UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

DOTTORATO DI RICERCA IN INGEGNERIA DELL'INNOVAZIONE INDUSTRIALE

Information fusion for location data analysis

Author:
Alket CECAJ

Supervisor:
Prof. Marco MAMEI

Dipartimento di Scienze e Metodi dell'Ingegneria

March 6, 2016

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Abstract

Dipartimento di Scienze e Metodi dell'Ingegneria

Doctor of Philosophy

Information fusion for location data analysis

by Alket CECAJ

Nowadays the world is more and more connected in different ways and our behavior affects in some way and by some extent that of other people. By understanding this behavior, a better coordination of the city life could be reached in terms of minor traffic congestion, better quality of air, availability of park places and less noise pollution. The applications that derive by this understanding come in a large number and those applications need data. As personal devices become smaller and smarter the quantity and dimensions of data they generate becomes larger. Data dimensions such as time and location capture different aspects of everyday life. By combining and integrating those dimensions on different data sets is possible to extend and improve the understanding of human behavior. On the one hand, this integration enables even more powerful applications, on the other it may result in serious harm for the individual's privacy. The study of this integration thus data fusion between data sets and its implications in different fields is also the main contribution of this thesis...

Acknowledgements

I would like to express my greatest thanks to Prof Marco Mamei which is an example for me, both as a researcher and in life. His advice, encouragement and patience helped me a lot in understanding the topics during my PhD studies.
Many thanks also to Prof Franco Zambonelli for his advice and for giving me the opportunity to participate in different interesting projects.
I don't want to forget here my colleague Andrea Sassi whose interests for the field is always contagious.
Finally, my greatest thanks to my big family and in particular to Simona and Enea. Being a partner and a father helps me learn a lot everyday.

Contents

Acknowledgements **v**

Preface **1**

1 Introduction **3**
1.1 Data and human behavior 3
1.2 Data combination, integration and fusion. 4
1.2.1 Traditional data fusion techniques 6
1.2.2 Cross domain data fusion 7
1.3 A first classification of the data fusion techniques 8
1.3.1 Algorithms of semantic meaning-based data fusion method 9
Multi-view learning-based 9
Similarity-based methods 10
Probabilistic dependency-based method 10
Transfer learning-based methods 11
1.4 Thesis contribution and its outline 11

2 Location data sources and applications **15**
2.1 Introduction 15
2.2 Geo-referenced data sources 15
2.2.1 Call Description Records- CDR data 16
2.2.2 Geo-referenced social network data 17
2.3 Application fields of CDR data 19
2.3.1 Social and economic development 20
2.3.2 Mobility and transport management 20
2.3.3 Land use and City management 21
2.3.4 Preventing epidemics and health care 21
2.3.5 Studies on Geo-tagged social network data 22
2.3.6 Business activities improved by the use of location data applications 22
2.4 Conclusions 23

3 Data Fusion for re-identification of anonymized Call Description Records **25**
3.1 Introduction 25
3.1.1 How to read (and understand) this chapter 25
3.2 Re-identification problem 26
3.3 Dataset 27
3.3.1 Social Network Data 28
3.4 Mobility traces analysis as extraction of knowledge 29
3.4.1 Uniqueness Test on CDR Datasets 30
3.4.2 Experiments 30
3.4.3 Matching CDR and social network data. 31

3.4.4 Experiments . . . 33
3.5 Probabilistic Approach . . . 39
3.6 Data fusion for re-identification of CDR records . . . 41
3.7 Conclusion . . . 42

4 Data fusion for Event Detection and Description in Urban Areas 45
4.1 Introduction . . . 45
4.2 City dynamics and events in urban spaces . . . 45
4.2.1 Other studies on event detection . . . 46
4.3 Dataset description . . . 47
4.3.1 Spatial and temporal aggregation of the Telecom Challenge dataset . . . 48
4.4 Extracting knowledge from the two datasets . . . 49
4.4.1 Statistical Event Detection Method . . . 49
4.4.2 Evaluating the results of event detection algorithm in terms of precision and recall . . . 53
4.4.3 Groundtruth dataset . . . 53
4.4.4 Precision and Recall of the Event Detection Method . 53
4.5 Improving results of event detection method using data fusion 54
4.5.1 Event description by data fusion . . . 55
4.6 More experiments : analysis of parameters . . . 55
4.7 Conclusions . . . 57

5 A multi-perspective view of the city life obtained by Open Data 65
5.1 Introduction . . . 65
5.2 Related work . . . 65
5.3 Open Data . . . 66
5.3.1 An overview of the data sets . . . 67
5.3.2 Data Set Summary Statistics . . . 69
Census dataset . . . 69
POI dataset . . . 69
Social network dataset . . . 69
5.4 Research questions and some answers . . . 70
5.4.1 Correlation analysis . . . 71
5.5 Conclusions . . . 74

6 Privacy and anonymity: computer science and legal aspects 77
6.1 Introduction . . . 77
6.2 Sensible information that can be inferred from personal location data . . . 78
6.3 Preventing privacy harm : anonymization and suppression . 80
6.3.1 Can part of the location data be suppressed and be useful at the same time . . . 82
6.4 Anonymity as a protection of privacy : differences between legal and computer science perspectives . . . 83
6.4.1 Formal models proposed in computer science for data anonymity . . . 83
6.4.2 The notion of anonymity in the European legislation on personal data . . . 84
6.5 Conclusions . . . 85

7 Conclusions **87**
7.1 Introduction 87
7.1.1 Data fusion process and results 87
7.1.2 Conclusions on privacy implications 88
7.1.3 Future work and open challenges 89

Bibliography **91**

List of Figures

1.1 **Traditional data fusion methods** 7
1.2 **Cross domain data fusion methods** 8

2.1 Records for each dataset. a) CDR dataset, b) Twitter dataset . 17
2.2 **(a)** Flickr location data **(b)** Instagram location data**(c)** Twitter location data **(d)** Geo-referenced Twitter data as placemarks 19
2.3 **(a)** Paths created on Flickr location data 21

3.1 Records for each dataset. a) CDR-DATA1 dataset, b) CDR-DATA2 dataset, c) FT dataset 28
3.2 **(a)** Events per user in CDR-DATA1 dataset. **(b)** Radius of gyration for a given percentile of users in CDR-DATA1.**(c)** Events per user in CDR-DATA2 dataset. **(d)** Radius of gyration for a given percentile of users in CDR-DATA2 dataset. . 29
3.3 **(a)** Twitter events per user. **(b)** Radius of gyration for a given percentile of Twitter users.**(c)** Flickr events per user. **(d)** Radius of gyration for a given percentile of Flickr users. 30
3.4 Uniqueness of mobility traces. After three sightings (geo-referenced data) we can uniquely pinpoint user A 31
3.5 Uniqueness of mobility traces. (left)Number of events needed for identification.(right)Percentage of users re-identified . . 32
3.6 Flickr/Twitter user FT_a (int red) is compared with CDR users u_x, u_y and u_z (in green). The most probable match is $FT_a \equiv u_x$ as they are consistently producing data at the same time and at the same place. 32
3.7 Percentage of users re-identified and number of points needed for re-identification 35
3.8 (right)Boxplot diagram summarizing the statistic of the number of CDR users having X matching events with FT users. (left)Percentage of FT users for that can be associated to X number of CDR users 35
3.9 **(top)** Matching mobility traces between Flickr events in blue squares and Twitter events in red round shape of the same social user. **(bottom)** CDR events from a user in black triangle. The social user in its both profiles as Twitter and as Flickr user matches with the same CDR. 38
3.10 (left)Distribution of events per mobile country code.(right). Distribution of social users languages 38
3.11 (left)Error bars plot of the number of CDR-DATA2 users from the sample chosen having X matching events with all other users.(right) The quartiles of the same distribution, 25th and 75th percentiles. 39

3.12 **(left)**. Probability $p(CDR = FT|n)$. Plots for different values of Δt used in matching process. **(right)**. Maximum probability obtained for a given user percentile (we do not plot below-50 percentile as the probability is almost constant 0) . 42

4.1 **Knowledge based data fusion methods** 46
4.2 **(a)** Part of the Milan grid and **(a)** Trento province grid. The cells in bright red and yellow color indicate a more intense activity then the green and blue ones. 48
4.3 **(a)** CDR Data activity also called SMS-call dataset **(b)** Twitter Data 48
4.4 **(a)** High threshold **(c)** Low threshold. In **(b)** and **(d)** there are boxplot diagrams representing the cell activity levels and outliers 50
4.5 **(a)** Data with no aggregation **(b)** Data aggregated by week days **(c)** Data aggregated by hour **(d) Data aggregated in hourly basis only by week days** 52
4.6 **(a)** SMS-Call Data. **(b)** Internet Data. **(c)** Social Data. **(left)** Milano. **(right)** Trento 59
4.7 Results in aggregating data from multiple source **(a)** Milan area **(b)** Trento area 60
4.8 Results of event description. **(left)** An artisans' fair. **(right)** A football match. From the world cloud it is possible to spot the words characterising the events. 60
4.9 **SMS-Call Data - (a)** IQR method for diff. k values **(b)** Median method for diff. k values **(c)** Q75 methods for diff. k values **(d)** best method for diff. cell sizes 61
4.10 **Internet Data - (a)** IQR method for diff. k values **(b)** Median method for diff. k values **(c)** Q75 methods for diff. k values **(d)** best method for diff. cell sizes 62
4.11 **Social Data - (a)** IQR method for diff. k values **(b)** Median method for diff. k values **(c)** Q75 methods for diff. k values **(d)** best method for diff. cell sizes 63

5.1 **(a)** How the administrative region of Campania looks divided in sections **(a)** Heat map of social data points 68
5.2 **(a)** Distribution of number of people for municipality **(b)** Age range expressed as percentage of the total population for municipality in a boxplot diagram representation. 70
5.3 **(a)** Distribution of number of POI per municipality **(b)** The twenty most frequent categories of POI 70
5.4 **(a)** Distribution of number of pictures and tweets per user **(b)** The distribution of the number of pictures and tweets per municipality 71
5.5 **(a)** 20 most important topics for the municipalities with mostly young people **(b)** 20 most important topics for the municipalities composed by mostly old people 73
5.6 **(a)** Word cloud graphics for municipalities with mostly young people **(b)** word cloud graphics for municipalities composed by mostly elderly people 73

5.7 Correlation graphic of the number of pictures and the different categories of POI per municipality 74

6.1 **(a)** Location history from Google Maps timeline. **(b)** Location history in an international context 80

List of Tables

Preface

In the ever changing technological landscape there are many new – born terms such as big data, smart mobility, anticipatory sensing, predictive analytics, machine learning and many others. Each of those terms, is due to the evolution of IT sector and to the many applications that can be created, by the huge quantity of data generated everyday by people and electronic devices.
This mole of data, increases together with other social and economic phenomena such urbanization, in particular in the developing countries. As technology become cheaper (as for example low cost smart phones or laptops) also access to it and to internet becomes more affordable. It is clear thus, that the phenomenon of data explosion is global and ubiquitous.
In this scenario the diversity of data sources and data representation in terms of scale, density, and distribution makes it even harder to store and analyse the data itself. Moreover, each devices measures a certain action or activity by translating it into a dataset. Each dataset on its own diversity, can tell a small story related to the single aspect of the activity or phenomena the data set itself represents.
These small stories can be combined by fusing the information in each data set, in order to provide a better view of the whole picture, gain knowledge and understand better why and how the single data sets relate or correlate to each other. But as the volume and variability of data sets increase, traditional data fusion methods such as schema matching or schema mapping become obsolete in terms of efficiency and efficacy. Moreover, given the diversity of data sources and the complexity of the phenomena they report the simple join on two data sets would be simply naive.
It is thus necessary, the experimentation and use of new methods i.e. machine learning methods, that extract high level information, in fact knowledge, from each data set. By combining the knowledge and not simply the data records it becomes possible to have better insights on the phenomena object of study.

Chapter 1

Introduction

This chapter explains different data fusion methods and algorithms and how they are applied. In particular, what the case studies brought in the following chapters can show in order to support the thesis topics. It also presents the data fusion process, why it is needed and how it is performed. More specifically there are illustrated the different concepts related to data fusion methods evolution and classification. Finally it illustrates the structure of the thesis, how to read the different chapters and more in detail their contents.

1.1 Data and human behavior

Unprecedented amounts of data are created each day as every person lives and documents his life by sharing information about himself or the surrounding environment.

For having an idea of the data generated let's consider the typical day of a normal person which sets his/her alarm clock on the phone for the next morning, goes for a run and registers his/her activity on a wearable fit device, buys coffee using his fidelity/credit card, chooses the public transportation using a dedicated app and shares a nice morning picture on a favorite photo sharing social network. The day continuous by responding to incoming calls, SMSs, instant messaging and emails thus creating more and more data.

There are also a very large number of devices (larger than humans according to the CISCO Report) [1] connected to the internet such as cameras, servers of any type, smart home devices and the numbers keeps increasing as the IoT becomes more present and efficient.

All these data created in different context, represent different activities as they measure very different physical phenomena as acceleration, position, orientation, online or real world physical interaction/participation, and so on and so forth.

Also the technology behind the devices generating the data, can be very different from case to case, resulting in high heterogeneity of the data sources Zambonelli, 2008. As a consequence the great majority of these data sets is unstructured data such as text, images, audio or video.

However those data, can power many interesting applications that exploit different crowd sourcing techniques to coordinate and organize people, especially in urban context, giving to the cities a smart dimension Sassi and Zambonelli, 2014. Most of these data generating activities come from mobile devices.

[1] http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

In particular the number of mobile subscribers keeps increasing also in the developing countries as of latest report from world bank [2]. Mobile means also a moving position thus giving to the data set created a mobility dimension into space and time.

Mobility data can be represented as a trajectory on time and space while social network activity and presence as a connected graph. From a data analyst perspective every data set can be seen as a single piece of puzzle providing insights on the single activity or phenomena observed and represented by the data set itself.

However to create interesting and useful applications, often we need to use different types of data or in some way put together two or more data sets.

1.2 Data combination, integration and fusion.

As mentioned in the previous section when creating an application that solves a problem, very often, data from different domain has to be analyzed Zheng et al., 2014b and then combined to generate results of higher level of abstraction.

In this sense data integration or data fusion is the process of combining multiple types of data that represent the same object into a consistent, accurate, and useful representation [3]. Another definition that considers how the data sets relate to each other can be : data fusion is the analysis of different data sets where every data set can interact, inform and complete the other data sets.

The result of the data fusion process is expected to be more informative and synthetic than the original inputs. As an ideal example of data fusion system can be considered the human brain where the most intuitive data fusion process is audio and video fusion from the two respective senses.

A less ideal example of data fusion, is multi-sensor data fusion which combines data from different sensors as in the following papers which are some of the pioneers in the field of multi-sensor data fusion Khaleghi et al., 2013 Durrant-Whyte and Henderson, 2008 Varshney, 1997 Hall and Llinas, 1997 Ulfarsson, Benediktsson, and Sveinsson, 2003 .

There are many other fields and more specifically human - machine Interaction, robotics Nigay and Coutaz, 1993, environmental studies such as meteorological ones Treinish, 2000 etc.. Another example is the work in De Bie et al., 2007 which uses data fusion for gene prioritization in the in bio-medical research field.

Recommendation systems can also make a fruitful use of data fusion methods as in Tan et al., 2014 where the authors bring a recommendation system based on multi-type media fusion.

For the same purpose of media processing the work in Weng and Chuang, 2012 uses data fusion methods for concept video indexing while the paper of Wang, Guan, and Venetsanopoulos, 2012 makes use of kernel cross-modal factor analysis for emotion recognition applications when analyzing cross-modal relationship between audio and visual features. For face and

[2] http://www.worldbank.org/en/news/press-release/2012/07/17/mobile-phone-access-reaches-three-quarters-planets-population

[3] https://en.wikipedia.org/wiki/Data_fusion

speech there is the work in Ben-Yacoub, Abdeljaoued, and Mayoraz, 1999 that applies data fusion for this purpose.

There is also the study in Gross et al., 2012 which describes the processing architecture designed and implemented for the fusion of hard and soft data where hard data are the ones coming from physical sensors and the soft data that are human observed data. Finally the work in the domain of security and privacy described in Eldardiry et al., 2013 makes use of data fusion methods for insider threat detection.

Data fusion as in Lahat, Adalý, and Jutten, 2014 is a challenging field for several reasons and the main ones are listed below :

1. The data to be fused or integrated is generated by very different data sources and in very different context.

2. Data fusion results have to answer every time a new research question as the application of it are very wide.

3. Fusing data of different types in order to maximize advantages and minimize drawbacks may be a difficult task to understand or even a tricky one

Data fusion processes may be divided in low, intermediate and high , depending on the processing stage at which fusion takes place. Low level data fusion, combines several sources of raw data to produce new raw data. In more simple terms it combines data using numerical fusion methods such as kalman filtering or probabilistic methods.
The intermediate and high processes are concerned with the extraction of high level knowledge such as situation awareness and the formulation of decisions and actions Durrant-Whyte and Henderson, 2008

Depending on the domain, data integration can be thought as a combination of data where the larger data set result of the combination itself is retained, whereas data fusion is a set reduction technique with improved confidence.

Inside of data fusion field itself, we can distinguish between data fusion as a conventional technique intended as a combination of records Bleiholder and Naumann, 2009 and new data fusion techniques, based on machine learning algorithms, thus combination of knowledge Zheng and Member, 2015.
Thus to have real knowledge about observed phenomena, and to use it in useful applications, two or more data set should be combined together in order to give better insight.

However, the process of fusing those data sets of different dimensions together is not straightforward as data may be generated in a complex context and by very different processes.

For example to measure air pollution we need the geographical maps of the environment we are going to monitor, weather conditions, population

density and traffic or polluting industrial activity.
Again for recommending a route to a traveler for a car sharing service we need the road network represented as graph, the data about the potential people interested in sharing their car with other people, people willing to pay for the car sharing service but only with people with whom they share common interests Guidotti et al., 2015.

All this requires a good knowledge of all the types of variables involved that usually is very unlikely to have in only one type of data set. As each of the considered data sets exhibit a different spatial and or sampling rate a simple combination would likely create erroneous assumptions and compromise the results of the analysis.

Thus the use of data fusion, should not only aim to bring together all data and attributes into a single view, but to combine the knowledge extracted from each data set, by creating in fact, a more complete picture of the object of study.

1.2.1 Traditional data fusion techniques

The traditional data fusion techniques (invented during the 80s) are more about data integration by simply joining the fields/attributes on the different data sets into a single data set or using a combination of fields as for example schema mapping or schema matching.

Schema matching is the process of identifying that two objects are semantically related while data mapping or schema mapping refers to the transformations between the objects.

The research in schema matching aims to provide automated support to the process of finding semantic matches between two schema but this process is made harder due to heterogeneity in data. Analogously data mapping aims to make a data mediation between a data source and a destination and consolidate multiple databases into a single data base. As a last step it tries to identify redundant columns of data for consolidation or elimination.

As in the graphics in Fig 1.1 the traditional data fusion methods combine the records through schema mapping and data matching, eliminate duplicates and merge data.

More in general, another disadvantage of the traditional data fusion methods, is that they were designed for a modest quantity of data compared with the big data volumes of today. In the 1980s, SQL, that was the standard for relational databases and querying, used to query databases which contained at most few million records.

Conventional data fusion techniques refer to the JDL/DFIG model that introduced a way to classify the data fusion processes in five levels that are:

1. Source Pre-processing /subject Assessment which is associated with pre-detection activities such as pixel or signal processing

2. Object Refinement which is concerned with estimation or prediction of continuous / discrete state of the object.

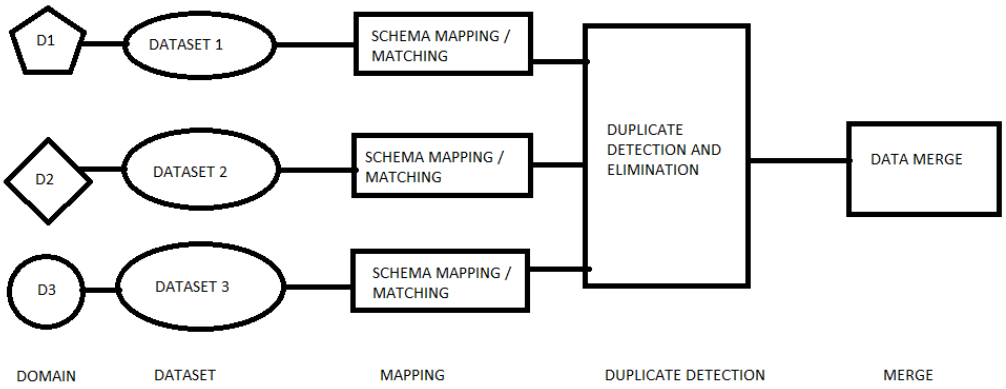


FIGURE 1.1: **Traditional data fusion methods**

3. Situation Assessment which examines the relations among the different entities

4. Impact Assessment that evaluates the effects that possible actions would have in the current situation.

5. Process Refinement (or Cognitive Refinement) which closes the loop by re-tasking resources.

This model it is often criticized for its implication that the levels necessarily happen in this order and also for its lack of adequate representation. In Steinberg, Bowman, and White, 1999 is proposed a revisions and expansions of the JDL Data Fusion model to remedy some of these deficiencies.

However the JDL / DFIG model is useful for visualizing the data fusion process and facilitating discussion and common understanding Azimirad, 2015.

1.2.2 Cross domain data fusion

With the cross domain data fusion techniques as in Fig 1.2 instead of combining records in a linear process, there is knowledge combined. From each data set that comes from different domains, a piece of information is extracted and then these chunks of information are combined to create a bigger container of knowledge.

The methods used for this kind of data processing rely on inferential statistics and data mining algorithms and techniques such as machine learning. Those techniques are the new SQL, and even their understanding and implementation is not easy there are many open projects like Hadoop or Spark that are making this kind of queries accessible to a vast number of people.

Revising, testing and using those methods of extracting and combining knowledge from different data sets in different context from re-identification of anonymized data sets to event detection in cities is also the main purpose of this thesis.

As the field of data fusion especially in the new conception of it as a machine learning field is a relatively young field, there is a lot of fermentation in the academic and industry world. There are different studies that

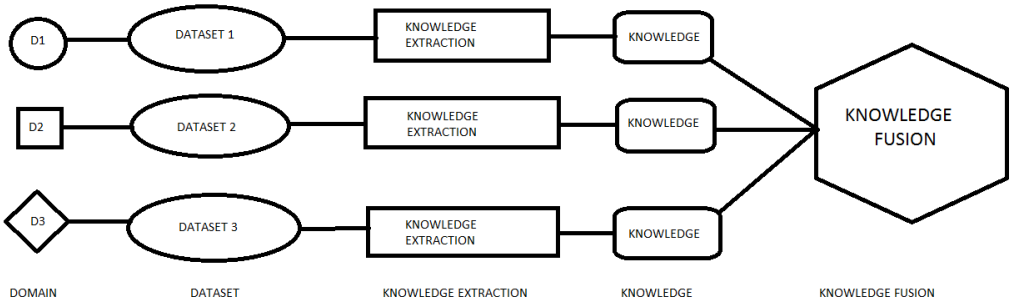


FIGURE 1.2: **Cross domain data fusion methods**

describe the methods of data fusion itself using different scales of classification and organization.

In the next subsection follows a summary of those works and possible classifications of the data fusion methods and processes.

1.3 A first classification of the data fusion techniques

Data fusion is an area of research that involves many fields and domains. Thus the existing techniques for the data fusion process could be intersecting different areas and this makes the classification subjective .

However a way that leads to possible classification as provided in Castanedo, 2013, could be that of considering :

1. Criteria as relations between different data sets as proposed by Durrant-Whyte and Henderson, 2008. The relations between the data sets could be seen as complementary, redundant, or cooperative data.

2. Input and output data types during the process of data fusion, also according to the input/output data types and their nature,

3. Level of abstraction in which the data used are provided. It means that a possible classification, can follow the abstraction level of the data in use which can be: raw measurement (numbers), signals, characteristics or decisions.

4. Architecture type, weather it is centralized or distributed architecture.

According to the work in Lahat, Adali, and Jutten, 2015 the approaches to the problem of data fusion can be divided in two categories:

- Model driven data fusion approaches which rely on an explicit realistic model.

- Data driven approaches that rely on data. However, in the context of cross domain data fusion where data is generated in complex context, and little is known about the underlying relationships between the data sets, data driven models can be of better use.

In the literature can be found plenty of examples of data fusion techniques and the thesis tries to bring the most relevant ones. One of those is re-assumed in Zheng and Member, 2015.
These new data fusion techniques can be of different types and can be classified in the following categories

1. Stage based methods. This is a category that uses different data sets that are in different stages of the process of data mining. Following this category, the data sets are loosely coupled without any requirements on their consistency.

2. Feature level-based. This method treats features extracted from different data sets and creates an array by concatenating them. This array can then be used in clustering and classification methods. An example of this kind of methods can be consulted in

3. Semantic meaning-based data fusion methods. These methods take in consideration the relations between features in different data sets. This implies that the data miner knows what each data set represents, and why they can be fused or why they re-inforce each other in terms of enrichment of information.

The data fusion studies and application methods presented in this thesis are part of the semantic meaning based data fusion methods. In particular Semantic meaning based method with a probabilistic dependency.Those methods are based in particular algorithms that are described in details in the next subsection.

1.3.1 Algorithms of semantic meaning-based data fusion method

Semantic meaning-based data fusion methods provide means for understanding the insight of each dataset and the relations between features of different data sets Zheng and Member, 2015. Those methods, wants the analyst to know why different data sets can be fused, how they complete each other and how they provide better information about the object of study .

Semantic meaning methods can be further divided into four groups, and in the following paragraphs, is provided a more detailed description of each group.

Multi-view learning-based

As different data sets may describe observed phenomena in different ways, in the same way they can be regarded as different views on the observed phenomenon. For instance, a person can be identified by the information obtained from multiple sources, such as face, mobility traces or signature.

All these data sets describe the object and are complementary to each other as they contain knowledge that other views do not have. As a result, combining multiple views can describe a phenomenon better and more accurately. As of Xu, Tao, and Xu, 2013, the multi-view learning algorithms are divided in three groups:

- co-training that aims to maximize the mutual agreement on two distinct views of the data
- multiple kernel learning, which exploits kernels that naturally correspond to different views and combine kernels either linearly or nonlinearly to improve learning.
- subspace learning. Subspace learning algorithms aim to obtain a latent subspace shared by multiple views, assuming that the input views are generated from this latent subspace.

Similarity-based methods

If two objects (A, B) are similar in terms of some metric, then the information of A can be leveraged by B when B is not described by any type of data. In the same way, if A and B are described by two data sets dA and dB respectively, then it is possible to learn multiple similarities between the two objects. Each of the similarities is calculated based on a pair of corresponding data sets. These similarities can mutually reinforce each other and the correlation between two objects gets stronger.

For example, the similarity learned from a dense data set can be used to verify those derived from other sparse data sets, thus filling the missing values.

There are two types of similarity based algorithms :

- Coupled matrix factorization: which is strictly related to collaborative filtering an matrix factorization algorithms. Coupled matrix factorization is used (or called context-aware matrix factorization) to represent different data sets with different matrices, which share a common dimension between one another. By decomposing these matrices collaboratively, we can transfer the similarity between different objects learned from a data set to another one, therefore complementing the missing values more accurately **Acar13**

- manifold alignment : Manifold alignment assumes that disparate data sets produced by similar generating processes will share a similar underlying manifold representation. By learning projections from each original space to the shared manifold, correspondences are recovered and knowledge from one domain can be transferred to another. Most manifold alignment techniques consider only two data sets, but the concept extends to arbitrarily many initial data sets.

Probabilistic dependency-based method

Those methods are based on the use of a probabilistic model for example a graph that expresses the conditional dependence structure between random variables or between different entities.

The graph-based representation, is used to model a complete distribution over a multi-dimensional space. The graph can be regarded as a compact or factorized representation of a set of independences that hold in the specific distribution.

There are two types of graphical representations of distributions that are commonly used, as Bayesian Networks and Markov Networks. Both the models encompass the properties of factorization and independences, and differ in the set of independences they can represent.

Transfer learning-based methods

Those methods are used when there is a classification task to be done. If this task regards one domain of interest, but there are only sufficient training data in another domain of interest, then the latter data may be used for the training process. Thus transfer learning, in contrast to semi-supervised learning, allows the domains, tasks, and distributions used in training and testing to be different.

In the real world, can be observed different examples of transfer learning as for example learning to drive a car may help driving a bus. In the same way by analyzing a user's transactions in Amazon, and discovered his interests, those interests may be transferred into another application for travel recommendation. The algorithm takes the name from the technique itself and is called Transfer learning.

1.4 Thesis contribution and its outline

The thesis aims to bring a two fold work. On the one hand investigate and apply the categories of the data fusion methods seen above to large mobility data sets such as CDR and geo-tagged social network data.
By providing use case examples and published studies in which those techniques give interesting results, it will focus on semantic meaning based methods with a probabilistic dependency.

On the other hand it will overview the privacy risks for personal location data that emerge from the data fusion process. In particular it will survey the possible ways of anonymizing personal location data, the robustness of such methods and the remaining utility of the data set once the anonymizing methods such as suppression are applied. Finally the legal aspects on privacy and anonymity will be considered with focus on a certain lack of collaboration between computer science experts and legal experts.

More in detail the application of data fusion methods regards the reidentification of anonymized CDR by using social network geo-referenced data, in the first case, and in the second case that of detecting and describing city life events using aggregated CDR data and location based social network data.

A third study presented in the thesis, regards the use of open data, combined with geo-referenced pictures and POI (Points Of Interest).

The contribution of this thesis is focused in applying data fusion methods to solve real problems and in particular the work is divided in three phases. More in detail :

1. Collection and analysis of geo-referenced urban data, data coming from social networks, GPS, cellular networks data but also urban open

data. The data for the works in this thesis are also obtained by participating in data challenges organized by different mobile phone companies.

2. The discovering of hidden patterns in the data, using algorithms and statistical techniques of machine learning to gain insights of user behavior, recommend and predict actions, identify peculiar areas of the city and undergoing events. The development of innovative algorithms to process mobility and social network data and use this information as the basis to create advanced applications.

3. Combination of different algorithms for exploring novel ways of data analysis such as data fusion on the basis of semantic information. Creation and use of methodologies to deal with the scale of data. Evaluation and testing of new technologies for contextualized research purposes and application of pattern analysis mechanism to pervasive data.

The content of the thesis is organized in seven chapters. The following paragraphs present a brief description of each chapter:

- The Chapter 1 presents an overview of the algorithms and techniques used in cross domain data fusion. It explains how the data fusion is evolving with the availability of large data sets. This chapter provides a comparison between traditional data fusion as for example records matching and more advanced cross domain data fusion techniques that make use of machine learning algorithms.

- Chapter 2 describes the location data and how this type of data is created during different human activities, from mobile phones to location based services. It also reviews relevant studies on the utility of mobile location data for studying human behavior at a large scale. Such studies range from density estimation and smart mobility applications to emergency situations handling and land use estimation. Furthermore in this chapter is given a detailed description of two main categories of location data such as LBS (Location Based Services) data and CDR (Call Description Records) and the technology behind each.

- Chapter 3 presents a use-case of data fusion from two different sources in which location data from social networks are used to re-identify CDR data from mobile phones. In this case, in order to evaluate the potential of the re-identification method, a probabilistic data fusion method was used.

- Chapter 4 shows a use-case in which mobile aggregated location data and geo-referenced social network data is used to detect city life events. In this case, the data is then integrated to describe the events. This work ,illustrates with detailed graphics how the process of event detection can be improved by using data fusion. Furthermore the data

fusion methods used in this chapter, can be effective also for event description.

- Chapter 5 presents different experiments in which Open Data is used to understand better the city dynamics in terms of demographic composition, social and economic activity.
  It is an attempt to link socio - economic and demographic characteristics from census data, geo-referenced social network data from Twitter, Flickr and Instagram and POI (Points Of Interest) data obtained from the Nominatim API. By doing so and by using a correlation analysis it aims to demonstrate / evaluate the utility of this type of data for research purposes.

- The chapter 6 shows how the data fusion approach specially on location data could have a serious impact on the privacy of individuals. In particular how data fusion may help re-identification of sensible anonymized data such as medical records or mobile phone data (CDR).

- Chapter 7: Presents the conclusions by addressing open challenges in data fusion methods implications on individual's privacy and future orientation of research.

Chapter 2

Location data sources and applications

2.1 Introduction

In this chapter are presented the different types of technologies and applications such as mobile communication, location based social networks, wearable devices etc.. that monitor populations behavior and register the activities thus translate this behavior in data.
Furthermore, this chapter explains and illustrates how the data can be used to extract useful information in location data analysis domain, smart mobility and more in general smart city domain.

The work represented in this thesis focus mainly on geo-referenced data, thus data that have a location dimension. Location, time, and proximity are powerful dimensions on a data set enabling studies on urban planning, emergency handling, traffic monitoring and mobility dynamics etc..
There are a considerable number of location based services, such as social network applications or car sharing applications.

2.2 Geo-referenced data sources

Research on human mobility have traditionally relied on expensive data collection methods, as for example direct observation, to have an insight into the way that people move in space and time.
The high cost method of such surveys typically results in infrequent data collection (for example census data collection happens only once every ten years) or in small sample sizes.
Instead, data from cellular telephone networks can make it possible to study human mobility cheaply, frequently, and on a large scale as billions of people keep a phone with them most of the time.

In the last five years there is a rapid technological speed up world-wide in terms of number of mobile subscribers and connectivity higher speed.
As of GSMA 2015 report [1] : "The mobile industry continues to scale rapidly, with a total of 3.6 billion unique mobile subscribers at the end of 2014.
Half of the world's population now has a mobile subscription—up from just one in five 10 years ago.

[1] http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf

An additional one billion subscribers are predicted by 2020, taking the global penetration rate to approximately 60%. There were 7.1 billion global SIM connections at the end of 2014, and a further 243 million machine-to-machine (M2M) connections."
Furthermore, in developing countries such as African ones the mobile market have only 70-80 % penetration so these countries should experience even more growth until 2020 as of [2].

While the 2015 Ericson Mobility Report [3] shows that the monthly data traffic has a CAGR (Compound Annual Growth Rate) of 30%.
Even the majority of mobile subscriptions use basic phones such as the ones for calling or texting, by 2016 the number of smartphone subscriptions will surpass those for basic phones as the same report forecasts.

As cellular networks need to know the rough location of all active phones to provide them with voice and data services, location information from these networks holds the potential to revolutionize the study of human mobility and many other fields of human activity.
More deeply with this point there are two types of data that are particularly interesting such as CDR and geo-referenced social network data.

2.2.1 Call Description Records- CDR data

In general, aggregated CDR data can be obtained during data challenges such as the ones organized by the French operator Orange or the Italian one Telecom TIM.
Mobile operators collect CDR data for billing purposes and each time a subscriber makes or receives a call or an SMS a record will be generated and stored in a database.There are two primary sources of these data: communication and internet usage.

The Call Description Records have a timestamp indicating the moment of the event and there is also an identifier (hash code) indicating the mobile number which is obviously anonymized.
There are no standards in gathering this kind of data and each telecommunication operator decides which information is collected and how it is formatted.
For example, there could be the time-stamp of the end of the call instead of the duration, and the sampling is done using different time slots.
Another interesting field is the MCC (Mobile Country Code) that indicates potentially the country provenience of the subscriber itself. The tower id field and the latitude – longitude coordinates indicate the tower handling the event and its location in geographic coordinates.

In the following table we show how each CDR data record is represented and what information contains each of the fields. In particular in a)

[2] http://www.ericsson.com/res/docs/2015/mobility-report/emr-nov-2015-regional-report-sub-saharan-africa.pdf
[3] http://www.ericsson.com/res/docs/2015/mobility-report/emr-mobile-business-trends-2015.pdf

a)

User id	MMC	Timestamp	Tower id	Coord	Radius
4dd285m	222	654628623749	345	(42.28,14.92)	590

b)

User id	Pic/Tweet	Timestamp	Coord
428205c	Text	984738678749	(41.28,13.92)

FIGURE 2.1: Records for each dataset. a) CDR dataset, b) Twitter dataset

there is given a CDR record while in b) a social network record from Twitter. They both have a temporal and geographic dimension.

CDRs are a very useful source of location information. The main reasons are that :

1. they are collected for all active cellular phones, which means several millions in the country the data set is released and potentially billions worldwide;

2. they have very little marginal cost as they are already being collected for billing purposes;

3. they are continuously collected as each voice call and text message completes, thus enabling timely analysis.

At the same time there have two significant limitations on CDR data which are due to:

1. they are sparse in time because they are created only when a phone generates an event (such as incoming/ outgoing SMS or call)

2. the data is coarse in space, as the location precision is the radius of the cell tower. With the actual 3G-4G technology it varies from 300-500m to several km while the incoming 5G technology should reduce this radius in 50 to 100 m, making the CDR data location dimension close to precise.

2.2.2 Geo-referenced social network data

As social media services such as Instagram, FourSquare, Twitter, Flickr and others, become increasingly popular platforms they also allow participants to post streams such as short status messages, pictures, and videos, in a highly connected social context.

Many of these items are tagged by location i.e. they contain coordinates of a certain location, in the form of latitude and longitude.
Their location might be associated as well with a business or venue that is in turn associated with a precise location. The number of "geo-tagged" events is likely to grow with the number of people using geo-enabled devices.

The technology at the heart of location based services is GPS. In particular, smart-phones are equipped with a GPS chip which uses satellite data to calculate the user's exact position.

The technology in smart-phones functions better with good meteorological conditions and usually outdoors. A service such as Twitter or Flickr can then map the position in a geo-referenced point and in the specific case as a geo-referenced picture. When the GPS signal is unavailable (if user is located indoors), Flickr can use information from cell towers to triangulate the user's approximate position.

This method is not as accurate as GPS but is has greatly improved in recent years. Another technique for user's location is given by a combination of GPS and cell-site triangulation or Wi-Fi networks.

From a privacy point of view the user have first to agree to disclose his/her position. Based on this information, the location provider returns the geographical coordinates, at the nearest granularity obtainable, together with the estimated accuracy of the position.

Mainly the geo-location apps do two things:

- they report the user's location to other users (typically user's friends that might be interested to know) such as Foursquare
- they associate real-world locations (such as restaurants, bars or hotels but also events or happenings) to the user's current location.

This vast amounts of location data offer unique opportunities for understanding local communities and people's attitudes in particular context, such as mobility and other interest related with the particular environment they live and work in.

The geo-referenced social data can be obtained through the querying of REST API services of the social network itself.
Social network services that provide such data records through their APIs are Twitter, Flickr, Instagram, Foursquare etc..

In the figure below is given a heat-map graphics over the Italian region of Emilia Romagna of the geo-referenced data from a) Flickr, b) Instagram and c)Twitter. There is a very little difference between Flickr and Twitter data in terms of density as they are both more dense around city center while the Instagram data set is more homogeneously distributed over the territory.
Moreover these social network services have very different number of users so the quantity of data per user is different from service to service.

Other fields of interest are the picture tags that describe the picture posted in Instagram or Flickr while in the Twitter case there is the text tweeted by the user.
The text is a very interesting data field as by analyzing it semantically can be inferred many useful types of information.
Even the tweet is not geo-referenced it is possible to establish with a certain precision the location of the user through content based geo-location as showed in Cheng, Caverlee, and Lee, October 26–30, 2010.

The main advantages of those data are that :

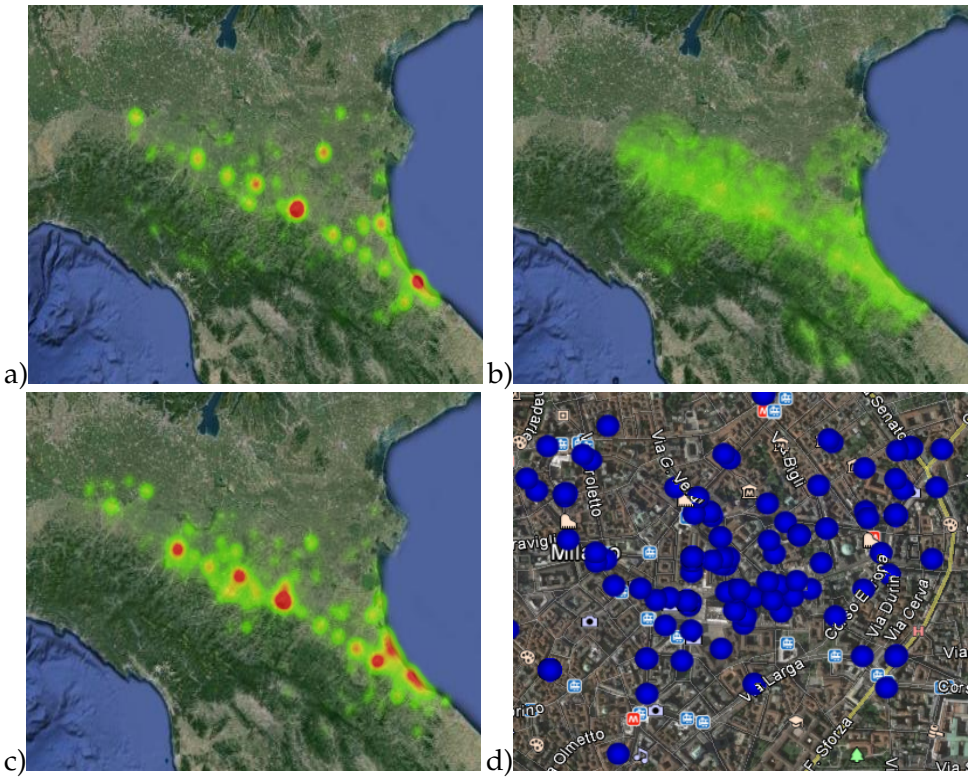


FIGURE 2.2: **(a)** Flickr location data **(b)** Instagram location data**(c)** Twitter location data **(d)** Geo-referenced Twitter data as placemarks

1. they are rich of semantic information about the social media user and abstract concepts or the location itself. They also provide a social context and show how this content is related to other users and / or physical entities

2. their location dimension might be relatively precise as it comes from WiFi geo-location, GPS or triangulation".

The dis-advantages are that : those data are quite sparse. This means that the biggest part of this data set has a large number of users but very little information about the user itself. Looking to it in a different way it contains too much information about a small number of users.

2.3 Application fields of CDR data

Mobile companies collect subscribed data for billing purposes and in collaboration with scientific laboratories or in data challenge contest they release those data in an aggregated or anonymized format.

During the last years there have been presented different works that evidence the impact of telecommunications on the economy.

However with the release of CDR data from the mobile companies such as French carrier Orange or the Italian one TIM (Telecom Italia Mobile) and the location based services of different social networks it becomes possible to discover more areas where the study of human behavior monitored by the data can be fruitfully used for different applications.

2.3.1 Social and economic development

As described previously the telecommunications sector is one of the corner stones of the economy as more people and information are involved .
Moreover, by studying the mobility patterns of the people in different context, could be inferred even more information about the cities where the people resides such as population density Saluveer et al., 7, energy consumptionAndrey Bogomolov and Pentland, 7 and ROSA et al., 2015, quantifying the urban economic activity and estimating poverty levels as in Adeline Decuyper et al., 7 etc..

More specifically the work in ROSA et al., 2015 aims to provide a possible way to the electrical grid operators to predict better the demand of electricity.
In Bogomolov et al., 2015/08/14 the authors predict the crime hotspot areas in the city studying human mobility. In other applications CDR data can be used for demographic census purpose such as the study in Saluveer et al., 7.

2.3.2 Mobility and transport management

By sequencing the CDR data in time and space it is possible to obtain mobility paths and daily routines of individuals.
More in general it is possible to obtain what is called a origin – destination matrix which is a method to describe the urban dynamics between the different areas of the city in different moments of the day.

As the sharing economy develops and prospers also do mobility models such ride sharing or car sharing and assessing its potential between different areas of the city can be one of the applications of the study of CDR data.

In the same domain could be interesting to predict users location in a certain moment of the day and offer her/him a car sharing service or other possible services Alexander, Toole, and Gonzalez, 7. As mobility of people happens in a routine way then the work in Sevtsuk and Ratti, 2010 analyzes the data by taking the routine of mobility as a key element.

While the authors of the work in estimate weekday and weekend travel patterns of a large metropolitan area with high accuracy by using an algorithm to analyze opportunistically collected geo-localized data.

An example of mobility data is produced also by the GPS carried by buses and taxis and represents trajectories and timetables of the urban transportation network. An example of study is given in Veloso, Phithakkitnukoon, and Bento, 2011 that studies urban mobility using taxi traces.

To track the evolution of the attractiveness of different areas of interest in New York city the work in Girardin F, 2009 provides a case study

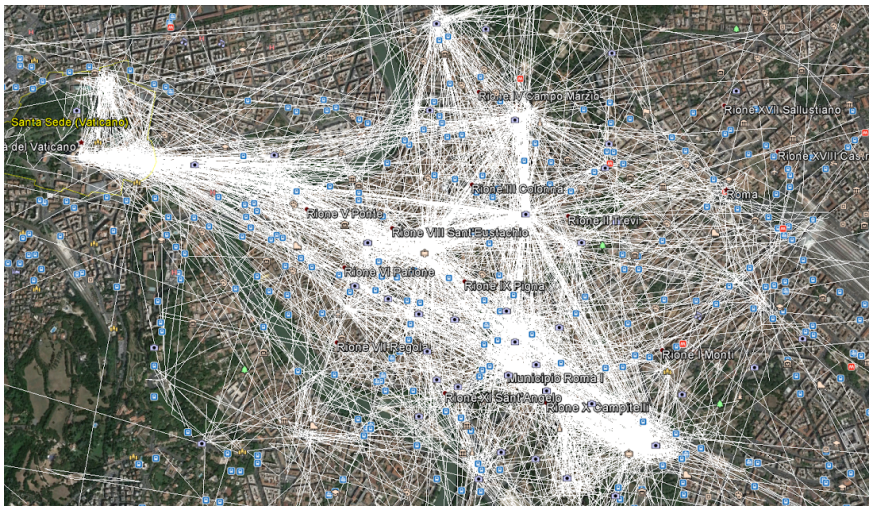


FIGURE 2.3: **(a)** Paths created on Flickr location data

where aggregate and anonymous cell phone network activity data and geo-referenced photos from Flickr are used for this purpose.

2.3.3 Land use and City management

CDR data are an invaluable source of information also for city managers as they allow to discover interesting properties of the different urban areas. In the work in Caceres, Rowland, and Urbanek., 2012 authors analyze when and where people use green spaces and how such behavior differs from other urban areas.

Another paper in Soto and Frias-Martinez., 2011 brings an example in which authors make use of Voronoi tessellation to automatically identify land uses from call detail records. The work is focused on different types of areas : industrial parks and office areas, commercial and business areas, leisure areas,transport hubs and residential areas.

In Reades et al., 2007 authors monitor the dynamics of Rome and obtain clusters of geographical areas measuring cell phone towers activity. Another interesting paper is also Dong et al., 2015 in which authors infer events happening in the city such as concerts, protest or gathering of unusual crowds of people from the analysis of CDR data.

2.3.4 Preventing epidemics and health care

Human interactions and human behavior in general is the main subject of study using the CDR data. In many cases the spread of diseases such as malaria or cholera may depend on human social ties and human mobility. Uncovering the underlying graph of such social connections may help epidemiologists understand better the spread of such diseases and eventually manage better situations of emergency or intervene with more efficacy against those diseases.
For example in Bengtsson et al., 2015 the researchers try to study the way

people relate to each other for understanding better their patterns of mobility that may be a vehicle for cholera in this case.

In a more recent study the authors use the mobility data from anonymized CDR to study the potential of spread of diseases such as rubella Wesolowski et al., 2015.
There is more investigation on geo-referenced social network data which are key for the applications studied.

2.3.5 Studies on Geo-tagged social network data

Analogously to CDR data, insights gained by geo-referenced social data can enable studies and have also interesting applications in different areas For example in Lambiotte et al., 2008 the authors translate the coordinates of geo-tagged content in a geographic map and study it from a statistical perspective.

In the Steiger et al., 2015 census data are used to explore the semantic association between geo-referenced tweets and their respective spatio-temporal whereabouts. By applying a semantic topic model classification and spatial autocorrelation analysis the authors detect tweets indicating specific human social activities.

A series of works of minor importance that use location social network data can be found in Noulas et al., 2012 Barchiesi et al., 2015 Illushka, Tobias, and Susannah, 2015 Chareyron, Da-Rugna, and Branchet, 2013 Vu et al., 2015.

2.3.6 Business activities improved by the use of location data applications

In particular the applications of geo-referenced data from social networks are similar to the CDR data. However those data sets though sparse have a rich semantic content such as tags and text describing pictures or status updates.
The text may express a personal opinion on politics, preferences, habits and so on. This information that have a location dimension can be used in business areas such as marketing and sales, business strategy and competitors analysis.

Location analytic is one of the fastest growing fields and companies are increasingly embracing this discipline to create ever more sophisticated applications. It is difficult to make sense of disjointed raw data that become quite meaningless if no such dimension as location is used to provide context. Fusing the raw data and drilling down the figures generates insight that allows for long-term strategies to be applied.
For example, location analytic can show a business how sales could be affected by a competitor opening a branch nearby.

It can also help in advertising products or services as the work in Quercia et al., 2011. In this work the authors provide a way for measuring audiences of outdoor advertising in specific areas based on the combination of mobile phone location estimations with Internet listings of social events.

Demographics, population density and distance from other points of interest can be assessed to see if there are any unique location advantages. This information may give the company a chance to create future plans to cope with their competitor's arrival or to compete the strategy of competitors.
In the next paragraphs there are described in detail how a business processes such as marketing may benefit from location social network data.

1. Analyze market and product demand: by filtering social media conversations by location the analyst can pinpoints where a company and its competitors, are most active on a specific kind of product or service.
   This would allow for a better brand monitoring and boost marketing efforts at that specific location.

2. Spot cultural trends: by segmenting social media data by country or region it is possible to understand better the 'language' and cultural orientation of customers. Understanding cultural differences, marketing teams can learn which trends and products are most popular in certain locations and optimise campaigns for their location-based popularity accordingly.

3. It also may help for sentiment analysis. For example, analyzing content of customers enjoying a product or disliking it at a certain location or region will increase or decrease a brand's popularity.

In the previous paragraph there is an example of how location data analysis may help marketing but the same may be applied to many other business processes such as predicting particular business costs, understanding the impact of pricing strategies etc..
Location analysis is increasing the chances of improved profitability in the future and it boosts business strategy by providing companies with a clearer picture of an area's potential for profit [4]

2.4 Conclusions

With the technological advance on electronic personal devices a huge mole of data is generated.
In this scenario of data explosion, with the presence in social networks or simply by owning a smart phone personal data is also released.
A main importance in this context has privacy and related issues which in this thesis has a dedicated chapter.

In our domain of research the most interesting type of data is of course mobility data but in general all geo-referenced data provide a way to understand how people use the city's infrastructure from the point of view of

[4] http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1019&context=icr

mobility (do they move by car, train or bicycle), consumption (energy, water,waste) and environmental impact (noise, pollution).

More importantly, mobility data offers new insights about the city, which are of great interest for city managers both from an economic and a political perspective.

Chapter 3

Data Fusion for re-identification of anonymized Call Description Records

3.1 Introduction

In this chapter of the thesis is presented an analysis of different anonymized mobility data sets. The analysis aims in highlighting re-identification and information fusion possibilities. In particular the analysis focus is on CDR data sets, released by mobile telecom operators and data sets comprising geo-localized messages released by social network sites.
Results shows that: *(i)* in line with previous findings, few (about 4) data points are enough to uniquely pin point the majority (90%) of the users, *(ii)* more than 20% of CDR users have a single social network user exhibiting a number of matching data points. Given this information, this work studies the probability that these two users might be the same person. *(iii)* The study derives an estimate of the probability of two users begin the same person given the number of data points they have in common, and estimate that for 3% of the social network users we can find a CDR user very likely ($> 90\%$ probability) to be the same person.

3.1.1 How to read (and understand) this chapter

The chapter uses as work flow the diagram 1.2 in Chapter 1 that shows how knowledge from different data sets is extracted and fused in order to describe better the observed object. In this case the object of study is the user. He / she and his activity in terms of mobility traces can be described by the CDR data set and by the the social data set.

Each data set provides a description of this activity as a mobile phone user and as a social network user respectively. Thus, once that this description is transformed into knowledge it can be integrated into a unique representation that can be used for different purpose. One of the most important parts of knowledge extraction is the evaluation of individual paths of a user and their being unique. For this approach of data fusion, this information is key in understanding how the data sets relate with each-other.
In general the work flow of this chapter follows the diagram in Fig. 1.2 in Chapter 1 where the extraction of knowledge happens as described in the previous paragraph through the uniqueness test of the mobility traces of

the users, and probabilistic matching of their traces.
More in detail the content of this chapter is organized as follows. Section 2 presents research at the state of the art in entity matching among multiple data sources. Section 3 illustrates the CDR and social network datasets used for this analysis. In Section 4 is presented the uniqueness evaluation analysis performed on mobility traces. Section 5 presents initial re-identification results based on counting the number of matches among events generated by users across the two data sets and some ground truth evidence. Section 6 presents the probabilistic model to assess whether different users are actually the same person and provides experiments in this direction. Finally in Section 7 are presented the conclusions of this work.

3.2 Re-identification problem

Mobile phones track the activities of their users along many dimensions. For example, telecom operators can access the location of their subscribers at a fine grain scale, e.g., via the use of CDR (Call Detail Record) data. Similarly, the vast majority of applications installed on smartphones (especially those related to social network sites) collects and stores on the Web information about users locations and activities (e.g., by means of geo-localized pictures, or geo-localized messages and tweets).

The analysis of multiple data sets on users' behaviors opens interesting information fusion possibilities and, at the same time, creates a potential for re-identification and de-anonymization of users' data. On the one hand, this kind of approaches can breach users' privacy despite anonymization. On the other hand, combining different data sets is a key enabler for advanced context-awareness in that information from multiple sources can complement and enrich each other.

As location and mobility are primary sources of context-information for several applications, the use of this kind of data is important and can have a strong impact in the fulfillment of the pervasive-anticipatory computing vision as described in Pejovic and Musolesi, 2013. Furthermore, in developing countries, there seems to be a direct connection between mobile telephone services penetration and economic development as location and context based services can improve productivity many sectors of economy as research shows in Abraham, 2006 Blondel et al., 2013.

In this scenario, an important but problematic activity consists in joining data sets by matching different users associated to the same real person. For example, it would be interesting to realize that user X in a CDR dataset is actually the same person as Twitter user Y, and then join the two data sets. Geo-located data presents a natural framework to evaluate users' similarity and enable re-identification (join) across multiple data sets.
The matching process across multiple data sets is in fact rather straightforward in principle and consists in identifying whether CDR user X and Twitter user Y consistently produce data at the same time and place. Once enough geo-referenced elements overlap, we can be reasonably sure that

the two users are actually the same person.

On the one hand, this could raise serious privacy issues, as relations between different types of data can be used to infer information of any kind from socio-economic status, to mobility and shopping patterns, to the user's social graph as illustrated in Wicker, 2012. This is particularly problematic once the process of matching users among data sources allows to bypass the anonymization of a given dataset.
As work in Wicker, 2012 describes it may in fact happen that: "The continued accumulation of location data may reach a point where a marketer can uniquely match an anonymous location trace to a named record in a separate database".
On the other hand – for the same reason – joining different data sets is the key for advanced forms of context awareness that could notably improve pervasive applications and services. In fact on the basis of such a combined dataset, it would be possible to infer what the users were doing in a given location and their general profile.

The contribution of this section is to conduct analysis and experiments in this direction. In particular:

1. to collect and synthesize knowledge about the two datasets. In more simple terms it means to analyze the uniqueness of mobility traces as a first essential step in order to quantify the re-identification potential. Uniqueness analysis, pioneered in Montjoye et al., 2013, consists in evaluating how many (geo-located) data points are necessary to uniquely identify a user in a dataset.

2. The next step is the re-identification method itself which can be understood as the data fusion process. Specifically, the data fusion process tries to answer the following question: *Can we use data from geo-referenced social networks, to re-identify mobile users from an anonymized Call Description Records data set?* To provide an answer to this question a probabilistic approach is used in order to evaluates the probability that users from multiple data sets are actually the same person.

3.3 Dataset

We got access to two CDR data sets describing mobility traces of a large user population over an extended period of time.The first data set (referred to as CDR-DATA1) has a time span of one month. It comprises records of each cell network event whether associated to incoming/outgoing call,internet connections and text messages made by a mobile device, its time-stamp and the geographic coordinates of the cell tower handling the event. This latter information is extracted from a table containing the coordinates of each cell tower and approximate area of coverage.
In Fig 3.1(a) - it is shown the structure of a a CDR record. Each record comprises a user (hashed) id , the MCC (Mobile Country Code), the time-stamp of the event, the code of the cell tower and the coordinates and coverage radius of the cell tower.
Thus, the spatial resolution of CDR localization is the cell radius. Figure

a)

User id	MMC	Timestamp	Tower id	Coord	Radius
3dd285b	222	734628648723	123	(41.28,13.92)	450

b)

User id	Timestamp	Tower id	Coord
12d285b	734628648723	123	(41.38,13.22)

c)

User id	Pic/Tweet	Timestamp	Coord
12d285b	Text	734628648723	(41.28,13.92)

FIGURE 3.1: Records for each dataset. a) CDR-DATA1 dataset, b) CDR-DATA2 dataset, c) FT dataset

3.2(a) illustrates the distribution of CDR-DATA1 events per user. Figure 3.2(b) illustrates the radius of gyration for a given percentile of users.
The radius of gyration is a synthetic and easy-to-compute parameter describing the spatial extent of user traces. It is defined as the deviation of user positions from the corresponding centroid position. It is given by:
$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_{centroid})^2}$
where p_i represents the i^{th} position recorded for the user and $p_{centriod}$ is the center of mass of the user's recorded displacements obtained by: $p_{centroid} = \frac{1}{n}\sum_{i=1}^{n}(p_i)$.
It is possible to see that the first quartile tends to be associated to sedentary people with $r_g < 5Km$. The bulk of the distribution (25^{th}-75^{th}) percentiles can be associated to urban mobile people as the diameter of peri-urban areas of main cities in the region is about 15Km. Users beyond the 75^{th} percentile are associated to commuters spanning on the wide region area.

The second dataset (referred to as CDR-DATA2) records have the same characteristics as those of CDR-DATA1 and are shown in Fig 3.1(b). The only difference is that there are not given the radiuses of the cells and the mobile country code of the user. This dataset is not that big compared to CDR-DATA1 as the number of mobile users is smaller.

The distribution of the events generated during a period of two weeks is given in Figure 3.2(c) as events per user. The radius of gyration is represented in Figure 3.2(d).
With regard to the radius of gyration it is possible to see that some users travel for longer distances than in the CDR-DATA1 dataset. This is because the CDR-DATA2 data spans in a much bigger area than CDR-DATA1.

3.3.1 Social Network Data

Social network data has been captured by using the REST API services provided by Flickr (*www.flickr.com*) and Twitter (*www.twitter.com*). Overall we refer to this dataset as FT data.
In Figure 3.1(c) is given how the records of this dataset are structured. Starting from a pool of 2456 users and having discarded from this dataset users with only one or two events, we consider for our analysis 965 (231 Flickr users and 739 Twitter ones) of them. We discarded users with less than 3 events as those provide very little information on their location history.
At the end, there is an average number of 21 events per user and a max/min

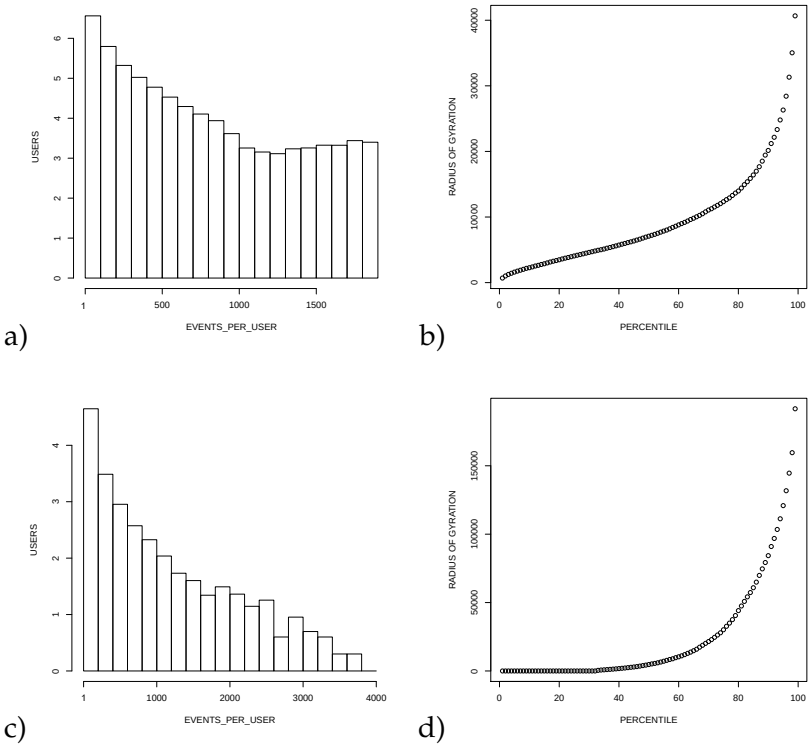


FIGURE 3.2: **(a)** Events per user in CDR-DATA1 dataset. **(b)** Radius of gyration for a given percentile of users in CDR-DATA1.**(c)** Events per user in CDR-DATA2 dataset. **(d)** Radius of gyration for a given percentile of users in CDR-DATA2 dataset.

value respectively of 686 and 3 events per user. Figure 3.3(a-b) shows the distribution and the radius of gyration of Twitter users.
Figure 3.3(c-d) shows the distribution and the radius of gyration of Flickr users. The radius of gyration in both Twitter and Flickr is rather low, since we extracted information only from a limited bounding box around the area of interest.

3.4 Mobility traces analysis as extraction of knowledge

A key property in any analytical model is uniqueness and it is necessary in order to achieve interpretability. The extraction of knowledge implies In this section, is evaluated the quantity of information, in terms of geo-referenced events, that is necessary to identify a mobility trace as unique.
Another way to deal with this problem is to find out for each user, what is the minimum number of events that set him/her apart from all the other users. This basically allows to answer the question: *who is the user who was in these locations at these times?* When the number of locations and times increases there is a *single* user compatible with all the sightings.
A simple representation of this idea is given in Figure 3.4 where the trace of user A is highlighted. In the figure, it is possible to see that as soon as we consider three events, we can pin point user A as no one of the other users have the same events.

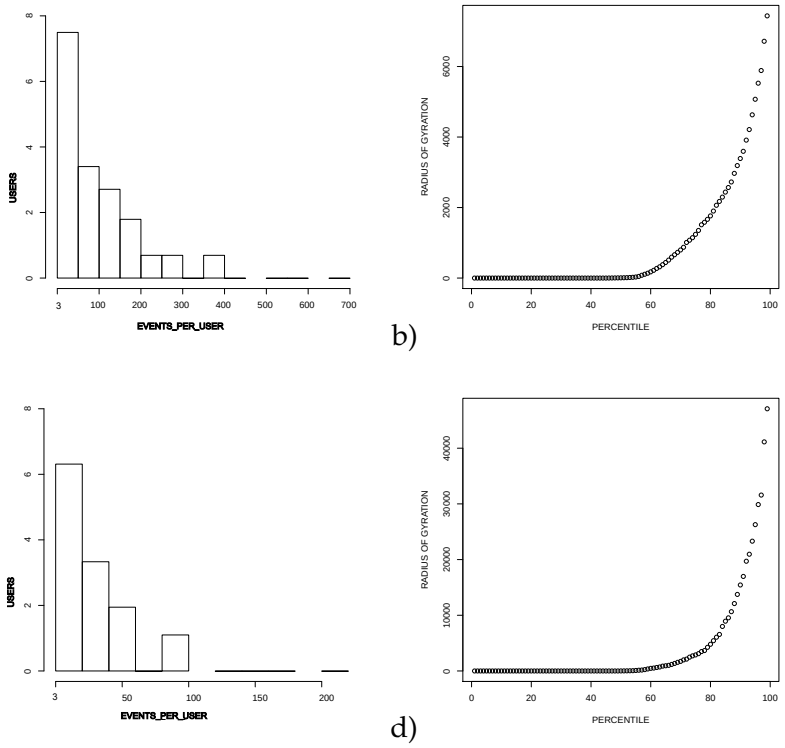


FIGURE 3.3: **(a)** Twitter events per user. **(b)** Radius of gyration for a given percentile of Twitter users.**(c)** Flickr events per user. **(d)** Radius of gyration for a given percentile of Flickr users.

3.4.1 Uniqueness Test on CDR Datasets

More formally, for each CDR user u_x we consider their data trace in terms of id of the network cell and $time$ of the CDR event. We call cdr_j^x the $j-th$ event generated by user x. Given a time interval Δt, we say that two CDR events from users x and y match if $cdr_j^x.id == cdr_i^y.id$ and $|cdr_j^x.time - cdr_i^y.time| < \Delta t$. We call M_j^x the set of users having at least one event (M)atching with cdr_j^x.
For each user x, we want to find the minimum number of events min so that $\bigcap_{j=0}^{min} M_j^x = \{u_x\}$. This is the minimum number of events necessary to set u_x apart from other users. The idea of a time limit is to take into account uncertainty on when the user created an event. So while there could be always just a single user generating an event at a given millisecond (thus $min = 1$), when we have uncertainty about the time of an event multiple users might be a match.

3.4.2 Experiments

We perform experiments with the CDR datasets (CDR-DATA1 and CDR-DATA2) to find the average number of points/events to pinpoint a user. To limit the computational effort, we take a sample of 1000 users from CDR-DATA2 and of 1000 users from CDR-DATA1 dataset. The uniqueness test is performed by an algorithm implementing the above definitions. In doing this type of analysis, we are interested basically in two types of results:

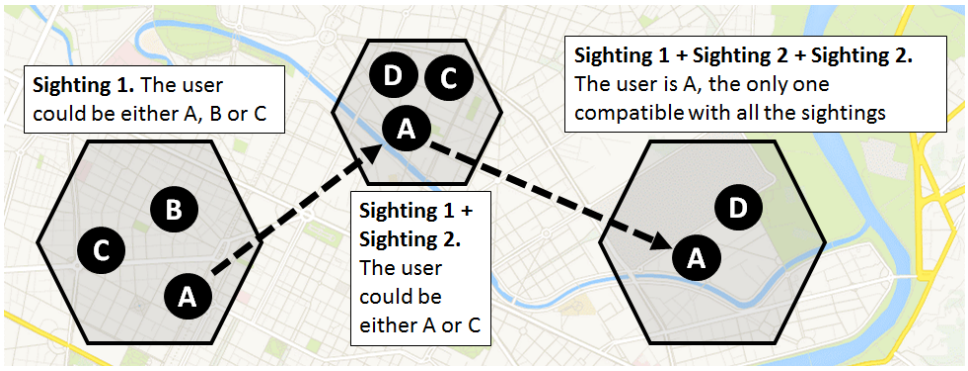


FIGURE 3.4: Uniqueness of mobility traces. After three sightings (geo-referenced data) we can uniquely pinpoint user A

1) the average number of points/events needed for a trace to be unique and 2) the percentage of users that could be re-identified in terms of identification as unique. The test for matching the events is repeated with different time limits Δt that go from 2 to 30 minutes.

The summary of this analysis is given in Figure 3.5 which provides results for the CDR-DATA1 and CDR-DATA2 datasets. These results are in line with those obtained by related studies summarized in
Looking at Figure 3.5-left it is possible to see that the number of CDR events to uniquely identify a user slowly grows with time interval. This is a rather natural phenomenon in that increasing the time limit loose the constraint about matching CDR. In the extreme case of time limit $\Delta t = 0$, all the users would be uniquely identified with only 1 event.
As pointed out also by other authors in Montjoye et al., 2013, the slow growth in the number of points required for uniqueness illustrates that anonymization techniques based on blurring the time of the events are not effective.
These considerations are also represented in Figure 3.5-right: the percentage of users that can be uniquely pin-pointed remains high (over 80%) independently of the limit interval. In both the graphs, it is possible to see that it is slightly easier to uniquely identify users in the CDR-DATA1 dataset.
This is probably because of the denser cell network in CDR-DATA1, rather than in CDR-DATA2 – (denser cell means higher spatial resolution in the dataset). Also in this case, however, the difference between the two results is not dramatic, showing that the spatial resolution does not impact uniqueness a lot as stated in Montjoye et al., 2013.
In general these results are important to highlight issues related to CDR datasets with regard users' privacy. However, to evaluate the potential for joining different datasets, this is not enough: even if a user is uniquely found in a dataset on the basis of their traces, it might be still difficult to find the same users also in other datasets, thus enabling the join.

3.4.3 Matching CDR and social network data.

As already introduced, matching users on the basis of mobility and geo-referenced data is in principle straightforward: it consists in identifying couples of users in the two datasets that often produce data at the same

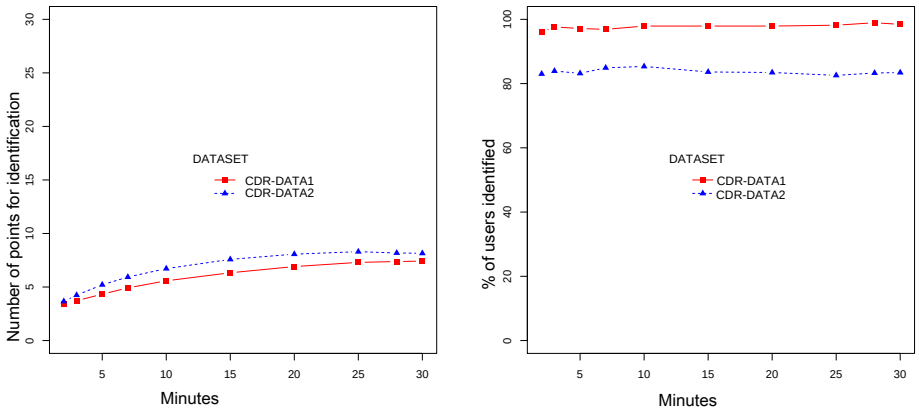


FIGURE 3.5: Uniqueness of mobility traces. (left)Number of events needed for identification.(right)Percentage of users re-identified

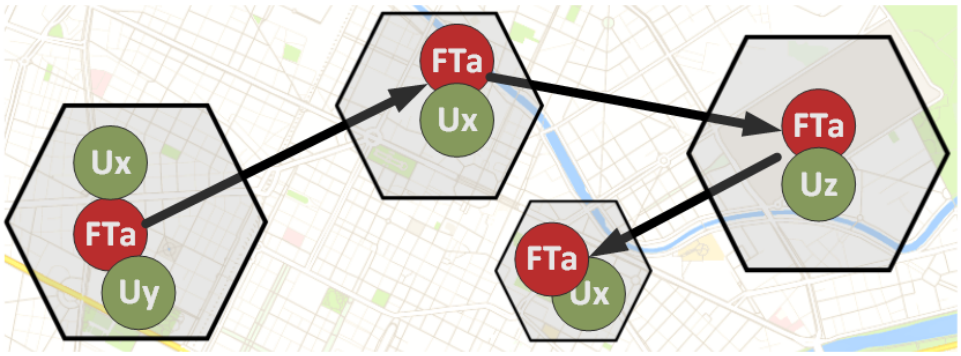


FIGURE 3.6: Flickr/Twitter user FT_a (int red) is compared with CDR users u_x, u_y and u_z (in green). The most probable match is $FT_a \equiv u_x$ as they are consistently producing data at the same time and at the same place.

time and at the same place. Once enough geo-referenced elements overlap, we can be reasonably sure that the two users are actually the same person.

Figure 3.6 exemplifies the matching process: Flickr/Twitter user FT_a is compared with CDR users u_x, u_y and u_z from the CDR-DATA1. The most probable match is $FT_a \equiv u_x$ as the two users appear almost always together.

The key novelty when considering multiple datasets is that the events from the same user in different datasets are *not* always in a 1-to-1 relationship. In contrast to the previous scenario, we can consider users in different datasets to be the same even if not all the events in one dataset have the corresponding events in the other dataset.
In more technical terms, the intersection we used in Section 4 to uniquely identify users can produce the empty set. Still, given a sufficient overlaps among the data we will consider two users to be the same.

More in detail, for each Flickr/Twitter user FT_a and for each CDR user

u_x we consider their respective data traces with a geographic and temporal reference.
We confront these traces in order to find matches among couples of events. More formally, we call ft_i^a the events generated by user FT_a, and cdr_j^x the events generated by user u_x.

We also call r_j^x the radius of the network cell associated to the event cdr_j^x. Two events ft_i^a and cdr_j^x match if $s_{dist}(ft_i^a, cdr_j^x) < r_j^x$ and $t_{dist}(ft_i^a, cdr_j^x) < \Delta t$, where s_{dist} and t_{dist} are the spatial and temporal distance respectively.

We set the threshold for the spatial distance to r_j^x as the radius of the cell where the CDR event originates (we conducted experiments with the CDR-DATA1 dataset where this radius is reported. In other situations similar measures might be derived from the Voronoi tessellation of the network cells as described in Blondel et al., 2013.

In particular, for each couple of users we count how many matching events are there. Since events are often produced in bursts, a naive matching approach between the traces could result in matching a number of CDR events with a single FT event, or vice versa.

This would over-count the number of matching pairs as a single CDR/FT event could match with multiples FT/CDR events.
To avoid such an issue, after recording an event cdr_j^x we skip all the events from the same users closer in time than Δt from cdr_j^x. In this way any given FT event can match with only one CDR event of a single user.

Exclusion Condition. It is also important to consider the situation in which a CDR user u_y generates an event cdr_j^y that is close in time to an event ft_i^a generated by the Flickr/Twitter user FT_a, but the two events are far away in space.
More formally $t_{dist}(ft_i^a, cdr_j^x) < \Delta t$ and $s_{dist}(ft_i^a, cdr_j^x) > 2 \cdot r_j^x$. In this case, it is impossible that the two users are the same in that we know that on that time they were physically apart.
We call this case the *exclusion condition* as these CDR users can be excluded from the analysis.
It is worth noticing that we considered $2 \cdot r_j^x$ in order to have a buffer-zone separating a match from a certain exclusion.

3.4.4 Experiments

Experiment 1 - Matching. In a first experiment we tested the above approach to find, for any FT user, the CDR user having the largest number of matching events. We say we re-identified the user if there is a *single* CDR user with such a number of events. The naive assumption is that such a single best-match is the same person of the FT user (hence the re-identification) – more on these aspects is in Section 6.
In Figure 3.7 we present the average number of points required for a re-identification and the percentage of the users than can be re-identified. We experiment with time intervals Δt from 10 seconds to 30 minutes. In general the figure shows that about 20% of the FT users can be re-identified

with about 4-5 matching events.

To support our study it is also interesting to consider the smartphones' usage pattern focusing on sms/calls and Internet connection requests for uploading a picture to Flickr or sending a Twitter message.
The transition from one activity to the other should be happening in a short interval of time that would permit us to link a CDR with a FT event. Of course, the presence and frequency of this behavior is very important for the results of our study.
Research in Verkasalo, 2010 shows that voice call/sms functions and Internet browsing are the two most used functionalities, and the largest part of Internet activity is spent on social network sites and is mainly conditioned from the places and social context in which the user happens to be as in Do, Blom, and Gatica-Perez, 2011.

According to statistics reported in Verkasalo, 2010 the voice call/sms functions are used 24-27 days per month by 90% of users. However, these data correspond approximately also to our CDR dataset statistics and we find an average of 3 events per day. On the other hand internet browsing is used 13-18 days per month by 85% of users spending 60% of the time in social network activities.
Considering the above statistics, we assume that the call/sms-social network pattern, happens very often even in a short interval of time, thus enabling our approach to detect correspondences.

In Kang, Seo, and Hong, 2011, authors report that: "though there exists large variation in usage patterns across users, it is possible to built a state-transition graph to model user behavior from a large dataset.
For doing that only a few significant states and transitions are required". As for our purpose of study, the significant states are call/sms activity and Internet data communication activity, as well as the transition state from one state to the other.
Experiment 2 - Matching Statistics. We run experiments to count, for each FT user, the number of CDR users having 1,2...n matching events with FT. Rather naturally, the number of CDR users that match with a FT user diminishes as we look for a larger number of events in common between them.
In many cases, it converges to one after a certain number of points n which means that there is only one CDR user that has n points in common with the FT user.

As reported in Montjoye et al., 2013; Zang and Bolot, 2011, the CDR user count (especially for a high-number of matching events) do not change significantly by changing Δt or by using a different threshold for the spatial distance.

Figure 3.8(left) illustrates the statistics of the matching process. For a given number of matching events, we create a box plot. Each box plot describes the distribution of the number of CDR users having n matching events with the FT users.
In the box plot, we report 25^{th} percentile, median, 75^{th} percentile. The boxplots extends 1.5 times the interquartile range. Such counts are in the log-scale to better appreciate the differences. It is easy to see that there are a lot

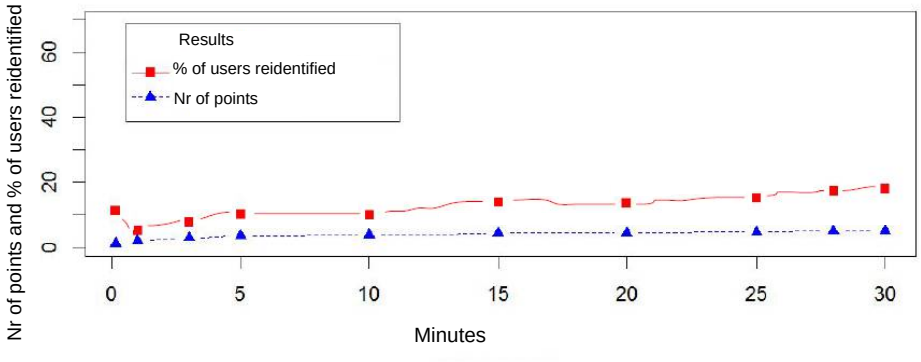


FIGURE 3.7: Percentage of users re-identified and number of points needed for re-identification

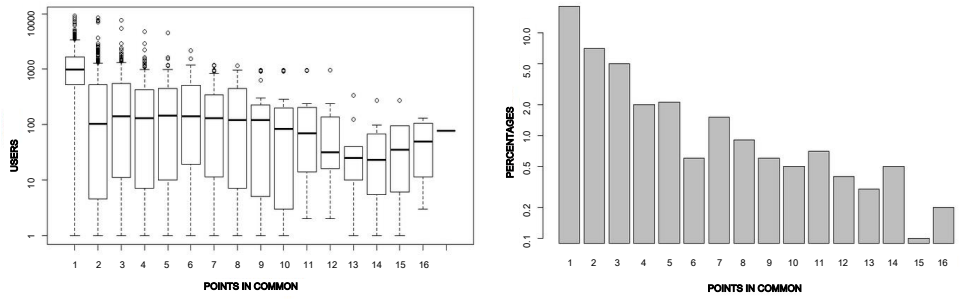


FIGURE 3.8: (right)Boxplot diagram summarizing the statistic of the number of CDR users having X matching events with FT users. (left)Percentage of FT users for that can be associated to X number of CDR users

of CDR users sparsely matching with FT users, while the number decreases for multiple matches.

Figure 3.8(right) shows the percentage of cases in which those correspondences converge to $1, 2, \cdots, n$. In particular 23% of the FT users have in common a certain number of points with only a single CDR user. *Does it mean that the two users are actually the same?*
Of course, if the number of matching events would be large, we could be confident of an affirmative answer. Otherwise, such coincidences could be made by chance. We will discuss further this question in the next Section making some more considerations.

Experiment 3 - Matching By Chance. We tried also to understand if the matching events can happen only by chance (this of course would invalidate our analysis). In particular, the Bonferroni's principle illustrated in Rajaraman and Ullman, 2011 states that in a large dataset, like the one we consider, there are *always* some events matching a given signature, even if the events are produced randomly. In our settings, this means that even considering random FT users there will be CDR users that will resemble (by chance) FT users.

In order to avoid incorrect conclusions, and to validate our results we made a comparative experiment. For this purpose, a new dataset of mock-up FT users has been created on the basis of an artificial distribution of values (geographic coordinates and timestamps).

1. We created a first mock-up dataset by adding some noise to the original FT users' traces. In particular, we create a clone of each FT by displacing both in space ($\pm 1Km$) and time ($\pm 30min$) the original user's events.

2. We created a second mock-up dataset by just randomly generating events within the area and time frame of the original data.

Testing with these fake users we find significantly less correspondences and less traces of series of points overlapping. More specifically, in the first case, we find 40% less correspondences per user and in the second up to 60% less correspondences than in the case with the real data. The fact that our data exhibits much larger correspondences means that most of the matches cannot be explained by chance only.

Experiment 4 - Validating Results. Since groundtruth about the real identity of the users is missing, sound validation of the results is difficult. Partial validation evidence can be found in matching Flickr social users with Twitter social users.
We have 231 users from the first social network and 734 from the second, and during the matching we notice that about 6% of Flickr users, other than having similar traces have the same name or the same very singular username with Twitter users.

Moreover those users with the same username often match with a single CDR user which is the same for Flickr and for Twitter users. Figure 3.9(top)

shows the traces of two exemplary social users with the same username. Figure 3.9(bottom) shows that a single CDR user well matches will both of them. The three users are thus likely to be the same person.

A similar analysis can be conducted by comparing the CDR mobile county code (MMC), typically indicating the nationality of the user, with the language used in Twitter messages and Flickr tags.
This type of analysis can also give an idea of the number of tourists that visit the region during the period of observation. In our CDR-IT datset about 15 million events have a mobile country code different from that of Italy and a time window permanence of 1-2 weeks (see Figure 3.10-left).

In the social dataset there are more than 200 users that use a language different from Italian. Figure 3.10(right) shows along the y axis in logarithmic scale, the number of events and along the x axis the different languages we found in analyzing the text.
To extract in an automatic way the languages of the social users we use an open Java library for language detection[1].

A common problem with this approach is that for example users from Switzerland can tweet or describe pictures in different languages such as German, French or Italian as well as people from nordic countries tend to tweet in English. Thus only 3% of the social users could be re-identified supported by the match between Mobile Country Code and social users'language.

Despite the scarce efficiency, this method functions well for the re-identification of users comming from small countries that remain faithful to their language while tweeting or describing pictures. From our point of view this provides one more clue in supporting ground truth evidence for the matching process.

Experiment 5 - Reidentification with CDR-DATA2 Dataset. In analyzing the CDR-DATA2 data set we realize that it presents a challenge for our re-identification method. The dataset contains different subsets created in subsequent periods of time where in each subset there is a new random numeric identifier for the users. We choose 1500 users from the first subset and confront these users with all the users in the other subsets.

Using the same approach as in the previous section, we repeat the matching process by time and tower id. While in the case of matching CDR-DATA1 and FT users we had data from the same period in this case we are matching data from different subsequent periods of time so we use time slots instead.
This means that every time two users - the one from the 1500 sample and every other user in the other subsets - made an event with the same tower id and in the same time slot we had a match. In Fig 3.11 is represented an error bars plot about the matching process up to 50 points/events in common, while the percentage of users we can re-identify this way is 88.7 there

[1]https://code.google.com/p/language-detection/



FIGURE 3.9: **(top)** Matching mobility traces between Flickr events in blue squares and Twitter events in red round shape of the same social user. **(bottom)** CDR events from a user in black triangle. The social user in its both profiles as Twitter and as Flickr user matches with the same CDR.

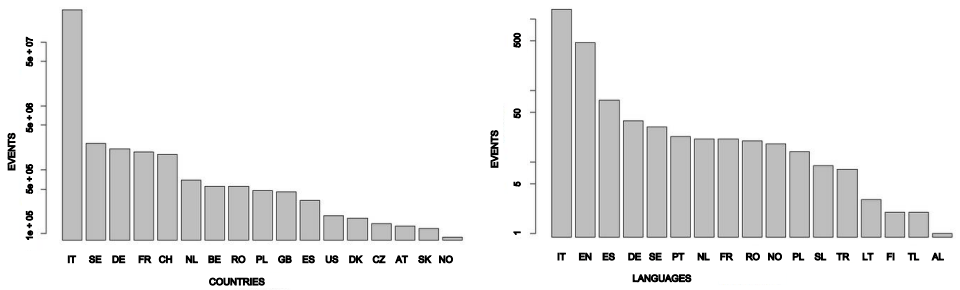


FIGURE 3.10: (left)Distribution of events per mobile country code.(right). Distribution of social users languages

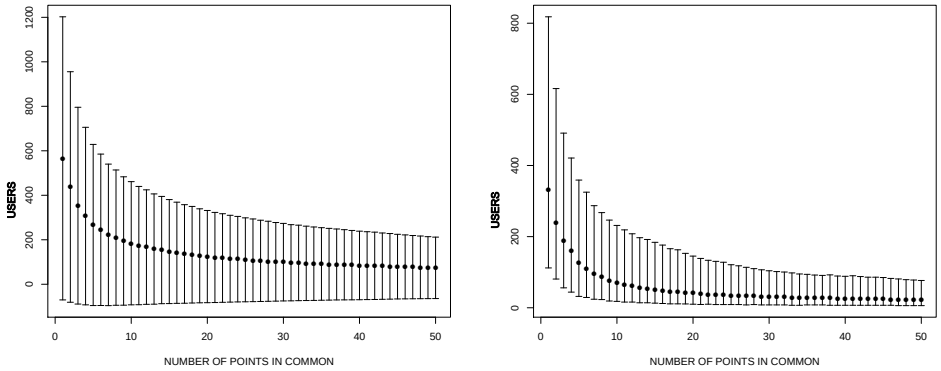


FIGURE 3.11: (left)Error bars plot of the number of CDR-DATA2 users from the sample chosen having X matching events with all other users.(right) The quartiles of the same distribution, 25th and 75th percentiles.

is a new assignment of random identifiers for the sampled users.

As in Blondel et al., 2013 is not stated that each period contains new users that are totally different from those of any other period, we decided to try and re-identify a sample of users. We choose 1500 of them from the first subset and confront these users with all the users in the other subsets. Using the same approach as in the previous section, we repeat the matching process by time and tower id.

While in the case of matching CDR-DATA1 and users we had data from the same period in this case we are matching data from different periods of time so we use time slots instead. This means that every time two users - the one from the 1500 sample and every other user from 1 to 9 subsets - made an event with the same tower id and in the same time slot we had a match.
In Fig 3.11 is represented an error bars plot about the matching process up to 50 points/events in common, while the percentage of users we can re-identify this way is 88.7 %.

3.5 Probabilistic Approach

In this section, is presented a model trying to answer the following question: *given that the CDR user u_i has n_i events matching with the events of the Flickr/Twitter user FT_i, how likely it is that the two users are the same? In other words, how likely it is that we actually de-anonymized CDR user u_i?*

This kind of question is both novel and fundamental for the practical applicability of our approach. In previous seminal works, as for example in Montjoye et al., 2013,this issue is not addressed. In Montjoye et al., 2013, it is assumed that for each FT event there is *always* the corresponding CDR event. Under such an assumption, re-identification reduces at finding the only CDR user that is *always* around the FT user.
In other words, once we found that a FT user has n matching events with a

single CDR user, we are 100% sure that the two users are actually the same, even if the number of matching events n is 1.
In our setting, since we removed the assumption that for each FT event there is *always* the corresponding CDR event, such a result is unreasonable in that a single common data point might be due to chance (as described in Section 5.1).

To answer this question in this work the approach follows the paper in Narayanan and Shmatikov, 2008 in which they tackle the problem of re-identifying users in movie-ratings databases. In the present work the approach starts by estimating the probability of a CDR and a FT users being the same person *given* they have n matching elements – $p(CDR \equiv FT|n)$. By definition of conditional probability, $p(CDR \equiv FT|n) = p(CDR \equiv FT, n)/p(n)$.

This way it is possible to compute $p(n)$ by counting in the dataset how many couple of users (CDR/FT) have n matching events. Finally this count is divided by the total number of couples (CDR/FT).

To illustrate this approach is used a simple numerical example. Suppose that in the the whole dataset there are only 3 CDR users (A,B,C) and 3 FT users (X,Y,Z). Assume that they have a number of matching events as in the following table:

Matches	**A**	**B**	**C**
X	30	1	3
Y	1	3	1
Z	1	1	3

For example CDR user A has 30 matching events with FT user X, 1 matching event with FT user Y and 1 matching event with FT user Z. In this setting we have:

$p(n = 1) = 5/9$, $p(n = 3) = 1/3$ and $p(n = 30) = 1/9$.
Computing $p(CDR \equiv FT, n)$ is more difficult as we do not have groundtruth information on the cases in which $CDR \equiv FT$. To estimate such a probability we *assume* that whenever a CDR user matches with a FT user such that there are not other CDR users having the same or greater number of matches, then the CDR user and the FT user are the same person (This is the same assumption made in Section 5.2 - Experiment 1).

Accordingly, $p(CDR \equiv FT, n)$ is the number of couples of users (CDR/FT) having n matching events in common *and* such that there are not other CDR users with a greater number of matches for the same FT user, *divided* by the total number of couples (CDR/FT).

In the example, $p(CDR \equiv FT, 30)$ = 1/9. In fact, looking at the table, 30 appears once associated to the couple $(A - X)$ and there are not other CDR users with a better match with X. $p(CDR \equiv FT, 3)$ = 2/9. In fact, looking at the table, 3 appears twice in situations in which there are not better matches (i.e., 3 appears in $B - Y$, $C - Z$).

We can then estimate $p(CDR \equiv FT|n) = p(CDR \equiv FT, n)/p(n)$. In the example, $p(CDR \equiv FT|n = 30) = 1$, $p(CDR \equiv FT|n = 3) = (2/9)/(1/3) = 2/3$.

In the example it is possible to see that although B and C are the users that best match Y and Z the re-identification probability is *not* 1. This is because matches with 3 events are rather common in the dataset, so we are less confident of the re-identification.
Considering the real dataset with a large number of CDR and FT users having multiple matches among each other, the approach becomes much more effective.

Since there are *a lot* of CDR users having few events matching with some FT users, the re-identification probability on the basis of few events drops almost to zero. Only users with an unusually high number of matching events would be considered properly re-identified.

3.6 Data fusion for re-identification of CDR records

The above model was applied in order to answer the motivating question: *given that the CDR user u_i has n_i events matching with the events of the Flickr/Twitter user FT, how likely it is that the two users are the same?*.

For each Flickr/Twitter user, are recorded all the CDR users that produced events nearby. Then the number of matches n_i among them was computed. With the above approach was estimated $p(CDR = FT|n)$. Figure 3.12-left shows the result obtained. We run experiments changing Δt (i.e., the maximum tolerable time-distance between matching events).

The figure shows that for very small Δt ($< 2min$), or large Δt ($> 10min$) the $p(CDR = FT|n)$ remains low and drops to zero as the number of matches increases. This is highly counterintuitive, we would expect that the more the matches, the more the probability of the two users being the same person. This can be explained considering that for very small Δt there are 0 users generating n matches with any FT users (for $n > 3$).

For large Δt, it is possible to have multiple matches also with different users. So, while $p(n)$ increases, $p(CDR = FT, n)$ does not keep up because (under our computational assumptions) there are few cases in which a single CDR user rises as the single most compatible one.
For $\Delta t > 2mins$ and $\Delta t < 10mins$ we have the expected behavior: the higher the number of matches the greater the probability. Probability that quickly rises to almost one after 4-5 matching events (as expected). Accordingly, we assume this is the proper range to consider for further experimentation.

It is worth noticing that this kind of measure "solves" the issue raised at the beginning of this section, even if there is a *single* CDR user with n matching events with a FT users, if n is small, the chance ($p(CDR = FT|n)$) of the two users being the same person is low. Viceversa, if there are *multiple* CDR users with n matching events with a FT users, if n is large, the chance of all

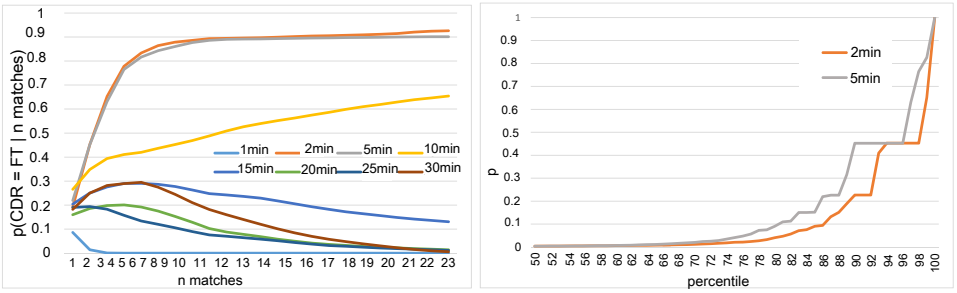


FIGURE 3.12: **(left)**. Probability $p(CDR = FT|n)$. Plots for different values of Δt used in matching process. **(right)**. Maximum probability obtained for a given user percentile (we do not plot below-50 percentile as the probability is almost constant 0)

the users being the same person is high (e.g., a user with multiple phones).

Finally, another experiment to measure the fraction of users in our CDR-IT dataset that can be reliably matched with some FT users. Figure 3.12-right shows the maximum probability obtained for a given user percentile (we do not plot below-50 percentile as the probability is almost constant 0). It is possible to see that only the top few percentiles (still amounting at a lot of users) can actually be reliably matched with a corresponding FT user.

This is rather expected given the inherent difficulty in matching among different sparse datasets having non-deterministically correlated entries. It also highlights the difference between this task and the uniqueness identification task addressed in the related works and in Section 4.

3.7 Conclusion

The main intent of this work is to present the potential and/or limits of a data fusion technique to re-identify users across multiple mobility datasets. Results illustrates that it is possible to identify some users across different datasets that are likely to be the same person, in the sense that the match among them is largely more probable than the match with any other person.
While correlation among datasets (i.e., CDR and social network data) represent a fundamental issue with regard to privacy issues, it also represents a big opportunity to enrich the information available to a pervasive applications.
In fact correlation is also the basic building block to fuse different perspectives and information together in order to obtain a multi-faceted representation of phenomena and events influencing the application.

For example fusing CDR and Flickr/Twitter data it is possible to better pinpoint where the user was at a given time and what (s)he was doing in there. Moreover, such information could provide a much more fine grained view on the user profile enabling high forms of personalization and context

awareness. Future research on this topic will continue into three main directions:

The probabilistic model presented in the above section is rather simple. On the one hand, it is based on a number of independence assumptions that can be hardly justified in the real data.
On the other hand, we think that further analysis on the eccentricities of the resulting probability distribution could give further insights on the re-identification process as research shows in Narayanan and Shmatikov, 2008. With regard to this latter aspect, we plan to investigate and adopt standard privacy measures like k-anonimity concepts presented in Sweeney, 2002; Zang and Bolot, 2011 and differential privacy described in Dwork, 2011. Another interesting aspect are social ties as work in Crandalla et al., 2010 which is worth investigating for our social users.

As privacy concerns are the main impeding factors preventing CDR data (and pervasive/mobility data in general) to be applied to pervasive applications, it would be fundamental to develop anonymization and privacy-preserving mechanism that preserve data usefulness in the context of pervasive application.
There are a number of researches addressing such issues as for example in Brickell and Shmatikov, 2008; Zang and Bolot, 2011. However, most general approaches conclude that:*even modest privacy gains require almost complete destruction of data-mining utility* as stated in Brickell and Shmatikov, 2008; Zang and Bolot, 2011.
We think that a viable approach could be based on application-specific anonymization mechanisms: mechanisms preserving privacy and data mining utility for a *single* aspect (the one useful of the specific application). We will try to investigate these issues in our future work.

Finally, we will try to leverage the proposed re-identification approach to join multiple complementary datasets together to develop advanced context-awareness services in smart city scenarios.
In particular a good source of inspiration, of possible applications as well as advances on balancing risks and rewards on data-driven applications is given in Pentland, 2014.
Overall, we think that the above research directions will have a strong role in increasing even further the impact of human mobility data on the achievement of the pervasive computing vision.

Chapter 4

Data fusion for Event Detection and Description in Urban Areas

4.1 Introduction

The pervasive diffusion of mobile phones and social networks provides a practical way to collect geo-located information from a large users population.
The analysis of such collected data is a fundamental asset in the development of geographic information systems and applications, including location-based services, traffic management, urban planning, and disaster response Ferrari, Mamei, and Colonna, 2012; Zambonelli, 2008.

This chapter presents a study of data fusion for improving the detection and description of events in cities and urban life in general.
The results show the application of statistical anomaly detection to two types of datasets comprising both the activities recorded by mobile phone operators (call detail records - CDRs) and those recorded by the Twitter online service.

The experiments show that with a wide range of events-to-be-detected, the data fusion of different data sources is effective in improving precision and recall. In addition there are presented results in automatically describing the events via the keywords appearing in Twitter messages.

The process of data fusion will follow the flow of the chart 4.1 that analyze data from different domain, extract knowledge and then eventually fuse this knowledge to better describe or recognize the object of study which in this case is the event itself.

4.2 City dynamics and events in urban spaces

Events can refer to a wide range of situations. For example, an event can be associated to a concert in the street or in an open square, or it could be a traffic jam in a road.
It also could be a protest/riot of a relatively large group of individuals or a situation of emergency forcing people to gather in specific areas of the city as in the following related papers Dobra, Williams, and Eagle, 2015 Nurwidyantoro and Winarko, 2013 Atefeh and Khreich, 2015 Chandola, Banerjee, and Kumar, 2009.

In all these cases, the automatic detection and interpretation of this kind of events from data opens innovative services from city management and

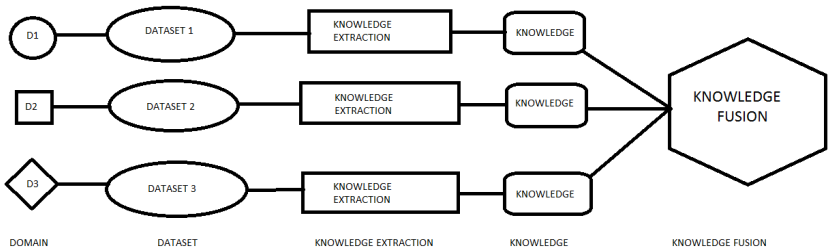


FIGURE 4.1: **Knowledge based data fusion methods**

emergency response, to social and entertainment applications.

Events in cities are registered by a wide range of sensors such as mobile phones, social networks, online media etc..
They are one of the best objects of study from a data fusion perspective as can be better detected and described by the data fusion process.
The basic approach followed to detect events is an application of statistical outlier detection by trying to model the distribution of data and identifying outliers (i.e., events) as extreme values with respect to that statistics.

The contribution of this chapter is also to show :

1. that there are many methods and algorithms used in detecting events (see e.g., Nurwidyantoro and Winarko, 2013 and Chandola, Banerjee, and Kumar, 2009 and Traag et al., 2011 and Dong et al., 2015 and Yin, Hu, and Yang, 2009 and Zhou and Chen, 2014 and "Towards Events Detection from Microblog Messages" and Popescu and Pennacchiotti, 2010 and Watanabe et al., 2011) but is by merging and fusing the different types of datasets that becomes possible to have a confirmation of the recognition of these events.
   We illustrate results obtained by combining and integrating together multiple kinds of data.

2. results to describe an event by automatically extracting keywords from social network posts generated at the event time and location.
   Anonymized CDR data are pervasive and probably the best for event detection but not for describing the event itself.
   For example the peak in the data collected by a mobile phone operator (CDR) in a given location due to a certain event may be described in real time from keywords appearing in concurrent geo-referenced Twitter messages that come from the same location.

   By fusion these two types of knowledge we have a double function : that of detecting with better precision the events, and that of describing the events in a semantic way.

4.2.1 Other studies on event detection

The literature on event-detection tends to focus more on novel methods and algorithms from single datasets and the problem data fusion is typically under-represented.
However, the recent availability of datasets from different data sources has emphasized the role of combining multiple data.

For example in Rozenshtein et al., 2014 there is given an attempt to relate a dataset from a public bicycling system with a twitter geo-located dataset.
In this work the problem of event-detection is formalized using graph-theoretic formulations and in particular as a problem of finding compact sub-graphs in graphs with vertex weights.

Another interesting work that combines different datasets to achieve event detection is Neumann et al., 2013.
Here the authors analyse aggregated data from cell phone towers and shared bicycle rentals for automatically detecting unusual behavior.
Furthermore they classify the events detected as events happening in rainy days, holidays and special days.
The recent work Botta, Moat, and Preis, 2015 analyzes the dataset releasd within a Telecom Data Challenge to estimate the number of participants to events. They show that data fusion can be useful to identify the peak of attendance to events.

4.3 Dataset description

The dataset considered was released in the context of the Telecom Big Data Challenge 2014 (openly accessible form *theodi.fbk.eu/openbigdata*) and it comprises telecom and social data associated to two cities in Italy (Milan and Trento province) during a 2 months period (from November to December 2013).

Telecom CDR data describes aggregated telecom activity of SMSs, calls and Internet connections.
From a location perspective the city area is covered with a grid in GeoJSON format of square cells of about 200m side for Milan and 1000m side for Trento province area. The Milan grid is square-shaped and covers an area of 552 km^2.
The Trento grid is bigger that the Milan one, covers an area of 6212 km^2 and it is of irregular shape. Figure 4.2 shows the grids associated to the two cities.

Figure 2(a) shows the structure of CDR data. For each cell, and for each 10 minutes time interval, the overall number of SMSs, calls and Internet connections i.e. cell activity is recorded an expressed as an activity level.
In more simple terms for every cell we have a certain value of activity at a certain hour of the day depending on time aggregation with respect to a certain type of dataset (SMSs, calls, Internet connections) and a certain group of subscribers (by nationality – country code).

Given the different measurement scale used between SMS/call activity and Internet data connection activity, these two subsets are considered as different data sets, in the following referred to as: *SMS-Call data* and *Internet connections data*.

Analogously the social dataset, describes individual users' activities on Twitter over the same cities and period.
The dataset contains all the geo-localized Twitter messages in the two cities

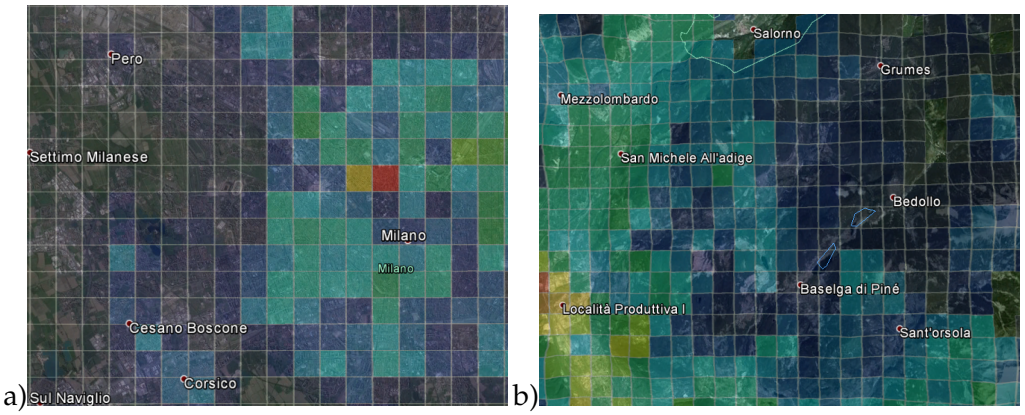


FIGURE 4.2: **(a)** Part of the Milan grid and **(a)** Trento province grid. The cells in bright red and yellow color indicate a more intense activity then the green and blue ones.

CDR dataset	Values
CellId	1
Timestamp	1383260400
CountryCode	39
SMSIn	0.1467
SMSOut	0.1523
CALLIn	0.1672
CALLOut	0.0512
Internet	11.3291

Social dataset	Values
UserId	38859420
TweetId	12319812398
TextOfTweet	Hi I'm in Milan
Language	en
Date	2013-11-08T16:22:34+00:00
Timestamp	1383924154
Location	Milan
Coordinates	(45.46282657, 9.12790569)

FIGURE 4.3: **(a)** CDR Data activity also called SMS-call dataset **(b)** Twitter Data

over the observation time period and in the same area.

Each tweet exhibits the id of the user, the unique id of the message, the time and the geo-coordinates where the tweet comes from, the text and the language being used. The figure 2(b) shows the structure of social data.

4.3.1 Spatial and temporal aggregation of the Telecom Challenge dataset

Spatial Aggregation. With spatial aggregation we group the cells tessellating the city area into larger cells.
For example, the initial Milan grid provided by the dataset contains 10000 (100 X 100) cells of 230 meters of side.

We implemented an aggregation mechanism to combine cells in larger regions. For example by aggregating the cells 4-by-4, we create a coarser resolution gird of (25 x 25) 625 regions. This kind of aggregation is useful for two main reasons:

- One problem with the original grid, is that the event area can be fragmented among multiple cells.

A partitioned event area is problematic for our approach in that the overall activity is diluted and activity spikes (on the basis of which we identify events) may become invisible.
Aggregating cells allow to cover the whole area with only one region, thus avoiding activity dilution and making events more visible. Of course, too much aggregation is also problematic, as in very large areas an event spike can be averaged-out by all the surrounding activity.

In other words, the intent of aggregating cells in regions is to adapt the analysis to the events density and magnitude.
For example by diminishing the spatial granularity (thus increasing the cell dimensions) it is possible to ignore small insignificant events and focus only in more relevant events. On the other hand by augmenting the spatial granularity even small events can be captured.
Accordingly, it is fundamental to identify the proper grid aggregation size.

- Aggregating cells simplify the analysis form a computational viewpoint. The fewer regions (i.e., the more the aggregation), the fewer operations this approach has to perform.

Temporal Aggregation. All the events we are interested span at least 1 hour so the temporal aggregation is done on the hourly basis.
This allows for better fragmentation of the monitored activity but also it will increase computational burden.

In the same way spatial and temporal aggregation reduces the numerosity of data considerably and thus computational times.
For example in Milan, for each data source, aggregating cells 4-by-4, data is reduced from a number of activity values equal to: 100 (cells) * 100 (cells) * 60 (days) * 24 (hours) * 6 (10 minutes time slots) = 86.4 million entries, to a much more manageable 25 (cells) * 25 (cells) * 60 (days) * 24 (hours) = 900K entries.

4.4 Extracting knowledge from the two datasets

In this section is presented the second step of the 4.1 diagram, that of knowledge extraction and eventually knowledge analysis.
More specifically, the knowledge regards events and their detection using a statistical analysis of the datasets described in the previous section. Knowledge extraction regards also event description.
First, is analyzed the aggregated CDR dataset(in its two parts CALL-SMS events and Internet connections) and then the social dataset.

4.4.1 Statistical Event Detection Method

The approach presented for detecting events in the dataset is an application of statistical outlier detection.
By modeling the distribution of data in terms of its quantiles it is possible to

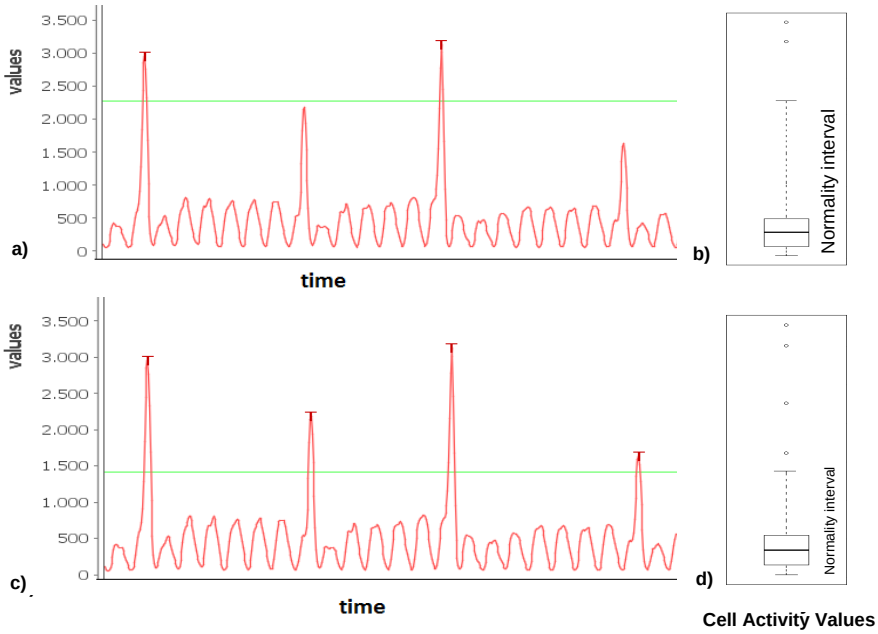


FIGURE 4.4: **(a)** High threshold **(c)** Low threshold. In **(b)** and **(d)** there are boxplot diagrams representing the cell activity levels and outliers

identify outliers (i.e., events) as unusual values with respect to that statistics.

For our analysis, every outlier i.e unusual high cell activity, will be a possible candidate for being an event. This concept is also illustrated in Figure 4.4 b) and d) showing cell activity values.

In the following sections, we present further analysis on how to discriminate and recognize if we have encountered an event or just a false positive.

In the literature, there are different examples of research papers using a statistical modeling approach similar to ours for detecting anomalies which in statistical analysis terms means detection of an outlier.
Similar variations of the statistical event detection method illustrated in this section have been applied also in other fields such as in Guttormsson et al., 1999 Paul S. Horn and Pesce, 2001

This type of approach works better for unimodal distributions, in which the mass of the probability distribution is concentrated in a compact space. Below are given some preliminary experiments to understand better if the distribution of activity values in a given cell is indeed unimodal.

Figure 4.5(a) shows the distribution of values of a given cell in the Milan dataset.
It is possible to see that the distribution is multi-modal, with a lower peak likely to be associated to night-time and an higher peak (with two bumps) likely to be associated to day-time.
In order to pursue unimodality, we tried to separate the dataset in different ways: *(i)* weekdays vs. weekends, *(ii)* on an hourly basis (i.e., a different

distribution for each hour), *(iii)* weekdays vs. weekends on an hourly basis (the combination of the former two).

The idea is that by reducing the variation in the samples being considered, the distribution will become more unimodal.
This kind of separation translates in having multiple models of the distribution of data in terms of its quantiles (e.g., one model for the weekdays and one model for the weekends). On a given day, we will identify outliers (i.e., events) as extreme values with respect to the associated (weekend or weekdays) statistics.

Figure 4.5(b) shows the distribution of values of a given cell in the Milan dataset for weekdays only (vice versa, we computed also the distribution for weekends only - not shown). Figure 4.5(c) shows the distribution of values of a given cell in the Milan dataset for a given hour (we computed also the distribution for all the other hours, with similar results). In both the cases, can be seen a bimodal distribution associated to nights vs days for Figure 4.5(b) and associated to weekdays vs. weekends for Figure 4.5(c). Finally, Figure 4.5(d) shows the distribution of values of a given cell in the Milan dataset for a given hour during weekdays only. This distribution is much more unimodal meaning that almost all the variation has been explained away.

Following this setting, for each cell, we computed 2 (weekdays vs. weekends) * 24 (hours) statistics to be used as baselines to detect events.
There is a trade-off between number of different statistics to be computed and the amount of data available for each statistics.
In this case, there are 45 samples for each weekday statistics and 20 samples for each weekends statistics.

Each data distribution is represented by using a boxplot that describes data summary attributes such as min/max value and Q25, Q50 (i.e., median), Q75, which are respectively lower, median and upper quartiles.
The difference Q75-Q25 is called interquartile range IQR and it gives a measure of how concentrated or how sparse the dataset is.
The Figure 4.4 illustrates better this concept. Thus by using the boxplot rule as a statistical technique for detecting outliers it is possible to detect anomalies in univariate data.

On the basis of computed quantiles and a constant k (serving as tunable parameter) are experimented different methods to compute the *normality* interval, i.e., the interval outside which a value is considered an outlier (and thus possibly representing a noteworthy event).
By representing such interval in terms of its lower bound (LB) and upper bound (UB) there are three possible approaches to define the interval [LB,UB]:

- Median method: [LB,UB] = [Q50 - k*Q50, Q50 + k*Q50]
- IQR method: [LB,UB] = [Q25 - k*IQR, Q75 + k*IQR]
- Q75 method[1]: [LB,UB] = [Q25 - k*Q25, Q75 + k*Q75]

[1]We use the name Q75 for this method as we are mainly interested in comparison with the upper bound to detect overcrowded events

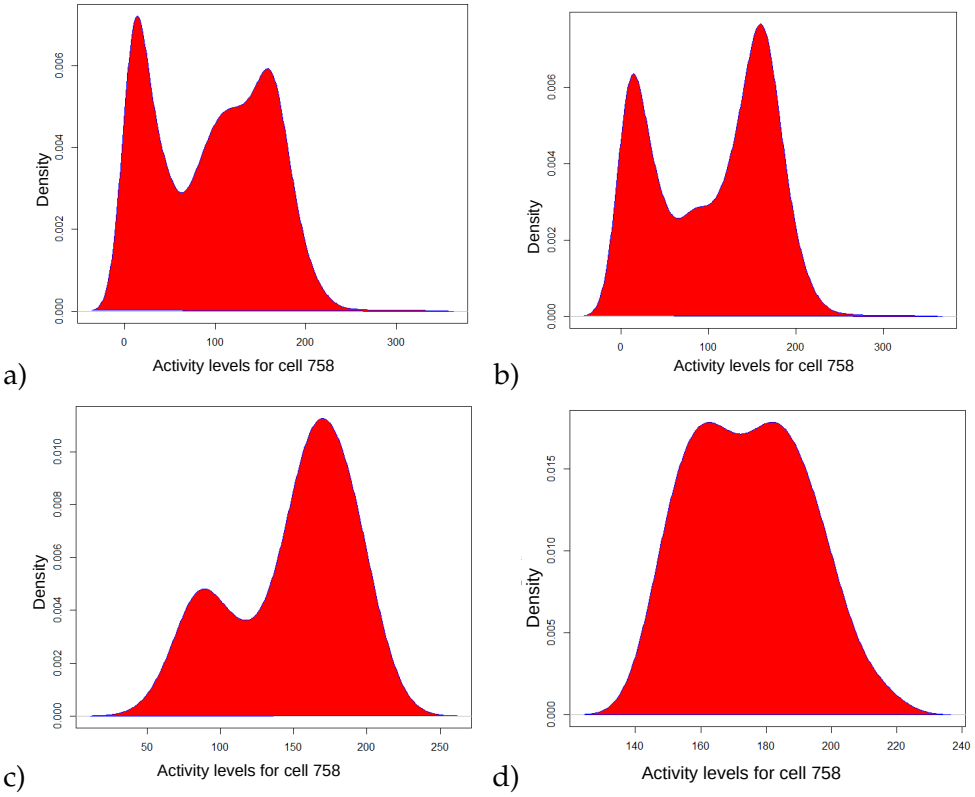


FIGURE 4.5: **(a)** Data with no aggregation **(b)** Data aggregated by week days **(c)** Data aggregated by hour **(d) Data aggregated in hourly basis only by week days**

We compute the normality interval for each cell in our dataset. Then, for each cell, we scan the activity time series and, at each time (e.g., Tuesday, 19 Nov 2013, 19:00), we compare the value with the associated normality interval (e.g. normality interval for weekdays at 19:00). On this basis, we identify values outside those intervals as candidate events as the activity in that cell at that time was unusual.

The main limitation of the approach described up to now is that it can consider at the same level both large and small events. For example, considering the median method (but the same applies to all the methods), in a cell at the periphery of the city where activity is almost always zero, even a small event attracting very few people can result in an activity detected as an outlier (e.g., a datapoint many times above the median – which is almost 0). The same deviation from the median in the city center would represent an event attracting thousands of people. Accordingly, an event detection system based on the above algorithm would report the two kind of events in the same way.

To better discriminate among events of different size, we introduce another parameter t or threshold that varies between zero and maximum cell activity level in the period of observation in the whole city. We use this parameter to ignore all the detected outliers whose absolute activity value is below t.

As for some kind of activities (e.s., outgoing calls) the activity level well

correlates with the number of people in the area Douglass et al., 2015, then t can be associated to the absolute number of people that must be at the event to be considered as meaningful. So, for example setting $t = 1000$, we do not consider as events those outliers at the periphery attracting less than 1000 people.

In this approach t becomes an application-dependent parameter and we use it for discriminating between low and high magnitude events. Should the user be interested in major events only, (s)he will set t to large values. Should the user be interested also in minor anomalies, (s)he will set t to small values.

Following this approach, the method can dynamically reports as meaningful events those outliers surpassing the t threshold. In the experiments reported in Section **??** we analyze the performance of this approach in terms of precision and recall for different parameters' configurations.

4.4.2 Evaluating the results of event detection algorithm in terms of precision and recall

In this section are present results of the event detection algorithm using the above methods to the data released in the context of the Telecom Big Data Challenge 2014. First are described collected ground-truth data used to validate the results of event detection. Then are present the results of the event detection algorithm applied to the individual datasets separately. Then we present results obtained by merging different datasets together to improve events' recognition. Finally, we show some results of the approach to describe events from Twitter data.

4.4.3 Groundtruth dataset

The ground-truth events data from local news and selected relatively major events such as football matches, local feasts, fairs, protests and other events that took place in the area and that resulted in consistent gathering of people. Overall there are 29 events, 17 in Milan e and 12 in Trento. For each event in the ground-truth data is defined the location into the grid and its start and end time.
There are not include global events (happening in all the city) like Christmas or New Year Eve in our ground-truth. Even if activity spikes in many cells during those days, it is difficult to precisely quantify precision and recall in those circumstances. Accordingly, those events would distort the results of the method.

4.4.4 Precision and Recall of the Event Detection Method

- **true positives (tp)** detected events having a corresponding groundtruth event in the same cell and at overlapping time (i.e., the detected outlier has to be within the groundtruth event's beginning and end)

- **false positives (fp)** detected events not having a corresponding groundtruth event in the same cell *or* in neighbor cells and at overlapping time. Initial experiments revealed a large number of false positives, as a

groundtruth event in a cell, typically generates outliers also in neighbor cells. Therefore they do not count as false positives events next to a ground-truth event

- **false negatives (fn)** groundtruth events not associated to detected ones.

We then computed:

- precision = tp / (tp + fp)
- recall = tp / (tp + fn)

By running a number of experiments to identify the best parameters of this approach a series of graphics are generated as in Figure 4.6 where are shown precision-recall graphs, with the best parameters configuration, obtained by varying the threshold t used to detect events (i.e., each point of the curve corresponds to a different t). When t is small, we typically have high recall, but low precision – as almost all the outliers are considered events. Vice versa when t is large, only outliers associated to a very large number of people are considered events thus typically yielding high precision, but low recall.

Specifically, Figure 4.6(a)-left shows results for SMS-Call data for the Milan area. Figure 4.6(a)-right shows results for SMS-Call data for the Trento area. Figure 4.6(b)-left shows results for internet-access data for the Milan area. Figure 4.6(b)-right shows results for internet-access data for the Trento area. Figure 4.6(c)-left shows results for social data for the Milan area. Figure 4.6(c)-right shows results for social data for the Trento area.

It is possible to see that SMS-Call data is by far the most useful source of data to detect events. It is also possible to see that results from Milan dataset are better than Trento ones. This is because groundtruth events in Milan tend to be associated to larger populations thus making the recognition task inherently easier.
More results on this event detection method can be found in the Appendix.

4.5 Improving results of event detection method using data fusion

In figure 4.7 are presented the results obtained by combining multiple data together with the approach described in Section **??**. In general, it is possible to distinguish an improvement for the recall - precision performance when joining the results of the two datasets.
In particular it is possible to notice how even a small contribution from the social dataset can result in an improvement when join the results from the two datasets. For example a given recall value in the fused knowledge dataset and in social dataset correspons to a much better value of precision and with respect to the social dataset.

Viceversa a given value of precision in the fused knowledge dataset corresponds to much better value of recall than in the social dataset alone. The same can be said when comparing the fused knowledge dataset with the

SMS-Call dataset.

Even though in the long run the precision of the joined data is limited by the best precision obtained with the SMS-call dataset, we are confident that by adding more information to the process of joining, it is possible to improve even more. The same improvement can be seen also when we joint the results of the internet dataset with those of the social one.

4.5.1 Event description by data fusion

For describing the events we use a simple though very efficient method that is a word cloud graphics. Once identified the events from SMS-call data and social dataset that have the same location and time we use the text of the tweets in social dataset to describe events detected in the SMS-call dataset. We created the graph using the wordcloud R package (*cran.r-project.org/web/packages/wordcloud*).

This subsection brings two examples of event description results in terms of word clouds. One is about a football match and one about an artisans fair. Thought the two events have different time duration -the football match lasts only a couple of hours while the fair several days- and participants number it is perfectly possible to spot the words characterising the events in the word cloud graphics. The results of this analysis are shown in Figure 4.8.

4.6 More experiments : analysis of parameters

In this section is shown a summary of the several experiments and results obtained to optimize the parameters involved in the event detection approach.
As described in the previous section our approach is based on four key elements:

- The granularity of the grid used to tessellate the environment.
- The approach used to compute the normality interval: IQR, Median or Q75
- The value k setting the length of the normality interval
- the threshold t used to filter out minor anomalies

In the following we report experiments testing precision and recall of our approach for different data and parameters configurations. Specifically:

- Figure 4.9 deals with experiments with SMS-Call data
- Figure 4.10 deals with experiments with Internet data
- Figure 4.11 deals with experiments with social network data

In each figure are reported experiments both from Milan and Trento provinces. In each figure, subfigures *a,b,c* illustrate results for IQR, Median and Q75 approaches respectively and shows results of precision and recall

for varying values of k in the interval 0.5 and 3.
Each plot is obtained by varying the t value. When t is small, almost all the outliers are considered events thus typically yielding high recall, but low precision.

Vice versa when t is large, only outliers associated to a very large number of people (i.e. high levels of activity associated to the given cell) are considered events thus typically yielding high precision, but low recall.
By analyzing these results it is possible to identify the best approach (among IQR, Median and Q75) and the best value of k to identify meaningful events. We consider instead the t value as a free parameter to be set according to the cost of false-positive and false-negative that is application dependent.

Subfigure *d* is obtained by selecting the best combination of the previous parameters and changing the grid resolution in areas of 4, 9, 16 and so on and so forth cells.
Results indicate the following best results:

- **SMS-Call data - Milan:** Median, k=3, cells of 690m of side
- **SMS-Call data - Trento:** Median, k=0.5, cells of 9000m of side
- **Internet Data - Milan:** Median, k=1.5, cells of 2070m of side
- **Internet Data - Trento:** Median, k=1.5, cells of 9000m of side
- **Social Data - Milan:** Median k= 3.0 cells of 2070m of side
- **Social Data - Trento:** Median k = any, cells of 9000m of side

An eye catching result, is the big difference between cell size for Milan and Trento.
In Trento case the social dataset is very sparse and the results of precision and recall tend to be the same with every method.
The cell size that gives better results is 9000m of side and it is due to the much different population density in Milan and in the province of Trento which differs by the same order of magnitude.

4.7 Conclusions

The purpose of this chapter is two fold and in particular the two main points to underline are :

- to show a simple and effective event detection method for aggregated mobile activity data and social network data and the description of events by using a semantic representation such as a word cloud

- most importantly to illustrate how knowledge extracted by two different datasets can be integrated and fused to improve sensibly the event detection results.

  By measuring the results of event detection in terms of precision and recall is shown that this method of event detection performs much better that other methods used in previous works such as Neumann et al., 2013 and Walther and Kaisser, 2013.

- Another important point of this work is the description of the events detected by the extraction of keywords describing the event from social networks. This is possible once that the events detected from the first dataset are fused with events detected in the second dataset. This fusion corresponds to a fusion of knowledge that is extracted in a statistical way.

- We compare the results of this work with those that appear in studies that implement systems for event detection, using different algorithms but similar datasets as ours i.e. human activity datasets.
  More specifically this study is compared with works as the ones in Neumann et al., 2013 and in Walther and Kaisser, 2013.

  In particular the work in Neumann et al., 2013 uses a LPE (Localized p-value estimator) algorithm. They evaluate its performance with a bicycling activity dataset but distinguishing only between two categories of event days: rainy days and holidays. Results obtained are: recall 0.38, precision 1.0, f-score 0.57.

  The authors of the work in Walther and Kaisser, 2013 use a clustering approach based on geographic proximity, text common theme and other semantic features on a dataset of tweets. The authors call the algorithm ClusterScorer. Results obtained with large-scale citywide events are: recall 0.85, precision 0.85, f-score 0.85.

  We bring here a comparison of the results in terms of precision, recall and F1-score as in the following table.

Algorithm	**Recall**	**Precision**	**F1-score**
CS Walther and Kaisser, 2013	0.856	0.858	0.85
LPE Neumann et al., 2013	0.38	1.0	0.57
M (our method)	0.33	1.0	0.598

Although a direct comparison is impossible (as the datasets are different), by reviewing these works it is possible to infer that: our approach is geared toward greater recall and lower precision than LPE. Differently than ClusterScorer, it does not consider the textual content of Tweet and do not require an explicit training phase (for which can be hard to identify labeled training samples).

In our future work we will try to compare these algorithms on common datasets and possibly to combine them in an improved and integrated solution.Also in future work we will investigate event detection further, by using other methods (e.g., Ihler, Hutchins, and Smyth, 2006), combining even more data sources, and conduction more experiment in other geographic areas.

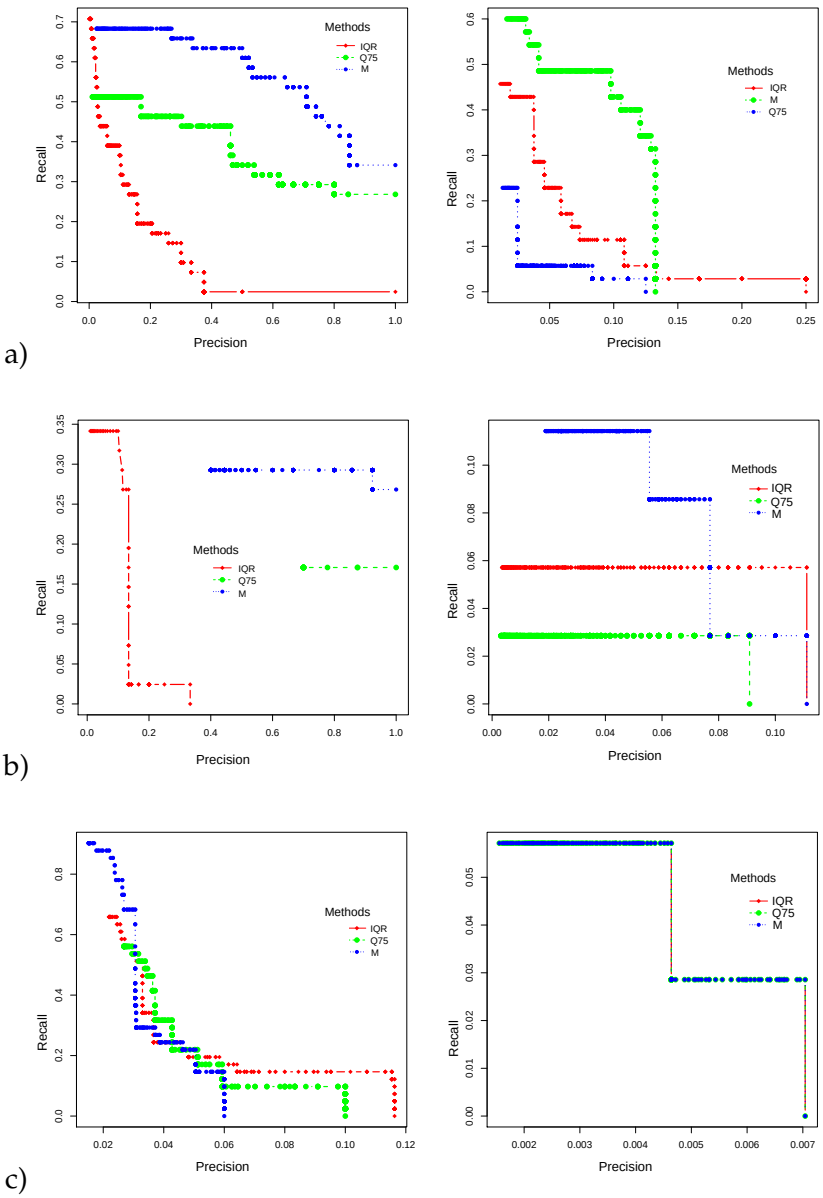
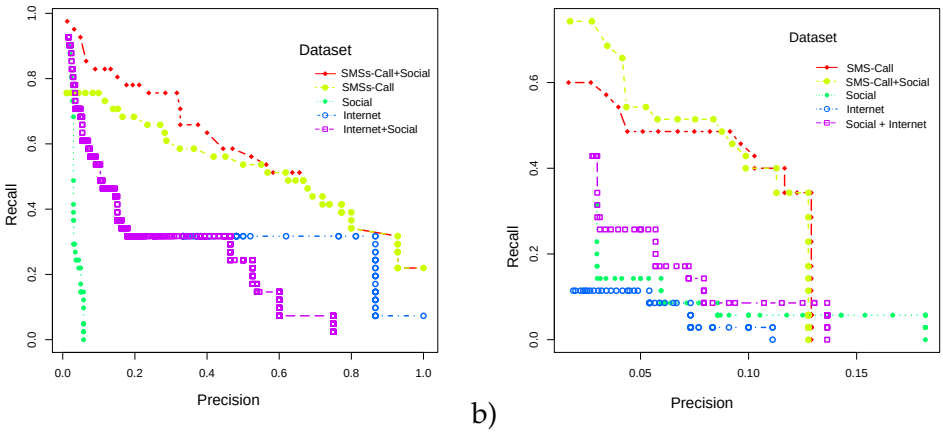


FIGURE 4.6: **(a)** SMS-Call Data. **(b)** Internet Data. **(c)** Social Data. **(left)** Milano. **(right)** Trento



a) b)

FIGURE 4.7: Results in aggregating data from multiple source **(a)** Milan area **(b)** Trento area

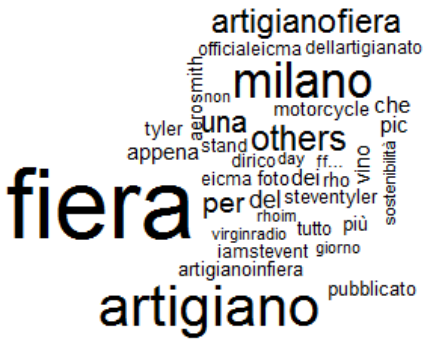
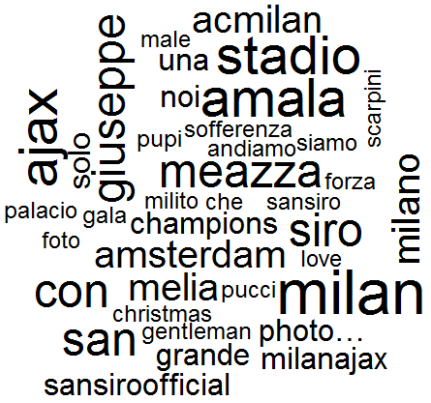
Artisan fair event held in Milano Fiere from 30/11 to 8/12/2013	Football match event Milan AC vs AFC Ajax held in San Siro stadium on 11/12/2013
	

FIGURE 4.8: Results of event description. **(left)** An artisans' fair. **(right)** A football match. From the world cloud it is possible to spot the words characterising the events.

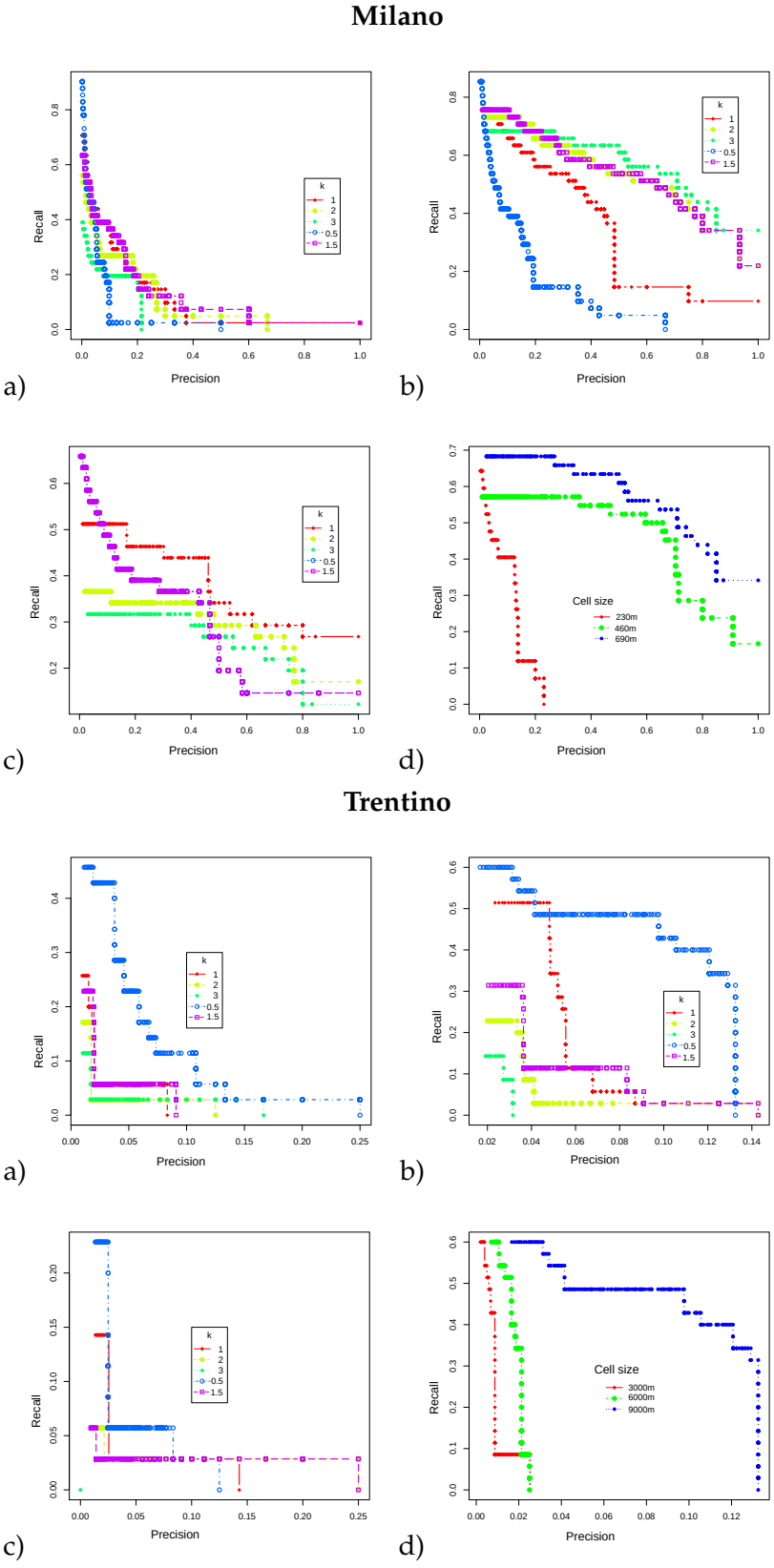


FIGURE 4.9: **SMS-Call Data - (a)** IQR method for diff. k values **(b)** Median method for diff. k values **(c)** Q75 methods for diff. k values **(d)** best method for diff. cell sizes

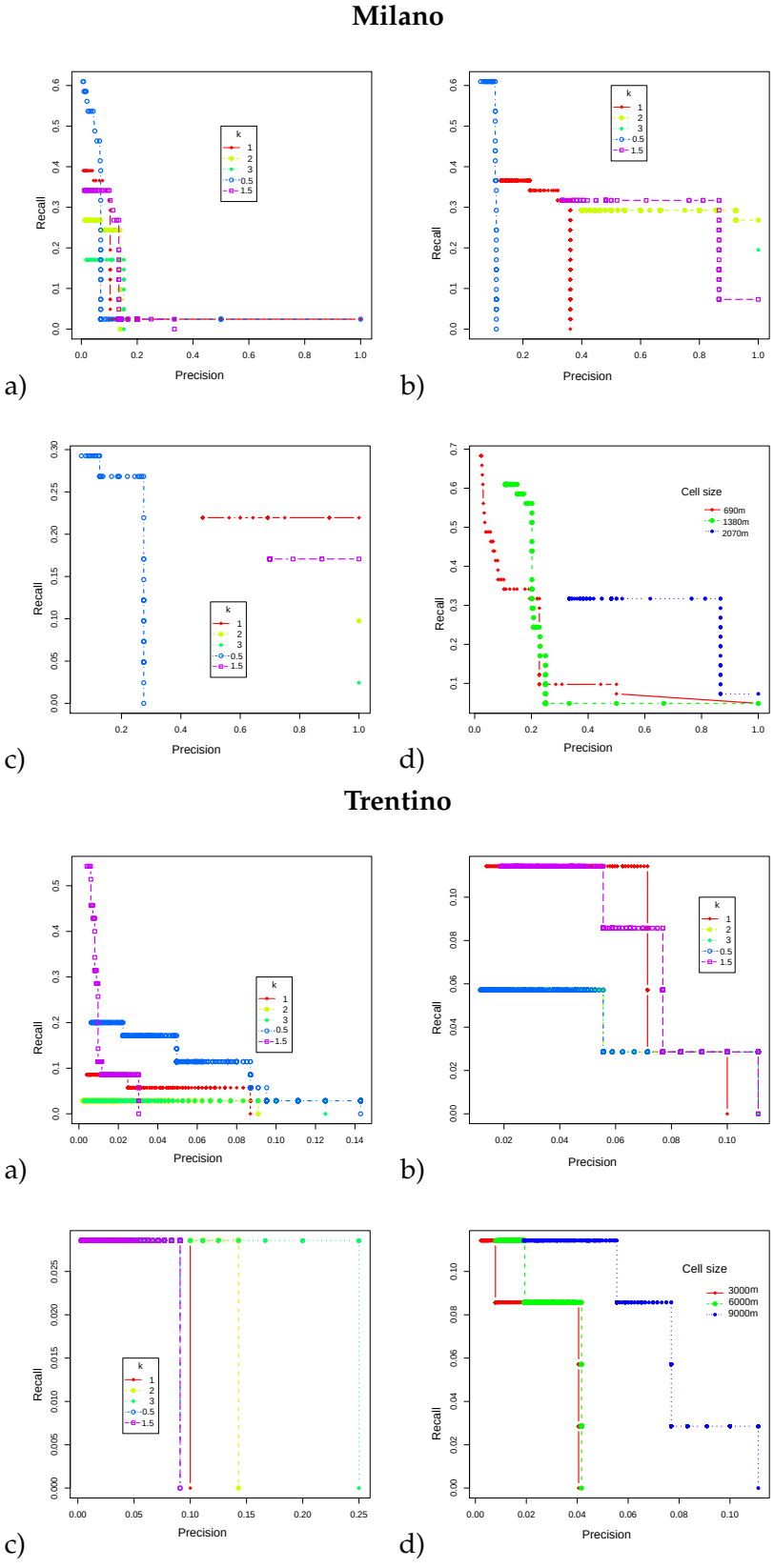


FIGURE 4.10: **Internet Data - (a)** IQR method for diff. k values **(b)** Median method for diff. k values **(c)** Q75 methods for diff. k values **(d)** best method for diff. cell sizes

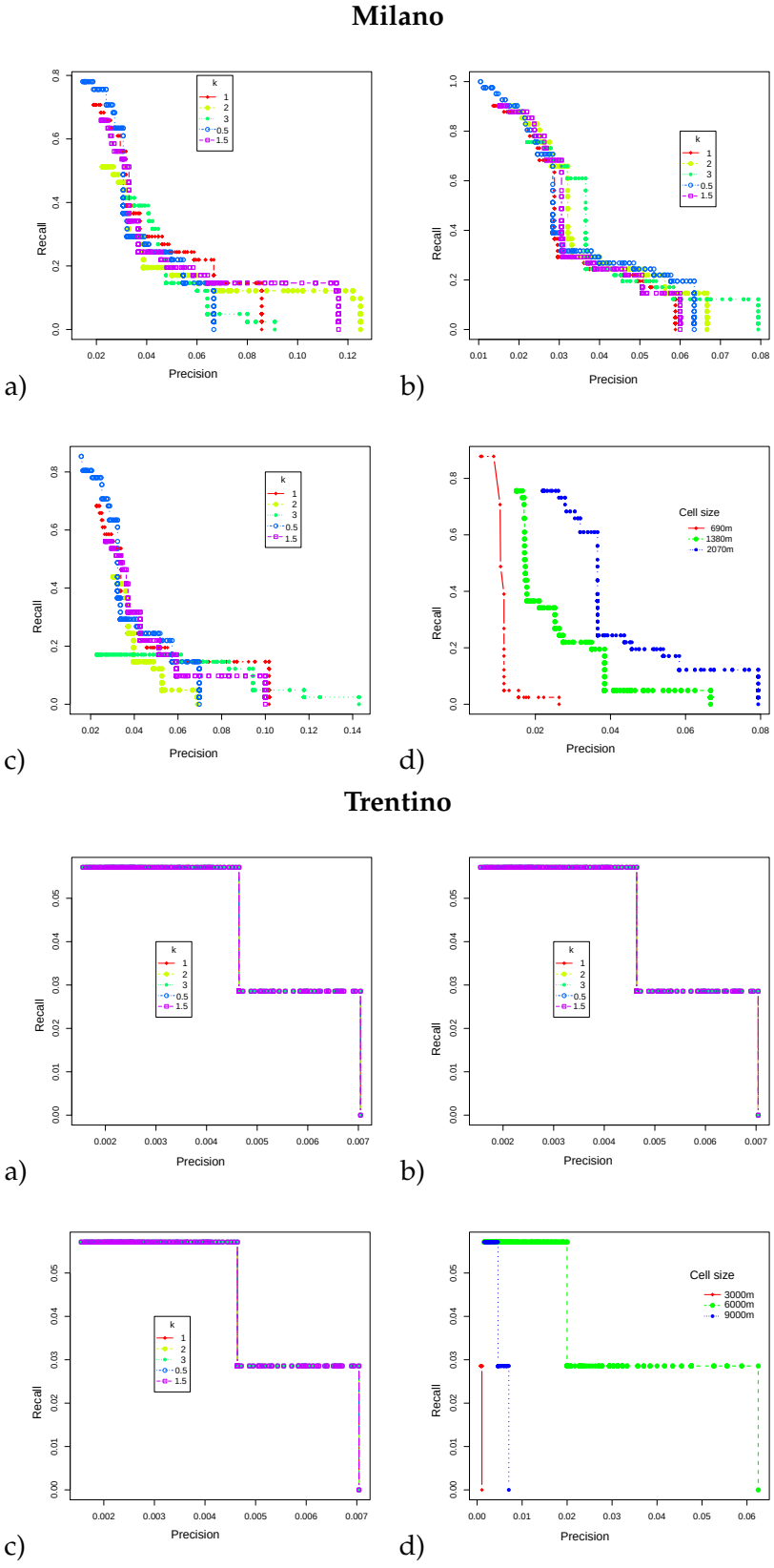


FIGURE 4.11: **Social Data - (a)** IQR method for diff. k values **(b)** Median method for diff. k values **(c)** Q75 methods for diff. k values **(d)** best method for diff. cell sizes

Chapter 5

A multi-perspective view of the city life obtained by Open Data

5.1 Introduction

This chapter brings a case study in which Open Data is analyzed to understand better the city dynamics in terms of demographic composition, social and economic activity. By linking socio - economic and demographic characteristics from census data to geo-referenced data from Twitter, Flickr and Instagram, it aims to evaluate the utility of this type of data for research and presents some results on this purpose.

Following one of the main topics of this thesis -that of cross domain data fusion- this chapter, presents the analysis of different data sets that can be obtained online from data services and are free to use. As each data set, provides a unique view, then through their examination and analysis and by combining the final results, it is possible to have a better description of the dynamics of city's socio-economic live.

5.2 Related work

On the one hand social networking services such as Twitter or Flickr that exist from many years now, provide an opportunity to study populations and human behavior in a large scale with their free available data. On the other hand national statistics institutes make available census data following the open data philosophy.

This data availability has inspired researchers of different fields and the following paragraphs overview some of the most relevant works in which census and social network data are used to gain insights on a country's socio-economic and demographic dynamics.

The literature contains a long list of works that use geo-referenced data, from location based social networks such as Foursquare, Flickr and others. Those works are about the creation of automatic methods and algorithms to estimate land use in urban areas, mining tourist routes and behavior, finding patterns of mobility between areas of a city or even quantifying the impact of scenic environments on health .

An array of such works can be found in Noulas et al., 2012 Barchiesi et al., 2015 Illushka, Tobias, and Susannah, 2015 Chareyron, Da-Rugna, and Branchet, 2013 Vu et al., 2015.

However those works are based only on one type of data -that of geo-referenced social network data- and lack of ground-truth information that would validate those works at a higher level of analysis.

More in line with the work done in this chapter, are the following studies which use census data combined with location based services data, as a way to have more reliable results on the one hand, and to gain better insights on the other.

In the Steiger et al., 2015 census data are used to explore the semantic association between geo-referenced tweets and their respective spatio-temporal whereabouts.
By applying a semantic topic model classification and spatial autocorrelation analysis the authors detect tweets indicating specific human social activities.

Furthermore they correlate the observed tweet patterns with official census data from London city. The results of semantic and spatio-temporal clustered tweets show an overall strong positive correlation in comparison with workplace population census data.

In a similar way, the paper in Mitchell et al., 2013 , in which the authors analyze Twitter's real-time expressions (comprising about 80 million words) of individuals made across the United States. Then they correlate those expressions with geographic,demographic, and health annually surveyed characteristics on 50 states and about to 400 urban populations.

By generating taxonomies of states and cities based on their similarities in word use, they estimate the happiness levels of different American states and urban areas.

They also correlate highly-resolved demographic characteristics with happiness levels, connect word choice and message length with urban characteristics such as education levels and obesity rates. The results show how social media, may potentially be used to estimate real-time levels and changes in population-scale measures.

5.3 Open Data

Open data - as defined on Wikipedia - is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.
Following this almost 8 years old philosophy the Open Data has created an ecosystem composed by many individuals which mainly can be classified in two broad categories.

- Governments and government controlled institutions specially the ones that study demographics and related statistics.
  An example is the Italian Institute of Statistics while in international context the Wold Bank data.

- Big digital companies that have as their core business the data. Twitter is one of the main companies that through an open architecture makes available part of its user's data to developers and researchers.

Even there is much more work to do from the governments and institutions until the Open Data becomes truly useful for scientific studies, there are examples of cities that provide data sets that can be of good utility in different fields of study, specially if combined by using data fusion methods.

One of those is NYC Open Data [1] which make available to everyone urban big data of any type, from transportation and social services to environment, public safety and city government.

5.3.1 An overview of the data sets

In this subsection are described the different types of data sets used for the study.
In particular the Open Data are obtained by government institutions, and private companies that provide open data through their REST web services. The data used in this study are three different types.

1. Census data from the last 2011 census of Istat, the Italian National Statistics Institute. This data set provides aggregated demographic data and from a spatial point of view has an hierarchical structure.
   For this study are considered three levels of granularity: administrative region, municipality and section. Each region contains municipalities and each municipality contains sections.

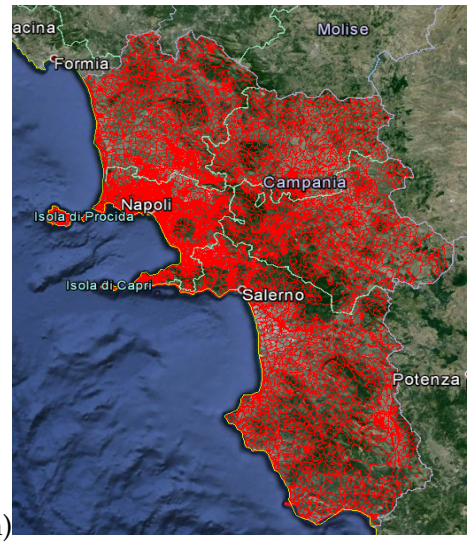
   The aggregated demographic data provides information about number of inhabitants, distribution of age in a certain range, sex, marital status,literacy and more for each section.

   A section is a specific area of the city or territory delimited by a geographic polygon. For each section there are provided many useful aggregated data such as number of inhabitants, number of foreign people, age distribution of the inhabitants of the section itself, income distribution, sex distribution, and many other demographic features.
   The table below shows how this dataset looks, while in fig 5.1 a) is given the administrative Campania region.The dataset containing the information can be downloaded from the ISTAT website [2]

Region	Municipality	section	Inhabitants	Age Range
Piemonte	Agliè	10050000001	352	8
Lombardia	Mozzo	161430000014	16	1
ER	Fidenza	340140000046	161	6

2. Geo-referenced data form social networks such as Instagram, Flickr and Twitter. These are user data from the users of the above social networking applications. The temporal dimension of the data set spans

[1] https://data.cityofnewyork.us
[2] http://www.istat.it/it/

a) 

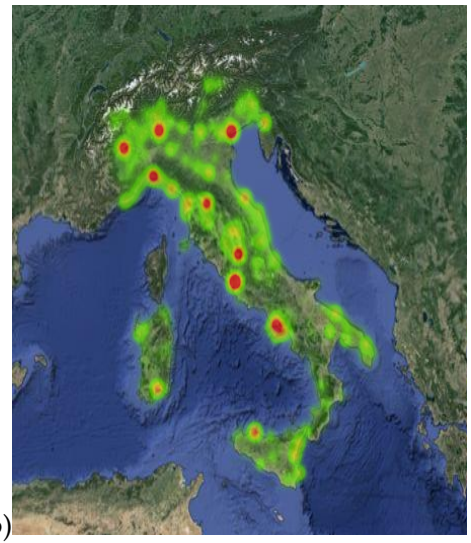
b) 

FIGURE 5.1: **(a)** How the administrative region of Campania looks divided in sections **(a)** Heat map of social data points

from the year 2006 which are mainly Flickr geo-tagged pictures until 2015 which are Twitter and Instagram data. Even this type of data is not proper "open data" due to restrictions in its use, however, its availability online for research purpose makes it very similar to open data.
More specifically each time a user makes a picture or a tweet a geo-referenced record is generated. These records are available to collect through the social network REST API.
In the table below are given some examples of those records. In Fig 5.1 b) is shown a heat map graphics of the social network data set.

Username	Description	Coordinates	Timestamp
AlketC	Great day!	40.5867733 15.6771344	1428613789
MarcoM	Hi there!	45.4858392 9.1681616	1427721154
AndreaS	Im in Liguria!	44.5833504 9.0231382	1426613153

3. Nominatim API data. This service allows to obtain geo-referenced data for example POI (Points of Interest) such as restaurants, bars,pubs, universities, libraries etc..
The table shows POI data records for different types of POI. For more information about Nominatim API [3]

POI id	Type	Coordinates	Importance
25963015	Bank	40.5867733 15.6771344	0.001.
1300581	Pub	45.4858392 9.1681616	0.001
1436033	Restaurant	44.5833504 9.0231382	0.001

[3] http://nominatim.openstreetmap.org

5.3.2 Data Set Summary Statistics

In order to understand better each of the data sets illustrated until here, are performed a set of descriptive statistics graphics. This statistics will help to understand better the type of distribution of each variable.

For computing the statistics there is performed a combination between the sections aggregated in municipalities as geographic polygons and the geo-referenced pictures / tweets that are made inside that municipality.

The same combination is done also for the POIs so that each time a POI is inside the geographic polygon delimiting a municipality, the POI is associated to that municipality.
In the following paragraphs are shown different summary statistics of the all the data set used for this study such as number of people and age range distribution per municipality, the distribution of pictures and POI per municipality etc..

Census dataset

In this subsection are illustrated the distribution of users for each municipality, and the distribution of each age range.

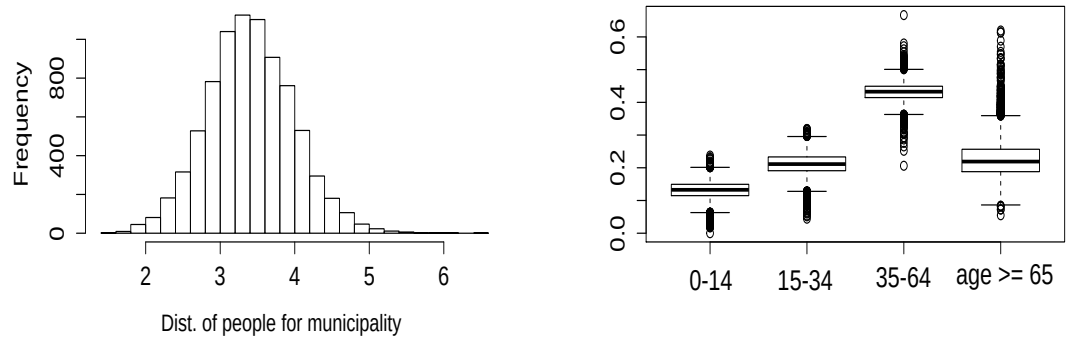
The original dataset provides the age distribution in each section thus municipality from 0 to 74 year old and 75 or older. For a clear representation the demographic data about the people living in each municipality is aggregated in ranges of 0-14 years old, 15-34 years old, 35-64 years old and 65 or older. For each age range in a municipality its percentage is calculated.
The following graphics in Fig 5.2 represents in a) the distribution of people per municipality as a histogram. In b) is represented the distribution of each age range expressed as percentage of the total population of the municipality by using a boxplot representation.

POI dataset

The POI dataset contains 69867 different POI allover Italy divided in 37 different categories. The graphics below in Fig 5.3 show the distribution of POI per municipality in a) and the twenty categories that have more entries in b).

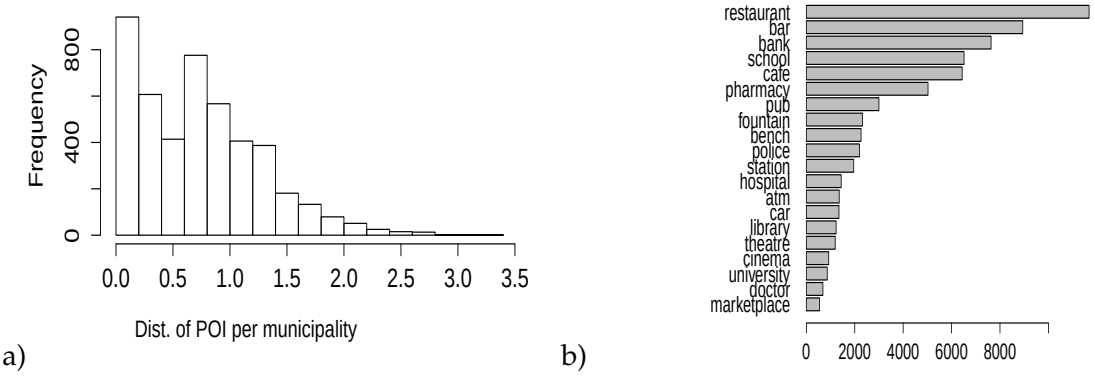
Social network dataset

The social data set contains 4.177.098 pictures and tweets made by 485.232 users all over Italy. As very often happens with these types of data it approximate an exponential distribution with a long tail.
This means that the majority of users have just a few pictures or tweets (from 1 to 5), while there is a low number of users that have a big number of pictures / tweets. The Fig 5.4 a) illustrates better this idea. In Fig 5.4 b) is given the distribution of the number of pictures and tweets per municipality.



a) b)

FIGURE 5.2: **(a)** Distribution of number of people for municipality **(b)** Age range expressed as percentage of the total population for municipality in a boxplot diagram representation.



a) b)

FIGURE 5.3: **(a)** Distribution of number of POI per municipality **(b)** The twenty most frequent categories of POI

5.4 Research questions and some answers

In this section we try to answer different questions that may be answered by linking census socio-demographic data with geo-tagged social network data and the POI.

As already mentioned in the introduction section, the data sets presented till here are expressions of urban (and rural) life in this case in Italy. They show not only the geographic distribution of different age ranges of people or buildings, POI and social activity but also some associations between them.
Those associations or correlations can tell something about urban dynamics and/or activities in rural areas as people move, go to work, attend events and interact with other people on different context. In this scenario some interesting questions can be :

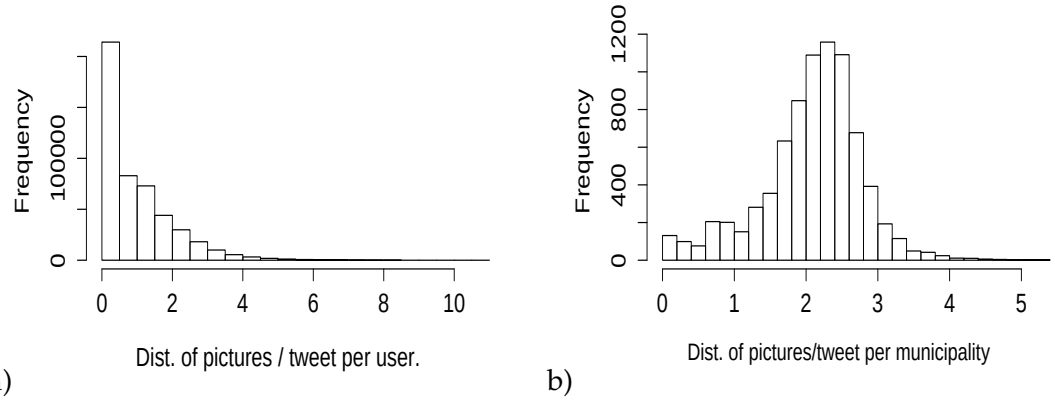


FIGURE 5.4: **(a)** Distribution of number of pictures and tweets per user **(b)** The distribution of the number of pictures and tweets per municipality

- How does the number of geo-referenced pictures / tweets in a municipality, relate with the total population and with the percentage of each age group living in that municipality?
- What can the text that describes geo-referenced pictures and tweets in a given municipality show about the demographic composition of that municipality in terms of age range ?
- Is there any correlation between the number of pictures / tweets in a certain municipality and the number and categories of POI that are present there?

For answering those question in the next section there is a correlation analysis showing different graphics obtained by the study. The analysis however, will try to answer the research questions in the most open way.

5.4.1 Correlation analysis

Before illustrating the results in this subsection, the author of this thesis points out how the correlations doesn't imply causation and aware of this mathematical fact the correlations showed below are only to be interpreted as such. The analysis however can explain possible relations between the variables object of study.

For answering the fist question about number of pictures / tweets and the number of people of a certain age group living in a municipality a correlation matrix is presented. First of all we correlate the total population of a municipality, expressed as a fraction of the country's total population, with the number of pictures in that municipality expressed as a fraction of the total number of pictures.

As of the table below the correlation between them is strong, c = 0.82. We show also the correlation of the number of pictures / tweets with different age groups where each age group is expressed as a percentage of the total population of the municipality. For example there are municipalities

that have up to 61 percent of the residents in an age range of 64 years old or older. Other municipalities have a strong presence of residents that are young, in the age range of 15 -34 years old.

The correlations coefficients of the number of pictures spans from 0.79 with the young group to 0.85 with the elderly group as shown from the table below.

	Pic-tw	0-14	15-34	35-64	65 or older
Pic-tw	1.00	0.79	0.79	0.82	0.85
0-14	0.79	1.00	1.00	1.00	0.91
15-34	0.79	1.00	1.00	0.99	0.90
35-64	0.82	1.00	0.99	1.00	0.90
65 or older	0.85	0.91	0.90	0.90	1.00

From this result, we can say that there is a difference between those two groups in terms of number of pictures and tweets even though it would have been expected an opposite result. Different hypothesis can be made in order to explain this fact.

The second question is on the number of pictures / tweets gathered in a municipality and how the text that describes the social geo-tagged content could describe also the demographic composition in terms of age groups of that municipality.

For answering this question, the municipalities that show a strong presence of a single age range are considered. We choose two main categories. The municipalities that have a strong presence of residents in an age range of 64 years old or older, and those that have a strong presence of young people, in the age range of 15 -34 years old. A total of 99 municipalities was considered, where 49 have a demographic composition of at least 45 percent in the age range 64 years old or older. The remaining 50 municipalities have a percentage of young people in the age range 15 - 34 years old of at least 29 percent.

Next the text in the social data set is analyzed for each municipality belonging to the two categories previously identified in order to extract key words and/or topics.

For the text analysis was used TAGME [4], a topic-annotator better described in Ferragina and Scaiella, 2010.

This text mining tool is able to identify meaningful sub-strings or spots in an unstructured text. Each spot has a weight or probability that measures the importance of that spot as entity of the text. It also can give a measure of importance of the topic in the text.

TAGME can annotate very short texts such as the text in tweets which is typically short (140 characters) thus it is the right tool to use for analyzing textual descriptions of social content such as tags and tweets.

By annotating the text with TAGME for each category of municipality (young and old ones) a list of topics was extracted.
In the following graphics in Fig 5.5 are shown the most relevant topics sorted by weight of importance i.e. the topics with the highest index of importance in the text.

[4] http://tagme.di.unipi.it

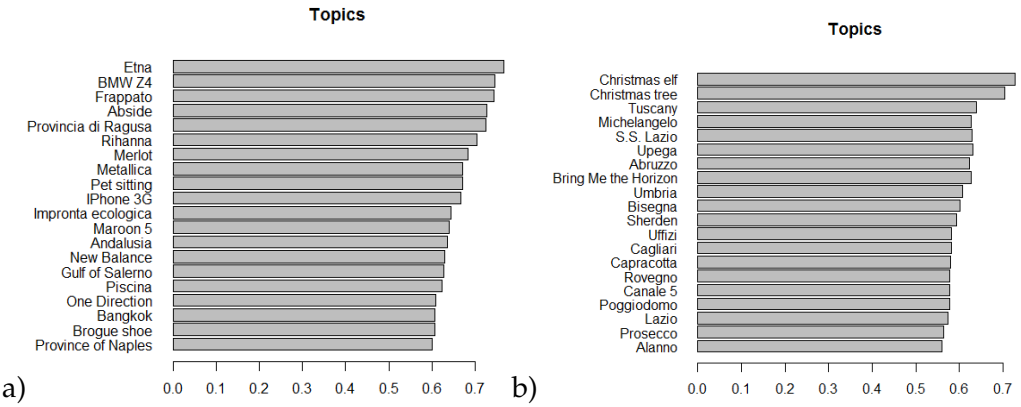


FIGURE 5.5: **(a)** 20 most important topics for the municipalities with mostly young people **(b)** 20 most important topics for the municipalities composed by mostly old people

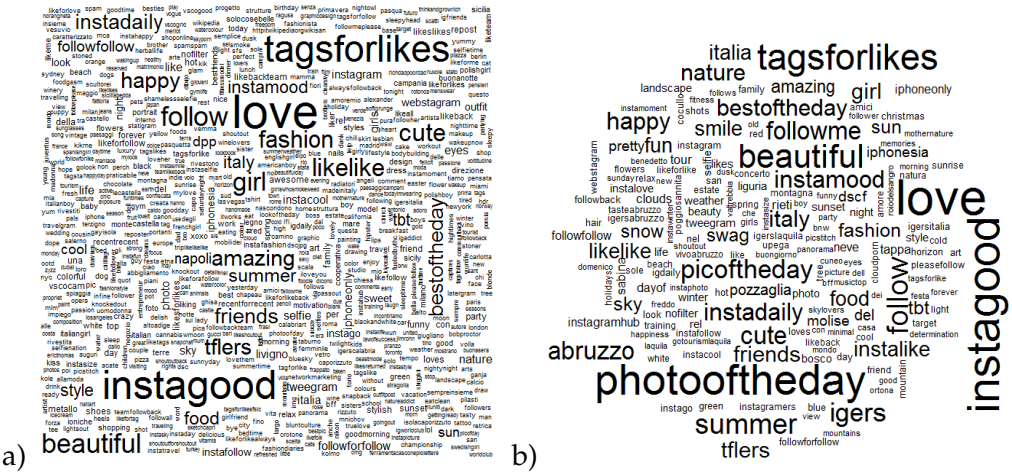


FIGURE 5.6: **(a)** Word cloud graphics for municipalities with mostly young people **(b)** word cloud graphics for municipalities composed by mostly elderly people

Again there is a difference between the types of topics discovered in the two categories of municipalities. However, the difference is quite thin and can be interpreted in a subjective way.

As a complementary text analysis a world cloud graphics was computed for both the categories of old and young people as in the graphics in Fig 5.6.

There are some key words that are common in both the categories but there is a difference that can be noted immediately. For the municipalities with a strong presence of young population, there are much more different words than for the ones that have a majority of elderly people.

Answering the third question: Is there any correlation between the number of pictures / tweets in a certain municipality and the number and categories of POI that are present there? How the POI relate with particular age groups of the population ?

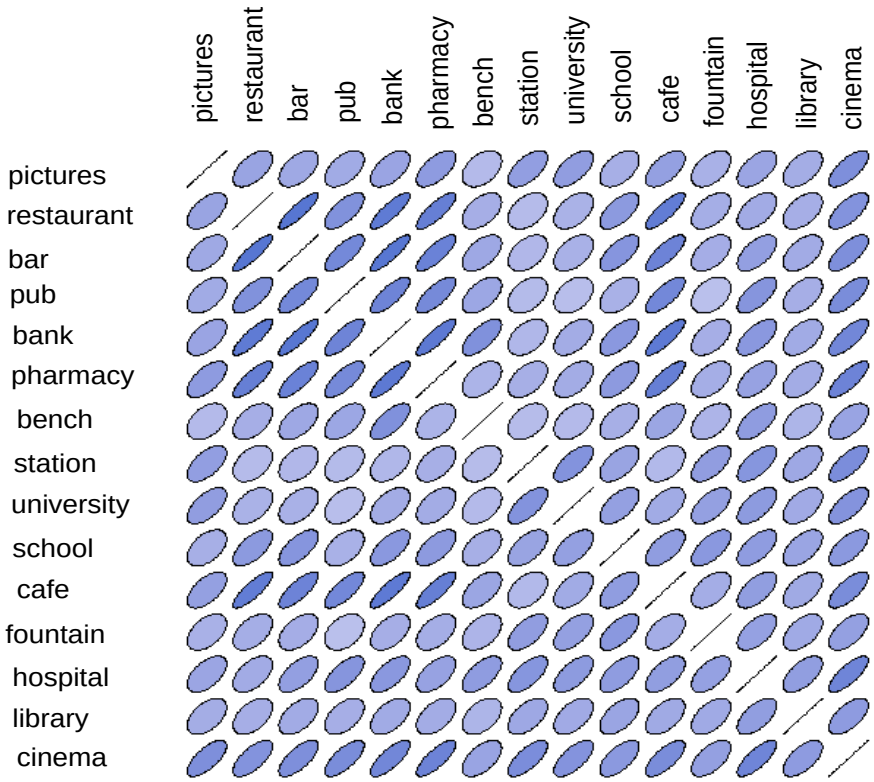


FIGURE 5.7: Correlation graphic of the number of pictures and the different categories of POI per municipality

By correlating the number of POI per municipality and that of the pictures per municipality we find that there is a strong correlation c = 0.74 as would have been expected. Next we present a correlation graphics as in Fig 5.7 ,trying to answer the question on the relation between the number of pictures and each category of POI by representing the correlation between the variables. The graphics represents the correlation in an graphical way that uses an ellipse as a set of points.

The graphics in Fig 5.7 shows that there is a strong correlation between the number of pictures in a municipality and the number of cinemas, universities and stations. There are also some interesting correlations between POIs them self. For example between bars and banks or restaurants and pharmacies. All the correlations showed above illustrate different links that exist between social, demographic and economic aspects in a urban environment. By analyzing better those links it is possible to discover many interesting dynamics as better explained in the conclusions section.

5.5 Conclusions

In this chapter was presented a study of different data sets that once combined can reveal much more information that taken separately. In particular, the study shows that there is a strong correlation between the number of people living in a certain area, and the number of pictures and tweets that where made in that area.

It also shows how certain categories of POI, are correlated to the areas where a certain category of people leaves.

The text describing the pictures and tweets is different from the areas where a certain age group is the majority compared to the total number of people living in the area. Given those correlations, the future work will continue in analyzing even further those links and enriching the analysis with more data from other social networks and POI.

The final purpose is that of building an algorithm that given the social data set pictures and tweets from a certain area can predict the POI type and / or the population age ranges living in that area.

Furthermore as each administrative region in Italy may present very different characteristics the analysis may be performed for each region separately. This could help in finding the structural differences between regions in general and between northern and southern administrative regions in particular.

Studying the data from a mobility prospective it can be interesting to analyze the paths of the social users and associate to each point of the path a set of characteristics in terms of demographic and POI composition. This kind of analysis would help to understand better the triggers of mobility in a large scale for urban and rural environments.

Chapter 6

Privacy and anonymity: computer science and legal aspects

6.1 Introduction

As already stated in the previous chapters the thesis focuses in two key points : data fusion in the first place and in second place the privacy issues that could emerge during the process of data fusion itself. This chapter explores in particular the potentially revelatory power of precise location data, then addresses the question of potential harm to individuals privacy specially in the case in which two or more data sets are analyzed and integrated for different purposes.

It presents the state of art in matter of privacy in three main areas an in particular in:

- estimating privacy risks form data release with main focus on geo-referenced data such as Location Based Services for example Foursquare and anonymized CDR data release in the context of data challenges such as Orange D4D (Data for Development) case. Other works covering areas such as privacy on medical data records will also be covered.

- privacy protection methods and systems proposed so far, in order to prevent privacy harm through two different methods : a) data anonymization and b) data suppression. For the anonymization will be over-viewed how it is done and how much it works with focus on location data fusion techniques which are the basis for de-anonymization attacks.
  It also argues the anonymity techniques and possible applications design that aim on preserving privacy by storing less data in the users device
  In the case of data suppression, questions on the remaining utility of the data set, will be reviewed.

- finally legal aspects on personal data privacy will be considered. In particular how different sounds the concept of anonymity in computer science perspective and legal perspective in European context.

6.2 Sensible information that can be inferred from personal location data

On the one hand the precision of cellular-location estimates means service providers are able to obtain location estimates with address-level precision. This means also a privacy problem, as it can reveal user behaviour, preferences, and beliefs as described below.

On the other hand geo-referenced data can enable the creation of a big number of applications known as LBS (Location Based Services). Those services may range from social network services such as Foursquare or Twitter to ride sharing applications such as Uber or Lyft and they've become essential to many users.
In the same time the users have shared a great number of check-ins and location/mobility data that are substantially available to everyone through the APIs of the above mentioned services.

A first set of works and surveys that considers the privacy harms that might result from the disclosure of personal data during the use of the above services is given in Perusco and Michael, 2007 Junglas, Johnson, and Spitzmüller, 2008-08-01T00:00:00 Minch, 2004 Krumm, 2008 Shokri et al., 2011 Zhou, 2011,

Considering the release of CDR data for example in a context of data challenges or on bilateral agreements between academic laboratories and mobile companies there are Shilton, 2009 Andrienko et al., 2013 while the work in Chen and Zhao, 2012 studies the privacy risks in cloud computing.

There are also a set of works dealing with other types of data such as medical and Genomics data as for example Heeney C, 2011 Malin, Karp, and Scheuermann, 2015 and the work in Cavoukian, Polonetsky, and Wolf, 2010 which presents a case of "SmartPrivacy for the Smart Grid: embedding privacy into the design of electricity conservation"

In the following paragraphs are described how powerful can be the location data and the different scenarios that may be a threat for the individuals privacy. Let's consider the following scenario where different kind of information can be inferred by correlating location data with publicly available information such as census data.

- Location of a person's home. What kind of city area do he lives in? What is his address? The study in Jahanbakhsh, King, and Shoja, 2012 in which the authors demonstrate the possibility of predicting people's hometowns by using their geotagged photos posted on Flickr website. Also the work in Zheng et al., 2014a which infers home location base on visual content.

  By discovering this information there are different privacy harms, for example : very often happens in cities there are different neighborhoods where people of different incomes leave.
  The so called "rich neighborhoods" and "poor" ones. By this information a person's socioeconomic status can be deduced.

- Locations of the stores a person frequent and from this information shopping patterns can be inferred. Those patterns may reflect also

preferences and in some cases one's beliefs, religious belief and belonging and from this information the individuals worldview.

- There are also other types of very sensitive data such as health records. These can be deduced by locations of doctors and hospitals the person visits and by the frequency and time permanence of a person in those places. On this issue there is the work in Rothstein, 2010 which asks the interesting question "Is De-identification Sufficient to Protect Health Privacy in Research?"
  Even the type of illness can be determined if the person visits a certain specialty clinics. Such information may be of great interest to insurance providers and the marketers of medical appliances and drugs, among others.

- The religious affiliation can also be inferred by analyzing the location of any building such as churches that the person frequents or not. From this can be deduced a person's worldview.

- Personal hobbies can also be deduced by analyzing frequently visited locations such as gyms or clubs. Entertainment venues also are part of the category. From such type of information the attacker can draw multiple conclusions and use this information in different ways. An interesting work that considers the problem of privacy harm in this context is Daries et al., 2014 .

- Furthermore by linking two or more locations on time and space, mobility paths may be inferred. By knowing those paths an advertiser could exploit this information for business profit purposes. The work in Hoh and Gruteser, 2005 illustrate a way to hamper the attacker from doing so.

There are many more examples that may be added to the list as location information can be used to determine a lot of information about an individual.
With time this information may be stored in databases accumulated and enriched by new data.

It is worth noticing that the risk of matching a location coming from a personal device and a name record in a database increases with the accumulation of data Wicker, 2012. Therefore the more can be done and kept within the personal handset, the greater the preservation of privacy .

An example of accumulated data is the picture in Fig 6.1 which shows personal location data from Google Maps timeline. Google tracks users location on time and shows it in a map as location history.

It is possible however, for the user change the location settings and disable geo-location function. In a) show a path in a local context inside a city and on an international context on the b) picture.

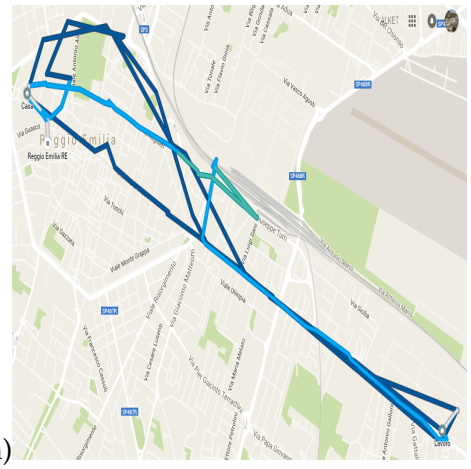
a)  b) 

FIGURE 6.1: **(a)** Location history from Google Maps timeline. **(b)** Location history in an international context

6.3 Preventing privacy harm : anonymization and suppression

A privacy threat occurs when an adversary can associate a user's identity with information that the user considers sensitive. Social network location data expose users to several privacy threats, due to the release of spatial-temporal information.

Even those LBS offer different options that allow users to use a certain granularity or hide their location information, research shows that many users are not aware of these privacy threats.

In Location Based Services in general, there are two categories of privacy threats for the user.

1. the release of sensitive location information. This is a category of threat that applies to cases in which the user's identity is known, but the location information pertaining to the user shouldn't be exposed to an adversary.

   Some Location Based Services allow the users release location information to other users at a coarser granularity than what's actually available.
   Thought it may be one way to avoid this type of threat, even if location data's sensitivity decreases at coarser granularity there are many studies that show that location data and mobility data prints are highly unique Montjoye et al., 2013. This can help the adversary distinguish between a user identifier among k other users.

2. re-identification through the user's location data . Re-identification through location refers to an adversary's ability to attack the anonymity of user's location information by reducing it or even re-identifying the user itself.

For example, by knowing that an anonymous CDR user was in a given place at a given time, an adversary could exclude several candidate individuals and possibly identify the user by using geo-referenced social network data as in Cecaj, Mamei, and Zambonelli, 2015.
There are also other datasets such as census data or open data of any kind which have a location dimension. As already extensively proved in the previous chapters by correlating the different datasets more information can be deduced.

Even the works in Cecaj, Mamei, and Zambonelli, 2015 and Zang and Bolot, 2011 shows that the anonymization of CDR data can always be attacked by using side information such as social network data there are different works that show methods such as k-anonymity Sweeney, 2002 could work.

For example the work in Kido, Yanagisawa, and Satoh, 2005 illustrates an anonymous communication technique using dummies (i.e. false locations) for location-based services while the paper in Tang et al., 2006 presents an anonymous and privacy-sensitive approach to collecting sensed data in location-based applications. In the same data domain the work in Gruteser and Grunwald, 2003 proposes a system for anonymous usage of LBS that uses a spatial and temporal cloaking.

Other works such as Hong et al., 2004 focus on preventing privacy threats by design. The authors in this paper propose privacy risk models for evaluating the risk of the service for users. By doing so it is possible to designing privacy-sensitive systems. In the same direction the work in Matthews and Harel, 2011 reviews a set of statistical methods in order to a) evaluate privacy risks from data disclosure and b) limit data disclosure before privacy harm could happen.

Regarding the identity of individuals the work in Claubeta, Kesdogan, and Kolsch, 2005 proposes a model for protection against re-identification and profiling. In the context of partecipatory sensing the paper in Christin et al., 2011 collects a set of different works on the privacy in mobile participatory sensing applications in which the privacy risk is even greater as the users are invited to participate during the process of sensing .

There are papers that use the idea of differential privacy in order to guaranty the privacy. For example the paper in Acs and Castelluccia, 2014 presents an anonymization scheme to release the spatio-temporal density of Paris.The scheme is differential private, and aims to provide privacy guarantee to each individual in the data set. The aim of the authors is to show that, even with large dimensional sensitive data, differential privacy can provide practical utility with meaningful privacy guarantee.

Using the same philosophy the work in Isaacman et al., 2013 presents "DP-WHERE" a system that preserves privacy on location data by adding controlled noise to achieve differential privacy. However the authors make no assumptions about the power or background knowledge of a potential adversary.

Always in the direction of privacy preserving methods the work in Aggarwal and Yu, 2008 surveys a set of privacy-preserving data mining models and algorithms.

6.3.1 Can part of the location data be suppressed and be useful at the same time

This subsection is inspired by Brickell and Shmatikov, 2008 and explores the possibilities of data anonymization and what it means for data utility. As described in previous chapters location data is powerful and useful for a wide range of applications. In the previous subsections there are also illustrated the different types of privacy threat due to personal location data release.

In this context the question is twofold : can location data be used and or released publicly so that other people can use it for useful applications on one hand, and on the other can we release it without making any harm to a user privacy ? Reviewing the actual research on this questions the answer seems to be also double fold.

There are studies that claim to anonymize data by clustering and perturbation as in Abul, Bonchi, and Nanni, 2010 or . These studies use privacy preservation techniques by making use of the k-anonymity approach.

Other works instead claim that location data cannot be anonymized in a definitive way and there is always a method to de-anonymized the users of a geo-referenced data set such as an anonymized CDR dataset for example or a movie data set Narayanan and Shmatikov, 2008.

This last view seems to be the most correct one. This prospective correctly concludes that the anonymization of location data using techniques such as generalization and / or suppression of identifiers will eventually destroy the utility of the data set itself.

In the best case a trade-off must be made between applied sanitization and remaining utility of the data set itself. The remaining utility of the data set can be measured as the accuracy of data-mining algorithms executed on the same sanitized records.

An interesting paper that provides a division on the control of data for an attempt of trade-off between the rewards of having available open data and the risk of privacy violations is described by Pentland in Pentland, 2014. In this paper the author brings a big data taxonomy within the spectrum of data control on three main data types which are :

1. data commons which are available to everybody
2. personal pr proprietary data typically controlled by individuals or companies
3. and government data intended as secret data.

However the final phrase for this subsection remains : "supposedly anonymous location traces can be de-anonymized through correlation with publicly available databases" Wicker, 2012.

6.4 Anonymity as a protection of privacy : differences between legal and computer science perspectives

Maybe the only possible way to release and use personal data publicly is that of anonymization. As already explained in the previous chapter personal data anonymization has advantages and disadvantages.

The advantages are those of assuring protection for the privacy of the individuals and the disadvantages come with the fact that anonymization techniques may destroy the utility of the data.

However as data of any type becomes more and more available also increases the will of releasing data for public use. This trend is confirmed by the open data philosophy which in five years of proliferation helped organizations being more and more open.
As for data handling and related privacy issues there are two opposing interests between different players of the IT sector. The ones whom interest is that of opening the data to the society by sharing and exploiting all the possible uses of it and the ones that claim the rights of the personal data protection conceded by the law.
Those are two different ways of thinking that represents often two different worlds. That of computer science on the one hand and that of Law.

An interesting study that focus on these differences is done in Mascetti et al., 2013.
The authors address the concept of anonymity from the two perspectives, that of computer science and that of Law regulations. They explaining the most important concepts, finding correspondences in the terminology and highlighting the points in common and main differences in the two areas.

They analyze the legal definitions of anonymous datum, as specified in the European regulations and describe the main models and techniques proposed in the Computer Science literature to target the problem of anonymity.

Also the study in Freiwald, 2011 analyzes the problem from a legal view. It from the one hand discusses an inappropriate court decision in the US regarding access to CDR data. On the other hand claims that a full understanding of the power of this technology is yet to be achieved by legal experts.

6.4.1 Formal models proposed in computer science for data anonymity

One of the most famous model for data anonymity is k-anonymity. This model deals with the problem of guaranteeing k-anonymity in spatial databases. The idea, is that a system provides k-anonymity protection if the information for each person contained in a data release cannot be distinguished from at least k-1 individuals also appearing in the data release as defined in Sweeney, 2002 and Sweeney, Abu, and Winn, 2013.

For example, techniques for k-anonymity in the case of mobility data are presented in Abul, Bonchi, and Nanni, 2010.

Similarly, in Parent et al., 2013 authors consider two approaches: one based on anonymizing each trajectory as a whole, and another which considers attackers who can link specific pairs of location and time to a person and re-identify him/her out using a certain number of points.

In this latter case the goal of anonymization is to hamper the attacker from associating a sequence of points to less than k individuals. This is done in Parent et al., 2013 by removing those points from a person's trajectories that would allow to single him/her out.

These approaches however, often tend to neglect the fact that most solutions for privacy preservation such as data suppression (removing critical data) or generalization (provide clusters of data rather than individual instances) highly compromise or destroy the utility of the data set itself.
Some approaches extending k-anonymity concepts are presented in Dwork, 2011 and in Zang and Bolot, 2011.

However simple k-anonymity may not be enough. In Bettini, Wang, and Jajodia, 2005 is given an example in which an attacker might look for patterns in the particular service requests on particular habits.
These patterns could be used to link otherwise un-linkable location reports. In this work is also introduced the notion of "historical k-anonymity" as a way to inject ambiguity into service requests to mitigate this type of attack.

For the privacy of location data to be maintained the most obvious way to is to replace the associated name with a pseudonym. This pseudonym may change on time frequently as proposed by Beresford and Stajano, 2003 and reduce the chance that an attacker could accumulate enough history on users to infer their habits or identity.
However they precise that if a user's preference data were stored on a server, an attacker could link all the pseudonyms attached to requests for a single user's data.

6.4.2 The notion of anonymity in the European legislation on personal data

Due to data explosion and need to anonymization for making them available also grows the importance that European legislation gives to the personal data protection.
The approach of the European legislation, is that of introducing the concept of personal data and then proceed to define anonymous data.

Standing to the definitions stated in the Directive 95/46/EC there are different actors of different roles involved in the process of personal data processing which are :

1. the controller is an entity (i.e., a natural or legal person, public authority, agency or any other body) that, alone or jointly with others, determines the purposes and means of personal data processing;

2. the processor is an entity that processes personal data on behalf of the controller;

3. the recipient is an entity to whom data are disclosed, whether a third party or not and finally, the data subject is the person to whom the personal data refer to.

As for privacy in matter of data regulations in EU a described in Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

6.5 Conclusions

The conclusions on this chapter are organized on three following points :

- the many case studies over viewed till here demonstrate that privacy risks for personal data will always exist.
- the anonymization techniques proposed by the literature have shown to have privacy breach while an attacker can use side data. It also shows how techniques of data suppression near-completely destroy the utility of the data.
  However privacy-aware design makes it possible to retain the full benefits of LBS on the one hand. On the other it can prevent the accumulation of address-level location traces for a given individual. This philosophy may reduce the potential for de-anonymization on personal data.
- finally it shows that there still exists a lack of comprehension and adequate collaboration between the two main responsible areas on privacy matters, that or computer science and that of legal representation.

Chapter 7

Conclusions

7.1 Introduction

In this thesis are introduced data fusion methods for geo-referenced data mining, by clearly distinguishing between traditional data fusion and cross domain data fusion methods. The thesis also presents different data fusion case studies in a smart city scenario.

The thesis puts the emphasis on two key points. On the one hand, the data fusion as a process which enables new applications for pervasive computing in smart cities scenarios. On the other, the privacy issues for individual's personal data that this process might rise.

7.1.1 Data fusion process and results

The data fusion methods presented in this thesis are inspired by the cross domain data fusion methods where high level information i.e. knowledge is integrated. In particular it presented an analysis of different anonymized mobility data sets.

Possibilities of re-identification and information fusion are evaluated through combinatorial and probabilistic analysis based on previous findings on patterns of human mobility. By analyzing CDR data sets released by mobile telecom operators and datasets of geo-tagged messages released by social network sites, two important findings are highlighted :

- Even a few geo-referenced data points are enough to uniquely pin point the majority of the CDR users and this is mostly due to the routine of everyday life.
- By using geo-referenced social network data as side information the study estimate the probability of re-identification of CDR users using a probabilistic model. It also considers the trade-off between the anonymization methods and reviews important works on the remaining utility of a data set after methods such as sanitization or suppression are applied to the dataset itself.

Another important example of data fusion is given by the case study in chapter 4 in which an aggregated CDR data set and a geo-reference social network one is analyzed. The purpose of this study is to build an event detection method in order to detect important events that result in massive gathering of people in urban environments.

The study evaluates a method of event detection based on a statistical method capable of detecting the events in a simple and robust way. The

results of the event detection system are then evaluated through measures of precision an recall of the system itself.

These measures show that by using a simple method of data fusion between the results of the two data sets - CDR and social- it is possible to sensibly improve the results of the event detection method. By comparing it with other systems of event detection we show that given the simplicity of our system it performs better and in a more robust way than the other two systems we compare with.

The benefits of the process of data fusion in this case are not only improving the accuracy of event detection itself. It also provides a simple and immediate method for event description. In fact, by using the text describing the geo-referenced social content it is possible to describe the event itself from a socio-cultural point of view.

Finally the study in chapter presents a case of data integration between census data, POI (geo-referenced point of interest) and geo-tagged pictures and tweets. By linking the three data sets it illustrates simple finding on the socio-economic and demographic structure of municipalities.

Regarding this chapter the future work will continue on applying advanced data fusion techniques in order to build systems of classification for the geo-tagged social data.

7.1.2 Conclusions on privacy implications

As of chapter 6 the data fusion process not only benefits in terms of applications created. It presents also a major downside that is about the privacy impact that the process might have. During the data fusion process, in fact, the data are linked with other data, enriching with information and giving to the initial data set much more meaning.

In a smart city context the data comes mainly from personal devices, thus personal data. If the personal data are fused with location information in particular, it is possible to infer many privacy harming facts about the individuals everyday routine. This might be information such as home and work location, hobby place etc..

The chapter 6 introduces different works in which privacy preserving methods for personal data are proposed.

However the anonymization techniques proposed by the literature have shown to have evident privacy breach while an attacker can use side data. A possible solution could come by privacy-aware design. This philosophy prevents the accumulation of address-level location traces for a given individual while retaining the benefits of a data fusion location based application. This philosophy may reduce the potential for de-anonymization on personal data.

Also location obfuscation, which is a technique that do not reveal the real location of the user, but still contains enough information to provide

a satisfactory service. The K-anonymity for trajectories gives also good results in privacy preserving policies.

As a finally conclusion on privacy the thesis point out that there still exists a lack of comprehension and adequate collaboration between the two main responsible areas on privacy preservation. That or computer science and that of legal representation.

7.1.3 Future work and open challenges

Future work will continue in two principal directions. In first place, data fusion techniques have a wide application in different fields and future work will focus in applying those techniques to data of different domain in order to compare the results of each technique / algorithm related to the data type or domain.

In the second place it will focus on testing the extent of the impact of the different data fusion methods on privacy on the one hand and the efficacy of privacy preserving methods such as K-anonymity on the other.

As for the open challenges on data fusion these are concentrated in the following points.

- One of the main challenges is the lack of common engineering standards for data fusion systems. It has been one of the main impediments to integration and data fusion. In general there is lack of standardized — or even well-documented — performance evaluation, system engineering methodologies, architecture paradigms, or multi-spectral models of targets and collection systems. It is an actual problem since even for current developments it is difficult to give a objective evaluation or comparison Steinberg, Bowman, and White, 1999 . The JDL model itself needs further improvement in this sense.

- As different methods of data fusion behave differently in different applications, it is not trivial to choose the best method for a specific task. Even there are given different guidelines on how to choose a data fusion method Zheng and Member, 2015 -based on the data volume and the goal of data fusion- a full mathematical classification lacks from the available literature.

- A more technical challenge is that of the acquisition and measurement of data from heterogeneous devices . As of Lahat, Adalý, and Jutten, 2014 "different instruments are sensitive to different physical phenomena, and consequently, report on different aspects of the underlying processes in a different way. A natural outcome is that the raw measurements may be represented by different types of physical units that do not commute. This situation is known as non-commensurability".

  There are many examples of such data fusion scenarios as in speech processing and more specifically speech recognition, speech activity detection, speaker extraction and separation. Another example is scene analysis in biometrics and monitoring, for safety and security applications.

  Finding a way for non-commensurable data sets, to inform each other and interact, is one of the great challenges encountered in a large number of cross domain data fusion scenarios.

- Again, in a data level, the different resolution of data can create one more difficulty. As different types of acquisition methods and observation setups provide data at different sampling, they also provide data in disparate resolutions. This will result in an additional challenge for the process of data fusion.

- Challenges during the data fusion design phase. This regards the different strategies to data fusion,in particular, at which level of abstraction, reduction and simplification the data should be fused. As mentioned above, the JDL model is more a model for facilitating the comprehension and communication and not a model that provides guidelines in this sense. The design phase also should evaluate which relations between data sets maximally exploit diversity, enhance interpretability and performance. Until now, the design of data fusion systems for this purpose, remains an open challenge. It remains also an open challenge, the lack of a framework that could provide a "structured data fusion" as in Lahat, Adalý, and Jutten, 2014.

Bibliography

In:
In:
Abraham, Reuben (2006). "Mobile Phones and Economic Development:Evidence from the Fishing Industry in India". In: *Information Technologies and International Development Journal*.

Abul, Osman, Francesco Bonchi, and Mirco Nanni (2010). "Anonymization of moving objects databases by clustering and perturbation". In: *Information systems* 35.8, pp. 884 –910.

Acs, Gergely and Claude Castelluccia (2014). "A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, New York, USA: ACM, pp. 1679–1688.

Adeline Decuyper, and Alex Rutherford et al. (7). "Estimating Food Consumption and Poverty Indices with Mobile Phone Data". In:

Aggarwal, Charu C. and Philip S. Yu (2008). "Privacy-Preserving Data Mining: Models and Algorithms". In: ed. by Charu C. Aggarwal and Philip S. Yu. Boston, MA: Springer US. Chap. A General Survey of Privacy-Preserving Data Mining Models and Algorithms, pp. 11–52.

Alexander, Lauren, Jameson Toole, and Marta Gonzalez (7). "Assessing the impact of ride sharing on congestion using mobile phone data". In:

Andrey Bogomolov Bruno Lepri, Roberto Larcher Fabrizio Antonelli Fabio Pianesi and Alex Pentland (7). "Energy Consumption Prediction using People Dynamics Derived from Cellular Network Data". In:

Andrienko, Gennady et al. (2013). "Report from Dagstuhl: The Liberation of Mobile Location Data and Its Implications for Privacy Research". In: *SIGMOBILE Mob. Comput. Commun. Rev.* 17.2, pp. 7–18.

Atefeh, Farzindar and Wael Khreich (2015). "A Survey of Techniques for Event Detection in Twitter". In: *Computational Intelligence* 31.1, pp. 132–164. ISSN: 1467-8640.

Azimirad, Ehsan (2015). "The Comprehensive Review on JDL Model in Data Fusion Networks: Techniques and Methods". In: *International Journal of Computer Science and Information Security* 13.1, pp. 53–60.

Barchiesi, Daniele et al. (2015). "Modelling human mobility patterns using photographic data shared online". In: *Royal Society Open Science* 2.8.

Ben-Yacoub, S., Y. Abdeljaoued, and E. Mayoraz (1999). "Fusion of face and speech data for person identity verification". In: *IEEE Transactions on Neural Networks* 10.5, pp. 1065–1074.

Bengtsson, Linus et al. (2015). "Using Mobile Phone Data to Predict the Spatial Spread of Cholera". In: *Scientific Reports* 5.

Beresford, Alastair R. and Frank Stajano (2003). "Location Privacy in Pervasive Computing". In: *IEEE Pervasive Computing* 2.1, pp. 46–55.

Bettini, Claudio, X.Sean Wang, and Sushil Jajodia (2005). "Protecting Privacy Against Location-Based Personal Identification". In: *Secure Data*

Management. Ed. by Willem Jonker and Milan Petković. Vol. 3674. Springer Berlin Heidelberg, pp. 185–199.

Bleiholder, Jens and Felix Naumann (2009). "Data Fusion". In: *ACM Comput. Surv.* 41.1, 1:1–1:41.

Blondel, Vincent D. et al. (2013). "Data For Development:the D4D Challenge on mobile phone data". In: *Orange Data for Development Challenge.*

Bogomolov, Andrey et al. (2015/08/14). "Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics". In: *Big Data* 3.

Botta, F., H. Moat, and T. Preis (2015). "Quantifying crowd size with mobile phone and Twitter data". In: *Royal Society Open Science* 2.150162.

Brickell, J. and V. Shmatikov (2008). "The cost of privacy: destruction of data-mining utility in anonymized data publishing". In: *International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA.

Caceres, C. Small, J. Rowland, and S. Urbanek. (2012). "Exploring the use of urban greenspace through cellular network activity." In: *In Proceedings of the Second Workshop on Pervasive Urban Applications (PURBA)*. Newcastle, UK.

Calabrese, Francesco et al. In:

Castanedo, Federico (2013). "A Review of Data Fusion Techniques". In: *The Scientific World Journal* 2013.704504, p. 19. DOI: 10.1155/2013/704504.

Cavoukian, Ann, Jules Polonetsky, and Christopher Wolf (2010). "SmartPrivacy for the Smart Grid: embedding privacy into the design of electricity conservation". In: *Identity in the Information Society* 3.2, pp. 275–294.

Cecaj, Alket, Marco Mamei, and Franco Zambonelli (2015). "Re-identification and information fusion between anonymized CDR and social network data". In: *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–14.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly Detection: A Survey". In: *ACM Comput. Surv.* 41.3, 15:1–15:58. ISSN: 0360-0300.

Chareyron, G., J. Da-Rugna, and B. Branchet (2013). "Mining tourist routes using flickr traces". In: *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference*, pp. 1488–1489.

Chen, D. and H. Zhao (2012). "Data Security and Privacy Protection Issues in Cloud Computing". In: *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*. Vol. 1, pp. 647–651.

Cheng, Zhiyuan, James Caverlee, and Kyumin Lee (October 26–30, 2010). "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users". In: *CIKM'10*. Toronto, Ontario, Canada.

Christin, Delphine et al. (2011). "A survey on privacy in mobile participatory sensing applications". In: *Journal of Systems and Software* 84.11. Mobile Applications: Status and Trends, pp. 1928 –1946.

Claubeta, Sebastian, Dogan Kesdogan, and Tobias Kolsch (2005). "Privacy Enhancing Identity Management: Protection Against Re-identification and Profiling". In: *Proceedings of the 2005 Workshop on Digital Identity Management*. DIM '05. Fairfax, VA, USA: ACM, pp. 84–93.

Crandalla, David J. et al. (2010). "Inferring social ties from geographic coincidences". In: *22436–22441, PNAS vol. 107 no. 52 December 28 2010*.

Daries, Jon P. et al. (2014). "Privacy, Anonymity, and Big Data in the Social Sciences". In: *Commun. ACM* 57.9, pp. 56–63. ISSN: 0001-0782.

De Bie, Tijl et al. (2007). "Kernel-based data fusion for gene prioritization". In: *Bioinformatics* 23.13, pp. i125–i132. eprint: http://bioinformatics.oxfordjournals.org/content/23/13/i125.full.pdf+html.

Do, Trinh Min Tri, Jan Blom, and Daniel Gatica-Perez (2011). "Smartphone usage in the wild: a large scale analysis of Applications and Context." In: *ICMI '11 Proceedings of the 13th international conference on multimodal interfaces Pages 353-360*. Alicante, Spain.

Dobra, Adrian, Nathalie E. Williams, and Nathan Eagle (2015). "Spatiotemporal Detection of Unusual Human Population Behavior Using Mobile Phone Data". In: *PLoS ONE* 10.3.

Dong, Yuxiao et al. (2015). "Inferring Unusual Crowd Events From Mobile Phone Call Detail Records". In: *CoRR* abs/1504.03643.

Douglass, Rex W et al. (2015). "High resolution population estimates from telecommunications data". In: *EPJ Data Science* 4.4.

Durrant-Whyte, Hugh and Thomas C. Henderson (2008). "Springer Handbook of Robotics". In: ed. by Bruno Siciliano and Oussama Khatib. Berlin, Heidelberg: Springer Berlin Heidelberg. Chap. Multisensor Data Fusion, pp. 585–610.

Dwork, C. (2011). "The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques". In: *IEEE Symposium on Foundations of Computer Science*. Palm Springs, CA, USA.

Eldardiry, H. et al. (2013). "Multi-Domain Information Fusion for Insider Threat Detection". In: *Security and Privacy Workshops (SPW), 2013 IEEE*, pp. 45–51.

Ferragina, Paolo and Ugo Scaiella (2010). "TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities)". In: CIKM '10. Toronto, ON, Canada: ACM, pp. 1625–1628. ISBN: 978-1-4503-0099-5.

Ferrari, L., M. Mamei, and M. Colonna (2012). "People Get Together on Special Events: discovering happenings in the city via cell network analysis". In: *IEEE International Workshop on the Impact of Human Mobility in Pervasive Systems and Applications*. Lugano (Switzerland).

Freiwald, Susan (2011). "Cell Phone Location Data and the Fourth Amendment: A Question of Law, Not Fact". In: *Maryland Law Review* 70.

Girardin F Vaccari A, Gerber A Biderman A Ratti C. (2009). "Quantifying urban attractiveness from the distribution and density of digital footprints." In: *International Journal of Spatial Data Infrastructures Research.* 4.175-200, pp. 5317 –5325.

Gross, G. A. et al. (2012). "Towards hard - soft data fusion: Processing architecture and implementation for the joint fusion and analysis of hard and soft intelligence data". In: *Information Fusion (FUSION), 2012 15th International Conference on*, pp. 955–962.

Gruteser, Marco and Dirk Grunwald (2003). "Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking". In: *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*. MobiSys '03. San Francisco, California: ACM, pp. 31–42.

Guidotti, R. et al. (2015). "Social or Green? A Data-Driven Approach for More Enjoyable Carpooling". In: *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pp. 842–847.

Guttormsson, S.E. et al. (1999). "Elliptical novelty grouping for on-line short-turn detection of excited running rotors". In: *Energy Conversion, IEEE Transactions on* 14.1, pp. 16–22.

Hall, D. L. and J. Llinas (1997). "An introduction to multisensor data fusion". In: *Proceedings of the IEEE* 85.1, pp. 6–23.

Heeney C Hawkins N, de Vries J Boddington P Kaye J (2011). "Assessing the Privacy Risks of Data Sharing in Genomics". In: *Public Health Genomics* 14, pp. 17–25.

Hoh, Baik and M. Gruteser (2005). "Protecting Location Privacy Through Path Confusion". In: *Security and Privacy for Emerging Areas in Communications Networks, 2005. SecureComm 2005. First International Conference on*, pp. 194–205.

Hong, Jason I. et al. (2004). "Privacy Risk Models for Designing Privacy-sensitive Ubiquitous Computing Systems". In: *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. Cambridge, MA, USA: ACM, pp. 91–100.

Ihler, Alexander, Jon Hutchins, and Padhraic Smyth (2006). "Adaptive Event Detection with Time-varying Poisson Processes". In: *International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA, USA.

Illushka, Seresinhe Chanuki, Preis Tobias, and Moat Helen Susannah (2015). "Quantifying the Impact of Scenic Environments on Health". In: *Scientific Reports* 5.

Isaacman, D. J. Mirand S. et al. (2013). "DP-WHERE: Differentially private modeling of human mobility". In: *Big Data, 2013 IEEE International Conference on*, pp. 580–588.

Jahanbakhsh, Kazem, Valerie King, and Gholamali C. Shoja (2012). "They Know Where You Live!" In: *CoRR* abs/1202.3504.

Jie Zhao, Xueya Wang and Zheng Ma. "Towards Events Detection from Microblog Messages". In:

Junglas, Iris A, Norman A Johnson, and Christiane Spitzmüller (2008-08-01T00:00:00). "Personality traits and concern for privacy: an empirical study in the context of location-based services". In: *European Journal of Information Systems* 17.4, pp. 387–402.

Kang, Joon-Myung, Sin seok Seo, and James Won-Ki Hong (2011). "Usage Pattern Analysis of Smartphones". In: *Network Operations and Management Symposium (APNOMS)13th*. Asia-Pacific.

Khaleghi, Bahador et al. (2013). "Multisensor data fusion: A review of the state-of-the-art". In: *Information Fusion* 14.1, pp. 28 –44.

Kido, H., Y. Yanagisawa, and T. Satoh (2005). "An anonymous communication technique using dummies for location-based services". In: *Pervasive Services, 2005. ICPS '05. Proceedings. International Conference on*, pp. 88–97.

Krumm, John (2008). "A survey of computational location privacy". In: *Personal and Ubiquitous Computing* 13.6, pp. 391–399.

Lahat, D., T. Adali, and C. Jutten (2015). "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects". In: *Proceedings of the IEEE* 103.9, pp. 1449–1477. ISSN: 0018-9219. DOI: 10.1109/JPROC.2015.2460697.

Lahat, D., T. Adalý, and C. Jutten (2014). "Challenges in multimodal data fusion". In: *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pp. 101–105.

Lambiotte, Renaud et al. (2008). "Geographical dispersal of mobile communication networks". In: *Physica A: Statistical Mechanics and its Applications* 387.21, pp. 5317 –5325.

Malin, Bradley, David Karp, and Richard H. Scheuermann (2015). "Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research". In: *Journal of Investigative Medicine* 58.1, pp. 11–18.

Mascetti, Sergio et al. (2013). "Anonymity: A Comparison Between the Legal and Computer Science Perspectives". In: *European Data Protection: Coming of Age*. Ed. by Serge Gutwirth et al. Springer Netherlands, pp. 85–115. ISBN: 978-94-007-5184-2.

Matthews, Gregory J. and Ofer Harel (2011). "Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy". In: *Statistics Surveys* 5, pp. 1–29.

Minch, R. P. (2004). "Privacy issues in location-aware mobile devices". In: *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, 10 pp.–.

Mitchell, Lewis et al. (2013). "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place". In: *PLoS ONE* 8.5, pp. 1–15.

Montjoye, Y. et al. (2013). "Unique in the crowd. The privacy bounds of human mobility". In: *Scientific Reports* 3, pp. 161 –180.

Narayanan, Arvind and Vitaly Shmatikov (2008). "Robust De-anonymization of Large Sparse Datasets". In: *IEEE Symposium on Security and Privacy*. Las Vegas, Nevada, USA.

Neumann, Joachim et al. (2013). "Event Detection in Communication and Transportation Data". In: *Pattern Recognition and Image Analysis*. Ed. by JoãoM. Sanches, Luisa Micó, and JaimeS. Cardoso. Vol. 7887. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 827–838. ISBN: 978-3-642-38627-5.

Nigay, Laurence and Joelle Coutaz (1993). "A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion". In: *Proceedings of the INTERACT ́93 and CHI ́93 Conference on Human Factors in Computing Systems*. CHI '93. Amsterdam, The Netherlands: ACM, pp. 172–178. ISBN: 0-89791-575-5.

Noulas, Anastasios et al. (2012). "A Tale of Many Cities: Universal Patterns in Human Urban Mobility". In: *PLoS ONE* 7.5, pp. 1–10.

Nurwidyantoro, A. and E. Winarko (2013). "Event detection in social media: A survey". In: *ICT for Smart Society (ICISS), 2013 International Conference on*, pp. 1–5.

Parent, Christine et al. (2013). "Semantic Trajectories Modeling and Analysis". In: *Journal ACM Computing Surveys (CSUR)* 45.42, pp. 161 –180.

Paul S. Horn a, Lan Feng Yanmei Li and Amadeo J. Pesce (2001). "Effect of Outliers and Nonhealthy Individuals on Reference Interval Estimation". In: *Clinical Chemistry* 47.12, pp. 2137–2145.

Pejovic, Veljko and Mirco Musolesi (2013). *Anticipatory Mobile Computing: A Survey of the State of the Art and Research Challenges*. ArXiv: arXiv:1306.2356.

Pentland, Alex (2014). "Big Data: Balancing the Risks and Rewards of Data-Driven Public Policy". In: *2014 World Economic Forum. The Global Information Technology Report 2014*.

Perusco, L. and K. Michael (2007). "Control, trust, privacy, and security: evaluating location-based services". In: *IEEE Technology and Society Magazine* 26.1, pp. 4–16.

Popescu, Ana-Maria and Marco Pennacchiotti (2010). "Detecting Controversial Events from Twitter". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. Toronto, ON, Canada: ACM, pp. 1873–1876. ISBN: 978-1-4503-0099-5.

Quercia, D. et al. (2011). "Mobile Phones and Outdoor Advertising: Measurable Advertising". In: *IEEE Pervasive Computing* 10.2, pp. 28–36.

Rajaraman, Anand and Jerey D. Ullman (2011). *Mining of massive datasets*. Berlin, Germany: Springer Verlag.

Reades, J. et al. (2007). "Cellular Census: Explorations in Urban Data Collection". In: *IEEE Pervasive Computing* 6.3, pp. 30–38.

ROSA, Mario LA et al. (2015). "ENERGY CONSUMPTION AND DEMAND ESTIMATION FROM CELLULAR NETWORK DATA: A REAL WORLD CASE STUDY". In:

Rothstein, Mark A. (2010). "Is Deidentification Sufficient to Protect Health Privacy in Research?" In: *The American Journal of Bioethics* 10.9, pp. 3–11.

Rozenshtein, Polina et al. (2014). "Event Detection in Activity Networks". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD 14. New York, New York, USA: ACM, pp. 1176–1185.

Saluveer, Erki et al. (7). "Measuring de facto populations with Mobile Network Operator's Call Detail Record data". In:

Sassi, Andrea and Franco Zambonelli (2014). "Coordination Infrastructures for Future Smart Social Mobility Services". In: *IEEE Intelligent Systems* 29.5, pp. 78–82. ISSN: 1541-1672.

Sevtsuk, Andres and Carlo Ratti (2010). "Does Urban Mobility Have a Daily Routine? Learning from the Aggregate Data of Mobile Networks". In: *Journal of Urban Technology* 17.1, pp. 41–60.

Shilton, Katie (2009). "Four Billion Little Brothers?: Privacy, Mobile Phones, and Ubiquitous Data Collection". In: *Commun. ACM* 52.11, pp. 48–53.

Shokri, R. et al. (2011). "Quantifying Location Privacy". In: *Security and Privacy (SP), 2011 IEEE Symposium on*, pp. 247–262.

Song, Yi, Daniel Dahlmeier, and St. In:

Soto, V. and E. Frias-Martinez. (2011). "Robust land use characterization of urban landscapes using cell phone data." In: *In Proceedings of the First Workshop on Pervasive Urban Applications(PURBA)*.

Steiger, Enrico et al. (2015). "Twitter as an indicator for whereabouts of people? Correlating Twitter with {UK} census data". In: *Computers, Environment and Urban Systems* 54, pp. 255 –265.

Steinberg, Alan N., Christopher L. Bowman, and Franklin E. White (1999). *Revisions to the JDL data fusion model*.

Sweeney, L. (2002). "k-Anonymity: A Model for Protecting Privacy". In: *International Journal on Uncertainty,Fuzziness and Knowledge-based Systems* 18.10, pp. 557–570.

Sweeney, L., A. Abu, and J. Winn (2013). "Identifying Participants in the Personal Genome Project by Name." In: *White Paper 1021-1, Harvard University. Data Privacy Lab.*

Tan, Shulong et al. (2014). "Cross domain recommendation based on multi-type media fusion". In: *Neurocomputing* 127, pp. 124 –134. URL: http://www.sciencedirect.com/science/article/pii/S0925231213009260.

Tang, Karen P. et al. (2006). "Putting People in Their Place: An Anonymous and Privacy-sensitive Approach to Collecting Sensed Data in Location-based Applications". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: ACM, pp. 93–102.

Traag, V.A. et al. (2011). "Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference". In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 625–628.

Treinish, Lloyd A. (2000). "Visual Data Fusion for Applications of High-resolution Numerical Weather Prediction". In: *Proceedings of the Conference on Visualization '00*. Salt Lake City, Utah, USA: IEEE Computer Society Press, pp. 477–480.

Ulfarsson, M. O., J. A. Benediktsson, and J. R. Sveinsson (2003). "Data fusion and feature extraction in the wavelet domain". In: *International Journal of Remote Sensing* 24.20, pp. 3933–3945.

Varshney, P. K. (1997). "Multisensor data fusion". In: *Electronics Communication Engineering Journal* 9.6, pp. 245–253.

Veloso, Marco, Santi Phithakkitnukoon, and Carlos Bento (2011). "Urban Mobility Study Using Taxi Traces". In: *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*. TDMA '11. Beijing, China: ACM, pp. 23–30.

Verkasalo, Dr. Hannu (2010). "Analysis of Smartphone User Behavior". In: *Ninth International Conference on Mobile Business / 2010 Ninth Global Mobility Roundtable*.

Vu, Huy Quan et al. (2015). "Exploring the travel behaviors of inbound tourists to Hong Kong using geotagged photos". In: *Tourism Management* 46, pp. 222 –232.

Walther, Maximilian and Michael Kaisser (2013). "Geo-spatial Event Detection in the Twitter Stream". In: *Proceedings of the 35th European Conference on Advances in Information Retrieval*. ECIR'13. Moscow, Russia: Springer-Verlag, pp. 356–367.

Wang, Y., L. Guan, and A. N. Venetsanopoulos (2012). "Kernel Cross-Modal Factor Analysis for Information Fusion With Application to Bimodal Emotion Recognition". In: *IEEE Transactions on Multimedia* 14.3, pp. 597–607.

Watanabe, Kazufumi et al. (2011). "Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. CIKM '11. Glasgow, Scotland, UK: ACM, pp. 2541–2544. ISBN: 978-1-4503-0717-8.

Weng, M. F. and Y. Y. Chuang (2012). "Cross-Domain Multicue Fusion for Concept-Based Video Indexing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.10, pp. 1927–1941.

Wesolowski, Amy et al. (2015). "Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data". In: *Proceedings of the National Academy of Sciences* 112.35, pp. 11114–11119.

Wicker, S. (2012). "The Loss of Location Privacy in the Cellular Age". In: *Communication of the ACM* 55.8, pp. 60 –68.

Xu, Chang, Dacheng Tao, and Chao Xu (2013). "A Survey on Multi-view Learning". In: *CoRR* abs/1304.5634.

Yin, Jie, Derek Hao Hu, and Qiang Yang (2009). "Spatio-temporal Event Detection Using Dynamic Conditional Random Fields". In: *Proceedings of the 21st International Jont Conference on Artifical Intelligence*. IJCAI'09. Pasadena, California, USA: Morgan Kaufmann Publishers Inc., pp. 1321–1326.

Zambonelli, F. (2008). "Toward sociotechnical urban superorganisms". In: *IEEE Computer* 45.8, pp. 76–78.

Zang, Hui and Jean Bolot (2011). "Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study". In: *MobiCom'11*. Las Vegas, Nevada, USA.

Zheng, Danning et al. (2014a). "Inferring Home Location from User's Photo Collections Based on Visual Content and Mobility Patterns". In: *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia*. GeoMM '14. Orlando, Florida, USA: ACM, pp. 21–26.

Zheng, Yu and Senior Member (2015). "Methodologies for Cross-Domain Data Fusion: An Overview". In: *IEEE Transactions on Big Data*.

Zheng, Yu et al. (2014b). "Urban Computing: Concepts, Methodologies, and Applications". In: *ACM Trans. Intell. Syst. Technol.* 5.3, 38:1–38:55.

Zhou, Tao (2011). "The impact of privacy concern on user adoption of location based services". In: *Industrial Management & Data Systems* 111.2, pp. 212–226.

Zhou, Xiangmin and Lei Chen (2014). "Event Detection over Twitter Social Media Streams". In: *The VLDB Journal* 23.3, pp. 381–400.