

**INFERÊNCIA DA LOCALIZAÇÃO DE  
RESIDÊNCIA DE USUÁRIOS DE REDES SOCIAIS  
A PARTIR DE DADOS PÚBLICOS**



TATIANA PONTES SOARES ROCHA

**INFERÊNCIA DA LOCALIZAÇÃO DE  
RESIDÊNCIA DE USUÁRIOS DE REDES SOCIAIS  
A PARTIR DE DADOS PÚBLICOS**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: JUSSARA MARQUES DE ALMEIDA GONÇALVES  
CO-ORIENTADOR: PONNURANGAM KUMARAGURU

Belo Horizonte

Março de 2013



TATIANA PONTES SOARES ROCHA

**INFERRING HOME LOCATION OF SOCIAL  
NETWORK USERS FROM PUBLIC DATA**

Dissertation presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Ciência da Computação.

ADVISOR: JUSSARA MARQUES DE ALMEIDA GONÇALVES  
CO-ADVISOR: PONNURANGAM KUMARAGURU

Belo Horizonte

March 2013

© 2013, Tatiana Pontes Soares Rocha.  
Todos os direitos reservados.

R672i Rocha, Tatiana Pontes Soares  
Inferring Home Location of Social Network Users  
from Public Data / Tatiana Pontes Soares Rocha. —  
Belo Horizonte, 2013  
xxiv, 78 f. : il. ; 29cm  
  
Dissertação (mestrado) — Universidade Federal de  
Minas Gerais  
Orientador: Jussara Marques de Almeida Gonçalves  
  
Co-Orientador: Ponnurangam Kumaraguru  
  
1. Computação - Teses. 2. Redes sociais on-line -  
Teses. 3. Foursquare - Teses. I. Título.

CDU 519.6\*04(043)



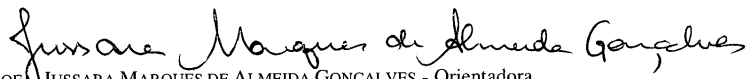
UNIVERSIDADE FEDERAL DE MINAS GERAIS  
INSTITUTO DE CIÊNCIAS EXATAS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO


## FOLHA DE APROVAÇÃO


Inferência da localização de residência de usuários de redes sociais a partir de dados públicos

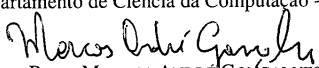
### TATIANA PONTES SOARES ROCHA

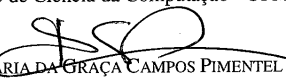
Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

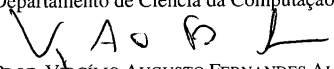
  
PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora  
Departamento de Ciência da Computação - UFMG

  
PROF. PONNURANGAM KUMARAGURU - Coorientador  
Departamento de Ciência da Computação - IIITD

  
PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR  
Departamento de Ciência da Computação - UFMG

  
PROF. MARCOS ANDRÉ GONÇALVES  
Departamento de Ciência da Computação - UFMG

  
PROFA. MÁRIA DA GRAÇA CAMPOS PIMENTEL  
Departamento de Ciência da Computação - USP

  
PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA  
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 11 de março de 2013.





*Aos meus pais, meus amores Gui e Jack.*



# Acknowledgments

Eu gostaria de agradecer, em primeiro lugar, à minha orientadora Jussara Almeida pelos últimos quatro anos em que trabalhamos juntas. Além da imensa admiração pela sua competência e comprometimento, sou muito grata por todas as oportunidades que ela me ofereceu durante toda a minha carreira acadêmica. Fica aqui o meu “muito obrigada” pela amizade, pela confiança depositada em mim, pela atenção constante, pelos nossos artigos, pelo grande aprendizado.

Um agradecimento especial ao professor Virgílio Almeida, que esteve sempre presente, apesar da agenda bastante requisitada. É um privilégio ter trabalhado com ele. *My sincere thanks also to my co-advisor, the professor Ponnurangam Kumaraguru, who has motivated the research topic of this dissertation and was actively involved in the development of this work.*

Gostaria de agradecer aos colegas, professores e funcionários do DCC/UFMG por serem tão atenciosos e prestativos comigo. Aos amigos do CAMPS, que fizeram meu dia-a-dia mais leve e divertido. Sem dúvida, os momentos mais agradáveis no DCC foram com Gabriel, Giovanni, Geraldo, Tiago, Pesce, Rapha, Las Casas, Saulo, Matheus, Fabrício, Evandro, Rauber e Emanuel. Agradeço também à Marisa, pelo companheirismo e pelas conversas no caminho de casa.

Não poderia deixar de expressar minha enorme gratidão aos meus pais, Gui e Jack, pelo imenso amor e por me motivarem todos os dias. Aos meus irmãos, Bruno e Kaká, e à minha cunhadinha Mari por acreditarem em mim. Às minhas vovós queridas, Jane e Darcy, pela torcida e por estarem sempre presentes. E aos tios e primos, pelo carinho e apoio.

Aos meus amigos, meu agradecimento pelos momentos de descontração e muitas risadas que aliviaram o stress.

A todos vocês, obrigada por fazerem parte desse sonho, agora concretizado.



*“Keep your thoughts positive  
because your thoughts become  
YOUR WORDS.*

*Keep your words positive  
because your words become  
YOUR BEHAVIOUR.*

*Keep your behaviour positive  
because your behaviour becomes  
YOUR HABITS.*

*Keep your habits positive  
because your habits become  
YOUR VALUES.*

*Keep your values positive  
because your values become  
YOUR DESTINY.”*

*(Mahatma Gandhi)*



# Resumo

A crescente acessibilidade às mídias sociais atrelada à facilidade de uso dos serviços de compartilhamento têm propiciado a geração voluntária de um grande volume de dados pessoais nesses ambientes. As informações compartilhadas, que variam de fotos do cotidiano a associações profissionais, podem ser exploradas para os mais diversos fins. Ao mesmo tempo em que esses dados criam oportunidades para os usuários fortalecerem seus laços nas redes sociais, eles também favorecem o desenvolvimento de mecanismos personalizados e estratégias de recomendação mais eficientes. Entretanto, esses mesmos dados podem ser manipulados de forma maliciosa e indesejada para promover marketing viral ou acessar informações confidenciais sobre os usuários. A violação de privacidade ocorre frequentemente devido ao desconhecimento e descuido das pessoas em relação àquilo que divulgam e tornam público. Com o aumento de serviços baseados em localização, um aspecto adicional é incluído ao dado referente à informação geográfica, o que torna a discussão sobre privacidade ainda mais incisiva, visto que tais dados podem colocar em risco a integridade física dos usuários, permitindo que eles sejam rastreados. Neste trabalho, analisamos uma das mais populares redes sociais baseadas em localização, o Foursquare, com o intuito de investigar como os seus membros exploram os recursos públicos do sistema (especificamente os atributos que possuem informação geográfica associada). A caracterização do comportamento humano no Foursquare consiste de um estudo que agrega cerca de 13 milhões de usuários e visa observar o potencial dos atributos geográficos do sistema em agir como fontes de vazamento de informação. Nesse contexto, propomos variados modelos de inferência na tentativa de revelar a localização da residência dos usuários a partir de dados geográficos publicamente disponibilizados. Apesar dos modelos serem genéricos e poderem gerar inferências em diferentes níveis espaciais, focamos nas inferências mais refinadas, nas granularidades de cidade e de coordenada geográfica, que, se bem sucedidas, representam riscos maiores à privacidade individual. Nossa avaliação experimental indica que os modelos propostos são capazes de inferir facilmente a cidade onde os usuários moram com uma precisão de cerca de 78% dentro de um raio de 50

quilômetros. Num grau ainda mais fino, acertamos a localização exata da casa dos usuários no nível de coordenada geográfica com aproximadamente 60% de acurácia em um raio de 5 quilômetros.

**Palavras-chave:** privacidade, inferência, redes sociais, foursquare, residência, localização.



# Abstract

The increasing access to social media, associated to the ease of use of sharing services, have fostered the voluntary generation of a large amount of personal data in these environments. The shared information, which vary from photos of everyday life to professional associations, can be exploited for various purposes. While these data provide opportunities for users to strengthen their ties in social networks, they also favour the development of personalised mechanisms and more efficient recommendation strategies. However, the same data can also be manipulated to promote malicious and unwanted viral marketing or access sensitive information about users. The privacy breach frequently occurs due to unawareness and carelessness of people about making information publicly available. With the rise of the location-based services, an additional aspect is added to the data related to geographic information, which makes the discussion about privacy even more incisive, since such data can endanger the physical safety of users, allowing them to be tracked. In this dissertation, we explore one of the most popular location-based social networks, Foursquare, aiming at investigating how its members exploit public system resources (specifically the attributes that are associated to geographic information). The characterisation of human behaviour in Foursquare consists of a study which aggregates about 13 million users and aims to observe the potential of geographic attributes in the system to act as sources of information leakage. In this context, we propose various inference models in an attempt to reveal the home location of users through their geographic data publicly available. Although the models are generic, being able to produce inferences at various scales, we focus on finer-grained inferences at the city and geographic coordinate levels that, if successful, represent greater risks to individual privacy. Our experimental evaluation indicates that the proposed models can easily infer the city where users live with an accuracy of about 78% within a radius of 50 kilometres. At an even finer scale, we correctly infer the coordinates of the users' home with approximately 60% accuracy within a 5 kilometres radius.

**Keywords:** privacy, inference, social networks, foursquare, residence, location.

# List of Figures

3.1	Snapshot of the Profile Page of a Foursquare User. . . . .	17
3.2	Cumulative Distribution of the Location-based Attributes per User. . . . .	22
3.3	Cumulative Distribution of the Location-based Attributes per City. . . . .	23
3.4	Global Distribution of Users and Venues Location across Cities. . . . .	24
3.5	Global Distribution of the Location-based Attributes across Cities. . . . .	25
3.6	Tag Cloud of the Words Present in Tips. . . . .	27
3.7	Cumulative Distribution of Displacements Between Consecutive Tips/Likes Posted per User. . . . .	28
3.8	Cumulative Distribution of Time Interval Between Consecutive Tips/Likes Posted per User. . . . .	29
3.9	Distribution of Returning Times. . . . .	30
4.1	Successful Example of <i>Iterative_MVS</i> Approach to Solve Ties among Lo- cations. . . . .	37
4.2	Summarisation of the Home Location Inference Models. . . . .	42
5.1	Impact of the Parameter <i>min_evidence</i> . . . . .	49
5.2	Impact of the Parameter <i>min_votes</i> . . . . .	50
5.3	Cumulative Distribution of the Distances Between the Declared and the Inferred User Home City for the <i>Mayorship</i> model. . . . .	54
5.4	Cumulative Distribution of the Number of Candidates for Neighbours for both <i>Local_KNN</i> and <i>Global_KNN</i> Approaches. . . . .	56
5.5	Cumulative Distribution of the Number of Users per Venue and Venues per Users. . . . .	56
5.6	Impact of the Parameter <i>K</i> for <i>Global_KNN</i> and <i>Local_KNN</i> Approaches. . . . .	58
5.7	Impact of the Parameter <i>min_friends</i> . . . . .	60
5.8	Impact of the Parameter <i>min_mutual</i> . . . . .	61

5.9 Cumulative Distribution of the Distances Between the Declared and the  
Inferred Geographic Coordinates of the User Residence. . . . . 64

# List of Tables

3.1	Summary of Statistics about our Foursquare Dataset. . . . .	18
3.2	Availability of Geographic Information (GI). . . . .	20
3.3	Quality of the Valid and Unambiguous Geographic Information. . . . .	21
5.1	Summary of the Results Obtained for the <i>Original_MVS</i> Approach for Home City Inferences. . . . .	47
5.2	Summary of the Results Obtained for the <i>Iterative_MVS</i> Approach for Home City Inference. . . . .	51
5.3	Results of the <i>Global_KNN</i> Approach for Home City Inference. . . . .	56
5.4	Results of the <i>Local_KNN</i> Approach for Home City Inference. . . . .	57
5.5	Summary of the Results Obtained for the Home Inference in the Geographic Coordinate Level. . . . .	63



# Contents

<b>Acknowledgments</b>	<b>xi</b>
<b>Resumo</b>	<b>xv</b>
<b>Abstract</b>	<b>xvii</b>
<b>List of Figures</b>	<b>xix</b>
<b>List of Tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Context . . . . .	2
1.2 Motivation / Purpose . . . . .	3
1.3 Objectives . . . . .	4
1.4 Contributions . . . . .	5
1.5 Organisation . . . . .	6
<b>2 Literature Review</b>	<b>7</b>
2.1 Location-Based Social Networks – LBSNs . . . . .	7
2.2 Location-Aware Recommendation Services . . . . .	9
2.3 Privacy in Online Systems . . . . .	10
2.4 Home Location Inference Strategies . . . . .	12
<b>3 Foursquare Dataset</b>	<b>15</b>
3.1 Foursquare: Key Elements and Features . . . . .	15
3.2 Crawling Methodology . . . . .	17
3.3 Dataset Overview . . . . .	18
3.4 Geographically Referenced Information . . . . .	19
3.5 Attribute Characterisation . . . . .	21
3.6 Spatial Analysis . . . . .	27

3.7	Temporal Analysis . . . . .	29
3.8	Summary of this Chapter . . . . .	30
<b>4</b>	<b>Home Location Inference Models</b>	<b>33</b>
4.1	Problem Statement . . . . .	33
4.2	Inference Models at the City Level . . . . .	34
4.2.1	The Majority Voting Scheme . . . . .	35
4.2.2	The K-Nearest Neighbour Approach . . . . .	38
4.3	Inference Models at the Geographic Coordinate Level . . . . .	40
4.4	Evaluation Methodology . . . . .	40
<b>5</b>	<b>Experimental Evaluation</b>	<b>45</b>
5.1	Inference Results at the City Level . . . . .	45
5.1.1	MVS Inference Models . . . . .	45
5.1.2	KNN Inference Models . . . . .	54
5.2	Inference Results at the Geographic Coordinate Level . . . . .	62
5.2.1	Experimental Setup . . . . .	63
5.2.2	Results . . . . .	63
5.2.3	Discussion . . . . .	65
<b>6</b>	<b>Conclusions and Future Work</b>	<b>67</b>
6.1	Main Contributions . . . . .	67
6.2	Limitations . . . . .	69
6.3	Future Work . . . . .	69
	<b>Bibliography</b>	<b>71</b>



# Chapter 1

## Introduction

The human history has always given evidence that human nature demands the creation of ties among people. In fact, life in society already means that the socialisation is expectable and needed. It is almost impossible to deny this instinct, once people live in contact with groups in the most varied environments such as at home with a family, in the work place with colleagues, or even in the bus with other passersby. Thus people end up establishing new relationships and eventually increasing their social network. Even in the Age of Technology, human nature remains in the virtual world finding alternative ways of communication and interaction, through *Online Social Networks* (OSNs).

Initially OSNs were designed to link close friends, but gradually new systems arose with different purposes, attracting users with different needs and reasons to sign up to this kind of system. Facebook <sup>1</sup>, Twitter <sup>2</sup>, LinkedIn <sup>3</sup> and Pinterest <sup>4</sup> are currently the most popular applications amongst Internet users [Alexa, 2013] allowing them to connect to a huge network of people spread throughout the world and to share an infinity of personal information, including photos, topics of interest, age, relationship status, and address. Due to the great popularity of these systems, becoming a member is a matter of time motivated by several reasons, including the efficient way to communicate and relate with others on an unprecedented rate, the possibility to share content in large scale, the opportunity of self-promotion, commercial interests, as well as the simple intent of socialisation [Tang et al., 2010].

---

<sup>1</sup>[www.facebook.com](http://www.facebook.com)

<sup>2</sup>[www.twitter.com](http://www.twitter.com)

<sup>3</sup>[www.linkedin.com](http://www.linkedin.com)

<sup>4</sup>[www.pinterest.com](http://www.pinterest.com)

## 1.1 Context

Due to the rapid growth in the use of smart devices equipped with Global Positioning System (GPS) receivers, location-based services (LBS) have become prevalent, raising the interest of the research community. Similarly, they have also motivated the creation of the *Location-Based Social Networks* (LBSNs) [Zheng, 2011], which are specialised systems in creating new means for online interaction based mostly on the geographic location of their registered users. LBSNs allow users to associate geographic information with the content they share, a feature that is being embedded also in OSNs.

Out of the various existing LBSNs, such as Yelp <sup>5</sup>, Google Latitude <sup>6</sup> and Instagram <sup>7</sup>, Foursquare <sup>8</sup> is currently among the most popular ones. <sup>9</sup> Its overall goal revolves around location sharing while users accumulate special awards for visiting specific places registered in the system. Such appeal encourages users to voluntarily make more personal information publicly available, such as their favourite places to visit, mobility patterns and behavioural habits. The availability of such data in the Web supports the design of several mechanisms and solutions that are of interest to the user, such as the development of personalisation mechanisms and tools to locate nearby friends according to the current user location [Berjani and Strufe, 2011], as well as urban planning models [Cranshaw et al., 2012], and more effective recommendation and advertisement strategies [Ye et al., 2010].

Meanwhile, improving the users' experience in the system and with the surrounding community through services and features which demand the disclosure of location information raises user exposure to a varied audience, possibly spread in many different systems. This overexposure potentially touches privacy concerns, creating opportunities for unauthorised usage of personal data [Ruiz Vicente et al., 2011]. Privacy attacks may occur in different fronts, as through the access of personal data [Friedland et al., 2011], or even via social network, due to information leakage provided by friends [Sadilek et al., 2012; Pesce et al., 2012]. These violations are typically result of privacy breaches which offer ways to gather user information from different sources unveiling sensitive data which, in turn, may contain revealing information such as the exact location and time of where and when a photo was taken.

Privacy breaches on LBSNs may favor positive and negative uses of personal

---

<sup>5</sup>[www.yelp.com](http://www.yelp.com)

<sup>6</sup>[www.google.com/latitude](http://www.google.com/latitude)

<sup>7</sup>[www.instagram.com](http://www.instagram.com)

<sup>8</sup>[www.foursquare.com](http://www.foursquare.com)

<sup>9</sup>The Foursquare community is estimated to include over 30 million people worldwide, according to the last census in January of 2013 (<https://foursquare.com/about/>).

data. At the point of view of police officers, criminals and suspect individuals could be investigated through the data disclosed on these kind of system, while overprotective parentes could take advantage of the availability of their kids' location information to track them being aware of where they have been and when. On the other hand, privacy violations may harm users making them more vulnerable to the action of robbers and kidnappers who previously observed a user habit to act in the most convenient time.

## 1.2 Motivation / Purpose

Social networking sites, in particular LBSNs, provide a range of increasingly more sophisticated security and privacy settings, aiming to empower their members as managers of their own exposition. Nevertheless, these settings are often confusing, unknown to all users, difficult to be applied, and often inefficient in controlling each possible sink-hole of information leakage [Gundecha et al., 2011]. Thereby, users also need to strike the right balance between concealment and disclosure in an attempt to meet their individual privacy requirements goals [Quercia et al., 2012]. Combining such scenario with the Web's tendency to "never forget" [Friedland et al., 2011], which can make shared data live forever, privacy violations may cause irreversible damage.

The diversity of user-supplied content – text, location, and photos – shared in a fast pace across systems creates a cloud of information about individuals. Taken together, the pieces of information disclosed on the Web can reveal a quite comprehensive picture of a person in ways that are hard to intuitively grasp, even if individually none of these pieces may be worrisome on their own [Friedland et al., 2011]. Therefore, users are constantly threatened by the potential of their own data in unveiling private information by the creation of inference chains formed with data correlation, friends that may reveal user data in their profiles, or simply by public attributes that are, by nature, not protected. This leakage of information can certainly tell much more about individuals than they are really aware of revealing.

Some studies have shown that the increasing amount of user data in online systems makes users more vulnerable to privacy violations. Through the analysis of historic data [Lieberman and Lin, 2009], textual messages [Mao et al., 2011], friend attributes [Gundecha et al., 2011], or user behaviour [Quercia et al., 2012], researchers are able to reveal an impressive range of sensitive information, allowing explicit data to unveil implicit information. Mao et al. [2011] have argued that it is possible to define if a Twitter user is under the influence of alcohol while tweeting only looking at the textual content and the time when the tweet was published. Also, Gundecha et al.

[2011] proved that a single vulnerable user can place all his friends at risk since he may be a source of information leakage, thus we can say that a user's privacy protection goes beyond his privacy settings and becomes a social networking problem. Privacy attacks are potentially more harmful in LBSNs in which the geographic aspect of the data can be explored. The analysis of location data permits the discovery of mobility patterns [Cheng et al., 2011b] and behavioural habits [Noulas et al., 2011]. Going further, the collation of geographic attributes may enable inferences about where a user lives. Once again, this kind of discovery is possible through user-generated data publicly available, such as the vocabulary in textual content [Cheng et al., 2010], friend characteristics [Davis Jr. et al., 2011], and the user's own traits [Mahmud et al., 2012].

Still, online privacy research in LBSNs clearly has not been sufficiently addressed in order to help developers to create safer systems or to alert users about protecting themselves properly from unexpected harm. This discussion has so far ignored an area that is poised to open up a new category of powerful privacy attacks based on global inferences resulting from automated content analysis, enabling cross-site correlation of personal information. The risks and implications of these attacks can also extrapolate the boundaries of the virtual world, making users physically vulnerable to robbery and kidnapping.

### 1.3 Objectives

In a world where confidentiality is something required in employment agreements that keep the company's information in secret from the public, finding a user's home location seems to be a much more invasive privacy breach, especially when it is uncovered by public data. In this context, the present dissertation aims to bring up a large-scale study on inferring the location of user homes in Foursquare through publicly available attributes.

Therefore, we guide a study that consists of three main steps. First, collecting the public information provided by users, building a dataset containing data on millions of users. Second, perform a basic characterisation of data that are potentially relevant in revealing private user information, in particular, the home location. Finally, we propose and evaluate alternative models that, based on data that are shared with everyone in the system, infer the user home location.

The inference assessment was conducted for different levels, from coarser granularities such as country and state to finer-grained approaches, restricted by the borders of a city or by the precision of geographic coordinates. We focus on the two more

refined inference levels which represent more serious privacy violations, since they are closer to revealing the exact location of the user residence.

## 1.4 Contributions

Since LBSNs are an emerging kind of system in the web, studies in this field are considered innovative and relevant. The novelty is due to the fact that a large amount of data produced by their users is associated to geographic information – something new compared to the data shared in the original OSNs. Thereby, exploring the locational aspect of users' content is valuable in the sense that it can be used for various purposes. Thus, knowing where a user lives, for instance, may provide improvements in various research areas, ranging from the development of models for human occupation in a city to the creation of more efficient strategies for recommendation and personalised advertisement systems – as reviewed in Chapter 2.

Our first efforts in inferring user home city (and also state and country) were presented in the Workshop on Location-Based Social Networks (LBSN'12) held in conjunction with the 2012 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'12) [Pontes et al., 2012b]. Then, an extension of this preliminary study was published in the Workshop of Privacy in Social Media (PinSoDa'12) in conjunction with the 2012 IEEE International Conference on Data Mining (ICDM'12) [Pontes et al., 2012a], in which we improve our last models to perform a finer-grained inference in the level of geographic coordinate of the user's residence. In both studies, we consider the same Foursquare dataset of millions of users and our inferences are applied in a global scale, not restrict to a specific region or country.

Given these considerations, the main contributions of this dissertation are: (1) the use of a large dataset containing data on millions of users of one of the currently most popular LBSNs – Foursquare; (2) the investigation of privacy breaches and information leakage; (3) the proposal of several models to infer the home location of a user exploring different system features and inference techniques; (4) and finally, the analysis of varied inference granularities from country to the exact user residence (represented by geographic coordinates). As part of this study, we intend to provide valuable insights about user behaviour, and an assessment on to which extent publicly available features can be exploited to uncover private user information, to drive future system designs and optimisations towards maintaining a fair and acceptable balance between the users' exposure and the quality of location-aware services.

## 1.5 Organisation

This dissertation is organised as follows. Chapter 2 presents the literature review, which includes studies on LBSNs and a discussion about privacy in online social networking systems. Chapter 3 explains some basic concepts and the terminology used in Foursquare, as well as a description of the crawling process and details about the dataset collected. A wide characterisation about the main features of the system is showed in the same Chapter. In Chapters 4 and 5, the proposed inference models are addressed, first with a thorough explanation about the methodology applied, and then the whole set of experiments, followed by discussion of results achieved. Finally, this dissertation is concluded in Chapter 6.

# Chapter 2

## Literature Review

User behaviour and information diffusion in online systems is a very active and relevant research area, with a rich set of studies. Since location-aware services and social networks are getting even more popular, there is a great interest in analysing how users explore the new resources available in such systems and which kind of information they are geographically tagging (topic discussed in Section 2.1). This increasing amount of user information on the web relights the investigation of contrasting uses of such data: for good faith purposes – as the creation of new recommendation strategies and more effective personalisation mechanisms (covered in Section 2.2), and for suspect purposes – as opportunistic actions (e.g., the development of methods capable of tracing users or to infer sensitive information through publicly available data) facilitated through privacy breaches caused by high exposure on the system (in Section 2.3). In this context, this Chapter is concluded with a deep discussion about an important privacy concern related with location: the home location inference. The state-of-the-art inference models (tackled in Section 2.4) are presented with an emphasis on the techniques and features used to reveal where a user lives.

### 2.1 Location-Based Social Networks – LBSNs

Sharing the current location or some other geotagged information like photos, status and videos associated with some specific place is the new tendency among users in the LBSNs. These systems provide services which allow users to associate location information to the shared data, creating a map of places where they have left some virtual footprints. Therewith, some questions arise: *Why do users share their own location in social networks?* and *What can researchers do with the increasing amount of geographic-tagged information available?*

In an attempt to understand the reasons why users share their location, Tang et al. [2010] proposed a study which discusses the purpose and social-driven aspects of sharing. Some location sharing applications like Reno [Iachello et al., 2005], WatchMe [Marmasse et al., 2004] and the Whereabouts Clock [Brown et al., 2007] are all motivated by scenarios which have a more utilitarian perspective – *purpose-driven*. Thus, the requests for a user location focus on pragmatic issues, including activities such as planning a meeting, checking for availability, coordinating transportation routes and estimating traffic delays. Hence, users detain the power to decide whether to share, for which reason, with whom, and in which level of precision. These applications are in distinct contrast from current LBSNs that support location sharing within social networks. Foursquare, Loopt<sup>1</sup>, Yelp<sup>2</sup>, and Locaccino<sup>3</sup> are some examples of *social-driven* systems which emphasise the “social factor” of sharing related to more subjective aspects. This is also confirmed by Lindqvist et al. [2011], which justifies that users might announce their current location not because someone needs to know but because it is a way of fun, boosts self-presentation, signals friends’ availability and sustains the social capital within one’s network. Thereby, we see a clear shift in location sharing from the *one-to-one* approach to the current LBSNs’ approach (*one-to-many* or *one-to-all*), where members can share location with a much wider and more diverse audience.

In this context, several recent studies have focused on investigating geographic user information to understand aspects related to human mobility [Cheng et al., 2011b; Cho et al., 2011; Sadilek et al., 2012], user behaviour patterns [Noulas et al., 2011; Vasconcelos et al., 2012], cross-cultural peculiarities in the usage of online systems [Magno et al., 2012], places modelling through geotagged photos [Crandall and Snavely, 2012], city dynamics [Silva et al., 2012b] and urban development [Cranshaw et al., 2012], and natural event detection [Sakaki et al., 2010]. These studies consider users to be like web sensors [Silva et al., 2012a; Pozdnoukhov and Kaiser, 2011; Lathia et al., 2012] that are potentially social indicators of topics associated with particular places and times. Observing users’ visits to different places, Cheng et al. [2011b] have shown that the movements performed by users follow simple reproducible patterns explained by social status, in addition to geographic and economic factors. Also, Cho et al. [2011] have proposed a human mobility model based on a combination of periodic short-ranged movements both geographically and temporally limited, and seemingly random jumps (due to long distance travels) highly influenced by user friendships. Noulas et al. [2011] and Vasconcelos et al. [2012] have focused on Foursquare, analysing the dynamics of

---

<sup>1</sup>[www.loopt.com](http://www.loopt.com)

<sup>2</sup>[www.yelp.com](http://www.yelp.com)

<sup>3</sup>[www.locaccino.com](http://www.locaccino.com)



collective user activity and uncovering distinct behaviour profiles, while Magno et al. [2012] have addressed Google+, showing the various usage patterns of the services available across different cultures. In particular, the usage of geographic tags in photos is explored by Crandall and Snavely [2012], who leverage such information to identify famous places and regions (highly photographed) and create 3D versions of landmarks. They also have observed that when two people are photographed at about the same place and time on five distinct occasions, they have nearly 60% of chances of being friends. Finally, Cranshaw et al. [2012] have proposed an online system able to portrait the rhythm of human steps, in near real time, throughout different parts of a city. This kind of monitoring also motivated Sakaki et al. [2010] to create mechanisms to detect earthquake promptly, to broadcast timely notifications.

The study proposed in this dissertation explores the Foursquare LBSN. We analyse the geographic and temporal aspects of user activity in the system through publicly available personal attributes. Such analysis of our dataset provides a proper understanding about how users behave in the system, in terms of how (and whether) they use these attributes, giving enough subsidies for us to propose and develop home location inference models – described in Chapter 4.

## 2.2 Location-Aware Recommendation Services

Making recommendations or offering suggestions to users in order to increase their degree of engagement is a very common practice for websites. Moreover, the phenomenal participation of users in OSNs, and specially in LBSNs, has given a tremendous hope for designing a new type of user experience based on both the *social* and the *spatial* aspects. While traditional recommenders provide default and generic results to everyone, social network-aware and location-aware systems can bring forth more targeted recommendations based on information gathered from friends and places visited [Vögele and Schlieder, 2003]. Particularly for users whose activity is little to none in a system (*cold starters*), social recommendations are notably interesting, since they enhance the input data for a recommender with more information, increasing the chances to achieve an appropriate and effective suggestion. Also, considering the location associated with the available data can be useful to drive the suggestions towards a geographically limited space where a user will probably move.

The task of recommending places has been tackled by some authors. For instance, Berjani and Strufe [2011] have proposed personalised recommendations of places in the extinct LBSN Gowalla, exploiting collaborative filtering techniques and the num-

ber of times a user has visited specific places. Ye et al. [2010] have also developed a recommender based on the social and spatial ties among users and their visited locations – they argue that friends share more common locations than non-friends, and nearby friends tend to share more commonly visited locations. Knowing this, Quercia and Capra [2009] have recommended friends using short-range technologies (e.g., bluetooth) on mobile phones based on social network theories of “geographic proximity” and “link prediction”. Furthermore, complementarily, Cheng et al. [2011a] have found that traffic patterns revealed through Foursquare history of visits to places can identify semantically related locations, thus favouring the creation of a traffic-driven location clustering algorithm to group semantically related locations with high confidence, which, in turn, may be naturally incorporated into location-based recommenders. Other types of recommendation are also possible, as the suggestion of social events proposed by Quercia et al. [2010]. Authors agree that there is a clear relationship between preferences for social events and geography and it also contributes to the recommendation of users to events. Saez-Trumper et al. [2012] show that individuals tend to go to a venue not only because they like it but also because they are close by.

Since this dissertation intends to present different ways to infer the real home location of users in Foursquare, we can say that the result offered by our proposed models may help researchers and system developers to improve the current recommendation strategies in this LBSN. Knowing the location where a user lives, whether at city level or at finer-grained granularities, favours the systems to perform nearby suggestions discarding misplaced options such as recommendations for places located too far from the user’s home location.

## 2.3 Privacy in Online Systems

Improving the user experience in online systems with sophisticated location sharing services and more accurate and personalised recommendations comes at a cost – it raises several concerns about privacy related issues. This occurs due to the rapid increase in the amount of personal and sensitive user information publicly available through a diverse range of social networks with different purposes, which is contributing to open privacy breaches and letting users even more exposed [Mao et al., 2011; Krishnamurthy and Wills, 2008]. Some researchers have shown how users face their own exposition in such systems, and which strategies or tricks they resort to an attempt to manage the visibility of their profiles, avoiding to be vulnerable to privacy violations. More sophisticated studies addressed the use of inferences applied in general data through the

collation of a set of public user attributes in a system to uncover private information. Recently, due to the amazing popularity of the LBSNs, studies which involve privacy related with the disclosure of some geographic information have also become a topic of interest for the research community [Wagner et al., 2010; Fusco et al., 2011].

Barkhuus and Dey [2003], as well as Gross and Acquisti [2005], have studied how users deal with privacy concerns in online systems. They observed that frequently there is not much concern regarding this topic, since users are more interested in the quality of the services offered in these environments than in protecting their data. In this context, and also due to the complexity of expressing privacy preferences on various applications [Benisch, 2011], only a small minority of users makes some effort to change the highly invasive privacy settings. Moreover, Li and Chen [2010] have shown that there are correlations among the user's vulnerability and his personal characteristics, which include factors like age, gender, friendships, mobility patterns, and others.

Although many users do not seem to worry about their high exposition in the system, Choudhury et al. [2010] agree that, many times, users are not aware of the risks involved. One of the main threats to user privacy is derived from inferences, which consist in combining pieces of explicit information in attempt to generate new conclusions, sometimes not so evident, which can reveal implicit and sensitive data and make users more vulnerable. It is known, for instance, that individual preferences can be deduced from friends which in general share similar preferences, as revised by Gundecha et al. [2011] and Li and Chen [2010]. Likewise, Mislove et al. [2010] and Zheleva and Getoor [2009] claim that user homophily does influence the information diffusion in social networks, suggesting that people with common characteristics and similar tastes are more likely to become friends, and therefore end up creating dense and homogeneous communities. Thus, this scenario where similar users are clearly grouped is quite favourable for the successful application of the inferences. Pesce et al. [2012] have demonstrated, in particular, that a simple tagged photo could reveal private user attributes that are extremely sensitive. Conclusively, users see themselves in a blind alley, since their efforts against information leakage may be insufficient to keep them protected. Indeed, Lam et al. [2008] and He et al. [2006] have shown that users are frequently unaware of information leaks through social relationships, which characterise involuntary violations.

The increasing number of users who are joining LBSNs has also attracted researchers towards the risks associated with privacy breaching in this kind of system. Annavaram et al. [2008] have created an application to guarantee privacy preservation of shared locations, whereas Ruiz Vicente et al. [2011] have punctuated scenarios which may result in problems for users when sharing location. Closely related to what this dis-

sertation proposes, Jin et al. [2012] have provided an analysis of user activity involving residential venues in Foursquare, aiming to identify system vulnerabilities and privacy risks. Although the numerous concerns associated with location sharing, Barkhuus and Dey [2003] have shown that users are often more worried about services that trace locations (*location-tracking*) than the ones that only request instant location (*position-aware*). Also, Lindqvist et al. [2011] have observed that only a small minority of users change their privacy settings and the ones who are concerned with privacy usually opt not to share their current location, omitting this data from their profiles.

In this dissertation, we discuss privacy-related issues regarding the geographic attributes present in Foursquare. Our study was carried out with a view to provide valuable insights to drive future systems designs and optimisations towards maintaining a fair and acceptable balance between the users' exposure and the quality of the services in LBSNs.

## 2.4 Home Location Inference Strategies

The literature has various studies on whether it is possible to infer a user's home location from various features (or attributes) with some geographical information associated. Backstrom et al. [2010] have measured the relationship between spatial and social proximity among Facebook users observing that the probability of a friendship drops monotonically as a function of distance – a finding that motivated the authors to introduce an algorithm based on a maximum likelihood approach to predict the location of an individual. Similarly, Davis Jr. et al. [2011] have also proposed a model for inferring the location of Twitter users assuming that reciprocal relationships in that system usually consists of people who are likely to be geographically close. The lack of geographic-based features on Twitter has fostered the design of inference models based on the tweet textual content. Cheng et al. [2010] have created a model based on the common vocabulary of users from the same geographical region, while Hecht et al. [2011] and Mahmud et al. [2012] have used machine learning strategies to infer the location where users live by only looking at what they tweet. Likewise, a recent study on Twitter demonstrated that distinct sets of relevant keywords may be associated with different locations [Ikawa et al., 2012], thus favouring guesses about the location of a tweet based only on the set of words it contains.

Other efforts have targeted more sophisticated and challenging models that combine many aspects present in a system to derive a user's location. Focusing, once again, on Twitter, Li et al. [2012] have integrated signals observed from both social network

(friends) and user-centric data (tweets) into a unified probabilistic framework to profile the users' home location. Similarly, Sadilek et al. [2012] have also explored patterns in friendship formation, but here in conjunction with the content of people's tweets and their reported locations, showing that the combination of all these predictors result in a stronger and more accurate inference model. Finally, Lieberman and Lin [2009] have suggested that a wealth of information about contributors on Wikipedia <sup>4</sup> can be gleaned from edit histories, revealing that it is often possible to associate contributors with relatively small geographic regions, usually corresponding to where they were born or where they presently live.

As in these previous studies, this dissertation aims at proposing models to infer user home location. However, unlike them, we here focus on one of the most popular LBSNs, Foursquare. To our knowledge, no previous work has addressed this problem in this social network before. Our proposed models are based only on publicly available attributes which are specific of Foursquare, namely mayorships, tips and likes (detailed in Chapter 3). Basically, all these attributes are related with web pages that represent real places registered in the system: mayorships are titles given to the most frequent visitor of a place, tips are comments left by users about their previous experiences and opinions about the place, while likes are a sign of approval marked in a previously posted tip. These three types of information are locatable, since they are associated with the location (geographic coordinates) of the place to which they refer. Note that although various similar efforts [Hecht et al., 2011; Ikawa et al., 2012; Mahmud et al., 2012] have mostly explored the textual content of attributes, we here consider the location associated with attributes, proposing inference models that use them either in isolation or jointly. The techniques explored here – Majority Voting Scheme and the machine learning K-Nearest Neighbour algorithm – have also been adopted by some previous studies [Davis Jr. et al., 2011; Hecht et al., 2011]. However, we here apply these techniques to a dataset of millions of Foursquare users, providing inferences at a global scale, as opposed to previous efforts [Cheng et al., 2010; Backstrom et al., 2010; Sadilek et al., 2012] that restrict their inferences to a specific region.

---

<sup>4</sup>[www.wikipedia.org](http://www.wikipedia.org)



# Chapter 3

## Foursquare Dataset

In this Chapter, we review the main elements and features of the LBSN Foursquare (Section 3.1) providing a description about some basic aspects of the system and also the terminology used throughout this dissertation. We also detail our crawling methodology (Section 3.2) and present some properties of the dataset used in our experimental evaluation (Section 3.3). Then, we characterise our Foursquare dataset in terms of the main user attributes, which are publicly available through the system’s API and are geographically-referenced. Our goal is to assess how users can use these attributes, which are the basic input to our home location inference models, presented in Chapter 4. First, we standardise the location information associated with these attributes and analyse the *quality* of the data in terms of the level of spatial granularity (Section 3.4). Then, in Section 3.5, we characterise those attributes assessing their usage around the world. We also perform a spatial and temporal analysis of the data to uncover human mobility patterns and common behaviours. These results are discussed in Sections 3.6 and 3.7 respectively. And finally, we summarise our findings in Section 3.8.

### 3.1 Foursquare: Key Elements and Features

Foursquare, currently one of the largest and most popular LBSNs, was launched in early 2009 providing support to members location sharing with friends through *check ins*. Check ins are only performed on devices equipped with GPS or other service reported location, in which a user may select a location from a list of places or he may create a page for his actual location, named *venue*. Thus, venues are places of a wide variety of categories, such as restaurants, airports or residences, that represent real (physical) locations previously registered in the system. Examples are Taj Mahal (Agra, India) and Paradiso Restaurant (Belo Horizonte, Brazil). Foursquare has a

playful aspect that gives incentives to users who share more locations. Thus, the larger the number of check ins a user does, the more incentives she may earn to continue sharing. As incentives, Foursquare offers, for instance, *badges* and *mayorships*. Badges are like medals earned if a user checks ins at specific venues or achieves some predefined number of check ins. Mayorships, in turn, are titles given to the most frequent visitor of a given venue in the last 60 days. Venue mayors are often granted rewards, promotions, discounts or even courtesies by business and marketing managers who own the venue. Although Foursquare was initially created with the primary intention of promoting a game between users competing for check ins as well as badges and mayorships, it also includes attributes (*tips* and *likes*) that favour the recommendation of places among users. Tips are comments posted by users on specific venues which reflect their experiences and opinions about some aspect of visited places (e.g., the quality of service or availability of parking space in a restaurant or even instructions about how to find the place). Users can also keep track of previously posted tips, marking them as “like” to signal their agreement with the content of a tip.

Users in Foursquare can be categorised as standard user, celebrity or brand page.<sup>1</sup> Standard users are common members, celebrities are standard users who achieved more than 1,000 friends<sup>2</sup> and brand pages represent companies or shops. The main difference among them is the type of social relationship they can have: friendship and/or relation of following and being followed. Thus, while celebrities can have both friends and followers, brand pages can have just followers and standard users can have only friends.

In the context of this dissertation, our focus is on publicly available attributes associated to some location information. Thus, for each user in our dataset, we consider the history of mayorships, tips and likes, all related with a venue, which in turn has necessarily a public geographic position declared in its Foursquare web page. Likewise, we also consider the user friends list, since they are associated with a home location, represented as a public and optional field in their profile page. Check ins are not considered, since they are a necessarily private attribute.

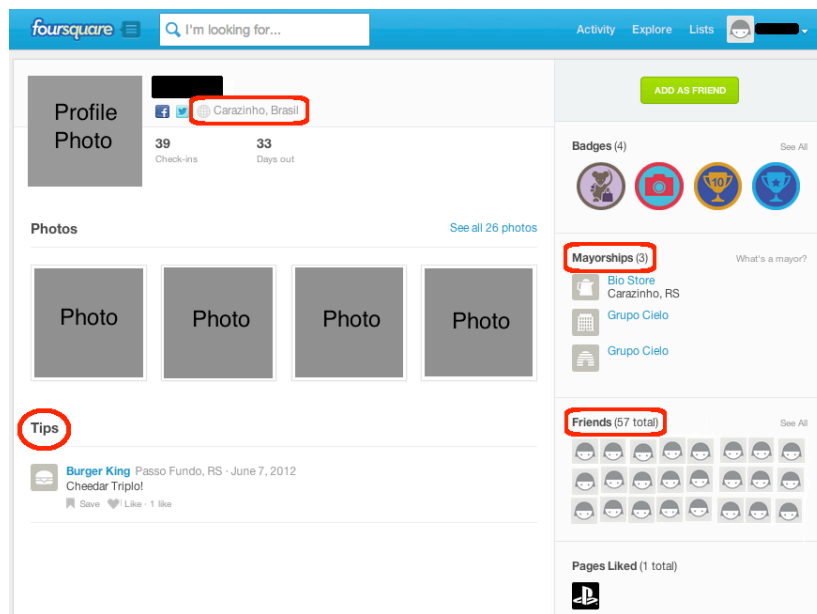
Note that mayorships, tips and likes are publicly available in Foursquare, and they are also broadcast to friends, appearing in their news feed. Figure 3.1 illustrates a system snapshot showing that the lists of mayorships, tips and friends (circled in red) are presented in the front page of a user profile – likes, on the other hand, only appear in feeds. Although optional, user home location is a public attribute and also appears in the profile page (next to the profile photo). Note that we cannot visualise the history

---

<sup>1</sup><http://aboutfoursquare.com/user-type-comparison/>

<sup>2</sup><http://aboutfoursquare.com/foursquare-converts-most-popular-users-to-celebrity-accounts/>





**Figure 3.1.** Snapshot of the Profile Page of a Foursquare User.

of check ins since it is a private information only visible by the user's friends, who also receive notifications about their friends' visits to venues.

## 3.2 Crawling Methodology

Our study is based on a large dataset crawled using Foursquare API from August to October 2011. We collected user profile data including home city, list of friends, mayorships, tips, likes, and total number of check ins. We also collected information associated with the venues visited by each user (i.e., venues linked to tips, likes and mayorships of the user), such as their location, category and total number of check ins and unique visitors. Recall that check ins were not collected, since they are not publicly available in the system, thus a private attribute.

Basically, our crawling strategy relies on a set of worker processes and a master process. The design of our crawler exploited the fact that each user in Foursquare receives a unique and sequential numeric identifier (ID). Thus, given (an estimate of) the largest ID assigned to a user in the system, the master process randomly selects an ID according to a uniform distribution in the range of IDs, and gives it to the next idle worker. We chose to perform a random selection of IDs, as opposed to sequentially selecting each possible value, to minimise the chance of bias towards older user accounts (which, we conjecture, have smaller IDs). The worker then sends a request to the

Foursquare API to gather information about the corresponding user.

We send HTTP GET requests to the pages of specific users identified by their IDs (we tried increasing values, starting with 0) to verify their existence. The largest ID for which we did get a response corresponding to a valid webpage was 20 million. Although we experimented with many IDs greater than that value, in all those cases, the response was *Not Found*. Therefore, we speculate that, at the time of our crawling, 20 million was about the largest user ID in Foursquare. We thus set this value as an estimate of the largest ID in the system and used it as input to our crawler.

### 3.3 Dataset Overview

# <b>Users</b>	13,570,060
# <b>Venues</b>	15,898,484
# <b>Mayorships</b>	15,149,981
# <b>Tips</b>	10,618,411
# <b>Likes</b>	9,989,325
# <b>Users with some activity</b>	4,140,434
# <b>Users with some friend</b>	6,973,727

**Table 3.1.** Summary of Statistics about our Foursquare Dataset.

Table 3.1 provides some statistics about our Foursquare dataset. Our entire data consists of 13,570,060 users, which we believe represent a large fraction of the total user population of the system by the time it was crawled.<sup>3</sup> It also includes 15,898,484 venues, which are in turn associated to users' attributes, namely mayorships, tips and likes. Although Foursquare was fairly new during the crawling process, the amount of data collected about the user activity in the system, represented by his attributes, was noticeably significant since, in total, we collected 35,757,717 attributes, including mayorships, tips and likes. In summary, 30.5% of the entire user population in our dataset have some activity (in terms of the collected features) in the system, with 2,873,883 unique users having at least one mayorship and 2,396,013 and 1,802,997 with some tip and like, respectively. We also looked into the user's list of friends and found that 51.4% of users are not isolated and have a social network.

<sup>3</sup>The overall number of registered users varied from 10 million in June 2011 (<https://foursquare.com/infographics/10million>) to 15 million in December of the same year (<http://www.socialmedianews.com.au/foursquare-reaches-15-million-users/>).

## 3.4 Geographically Referenced Information

The attributes considered in the proposed home location inference models are all geographically-referenced and publicly available in Foursquare. These attributes – described in detail in Section 3.1 – are the *home city* field, related to the location where a user lives, and the list of *mayorships*, *tips* and *likes* which, in turn, are associated with the location of specific venues in the system.<sup>4</sup> Recall that the information supplied as the home location of users and venues are both free-form, which means that they are open text fields whose validity is not enforced by the system. Indeed, they may carry noise and invalid locations.

The user’s home city, in particular, is limited to 100 characters and is not required to be filled. It is expected that users provide the name of the city where they live, although the system provides neither a rule to guarantee it nor any automatic tool to help users fill out the field (e.g., a predefined list of cities from which the user could choose one). Thus, users are free to provide this location information at various granularities, ranging from specific addresses, to city, state and country names, or even regions of the planet (e.g., “North Pole”). We also observed some home city fields filled with emails, phrases, or even numbers in our dataset. Similarly, the location associated with a venue, and thus, indirectly, with mayorships, tips and likes of that particular venue, is also an open text field. Unlike the user home city, the address and the city of a venue must be filled at the moment of the creation of the venue page. Moreover, it is necessary to set a pin in a map to update the venue’s location, in which a pin is essentially a point in geographic coordinates.<sup>5</sup> However, once again, users may choose to provide invalid addresses and city names, or mark arbitrary locations in the map.

To standardise and filter location names of users and venues, we used the *Yahoo! PlaceFinder* geocoding API.<sup>6</sup> The tool was used to perform disambiguation, that is, to uniquely identify a city despite the existence of multiple name variations (e.g., NY, New York City, etc). It was also used to verify whether some locations provided by users are actually valid. For example some users claim to live in imaginary or non-locatable places such as “around”, “everywhere”, or even “at Justin Bieber’s heart” [Hecht et al., 2011]. The use of the *Yahoo! PlaceFinder* tool allowed us to identify and disregard those non-valid places.

---

<sup>4</sup>*Check ins* and *Badges* are private attributes, and thus, it is not possible to access the geographic location associated with them.

<sup>5</sup>The availability of location information in the form of coordinates opens an opportunity for more specific inferences regarding user home location, such as the inference of the user residence location, as discussed in Section 4.3.

<sup>6</sup><http://developer.yahoo.com/geo/placefinder/>

Statistics	User Home City	Venue Home City
# in dataset	13,570,060 (100.00%)	15,898,484 (100.00%)
# valid unambiguous GI	12,939,569 (95.35%)	8,815,177 (55.45%)
# valid ambiguous GI	359,543 (2.65%)	2,868,636 (18.04%)
# non-GI	244,233 (1.80%)	4,214,671 (26.51%)
# empty entries	26,715 (0.20%)	0 (0.00%)

**Table 3.2.** Availability of Geographic Information (GI).

Basically, for a given query (text), the tool either returns some geographic data, in case the query consists of a valid location, or an error, otherwise. For queries consisting of valid locations, the tool’s response depends on the *quality* indicator of the query, which, in turn, is an integer value between 0 - 99 that represents the finest spatial granularity (e.g., street, city, state, country) that matched the corresponding location information provided in the query. For instance, for the query “Belo Horizonte”, *Yahoo! PlaceFinder* would provide the query’s quality (equal to 40, indicating that it is at the city level), the corresponding default geographic coordinates (the pair of latitude and longitude: -19.945360, -43.932678), a standardised city name (“belo horizonte”, in this case) as well as the state and country names (“minas gerais” and “brazil”).

Table 3.2 provides the distribution of the geographic information (GI) of all considered attributes in the dataset. We present the total number of users and venues in the dataset, as well as the percentages of those users and venues that correspond to valid geographic information (real location), non-geographic information (e.g., emails, phrases) and no information declared (empty entries). The valid geographic information can be unambiguous or ambiguous. Ambiguous information correspond to location names that, though identified as valid, can refer to multiple places. One example is “Springfield” which is the name of ten different cities in the United States. In those cases, *Yahoo! PlaceFinder* is unable to decide which one is correct. Observe that tips, likes and mayorships were grouped as venue attributes, while user attributes correspond only to the home city field.

Note that, perhaps surprisingly, the vast majority of the Foursquare users in our dataset (98% of a total of 13,570,060) do provide valid place names as home locations, with only a tiny fraction leaving it blank (0.2%) or filling it with non-geographic information (1.8%). Moreover, 11.6 million venues have valid associated locations, although a substantial fraction of all venues have non-valid locations (26.51%). This large fraction of non-valid venue locations comes as a surprise, particularly considering that, unlike the user home city field, the venue location information is a mandatory attribute. Since ambiguous locations do not represent a unique place in the globe, we

decided to disregard users and venues with ambiguous geographic information as home city, which correspond to 2.65% and 18.04% of all users and venues, respectively, in our dataset.

Quality	# Users	# Venues
<b>Continent</b>	107 (0.0008%)	61 (0.0007%)
<b>Country</b>	602,932 (4.66%)	294,596 (3.34%)
<b>State</b>	390,224 (3.02%)	93,513 (1.06%)
<b>County</b>	251,383 (1.94%)	276,097 (3.13%)
<b>City</b>	10,354,058 (80.02%)	6,937,523 (78.70%)
<b>Neighbourhood</b>	981,139 (7.58%)	1,060,124 (12.03%)
<b>Area of Interest</b>	27,307 (0.21%)	47,896 (0.54%)
<b>Street</b>	326,751 (2.53%)	95,543 (1.08%)
<b>Point of Interest</b>	5,607 (0.04%)	9,792 (0.11%)
<b>Geographic Coordinate</b>	61 (0.0005%)	32 (0.0004%)

**Table 3.3.** Quality of the Valid and Unambiguous Geographic Information.

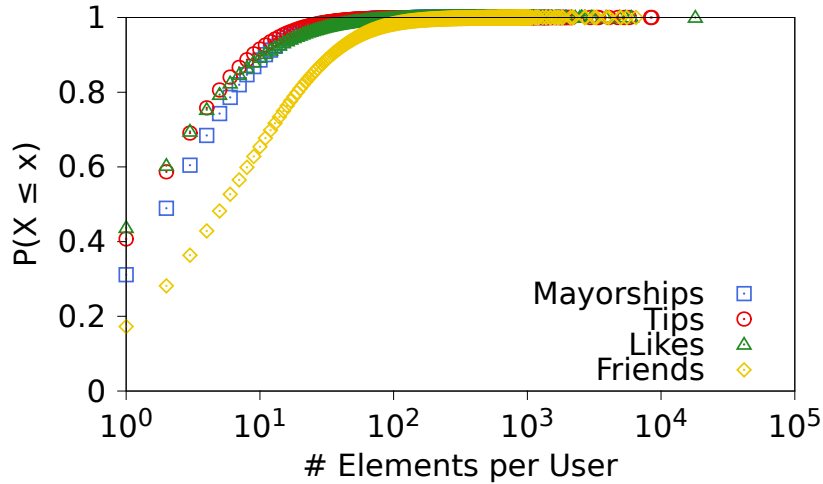
Next, we analyse the *quality* indicator of the valid (and unambiguous) geographic information available in the dataset. In Table 3.3, we present the distributions of the home location of users and venues across ten different quality levels, ranging from continent to specific geographic coordinates. It is clear that the vast majority of Foursquare users and venues have location information at the city level or at finer granularities. However, over 1.2 million users provide home location information at coarser granularities (often at country level) letting inferences at finer-grained levels even more invasive in revealing sensitive information. Thus, in sum, for our inferences we consider locations in the level of city or in finer granularities. Since each location is modelled for our inferences as a triple with the corresponding names of its city, state and country, 11,522,201 users and 8,127,790 venues are part of our experimental analysis.

Although we check the validity of the geographic information associated to the public user attributes, in this study we do not verify the veracity of the information. This means that fake attributes may exist in our dataset. Since the task to detect these attributes is not trivial, we do not apply any filter to remove them.

## 3.5 Attribute Characterisation

We here focus on the usage of the publicly available attributes present in our dataset. For each considered user, we analyse his home city field, the history of mayorships, tips, likes, and list of friends. Our goal is to assess the potential of exploiting these

attributes for inference purposes in terms of the fraction of users we would cover as well as how those users and locations are spread around the world.

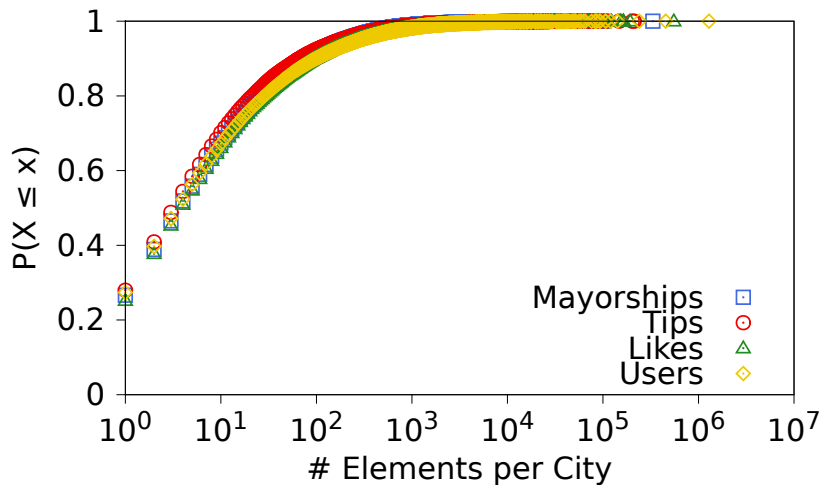


**Figure 3.2.** Cumulative Distribution of the Location-based Attributes per User (log scale in the x-axis).

Considering the entire dataset (without any filtering), we observe that almost 4.2 million users, or around 30% of all users in our dataset, have at least one mayorship, tip or like. Out of these, around 1 million have only mayorships, 670 thousand have only tips and 367 thousand have only likes, whereas 890 thousand users have all three attributes. Beyond that, almost 7 million users (about 51% of the total number of users in the dataset) have at least one friend in the system, being around 47% of them with some activity represented by the attributes mentioned (mayorships, tips and likes). Thus, exploiting all these attributes to infer a user home city is promising as the required information is available in a large fraction of the whole dataset. Moreover, as shown in Figure 3.2 and consistent with previous analyses of Foursquare [Noulas et al., 2011; Vasconcelos et al., 2012], the distributions of the numbers of these attributes per user are very skewed, with a heavy tail, implying that few users have many mayorships (tips, likes or friends) while the vast majority have only one mayorship (tip, like or friend). Indeed, for users that have one mayorship (tip, like or friend), we find that 69% (59%, 56% and 83%) of the users have 2 or more mayorships (tips, likes and friends).

Figure 3.3 shows the distributions of numbers of mayorships, tips, likes and users per city <sup>7</sup>, considering only cities with at least one instance of the attribute – in total,

<sup>7</sup>To compute the numbers of mayorships, tips and likes per city, we considered all venues located in each city and counted all the elements (mayorships, tips and likes) of those venues. Similarly, to



**Figure 3.3.** Cumulative Distribution of the Location-based Attributes per City (log scale in the x-axis).

our dataset includes references to 100,629 different cities around the world. For such analysis, we only considered those attributes which are associated with real locations – validated by the process described in Section 3.4. As shown, the distributions are also very skewed, with a few cities having as many as 100 mayorships, tips, likes or users.

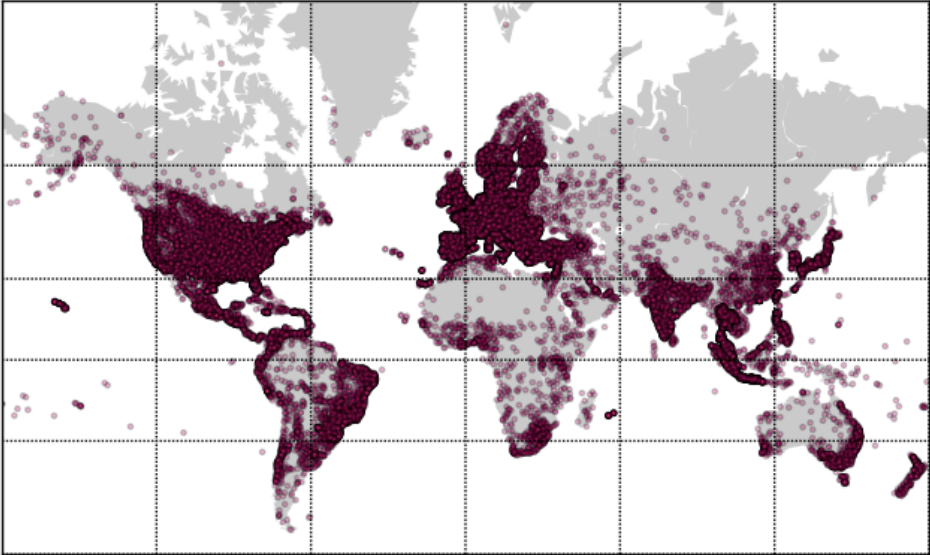
We now discuss the distribution of cities with users and venues around the world. Furthermore, we also show in distinct maps the cities with venues where users have mayorships, tips and likes. Once again, we only consider users and attributes associated with real cities as location (validated by *Yahoo! PlaceFinder*). Figures 3.4 and 3.5 show these distributions in maps of the globe<sup>8</sup>, with each point representing a city.<sup>9</sup> As the maps show, Foursquare users and venues are spread all over the world, including remote places such as Svalbard, an archipelago in the Arctic Ocean, with coordinates (78.218590,15.648750). Moreover, all five maps are very similar, with most incidences of points in America, Europe and Southeast Asia. The distribution of venues, in Figure 3.4(a), aggregates the highest number of distinct cities (82,248) among all distributions, showing that there are cities which host venues, but have no user registered in the system (since users comprise less cities, 76,918 in total – Figure 3.4(b)). We also observe that the distribution of mayorships, shown in Figure 3.5(a), is denser, with a total number of unique cities (75,169) much larger than in the distributions

---

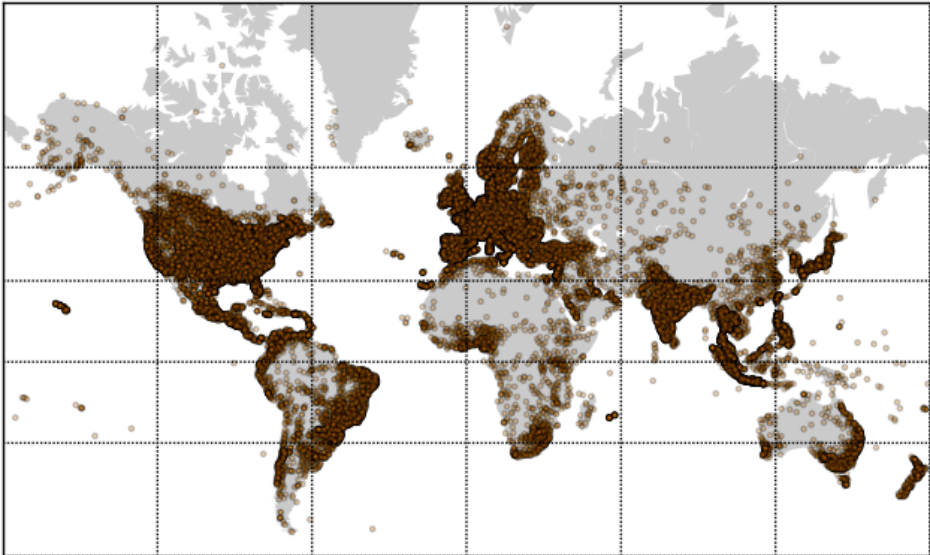
compute the number of users per city, we counted the amount of all users with home location in each city.

<sup>8</sup>The maps were plotted using the *Basemap* package from Python’s library (<http://matplotlib.org/basemap/users/geography.html>)

<sup>9</sup>The Antarctica continent was omitted because there was no point on it.



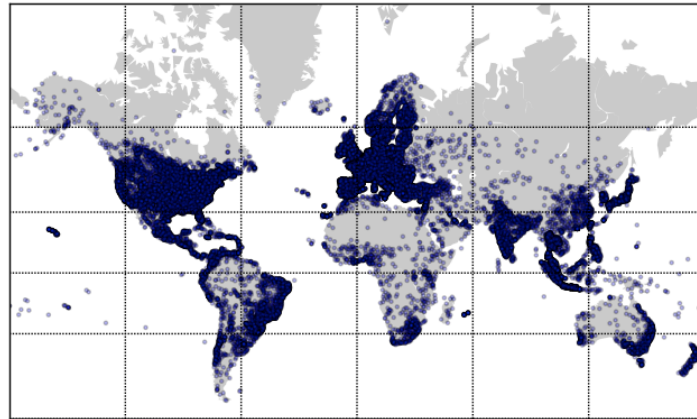
(a) Venues.



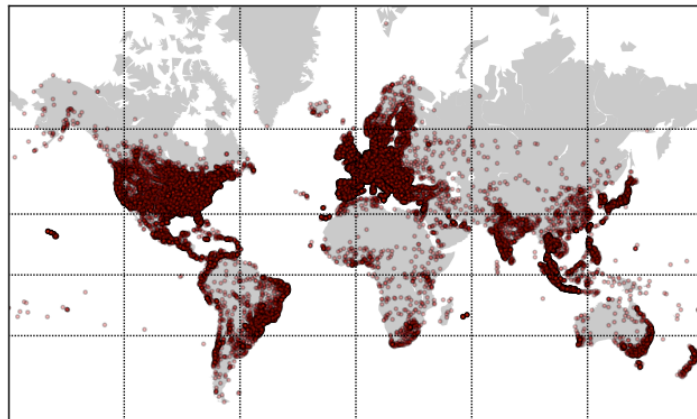
(b) Users.

**Figure 3.4.** Global Distribution of Users and Venues Location across Cities.

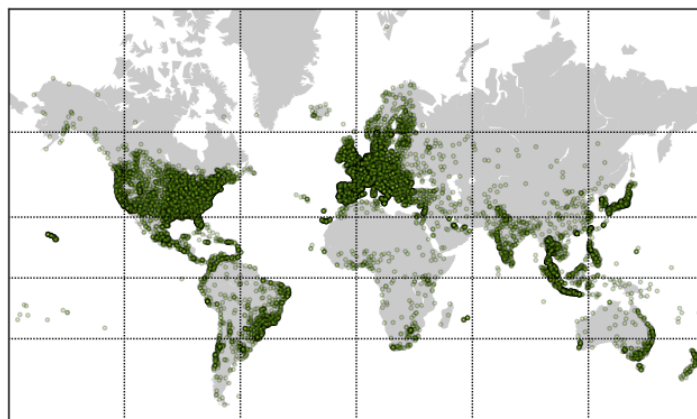




(a) Mayorships.



(b) Tips.



(c) Likes.

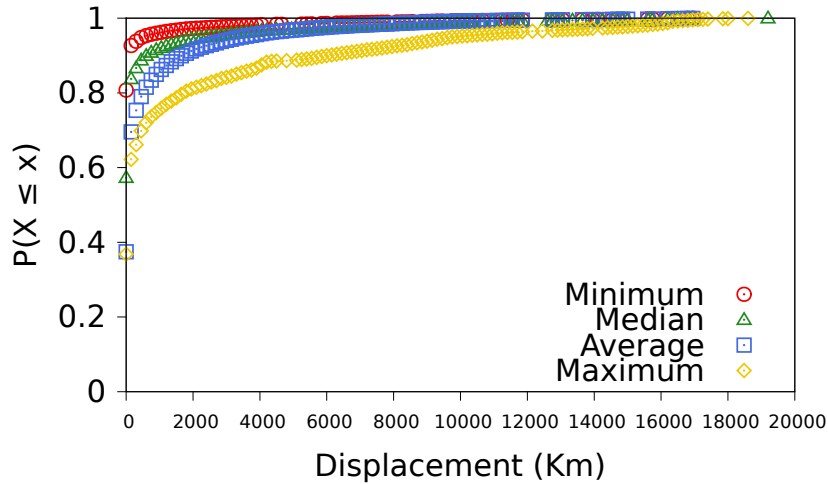
**Figure 3.5.** Global Distribution of the Location-based Attributes across Cities.

of tips and likes, which cover a total of 51,394 and 29,092 unique cities, respectively. The somewhat sparser map built from tips (Figure 3.5(b)) indicates that there are many cities, particularly in Canada, Australia, central Asia and Africa, where, despite the existence of venues and mayors, users do not post tips. The distribution of likes, shown in Figure 3.5(c), reveals an even sparser map, with most activity concentrated in touristic or developed areas, such as USA, western Europe and southeast Asia. We note that a similar map was produced for check ins in [Cheng et al., 2011b]. Even though both datasets were collected at different times, we find that their main areas of concentration do overlap.

We found that the cities with the largest numbers of mayorships tend also to have large numbers of tips and likes, although some interesting differences are worth noting. For instance, mayorships are more concentrated in Southeast Asia, in cities like Jakarta, Bandung and Singapore, which are the top three cities in number of mayorships, jointly having more than 500 thousand mayorships. Tips, in turn, are concentrated in different locations around the Earth: the top three cities in number of tips are New York, Jakarta and São Paulo, with a total of about 600 thousand tips. Likes, on the other hand, tend to be concentrated in venues in the United States, in cities like New York, Chicago and San Francisco, which jointly received around 1 million likes. Just as these attributes, users are also spread through the globe, being New York the city with the greatest concentration of members, around 1.3 million users – which is nearly three times more users than the second city with more users in the rank (Jakarta). Somewhat unexpected, New York is not the city which hosts the highest number of venues, being the fifth in the rank after three asian cities (Jakarta, Bandung, Singapore) and São Paulo.

Next, we analyse the correlation between the numbers of mayorships, tips, likes, users, and venues per city. To that end, we use the Spearman’s correlation coefficient  $\rho$  [Zwillinger and Kokoska, 2000]. In statistics, it is a nonparametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. Thus, the  $\rho$  coefficient varies between -1 and +1, with 0 implying no correlation and correlations of -1 or +1 implying an exact linear relationship. Positive correlations imply that as  $x$  increases, so does  $y$ , while negative correlations imply that as  $x$  increases,  $y$  decreases. Such correlation is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size  $n$ , the  $n$  raw scores are converted to ranks, and  $\rho$  is computed according to the Equation 3.1. We observed that there is a very strongly positive correlation between the number of mayorships and the number of venues across cities ( $\rho = 0.96$ ). Similarly, the correlation is also high between the number of tips and likes in relation





**Figure 3.7.** Cumulative Distribution of Displacements Between Consecutive Tips/Likes Posted per User.

We analyse the displacement between two venues visited in sequence by the user, as indicated by consecutive tips and/or likes of the user. For this analysis, we consider only users with at least two activities, provided that the venues associated with these activities have valid locations – validated by the procedure detailed in Section 3.4, with *quality* of city level or finer granularities. Our dataset contains almost 1.5 million users in this group. For these users, we computed the displacements between consecutive tips/likes by taking the difference between the geographic coordinates of the associated venues. We summarise user activity computing the minimum, median, average and maximum displacement per user. Figure 3.7 shows the distributions of these measures for all analysed users.

Around 37% of the users have average and maximum displacements of 0 kilometres, indicating very short distances (within a few meters). Moreover, 90% of the users have minimum displacements of up to 40 kilometres, which could be characterised as within the metropolitan area of a large city. Also, 70% of the users have an average displacement of at most 150 kilometres, possibly the distance between neighbouring cities. However, there are exceptions. Note that about 10% of the users have a maximum displacement of at least 6,000 kilometres.<sup>11</sup>

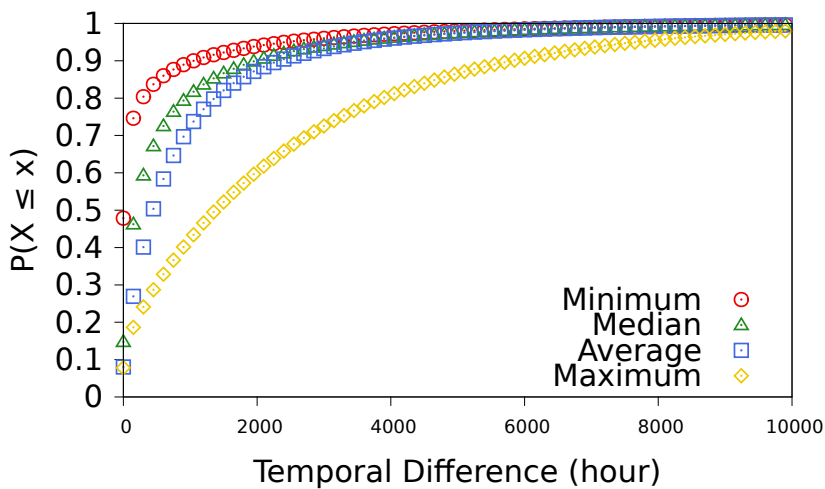
Thus, overall, consecutive tips/likes of a user are often posted at places near each other. Such finding motivates the use of tips and likes in the proposed inference models as these attributes tend to be concentrated in specific regions. Although the exceptions

<sup>11</sup>Note that the maximum displacement between two points in the Earth is the distance between antipodes (two diametrically opposed points) that is about 20,000 kilometres.

represented by huge displacements beyond 6,000 kilometres (a possibly indication of a travel), on average, the majority of user displacements consist of short distances.

## 3.7 Temporal Analysis

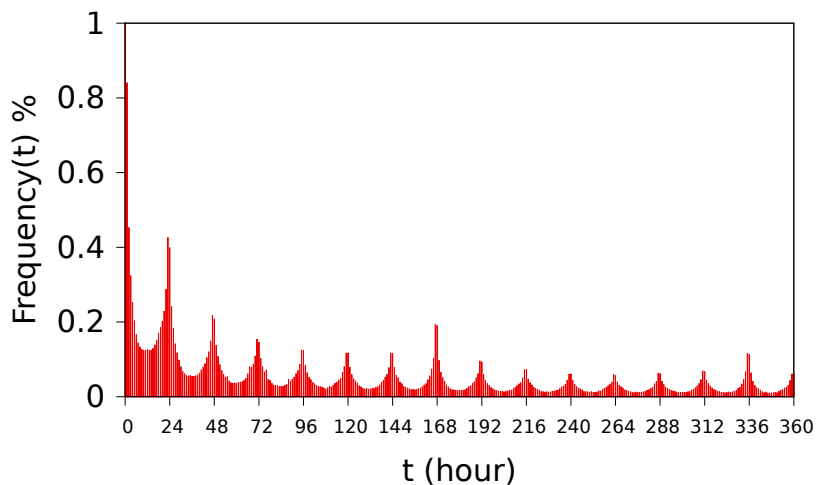
Here, we perform a temporal analysis of user activity in terms of tips and likes aiming to analyse how often users leave tips/likes. To that end, we again make use of the timestamp associated with each tip and like to identify consecutive events in time and measure the interval between such events.



**Figure 3.8.** Cumulative Distribution of Time Interval Between Consecutive Tips/Likes Posted per User.

We start by analysing the time interval between consecutive activities (be it a tip or a like) of the same user. Thus, we consider only users with at least two activities, covering a total of 1,959,644 users. Once again, we summarise user activity by the minimum, median, average and maximum inter-activity times. Figure 3.8 shows the cumulative distributions of these four measures computed for all considered users. We note that the distribution of minimum inter-activity times is very skewed towards short periods of time, with almost 50% of the users posting consecutive tips/likes 1 hour apart. However, on average, median and maximum, users do tend to experience very long periods of time between consecutive tips and likes. For instance, around 50% of the users have an average inter-activity time of at least 450 hours, whereas around 80% of the users have a maximum inter-activity time above 167 hours (roughly a week).

Finally, we analyse how often users return to the same venue for tipping or marking tips as like. That is, we analyse the returning times, defined as the time interval



**Figure 3.9.** Distribution of Returning Times.

between consecutive tips/likes posted at the same venue by the same user. This analysis is focused on 813,606 users, who have at least two tips/likes in the same venue, and cover more than 3 million returns. We here choose to show the distribution of all measured returning times, as opposed to summarising them per user first, so as to compare our results against previous findings of check in patterns [Cheng et al., 2011b]. Figure 3.9 shows the distribution, focusing on returning times under 360 hours, which account for 69.7% of all measured observations. The curve shows clear daily patterns with returning times often being multiples of 24 hours, which is very similar to the distribution of returning times computed based on check ins [Cheng et al., 2011b]. We note, however, that 50% of the measured returning times are within 1 hour, which cannot be seen in the Figure as its y-axis is truncated at 1% so that the rest of the curve could be distinguished. Moreover, out of these observations, 90% of them are at most 10 minutes. Thus, returning times, in general, tend to be very short. If we analyse the behaviour per user, we note that most users have very short minimum returning times, which is below 1 hour for 62% of the users. However, consistently with results in Figure 3.8, on average, median and maximum, users do tend to experience longer returning times. For instance, 52% of the users have average returning times of at least 168 hours.

## 3.8 Summary of this Chapter

In this Chapter, we introduced the main attributes in Foursquare, detailing the crawling methodology and the amount of data collected. We also presented a characterisation of

Foursquare users in relation to their publicly available attributes in the system, namely home city field, history of mayorships, tips and likes, as well as list of friends. Since our focus is on the location information associated with attributes to build models to infer a user's home location, we evaluate the validity and the quality of the locations declared for users and venues. Our findings show that:

- the vast majority of users provide valid locations in the home city field (98% of our dataset);
- despite the large fraction of non-geographic locations associated with venues (27%), about 73% of them are associated with valid geographic locations;
- 90% and 92% of user and venue locations, respectively, are in the level of city or in finer granularities.

We also characterised user activity in the system, showing that mayorships, tips, likes as well as friends are promising sources of information about a user's home location as:

- 30% of the users in our dataset have some activity;
- almost 7 million users in our dataset have a non-empty list of friends;
- there are friends, mayorships, tips and likes spread all over the world, thus inferences using such attributes may be performed at a global scale (not limited to a specific region).

Finally, the study of the patterns followed by users in terms of their attributes can tell much about one's behaviour whereas common observations reflect relevant insights to the development and analysis of home location inference models. Our main conclusions in this direction are:

- users tend to post tips and likes in a limited region, exhibiting short distances among displacements between consecutive attributes. This spatial locality might imply that these attributes are somewhat related with the location where the user lives;
- in general, users are not very active in the system considering tips and likes, as the time interval between consecutive tips (or likes) posted by the same user tends to be very long (in the order of weeks).





# Chapter 4

## Home Location Inference Models

In this Chapter, we start by presenting our problem statement in Section 4.1. In Sections 4.2 and 4.3, we describe the techniques used to make inferences at the city and geographic coordinate levels, respectively. Then, in Section 4.4, we discuss the methodology adopted in our study, focusing on how we defined our *ground truth* and the metrics used to evaluate the proposed inference models.

### 4.1 Problem Statement

Location sharing can be exploited for several new services, such as targeted recommendation and advertisement, as well as to improve existing services (e.g., personalised search). Our focus here is on the implications of location sharing as to privacy violation. Our main goal, in this dissertation, is to investigate whether it is possible to infer, with reasonable accuracy, the home location of Foursquare users using only publicly available attributes. Since Foursquare is in essence a LBSN, several features in the system have embedded geographic information, creating opportunities to explore the spatial aspect associated with each member. In this sense, we aim to aggregate every single piece of location information associated with a user (notably information associated with her lists of mayorships, tips, likes and friends) in an attempt to show the potential of the shared data in revealing virtual footprints, which in turn can guide us to where a user actually lives in the physical world. Moreover, we intend to exploit only publicly available user attributes to show to which extent such inferences can be made by anyone (as opposed to a friend of the user in the system).

The key assumption behind this study is that users tend to have mayorships, tips and likes in venues at the same location (e.g., city) where they live, and they also are likely to have friends living nearby. At first, one might think that the mayorship

locations are perhaps the strongest pieces of evidence about a user’s home location, as the former represent places the user possibly goes very often. Recall that a user only becomes a mayor if she is the most frequent visitor of a venue in the last 60 days. However, tips may also reveal places where a user has been, since when posting tips users are often sharing experiences.<sup>1</sup> Along the same lines, likes may also provide some evidence about a user’s home location, although perhaps not as strong as tips and mayorships. Our conjecture is that users often mark as *liked* tips about physical places where they have been to or intend to go soon. We note however that, despite being intuitive and supported by some of our characterisation findings, the aforementioned assumption is not guaranteed to hold for all users. As discussed in Section 3.6, 10% of the users in our dataset have a maximum displacement of at least 6,000 kilometres between consecutive tips and likes. Finally, as many studies claim, users tend to maintain friendships with people who live near them [Davis Jr. et al., 2011; Backstrom et al., 2010], and thus the home locations of friends are promising sources of evidence with regard to where a user actually lives.

We thus propose several home location inference models, distinguished by the considered attributes (mayorships, tips, likes and friends) and the employed technique (Majority Voting Scheme and K-Nearest Neighbour – both described in Section 4.2). Our inferences are made at different granularities, ranging from the country where a user lives up to the exact geographic coordinate of her residence. Motivated by the privacy implications of this kind of inference, we have focused our efforts on the finest location granularities, namely city and geographic coordinates, as finding the home location of a user in such granularities may represent a major concern when compared, for instance, with the discovery of the country where someone lives. We note that inference results for less specific geographic levels (country and state) are provided in [Pontes et al., 2012b]. Similarly, results for other applications – Google+ and Twitter – exploiting different attributes are discussed in [Pontes et al., 2012a].

## 4.2 Inference Models at the City Level

We here propose different models to infer the home city of Foursquare users in attempt to explore the potential of distinct attributes in revealing where a user lives, as well as to investigate which strategy is more effective to gather all the information available. In our approaches, we make use of the city of venues where a user has may-

---

<sup>1</sup>Although users may post tips at unknown venues to, for instance, inquire about driving directions, operation time, or parking conditions, we believe that this does not occur very often and may possibly signalise an intention for a future visit.

orships, tips, likes, and the cities where the user’s friends live, as indicated in their home city attribute. We first propose a simple Majority Voting Scheme, explained in Section 4.2.1. Next, we introduce a more sophisticated method that is based on the K Nearest Neighbours algorithm, which is described in Section 4.2.2.

### 4.2.1 The Majority Voting Scheme

One of the simplest way of aggregating items to decide one to choose consists in counting votes and taking the majority [David and Jon, 2010]. We refer to this approach as majority voting scheme (MVS). MVS is an efficient strategy that can be very democratic. Not coincidentally, many people adopt such strategy to make decisions as in political elections, or while selecting which place to go with friends, or even to decide which team deserve to win a championship. In all these scenarios, all options eligible to receive votes are considered equal, with no priority or privilege in the dispute to win votes. Thus, in sum, the MVS consists of three major components: the *electable* – which are able to be elected, the *votes* – which count points in favour of some of the electable, and the *winner* – represented by the majority of the votes.

In the context of this dissertation, we apply this strategy to infer the user home location. To that end, we initially select a set of user attributes, which can be the list of mayorships, tips, likes, or friends. We then take all the unique cities to which these attributes are related, being these locations the *electable* candidates for the user home location. Next, we get the *votes*, which are here represented by the location associated to each user attribute (e.g., the city of the venue where a tip was posted). Finally, we count the votes, and if only one majority is achieved, we have a *winner*, which is the city where the user has more attributes and which will be set as his inferred home location. We call such approach **MVS** and it includes 15 distinct models which differ in terms of the attributes used for the inference. We take into account models that exploit single attributes (here referred as *Mayorship*, *Tip*, *Like* and *Friend*), as well as all possible combinations of them (namely, *Mayorship+Tip*, *Mayorship+Like*, *Mayorship+Friend*, *Tip+Like*, *Tip+Friend*, *Like+Friend*, *Mayorship+Tip+Like*, *Mayorship+Tip+Friend*, *Mayorship+Like+Friend*, *Tip+Like+Friend* and *All*, the latter representing the combination of all four attributes). We build combined models aiming at assessing the potential of each attribute to improve the overall accuracy of the single-attribute models and to increase the number of users covered in the inference.

#### 4.2.1.1 MVS Limitations

Note that the MVS has two clear limitations: the lack of enough votes and the possibility of ties (multiple winners). Inferences for users with too few attributes (e.g., just a few mayorships) might be based on weak evidence, and thus, are more susceptible to errors. On the other hand, the MVS may lead to tied results with multiple locations, and therefore a winner must be picked through some other criteria.

In an attempt to overcome these limitations, we have proposed variations for the *MVS* approach. To avoid lack of evidence for the inference, we have suggested the *Filtered\_MVS* approach, which implies in applying *MVS* only for active users regarding the attributes considered in each model. By active users we mean users that have at least a minimum number of attributes (evidence) exploited by the model. This minimum is determined by the parameter *min\_evidence* (e.g., *min\_evidence* equal to 5 for the *Tip* model means that only users with at least 5 tips are considered in the inference). Since Foursquare was relatively new at the time our dataset was collected, many users do not have many pieces of evidence (e.g., many mayorships, tips, likes and friends) to be explored. Thus, we focus only on more active users<sup>2</sup>, conjecturing that they might reflect the behaviour of a larger user population in the near future, as Foursquare continues to grow in popularity. We also explore another parameter, proposed by Davis Jr. et al. [2011], to avoid elections based on weak evidence, including a parameter (*min\_votes*), which is the minimum number of votes for an electable location to be set as inferred home location of a user. The assumption here is that locations elected with a small number of votes do not represent reliable inferences.

We have also treated the cases when *MVS* faces a tie when deciding which location would be set as the user inferred home city. Since only the majority rule is not enough to solve the problem of multiple winners, we have introduced an iterative approach, the *Iterative\_MVS*, to try to reduce the number of winners to only one location. It works as follows. We take the winner locations as the new electable locations; then, as previously explained, the attributes that can be found at one of these locations add votes to them. For the attributes which are not located in any of the electable locations, we calculate the distance between the attribute's location (e.g., the city of the venue where a like was posted) and the electable locations, the smallest distance will define the electable location for which the vote of that attribute will be assigned.

---

<sup>2</sup>This was motivated by the study conducted in [Sadilek et al., 2012], where the authors proposed a probabilistic model of human mobility for Twitter users based on a large sample of highly active users with more than 100 GPS-tagged tweets. That study shows that prediction accuracy degrades gracefully as the amount of data available is limited.

Observe that this process can have many iterations, since a new election may result in a new tie. There may also be intractable impasses in which it is not possible to reduce the number of winners to one. In summary, the assumption behind this strategy is that in face of a tie, for instance between two locations, we believe that probably the most likely location to be set as the user inferred home location is the one which present a denser cluster of attributes located nearby.

Locations	1 <sup>st</sup> Iteration		2 <sup>nd</sup> Iteration	
Belo Horizonte	Locations	# Votes	Locations	# Votes
Belo Horizonte	<b>Belo Horizonte</b>	2	<b>Belo Horizonte</b>	4
New Delhi	<b>New Delhi</b>	2	New Delhi	2
New Delhi	Ouro Preto	1		
Ouro Preto	Tiradentes	1		
Tiradentes				

**Figure 4.1.** Successful Example of *Iterative\_MVS* Approach to Solve Ties among Locations.

Figure 4.1 illustrates how the *Iterative\_MVS* works showing a situation where a user have six attributes associated with four locations: Belo Horizonte, New Delhi, Ouro Preto and Tiradentes. Applying the original majority voting strategy (exemplified by the first iteration), we see that there is a tie between two locations, Belo Horizonte and New Delhi, with two votes each. Thus, in face of this tie we apply a second iteration (based on our *Iterative\_MVS* approach) to try to choose the location (Belo Horizonte or New Delhi) which contains more attributes nearby. Note that the majority of the remaining attributes are located in the same state as Belo Horizonte, thus Ouro Preto and Tiradentes count votes to Belo Horizonte since they are closer to it in comparison to New Delhi. In this case, Belo Horizonte, and not New Delhi, is chosen as the user home city. For this approach, one parameter is considered –  $\alpha$ , which represents the maximum distance (in kilometres) between an attribute’s location and an electable location such that attribute’s vote is accounted for (e.g., if  $\alpha = 100Km$ , an attribute will only be considered as a vote for an electable location in the inference if its location is at most 100km away from it). The example of Figure 4.1 considers *alpha* to be unlimited. The reason why we include such parameter is to avoid that an attribute located very far from all electable locations accounts votes to one of them, eventually adding noise to the inference – for example, in a tie between New York and

Hong Kong, an attribute in Abu Dhabi would not be useful in this election as it is thousands of kilometres far from both electable locations.

We refer to the basic MVS approach, which does not deal with neither weak evidence nor ties as *Original\_MVS* so as to more clearly distinguish it from the *Filtered\_MVS* and *Iterative\_MVS* approaches.

## 4.2.2 The K-Nearest Neighbour Approach

The K-Nearest Neighbour (KNN) is a machine learning classification algorithm that is based on lazy learning, in which the classification model is only approximated locally and all computation is deferred until classification. Basically, each object is represented by a vector containing the values of attributes associated with it. The vector of the target object that we want to classify is compared to all other objects with some similarity (or distance) function to ultimately determine to which class or category that object will be assigned. The  $K$  objects which achieve the highest values for the similarity function are considered to be the closest neighbours of the target object and will be responsible to define its class. The classification takes into account the majority of the neighbours' classes, thus the target object is assigned to the most common class amongst its  $K$  nearest neighbours [Manning et al., 2008].

In light of our present focus, we propose the following KNN based approach to infer the home location of Foursquare users. Basically, we apply inferences considering two different approaches: global (*Global\_KNN*) and local (*Local\_KNN*). Given a user for whom we want to infer the home location (the target user), the *Global\_KNN* approach takes all users of our dataset to serve as neighbours while applying the technique. In contrast, the *Local\_KNN* takes only the target user's friends into account. These approaches were motivated by the location prediction methods proposed by Li et al. [2012] for Twitter. Authors have developed two models which consider the social network to infer a user's location based on different scopes: a more restricted with fewer users as predictors (local), and a broader one with comparatively more predictors (global).

For both approaches, all users are modelled by a  $3\mathbf{V}$ -dimensional vector, where  $\mathbf{V}$  is the number of distinct venues where the **target user** (i.e., the user whose home city is being inferred) has some activity.<sup>3</sup> Each position in the vector expresses the number of mayorships (tips or likes) that the user has in a particular venue. Having all vectors in hand (from the target user and her neighbours), we calculate the similarity

---

<sup>3</sup>Note that the number of unique venues where users have some activity tend to be small. As shown in Section 3.5, the distributions of mayorships, tips and likes per user are heavy tail, with few users having lots of attributes but most users having only a few attributes.

between each neighbour and the target user. Here, we opt to use the cosine measure as similarity metric, although other metrics could be adopted. The cosine similarity function (*cosine\_similarity*) between the target user ( $t_{user}$ ) and a neighbour ( $n_{user}$ ) is defined in Equation 4.1, where  $\mathbf{T}$  and  $\mathbf{N}$  are the  $3\mathbf{V}$ -dimensional vectors that represent  $t_{user}$  and  $n_{user}$ , respectively – being  $T_{v,i}$  and  $V_{v,i}$  references to the  $v$  dimension and the position  $i$  of the vectors  $\mathbf{T}$  and  $\mathbf{V}$ , respectively. Finally, for the  $K$  users with the highest similarity with the target user (the  $K$  nearest neighbours), we apply the *Original\_MVS* to define the inferred user home location. Note that, *Global\_KNN* and *Local\_KNN* do not solve ties since we use *Original\_MVS* to infer the target user home location. However, other approaches (such as *Iterative\_MVS*) could be used.

$$cosine\_similarity(t_{user}, n_{user}) = \frac{\sum_{v=1}^V \sum_{i=1}^3 T_{v,i} N_{v,i}}{\sqrt{\sum_{v=1}^V \sum_{i=1}^3 (T_{v,i})^2} \sqrt{\sum_{v=1}^V \sum_{i=1}^3 (N_{v,i})^2}} \quad (4.1)$$

#### 4.2.2.1 KNN Limitations

One limitation of the *Local\_KNN* approach in the context of this dissertation is related to the eligible users for inferences and candidates for neighbours. Initially, the algorithm considers all friends of the target user as candidates – recall that the target user is the one for whom we intend to infer a home location. Thus, mere acquaintances may participate of the inference and might potentially impact the results negatively. We thus propose two variations of the *Local\_KNN* approach to try to reduce this impact.

Our assumption behind these variations is that *close* friends are more likely to live near the target user’s residence than other people. We here consider a friend as *close* to a target user if they have friends in common. Thus, a friend with some friend in common is a better predictor for inferring the target user’s home city than other friends.

In the first variation of *Local\_KNN*, we only consider eligible for inference users that have at least a minimum number of friends with some friend in common with them. We use the *min\_friends* parameter to determine this threshold. In the second variation, we only consider as neighbours those friends that have at least a minimum number of friends in common with the target user. This minimum number of mutual friends is defined by parameter *min\_mutual*.

### 4.3 Inference Models at the Geographic Coordinate Level

Now we focus on inferring home location at a finer granularity: we aim at inferring the pair of geographic coordinates (i.e., latitude and longitude) of the residence of a Foursquare user. To that end, we explore the history of mayorships, tips and likes, but not friends <sup>4</sup>, analysing the single-attribute models (*Mayorship*, *Tip* and *Like*) and the combined model *Mayorship+Tip+Like*. As the ground-truth is in the level of geographic coordinates, only users who are mayors of venues of the Residence category are considered for inference (i.e., are eligible for inference).

We adopt a two-phase approach to infer the geographic coordinates of a user residence. First, we apply the majority voting scheme (*Original\_MVS* approach) to infer the user’s home city. <sup>5</sup> Afterwards, for those attributes that are located in the inferred city, we compute the coordinates centroid (pair with the mean latitude and longitude of unweighted points in the space) of the venues to which they are associated. The assumption is that all the attributes located at the inferred home city are possibly associated with places that are likely to be near the user’s residence. Thus, the coordinates centroid of these places would be a reasonable estimation of where a user actually lives taking into account places which are probably within a user’s mobility area. To evaluate how good our inference is, we plot the cumulative distribution of the distances between the inferred coordinates and the ones associated with the ground-truth – which represent the exact location of the user’s home. <sup>6</sup> We discuss how we obtain the ground truth in the next section.

### 4.4 Evaluation Methodology

In this Section, we discuss key aspects related to the methodology adopted to evaluate our proposed inference models. We start by discussing how we defined the ground truth for inferences and which users are eligible for our proposed models. Then, we introduce a categorisation of eligible users into three classes and discuss for which of them we are

---

<sup>4</sup>Since the friends location is defined as a default coordinate of their home city, we do not consider friends in our geographic coordinate inference models.

<sup>5</sup>Although we adopted the *Original\_MVS* approach in the first phase of the home location inference at the geographic coordinate level, other approaches (such as *Filtered\_MVS* and *Iterative\_MVS*) could be used.

<sup>6</sup>In the case of Foursquare users with multiple mayorships in venues of the Residence category, we decided to report the lowest distance between them and the inferred geographic coordinate.



able to provide inferences. Finally, we conclude this chapter with a description of our evaluation metrics to measure the effectiveness of our inference models.

To define the ground truth, we first need to determine the inference granularity. For inferences at the city level (or coarser granularities, such as country and state), we take the information provided in the user’s home city field as the location where she lives. Thus, when inferring home city locations, we consider only users whose home city attributes contain valid locations at the city level or at a finer granularity – as validated by *Yahoo! PlaceFinder* in the process described in Section 3.4. Although users are free to enter whatever they want in the home city field, we found that the majority of Foursquare users in our dataset do enter valid locations (see Table 3.2).<sup>7</sup> To provide inferences in a finer degree – as the geographic coordinate level – a more specific ground truth is needed. We then make use of the location associated with a venue of the Residence category for which a user is mayor in Foursquare, since such location is represented with a pin in a world map, thus necessarily having a pair of coordinates associated.<sup>8</sup> Therefore, only users who are mayors of Residence venues are considered in the inferences of the home location in the level of geographic coordinate.

In our experimental evaluation, our first step is to select the *eligible users*, that is, users who are qualified to participate of the inference process for a specific proposed model, or, in other words, users who have filled the public attributes exploited by the model. Users who are not in this group are disregarded since it is impossible to try to infer something about them applying the considered model. After inferring the home location, we group users into three classes. *Class 0* consists of users who have their home location inferred through only one piece of evidence, being in this case the unique alternative to be chosen. *Class 1* and *Class 2*, on the other hand, consist of users who have multiple pieces of evidence (e.g., a tip and a mayorship, or multiple tips). However, the proposed models can define one single location for users who fall into *Class 1*, but it cannot be done for users in *Class 2* since models are unable to decide which location is the best choice for that specific user. In other words, users in *Class 2* are those for which there are ties among the electable locations. We first consider the *Original\_MVS* approach, which does not handle users in *Class 2*, who are thus considered intractable. Later, we evaluate the *Iterative\_MVS* solution which tackles tied results, in which case some users originally in *Class 2* move to *Class 1*. In both cases, only users in classes 0 and 1 are treated by our proposed models.

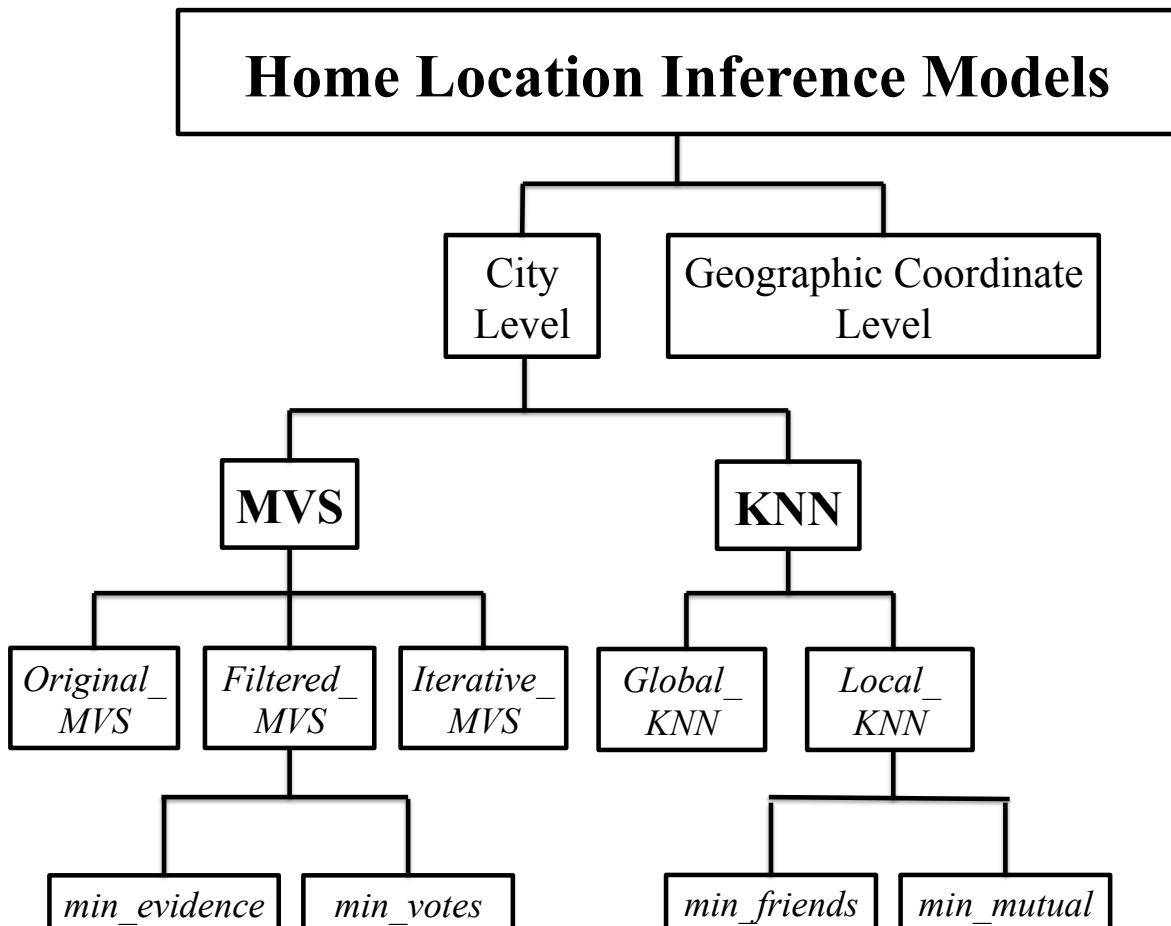
As part of this study, we evaluate and contrast the proposed models in terms of

---

<sup>7</sup>Note that the veracity of this information cannot be confirmed.

<sup>8</sup>Residence is a venue category related to real homes. Their coordinates are omitted in the venue’s page, but are accessible via Foursquare API.

their accuracy, stressing the pros and cons of the different approaches as well as their main limitations. We assess the *accuracy* of the inferences carried out for users of both *Class 0* and *Class 1*, who are called *covered users* and represent those for which it was possible to infer a home location. The accuracy corresponds to the percentage of correctly inferred locations out of all users of each aforementioned class. Moreover, we also report the overall accuracy of each model, considering all the covered users for the inference being applied. Basically, an accurate and efficient model has a reasonable balance between the number of covered users and the inference accuracy. At the same time we look for a model that can generate inferences for a large fraction of users in our dataset, we aim to have high percentages of correct inferences. Such a balance is represented by the product of the number of covered users and the overall accuracy, which results in the total number of *correct inferences*.



**Figure 4.2.** Summarisation of the Home Location Inference Models – Methods Applied and Levels of Inference.

For purposes of summarisation, Figure 4.2 presents a diagram that illustrate the methods applied in this study for the home location inference, as well as the levels of inference considered in the experimental evaluation. As previously detailed, MVS and KNN are the methods used. Both of them branch out into adaptations of the original methods, for the MVS approach the branches are *Original\_MVS*, *Filtered\_MVS* and *Iterative\_MVS*, while for the KNN approach we have *Global\_KNN* and *Local\_KNN*. Note that *Filtered\_MVS* and *Local\_KNN* ramify into two more method variations each with respect to the use of specific parameters. Finally, the inferences are made in the finer-grained levels of city and geographic coordinates, although coarser granularities (as state and country) are also possible.



# Chapter 5

## Experimental Evaluation

This Chapter discusses the experimental evaluation of our proposed models for inferring the home location for the Foursquare users in our dataset, detailed in Chapter 3. We present the results of applying location inferences at both the city (Section 5.1) and geographic coordinate levels (Section 5.2) discussing our main findings when applying the considered techniques for inferring user home location on Foursquare.

### 5.1 Inference Results at the City Level

The results of our experimental evaluation of both MVS and KNN based models for inferring user home city are discussed in Sections 5.1.1 and 5.1.2, respectively. The purpose of applying different techniques for inferring location is to assess which method is more effective in terms of both inference accuracy and user coverage, as defined in Section 4.4.

#### 5.1.1 MVS Inference Models

Recall that, as discussed in Section 4.2.1, we propose several variations of the general MVS model for inferring the user’s home city. These models differ in terms of the attributes considered (mayorships, tips, likes and/or friends) and the approach used (*Original\_MVS*, *Filtered\_MVS* and *Iterative\_MVS*). In the following, we start by discussing the set of experiments performed to explore the considered attributes and assess the impact of key model parameters (Section 5.1.1.1). Next, we present and discuss our most representative results (Section 5.1.1.2), and summarise our main findings and conclusions (Section 5.1.1.3).

### 5.1.1.1 Experimental Setup

For the *Original\_MVS* approach no parameter is required, being the models subject only to the impact of the attributes considered. Our experiments are performed exploring all the possible combinations of attributes in 15 different models. Thus, we propose four single-attribute models which take attributes in isolation (*Mayorship*, *Tip*, *Like* and *Friend*), six models that explore pairs of attributes (*Mayorship+Tip*, *Mayorship+Like*, *Mayorship+Friend*, *Tip+Like*, *Tip+Friend* and *Like+Friend*), four models with combinations of three attributes (*Mayorship+Tip+Like*, *Mayorship+Tip+Friend*, *Mayorship+Like+Friend* and *Tip+Like+Friend*), and a final model which combines all four attributes (referred to as *All*).

The *Filtered\_MVS* approach, in turn, which was designed with the goal of avoiding the use of limited evidence in inferences, has two key parameters, namely *min\_evidence* and *min\_votes*. Recall that *min\_evidence* consists of the minimum number of attributes a user must have to be eligible for inference (e.g., for the model *Tip+Like*, the number of attributes is represented by the number of tips and likes a user has). Parameter *min\_votes*, on the other hand, expresses the smallest number of votes allowed for the inference of a user’s home location, being a location elected only if it achieves at least *min\_votes*. We analyse the impact of each parameter separately by varying it from 1 to 200, keeping the other parameter fixed at 1 (default value that implies in no restriction to the inference task). For this analysis, we consider only the single-attribute models and the *All* model.

Finally, regarding the *Iterative\_MVS* approach, which was proposed to deal with ties in the original majority voting scheme, parameter  $\alpha$  defines the maximum distance allowed for a city to be considered when counting votes to break ties. We here experiment with values of  $\alpha$  equal to 100 km, 200km as well as unlimited distances (i.e.,  $\alpha = \infty$ km). Once again, to study the impact of  $\alpha$  on model effectiveness, we consider only the single-attribute and the *All* models.

### 5.1.1.2 Results

The experimental results for the *Original\_MVS* approach are presented in Table 5.1. For each proposed model, the table shows the number of eligible users for the inference task (i.e., users who have at least one of the attributes required by the specific model), the distribution of users across the three previously defined classes, and the model accuracy (per class and the overall accuracy). Recall that both classes 0 and 1 (users with one majority defined by a single and multiple pieces of evidence, respectively) represent the users covered by the inference task, that is, users for which the model

**Table 5.1.** Summary of the Results Obtained for the *Original\_MVS* Approach for Home City Inferences.

Inference Model	# Eligible	Distribution (%)			Accuracy (%)		
		Class 0	Class 1	Class 2	Class 0	Class 1	Total
<i>Mayorship</i>	1,814,184	40.08	46.55	13.37	49.86	65.05	<b>58.03</b>
<i>Tip</i>	1,589,429	45.62	42.09	12.30	49.83	65.20	57.21
<i>Like</i>	1,194,907	45.76	45.17	9.06	48.25	59.55	53.86
<i>Friend</i>	<b>6,973,727</b>	17.27	61.23	21.50	32.47	58.71	52.93
<i>Mayorship+Tip</i>	2,521,337	35.63	52.27	12.11	49.93	64.07	<b>58.34</b>
<i>Mayorship+Like</i>	2,309,900	35.72	52.35	11.93	49.35	63.05	57.49
<i>Mayorship+Friend</i>	7,013,106	16.79	62.65	20.56	32.87	59.79	54.10
<i>Tip+Like</i>	2,093,119	39.74	49.43	10.83	49.50	62.06	56.46
<i>Tip+Friend</i>	7,082,095	17.16	62.14	20.70	33.95	59.52	53.99
<i>Like+Friend</i>	7,027,402	17.11	62.00	20.89	33.04	58.94	53.34
<i>Mayorship+Tip+Like</i>	2,823,403	33.29	55.55	11.16	49.83	62.74	57.90
<i>Mayorship+Tip+Friend</i>	7,112,548	16.79	63.33	19.88	34.24	60.26	54.81
<i>Mayorship+Like+Friend</i>	7,062,524	16.70	63.22	20.08	33.41	59.84	54.32
<i>Tip+Like+Friend</i>	7,124,687	17.03	62.77	20.21	34.34	59.56	54.18
<i>All</i>	<b>7,153,077</b>	16.69	63.82	19.49	34.61	60.24	54.93

can infer a home city location. Users in Class 2 cannot be covered by the *Original\_MVS* approach as they represent cases where there are ties in the majority voting. We discuss the inferences for these users, using the *Iterative\_MVS* approach, later. Note that the number of correct inferences is derived from the product of the number of covered users and the model’s overall accuracy.

We start by noting that the number of users who have friends in our dataset (eligible users for the models which include the *friend* attribute) is much larger than the amount of users who have the other attributes (mayorship, tip or like). We also observe that the vast majority of the eligible users of all *Original\_MVS* models (79% - 91%) are in classes 0 and 1. Thus, for most users, either they have a single evidence (17-46%) or they have multiple pieces of evidence indicating a single predominant city (42-64%), and thus their home city can be inferred by the simple *Original\_MVS* approach. Not surprisingly, on average, the models produce better accuracies for Class 1 (59-65%) than for Class 0 (32-50%), fact possibly justified by the larger number of sources of evidence supporting the inference.

Considering the overall accuracy, we find that there are not large differences across models. Recall that mayorships are obtained for venues where the user often checks in and thus have a higher probability of being located in the same city where she lives. Thus, as expected, *Mayorship* is the best single-attribute model to infer home city location, although, perhaps surprisingly, *Tip* is only marginally worse. *Friend*, in turn, produces the lowest accuracy among the four attributes when used in isolation, possibly because the attributes associated with visits to places is more strongly related

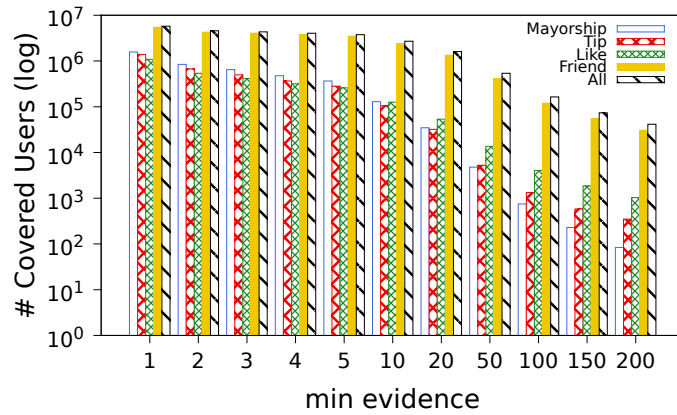
to short geographic distances than the friendship connections. Moreover, the combination of attributes usually hurts the overall accuracy in comparison with the best single-attribute model (being *Mayorship+Tip* the exception), probably because likes and friends add some noise. However, note that, despite a somewhat lower accuracy, these combined models actually cover a much larger user population. For instance, the *Like* model can only be applied to 1,194,907 users, whereas the *All* model is applicable to 7,153,077. Thus, considering the actual number of users for which each model is able to correctly predict the home city, we find that *All* is the best model, with 3,163,386 correct inferences.

Our *Friend* model of the *Original\_MVS* approach is comparable with the model proposed by Davis Jr. et al. [2011] for Twitter users. We observe that we achieved better accuracies while inferring the user home location (52.93% against about 40.47%), although the percentage of intractable users (Class 2 users) is bigger for our approach (21.5% while in Twitter it is around 8.63%). These findings show that Foursquare users are probably more connected to friends that live nearby, thus favouring more accurate inferences for home location. But results may also be justified by the fact that our dataset is much larger than the Twitter dataset considered in the referred study – we consider almost 7 million users for inferences while Davis Jr. et al. [2011] consider 24,767.

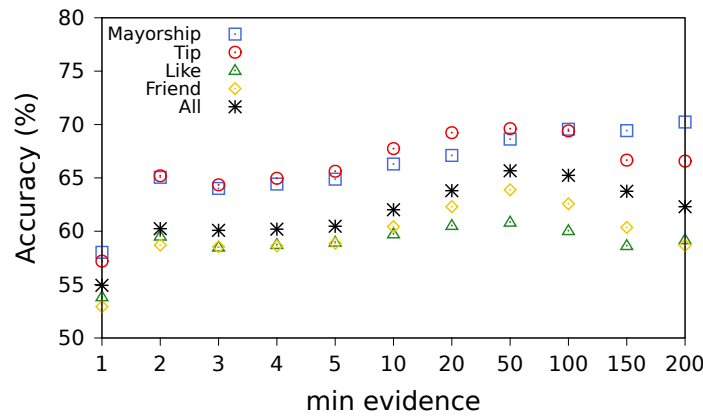
We now discuss the results of the *Filtered\_MVS* approach, which corresponds to the use of the *Original\_MVS* approach restricted to only more *active* users. The criterion to choose the active users depends on the inference model being analysed and the value of *min\_evidence*. For example, for the *Mayorship+Like* model with *min\_evidence* equal to 10, active users are considered to be those who have at least 10 mayorships and/or likes in total.

We start by analysing the impact on the inferences of parameters *min\_evidence* and *min\_votes* separately, showing results for the single-attribute and *All* models. Our aim is to evaluate how each parameter affects user coverage and overall accuracy of each model. Figure 5.1 shows the user coverage in number of users (Figure 5.1(a), with the y-axis in logarithm scale) and the overall accuracy in percentage (Figure 5.1(b), with the y-axis in the 50% to 80% range) achieved with each *Filtered\_MVS* model for various values of *min\_evidence* while *min\_votes* is fixed at 1. We can see that as *min\_evidence* increases, the model becomes more restrictive and fewer users are eligible for inference, since users with less evidence than the stipulated by such parameter are excluded. This indirectly impacts the user coverage by each model which decreases by 20.7% (*All* model) to 52% (*Tip* model) when *min\_evidence* increases to 2, for example. Visibly, the losses are particularly significant for the *Mayorship*, *Tip* and *Like* models, although





(a) User Coverage



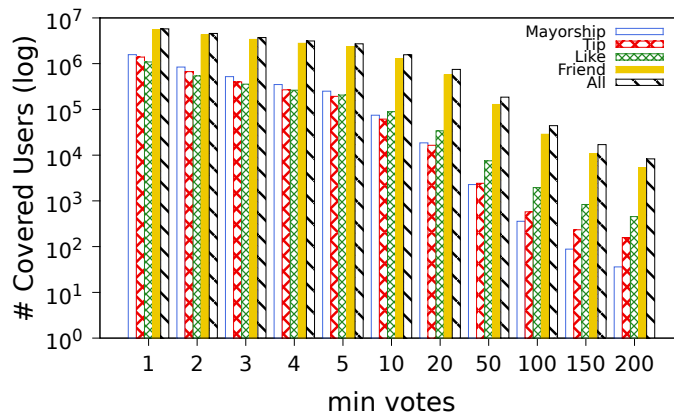
(b) Overall Accuracy.

**Figure 5.1.** Impact of the Parameter  $min\_evidence$  ( $min\_votes = 1$ ).

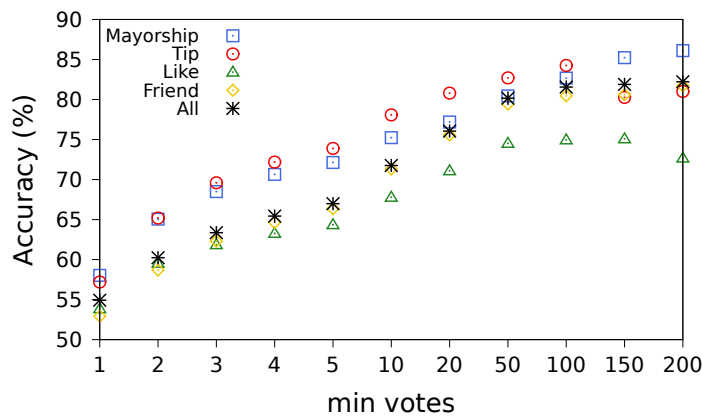
in numbers the *Friend* and *All* model suffer more with more users disregarded of the inference. In contrast, model accuracy tends to improve as  $min\_evidence$  increases: the gains reach 14% for  $min\_evidence$  equal to 2. Except for the *Mayorship* model, the largest overall accuracies (above 60% for all models) are achieved for  $min\_evidence$  equal to 50, with gains over the corresponding *Original\_MVS* models reaching 21.7% (for the *Tip* Model). For larger values of  $min\_evidence$ , accuracies show an inverse tendency, diverging as the inferences become more restrictive. This is possibly because the excess of sources of evidence (mayorships, tips, likes or friends) may reveal a compulsive behaviour of users (like celebrities, brands or travellers) that have such attributes spread in places all over the world. As consequence the number of votes per candidate city tends to be small, and thus elections tend to be based on few votes to

define the inferred home city of the user.

The parameter  $min\_evidence$  is also explored by authors in [Davis Jr. et al., 2011], although they associate it to another parameter that also limits the maximum number of friends, to avoid providing inferences for very popular users (those who have a large social network, celebrity-like users). Thus, comparing our *Friend* model with the referred study for  $min\_evidence$  equal to 20, we observe that we achieve gains in accuracy of 17.7% while Davis Jr. et al. [2011] reach gains of 10% in comparison to the simple MVS strategy. In both scenarios, just few users (about 5% of all eligible) fall in Class 2 and, thus, are considered intractable.



(a) User Coverage.



(b) Overall Accuracy.

**Figure 5.2.** Impact of the Parameter  $min\_votes$  ( $min\_evidence = min\_votes$ ).

We turn our attention now to the  $min\_votes$  parameter. Figure 5.2 shows the impact of varying this parameter on user coverage (in number of users) and overall

**Table 5.2.** Summary of the Results Obtained for the *Iterative\_MVS* Approach for Home City Inference (varying the parameter  $\alpha$ ).

Attributes	# Covered Users			Total Accuracy (%)		
	$\alpha=100Km$	$\alpha=200Km$	$\alpha=\infty Km$	$\alpha=100Km$	$\alpha=200Km$	$\alpha=\infty Km$
<i>Mayorship</i>	1,587,572	1,588,648	1,591,979	57.75	57.73	57.68
<i>Tip</i>	1,406,388	1,407,471	1,411,236	56.97	56.95	56.87
<i>Like</i>	1,094,336	1,094,973	1,098,099	53.65	53.63	53.53
<i>Friend</i>	5,684,134	5,697,343	5,732,271	51.97	51.91	51.73
<i>All</i>	5,971,388	5,984,306	6,019,011	53.97	53.91	53.74

accuracy (in percentage). Note that by varying *min\_votes*, we are indirectly varying *min\_evidence*, which has to be assigned the same value since in order to achieve a certain number of votes, at least the same number of sources of evidence is needed. Similar findings are observed, although the impact on both user coverage (Figure 5.2(a), with the y-axis in logarithm scale) and overall model accuracy (Figure 5.2(b), with the y-axis in the 50% to 90% range) is even stronger. If we increase *min\_votes* from 1 to 2, overall accuracy improves by as much as 14% (*Tip* model), but this comes at the cost of a reduction of 52% in the user coverage. For *min\_votes* set at 20 all models present accuracies above 70%, despite of the greater reductions in coverage (87.0-98.8%). Note that for values of *min\_votes* larger than 100, the improvements in model accuracy are less evident, while some approaches (*Tip* and *Like*) experience accuracy losses. Such losses are possibly justified by the presence of users with lots of tips or likes (much more than the value of *min\_votes*) in venues of different places all over the world, thus presenting many locations with a large number of votes. Once tips and likes may not represent physical visits to places, users are able to have a great amount of these attributes in various spread locations. In sum, we observe that the best approach is to set both *min\_evidence* and *min\_votes* equal to the same value as choosing larger values of *min\_evidence* imply in even more restrictive models (in terms of covered users). We here choose *min\_evidence* (and *min\_votes*) equal to 5 as it reaches a good tradeoff between model accuracy and user coverage.

Once again, the results achieved for *Filtered\_MVS* are comparable to numbers reported in [Davis Jr. et al., 2011]. Unlike the inferences in Twitter, our *Friend* model for Foursquare users shows great improvements in relation to *Original\_MVS* approach. Our gains in accuracy reach 34.7% for *min\_votes* set to 10 whereas Davis Jr. et al. [2011] achieve 12.5%. However, despite such gains, this parameter provides reductions in user coverage, and inferences are made only for about 22% of all eligible users for both systems, thus Class 2 is inflated in 71.8% and 75.8% in comparison with the original majority voting strategies for both Foursquare and Twitter inferences, respectively.

Finally, we discuss the results of the *Iterative\_MVS* approach, for values of  $\alpha$

equal to 100Km, 200Km and unlimited ( $\alpha = \infty$ Km). Results are shown in Table 5.2. Note that results are the same as those for *Original\_MVS* (shown in Table 5.1), with respect to users in Class 0. The differences occur for users in classes 1 and 2. The latter are simply disregarded of the inference task when the *Original\_MVS* models are used, but are considered by the *Iterative\_MVS* method. The use of the *Iterative\_MVS* approach may eventually lead to a single location as inferred home city for those users, thus impacting Class 1. As shown in Table 5.1, when the *Original\_MVS* approach is applied, the fraction of users that fall into Class 2 is relatively small, ranging from 9% to 22% from all eligible users. The *Iterative\_MVS* approach is able to provide inferences to some of these users, while the others remain uncovered as the *Iterative\_MVS* approach is not able to break the ties. This may happen either because the tie is between all locations associated to a user’s attributes (and thus there is no new vote to be accounted for), or due to new ties caused by the iterative strategy (e.g., votes equally distributed among electable locations). Indeed, we find that, for unlimited  $\alpha$ , the fraction of users in Class 2 for which an inference can be made varies from 8% (*Mayorship*) to 19% (*All*). Obviously, this fraction decreases as we reduce  $\alpha$ . In any case, in comparison with the *Original\_MVS* approach, *Iterative\_MVS* leads to an increase in the user coverage by as much as 4.72% (*Friend* model). In absolute terms, this implies in 258,233 more inferences. Moreover, even though model accuracy suffers a slight decrease (up to 2.27% reduction for the *Friend* model with unlimited  $\alpha$ ), we find that the number of correct inferences increases by as much as 67,704 (2.34%).

### 5.1.1.3 Discussion

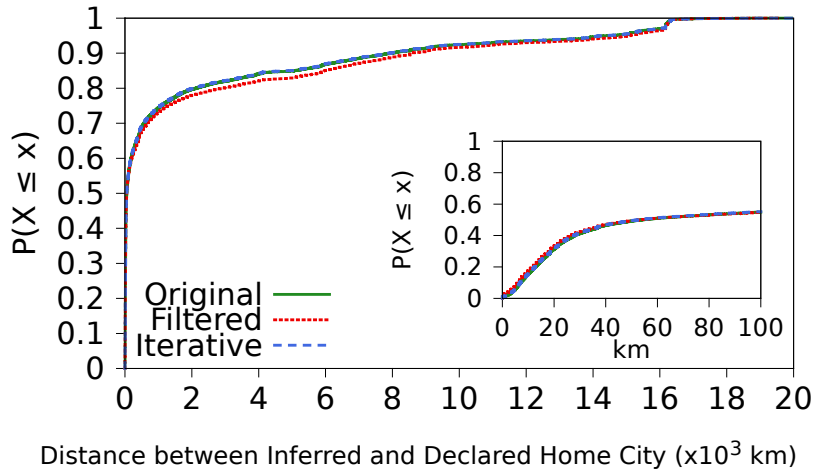
Our experimental evaluation of the use of the MVS-based models for inferring user home city in Foursquare presented satisfactory results. The inferences performed with the *Original\_MVS* models pointed out mayorships as the best public attribute in revealing a user’s home location: the single-attribute model obtained 58.03% of overall accuracy. Friends, on the other hand, stood out as the attribute that contributed the most for user coverage, since almost 7 million users in our dataset have friends. Since our goal is to achieve a good tradeoff between user coverage and overall accuracy, combining all attributes into the *All* model definitely leads to the best solution: the *Original\_MVS* approach was able to produce 3,163,386 correct inferences.

Given that Foursquare was a relatively new social network by the time our dataset was crawled, we conjecture that some of our inference errors may have occurred due to inferences based on weak evidence. The distribution of the number of attributes (mayorships, tips, likes and friends) per user, presented in Figure 3.2 (Section 3.5),

supports this argument, showing that a large share of users considered in our analysis has few friends and low activity in the system. Having that in mind, we applied the *Filtered\_MVS* approach for the active users in the system (those who have at least a minimum number of attributes) finding that such constraint can improve model accuracy. We found that with *min\_evidence* equal to 5 all models achieve an overall accuracy of around 60% (an improvement of 9-15% over *Original\_MVS*), and the maximum accuracies are achieved for *min\_evidence* equal to 50. This strategy also revealed another important source of error of our inferences related to the elections of locations using only a few votes. Thus, imposing restrictions on the minimum number of votes further improves model accuracy, but at the cost of large reductions in user coverage. Thus we here propose to set both *min\_votes* and *min\_evidence* equal to 5, as this value leads to a good tradeoff between both accuracy and user coverage for all models.

In an attempt to treat those users whose inferences were initially considered intractable by the *Original\_MVS* approach, we applied the *Iterative\_MVS* approach. Although the models' overall accuracies remain almost the same compared to the *Original\_MVS*, the number of users for whom we could make inferences increases. For the *All* model and with no limitation on parameter  $\alpha$ , we could infer a location for 260,108 users that used to be disregarded of the inference (27% of such inferences were correct). We observed that the parameter  $\alpha$  has little impact on the results, being the accuracies for these new inferences pretty much the same for different values of  $\alpha$ . In sum, considering the *All* model with unlimited  $\alpha$ , the *Iterative\_MVS* approach produced an additional 70,948 correct inferences (2.24% improvement), compared to the same model using the *Original\_MVS* approach.

Finally, to further analyse the inference errors produced by our models, we computed for each *incorrect inference* the spatial distance between the inferred city and the declared user home city. Figure 5.3 shows the distributions of these distances for the best isolated model in terms of overall accuracy, the *Mayorship* model, for each approach: *Original\_MVS*, *Filtered\_MVS* with the parameters *min\_evidence* and *min\_votes* set to 5, and *Iterative\_MVS* with unlimited  $\alpha$ . Observe that the inner graph is a zoom of the outer graph. We found that all three approaches lead to very similar results, particular *Original\_MVS* and *Iterative\_MVS*. We see that around 49% of the distances for all approaches are under 50 kilometres, which is a reasonable distance between neighbouring (in conurbations) cities and metropolitan areas. Thus, combining these results with the correct inferences produced by the corresponding models, we find that, using the *Mayorship* model, we can correctly infer the city of around 78.6% (85.9% and 78.4%) of the users considered by the *Original\_MVS* (*Filtered\_MVS* and



**Figure 5.3.** Cumulative Distribution of the Distances Between the Declared and the Inferred User Home City for the *Mayorship* model.

*Iterative\_MVS*) approach within 50 kilometres of distance.

## 5.1.2 KNN Inference Models

In this Section we discuss the results for the two KNN-based models proposed to infer the user’s home city – the *Global\_KNN* and the *Local\_KNN* approaches, presented in Section 4.2.2. We describe the set of experiments conducted to evaluate both approaches in Section 5.1.2.1 whereas the results for the experimental evaluation are discussed in Section 5.1.2.2. We conclude this section discussing the possible errors and findings (Section 5.1.2.3).

### 5.1.2.1 Experimental Setup

Parameter  $K$  is intrinsic to the KNN technique since it determines the neighbourhood size. In our home city inference, parameter  $K$  represents the number of neighbours selected. Thus, users with a neighbourhood of size smaller than  $K$  are not part of the inferences. Defining  $K$  to reach the best results for a specific application is not trivial. Thus we experiment with values of  $K$  varying from 1 to 100, and assess their impact on inferences results of both approaches, *Global\_KNN* and *Local\_KNN*.

Recall that, as discussed in Section 4.2.2, we proposed two variations of *Local\_KNN* that restrict the users eligible for inference and the friends that can be considered as neighbours based on the number of friends they have in common. These variations are defined by parameters *min\_friends*, representing the minimum number of friends with at least one friend in common that a user must have to be eligible for

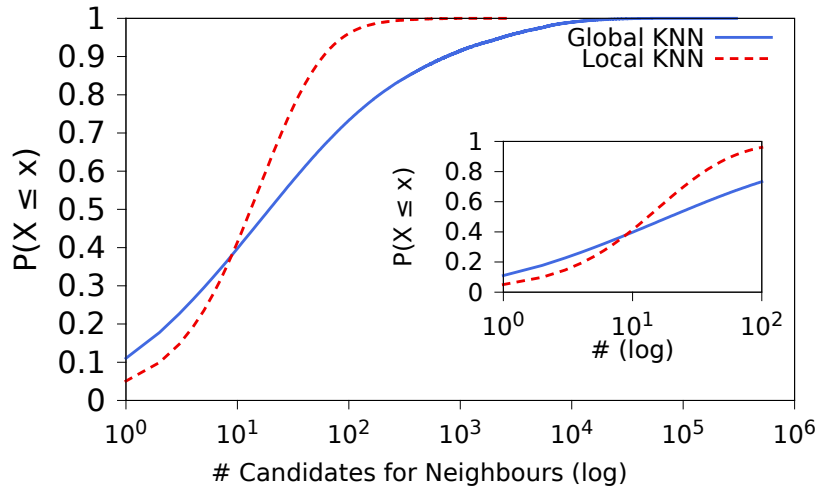
inference, *min\_mutual*, which defines the minimum number of friends in common with the target user a friend must have to be considered neighbour. We evaluate the impact of these restrictions on the performance of *Local\_KNN* by varying both parameters from 1 to 20.

Recall that a user is represented by a  $3\mathbf{V}$ -dimensional vector, as discussed in Section 4.2.2, where  $\mathbf{V}$  is the number of distinct venues where the user target of the inference has some activity. Each position in this vector contains the number of mayorships (tips or likes) that the user has in one of these venues. These values are normalised by taking the logarithm of all values plus one and then dividing them by the maximum logarithm of the corresponding activity (mayorship, tip or like). For example, the numbers of mayorships in different venues are normalised by first taking the logarithm of the number of mayorships plus one and then dividing them by the logarithm of the largest number of mayorships the user has in any venue plus one (note that we sum one to the logarithm to avoid  $\log 0$ , which is not defined).

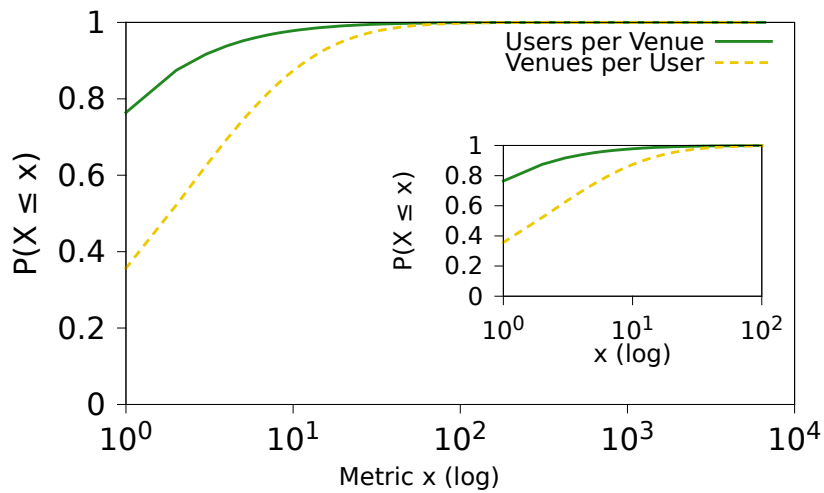
### 5.1.2.2 Results

Note that, considering the user representation adopted, both approaches, particularly *Global\_KNN*, may require a lot of memory space to represent the target user and her neighbours, particularly when the target user has activities in a large number of venues. Indeed, memory consumption grows linearly with the product of the number of venues  $\mathbf{V}$  and the number of neighbour candidates. To support our argument, Figure 5.4 presents the cumulative distribution of the number of candidates for neighbours of both approaches, *Global\_KNN* and *Local\_KNN*, whereas Figure 5.5 shows the cumulative distribution of the number of unique venues where users have some activity (mayorship, tip and/or like). Note the logarithmic scale in the x-axis and that the inner graphs provide a zoom of the distribution curves. Although the distributions are all very skewed, with a large fraction of users having no more than 100 candidates for neighbours (for 96% and 73% of users for local and global approaches, respectively) or associated to fewer than 7 unique venues (around 80% of users), a small fraction of users have very large values for both amounts (up to 300,692 neighbours and 5,762 venues per user), resulting in a product of the number of venues and the number of candidates that reaches the order of hundreds of billions. Ultimately, this implies in very large memory requirements, which indirectly makes the time required for performing the inference too long. Thus, due to practical reasons, we choose to disregard the 0.7% users with largest memory demands.

Inference results for *Global\_KNN* and *Local\_KNN* approaches are presented in



**Figure 5.4.** Cumulative Distribution of the Number of Candidates for Neighbours for both *Local\_KNN* and *Global\_KNN* Approaches.



**Figure 5.5.** Cumulative Distribution of the Number of Users per Venue and Venues per Users.

**Table 5.3.** Results of the *Global\_KNN* Approach for Home City Inference.

<i>K</i> value	# Eligible	Distribution (%)		
		Class 0	Class 1	Class 2
1	<b>2,201,874</b>	100.00	0.00	0.00
2	1,957,296	0.00	36.47	<b>63.53</b>
3	1,807,532	0.00	60.61	39.39
4	1,699,322	0.00	72.13	27.87
5	1,615,677	0.00	78.69	21.31
10	1,356,421	0.00	89.70	10.30
20	1,099,294	0.00	94.63	5.37
50	782,976	0.00	97.65	2.35
100	579,616	0.00	<b>98.77</b>	1.23

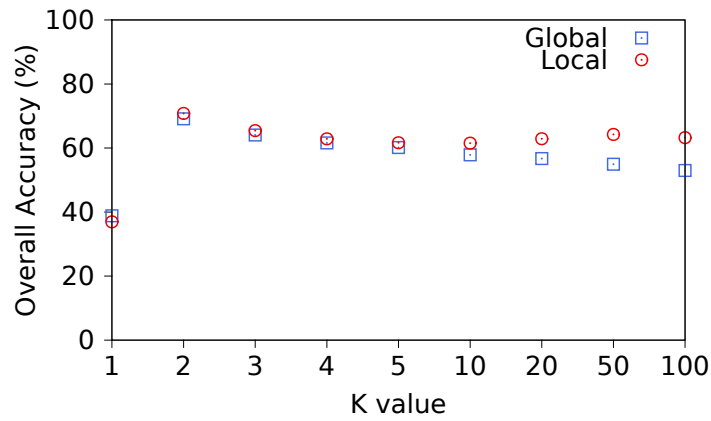


**Table 5.4.** Results of the *Local\_KNN* Approach for Home City Inference.

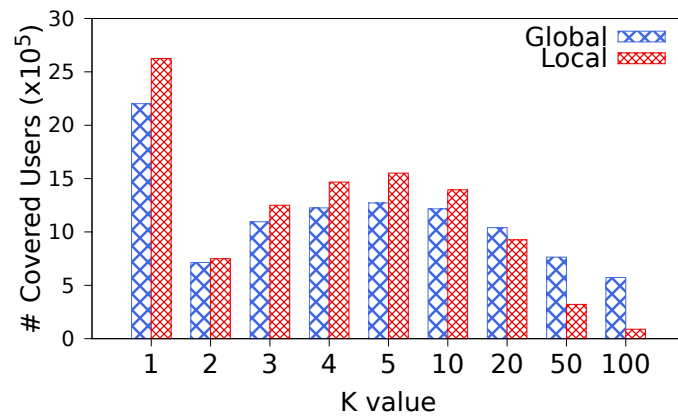
<i>K</i> value	# Eligible	Distribution (%)		
		Class 0	Class 1	Class 2
1	<b>2,626,047</b>	100.00	0.00	0.00
2	2,492,320	0.00	30.06	<b>69.94</b>
3	2,363,316	0.00	52.89	47.11
4	2,237,508	0.00	65.58	34.42
5	2,116,596	0.00	73.30	26.70
10	1,613,324	0.00	86.48	13.52
20	994,759	0.00	93.11	6.89
50	330,511	0.00	97.15	2.85
100	90,674	0.00	<b>98.15</b>	1.85

Tables 5.3 and 5.4, respectively. We show, for varied values of  $K$ , the number of eligible users for the inferences as well as the distributions of users among the three classes. Recall that, similar to the *Original\_MVS*, we do not treat tied results (i.e., users in Class 2) with the KNN based approaches, although a strategy similar to *Iterative\_MVS* could be applied. The eligible users for both approaches need to have some activity in the system (some mayorship, tip or like). For *Global\_KNN*, eligible users also need to have users with activities in common venues, whereas for *Local\_KNN* the requirement is to have at least one friend in the system. Note that for  $K$  equal to 1, all users have a single source of evidence (one neighbour) to be used in the inference task, that is, all users are in Class 0. As we can see the number of users eligible for inference by the local approach is larger than those eligible for the *Global\_KNN* approach. This basically implies that there are users in our dataset with activities in venues that no other user has explored. This observation is illustrated in Figure 5.5, which also shows the cumulative distribution of the number of users per venue in our dataset. We clearly observe that the majority of the venues (76.4%) have activities of only one user, and only 3% of them have mayorships, tips and/or likes of more than 8 users. Note also that the number of eligible users decreases as  $K$  increases. However, the fraction of users that are not covered by the inference (Class 2) decreases significantly as  $K$  increases, reaching only 1.23% and 1.85% of all eligible users for  $K$  equal to 100, for *Global\_KNN* and *Local\_KNN* respectively.

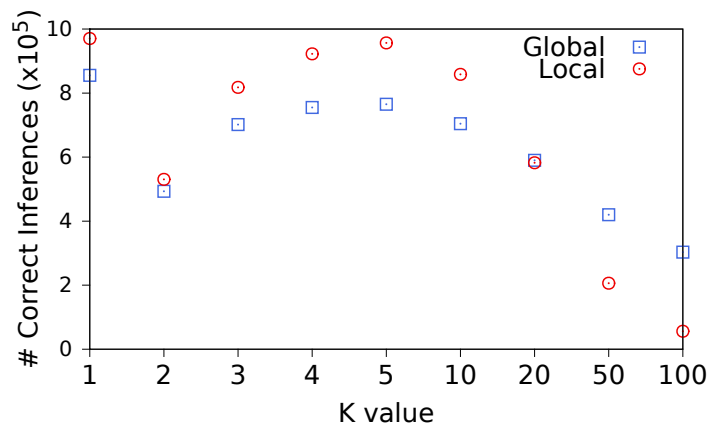
To evaluate the effectiveness of the proposed approaches, Figure 5.6 shows the overall accuracy and user coverage (in number of users) for varied values of  $K$ , along with the number of correct inferences. As shown in the figure, both approaches have similar behaviour with respect to all metrics. Both approaches have very close accuracy (in Figure 5.6(a)) for values of  $K$  up to 10: the lowest accuracy is achieved with  $K$  equal to 1, while the best results are achieved for  $K$  equal to 2. For values of  $K$  larger than 10, the two curves diverge, and *Local\_KNN* outperforms *Global\_KNN*. This divergence is



(a) Overall Accuracy.



(b) User Coverage.



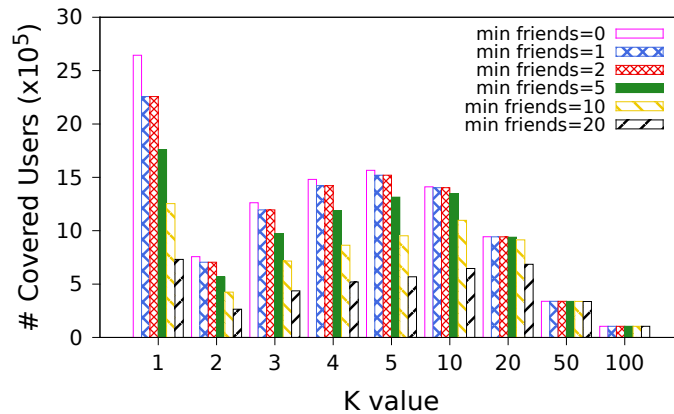
(c) Number of Correct Inferences.

**Figure 5.6.** Impact of the Parameter  $K$  for *Global\_KNN* and *Local\_KNN* Approaches.

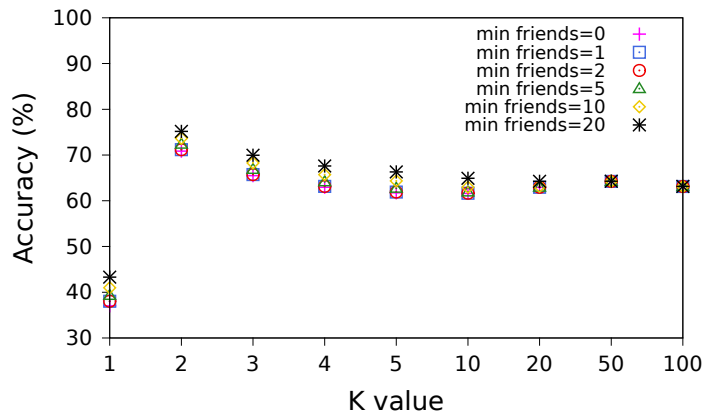
also observed for user coverage, in Figure 5.6(b): *Local\_KNN* presents greater values in comparison with *Global\_KNN* for values of  $K$  up to 20, but the relation reverses for larger values. Once again, this is explained by the distribution of the number of candidates for neighbours of both approaches, with *Local\_KNN* presenting a high concentration (90%) in amounts of up to 60 candidates, while *Global\_KNN* has a less skewed distribution with 37% of users having more than 60 candidates. Note that user coverage sharply decreases for  $K$  from 1 to 2, somewhat explained by the large percentage of intractable users in this scenario and the fact that, for  $K$  equal to 1, all users are in Class 0 and thus are covered by the inference models. After this point, another peak is reached for  $K$  equal to 5. Finally, Figure 5.6(c) also shows that, in terms of the total number of correct inferences, *Local\_KNN* outperforms *Global\_KNN* for  $K$  lower than 20, reaching the largest number of correct inferences for  $K$  set at 1 and 5. Both approaches have similar performance for  $K$  equal to 20, while, for larger values of  $K$ , the global approach becomes the best one. Thus, we choose 5 as an good value for  $K$ , since it establishes a reasonable trade-off between overall accuracy and user coverage, and produces a large number of correct inferences (764,988 for global approach and 956,715 for local).

We now investigate the impact on *Local\_KNN* of using as neighbours users (friends) that have friends in common with the user for whom the inference is made. We do so by evaluating the impact of parameters *min\_friends* and *min\_mutual* separately, varying them between 1 and 20, and comparing results against setting both parameters to 0.

Figure 5.7 shows results for varied values of *min\_friends* and  $K$ , keeping *min\_mutual* equal to zero. Note that the restriction on the number of eligible users leads to gains in overall accuracy but at the cost of large reductions in user coverage in comparison with the basic *Local\_KNN* (no restriction). As we can see, as *min\_friends* increases the model becomes more restrictive, since this parameter imposes a limit on the minimum number of friends with some mutual friend a user must have to participate of the inference. Thus, fewer users are eligible for the inferences. As expected, this causes reductions in user coverage (shown in Figure 5.7(a), with the y-axis in number of users), despite gains in overall accuracy (Figure 5.7(b), with the y-axis in the 30% to 100% range). For example, results for *min\_friends* equal to 1 and 2 (overlapped in both graphs) show gains in accuracy of up to 2.7% for all values of  $K$  tested and reductions in user coverage lower than 15%. Also, note that larger values of *min\_friends* lead to more noticeable impact, with improvements in accuracy of 17% in relation to the *Local\_KNN* but reduction in user coverage of at most 72%. However, like observed for the basic *Global\_KNN* and *Local\_KNN* (with no restrictions),  $K$  equal to 5 pro-



(a) User Coverage.

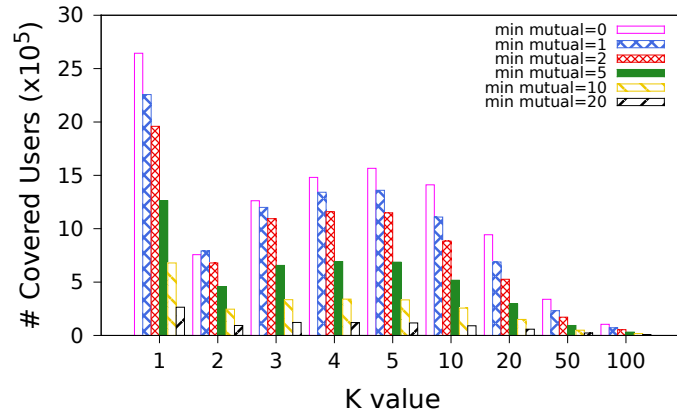


(b) Overall Accuracy.

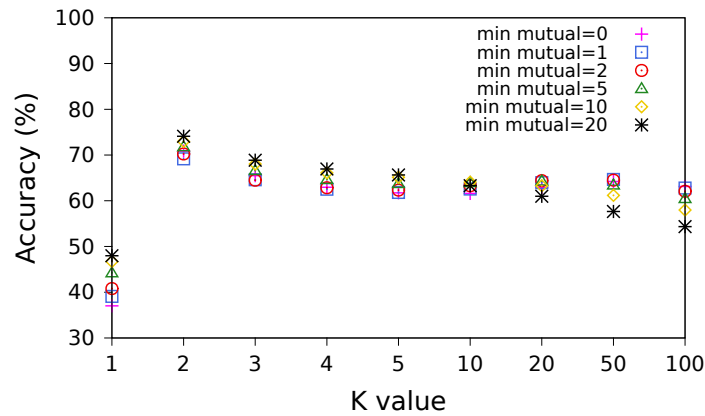
**Figure 5.7.** Impact of the Parameter  $min\_friends$  ( $min\_mutual = 0$ ).

vides one of the most efficient models with accuracies of 61.9% to 66.31%, although the big losses in user coverage of up to 63.7% in relation to the local approach (with no restriction) – specially for larger values of  $min\_friends$ . We here choose  $min\_friends$  equal to 5 as it leads to a reasonable balance between accuracy gains and losses in user coverage (producing 824,415 correct inferences for  $K$  equal to 5).

Regarding parameter  $min\_mutual$ , Figure 5.8 shows results for various values while keeping  $min\_friends$  equal to zero. We observe in Figure 5.8(b) that, similarly to  $min\_friends$ , when compared with the basic  $Local\_KNN$  (no restriction), the parameter  $min\_mutual$  has a positive impact on the model accuracy, with gains of up to 30%, but a detrimental impact on user coverage, that decreases by as much as 94%, as shown in Figure 5.8(a). As  $min\_mutual$  increases, fewer users are able to participate



(a) User Coverage.



(b) Overall Accuracy.

**Figure 5.8.** Impact of the Parameter  $min\_mutual$  ( $min\_friends = 0$ ).

of the inferences as a few users might meet the minimum number of mutual friends stipulated by  $min\_mutual$ . On the other hand, inferences based on such restricted set of friends leads to more accurate results. Note that the lowest losses occur for  $K$  equal to 2 and the biggest gains appear for  $K$  set at 1. Somewhat surprising, the gains in accuracy oscillate with peaks for  $K$  set in 1, 5 and 100. Thus, once again 5 is chosen as a good value for the parameter being analysed as it represents an average behaviour in comparison to the other values of  $min\_mutual$  tested (for  $K$  equal to 5, the *Local\_KNN* model with  $min\_mutual$  set at 5 reaches 437,491 correct inferences).

### 5.1.2.3 Discussion

As we observed in the experiments performed for the KNN technique, the *Global\_KNN* and *Local\_KNN* present very similar results. However, the local approach outperforms the global one, producing a larger number of correct inferences, especially for  $K$  set to 1 or 5. Moreover, the *Global\_KNN* takes too long to execute and suffers with space constraints, as users may have up to 300,692 candidates for neighbours against the maximum of 2,677 candidates observed for *Local\_KNN*. Regarding the parameters *min\_friends* and *min\_mutual*, we observed that both constrain the *Local\_KNN* approach, providing a more accurate strategy to select neighbours from the candidates. For both cases, the assumption is that friends with other friends in common have a higher chance to live nearby. However, the increase in overall accuracy provided by such restrictions comes at the cost of large reductions in user coverage. We choose to set each parameter in 5 as it represents a reasonable balance between both metrics and yields large numbers of correct inferences.

Comparatively, the KNN based models cover a much smaller number of users in relation to the MVS based models. This is due to the characteristics required for both target users and candidates for neighbours. In both *Global\_KNN* and *Local\_KNN*, eligible users must have some activity (mayorship, tip or like) in the system and need to have at least  $K$  candidates for neighbours. Also, neighbours must present some activity and have some venue in common with the target user (for the global approach) or be friends with him (for local). Thus, many users are disregarded of inferences as they cannot achieve these requirements. Consequently, although KNN models may achieve greater accuracies, the number of correct inferences provided by MVS models is much larger (especially for *All* models, which aggregate all possible attributes into one model).

## 5.2 Inference Results at the Geographic Coordinate Level

In this section we discuss representative results for our inference models at the geographic coordinate level, which are applied only to users who are mayors of their own Residential venues. We start by presenting our experimental setup in Section 5.2.1, and then discuss our results in Section 5.2.2. We summarise our main findings in Section 5.2.3.

### 5.2.1 Experimental Setup

Since only venues have location information available at the geographic coordinate level, only inferences for users who are mayors of their own homes may be evaluated in such granularity. In total, there are 832,191 users in our dataset (6.13%) that are associated to Residential venues through mayorships with valid city names and geographic coordinates, as validated by *Yahoo! PlaceFinder*. Our experimental evaluation of inferences at the geographic coordinate level is performed over these users for four different models, namely the single-attribute models (*Mayorship*, *Tip*, and *Like*) and the model that combines all three attributes (*Mayorship+Tip+Like*). Recall that we do not exploit the friend attribute in this inference task since the geographic information associated with friends are not available at the geographic coordinate level. To evaluate how good our inferences are, we plot the cumulative distributions of the distances between the inferred coordinate and the one associated with the user home (ground truth).

### 5.2.2 Results

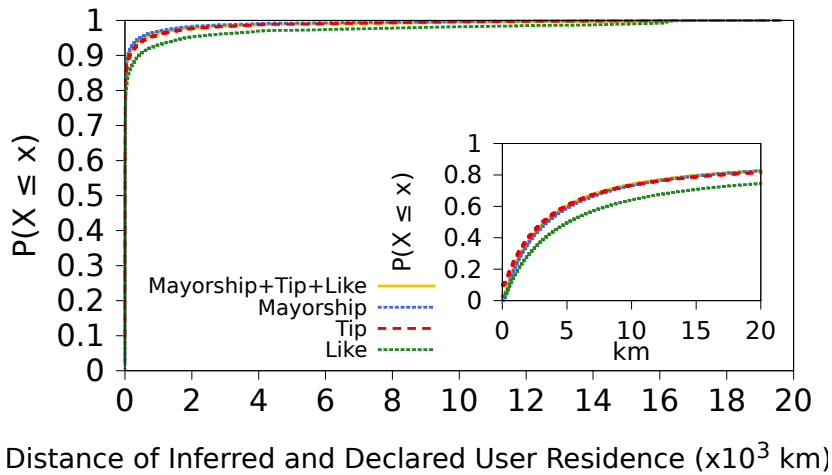
Table 5.5 summarises the results of the inferences at the residence level for the proposed models. For each model, we show the number of eligible users, which represent users in our dataset that have the required attributes (mayorships, tips and/or likes) and an associated ground truth (i.e., valid geographic coordinates associated with a Residence venue where the user is a mayor). We also show the distribution of users across the three classes: once again our inferences are performed for users in classes 0 and 1, for which a single majority location can be defined.

**Table 5.5.** Summary of the Results Obtained for the Home Inference in the Geographic Coordinate Level.

Attributes	Eligible	Distribution (%)		
		Class 0	Class 1	Class 2
<i>Mayorship</i>	562,922	29.68	56.29	14.04
<i>Tip</i>	467,915	33.24	52.56	14.20
<i>Like</i>	337,653	34.02	55.21	10.76
<i>Mayorship+Tip+Like</i>	<b>695,012</b>	<b>16.58</b>	71.66	11.77

As expected, the model that considers all activities aggregates the largest number of users (695,012 in total): mayorship is the most common activity among those users, and likes the least used. Note also that, in comparison with the three single-attribute models, the fraction of inferences made using the *Mayorship+Tip+Like* model based on a single source of evidence (Class 0 users) is much smaller. Moreover, the fraction

of tied results (Class 2 users), which are not treated by our approach, is only slightly larger than the smallest one obtained with any considered model, which is obtained with the *Like* model. This implies that the *Mayorship+Tip+Like* model produces the best results in terms of fraction of covered users. Moreover, note that the numbers of eligible users for the single-attribute models are much smaller than the number of users eligible for inference by the *Mayorship+Tip+Like*. This implies that the combination of the three attributes is the best solution in terms of the number of covered users (613,240).



**Figure 5.9.** Cumulative Distribution of the Distances Between the Declared and the Inferred Geographic Coordinates of the User Residence.

Figure 5.9 shows the cumulative distribution of the distances between the declared and inferred geographic coordinates of the residence for all inferences made – the inner graph is a zoom in the outer one. Recall that our inferences have two steps: (1) we use the *Original\_MVS* approach to infer the user home city, and (2) we compute the mean coordinate of all venues with some activity of the user which are located in the inferred city. Even though any of the considered models for inferring user home city could have been adopted in step 1, we here consider only the use of the *Original\_MVS* technique. Figure 5.9 shows one curve for each single-attribute model as well as one curve for the *Mayorship+Tip+Like* model. Although the curves are all very similar, note that the *Like* model presents the worst results, producing location inferences that are farther away from the declared user residence than the other models. Considering the other three models, we find that around 60% of the inferred locations are within a radius of up to 5 kilometres of the user residence, while 22-27% of the inferred locations are within only 1 kilometre of the user residence. Considering the much larger user coverage, we find that the *Mayorship+Tip+Like* model is the best considered approach.



### 5.2.3 Discussion

Recall that our proposed models are applied for those users who have a ground truth available in the level of coordinates and have some activity in the system, represented by mayorships, tips and/or likes. In this sense, the *Mayorship+Tip+Like* is the model that covers the largest number of users. In relation to the accuracy of our inferences, we find that all considered models present similar results, except for the *Like* model which is slightly worse. Thus, disregarding the *Like* model, all other models achieved very interesting results showing that we are capable of inferring the exact location where a user lives for around 60% of users within a radius of just 5 kilometres, representing an area within the same neighbourhood. We believe that the greater distances amongst the coordinates referred to the user residence and the point inferred by our proposed models are possibly justified by inferences performed based on a few sources of evidence (only a few mayorships, tips and/or likes), or also due to the fact that users may have mayorships, tips or likes in places (venues) far from their residences while traveling, thus including noise to our inference results.



# Chapter 6

## Conclusions and Future Work

The availability of geographic information associated with data shared by Foursquare users raises various concerns about privacy violation. The knowledge of locations related to a user (where she has been, for example) may facilitate inferences about behavioural patterns and habits. Going one step further, gathering all pieces of information leaked from publicly available sources may reveal the location where an individual lives. In this dissertation, we have proposed several models to infer the home location of Foursquare users exploiting public attributes with embedded geographic information. Our goal was to show the potential of each attribute in uncovering the user's home location while exploring the effectiveness of different techniques for the inference task. We considered inferences at both the city level and at the finer granularity of geographic coordinates.

### 6.1 Main Contributions

The most valuable contribution of this dissertation is that it is a pioneer study in trying to infer the user home location in Foursquare only exploiting public available attributes associated with users. We have proposed several inference models which differ in terms of the attributes considered and the technique applied. These different approaches were exploited aiming to investigate the potential of different attributes, used in isolation or jointly, in revealing sensitive information of a user, as well as to detect which inference technique would be more accurate for the context of the study developed in this dissertation.

The proposed models are generic in relation to the spatial degree of the inference, capable of generating responses for the inferred home location of a user in levels that vary from country to geographic coordinates. However, since our main motivation here

is about privacy violation, we have focussed our experimental evaluation on the finest granularities – city and the exact coordinates of the house of a user, as they represent the most concerning levels of inferences. The effectiveness of the proposed models was assessed through the analysis of the percentage of correct inferences and user coverage (i.e., the number of users for whom the model could infer some place as home location). Our goal was to determine the approach that delivers a good tradeoff between both metrics.

Our evaluation of the proposed model indicated that the mayorship attribute is the most accurate in revealing a user’s home location, whereas friends and likes are the worst, i.e., the vanity of becoming a mayor has a privacy cost. However, the model that jointly exploits all user attributes is indisputably the best in terms of user coverage which ultimately leads to the largest number of correct inferences. We have also found that inferences applied in a select group of active users have more chances to reach success, showing how our proposed models may perform in Foursquare in a near future (when the system would accumulate more information about members). For highly active users, this strategy avoids inferences based on weak evidence, reaching accuracies of about 72% for the *Mayorship* model with parameters (*min\_evidence* and *min\_votes*) equal to 5 for the city-level inference. We also lead to ties in the majority voting in the *Iterative\_MVS* approach which is the model that provides the highest number of correct inferences among all models proposed in this dissertation (3,234,334 for *All* model and  $\alpha$  unlimited).

Considering the KNN based models, we find that the *Global\_KNN* is much more costly than *Local\_KNN*, although they are very similar regarding overall accuracy and user coverage. Thus, we choose local approach as the best KNN model, which may become more accurate through the use of parameters *min\_friends* and *min\_mutual* in spite of the reductions in user coverage. Now, comparing the MVS and KNN based approaches, we find that the KNN is worse in our context. Besides the fact that KNN-based models are more costly in terms of execution time, they also cover a smaller fraction of users. Thus, even with accuracies slightly larger than MVS-based models, the latter present a larger number of correct inferences – nearly three times more. Thus, we found that the MVS technique is the most accurate for home inferences in Foursquare.

Finally, our most refined inferences at the geographic coordinates level revealed that we are able to uncover the exact location where 60% of users live in a radius of 5Km, which represents a serious concern from the privacy violation perspective.

## 6.2 Limitations

Although the study carried out presented a wide comprehensiveness about the system analysed and the use of the public features provided, it also contains some limitations. As already discussed in this dissertation, the young age of Foursquare at the time we crawled our dataset was a complicating factor with respect to the amount of data we gathered for each user. Since most user accounts we collected were from recently associated members, our proposed models for inferring the home location of Foursquare users suffered with inferences based on weak evidence due to lack of information (we tried to overcome this limitation with the *Filtered\_MVS* models applied for active users). In the same vein, as only few users are mayors of their own houses, venues of the Residence category, our inferences in the finest granularity, the geographic coordinate level, were applied for a reduced set of users of our dataset.

The impossibility of using check ins as an attribute to be exploited by our inference models is an obvious limitation, as check ins are possibly the most explored Foursquare feature by its users. However, we note that check ins are a private attribute, visible only to the user's friends. Thus any privacy breach through check ins might be more contained, in comparison to the information revealed by tips, mayors, and likes, which are visible to anyone.

At last, although Foursquare developers are continuously improving the system to avoid misuse of its functionalities, it is known that fake attributes may exist. Since the detection of users who fool the system is not trivial, we do not address this problem in our study.

## 6.3 Future Work

There are several directions which this work can evolve, specially aiming at overcome the limitations mentioned. Since Foursquare was a recently launched social network at the time we collected it, we agree that a new collection would provide a wealthier dataset. Now, Foursquare is more mature and it is likely to present more pieces of evidence per user, and also a larger number of residential venues – implying in a greater set of users to apply the most concerning inferences at the geographic coordinate level.

One promising effort is to include new features to the proposed models in attempt to reach results even more accurate and capable of covering a larger fraction of users. One possibility is to explore the mobility area of users in Foursquare, an aspect that can be valuable in revealing the users home location, especially the exact coordinates of one's home – an initiative already in progress. Along the same line, another option is to

explore check ins, as they represent the most typical attribute of Foursquare carrying a powerful meaning associated with physical visits of users to specific places previously registered in the system. Check ins may be exported to other social networks, thus we could crawl this attribute via Twitter. But once again, this implies in other limitations as it is not guaranteed that the entire user history of check ins would be exported. Thus, include check ins in our inferences may possibly not improve results.

In a very low level, some new approaches or combined tactics can be tested regarding the methods used for the home location inference. Machine Learning algorithms as Support Vector Machine – SVM and Weighted Majority Algorithm – WMA, for instance, are suitable for the problem in focus. Also, new strategies to break ties may be tried based on characteristics of the social network or of similar users present in the system; as well as different ways to calculate distances (Euclidean distance,  $L_1$  norm) may be used to compare the accuracy of the results. For the presentation of results, more intuitive considerations may be taken into account as the size of a city (e.g., infer that a user lives in a very small town seems to be more accurate/specific than suggest that she lives in a huge city) – here, the size of a city can be measured through the number of citizens or by the physical area.

Another direction for future work is the design of inference models to discover other pieces of information about the users such as their preferences, interests and tastes. Such inferences along with home location inferences can provide useful sources of data for target information services such as personalised recommendation and advertisement. In terms of privacy awareness, the implementation of an application with all the gathered public information about a user, together with the possible inferences about him, would certainly increase users' consciousness about their own exposure and personal information leakage in the system.

Note that our motivation for the study developed in this dissertation extrapolates the Foursquare system. There are currently several web and mobile applications, social networks in particular like Google+ and Instagram, growing at incredible rates and encouraging huge masses to share everything about themselves and to connect with even more people in the world. In times when the individual privacy is notably a major concern, preserved for many though coveted for millions while watching reality shows or prying a friend's profile, many questions arise, stimulating us to investigate possible privacy breaches, sources of information leakage, and opportunities for the application of inferences.

# Bibliography

- Alexa (2013). Top global sites. <http://www.alexa.com/topsites/global>. Accessed: 2013-03-01.
- Annavaram, M., Jacobson, Q., and Shen, J. (2008). HangOut: A Privacy Preserving Social Networking Application. In *Proceedings of the International Workshop on Mobile Device and Urban Sensing*, St. Louis, MO, USA. ACM.
- Backstrom, L., Sun, E., and Marlow, C. (2010). Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, NC, USA. ACM.
- Barkhuus, L. and Dey, A. (2003). Location-Based Services for Mobile Telephony: a Study of Users' Privacy Concerns. In *Proceedings of the 10th International Conference on Human-Computer Interaction*, Crete, Greece. ACM.
- Benisch, M. (2011). *Using Expressiveness to Increase Efficiency in Social and Economic Mechanisms*. PhD thesis, Carnegie Mellon University.
- Berjani, B. and Strufe, T. (2011). A Recommendation System for Spots in Location-Based Online Social Networks. In *Proceedings of the 4th Workshop on Social Network Systems*, Salzburg, Austria. ACM.
- Brown, B., Taylor, A. S., Izadi, S., Sellen, A., Kaye, J. J., and Eardley, R. (2007). Locating Family Values: a Field Trial of the Whereabouts Clock. In *Proceedings of the 9th International Conference on Ubiquitous Computing*, Innsbruck, Austria. Springer-Verlag.
- Cheng, Z., Caverlee, J., Kamath, K. Y., and Lee, K. (2011a). Toward Traffic-Driven Location-Based Web Search. In *Proceedings of the 20th International Conference on Information and Knowledge Management*, Glasgow, Scotland, UK. ACM.

- Cheng, Z., Caverlee, J., and Lee, K. (2010). You are Where You Tweet: a Content-Based Approach to Geo-Locating Twitter Users. In *Proceedings of the 19th International Conference on Information and Knowledge Management*, Toronto, ON, Canada. ACM.
- Cheng, Z., Caverlee, J., Lee, K., and Sui, D. Z. (2011b). Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the 5th International Conference on Weblogs and Social Media*, Menlo Park, CA, USA. AAAI.
- Cho, E., Myers, S. A., and Leskovec, J. (2011). Friendship and Mobility: User Movement in Location-Based Social Networks. In *Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA. ACM.
- Choudhury, M., Sundaram, H., John, A., Seligmann, D., and Kelliher, A. (2010). “Birds of a Feather”: Does User Homophily Impact Information Diffusion in Social Media? *CoRR*.
- Crandall, D. and Snavely, N. (2012). Modeling People and Places with Internet Photo Collections. *Commun. ACM*, 55(6):52--60.
- Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N. (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proceedings of the 6th International Conference on Weblogs and Social Media*, Dublin, Ireland. ACM.
- David, E. and Jon, K. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA.
- Davis Jr., C. A., Pappa, G. L., de Oliveira, D. R. R., and de Lima Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. *Journal of Geographic Information System*, 15(6):735–751.
- Friedland, G., Maier, G., Sommer, R., and Weaver, N. (2011). Sherlock Holmes’ Evil Twin: On the Impact of Global Inference for Online Privacy. In *Proceedings of the Workshop on New Security Paradigms*, Marin County, CA, USA. ACM.
- Fusco, S. J., Michael, K., Aloudat, A., and Abbas, R. (2011). Monitoring People using Location-Based Social Networking and its Negative Impact on Trust: An Exploratory Contextual Analysis of Five Types of “Friend” Relationships. *IEEE International Symposium on Technology and Society*.



- Gross, R. and Acquisti, A. (2005). Information Revelation and Privacy in Online Social Networks. In *Proceedings of the Workshop on Privacy in the Electronic Society*, Alexandria, VA, USA. ACM.
- Gundecha, P., Barbier, G., and Liu, H. (2011). Exploiting Vulnerability to Secure User Privacy on a Social Networking Site. In *Proceedings of the 17th SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA. ACM.
- He, J., Chu, W. W., and Liu, Z. V. (2006). Inferring Privacy Information from Social Networks. In *Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics*, San Diego, CA, USA. Springer-Verlag.
- Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, BC, Canada. ACM.
- Iachello, G., Smith, I., Consolvo, S., Abowd, G. D., Hughes, J., Howard, J., Potter, F., Scott, J., Sohn, T., Hightower, J., and LaMarca, A. (2005). Control, Deception, and Communication: Evaluating the Deployment of a Location-Enhanced Messaging Service. In *Proceedings of the 7th International Conference on Ubiquitous Computing*, Tokyo, Japan. Springer-Verlag.
- Ikawa, Y., Enoki, M., and Tatsubori, M. (2012). Location Inference using Microblog Messages. In *Proceedings of the 21st International Conference Companion on World Wide Web*, Lyon, France. ACM.
- Jin, L., Long, X., and Joshi, J. B. (2012). Towards Understanding Residential Privacy by Analyzing Users' Activities in Foursquare. In *Proceedings of the Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, Raleigh, NC, USA. ACM.
- Krishnamurthy, B. and Wills, C. E. (2008). Characterizing Privacy in Online Social Networks. In *Proceedings of the 1st Workshop on Online Social Networks*, Seattle, WA, USA. ACM.
- Lam, I.-F., Chen, K.-T., and Chen, L.-J. (2008). Involuntary Information Leakage in Social Network Services. In *Proceedings of the 3rd International Workshop on Security: Advances in Information and Computer Security*, Kagawa, Japan. Springer-Verlag.

- Lathia, N., Quercia, D., and Crowcroft, J. (2012). The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility. In *Proceedings of the 10th International Conference on Pervasive Computing*, Newcastle, UK. Springer.
- Li, N. and Chen, G. (2010). Sharing Location in Online Social Networks. *Network, IEEE*, 24(5):20–25.
- Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012). Towards Social User Profiling: Unified and Discriminative Influence Model for Inferring Home Locations. In *Proceedings of the 18th SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China. ACM.
- Lieberman, M. D. and Lin, J. (2009). You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories. In *Proceedings of the 3rd International Conference on Weblogs and Social Media*, San Jose, CA, USA. AAAI.
- Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J. (2011). I’m the Mayor of My House: Examining why People use Foursquare - a Social-Driven Location Sharing Application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, BC, Canada. ACM.
- Magno, G., Comarela, G., Saez-Trumper, D., Cha, M., and Almeida, V. (2012). New Kid on the Block: Exploring the Google+ Social Graph. In *Proceedings of the Conference on Internet Measurement Conference*, Boston, MA, USA. ACM.
- Mahmud, J., Nichols, J., and Drews, C. (2012). Where Is This Tweet From? Inferring Home Locations of Twitter Users. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *Proceedings of the 6th International Conference on Weblogs and Social Media*, Dublin, Ireland. AAAI.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- Mao, H., Shuai, X., and Kapadia, A. (2011). Loose Tweets: an Analysis of Privacy Leaks on Twitter. In *Proceedings of the 10th Workshop on Privacy in the Electronic Society*, Chicago, IL, USA. ACM.
- Marmasse, N., Schmandt, C., and Spectre, D. (2004). WatchMe: Communication and Awareness Between Members of a Closely-Knit Group. In *Proceedings of the 6th International Conference on Ubiquitous Computing*, Nottingham, England. Springer.

- Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. (2010). You are Who You Know: Inferring User Profiles in Online Social Networks. In *Proceedings of the 3rd International Conference on Web Search and Data Mining*, New York, NY, USA. ACM.
- Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. (2011). An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proceedings of the 5th International Conference on Weblogs and Social Media*, Barcelona, Spain. AAAI.
- Pesce, J. a. P., Casas, D. L., Rauber, G., and Almeida, V. (2012). Privacy Attacks in Social Media Using Photo Tagging Networks: a Case Study with Facebook. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, Lyon, France. ACM.
- Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., and Almeida, V. (2012a). Beware of What You Share: Inferring Home Location in Social Networks. In *Proceedings of the 12th International Conference on Data Mining Workshops*, Brussels, Belgium. IEEE.
- Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., and Almeida, V. (2012b). We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of the 14th International Conference on Ubiquitous Computing*, Pittsburgh, PA, USA. ACM.
- Pozdnoukhov, A. and Kaiser, C. (2011). Space-time Dynamics of Topics in Streaming Text. In *Proceedings of the 3rd SIGSPATIAL International Workshop on Location-Based Social Networks*, Chicago, IL, USA. ACM.
- Quercia, D. and Capra, L. (2009). FriendSensing: Recommending Friends using Mobile Phones. In *Proceedings of the 3rd Conference on Recommender Systems*, New York, NY, USA. ACM.
- Quercia, D., Casas, D. B. L., Pesce, J. P., Stillwell, D., Kosinski, M., Almeida, V., and Crowcroft, J. (2012). Facebook and Privacy: The Balancing Act of Personality, Gender, and Relationship Currency. In *Proceedings of the 6th International Conference on Weblogs and Social Media*, Dublin, Ireland. AAAI.
- Quercia, D., Lathia, N., Calabrese, F., Di Lorenzo, G., and Crowcroft, J. (2010). Recommending Social Events from Mobile Phone Location Data. In *Proceedings of the 10th International Conference on Data Mining*, Sydney, Australia. IEEE.

- Ruiz Vicente, C., Freni, D., Bettini, C., and Jensen, C. (2011). Location-Related Privacy in Geo-Social Networks. *Internet Computing, IEEE*, 15(3):20–27.
- Sadilek, A., Kautz, H., and Bigham, J. P. (2012). Finding Your Friends and Following them to Where You Are. In *Proceedings of the 5th International Conference on Web Search and Data Mining*, Seattle, WA, USA. ACM.
- Saez-Trumper, D., Quercia, D., and Crowcroft, J. (2012). Ads and the City: Considering Geographic Distance goes a Long Way. In *Proceedings of the 6th Conference on Recommender Systems*, Dublin, Ireland. ACM.
- Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, NC, USA. ACM.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M. d., and Loureiro, A. A. F. (2012a). Uncovering Properties in Participatory Sensor Networks. In *Proceedings of the 4th International Workshop on Hot Topics in Planet-Scale Measurement*, Low Wood Bay, Lake District, UK. ACM.
- Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M. d., and Loureiro, A. A. F. (2012b). Visualizing the Invisible Image of Cities. In *Proceedings of the International Conference on Cyber, Physical and Social Computing*, Besancon, France. IEEE.
- Tang, K. P., Lin, J., Hong, J. I., Siewiorek, D. P., and Sadeh, N. (2010). Rethinking Location Sharing: Exploring the Implications of Social-Driven vs. Purpose-Driven Location Sharing. In *Proceedings of the 12th International Conference on Ubiquitous Computing*, Copenhagen, Denmark. ACM.
- Vasconcelos, M. A., Ricci, S., Almeida, J., Benevenuto, F., and Almeida, V. (2012). Tips, Dones and Todos: Uncovering User Profiles in Foursquare. In *Proceedings of the 5th International Conference on Web Search and Data Mining*, Seattle, WA, USA. ACM.
- Vögele, T. and Schlieder, C. (2003). Spatially-Aware Information Retrieval with Graph-Based Qualitative Reference Models. In *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference*, St. Augustine, FL, USA. AAAI.

- Wagner, D., Lopez, M., Doria, A., Pavlyshak, I., Kostakos, V., Oakley, I., and Spiliotopoulos, T. (2010). Hide and Seek: Location Sharing Practices with Social Media. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, Lisbon, Portugal. ACM.
- Ye, M., Yin, P., and Lee, W.-C. (2010). Location Recommendation for Location-Based Social Networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, CA, USA. ACM.
- Zheleva, E. and Getoor, L. (2009). To Join or not to Join: the Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain. ACM.
- Zheng, Y. (2011). Location-Based Social Networks: Users. In Zheng, Y. and Zhou, X., editors, *Computing with Spatial Trajectories*, pages 243–276. Springer.
- Zwillinger, D. and Kokoska, S. (2000). *CRC Standard Probability and Statistics Tables and Formulae*. Chapman & Hall.

