Mining, Modeling and Predicting Mobility

THÈSE N^{O} 6662 (2015)

PRÉSENTÉE LE 18 SEPTEMBRE 2015

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE POUR LES COMMUNICATIONS INFORMATIQUES ET LEURS APPLICATIONS 4

PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Mohamed KAFSI

acceptée sur proposition du jury:

Prof. W. Gerstner, président du jury
Prof. M. Grossglauser, Prof. P. Thiran, directeurs de thèse
Prof. A. Chaintreau, rapporteur
Prof. . J. Crowcroft, rapporteur
Prof. P. Vandergheynst, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

It is good to have an end to journey toward;
but it is the journey that matters, in the end
— Ursula K. Le Guin

Acknowledgements

Before expressing my gratitude to the people whose names appear hereafter, I would like to sincerely thank all those who are not mentioned, including my friends in Tunisia and Switzerland, the numerous colleagues at EPFL, Nokia, and Yahoo!, the students I advised and assisted, the "fantastic four" of the administrative staff, and our IT team.

I had the good fortune to be supervised by two truly amazing professors, Matthias Grossglauser and Patrick Thiran. I would like express my affection and admiration for them as scientists but also as friends. As great mentors, they knew how to guide me, while giving me enough freedom to grow as a researcher. Thanks to them, I will always remember Helsinki, not as the coldest and darkest city I have ever lived in, not as the city where I learned to love saunas, but as the city where our first scientific collaboration took place and, most importantly, the city where I decided to begin this intellectual journey. Patrick and Matthias, this thesis is inscribed to you because you have been a part of its birth and growth.

I would also like to thank the great minds I met during my journey as a student and a researcher. Special thanks to Prof. Emre Telatar, the best teacher I have ever had, Olivier Lévêque and Olivier Dousse.

Naturally, I would like to thank my fellow labmates and friends in *LCA*, especially the international coffee-break crowd for the stimulating discussion around our daily intakes of caffeine. I am sure that I will miss these moments with Vincent, Julien, Mathias, Igor, Christina, Kevin, Lucas, Sébastien, Brunella, Ehsan, Farid, Lyudmila, Dan, Ghid and Adel. Special thanks to Julien (my former and future office mate) Vincent (the Coder) and Mathias (the Red) for the discussions and fun shared during our impromptu meetings in BC 202.

I am especially grateful to Sylviane, the "boss" of IC, for her support and kindness, to Holly for "hollifying" with a smile, my papers and my thesis, and to Patricia for her support.

Believe it or not, I still have friends outside LCA: Two friends I must mention are Christophe and Yann. Our friendship began during our year at Carnegie Mellon University, we shared amazing moments during our student life in Pittsburgh, American road trip, internship in Helsinki, and trips in Europe. I am lucky to have them in my life and I am sure that they will be there for me whenever I need a friend. I would also like to thank my friends from the *Tunes* gang, with whom I had amazing moments such as *Vivapoly* and *Tunnovation*.

Acknowledgements

Of course no acknowledgments would be complete without giving thanks to my whole family, especially my parents, Amel and Moncef, who taught me about the values of education ("Have you graduated yet?"), hard work and dedication to family. I would also like to thank, my sister Hela, my parents-in-law, Houria and Taoufik, and my brothers and sisters-in-law, Raja, Rifka, Bilel and Zakaria.
Last, but not least, I would like to thank my wife Zakia for her love, smile, and unconditional support throughout these years. I would like to believe that meeting her, for the first time a few days before the beginning of my thesis, happened for a reason. Since then, we have laughed and cried, traveled and played, and discussed and dreamed. Zakia, I could not have completed this journey without you by my side, and as The Beach Boys say, "God only knows what I'd be without you".

Lausanne, 2015 Mohamed

Abstract

Mobility is a central aspect of our life, and our movements reveal much more about us than simply our whereabouts.

In this thesis, we are interested in mobility and study it from three different perspectives: the modeling perspective, the information-theoretic perspective, and the data mining perspective.

For the modeling perspective, we represent mobility as a probabilistic process described by both observable and latent variables, and we introduce formally the notion of individual and collective dimensions in mobility models. Ideally, we should take advantage of both dimensions to learn accurate mobility models, but the nature of data might limit us. We take a data-driven approach to study three scenarios, which differ on the nature of mobility data, and present, for each scenario, a mobility model that is tailored for it. The first scenario is individual-specific as we have mobility data about individuals but are unable to cross reference data from them. In the second scenario, we introduce the collective model that we use to overcome the sparsity of individual traces, and for which we assume that individuals in the same group exhibit similar mobility patterns. Finally, we present the ideal scenario, for which we can take advantage of both the individual and collective dimensions, and analyze collective mobility patterns in order to create individual models.

In the second part of the thesis, we take an information-theoretic approach in order to quantify mobility uncertainty and its evolution with location updates. We discretize the user's world to obtain a map that we represent as a mobility graph. We model mobility as a random walk on this graph —equivalent to a Markov chain —and quantify trajectory uncertainty as the entropy of the distribution over possible trajectories. In this setting, a location update amounts to conditioning on a particular state of the Markov chain, which requires the computation of the entropy of conditional Markov trajectories. Our main result enables us to compute this entropy through a transformation of the original Markov chain. We apply our framework to real-world mobility datasets and show that the influence of intermediate locations on trajectory entropy depends on the nature of these locations. We build on this finding and design a segmentation algorithm that uncovers intermediate destinations along a trajectory.

The final perspective from which we analyze mobility is the data mining perspective: we go beyond simple mobility and analyze geo-tagged data that is generated by online social medias and that describes the whole user experience. We postulate that mining

Abstract

geo-tagged data enables us to obtain a rich representation of the user experience and all that surrounds its mobility. We propose a hierarchical probabilistic model that enables us to uncover specific descriptions of geographical regions, by analyzing the geo-tagged content generated by online social medias. By applying our method to a dataset of 8 million geo-tagged photos, we are able to associate with each neighborhood the tags that describe it specifically, and to find the most unique neighborhoods in a city.

Keywords: Mobility, probabilistic mobility models, individual, collective, Markov trajectories, entropy of Markov trajectories, data mining, hierarchical probabilistic model.

Résumé

La mobilité occupe une place prépondérante dans notre vie et les endroits que nous fréquentons sont le miroir de notre identité et de notre personnalité.

Ce travail de thèse a pour but l'étude de la mobilité, que nous analysons sous trois angles certes différents mais complémentaires : l'angle de la modélisation, celui de la théorie de l'information, et enfin celui de l'analyse des données.

Nous commençons par la modélisation et représentons la mobilité d'un individu en tant qu'un modèle probabiliste. Nous supposons que ce modèle est caractérisé par deux dimensions que nous définissons formellement : la dimension individuelle, qui reflète la spécificité de l'individu (personnalité, centres d'intérêt), et la dimension collective, qui reflète des notions partagées par tous les individus (normes sociales, contraintes géographiques). Il s'agit de présenter trois types de modèles probabilistes en expliquant les scénarios et le type de données pour lesquelles ils conviennent. Le premier scénario est celui du modèle purement individuel qui convient dans le cas où les données sur les individus sont indépendantes, ce qui implique qu'on ne peut pas recouper les données entre individus. Le deuxième scénario est celui du modèle collectif, dans lequel nous supposons que les individus peuvent être regroupés de telle sorte que chaque groupe est composé d'individus ayant des mobilités similaires. Le troisième scénario est le scénario idéal car nous pouvons exploiter aussi bien la dimension individuelle que la dimension collective.

L'approche théorie de l'information se focalise sur la mesure de l'incertitude sur la mobilité d'un individu et l'évolution de celle-ci lorsque cet individu révèle des endroits par lesquels il est passé. Notre approche se base sur une discrétisation du monde de l'utilisateur où sa mobilité est représentée comme une marche aléatoire sur un graphe de mobilité. Un nœud du graphe représente un endroit sémantique comme la maison ou le lieu de travail, alors qu'une arrête représente la possibilité de transition entre deux endroits. Étant donné que la trajectoire d'un individu représente sa mobilité, le degré d'incertitude sur sa mobilité est quantifié par l'entropie de la distribution de trajectoires. Ce degré d'incertitude évolue quand l'utilisateur partage un sous-ensemble des endroits visités, ce qui correspond formellement à l'entropie de sa trajectoire conditionnée sur ces endroits intermédiaires. L'une des contributions principales de cette thèse est une méthode qui permet de calculer l'entropie des trajectoires markoviennes conditionnelles. Cette méthode est basée sur la transformation de la chaîne de Markov originale, de sorte que la distribution de trajectoires dans la chaîne de Markov obtenue est égale à la distribution

conditionnelle de trajectoires dans la chaîne de Markov originale. Nous appliquons cette méthode afin analyser les trajectoires de milliers d'utilisateurs et montrons un lien entre la position d'un endroit dans le graphe de mobilité et son impact sur l'entropie de trajectoires. Nous montrons également que les points intermédiaires qui augmentent l'entropie conditionnelle sont plus enclins à être des destinations intermédiaires.
La troisième et dernier angle, est celui de l'analyse de données sous lequel on considère la mobilité non plus comme une simple séquence d'endroits visités, mais comme une riche expérience vécue par l'utilisateur. Analyser les traces numériques relatives à cette expérience nous permet non seulement de la décrire mais aussi de décrire l'espace dans lequel elle se manifeste. Nous concrétisons cette idée en développant une méthode probabiliste qui nous permet de retrouver ce qui rend une zone géographique unique. En utilisant cette méthode, nous analysons des millions de photos géo-taguées et obtenons une description spécifique de chaque quartier de New York et de San Francisco.
Cette méthode nous permet également de quantifier à quel point un quartier est unique et de comparer des quartiers de deux villes différentes.

Mots clefs : Mobilité, mobilité individuelle, mobilité collective, modèle probabiliste de mobilité, trajectoires Markoviennes, entropie de trajectoires Markoviennes, analyse de données, modèle probabiliste hiérarchique.

Contents

1 Introduction

Your imagination leaves digital traces.
— Bruno Latour

1.1 Motivation

A major characteristic of modernity, as well as post-modernity, is spatial mobility: "Modern society is a society on the move" [6]. Spatial mobility is an essential component of the way societies organize space, and it traditionally refers to a geographical displacement, i.e, "the movement of entities from an origin to a destination along a specific trajectory that can be described in terms of space and time" [43]. These entities can be tangible (e.g., objects, animals or people) or abstract (e.g., information, ideas or social norms). With the technological developments in transport and communication systems, the speed of displacements has significantly increased, which has compressed distances and globalized the mobility of ideas and people.

Spatial mobility is much more than a link from an origin to a destination; it is a structuring dimension of life in society. At the collective level, the change in mobility patterns is at the base of fundamental societal changes, which explains the fact that the social, cultural, economic and political consequences of these dynamics are much debated in the social sciences. Human geographers, for example, study the spatio-temporal patterns in the flow of migrant populations, and historians study military campaigns and analyze the movements of armies (the Russian military campaign of Napolean is depicted in Figure 1.1) and populations. Urban planners study the mobility in urban spaces to detect patterns and support decision makers with the knowledge needed to enhance urban mobility and reduce congestion, accidents and pollution.

At the individual level, mobility reveals much more about individuals than simply their whereabouts, as people tend to favor places where they feel comfortable and avoid areas where they do not fit in [60]. In fact, our daily mobility routine reveals our social status, and the locations we visit reflect our habits, tastes and certain personality traits such as our degree of extraversion and openness. Many location-based services, such as personal navigation systems or intelligent personal assistants, benefit from learning our mobility patterns and predicting our future whereabouts. Moreover, these services offer a unique economic opportunity as merchants and companies can target precisely a given audience (based on age, gender, and consumption habits) and promote their product with high effectiveness.

Studying both individual and collective mobility patterns is therefore crucial in many fields including social sciences, history, urban planing and economy. In the twentieth century,

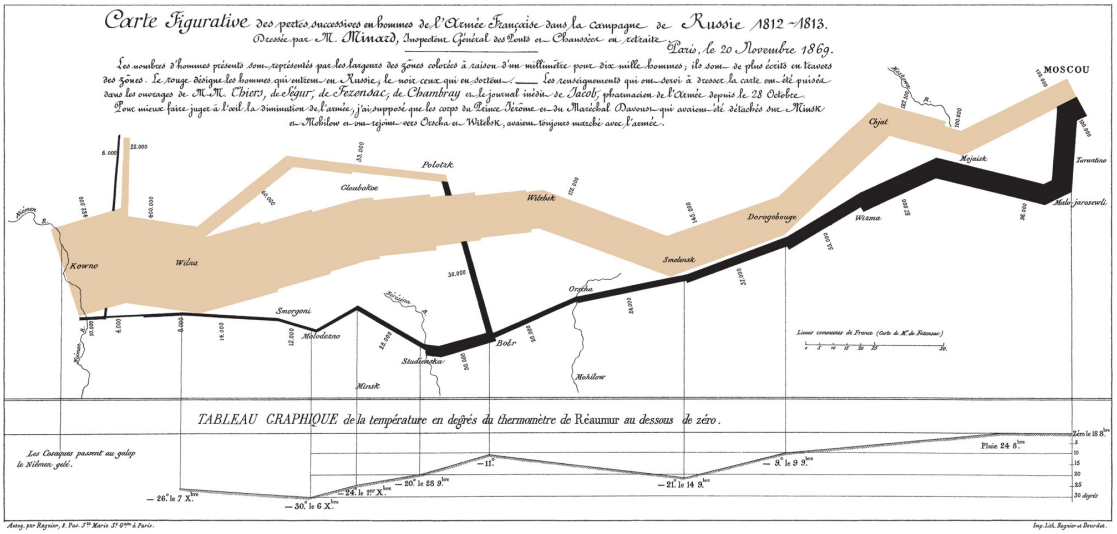


Figure 1.1 – Charles-Joseph Minard created this cartographic depiction of Napoleon's disastrous Russian campaign of 1812. The width of the line represents the size of his army at specific locations during their advance (brown line, from left to right) and retreat (black line from, from right to left). It displays different types of data that include the temperature and location relative to specific dates.

studies about human behavior were based on two types of data: *surface data* about many people and *deep data* about a few individuals or a small group [7]. Surface data is used in all fields that adopted quantitative approaches (i.e., statistical or computational techniques for data analysis), whereas deep data is used in the humanities such as qualitative schools of psychology, anthropology and ethnography. For example, a quantitative sociologist works with surface data as he analyzes census data that describes most of a country's citizens. However, this data is collected only every ten years and describes each individual macroscopically, without her opinions, emotions, tastes and motivations. In contrast, a psychologist studies the behavior of a few individuals over a long period of time

and collects personal data that a census is unable to capture. In between these two methodologies, surface data and deep data, we find sampling methods. By carefully choosing a sub-population (sample), researchers take a result found for a few and generalize it into knowledge about many. For example, the Swiss Federal Statistical Office [4] conducts, every 5 years, a survey about mobility in Switzerland. The most recent survey was based on a sample of 62,868 individuals, nearly 1% of the population of Switzerland. Based on data obtained through phone interviews, the survey describes the mobility of the Swiss population and analyzes, for example, the link between mobility and socio-demographic attributes.

Data about this sub-population, however, does not tell us anything about the actual day-to-day mobility patterns of every individual or every household. Moreover, even for the individuals who are in the studied sub-population, data does not tell us everything about their individual mobility patterns. It does not tell us about their favorite places, their favorite neighborhoods or the regions where they like to spend the weekends.

With the rise of social media comes the promise of deep data about millions of individuals: we no longer have to choose between depth and scale. There are many reasons for the data explosion. The first reason is technology, as the capabilities of digital devices soar and the number of people that have access to powerful mobile devices increases. The second reason is that practically everything on the Internet is recorded. In fact, we leave digital traces as we are moving, communicating, maintaining social relationships, making purchases or traveling. The field that can benefit the most from this data deluge is the study of human behaviors, and in particular human mobility. In fact, with the democratization of GPS-enabled smart phones, every digital experience can be associated with a location (geo-tagged). As a consequence, deep data, which used to describe the behavior of a few individuals, is now about millions of mobile individuals: we have access to a geo-tagged description of the experiences of millions of individuals as they are moving.

In conjunction with the tremendous data explosion, our capacity to extract knowledge from unstructured data is continuously increasing because of the emergence of new computational tools [1–3, 19, 23, 30] that can handle massive amounts of data. The development of these computational tools is driven by the open-source software community. The Hadoop [1] open-source implementation of Map-Reduce [23], for example, is playing a crucial role in the "Big data" ecosystems, and is adopted by numerous startups but also industry heavyweights such as IBM and Oracle.

In 2007 [48], Bruno Latour summarized the opportunity offered by the digitization of our life as follows: "The precise forces that mould our subjectivities and the precise characters that furnish our imaginations are all open to inquiries by the social sciences. It is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable."

1.2 Thesis Objectives

This thesis addresses a number of important questions regarding mobility. We identify fundamental research challenges that can be summarized with the three following points

Modeling and predicting individual mobility. *How can we learn accurate and interpretable mobility models?*

Quantifying mobility uncertainty. *How can we rigorously quantify mobility uncertainty and its evolution with location updates?*

Mining geo-tagged data. *How can we take advantage of geo-tagged data in order to characterize the whole experience surrounding mobility?*

Each of these questions is associated with a specific perspective from which we study mobility.

1.3 Dissertation Outline

1.3.1 The modeling perspective

The first perspective from which we study mobility is the modeling perspective. We assume that mobility data is a realization of a probabilistic process that is described by both observable and latent variables. For example, the observable variables might correspond to location or time, whereas latent variables might correspond to abstract concepts such as individual emotion or tastes. We postulate that, in order to learn accurate individual mobility models, we need to take advantage of the interplay between the individual (e.g., taste, habits) and collective (e.g., social groups, urban environment) dimensions that influence mobility. To do this, we introduce formally, in Chapter 2, the notion of individual and collective dimensions in a mobility model, and we link these dimensions to the nature of the parameters that describe the model. Moreover, we explain how the nature of the data from which we learn the mobility model might limit the dimensions we can take advantage of; ideally, we should take advantage of both individual and collective dimensions to learn accurate individual mobility model.

First, we study the case where the nature of data limits us to individual specific models. In Chapter 3, we illustrate this situation by describing our participation in the Nokia Mobile Data Challenge [28, 29]: The data provided consists of a rich set of features —actual deep data —recorded from the smartphones of 170 individuals, but all sensitive data was processed to make it individual specific thus preventing cross-referencing people and places between users. We show that, given enough data at hand, we are still able to predict accurately the future whereabouts of an individual but the predictive power of individual-specific models diminishes as soon as data about an individual is sparse.

We then present, in Chapter 4, the collective models that we use to overcome the sparsity of individual traces, and for which we assume that individuals can be assigned to mobility groups so that individuals of the same group share the same mobility patterns. We illustrate this approach by our work on a mobility model [42] of Ivory Coast population learned from Call Data Records of 5 Million Orange customers in Ivory Coast. This mobility model is a central component of epidemic mitigation strategies that are based on personalized mobility recommendation to individuals.

Finally in Chapter 5, we study the ideal scenario where we can take advantage of both the individual and collective dimensions in order to learn individual mobility models. We illustrate this scenario with our work on modeling the behaviors of ants, which is based on the analysis of a dataset that describes, with high precision, the behaviors of nearly 1000 ants distributed in 6 colonies. We show that our approach enables us not only to uncover the fundamental collective behaviors in the colony but also to express the mobility of each ant as a time-dependent random combination of these collective behaviors.

1.3.2 The information-theoretic perspective

In the second part of the thesis, we take an information-theoretic approach to quantify mobility uncertainty and its evolution as additional information becomes available. In contrast to the tailored models of the first part of this thesis, the mobility model we consider is general enough to represent most situations while still capturing mobility patterns. We discretize the user's world to obtain a map that we represent as a mobility graph: A vertex represents a branch point where the user takes the decision about where to move next and it can correspond, for example, to a semantic place (e.g., home, work place); an edge represents a direct transition between two vertices. We model mobility as a random walk on this graph, or equivalently a Markov chain; and in order to quantify the uncertainty about the user's mobility, we choose to compute the entropy of the distribution over all possible trajectories between the source and destination vertices. To do so, we use the result of Ekroot and Cover [26], which enables us to compute the entropy of Markov trajectories. For this model, a location update amounts to conditioning the trajectory on a given sequence of vertices. Hence, we need to compute the entropy of Markov trajectories *conditional* on a set of intermediate states.

In Chapter 6, we introduce one of the main contributions of this thesis [41]: We propose a method for computing the entropy of conditional Markov trajectories through a transformation of the original Markov chain into a Markov chain that exhibits the desired conditional distribution of trajectories. Computing the entropy of conditional Markov trajectories enables us to quantify the change of entropy for each location update, given the model and the previously revealed locations.

In Chapter 7, we apply this result to quantify the evolution of the uncertainty about a user's mobility as she reveals intermediate locations. Furthermore, we empirically link the evolution of conditional trajectory entropy as we reveal a location along a trajectory and the nature of this intermediate location. We build on this finding and design an algorithm that uncovers intermediate destinations along a trajectory.

1.3.3 The data mining perspective

The last part of this thesis is dedicated to the data mining perspective: We postulate that mining the geo-tagged digital traces of a user enables us to gain more insight, not only into her personal experience, but also into the nature of the environment where she moves.

We propose, in Chapter 8, a method [40] that enables us to uncover a specific description of a geographical region, by analyzing the geo-tagged content generated by online social medias. The method is based on a hierarchical probabilistic model that encodes the assumption that the data observed in a region is a random mixture of terms generated by different levels of a geographical hierarchy. Our model further quantifies the diversity of local content, for example, by allowing for the identification of the most unique or most generic regions amongst all regions in the hierarchy. We apply our method to a dataset of 8 million geo-tagged photos taken in the neighborhoods of San Francisco and New York City and described by approximately 20 million tags. We are able to associate with each neighborhood the tags that describe it specifically and coefficients that quantify its uniqueness. This enables us to find the most unique neighborhoods in a city and to find mappings between similar neighborhoods in both cities.

1.4 Related work

As modeling mobility is at the heart of this thesis, it is important that we present a selection of articles that are representative of the evolution of the research on mobility modeling. Mobility is intrinsically related to human behavior which explains that studying and modeling human mobility sparked interest in different research communities that include wireless networks [18, 45], social networks [70] and statistical physics [34, 63]. From this rich literature, we distinguish two main approaches to mobility modeling: descriptive and predictive.

The descriptive approach, which originates from the physics community, focuses on describing the statistical properties of human mobility in order to find universal laws or probability distributions that characterize human mobility. In migration theory and urban planning, the gravity model of mobility [5, 75] and the intervening opportunity model [68] are used to model mobility flows between an origin and a destination. On one

hand, the gravity model of mobility, inspired from Newton's law of universal gravitation, assumes that humans move like particles whose behaviors are governed by a law similar to the physical law of gravitational attraction. The mobility flow between an origin and a destination areas is proportional to their importance (e.g., population size or gross domestic product) and inversely proportional to the distance between them. On the other hand, the intervening opportunity model, introduced by Samuel Stouffer [68], states that migration is influenced most by the opportunities to settle at the destination (e.g., jobs), less by distance or population pressure at the starting point.

Until recently, a large scale empirical validation of these models suffered from the lack of large-scale datasets that describe the movements of individuals. The situation, however, changed when Call Data Records (CDRs) became available to a few research groups. In fact, mobile phone operators record the nearest antenna each time a user initiates a call or send a text message. By associating, for each communication event, the location of a user with the location of the antenna that handles this communication, we obtain a relatively low-resolution version of the mobility of all individuals. Using this approach, Gonzalez et al. [34] analyzed the mobility of 100,000 mobile phone users whose whereabouts are tracked for a six-month period. To model mobility, they analyzed the empirical distributions of mobility-related statistics such as the distance between consecutive locations visited or the radius of gyration, and found that the power-law distribution characterizes well the distribution of these quantities. The power-law distribution of the travel distance reflects, for instance, the frequent existence of short movement and the rare presence of very long movements.

The universality of these results, however, have to be taken with a grain of salt because they depend heavily on the dataset analyzed. For instance, Isaacman et al. [39] study the mobility of hundreds of thousands mobile-phone users in Los Angeles and New York and demonstrate clearly different mobility patterns between the two cities (for example different distributions of travel distance).

The descriptive approach to mobility focuses on the properties of collective mobility and is therefore suitable for the applications that depend on mobility-related quantities. This is the case, for example, of opportunistic wireless networks as the distribution of inter-contact times impacts the performance in terms of delivery delay [18]. The descriptive approach, however, does not enable us, as opposed to the predictive approach, to predict the future locations an individual will visit.

The predictive approach focuses on the implementation of methods that predict accurately the locations an individual will visit in the future. We distinguish two groups of mobility predictors: The first group is composed of predictors that are based on probabilistic models [11, 20, 65] such as Markov chains or Bayesian networks, whereas the second group is composed of "black box" approaches such as support vector machine or artificial neural networks. Among the probabilistic models, the Markov chain is a very popular model for

mobility. Song et al. [65] evaluate several location predictors using a two-year trace of the mobility of over 6000 users, represented as the sequence of Wi-Fi access points detected. The most accurate predictor is a second order Markov chain with a fallback mechanism for unseen contexts. Cho et al. [20] model individual mobility as a time-dependent mixture of Gaussian whose components represent latent semantic locations such as home and work place.

In this thesis, we take a predictive approach, based on probabilistic models, to model human mobility. This approach enables us not only to model individual mobility accurately, but also to have interpretable results as we encode the relationships between different mobility-related variables. Again, analyzing the statistical properties of human mobility does not enable us to predict accurately the future whereabouts of individuals. The opposite is, however, true. In fact, as collective mobility is the sum of individual mobilities, accurate individual models enable us to model collective mobility, and capture the statistical properties that are the focus of the descriptive approach to mobility.

1.5 Contributions

- The individual-specific mobility model we created is an important component of the mobility-prediction algorithm that enabled our team to win the Nokia Mobile Data challenge (7% more accurate than the runner-up).
- The model we use to analyze ant behaviors enables us not only to uncover the fundamental behaviors in ant colonies but also to express the behavior of each as a function of these behaviors. Moreover, the results of our model match the results that are hand-annotated by domain experts.
- We propose a closed-form expression that enables us to compute the entropy of Markov trajectories under conditions weaker than those assumed in [26].
- We express the entropy of Markov trajectories —a global quantity —as a linear combination of *local entropies* associated with the Markov chain states.
- We propose a method to compute the entropy of conditional Markov trajectories through a transformation of the original Markov chain so that the transformed Markov chain exhibits an (unconditional) distribution of trajectories equal the desired conditional distribution of trajectories in the original Markov chain.
- We use the trajectory entropy framework to analyze individual trajectories. We quantify the evolution of the uncertainty about the mobility of a user as intermediate locations are revealed. We also design a recursive algorithm, based on trajectory entropy, that uncovers intermediate destinations along a trajectory.
- We propose a probabilistic hierarchical model that enables us to find a specific description of a region within a given geographical hierarchy. We apply our method

to a dataset of 8 million geo-tagged photos described by approximately 20 million tags. This enables us to find specific descriptions of the neighborhoods of San Francisco and New York, and to find mapping between similar neighborhoods in both cities.

The Modeling Perspective Part I

2 From Collective to Individual Mobility

Society exists only as a mental concept;
in the real world there are only individuals.
— Oscar Wilde

2.1 Introduction

Collective and individual mobility are interdependent. On one hand, we can see collective mobility as the aggregation of complex individual mobilities that are influenced by personal attributes such as culture, home and work places. On the other hand, individual mobility is shaped by social groups (friends and family) and collective dynamics (urban transportation, cultural events). Ideally, we should take advantage of both individual and collective dimensions in order to develop accurate mobility models for individuals. However, we are often limited by the nature of the data available for learning these models. This limitation is dictated mainly by the trade-off between the richness of the data and the privacy of individuals.

In this chapter, we distinguish three general scenarios that differ in the characteristics of the available data, and we present different modeling approaches tailored for each scenario. We begin by studying the case of individual-specific data. In such a scenario, the inability to cross reference data from different users limits our arsenal to individual mobility models that are independent of each other. In the second scenario, the data about the mobility of an individual is extremely sparse. We can, however, take advantage of the individual's attributes in order to create homogeneous groups whose individuals exhibit similar mobility patterns. Finally, we explore the ideal scenario for which we have complete information about the behaviors and attributes of individuals. We will

show that in such a scenario, we are able not only to uncover the fundamental collective behaviors but also to express the mobility of each individual as a combination of these collective behaviors.

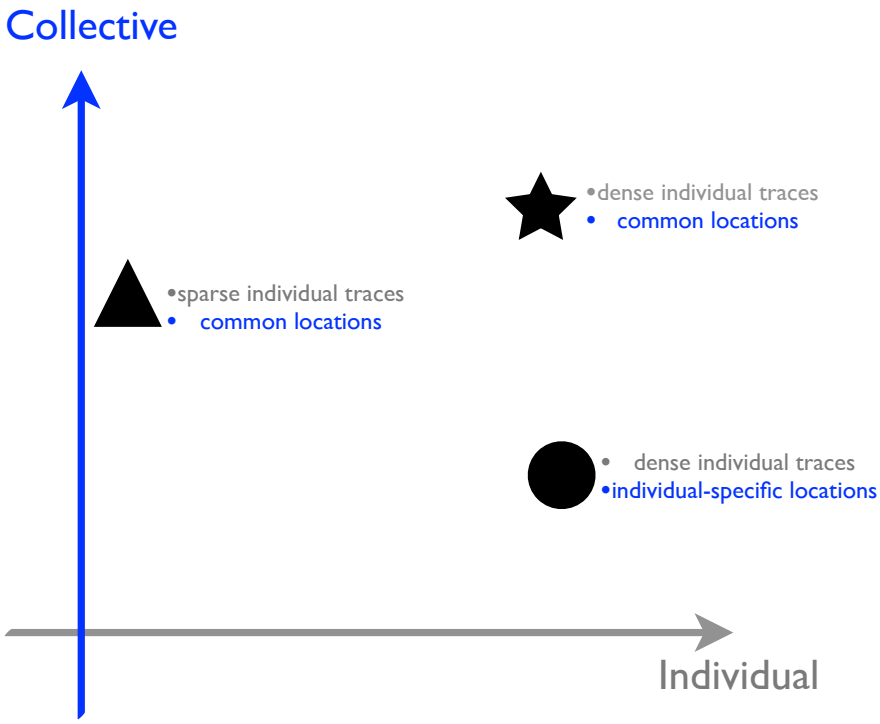


Figure 2.1 – In this thesis, we propose mobility models that are tailored to the position of the dataset in the individual-collective space. A dataset that is characterized by dense but individual-specific mobility traces (circle) is adequate for a individual specific model whereas a dataset composed of numerous but very sparse individual traces fits well to collective models. The ideal situation is the situation for which the dataset (star) enables us to learn a model that takes advantage of both individual and collective dimensions. An example of such a dataset includes dense individual traces in addition to traces of interactions between individuals.

2.2 Models

In this section, we introduce formally the notion of individual and collective dimensions in mobility models. The abstract setting we consider is as follows: we observe the behavior of a set of N individuals that are indexed by the integer $i \in \{1, \ldots, N\}$. We assume that data about individuals is not fully observed: the behaviors of these users can be described using both observable and latent (i.e., non observable) variables. For example, the observable variables might correspond to the locations visited by an individual whereas the latent variables represent abstract concepts that cannot be directly observed, such as psychological state, individual taste or the membership to a social category. We denote

the set of all observable variables by $\boldsymbol{X}$, in which $\boldsymbol{X}_i$ represents data for individual i, and similarly we denote the set of latent variables by $\boldsymbol{Z}$, in which $\boldsymbol{Z}_i$ represents the latent variables for individual i. Without loss of generality, we assume that a latent variable corresponds to a discrete component label in $\{1, \ldots, C\}$, which implies that the distribution of the observed variables is a mixture of distributions. Introducing latent variables in our models enables us to describe the individual behaviors by using relatively complex distributions that are formed from simple components.

Each model is described by a set of parameters: We denote the set of model parameters that are linked to the observed variables $\boldsymbol{X}$ by $\boldsymbol{\Theta}$, and we denote the set of parameters linked to the latent variables $\boldsymbol{Z}$ by $\boldsymbol{\Pi}$. The parameters $\boldsymbol{\Theta}$ represent the individual components of the mixture and might consist, for example, of normal distributions each with its own mean and independent covariance matrix. The parameters $\boldsymbol{\Pi}$ are called the mixture coefficients and can be interpreted as the prior probabilities for the value of the latent variable.

Throughout this thesis, we use graphical models (Bayesian networks) in order to represent formally the structure of the probabilistic models and the dependence between their components. In order to estimate the parameters of our model, we maximize the log-likelihood of the observed data

$$\log p\left(\boldsymbol{X}|\boldsymbol{\Theta}, \boldsymbol{\Pi}\right) = \log \sum_{\boldsymbol{Z}} p\left(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta}, \boldsymbol{\Pi}\right). \tag{2.1}$$

A key observation is that the summation over the latent variable is inside the logarithm, which prevents the logarithm from acting directly on the joint distribution, resulting in complicated expressions for the maximum-likelihood solution. A powerful method for finding maximum likelihood solutions for models with latent variables is called the *Expectation-Maximization* or *EM* algorithm [24]. It is based on the idea that our state of knowledge about the latent variables $\boldsymbol{Z}$ is given only by the posterior distribution $p\left(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}\right)$, and hence we need to maximize the expected data log-likelihood under this posterior distribution. Throughout this thesis, we will apply this algorithm to find the best parameters for different models with latent variables.

The individual dimension of a model is captured by the parameters that are individual-specific, whereas the collective dimension is captured by the parameters that are shared by a group of individuals. We therefore categorize a model as being individual, collective or a combination of both as a function of the nature of its parameters. In this thesis, we distinguish the three following types of model.

Individual model In Figure 2.2, we show a Bayesian network that represents an individual-specific model: we suppose that the data $\boldsymbol{X}_i$ that describe the behavior of individual i is explained by the parameters ($\boldsymbol{\Theta}_i$ and $\boldsymbol{\Pi}_i$) that are specific to this user,

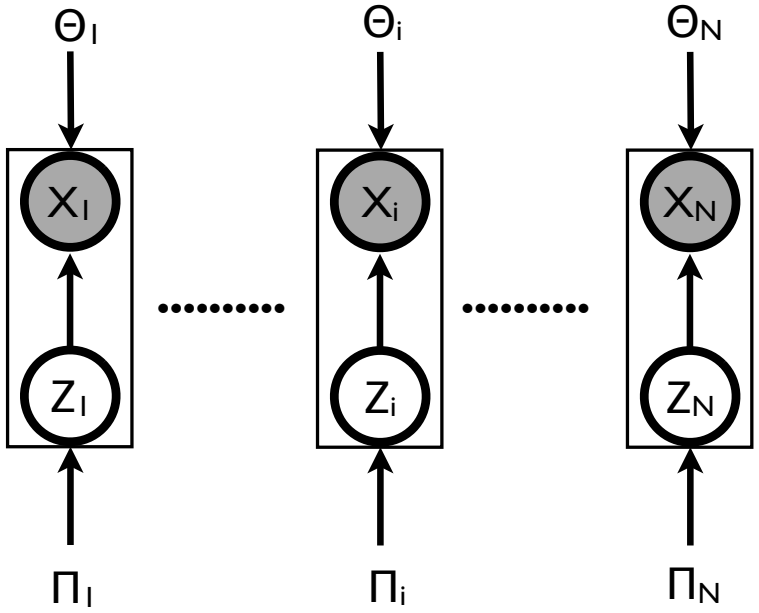


Figure 2.2 – Bayesian network that represents independent individual models with individual specific parameters only. The behavior of each individual is described by a mixture of distributions whose components and mixture coefficients are individual specific.

which implies that these parameters are As a consequence, these parameters are learnt from the data associated with individual i only, and the model has a high predictive power if the data about this individual is dense enough.

We illustrate this scenario in Chapter 3 through examples taken from our participation in the *Nokia Mobile Data Challenge* [49]. In this challenge, the dataset is characterized by dense individual traces that are not expressed as a function of the same set of locations.

Collective model In Figure 2.3, we show a Bayesian network that represents a collective model. We group the individuals in $K \ll N$ disjoint groups where the observations associated with each group k are explained by the parameters ($\mathbf{\Theta}_k$ and $\mathbf{\Pi}_k$). These parameters are shared by the individuals that form group k and learnt from the data associated with them. As a consequence, this model is adequate when (a) individual traces are very sparse, and (b) we are able to group users given their attributes. These attributes can correspond, for instance, to socio-demographic categories if we assume that individual from the same socio-demographic category exhibit similar behaviors. Naturally, when the number of groups K is equal to the number of individuals N, the collective model boils down to the individual-specific model.

In Chapter 4, we illustrate such a model with a scenario for which we overcome the

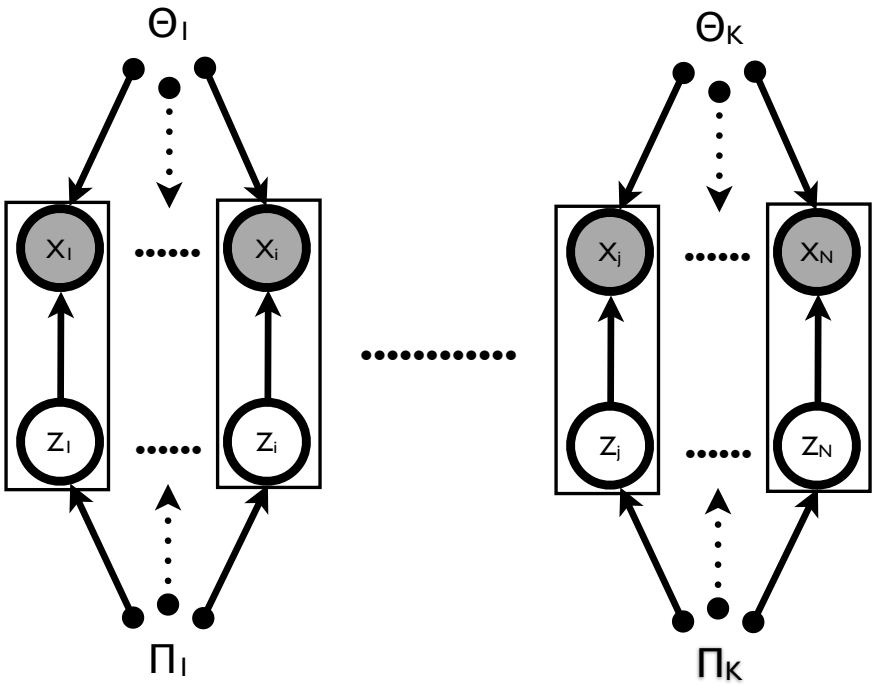


Figure 2.3 – Bayesian network that represents collective models composed of K groups. The behavior of each individual is described by a mixture of distributions whose components and mixture coefficients are specific to the group this individual belongs to.

sparsity of individual traces by taking advantage of the collective dimension and by assuming that individuals who share the same home region exhibit similar mobility patterns.

Individual-Collective model In Figure 2.4, we show a Bayesian network that represents an individual-collective model. With such a model, the observations associated with each individual are explained by both individual-specific parameters $\mathbf{\Pi}_i$ and collective parameters $\mathbf{\Theta}$. The collective parameters $\mathbf{\Theta}$ are shared by all individuals and learnt from the whole dataset, whereas each individual set of parameters $\mathbf{\Pi}_i$ is learnt from the data associated with individual i only. In such a model, the mobility of an individual is represented by a mixture of distributions whose components ($\mathbf{\Theta}$) are shared collectively and mixture coefficients $\mathbf{\Pi}_i$ are individual-specific. This model is adequate when two conditions are met (a) the individual traces are very dense, and (b) the individual traces are expressed in the same vocabulary.

In order to study the behaviors of ant colonies, we introduce such a model in Chapter 5. We show how it enables us to uncover collective behaviors in ant colonies and to describe the behavior of each individual ant as a combination of these behaviors.

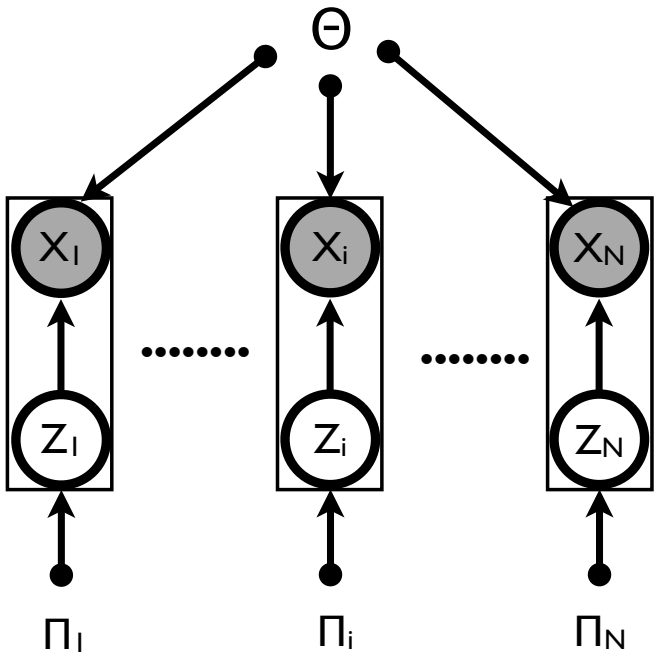


Figure 2.4 – Bayesian network that represents an individual-collective model. The behavior of each individual is represented by a mixture of distributions whose components $\boldsymbol{\Theta}$ are shared by all individuals, and mixture coefficients $\boldsymbol{\Pi}_i$ are individual specific.

Parameters Comparing between the number of parameters associated with different models enables us to compare between the predictive abilities of these models for the same number of observations: the larger the number of free parameters, the more complex is the model and the higher the risk of overfitting for the same number of observations. Without loss of generality, we assume that the observed variables are discrete random variables that take value in $\{1, \ldots, V\}$. As a consequence, the distribution of the observed data is a mixture of C multinomials each described with $V - 1$ parameters.

For the individual model, the behavior of each individual is described by $C(C-1)(V-1)$ parameters, making $NC(C-1)(V-1)$ parameters in total.

For the collective model, the behavior of each group is described by $C(C-1)(V-1)$ parameters, making $KC(C-1)(V-1)$ parameters in total. This suggests that we can compensate the sparsity of individual traces by a large number of individuals in each group.

For the individual-collective model, the behavior of each individual is described by $(C-1)$ parameters and each collective parameter is described by $(V-1)$ parameters, making $N(C-1) + C(V-1)$ parameters in total.

For the same number of data points, the individual collective model generalizes better

than the individual specific model: the number of free parameters for each individual is $C(C-1)V$ for the individual model, whereas it is $(C-1)$ for the individual-collective model. Moreover, the collective parameters acts as a regularization procedure because the estimation of the model's parameters improves by learning from the data of all individuals.

3 Individual Mobility

Modeling individual mobility would certainly improve by taking advantage of mobility correlation between individuals. However, in some situations, data is made user specific in order to protect the privacy of users. The data released should not, for example, enable an attacker to infer that two individuals were in the same location at the same time. This limits the type of models to individual specific models for which we can take advantage of the individual dimension only. In this chapter, we explore this scenario through examples taken from our participation in the Nokia Mobile Data Challenge (NMDC) [46]. For this challenge, the data publicly released has been anonymized and made user specific in order to ensure that the privacy of the individuals is respected.

3.1 Mobile Phone Dataset

The NMDC is [49] "a large-scale research initiative aimed at generating innovations around smartphone-based research, as well as community-based evaluation of related mobile data analysis methodologies". It was organized by Nokia and took place from January to April 2012. It featured an open track, in which participants were able to propose their own problems to study, and three dedicated tracks, each defining a specific problem for teams to solve: semantic place prediction, next-place prediction and demographic attributes prediction.

At the heart of this challenge is the dataset gathered during the Lausanne Data Collection Campaign (LDCC) [46]. This dataset is an example of deep data as we describe it in Chapter 1: Smartphones are allocated to nearly 170 participants for periods of time ranging from a few weeks to almost two years. A data collection application runs on the background of these phones, yielding a digital representation of the participants life: social interactions (e.g., phone calls, text message), location data, media creation and usage (e.g., images and videos) and behavioral data (e.g., accelerometer and application usage). In order to preserve the privacy of the participants when sharing the data with

the research community, the location identifiers were encrypted using user-specific keys. As a consequence, researchers are unable to assess whether two users have visited the same location.

The *Next-Place Prediction* challenge, the focus of this chapter, was assigned a subset of 80 users. For each user, the last 50 days of data were kept as a test set for the evaluation of each team's submissions, and the rest was used as training data.

For privacy reasons, all identifiers (phone numbers, WLAN SSIDs, contact names, etc.) were encrypted, but more importantly, physical locations were not released. Instead, for each user, the organizers of the NMDC first identified *places* - corresponding to discs with a 100-meter radius - by using both GPS and WLAN data. Then, they represented each place by a unique identifier. Consequently, a sequence of geographic coordinates is represented as a sequence of place identifiers.

These visits represent the basic unit for the prediction task. They are defined by their starting and ending times, and the corresponding place. In addition, several types of data are available: accelerometer, application usage, GSM, WLAN, media plays, etc. The complete list can be found in the dataset description [49].

We present below two major constraints (imposed by the rules of the NMDC) that restrict the range of methods we could use, which makes our task more challenging:

User specificity for privacy. To prevent cross-referencing people and places between users, all sensitive data are user-specific: The identifiers are encrypted using different keys, and places are independently defined and numbered for each user. Moreover, the rules of the challenge explicitly forbade all participants to reverse this process, or to make some links between users. We were therefore not allowed to build joint models over the user population, *i.e.*, to learn from one user to make a prediction about another. For this reason, we build individual specific mobility models that are independent of each other.

Memory-less predictors. The input for the Next-Place Prediction task is a *current* visit, along with all additional data recorded from a user's phone during a time. However, we do not have access to the *history* of the user, *i.e.*, the sequence of previous visits. If we did, we could develop higher-order predictors that not only take into account the current place but also the sequence of places visited just before. Indeed, such information is very useful: If a user is currently at a transportation hub, e.g., a bus station, knowing whether he was home or at work just before greatly helps in predicting his next move. Because this information is not available to us in this challenge, we limit ourselves to *memoryless* predictors, *i.e.*, methods that take into account only the current context, without any knowledge of the past.

We now explore some characteristics of the dataset, and define the framework within which we develop our model.

Dataset characteristics We show in Figure 3.1 an intuitive representation of the mobility traces of three users selected from the dataset. The figure depicts a user's behavior over one year as a matrix, where each column is a day of the year and each line an interval of one hour. We map each place to a color and leave blank intervals of time during which we have no information about the user's location.

User 143, whose mobility is represented in Figure. 3.1a, has a very regular behavior, which seems to support the results (such as those presented by Song et al. [64]) claiming that human mobility is very predictable. However, similarly to User 1 (Figure 3.1c), the majority of users show no clear regular pattern in their behavior. Of course, a lack of visual regularity does not imply that there is no latent structure in a user's mobility. We will see in Section 3.3 that we can still predict the behavior of such users with reasonable accuracy.

We summarize below some salient characteristics of the data that we believe are critical to the prediction task:

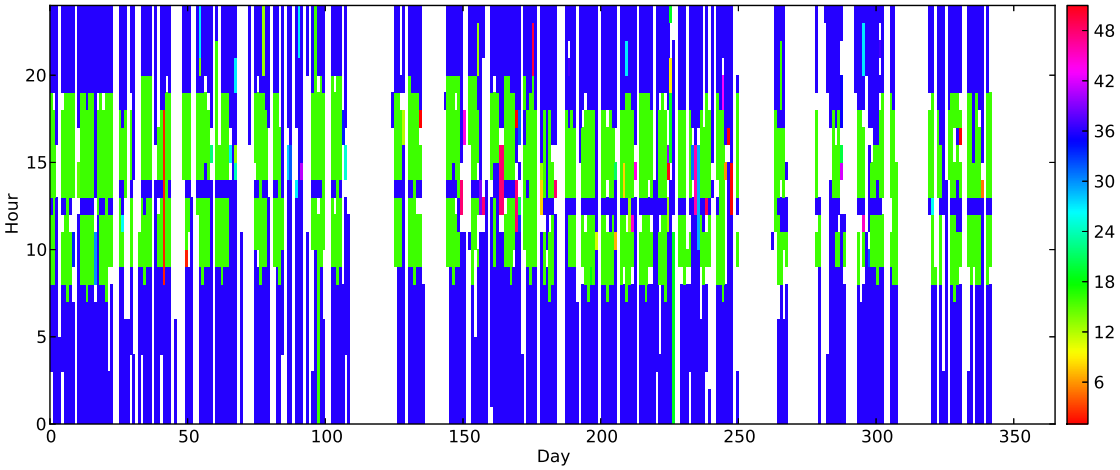
Non-stationarity. We often observe a significant change in users' habits over time, as illustrated in Figure 3.1b. The fact that some users change their home or work location right at the end of the observation period complicates the prediction task. To overcome this, we implement aging mechanisms, as described in Section 3.3. Moreover, to get a realistic estimation of our predictors' performances, we keep the last part of the dataset as testing data, as explained in Section 3.3.1.

Data gaps. We experience, for some users, periods (ranging from a few hours up to a few months) with no information about their behavior. Moreover, as shown in Figure 3.1c, these gaps are sometimes followed by a change of mobility habits. To limit the effect of such transitions, our model takes into account the possibility that we have missed some data between two detected visits.

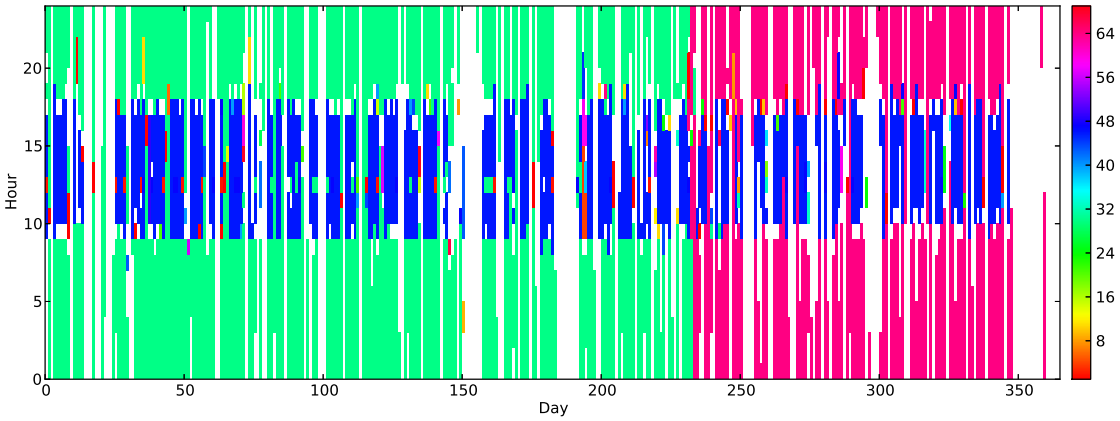
Sparsity. The period of observation for some users is too short (less than 15 days) to reflect faithfully their mobility patterns.

We believe that taking the above observations into account in the design of predictors has a significant effect on their prediction accuracy.

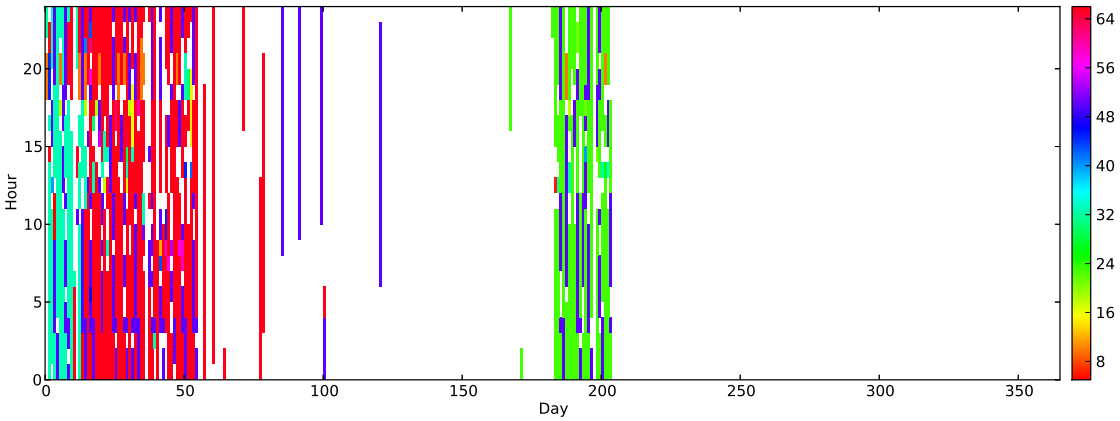
Before formally introducing our mobility model, we need to define the variables that describe the dataset. During the study period, a user makes a certain number of visits of variable duration to L distinct places, represented by the set $\mathcal{L} = \{1, \ldots, L\}$.



(a) User 143: regular mobility pattern.



(b) User 13: regular mobility pattern with change of home.



(c) User 1: irregular mobility pattern characterized by data gaps and non-stationarity.

Figure 3.1 – Mobility of users over a year, shown as matrices where each column is a day of the year and each line an interval of one hour. We map each place to a color, and leave blank intervals of time during which we have no information about the user's location. Figure (a) illustrates the behavior of a very regular user, Figure (b) a home change, and Figure (c) data gaps and non-stationarities.

Definition	Domain	Explanation
L	$\mathbb{N}$	Number of distinct places
$\mathcal{L}$	$\{1, \ldots, \mathrm{L}\}$	Set of visited places
k	$\mathbb{N}$	Time resolution
$X(n)$	$\mathcal{L}$	Place
$T_s(n)$	$\mathbb{N}$	Absolute starting time
$H_s^k(n)$	$\{1, \ldots, k\}$	Quantized starting hour
$D_s(n) = \mathrm{day}(T_s(n))$	$\{1, \ldots, 7\}$	Starting day
$W_s(n) = \mathrm{weekday}(T_s(n))$	$\{0, 1\}$	Indicates whether the visit starts on a weekday
$T_e(n)$	$\mathbb{N}$	Absolute ending time
$H_e^k(n)$	$\{1, \ldots, k\}$	Quantized ending hour
$D_e(n) = \mathrm{day}(T_e(n))$	$\{1, \ldots, 7\}$	Ending day
$W_e(n) = \mathrm{weekday}(T_e(n))$	$\{0, 1\}$	Indicates whether the visit ends on a weekday
$U(n)$	$\{0, 1\}$	Indicates whether there might be an unobserved place between $X(n)$ and $X(n+1)$

Table 3.1 – List of the definition and domain of the variables relative to an individual, as well as those describing his n^{th} visit.

In Table 5.2, we list the variables corresponding to an individual, as well as those relative to his n^{th} visit. All time-relative variables are derived from the starting and ending times, which are given as absolute times. The binary variable $U(n)$, which is given as a feature in in the NMDC dataset, indicates whether there might be an unobserved place between $X(n)$ and $X(n+1)$. This situation arises typically when location data are partly available between the two visits. In such cases, we say that the transition from $X(n)$ to $X(n+1)$ is not necessarily *direct*.

To allow for various quantizations of the day, we introduce a *time resolution* parameter k. This lets us consider a coarser segmentation of the day: instead of always splitting a day into 24 hours, we can choose to split it into k time periods. For instance, if $k = 2$, $H_s^k(n) \in \{1, 2\}$, with $H_s^k(n) = 1$ corresponding to the n^{th} visit starting between midnight and noon. Such a coarse segmentation can be helpful when training predictors for a user for which few data are available.

3.2 Model

We model the mobility patterns of individuals as a Dynamical Bayesian Network (DBN). A DBN involves a modeling phase where we express causal relationships and independence

assumptions between the features that describe the mobility of an individual. The assumptions in our model are as follows: The next place a user will visit depends on his current place and on the time at which he leaves it. The dependence between the current and next place is strong when the difference between the ending time of the current visit and the starting time of the next one is small (typically the case for direct transitions). However, as this time difference increases, the influence of the present place on the next one fades away and the starting time of the next visit bears increasing importance.

As we do not know the starting time of the next visit, the main challenge is to model its randomness, given carefully chosen information about the current visit.

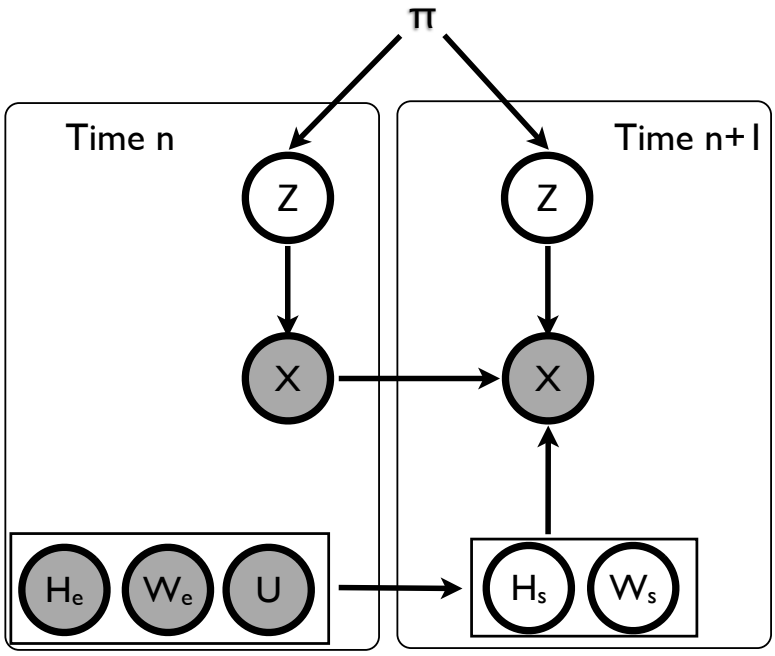


Figure 3.2 – Graphical model representing the DBN associated with a user. The conditional distribution of the next place $p\left(X(n+1)|X(n), H_e(n), U(n), W_e(n)\right)$ is a random mixture of place-dependant $p\left(X(n+1)|X(n)\right)$ and time-dependent $p\left(X(n+1)|H_e(n), W_e(n), U(n)\right)$ distributions. Note that the structure of the DBN reflects the conditional independence of $X(n+1)$ and $(H_e(n),\ W_e(n), U(n))$ given $(H_s(n+1), W_s(n+1))$.

As shown in Figure 3.2, our DBN captures these intuitions: The conditional distribution of the next place $p\left(X(n+1)|X(n), H_e(n), U(n), W_e(n)\right)$ is a random mixture of place- and time-dependent distributions

$$\pi p\left(X(n+1)|X(n)\right) + (1-\pi) p\left(X(n+1)|H_e(n), W_e(n), U(n)\right),$$

where $0 \leq \pi \leq 1$ is the mixture parameter that governs the contribution of each

distribution.

For ease of notation, we omit the time resolution parameter k and assume that it is fixed. The place-dependent distribution

$$p(X(n+1)|X(n)) \tag{3.1}$$

is simply a first order Markov chain that encodes the frequency of transitions between places.

Using Bayes' rule, we express the time-dependent distribution

$$p\left(X(n+1)|H_e(n), W_e(n), U(n)\right) \tag{3.2}$$

as

$$\sum_{W_s}\sum_{H_s}\Big\{p\left(X(n+1)|H_s(n+1), W_s(n+1), H_e(n), W_e(n), U(n)\right) \\ p\left(H_s(n+1), W_s(n+1)|H_e(n), W_e(n), U(n)\right)\Big\}.$$

Note that the conditional distribution

$$p\left(H_s(n+1), W_s(n+1)|H_e(n), W_e(n), U(n)\right) \tag{3.3}$$

models the randomness of the starting time of the next visit $(H_s(n+1), W_s(n+1))$ given the ending time of the current one $(H_e(n), W_e(n))$ and the directness of the transition $U(n)$. In addition to reflecting the temporal rhythm at which a specific user moves from one place to another, the conditional distribution (3.3) also captures the randomness of the data gaps. Empirically, we observe that direct transitions usually imply a shorter time interval between the visits. This is not surprising: If the transition between the n^{th} and $(n+1)^{\text{th}}$ visits is direct $(U(n)=0)$, then we are sure that there are no intermediate visits between them. The main assumption we make when designing our DBN is that $X(n+1)$ is independent of $H_e(n)$, $W_e(n)$ and $U(n)$ given $H_s(n+1)$ and $W_s(n)$. We can therefore write (3.2) as

$$\sum_{W_s}\sum_{H_s} p\left(X(n+1)|H_s(n+1), W_s(n+1)\right)\, p\left(H_s(n+1), W_s(n+1)|H_e(n), W_e(n), U(n)\right).$$

The assumption of independence makes sense, as knowing the time $(H_e(n), W_e(n))$ at which a user leaves his current place is not informative (with respect to the next place $X(n+1)$) if we know the starting time of the next visit $(H_s(n+1), W_s(n+1))$.

We show in Figure 3.2 a graphical model that represents our DBN. We formulate the mixture of distributions with respect to a latent variable $\boldsymbol{z}$: We introduce a binary random vector $\boldsymbol{z}$, whose dimension is the number of visits, that indicates, for each visit, the distribution from which it was sampled. In other words, $\boldsymbol{z}_i = 1$ means that the i^{th} visit is sampled from the place-dependent distribution (3.1), whereas $\boldsymbol{z}_i = 0$ implies that it is sampled from the time-dependent distribution (3.2).

The choice of the model structure and variables is driven by our intuition and confirmed by empirical evidence. We tested several variants of our model: For example, we incorporated in our DBN the distribution $p(X(n+1)|X(n), U(n))$ instead of the distribution $p(X(n+1)|X(n))$ to check whether the directness of the transition contains information about the next place. However, the prediction accuracy decreased. Furthermore, the lack of data prohibits us from learning more sophisticated distributions, as over-fitting a small training set leads to very poor generalization.

Maximum likelihood solution To predict the user's next place using our model, we need to estimate, for each user of our dataset, the corresponding DBN's parameters. Estimating the joint distribution (3.3) about the start time of the next visit is straightforward: we use a maximum likelihood estimator on the data about the transitions from one place to the other. However, the learning procedure for the parameters π, $p(X(n+1)|X(n))$, $p\left(X(n+1)|H_s(n+1), W_s(n+1)\right)$ is more complicated because of the dependence between them: We then use an *Expectation-Maximization* algorithm [13] to maximize the likelihood of the data with respect to the model parameters. Moreover, the structure of the DBN enables us to derive closed-form expressions for the update of the model parameters.

3.3 Mobility Prediction from Instantaneous Information

Overcoming non-stationarity

The idea of introducing *aging* mechanisms in the learning process is based on the observation that, when a user changes his habits, recent history is more representative of his future behaviour than the accumulated information. The first method we use to reduce the negative effect of non-stationarity on the prediction performance is the introduction of an aging mechanism, governed by an aging parameter. An example of an aging process is to introduce a multiplicative parameter that intervenes in the learning process to reduce the count (contribution) of old samples. As a result, recent samples will have more of an effect on the user's mobility model. The second method is an algorithm that detects changes in home locations and adapts the learning process accordingly. At any moment t, we define home as the place where the user spends more than $T_{threshold}$ hours of his sleeping periods during the interval of time $[t - T_{history}, t]$. The parameter

Algorithm 1: Home-change detection algorithm

Input: *visits*, $T_{threshold} > 0$, $T_{history} > 0$, *sleeping period*.
Output: visits.

```
for v ∈ visits do
    t ← starting time of the visit v;
    home candidate ← place where the user spent most of his sleeping period in
    [t − T_history, t] ;
    T_candidate ← time spent in home candidate during the sleeping period in
    [t − T_history, t] ;
    if (T_candidate ≥ T_threshold) then
        add home candidate to home list;
final home ← last element of home list ;
for v ∈ visits do
    if (v.place belongs to home list) then
        (v.place belongs to home list)
return visits
```

$T_{history}$ controls to which extent we keep in memory the past behavior of the user. At the end of the observation period corresponding to the training set, the user who changes his habits will have at least two places flagged as home. We declare the last place flagged as such as his *final home*. More importantly, the history of visits is modified as if the user's home has always been his *final home*. Such modification enables us to capture the user's habits while avoiding the lengthy process of adapting to a home change. The pseudo-code of the home-change detection algorithm[1] is shown in Algorithm 1. Empirical results show that applying our home-change detection algorithm results in a significant improvement in the prediction accuracy for the users who change their habits during the observation period.

3.3.1 Training

The training procedure is as follows: we separate the data about an individual into three parts that are illustrated in Figure 3.3. We define the first 80% of the visits as the training set (set A), the following 10% as the validation set (set B), and the last 10% as test set (set C). Finally, we denote set D the undisclosed part of the dataset, on which the performance of our algorithm for the Next-Place Prediction Challenge is computed. The reason we divide the dataset deterministically is that our goal is to maximize the prediction accuracy on set D. In fact, if we take into account the non-stationarity of the data, we expect set D to be much more similar to the end of the dataset than to its beginning. Moreover, even if an individual's behavior is globally non-stationary, it

[1] Based on empirical evidence, we choose $T_{history} = 14$ days, $T_{threshold} = 18$ hours and *sleeping period* to be between 3 a.m. and 6 a.m.

usually shows regular patterns over smaller time intervals. Having set C as close in time as possible to set D maximises the likelihood of their samples belonging to the same "stationary" period. Moreover, by training our models on "past" data and evaluating them on very recent data (set C), we can test whether they are able to adapt to users' changes of habit.

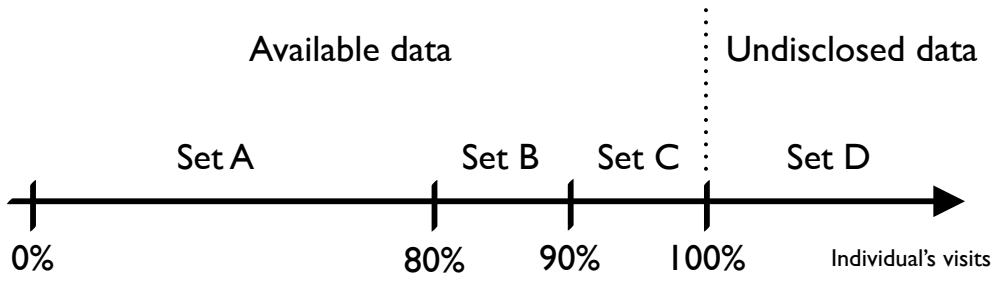


Figure 3.3 – Separation of the data associated with an individual. We define the first 80% of the visits as set A, the following 10% as set B, and the last 10% as set C. Finally, we denote set D the undisclosed part of the dataset, on which the final performance for the NMDC are computed.

3.3.2 Evaluation

Prediction accuracy To evaluate the performance of a predictor, we measure its prediction accuracy, *i.e.*, the proportion of samples for which it successfully predicts the next place. First, consider a predictor ϕ: It takes as input $\boldsymbol{v}^{(n)}$, the data corresponding to the n^{th} visit, and outputs a probability distribution over the possible next places. More formally, a predictor is a function

$$\phi \colon \mathcal{V} \to \{\boldsymbol{x} \in [0,1]^L \colon \sum_{l=1}^{L} x_l = 1\},$$

where $\mathcal{V}$ is the space of data corresponding to a visit.

The place $\hat{X}_n^\phi$ predicted by ϕ for the visit $\boldsymbol{v}^{(n)}$ is hence the most likely next place

$$\hat{X}_n^\phi = \arg\max_{l \in \mathcal{L}} (\phi(\boldsymbol{v}^{(n)}))_l,$$

where $(\phi(\boldsymbol{v}^{(n)}))_l$ is the l^{th} component of the vector outputted by ϕ when given the data corresponding to the n^{th} visit as input, *i.e.*, the probability that the next visited place is l.

Finally, we define the prediction accuracy $A_S(\phi)$ of the predictor ϕ over the samples in

set S as

$$A_S(\phi) = \frac{1}{|S|} \sum_{i \in S} I_{\left\{\hat{X}_i^{\phi} = X(i+1)\right\}},$$

where $|S|$ is the number of samples in set S, $X(i+1)$ is the true next place corresponding to the ith visit, and $I_{\{A\}}$ is the indicator function, taking value 1 if the event A is true, and 0 otherwise.

Baseline methods In addition to our model we consider the following baseline predictors

Most visited
Always predicts that the next visit is to the most visited place.

First order Markov chain (MC)
First order Markov chain that encodes the probability of transitions between places. It predicts that the next visit is to the most likely location given the current location.

Artificial neural network (ANN)
We train for each individual a two-layer artificial neural network with different sets of features that include, among others, location, time of the day, day of the week, and charging. We combine these different features in an exhaustive way and obtain more than 200 ANNs for each user. We then choose the ANN that maximizes the prediction accuracy on the validation set C.

Results We show in Table 3.2, for different predictors, the prediction-accuracy on set C averaged over all users. With a prediction accuracy of 0.52, our DBN outperforms all the other predictors. In spite of their very close prediction accuracies, DBN has a crucial advantage over ANN: we can relate its parameters to the behavior and habits of users. For example, the distribution $p\left(H_s(n+1), W_s(n+1) | H_e(n), W_e(n), U(n)\right)$ encodes the randomness of the starting time of the next visit, given the ending time of the current one, whereas the mixture coefficient π reflects the extent to which the current place and time have an influence the next place to visit.

NMDC results The accuracy of the predictors participating in the Nokia Next-Place Prediction Challenge were evaluated on the undisclosed set D. We proposed a predictor [28, 29] that is a combination of the DBN introduced here, an ANN and a Gradient Boosted Decision Tree (GBDT). As shown in Table 3.3, our predictor outperformed all the other competitors with an average prediction accuracy of 0.56. The runner-up predictors were a periodicity based model [71] and a smoothed spatio-temporal model (HPHP) [37].

	Average accuracy
Most visited	0.35
MC	0.44
ANN	0.51
DBN	**0.52**

Table 3.2 – Prediction accuracy on set C.

	Average accuracy
Periodicity model	0.52
HPHD	0.52
Our method	**0.56**

Table 3.3 – Prediction accuracy on set D.

Limitations Individual specific models has severe drawbacks when data is sparse. Indeed, a model with individual specific parameters that are learnt from a few samples leads inevitably to overfitting and therefore poor prediction performance. We show in Figure 3.4 the histogram of accuracy on set C. We observe a high variance in the predictability of users: We reach a prediction accuracy of 100% for the most predictable user, whereas we predict correctly 0% of the time for the least predictable one. In addition to the intrinsic unpredictability, the major factor causing such a poor performance is the conjunction of individual specific models and the lack of data: the data available about the user, for which we make no correct prediction, span over a period of only 12 days. In Chapters 4 and 5, we will present individual mobility models that take advantage of the collective dimension in order to predict accurately the mobility of individuals, even if the individual traces are sparse.

3.4 Conclusion

We have introduced in this chapter an example of individual specific models that are tailored to scenarios where mobility data is user specific. We have showed, through the description of our participation in the Next-Place Prediction Challenge organized by Nokia, that we are able to learn accurate individual models given dense individual traces. However, the Achilles heel of these models is data sparsity as a few data samples about an individual leads inevitability to overfitting and results in models that have poor predictive power. In Chapter 4 and 5, we present a solution to overcome this data sparsity; we enhance the individual model by taking advantage of the collective dimension of mobility.

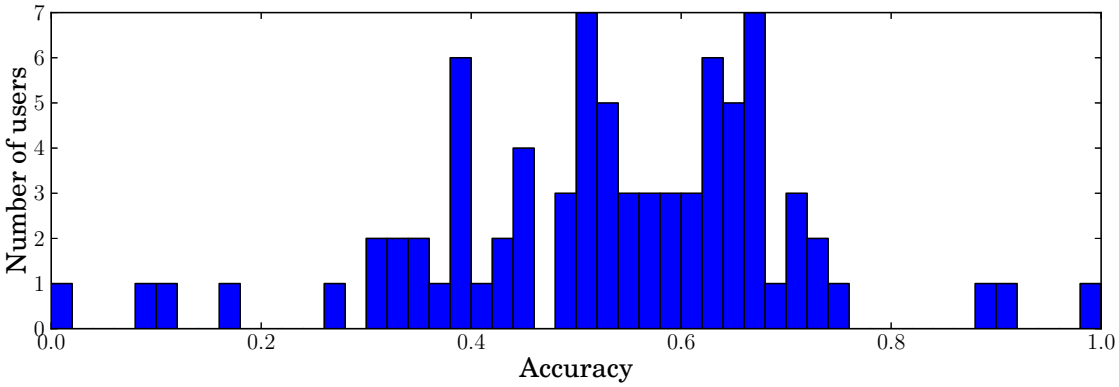


Figure 3.4 – Histogram of the accuracy on set C, for all users. For each user, we chose the best predictor on set B. This figures shows that there is a high variance in the predictability of users.

4 Collective Mobility

In Chapter 3, we have seen that individual-specific models are suitable when individual traces are very dense but fail to predict the behavior of individuals for which we have few data samples. In this chapter, we are interested in the modeling challenges faced when data about the behavior of *each* individual is very sparse. This can be the case, for example, when the sampling rate is very low or the observation period very short. In such scenarios, the naive approach of building independent individual models fails because generalizing from few samples leads to severe over-fitting. We can, however, take advantage of the collective behavior in order to enhance individual models. This solution comes at the price of a simplifying yet reasonable hypothesis: We assume that individuals can be assigned to mobility groups so that individuals of the same group share similar mobility patterns. To illustrate this approach, we present our work about the modeling of population mobility based on the analysis of Call-Data Records of Orange customers in Ivory Coast. This mobility model is then used to anticipate and influence the mobility of individuals in order to mitigate effectively the spread of an epidemic.

4.1 Mobility and Epidemics

The effective mitigation of the spread of infectious diseases is a long standing public health goal. The stakes are high: throughout human history, epidemics have had significant death tolls. In 430 BCE, the overcrowded city of Athens was devastated by a plague that killed an estimated one-third to two-thirds of the population. More recently, the region of West Africa has witnessed the largest Ebola outbreak ever documented, with around 24,000 cases and 9400 deaths[1]. For these epidemics, human mobility clearly plays a crucial role as it enables the epidemic to spread geographically. In fact, the mobility of infectives defines the locations where they interact with healthy individuals, and therefore the epidemic propagation network. Consequently, an accurate mobility model is a mandatory step towards an accurate modeling of the spread of a disease.

[1] http://apps.who.int/ebola/ accessed February 2015

4.2 Call-Data Records Dataset

With around five million customers, Orange has a significant market share in Ivory Coast, whose population is estimated to be around 20 million individuals. The Orange "Data For Development" dataset [16] is based on anonymized call-data records of phone calls and SMS exchanges between five million of Orange's customers in Ivory Coast between December 1, 2011 and April 28, 2012. This amounts to 2.5 billion calls and SMS exchanges between five million customers. The format of CDRs data is

```
time, caller id, callee id, call duration, antenna id
```

Moreover, we have for each antenna the corresponding, but slightly blurred, location. The dataset of interest contains high resolution trajectories of 50,000 randomly selected individuals over two-week periods. The original raw data was split into consecutive two-week periods. For each period, 50,000 of the customers are randomly selected and then assigned anonymized identifiers. To protect their privacy and avoid the identification of customers based on observing their trajectories over a long time period, identifiers are regenerated every two weeks. As a consequence, the mobility traces of each individual are very sparse: they are composed of the locations of the antennas from which they made a phone call or sent an SMS over a two-weeks period only.

4.3 Collective Mobility Models

The challenge here is to create accurate individual mobility models, despite the sparsity of the individual traces. To do this, we explore the dimension of collective mobility by assuming that the population can be divided in mobility groups. Each mobility group is composed of individuals who share common attributes and exhibit similar mobility characteristics. Ideally, these attributes could include workplace, gender, age, ethnicity, socio-economic status and hobbies. However, in our situation, we have no personal information about individuals but only a description of their call activity. Hence, the attribute that defines mobility groups is limited to statistics that are related to CDRs.

We choose home location as the attribute that defines mobility groups: the home location of an individual strongly shapes her mobility patterns because the places she visits regularly (e.g., workplace, school or shopping center) depend on their proximity to home. Typically, we expect the most visited location (home) and the second most visited location (school, university or work) to be geographically close to each other. In addition to this geographical component, mobility is strongly time-dependent: individuals commute between home and work during the weekdays, with a substantial change in travel behavior during the weekends.

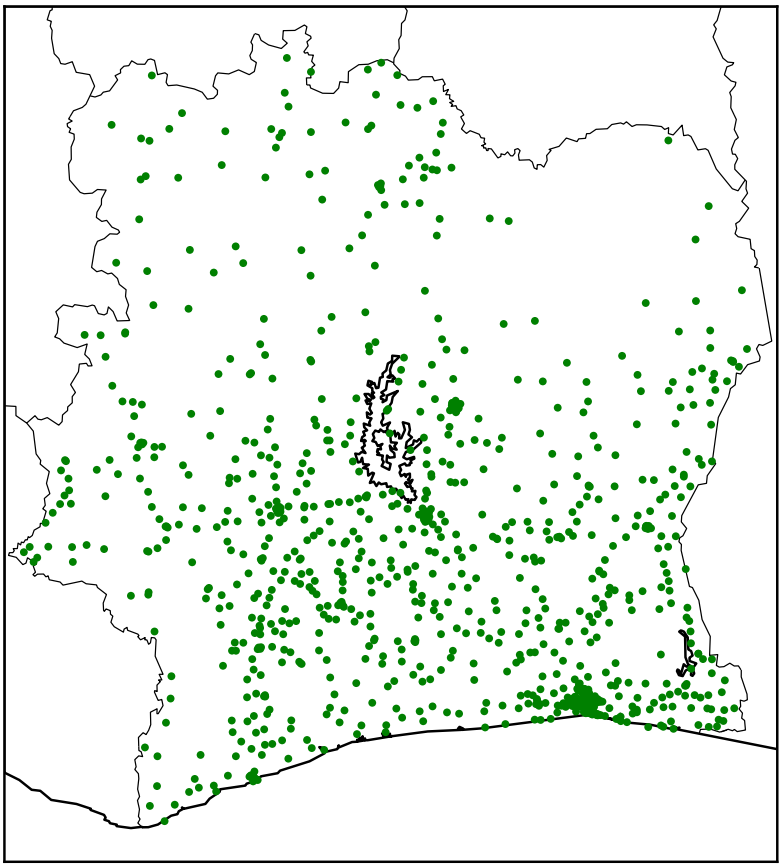


Figure 4.1 – The geographical reparation of the Orange antennas in Ivory Coast.

We therefore make the assumption that the individuals who share the same home-location exhibit a similar time dependent mobility pattern, and we construct a location and time-based mobility model that depends on the variables presented in Table 4.1. The conditional distribution of the location $X(n)$ of user u depends on his home antenna $a_{\text{home}}(u)$, but also on the time of the visits $(h^k(n), w(n))$:

$$p\left(X(n)|u, t(n)\right) = p\left(X(n)|h^k(n), w(n), a_{\text{home}}(u)\right). \tag{4.1}$$

First, we choose the time resolution $k = 3$, in order to divide a day in 3 distinct periods: morning (6 am to 1 pm), afternoon (1 pm to 8 pm) and night (8 pm to 6 am). Second, conditioning on the day type $w(n)$ enables us to distinguish between weekdays and weekends. Finally, the home antenna $a_{\text{home}}(u)$ of user u is defined as the most visited antenna during the night period. Consequently, given the period of the day, the day type and the home antenna of user u, the distribution of the location she might visit (4.1) is a multinomial distribution with $|\mathcal{A}|$ categories. The correponding graphical model is shown in Figure 4.2.

Definition	Domain	Explanation
$\mathcal{A} = \{1, \ldots, 1231\}$	-	Set of antennas
$\mathcal{SP} = \{1, \ldots, 255\}$	-	Set of sub-prefectures
k	$\mathbb{N}$	Time resolution
$sp_{\text{home}}(u)$	$\mathcal{SP}$	Home sub-prefecture
$a_{\text{home}}(u)$	$\mathcal{A}$	Home antenna
$X(n)$	$\mathcal{A}$	Antenna
$t(n)$	$\mathbb{N}$	Absolute time
$h^k(n)$	$\{1, \ldots, k\}$	Period of the day
$d(n) = \text{day}(t(n))$	$\{1, \ldots, 7\}$	Day of the week
$w(n) = \text{weekday}(t(n))$	$\{0, 1\}$	Day type: weekday or weekend

Table 4.1 – List of the definition and domain of the variables relative to user u, as well as those describing his n^{th} visit.

Maximum likelihood solution In order to assign non-zero probability to locations that were not observed for a given group, we assume that each multinomial distribution (4.1) is drawn from an exchangeable Dirichlet distribution, which implies that the inferred distribution is a random variable drawn from a posterior distribution conditioned on the training data. A more detailed description of this smoothing procedure is given by [15].

Evaluation In order to evaluate our mobility model, we separate the data into two parts: For each user, we put 90% of the samples in the training set and the remaining 10% in the test set. First, we build a mobility model by learning from the training set by using a maximum likelihood estimator. Then, we test the accuracy of our model by computing the average log-likelihood of the samples in the test set; the average log-likelihood then reflects how well our model generalizes to unseen data.

We tested several variants of mobility models by varying their structure and parameters (time resolution, day of the week, etc). Among these predictors, we present three representative baseline models

Time-based Mobility (TM)

The first baseline model is a time-based mobility model defined by

$$p\left(X(n)|u, t(n)\right) = p\left(X(n)|h^k(n), w(n)\right), \tag{4.2}$$

which implies that all users exhibit the same time-dependent mobility pattern.

Markov Chain (MC)

the second baseline method is a location-dependent first order Markov Chain defined

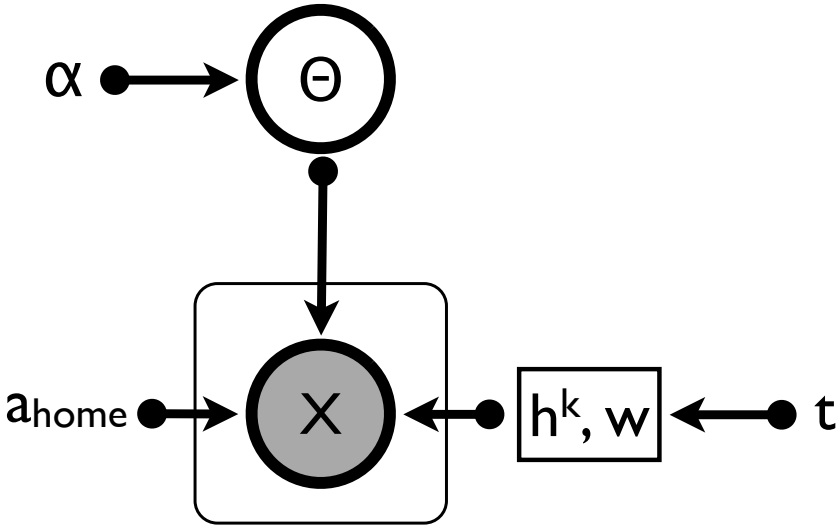


Figure 4.2 – Graphical model representing the collective mobility model for user whose home antenna is a_{home}. The location of this user, at time step n, depends on his home antenna a_{home}, the period of the day $h^k(n)$ and the day type $w(n)$.

by

$$p\left(X(n)|u, t(n), X(n-1), \ldots, X(0)\right) = p\left(X(n)|X(n-1)\right) \tag{4.3}$$

where the current location of a user depends only on the location he visited just before.

Sub-Prefecture Mobility (SPM)

The third baseline is a time and sub-prefecture dependent Mobility model defined by

$$p\left(X(n)|u, t(n)\right) = p\left(X(n)|h^k(n), w(n), sp_{\text{home}}(u)\right). \tag{4.4}$$

In other words, all users who share the same home sub-prefecture exhibit a similar mobility pattern. This is similar to our approach, as it assigns users to mobility groups depending on their personal attributes, but is different in the granularity of aggregation level.

The experimental results are shown in Table 4.2. It is not surprising that the first-order Markov chain (MC) performs the worst, because the time difference between two call-records varies greatly, ranging from a few minutes to a few days. The location associated to a call made a few hours or days ago does not necessarily have an effect on the current location. As the location data is sporadic, it is not surprising that time-based mobility models perform better than any model that is based on learning from transitions. Our model performs the best, and by comparing it to the time-based model (TM), we realize that knowing the home-locations of users contributes the predictive power of the mobility

Mobility model	Average log-likelihood
MC	-6.49
TM	-2.9
SPM	-1.67
Our model	**-1.07**

Table 4.2 – Log-likelihood of the unseen data from the test set. Our mobility model significantly outperforms the baseline models since its predictive power, with respect to the test set, is higher.

model. Moreover, the granularity of this home location is crucial: Our model significantly outperforms the sub-prefecture dependent mobility because it has a finer home-location granularity.

4.4 Conclusion

In this chapter, we have focused on the collective mobility model for which we assume that individuals can be assigned to groups whose members exhibit homogeneous behaviors. We illustrate such an approach with our work on modeling the mobility of a population in order to anticipate the propagation of an epidemic. Despite the sparsity of the individual traces, we are able to create accurate individual mobility models by grouping the data of individuals based on their home antenna. Our empirical analysis, conducted on the data of Orange customers in Ivory coast, reveals the importance of a wise choice of the attribute that defines mobility groups. The approach we take is to consider different options (home antenna vs. home sub-prefecture) and to choose the one with the maximum predictive power on a test set. As it is unfeasible to exhaustively test all possible attributes, our choice might be sub-optimal.

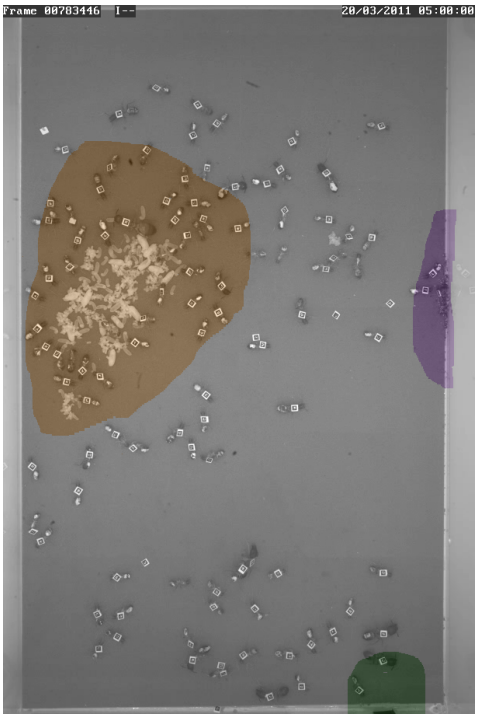
In the next chapter, we will show that, given enough data about each individual, we are able to automatically detect these homogeneous mobility groups that explains the best the variance observed in the data.

5 From Collective to Individual Mobility

In this chapter, we study the scenario where we take advantage of both the individual and collective dimensions in order to learn an individual mobility model. This mobility model goes further than the models introduced in Chapter 3 and 4 as it captures not only the individual mobility patterns but also the correlation between the behaviors of different individuals. It takes advantage of the collective dimension in order to improve the modeling of individual behaviors. We illustrate this scenario with our work about modeling the behaviors of ants. We show that our approach enables us not only to uncover the fundamental collective behaviors in the colony but also to express the mobility of each ant as a time-dependent random combination of these collective behaviors.

5.1 Camponotus Fellah Dataset

The dataset we use in this section was collected by Mersch et al. [54] who recorded, using an automated video tracking system, the activity of six colonies that are composed of a total of 956 *Camponotus fellah* ants. The set-up is divided in a nest area (Figure 5.1a) and a foraging area (Figure 5.1b) made of Plexiglas. Each area is rectangular and has on its short side an exit hole of 10 mm diameter that is connected to a tunnel and that enables ants to move from one area to the other. The foraging area contains a water source, food and liquid honey. Both areas are filmed from above with a high resolution camera (4560×3048 pixels) equipped with an enlarging lens and infrared light flash. The queens and all workers are marked with a unique barcode that allows for their identification in each image taken by the camera: their locations are estimated with a mean precision of 2.37 pixels (i.e., 0.14 mm, $0.8\% - 2\%$ of a *Camponotus fellah* ant). The continuous recording (2 images per second) results in a rich dataset that describes with great accuracy and resolution the mobility of each ant. Table 5.1 presents different statistics of the dataset.



(a) Nest area of colony 4.

(b) Foraging area of colony 4.

(c) Nest area of colony 18.

(d) Foraging area of colony 18.

Figure 5.1 – We show the nest and foraging areas of colony 4 and 18. The important areas of the nest, as delimited by domain experts, are colored manually: The brood (brown), the rubbish pile (violet) and nest entrance (green).

No. of colonies	6
No. of ants	956
No of days	11
Observations frequency	2 per second
No of time steps	1,900,800
No of observations	1,817,164,800

Table 5.1 – Statistics of the *Camponotus fellah* dataset.

5.2 Model

Uncovering the collective behaviors that govern the colony of ants is challenging yet necessary to understand and describe the colony behavior. Hence, our model should enable us to uncover latent collective behaviors and describe the behavior of each individual ant as a function of these collective behaviors. But, first we need to define the notion of behavior. In this work, we describe the behavior of an ant as a function of the following variables.

Location represents a location within the nest or foraging area, and is indexed by an integer $x \in \mathcal{X}$. We further distinguish the locations that are within the nest $x \in \mathcal{X}_n$ from those that are within the foraging area $x \in \mathcal{X}_f$.

Activity is a binary variable $a \in \{0, 1\}$ that indicates whether an ant is moving or not. An ant is declared to be active if it moves more than the camera precision per frame for a significant number of time frames.

We describe the state of ant i at time t using two time-dependent stochastic processes $\{(X_i(t), A_i(t)), t \in \mathcal{T}\}$. The process $X_i(t) \in \mathcal{X}$ indicates the location of ant i at time t, and the process $A_i(t) \in \{0, 1\}$ indicates whether this ant is active at that time. As we are interested in describing macroscopically the geographical distribution of ants and their activity, we assume that

$$\begin{aligned} p(\boldsymbol{X}_i, \boldsymbol{A}_i) &= p(X_i(t_1), A_i(t_1), \dots X_i(t_n), A_i(t_n)) \\ &= \prod_{t \in \mathcal{T}} p(X_i(t), A_i(t)), \end{aligned} \tag{5.1}$$

where the joint distribution $p(X_i(t), A_i(t))$ describes probabilistically the state of the ant i at time t. As stated above, we assume that the individual behavior of an ant can be expressed as a random combination of K collective behaviors that are learnt from the behaviors observed in the whole colony. This translates to the fact that we can express the joint distribution $p(X_i(t), A_i(t))$ with respect to a latent variable $z \in \{1, \dots K\}$ that indicates for each state (x, a) the behavior it was sampled from. More formally, we can

write

$$p(X_i(t) = x, A_i(t) = a) = \sum_{z=1}^{K} p(x, a|z)p(z|i, t),$$
$$= \sum_{z=1}^{K} \theta_z(x, a)\, p(z|i, t). \tag{5.2}$$

The K collective behaviors represent the collective dimensions of our model and are described by the shared multinomial distributions $\theta_k(x, a)$. The individual dimension is captured by the mixture coefficients $p(z|i, t)$ that quantify to which extent the behavior of ant i at time t samples from the collective behavior k. Given that we are interested in analyzing the individual ontogeny (mid-long term development) of ants while removing the variations that are due to the circadian rhythm (short term) of ants [38], we choose 24 hours as a temporal unit. Therefore, we assume that the behavior of an ant is stationary over a day i.e.,

$$p(X_i(t) = x, A_i(t) = a) = \sum_{z=1}^{K} \theta_z(x, a)\, p(z|i, t)$$
$$= \sum_{z=1}^{K} \theta_z(x, a)\, \pi_i(d(t), z), \tag{5.3}$$

where $d(t)$ is the day that corresponds to time t. Hence, the generative process, for ant i at time t, is as follows:

1. Adopting randomly the behavior z according to the mixture coefficient $\pi_i(d(t), z)$
2. Selecting randomly a state (x, a) from the the joint distribution distribution $p(x, a|z)$

We show in Figure 5.2 the corresponding graphical model: The collective dimension is captured by the multinomial distributions θ_k that are shared by all individuals of the colony whereas the individual dimension is captured by the mixture coefficients Π_i that are specific to individual ants.

Training The parameters of our model are the multinomial distributions θ_k associated with each collective behavior and the mixture coefficients $\Pi_{ik}(t)$. In order to learn the model parameters that maximize the likelihood of data, we use the Expectation-Maximization (EM) algorithm for which we derive closed-form expressions of the E and M updates.

Finding the number of canonical behaviors In order to find the number of collective behaviors that explains the data the best, we repeat the following process 100

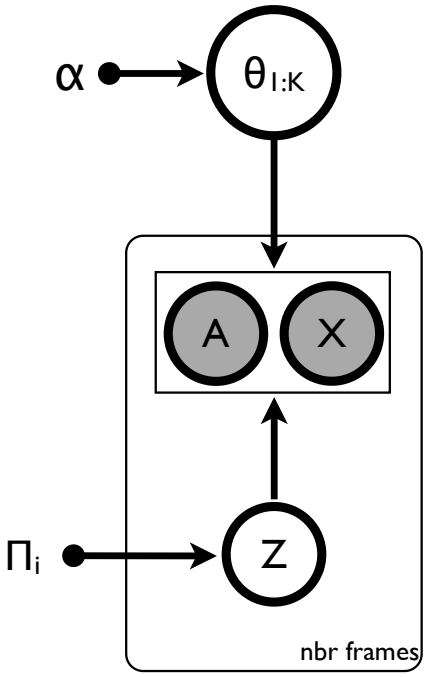


Figure 5.2 – Graphical model representing the behavior of ant i at a given day.

times for each colony c: We divide randomly the dataset of colony c in a training (95%) and a test set (5%), and train our model for different values of the number of collective behaviors $K \in \{1 \dots 7\}$. We then compute the log-likelihood of the test set given the parameters learnt from the training set. We obtain the results shown in Figure 5.3 by averaging the log-likelihoods of the 600 (6 colony and 100 random splits) test sets obtained by randomly splitting the dataset of each colony. We observe that the number of behaviors that maximizes the average likelihood of the test sets is $K = 3$. This implies that having 3 collective behaviors in the colony is the configuration that generalizes the best to unseen data.

Describing the collective behaviors In order to be able to map the behaviors uncovered in different colonies between each other, we assign labels to these behaviors according to the probability of being in the nest.

- **Nest (N)** the behavior z that maximizes the probability of being in the nest $p(x \in \mathcal{X}_n|z)$.
- **Foraging (F)** the behavior z that maximizes the probability of foraging $p(x \in \mathcal{X}_f|z)$.
- **Intermediate (I)** the remaining behavior.

Definition	Domain	Explanation
t	$\mathcal{T} \subset \mathbb{R}_+$	Time step
c	$\mathcal{C} = \{4, 21, 78, 58, 29, 18\}$	Colony id
i	$\{\mathcal{I}_c, c \in \mathcal{C}\}$	Ant id
$\mathcal{X}$	$\mathbb{N}$	Set of locations
$\mathcal{X}_f$	$\mathcal{X}_f \subset \mathcal{X}$	Set of locations in the foraging area
$\mathcal{X}_n$	$\mathcal{X}_n \subset \mathcal{X}$	Set of locations in the nest area
$d(t)$	$\mathbb{N}$	Day associated with time t
k	$\{1, \dots K\}$	Collective behavior id
K	$\mathbb{N}$	Number of collective behaviors
$X_i(t)$	$\mathcal{X}$	Location of ant i at time t
$A_i(t)$	$\{0, 1\}$	Indicates whether ant i is active at time t
$\pi_i(d, z)$	$K-1$ simplex	Mixture coefficients of behavior z at day d for ant i
$\boldsymbol{\pi}_i(d)$	$-$	Vector $[\pi_i(d, 1), \dots, \pi_i(d, K)]$ for ant i
$\beta_i(t)$	$\{1, \dots K\}$	Dominant behavior of ant i at time t
$\theta_k(x, a)$	$\mathcal{X} \times \{0, 1\}$	multinomial distribution representing behavior k

Table 5.2 – List of the definition and domain of the variables relative to an ant, as well as those describing its state at time t.

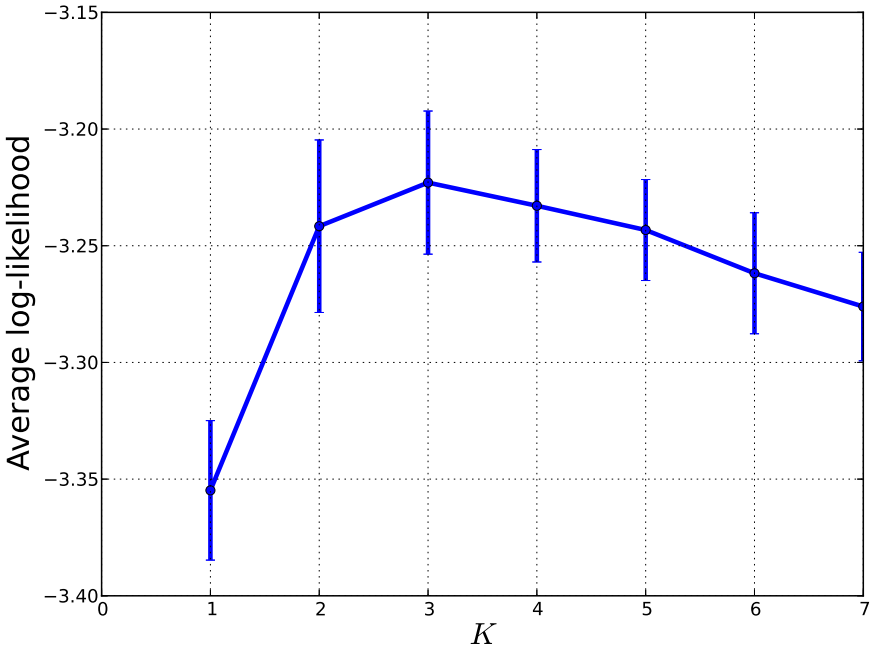


Figure 5.3 – Average log-likelihood vs. the number of collective behaviors K. These values are obtained by averaging the log-likelihoods of the 600 test sets obtained by randomly splitting the dataset of each colony

Mobility In order to visualize the location distributions associated with the behaviors uncovered by our model, we plot in Figure 5.5 and Figure 5.7 the top 1000 locations of each behavior z ranked according to the probability $p(x|z)$. These locations represent, for each behavior, the area where an ant that adopts this behavior would spend most of its time. Moreover, we show in Figure 5.8 the locations of each colony colored according to the mixture coefficients $p(z|x)$. We focus on Figure 5.5 and note that the top locations associated with the behavior N capture well the shape of the brood pile, as shown in Figure 5.5c. Moreover, the most likely locations associated with behavior F captures the entrance of the foraging area and food source (Figure 5.5d) where foragers spend most of their time. The same observation is valid when we show, in Figure 5.7, the 1000 top locations associated with all other colonies. In order to confirm these observations, we

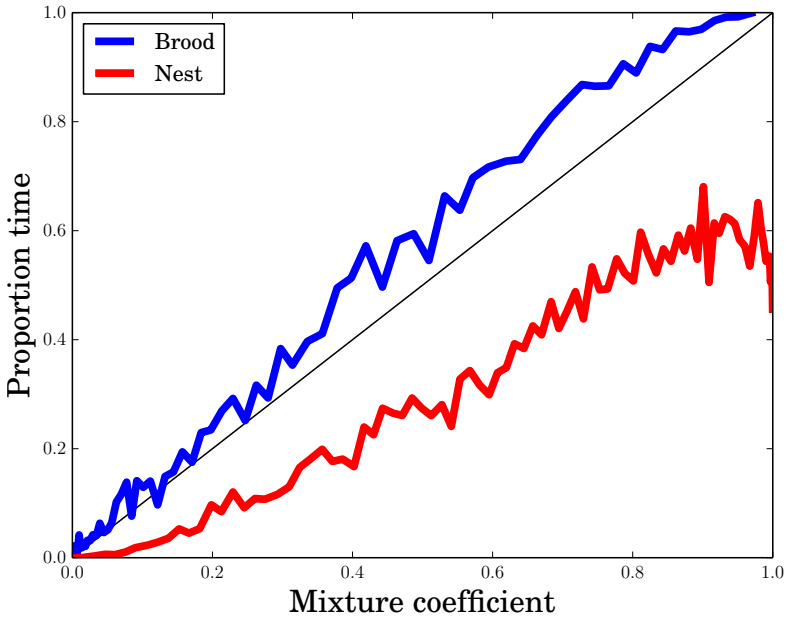


Figure 5.4 – The blue (dark) curve represents the average proportion of time in the brood pile as a function of the mixture coefficient $\pi_i(d, z = N)$ for behavior N, and the red (light) curve represents the average proportion of time in the nest as a function of the mixture coefficient $\pi_i(d, z = I)$ for behavior I.

compare the mixture coefficients $\pi_i(d, z)$ for behaviors N and I to the proportion of time an ant would spend in the correponding area as delimited by domain experts. In fact, Mersch et al. [54] annotated manually the most important region in the nest, namely the brood pile (brown region in Figure 5.5c), and then measured the time each ant spends inside and outside this region. The purpose of these measurements is to associate ants with tasks in the colony, for example, nurses around the brood pile. In Figure 5.4, we plot these quantities as a function of the mixture coefficients $\pi_i(d, z)$ associated with behaviors N and I: The proportion of time an ant spends in the brood pile increases with the mixture coefficient associated with behavior N while the proportion of time spent

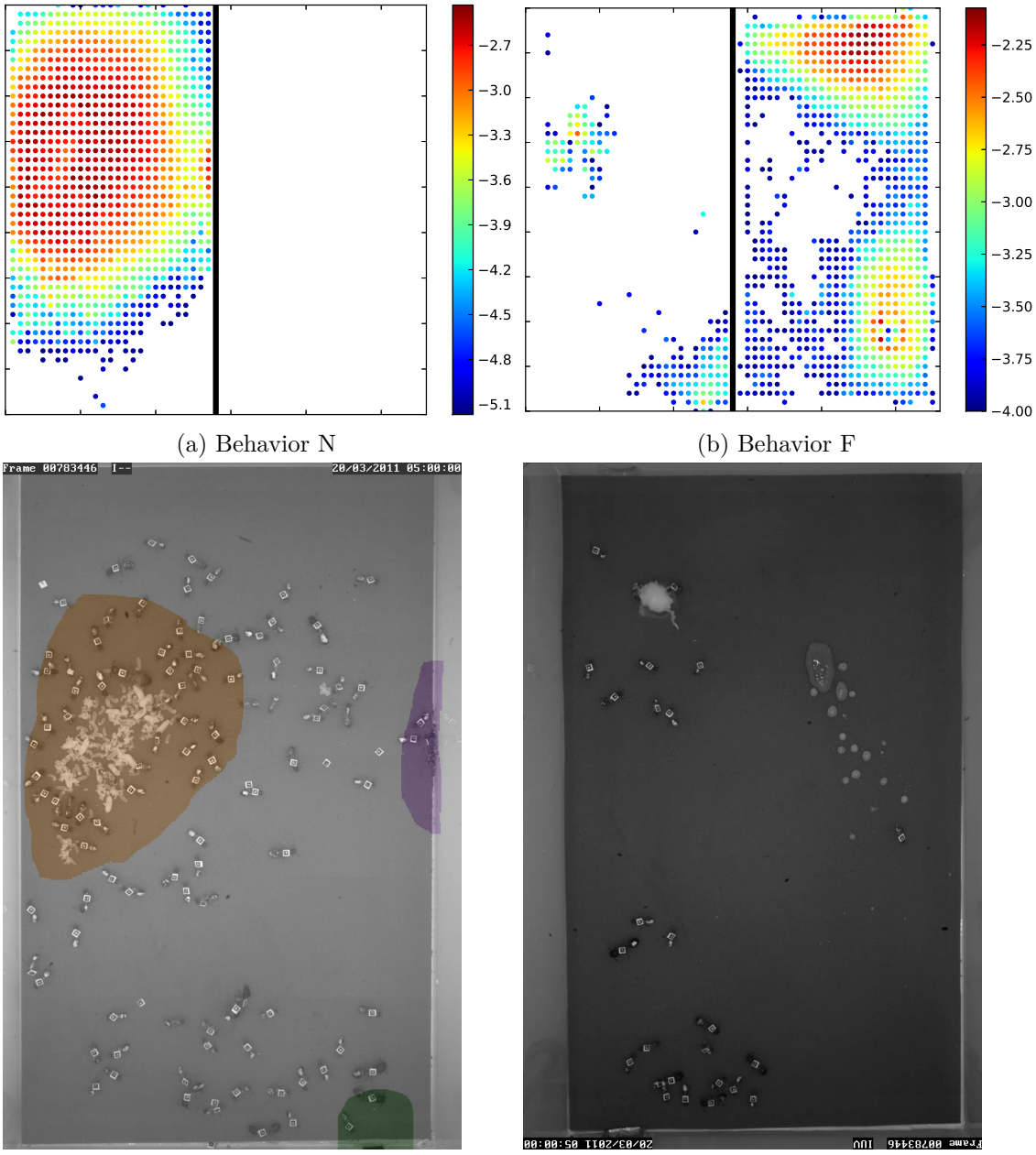
	Behavior N	Behavior I	Behavior F
Colony 4	0.26	0.36	0.51
Colony 18	0.27	0.28	0.4
Colony 21	0.28	0.33	0.42
Colony 29	0.29	0.4	0.4
Colony 58	0.22	0.39	0.52
Colony 78	0.26	0.28	0.46
Average	**0.26 (0.02)**	**0.34 (0.04)**	**0.45 (0.05)**

Table 5.3 – The probability of being active $p(a = 1|z)$ for each colony and each behavior.

in the nest —but outside the brood pile —increases with mixture coefficient associated with behavior I. More importantly, the fact that the average proportion of time in the brood pile is very close to the mixture coefficient of behavior N confirms that the area associated with behavior N matches accurately the area of the brood pile, as delimited manually by domain experts. Our approach would therefore enable biologists to detect the tasks performed by the ants in a colony without going through the tedious process of manual annotation.

Activity In Table 5.3, we show the probability of being active for each behavior and each colony. We notice that (a) inactivity prevails in the colony, and (b) the probability of being active increases as we move outside the nest. In fact, independently of the behavior adopted, an ant spends most of its time inactive. In fact, for all behaviors, the probability of being inactive is always higher than the probability of being active: the probability of being active is $p(a = 1|z) = 0.35$ on average. However, this probability increases as we ants move away from the brood pile. For all colonies, the ants that adopt behavior N are less likely to be active than ants that adopt the intermediate behavior. The foragers are clearly the most active individuals, as their probability of being active is, on average, $p(a = 1|z) = 0.45$.

Ant behavior We show in Figure 5.6 a scatter plot of the mixture coefficients $\pi_i(d, z)$ for each behavior and each colony. Most of these coefficients are concentrated on the edges of the triangle that represents the 3 dimensional simplex. This suggests that the majority of ants adopt a mixture of two behaviors. For example, the mixture coefficient that is close to the edge from N to I corresponds to an ant whose behavior is a combination of behavior N and I but exhibits no signs of behavior F.



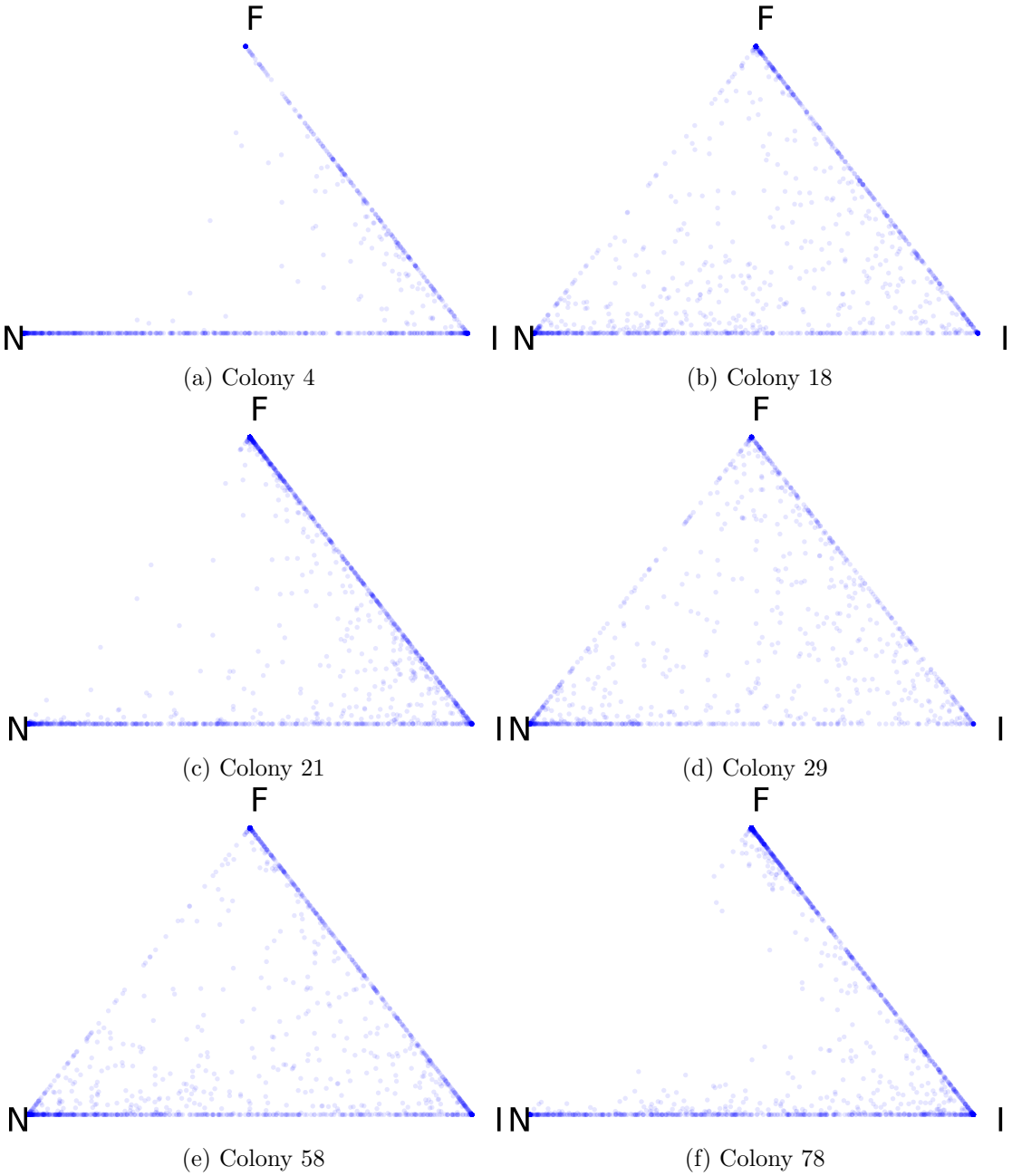
(a) Behavior N

(b) Behavior F

(c) Nest area

(d) Foraging area

Figure 5.5 – We plot the top 1000 locations of each behavior z ranked according to the probability $p(x|z)$. Behavior N is concentrated around the brood pile, whereas behavior F is concentrated around the water source and food.



(a) Colony 4

(b) Colony 18

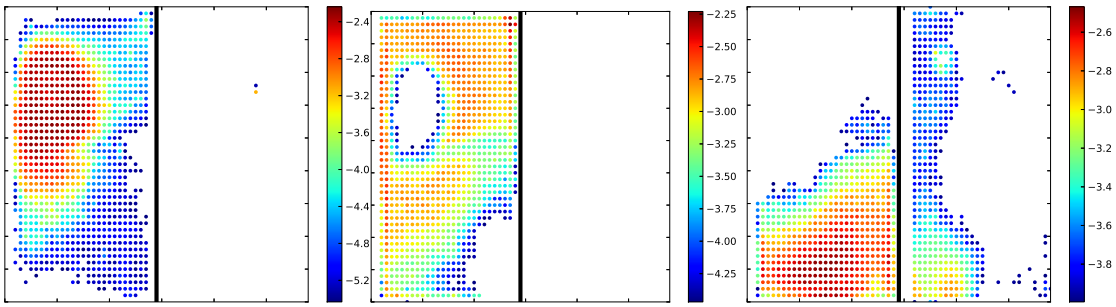
(c) Colony 21

(d) Colony 29

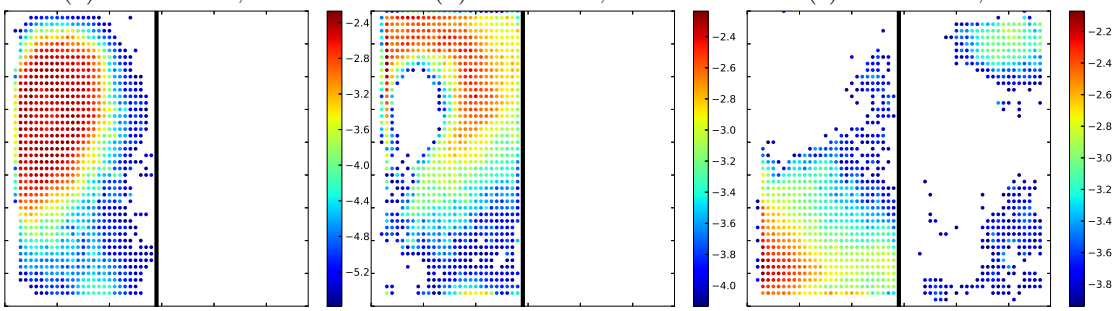
(e) Colony 58

(f) Colony 78

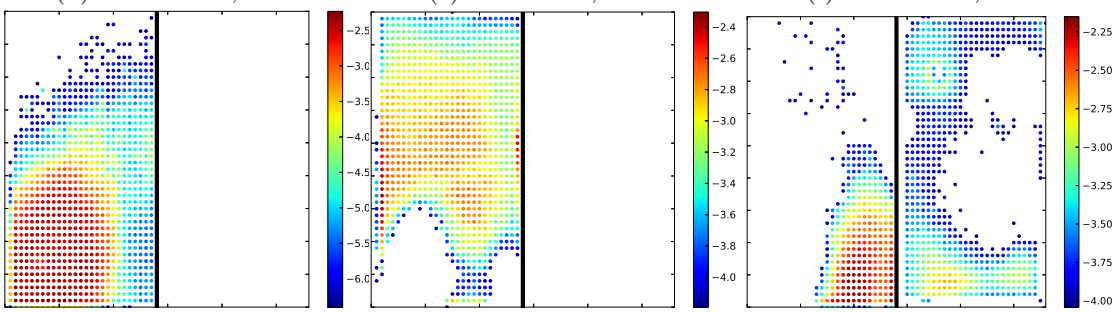
Figure 5.6 – Each dot plotted in the simplex represents a mixture of behaviors (coefficient $\pi_i(d, z)$). Note that these coefficients are concentrated on the edges of the simplex, which implies that one distribution dominates the others.



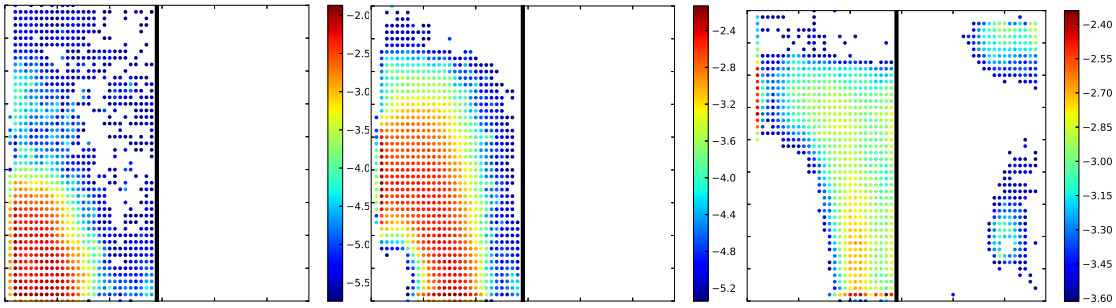
(a) Behavior N, col 4 (b) Behavior I, col 4 (c) Behavior F, col 4



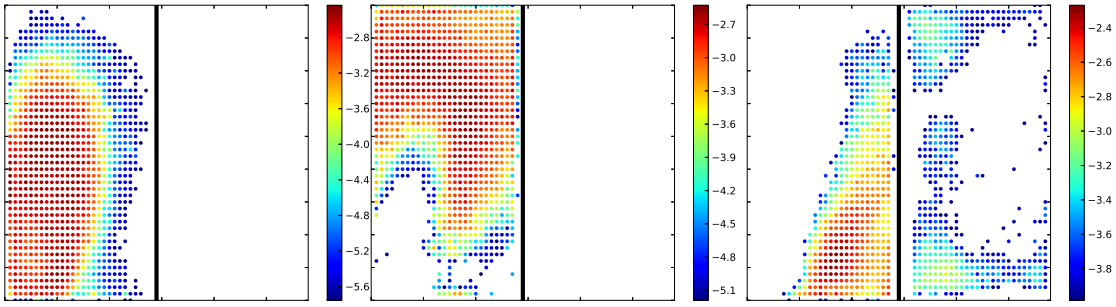
(d) Behavior N, col 29 (e) Behavior I, col 29 (f) Behavior F, col29



(g) Behavior N, col 58 (h) Behavior I, col 58 (i) Behavior F, col 58

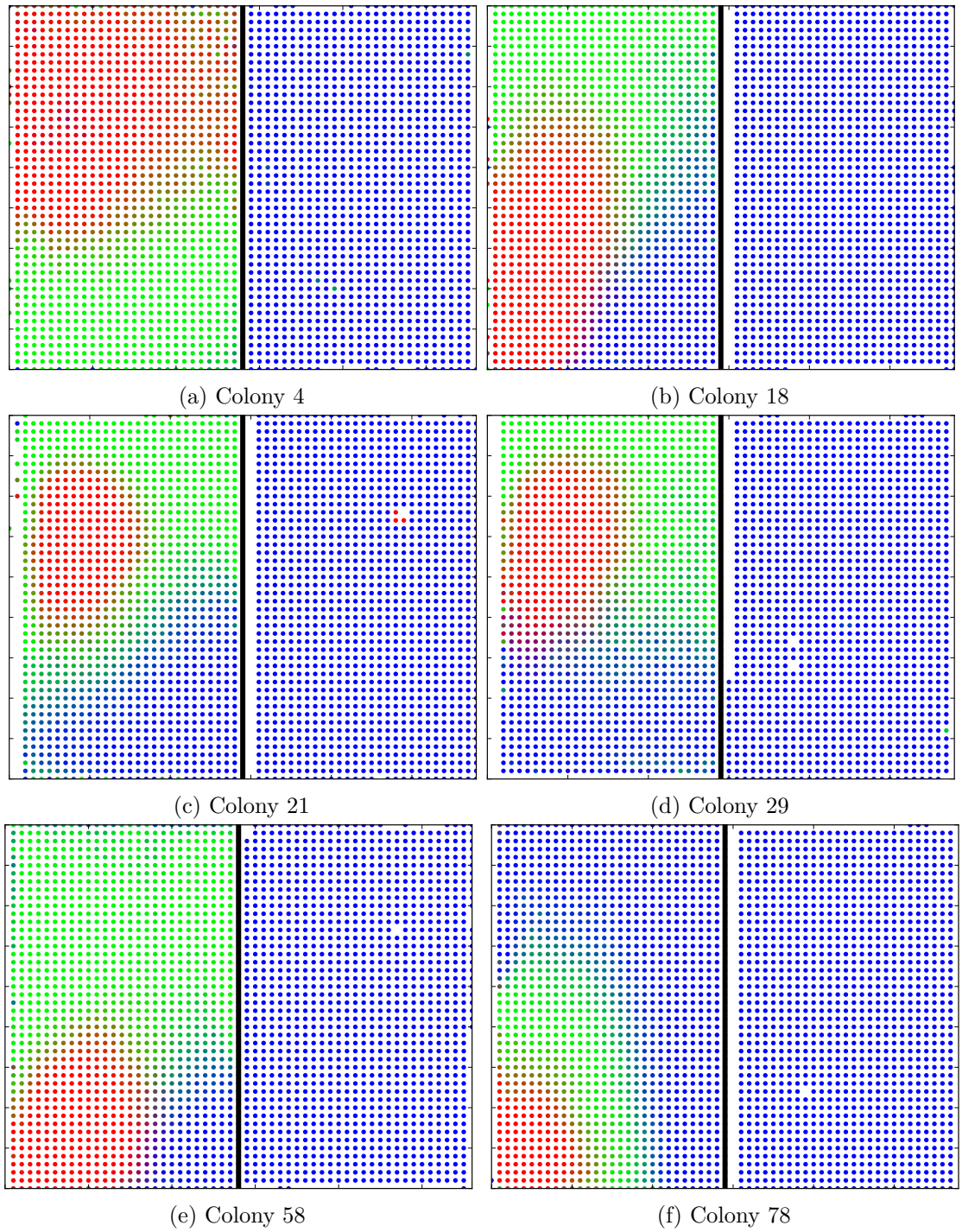


(j) Behavior N, col 78 (k) Behavior I, col 78 (l) Behavior F, col 78



(m) Behavior N, col 18 (n) Behavior I, col 18 (o) Behavior F, col 18

Figure 5.7 – For each colony and behavior z, we plot the top 1000 locations x ranked according to the probability $p(x|z)$.



(a) Colony 4 (b) Colony 18

(c) Colony 21 (d) Colony 29

(e) Colony 58 (f) Colony 78

Figure 5.8 – The color of each location x is a mixture of colors that is a function of the posterior probability $p(z|x)p(z)$. The colors red, green and blue are mapped to the behaviors N, I and F respectively.

5.3 Conclusion

In this chapter, we study the ideal scenario for which the mobility model expresses both the individual and collective dimensions. We model the behaviors of *Camponotus fellah* ants by analyzing large-scale digital traces that describe their mobility. Our model take advantage of the correlation between the behaviors of individual ants to uncover the fundamental collective behaviors in a colony, and to express the behavior of each ant as a time-dependent combination of these behaviors. More importantly, the collective behaviors found by our model correspond to actual functional behaviors in ant colonies: the spatial distributions associated with them match well the spatial distributions as defined by domain experts. This example illustrates well the predictive power of mobility models that take advantage of both the individual and the collective dimensions of mobility.

Part II

The Information Theoretic Perspective

Introduction

In the first part of this thesis, we studied mobility from the modeling perspective: We learnt sophisticated models that are tailored to the data and scenario of interest; then we use them in order to predict individual behaviors. This also enabled us to quantify mobility predictability by measuring the proportion of accurate predictions made by our model on unseen data.

In this chapter, we study mobility from a different yet complementary perspective: We take an information theoretic approach to rigorously quantify mobility uncertainty and its evolution with additional information. The mobility model we consider has the advantage, as opposed to the tailored model learnt in the first part of this thesis, of being general enough to be representative of most scenarios. This parsimonious model represents mobility in its simplest form —sequence of decision points —but still captures mobility patterns. We discretize the user's world to obtain a map that we represent as a mobility graph G: A vertex represents a branch point where the user takes the decision about where to move next, and an edge represents a direct physical path between two vertices. A vertex typically corresponds to a semantic place such as home or work place. The advantage of this representation, over full geographical trajectories, is that it encodes the space of possible user decisions, while abstracting away any finer but irrelevant details of the mobility process.

In this model, the mobility of a user is simply a sequence of vertices (trajectory) generated by a random walk on the mobility graph G. The randomness of the user mobility is captured by the distribution over all possible trajectories. As our goal is to quantify mobility uncertainty , we need a measure that enables us to quantify the randomness of the trajectory taken by a user; we choose to compute the entropy of the trajectory distribution. To do so, we use the result of Ekroot and Cover [26] in order to compute the entropy of Markov trajectories with fixed initial and final states. For this model, a location update amounts to conditioning on a particular state of the Markov chain. Hence, we need to compute the entropy of Markov trajectories *conditional* on a set of intermediate states.

In this chapter, we introduce one of the main contributions of this thesis: We propose a method for computing the entropy of conditional Markov trajectories through a

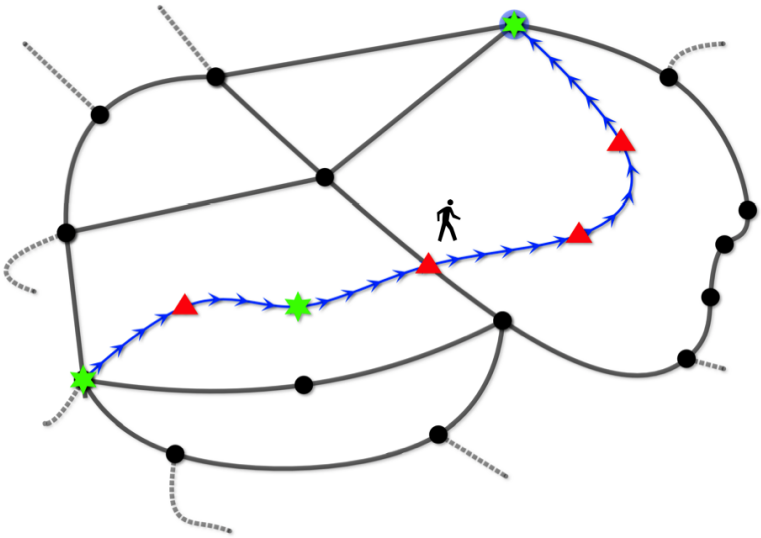


Figure 5.9 – We discretize the user's world and assume that she moves on a mobility graph: A vertex of this graph represents a decision point where the user can choose their next move, and an edge represents a direct physical path between two vertices. We model the user mobility as a random trajectory on the mobility graph. Our goal is to quantify the impact of the locations revealed by the user (green stars) on the uncertainty about his trajectory and hence the undisclosed locations (red triangles).

transformation of the original Markov chain into a Markov chain that exhibits the desired conditional distribution of trajectories. Computing the entropy of conditional Markov trajectories enables us to quantify the change of entropy for each location update, given the model and the previously revealed locations. We apply this result in order to quantify the evolution of the uncertainty about mobility of a user as intermediate locations are revealed. We also find an empirical relation between the change in trajectory entropy brought by the disclosure of a particular intermediate location and the nature —is it an intermediate destination? —of this location. We build on this finding and design an algorithm that uncovers intermediate destinations along a trajectory.

6 The Entropy of Conditional Markov Trajectories

6.1 Introduction

Quantifying the randomness of trajectories over Markov chains has applications in graph theory [73], statistical physics [51] and the study of random walks on graphs [17, 61]. The need to quantify the randomness of Markov trajectories first arose when Lloyd and Pagels [51] defined a measure of complexity for the macroscopic states of physical systems. They examine some intuitive properties that a measure of complexity should have and propose a universal measure called *depth*. They suggest that the depth of a state should depend on the complexity of the process by which that state arose, and they prove that it must be proportional to the Shannon entropy of the set of trajectories leading to that state. Subsequently, Ekroot and Cover [26] studied the computational aspect of the depth measure. In order to quantify the number of bits of randomness in a Markov trajectory, they propose a closed-form expression for the entropy of trajectories of an irreducible finite state Markov chain. Their expression does not allow, however, for computing the entropy of Markov trajectories conditional on the realisation of a set of intermediate states. Computing the conditional entropy of Markov trajectories turns out to be very challenging yet useful in numerous domains, including the study of mobility uncertainty and its dependence on location side information.

Consider a scenario where we are interested in quantifying the uncertainty about user mobility. We discretize the world and assume that the user moves on a mobility graph G for which a trajectory is a realization of a random walk from a starting vertex s to a destination vertex d. The randomness of the trajectory a user would follow is represented by the distribution of trajectories whose randomness is captured by the entropy of Markov trajectories between the source and destination vertices. Now, if we obtain side information stating that the user went (or has to go) through a set of intermediate vertices, quantifying the evolution of the uncertainty about her mobility requires the computation of the trajectory entropy conditional on the set of known intermediate states. For example, if the entropy conditional on the set of known intermediate states is

zero, then this set reveals the whole trajectory of the user.

In this chapter, we introduce the notion of Markov trajectories and propose a method for computing the entropy of Markov trajectories conditional on a set of intermediate states. The method is based on a transformation of the original Markov chain so that the transformed Markov chain exhibits an (unconditional) distribution of trajectories equal the desired conditional distribution of trajectories in the original Markov chain. We also derive an expression that enables us to compute the entropy of Markov trajectories, under conditions weaker than those assumed in [26]. Moreover, this expression links the entropy of Markov trajectories to the local entropies at the Markov chain states.

6.2 Model

Let $\{X_i\}$ be a finite state irreducible and aperiodic Markov chain (MC) with transition probability matrix P whose elements are the transition probabilities

$$\begin{aligned} P_{x_n x_{n+1}} &= p(X_{n+1} = x_{n+1} | X_n = x_n) \\ &= p(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1). \end{aligned}$$

This MC admits a stationary distribution Π, which is the unique solution of

$$\Pi = \Pi P.$$

The entropy rate $H(X)$ is a measure of the average entropy growth of a sequence generated by the process $\{X_i\}$ and is defined as

$$H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, ..., X_n).$$

For the particular case of an irreducible and aperiodic MC, the limit above is equal to [21, p. 77]

$$H(X) = \sum_i \Pi(i) H(P_{i\cdot}),$$

where $P_{i\cdot}$ denotes the i^{th} row of P and where $H(P_{i\cdot}) = -\sum_j P_{ij} \log(P_{ij})$ is the *local entropy* of state i. Note that, throughout this chapter, we use MC_P as shorthand for the Markov chain whose transition probability matrix is P.

The entropy of markov trajectories We follow the setting of [26] closely. We define a *random trajectory* T_{sd} of a MC as a path with initial state s, final state d, and no intermediate state d, in other words, the trajectory is terminated as soon as it reaches state d. Using the Markov property, we express the probability of a particular trajectory

$t_{sd} = sx_2...x_kd$ given that $X_1 = s$ as

$$p(t_{sd}) = P_{sx_2} P_{x_2x_3} \dots P_{x_kd}. \tag{6.1}$$

Let $\mathcal{T}_{sd}$ be the set of all trajectories that start at state s and end as soon as they reach state d. As the MC defined by the matrix P is finite and irreducible, we have

$$\sum_{t_{sd} \in \mathcal{T}_{sd}} p(t_{sd}) = 1 \qquad \text{for all } s, d.$$

So the discrete random variable T_{sd} has as support the set $\mathcal{T}_{sd}$, with the probability mass function $p(t_{sd})$. Subsequently, we use $p(t_{sd})$ as shorthand for $p(T_{sd} = t_{sd})$. We can now express the entropy of the random trajectory T_{sd} as

$$H_{sd} \equiv H(T_{sd}) = -\sum_{t_{sd} \in \mathcal{T}_{sd}} p(t_{sd}) \log p(t_{sd}).$$

We define the matrix of trajectory entropies H where $H_{ij} = H(T_{ij})$. Ekroot and Cover [26] provide a general closed-form expression for the matrix H of an irreducible, aperiodic and finite state MC. They take advantage of the convergence properties of the matrix of transition probability P in order to compute the matrix of trajectory entropies without expressing explicitly the distribution of trajectories.

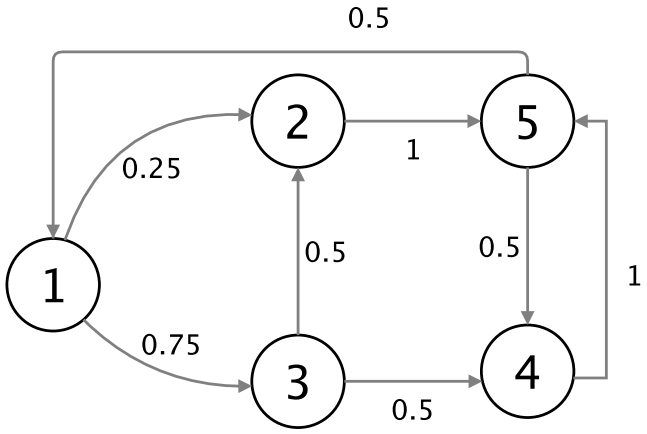


Figure 6.1 – An irreducible, 5-state, Markov chain annotated with the transition probabilities.

Example In order to illustrate the concept of trajectory entropy, we compute the matrix of trajectory entropies H for the finite-state irreducible and aperiodic MC shown

in Figure 6.1. The transition matrix associated with this MC is

$$P = \begin{pmatrix} 0 & 0.25 & 0.75 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 & 0 \end{pmatrix},$$

and its stationary distribution is

$$\Pi = \begin{pmatrix} 0.17 \\ 0.11 \\ 0.13 \\ 0.24 \\ 0.35 \end{pmatrix}.$$

Given the vector of local entropies $(0.81, 0, 1, 0, 1)$, we compute the entropy rate using (6.2) and find that $H(X) = 0.61$ bits per transition. Note that the presence of cycles implies that the set of trajectories between some pair of states might have infinite cardinality ($|\mathcal{T}_{14}| = \infty$, for example). Therefore, in addition to being complex, the naive approach of enumerating all trajectories is not always possible. Now we compute the closed-form expression [26] to obtain the associated trajectory entropies

$$H = \begin{pmatrix} 3.56 & 3.69 & 1.74 & 3.18 & 1.56 \\ 2 & 5.69 & 3.74 & 2.59 & 0 \\ 3 & 3.84 & 4.74 & 2.29 & 1 \\ 2 & 5.69 & 3.74 & 2.59 & 0 \\ 2 & 5.69 & 3.74 & 2.59 & 1.78 \end{pmatrix}.$$

The zero elements of the matrix H correspond to deterministic trajectories such as T_{25}, which is equal to the path 25 with probability 1 because no other path allows a walk to go from 2 to 5. The entropy of the random Trajectory T_{35} is 1 bit because this trajectory takes two values —$\{3, 2, 5\}$ or $\{3, 4, 5\}$ —with equal probability. Despite the fact that the set $\mathcal{T}_{14}$ has an infinite number of members, we are able to compute the entropy of the random trajectory $T_{14} = 3.18$ bits without having to explicitly express its distribution.

Hitting time and trajectory entropy in d-regular graphs We show here an interesting link between trajectory entropy and hitting times for a random walk on a regular graph. We consider a random walk on δ-regular graph $G(V, E)$: given that we are at a given vertex v, we move to a neighborhood of v with probability $1/\delta$. The sequence of vertices visited by the random walk is a Markov chain whose matrix of transition

probability is

$$P_{ij} = \begin{cases} 1/\delta & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

We define the access time or hitting time S_{ij} as the expected number of steps before the vertex j is visited, given that the walk starts at vertex i. For all $i, j \in V$, we prove the following equality

Proposition 1.

$$H_{ij} = \log(\delta)\, S_{ij}.$$

Proof. By definition, we have

$$H_{sd} = -\mathbb{E}\left[\log p(T_{sd})\right].$$

Using the Markov property, we express the probability of a particular trajectory $t_{sd} = sx_2...x_kd$ as

$$p(t_{sd}) = P_{sx_2} P_{x_2x_3} \dots P_{x_kd}.$$

Since the trajectory is generated by a random walk on a δ-regular graph, we have

$$p(t_{sd}) = \frac{1}{\delta^{l(t_{sd})}},$$

where $l(t_{sd})$ be the length of the trajectory t_{sd}. Thus

$$H_{sd} = -\mathbb{E}\left[\log\left(\frac{1}{\delta^{l(T_{sd})}}\right)\right] = -\mathbb{E}\left[l(T_{sd}) \log \frac{1}{\delta}\right] = \log(\delta)\, \mathbb{E}\left[l(T_{sd})\right].$$

The expected length $\mathbb{E}\left[l(T_{sd})\right]$ of the random trajectory T_{sd} is equal to the hitting time S_{sd} because the trajectory T_{sd} is a path from s to d that terminates as soon as it reaches vertex d. Therefore

$$H_{sd} = \log(\delta)\, S_{sd}.$$

□

In other words, as all vertices have the same degree, the entropy of a trajectory generated by a random walk on a regular graph is proportional to its expected length. We show in Figure 6.2 a grid for which we consider a random walk that starts at the center of the grid $s = (6, 6)$. Having a closed form expression for the hitting times would enable us to use the result of Proposition 1 in order to compute the trajectory entropy. However, despite the symmetry that characterizes this random walk, finding a closed-form expression of

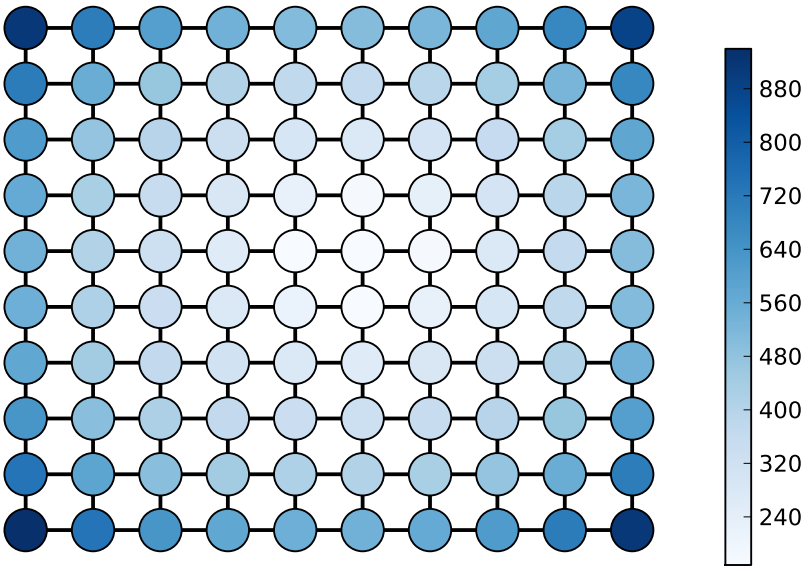


Figure 6.2 – We consider a random walk on 10×10 grid and color each vertex d according to the trajectory entropy H_{sd} (bits), where s is the central vertex $(6, 6)$.

the hitting times for this random walk is hard. Consequently, we compute directly the matrix of trajectory entropy H and color each vertex v according to the value of the entropy H_{sv}. We see in Figure 6.2 that the trajectory entropy H_{sv} increases as we move away from the starting vertex s, which is not surprising given that the entropy H_{sv} is proportional to the expected length of the trajectory T_{sv}.

In the next section, we study the entropy of Markov trajectories conditional on *multiple* intermediate states and derive a general expression for this entropy.

6.3 The Entropy of Conditional Markov Trajectories

We are interested in computing the entropy of the random trajectory T_{sd} given additional information about intermediate states. We denote by $H_{sd|u}$ the entropy of the trajectory

from s to d, given that it goes through u. The definition of this entropy is

$$H_{sd|u} \equiv H(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^{u}) = -\sum_{t_{sd} \in \mathcal{T}_{sd}^{u}} p(t_{sd}|T_{sd} \in \mathcal{T}_{sd}^{u}) \log p(t_{sd}|T_{sd} \in \mathcal{T}_{sd}^{u}), \tag{6.2}$$

where $\mathcal{T}_{sd}^{u}$ is the set of all trajectories in $\mathcal{T}_{sd}$ with an intermediate state u

$$\mathcal{T}_{sd}^{u} = \{t_{sd} \in \mathcal{T}_{sd} : t_{sd} = s \dots u \dots d\}.$$

The major challenge is to compute efficiently the entropy $H_{sd|u}$. Even the costly approach of computing all the terms of the sum (6.2) is not always possible because the set $\mathcal{T}_{sd}^{u}$ has an infinite number of members in the case where, after removing state d, the transition graph of the MC is not a DAG. It is important to emphasize that the entropy $H_{sd|u}$ is not the entropy of the random variable T_{sd} given another random variable—a quantity which is easy to compute—but rather the entropy of T_{sd} conditional on the realization of a dependent random variable. We return to the example of the MC shown in Figure 6.1: the entropy of the random trajectory T_{15} is 1.56 bits. Now imagine that we have an additional piece of information stating that the trajectory T_{15} goes through state 4. Intuitively, we would be tempted to argue that the entropy $H_{15|4}$ of the trajectory T_{15} conditional on going through state 4 is equal to $H_{14} + H_{45}$, but this additivity property does not hold. Indeed, the conditional entropy $H_{15|4}$ is zero because the trajectory T_{15}, conditional on the intermediate state 4, can only be equal to the path 1345, whereas $H_{14} = 3.18$ bits, hence $H_{14} + H_{45} = 3.18 + 0 = 3.18 \neq H_{15|4}$ bits.

Let α_{sud} denote the probability that the random trajectory T_{sd} goes through the state u at least once:

$$\alpha_{sud} = p(T_{sd} \in \mathcal{T}_{sd}^{u}).$$

This is also equal to the probability that a walk reaches the state u before the state d, given that it started at s. In order to compute α_{sud}, the technique from [44,67] is to make the states u and d absorbing (a state i is absorbing if and only if $P_{ii} = 1$) and compute the probability of being absorbed by state u given that the trajectory has started at state s. In Figure 6.3, we show an example of a biased random walk and the associated probabilities α_{sud}.

Our first step towards computing $H_{sd|u}$ is to express it as a function of quantities that are much simpler to compute. The idea is to relate the entropy of a trajectory conditional on a given state to its entropy that is conditional on *not* going through that state. Therefore, we define the entropy $H_{sd|\bar{u}}$ of a trajectory from s to d given that it does *not* go through u to be

$$H_{sd|\bar{u}} \equiv H(T_{sd}|T_{sd} \notin \mathcal{T}_{sd}^{u})$$

Using the chain rule for entropy, we can derive the following equation that relates $H_{sd|u}$

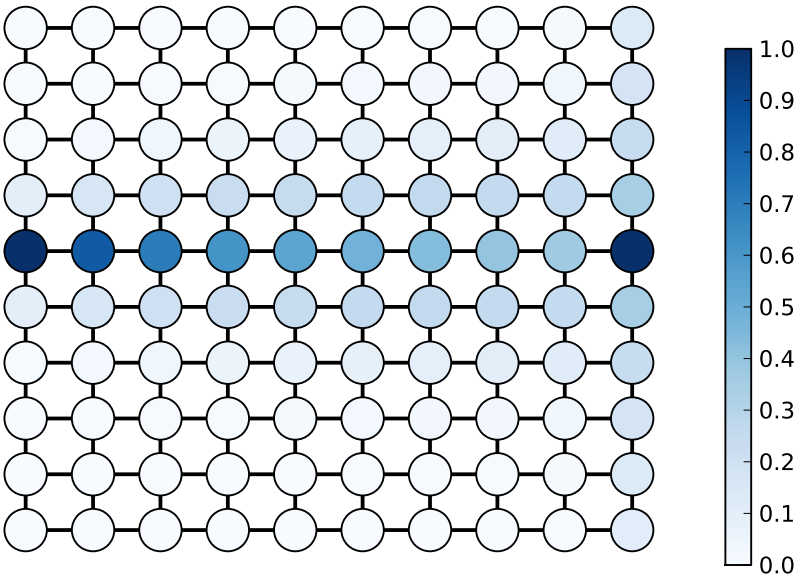


Figure 6.3 – We consider a biased random walk on 10×10 grid for which the probability of moving to right, up and down are 0.6, 0.2, and 0.2, respectively. We focus on the random trajectory T_{sd} where s is the vertex $(1, 6)$ and d the vertex $(10, 6)$, and we color each vertex u according to the probability α_{sud} following the color map displayed on the right. Naturally, we have that the probability $\alpha_{sdd} = \alpha_{ssd} = 1$, because we condition on starting at vertex s and ending at vertex d. Moreover, the vertices that are in the direct path from s to d are the most likely.

to $H_{sd}, H_{sd|\bar{u}}$ and α_{sud}:

$$H_{sd} = \alpha_{sud} H_{sd|u} + (1 - \alpha_{sud}) H_{sd|\bar{u}} + h(\alpha_{sud}) \tag{6.3}$$

for all u, where $h(\alpha_{sud})$ is the entropy of a Bernoulli random variable with success probability α_{sud}.

Proof. First, we define the indicator variable I by

$$I = \begin{cases} 1 & \text{if } T_{sd} \in \mathcal{T}_{sd}^{u}, \\ 0 & \text{otherwise.} \end{cases}$$

Using the chain rule for entropy, we express the joint entropy $H(T_{sd}, I)$ in two different ways,

$$H(T_{sd}, I) = H(I) + H(T_{sd}|I),$$
$$H(T_{sd}, I) = H(T_{sd}) + H(I|T_{sd}) = H(T_{sd}),$$

where the last equality follows because I is a deterministic function of T_{sd}. So the entropy of the random trajectory T_{sd} can be expressed as

$$\begin{aligned} H(T_{sd}) &= H(I) + H(T_{sd}|I) \\ &= H(I) + H(T_{sd}|I=1)p(I=1) + H(T_{sd}|I=0)p(I=0) \\ &= H(I) + H(T_{sd}|T_{sd} \in \mathcal{T}^u_{sd})p(T_{sd} \in \mathcal{T}^u_{sd}) + H(T_{sd}|T_{sd} \notin \mathcal{T}^u_{sd})p(T_{sd} \notin \mathcal{T}^u_{sd}). \end{aligned}$$

Since $\alpha_{sud} = p(T_{sd} \in \mathcal{T}^u_{sd}) = p(I=1)$, we obtain

$$H(T_{sd}) = \alpha_{sud} H(T_{sd}|T_{sd} \in \mathcal{T}^u_{sd}) + (1-\alpha_{sud}) H(T_{sd}|T_{sd} \notin \mathcal{T}^u_{sd}) + h(\alpha_{sud}).$$

□

As we know how to compute H_{sd} and α_{sud} [26, 44, 67], if we are able to compute $H_{sd|\bar{u}}$, we can use (6.3) to find $H_{sd|u}$. However, generalizing (6.3) to trajectories conditional on passing through *multiple* intermediate states turns out to be difficult. Therefore we propose an approach that circumvents this problem. As we will see, the difficulty of our approach also boils down to computing the entropy of a trajectory conditional on *not* going through a given state.

First, we define $\mathcal{T}^{\boldsymbol{u}}_{sd}$, the set of all trajectories in $\mathcal{T}_{sd}$ that exhibit the sequence of intermediate states $\boldsymbol{u} = u_1 u_2 \dots u_l$, i.e.,

$$\mathcal{T}^{\boldsymbol{u}}_{sd} = \{t_{sd} \in \mathcal{T}_{sd} : t_{sd} = s \dots u_1 \dots u_2 \dots u_l \dots d\}.$$

For an arbitrary sequence of states $\boldsymbol{u} = u_1 u_2 \dots u_l$ satisfying $p(T_{sd} \in \mathcal{T}^{\boldsymbol{u}}_{sd}) > 0$, we prove the following lemma.

Lemma 1.

$$H(T_{sd}|T_{sd} \in \mathcal{T}^{\boldsymbol{u}}_{sd}) = \sum_{k=0}^{l-1} H_{u_k u_{k+1}|\bar{d}} + H_{u_l d}, \tag{6.4}$$

where $u_0 = s$.

Proof. First, given $T_{sd} \in \mathcal{T}^{\boldsymbol{u}}_{sd}$, the random trajectory T_{sd} can be expressed as a sequence of random sub-trajectories $(T_{su_1}, T_{u_1 u_2}, \dots, T_{u_{l-1} u_l}, T_{u_l d})$. Therefore, the conditional entropy $H(T_{sd}|T_{sd} \in \mathcal{T}^{\boldsymbol{u}}_{sd})$, which we denote by $H_{sd|u_1 \dots u_l}$, can be written as a joint sub-trajectory

entropy

$$H_{sd|u_1\dots u_l} = H(T_{su_1}, T_{u_1u_2}, \dots, T_{u_ld}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}).$$

By applying the chain rule for entropy, we obtain successively

$$\begin{aligned} H_{sd|u_1\dots u_l} &= H(T_{su_1}, T_{u_1u_2}, \dots, T_{u_ld}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) \\ &= H(T_{su_1}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) \\ &\quad + H(T_{u_1u_2}|T_{su_1}; T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) \\ &\quad \vdots \\ &\quad + H(T_{u_ld}|T_{su_1}, \dots, T_{u_{l-1}u_l}; T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}). \end{aligned}$$

The Markovian nature of the process generating the trajectory T_{sd} implies that each of the sub-trajectories $T_{u_ku_{k+1}}$ is independent of those preceding it, given its starting point u_k. Since the sequence $s\boldsymbol{u} = su_1u_2\dots u_l$ defines the starting point of each sub-trajectory, we can therefore write that

$$H(T_{u_ku_{k+1}}|T_{su_1}, \dots, T_{u_{k-1}u_k}; T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) = H(T_{u_ku_{k+1}}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}). \tag{6.5}$$

Using (6.5), the expression for the conditional entropy becomes

$$\begin{aligned} H_{sd|u_1\dots u_l} &= H(T_{su_1}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) \\ &\quad + H(T_{u_1u_2}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}) \\ &\quad \vdots \\ &\quad + H(T_{u_ld}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}). \end{aligned}$$

Note that for each trajectory $T_{u_ku_{k+1}}$, the only restriction imposed by the event $\{T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}\}$ is that the final state d cannot be an intermediate state of any of the first l trajectories $T_{su_1}, T_{u_1u_2}, \dots, T_{u_{l-1}u_l}$. As a result,

$$\begin{aligned} H_{sd|u_1\dots u_l} &= H(T_{su_1}|T_{su_1} \notin \mathcal{T}_{su_1}^{d}) \\ &\quad + H(T_{u_1u_2}|T_{u_1u_2} \notin \mathcal{T}_{u_1u_2}^{d}) \\ &\quad \vdots \\ &\quad + H(T_{u_ld}) \\ &= \sum_{k=0}^{l-1} H_{u_ku_{k+1}|\bar{d}} + H_{u_ld}, \end{aligned}$$

where $u_0 = s$. □

Now, if we are able to compute $H_{u_ku_{k+1}|\bar{d}}$, we can use (6.4) to derive $H(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}})$. The following lemma shows how the conditional entropy $H_{s'd'|d}$ can be obtained by a

simple modification of the MC.

We consider a MC whose transition probability matrix is P, and three distinct states s', d' and d such that $\alpha_{s'dd'} = p(T_{s'd'} \in \mathcal{T}^d_{s'd'}) < 1$. Let $\bar{P}$ be the transition matrix of the same MC but where both states d and d' are made absorbing, and whose entries are thus

$$\bar{P}_{ij} = \begin{cases} 0 & \text{if } i = d, d' \text{ and } i \neq j, \\ 1 & \text{if } i = d, d' \text{ and } i = j, \\ P_{ij} & \text{otherwise.} \end{cases} \tag{6.6}$$

Next, we define a second matrix P', obtained by a transformation of the matrix $\bar{P}$

$$P'_{ij} = \begin{cases} \frac{\alpha_{jd'd}}{\alpha_{id'd}} \bar{P}_{ij} & \text{if } \alpha_{id'd} > 0, \\ \bar{P}_{ij} & \text{otherwise.} \end{cases} \tag{6.7}$$

Lemma 2. (*i*) *The matrix P' is stochastic and* (*ii*) *If T'_{sd} is a random trajectory defined on the MC whose transition probability matrix is P' then*

$$H(T_{s'd'} | T_{s'd'} \notin \mathcal{T}^d_{s'd'}) = H(T'_{s'd'}).$$

Proof. (i) The matrix $\bar{P}$ is the transition probability matrix of a MC where the states d and d' are absorbing. We can therefore introduce the vectors of absorption probability $\boldsymbol{a_d} = (a_{1d}, a_{2d}, \ldots, a_{nd})$ and $\boldsymbol{a_{d'}} = (a_{1d'}, a_{2d'}, \ldots, a_{nd'})$ where a_{id} and $a_{id'}$ are, respectively, the probability of being absorbed by d and d', given that the trajectory starts at i. These vectors are eigenvectors of $\bar{P}$ associated with the unit eigenvalue [67, p. 227]

$$\bar{P}\boldsymbol{a_d} = \boldsymbol{a_d} \qquad \bar{P}\boldsymbol{a_{d'}} = \boldsymbol{a_{d'}}. \tag{6.8}$$

Moreover as $MC_{\bar{P}}$ has only two absorbing states d and d', for all i, $a_{id} = 1 - a_{id'}$. Recall that for all i, $\alpha_{id'd} = a_{id'}$ hence (6.7) can be written as

$$P'_{ij} = \begin{cases} \frac{a_{jd'}}{a_{id'}} \bar{P}_{ij} & \text{if } a_{id'} > 0, \\ \bar{P}_{ij} & \text{otherwise.} \end{cases}$$

Note that all transitions leading to state d in $MC_{\bar{P}}$ will have zero probability in $MC_{P'}$. In fact, consider a state i such that $\bar{P}_{id} > 0$ and $a_{id'} > 0$. In the new matrix P', the probability of transition from i to d will be $P'_{id} = a_{dd'}\bar{P}_{id}/a_{id'}$, which is zero because $a_{dd'} = 0$. Proving that P' is stochastic is now straightforward: First, the entries of P' are positive; Second, they are properly normalized and sum up to one. Indeed, if we consider a state i such that $a_{id'} = 0$, we have that $\sum_j P'_{ij} = \sum_j \bar{P}_{ij} = 1$, whereas if $a_{id'} \neq 0$, we

have that

$$\begin{aligned}\sum_j P'_{ij} &= \sum_j \frac{a_{jd'}}{a_{id'}}\bar{P}_{ij} \\ &= \frac{1}{a_{id'}}\sum_j \bar{P}_{ij}a_{jd'} \\ &= \frac{1}{a_{id'}}(\bar{P}\boldsymbol{a_{d'}})_i = \frac{1}{a_{id'}}a_{id'} = 1\end{aligned}$$

because of (6.8).

(ii) Let p and p' be the probability measures defined, respectively, for MC_P and $MC_{P'}$ on the same sample space $\mathcal{T}_{s'd'}$. Any trajectory from the set $\mathcal{T}_{s'd'}$ has the form $t_{s'd'} = s'x_2...x_kd'$.

If $t_{s'd'} \in \mathcal{T}^d_{s'd'}$,

$$p'(t_{s'd'}) = 0 \tag{6.9}$$

since we have constructed $MC_{P'}$ such that all transitions leading to state d have zero probability.

If $t_{s'd'} \notin \mathcal{T}^d_{s'd'}$, we have

$$\begin{aligned}p'(t_{s'd'}) &= P'_{s'x_2}P'_{x_2x_3}\dots P'_{x_kd'} \\ &= \frac{a_{x_2d'}}{a_{s'd'}}\bar{P}_{s'x_2}\frac{a_{x_3d'}}{a_{x_2d'}}\bar{P}_{x_2x_3}\dots\frac{a_{d'd'}}{a_{x_kd'}}\bar{P}_{x_kd'} \\ &= \frac{a_{d'd'}}{a_{s'd'}}\bar{P}_{s'x_2}\bar{P}_{x_2x_3}\dots\bar{P}_{x_kd'},\end{aligned} \tag{6.10}$$

but $a_{d'd'} = 1$ as the probability to be absorbed by state d', given that we have started at this same state, is 1. Moreover, we know from (6.6) that $P_{ij} = \bar{P}_{ij}$, for all $i \neq d, d'$. As we have supposed that the trajectory $t_{s'd'}$ does not admit either d or d' as intermediate states, $\bar{P}_{s'x_2}\bar{P}_{x_2x_3}\dots\bar{P}_{x_kd'} = P_{s'x_2}P_{x_2x_3}\dots P_{x_kd'}$. Rewriting (6.10) yields

$$\begin{aligned}p'(t_{s'd'}) &= \frac{1}{a_{s'd'}}P_{s'x_2}P_{x_2x_3}\dots P_{x_kd'} \\ &= \frac{p(t_{s'd'})}{1-a_{s'd}} \\ &= \frac{p(t_{s'd'})}{1-p(T_{s'd'}\in\mathcal{T}^d_{s'd'})} = p(t_{s'd'}|T_{s'd'}\notin\mathcal{T}^d_{s'd'}).\end{aligned} \tag{6.11}$$

Combining (6.9) and (6.11), we have therefore proven, for all $t_{s'd'} \in \mathcal{T}_{s'd'}$, that

$$p'(t_{s'd'}) = p(t_{s'd'}|T_{s'd'}\notin\mathcal{T}^d_{s'd'}). \tag{6.12}$$

Consequently, if the random variable $T'_{s'd'}$ describes the trajectory between s' and d' in

$MC_{P'}$, (6.12) implies that

$$H(T_{s'd'}|T_{s'd'} \notin \mathcal{T}^d_{s'd'}) = H(T'_{s'd'}).$$

□

For the particular case where $s' = d'$, we still can use Lemma 2 to express the conditional entropy $H_{s's'|\bar{d}}$: We modify the MC by removing the incoming transitions of s' and creating a new state s'' that will inherit them. The conditional entropy $H_{s's'|\bar{d}}$ in the original MC is equal to $H_{s's''|\bar{d}}$ in the modified one and, since $s' \neq s''$, we can use Lemma 2 to express it.

Building on Lemma 1 and Lemma 2, we can now state the main result of this chapter: a general expression for the entropy of Markov trajectories conditional on multiple intermediate states.

Theorem 1. *Let P be the transition probability matrix of a finite Markov chain and $s\boldsymbol{u}d = su_1 \dots u_l d$ a sequence of states such that $p(T_{sd} \in \mathcal{T}^{\boldsymbol{u}}_{sd}) > 0$. Then, we have the following equality*

$$H(T_{sd}|T_{sd} \in \mathcal{T}^{\boldsymbol{u}}_{sd}) = \sum_{k=0}^{l-1} H(T'_{u_k u_{k+1}}) + H(T_{u_l d}), \tag{6.13}$$

where $u_0 = s$, and $T'_{u_k u_{k+1}}$ is a random trajectory defined on the Markov chain whose transition probability matrix P'_k is defined as follows

$$(P'_k)_{ij} = \begin{cases} 0 & \text{if } i = u_{k+1}, d \text{ and } i \neq j, \\ 1 & \text{if } i = u_{k+1}, d \text{ and } i = j, \\ P_{ij} & \text{if } i \neq u_{k+1}, d \text{ and } \alpha_{idu_{k+1}} = 1, \\ \frac{1-\alpha_{jdu_{k+1}}}{1-\alpha_{idu_{k+1}}} P_{ij} & \text{if } i \neq u_{k+1}, d \text{ and } \alpha_{idu_{k+1}} < 1. \end{cases} \tag{6.14}$$

Proof. The matrix P'_k is obtained from P using (6.14), which is equivalent to applying successively (6.6) and (6.7) where the starting, intermediate and ending states are, respectively, u_k, d and u_{k+1}. Therefore, using Lemma 2, we have

$$H(T'_{u_k u_{k+1}}) = H(T_{u_k u_{k+1}}|T_{u_k u_{k+1}} \notin \mathcal{T}^d_{u_k u_{k+1}})$$

for all $0 \leq k \leq l-1$. Consequently, we can write that

$$\sum_{k=0}^{l-1} H(T'_{u_k u_{k+1}}) + H(T_{u_l d}) = \sum_{k=0}^{l-1} H(T_{u_k u_{k+1}}|T_{u_k u_{k+1}} \notin \mathcal{T}^d_{u_k u_{k+1}}) + H(T_{u_l d}),$$

where $u_0 = s$. Using Lemma 1, we finally obtain

$$\sum_{k=0}^{l-1} H(T'_{u_k u_{k+1}}) + H(T_{u_l d}) = H(T_{sd} | T_{sd} \in \mathcal{T}_{sd}^{\boldsymbol{u}}).$$

□

Now that we have derived a general expression for the entropy of Markov trajectories conditional on multiple states, we introduce, in the next section, a method that enables us to compute this expression.

6.3.1 Entropy computation

The closed-form expression for the entropy of Markov trajectories proposed by Ekroot and Cover [26] is valid only if the Markov chain studied is irreducible. However, the Markov chain $MC_{P'}$ obtained from MC_P after the transformations (6.6) and (6.7) is not necessarily irreducible: All transitions leading to state u have zero probability, which implies that possibly many states do not admit any path leading to d. Therefore, we need an expression for the entropy of Markov trajectories that is valid under milder conditions. In order to identify these conditions, we study the properties of $MC_{P'}$. Let $\mathcal{S}$ be the set of all states in $MC_{P'}$ and let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subsets that partition $\mathcal{S}$ in the following manner

$$\mathcal{S}_1 = \{i \in \mathcal{S} : a_{id} > 0\} \qquad \mathcal{S}_2 = \{i \in \mathcal{S} : a_{id} = 0\}.$$

The set $\mathcal{S}_1$ is closed as no one-step transition is possible from any state in $\mathcal{S}_1$ to any state in $\mathcal{S}_2$. In fact, if $i \in \mathcal{S}_1$ and $j \in \mathcal{S}_2$, (6.7) yields that $P'_{ij} = \bar{P}_{ij} a_{jd} / a_{id} = 0$. Clearly, all trajectories leading to state d are composed of states belonging to $\mathcal{S}_1$. Now, we propose a closed-form expression for the entropy of Markov trajectories that is valid under the weaker condition that the destination state d can be reached from any other state of the MC. Moreover, we prove that the trajectory entropy can be expressed as a weighted sum of local entropies. We also provide an intuitive interpretation of the weights.

Lemma 3. *Let P be the transition probability matrix of a finite state MC such that there exists a path with positive probability from any state to a given state d. Let Q_d be a sub-matrix of P obtained by removing the d^{th} row and column of P.*

$$P = \left(\begin{array}{c|c} Q_d & \begin{matrix} P_{1d} \\ \vdots \end{matrix} \\ \hline \begin{matrix} P_{d1} & \cdots \end{matrix} & P_{dd} \end{array} \right).$$

For any state $s \neq d$, the trajectory entropy H_{sd} can be expressed as

$$H_{sd} = \sum_{k \neq d} ((I - Q_d)^{-1})_{sk} H(P_{k\cdot}), \tag{6.15}$$

where $H(P_{k\cdot})$ is the local entropy of state k.

Proof. First, observe that the matrix Q_d is a sub-matrix of P corresponding to all states except state d and that we use Q_d to derive the entropy of all trajectories ending at d. Applying the chain rule for entropy, we express the entropy of a trajectory as the entropy of the first step plus the conditional entropy of the rest of the trajectory, given this first step

$$H_{sd} = H(P_{s\cdot}) + \sum_{k \neq d} P_{sk} H_{kd}.$$

We expand this equality further by recursively expanding the entropy H_{kd} as follows

$$\begin{aligned}
H_{sd} &= H(P_{s\cdot}) + \sum_{k \neq d} P_{sk} \left(H(P_{k\cdot}) + \sum_{k' \neq d} P_{kk'} H_{k'd} \right) \\
&= H(P_{s\cdot}) + \sum_{k \neq d} P_{sk} H(P_{k\cdot}) + \sum_{k \neq d} P_{sk} \sum_{k' \neq d} P_{kk'} H_{k'd} \\
&= H(P_{s\cdot}) + \sum_{k \neq d} P_{sk} H(P_{k\cdot}) + \sum_{k \neq d} P_{sk} \sum_{k' \neq d} P_{kk'} \cdot \left(H(P_{k'\cdot}) + \sum_{k'' \neq d} P_{k'k''} \left(H(P_{k''\cdot}) + \dots \right) \right) \\
&= H(P_{s\cdot}) + \sum_{k \neq d} \left(\sum_{i=1}^{\infty} ({Q_d}^i)_{sk} \right) H(P_{k\cdot}) = \sum_{k \neq d} \left(\sum_{i=0}^{\infty} ({Q_d}^i)_{sk} \right) H(P_{k\cdot}),
\end{aligned} \tag{6.16}$$

with ${Q_d}^0 = I$.

Observe that the matrix Q_d describes the Markov chain as long as it does not reach state d. Moreover, the matrix Q_d has a finite number of states and there is a path with positive probability from each state to state d. As a consequence, the Markov process will enter state d with probability 1, i.e., $\lim_{n \to \infty} {Q_d}^n = O$ (zero matrix). In addition, since

$$(I - Q_d)(I + Q_d + {Q_d}^2 + \dots + {Q_d}^{n-1}) = I - {Q_d}^n,$$

we can easily verify that

$$\sum_{i=0}^{\infty} {Q_d}^i = (I - Q_d)^{-1}. \tag{6.17}$$

Replacing (6.17) in (6.16), we have

$$H_{sd} = \sum_{k \neq d} ((I - Q_d)^{-1})_{sk} H(P_{k\cdot}).$$

□

We have shown that the entropy of a family of trajectories can be expressed as a weighted sum of the states' local entropies. The weights are given by the matrix $(I - Q_d)^{-1}$. In the Markovian literature, the matrix $(I - Q_d)^{-1}$ is referred to as the fundamental matrix [44, 67]. In fact, the $(sk)^{\text{th}}$ element of the fundamental matrix (defined with respect to the destination state d) can be seen as the expected number of visits to the state k before hitting the state d, given that we started at state s. As a result, the entropy of the random trajectory T_{sd} is the sum over the chain states of the expected number of visits to each state multiplied by its local entropy. This is a remarkable observation as it links a global quantity, which is the trajectory entropy, to the local entropy at each state.

Example I Recall that in the example shown in Figure 6.1, we found that the entropy of the trajectory T_{15} is equal to 1.56 bits. We can retrieve this result by computing the fundamental matrix with respect to state 5. First, we remove the 5^{th} row and column of matrix P to extract the submatrix

$$Q_5 = \begin{pmatrix} 0 & 0.25 & 0.75 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and then compute the fundamental matrix

$$(I - Q_5)^{-1} = \begin{pmatrix} 1 & 0.625 & 0.75 & 0.375 \\ 0 & 1 & 0 & 0 \\ 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The $(ij)^{\text{th}}$ element of this matrix is equal to the expected number of visits to state j before hitting state 5, given that we started at state i. For example, the expected number of visits to state 3, given that we start at state 1, is equal to 0.75 as a trajectory T_{15} can go through state 3 only once and this happens with probability 0.75. Multiplying the fundamental matrix by the matrix of local entropies

$$\begin{pmatrix} 0.81 & 0.81 & 0.81 & 0.81 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

yields the column vector of trajectory entropy

$$\begin{pmatrix} 1.56 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

As expected, we retrieve the results obtained in Section 6.2. For example, the entropy of the trajectory $T_{15} = 1.56$ bits whereas the entropy of the trajectory $T_{35} = 1$ bit.

Example II In Figure 6.4, we show a 7×7 grid characterized by an upper half that has more edges than the lower half. We created these irregularities by randomly removing vertical edges from the lower half of grid. We consider a biased random walk on this grid: the probability of moving to the right, up and down are 0.8, 0.1, and 0.1, respectively. We are interested in the trajectories between the starting vertex $s = (1, 4)$ and destination $d = (7, 4)$, and we color each vertex u of the graph according to the probability α_{sud} that the trajectory T_{sd} goes through this vertex. The result is shown in Figure 6.4.

For each vertex u of the same graph, we compute the conditional trajectory entropy $H_{sd|u}$, and we color each vertex according to the normalized entropy $H_{sd|u}/H_{sd}$. As shown in Figure 6.5, revealing a vertex along the most likely path from s to d decreases trajectory entropy, as uncertainty diminishes and as we become more confident that the trajectory follows the direct path from s to d. However, revealing a vertex might increase trajectory entropy. For example, revealing that the trajectory T_{sd} went through vertex $(7, 5)$ increases trajectory entropy because the posterior distribution of trajectories becomes concentrated in the upper part of the grid that is much richer that the other parts of the graph. We emphasize the fact the the probability of the revealed intermediate vertex does not determine its influence on trajectory entropy: the vertices $(1, 1)$ and $(1, 7)$ have the same probability of belonging to the trajectory T_{sd} but have opposite effects on trajectory entropy. On one hand, conditioning on vertex $(1, 7)$ increases trajectory entropy because the family of possible trajectories is much richer in the upper part of the grid than in the rest of the grid. On the other hand, conditioning on vertex $(1, 1)$ decreases trajectory entropy because the lower part of the graph has fewer edges, which reduces the uncertainty about the random trajectory.

6.3.2 Algorithm

In Algorithm 2, we define the set of steps needed to compute the entropy of Markov trajectories conditional on a set of intermediate states. The worst-case running time for the algorithm is $\mathcal{O}(lN^3)$, where N is the number of states of MC_P, and l the length of the sequence of intermediate states $\boldsymbol{u}$. This complexity is dominated by the cost of computing the inverse of the matrix $(I - Q_d)$, which is needed to compute the entropy H_{sd}

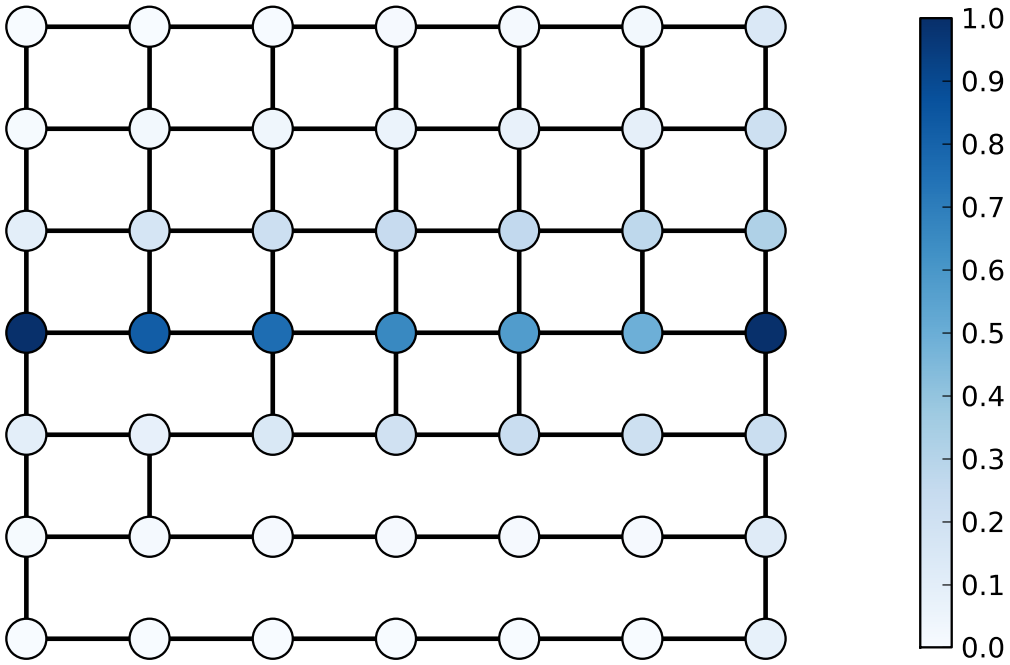


Figure 6.4 – We consider a biased random walk on 7×7 grid and color each vertex u according to the probability α_{sud}, where $s = (1, 4)$ and $d = (7, 4)$.

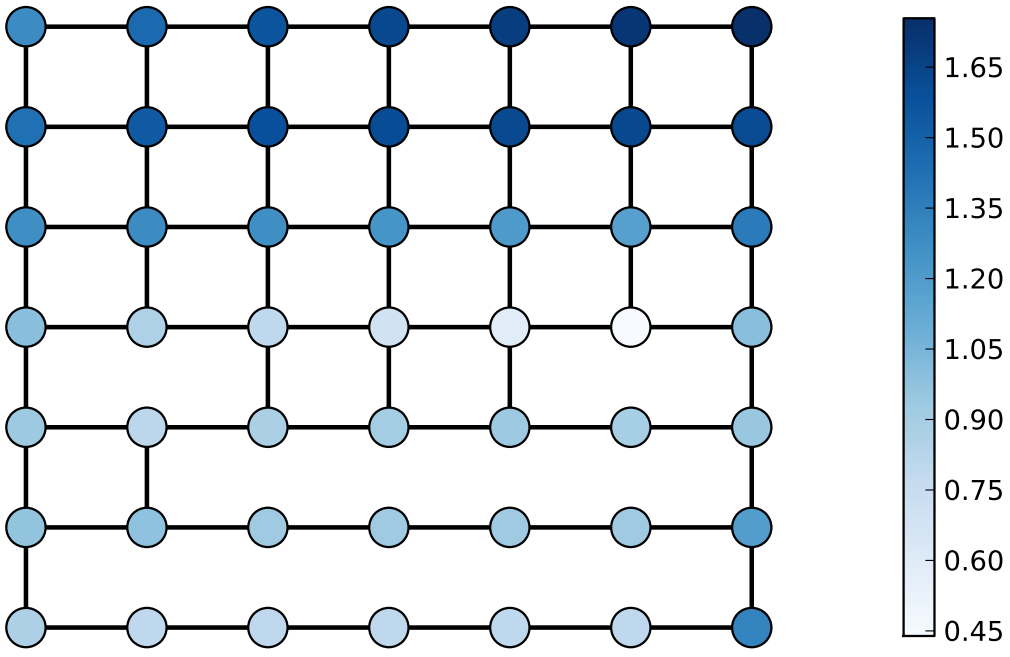


Figure 6.5 – We consider a biased random walk on 7×7 grid and color each vertex u according to the normalized entropy $H_{sd|u}/H_{sd}$.

Algorithm 2: Computing the conditional trajectory entropy

Input: Matrix of transition probability P,
source state s, destination state d,
sequence of intermediate states $\boldsymbol{u} = u_1 \dots u_l$.
Output: $H_{sd|u_1 \dots u_l}$

$u_0 \leftarrow s$;
for $k = 0$ *to* $l - 1$ **do**
  Compute P'_k from P using (6.14) ;
  Compute $H(T'_{u_k u_{k+1}})$ from P'_k using Lemma 3 ;
  $H_{u_k u_{k+1}|\bar{d}} \leftarrow H(T'_{u_k u_{k+1}})$;
Compute $H_{u_l d}$ from P using Lemma 3;
$H_{sd|u_1 \dots u_l} = \sum_{k=0}^{l-1} H_{u_k u_{k+1}|\bar{d}} + H_{u_l d}$;
return $H_{sd|u_1 \dots u_l}$;

in (6.15). However, since we only need the s^{th} row of the matrix $(I - Q_d)$ to compute the trajectory entropy H_{sd}, we can solve a system of—potentially sparse—linear equations. Moreover, many iterative methods [33, p. 508] take advantage of the structure of the matrix representing the system of linear equations in order to solve them efficiently.

Coming back to the example shown in Figure 6.1, we use the algorithm above to compute the conditional entropy $H_{15|3} = 1$ bit. We leave no ambiguity about the trajectory T_{15} when we condition on both states 3 and 2 and find that $H_{15|3,2} = H_{13|\bar{5}} + H_{32|\bar{5}} + H_{25} = 0$ bits.

6.3.3 Divergence between the prior and posterior distribution of trajectories

We are interested in quantifying the divergence between the prior distribution of trajectories $p(T_{sd})$ and the posterior distribution of trajectories given an additional information that indicates that the trajectory T_{sd} belongs to a set $\mathcal{A} \subset \mathcal{T}_{sd}$. We are therefore interested in the divergence between the distributions $p(T_{sd})$ and $p(T_{sd}|T_{sd} \in \mathcal{A})$. In order to quantify this divergence, we compute the *relative entropy* or *Kullback-Leibler divergence* [21] between these distributions

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{A}||T_{sd}) = \sum_{t_{sd} \in \mathcal{T}_{sd}} p(t_{sd}|T_{sd} \in \mathcal{A}) \log \frac{p(t_{sd}|T_{sd} \in \mathcal{A})}{p(t_{sd})}.$$

We show the following result

Proposition 2.

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{A}||T_{sd}) = -\log p\left(T_{sd} \in \mathcal{A}\right).$$

Proof. We model the additional information as an indicator variable of the set of trajectories $\mathcal{A} \subset \mathcal{T}_{sd}$

$$I_{\{\mathcal{A}\}} = \begin{cases} 1 & \text{if } T_{sd} \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases}$$

Using the definition of the Kullback-Leibler divergence, we can write

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{A}||T_{sd}) = \mathrm{KL}(T_{sd}|I_{\{\mathcal{A}\}} = 1||T_{sd}) = \sum_{t_{sd} \in \mathcal{T}_{sd}} p(t_{sd}|I_{\{\mathcal{A}\}} = 1) \log \frac{p(t_{sd}|I_{\{\mathcal{A}\}} = 1)}{p(t_{sd})}.$$

For ease of notation, we replace $I_{\{\mathcal{A}\}}$ by I. Because $p(t_{sd}|I = 1) = 0$ if $t_{sd} \notin \mathcal{A}$, we have

$$\mathrm{KL}(T_{sd}|I = 1||T_{sd}) = \sum_{t_{sd} \in \mathcal{A}} p(t_{sd}|I = 1) \log \frac{p(t_{sd}|I = 1)}{p(t_{sd})}.$$

Applying Bayes' theorem, we have

$$p(t_{sd}|I = 1) = \frac{p(I = 1|t_{sd})p(t_{sd})}{p(I = 1)}.$$

However, for all $t_{sd} \in \mathcal{A}$,

$$p(I = 1|t_{sd}) = 1,$$

which implies

$$p(t_{sd}|I = 1) = \frac{p(t_{sd})}{p(I = 1)} \qquad \forall t_{sd} \in \mathcal{A}.$$

It follows that

$$\begin{aligned} \mathrm{KL}(T_{sd}|I = 1||T_{sd}) &= \sum_{t_{sd} \in \mathcal{A}} p(t_{sd}|I = 1) \log \frac{p(t_{sd})}{p(t_{sd})p(I = 1)} \\ &= -\log p(I = 1) \sum_{t_{sd} \in \mathcal{A}} p(t_{sd}|I = 1) \\ &= -\log p(I = 1), \end{aligned}$$

which is equivalent to

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{A}||T_{sd}) = -\log p\left(T_{sd} \in \mathcal{A}\right).$$

□

We are therefore able to compute the divergence between the prior distribution $p(T_{sd})$ and the posterior distribution of this trajectory, given additional information represented by the event $\{T_{sd} \in \mathcal{A}\}$: the less likely the event $\{T_{sd} \in \mathcal{A}\}$ is, the larger the divergence

between the distributions $p(T_{sd})$ and $p(T_{sd}|T_{sd} \in \mathcal{A})$ is. This result enables us, for example, to use the bound [21, p. 300] in order to bound the L_1 norm between these two distributions

$$\|p(T_{sd}) - p(T_{sd}|T_{sd} \in \mathcal{A})\|_1 = \sum_{t_{sd} \in \mathcal{T}_{sd}} |p(t_{sd}) - p\left(t_{sd}|T_{sd} \in \mathcal{A}\right)| \leq -2\ln(2) \log p\left(T_{sd} \in \mathcal{A}\right).$$

More importantly, we will use this result in order to compute the divergence between the conditional random trajectory $T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u$ and the random trajectory (T_{su}, T_{ud})

Proposition 3. $\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u || (T_{su}, T_{ud})) = -\log p\left(T_{su} \notin \mathcal{T}_{su}^d\right)$.

Proof. By definition, we have

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u || (T_{su}, T_{ud})) = \mathbb{E}\left(\log \frac{p(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u)}{p(T_{su}, T_{ud})}\right).$$

The random trajectory T_{sd} can be expressed as a sequence of random sub-trajectories (T_{su}, T_{ud}). Moreover, the Markovian nature of the process that generates the trajectory T_{sd} implies that the sub-trajectories T_{su} and T_{ud} are independent of each other, given the intermediate point u.

$$\begin{aligned}
\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u || (T_{su}, T_{ud})) &= \mathbb{E}\left(\log \frac{p(T_{su}, T_{ud}|T_{sd} \in \mathcal{T}_{sd}^u)}{p(T_{su})p(T_{ud})}\right) \\
&= \mathbb{E}\left(\log \frac{p(T_{su}|T_{sd} \in \mathcal{T}_{sd}^u)p(T_{ud}|T_{su}; T_{sd} \in \mathcal{T}_{sd}^u)}{p(T_{su}, T_{ud})}\right) \\
&= \mathbb{E}\left(\log \frac{p(T_{su}|T_{su} \notin \mathcal{T}_{su}^d)p(T_{ud})}{p(T_{su})p(T_{ud})}\right) \\
&= \mathbb{E}\left(\log \frac{p(T_{su}|T_{su} \notin \mathcal{T}_{su}^d))}{p(T_{su})}\right) \\
&= \mathrm{KL}(T_{su}|T_{su} \notin \mathcal{T}_{su}^d || T_{su}).
\end{aligned}$$

Using Proposition 2, we can write

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u || (T_{su}, T_{ud})) = -\log p\left(T_{su} \notin \mathcal{T}_{su}^d\right).$$

□

This result implies that, if $p(T_{su} \notin \mathcal{T}_{su}^d) = 1$, the random trajectories $T_{sd}|T_{sd} \in \mathcal{T}_{sd}^u$ and (T_{su}, T_{ud}) are identically distributed. Exploiting the fact that $H(T_{su}, T_{ud}) = H_{su} + H_{ud}$, we obtain that

$$H_{sd|u} = H_{su} + H_{ud}. \tag{6.18}$$

Recall that, in the beginning of Section 6.3, we gave an example that shows that this equality does not hold in general. We have now expressed a condition under which this

equality holds. More importantly, we are now able to quantify the divergence between the distributions $T_{sd}|T_{sd} \in \mathcal{T}_{sd}^{u}$ and (T_{su}, T_{ud}). A direction of future research is to find an upper bound for the difference between the entropy $H_{sd|u}$ and the sum of entropies $H_{su} + H_{ud}$.

Conditioning on a set of states

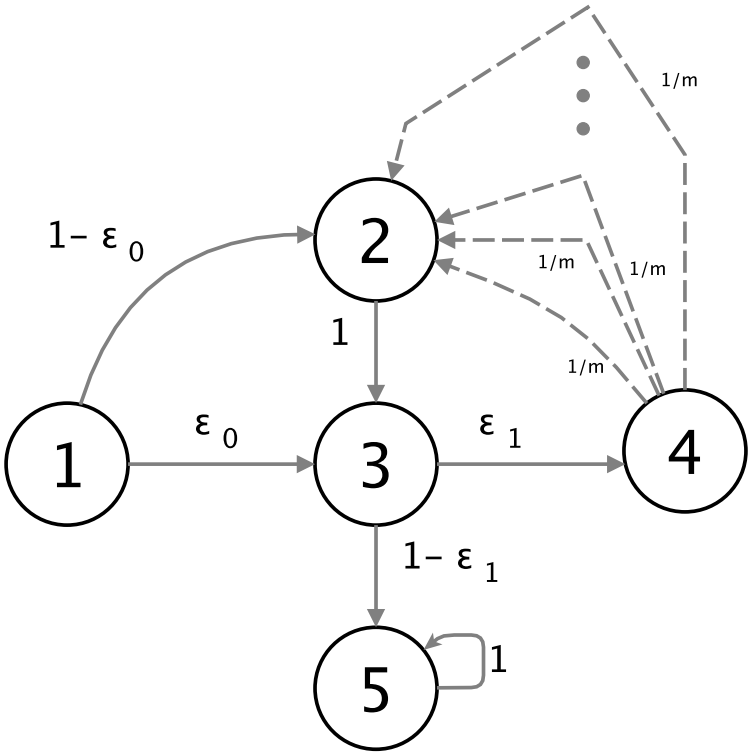


Figure 6.6 – A Markov chain annotated with the transition probabilities. The dashed lines between states 4 and 2 represent the m equiprobable paths leading from state 4 to state 2. We choose $0 < \epsilon_1 < 1$ and $m \geq 1$ to guarantee that $|\mathcal{T}_{15}| > 0$ and that $p(T_{15} \in \mathcal{T}_{15}^{3,2}) > 0$.

So far, we have focused on computing the entropy of Markov trajectories conditional on a *sequence* of states. A natural extension is the computation of this entropy conditional on a *non-ordered* set of states. Finding a general expression for this conditional entropy appears to be very hard, and there is no simple relation linking it to the entropy conditional on a sequence. In Figure 6.6, we show an example that illustrates an interesting and counter-intuitive result about conditioning on a set of states. Intuitively, we would expect that the entropy of a random trajectory conditional on a sequence of states is always less than the entropy of the same trajectory conditional on the set formed by these states. However, this is not true. We take the MC shown in Figure 6.6 as an example and we compute, using Theorem 1, the entropy of the random trajectory T_{15} conditional on

going through the sequence of intermediate states $(3, 2)$

$$H_{15|32} = H_{13|\bar{5}} + H_{32|\bar{5}} + H_{25} = h(\epsilon_0) + \log m + H_{35},$$

where $h(\epsilon_0)$ is the entropy of a Bernoulli random variable with success probability ϵ_0. To compute the entropy of the random trajectory T_{15} conditional on going through the set of states $\{2, 3\}$, we apply the chain rule for entropy and express the entropy of a trajectory as the entropy of the first two steps plus the conditional entropy of the rest of the trajectory given these first two steps

$$H_{15|\{2,3\}} = h\left(\frac{\epsilon_0\epsilon_1}{1-\epsilon_0(1-\epsilon_1)}\right) + \frac{\epsilon_0\epsilon_1}{1-\epsilon_0(1-\epsilon_1)}H_{45} + \frac{1-\epsilon_0}{1-\epsilon_0(1-\epsilon_1)}H_{35}.$$

Since $H_{45} = \log m + H_{25} = \log m + H_{35}$, we have that

$$H_{15|\{2,3\}} = h\left(\frac{\epsilon_0\epsilon_1}{1-\epsilon_0(1-\epsilon_1)}\right) + \frac{\epsilon_0\epsilon_1}{1-\epsilon_0(1-\epsilon_1)}\log(m) + H_{35}. \tag{6.19}$$

Using (6.3.3) and (6.19), we can write

$$H_{15|32} - H_{15|\{2,3\}} = h(\epsilon_0) - h\left(\frac{\epsilon_0\epsilon_1}{1-\epsilon_0(1-\epsilon_1)}\right) + \frac{1-\epsilon_0}{1-\epsilon_0(1-\epsilon_1)}\log m.$$

This difference can therefore be lower bounded by

$$H_{15|32} - H_{15|\{2,3\}} \geq -1 + \frac{1-\epsilon_0}{1-\epsilon_0(1-\epsilon_1)}\log m.$$

As a consequence, if $\log m > 1 + \epsilon_0\epsilon_1/1 - \epsilon_0$, the entropy of the random trajectory T_{15} conditional on going through the sequence $(3, 2)$ is strictly greater than the entropy of the same trajectory conditional on going through the set of states $\{2, 3\}$. The reason is that conditioning on the sequence $(3, 2)$ implies that the random trajectory T_{15} is composed of a random sub-trajectory T_{42} whose entropy can be made arbitrary large by increasing the parameter m. More generally, this example illustrates the absence of a simple relation between the entropy of random trajectories conditional on a sequence of states and the entropy of the same trajectory conditional on the set formed by these same states.

6.4 Conclusion

In this chapter, we have taken an information theoretic approach in order to quantify rigorously mobility uncertainty and its evolution with additional information. We have modeled individual mobility as a random trajectory on a graph that represents a discretized map of an individual's world. We quantify mobility uncertainty as the entropy of the distribution of possible trajectories and addressed the problem of computing the

entropy of *conditional* Markov trajectories. Our method is based on a transformation of the original Markov chain into a Markov chain that yields the desired conditional entropy. We also derived an expression that enables us to compute the entropy of Markov trajectories, under conditions weaker than those assumed in [26]. Furthermore, we have shown the link between the entropy of Markov trajectories —a global quantity—to the local entropy of states. We also computed the divergence between the prior and posterior distribution of trajectories. These results have applications in various fields including mobility mining and trajectory segmentation. In fact, we show in the next chapter, how our framework enables us to quantify the uncertainty about user mobility and its evolution with locations updates, and to uncover intermediate destinations along a given trajectory in the absence of time information.

7 Applying Trajectory Entropy to Mobility

7.1 Introduction

In their seminal paper [26], Ekroot and Cover presented a method to compute the entropy of Markov trajectories for irreducible finite-state Markov chains. This work was based on the need to quantify the thermodynamic depth as defined by [51]: the descriptive complexity of the process (or path) by which a state of a physical system arises. In Chapter 6, we have extended this work and have presented a method that enable us to compute the entropy of Markov trajectories as additional information about intermediate states becomes available. More generally, quantifying the randomness of Markov trajectories offers us powerful tools that enable us to study the properties of dynamics on graphs.

In this chapter, we present two mobility-related applications for which we use tools borrowed from the trajectory entropy framework introduced in Chapter 6. First, we quantify mobility uncertainty and its evolution with location updates and link the evolution of conditional entropy to the nature of the intermediate locations revealed. Then, we propose a segmentation algorithm that infers the likely intermediate destinations along a trajectory, based on the conditional trajectory entropy. For both applications, we take an empirical approach that is based on the analysis of actual city maps and a large scale dataset of $20,000$ GPS trajectories.

7.2 Datasets

In this section, we introduce the datasets we use throughout this chapter. The first dataset is composed of graphs that represent city maps, and the second dataset is composed of GPS trajectories collected for Geolife [74], an experiment that involved 182 users over a five-year period.

OpenStreetMap We can describe the urban environment in which we evolve by a graph whose vertices and edges represent the geometrical shape of the roads. To obtain this representation, we use the data from OpenStreetMap (OSM), a collaborative project, with over 1.9 million registered users, that was created in order to provide free geographic data. The geo-spatial databases of OSM are built by a community of mappers who contribute to maintain up to date data about roads, trails, and points of interest. The maps provided by the OSM project are represented using a standard geospatial vector data format called shapefile. We download the shapefiles that encode these maps, and we process them to represent the city road map as a graph $G(V, E)$ similar to the one shown in Figure. 7.1. In addition to being connected, the graph G is weighted: we associate with each edge $(i, j) \in E$ a cost $c_{ij} > 0$ equal to its length.

Figure 7.1 – The graph extracted from the geospatial data of OSM about the city of Lausanne, Switzerland. The vertices and edges of the graph enable us to represent the geometrical shape of the roads.

Geolife The Geolife Project [74] consisted in collecting the mobility traces of 182 users over a five-year period. The collected dataset contains around $18,000$ trajectories (more than $1,300,000$ km), mainly located in China. A trajectory of this dataset is represented as a sequence of time-stamped points described by their latitude, longitude and altitude. The sampling rates vary between users but remain very high in general: 91.5% of the trajectories are logged in a dense representation (e.g., every $1 - 5$ seconds). The range of activities associated with these trajectories is also quite broad: some trajectories are associated with home-to-work routines, whereas others are associated with shopping and sightseeing. We process the data as follows: First, we discretize the GPS records

by dividing the surface of the globe into identical areas (squares whose side lengths are 1km). A square basically represents the set of locations enclosed within the area it covers. Then we represent each trajectory in the dataset as the sequence of areas visited and the associated *residence* time in each area i.e., the total time spent by the user in this area. In Figure. 7.2, we show the empirical distribution of trajectory length given as the number of areas covered by the trajectory. The trajectory lengths range from very short trajectories (1 or 2 locations) that correspond to short urban trips to very long trajectories that correspond to inter-city trips. In fact, in the raw dataset, 36% of the trajectories span a distance that is less than 5 km, whereas 5% of the trajectories span a distance superior to 100 km. Moreover, as the majority of the trajectories are geographically within city of Beijing [74], we choose to focus on the data produced in this capital.

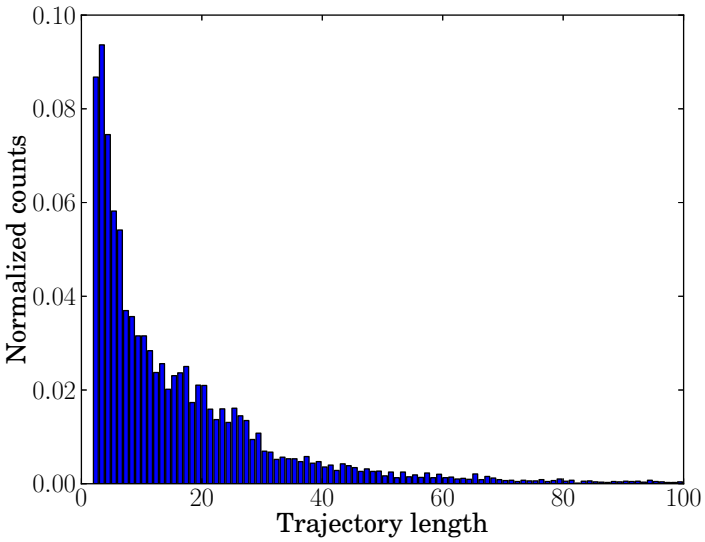


Figure 7.2 – Histogram of the length of Geolife trajectories after pre-processing. The trajectory lengths range from very short trajectories that correspond to short urban trips to very long trajectories that correspond to inter-city trips.

7.3 Mobility Uncertainty and Its Evolution with Location Updates

Mobility is a central aspect of our life and the locations we visit reveal our habits, tastes, occupations and personal traits. This has led to a wide spectrum of location-based services and applications (maps, social networks, local search, etc.) that collect mobility patterns: Mobility data offers a unique economic opportunity, as companies and local businesses are able to precisely target a given audience (based on age, gender, consumption habits) and promote their product with high effectiveness.

The collateral collection of this information raises strong concerns and vigorous debates about the privacy implications of mobility data mining. A particular problem arises in mobility inference: some services can benefit from inferring the locations visited by a user, but obviously this predictability also carries major risks to privacy. This is especially important given that users tend to overestimate their privacy level when sharing information [8]. Quantifying the evolution of the uncertainty about our mobility if we share our location lets us situate the subjective frontier above which we consider that sharing becomes over-sharing. This empowers users with an objective technique to protect their mobility privacy: they are able to anticipate the evolution of their mobility uncertainty as they reveal a subset of the locations they visited. This is especially important in the light of the results we find: The informational value of locations varies greatly, and revealing one intermediate location along a trajectory might be as threatening to privacy as revealing the whole trajectory would be.

In this section, we use the tools borrowed from the trajectory entropy framework in order to measure the level of mobility uncertainty and its evolution with location updates. We assume that we have a Markov chain that captures the users' mobility patterns. As seen in Chapter 2, this MC can be an individual-specific, collective or a combination of both (e.g., a random mixture of individual and collective Markov chains). A user then provides selective location updates based on her needs (e.g., a search request or a check-in); these updates, in combination with the mobility model, enable the service provider to infer the user's trajectory with a level of uncertainty.

7.3.1 Analysis of trajectories on city maps

In this section, we model urban mobility by applying a route-choice model that was developed by the community of research on transportation, on city maps. We then show how knowing the structure of a city and a few locations visited might be sufficient to make mobility very predictable.

Route choice model For a given source vertex s and destination vertex d, we associate with each edge $(i,j) \in E$ a weight $\omega_{(i,j)|s,d}$ defined as

$$\omega_{(i,j)|s,d} = 1 - \left(1 - \left(\frac{D_{sd}}{D_{si} + c_{ij} + D_{jd}}\right)^{b_1}\right)^{b_2}, \tag{7.1}$$

where D_{ij} is the cost of the shortest path between vertices i and j, and c_{ij} the length of edge (i,j). The weight (7.1) is inspired from the cumulative distribution function of Kumaraswamy's double-bounded distribution [47], defined on the interval $[0,1]$ and having two non-negative shape parameters b_1 and b_2. It indicates to which extent the cost of a path going through the edge (i,j) deviates from the cost of the shortest path between

the vertices s and d. The weights definition is based on the paper [31] in which the authors propose a method to stochastically generate paths for a given origin-destination pair, without having to enumerate all paths between these points. Considering that the choice of a route is only justified by its length is a simplifying yet realistic assumption: Golledge et al. [32] are interested in the criteria people use to choose a route to take between two given locations. Using both laboratory and field experiments, the authors conclude that a person's perception of distance and environment configuration is very subjective and varies greatly between individuals. Nevertheless, the predominant criteria for route choice remains shortest time or distance.

Now that the mobility graph is well defined, we model the user mobility as a second order MC whose state space is the set of vertices V. Equivalently, we can represent it as a first order Markov chain X_i with an extended state space E: a state represents a directed edge $(i, j) \in E$.

Therefore, the transition probabilities are given by

$$P_{(i,j),(k,l)} = P(X_{n+1} = (k, l)|X_n = (i, j))$$

As we are interested in the mobility between two fixed vertices s and d, we define the transition probabilities between the states $(i, j), (k, l) \in E$ as

$$P_{(i,j),(k,l)} = \begin{cases} \dfrac{w_{(k,l)|s,d}}{\sum_{l' \in \Gamma(k) \setminus l} w_{(k,l')|s,d}} & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{7.2}$$

Note that we choose a *second* order MC because we want to keep in memory the momentum of the mobility and to have a realistic behavior of the random walker: The next location he will visit is different from the one he just left.

In Figure 7.3, we plot the road map $G_L = (V_L, E_L)$ of an area surrounding the train station in the city of Lausanne, Switzerland. We are interested in the trajectories between two locations represented by the vertices s (green star) and d (red triangle). Using the mobility model defined in (7.2) with $b_1 = 2$ and $b_2 = 1$, we obtain a second order MC, where a state represents a sequence of two vertices or, more simply, a directed edge in E_L.

The entropy H_{sd} is equal to 7.13 bits, the expected number of bits needed to represent the random trajectory T_{sd}. Equivalently, the trajectory T_{sd} is as predictable as a random choice made between $2^{7.13} = 140$ objects. However, as we will see below, revealing only one intermediate location might be enough to make this trajectory as predictable as a random choice between two values.

In order to quantify the effect of revealing an intermediate location along a given edge,

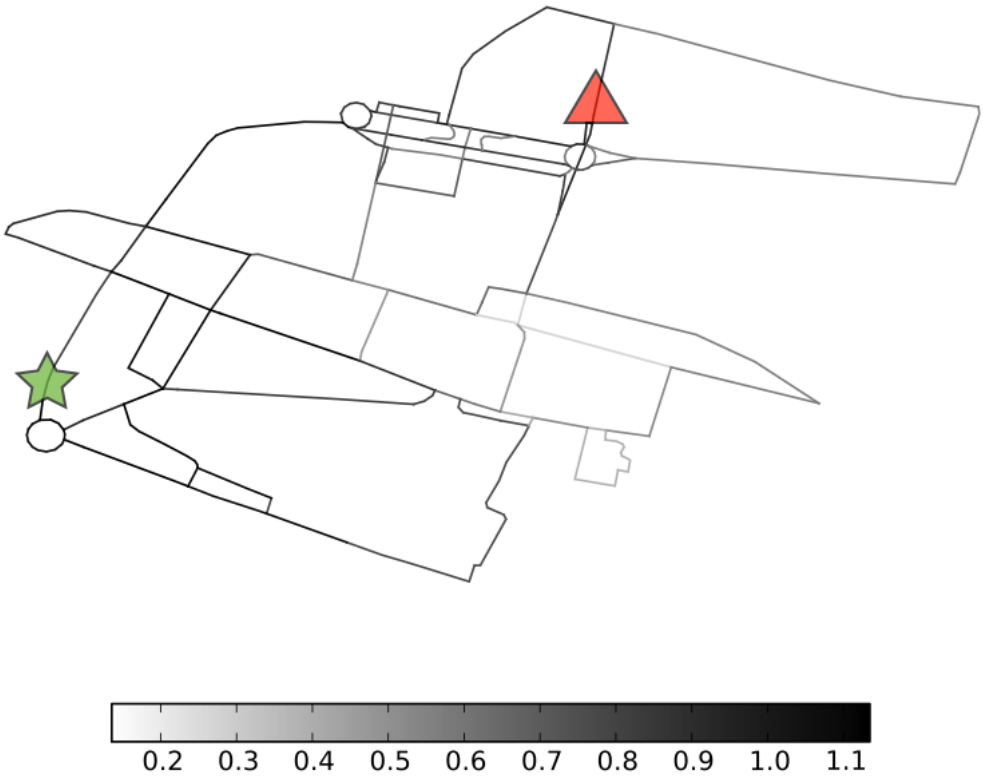


Figure 7.3 – The graph $G_L = (V_L, E_L)$ represents an area around the train station of the city of Lausanne. We focus on the trajectory between the vertices s (green star) and d (red triangle), and color each directed edge (u, v) proportionally to the value of the conditional entropy $H_{sd|(u,v)}/H_{sd}$. Light gray represents a low value of entropy and hence a low uncertainty about trajectory.

we associate with each directed edge (i, j) the value of the conditional entropy $H_{sd|(i,j)}$, which represents our level of uncertainty the trajectory between the locations s and d given that it goes through the edge (i, j). We plot in the Figure 7.3 the graph G_L and color each edge $(i, j) \in E_L$ with a color that is proportional to the value $H_{sd|(i,j)}/H_{sd}$. First, we notice a high variability of the quantity $H_{sd|(i,j)}/H_{sd}$ whose range is the interval $[0.13, 1.1]$. Unsurprisingly, this means that we cannot consider location updates as having an equal effect on the trajectory uncertainty: revealing one location can have almost no effect on the uncertainty about a trajectory, whereas revealing another location can be more threatening to privacy as it drastically decreases trajectory uncertainty.

To understand the reason behind an important decrease of the entropy value, we have to dig a bit deeper and study the trajectory conditional distribution. We observe that the distribution of trajectories conditioned on the directed edge that minimizes entropy is dominated by two trajectories with very close probabilities. If we reveal this intermediate edge, the randomness of the trajectory would be equivalent to the randomness of Bernoulli

random variable with $p \simeq 0.5$.

Revealing an intermediate location might increase trajectory entropy. In fact, the maximal conditional entropy is slightly higher than the entropy H_{sd} because conditioning on a intermediate location might yield a posterior distribution of trajectories that is very different from the prior. This posterior distribution of trajectories might have an entropy larger than the unconditional entropy. In Section 7.4.2, we will explore the characteristics of intermediate locations that increase trajectory entropy and we will show their links with intermediate destinations.

7.3.2 Analysis of GPS trajectories

In this section, we analyze the GPS traces collected for the Geolife Project [74] to gain more insight into the link between conditional entropy, mobility uncertainty, and the position of intermediate locations. Using the Geolife datatset, we focus on the pair of squares (s, d) with the largest set of trajectory realizations, in other words, the number of trajectories starting at square s and ending at square d is larger than the number of trajectories between any other pair of source-destination. Limiting our analysis to this set of trajectories enables us to interpret and understand the link between the evolution of conditional entropy and the geographical position of intermediate locations. In Figure 7.4, we plot the *raw* trajectories and observe that two main roads, with similar lengths, allow for reaching the destination d. As we will see in this section, a side information indicating which one of these roads is followed to reach the destination has an important effect on the uncertainty about chosen trajectory. Finally, we infer, using a maximum likelihood estimator, the first order MC that have generated the observed trajectories. In such a model, an observed trajectory t_{sd} is a sequence of states —squares of 1 km^2 —which starts with the state s, ends with the state d, and admits no intermediate state d. We remove self transitions because they reflect only the number of samples within an area. Note that we choose a first order MC because the training data available is not sufficient for training a higher order MC, similar to the one presented in Section 7.3.1. Having constructed the MC that represents the mobility between two fixed areas, we can apply the trajectory entropy framework in order to quantify the evolution of trajectory uncertainty.

In Figure 7.5, we show the set of locations that, when revealed, decreases the most significantly trajectory entropy. The fact that this area is along one of the two main roads leading from s to d explains its high likelihood: the probability of going through it is 0.56. Knowing this, it is not surprising that the entropy decreases to 73% of its initial value by just conditioning the trajectory on going through this blue square. Revealing this intermediate location excludes the trajectories that go through the second main road leading to d, thus decreasing significantly the uncertainty about the trajectory taken. In contrast, revealing some other locations has a small effect on trajectory uncertainty:

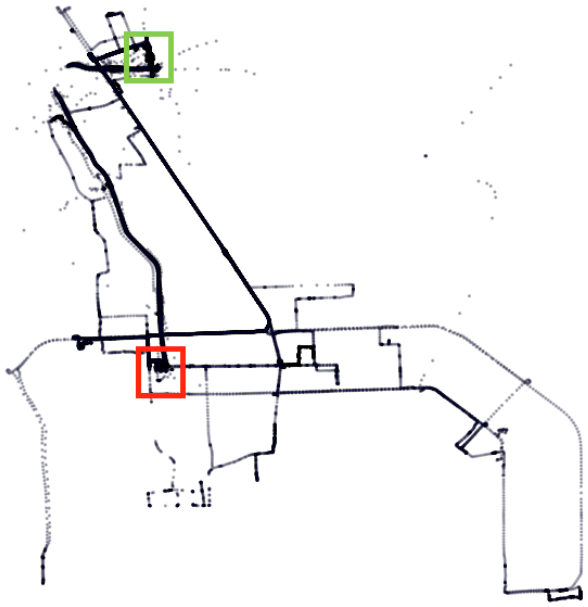


Figure 7.4 – We plot all the *raw* trajectories, starting inside the area delimited by the green square (upper part) and ending inside the area delimited by the red square (lower part). Observe that the starting and ending areas are connected by two main roads.

The probability of going through an intermediate location shown in Figure 7.5 is equal to 0.96 as this square is just next to the starting location when heading towards the destination. As a result, revealing such information has a small effect on the trajectory distribution and decreases the entropy by only 2%. Similarly to the results shown in Section 7.3.1, we observe that revealing a location might yield a posterior distribution of trajectories that is completely different from the prior distribution. As a consequence, the entropy of the conditional trajectory might be larger than the initial entropy. Figure 7.5 shows how revealing an intermediate location —that is on a lengthy path between the source and destination —increases the entropy by 70%. Visualizing the raw trajectories enables us to see that this intermediate destination is not on the most popular paths between the source and destination. In the next section, we will show empirically that the intermediate locations that increase trajectory entropy are likely to be intermediate destinations much more than the other locations.

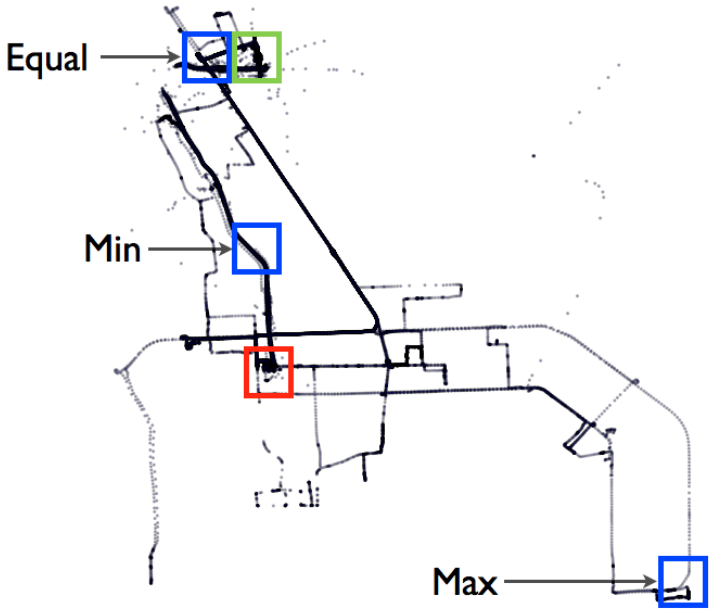


Figure 7.5 – It is not surprising that revealing an intermediate location that is adjacent to the starting location has a small effect on the trajectory distribution and decreases the entropy by only 2%. However, revealing an intermediate location in an area along one of the two main paths decreases the trajectory entropy to 73% of its initial value, which significantly increases trajectory uncertainty. Moreover, revealing the location in the lower-right corner changes completely the distribution of trajectories and maximizes conditional entropy.

7.4 Conditional Entropy and Trajectory Segmentation

Revealing some intermediate locations along a trajectory increases entropy because the conditioning of trajectories to pass through these locations drastically changes the distribution of trajectories. In a sense, these intermediate locations are outliers given the prior distribution of trajectories. In this section, we take advantage of this observation in order to develop a method that infers the set of waypoints given a trajectory. We formulate it as an optimization problem, for which we classify every vertex on a trajectory as either a waypoint or an intermediate point. First, we find an empirical connection between the class of a vertex on a trajectory (waypoint or intermediate point) and the conditional entropy of that vertex. More specifically, we show that, through an extensive analysis of real mobility traces, waypoints are those with a high ratio $H_{sd|u}/H_{sd}$. It is remarkable that waypoints can be found from trajectories in the mobility graph G alone

without requiring any timing information nor any absolute geographic locations.

Next, we build on this finding and develop a segmentation algorithm, based on the conditional trajectory entropy, that infers the likely waypoints for a given trajectory. We evaluate this algorithm over the Geolife dataset introduced in Section 7.2: We show that the points corresponding to high conditional entropy tend to be those with high residence time, which is much more likely for a waypoint than an intermediate point. The entropy-based heuristic used by our algorithm outperforms alternative approaches, as it is 43% more accurate than a geometric approach and 20% more accurate than path stretch based approach. Moreover, it is computationally efficient for online segmentations of trajectories, given that offline computations of conditional entropies are performed only once.

7.4.1 Mobility Model

After the pre-processing phase described in Section 7.2, we construct a weighted graph $G(V, E)$ whose vertices represent geographical areas and edges *direct* —no loss of GPS signal between two observations —transitions. As we are interested in actual transitions between areas, we exclude jumps that are due to a loss of GPS signal, as well as self-transitions that would only reflect repeated location samples within the same area. As a result, the weight of an edge $(i, j) \in E$ is equal to the number of direct transitions from area i to area j.

We infer, using a maximum likelihood estimator, the first order MC that has generated the observed data. The training set $\mathcal{T}_{\text{train}}$ contains the trajectories of all users and the resulting MC is therefore a low order mobility model that captures the population mobility pattern. We choose a low order MC because the training data available is too sparse to train a higher order MC, similar to the one we present in Section 7.3.1. In fact, for the same predictive accuracy, the number of samples needed to train a MC increases exponentially with the order of the MC. Having fewer samples than the number needed for a correct training of a high order MC leads to severe over-fitting. Having constructed the MC that captures the patterns of population mobility, we can analyze the link between trajectory entropy and waypoints.

7.4.2 Waypoints increase trajectory entropy

As discussed in Section 7.3, conditional trajectory entropy can be larger than unconditional trajectory entropy. A plausible explanation for such a Bayesian surprise —the posterior distribution of trajectory is very different from the prior distribution—is that the location revealed is not simply an intermediate location: It is a waypoint in itself. To explore this hypothesis, we conduct the following experiment: we observe the mobility of a user whose trajectory t_{sd} starts at location s and ends at location d. Suppose that this user

has a waypoint u along his trajectory. As this waypoint is more important than the intermediate locations that lead to it, the time the user would spend at this waypoint u should presumably be larger than the average time spent at the other intermediate locations. If the locations that increase the entropy are more likely to be waypoints, the average time spent by the users at these locations —a proxy for their importance —should be larger than the average time spent at other intermediate locations.

If our hypothesis is true, we would be able to quantify more accurately the importance of a location, even when the location records are not associated with time steps; observing the evolution of trajectory entropy would enable us to detect these important waypoints.

To test this hypothesis, we conduct the following experiment: we associate with each trajectory t_{sd} the set of locations $\mathcal{U}_\alpha(t_{sd})$ that depends on the parameter $\alpha \in [0, 1]$, and is defined as

$$\mathcal{U}_\alpha(t_{sd}) = \{u \in t_{sd} | H_{sd|u} > \alpha H_{sd}\}. \tag{7.3}$$

This set contains the intermediate locations u whose conditional entropy $H_{sd|u}$ is larger than αH_{sd}. For $\alpha = 0$, the set $\mathcal{U}_\alpha(t_{sd})$ is equal to the trajectory t_{sd}, whereas for $\alpha = 1$, the set $\mathcal{U}_\alpha(t_{sd})$ is the set of all locations in t_{sd} that strictly increase the trajectory entropy.

We introduce the continuous random variable $R(u)$ that represents the residence time at location u. We are interested in analyzing the evolution of the expected residence time at a location, as a function of the change of trajectory entropy if we reveal this location. More formally, we analyze the evolution of

$$\mu_\alpha = \mathbb{E}[R(u) | u \in \mathcal{U}_\alpha(T_{sd})] \tag{7.4}$$

as we increase the value of the parameter α.

We approximate the quantity (7.4) by its empirical average

$$\hat{\mu}_\alpha = \frac{\sum_{t_{sd} \in \mathcal{T}_{\text{train}}} \sum_{u \in t_{sd}} r(u, t_{sd}) \mathbb{1}_{u \in \mathcal{U}_\alpha(t_{sd})}}{\sum_{t_{sd} \in \mathcal{T}_{\text{train}}} \sum_{u \in t_{sd}} \mathbb{1}_{u \in \mathcal{U}_\alpha(t_{sd})}}, \tag{7.5}$$

where $r(u, t_{sd})$ is the residence time at location u along the trajectory t_{sd}.

Figure 7.6 illustrates the evolution of $\hat{\mu}_\alpha$ as a function of α. As we increase the value of α, we become increasingly restrictive and consider only the intermediate locations that satisfy the inequality $H_{sd|u} > \alpha H_{sd}$. We notice that the average time $\hat{\mu}_\alpha$ increases with α. More importantly, as soon as we consider only the points that increase the trajectory entropy, we observe a sharp transition in the value of $\hat{\mu}_\alpha$. In fact, the average time a user spends at a location is less than 4 minutes, whereas the average time spent at locations that increase the trajectory entropy ($H_{sd|u} > 1.3\, H_{sd}$) is larger than 8 minutes.

We take a step further and study the evolution of the distribution of the conditional

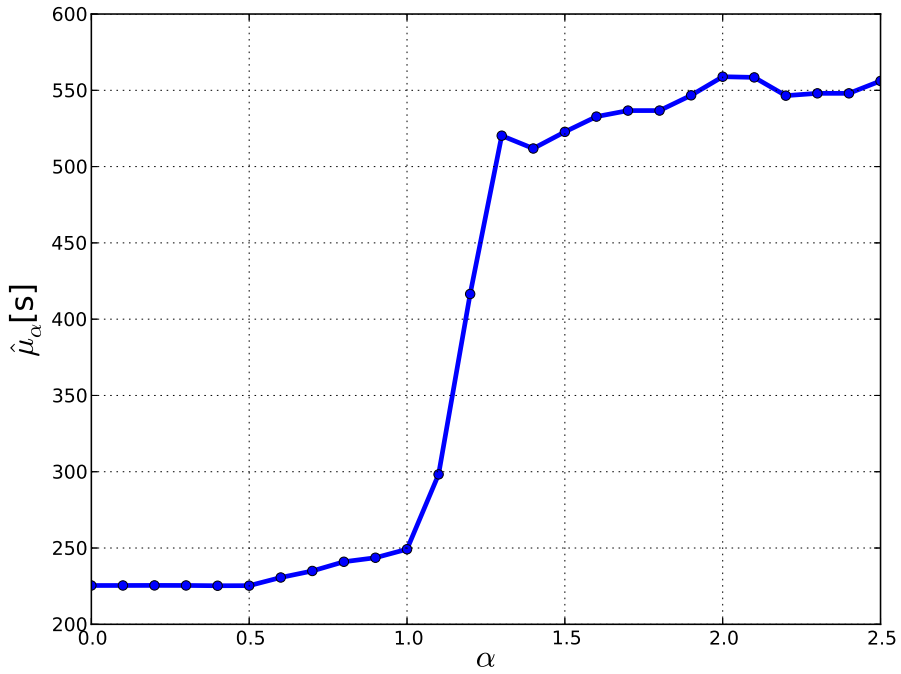


Figure 7.6 – Evolution of the average time spent at a location $\hat{\mu}_\alpha$ as a function of the entropy ratio α. We observe a sharp transition in the value of $\hat{\mu}_\alpha$ as we consider only the points that increase trajectory entropy. This strongly supports our hypothesis stating that locations that increase trajectory entropy are more likely to be waypoints.

residence time $R(u)|u \in \mathcal{U}_\alpha(T_{sd})$ for different values of α: In Figure 7.7, we show the evolution of the empirical distribution of the conditional residence time as a function of α. Each column of the matrix is normalized count over intervals of residence time for fixed values of α: each square is colored according to the density of values within the corresponding interval. Note that, for sake of readability, we truncate the residence time values so that the interval $[900, 1000]$ includes the values that are larger than 1000. We observe a change in the distribution of residence time that explains the behavior of the average residence time shown in Figure 7.6. Indeed, the probability of observing a large residence time increases as soon as we consider only the points that increase trajectory entropy. Naturally, the probability to observe intermediate locations with low residence time is still the highest because (a) the majority of the trajectories don't admit intermediate locations with high residence time, and (b) even if a trajectory admits an intermediate location with high residence time, the locations adjacent to it will have a similar conditional entropy but a much lower residence time. In Section 7.4.4, we will provide empirical evidence about these claims using an method that detects, for a given trajectory, intermediate waypoints.

To explore further this direction, we compare the distribution of the residence time $R(u)$ at locations that decrease entropy or leave it unchanged ($H_{sd|u} \leq H_{sd}$) with the distribution of residence time at locations that increase it ($H_{sd|u} > H_{sd}$). Figure 7.8 shows

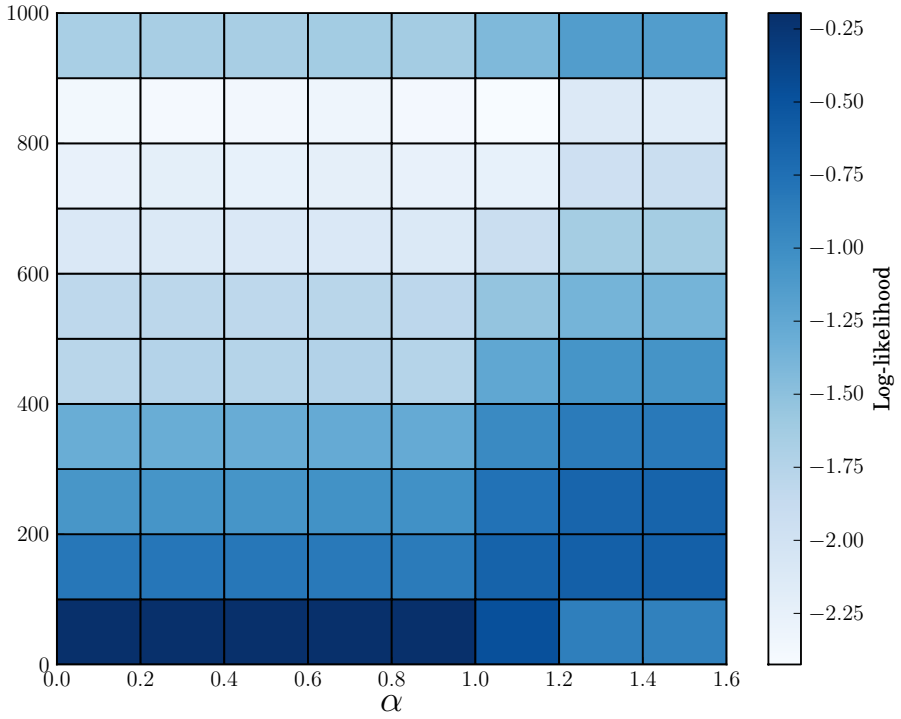


Figure 7.7 – Empirical distribution of the conditional residence time $R(u)|u \in \mathcal{U}_\alpha(T_{sd})$ as function of the value of α. Each column of the matrix is a normalized count over intervals of residence time for a fixed values of α, and each square is colored according to the logarithm of the normalized density of values within the corresponding interval.

two CCDFs (complementary cumulative distribution functions) of the residence time $R(u)$ for both situations. We clearly observe that the CCDFs have the same evolution for low values of residence time but they diverge starting at $r = 8$ minutes. Above this value, observing a large residence time is much more likely in a location that increases the entropy than in a location that decreases it. For example, observing a user that spends more than 30 minutes at a location is 10 times more likely at a location that increases the entropy than at a location that decreases it.

Taken all together, our results strongly support our hypothesis stating that locations that increase trajectory entropy are more likely to be waypoints where a user might spend more time. Furthermore, these results imply that, using a low order mobility model that is based on the patterns of population mobility, we are able —with no time information —to segment individual trajectories by detecting waypoints.

In the next section, we build on these findings and propose an algorithm that uses a entropy-based heuristic in order to automatically segment a given trajectory.

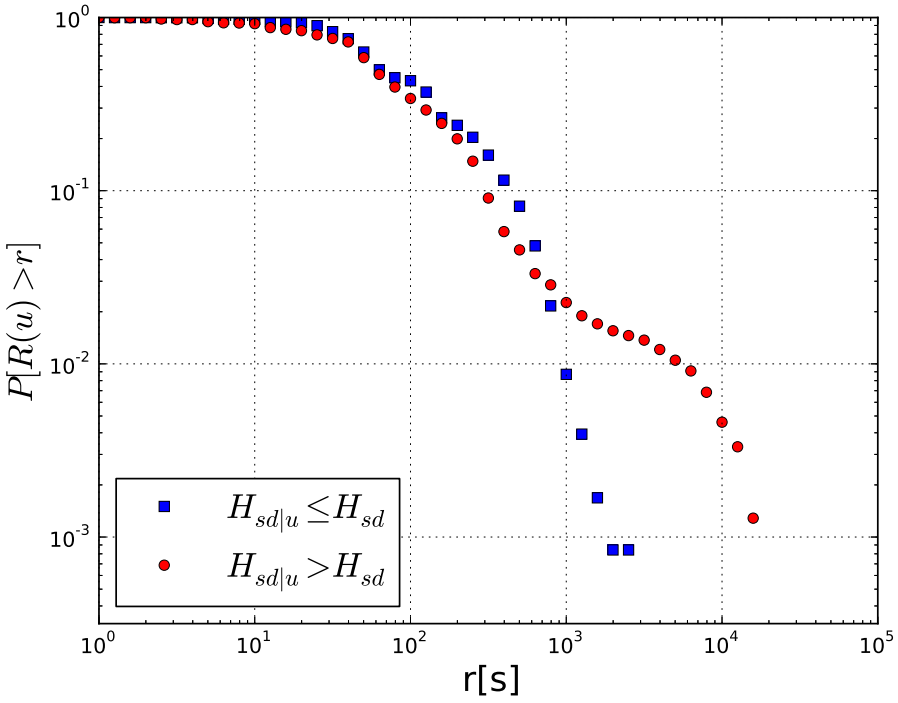


Figure 7.8 – Log-log plot of the complementary cumulative distribution function (CCDF) of the residence time R.

7.4.3 Trajectory segmentation algorithm

We propose a recursive algorithm (pseudo-code in Algorithm 3) that segments a trajectory by finding intermediate locations that increase the conditional entropy. The input of the algorithm is a trajectory t_{sd}, a MC transition probability matrix P and the *sensitivity* parameter $\alpha > 0$. The algorithm recursively segments the trajectory t_{sd} by finding the intermediate point u that maximizes the conditional entropy $H_{sd|u}$ (line 21). If this conditional entropy $H_{sd|u}$ is larger than αH_{sd}, the point u is added to the sequence of waypoints U. Note that the sensitivity parameter α controls to which extent the segmentation process is conservative: The higher alpha is, the more selective the algorithm is at declaring a point as a waypoint. If a waypoint u is chosen, the segmentation algorithm continues by applying the same procedure to the two sub-trajectories t_{su} and t_{ud}. Our algorithm bears similarity with the Ramer-Douglas-Peuker algorithm [57] that is used for polygonal approximation of plane curves. The conditional trajectory entropy in our algorithm is analogous to Euclidean distance between the original curve and the simplified curve in [57].

Complexity We study the average case complexity of our algorithm for a N states MC and a trajectory of length l. The expected number of nested calls is upper bounded by $\log l$: in the most balanced situation, we divide the trajectory into two sub-trajectories

Algorithm 3: Trajectory segmentation

Input: trajectory $traj$, transition probabilities matrix P, sensitivity α
Output: indices of waypoints U

```
1  begin
2      U ⟵ ∅                                  // global variable
3      if len(traj) > 2 then
4          segment(traj, 0, len(traj) − 1)
5      end
6      return U
7  end
8  Function segment(traj, i, j)
9      k ⟵ partition(traj, i, j)
10     if k ≥ 0 then
11         U ⟵ U ∪ {k}
12         if i + 1 < k then
13             segment(traj, i, k)
14         end
15         if k + 1 < j then
16             segment(traj, k, j)
17         end
18     end
19 Function partition(traj, i, j)
20     s ← traj[i], d ← traj[j]
21     k ⟵ arg max_{i<k<j} H_{sd|traj[k]}      // finding the element that maximizes
       conditional entropy
22     u ⟵ traj[k]
23     if H_{sd|u} > αH_{sd} then
24         return k
25     else
26         return −1
27     end
```

with approximately the same length. Typically, the number of nested calls is much lower than $\log l$ because the number of waypoints in a trajectory of length l is much lower than l. For each call, we compute the conditional entropy for $\mathcal{O}(l)$ candidates; this requires the computation of the fundamental matrix introduced in Section 6.3.1. Computing the fundamental matrix has a $\mathcal{O}(N^3)$ complexity because of the inversion of a matrix of size $\mathcal{O}(N)$. However, for a given MC, we can pre-compute the conditional entropies $H_{sd|u}$ offline and then use these results in order to segment all the trajectories on this MC. In such a situation, the expected time complexity of segmenting a trajectory of length l is $\mathcal{O}(l \log l)$ which enables for very efficient online segmentation of trajectories.

Another interesting direction we explore is the approximation of the conditional entropy

$H_{sd|u}$ by the sum of entropies $H_{su} + H_{ud}$. Such an approximation would reduce the complexity of computing the conditional entropies $H_{sd|u}$ because we would be able to use the matrix of trajectory entropies H —$\mathcal{O}\left(N^3\right)$ complexity to compute the entropy between all pairs $s, d \in V^2$ —to approximate the conditional entropy $H_{sd|u}$. In fact, we prove in Proposition 3 that

$$\mathrm{KL}(T_{sd}|T_{sd} \in \mathcal{T}_{sd}^{u}||(T_{su}, T_{ud})) = -\log p\left(T_{su} \notin \mathcal{T}_{su}^{d}\right),$$

which implies that the larger the probability $p\left(T_{su} \notin \mathcal{T}_{su}^{d}\right)$ is, the closer the distributions of the random trajectories $T_{sd}|T_{sd} \in \mathcal{T}_{sd}^{u}$ and (T_{su}, T_{ud}) are, and hence the closer the quantities $H_{sd|u}$ and $H_{su} + H_{ud}$ are. Moreover, in a typical mobility graph, the probability $p\left(T_{su} \notin \mathcal{T}_{su}^{d}\right)$ should be very low because going through the final location d in order to reach an intermediate location u is very unlikely. If we use this approximation, the pre-computation, which is performed once, has a complexity $\mathcal{O}\left(N^3\right)$, and the expected time complexity of segmenting a trajectory of length l becomes $\mathcal{O}\left(l \log l\right)$. For future work, we plan to work on this approximation and provide bounds for the difference between the $H_{sd|u}$ by the sum of entropies $H_{su} + H_{ud}$.

7.4.4 Experimental evaluation

In this section, we apply the entropy-based segmentation to the GPS trajectories of the Geolife project and show that it is able to accurately uncover waypoints along a trajectory, without having access to time information.

Detecting waypoints As we do not have data that classifies intermediate locations along a trajectory as waypoints, we use the available time information in order to detect potential waypoints. This bears similarity with the approach taken by Zheng et al. [74] who analyze the same dataset and classify a location as a stay point if an individual stays within an area around it for more than 20 minutes. We take this idea further and improve it by comparing the individual behavior to the collective behavior: We assume that a location visited by a user along his trajectory is likely to be a waypoint if this user spends *significantly* more time at this location than the other users typically do. More formally, we associate with each location x a Gaussian distribution of residence time $\mathcal{N}\left(\mu_x, \sigma_x\right)$, whose parameters are learnt from behavior of the whole observed population. For a user moving along a given trajectory t_{sd}, an intermediate location u is considered to be an intermediate destination if the time this user spends at u is classified as an outlier by the Chauvenet's criterion [72] applied on the distribution $\mathcal{N}\left(\mu_x, \sigma_x\right)$. This criterion states that, given a dataset of n observations produced by a Gaussian distribution, we consider a data point as an outlier only if the probability of observing its deviation from the mean is less than $\frac{1}{2n}$. In order to check whether our results are consistent independently of the choice of outlier detection method, we tested different outlier detection methods and

obtained consistent results. We denote by $\mathcal{W}(t_{sd})$ the set of waypoints associated with the trajectory t_{sd}.

We apply this waypoint-detection procedure to the Geolife GPS trajectories and observe that the majority (more than 87%) of the trajectories has an empty waypoints set. Among the trajectories that admit at least one waypoint, the clear majority (around 90%) has only one waypoint. We will therefore focus on assessing the performance of different segmentation methods on finding, for a given trajectory t_{sd} (a) whether the trajectory admits waypoints, and (b) if yes, finding the waypoint where the user spends most of her time

$$w = \arg\max_{u \in \mathcal{W}(t_{sd})} r(t_{sd}, u).$$

Baseline methods In order to assess the performance of our approach at trajectory segmentation, we consider different baseline methods that rely on different heuristics for trajectory segmentation. As the challenge is to uncover waypoints with no information about time, all the methods presented here share the fact that their heuristics are based on the structure of the trajectory only. Each method first constructs a set of candidate waypoints $\hat{\mathcal{W}}(t_{sd})$, and then chooses the waypoint $\hat{w}$ that maximizes a given heuristic (e.g., the conditional entropy in line 21 of our segmentation algorithm). The baseline methods are as follows:

Random (R) This method assumes that each trajectory admits waypoints and selects uniformly at random one of the points of the trajectory t_{sd}.

Geo Stretch (GS) The set of candidate waypoints $\hat{\mathcal{W}}(t_{sd})$ is composed of the intermediates locations that are not along the direct line from s to d. The waypoint is the intermediate location that is the furthest from the segment with s and d as end points. This simple yet strong baseline is used in the very popular Ramer-Douglas-Peuker algorithm [57] to select the point on a trajectory that is the furthest from the approximating line segment between s and d.

Path Stretch (PS) We consider the weighted mobility graph introduced in Section 7.4.1. The weight associated with an edge (i, j) is equal to $-\log(P_{ij})$ where P_{ij} is the probability of visiting location j given that we are at location i. These weights favor the transitions that are the most frequently observed. A simple computation gives that the cost of a trajectory t_{sd} is equal to $-\log p\,(t_{sd})$, which implies that the more probable a path is, the less costly it is. The set of candidate waypoints $\hat{\mathcal{W}}(t_{sd})$ is composed of the intermediate locations that are not along the shortest path from s to d. The waypoint is the candidate location $\hat{w}$ that maximizes the path cost.

Entropy (E) The set of candidate waypoints $\hat{\mathcal{W}}(t_{sd})$ is composed of the intermediate locations whose conditional entropy $H_{sd|u}$ is larger than the trajectory entropy H_{sd}.

The waypoint $\hat{w}$ is the candidate location u that maximizes conditional entropy

$$\hat{w} = \arg\max_{u \in \hat{\mathcal{W}}(t_{sd})} H_{sd|u}.$$

Empirical evaluation In order to evaluate the performance of the different segmentation methods, we repeat the following process 100 times: we divide randomly the dataset of trajectories in a training set (90 % of the data) and a test set (10 % of the data). Then, we train a Markovian mobility model based on the trajectories of the training set, and we apply the different segmentation methods to the trajectories of the test set. As we do not have access to a ground truth about waypoints, we consider the results produced by the time-based classification procedure, introduced in the beginning of Section 7.4.4, as target values. We assess the performance of each segmentation method by measuring (*a*) its average classification accuracy and the average F_1-score (harmonic mean of precision and recall), (*b*) the average residence time $r(t_{sd}, \hat{w})$ at the location classified as waypoint, and (*c*) the average distance (number of hops) between the waypoint guess $\hat{w}$ and the actual waypoint w.

	Residence time (std) [s]	Distance (std) [hops]
R	209 (73)	6.1 (1.2)
GS	570 (121)	2.5 (0.6)
PS	700 (114)	1.76 (0.27)
E	**1151** (150)	**1.41** (0.28)

Table 7.1 – The average performance of the entropy based segmentation compared to baseline methods.

	Accuracy	F_1 score
R	0.1	0.16
GS	0.12	0.18
PS	0.47	0.25
E	**0.7**	**0.60**

Table 7.2 – Average classification accuracy and average F_1 score of the entropy based segmentation compared to baseline methods.

We report the results, obtained by averaging the results of the process presented above, in Tables 7.1 and 7.2. The methods whose heuristics are based on the statistics related to the population mobility (PS and E) perform better than purely geographical heuristics (GS). This is not surprising as heuristics that are based on geographical distances fail

to capture paths that are geographically costly but very popular (i.e., a long route that includes many points of interests is more popular than a short route that has none).

Among the methods that are based on the mobility graph, the entropy-based segmentation is clearly the best: It takes advantage of the entropy-based heuristic that describes the evolution of whole the distribution of trajectories, as opposed to PS that is based on the evolution of path probability only. This also confirms that trajectory entropy captures much more than simply the evolution of the cost of the shortest path.

By looking at Table 7.1, we see that the average residence time at the locations classified as way-points by our method is larger than the residence time at the locations classified as such by the baseline methods (102% larger than GS, 64% larger than PS). This indicates clearly that the entropy based segmentation is the best at retrieving locations where users spend a significant amount of time. More importantly, the average distance between our method's guess and the actual waypoint is 1.4 on average, which is a 43% improvement over the GS segmentation and a 20% improvement over the PS segmentation. We can see this average distance as a measure of the waypoints privacy [62], which implies that the privacy of waypoints is the lowest when the adversary's estimate of the waypoint is based on trajectory entropy.

Table 7.2 shows the average accuracy and F_1 score of each method. The fact that R and GS perform poorly is not surprising if we know that the majority of the trajectories in our dataset admits no waypoints: the random segmentation method declares systematically that a trajectory admits a waypoint, and GS declares the same as soon as the trajectory t_{sd} deviates from the direct line from s to d. Our method is the most accurate —48% more accurate than PS —and is able to classify correctly an important proportion (75%) of the trajectories that have a waypoint. Moreover, it offers the best trade-off between precision and recall, with a F_1 score equal to 0.60. Note that, for all methods, the precision is lower than the recall: an intermediate location that deviates significantly from the most probable path, or increases trajectory entropy does not always imply, with certainty, that this location is a waypoint.

Taken together, these results strongly support the possibility of uncovering waypoints of a trajectory without having access to time information: computing the conditional trajectory entropy associated with a parsimonious mobility model enables us to detect structural outliers that are very likely to be waypoints and not simple intermediate locations.

7.5 Conclusion

In this chapter, we present two mobility-related applications for which we use tools borrowed from the trajectory entropy framework. First, we quantify mobility uncertainty

and its evolution with location updates and link the evolution of conditional entropy to the nature of the intermediate locations revealed. Second, we propose a trajectory segmentation method based on the computation of the entropy of conditional Markov trajectories. We showed empirically that the entropy of a trajectory conditioned on a particular location is a powerful metric to estimate whether this location is likely to be a waypoint or not, and more generally to reveal whether knowing this location makes the trajectory more or less predictable. Based on this observation, we develop an algorithm that is able to efficiently segment trajectories: we take advantage of a model of population mobility and quantify to which extent this individual trajectory deviates from the plausible behaviors. The advantage of our approach is that little information about the individual trajectory is needed. In particular, no timestamps nor absolute geographic locations are used. More generally, we believe the entropy of conditional trajectories is a powerful tool to study dynamics on graphs because it is able to capture the evolution of the distribution of trajectories as intermediate locations are revealed.

The Data Mining Perspective Part III

Introduction

In the first and second parts of this thesis, we considered mobility as a simple time-stamped sequence of locations visited. However, mobility, as experienced by an individual, is a much more than this; it is an experience that is influenced by both personal (e.g., interests, mood, environment perception) and environmental factors (e.g., roads, noise, aestheticism). For example, visiting a city cannot be reduced to visiting some particular venues within it, but includes the atmosphere and ambiance of the streets, the interactions with people and the architecture of the buildings.

With the democratization of GPS-enabled smartphones and the rapid development of location-based services, we have witnessed an increasing availability of geo-tagged digital traces that describe the user experience as she moves. These digital breadcrumbs take different forms (e.g., phone calls, status updates, geo-tagged photos or check-ins), thus capturing different facets of the user's experience: Photo-sharing websites, such as Flickr and Instagram, contain photos that link geographical locations with user-generated annotations; and social networks, such as Facebook and Yelp, associate check-ins with crowd-sourced reviews that describe the local experience. In Figure 7.9, we show an example that illustrates the abundance of geo-tagged data: We obtain a sketch of the world map by simply plotting the locations of geo-tagged tweets of a subset of Twitter users over a four-month period only.

With this abundance of geo-tagged data comes the promise of a better understanding of users' experiences and of an increased ability to describe the environment where they take place. In this chapter, we study human mobility from the data-mining perspective and postulate that mining geo-tagged data enables us to gain more insight into the experience that surrounds mobility. We illustrate this with our work about characterizing geographical regions: We show how mining millions of geo-tagged photos enables us to uncover terms that are specifically descriptive of region within a geographical hierarchy.

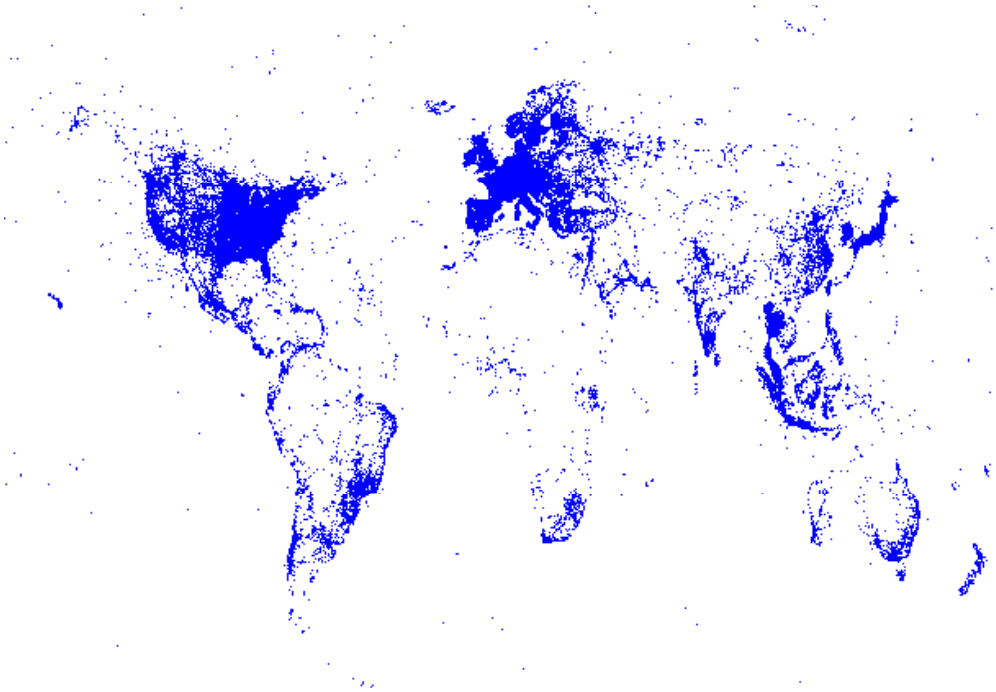


Figure 7.9 – We plot 22, 506, 721 geo-tagged tweets posted on Twitter by 225, 098 users over a four-month period. The fact that the location cloud draws the world map silhouette is a good illustration of the abundance of geo-tagged data

8 Describing the Characteristics of Geographical Regions

8.1 Introduction

Finding characteristics that are specific to a geographic region is challenging because it requires local knowledge to identify what is particularly salient in that area. Knowledge of regional characteristics becomes critical when communicating about or describing regions, for instance in the context of a mobile travel application that provides country, city and neighborhood summaries. This knowledge is especially useful when comparing regions to make geo-based recommendations: tourists visiting Singapore, for example, might be interested in exploring the Tiong Bharu neighborhood if they are aware that it is known for its coffee in the same way the San Francisco Mission district is.

Photo-sharing websites, such as Instagram and Flickr, contain photos that connect geographical locations with user-generated annotations. We postulate that local knowledge can be gleaned from these geo-tagged photos, enabling us to discriminate between annotations (i.e., tags) that specifically characterize a region (e.g., neighborhood) and those that characterize surrounding areas or more general themes. This is, however, challenging because much of the data produced within a region is not necessarily specifically descriptive for that area. Is the word "desert" specifically descriptive of Las Vegas, or rather of the surrounding area? Can we quantify to what extent the word "skyscraper" is descriptive of Midtown, Manhattan, New York City or the United States as a whole?

In this chapter, we propose the geographical hierarchy model (GHM), a probabilistic hierarchical model that enables us to find terms that are specifically descriptive of a region within a given hierarchy. The model is based on the assumption that the data observed in a region is a random mixture of terms generated by different levels of the hierarchy. Our model further gives insight into the diversity of local content, for example by allowing for identification of the most unique or most generic regions amongst all regions in the hierarchy.

To investigate how well the descriptive terms surfaced by our model for a given region correspond with human descriptions, we focus on annotations of photos taken in neighborhoods of San Francisco and New York City. We apply our method to a dataset of 8 million geo-tagged photos described by approximately 20 million tags. We are able to associate each neighborhood with the tags that describe it specifically, and coefficients that quantify its uniqueness. This enables us to find the most unique neighborhoods in a city and to find mappings between similar neighborhoods in both cities.

We contrast the neighborhood characteristics uncovered by our model with their human descriptions by conducting a survey and a user study. This allows us not only to assess the quality of the results found by the GHM, but mainly to understand the human reasoning about what makes a feature distinctive (or not) for a region. Beyond highlighting individual differences in people's local experiences and perceptions, we touch topics such as the importance of supporting feature understanding, and consideration of adjacency and topography through porous boundaries.

8.2 Related Work

Human activities shape the perception of coherent neighborhoods, cities and regions; and vice versa. As Hillier and Vaughan [35] pointed out, urban spatial patterns shape social patterns and activity, but in turn can also reflect them. Classic studies such as those by Milgram and Lynch [52] on mental maps of cities and neighborhoods reflect that people's perceptions go well beyond spatial qualities. They illustrate individual differences between people, but also the effects of social processes affecting individual descriptions. People's conceptualization of region boundaries are fuzzy and can differ between individuals, even while they are still willing to make a judgement call on what does or does not belong to a geographic region [55]. Urban design is argued to affect the imageability of locales, with some points of interests for example may be well known, whereas other areas in a city provide less imageability and are not even recalled by local residents [52]. The perceived identity of a neighborhood differs between individuals, but can relate to their social identity and can influence behavior [66]. Various qualitative and quantitative methods, from surveys of the public to trained observation, have been employed over the years to gain such insights [25, 55, 66], with the fairly recent addition of much larger datasets of volunteered geographic information [27] and other community-generated content to benefit our understanding of human behavior and environment in specific regions.

We differentiate our approach from those aiming to discover new regions [69] or redefine neighborhood boundaries [22] using geo-tagged data. Backstrom et al. [12] presented a model to estimate the geographical center of a search engine query and its spatial dispersion, where the dispersion measure was used to distinguish between local and global queries. Ahern et al. [9] used k-means clustering to separate geo-tagged photos

into geographic regions, where each region was described by the most representative tags. This work was followed up by Rattenbury and Naaman [58], in which the authors proposed new methods for extracting place semantics for tags. While these prior works principally attempted to identify new regions or arbitrary spaces, we instead aim to find the unique characteristics of regions in a known hierarchy. For each region at each level in the hierarchy we aim to find a specifically descriptive set of tags, drawn from the unstructured vocabulary of community-generated annotations.

Hollenstein and Purves [36] examined Flickr images to explore the terms used to describe city core areas. For example, they found that terms such as `downtown` are prominent in North America, `cbd` (central business district) is popular in Australia and `citycenter` is typical for Europe. The authors further manually categorized tags in the city of Zürich according to a geographical hierarchy; even with local knowledge they found that doing so was tedious because of tag idiosyncrasies and the diversity of languages. In contrast, our method enables us to automatically obtain such classifications. Moreover, we are not just able to distinguish between local and global content, but can classify a tag according to a geographical hierarchy with an arbitrary number of levels.

Hierarchical mixture models can be used for building complex probability distributions. The underlying hierarchy, that encodes the specificity or generality of the data, might be known a priori [53] or may be inferred from data by assuming a specific generative process, such as the Chinese restaurant process [14] for Latent Dirichlet Allocation [10,15], where the regions and their hierarchy are learned from data. We emphasize that, in this chapter, we do not try to learn latent geographical regions, but rather aim to describe regions that are known a priori. Moreover, these regions are structured according to a known hierarchy.

8.3 Geographical Hierarchy Model

Finding terms that are specifically descriptive of a region is a challenging task. For example, the tag `paname` (nickname for Paris) is frequent across many neighborhoods of Paris but is specific to the city, rather than to any of the neighborhoods in which the photos labeled with this tag were taken. The tag `blackandwhite` is also a frequent tag, which describes a photography style rather than any particular region in the world where the tag was used. The main challenge we face is to discriminate between terms that (i) specifically describe a region, (ii) those that specifically describe a sibling, parent or child region, and (iii) those that are not descriptive for any particular region. To solve this problem, we introduce a hierarchical model that is able to isolate the terms that are specifically descriptive of a given region by considering not only the terms used within the region, but also those used elsewhere. Our model goes beyond the distinction between only global and local content [56] by probabilistically assigning terms to a particular level in the hierarchy. In this chapter we present our model from a geographic perspective,

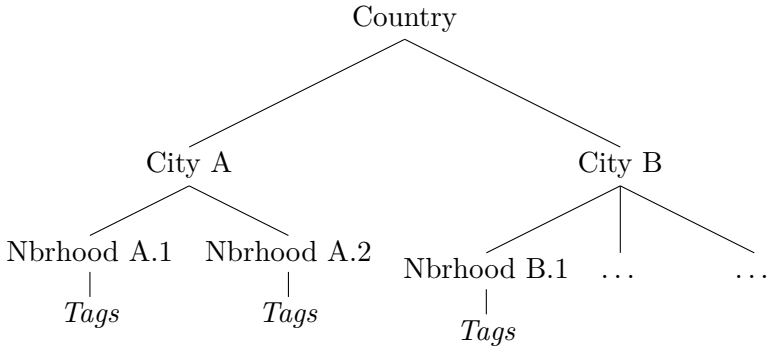


Figure 8.1 – We represent the geographical hierarchy Country → City → Neighborhood as a geo-tree. Each node v of the geo-tree is associated with a multinomial distribution over tags θ_v.

even though it is generic in nature: in principle any kind of hierarchy can be used where labeled instances are initially assigned to the leaf nodes.

8.3.1 Definitions

tag is the basic semantic unit and represents a term from a vocabulary indexed by $t \in \{1, \ldots, T\}$.

neighborhood is the basic spatial unit. The semantic representation of a neighborhood is based on the collection of tags associated with the geo-tagged photos taken within the neighborhood. Consequently, we associate with each neighborhood $n \in \{1, \ldots, N\}$, where N is the number of neighborhoods, the vector $\boldsymbol{x}_n \in \mathbb{N}^T$, where x_{nt} is the number of times tag t is observed in neighborhood n.

geo-tree is the tree that represents the geographical hierarchy. Each node v of the geo-tree is associated with a multinomial distribution θ_v such that $\theta_v(t)$ is the probability to sample the tag t from node v. We designate the set of nodes along the path from the leaf n to the root of the geo-tree as R_n, whose cardinality is $|R_n|$. We use the hierarchy Country → City → Neighborhood, illustrated in Figure 8.1, as the leading example throughout this chapter.

8.3.2 Model

The principle behind our model is that the tags observed in a given node are a mixture of tags specific to the node and of tags coming from different levels in the geographic hierarchy. We represent the tags as multinomial distributions associated with nodes that are along the path from leaf to root. This model enables us to determine the tags that are specifically descriptive of a given node, as well as to quantify their level of specificity

or generality. The higher a node in the tree, the more generic its associated tags, whereas the lower a node, the more specific its tags. The tags associated with a node are shared by all its descendants. We can formulate the random mixture of multinomial distributions

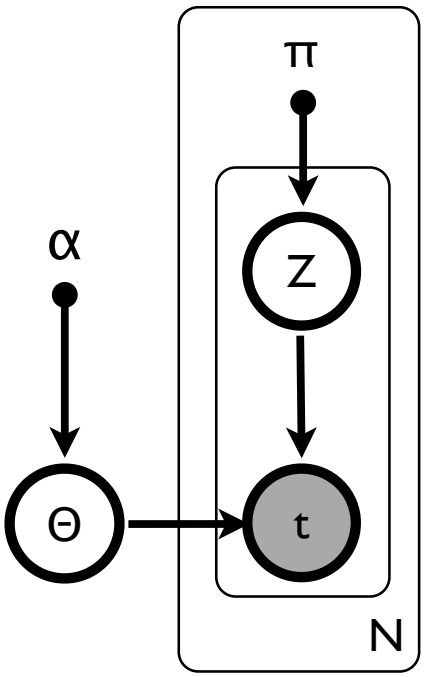


Figure 8.2 – Graphical model of the GHM. The outer plate represents the leaves (neighborhoods) of the geo-tree while the inner plates represent words (tags)

with respect to a latent (hidden) variable $z \in \{1, \ldots, |R_n|\}$ that indicates for each tag the level in the geo-tree from which it was sampled. For a tag t observed in neighborhood n, $z = 1$ means that this tag t was sampled from the root node corresponding to the most general distribution θ_{root}, whereas $z = |R_n|$ implies that the tag t was sampled from the most specific neighborhood distribution θ_n. This is equivalent to the following generative process for the tags of neighborhood n: (i) randomly select a node v from the path R_n with probability $p(v|n)$, and (ii) randomly select a tag t with probability $p(t|v)$. We suppose that tags in different neighborhoods are independent of each other, given the neighborhood in which they were observed. Consequently, we can write the probability of the tags $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ observed in the N different neighborhoods as

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = p(\boldsymbol{x}_1) \ldots p(\boldsymbol{x}_N). \tag{8.1}$$

We further assume tags are independent of each other given their neighborhood, so that we can write the probability of the vector of tags $\boldsymbol{x}_n$ as:

$$p(\boldsymbol{x}_n) = \prod_{t=1}^{T} p(t|n)^{x_{nt}}. \tag{8.2}$$

The probability of observing tag t in neighborhood n is then:

$$p(t|n) = \sum_{v \in R_n} p(t|v)\, p(v|n) = \sum_{v \in R_n} \theta_v(t)\, p(v|n), \tag{8.3}$$

which expresses the fact that the distribution of tags in neighborhood n is a random mixture over the multinomial distributions θ_v associated with the nodes along the path from the leaf n up to the root of the geo-tree. The random mixture coefficients are the probabilities $p(v|n)$. By combining (8.1), (8.2) and (8.3) we obtain the log-likelihood of the data:

$$\begin{aligned} \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) &= \log \prod_{n=1}^{N} \prod_{t=1}^{T} p(t|n)^{x_{nt}} \\ &= \sum_{n=1}^{N} \sum_{t=1}^{T} x_{nt} \log p(t|n) \\ &= \sum_{n=1}^{N} \sum_{t=1}^{T} x_{nt} \log \sum_{v \in R_n} \theta_v(t)\, p(v|n). \end{aligned} \tag{8.4}$$

Classification For a tag t observed in neighborhood n, we can compute the posterior probability that it was generated from a given level z of the geo-tree: we apply Bayes' rule to compute the posterior probability of the latent variable z

$$p(z|t, n) = \frac{p(t|n, z)p(z|n)}{\sum_{z=1}^{|R_n|} p(t|n, z)p(z|n)}. \tag{8.5}$$

Since we assume that the distribution of tags in neighborhood n is a random mixture over the distributions θ_v associated with nodes that forms the path R_n from leaf n up to the root of the geo-tree, the probability (8.5) is equal to

$$p(v'|t, n) = \frac{\theta_{v'}(t)\, p(v'|n)}{\sum_{v \in R_n} \theta_v(t)\, p(v|n)},$$

where v' is the node in R_n that is at level z. Classifying a tag t observed in leaf n amounts to choosing the node $v' \in R_n$ that maximizes the posterior probability $p(v'|t, n)$.

8.3.3 Learning

The parameters of our model are the multinomial distributions θ_v associated with each node v of our geo-tree and the mixture coefficients $p(v|n)$. In order to learn the model parameters that maximize the likelihood of data, given by (8.4), we use the Expectation-Maximization algorithm. This iterative algorithm increases the likelihood of the data by updating the model parameters in two steps: The E-steps and the M-steps. The algorithm converges to a solution as soon as the change in likelihood between two consecutive

iterations is smaller than a given threshold.

For the E-step, we evaluate, for each tag t and neighborhood n, the posterior probability

$$p(v|t,n) = \frac{p(v|n)\,p(t|v)}{\sum_{v' \in R_n} p(v'|n)\,p(t|v')}.$$

The probability $p(v|t,n) = 0$ for all $v \notin R_n$ because we assume that the tags observed in neighborhood n are generated from the nodes that form the path R_n in the geo-tree.

For the M-step, we re-estimate the parameters of the model using the posterior probabilities computed in the E-step. We update the multinomial distributions as follows:

$$\theta_v(t) = p(t|v) = \frac{\sum_{n=1}^{N} I_{\{v \in R_n\}}\, x_{nt}\, p(v|t,n)}{\sum_{t'=1}^{T} \sum_{n=1}^{N} I_{\{v \in R_n\}}\, x_{nt'}\, p(v|t',n)},$$

where $I_{\{A\}}$ is the indicator function that takes values 1 if event A is true and 0 otherwise. The numerator is the expected number of tags t generated by node v, and the denominator is the expected number of total tags generated by the same node. Both expectations are computed with respect to the posterior distribution of the latent variable $p(v|t,n)$.

We also update the mixture coefficients

$$p(v|n) = \frac{\sum_{t=1}^{T} x_{nt}\, p(v|t,n)}{\sum_{t=1}^{T} \sum_{v' \in R_n} x_{nt}\, p(v'|t,n)} = \frac{\sum_{t=1}^{T} x_{nt}\, p(v|t,n)}{\sum_{t=1}^{T} x_{nt}}.$$

The numerator is the expected number of tags generated by node v in neighborhood n, and the denominator is the total number of tags generated in neighborhood n.

If a node in the geo-tree contains a tag that was not observed in the training set, maximum likelihood estimates of the multinomial parameters would assign a probability of zero to such a tag. In order to assign a non-zero probability to every tag, we "smooth" the multinomial parameters: we assume that the distributions θ_v are drawn from a Dirichlet distribution. The Dirichlet distribution is a distribution of T-dimensional discrete distributions parameterized by a vector $\boldsymbol{\alpha}$ of positive reals. Its support is the closed standard $(T-1)$ simplex, and it has the advantage of being the conjugate prior of the multinomial distribution. In other words, if the prior of a multinomial distribution is the Dirichlet distribution, the inferred distribution is a random variable distributed also as a Dirichlet conditioned on the observed tags. In order to avoid favoring one component over the others, we choose the symmetric Dirichlet distribution as a prior. We also assume a symmetric Dirichlet prior for the mixture coefficients $p(v|n)$.

Complexity Learning the parameters of our model using EM algorithm has, for each iteration, a worst case running-time complexity of $\mathcal{O}(NTD)$, where N is the number of leaves of the tree, T the vocabulary cardinality and D the tree depth. GHM has therefore an important strength, as the time-complexity of training GHM scales linearly with respect to the number of leaves of the tree and the number of *unique* tags rather than the number of tag instances. Furthermore, the number of iterations needed for EM to converge, when trained on the Flickr dataset introduced in Section 8.4.1, is typically around 10.

8.3.4 Geographical hierarchy model with adjacency

Geographical hierarchy model with adjacency (GHMA) is an extension of GHM that takes into account the porosity of the frontiers that separates adjacent neighborhoods. In fact, neighborhood boundaries are not set in stone thus we can observe tags that are specific to neighborhood n in a adjacent neighborhood n' (e.g., photo of a POI in neighborhood n' taken from a distant location in neighborhood n). Moreover, this addresses that there is not necessarily a long-term consensus about the exact boundaries of a neighborhood, as illustrated by the findings of our user study.

Let A_n be the set of neighborhoods adjacent to neighborhood n. With GHMA, the probability of observing tag t in neighborhood n is

$$\begin{aligned} p(t|n) &= \sum_{v \in R_n \cup A_n} \theta_v(t)\, p(v|n) \\ &= \sum_{v \in R_n} \theta_v(t)\, p(v|n) + \sum_{v \in A_n} \theta_v(t)\, p(v|n). \end{aligned} \tag{8.6}$$

Note that the term $\sum_{v \in A_n} \theta_v(t)\, p(v|n)$ in (8.6) is not present in the sum (8.3), and accounts for the possibility of sampling tags from nodes $v \in A_n$ representing adjacent neighborhoods. Using GHMA, we are able to quantify the porosity of the frontier between neighborhoods n' and n: We simply compute the probability $p(n'|n)$ of sampling tags from the local distribution of neighborhood n' given that we are at neighborhood n. A high probability $p(n'|n)$ indicates that neighborhood n' has a strong influence on neighborhood n because it is very likely to observe tags that are specific to neighborhood n' in neighborhood n.

8.4 Uncovering the Characteristics of Geographical Regions

In this section, we apply our model to a large collection of geo-tagged Flickr photos taken in neighborhoods of San Francisco and New York City. The quality of the descriptive tags found by the GHM strongly supports the validity of our approach, which we further confirm using the results of the user study in Section 8.5: we approximate the probability

that GHM classifies a tag "correctly" by the average number of times its classification matches the experts' classification. Moreover, the GHM allows us to quantify the uniqueness of these neighborhoods and to obtain a mapping between neighborhoods in different cities (Section 8.4.1), enabling us to answer questions such as "How unique is the Presidio neighborhood in San Francisco?" or "How similar is the Mission in San Francisco to East Village in New York City?". Finally, we compare the performance of our model with the performance of other methods in classifying data generated according to a given hierarchy (Section 8.4.2).

8.4.1 Dataset and classification

We now apply our model to a large dataset of user-generated content to surface those terms considered to be descriptive for different regions as seen through the eyes of the public.

Flickr dataset

Describing geographical areas necessitates a dataset that associates rich descriptors with locations. Flickr provides an ample collection of geo-tagged photos and their associated user-generated tags. We couple geo-tagged photos with neighborhood data from the city planning departments of San Francisco[1] and New York City[2], and focus on the neighborhoods of San Francisco (37 neighborhoods) and Manhattan (28 neighborhoods). Flickr associates an accuracy level with each geo-tagged photo that ranges from 1 (world level) to 16 (street level). In order to correctly map photos to neighborhoods, we only focus on photos whose accuracy level exceeds neighborhood level. We acquired a large sample of geo-tagged photos taken in San Francisco (4.4 million photos) and Manhattan (3.7 million photos) from the period 2006 to 2013. We preprocessed the tags associated with these photos by first filtering them using a basic stoplist of numbers, camera brands (e.g., Nikon, Canon), and application names (e.g Flickr, Instagram), and then by stemming them using the Lancaster[3] method. We further removed esoteric tags that were only used by a very small subset of people (less than 10). To limit the influence of prolific users we finally ensured no user can contribute the same tag to a neighborhood more than once a day. The resulting dataset contains around 20 million tags, of which 7,936 unique tags, where each tag instance is assigned to a neighborhood. These tags form the vocabulary we use for describing and comparing different regions. The geo-tree we build from this dataset has 3 levels: level one with a single root node that corresponds to the United States, level two composed of two nodes that correspond to San Francisco and Manhattan, and level three composed of 65 leaves that correspond to the neighborhoods of these cities.

[1] https://data.sfgov.org (Accessed 04/2015)
[2] http://nyc.gov (Accessed 04/2015)
[3] http://comp.lancs.ac.uk/computing/research/stemming/(Accessed 04/2015)

	Number
Geo-tagged photos	430,329,461
Geo-tagged photos in the US	128,238,326
Photos in Manhattan	4,477,655
Photos in San Francisco	3,711,194
Tags in San Francisco and Manhattan	16,892,259
Unique tags	7936
Neighborhoods in Manhattan	28
Neighborhoods in San Francisco	37

Table 8.1 – Flickr dataset statistics. The number of unique tags, after filtering and stemming, is 7963.

Mission	**GG Park**	**Battery Park**	**Midtown**
california ‡	california ‡	newyork ‡	newyork ‡
mission	**flower**	manhattan ‡	manhattan ‡
sf ‡	**park**	*usa* †	*usa* †
usa †	sf ‡	**wtc**	**midtown**
graffiti	*usa* †	**brooklyn**	**skyscraper**
art	**museum**	**downtown**	**timessquare**
mural	**tree**	**bridge**	*light* †
valencia	**ggpark**	gotham ‡	**moma**
food	**deyoung**	**memorial**	**broadway**
car ‡	architecture ‡	*light* †	**rockefeller**

Table 8.2 – For each neighborhood, we show the 10 most likely tags ranked according to their probability $p(t|n)$. We use the posterior probability $p(z|t,n)$ in order to assign each tag to the distribution that maximizes this posterior probability: country (†), city (‡) or neighborhood (**bold**).

Classifying tags

We applied our model to the Flickr dataset in order to find tags that are specifically descriptive of a region. We show the results for two neighborhoods in San Francisco—Mission and Golden Gate Park— and two neighborhoods in Manhattan—Battery Park and Midtown—in particular. In Table 8.2 we show the top 10 *most likely* tags for each of these four neighborhoods, where each tag is further classified to the *most likely* level that have generated it. Effectively, each tag observed in a neighborhood is ranked according to the probability $p(t|n)$. Then, we apply (8.5) to compute the posterior probability of the latent variable $p(z|t,n)$ and classify the tag accordingly. Recall that the value of the latent variable z represents the level of the geo-tree from which it was sampled. For example, a tag observed in neighborhood n is assigned to the neighborhood level if the most likely distribution from which it was sampled is the neighborhood distribution θ_n. We consider such a tag as being specifically descriptive of neighborhood n.

Despite the fact that the tag `california` is the most likely (frequent) tag in both Mission and Golden Gate Park, it is not assigned to the neighborhood distribution but rather to the city distribution (we presume, if we were to add a new State level to the geo-tree, that the tag `california` would most likely be assigned to it). This confirms the notion that the most frequent tag in a neighborhood does not necessarily describe it specifically. We see that `architecture` is applicable not only to the buildings in Golden Gate Park—notably the De Young—but that it is also is a descriptor for San Francisco in general. Our method is able to discriminate between frequent and specifically descriptive tags, whereas a naive approach would still consider a very frequent tag in a neighborhood as being descriptive. The same observation is valid for the tags `usa` and `light`, which are tags that are too general and therefore not specifically descriptive of a neighborhood. The tag `car` may seem misclassified at the city level in San Francisco, until one realizes the city is well known for its iconic cable cars. Considering that the tags `art`, `mural` and `graffiti` are classified as being specific to Mission is not surprising, because this neighborhood is famous for its street art scene. The most probable tags that are specifically descriptive of Midtown in Manhattan include popular commercial zones, such as Rockefeller Center, Times Square and Broadway, as well as the museum of modern art (MoMa). The tag `gotham`, one of the most observed tags in Battery Park, is assigned to city level, which is not unexpected given that Gotham is one of New York City's nicknames.

Neighborhood uniqueness

If you are interested in visiting the most unique neighborhoods in San Francisco, which ones would you choose? With our framework, we can quantify the uniqueness of neighborhood n by using the probability $p(z|n)$ of sampling local tags, where $z = |R_n|$. In fact, a high probability indicates that we sample often from the distribution of local tags θ_n, and can therefore be interpreted as an indicator of a more unique local character. We show a map of the city of San Francisco in Figure 8.3, where each neighborhood is colored proportionally to its local-mixture coefficient. The darker the color , the more unique the personality of the neighborhood. The four most unique neighborhoods in San Francisco are Golden Gate Park (0.70), Presidio (0.67), Lakeshore (0.65) and Mission (0.60), which is not surprising if you know these neighborhoods. The Golden Gate park is the largest urban park in San Francisco, famous for its museums, gardens, lakes, windmills and beaches. The Presidio is a national park and former military base known for its forests, scenic points overlooking the Golden Gate Bridge and the San Francisco Bay. Lakeshore is known for its beaches, the San Francisco zoo and also for San Francisco State University. The Mission is famous for its food, arts, graffiti, and festivals.

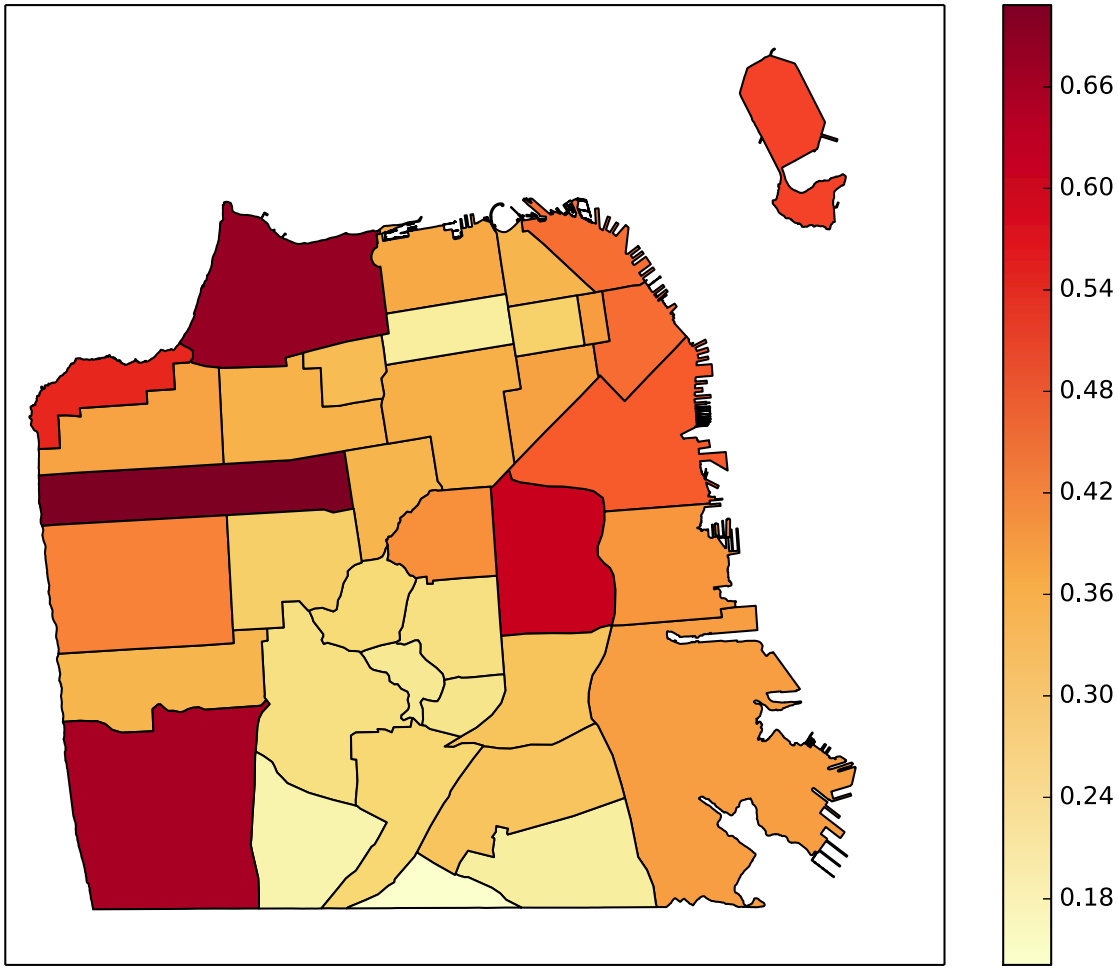


Figure 8.3 – Neighborhoods of San Francisco colored according to their local mixture coefficient $p(z|n)$, where $z = |R_n|$. A darker color indicates a larger local mixture coefficient ('uniqueness').

Mapping neighborhoods between cities

Given a neighborhood in San Francisco, what is the most similar neighborhood in Manhattan? To answer such a question, we can use our framework to find a mapping between similar neighborhoods that are in different cities and even different countries. Recall that each neighborhood n is described by its local distribution θ_n. In order to compare two neighborhoods n and n', we compute the cosine similarity between their respective local distributions θ_n and $\theta_{n'}$, given by:

$$\mathrm{sim}(\theta_n, \theta_{n'}) = \frac{\sum_{t=1}^{T} \theta_n(t)\, \theta_{n'}(t)}{\sqrt{\sum_{t=1}^{T} {\theta_n}^2(t)}\sqrt{\sum_{t=1}^{T} {\theta_{n'}}^2(t)}}. \tag{8.7}$$

The similarity range is $[0, 1]$, with $\text{sim}(\theta_n, \theta'_n) = 1$ if and only if $\theta_n = \theta'_n$. Table 8.3 shows the mapping from six San Francisco neighborhoods to the most similar neighborhoods in Manhattan respectively. We also include the second most similar neighborhood when the similarities are very close. In order to give some intuition about the mapping obtained, we also show the top five common local tags obtained by ranking the tags t according to the product of tag probabilities $\theta_n(t)\,\theta_{n'}(t)$. For example, East Village in Manhattan is mapped to Mission in San Francisco; the strongest characteristics they share are graffiti/murals, food, restaurants and bars. Moreover, despite the major differences between San Francisco and Manhattan, their Chinatowns are mapped to each other and exhibit a highly similar distribution of local tags (cosine similarity of 0.85). Finally, it is not surprising that Treasure Island for San Francisco and Roosevelt Island for Manhattan are mapped to each other, since both are small islands located close to each city. We however emphasize that the top *common* local tags between neighborhoods are not necessarily the most descriptive for the *individual* neighborhoods. The top tags may rather provide a shallow description (e.g., dragons in Chinatown, boats for island) and are useful to gain some insight into the mapping obtained, but we are aware that every Chinatown is unique and not interchangeable with another one.

San Francisco	Manhattan	Top common local tags
Mission	East Village (0.23)	graffiti, food, restaurant, mural, bar
Golden Gate Park	Washington Heights (0.26), Upper West Side (0.22)	park, museum, nature, flower, bird
Financial District	Battery Park (0.29), Midtown Manhattan (0.27)	downtown, building, skyscraper, city, street
Treasure Island	Roosevelt Island (0.38)	bridge, island, water, skylines, boat
Chinatown	Chinatown (0.85)	chinatown, chinese, downtown, dragons, lantern
Castro	West Village (0.06)	park, gay, halloween, pride, bar

Table 8.3 – Mapping from San Francisco neighborhoods to the most similar ones in Manhattan. For each neighborhood pair n and n', we give the cosine similarity between their local distributions θ_n and $\theta_{n'}$, and list the top "common" local tags ranked according to the product of tag probabilities $\theta_n(t)\,\theta_{n'}(t)$.

	Mission	Castro	North Beach
Tag count	32	32	32
Tag classications	561	381	349
GHM alignment	0.84	0.81	0.66
Misaligned tag count	5	6	11
Misaligned tags	night, coffee, sidewalk cat, brannan	mission, streetcar, dolorespark church, sign, night	embarcadero, bridge, water alcatraz, seal, sea, crab, wharf, boat, bay, pier

Table 8.4 – Alignment of GHM tag assignments with the majority of survey respondents' assignment for each survey neighborhood. Misalignment would for example be a tag classified as neighborhood level by the GHM while the majority of survey respondents had assigned it to another neighborhood or another level.

8.4.2 Experimental evaluation

Evaluating models such as the GHM on user-generated content is hard because of the absence of ground truth. There is no dataset that associates, with objectivity, regions with specifically descriptive terms, or assigns terms to levels in a geographic hierarchy. This is due to the intrinsic subjectivity and vagueness of the human conception of regions and their descriptions [36]. In the absence of ground truth, a classic approach is to generate a dataset with known ground truth, and then use it to evaluate the performance of different classifiers. We follow the generative process presented in Algorithm 4 using the geo-tree (3 levels, 68 nodes) built from the dataset presented in Section 8.4.1. For each node v in our geo-tree, we sample the distribution of tags θ_v from the symmetric Dirichlet distribution $Dir(\alpha)$. We set $\alpha = 0.1$ to favor sparse distributions θ_v. If the node is a leaf, i.e., it represents a neighborhood, we also sample the mixture coefficient $p(z|v)$ from a symmetric Dirichlet distribution $Dir(\beta)$. We set $\beta = 1.0$ to sample the mixture coefficients uniformly over the simplex and have well-balanced distributions. Now that we have the different distributions that describe the geo-tree, we can start generating the tags in each neighborhood. We vary the number of samples per neighborhood in order to reproduce a realistic dataset that might be very unbalanced: data is sparse for some neighborhoods while very dense for some others. For each neighborhood n, we sample uniformly the continuous random variable γ, which represents the order of the number of tags in neighborhood n, from the interval $[3, 6]$. The endpoints of the support of γ are based on the the minimum and maximum values observed in the Flickr dataset presented in Section 8.4.1, such that an expected number of 18.8×10^6 tags will be generated, similar to the quantity of tags available in the Flickr dataset. Once the number of tags ν to be generated is fixed, we sample, for each iteration, a level of the geo-tree and then a tag from the distribution associated with the corresponding node.

We can now assess the performance of a model by quantifying its ability to correctly predict the level in the geo-tree from which a tag observed in a neighborhood was sampled. In addition to our model, we consider the following methods:

Naive Bayes (NB) is a simple yet core technique in information retrieval [50]. Under this model, we assume that the tags observed in a class (node) are sampled independently from a multinomial distribution. Each class is therefore described by a multinomial distribution learnt from the count of tags that are observed in that class. However, since we do not use the class membership to train our methods, we assign a tag t, observed in neighborhood n, to all the classes (nodes) along the the path R_n, which amounts to having a uniform prior over the classes.

Hierarchical TF-IDF (HT) is a variant of TF-IDF that incorporates the knowledge of the geographical hierarchy. This variant was used in the TagMaps method [59] to find tags that are specific to a region at a given geographical level. The method assigns a higher weight to tags that are frequent within a region (node) compared

Algorithm 4: Generating tags in neighborhoods

Input: Geo-tree V, neighborhoods $n \in \{1, \ldots, N\}$, tags $t \in \{1, \ldots, T\}$, hyper-parameters α, β.
Output: Tags observed in each neighborhood $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$.

for $v \in V$ **do**
  Sample distribution $\theta_v \sim Dir(\alpha)$;
  if *v is a leaf* **then**
    Sample mixture coefficients $p(z|v) \sim Dir(\beta)$;
for $n \in \{1, \ldots, N\}$ **do**
  Initialize $\boldsymbol{x}_n = \mathbf{0}$;
  Sample order $\gamma \sim \mathcal{U}[3, 6]$;
  Set $\nu = \lfloor 2 \times 10^{\gamma} \rfloor$;
  while $\sum_t x_{nt} \leq \nu$ **do**
    Sample tree level $z \sim p(v|n)$;
    Sample tag $t \sim p(t|z, n)$;
    Increment tag count $x_{nt} \leftarrow x_{nt} + 1$;
  Sample tree level $z \sim p(v|n)$;
  Sample tag $t \sim p(t|z, n)$;
  Increment tag count $x_{nt} \leftarrow x_{nt} + 1$;

to the other regions at the same level in the hierarchy. We are able to represent each node with a normalized vector in which each tag t has a weight that encodes its descriptiveness.

For the classification, we map a tag t observed in the leaf n to the level $\hat{z}$ that maximizes the probability $p(z|t, n)$ (NB and GHM), or the tag weight (HT). Using our ground truth, we can then approximate the probability of correct classification $p(\hat{z} = z)$ by the proportion of tags that were correctly classified. In our evaluation, we repeat the following process 1000 times: we first generate a dataset, hold out 10% of the data for test purposes and train the model on the remaining 90%. For fair comparison, we initialize and smooth the parameters of each method similarly. Then, we measure the classification performance of each method. The final results, shown in Table 8.5, are therefore obtained by averaging the performance of each method over 1000 different datasets.

Our GHM model is the most accurate at classifying the tags to the correct level, greatly outperforming NB by 47% and HT by 27%. Even though both GHM and HT take advantage of the geographical hierarchy in order to classify the tags, the probabilistic nature of GHM enables a more resilient hierarchical clustering of the data, while the heuristic approach of HT suffers from overfitting. For example, if the number of samples available for a neighborhood is low, HT might overfit the training data by declaring a frequent tag as being characteristic, although not enough samples are available to conclude this. This is not case for GHM, because the assumptions of random mixture

	Classification Accuracy (std)
Random	0.33 (0.00)
NB	0.51 (0.02)
HT	0.59 (0.02)
GHM	**0.75** (0.01)

Table 8.5 – The average classification accuracy is computed, for each method, over 1000 generated datasets. We also indicate this accuracy if we classify tags uniformly at random.

enable us to obtain a resilient estimate of the distributions, which declare a tag as characteristic of a given level only if it has enough evidence for it. This observation is strengthened if we choose the maximum order γ of the number of tags per node to be 4 instead of 6: the classification accuracy of GHM decreases by 5% only (0.71), whereas the performance of HT decreases by 13% (0.51). Taken all together, these results suggest that, if the data observed in a neighborhood is a mixture of data generated from different levels of a hierarchy that encodes the specificity/generality of the data, our method will be successfully able to accurately associate a tag with the level from which it was generated.

8.5 Perception Focused User Study

The results of our model might not necessarily be intuitive given people's differing individual geographic perspectives [25]. Trying to objectively evaluate these, without taking into account human subjectivity and prior knowledge, could be misleading. For the Castro neighborhood in San Francisco, for example, the GHM classified the tag `milk` as specifically descriptive. Someone who is not familiar with Harvey Milk, the first openly gay person to be elected to public office in California and who used to live in the Castro district, would most probably not relate this tag to the neighborhood. It is therefore important to understand the correspondences and gaps between the results of our model and human reasoning about regions.

We conduct a user-focused study to explore the premise that the posterior probability of a tag being sampled from the distribution associated with a region is indicative of the canonical descriptiveness of this region. We further aim to identify potential challenges in user-facing applications of the model, and to uncover potential extensions to our model. We held ten interviews and conducted a survey with local residents of the San Francisco Bay Area focusing on their reactions to the tags that our model surfaced. To assess the performance of our model while reducing the bias of subjectivity, we used the results of our user survey to obtain the human classification of tags to nodes in the hierarchy,

allowing us to approximate the probability that our model classified a tag correctly by the average number of times it corresponded with human classification. We use the interviews to understand the reasons behind matches and mismatches.

8.5.1 Interview and survey methodology

Interviewees and survey respondents reacted to a collection of 32 tags per neighborhood. To ensure a certain diversity among the tags presented to the users, we selected randomly a subset of tags that are classified by the GHM as being descriptive of (i) neighborhood level (e.g., `graffiti`), (ii) city level (e.g., `nyc`), (iii) country level (e.g., `usa`), or (iv) another neighborhood (e.g., `mission` for the Castro neighborhood). We selected these tags randomly, with the probability of choosing a given tag t proportional to the probability $p(t|n)$. This survey highlights to which extent the locals' perspectives match the results produced by our model, while the interviews enable us to better understand the human perception of descriptiveness.

Interview procedure Our semi-structured interviews focused on how people describe neighborhoods. We investigated their reasoning behind the level of specificity they associate with a tag in a given neighborhood. Each one-on-one interview lasted 25–45 minutes. To have the perspectives of (former) locals to newcomers, we interviewed 10 people (5F, 5M; ages 26–62, $\mu = 37$, $\sigma = 12$) who (had) lived in the San Francisco Bay Area from 2 months to 62 years. Three of the participants worked in the technology industry. There was also two students, one real-estate agent, one building manager and one photographer. Each interview covered three different neighborhoods chosen by the participant out of 11 well-known San Francisco neighborhoods. However, one participant described only one neighborhood (due to time constraints), and another participant described four of them. Our interviews addressed the following points:

1. The participants' characterization of the neighborhoods using their own words, to get an understanding of the factors that are important to them.
2. Their considerations about whether a tag is perceived as specifically descriptive or not. Interviewees were first asked to classify the 32 tags presented to them as (not) specifically descriptive for the neighborhood and to explain the reasons. Then, they were shown the subset of tags that were classified by our model as specifically descriptive.

We emphasize the fact that the participants were not told about our model, nor that the terms presented to them were actually Flickr tags. This helps us identify the factors that led them to classify terms as (not) specifically descriptive of a neighborhood, and enables us to identify the factors not yet addressed by our model, without biasing their judgment

	Mission		Castro		North Beach	
Neighborhood-level tag count	17		17		17	
Neighborhood-level tag classifications	287	(100%)	201	(100%)	185	(100%)
Tags not understood	22	(8%)	3	(1%)	0	(0%)
Tags seen as non-descriptive for any level	116	(40%)	50	(25%)	62	(33%)
(Part of) neighborhood	109	(39%)	67	(33%)	37	(20%)
Higher-level node (CA/USA)	24	(8%)	38	(19%)	7	(4%)
Other neighborhood	16	(5%)	43	(22%)	79	(43%)

Table 8.6 – The distribution of answers given by survey participants. The (rounded) percentages are computed with respect to the total number of answers given.

towards our assumptions (e.g., hierarchy of tags). The interviews were recorded and the transcriptions were iteratively analyzed, with a focus on the identification of themes in the reasoning behind interviewees' classifications of tags.

Survey procedure A total of 22 San Francisco Bay Area residents (5F, 17M; ages 22–39, $\mu = 33$, $\sigma = 4.8$), who had lived there for an average of 5.6 years ($\sigma = 4.2$), participated in our survey about San Francisco and three of its neighborhoods (Mission, Castro and North Beach). Of these 22 respondents, 18 provided tag classifications for the Mission neighborhood, 12 for the Castro and 11 for North Beach. This resulted in 1291 tag classifications, of which 561 were for the Mission, 381 for the Castro and 349 for North Beach. The survey asked the participants to describe each neighborhood with their own words using open text fields and then to classify the 32 tags presented to them as descriptive for a given neighborhood, for a higher geographical level (the city or country), or for another neighborhood. They have also the possibility to indicate if they did not find the tag descriptive for any level, or did not understand its meaning.

8.5.2 Results

In this section, we present the results of the survey and provide examples from the interviews to understand the reasoning processes about whether a tag is descriptive for a specific region. We compare the tag classifications provided by the GHM with those supplied by the participants, and we specifically focus on disagreements between neighborhood-level tag classifications in order to identify the difficulties people have interpreting modeling results and potential extensions to our model.

Participant and model congruency

The model's premise that locally frequent content is not necessarily specific to a locale was strongly supported by the interviews and the survey. For the tags that occurred

frequently in a given neighborhood, none of the participants classified all of them as specifically *descriptive* of this neighborhood. This supports the results of the GHM in classifications of very frequent but wide-spread tags as not being specifically descriptive of the neighborhood. Without prompting, interviewees mentioned terms as being too generic or specific for a given neighborhood. For example, one interviewee (F 28), when describing the Western Addition neighborhood, picked `haight` as a descriptive tag, "because there's Haight Street in this neighborhood", but not the tag `streets`, as "there's [sic] streets everywhere". Similar interview examples included: "`california` or `usa` is a generic, or general term" (F 28), "I don't think of ever describing Golden Gate Park as in the USA. Unless I'm somewhere far away, but then I wouldn't even say USA, I would say California or San Francisco" (F 41). This result implies that participants tend to classify tags according to a geographical hierarchy, which supports the validity of the assumptions we make about the hierarchy of tags: tag specificity/generality depends on the hierarchical level from which it was sampled.

We are aware that people will not agree with every classification made by our model. Tags classified by the GHM as specifically descriptive of a neighborhood, were not necessarily perceived as such by all respondents; variations occurred between neighborhoods and between participants. As a consequence, evaluating the results of an aggregate model 'objectively', as if there were only a single correct representation of a neighborhood, is difficult. However, to place the GHM classification into a context with human classification, we use the results of our survey to reduce subjective biases: we obtain a majority-vote human classification by assigning each tag to the class that users have chosen most often. We then approximate the probability that the GHM classifies a tag 'correctly' by the average number of times its classification matches this human majority classification. For most tags the majority assignment is aligned with the assessment of the model (Table 8.4). We obtained an average classification correspondence of 0.77, with a highest classification accuracy of 0.84 for the Mission neighborhood. The alignment between the model and human classification for the Castro neighborhood was 0.81. Alignment was lowest for the North Beach neighborhood, with a correspondence of 0.66 mainly caused by tag classifications as 'another neighborhood' (Table 8.6). Such mismatches occur for a multitude of reasons. First of all, as a very basic requirement, participants have to understand what a tag refers to, before they can assign it to a specific level. For example, in the survey, 8% of the answers given for the Mission neighborhood were "I don't know what this is" (Table 8.6). This issue occurred less for the selection of tags for the Castro (1%) and North Beach (0%). The terms that users understood were necessarily perceived as either descriptive (i.e., assignable to a level in the geographical hierarchy) or non-descriptive (i.e., not belonging to any level in the hierarchy). The proportion of individual answers that are "non-descriptive" is around 34%, which included answers to tags such as `cat`, `sticker` and `wall` for the Mission. The fact that these tags indeed describe content that occurs frequently in the Mission does not imply that they are perceived by *all* users as descriptive.

People's local experiences shape and differentiate their perceptions. The interviews illustrated how tags were interpreted in multiple ways; `church` was taken to refer to a church building, or to the streetcars servicing the J-Church light-rail line, and it was not seen as descriptive by the majority of participants (see Castro in Table 8.4). Local terms surfaced by our model, such as `walls` in the Mission, represented the neighborhood's characteristic murals to a long-time local interviewee (M 49), whereas this term was a meaningless for others. Although the majority of survey respondents classified `night` and `coffee` as unspecific for the Mission, the same interviewee (M 49) for example saw `night` as descriptive for its bars, restaurants and clubs and thought `coffee` referred to the copious amounts of coffee shops. Similarly, one interviewee described North Beach as "party land" (M 46), but another claimed "I don't believe there's much of a nightlife there" (F 26). These results highlight an opportunity for discovery and recommendation of local content that users might not be aware of but also means that a careful explanation might be necessary.

Model extensions

Beyond misunderstanding tags or finding them not specifically descriptive of a certain neighborhood, the interviews provided additional clues about the mismatches identified in our survey, as well as individual differences. These findings are organized below as potential opportunities for extensions of the GHM model.

Sub-region detection According to the majority of our survey participants, 11 tags classified by the GHM as specifically descriptive of North Beach were actually descriptive of another neighborhood. From our interviews, we learned that the terms surfaced by our model for the North Beach neighborhood included references to the Bay's waterside and to the tourist attraction Fisherman's Wharf (see for example `wharf`, `seal`, `crab`, `bay`, `pier` in Table 8.4). The locals that responded to our survey (see Table 8.6) and the interviewees clearly made a distinction between Fisherman's Wharf and North Beach, whereas the set of administrative regions we used for our model did not: According to the data from the city planning department of San Francisco, Fisherman's Wharf is not a neighborhood in itself but simply a sub-region of North Beach. Clarifications about region borders and the detection of emerging sub-neighborhoods (e.g., "as you go further down south it's a totally different neighborhood") would improve the quality of the neighborhood tags presented to the user. From the modeling perspective, the GHM already allows for considering predefined sub-regions in certain neighborhood, as the geo-tree can be unbalanced and have a different depth for certain sub-trees.

Permeable adjacency and topography The GHMA model, introduced in Section 8.3, is an extension of GHM that takes into account the porosity of the frontiers that

separate adjacent neighborhoods. Using *permeable adjacency* by considering adjacent neighborhoods appears promising. For example, the mismatching tag `dolorespark` (see Table 8.4) refers to the park right on the edge of the Mission and Castro, but it was placed outside the Castro neighborhood by the majority of our survey respondents. Interviewees defined neighborhood boundaries differently and sometimes indicated they did not know neighborhoods well "It's a little difficult because it's right next to the Presidio. I kind of, maybe confuse them from each other" (F 26). Interviewees extended their reasoning about activities or points of interest that could spill over into adjacent neighborhoods. For example, the Golden Gate Bridge, officially part of the Presidio neighborhood, but photographed from a wide range of other neighborhoods was mentioned as a distant-but-characteristic feature: "you can see the Bay Bridge from there" (M 46).

Temporality Time of day, shifting character of neighborhoods, events, long-term history, and even change itself were referenced by interviewees: "I think of night, ... there's a lot of activity during night time with the bars and the clubs... but before... you wouldn't be caught there at nighttime... 20 years ago it was a different neighborhood." (M 49). Because the Flickr tag collection we used spanned multiple years, some aspects of neighborhoods that were characteristic at one time but no longer as prominent at present (such as the Halloween celebrations in the Castro neighborhood), were considered out of place (M 32). Yet other interviewees still found such tags characteristic and freely associated: "I think of outrageous costumes, and also the costumes that people see when it's not Halloween" (F 26). We can enhance our model by including a time dependent component that captures the time evolution of the neighborhoods character. However, we must be aware that this necessitates a dataset that is much more prolific than the current one: we need to have a sufficient number of geo-tagged samples per time period.

8.6 Conclusion

With this abundance of geo-tagged data comes the promise of a better understanding of users' experiences and an increased ability to describe the environment where they take place. We illustrated this by mining millions of geo-tagged photos in order to construct a specific description of neighborhoods in different cities. We proposed a probabilistic model that enables for uncovering terms that are *specifically descriptive* of a region within a given geographical hierarchy. By applying our model to a large-scale dataset of 20 million tags associated with approximately 8 million geo-tagged Flickr photos taken in San Francisco and Manhattan, we were able to associate each node in the hierarchy to the tags that specifically describe it. Moreover, we used these descriptions to quantify the uniqueness of neighborhoods, and find a mapping between similar but geographically distant neighborhoods. We further conducted interviews and a survey with local residents in order to evaluate the quality of the results given by the GHM and

its ability to surface neighborhood-characteristic terms, which span both terms known and unknown to locals. The classification accuracy of GHM, measured with respect to the classification of tags made by locals, provides strong support for the validity of our approach. However, the results of the interviews also highlighted the difference between the performance of a model at classifying community-generated data, and its performance as judged subjectively by an individual. This highlights the importance of taking the subjectivity of the user experience into account, and the need for explanation and framing. As a consequence, the traditional evaluation of modeling approaches that are fed with user-generated data faces the challenge of both the representation and the subjectivity inherent to vernacular geography.

9 Conclusion

In this thesis, we studied mobility from different perspectives and explored three fundamental points (a) How can we learn accurate mobility models?, (b) How can we rigorously quantify mobility uncertainty and its evolution with location updates?, and (c) How can we take advantage of geo-tagged data in order to characterize the whole experience surrounding mobility?

In the first part of this thesis, we studied mobility from the modeling perspective and formalized the individual and collective dimensions in mobility models. Ideally, we should take advantage of both dimensions in order to learn accurate mobility models, but the nature of mobility data might limit us. We took a data-driven approach to study three scenarios, which differ on the nature of the mobility data to analyze, and developed mobility models that are suitable for each scenario. We showed empirically that (a) we are able to learn accurate individual-specific models given enough data, (b) the performance of individual-specific diminishes if individual traces are sparse, and (c) we can overcome sparsity by taking advantage of the collective dimension.

In the second part of this thesis, we proposed an information-theoretic approach to rigorously quantify mobility uncertainty and its evolution with location updates. We formalize the problem of mobility uncertainty as follows (a) a user moves stochastically on a mobility graph whose vertices represent branch points, and edges represent the link between these points, (b) the entropy of the distribution over all possible trajectories quantify mobility uncertainty, and (c) location updates amount to condition the trajectory on going through a set of vertices. In order to quantify the evolution of mobility uncertainty with location updates, we need to compute the entropy of Markov trajectories conditional on a set of intermediate states. We developed a method to compute the entropy of conditional Markov trajectories through a transformation of the original Markov chain into a Markov chain that exhibits the desired conditional distribution of trajectories. Moreover, we expressed the entropy of Markov trajectories as a function of the expected number of visits to each state and the local entropy associated with each

one of them. We used the trajectory entropy framework to quantify mobility uncertainty and its evolution with location updates, and to develop a segmentation algorithm that infers the likely intermediate destinations along a trajectory, based on the conditional trajectory entropy.

With the increasing availability of geo-tagged data, generated by online social medias, comes the promise of a better understanding of users experiences as they move. In the third part of this thesis, we studied mobility from the data mining perspective and showed how mining geo-tagged data enables us to gain more insight into the environment where users move. We illustrated this with a hierarchical probabilistic model that enables us to obtain a specific description of a region within a geographical hierarchy. This model, applied to a dataset of millions of geo-tagged photos, uncovers the terms that are specifically descriptive of the neighborhoods of San Francisco and New York.

9.1 Future Research and Challenges

Trajectory entropy and dynamics on graphs The entropy of conditional Markov trajectories quantifies the uncertainty about the trajectory followed by a random walk conditioned on going through a given vertex. We can think of trajectory entropy conditional on a given vertex as a measure of centrality that quantifies the importance of this vertex. This centrality is not only defined with respect to the structure of the graph but also with respect to the transition probabilities of the random walk. We believe that it interesting to explore the usage of conditional trajectory entropy as a measure of vertex centrality and to compare it to other classic measures of centrality such as random-walk based centrality or betweenness centrality. This would have applications in scenarios that study dynamics on graphs such as information propagation on social networks.

Approximating trajectory conditional entropy We showed in Chapter 6 that computing the conditional trajectory entropy $H_{sd|u}$ is costly because it requires the inversion of a $(N-1) \times (N-1)$ matrix, where N is the number of states of the Markov chain. To reduce this complexity, an interesting direction to explore is the approximation of the conditional $H_{sd|u}$ by the sum of entropies $H_{su} + H_{ud}$. We showed that the equality $H_{sd|u} = H_{su} + H_{ud}$ holds if $p(T_{su} \notin \mathcal{T}^d_{su}) = 1$, but we are still not able to compute or bound the difference $H_{sd|u} - (H_{su} + H_{ud})$. Being able to do so would enable us to approximate conditional trajectory entropy by a quantity that is much less costly to compute. In fact, we would be able to pre-compute once the matrix of trajectories entropies H, and then compute the approximation of the conditional entropy $H_{sd|u}$ by adding two elements of matrix H.

The data divide In the introduction, we mentioned that we are witnessing the data revolution that is characterized by a deluge of data, or what is commonly called "Big data". However, only the Internet companies, such as Google or Facebook, have access to really large datasets. A sociologist working for Facebook will have access to data that the rest of the research community will not, and a multimedia researcher working for Yahoo will have access to *all* photos on Flickr. Even the largest dataset presented in this thesis (Flickr photos in Chapter 8) was analyzed during an internship within Yahoo labs. We can naturally sample from this data by using the public APIs provided by social media companies, but this does not provide us with all the data these companies are collecting. This created *data inequality* between researchers that collaborate with Internet companies and the rest of the scholarly community, which is detrimental to research. It contradicts the principle of reproducible research because the research community is not able to verify the claims made by the researchers with exclusive access to the data of Internet companies.

Bibliography

[1] Apache hadoop. http://hadoop.apache.org. Accessed: 2015-04-01.

[2] Apache pig. http://pig.apache.org/. Accessed: 2015-04-01.

[3] Apache spark. https://spark.apache.org/. Accessed: 2015-04-01.

[4] The federal statistical office. http://www.bfs.admin.ch/. Accessed: 2015-04-01.

[5] *Principles of Social Science*. J. B. Lippinco tt & Co., Philadelphia, PA, 1959.

[6] *Economies of Signs and Space*. SAGE Publications, 1994.

[7] *Debates in the Digital Humanities*. University of Minnesota press, 2012.

[8] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy Enhancing Technologies*, volume 4258 of *Lecture Notes in Computer Science*, pages 36–58. Springer Berlin Heidelberg, 2006.

[9] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of Digital Libraries*, pages 1–10. ACM Press, 2007.

[10] A. Ahmed, L. Hong, and A. Smola. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of WWW'13*, 2013.

[11] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, Oct. 2003.

[12] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *Proceedings of WWW'08*, pages 357–366. ACM, 2008.

[13] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, New York, 2006.

[14] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *NIPS*, 2003.

[15] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[16] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for development: the D4D challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[17] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw. Localization of the maximal entropy random walk. *Phys. Rev. Lett.*, 102:160602, 2009.

[18] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, June 2007.

[19] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.*, 26(2):4:1–4:26, June 2008.

[20] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1082–1090. ACM, 2011.

[21] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[22] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The livehoods project: utilizing social media to understand the dynamics of a city. In *Proceedings of ICWSM 2012*. AAAI, 2012.

[23] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[25] M. Egenhofer and D. Mark. *Naive geography*. Springer, 1995.

[26] L. Ekroot and T. Cover. The entropy of markov trajectories. *IEEE Transactions on Information Theory*, 39(4):1418 –1421, jul 1993.

[27] S. Elwood, M. Goodchild, and D. Sui. Prospects for VGI research and the emerging fourth paradigm. In *Crowdsourcing Geographic Knowledge*, pages 361–375. Springer, 2013.

[28] V. Etter, M. Kafsi, and E. Kazemi. Been there, done that: What your mobility traces reveal about your behavior. *The Proceedings of Mobile Data Challenge by Nokia Workshop at the Tenth International Conference on Pervasive Computing*, June 2012.

[29] V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser, and P. Thiran. Where to go from here? mobility prediction from instantaneous information. *Pervasive and Mobile Computing*, 9(6):784 – 797, 2013. Mobile Data Challenge.

[30] M. Franklin, A. Halevy, and D. Maier. From databases to dataspaces: A new abstraction for information management. *SIGMOD Rec.*, 34(4):27–33, Dec. 2005.

[31] E. Frejinger, M. Bierlaire, and M. Ben-Akiva. Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, 43(10):984 – 994, 2009.

[32] R. G. Golledge. Path Selection and Route Preference in Human Navigation: A Progress Report. *Spatial Information Theory A Theoretical Basis for GIS, Lecture Notes in Computer Science, Springer Verlag*, 988:207–222, 1995.

[33] G. H. Golub and C. F. van Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. The Johns Hopkins University Press, 3rd edition, Oct. 1996.

[34] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[35] B. Hillier and L. Vaughan. The city as one thing. *Progress in Planning*, 67(3), 2007.

[36] L. Hollenstein and R. Purves. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1):21–48, 2010.

[37] G. Huiji, T. Jiliang, and L. Huan. Mobile location prediction in spatio-temporal context. *the Procedings of Mobile Data Challenge by Nokia Workshop at the Tenth International Conference on Pervasive Computing*, June 2012.

[38] K. Ingram, S. Krummey, and M. LeRoux. Expression patterns of a circadian clock gene are associated with age-related polyethism in harvester ants, pogonomyrmex occidentalis. *BMC Ecology*, 9(1):7, 2009.

[39] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, HotMobile '10, pages 19–24, New York, NY, USA, 2010. ACM.

[40] M. Kafsi, H. Cramer, B. Thomee, and D. A. Shamma. Describing and understanding neighborhood characteristics through online social media. In *Proceedings of WWW'15*, 2015.

[41] M. Kafsi, M. Grossglauser, and P. Thiran. The entropy of conditional markov trajectories. *Information Theory, IEEE Transactions on*, 59(9):5577–5583, Sept 2013.

[42] M. Kafsi, E. Kazemi, L. Maystre, L. Yartseva, M. Grossglauser, and P. Thiran. Mitigating epidemics through mobile micro-measures. In *NetMob 2013*, 2013.

[43] V. Kaufmann, M. M. Bergman, and D. Joye. Motility: mobility as capital. *International Journal of Urban and Regional Research*, 28(4):745–756, 2004.

[44] J. Kemeny and J. Snell. *Finite Markov chains*. University series in undergraduate mathematics. VanNostrand, New York, repr edition, 1969.

[45] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *Proceedings of INFOCOM 2006. 25th IEEE International Conference on Computer Communications*, pages 1–13. IEEE Computer Society Press, apr 2006.

[46] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proceedings of the ACM International Conference on Pervasive Services (ICPS), Berlin*. ACM, jul 2010.

[47] P. Kumaraswamy. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1 - 2):79 – 88, 1980.

[48] B. Latour. Beware, your imagination leaves digital traces. http://www.bruno-latour.fr. Accessed: 2015-04-01.

[49] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. The mobile data challenge: Big data for mobile computing research. In *Mobile Data Challenge by Nokia Workshop*, Newcastle, UK, 2012. Springer.

[50] D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 4–15, London, UK, UK, 1998. Springer-Verlag.

[51] S. Lloyd and H. Pagels. Complexity as thermodynamic depth. *Annals of Physics*, 188(1):186 – 213, 1988.

[52] K. Lynch. *The Image of the City*. MIT Press, 1960.

[53] A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 359–367, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[54] D. P. Mersch, A. Crespi, and L. Keller. Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science*, 340(6136):1090–1093, 2013.

[55] D. Montello, M. Goodchild, J. Gottsegen, and P. Fohl. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3):185–204, 2003.

[56] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang. Summarizing tourist destinations by mining user-generated travelogues and photos. *Comput. Vis. Image Underst.*, 115(3):352–363, Mar. 2011.

[57] U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1(3):244 – 256, 1972.

[58] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web*, 3(1), 2009.

[59] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web*, 3(1), 2009.

[60] P. J. Rentfrow, S. D. Gosling, D. J. Stillwell, M. Kosinski, and J. Potter. Divided we stand: Three psychological regions of the united states and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, 2014.

[61] M. Saerens, Y. Achbany, F. Fouss, and L. Yen. Randomized shortest-path problems: two related models. *Neural Computation*, 2009.

[62] R. Shokri, G. Theodorakopoulos, J. Le Boudec, and J. Hubaux. Quantifying location privacy. In *2011 IEEE Symp. on Security and Privacy (SP)*, pages 247 –262, May 2011.

[63] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, Sept. 2010.

[64] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, feb 2010.

[65] L. Song, D. Kotz, R. Jain, and X. He. Evaluating next-cell predictors with extensive wi-fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12):1633–1649, dec 2006.

[66] R. Stedman. Toward a social psychology of place: Predicting behavior from place-based cognitions, attitude, and identity. *Environment & Behaviors*, 34(5), 2002.

[67] W. J. Stewart. *Probability, Markov Chains, Queues, and Simulation*. Princeton University Press, 2009.

[68] S. A. Stouffer. Intervening opportunities: A theory relating mobility and distance. *American Sociological Review*, 5(6):845–867, 1940.

[69] B. Thomee and A. Rae. Uncovering locally characterizing regions within geotagged data. In *Proceedings of WWW'13*, pages 1285–1296, 2013.

[70] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM.

[71] J. Wang and B. Prabhala. Periodicity based next place prediction. *the Procedings of Mobile Data Challenge by Nokia Workshop at the Tenth International Conference on Pervasive Computing*, June 2012.

[72] Wikipedia. Chauvenet's criterion, 2015. Online; accessed 20-February-2015.

[73] L. Yen, M. Saerens, A. Mantrach, and M. Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[74] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 791–800, New York, NY, USA, 2009. ACM.

[75] G. K. Zipf. The p1 p2/d hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):677–686, 1946.

Mohamed Kafsi

24 C avenue d'ouchy – 1006, Lausanne, Switzerland
(+41) 786297399 • mohamed.kafsi@epfl.ch • www.kafsi.ch
Swiss permit C - Permis C

Education

École Polytechnique Fédérale de Lausanne (EPFL) — **Switzerland**
Ph.D. student, supervised by Prof M.Grossglauser and Prof P.Thiran — *September 2009–May 2015*
Mining, Modeling and Predicting Mobility

École Polytechnique Fédérale de Lausanne (EPFL) — **Switzerland**
Master's degree in communication systems — *2006–2009*

Carnegie Mellon University (CMU) — **Pittsburgh, USA**
Exchange year (top 4 students selected for this exchange) — *2005–2006*

École Polytechnique Fédérale de Lausanne (EPFL) — **Switzerland**
Bachelor's degree in communication systems — *2003–2005*

Lycée Pères Blancs — **Tunis, Tunisia**
Scientific baccalaureat — *2002*

Research Interests

Researcher with a passion for modelling [human] behaviours by mining large scale datasets. My expertise includes probabilistic models, machine learning, graph theory and mining large scale datasets using parallel computing approaches such as Map-Reduce.

Work Experience

Yahoo! Labs — **California, USA**
Researcher, Mobile Sensing and User Behavior Research group — *August 2013–December 2013*
Mining geo-tagged data in order to describe specifically and compare neighborhoods of a city. Our approach, which is based on a probabilistic hierarchical model, uncovers features that are characteristic of a neighborhood. We then use these features to compare neighborhoods and to quantify their uniqueness.

Nokia Research Center — **Helsinki, Finland**
Research Intern — *September 2008–April 2009*
We worked on two issues inherent in GPS mobility traces: The impact of the GPS error on the accuracy of estimation of mobility-related data, such as speed and distance, and the need to compress mobility traces in order to allow for its storage, computation and display. We studied smoothing methods such as *spline smoothing* and showed, using a Nokia mobile application, how our methods can minimize the impact of GPS errors on the estimation of crucial metrics.

Deutsche Telekom (T-Labs) — **Berlin, Germany**
Research Intern — *April–July 2007*
The goal of this internship was to provide a framework to assess the feasibility and performance of future applications that rely on vehicular connectivity in urban scenarios. We conducted a thorough analysis of the connectivity of such networks by using a model inspired from percolation theory. We quantified the influence of a number of parameters, including vehicle density, proportion of equipped vehicles, and radio communication range. We also studied the influence of traffic lights and roadside units. Our results provided insights on the behavior of connectivity.

Awards

2012: Member of the winning team of the Nokia Mobile Data Challenge (Next-Location Prediction Challenge)

Technical Skills

Stochastic models, machine learning, information theory, graph theory, cryptography and security

Programming languages: Python, Java, Scala, C/C++, Matlab, R, Pascal, LaTeX

Computing: Hadoop, Pig, Hive, Spark, Condor

Databases: MySQL, SQLite, PostgreSQL

Operating Systems: Windows, Mac OS, Linux

Editing: Adobe Photoshop

Main Publications

Describing and Understanding Neighborhood Characteristics through Online Social Media

M. Kafsi, H. Cramer, B. Thomee and D. A. Shamma,*ACM WWW*, 2015

The Entropy of Conditional Markov Trajectories

M. Kafsi, M. Grossglauser and P. Thiran, *IEEE Transactions on Information Theory*, 2013

Where To Go from Here? Mobility Prediction from Instantaneous Information

V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser and P. Thiran, *Elsevier Pervasive and Mobile Computing Journal*, 2013

Mitigating Epidemics through Mobile Micro Measures

M. Kafsi, E. Kazemi, L. Maystre, L. Yartseva, M. Grossglauser and P. Thiran, *Third conference on the Analysis of Mobile Phone Datasets, NetMob, 2013*

Been There, Done That: What Your Mobility Traces Reveal about Your Behavior

V. Etter, M. Kafsi and E. Kazemi, *International Conference on Pervasive Computing (Pervasive)*, 2012
Winner of the Nokia Mobile Data Challenge about Mobility Prediction

Languages

Arabic: Mother tongue

English: Fluent

French: Mother tongue

German & Italian: Basic

Popular Press Coverage

PhysOrg.com, 24h.ch, lacote.ch and letemps.ch on *Mobility Prediction (Nokia Mobile Data Challenge)*

LesEchos.fr interview on *When the Phone will Predict our Behavior*

Volunteer Experience

Co-founder and General Secretary of TUNES: Co-founded the association of Tunisian students in Switzerland, which has around 300 members and counts among its partners the UNICEF and the Swiss confederation.

Radio reporter for the Swiss radio: Interviews broadcasted live.