Applications of Linked Document Topic Models to Spatially Embedded Social Networks

Andrew Campbell, *UC Berkeley*, in collaboration with:
Rory McGrath, *Amazon*, Prof. Alexey Pozdnukhov, *UC Berkeley*

***Abstract*—An open problem in transportation demand modeling is predicting locations of discretionary activities. Social networks impact activity location choice directly, through the coordination of group activities, and indirectly through latent social influence. We model the influence of social networks and the locations where activities are conducted by reframing the semantics of hyperlinked document topic models. Instead of documents, words, and hyperlinks, we model people, locations, and social network connections. Topics are latent communities, defined by the locations where people and their friends conduct activities. We apply the Linked-PLSA-LDA (L-PLSA-LDA) model to this problem. A spatially embedded population of socially networked individuals is synthesized from geotagged Twitter postings and inter-user connections. We demonstrate that the Linked-PLSA-LDA approach is capable of identifying latent communities with coherent spatial structure.**

Keywords*—LDA, L-PLSA-LDA, spatially embedded networks, social networks

I. Introduction

In this paper we propose using hyperlinked document topic modeling as tool for building meaningful models of the influences between activity locations and social networks. Specifically, we apply the L-PLSA-LDA (Linked - Probabilistic Latent Semantic - Latent Dirichlet Allocation) model to a spatially embedded social network of travelers in San Francisco. The project addresses three challenges in transportation demand modeling: 1) the need for up-to-date data and model estimates, 2) a desire to identify latent classes of travelers, and 3) addressing the open problem of predicting demand for discretionary activities.

This method supports our on-going effort to enable the use of large-scale passively collected data to build temporally up-to-date travel demand models. Informed transportation policy, planning, and operations decision making requires accurate forecasting of travel demand. Traditional data collection methods involve costly one-off travel surveys. Though the depth of data collected is great, these surveys are expensive, involve small samples, and are conducted very infrequently. The most recent household travel survey of the nine counties of the San Francisco metropolitan region was conducted in 2000 (BATS, the Bay Area Travel Survey). This data set still serves as the modeling baseline for many county and city agencies (though the data are augmented with even smaller project-specific studies). The effects of the Great Recession of the last seven years are completely missing from the BATS data! Given the fast pace that economies and land-use evolve, especially in the Bay Area, transportation practitioners need more temporally relevant data sources. Passively collected signals, such as geo-tagged social media posts and cellular network data, provide a continuous stream of travel behavior information. However, they are not appropriate for traditional travel demand models, which require detailed socio-demographic data for individuals. We demonstrate that L-PLSA-LDA is a viable method for extracting meaningful spatial behaviors from these types of data.

Paper submitted: December 4, 2015

Travel demand modeling is rooted in the microeconomic theories of random utility maximization and the methods of discrete choice modeling [1]. The state of the art in travel demand modeling incorporates latent classes of travelers within the discrete choice framework [2]. These latent "modality styles" are manifestations of a person's preferences, socio-demographics, and other unobserved factors. The modeling incorporated in this project also reveals latent communities (topics) of travelers. They are defined by the places where a person and his or her social network peers are observed.

Finally, an open problem in transportation is accurately predicting demand for discretionary (non-work, non-education trips) activities. Thanks to the wealth of population and economic census data, the home-work commute patterns are fairly well understood (though again, in the fast-moving Bay Area, even the 1-5 year American Community Survey samples are outpaced by the economy). The fact that home and work locations and schedules tend to be relatively fixed also helps stabilize commute patterns. For activities like shopping, eating, and leisure, travelers have much looser constraints on the choice of timing and location. Consequently, traffic volume predictions for midday, evening, and weekends is challenging. We hypothesize that in the context of discretionary activity location choices, the influence of a traveler's social network plays a significant role. The L-PLSA-LDA formulation captures this influence.

II. Previous Work

The model we employ, L-PLSA-LDA, was introduced by Nallapati and Cohen, 2008 [3]. It builds on previous efforts to model document topics and hyperlinks between relevant documents.

The PLSA model, Figure 1, was introduced by Hofmann in 1999 [4]. It is a factored model where each document, a bag of N words, is produced by a mixture of source topics, π. Each word is conditioned on a unique latent topic, z. Thus

a document does not belong to any single topic, but is a combination of topics. In the PLSA model, all documents in a corpus follow the same mixture of topics, π.

In 2003, Blei et al. introduced LDA, Figure 2, a model for corpora of related documents. In their formulation, each document follows a unique topic mixture, θ. The mixture follows a Dirichlet distribution, conditioned on the parameter α. Words, w, are conditioned on a set of multinomial parameters, β, selected by the topic, z.

The L-PLSA-LDA model, Figure 3, introduced by Nallapati and Cohen in 2008 combines the PLSA and LDA models in order to capture hyperlinks between related documents [3]. The Link-PLSA and Link-LDA models, which were developed earlier, also shared this objective [5], [6]. The main innovation of the L-PLSA-LDA model is that it allows for a flow of influence from the topic of the cited document back to the citing document. In L-PLSA-LDA, the citing documents obey a process identical to LDA except that in addition to a bag of words, a document also contains a bag of outgoing directed links, L. The probability of linking to a cited document is a function of the topics of the words in that document.

III. L-PLSA-LDA: From Hyperlinked Documents to Spatially Embedded Social Networks

We reframe the semantics of the L-PLSA-LDA model to describe the locations and connections of a spatially embedded network of people. In Table I, we define the meaning of the variables and notations in the context of our project. The original meanings are in parentheses.

A person (document) is defined by the bag of locations (words) where he or she is observed, and the links to other people. Unique locations are represented by categorical identifiers.There is *no measure of space* in the model's representation. Social network links are directed from from the citing person (document) to the cited friend (document). Latent communities (topics) are a function of the locations a person and his or her friends frequent. We believe that these communities reflect preferences and socio-demographic factors. The number of communities is fixed by the modeler.

Without considering social influence, the probability of visiting a location is conditioned on the latent community (topic) that the person is operating in. For example, I am a father, a graduate student, and an outdoors enthusiast. I can be viewed as a mix of three communities. From the modeling perspective, these personal aspects of myself are not known, so the communities are latent. The set of locations I am likely to visit is function of which community I am operating under. Since I go to UC Berkeley, I spend most of my time in the student community. In this mode, my activity locations are constrained to the campus and my home. On Saturdays, I am in the parent community and my location set shifts to capture playgrounds, parks, and again, my home. On on the rare occasion I can act as an outdoors community member, my location set ranges over a variety of parks and wilderness areas.

We believe that location of discretionary activities is also influenced by the preferences and communities of a person's friends. Since I am graduate student, it is likely that some of my friends and colleagues are also students. But perhaps one of them attends Stanford instead of UC Berkeley. This might expand my location set, when operating in the student community, to include the occasional visit to Stanford for a lecture or meeting.

TABLE I. L-PLSA-LDA Variables and Notation

α	Dirichlet parameter	d	person (citing document)
θ	community mixture (topic mixture)	d'	person (cited document)
z	community (topic)	Ω	link prob' conditioned on community
w	location (word)	β	word prob' conditioned on community
π	community (topic) probability	$N_{\leftarrow}$	number of locations (words) observed for cited person (doc)
$M_{\rightarrow}$	number of citing people (docs)	$M_{\leftarrow}$	number of cited people (docs)
N	number of locations (words) observed for citing person	L	number of links from citing person (doc)

IV. Method

A. Data Synthesis

The data were synthesized by querying the Twitter API for geotagged posts in San Francisco. Directed social network links were created between users when they followed someone else in our sample. A training set of 2,058 users was generated with 6,072 internal links. A test set of 272 users was withheld.

The users' home and work locations were identified by fitting a bi-modal Gaussian. Other locations were mapped to venues using the Google Places API. Thus, locations were mapped to venue identifiers [7].

B. Model Implementation

We used an open source C-package published by Nallapti for L-PLSA-LDA estimation, inference, and prediction [8]. This software extends the original LDA package published by Blei [9].

Closed form solutions for estimation and inference do not exist for this model, and the likelihood step of the EM algorithm is computationally intractable. Thus, both Blei's and Nallapati's software use variational approximation to estimate bounds on the likelihood [10]. This is accomplished by simplifying the structure of the graph, Figure 4. For a full description of the variational approximation, please refer to Nallapati and Cohen, 2008 [3].

For the purposes of this project, it is important to note that the γ term in Figure 4 is proportional to the posterior probability of community membership for a citing person. One of the outputs of the variational estimation is a $2,058 \times k$, Gamma matrix. The rows index people and the columns index communities. Each row is a vector describing the mixture of community membership for a person. These are the values we use to describe community membership in the following sections.

V. Results

The number of latent communities is fixed by the modeler. We estimated and tested L-PLSA-LDA for a range of two to fifty communities. A measure of the predictive power of the model is the log-likelihood of cited and citing populations. In Figure 5, we observe that the log-likelihood of the held-out test data increases with the number of latent topics, but that rate of increase decays with the number of topics. We expect that for a very large number of communities, we will see the log-likelihood decrease due to over-fitting. Unfortunately I was not able to test this due to overheating of my laptop.

As mentioned in the previous section, we can use the γ output to measure the mixture of community membership. In Figure 6 and Figure 7 we created scatter plots of the home locations of members of the training data and colored them by likelihood of community membership, as defined by γ. For Figure 6, we estimated three communities. For Figure 7, we estimated 15 communities and plotted a sample of membership in three of them. Intuitively, we see that the spatial extent of the community decreases with the number of communities.

VI. Limitations and Future Research

There are several aspects of the L-PLSA-LDA model that conflict with our understanding of real spatially embedded social networks and travel behavior. An obvious weakness is that the model imposes a bipartite graph separation between cited and citing documents. Fortunately there does exist a workaround: we include the full sample in both the citing and cited data sets. However, this can introduce some undesirable artifacts such as loop links between a person and his or herself.

Another weakness is that population of citing and cited people is fixed. Transportation modelers often need to forecast future scenarios where the population has grown. McGrath and Pozdnukhov, 2014, do propose a generative algorithm to allow a modeler to grow the population and their social networks [7].

The biggest weakness in my mind is that this model uses a bag of locations approach. To model traveler volumes, we need to know where people are going as well as when they are going there. The question of timing is partially a function of the sequence of activities. Certain types of locations are more likely to be visited proceeding another type, depending on the time of day. For example, if a person is observed at home in the morning, there is a high likelihood the next observation will be at work.

One of the reasons topic modeling often takes a bag of words approach is that the set of syntactically valid word sequences is combinatorially intractable. However, the viable sequences of locations for an individual is relatively small. This is especially true if we redefine locations as the types of activities that are conducted. I am currently researching ways to extend the L-PLSA-LDA to use semi-Markov techniques for building location models that have some memory of recent history.

VII. Conclusion

The most striking result of this project is that there is a strong spatial structure for the latent communities, *even though space is abstracted out of the model*! As mentioned previously, locations enter the model as categorical indicators. Measures of proximity, neighborhoods and districts are completely absent. Yet we see they emerge through the latent community detection. In Figure 6, the left most community seems to radiate out from the Mission District. The middle community is centered on the Financial District, and the rightmost captures residential neighborhoods that encompass the rest of San Francisco.

Although we have not yet applied this model to transportation demand modeling, the results are promising. As mentioned in the introduction, the start of art in travel demand modeling involves identifying latent classes of travelers based on preferences and lifestyle. At the risk of gross generalizations, we know there is a "type" that lives in the Mission District: young urbanite, technology oriented, bicycle, pedestrian and transit friendly. The Financial District caters to the banker types: middle aged, wealthy, auto-dependent. These are the same types of profiles created by state of art latent class discrete choice models for transportation demand.

Thus, we have demonstrated that there is reason to believe passively collected can be used to build travel demand models on par with the state of the art. Reflecting on the objectives stated in the opening paragraph of this paper, we see that data that be collected daily, as opposed to the rare and expensive one-off small samples normally used, may produce similarly useful models. The results of the latent community detection also appear meaningful. It is less clear how effective this technique is for discretionary activities. We have taken a first step in implementing a location model that includes social network effects. Future research is required to evaluate the efficacy of using L-PLSA-LDA for discretionary activity location modeling.

VIII. Authors' Contributions

Professor Alexey Pozdnukhov has acted as research adviser to both Rory McGrath and myself, contributing guidance in theory and analysis. Rory McGrath conducted the data collection and formatting. I conducted all other aspects mentioned in this paper: background research, configuring and running the L-PLSA-LDA package, graphics creation and writing.

It should be noted that some of my work for the class project overlaps McGrath and Pozdnukhov, 2014 [7]. Since the conference at which that work was presented, Rory McGrath left UC Berkeley and most of his work was lost due to a failed hard drive (and of course, not backing up). Only a copy of the original data files he created remained.

REFERENCES

[1] D. McFadden, "Conditional logit analysis of qualitative choice behavior," 1973. [Online]. Available: https://elsa.berkeley.edu/reprints/mcfadden/zarembka.pdf

[2] A. Vij, A. Carrel, and J. L. Walker, "Incorporating the influence of latent modal preferences on travel mode choice behavior," *Transportation Research Part A: Policy and Practice*, vol. 54, pp. 164–178, Aug. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0965856413001304

[3] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 542–550. [Online]. Available: http://dl.acm.org/citation.cfm?id=1401957

[4] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57. [Online]. Available: http://doi.acm.org/10.1145/312624.312649

[5] D. C. T. Hofmann, "The missing link-a probabilistic model of document content and hypertext connectivity," in *Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems. The MIT Press*, 2001, pp. 430–436.

[6] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5220–5227, 2004. [Online]. Available: http://www.pnas.org/content/101/suppl$_1$/$5220.short$

[7] R. McGrath and A. Pozdnukhov, "A generative model of urban activities: simulating a population." [Online]. Available: http://www.rorymcgrath.ie/papers/simulate$_communities.pdf$

[8] "Linked-pla-lda software - Ramesh Nallapati's homepage." [Online]. Available: https://sites.google.com/site/rameshnallapati/software

[9] "Blei-Lab/lda-c." [Online]. Available: https://github.com/Blei-Lab/lda-c

[10] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1561/2200000001

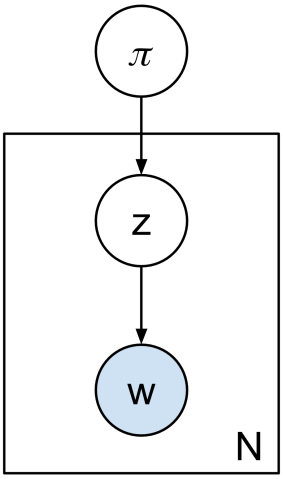

Fig. 1. PLSA model

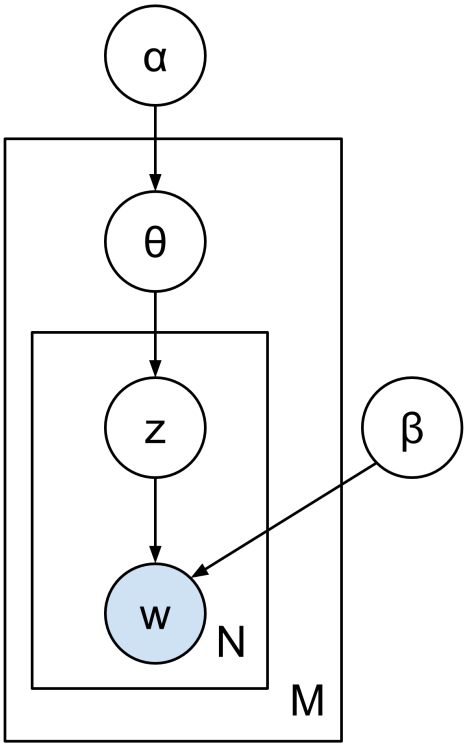

Fig. 2. LDA model

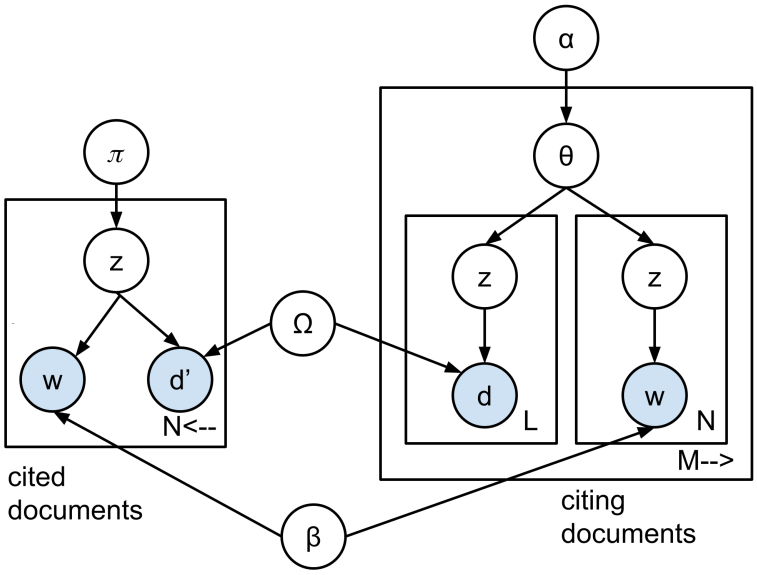

Fig. 3. LDA model

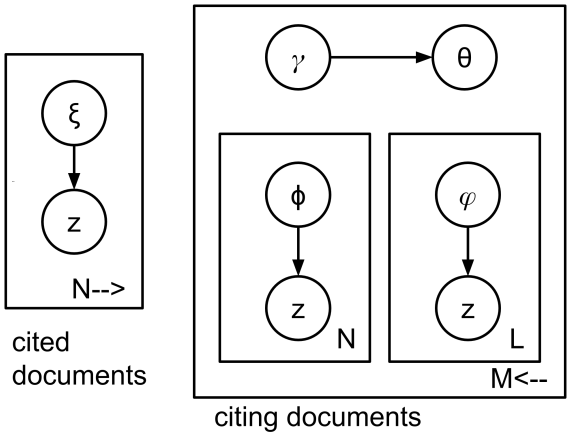

Fig. 4. Variational Estimation of L-PLSA-LDA

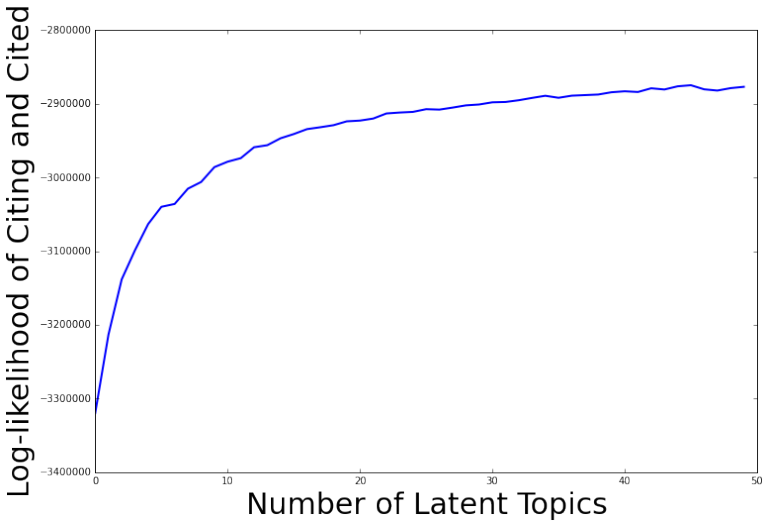

Fig. 5. Latent Community Detection in San Francisco

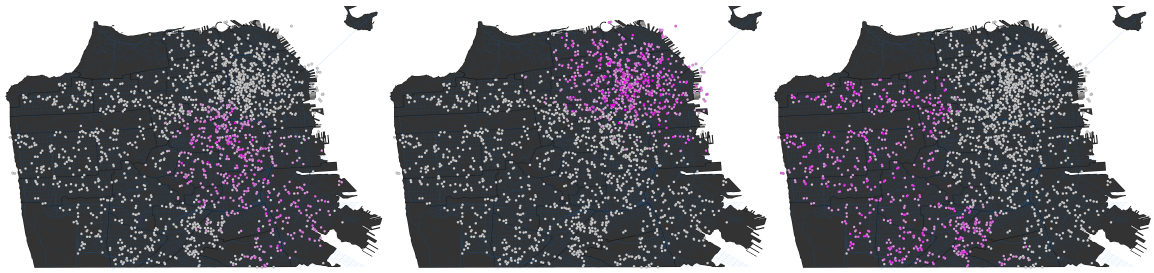

Fig. 6. Latent Community Detection in San Francisco: 3 out of 3 Communities

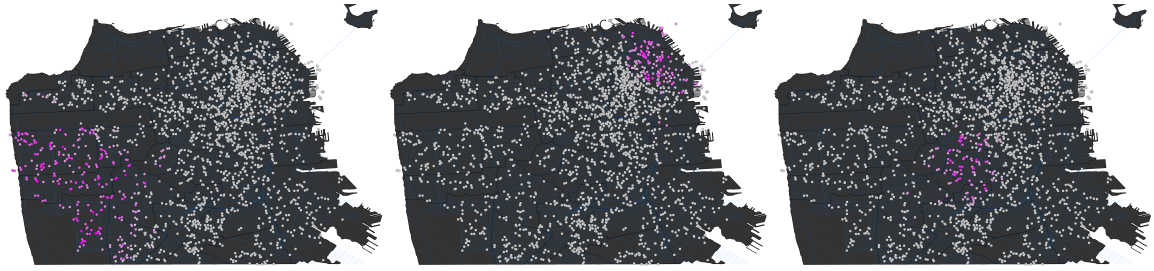

Fig. 7. Latent Community Detection in San Francisco: 3 out of 15 Communities