Los Angeles based on the geolocation information of the tweets. For this we use the spatial and temporal **Fractional LP-norm** defined as

$$LP_s = \frac{\|f_j^s\|_p}{\|f_j^s\|_1}, \quad LP_t = \frac{\|f_j^t\|_p}{\|f_j^t\|_1}$$

where f_j^s is the pdf of the **spatial distribution** and f_j^t the the pdf of the **temporal distribution** of tweets belonging to topic j. Note that for any function f, $\|f\|_1$ denotes the "mass" of the function and as $p \to 0^+$, $\|f\|_p$ becomes the volume of the support of f. Thus a small Fractional LP-norm indicates a topic which has large mass focused in a small region. We use this measure to decide which topics are very specially localized in either space or time.

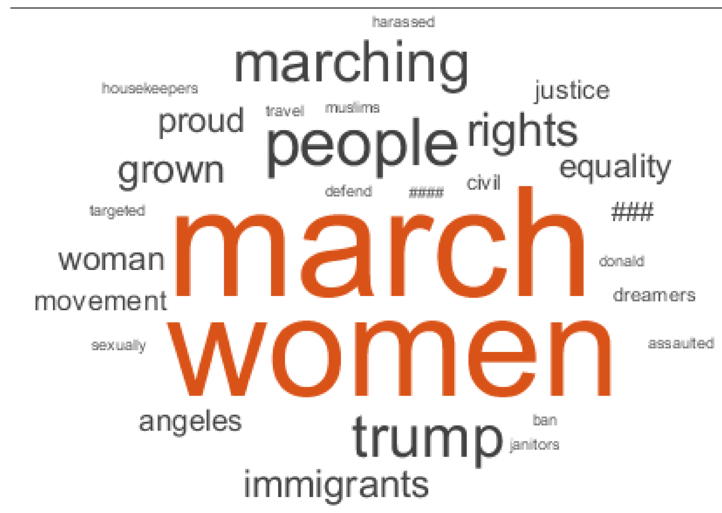
	LP_s	LP_t
	74.16	517.67
	223.80	599.73

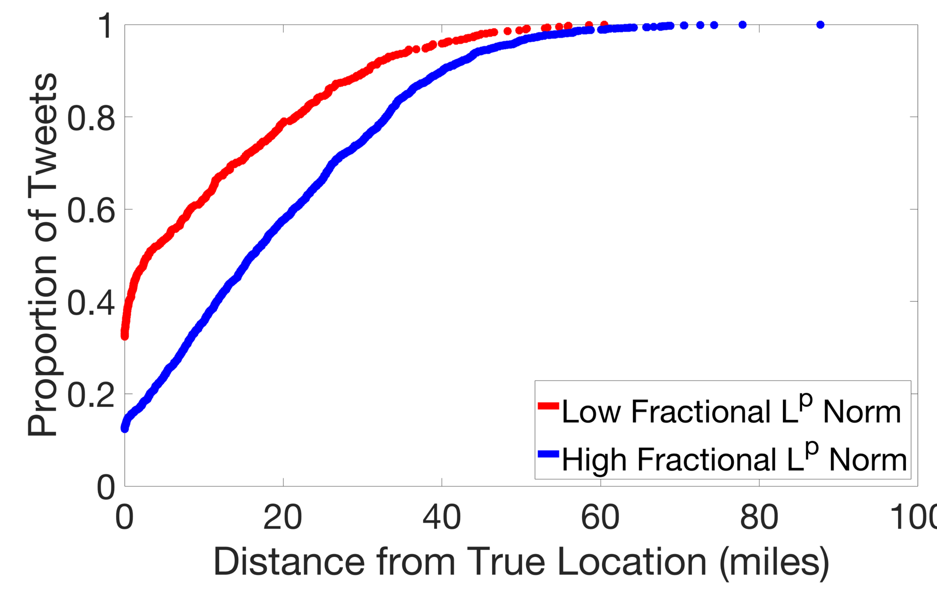
Table: Spatial and temporal LP norms for two topics.

Location Inference

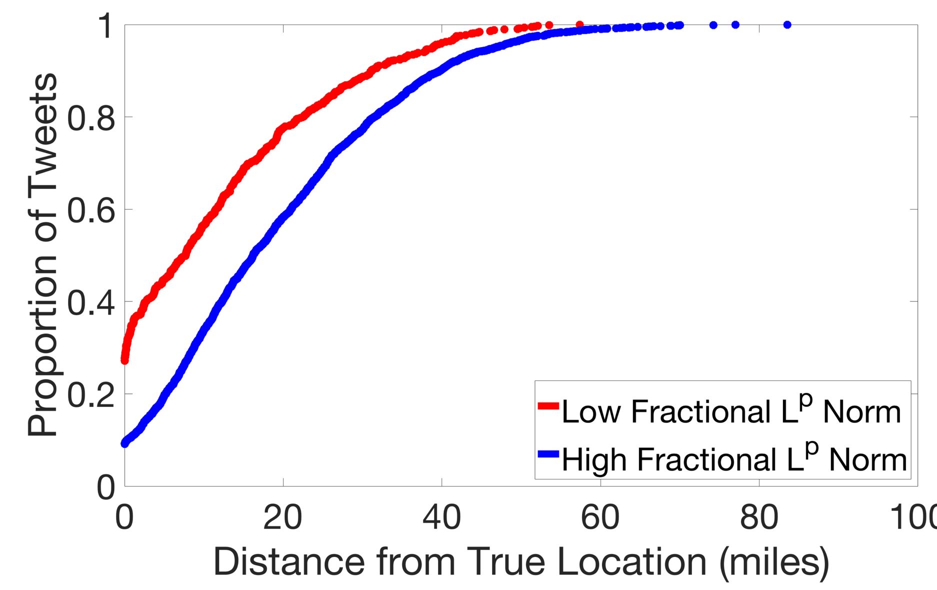
Only part of the tweets are label with geolocation information. One can assign such a tweet to the location of the closest matching tweet. To compute similarity between tweets we examined using the cosine similarity of the

1. word vectors, i.e. the columns of X, and
2. topic assignment vectors, i.e. the columns of H.

Not all tweets/topics are related to a certain location which makes inference difficult. Fractional LP-norm can be used to express the uncertainty of the inference.



(a) Tweet Similarity



(b) Topic Assignment Similarity

Figure: Comparing the accuracy of location inference for low and high LP_s tweets

References

[1] E. Lai, D. Moyer, B. Yuan, E. Fox, B. Hunter, A.L. Bertozzi, P.J. Brantingham, *Topic time series analysis of microblogs*, IMA Journal of Applied Mathematics (2016).

[2] T. Meyer, D. Balagué, M. Camacho-Collados, H. Li, K. Khuu, P.J. Brantingham and A.L. Bertozzi, *A year in Madrid as described through the analysis of geotagged Twitter data*, Environment and Planning B: Urban Analytics and City Science (2018).

[3] Y. Chen, H. Zhang, J. Wu, X. Wang, R. Liu, M. Lin, *Modeling emerging, evolving and fading topics using dynamic soft orthogonal nmf with sparse representation*, 2015 IEEE International Conference on Data Mining (2015).

Acknowledgements

This research was supported by funding from the National Science Foundation, grant DMS-1737770, and by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD).