NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Temporal decomposition and semantic enrichment of mobility flows

C. Coffey, B.Sc.

Supervisor: Dr. Alexei Pozdnoukhov

Thesis submitted in fulfilment of the requirements for the degree of Master of Science by Research

National Centre for Geocomputation, Faculty of Science
National University of Ireland, Maynooth
Maynooth, Ireland

September, 2013

Abstract

Mobility data has increasingly grown in volume over the past decade as localisation technologies for capturing mobility flows have become ubiquitous. Novel analytical approaches for understanding and structuring mobility data are now required to support the back end of a new generation of space-time GIS systems. This data has become increasingly important as GIS is now an essential decision support platform in many domains that use mobility data, such as fleet management, accessibility analysis and urban transportation planning. This thesis applies the machine learning method of probabilistic topic modelling to decompose and semantically enrich mobility flow data. This process annotates mobility flows with semantic meaning by fusing them with geographically referenced social media data. This thesis also explores the relationship between causality and correlation, as well as the predictability of semantic decompositions obtained during a case study using a real mobility dataset.

Dedication

This thesis is dedicated to:

my family; for teaching me to climb high,

my friends; for catching me when I fall,

my love; for encouraging me to do it all over again.

Acknowledgements

This thesis would have been impossible had I not been surrounded by people much smarter than myself. I would like to thank the following individuals for their support and encouragement of the past few years.

Alexei Pozdnoukhov for teaching me to read, write and navigate the academic world. For never closing his office door. For midnight paper submission deadlines and last minute corrections. For teaching me everything I know about Machine Learning.

Carson Farmer for being a father figure to all research students in the lab and for his friendship, advise and honest feedback at all times.

Fergal Walsh for so many things: introducing me to the team, teaching me that mapping can be cool, for i2maps, pico and dropmap. For always being one step ahead and leading the way.

Felix Kling for Javascript, JSNetworkX, long nights playing Munchin and driving me all the way from L.A to San Francisco.

Aodhan (my twin brother) for always running along beside me and encouraging me to keep up. For keeping me sane during the most difficult of times. For arcade machines, robosumo and flying machines.

My partner Cristina for finding me when I get lost. For always encouraging me to continue. For doing something much more difficult than me and still finding time to help. For being a constant source of inspiration.

Funding

Research presented in this thesis was funded in part through the Strategic Research Cluster (SRC) grant (07/SRC/I1168) by Science Foundation Ireland (SFI) under the National Development Plan (NDP), and an IBM PhD Fellowship award. This support is gratefully acknowledged.

List of previous publications

Portions of the work discussed in this thesis have previously been published in the following peer reviewed papers;

- Nair R., Coffey C., Pinelli F., Calabrese F., Large-Scale Transit Schedule Coordination Based on Journey Planner Requests, 92nd Annual Meeting of the Transportation Research Board 2013.
- Coffey C., Nair R., Pinelli F., Pozdnoukov A., Calabrese F., Missed Connections: Quantifying and Optimizing Multimodal Interconnectivity in Cities, Computational Transportation Science workshop at 20TH ACM SIGSPATIAL GIS'2012 2012.
- Lawlor A., Coffey C., McGrath R., Pozdnoukhov A., Stratification structure of urban habitats, Pervasive Urban Applications workshop at PERVASIVE'2012, 2012.
- McGrath R., Coffey C., Pozdnoukhov A., Habitualisation: localisation without location data, Nokia MDC challenge at PERVASIVE'2012, 2012
- Coffey C., Pozdnoukov A., Calabrese F., Time of Arrival Predictability Horizons for Public Bus Routes, Computational Transportation Science workshop at 19TH ACM SIGSPATIAL GIS'2011, 2011

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

Geographic Information Systems (GIS) have become essential decision support platforms in many domains that use mobility data, such as vehicle fleet management, accessibility analysis and urban transportation planning. Over the past decade these domains have accrued immense collections of mobility data. Now more than ever, novel analytical approaches for understanding and structuring mobility records are required to process these ever growing volumes of data.

This thesis advocates semantic enhancement of mobility flows as an essential component for next generation GIS. One example application from the domain of urban mobility analysis is the ability to infer trip purpose from aggregate data. This application would produce reliable volume estimates for different user groups, such as commuters, leisure travellers or tourists. This decomposition of aggregate into multiple trip purpose would simultaneously benefit both transportation operators and passengers. The operator can now optimise transportation operation at a systems level and the passenger can now benefit from a host of newly enabled context-aware smart mobility services.

A common type of mobility data is a record of the number of trips between predefined spatial zones known as origin-destination (OD) matrices. The same type of data is generated by the majority of urban transportation systems which log the number of trips between stations on their network from ticket sales or via swipe cards records. Despite the fact that OD tables are one of the main type of mobility data and that they have been studied in transportation science for decades, there is a surprising knowledge gap in methods that allow inferring trip purpose from i to j from non-direct mobility observations and/or other related data. In the transportation domain, trip purpose data is collected by conducting costly and time-consuming travel surveys. In geographic information science and recent data analytics studies, trip purpose identification methods have only been proposed for detailed records such as GPS tracks where exact types of origin destination locations can be defined with sufficient certainty (Andrienko et al., 2011). At the same time, the limits of applicability of such methods are constrained by location privacy considerations. Spatially aggregated data is often used instead of precise GPS locations, increasing the need for relevant methods even more.

1.2 Methodology

This thesis develops a machine learning methodology for the decomposition of mobility flows, provided in the aggregate form of origin destination matrices, into multiple modes of specific trip purpose.

Semantic enrichment is made possible by considering the space-time context of mobility flow modes as well as using available geo-referenced social media data for an overlapping period. This overlapping social media data details both periodic and non-periodic activities that influence the mobility of individuals living in the region of study.

The causality relations and predictability across two datasets (mobility and social media) are investigated to support the semantics assigned to each mode of specific trip purpose.

1.3 Contributions

The main contribution of this thesis is that it resolves a common drawback of intuitive ad-hoc semantic annotation of the obtained modes solely from space-time context. Our approach to the issue is based on causality analysis (Granger, 1969), as well as predictability of the obtained semantic decompositions of mobility flows. Using a real mobility dataset from a bike sharing network, we quantitatively show that temporal and/or spatial coincidence of seemingly related processes is not sufficient to attribute semantic labels across datasets. We also build a conditional dependency graph to reason about trip purposes and advance data-driven predictive systems (Breiman, 2001) of urban dynamic processes.

The advantages of our approach are twofold. First, it increases the credibility of semantic annotation of mobility flows. Secondly, it enhances the predictability of the components of mobility flows related to specific trip purposes. It can therefore be used for better network management and optimization as well as for providing new location-based and activity-aware services to the users of a new generation of smart transportation networks.

1.4 Case Study

In this thesis, we present a case study using two real world datasets (mobility and social media) from the city of Washington D.C, US. The case study re-enforces our claim that semantic enhancement of mobility flows is a powerful and important technique that belongs in the next generation of GIS. Al-

though the case study utilises a specific mobility dataset; a bike sharing dataset provided by Captial Bikeshare (see Section 2.1). Every attempt has been made in this thesis to frame the technique in a more general light. The methodology outlined can be used to decompose and semantically label almost any origin-destination based mobility dataset into multiple modes of specific trip purpose.

1.5 Thesis Structure

Chapter 1: Introduction

In this chapter we frame our work in the context of mobility analysis. We outline the main goal of the thesis, the methodology followed, contributions to the field and finally a case study preformed using two dataset (mobility and social media) from the city of Washington D.C, US.

Chapter 2: Data

This chapter describes two datasets (mobility - Capital Bikeshare) and (social media - Twitter) used throughout this thesis. The mobility dataset spans a much longer time period and so it is described in much more detail. We analyse the mobility network's growth over time, the effects of weather and tourism, we describe the results of a decomposition experiment which motivated the work and, ultimately, this thesis. Finally we present some basic statistics about the social media dataset.

Chapter 3: Methods

This chapter begins with a brief introduction to the field of topic modelling. After this, we explain in detail an algorithm used later in the thesis, Latent Dirichlet Allocation (LDA). We offer some intuitions about LDA, then we describe its formulation, first using a graphical model and then more formally using strict mathematical notation. We complete this chapter

with a practical discussion on fitting the LDA model to a real dataset using approximate posterior inference.

Chapter 4: Temporal Decomposition

In this chapter we decompose the abstruse mobility flows of the Capital Bikeshare dataset into simpler temporal components using LDA. We describe the process of creating documents from records, then we preform an initial decomposition and discuss the results.

Chapter 5: Semantic Labelling

In this chapter we annotate, with semantic meaning, the temporal decomposition preformed in the previous chapter. We then explore both the spatial and textual components of the decompositions.

Chapter 6: Correlation & Causation

This chapter discusses the important difference between correlation and causation. We show that even though many of the Twitter and bikeshare topics are highly correlated only some of them exhibit a true causal relationship. We use the information gained by this analysis to semantically annotate the mobility topics.

Chapter 7: Towards Demand Forecasting

This chapter details a proof of concept framework for predicting OD demand flows. We show that one can predict, using a neural network, bike rental topic intensities from Twitter topic intensities. We the show how these predicted topic intensities can be used to forecast mobility flows.

Chapter 8: Discussion and Conclusions

In the final chapter, we discuss the implications of our work and the limitations of our technique.

Chapter 2

Data

2.1 Data

To make our results reproducible we provide all of the data used by this thesis. We also provide high resolution versions of all figures.[1].

2.2 Capital Bikeshare

As bike sharing systems gained popularity with city dwellers, the data produced by bike sharing attracted the attention of researchers studying human mobility (Padgham, 2012; Montoliu, 2012; Borgnat et al., 2011). The bikeshare dataset used in this thesis is a subset of the historic trip data provided by Capital Bikeshare[2]. Capital Bikeshare package and release a complete dataset from their network every quarter.

2.2.1 System growth over time

At the time of this writing, Capital Bikeshare had released 9 quarters of data from their bikeshare network. This data spans from the final quarter of 2010

[1] `https://github.com/ccoffey/masters-thesis`
[2] `http://www.capitalbikeshare.com/trip-history-data`

until the final quarter of 2012. Figure 2.1 plots the number of bike rentals per month for the complete 28 month dataset. This figure clearly depicts increasing popularity in the Capital Bikeshare network. Each year the number of rentals across the network increased dramatically. One can also see from this figure the existence of a macro level trend. This trend is likely being caused by one of two things: seasonal change in weather or tourism.

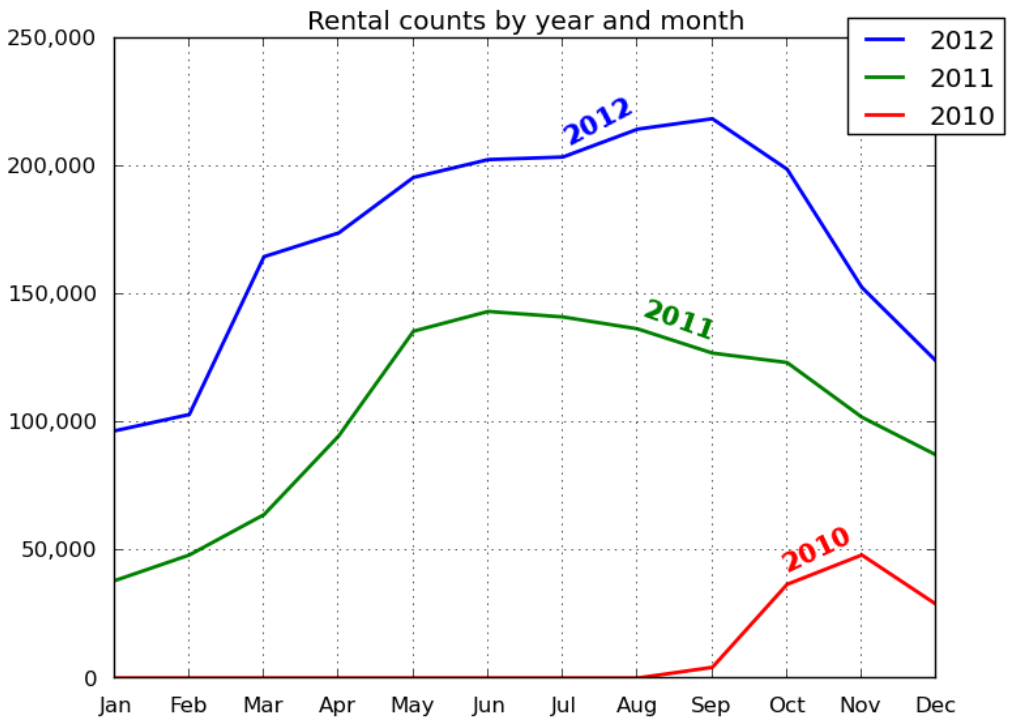


Figure 2.1: Number of bike rentals per month for the complete 28 month dataset.

The Capital Bikeshare network is not a static network. To meet growing popularity, new stations and bikes have been added over time (see Figure 2.2). Not surprisingly these two graphs display the same overall trend at different magnitudes. This is the case because a new station is never added without also adding new bikes to service it.

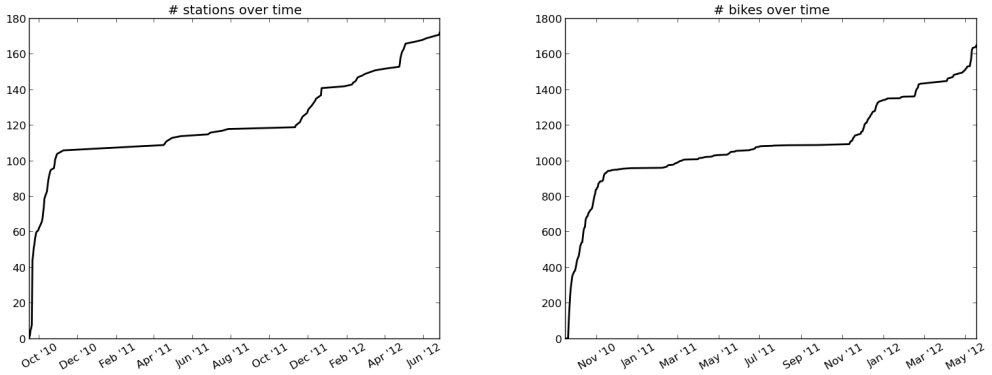


Figure 2.2: Station growth over time (left) and bike growth over time (right).

2.2.2 Tourism and Weather

It is important to understand the effects of both weather and tourism on the bike sharing network as they may have profound implications on the mobility flows of the underlying network. Figure 2.3, a decomposition of rentals into registered and casual, was made possible by the member type category collected by Capital Bikeshare. A registered user is one who pays for the bike system on an annual or monthly basis. A casual user is one who purchases shorter term access (1 to 5 days) membership. According to Capital Bikeshare, the casual user membership is primarily utilised by tourists. Figure 2.3 seems to support this claim as the Casual rental curve is highest in the summer months: May, June and July.

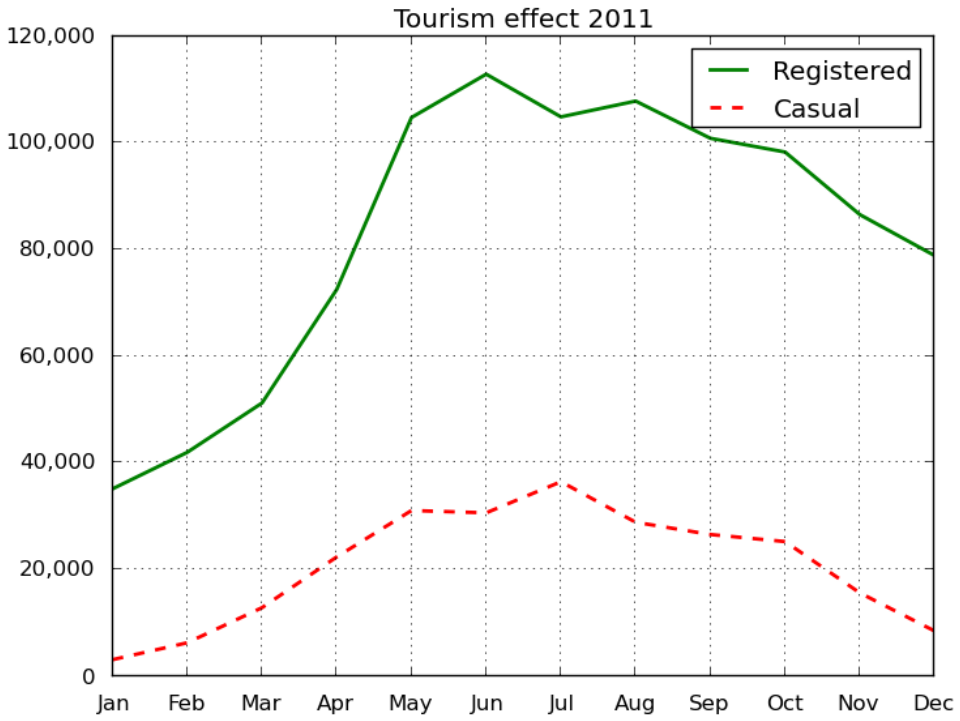
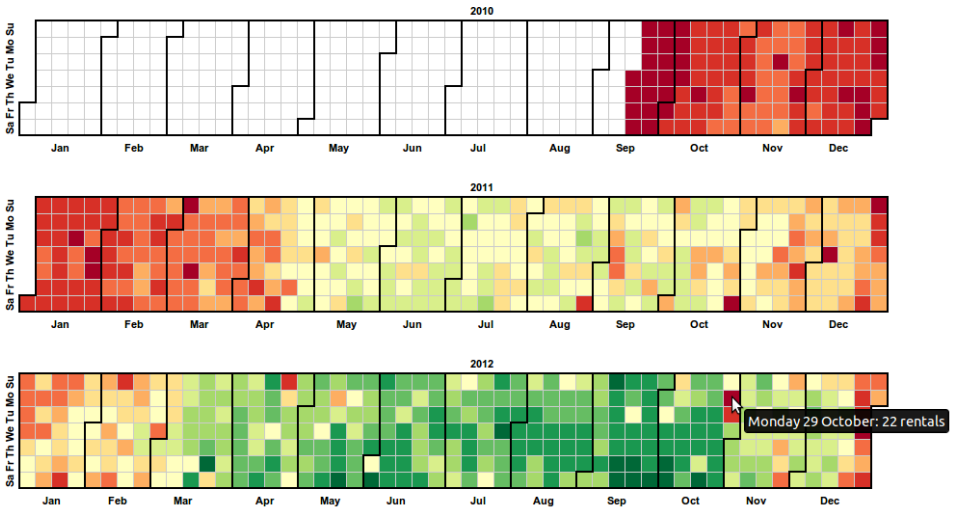


Figure 2.3: 2011 rental counts split into two components registered users (green) and casual users (red)

To investigate the effects of weather on bikeshare, we built an interactive calendar application (see Figure 2.4). This application displays the entire Capital Bikeshare dataset in a unique and informative manor. Each square in the calendar represents a single day. Days are arranged into columns by week, then grouped by month and year. The color of each square ranges from dark red to dark green and represents low to high rental counts respectively. Note that in Figure 2.4, a dark red square has been highlighted displaying the label "Monday 29 October: 22 rentals". There were only 22 rentals on this day, an incredibly low number considering this includes all 191 bike stations in the Capital Bikeshare network. In fact on the previous day the number of rentals was 4,460. The reason for this incredibly low rental count was Hurricane Sandy which hit Washington D.C, US on Monday October 29th 2012. The weather table at the bottom of the figure confirms this by displaying an average wind

speed of 39 kmph and a total rainfall of 97.79 mm. This was by far the highest recorded (wind speed and rainfall) for the entire dataset. This calendar allows us to easily identify rental anomalies and corresponding weather information. We have found by inspection that bad weather effects rental counts much more on weekends than weekdays. This suggests that cyclists who routinely use the bike network on weekdays, perhaps for commuting purposes, are not deterred by bad weather.



Mean Temp	Max Temp	Min Temp	Dew Point	Mean Hum	Max Hum	Min Hum	Rain	Snow	Snow Depth	Pressure	Wind	Max Wind	Max Gust	Vis
11 C	13 C	8 C	8 C	87	93	80	97.79 mm	0.00 cm	0.00 cm	986 hPa	39 km/h(NW)	66 km/h	98 km/h	7 KM

Figure 2.4: An interactive calendar application displaying rental counts and weather information for the entire 28 month Capital Bikeshare dataset.

A note to the reader: An interactive version of the calendar application is available online [3].

2.2.3 Station Clustering

Weekdays on the bikeshare network are very different from weekends. This becomes very obvious when you plot the total number of rentals, per hour and per day. In the rainbow plot (see Figure 2.5), it is clear that Monday

[3] `http://ccoffey.github.io/bike-weather/index.html`

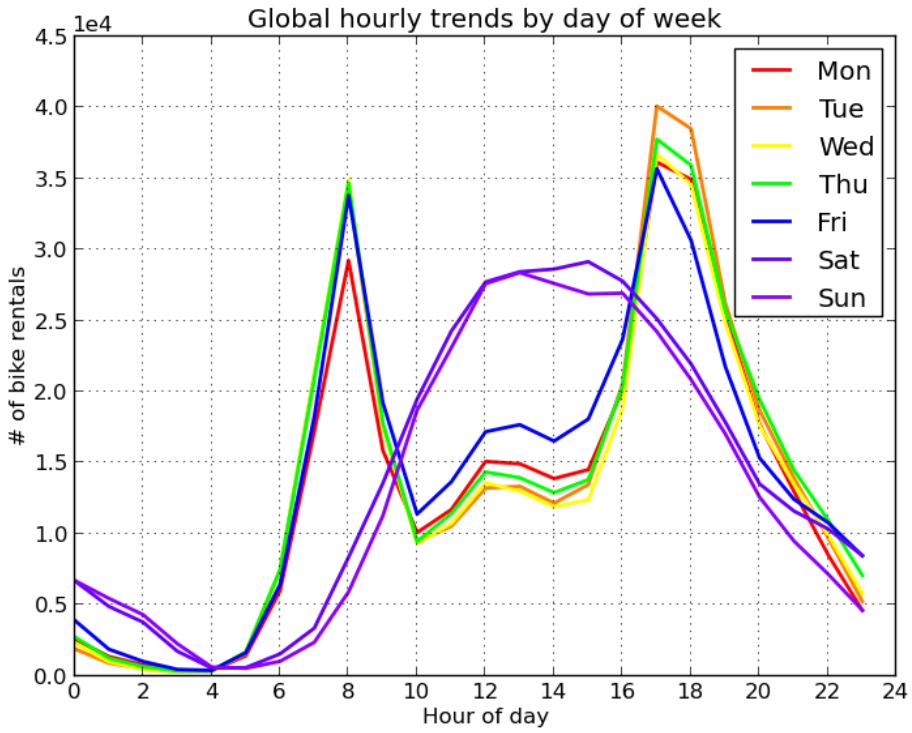


Figure 2.5: Global hourly trends by day of the week

to Friday are incredibly similar. Two sharp peaks are present, the first at 8am (most likely the morning commute) and another at 6pm (most likely the evening commute). Lunchtime (12 to 1) also shows increased activity. The weekend on the other hand is a completely different graph, there are no sharp peaks, simply a constant increase in activity until midday and then a constant decrease.

The rainbow plot presented in Figure 2.5 is an average taken over all stations in the network. In this sense, the plot is slightly misleading as it suggests that this is the average profile for a station in the network. To show that this is not the case, we generated an average 24 hour profile for each of the 191 stations and then clustered the results using **kmeans** with parameter $k = 3$.

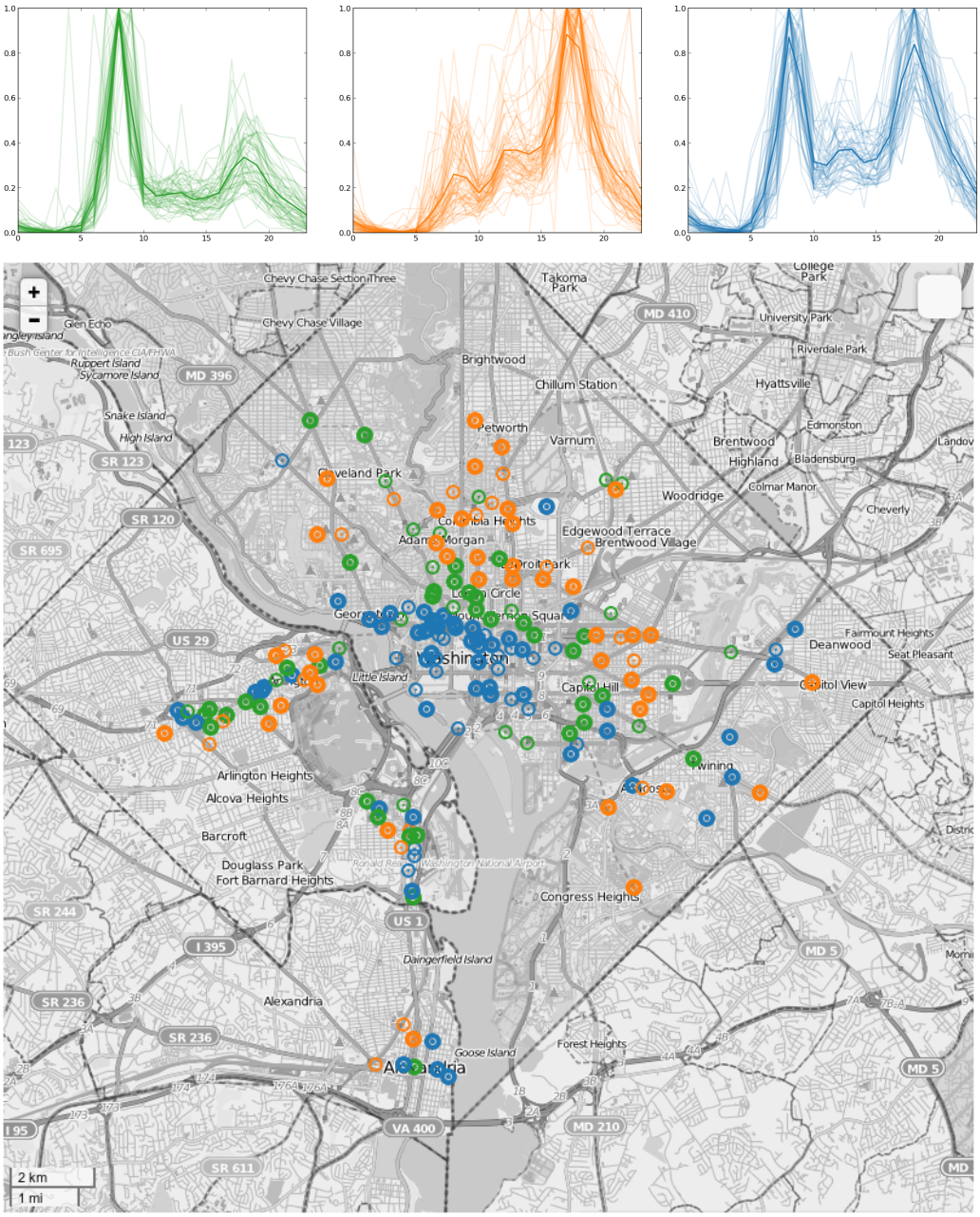


Figure 2.6: Clustering results of kmeans with parameter k=3. Station profiles (top) and stations positions (bottom).

The results of this clustering are shown in Figure 2.6. Plotted at the top of this figure are the 191 station profiles, each assigned to one of 3 clusters. The first of these clusters contains stations which are very active in the mornings but not at any other time. The second of these clusters contains stations which are very active in the evenings but not at any other time and the third cluster

contains stations that are active both in the morning and evening. Figure 2.6 largely motivated the rest of this thesis. This decomposition of an aggregate network signal, into station specific signals inspired many questions. Why do these stations have such different temporal profiles? What are the underlying motivations influencing mobility on the bike sharing network? Are the green and orange clusters morning and evening commutes? 3 was an arbitrary choice for k, how many significant components is the aggregate network signal composed of? Is time-series clustering the best way to preform this decomposition? If we do extract n distinct components, how can we then assign semantic meaning to these components?

2.2.4 Case Study Data

In our case study we use a two week subset of the complete Capital Bikshare dataset described above. Unfortunately we are forced to use only a subset due to limitations in the second dataset (Twitter) which is described below.

The case study dataset contains 43,636 unique bike rentals completed on a 191 station network (see Figure 2.7) during a one week period (2012-07-16 to 2012-07-23 inclusive). The dataset contains a timestamped record of every trip between a pair of rental stations. Although the final destinations of bike users are unknown, it is reasonable to assume that destination venues are located in the vicinity of rental stations. The trip records also contain a significant amount of self-loops, potentially corresponding to leisure trips and recreational cycling.

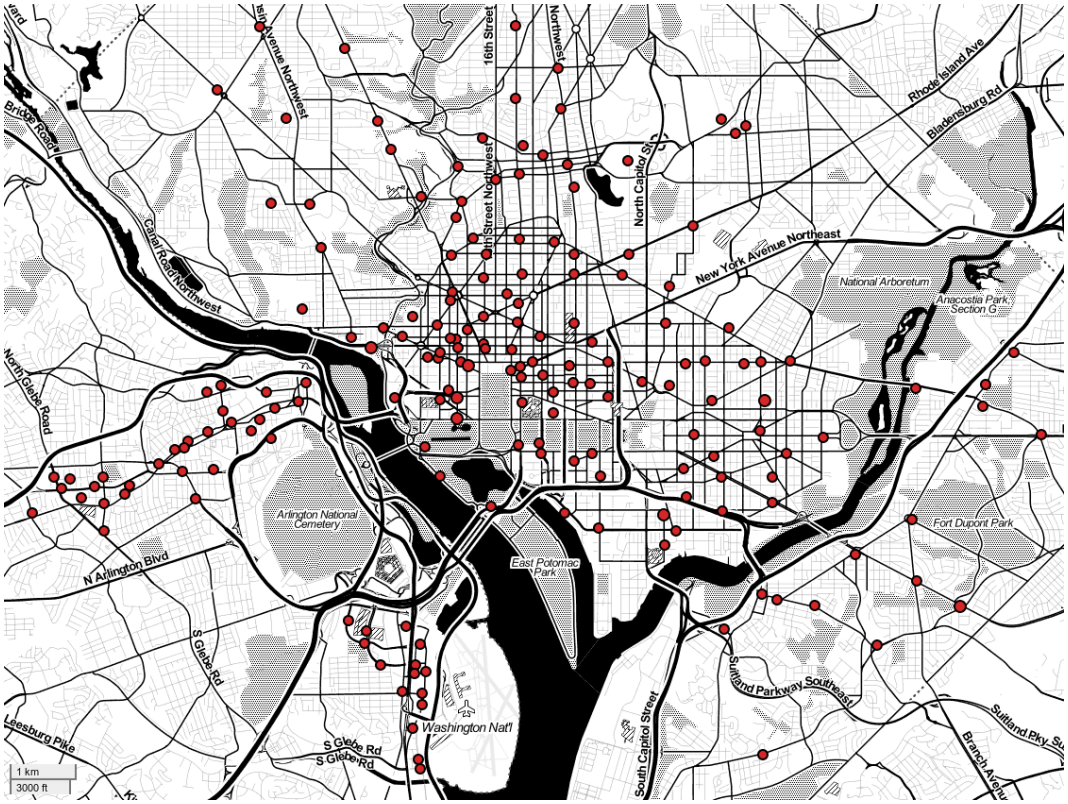


Figure 2.7: Capital Bikshare network map: locations of 191 stations in Washington D.C, US

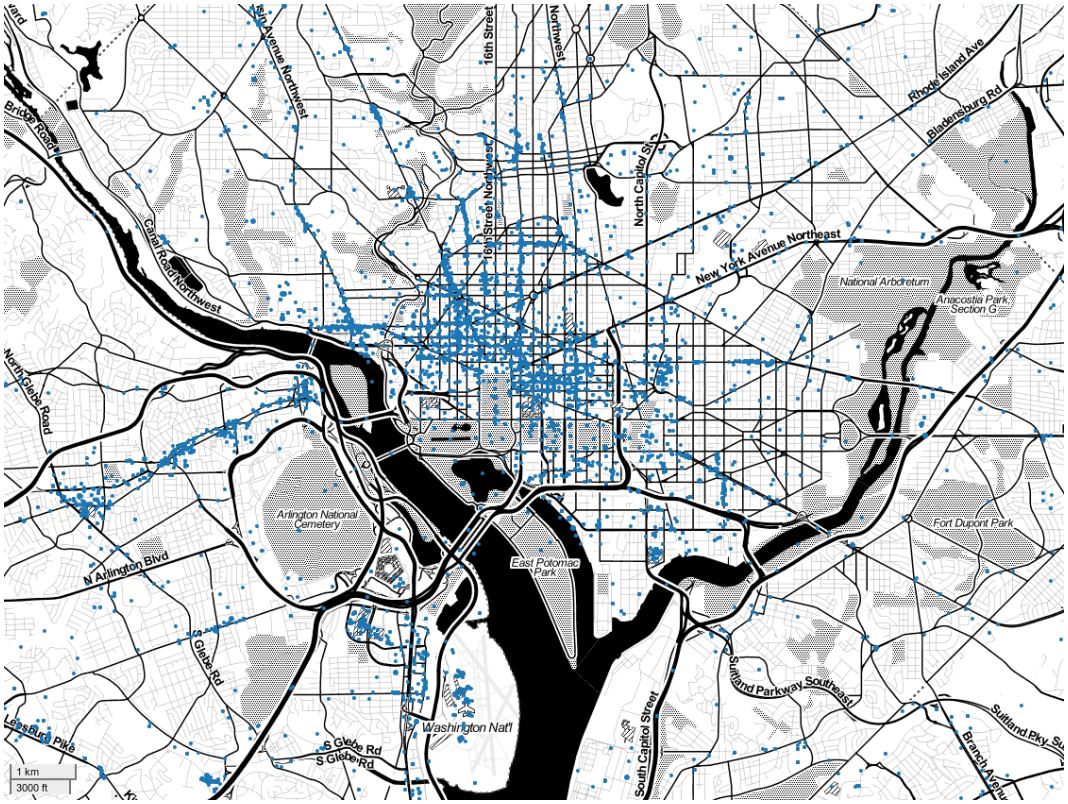


Figure 2.8: 37,335 tweets, generated by 9,170 unique users, during the 7 day period (2012-07-16 to 2012-07-23 inclusive).

2.3 Twitter

The Twitter dataset used in this thesis was collected using the Twitter streaming API [4]. This dataset contains 37,335 geo-tagged tweets. These tweets were generated by 9,170 unique users, checking into 15,860 unique venues during the same one week period as the bikeshare dataset (2012-07-16 to 2012-07-23 inclusive). The locations of check-ins are presented in Figure 2.8. Check-in messages are popular data sources to determine the semantics behind user activities (Kling & Pozdnoukhov, 2012; Lian & Xie, 2011; Ye et al., 2011) and detect significant events (Pozdnoukhov & Kaiser, 2011) as they contain user generated content and venue information.

[4] `http://dev.twitter.com/docs/streaming-api`

Chapter 3

Methods

3.1 Topic modelling

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the latent 'topics' that pervade a collection of documents. By discovering patterns of word use and connecting documents that exhibit similar patterns, topic models have emerged as a powerful new technique for finding useful structure in an otherwise unstructured collection. Topic models operate under the belief that documents are created by a generative process. This 'imaginary' process constructs new documents in the following way. First, it chooses a distribution over topics. Then, for each word, it chooses a topic at random according to this distribution, and draws a word from that topic. Topic modelling algorithms attempt to invert this 'imaginary' process; inferring the set of topics that were responsible for generating a collection of documents.

Figure 3.1 shows four example topics that were derived from the TASA corpus, a collection of over 37,000 text passages from educational materials (e.g., language & arts, social studies, health, sciences) collected by Touchstone Applied Science Associates (see Landauer et al. (1998)). For each of the four

topics depicted, only the 16 most probable words are displayed. By examining these words, one might conclude that *Topic 247* is about medicine, *Topic 5* is about colors, *Topic 43* is about cognition and *Topic 56* is about medical care.

The act of explicitly naming topics is considered subjective and is therefore generally regarded as bad practice. We do so here only to aid the reader in understanding an abstract concept.

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 3.1: An illustration of four (out of 300) topics extracted from the TASA corpus: Reproduced from (Steyvers & Griffiths, 2007a).

3.2 Latent Dirichlet Allocation

LDA (Blei et al., 2003b) was an important advancement in the area of topic modelling; it reinvigorated research in the field and ultimately acted as a catalyst for the development of many other topic models (Teh et al., 2006; Blei et al., 2003a; Blei & Lafferty, 2007; Li & McCallum, 2006; Reisinger et al., 2010; Wang & Blei, 2009; Doyle & Elkan, 2009).

LDA was originally developed to fix an issue with a previous topic modelling algorithm; probabilistic latent semantic analysis (pLSI) introduced by Hofmann (1999). pLSI was in turn a probabilistic implementation of the seminal work on latent semantic analysis (LSI) (Deerwester et al., 1990). The relationship between these techniques is clearly described in (Steyvers & Grif-

fiths, 2007b).

3.2.1 Intuition

The idea behind LDA is that documents exhibit multiple topics. Figure 3.1 describes this idea with an example document. This document entitled "Seeking Life's Bare (Genetic) Necessities" is about computing the approximate number of genes an organism needs to survive the process of evolution. To the left of this Figure, 4 (out of 100) topics are depicted. The document itself has certain words highlighted in different colors. These colors correspond with the topic that each word belongs to with the highest probability. Finally, to the right of this diagram is a histogram of topic proportions for this document. This histograms informs us that the document is primarily formed by words from 3 of the 100 topics (yellow, pink and blue).

Latent Dirichlet allocation gets its name from the distribution that is used to draw the per-document topic distributions (the histogram in Figure 3.2). In the generative process for LDA, the result of the Dirichlet is used to allocate the words of the document to different topics. The keyword, latent, is present in the title to emphasise the fact that the actual topics are never observed i.e. they are not provided as input to the algorithm. They are inferred by the algorithm.

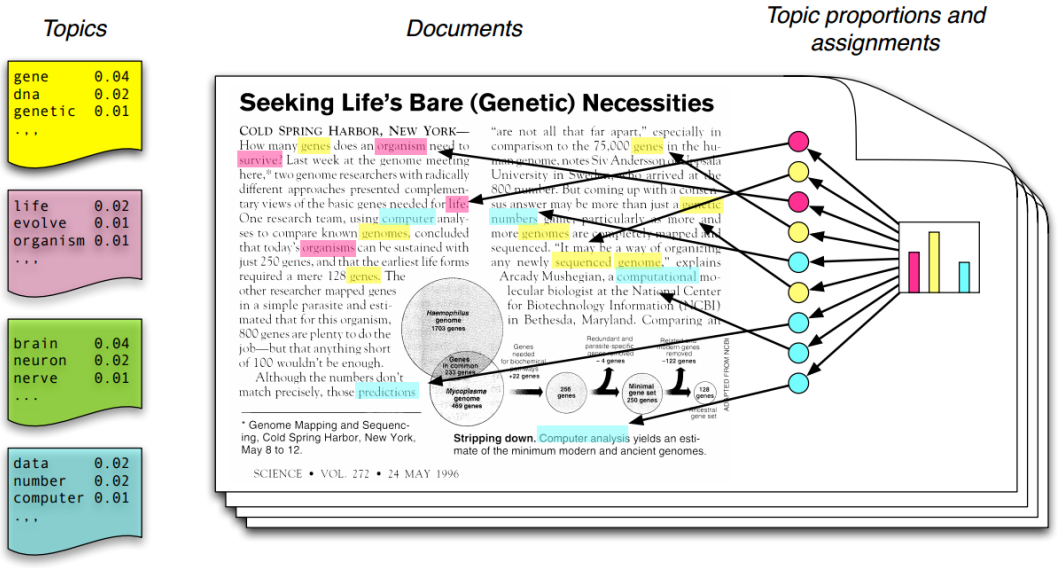


Figure 3.2: Intuitions behind latent Dirichlet allocation. Reproduced from (Blei 2012[1]).

3.2.2 Graphical Model

A graphical model is a probabilistic model for which a graph denotes the conditional dependence structure between random variables. LDA is very elegant when depicted as a graphical model (See Figure 3.3). In a directed graphical model, nodes represents random variables. If a node is shaded then it is observed, otherwise it is a latent variable. Edges between nodes denote possible dependence between random variables. The enclosing rectangles (plates) are a really compact way of denoting replicated structure.

The graphical model for LDA is best understood if one works from the outside in. Initially we will ignore α and η. Instead we begin our explanation with the rightmost plate, the K plate. The variable β_k here represents the topics, each β is a distribution over words and there are K of these distributions. β lives on the vocabulary simplex, the space of all possible solutions. β comes from a Dirichlet distribution. Next we describe the document plate, this plate is replicated once for each of the D documents. The only variable in

[1] `http://www.cs.princeton.edu/~blei/papers/icml-2012-tutorial.pdf`

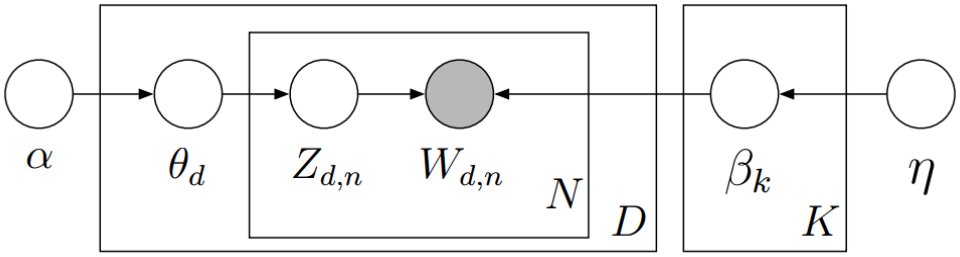


Figure 3.3: A graphical model representation of the latent Dirichlet allocation (LDA). Reproduced from (Blei & Lafferty, 2009).

this plate is θ_d which represents the topic proportions (the histogram in Figure 3.2). Each θ_d is of dimensionally K. The final plate which is replicated once for each word in each document contains two variables. The first is denoted $z_{d,n}$, the topic assignment (the coloured coin in Figure 3.2). We can see that $z_{d,n}$ depends on θ_d because it is drawn from a distribution with parameter θ_d. The second is denoted $w_{d,n}$, the nth word in the d^{th} document. We can see that $w_{d,n}$ depends on both θ_d and all of the β_k variables. $w_{d,n}$ is the only observed variable in the entire model. All LDA ever observes is a collection of words, grouped into documents.

3.2.3 Mathematical Model

The joint distribution of all variables, observed and hidden, according to the LDA model is given by Equation 3.1.

$$p(\beta, \theta, Z, W|\alpha, \eta) = \big(\prod_{i=1}^{K} p(\beta_i|\eta)\big)\big(\prod_{d=1}^{D} p(\Theta_d|\alpha)\big(\prod_{n=1}^{N} p(Z_{d,n}|\Theta_d)p(W_{d,n}|\beta_{1:K}, Z_{d,n})\big)\big) \tag{3.1}$$

where

K = the number of topics,

D = the number of documents,

N = the number of words in the d^{th} document,

β_i = topic i (a distribution over words),

θ_d = topic proportions for the d^{th} document,

η = topic hyper-parameter,

α = Dirichlet parameter,

$Z_{d,n}$ = the topic assignment Z for the n^{th} word of the d^{th} document,

$W_{d,n}$ = the n^{th} word of the d^{th} document.

The first section of this equation describes each topic which comes from some distribution appropriate over topics. The Dirichlet distribution. This equation is equivalent to the right most plate in the graphical model representation. We wrap this section in parentheses to emphasise that it is independent of anything else because the β values are only dependant on η.

$$\Big(\prod_{i=1}^{K} p(\beta_i|\eta)\Big) \tag{3.2}$$

The second section of this equation, introduces the second use of a Dirichlet distribution in the LDA model, to describe the topic proportions for each document. The topic proportions are only dependant on α.

$$\prod_{d=1}^{D} p(\Theta_d|\alpha) \tag{3.3}$$

The final section of the equation describes each word in each document. First we draw a topic assignment $Z_{d,n}$ from the specific topic proportions for this document θ_d, then we draw a word for this document $W_{d,n}$ conditioned on both the topics $\beta_{1:K}$ and the specific topic assignment Z for word n in this document d $Z_{d,n}$.

$$\Big(\prod_{n=1}^{N} p(Z_{d,n}|\Theta_d)p(W_{d,n}|\beta_{1:K}, z_{d,n})\Big) \tag{3.4}$$

3.2.4 The Dirichlet Distribution

The Dirichlet distribution is an exponential family distribution over the simplex. As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as prior, simplifying the problem of statistical inference. The probability density of a T dimensional Dirichlet distribution over the multinomial distribution $p = (p_1, ..., p_r)$ is defined by:

$$Dir(\alpha_1, ..., \alpha_t) = \frac{\Gamma(\Sigma_j \alpha_j)}{\Pi_j \Gamma(\alpha_j)} \prod_{j=1}^{T} {P_j}^{a_j - 1} \tag{3.5}$$

The parameters of the Dirichlet distribution are specified by $\alpha_1 ... \alpha_T$. Each hyper-parameter α_j can be interpreted as a prior observation count for the number of times topic j is sampled in a document, before having observed any actual words from that document. It is convenient to use a symmetric Dirichlet distribution with a single hyper-parameter α such that $\alpha_1 = \alpha_2 = ... = \alpha_t = \alpha$.

By placing a Dirichlet prior on the topic distribution Θ, the result is a smoothed topic distribution, with the amount of smoothing determined by the α parameter. Figure 3.4 illustrates the Dirichlet distribution for three topics in a two-dimensional simplex. The simplex is a convenient coordinate system to express all possible probability distributions. The smaller the value of the α parameter, the more spread out the distribution is.

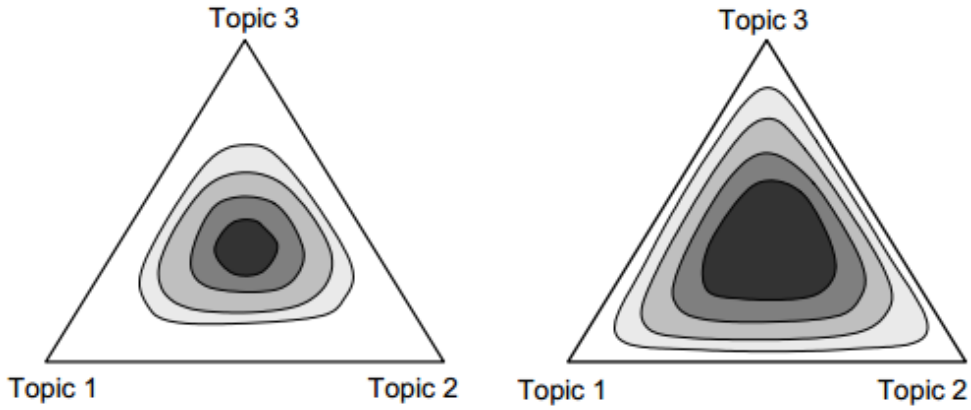


Figure 3.4: Illustrating the symmetric Dirichlet distribution for three topics on a two-dimensional simplex. Darker colors indicate higher probability. Left: $\alpha = 4$. Right: $\alpha = 2$. Reproduced from (Steyvers & Griffiths, 2007a).

The Dirichlet prior on the topic distributions (see equation 3.3) can be interpreted as forces on the topic combinations with higher α moving the topics away from the corners of the simplex, leading to more smoothing (compare the left and right panel). For $\alpha < 1$, the modes of the Dirichlet distribution are located at the corners of the simplex. In this regime (often used in practice), there is a bias towards sparsity, and the pressure is to pick topic distributions favouring just a few topics.

3.2.5 Model Fitting

In reality, fitting the LDA model to real data is computationally intractable. This is easiest to see by examining the per-document posterior distribution (see equation 3.6). For the rest of this section, we are assuming that the topics $\beta_{1:K}$ are fixed. The posterior distribution is the conditional distribution of the hidden variables given the observations. The hidden variables for one document are: the topic assignments Z and topic proportions θ. So the per-document posterior $p(\theta, Z|W_{1:N})$ which is the conditional distribution of one set of topic proportions θ and the topic assignments Z, given the observations which are the words in the document, is just the joint distribution of the hidden

variables divided by the marginal probability of the words.

$$p(\theta, Z|W_{1:N}) = \frac{p(\Theta|\alpha)\prod_{n=1}^{N} p(Z_n|\Theta)p(W_n|Z_n, \beta_{1:K})}{\int_{\Theta} p(\Theta|\alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} p(Z_n|\Theta)p(W_n|Z_n, \beta_{1:K})} \tag{3.6}$$

Equation 3.6 is intractable due to its denominator. There are two ways of seeing this. The first way is to recognise that the denominator is a multiple hypergeometric function (see Dickey (1983)). The second is to recognise that the denominator is also the sum of N^k (tractable) Dirichlet integral terms. These individual terms are themselves computationally tractable but there are (N^k) of them which leads us back to an computationally intractable denominator.

To fit the LDA model to a real dataset we must utilise approximate posterior inference of the posterior. There are many methods we can use for this task: Gibbs sampling, Variational methods, Partical filtering, Expectation propagation, etc. We describe Gibbs sampling in the next section as it is utilised by **plda** our preferred implementation of LDA developed by (Liu et al., 2011).

3.2.6 Gibbs Sampling

Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximately from a specified multivariate probability distribution (i.e. from the joint probability distribution of two or more random variables), when direct sampling is difficult. The basic steps of Gibbs sampling are:

- Define a Markov chain whose stationary distribution is the posterior of interest.
- Collect independent samples from that stationary distribution and approximate the posterior with them.

- In Gibbs sampling, the space of the MC is the space of possible configurations of the hidden variables.
- The chain is run by iteratively sampling from the conditional distribution of each hidden variables given observations and the current state of the other hidden variables
- Once a chain has 'burned in', collect samples at a lag to approximate the posterior

A very basic Gibbs sampler for LDA can be defined as follows. Let $n(z_{1:N})$ be a counts vector. The first step of the Gibbs sampler is to compute the conditional distribution of θ given $Z_{1:N}$ (the current state of the other hidden variables) and $W_{1:N}$ (the observations). From the graphical model (see Figure 3.3) we know that given Z, θ is independent of W. So θ is only dependant on Z and because its a Dirichlet and conjugate, the posterior distribution of θ given n draws from θ is just a Dirichlet distribution with parameter α plus the counts vector:

$$p(\theta|Z_{1:N}, W_{1:N}) \sim Dir(\alpha + n(Z_{1:N})) \tag{3.7}$$

The second step of the Gibbs sampler is to sample each Z_i again. Again from the graphical model (see Figure 3.3) we know that Z_i is only dependant on W and θ and so the posterior probability is proportional to the join distribution $p(Z|\theta)p(W_i|Z_i)$ (see equation 3.8)

$$p(Z_i|Z_{i-1}, W_{1:N}, \theta) \sim p(Z|\theta)p(W_i|Z_i) \tag{3.8}$$

So a basic Gibbs sampler for LDA iterates between equations 3.7 and 3.8 until convergence.

Chapter 4

Temporal Decomposition

4.1 Temporal Decomposition

The first goal of the presented analysis is to decompose the mobility flows of a bike sharing network into simpler temporal components that one can reason about. Given an assumption that individuals utilize the bike sharing network for different reasons: communing, leisure, exploration, health, etc., the objective is to identify these different motivations or 'topics'.

A methodology of this type for discovering the abstract topics that occur in a collection of documents is known as probabilistic topic modelling in machine learning and natural language processing. Various signal decomposition methods that can be applied to the problem are available, including the classical Principle Component Analysis, Independent Component Analysis, etc. PCA have been used in applications of urban dynamics (Reades et al., 2009; Toole et al., 2012). However, given the discrete nature of mobility flows, methods such as LDA have been shown to be more appropriate for the task and were applied to uncover hidden topics in generic urban activities (Ferrari & Mamei, 2013; Yuan et al., 2012; Kling & Pozdnoukhov, 2012). The other advantage of LDA is the probabilistic nature of a decomposition and an ability to deal with

overlapping topics. In the context of mobility flows, this is required as trips can have multiple purposes. Also, semantic attribution cannot be uniquely defined for all trips with certainty.

4.2 Document composition

To apply LDA as a decomposition technique, one requires an alternate representation of the bike sharing network data described in Section 2.2. Topic modelling operates on discrete dictionaries of atomic units conventionally called words. One's first objective is therefore to convert mobility flows into words, compose meaningful documents and ultimately process the obtained corpus with a topic modelling method.

With respect to the origin-destination flow dataset, a bike journey from station i to station j is therefore represented by the word i_to_j. Inserting _to_ between the stations names results in unique human readable words.

This collection of words needs to be grouped into documents to produce a corpus. To motivate a meaningful document composition for topic analysis, one asks the following question: what should a single document, read in isolation, tell about the bike sharing network? Such a document should contain a fuzzy account of the activity on the bike sharing network for a specific period of time. With respect to our bike sharing network, we take the stance that there is a set of hidden topics that motivate the transitions of bicycles between stations. For example, intuition suggests that perhaps one of these topics could be the morning commute. We will discuss further the caveats of explicitly naming these topics in Section 5.7.

By experimentation we have found that using any period less than 1 hour causes the documents to contain too few words to be truly descriptive. A document is thus simply an hours worth of (word, count) tuples where each

origin	dest	start	end
12	17	2010-09-15 14:05	2010-09-15 14:27
23	105	2010-09-15 14:07	2010-09-15 14:35
.	.	.	.
3	57	2010-09-25 23:37	2010-09-25 23:45

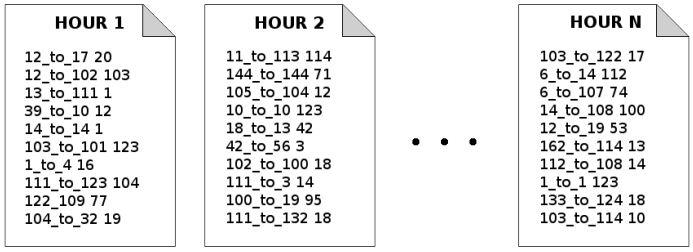


Figure 4.1: Tabular (above) and document (below) based representations of bike sharing rental data.

tuple represents the number of i to j trips observed in that hour. A summary view of word representation and document composition within a corpus is depicted in Figure 4.1.

4.3 Temporal regularities

By examining Figure 4.2, where a grey dotted line denotes total number of trips within each hour, one learns that traffic on the bike sharing network is very regular in nature. The working days, Monday to Friday, trace out an almost identical temporal profile. The double pronged heart beat on the working days is likely synonymous with urban commuting.

Figure 4.3 isolates and magnifies a typical weekday (grey dotted line). The morning commute now is clearly visible each day between (06:00 - 10:00) with a sharp peak at 08:00. The reverse evening commute is equally clear between (16:00 - 20:00). It also has a strong peak, this time at 17:30 but its decline is not as steep as the morning commute. One hypothesis for this is that workers feel strongly obliged to be at their desk by a specific hour in the morning. In the evening employers are not so keen to encourage the departure

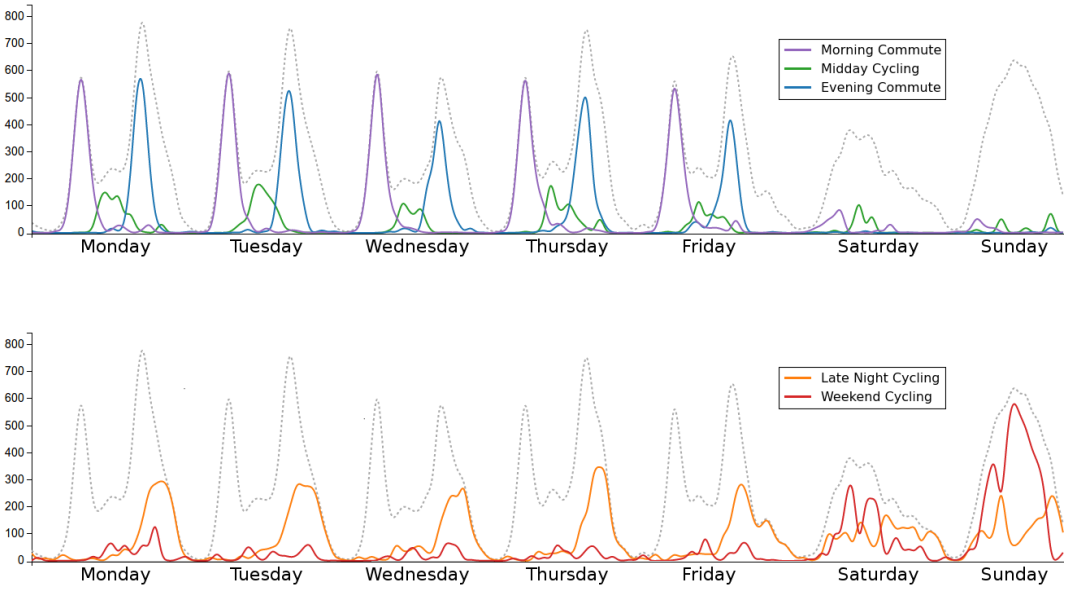


Figure 4.2: Temporal decomposition of the bike sharing rental data for the week (2012-07-16 to 2012-07-22 inclusive) into 5 named topics. See Figure 4.3 for a close-up view on a weekday.

of their workforce. It is equally plausible that cyclists try to avoid the evening rush hour, there is, after all, nothing worse for a cyclist than sitting in heavy traffic. Another thing that is evident from the daily temporal pattern of the bike sharing network is the slight peak in rentals around midday, this occurs each day between (12:00 and 1:30); lunchtime. The weekend profile of the bike sharing network is very different from the repetitive weekday profile. The rigid double pronged pattern is completely missing, instead it has been replaced with a jagged, bell shaped curve, further reinforcing the hypothesis that the weekday peaks are synonymous with urban commuting.

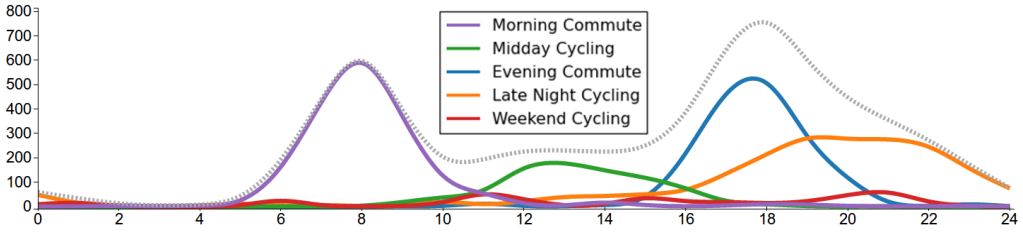


Figure 4.3: Temporal decomposition of bike sharing rental data for a typical weekday into 5 named topics.

4.4 Initial decomposition results

Figure 4.2 displays the results of running LDA on the bike sharing dataset with the following parameters: $K = 5$, $\alpha = 0.1$ and $\beta = 0.1$. We show the temporal decomposition here into 5 topics. We do this to motivate our ideas and to set the stage for the next section; Section 5.7 which contains a more in-depth decomposition and analysis.

The first thing one could notice when studying Figure 4.2 is that LDA has segmented the two peaks, described earlier as the morning and evening commutes, into separate and distinct topics. This is interesting for two reasons, firstly because it supports our intuition/hypothesis. Secondly, the LDA model applied here includes no temporal dependencies between documents. A baseline LDA does not model transition probabilities internally between words in a document nor does it model transition probabilities between adjacent documents. If one randomly re-orders the documents and/or the words inside each document one will get an equivalent decomposition. This decomposition will only differ slightly due to inherent randomness of inference based on Gibbs sampling which is used to approximate the otherwise intractable computations required for parameter estimation in LDA. How then is LDA distinguishing between what was superficially labelled the morning and evening commutes? LDA can do so because we have explicitly encoded time by forming documents

that represent discrete hours. Furthermore,d we know from Figure 4.2 that rental activity on the bike sharing network is very regular. LDA is uncovering and exploiting patterns of words it recognises as co-occurring on a regular basis.

Chapter 5

Semantic Labelling

5.1 Semantic Labelling

In Section 4.4 we demonstrated the ability to decompose a temporal profile into a number of distinct topic components. This is a very interesting and powerful idea but it raises an important question. How does one interpret these seemingly simpler subcomponents? If one cannot attribute meaning to them, then all one has done is swap a large enigma for many smaller enigmas. Indeed, all semantic attribution so far was superficial and based on intuition. For example, how does one make a distinction between evening commute and late night cycling topics (Figure 4.2) or claim with certainty that a midday topic is related to a lunch break and that those trips are not generated by schoolchildren coming back after classes?

5.2 The spatial component

So far only the temporal aspect of the decomposition has been examined. Extra information can be gained by examining the actual topic-specific trajectories between bike stations. Figure 5.1 displays the spatial representation of two topics. While each bike rental i_to_j is assigned a probability of belonging to

each of the K topics, we simply visualise each rental trajectory by its dominant topic i.e. the topic it has the highest probability of belonging to. Each of the maps displayed in Figure 5.1 is a snapshot of a different topic at its highest intensity on Monday. The first map, 'morning commute' was taken at 08:00. The second map, 'evening commute' topic, was taken at 18:00.

A detailed inspection of the maps supports the hypothesis that the considered topics are primarily related to commuting. This is evident by the overall flow directions, which appears to be in general towards the city centre on the morning map. Likewise a flow snapshot in the second map of Figure 5.1 appears to have the reverse flow direction, most of the flows in this topic are from the city centre towards the suburbs. However, it is less coherent compared to the morning as evening time is composed of several different processes. (Figure 4.2).

Arguably space is as important a descriptor as time. Consider, for example, increasing the number of topics to a number high enough to isolate individual events such as an important concert. Then one should be able to identify complementary pairs of topics, the first being described by many flows pointing towards the spatial center of an event and the second proceeding it with opposite flow directions.

Even with both spatial and temporal descriptors it is still however very difficult to associate convincing meaning to topics. The deductions on semantic meaning of topics above were still mainly based on background knowledge and common sense. One needs to investigate a more descriptive source of information.

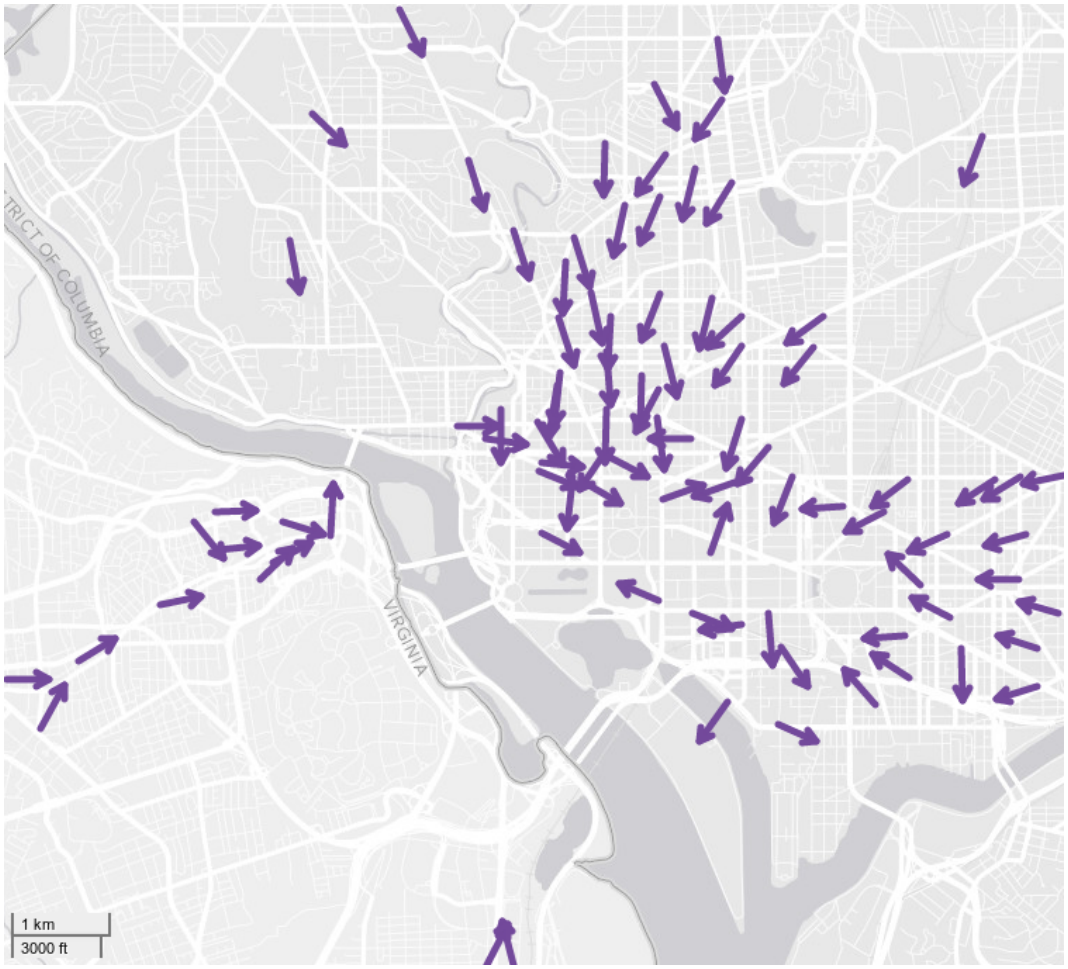


Figure 5.1: The morning commuting topic plotted in space.

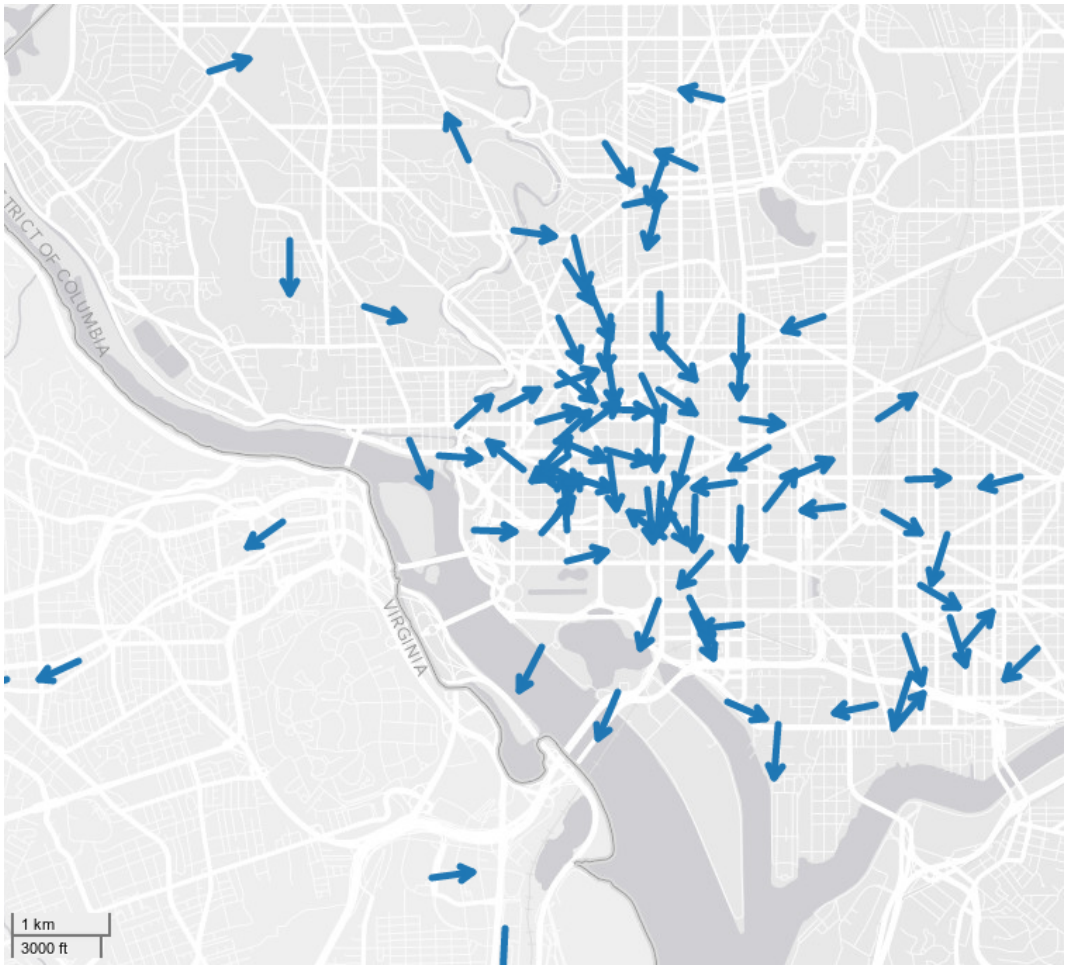


Figure 5.2: The evening commuting topic plotted in space.

5.3 The textual component

One such source that is readily available and abundant with textual information is social media streams such as Twitter. The enhanced geo-tagged Foursquare check-in messages posted via Twitter is a particularly promising source of data. The Foursquare check-in service transforms tweets from seemingly random, non-spatially constraint, temporal opinions; normally expressed with poor spelling and grammar into targeted bullets of spatially and temporally contextual, targeted record of location, activity and often a personal opinion.

Foursquares most recent feature actually makes the check-in process almost automatic 4Square (2013). The Foursquare application, once installed on a mobile device, continuously tracks location and provides a user with a best guess spatial description at all times. It is still however up to the user to broadcast this location information with an optional short textual comment, i.e. to 'check in'. A check-in is therefore a self-reported timestamped user location carrying the semantics of the intended user actions.

When many people create these high fidelity breadcrumb traces (Cheng et al., 2011) (and they do, see Figure 2.8), they offer a space/time window into the life and events of a specific geographic area, in our case an entire city. Figure 2.8 was generated using only 7 days of geo-tagged, Foursquare integrated tweets. In these 7 days, 9244 unique users managed to generate 37950 tweets densely covering the city center and a lot of the suburbs of Washington D.C, US This dense covering tells the story of a city, encompassing many places, events, people and opinions Kling & Pozdnoukhov (2012); Pozdnoukhov & Kaiser (2011).

These types of traces do more than just tell us about large scale social events; such as parades, protests, concerts, etc. They also contain the seemingly mundane details of everyday life; people check in at bus-stops, metro-

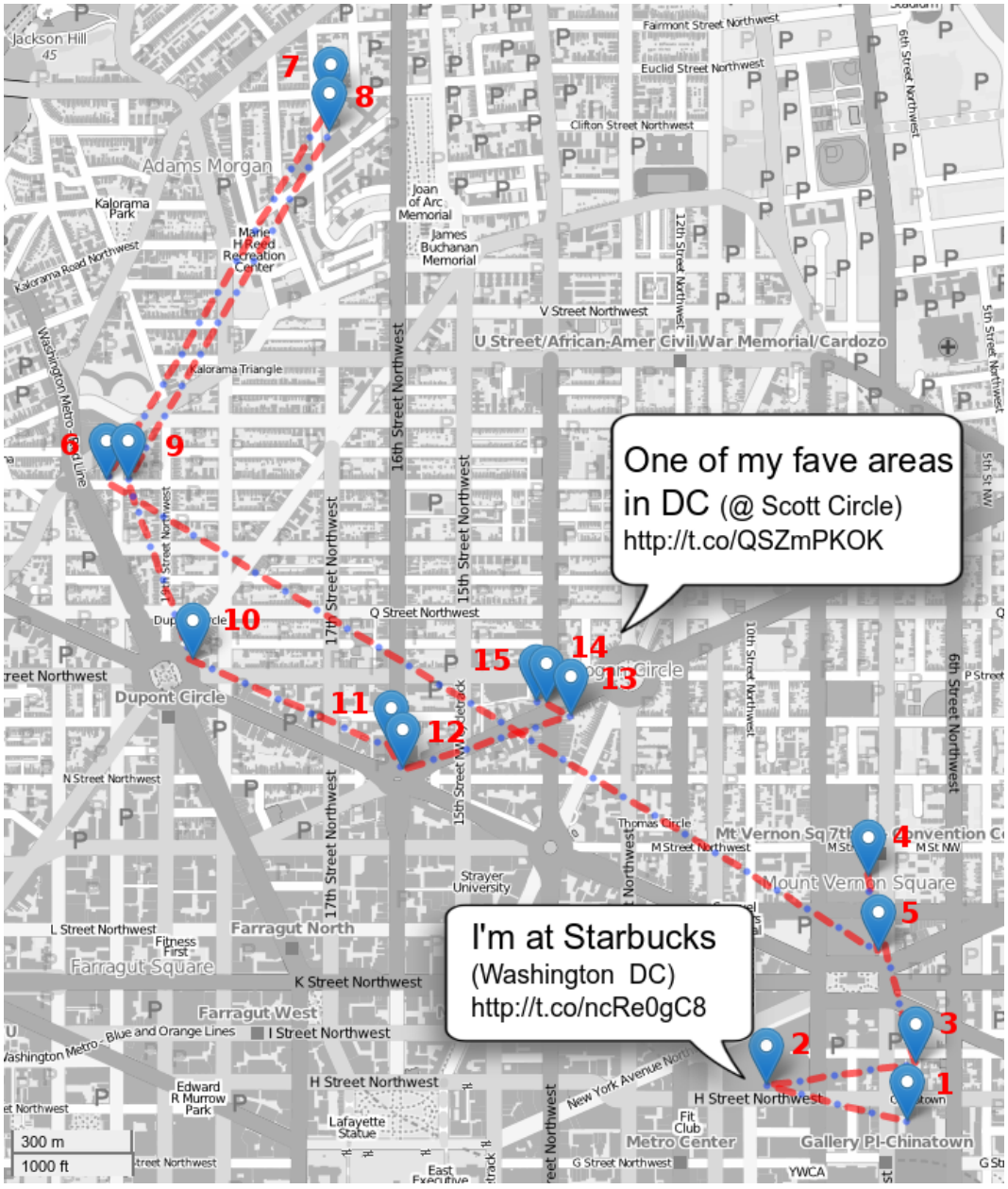


Figure 5.3: Digital breadcrumbs left behind by user id:96620504 on 2012-07-22. These breadcrumbs trace the entire day of the user, where he/she went, what he/she did and his/her opinions on everything that occurred during the day.

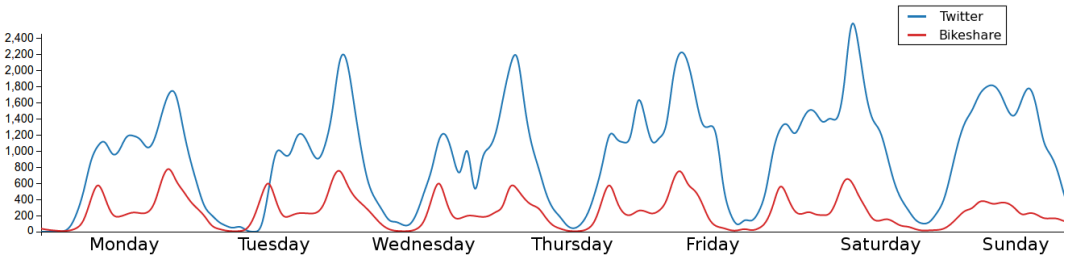


Figure 5.4: Overlapping Bike Sharing and Twitter time series. The peaks in each signal approximately align but are of different magnitudes.

stations, bike-kiosks, they check in when they buy coffee, visit the bank or get a haircut, they even check in at home and at work. These digital breadcrumbs, (trajectories) represent, in a similar vein to the bike trajectories, details about the mobility of everyday life. It is therefore likely that both forms of mobility are motivated by the same underlying needs. Though not every cyclist is a Twitter user and not ever cyclist tweets while cycling (although surely some do), these users' mobility needs and motivations are overlapping as they both live in the same city, the same time, the same network of places, activities and events.

A research goal we aim to achieve is to utilize tweets with Foursquare check-ins integration as an additional channel of semantically rich mobility information.

5.4 Descriptive analysis of the textual data

Before we coalesce these two mobility datasets, we first examine the Twitter check-in data in isolation. Figure 5.4 presents the temporal profile of bike sharing and check-in hourly counts for the exact same time period (a typical week) of 2012-07-16 to 2012-07-23. Both similarities and differences can be noticed by comparing the time lines. The most obvious similarity is the segmenting effect night has on both temporal profiles. The seven days of the week are

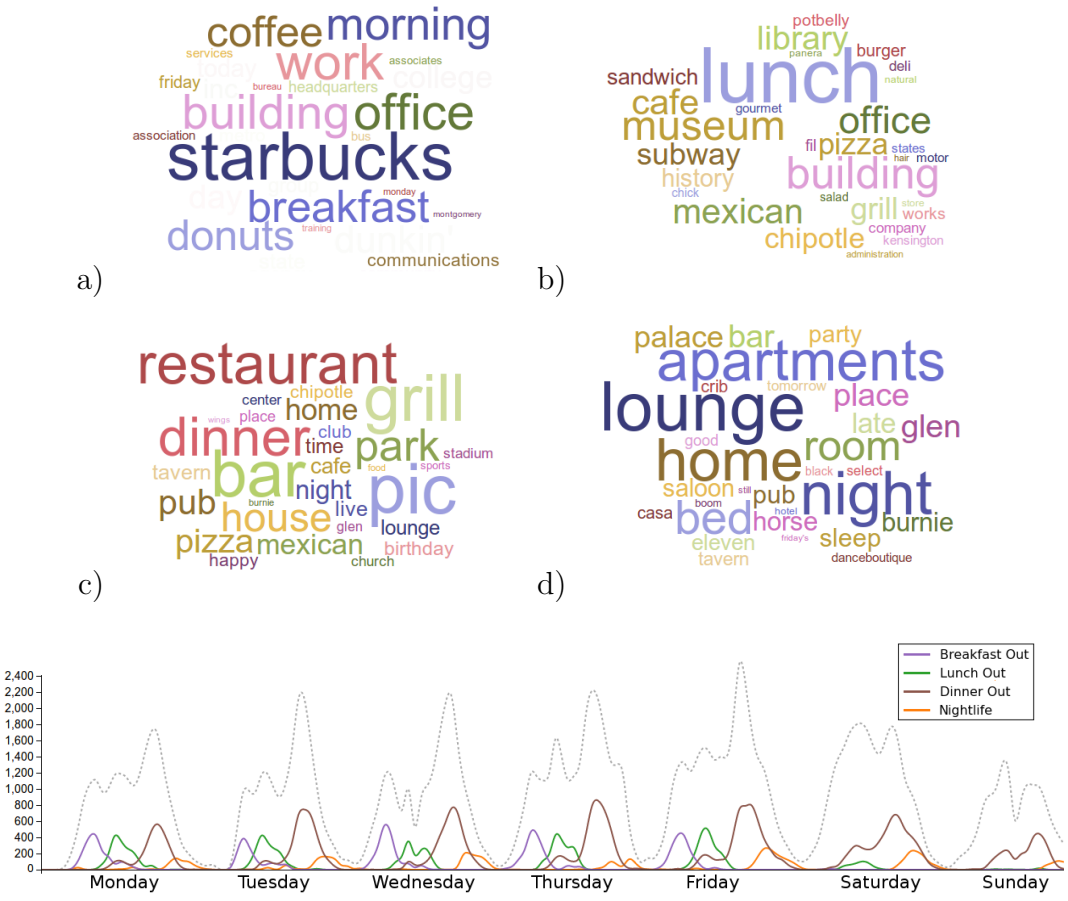


Figure 5.5: LDA decomposition: 4 topics related to breakfast (a), lunch (b), dinner (c), and nightlife (d). The dashed line represents the original count time series.

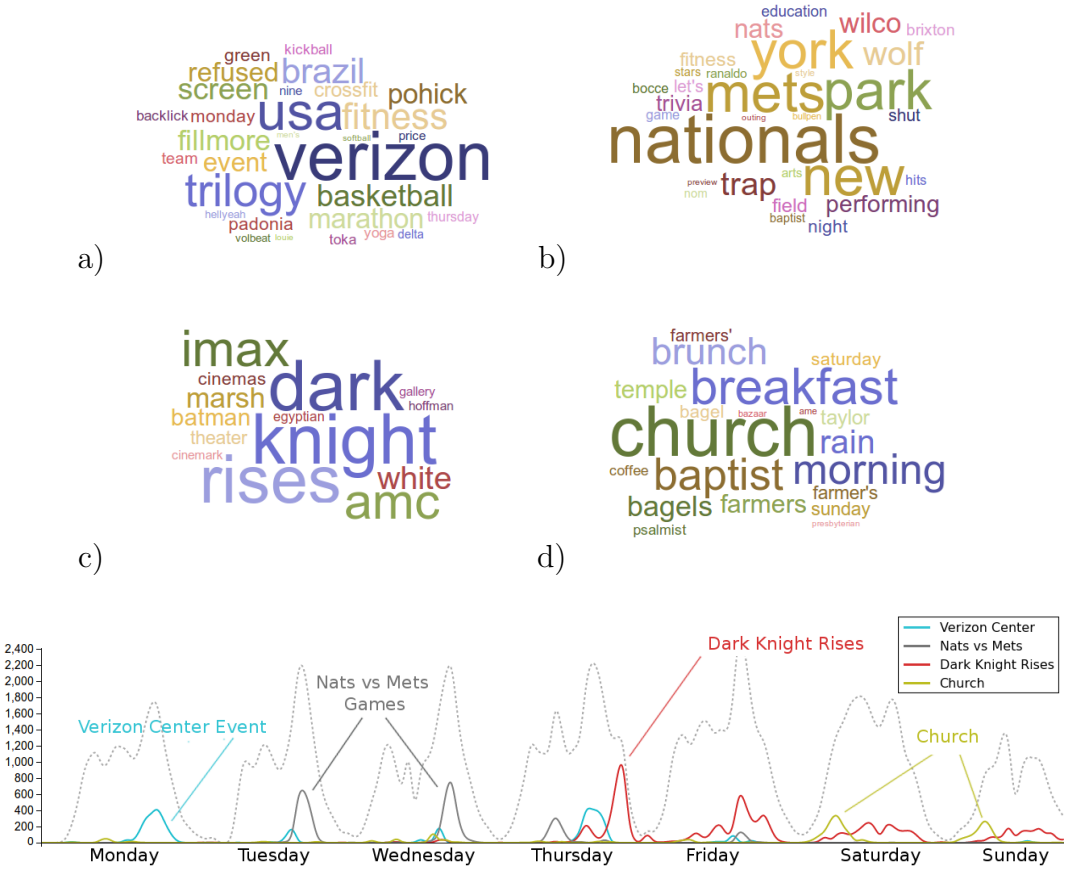


Figure 5.6: LDA decomposition: 4 topics related to events at Verizon Centre (a), baseball championship (b), a premier of a big box office movie: Dark Knight Rises (c), and church activities on a weekend (d). The dashed line represents the original count time series.

clearly separated in both plots as most individuals have regular sleep patterns. It also allows one to compare the two mobility profiles on a day by day basis. The second most striking similarity is the division of both signals into weekday and weekend patterns. Each weekday appears to trace out a common pattern which is noticeably different from that traced by a Saturday or a Sunday.

We have discussed similarities so to be thorough we must now discuss differences. The most significant difference between the two signals is the position and number of peaks. The Twitter time series clearly does not display the same double pronged activity profile as the bikeshare time series. Instead this pattern has been replaced by a three pronged variant. The first and second peaks of the Twitter time series approximately align with the so called commuting peaks in the bikeshare time series. The evening peak is however consistently greater than the morning peak. There is also a more prominent mid day peak in the Twitter time series. One final and obvious difference between the time series, is the relative importance of Saturday and Sunday. In the bikeshare time series Sunday is more active than Saturday, the reverse is however true in the Twitter dataset.

5.5 LDA analysis of check-ins

Now that we have identified and studied a new source of textually abundant data in isolation it is time to revisit LDA. The Twitter dataset was converted into a set of documents by grouping the check-ins into individual documents by hour. Some Simple filtering steps were performed on the content of the tweets. First, Twitter specific tokens such as mentions (@username) and hash-tags (#hashtag) were removed. Then the Foursquare specific phrases such as 'I just became the mayor of' and the URLs were removed. Finally, the remaining words were stemmed. As an example, the following Tweet 'Shipping out some

jewelry! #fb (@ The UPS Store) http://t.co/prudc1ve‘ simply becomes the list ['shipping', 'jewelry', 'ups', 'store']. This filtering process results in 168 (24 * 7) Twitter based documents that we decompose and analyse in the remainder of this section. We attempt to name each topic by examining its most likely words and temporal profile in isolation.

5.6 Periodic Topics

The optimal choice for the number of topics K when preforming LDA is an open question. By experimentation we found that a value of $K = 5$ consistently decomposed the data into 5 clear, periodic components. We named these 5 distinct and stable components: 'Routine', 'Breakfast Out', 'Lunch Out', 'Dinner Out' and 'Nightlife'. Routine can be thought of as general chatter or background noise. It contains a non-coherent jumble of key words. We have therefore deemed it non interesting with respect to our analysis and shall not consider it further. The other 4 topics were named using the same methodology as in (Kling & Pozdnoukhov, 2012; Pozdnoukhov & Kaiser, 2011). By examining the keywords and temporal profiles for these 4 topics (see Figure 5.5) our choices of appellation seems appropriate.

5.7 Event Topics

It was found empirically that increasing K beyond the value 5 enabled LDA to discover interesting, non-periodic sub components. The value K was incremented further until a newly created topic could no longer be explained in isolation. This process enabled the discovery of 4 specific event topics that took place during the period encompassed by our twitter dataset. The temporal profile for these events and their keywords can be seen in Figure 5.6.

The first of the newly discovered topics is the 'Verizon Center‘ topic. The

Verizon Center is a 20,000 seat multi-purpose sports and entertainment venue, owned and operated by Monumental Sports & Entertainment, in the Penn Quarter neighbourhood of Chinatown in down town Washington D.C, US This topic contains two significant peeks which coincide with two distinct events at the Verizon Center. The first event was a basketball game on the evening of Monday July 16^{th} (USA vs Brazil). The second was a smaller family event on the evening of Thursday July 19^{th} (How to Train Your Dragon live spectacular). The top four keywords in this topic are: verizon, usa, brazil and basketball.

The second of these topics, is the 'Nats vs Mets' topic. This topic peeks 3 times coinciding with the start times for a 3 day series baseball game that took place on (Tuesday July 17^{th}, Wednesday July 18^{th} and Thursday July 19^{th}) between the Washington Nationals and the New York Mets. The top 4 words in this topic are: Nationals, new, york and mets.

It would be very hard to refute that the third topic is about the premier of a long anticipated film in the Batman series "The Dark Knight Rises". This topic peeks dramatically at midnight on Thursday July 19^{th} coinciding with the premiere of the film "The Dark Knight Rises" which was held simultaneously by many cinemas in Washington D.C, US Among the top words in this topic apart from the obvious: Dark, Night, Rises are the names of 3 popular cinema chains: imax, acm and cinemark.

The forth and final topic has been named 'Church'. After the inspection of many tweets it became clear that this topic is being created by church goers checking into mass on Saturday and Sunday mornings. The key words in this topic are: church, baptist, temple, morning, Saturday and Sunday.

Increasing the value of K any higher than 9 caused the Routine topic to split into sub components. These sub topics are as difficult to annotate as the original routine topic itself. We therefore conclude that our Twitter dataset contains 8 meaningful topics and further decomposition is relatively fruitless.

Chapter 6

Correlation & Causation

6.1 Correlation & Causation

Throughout this thesis we have advocated the usefulness of decomposition methodology for semantic enrichment of mobility data from a related geo-referenced social media dataset. We warned against over interpretation of topics based on temporal coincidence of seemingly related processes. For example, while the morning commute is quite regular and self-evident, the validity of LDA decomposition of evening travel behaviours (Figure 4.2) is not as straightforward and can not be inferred solely by intuition. The reverse order is quite likely for the topics that we provisionally labelled as 'Evening commute' and 'Late Night Cycling' in Section 4.4. We further show how this issue can be resolved by exploring correlation and causal relationships between the topics found in both datasets.

6.2 Causation relationship

Discriminating true causation from correlation is a long-standing problem in statistics and data analysis. The common tools for causality verification include the Granger test Granger (1969), as well as the recently developed meth-

ods by Sugihara et al. (2012). We have opted to use the Granger's test to find relationships between urban activities (as detected from Twitter check-in topics) and intensities of bikeshare mobility modes detected with the LDA decomposition. The variation of the test that we have utilised (Ding et al., 2006) builds a multi-variate auto-regressive model which tests whether the addition of a particular topic identified in another dataset improves the prediction of the topic under consideration for the next hour.

Figure 6.1 shows correlation coefficients between pairs of topics from two datasets. For example, one can note high correlation between the 'Morning commute' and 'Breakfast Out' topics amongst others. Also, note a high correlation between 'Dinner Out' and a baseball game topic 'Nats vs Mets' which appears due to temporal co-occurrence of the two activities.

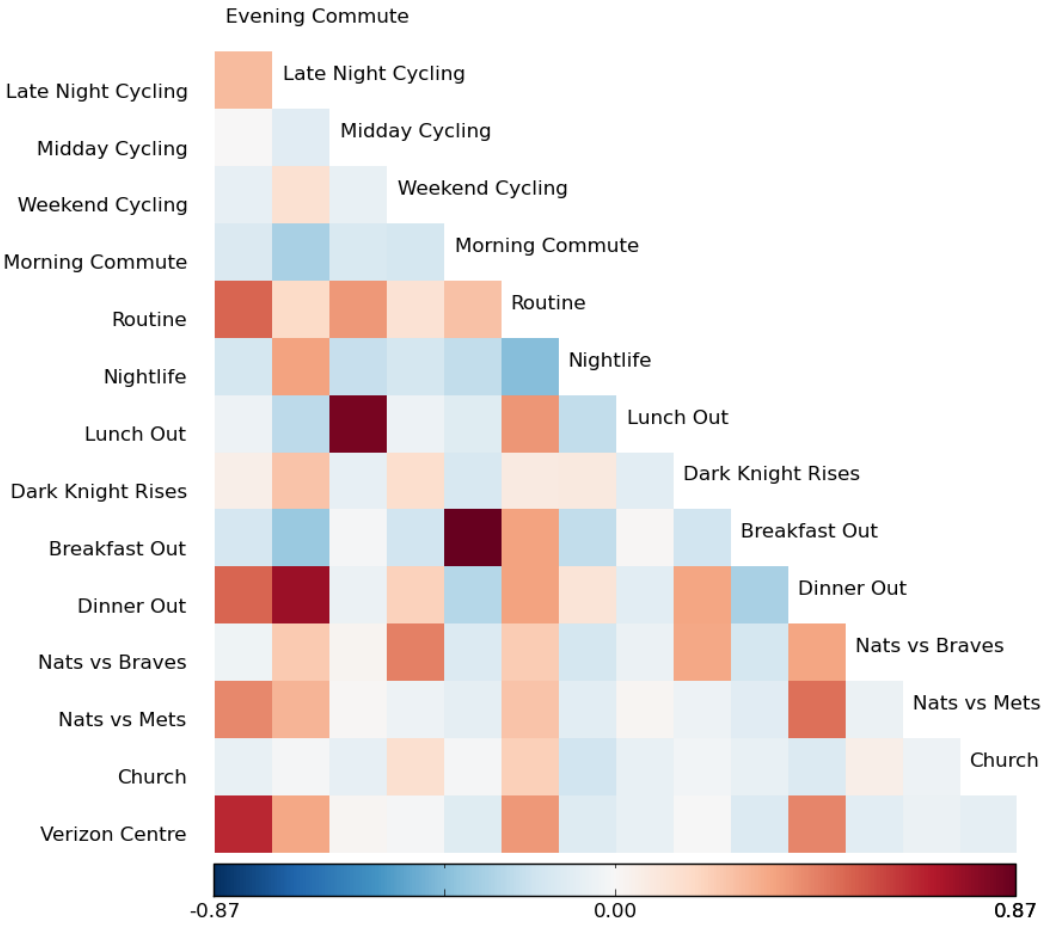


Figure 6.1: Correlation coefficient.

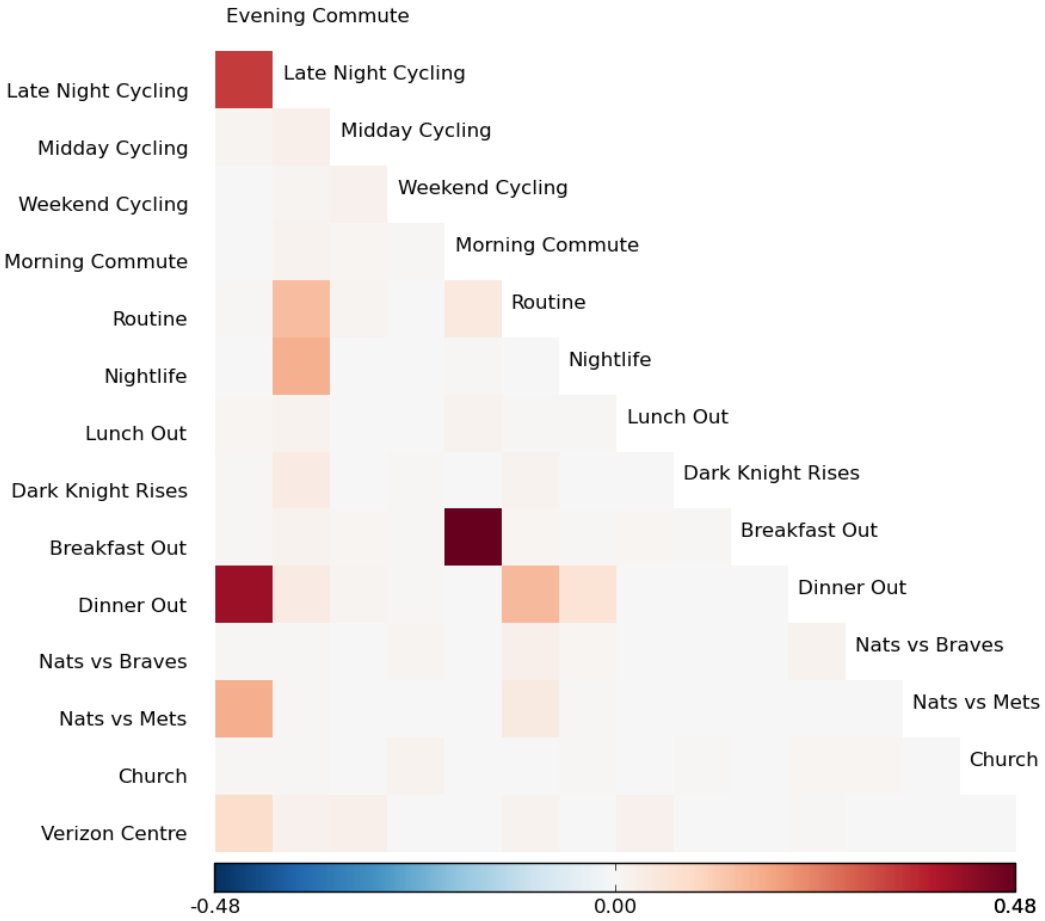


Figure 6.2: Causality index from Granger's test.

The Granger's causation index presented in Figure 6.2 reveals a very different pattern. While highlighting the true causal dependencies ('Morning commute' and 'Breakfast Out'), there is no significant causal relationship between many of the temporally co-occurring events such as the mentioned 'Dinner Out' and a baseball game.

6.3 Dependence graph

Figure 6.3 presents the causal relationship identified between topics in a form of a directed graph. To guide the eye, the topics corresponding to bikeshare mobility and Twitter check-ins are marked with different colors and icons. The size of a node in the graph is proportional to the mean intensity of the topic

and the link widths are proportional to the Granger index value (Figure 6.2).

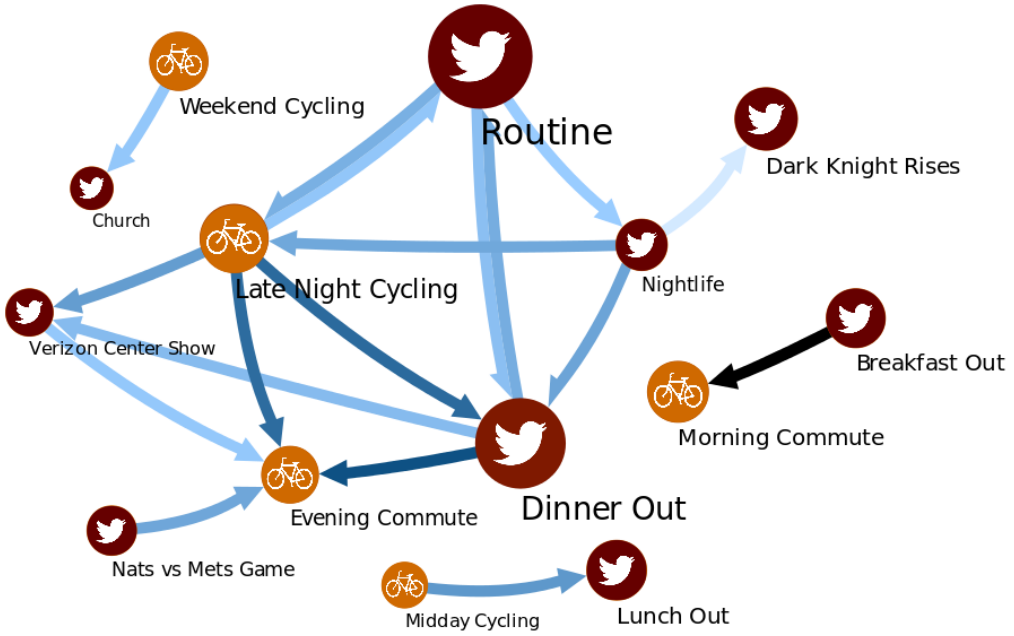


Figure 6.3: Causal relationships detected between estimated bikeshare mobility modes and social media topics.

We have previously mentioned that the morning commute is relatively apparent in all aspects (time, space and the content of the relevant 'Breakfast Out' topic), while the evening commute is far more convoluted. Indeed, the intensity of the evening commute is dependent on several Twitter topics corresponding both to the regular activities ('Dinner Out') and one-off events ("Nats vs Mets' game and a 'Verizon Centre' shows). It is also affected by the 'Late Night Cycling' topic.

An interesting relationship is the one between 'Breakfast Out' and 'Morning Commute'. The direction of causation suggests that people eat breakfast and then cycle to work rather than the reverse relationship which would suggest that people cycle in order to purchase or eat breakfast. In contrast, 'Midday Cycling' precedes the 'Lunch Out' suggesting that bikeshare users cycle to reach lunch venues.

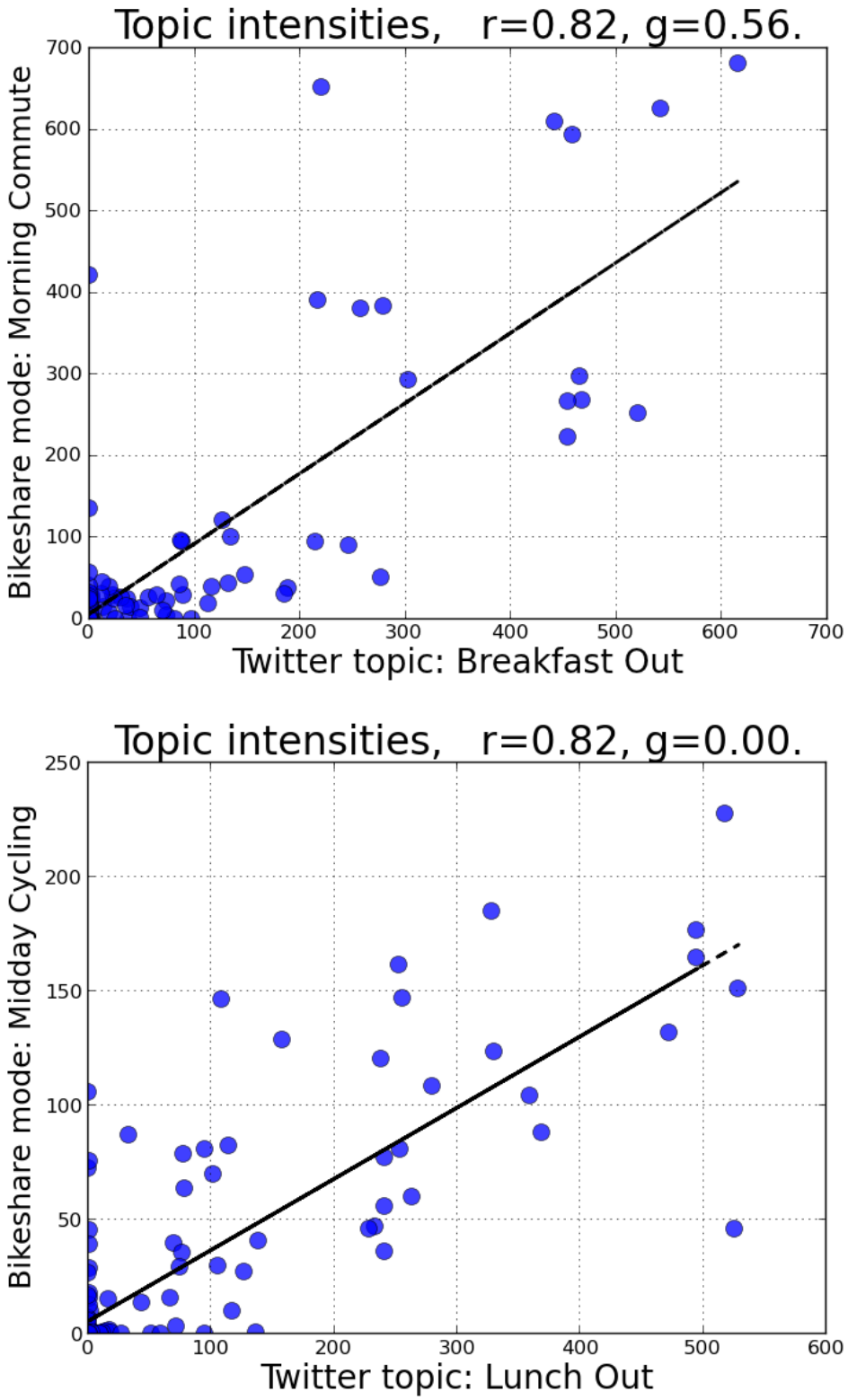


Figure 6.4: Scatter plots for the 'Breakfast Out' and 'Morning Commute' (top) and 'Midday Cycling' and 'Lunch Out' (bottom).

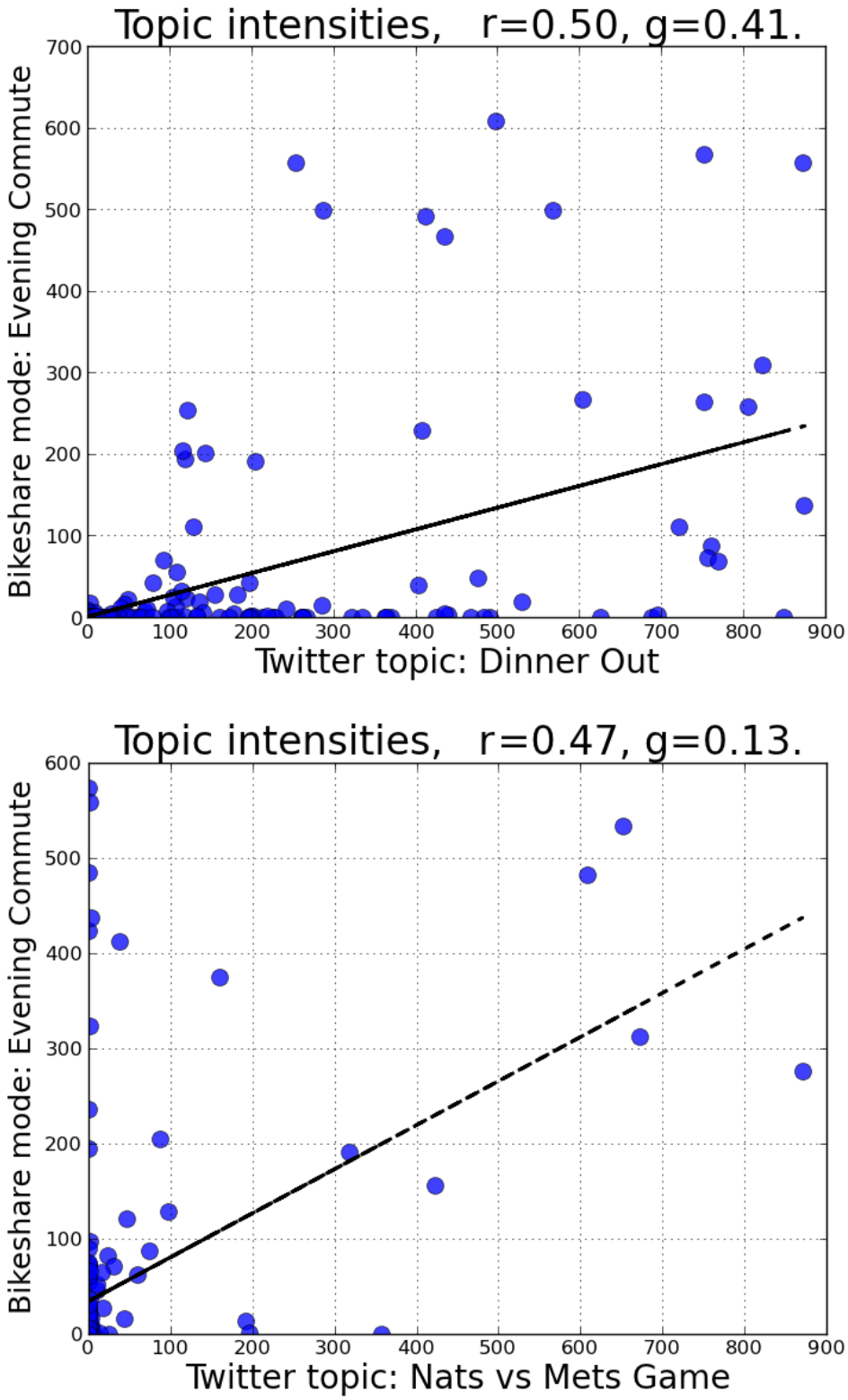


Figure 6.5: Scatter plots for the ‘Evening Commute‘ and ‘Dinner Out‘ (top) and ‘Evening Commute‘ and ‘Nats vs Mets‘ game (bottom).

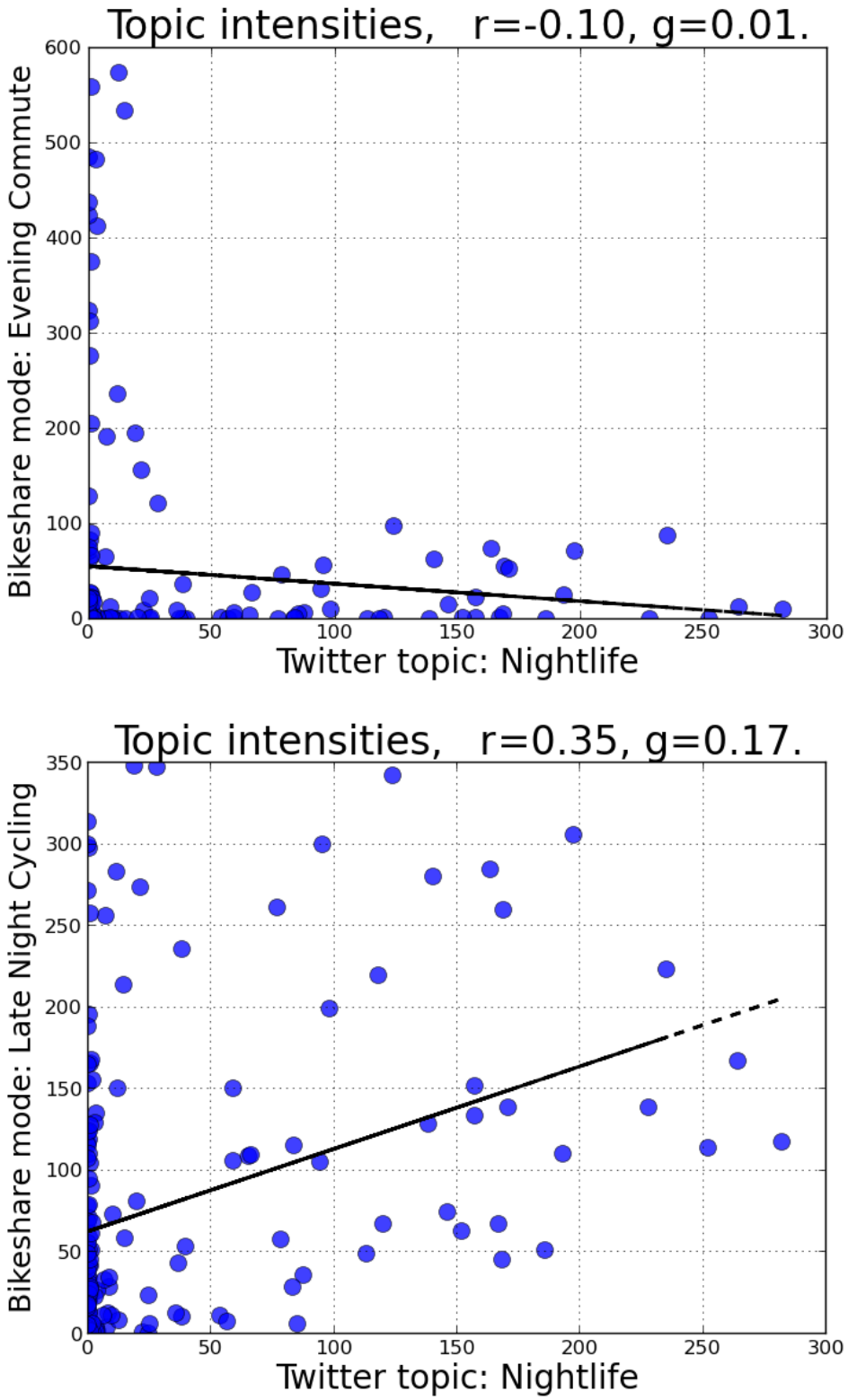


Figure 6.6: Scatter plots for the "Evening Commute" and "Nightlife" (top) and "Late Night Cycling" and "Nightlife" (bottom).

Chapter 7

Towards Demand Forecasting

7.1 Predictability

The dependence graph described in the previous chapter (see Figure 6.3) encodes conditional independence between variables and can be used to build predictive models. However, despite the seemingly simple relationships depicted in the graph, the actual dependencies between bike rentals and Twitter messages are quite complicated. Figures 6.4-6.6 provide scatter plots for several pairs of topics illustrating cases where correlation does not correspond to causation (Figure 6.4, see the Pearson's correlation coefficient r and the Granger index g values in the Figure title). Both of these values range from 0 to 1. A value of 0 indicates absolutely no correlation or causation whereas a value of 1 indicates perfect correlation or causation.

The rest of the chapter details a proof of concept framework, for forecasting mobility flows from social media streams. There are three steps in the forecasting process. At this point in the thesis we have already described step 1 in detail. The remaining two steps will now be described.

Step 1

Decompose the mobility and social media temporal profiles into (n and m)

topic components respectively.

Step 2

Train a system to predict the n mobility topic intensities from the m social media topic intensities.

Step 3

Use LDA to infer new documents, individual mobility flows (words), given topic intensities predicted by Step 2.

LDA is a generative model, once it has been trained on a corpus of documents, the model itself can be used to generate new documents. In our application a document is a collection of unordered bike rentals. Predicting individual bike trajectories with LDA is impractical, especially for stations with low rental counts. However the technique can be used to approximate aggregate flow between OD pairs. Aggregate flows are useful in many transportation applications, such as: optimizing multimodal interconnectivity (Coffey et al., 2012), and, large scale transit schedule coordination (Nair et al., 2013). Moreover, because this technique utilises social media streams, it can react to large scale events, providing real-time flow forecasts for a dynamic city.

7.2 Predicting mobility topic intensities

There are many models from the field of machine learning: Neural Networks, Support Vector Machines, Linear Regression, that could be used to learn a mapping from mobility topic intensities to social media topic intensities. We have opted to train a multilayer perceptron (MLP) (Rosenblatt, 1961). A MLP is a feedforward artificial neural network model that maps sets of input data onto sets of output data. An MLP consists of multiple layers of nodes

in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network (Rumelhart et al., 1985).

Our MLP instance (see Figure 7.1) has ten input neurons (the Twitter topic intensities) and 5 output neurons (the bikeshare topic intensities). We chose the number of hidden layers (2) and the number of nodes per hidden layer (100) experimentally. The validation protocol for the predictor was as follows. The 'one-day-out' procedure was applied, with one day of data taken out of the training set and predicted based on the data from the remaining days.

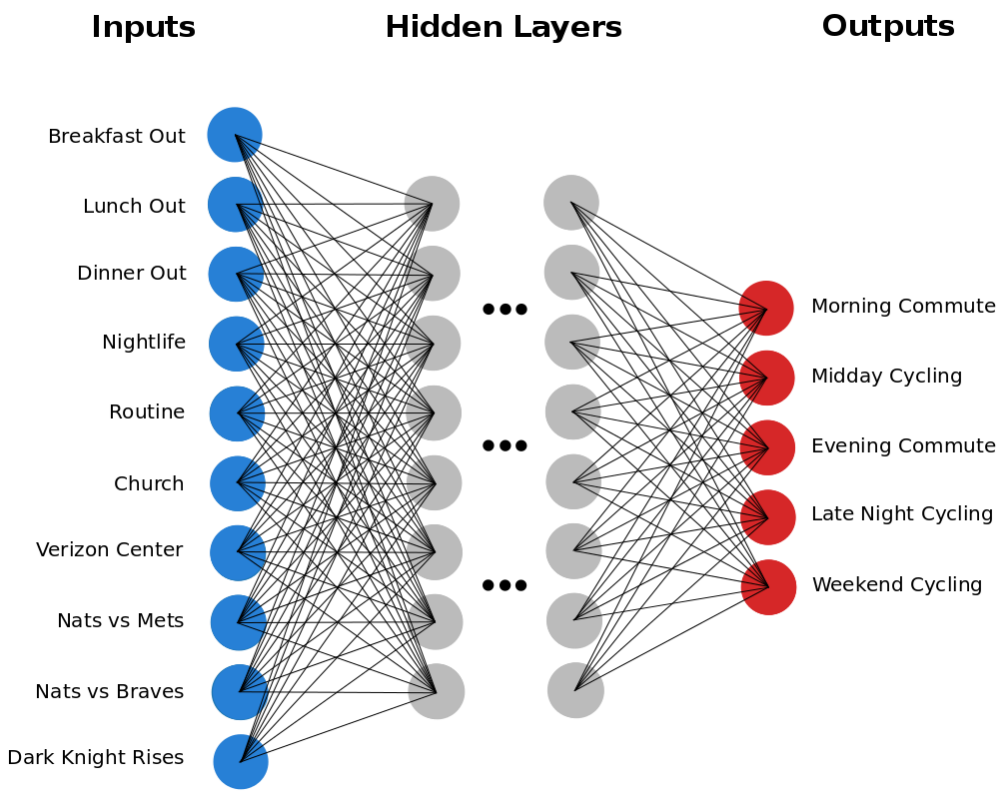


Figure 7.1: MLP with social media (Twitter topics) as input nodes and mobility (bikeshare topics) as output nodes.

Once the MLP has been trained we have a function mapping Twitter topic

intensities to bikeshare topic intensities (see equation 7.1). This function will be used in step 3 to predict actual bike flows on the Capital Bikeshare network.

$$\lambda(t_1, t_2, ..., t_{10}) = [b_1, b_2, ..., b_5] \tag{7.1}$$

7.3 Predicting mobility flows

Given an LDA model M, trained on a corpus of historical documents, we can generate a new document D using a vector of topic intensities T. Figure 7.2 displays the predicted topic intensities $T = [0.73, 0.21, 0.03, 0.02, 0.01]$ for Tuesday July 24th at 18:00.

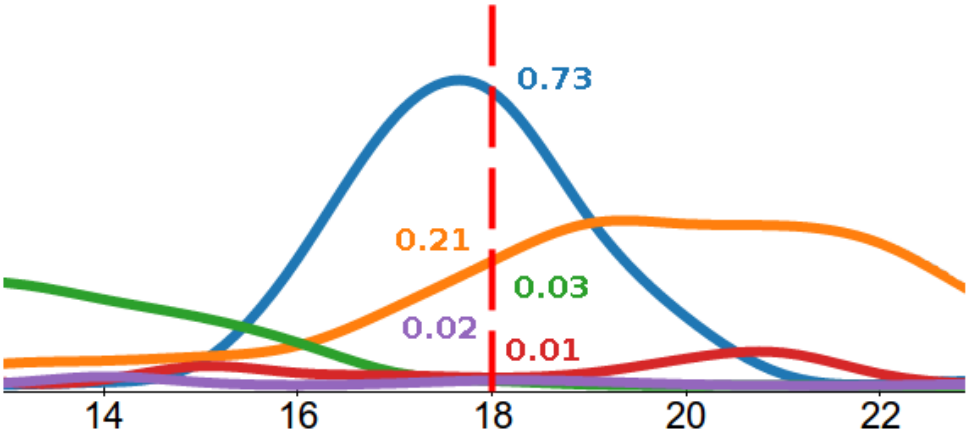


Figure 7.2: Predicted topic intensities (T vector) for Tuesday July 24th at 18:00.

The generated document contains 672 (the sum of the topic intensity vector T before normalisatio) bike rentals. LDA generated these bike rentals by first choosing a topic with probability proportional to the topic intensity vector T. Then given that topic, it drew station pairs (bike trajectories from origin to destination) with probabilities learnt during (step 1) the modelling stage. Figures 7.3-7.7 display per topic bike rentals in space.

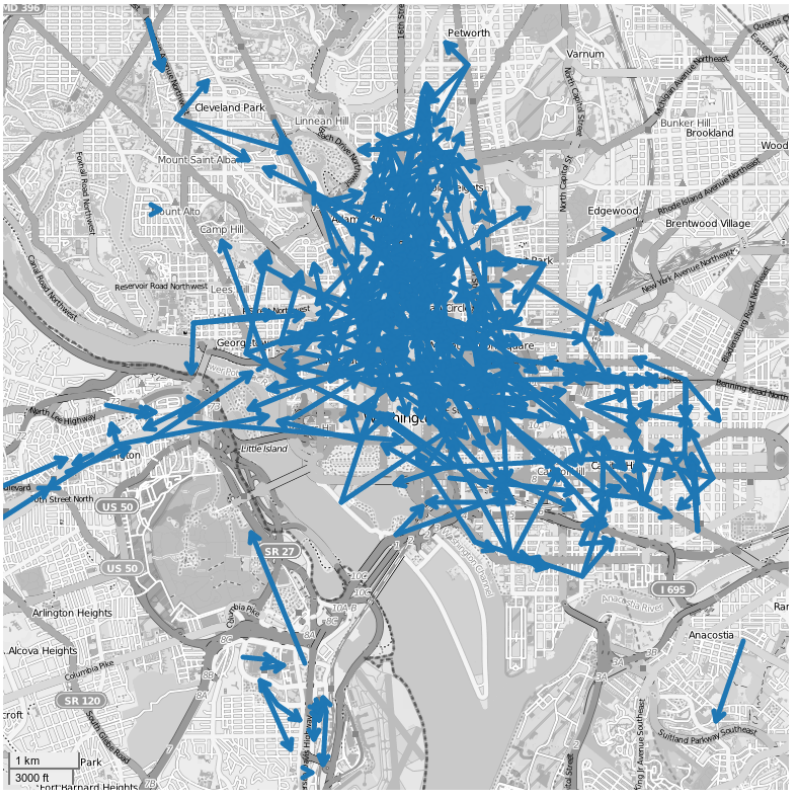


Figure 7.3: Topic 1 accounts for 489 of the 672 bike rentals for this hour

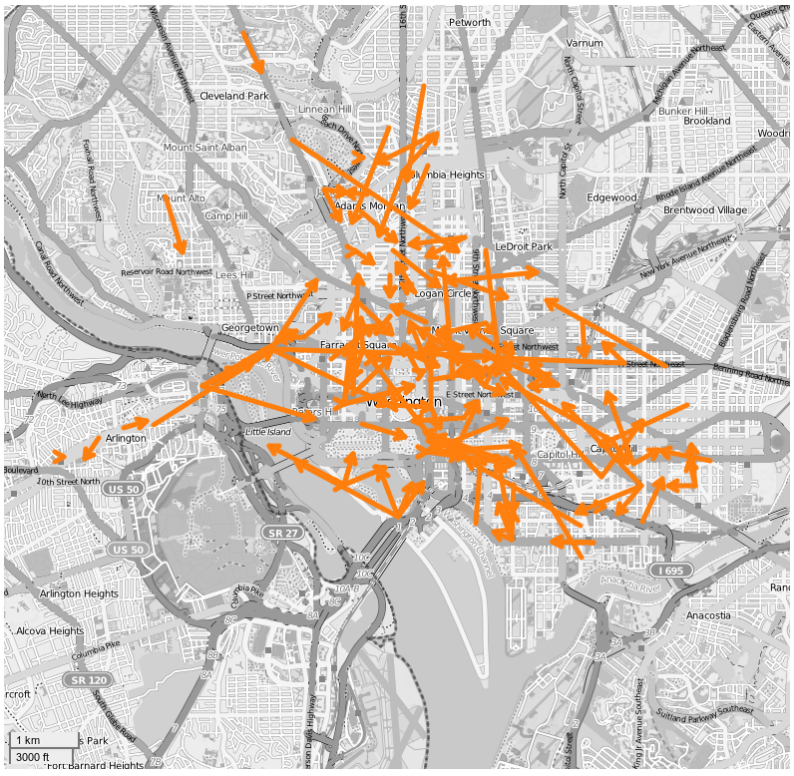


Figure 7.4: Topic 2 accounts for 142 of the 672 bike rentals for this hour

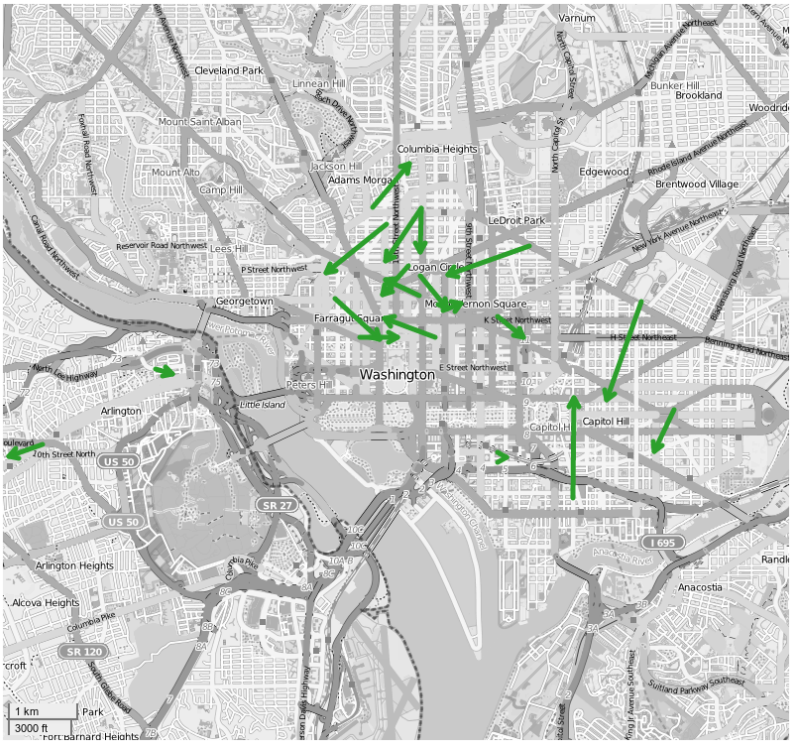


Figure 7.5: Topic 3 accounts for 21 of the 672 bike rentals for this hour

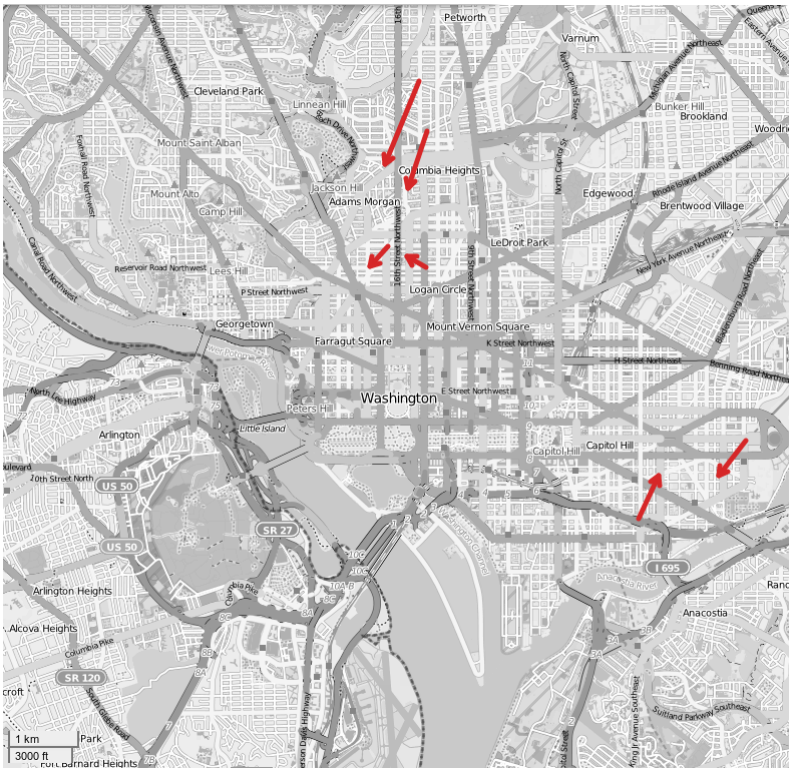


Figure 7.6: Topic 4 accounts for 6 of the 672 bike rentals for this hour

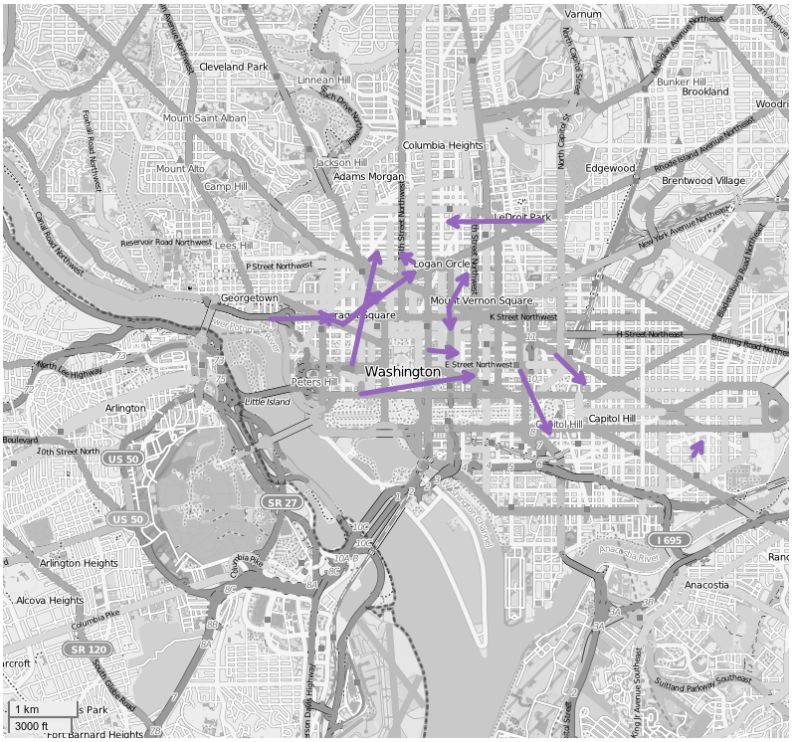


Figure 7.7: Topic 5 accounts for 14 of the 672 bike rentals for this hour

With respect to bikeshare, demand forecasting enables an operators to strategically redistribute bikes. Accurate forecasting can therefore help alleviate a crucial issue in bike-sharing systems; the unbalanced distribution in space and time of the bikes among the stations (Schuijbroek et al., 2013). For the cyclist, this would ensure that there is always an available bike at the beginning of a journey and always an empty docking station at the end of a journey.

Chapter 8

Discussion and Conclusions

We presented an approach that increases the credibility of semantic annotation of mobility flow datasets. It begins with a decomposition of aggregate flow data into several mobility modes. It then utilizes content-rich geo-referenced social media data to enrich the semantics behind the modes related to the trip purpose and user activities at destinations.

We have highlighted that the usual ad-hoc annotation based on co-occurrence, background knowledge and intuition may be misleading. As an alternative, we have introduced a rigorous approach based on causality testing. This approach opens novel perspectives for forecasting individual components of mobility flows related to specific trip purposes from relevant social media streams. This technique could enable new revenue opportunities for bikeshare operators; targeted advertisement at bike stations. If trip purpose is known ahead of time then advertisements for restaurants or services near destination stations could prove very effective.

We have also demonstrated a framework for forecasting mobility flows which illustrates the feasibility of using crowd-sourced social media data to forecast actual travel behaviours in a city. Such a system would allow bikeshare operators to minimise bike reallocation costs and maximise network through-

put by removing bottlenecks. Another benefit for the operator is the flexibility to react to large scale events announced by social media. This would enable the operator for example to provide extra bikes at unusually busy locations for the duration of a specific event. For the cyclist this means a better user experience and reassurance that their chosen bike stations will have bikes and empty docking ports available even in peak operating times.

The construction and validation of an actual prediction system is part of our ongoing work. We intend to collect social media data ranging a much larger time period than the one week case study preformed by this thesis. Such a dataset would enable us to rigorously test and compare our prediction technique against state of the art time-series forecasting approaches.

One limitation of the methodology presented here is the need for intensive human interaction; during the process of topic modelling, specifically the selection of LDA parameters (α, β and $nTopics$). The Bayesian nonparametric topic model (Teh et al., 2006), an extension of LDA, provides an elegant solution to at least the number of topics parameter. It does this by determining $nTopics$ during posterior inference, and furthermore, new documents can exhibit previously unseen topics. This parameter reduction comes at the cost of computational complexity.

A thorough critical view of our work would highlight the following. We expect mobility data to contain at least one pair of dependent topics; the morning and evening commute. If an individual participates in the morning commute it is highly likely they will also participate in the evening commute. However, the baseline implementation of LDA used does not search for volume dependent pairs of topics. The temporal decomposition is therefore naive in that it does not understand any dependency between topics. During semantic enrichment, it is up to the analyst to make connections between pairs, possibly even groups of topics. A more sophisticated technique would recognise

dependency between topics during the modelling phase.

Furthermore LDA is typically used in the domain of natural language processing. Documents from this domain do not contain the same sort of temporal dependency as the mobility based documents we have created. We expect to find both periodic topics (commuting, lunch, dinner, etc) and non periodic topics (baseball games, movie premiers, etc) within mobility data. However, again the baseline implementation of LDA used does not explicitly model topic periodicity.

In spite of the limitations noted above; we have shown through this thesis that modern techniques for spatial data analysis can greatly enhance our understanding of urban dynamics. In this light, we believe our work on temporal decomposition and semantic labelling moves us one step closer to our ultimate goal; engineering smarter transportation systems.

Bibliography

4Square (2013). 'Foursquare API.'. `https://developer.foursquare.com/` (last accessed 25.06.2013).

G. L. Andrienko, et al. (2011). 'From movement tracks through events to places: Extracting and characterizing significant places from mobility data.'. In *IEEE VAST*, pp. 161–170. IEEE.

D. Blei, et al. (2003a). 'Hierarchical topic models and the nested Chinese restaurant process'. *Advances in neural information processing systems* **16**.

D. M. Blei & J. D. Lafferty (2007). 'A correlated topic model of science'. *The Annals of Applied Statistics* pp. 17–35.

D. M. Blei & J. D. Lafferty (2009). 'Topic models'. *Text Mining: Theory and Applications* pp. 71–93.

D. M. Blei, et al. (2003b). 'Latent dirichlet allocation'. *J. Mach. Learn. Res.* **3**:993–1022.

P. Borgnat, et al. (2011). 'SHARED BICYCLES IN A CITY: A SIGNAL PROCESSING AND DATA ANALYSIS PERSPECTIVE'. *Advances in Complex Systems* **14**(03):415–438.

L. Breiman (2001). 'Statistical modeling: The two cultures'. *Statistical Science* **16**:199–231.

Z. Cheng, et al. (2011). 'Exploring Millions of Footprints in Location Sharing Services'. In *Proceedings of the 5th International Conference on Weblogs and Social Media*. AAAI.

C. Coffey, et al. (2012). 'Missed Connections: Quantifying and Optimizing Multi-modal Interconnectivity in Cities'. *ACM SIGSPATIAL, International Workshop on Computational Transportation Science* .

S. Deerwester, et al. (1990). 'Indexing by latent semantic analysis'. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* **41**(6):391–407.

J. M. Dickey (1983). 'Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses'. *Journal of the American Statistical Association* **78**(383):pp. 628–637.

M. Ding, et al. (2006). *Granger Causality: Basic Theory and Application to Neuroscience*, pp. 437–460. Wiley-VCH Verlag GmbH & Co. KGaA.

G. Doyle & C. Elkan (2009). 'Accounting for burstiness in topic models'. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 281–288, New York, NY, USA. ACM.

L. Ferrari & M. Mamei (2013). 'Classification and prediction of whereabouts patterns from the Reality Mining dataset'. *Pervasive and Mobile Computing* **9**(4):516 – 527.

C. W. J. Granger (1969). 'Investigating Causal Relations by Econometric Models and Cross-spectral Methods'. *Econometrica* **37**(3):424–438.

T. Hofmann (1999). 'Probabilistic latent semantic indexing'. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and*

development in information retrieval, SIGIR '99, pp. 50–57, New York, NY, USA. ACM.

F. Kling & A. Pozdnoukhov (2012). 'When a city tells a story: urban topic analysis'. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '12, pp. 482–485, New York, NY, USA. ACM.

T. K. Landauer, et al. (1998). 'An Introduction to Latent Semantic Analysis'. *Discourse Processes* (25):259–284.

W. Li & A. McCallum (2006). 'Pachinko allocation: DAG-structured mixture models of topic correlations'. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pp. 577–584, New York, NY, USA. ACM.

D. Lian & X. Xie (2011). 'Learning location naming from user check-in histories'. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pp. 112–121, New York, NY, USA. ACM.

Z. Liu, et al. (2011). 'PLDA+: Parallel Latent Dirichlet Allocation with Data Placement and Pipeline Processing'. *ACM Transactions on Intelligent Systems and Technology, special issue on Large Scale Machine Learning* Software available at `http://code.google.com/p/plda`.

R. Montoliu (2012). 'Discovering Mobility Patterns on Bicycle-Based Public Transportation System by Using Probabilistic Topic Models'. In P. Novais, K. Hallenborg, D. I. Tapia, & J. M. C. Rodríguez (eds.), *Ambient Intelligence - Software and Applications*, vol. 153 of *Advances in Intelligent and Soft Computing*, pp. 145–153. Springer Berlin Heidelberg.

R. Nair, et al. (2013). 'Large-Scale Transit Schedule Coordination Based on Journey Planner Requests'. *Annual Meeting of the Transportation Research Board* .

M. Padgham (2012). 'Human Movement Is Both Diffusive and Directed'. *PLoS ONE* **7**(5):e37754.

A. Pozdnoukhov & C. Kaiser (2011). 'Space-time dynamics of topics in streaming text'. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pp. 1–8, New York, NY, USA. ACM.

J. Reades, et al. (2009). 'Eigenplaces: analysing cities using the space-time structure of the mobile phone network'. *Environment and Planning B: Planning and Design* **36**(5):824–836.

J. Reisinger, et al. (2010). 'Spherical topic models'. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 903–910.

F. Rosenblatt (1961). 'Principles of neurodynamics. perceptrons and the theory of brain mechanisms'. Tech. rep., DTIC Document.

D. E. Rumelhart, et al. (1985). 'Learning internal representations by error propagation'. Tech. rep., DTIC Document.

J. Schuijbroek, et al. (2013). 'Inventory rebalancing and vehicle routing in bike sharing systems' .

M. Steyvers & T. Griffiths (2007a). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

M. Steyvers & T. Griffiths (2007b). 'Probabilistic topic models'. *Handbook of latent semantic analysis* **427**(7):424–440.

G. Sugihara, et al. (2012). 'Detecting Causality in Complex Ecosystems'. *Science* **338**(6106):496–500.

Y. W. Teh, et al. (2006). 'Hierarchical dirichlet processes'. *Journal of the American Statistical Association* **101**(476):1566–1581.

J. Toole, et al. (2012). 'Inferring land use from mobile phone activity'. In *Proceedings of the UrbComp'12*. ACM.

C. Wang & D. M. Blei (2009). 'Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process'. In *Advances in neural information processing systems*, pp. 1982–1989.

M. Ye, et al. (2011). 'What you are is when you are: the temporal dimension of feature types in location-based social networks'. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pp. 102–111, New York, NY, USA. ACM.

J. Yuan, et al. (2012). 'Discovering regions of different functions in a city using human mobility and POIs'. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pp. 186–194, New York, NY, USA. ACM.