Analysis of Spatio-Social Relations in a Photographic Archive (Flickr)

Nazanin Khalili Shavarini

Department of information science
School of Informatics
City University, London

A thesis submitted for the degree of
Doctor of Philosophy in Information Science
October 2011

Table of Contents

List of Figures

List of Equations

List of Tables

List of Appendices

Acknowledgements

In completing this research I owe my debts to many people. I would like to thank Jo Wood and Jason Dykes for their direction, assistance and guidance. In particular I wish to thank Jo who knows perfectly when to push me to work and when to leave me on my own. He taught me the techniques of programming and systematic thinking and for that I will always remain grateful.

Special thanks should be given to my friends in Information Science department, in particular Stelios Papakonstantinou and Aidan Slingsby, for their support, understanding and encouragement when I needed them the most.

I also owe a huge debt of thanks to my family who always been unreasonably generous to me. Lots of thanks are due to my parents, Noshin, Farhad and Nikki.

I must also thank my husband, Homy, for his unconditional love and support during the long hours of work I dedicated to this research. I am deeply indebted for his support and distraction of course.

Declaration

I grant Powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Abstract

This thesis aims to study and analyse the complex spatio-social relations among social entities who interact together in a spatially structured social group. This aim is approached in three steps:

1. Collecting and classifying spatio-social data,
2. Disambiguating place names that people use to refer to their homes and
3. Analysis of data of this kind (numerical and visual).

The source of spatio-social data used in this work is Flickr. Flickr is a yahoo photo sharing site. Users have a social network of friends and a collection of photos on their profiles. According to available statistics[1] the Flickr database contains more than three billion photos, out of which a hundred million are geo-tagged. In retrieving data from Flickr database two different samples have been explored. Initially a random collection of photos that have been uploaded in Flickr during the examined periods has been collected on a daily basis. This is followed by much narrower and more precise criteria for the second data sampling that resulted in Flickr sample GB data.

The thesis concludes that location dominates a significant pattern in online behavior of social entities who interact together via internet. The core contributions of this thesis are in the areas of:

1. Extracting indicative sample from very large data sets,
2. Disambiguation of place names that people use in their natural language to refer to their home locations and
3. Proposing potential new insights into behaviors of social entities with spatio-social relations.

Overall, the popularity of social networking sites and availability of data that can be obtained from the web (whether people provide voluntarily or can be retrieve as a consequence of online interactions) are likely to continue the increasing trend in future. In addition, the realm of spatio-social data analysis and its visualization also continue to expand, as do the types of maps that are achievable, the visualization packages that the maps can be built with, the number of map users and improved gazetteers with more comprehensive coverage of vague terms. Therefore, the developed methods, algorithm and applications in this study can be beneficial to researchers in social and e-social sciences, those who are interested in developing and maintaining social networking sites, geographers who work on disambiguation of fuzzy vernacular geographic terms, visualization and spatial data analysts in general and those who are looking for development and accommodation of better business strategies (i.e. localization and personalization).

[1] (http://www.Flickr.com, retrieved 20/07/09)

1 Introduction

1.1 Introduction

Social relations and interaction patterns are visualized using node link graphs in traditional social network maps (Scott 1991, Wasserman and Faust 1994, Wellman and Berkowitz 1988). The resultant network graphs frequently alter the geometric relations present in the real world in order to emphasize the connectivity and overall view of the networks. Whilst a node's position has considerable potential for carrying information regarding network pattern and structure, no spatial information is usually encoded: the node's position does not contain any spatial attribute of the data, while it has considerable potential for carrying information regarding network patterns, nature and structures. Location is an important spatial property of social entities (Wellman 1996) and integrating that information with social network data has potential for revealing insights into hidden patterns behind communities.

The advent of Web 2.0 and the popularity of online social networks have resulted in masses of voluntarily submitted locational information being available for study and analysis. Much of this has a geographic component, described as Volunteered Geographic Information, or VGI, by Goodchild (2007). The abundance of VGI provides opportunities for analysing the geography of social networks that have not yet been used to their full potential. Gaining insights into the patterns, nature and structure of large spatio-social data sets is a demanding job in different domains (Guo et al. 2006). Consequently, to date relatively little research has used locational information in a social network context. While attempts have been made in the past, (Liben-Nowell et al. 2005, Escher 2007, Wellman 1996) they have tended to focus on one specific location for each node, confining members to a bounding box of a city. Although this naïve geography is useful for variety of purposes (i.e. small world phenomena, analysis of scientific collaboration and friendship network) it is of limited or no use for analysis of multiple locations associated with a social entity.

Whilst there is scope for determining multiple geographies from large numbers of volunteers, the process is by no means straightforward. In reality, people are associated with more than one physical location and VGI provides data by which these might be determined. For example, online geo social networks such as yelp[2], Flickr[3], Gypsii[4], etc. enable us to associate users with home location, point of interest, work place, and geo located digital documents, etc. This variability in individuals' locations that are accessible through VGI raises a number of issues in visualization and network analysis.

The massive unformatted VGI available online, although having the potential to be representative of the complex geography in social interactions and relations, is of limited or no use for most of the time in the existing format.

[2] http://www.yelp.com

[3] http://www.Flickr.com

[4] http://www.gypsii.com

Therefore, this research aims to extend the existing work to develop methods for study and analysis of spatially variable attributes of entities in social networks. In other words this research takes advantage of the wasted geographies in graph presentation of social networks (Viegas and Donath, 2004) and will develop a suitable visualization solution for studying and modeling the multi geographies accessible through VGI. In doing so, a case study of Flickr (Yahoo photo-sharing site) will be used.

1.2 Motivation and Application

The motivation for this work comes from the fact that the web is still on exponential increase and so is the amount of digital information accessible for users. In December 2011 it has been estimated that there were 2.5 billion Internet users that make up about 33% of the world population[5]. As mentioned in previous section each online member knowingly or inadvertently leave social, temporal or spatial information on web for data aggregators to collect and study. Despite this wealth of information our tools and ability for collecting the data is far more advanced than our ability to make sense of it. Study and analysis of online users have potential to be beneficial to several fields of research, business and industry. That indicates the urgent need for effective way of managing that large volume of data. There have been several studies conducted by different scholars in the field regarding the effects of Internet on business, industry and every day routine of social entities (Dorogovtsev and Mendes, 2003; Watts and Strogatz, 1998; Huberman et al. 1998; Barabasi and Albert, 1999; Menczer, 2004; Dorogovtsev and Mendes; 2003). Although the way that individuals behave on the web is ultimately personal and unpredictable, there are certain factors that can be studied and generalized according to the most common behavior (Watts, 2007; Borgatta and Montgomery; 2000). Although, still it is not clear how Internet users interact with the web 2.0 technologies in respect of several properties e.g. age, culture, physical location, education and social role, etc., it has been agreed that there are significant impact on culture and businesses regarding Internet and especially web 2.0 (Berners-Lee et al. 2006; Watts, 2007; Pastor-Satorras and Vespignani; 2004, Goker et al. 2009). Accordingly, in regards to these implications, a brief overview of the fields and areas that can be benefitted from this study are discussed.

1. This study has implications on how businesses can plan their marketing strategies according to online behavior of social entities. In simpler terms, understanding of how people communicate together on the web according to their physical footprints and home locations can help in improving the strategies of marketing from target advertising to free and paid advertisement methods.

   It has been identified that web data has specific characters and users also have distinctive behaviors (Craswell and Hawking, 2009). Therefore understanding this distinctive behavior of online users and recognizing any pattern especially in their geographic footprints might be useful for providing context for context aware systems in order to deliver relevant information to the user (Goker et al. 2009). For instance there might be possibility of estimating the home location of a user by study and analysis

[5] http://www.internetworldstats.com/

of their geographic distribution of their geo-tagged photo collections. Consequently the localized information can be delivered with higher priority. For example relevant advertisements in the local market can be presented first (localization). Moreover, the search results can also be ranked according to the distance of the retrieved services to the place the user live or is present at the time of the search (mobile location based services/advertising (Craswell and Hawking, 2009). Other examples can widen the scope of the potential application of analysis of the online behavior of social entities to systems like context aware tourist applications (Cheverst et al. 2000). In the same line as the examples above, by estimating/disambiguating home location of users, tourist can be distinguished from locals and therefore, different appropriate information can be delivered accordingly.

2. Today the advent of Web 2.0 and related technologies and popularity of 'User Generated Content', have affected the way that businesses plan their strategies. Nowadays, the attention of large businesses has moved from consuming activities to producing activities. In doing so they expect to receive business value by allowing users to produce content (Jose van Dijck, 2008). Therefore, analysis of the users' behavior on the web is of great importance and potentially can help businesses to plan their strategies more efficiently. In other words, analysis of VGI and the associated spatio-social attributes can help in accommodating products and services to better fit the users' requirements. This has been called 'personalization' and has attracted attention of several researchers in the field (Prestchener and Gauch, 1999; Khopkar et al. 2003; Pitkow et al. 2002; Pretschner and Gauch, 1999; Lieberman et al. 2001). For example a retailer with an online feedback facility form can alter the design or quality of their products according to the customers feedback for individual items. Likewise users' reviews allow a holiday company to amend their services to better satisfy the customers.

   Saving the previous interaction of the user with the system (i.e. blog postings, bookmarks, favourite, preferences, the links that have been clicked) can help in making intelligent predictions for their current or future needs (optimizing the search result). For example a holiday finder company can advertise its facility by estimating users favorite places that can be retrieved by previous interaction of the users with system. The holidays can also be customized based on the estimation of the users' whereabouts (that can lead in identifying language, culture, specific preference, etc.). Making intelligent recommendations to a user according to the preference of previous users with the same information need is also beneficial to the businesses (Teevan et al. 2005). The obvious example of this is the way that Amazon[6] recommends the user other appropriate products based on shopping pattern of previous users who bought the same products. The distribution of book sales also has been used to design better online stores and recommendation engines (Newman, 2005b).

[6] www.amazon.co.uk

3. Information retrieval (IR) and specifically Image Retrieval (as a sub section for the Information retrieval field) can also be benefitted from the methods and techniques developed and discussed in this study. There are several studies about identifying the users' need and predict people behaviors in online environments. They all relate to understanding of 'context' that has been identified as a fundamental basis for study and analysis of people's information needs and 'user needs analysis' (Goker et al. 2009). Zwol et al. (2008) and Lew et al. (2006) worked on image retrieval based on user generated content and visual context. Overell (2009) has proposed a method for combining the textual meta data provided by users with the visual information in order to improve the image retrieval on the web. Olivaries (2010) has proved that use of UGC can significantly improve the image retrieval process. Study of the geographic footprints of photos for posters and identifying where the posters come from can help in ranking and categorization of their images. Moreover, the study and analysis of UGC has been found useful for dealing with uncertainty involved in IR interactions (Zwol, 2008). Therefore, the disambiguation algorithm and techniques developed (section 5.2-3) for dealing with ambiguity and uncertainty in fuzzy vernacular geographic terms that people use to refer to their home locations might also be useful for information retrieval interactions.

The methods and techniques developed in this study for analysis of the attributes of Flickr users in regards to their home location, geographic footprints of their photos, their friends' home locations and their geo-photo collections have potential to help in designing better systems that rely on human online behaviors (e.g. web sites, advertisements, personalized maps, calendars, tourist applications etc.). In addition, the methods are expected to be also useful in some elements of natural language processing and those interested in analysis and disambiguation of location names. This work also can be useful for those who are interested in developing personalized information systems (Brusilovsky 1996) through social filtering (Shardanand and Maes 1995) or location based services for personalization of geographic maps (Beigl, 2002).

There are also evidence that the majority of search engines apply disambiguation methods and techniques to improve the accuracy of their search results (Russell-Rose and Stevenson, 2009). Therefore disambiguation of place names is of potential relevance to information retrieval and natural language processing techniques. Consequently, the developed algorithm and techniques for disambiguation of the fuzzy vernacular geographic terms that people refer to their home locations on the web can be useful for search engine in the case of dealing with ambiguous place names as search keywords.

Accordingly, the analysis and study of the spatial (geographic distribution of geo-tagged photos, home location) and social (online friendship network) attributes of group of social entities who interact together on the web can be beneficial for several context aware applications and search engines. Overall, the methods and techniques developed and applied in this study are expected to result in crucial understanding of UGC systems and can provide valuable information to Internet service providers, the administrators and content owners with major commercial and technical implications.

1.3 Aims

In summary, this research aims to:

1. Identify concerns regarding locational uncertainties of online social entities and their implications.
2. Develop new forms of spatial social network presentations that support the visual synthesis of locational data and online relations.
3. Identify the challenges of visualizing large spatio-social information in a spatially structured social group.

1.4 Objectives

During this research the following objectives are to be achieved:

1. Develop techniques to discover and analyze spatio-social relations within a large spatially structured social group.
2. Assess the strengths and weaknesses of the developed visualizations in identifying the role of geography in online interactions and friendship patterns.

1.5 Research Questions

In accordance with the aims and objectives mentioned above, the following questions are to be answered during the course of this research:

1. What limitations are involved in applying VGI in spatial social network presentation?
2. Can geographic distribution of photos contribute to prediction of FHLI?
3. Can geographic distribution of friends contribute to prediction of FHLI?
4. To what extent locational information dominates a pattern in online communities?
5. Are spatial social network maps useful for better understanding of the geographic distribution of individuals and their online behaviors?
6. And finally is there any uncertainty involved in interpretation of spatio-social relations?

1.6 Contributions

This work is the first attempt to consider and visualize the variability in locations associated with social entities in online environments. It differs particularly from others in that it works with large volume of spatio-social data in regards to multi-geographies (obtainable from the web) in a social network context. The aim is to uncover the unknown patterns and relationships in complex spatio social relations among social entities.

Since the large complex spatio-spatial data have been increasingly become available on the net, this research will contribute to the knowledge and clarification of how to come up with appropriate methods for exploratory analysis of such data. Therefore, this study will involve in study, analysis and quantifying the ambiguity, uncertainty and effects of spatial data in online behavior of social entities in the scale that have not been conducted before. In doing so, this research aims to map the geography of online social communities in order to understand the effects of geography on online communities and extract the relevant information from the multiple geographies associated with each social entity. The developed classification methods, uncertainty measurements, disambiguation algorithms and visualization application are expected to contribute to the existing knowledge in the field of social and e-social sciences, development and maintenance of social networking sites, analysis of vague spatial terms, visualization and spatial data analysis.

1.7 Outline

In the next chapter a review of literature related to this research is given. It briefly describes social network analysis and examines its position regarding visualization, geography, ambiguity and privacy.

Chapter 3 summarises the process of data selection for this study. It reviews the data requirements and selects some essential criteria for an appropriate data set for this study.

The 4th chapter explains the characteristics of the data used in this research. It reviews the attributes and specifications of two sets of Flickr data collected for this study.

Chapter 5 covers the steps followed for classification, Disambiguation and visualization of the examined spatio-social data.

Chapter 6 covers the numerical analysis of selected data. The first part reviews the 'Flickr sample world data' and the second section looks into the 'Flickr sample GB data'.

Chapter 7 demonstrates how the developed visualization application can be used in exploring the data.

Chapter 8 applies the visualization application to the sampled data in order to answer the research questions.

Chapter 9 summarises the conclusions of the research and provides recommendations on how to improve the future research on the developed applications.

The References section contains a comprehensive list of reference material and literature sources related to the study.

The Appendices include material that is pertinent to this study but would interrupt the flow of the discussions if presented within the main body of the text. Java application for data collections and the interactive software and Figures provided in a separate appendices (on attached CD) form a major element of the work, in terms of illustration, demonstration and specification of the methods used.

2 Literature

2.1 Introduction

This chapter provides a brief overview of literature in the context of this study. The first section reviews the meaning and definition of online social networking sites in general. Afterwards, it studies the social networking sites from perspectives of visualization, geography, locational ambiguity and privacy.

2.2 User Generated Content

Today the 'web 2.0' has brought a new way of interactions and content creation. Users upload information on the web through variety of facilities. Wikis, tagging, rating, blogging, social networking, and etc. all enable users to communicate and interact on the web directly (photos, blogs, social profiles, etc.) or indirectly (digital footprint, IP address). As a result of these technologies the way that information can be provided and disseminated has been changed drastically. The voluntarily uploaded information can be used to add value to the existing content on the web (user review on products) or produce the contents (photos uploaded on Flickr or news on dig.com).

In other words, the most notorious effect of web 2.0 is on the way that interactions on the web can take place. As Cha et al. (2007) noted today's Internet users are self-publishing consumers as well. Unlike the traditional method of communication that is bounded to the same time and same place with the same device today there is access to personalized behavior of social entities that interact, participate and communicate together in their convenient time and manner. In the era of web 2.0 the content creation is no more exclusive to the information providers who are obliged to follow the policies, terms and conditions of their agencies (Cha et al. 2007). However, in majority of cases users are required to register with the sites before they can upload content. Consequently, knowingly or naively they provide important information about their whereabouts and behaviors to site owners and their consequent data aggregators. This realm of voluntarily uploaded information that users provide on the web is referred to as user generated content (UGC). According to the existing literature there is not any universally accepted definition for UGC (Koskinen, 2003 and Miller, 2005). Overall, any kind of information (i.e. files, audio, video, photos or text) that users provide on the web can be referred to as user generated content (Ochoa and Duval, 2008).

UGC can be studied from two different angles: from perspective of consumption (Cha et al. 2007) or from production point of view (Ochoa and Duval, 2008). The data set of this study has been collected in regards to users who 'produce' content on the web (geo-tagged photo and home locations). In the same line in regards to consumption and production there is participation issue that needs to be reviewed as well.

2.2.1 Participation

Despite what the web 2.0 technologies implied and have been expected -to transform the users from mere consumers to potential data creators- research has been revealed the unbalanced nature of the users' participation on the web (Jose van Dijick 2008, Nielsen 2006). Although it's been theorized that each user on the web also create content, it's been proved that the majority of postings come from small minority of users. Will et al. (1992) have studied this phenomenon for the first time and their findings gradually have resulted in the introduction of the '90-9-1' principle. This principle implies that the majority of the content on the web come from a small minority of the users. In precise terms as Neilson (2006) has described 90% of users are lurkers[7] who only observe the data and interactions on the web. 9% are intermittent contributors who leave comments or review for the data that are already on the web. Finally only the remaining 1% of the users are heavy contributors who create majority of the contents. Social scientists have observed this pattern since early nineties as 'participation inequality' and it has been found in every online environment and services that has been used by several social entities (Simon, 2010). Accordingly, a recent study by (Simon, 2010) shows that in general 24% of people who engage in the web 2.0 interactions have contributed to the site at some point. However, this number is considerably smaller for any specific examined site. For example a simple statistics on favourite video uploading site 'you tube[8]' showed that only 0.16% of viewers of videos have ever uploaded any video on the site (Simon, 2010).

This wiki style architecture of participation on the web has drastic impact on generalizing habits and patterns of online users. According to the participation inequality rule the behavior of a random sample of online users (between 9% to 1%) is just representative of small population of users who contribute a lot. The challenge here comes from the fact that if the 90% of users are lurkers who only observe the content and activities then how their behaviors can be studied. In other words how lurkers can be encouraged to participate more. This inequality comes to its most impact when companies and agencies make their business plan according to online behavior of users. Making decision about a product according to the 1% who wrote a review is not an efficient way of assessing users' feedback (Jose van Dijick, 2008). In doing so, participation inequality has been recognized as social media massive dispute. There are interesting research going on about the characteristics of lurkers, why they lurk and methods and technique on how to encourage more participation in online environments and develop methods and techniques to activate the silent observes of the web: Nonnecke and Preece, 2001; Neilsen, 2006; MacDonald, 2003; Rafaeli et al. 2004; Salmon, 2003; and Nonnecke et al. 2004. However, this topic is not in the scope of the research conducted here.

After all, despite the significant inequality in users' participation, the large amount of available data on the web still carry considerable potential for revealing interesting patterns about online behavior of social entities. However, it is important to be aware of this unbalanced nature of the web, while collecting data and generalizing findings. Consequently, in order to reduce the effect of this inequality, the data collection process in this study (chapter 4) is conducted by defining some specific criteria on contribution of the users. By having the data collected according to the minimum required participation, although generalizations are still at slight inevitable risk the lurkers who have no role in content creations have been excluded from analysis in this thesis. Efforts have

[7] Alternative description have been assigned to the term lurker in literature 'browser' with Salmon (2003), 'legitimate peripheral participant', McDonald (2003), 'Read Only Participants, ROPs' by William (2004) and 'vicarious learner' by Lee (1999).

[8] www.youtube.com

been made to collect data from the participants and therefore, the findings of this research are generalizable to participants and cannot be applied to lurkers who have not uploaded any photos at all.

2.2.2 Volunteered Geographic Information

In spite of the mentioned 'participation inequality' in the web 2.0 the online behavior of social entities has led to a great variability in user behavior (Durand et al. 2002). That variability can be examined through the analysis of freely available UGC. The popularity of web 2.0 and the large number of people who interact together on a regular basis in online environment provide opportunity and raise possibility of better understanding and analysis of several aspects of human social behavior (Goncalves, 2008).

It is quite clear that the size and volume of the available data is unprecedentedly large and the number of Internet users and service providers are exponentially increasing (Goker et al, 2009). It has been also realized that social entities in online environments have much more unpredictable behavior than real life (Cha et al. 2007). However, still there are uncertainty regarding the quality of the data and also about how these technologies affect business, industry and every day life. Analysis of social behaviors of online entities have been conducted from different perspectives by different scholars. For example Albert et al. (1999) and Dill et al. (2002) have focused on structure of the web from statistical perspectives. Meiss et al. (2005) and Meiss (2008) analysed the number of clicks on each hyperlinks without any attention to the number and identity of the members. Cattuto et al. (2005) worked on group of users who have been acting in online environments voluntarily (i.e. social networking sites). Goncalves (2008) looked into the structure of how users interact with web sites and its implications on web designs and online interaction cultures. Overall, all the mentioned studies conclude that the diversity and volume of the data produced on the web has made its study and analysis a challenge requiring new methods and techniques to be developed and applied.

Nowadays, UGC has brought lots of advantages to business, industry, news companies and also socializing in everyday life of people. However, the quality of the UGC has always been a subject of debate (Giles, 2005; Waters, 2007; Young, 2006). As much as its advantages and opportunities there are some inconsistencies and risks also involved in using and application of available UGC. Although it has lots of potential in revealing users' behavior and interactions there are some risks involved in analysis as well. Firstly there is a risk of making conclusion and generalizing findings that can come from 10% of users who do the 90% of the contents. Moreover, if the UGC is to be used as indicator of online behavior of social entities, reliability, consistency and quality of the UGC are also factors that need thorough consideration (Mooney et al. 2011). Overall, There is no formal quality assurance procedures for study and analysis of UGC but several different studies conducted in analysis of this issue.

Interestingly, several studies have proved that the quality of the information that can be produced through the collective intelligence of contributors compete with the information that is provided by expert official companies and professional agencies. For example the research that has been conducted on the quality of Wikipedia[9] and another encyclopedia (Britannia) revealed that the accuracy of the Wikipedians articles are relatively the same level as Britannia's (Chesney, 2006; Giles, 2005; Pirolli et al. 2009). Another study with the same aim and for different data set has been revealed that the most accurate and complete CD album and track information database has not been created by recording companies but through combining individual personal catalogues (Pachet, 2005). Likewise Google's 'page rank' has been recommended as the best way of getting more accurate results than searching the individual documents (O'Reilly, 2005a). 'Page rank' is a Google's way of calculating a page's importance. It is not the only factor in Google's search results but it is one of the important ones. Basically it is a numeric value that is assigned to a page according to the number times that other pages link to the examined page[10]. Blogging also has been introduced as a method for creating collective intelligence (O'Reilly, 2005a). The quality of the data that can be provided by group of participants can lead to what James Suriowecki (2004) referred to as 'wisdom of the crowd'. In other words if large enough people contribute towards a data creation process, the quality of the final result is as accurate as producing the same result through the traditional procedure (data collection by professionals or data production agencies).

Special type of UGC that contains spatial information about social entities who interact together in online environments is called VGI or Volunteered Geographic Information (Goodchild, 2007). Web 2.0 technologies that have brought the possibility of spatial data collection and sharing among volunteer online users require special attention (Haklay et al. 2008; Elwood 2009). Examples of sites that collect VGI are OpenStreetMap[11], Geonames[12], Google Maps Mashups, or WikiMapia[13]. In simple terms VGI is any kind of spatial data that online users provide on the web (e.g. home locations, point of interest, geo-tagged digital photographs etc.).

As mentioned before the quality and accuracy of the UGC and in line with this section VGI has attracted lots of attentions among scholars (Haklay et al. 2008; GoodChild and Hunter, 1997; Kounadi, 2009). There are several studies in the accuracy and validity of the VGI. Since the spatial data in this case is provided by volunteered posters it can be argued that the data collection is not necessarily conducted with the standard data collection measures and also the providers of VGI are not necessarily expert geographical data collectors. In other words no assumption can be made about the background knowledge or data collection ability of the volunteers. There have been several studies conducted to compare the accuracy of the spatial data that users upload on the web with the spatial data provided by mapping agencies. Research by (Haklay, 2010; Ather, 2009;) that focused on the evaluation of the road network and compared it to other sources of information, have demonstrated that, in terms of positional accuracy, the quality of open street map data is comparable to traditional geographical data sets that are maintained by national mapping agencies and commercial providers. Goodchild and Hunter (1997) developed a model that

[9] www.wikipedia.org

[10] Google PageRank Explained and How to Make Most of it by Phil Craven, WebWorkshop available at http://www.webworkshop.net/pagerank.html Accessed 16th Feb 2012

[11] openstreetmap.cp.uk

[12] geonames.co.uk

[13] wikimapia.com

provides an estimation of the overlap between a reference data set and a test data set. In all these comparisons, the national mapping agency was used as reference data set and open street map as the test data set. The results show overlap of about 80% in most cases but the values range from 100% down to 50% and below (Haklay, 2008). Moreover, there are some rules that imply the quality of the VGI based on some measures i.e. 'Linus law' (Raymond, 2001). They indicate that as the number of the contributors increases so dose the quality of the data (Haklay, 2010). Therefore, the large volume of spatial data that people upload voluntarily on the web has made a valuable collection of data that have the potential for revealing interesting patterns about online behavior of social entities that has not been examined to its full potential yet.

2.3 Social Networks

Traditionally a social network was made up of a group of people and their social relations, but today the large number of people interacting together via computers is also creating social networks, although in a quite different fashion. Social network sites are rich sources of information regarding people, their interactions and locations. This kind of information has been difficult, time consuming and expensive to obtain previously. To date, social network sites have brought an abundance of relational, social and locational information recorded automatically in digital format. This wealth of information enables researchers to study and examine fields and topics that were not feasible previously.

The historian Crosby (1997) has identified two crucial factors in the development of modern science: visualization and measurement (Freeman 2000). Interestingly SNA[14] has borrowed both of the factors. In the case of measurement it supports statistical analysis and mathematical theories and to satisfy the latter factor it uses graph theories to visualise the social entities and relations via the node link graphs. This might be the reason for the recent popularity of SNA in analysis of online communities e.g. (Flickr, Facebook, Twitter) where people develop a personal profile, make friends, interact and upload digital documents.

SNA is the analysis of network data in a social context (Scott, 1991). It is a set of tools, techniques and theories for the study and analysis of hidden patterns in social communities (Wasserman and Fuast, 1994). It has become identified as an interdisciplinary analysis method after the 1970s (Freeman, 2000). SNA techniques have been used in different domains e.g. organizational behavior and inter-organizational relations (Tichy et al. 1979), the spread of contagious diseases (Rapoport, 1958), mental health (Kawachi and Berkman, 2001) and the diffusion of information (Yamaguichi and Buskens, 1999). SNA has borrowed methods from sociology, psychology, anthropology, mathematics and statistics and has developed interdisciplinary techniques for collecting data, statistical analysis and visual representation of social networks (Scott, 1991; Wasserman and Faust, 1994). Traditionally, much of the literature on the social theories considered the individuals without their social context (Knoke and Kuklinski, 1982), whereas one of the distinguishing features of the SNA has been identified as the ability to deal with relational data (Otte and Rousseau, 2002). Relational data as defined by

[14] Social Network Analysis

Wasserman & Faust (1994) is the interacting relation and its properties among a number of social entities. SNA techniques have the ability to depict the presence, absence or strength of relationships among actors (Scott, 1991). Therefore, it has been suggested that analysis of relational data should be carried out with SNA techniques, and applying other techniques to relational data is of little benefit (Scott et al. 2005).

While some believe that social network analyses should be considered mainly as formal statistical and mathematical techniques (Hanneman and Riddle, 2005) there are examples in the literature demonstrating ideas, which imply that quantitative analysis is not the main aspect of SNA. Kleinberg (1999) argues that SNA is a strategy for exploring social patterns and structures. Wassermann and Faust (1994) focus on the relationship among actors as the most important aspect. In the same line, the main goal of SNA has been pointed out by Nooy et al. (2005) as to discover and analyze social relations in a social group. Likewise, Scott (1991) puts the emphasis on relationships between actors rather than quantitative analysis and statistical attributes. He describes it as the analysis of relational data in a social context to depict structure of a group. From another perspective, Kilkenny and Nalbarte (2000) focused on the ability of SNA to use relational information to study and test hypotheses. Thompson (2003) has marked the distinguishing factor of SNA as the ability to use sophisticated data and statistical techniques for discovering sets of patterned relations in the network. Newman's definition (2001a) is concentrated on the visualization of the social relations. In some literature it has even been introduced solely as a technique to provide vocabulary to identify and measure the network properties (Monge and Contractor, 2003).

Overall, despite all the different and sometimes conflicting definitions, SNA has been accepted as a systematic approach in identifying the nature and structure of social networks. Accordingly, sometimes it has been defined as 'structural analysis' (Wellman and Berkowitz, 1988; Scott, 1991)which comprises the understanding of individuals (nodes or vertices), social relations (ties, edges and arcs) and social and quantitative properties of the network (Borgatti, 2002).

2.4 Social Network and Visualization

An interesting connection has been identified between the visualization and social networks (Freeman, 2000). Visualization and social networks can be replaced by 'structural analysis' and 'structural concepts' respectively. Visualization plays a significant role in the analysis of social networks. It helps in investigating and illustrating patterns laid in relational data (Freeman, 2000). According to Scott (1991), Moreno (1934) was the first to use visual images to explore the substantial features of social patterns among members of a group. Focusing on analysis of social relations by means of diagrams was the central point of consideration in his approach. Today, Moreno's idea is still in use and imagery is a popular tool for analysis of social networks (Freeman, 1988). This might be defined by the fact that visualization facilitates the exploration of relational data upon which the networks are built (Herman et al. 2000). Visualized relational data provide new views for analyzing the properties of social networks and may reveal insights and communicate findings to others through visual means (Freeman, 1988). Traditionally graphical presentations of social relations have been represented using sociograms (Nooy et al. 2005). Accordingly, sociometrists have laid the basis of social network visualization (Scott, 1991). The relational data visualized in a sociogram can reveal the interaction patterns of a social network (Wasserman and Faust, 1994).

Since visualization provides opportunities for explanatory analysis and testing initial hypotheses (Thomas and Cook, 2005), it has recently become a popular research topic in the study of social entities and relations. There are real potential opportunities to take advantage of network visualization methods (Herman et al. 2000) and graph drawing techniques (Battista et al. 1998) to improve social network analysis methods. Many efforts have been devoted to conduct statistical techniques and hypothesis testing by using the social network analysis methods (Wetherell et al. 1994; Kleinberg, 1999; Borgman, 2000 and Lynch, 1998). However, few attempts have been made to improve the traditional node link picture of the social networks (Figure 2.1). Some researchers have currently become involved in attempts to improve SNA methods and theories by applying graph theory and visualization methods. Efforts have been made to improve the layout of the sociograms and add methods for exploratory analysis (Viegas and Donath, 2004; Bender-deMoll and McFarland, 2006; Henry and Fekete, 2006; Henry et al. 2008). However, the graph metaphor is inadequate for taking advantage of visualization methods and to visualize relations analytically in the social network context (Figures 2.1-2). Therefore, it has been argued that if graphs are to be considered as a methodological tool, there are certain factors that have not yet been examined (Bender-deMoll and McFarland, 2006).

According to Barabasi and Albert (1999) the commonly found structure of social relations is a network with most nodes of least degree and a few of higher degree (Figure 2.1). Consequently, even the most effective force directed approaches (e.g. Frutcherman and Reingold, 1991; or Kamada and Kawai,1989) result in cluttered layouts when applied to large relational data sets (Jia et al. 2008). Figure (2.2) demonstrates the inability of graph layout in demonstrating social relations. As can be seen the layout is too complicated to discern any patterns. Therefore, graphical presentation of social networks has raised serious challenges in the network visualization domain. McGrath et al. (1997) have proved scientifically that the position of nodes (node duplication or overlapping) significantly affects the understanding of the communities. Purchase (1994) and Ware et al. (2002) worked on the edge layout and proposed a solution for effortlessly finding the shortest path in the networks. Another recommended solution is to split the network into separate parts and use tree layout for each node to reduce the crossing edges and overlapping nodes (Lee, 2006). Although useful for finding properties of individual nodes, understanding the overall view of the community is difficult. Henry and Fekete (2006) applied two different layouts to provide an explanatory analysis. However, switching between node link graph and adjacency matrices has made the application inefficient and confusing. There are some efforts in clustering of nodes to reveal the network structures (Jain, 1999) and in different hierarchies simultaneously (Eades and Feng, 1996). Hence, the developed methods were found to be effective in reducing the network complexity in just small world networks (Watts and Strogatz, 1998).

Recent solutions in the field have aimed to improve the graph readability (Henry et al. 2007) by using adjacency matrices instead of node link graphs (Figure 2.3). Their recent proposal in reducing the crossing edges of the matrix layout is to duplicate the nodes shared between different clusters (Henry et al. 2008). As is shown in Figure (2.3) each cluster in the network has an associated matrix in NodeTrix layout. Each member can appear in as many matrices as required according to the number of their clusters. Geometry based edge clustering techniques (Cui et al. 2008) have also been applied to reduce the complexity of visualization of social networks in graph format (Figure 2.4).

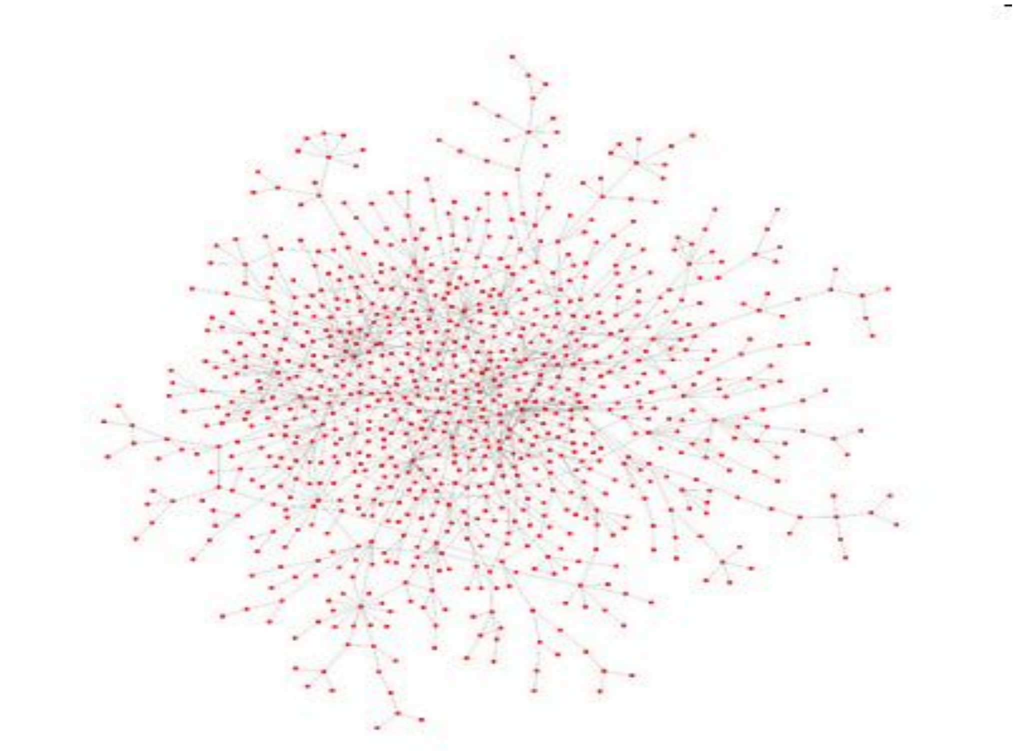


Figure 2.1 Traditional social graphs in node-link presentation

All in all, visualization of social networks, the graph readability issues and improvement in the layout algorithm for easy exploration and analysis of relational data have recently attracted the attention of many researchers in the field of social sciences, visualization, statistics and geography. All the existing solutions in the field frequently alter the geometric relations present in the real world in order to emphasize the connectivity and overall view of the network. Whilst a node position has considerable potential for carrying information regarding network pattern and structure no spatial information is usually encoded. Therefore, this thesis attempts to take the advantage of the wasted geographies in graph presentation of social network in order to develop techniques to discover and analyze spatio-social relations in a spatially structured social group.

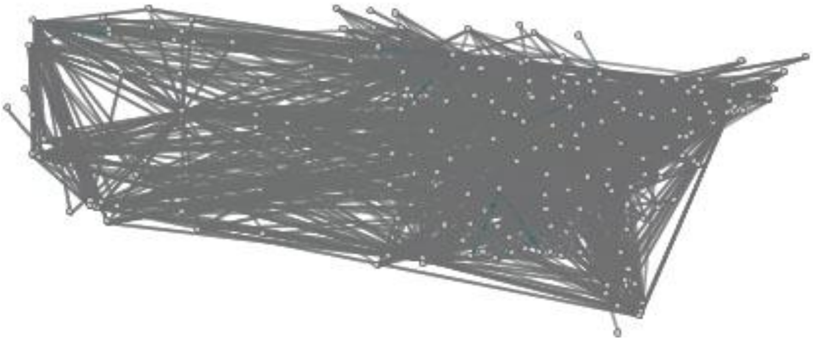


Figure 2.2 Graph layout of social relations (Cui et al. 2008)

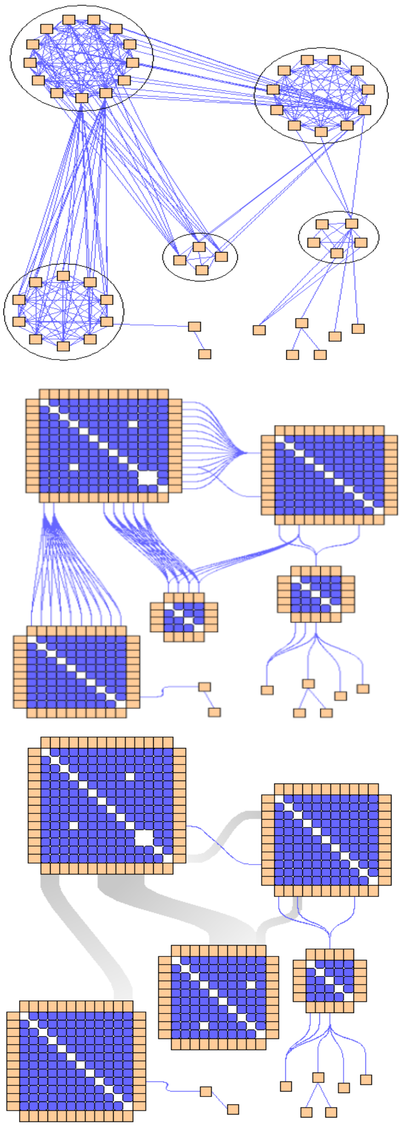


Figure 2.3 Visualization of social graphs by NodeTrix method (adapted from Henry et al. 2007)

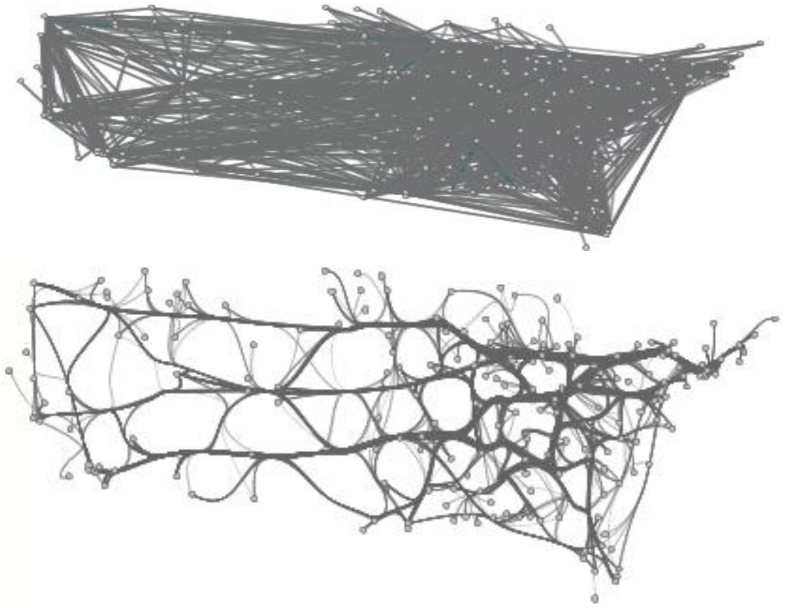


Figure 2.4 Cluttered graph layout and the simplified version by edge clustering techniques (adapted from Cui et al. 2008)

2.5 Social Network and Geography

Considering the fact that social networks are made up of social entities, and social entities are associated with multiple physical locations, one could argue that locational information is an important property of social networks that has not been extensively examined yet (Wellman, 1996; Katz 1993). The majority of current works on social networks have been conducted without including geographical data. This wasted information in geo social networks could potentially reveal hidden patterns and structural properties of communities. While social entities and their locations are indispensable parts of social networks, in addition to all conventional attributes examined in social networks before, this section looks into the geography of social networks in the literature.

It has been a long time since social science researchers proved that distance matters for social relations (Escher, 2007) and social relations are not independent of geography (Wellman, 1996). The majority of social ties are between neighbours in geographical closeness of each other (Cummings et al. 2006) and people have the highest interactions with their neighbours (Mok et al. 2007). However, in the computer-based society many people interact together via computers and therefore the effects of traditional distance on social interactions of people have become a subject of debate (Cairncross, 1997). For example, unlike what has been believed before, Beck (2002) in his work on online friendship concluded that the physical distance does not play a key role in friendship anymore.

A social network structure in graph format that assigns a two-dimensional position to each node also suggests some sort of geography (Viegas and Donath, 2004). Each node can be associated with an (x,y) value on the screen. This kind of network visualization in the best situation could play a role similar to cartography with additional challenges (Bender-deMoll and McFraland, 2006). Geography principles and cartography techniques allow real potential to be applied to visualization of non geographic information (Skupin and Fabrikant,

2003) and it is necessary to go beyond the graph drawing techniques for a novel network visualization (Viegas and Donath, 2004). As VanWey et al. (2005) describe there are standard rules for depiction of locational data. There are also standards for presentation of relational data (e.g. Wasserman and Faust, 1994; Scott, 1991). However, little has been done on mapping the geographic distribution of relational data with multiple spatial features (Gutmann and Stern, 2007). According to the existing limited attempts in relating geography to social networks, the effects of geographical data in online communities is a controversial topic still under analysis. Geography is an important property of social entities and applying that information in social networks could bridge the gap of not having enough research in that respect. The following are examples of current research on geography and social networks.

The effects of geographical data on social networks have been studied by repeating the Milgram experiment of 'six degrees of separation' with extra information about the location and occupation of individuals (Liben-Nowell et al. 2005). Six degrees of separation is referred to an interesting experiment that sociologist Stanely Milgram (1967) conducted. He asked the 96 participants to send the letters to someone they think can make each letter closer to its target. Interestingly it was found that the letters that reached their destinations had passed average of 6 people. However, Milgram's experiment was small and restricted to one country and since then several sociologists have questioned its validity[15]. Repeating this experiment by (Liben-Nowell et al. 2005) indicates that one third of friendships are independent of geography and the probability of being a friend of a particular person is proportional to the inverse of the number of closer friends. Katz (1993) investigated the impact of distance on scientific collaboration. His findings indicate that the collaboration reduces exponentially with increased geographic distance among scholars. Escher (2008) claims that the examined online social networks (Myspace and Facebook) have a centrality around the geographic proximity of the users. Wellman (1996) takes a different approach by emphasizing the number of contacts (weight of ties) instead of the existence of ties solely. He concludes that social relations are not independent of 'geography' and claims that locality is a factor behind the online interactions.

According to the existing literature there are few studies examining the possibility of geographic proximity as an idea behind online communities. Those that exist apply the naïve geography of confining each social entity in a bounding box of a city. Whilst a node's position has considerable potential for carrying information regarding complex geography of social interactions, no spatial information is usually encoded. As such, the existing geo social network visualizations are incapable of including locational information in a network presentation. All the works mentioned above have altered the geometry and physical location of the nodes in order to emphasize integration and connectivity of interactions.

2.6 Social Networks and Ambiguity

As mentioned briefly before, this research aims to work with spatial information that people provide voluntarily on the web. Since in most of the sites there is no

[15] Peter Sheridan Dodds and his colleagues at Columbia University and study co-author Duncan J. Watts 2003.

instruction or limitation to the format, size, length and accuracy of the information uploaded by the users, dealing with ambiguous location names is expected to be part of the methods required for this study. Therefore, this section briefly discusses different types of spatial information that have been introduced and studied in the literature. It also reviews the potential ambiguities that might occur in spatial information[16]. Overall there are two types of spatial information retrievable from existing social network sites:

- Formal format (latitude, longitude): that imposes some must meet criteria for spatial entries lat, lon (geo-tagged photos, Wikimapia[17] entries)
- Free format (place name in natural language of people) that does not have any restriction on type, format, and length etc. of the uploaded data.

The first type come with no ambiguity and can be associated with a specific place on the map, but the second category comes in different formats and different levels of detail that makes the geo coding a challenging task with several ambiguities and uncertainties that remain to be dealt with. In general, vague terms are not exclusive to geographic locations. Even the simple terms that are used in every day life such as 'hot' and 'cold' have been considered as poorly defined terms (Montello et al. 2003; Fisher, 2007). This kind of ambiguity has been attributed to the lack of a universally accepted temperature for definition of hot and cold (Fisher, 2007). In a more relevant concept, imprecise location names have been considered in different fields of science: linguistic, machine understanding and artificial Intelligence, information retrieval and knowledge management (Rauch et al. 2003). The existing solutions either describe the place according to their structure or context in which it is examined or rely on glossaries and gazetteers. One name per discourse (Gale et al, 1992), Named entity recognition (NER) (Malouf, 2002) and Natural Language Processing (NLP) (Smith and Crane, 2001) all rely on the context, structure and other similar words in the text in order to disambiguate the vague terms. Russell-Rose and Stevenson (2009) introduce four ambiguities that make the NLP as a challenge.

1. One word can have more than one meaning (bat)
2. Two phrases with the same terms but different order can have different meanings. New York and York New Festivals
3. Natural language allows the same idea to be expressed in several different ways. For example: Every day I have two coffees. I satisfied my caffeine addiction couple of times a day.
4. The meaning of a sentence is not simply reflected by the terms it contains. For example: 'A cricket bat is not a nocturnal mammal or any other types of animal' although has the terms animal and nocturnal mammal it is used as a cricket equipment.

Considering the nature of the spatial data in this thesis, it has been assumed that all words in a phrase that a person use to refer to their home location directly or indirectly can be associated to a place in the world. Therefore, the disambiguation here is relatively simpler than a natural language document discussing in different subjects. The subject of the phrases here has been narrowed down to a place name somewhere on the earth. However, the

[16] More detailed information on this topic is covered in chapter 5.

[17] www.wikimapia.org

methods and techniques in NLP are expected to be beneficial for developing a disambiguation algorithm for FHLI. As a result of the existing studies, the most common methods of dealing with imprecise definitions is to replace the vague terms with the closest precisely defined term, or completely ignore them (Montello et al. 2003). In addition to the mentioned studies, the vagueness in geographic terms has been tackled through the existing general name disambiguation and word sense disambiguation algorithms (Li et al. 2002; Li et al. 2003; Bilhaut et al. 2003; Hirst, 1987; McRoy, 1992; Ng and Lee, 1996).

As has been recognized by other researchers in the field, the unstructured data (like user generated data on the web) contained significantly more information than well-edited texts such as news articles (Rauch et al. 2003). The web has been found as a valuable source of geographic information and our ability to make sense of it is far less than our tools and techniques for producing and collecting data of that kind. Therefore, developing methods for disambiguating vague geographic terms in online unstructured data has the potential to reveal a considerable amount of information that is otherwise difficult to understand (Rauch et al. 2003). While the spatial terms often used in GIS (Geographic Information Systems) are officially assigned sharp boundaries, the freely available spatial information on the web is inherently imprecise and fuzzy (Silva et al. 2005). As can be predicted, this study needs to intelligently analyse the FHLI and disambiguate the location names, therefore, here several relevant steps of the NLP techniques (Russell-Rose and Stevenson 2009) are briefly mentioned:

Named Entity Recognition: it is the process that the words are assigned to a predefined classes of terms i.e. people, company, place, date, etc.

Information Extraction: it is an NLP technology that identifies pieces of valid information from a document i.e. movements of company executives or riot in a specific place. The process in the second step fit the information into object oriented structured templates for further reference.

Word Net (Fellbaum, 1998): is based on the sets of synonyms according to human mental lexicon. Words that have more than one meaning are associated to different categories. Each category is associated with other with sets of relations e.g. is a kind of.

Word Sense Disambiguation: is a process developed in NLP to automatically identify meanings of words in text. It is specifically beneficial in cases where words have more than one meaning (known as polysemy).

Evaluation: entails the evaluation of the results produced by machine against human judgments. This approach although is the most common methods within NLP it brings some sense of complications. There are ambiguities in agreeing on correct standard annotations and uncertainties in comparing the annotations.

There is growing body of work, during the last decades, disambiguating vague, indeterminate location information (Jones et al. 2008; Jacquez et al. 2000; Waters and Evans, 2003; Montello et al. 2003). McCurely (2001) studied all the information obtainable from a webpage that can be associated with a geographic place e.g. postcode, URL, IP address. His study was heavily dependent on the gazetteer and directories, which are not available and reliable for all places. Accordingly, his method has been found inefficient, inconsistent and even sometimes impractical for locations that are not documented in official glossaries. Montello et al. (2003) focused on how vague regions are understood among people. Their study analyzed the regions, which inherently have fuzzy boundaries. They attribute this fuzziness to the lack of any universally accepted geographic boundaries for regions. They draw boundaries for 'downtown of Santa Barbara' according to people's mental map and reached a conclusion of how people map the vague regions in their minds. With quite the same aim, Purves et al. (2008) focused on describing the city of Zurich through analysis of VGI (Volunteered Geographic Information, Goodchild 2007a) and an understanding of people's perception of Zurich city centre.

According to Smith and Mann (2003), there are three categories of ambiguity in geo-referencing:

- Referent ambiguity, the given name can be associated with more than one place.
- Reference ambiguity, the same place can be referred to with several names.
- Referent class ambiguity, place name has a non-geographic use as well.

In the majority of cases a place name on its own without any other information can refer to multiple locations. This ambiguity has been called as 'referent ambiguity' and can occur for places in different countries (e.g. Sheffield in UK and USA) or within the same country. For example Cambridge in the UK appears four times in the OS gazetteer list: Scottish Borders, Leeds, Cambridgeshire, and Gloucestershire (Arampatzis et al. 2006). There are also plenty of places around the world that people refer to with different names. For instance America, U of S, States, United States of America, they all refer to one place. Accordingly, the referent and reference ambiguities are very likely to occur in spatial information of social entities and therefore are expected to be dealt with during this research. However, the 'referent class ambiguity' is less expected to be the case in this work. Unlike text mining, in this research there is no need to recognize the place names and differentiate them from other names. This assumption is based on the fact that this study aims to work with the spatial information of social entities on the ground that regardless of the ambiguity and uncertainty the collected data are merely place names referring to a location on the earth (e.g. home location of posters, point of interest of users, etc).

As mentioned earlier the web is a rich source of geographic information that is being constantly updated by volunteer users. It is conventional for web users to describe the place names with natural language and vernacular geographic terms (Jones et al. 2008). Although these data have the potential to reveal large amounts of information about social entities and their geographic locations, the existing gazetteers are of limited use for searching vague imprecise place names (Popescu et al. 2008). In the majority of cases the given name described in natural language does not match with the name referring to the same place in official gazetteers (Jones et al. 2008) and this is a serious concern in mapping spatial information of social entities on the web.

In order to show the extent of the ambiguity in place names it is interesting to mention that for example there are 18 cities named Jerusalem and 24 cities named Paris. Accordingly, Smith and Crane (2002) concluded that for each name that at least has one match in the gazetteer, 92% are associated with more than one place on the earth. Likewise, Amitay et al. (2004) found that 37% of the geographic names mentioned on the web can be associated with more than one place on the earth. Their conclusion indicates that each location name on the web can be associated with at least two different places. Li et al. (2003) have conducted research assessing the performance of the Tipster Gazetteer[18]. They found that the Tipster Gazetteer does not assign default values in the majority of cases and 30,711 location names out of the 171,039 entries are ambiguous. Waters and Evans (2003) argued that the fuzzy vernacular geographical terms have not been taken into serious consideration for updating the gazetteers for two reasons:

- Difficulty in retrieving and storing the data systematically
- Difficulty in interpretation and use

Although the publicly available spatial information on the web is unstructured, vernacular and fuzzy, the majority of the gazetteers are based on the official and administrative geography developed by national mapping agencies (Hill et al. 1999). Consequently, in most cases the given vague vernacular geographic name accessed from the web does not reference in any gazetteer at all (e.g. Southern England, Midland). To wrap up we can conclude that in disambiguating the online spatial information three factors play important roles:

- The freely available online spatial information contains geographic information, described in the natural language of web users, which does not necessarily match with the official geographic names and boundaries.
- Gazetteers do not include the vague terms that have been used ubiquitously to describe place names by people in everyday life.
- The existing successfully tested disambiguation methods are only applicable to structured databases and/or specific areas.

Subsequently, unless being specifically precise about a place name and its locational hierarchy, there are strong chances of ambiguity and uncertainty in geo coding the available spatial information that need to be properly handled before analysis and visualization in this research. Overall, the results of the previous work on analysis of location names indicate that there is an inherent complexity in spatial information described and used in the natural language of people. All these proposed solutions are applicable to structured databases and well-edited texts (Jones et al. 2008). Consequently, they are of limited use for the tremendous amount of unstructured spatial data that has become freely available through the web, which is the main data set for this study. Accordingly, the difficult and fuzzy nature of the available spatial information on the web that can come in the natural language of users requires specific adaption strategy to the existing GIR (Geographic Information Retrieval), web mining and geo-ontology methods and techniques.

2.7 Social Network and Privacy

Since this research aims to study the locational uncertainties in freely available spatial information on the web, it involves detailed study and analysis of spatial information associated with individuals. This locational data could be the home location, point of interest, current locations, geo located documents, etc.

[18] http://crl.nmsu.edu/cgi-bin/Tools/CLR/clrcat

Although the locational information in social network sites is submitted voluntarily, it might involve implicit privacy concerns. Therefore, it is expected that the knowledge and awareness of the online and locational privacy have potential advantages in data set selection and data analysis phases of this research.

2.7.1 Online Privacy

Since the advent of Internet and online interactions, a significant amount of personal information has been recorded automatically in digital format that has enabled researchers to study and examine fields and topics that were not possible previously. This unprecedented wealth of information embraces some emergent and unexpected privacy issues. The electronic data available and accessible through computers, networks and databases has resulted in a considerable amount of privacy concerns (Berry and Linoff, 1997).

Although the ultimate goal of web 2.0 has been described as to develop a collective intelligence and global brain (O'Reilly, 2005a) and while it's been proved that society can get benefit from UGC there are also some concern regarding privacy and exploitation of online social entities.

There are some debates in the literature about the possibility of online privacy. Sanchez (2007) argues that online privacy at its best could offer 'selective anonymity' and not control over personal information. In his opinion anonymity and control are two impossible factors for online environments. This argument is supported by the fact that privacy is not consistent with 'transferability' and 'malleability' of digital information. Laws, ISPs, ethics and technology are all found to be unreliable and inadequate for devising an effective online privacy policy (Louden, 1996; Sanchez, 2007). However, the majority of the scholars (e.g. Adams, 2000; Sanchez, 2007; Sholtz, 2001) agree that online privacy is needed and social entities sharing personal information on the web are at risk. Consequently Scott McNealy, (CEO of Sun Microsystems 1999) declared that online members have zero privacy. As for this study, Armstrong and Ruggles (2005) foresee that with advances in geo spatial technologies the term 'zero privacy' will also be applicable to locational privacy in the very near future.

2.7.2 Locational (Spatial or Geo) Privacy[19]

The literature points out that geo spatial technologies have noticeably changed the way privacy is being perceived. There are certain discussions on the level of sensitivity of location for online or offline interactions (Raper 2002; Onsrud et al. 1994). A survey at personalisation.org indicated that more than half of the population is not willing to release their location to service providers (Raper, 2002). However, in virtual environments statistics indicate that people are less sensitive about their locations. According to Escher (2008), only one out of five profiles on popular social networking sites (i.e. Myspace and Facebook) lacks the locational information at city level.

Today 'Geography' has become an indispensable dimension of online social network sites, and the information from geography, linked and unified with the digital information of online social networking sites, could reveal considerable

[19] Defined as human right to respect their private, their homes and their correspondence (Raper 2002) and or control over geographic information (Sanchez 2007)

detail about individuals (Morgan, 2004). As such, Armstrong and Ruggles (2005) have named geography as ' unifying glue ' for online social identities. The invasive potential of geo spatial technologies has a significant threat for online socialisers in spatial domains (Kwan et al. 2004). Linking the social spatial information to a specific address on the earth could be a key issue in relevant studies (Morgan, 2004). Relating social and spatial information to a geographical coordinate (from GIS) leads to numerous personal information being available, and consequently intrusion of privacy (Gutmann and Stern, 2007; Morgan, 2004). In a computer-based society, individuals' personal spatial information can be stored, collected and disseminated in online environments with no concern for spatial privacy. Availability of identifiable locational information of individuals in complex databases is felt to be a major threat to individual privacy (Morgan, 2004). Since the majority of interactions (on/off line) generally carry some sort of spatial attribute (Morgan, 2004), locational privacy has been considered as the most extensive privacy concern, including having the highest possibility of intrusion. The recent advances in geo spatial technologies, as well as the availability of social spatial data, have created significant uncertainties about online confidentiality (Gutmann and Stern, 2007).

Geo-privacy as mentioned by Raper (2002) is a new challenge in e-social science and geographic information systems. Locational information and spatial privacy have been mentioned as unique and incomparable with other kinds of personal data and privacy issues (Kwan et al. 2004). In supporting that idea, Caslon (2004) argues that this uniqueness comes from the fact that traditionally privacy issues were concerned about whom, and not where and one's whereabouts. It is believed that traditional privacy techniques for anonymisation and control over personal information are not applicable to spatial digital data without noticeably changing and or distorting the spatial relationships among entities (AbdelMalik et al. 2008).

2.7.3 Geo-privacy and Spatio-Social Data set

According to the above literature, the following are key issues regarding geo-privacy and spatio-social data set:

- Data quality: there might be some data missing in location of members. While members might submit significant amounts of locational information e.g. point of interest, geo-tagged digital information, etc., not necessarily all of them are willing to release their current location or their home location.
- Disclosure: there is some concern that the detailed study of the available spatial information on the web, regarding the geographic distribution, might lead to the possibility of disclosure. If, for example, a member has a geo-tagged collection of photos fairly concentrated around limited places, this could lead to the exposure of the geographic location of that member, although not stated explicitly.

In this context, it is essential to consider the availability of the locational information in any potential data set for this study. This is due to the fact that data quality and disclosure concerns (as explained above) may limit the availability and accuracy of the available data. Consequently one of the main data set requirements of this study relates to geo spatial data accuracy and availability. Privacy issues that might occur as a consequence of developing new knowledge through the study, analysis and synthesis of spatio-social information will also be considered during the course of this research.

2.8 Summary

This chapter provided a thorough review of the literature associated with the key subject matter, in line with the aims and objectives of this study. It explained the social network concept and its relation to visualization, geography, ambiguity and privacy. The next two chapters consist of information on data set for this study. Chapter three explains the specific requirements needed for an appropriate data set that can be used in answering the developed research questions. According to the identified requirements potential sites have been examined against the assessment criteria. The fourth chapter is about the selected site and its characteristics and specifications. In addition, it describes data sampling strategy and final selected data set for further study and analysis.

3 Data requirements

3.1 Introduction

As discussed in the context chapter, social networks are made up of social entities and social entities are associated with several locations. Therefore, analysis of spatial information of this kind is involved in analysis of spatial social network data. Consequently, the data set of this study carries the features of relational data as well as spatial data, bearing inherent complexity and requiring an interdisciplinary approach to their consideration.

According to the volume of the research conducted in regards to data set of this study, two chapters are assigned for covering different aspects of appropriate data set. This chapter reviews the characteristics of spatio-social data. It describes the data set requirements and conducts thorough examinations of potential spatio-social sites. Considering the geographic richness, ease of access and spatial and social elements, potential sites are examined and the final decision is justified in terms of the proposed criteria. The next chapter introduces the selected spatio-social data for this study and studies its features and characteristics. Data sampling strategy is described and finally the appropriate sampled data sets are collected.

3.2 Social Network Data

The academic literature identifies a number of key characteristics of the relational data used for social network analysis (Wasserman and Faust, 1994; Scott, 1991):

- Social networks are inherently difficult to understand, because nodes are social entities and connectivity relies on the definition of relations, which are not something tangible as in conventional networks, e.g. wires in computer networks (Scott, 1991).
- Relational data are quite different from the conventional tabulated data sets and that makes them a challenge in terms of data collections, study and analysis (Borgatti and Foster, 2003; Wasserman and Faust ,1994).
- A social network is an information network without any geographic (metric or real) distance concept (Porta et al. 2006). It is the author's definition that specifies how close or far from each other nodes should be considered.
- In social networks nodes are as important as ties. Therefore, integration and connectivity play a key role (White and Harary, 2001).
- Sampling in social networks (relational data) is far too complicated compared with other types of data. Connectivity and clustering in relational data make the data set more vulnerable against bias (Kossinets, 2008).
- In terms of visualization, social networks often contain few nodes with high degree and many of low degrees. Therefore, even the most efficient layout algorithm results in networks with many crossing edges even in medium sized networks (Henry, et al., 2008).

3.3 Spatial Relational/Social Data set

The spatial social data set in this study is defined as a data set with characteristics of relational as well as locational data. This kind of data set has the challenging properties of being large and multi variate with a spatial structure, while likely to contain patterns relating geography to online interactions. Such large and complex information, although freely available on the web, is inherently difficult to browse and too complicated to discern using current techniques.

3.4 Data set Requirements

According to the aims, objectives and research questions discussed in the first chapter (sections 1.2-5) there are certain concerns regarding an appropriate data set for this study.

Firstly, in order to overcome the challenge of visualizing large relational data in a spatial social network concept, and to get a representative sample of online socialisers, it was essential for the data set to contain a large number of members.

Secondly, in order to reinforce generalization of inferences, local sites (e.g. limited to a city or a country) were found to be inappropriate. Those sites and services which were limited to specific organizations or people, or those that were only available in a city or country, are likely to mirror the behavior of a special group of people affected by the same rules, facilities and environment. The findings from that kind of data set are vulnerable against possible generalization of any hypotheses derived. Therefore, the appropriate data set of this study should be freely available and accessible to the majority of people around the world.

In the third place this research aims to study and visualize spatially variable attributes of entities in online networks. Therefore, the availability of variable locational information about individuals was essential. As a result, a suitable data set for this study should make the spatial information of the members (e.g. hometown, point of interest (POI), work place, geo located digital photographs, etc.) publicly available.

Bearing in mind that the data set of this study has been defined in the previous section as a spatial social data set, the fourth requirement is the availability of relational data (e.g. public friendship network, online interactions, discussion forums).

And the last requirement that the data set here should be able to satisfy is to provide an Application Programming Interface (API). This was based on the fact that even a large amount of locational information, without a formal way of retrieving, can be of a limited use for this study.

In accordance with the requirements discussed above and the aims and objectives of the study (chapter 1: sections 1.2-5) the data set requirements are summarised as:

1. Large number of members (REQ1)
2. Geographical distribution of members (scattered around the world, more than one country at least) (REQ2)
3. Availability of members' locational information (REQ3)
4. Availability of multiple spatial information in addition to members' home locations (e.g. POI, work place, geo located digital photos, etc.) (REQ4)
5. Availability of members' friendship network (REQ5)
6. Availability of usable API through which data can be accessed and visualization achieved (REQ6)

According to the criteria mentioned above, an appropriate site for this study is an online social network with the features and capabilities for developing personal profiles, online friends and online communication, while supporting locational information as well. In other words, a geo social network highly equipped with online friendship and interaction features.

3.5 Potential Sites

Among the 24 recent geo social sites listed by Raper in his blog[20], the following five sites were selected as potential sources due to their collection and provision of data with relational and spatial elements.

3.5.1 Yelp (http://www.yelp.com)

Yelp is an online social network site where members would leave their own personal reviews and experiences about local businesses, places, events, etc. Therefore, visitors with a particular interest could search the site according to a geographic location. As with other sites, *Yelpers* have a personal profile and a network of friends. *Yelpers* join the site according to the category of the cities where they live. Therefore, location of members (at city level) is publicly available.

REQ1. According to the communication director, Yelp had 11 million unique visitors and more than 3 million reviews in June 2008 (Ichinose, 15/07/08, PERS.COM).

REQ2. In terms of geographic distribution of members, although there were plans for extending the service overseas, at the time of this research only Americans were *Yelpers*. The service was not available in any other place e.g. London, Toronto (Ichinose, 15/07/08, PERS.COM).

REQ3. Details of the home location of members were available.

REQ4. Each place, event or service review submission to the site was attached with spatial information.

REQ5. Each member had a friendship network, which was publicly available.

[20] http://isblogs.soi.city.ac.uk/staff/raper/ Accessed on 17/06/08 Appendix 8.

REQ6. At the time (July 2007) the site did not provide users with API for systematic information retrieval.

3.5.2 TwitterVision (http://twittervision.com)

Twitter is a social networking site in which members communicate together via mini blogging (short messages of less than 140 characters). That is, they update their status by saying what they are doing. The updates are accessible to the other members through the site or on their mobile phones. The site mashes up Google maps and micro bloggings. It connects friends, families and colleagues through real time updated short messages. The site has been introduced as a remedy for information overload and a telegram of web 2.0. (Nicholas Car, 2008). *TwitterVision* is a site that demonstrates the real time geographic distribution of the posts to *Twitter*. All members of *Twitter* are automatically registered with *Twittervision*. As long as they have a picture and a location, their messages are shown on the map.

REQ1. According to the *Twitter* directory, the site had the 944,773 members at June 2008 and the number was constantly increasing.

REQ2. The *Twittervision* map demonstrates the fact that there is no limitation in the availability of the service in a specific place.

REQ3. Since messages were laid on the map according to the members' locations, the geographical proximity of members was publicly available.

REQ4. Apart from friendship networks and members' locations *TwitterVision* did not support any other spatial information.

REQ5. Each member had followers and those they followed. Therefore, the friendship network was clearly accessible.

REQ6. The site contained specially designed API for providing accurate current locational information of users.

3.5.3 Flickr (http://www.Flickr.com/)

Flickr is Yahoo's photo sharing site. It is an online photographic community where people upload their own pictures and can view others'. *Flickr* members have the option of geo-tagging their pictures and they can also attach pictures on their actual place on the geographic map. Accordingly it can be considered as a minimal geo social network.

REQ1. The number of members and their geographical distribution were not publicly accessible and *Flickr* customer care was not willing to release that kind of information (17/07/08, PERS.COM). But according to the high number of submitted photos in May 2008 (more than 2.5 million), and the variety of the locations of the geo-tagged photos (USA, Europe, Asia, Africa), it was clear that it contained a high number of members distributed around the world.

REQ2. *Flickr* members were not limited to a specific place (city or country). The *Flickr* service was available all around the world.

REQ3. Members' locational information was partly missing. Not all members provided information about their home locations in the profiles. In other words the home location field of the user profile is not mandatory.

REQ4. According to Dan Catt, a member of *Flickr* geographic works, at Where 2.0. Conference (2008), the *Flickr* geo database had more than 68 million photos that have been geo-tagged[21] with latitude and longitude coordinates. The collections also contain geographic information about members' locations and their uploaded pictures which could bring up lots of potential possibilities for further testing and analysis.

REQ5. Each member had a set of publicly available contacts, friends and family network in the profiles.

REQ6. In terms of retrieval methods it provided an API with lots of potential for mashing up data and retrieving useful information for different needs.

3.5.4 Gypsii (http://www.gypsii.com/)

Gypsii is a location aware mobile social network service combining social networking features with GPS[22]. *Gypsii* users can take photos, upload and tag them on a map in real time as well as locating friends in a network via GPS. In other words *Gypsii*'s target is to create real time user generated information via mobile phones. It helps users to create new places on the map, find friends' whereabouts according to users' own current location and search places of interest (updating constantly as they move). By nature *Gypsii* is designed to go beyond a static online social networking process. It is designed to overcome the static features of F*acebook* (Harpel, 15/07/08, PERS.COM).

REQ1. As a new application it just had 56 users at the time (May 2008), but the trend was constantly increasing.

REQ2. *Gypsii* was mainly based on American users, but it was planned to be extended into Asia, China and Europe. At the time it was too early to examine and conclude anything on the basis of the number of *Gypsii* users (Harpel, 15/07/08, PERS.COM).

REQ3. Locational information of members was available. Each member was associated with a different location e.g. home, work, favorite place, holidays.

REQ4. In addition, each member had a POI list with the exact location depicted on the map. Therefore, in addition to members' locations, each member could be related and studied according to their proximity of POIs.

REQ5. Members had a friendship network in which they shared their private locational information e.g. address and details of their wedding party.

REQ6. In July 2008 the site did not provide API for retrieving available information (Lennon, 15/07/08, PERS.COM).

3.5.5 POI friend (http://www.poifriend.com/)

POI friend is a Point of Interest (POI) community that lets users develop, maintain, comment and share their POI information with GPS, maps and driving directions. The developers' first and most important aim is to satisfy the GPS users who

[21] Adding information called geo spatial meta data that records the location where a photo was taken (http://en.wikipedia.org/wiki/Geotagged).

[22] Global Positioning System

always blame pre loaded POI on GPS devices for not being personally relevant. It applies the technologies of location based services and social networking sites.

REQ1. The site had 20,000 registered members and many visitors (Dave Krawczyk, 07/08, PERS.COM).

REQ2. The majority (90%) of the members were from North America (Dave Krawczyk ,07/08, PERS.COM).

REQ3. Members' locational information (approximate geographical location) was available.

REQ4. Members in addition to their own location could be studied according to the geographical proximity of their POIs list and also their friends' POI lists. This provided additional geo spatial information for study and analysis.

REQ5. In *POI friend* the friendship networks of members were only visible for friends and not accessible for the others.

REQ6. The site did not provide members with an API for systematic retrieval of huge locational information about members and their POIs.

3.5.6 *Wikipedia (http://wikipedia.com)*

In addition to the sites studied above, a non-geo social network was also assessed. The aim was to examine the possibility of retrieving relational and spatial information contained in online social sites.

Wikipedia is a free online encyclopedia for collaborative problem solving and knowledge gathering. *Wikipedia*'s contents rely heavily on users' participation to provide further ideas and information for enrichment of articles. It is based on Wiki technology where each member could also be an editor. Each page has an 'edit this page' link and a full history of different versions of an article is stored by the system and could be retrieved by users.

REQ1. *Wikipedia* had 5.77 million contributors at the time (May 2008).

REQ2. *Wikipedians* were distributed across the world and the articles were in 250 different languages.

REQ3. Locations of the registered users were secure unless indicated voluntarily in the profiles. Therefore, locational information was partly missing.

REQ4. As a non-geo social network, the site naturally did not have any other explicit spatial information. But, from the content analysis perspective, it was possible to assign location to the articles. The available geo spatial data available in that context have a different quality (from region or area to exact coordinates).

REQ5. Friendship network was not applicable to the site interactions. Members interacted together by editing each other's articles. Friendship had no definition in the site.

REQ6. The site did not provide users with API for systematic information retrieval.

3.6 Comparisons and Conclusion

Table 3.1 summarises the assessment results of the sites in terms of the developed criteria. The first attempt was given to *Wikipedia*, as the site had the following unique attributes:

1. Comprehensive, strong and constantly updating statistical support, updating monthly (e.g. distribution of article edits over *Wikipedians*, 50 recently active *Wikipedians*, etc.)
2. Accessible full history of pages and their edits
3. Specially designed tools for retrieving information and downloading databases in MySQL format

Although the above features sounded promising, the site did not provide any support for retrieving spatial information (strictly secured). Even the providers were not happy to release the randomized location of members with Latitude/Longitude +/- 5km (Maxwell, 03/06/08, PERS.COM). As has been predicted in chapter 2 (section 2.7) privacy has considerable effects on the data set of this study. This has reconfirmed here that geo privacy is a serious concern in spatial data analysis and has raised new challenges (Raper, 2002). Although people are less sensitive about releasing their locations online (Escher, 2008), even so obtaining spatial information about online members is strictly secured in some sites (e.g. *Wikipedia*). Therefore, as discussed in chapter 2 (section 2.7), affected by online privacy issues, the attempt has been made to select a site that has the spatial information publicly accessible. Accordingly, a sample of the available geo social networks at the time was assessed against the developed criteria in section 3.4.

Before, getting into details of the results of assessments in table 3.1 it should be noted that the selected criteria for the suitability of the examined sites are not of the same importance. In other words some of the criteria are much more vital than the others. In other words RQ3-6 have been assigned higher weight than RQ1-2. This inequality has been taken in consideration in the following comparison and conclusion.

The limited local distribution of members in *Yelp, Gypsii and POI friend* indicated that the result of analysis would be geographically biased and that would affect the stability of the future generalization of the result. 65 and 20,000 users of *Gypsii* and *POIfriend* respectively and 11 million *Yelp* visitors all were bounded in North America. In addition, none of the mentioned sites provided API for systematic information retrieval. API was the property that sounded essential for data gathering and analysis of this study. Moreover, *POI friend* did not possess the necessary relational data. The friendship networks of members in *POI friend* are not visible to the public. As a result, these three sites, *Yelp, Gypsii* and *POI friend*, did not seem to be quite appropriate for this study. Remaining sites with similar features were *TwitterVision* and *Flickr*. They both had a large number of users distributed across the world with locational information available. They provided users with API for locational information retrieval and free map services to locate information. The only feature that prioritized *Flickr* was the extra geo spatial information that was available for users through geo- tagged photo collections, whereas *TwitterVision* did not contain any other spatial information associated with members. The *Flickr* data set contained complex relational, temporal and spatial information that has been uploaded and submitted by millions of photographers voluntarily. Geo-referenced photo collections of members provided interesting multiple spatial information associated with individuals. From the social network analysis perspective it also included dynamic relational social data. Members had a social network of contacts, friends and family.

Consequently, although the locational information of all *Flickr* members might not be accessible (as predicted in chapter 2: section 2.7), the geo-tagged photo galleries and the existence of those members with their location available, made a stronger, richer and versatile data set with more potential for developing and testing genuine hypotheses. As discussed in chapter 2 (section 2.7.2) there is also a possibility of finding the approximate location of members with an unknown location by study and analysis of their geo-tagged photo collections.

Sites	Large number	Geo distribution	Home location	Multiple spatial information	Relational data	API
Yelp	+	-	+	+	+	-
TwitterVision	+	+	+	-	+	+
Flickr	+	+	+/-	+	+	+
Gypsii	-	-	+	+	+	-
POI friend	+	-	+	+	-	-
Wikipedia	+	+	+/-	+	-	-

Table 3.1. Assessment results of the selected sites

3.7 Summary

According to the aims and objectives of the research, this chapter discussed the essential properties of an appropriate data set. Potential social networks were examined (May 2008) and tested in terms of the data set requirements. *Flickr,* an online photographic community was found to be the most promising data set against the established criteria. The next chapter covers an overview of *Flickr* data in general. It also explains the sampling strategy for selecting 'Flickr Sample World data' as well as 'Flickr Sample GB data'.

4 Flickr data

4.1 Introduction

As described in chapter 3 the analysis of freely available spatial information on the web in this study is combined with an analysis of social networked data. Consequently, the appropriate data set for this study should carry the features of relational data as well as spatial data, bearing inherent complexity and requiring an interdisciplinary approach to their consideration. In order to collect a rich, unbiased and indicative sample of the Flickr data set, an initial analysis is conducted to assess the specification and characteristics of Flickr data in general. Sample data set is collected and has been used to assess the properties of the potential data. According to the results of the analysis of the pilot data, a final decision is made and justified for collecting a suitable, rich spatio-social data for developing and testing hypotheses.

4.2 Flickr

As mentioned in the previous chapter (section 3.1) Flickr is a photo-sharing site. Users have a social network of friends and a collection of photos on their profiles. According to available statistics (http://www.Flickr.com, retrieved 20/07/09) the Flickr database contains more than three billion photos, out of which a hundred million are geo-tagged. This indicates that approximately 3.3% of the submitted photos have locational information. In simpler terms, one out of every thirty photos in Flickr has been tagged with geographical information, although with different levels of accuracy. As mentioned by the Flickr developers[23], during the last two and a half years there have been as many geo-tagged photos as the total number of submitted photos during the first two years of the Flickr launch. This indicates the increasing popularity of geo-tagging behavior among Flickr users and the importance of the analysis of a huge amount of available spatial information that is of limited or no use in the existing format on the web. Some possible reasons for increases in spatially referenced submissions are considered in chapter 6, section 6.2.4. A simple search of the available Flickr data reveals that more than two thirds of the geo-tagged photos are public and available on the map, or accessible through the API (Table 4.1). This reconfirms the suitability of Flickr as a rich source of freely available spatial information (chapter3: section 3.6).

	Total (Public + Private)	**Public**
All Photos	3.6 billion	1.3 billion
Geo-Tagged Photos	100 million	74 million

Table 4.1 Number of photos uploaded to Flickr between February 2004 and February 2009

Interestingly, as the numbers in table 4.1 show the geo-tagged photos have significantly higher portion of public photos while around 64% of photos in 'all photos' category have been marked as private (available to view by specific authorized users and/or friends). In other words significantly higher proportion of geo-tagged photos are public compared to the Flickr public photos in total. This indicates that users have higher tendency towards sharing their

[23] Flickr developer blog available at code.Flickr.com retrieved on February 2009.

geo-tagged photos by everyone on the net than their non geo-tagged photos.

This finding shed light to an interesting fact regarding sampling data strategy and quality of the collected data in this study. Considering the fact that private photos cannot be retrieved for sampling, and being aware of the amount of public geo-tagged photos (74%) one can concludes that data sampling from geo-tagged photos results in better representative of the whole population. In other words, data sampling strategy and focusing on collecting geo-tagged photos only, sound as justified decision that can lead to more robust and indicative sample data.

4.2.1 Flickr geo-tagged photos

Flickr was designed especially to store, sort, search and share photos online. Users can sort their photos by grouping them into Sets, Collections or Galleries. According to the nature of this research, the focus is only on Flickr geo-tagged photos. The geo-tagged photos in Flickr are photos with additional locational (latitude, longitude) Meta data. Therefore, the geo-tagged photos have been used for sampling and analysis.

Before getting into more detail on geo-referenced photos in Flickr it is worth distinguishing between Volunteered Geographic Information (VGI) and User Generated Content (UGC). While the exact difference is a subject of debate, UGC can be used to refer to any information in the form of audio/video files, pictures, blogs, and discussion forums that users voluntarily upload on the web. However, the VGI can be seen as a specific kind of UGC that has been referred to spatial information that people provide voluntarily on the web. VGI has been defined as the most drastic phenomenon of the advent of Web 2.0. With reference to what has been discussed here, the online social networking sites can be divided in two categories in respect of UGC and VGI.

- Sites that collect content provided by users (Facebook, Flickr, Twitter)
- Sites for users to be able to populate a database (OpenStreetMap, Geograph)

The above categories are partly relevant to the role social networking plays in the use of the sites. In sites where users provide information to develop a shared final product (maps, driving directions, tourist information), the social behavior and interactions of the users are more likely to have relatively close patterns towards achieving a collaborative task. However, the first category has a higher chance of revealing general patterns and attitudes of users, without having a shared specific target to achieve.

According to the nature of the Flickr site and how it has been used by the users it is closer to the first category. That is, although people interact together and upload spatial information, there is no predefined collaborative task that is supposed to be satisfied with the user-generated contents.

In Flickr, users have the option for geo-tagging their photos. They can easily drag and drop their photos on the desired places on the map. Users can set the geo data for each photo. The geo data in this context includes the latitude, longitude and accuracy level. Therefore, each geo-tagged photo in Flickr, in addition to Photos Id, contains the following attributes:

- **Lat** (Mandatory): Valid range is -90 to 90 (maximum 6 decimal places).
- **Lon** (Mandatory): Valid range is -180 to 180 (maximum 6 decimal places).
- **Accuracy** (Optional): Accuracy level of the location information (current range is 1-16). World level is 1; Country is ~3; Region ~6; City ~11 and Street ~ 16.

There is also another optional attribute for geo-tagging photos, which is called context. Context adds extra spatial information in addition to latitude and longitude. For example, 1 is used for 'indoor', 2 for 'outdoor' and 0 for 'undefined'. Considering the nature of this research and its aims and objectives (chapter 1: sections 1.3 and 1.4) this attribute is not applicable and will not bring extra information for answering research questions. Therefore, it has not been used as part of the location Meta data for geo-tagged photos.

It is worth mentioning that before assigning any location information to a photo, users should define the privacy level for the photo to clarify who can get access to that information, otherwise the geo-tagging process cannot proceed. Therefore, as mentioned before, this research only focuses on the public photos, which by default can be accessed by anyone exploring the Flickr site.

Finally, in order to get more precise spatial Meta data for geo-tagged photos, the data set is restricted to those photos with the highest available 'Accuracy' in Flickr (level 16).

Since accuracy and precision are often used interchangeably, it is important to clarify the difference between the two terms. Accuracy and precision are two terms that are in use in most of the scientific methods, from science, engineering, industry, statistics and information systems to psychology[24]. Although each field has a specific definition and meaning for the mentioned terms, the general definitions that cover all the fields describe precision and accuracy as follows: Accuracy is the degree to which a measurement is close to the true value (Taylor 1999). Precision, known as reproducibility or repeatability, is the degree to which the same measurement can be achieved by repeating the experiment several times under the same conditions (Taylor 1999).In this study, accuracy for Flickr users' home locations means the degree to which available locations are indicative of a place where people live. It considers how close the information is to people's real home address. On the other hand, accuracy of the photos refers to the degree to which the photo's location on the map is close to the place at which the photo was really taken. According to the above definitions, it can be argued that Flickr uses the term 'accuracy' to measure the 'precision' of the georeferencing process[25]. Precision for home locations demonstrates, the level of detail that has been described on the profiles. For photos, precision shows how precisely (higher zooming level of the map) the photo is placed on the map.

[24] *JCGM 200:2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM), Available at http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf Accessed on 18/08/2011*

[25] Whether this has been done intentionally for a specific reason or not is not clear to the author.

According to the nature of this study, the home locational information of users can be classified according to different levels of detail, e.g. 'London, UK' can be ranked as a higher precision than 'UK' or 'Europe'. However, measuring the accuracy of this information is not quite feasible since it needs to assess if users provide valid and true information regarding their locations in online communities, and this can only be done with qualitative analysis (interviewing users, etc.). Hence, there is the possibility of assessing the accuracy of the home location of users by study and analysis of the geographic distribution of available spatial information (other than home locations) in later steps of this research (chapter 8).

Last but not the least point about the geo-tagged photos of this study that should be mentioned here is that according to Westman (2009) there are three different main levels of attributes for describing, indexing and querying images:

- Non-visual information (meta data attached to the image)
- Syntactic image information (visual information in the actual image)
- Semantic image information (conceptual content that requires previous personal or cultural knowledge)

In this study the Flickr geo-tagged photos were collected based on 'non-visual image information'. In other words the attributes collected for the photos of this study are not present in the image itself. The Meta data associated with the photos of this study were retrieved from the following proposed attributes for non-visual image information by (Westman, 2009):

- Biographical attributes (e.g. poster, date, title)
- Physical attributes (e.g. type, technique, location)
- Contextual attributes (e.g. caption, reference)

4.2.2 Flickr Friendship

Flickr is a photo-sharing site, but it also has functionality for people to interact together. This is mainly used among users to comment on each other's photos. It does not use the term 'friend', instead it uses 'contact', and therefore the friend in this study is the same as contact in Flickr. Moreover, unlike most of the social networking sites, e.g. Facebook and MySpace, friendship in Flickr is not reciprocal. Therefore, there are three different statuses of friendship for Flickr users:

- Non-reciprocal contacts: those 'the user' chooses as friends.
- Reverse contacts: those that choose 'the user' as a contact.
- Reciprocal contacts: when there is both a non-reciprocal and reverse contact between two users.

In addition to adding contacts, users in Flickr can define their contacts in three different categories, as:

- Family
- Fiends
- Friends and family

This level of detail among the users' contacts is not considered in this study. In addition, the above category can also be assigned a privacy level as private or public. The public friends are visible while the private contacts are hidden. In this study, friendship has been considered as a general term of having a relation between two users. Therefore, no difference is applied to the three different kinds of the aforementioned friendship types. This assumption was made based on the fact that, although there are different types of friendship on Flickr and the differences might have some effects on the friendship network of users, considering the research questions the simplified definition of friendship is sufficient to develop and test hypotheses. By making that decision one can argue that special members (i.e. famous people, celebrities, singers, actors, etc.) might get a very high number of one directional friendship according to the number of their fans who follow them. Consequently, dismissing the direction of friendship might lead to vulnerable conclusions. Here, the author argues that existence of a relation between two users plays a more important role in this study than the direction of the friendship. The developed research questions here examines where the friends live and where they take photos and therefore, if they are followers or being followed is not a subject of analysis. Accordingly, in testing hypotheses and making assumptions in chapters 7&8, the social interactions and patterns are examined by bearing this simplified version of friendship and its potential effects in mind. In order to achieve results that can be generalized to wider sets of spatio-social data, it is possible to examine and study the whole friendship network of GB posters (holistic view instead of the sub-network studied here). In that case although wider sets of home locations would need to be disambiguated, the analysis in answering the RQ3 might lead to a more general outcome.

4.3 Retrieving Spatial and Relational Data from Flickr Database

The Flickr site provides API that could be used in different programming languages e.g. PHP, Perl, Delphi, Java, etc. *Flickrj* is a Java API, which wraps the REST-based Flickr API in Java language (information available at http://www.Flickr.com/services/api/) and has been used for retrieving the information required for this study. In order to use the services, methods and classes available in Flickr API, Java codes were developed through *Flickrj*. The developed codes send requests through the appropriate calling methods, with relative arguments, to the Flickr database and the required information will be sent in the *Flickr* response format. This process can be followed through the code (Appendices A-B) developed in the Eclipse[26] software development kit (SDK) and the resultant data are summarized as:

Spatial Data

- (Lat, long) for geo-tagged photos
- Home location of posters
- Photo accuracy

Non-spatial Data

- Number of geo-tagged photos
- Photo ID
- Poster ID

[26] http://www.eclipse.org/

- Poster user name

Relational/social Data

- List of public friends for the given poster

To justify the approach taken for the data extraction process, it is worth mentioning the following two issues that influenced considerably the data collection application (Appendices A-B):

- **Flickr Queries**
  The speed of the response from the Flickr database is heavily dependent on the quality of the network at that time. Therefore, queries to Flickr can be slow, e.g. three queries per minute. In order to speed up the process, Hash Maps[27] were used to store and process data locally.
- **Flickr Bug**
  During the experimentation with Flickr API, it was found that queries that return a large numbers of results could have repeats. To overcome this problem, the retrieval period was broken to small time slots, e.g. 15 geo-tagged photos per day during the last two-year sampling period instead of 12,000 geo-tagged photos during the last two-year sampling period. This method minimizes the risk of processing repeated data while collecting sample data.

4.3.1 World Data Sampling

Having introduced the spatial, social and temporal aspects of the Flickr data set, this section briefly studies the attributes of the Flickr database. The findings and conclusion of this section have been used when developing a suitable sampling method. Before selecting a final data set for testing and developing hypotheses and answering research questions, a sample of data is required to assess and evaluate the specifications of the data set. Due to the volume of the available VGI and in order to work with a representative but manageable size of Flickr data, the analysis should be limited to a collection of the available geo-tagged photos. Accordingly, some criteria are defined for sampling a representative collection.

Firstly, since this research aims to study the multi-geographies associated with individuals in online communities, it is essential to focus on users with geo-tagged photos in their collections.

Secondly, for better analysis and ease of mapping, photos with the highest level of accuracy are considered. In the case of Flickr, it is accuracy 16 at street level.

Thirdly, in order to get a thorough view of Flickr geo-tagging behavior and friendship networks from the early days of Flickr to date, sampling is conducted for photos that have been uploaded between 01/02/04 to 20/07/09, covering the entire Flickr life (since its initial launch until the date of initial sampling).

In order to split this five and a half year period into sensible, manageable and comparable sampling intervals, it was decided to divide it into a number of twelve-month duration slots. This sampling process has been applied from the

[27] "A dynamic collection in Java that allows the number of elements to grow and shrink during running the program. This implementation provides all of the optional map operations". http://java.sun.com/javase/6/docs/api/

dates 19/07/04 to 20/07/09 (Table 4.2). In order to cover all the periods for which Flickr has been in use, the first six months of Flickr's life (from 01/02/04 to 20/07/04) have also been considered. This period is shorter than the other sampling periods and is expected to reveal the early behavior and statistics of Flickr and its users, which most likely would have been affected by rudimentary users, and the basic functionality of the site.

This finally led to a sample set, including five twelve-month periods and the initial first six months of Flickr's life, to be indicative of separate periods of Flickr's life. An attempt has been made to choose the best possible intervals for analysis of the changing behavior of users including online friendship, geo-tagging photos and accuracy level of their home locational information. Accordingly, in order to get random but indicative data, which can be rich enough for further analysis, fifteen photos were randomly selected on a daily basis during the selected time intervals. The results are summarized in Table 4.2.

Intervals	Photos	Posters	Photos Per Poster
19/07/08- 20/07/09	5,729	1,142	5
19/07/07- 20/07/08	5730	991	6
19/07/06- 20/07/07	5,730	1043	5.5
19/07/05- 20/07/06	5,730	947	6
19/07/04- 20/07/05	5,715	846	6.7
01/02/04 – 20/07/04	2,670	411	6.5

Table 4.2 Total number of randomly selected photos, with the unique number of their posters and average photos per person for all selected sampling intervals

The sampling process resulted in 31,304 photos with 5,380 unique posters and their variable home locational information, friendship networks and geo-tagged photo collections.

4.3.2 Great Britain Data set

After a general assessment of the Flickr database (section 5.2) and according to the findings of the pilot study and analysis of Flickr sample world data (chapter 6: section 6.2), in order to produce a representative and manageable data set for better analysis and ease of mapping, whilst also considering the ambiguity, uncertainty and bias in the data (chapter 6: section 6.2.6), the data set of this study is collected based on the following criteria[28]:

- Highest available accuracy (as measured in Flickr)
- Photos uploaded within GB borders
- Photos uploaded from 1st February 2004 until 15th July 2009
-

Disambiguation, classification and uncertainty measurements have also been adapted accordingly to handle the GB data set. For doing the above, the data

[28] More on how the pilot study influenced the full data collection and design in chapter 6.

collocation application (section 4.1 and Appendix 4) has been modified to apply restriction on the location of photos (Appendix 5). Therefore, in the first place all photos taken within GB borders were collected during the sampling period. 1,827,400 photos were collected within the GB borders. During the course of the data collection, a bug was noticed within the Flickr API service. This generally relates to cases where the level of the query exceeds ~4,250 photos, resulting in the replication of the output. In this case out of the 3,357,232 photos it has produced 144,042 duplicates, i.e. (%4.2) in world photos and 2,631 repeats out of 1,830,031 (%0.15) for GB photos. All replications were removed from the data prior to further analysis. This is the same problem mentioned in chapter 3 but according to the nature of GB data set this time the bug had a more severe effect. For the collected photos a list of unique posters was produced. Afterwards, in order to gain a better understanding of GB posters' photo-sharing behavior, in addition to GB photos a complete list of all GB posters' photos was also collected without any geographic restriction. This was carried out during the entire sampling period. Overall,[29] the data set comprised 3,245,866[30] photos, out of which 1,827,400 were within GB borders and taken by 19,782 posters. Out of 19,782 posters, 10,323 have their home location available in their profiles. The remaining posters (9,459) have declared their home as null or unknown. Friendship collection was the same as the one conducted for world sample in previous section. 74,840 posters have at least one GB friend. This finalized data set is believed to be an indicative sample, socially and spatially rich enough for presenting the complex spatio-social relation among Flickr posters. Each GB poster has the following attributes:

- Collection of GB photos (ID, poster ID, lat, long)
- Collection of world photos (ID, poster ID, lat, long)
- Collection of GB friends (same as poster)
- Home location

Java applications, developed to send queries to the Flickr database and retrieve the above required information, are attached to the Appendices (A, B, C). In the first place, all photos taken inside GB borders were collected. After that a list of unique posters of those photos was produced. According to those GB posters, a collection of their GB friends was created and saved. In addition to GB photos for each poster, a complete collection of each GB poster's photos was also gathered. Unlike the friendship network, which is limited to GB posters only, the photo collection for each user was gathered comprehensively. Considering the research questions, for each GB poster, in addition to a collection of his/her GB photos and in order to achieve a broad view of the photo-sharing behavior of posters, a collection was made of all their photos around the world without any restriction on the borders of any specific country. This decision can be supported by the fact that the geographic distribution of world photos for each

[29] The author here is obliged to inform the reader that one day of the 5 year sampling period has been missed from analysis (date: 19/07/08). On that day there were 3,129 photos uploaded. 23 of the posters uploaded photos in Britain only on that day and therefore, their behaviours are not crucial to the network of this study. They only uploaded photos once. Moreover, 12 of them have no friends in GB collection. Accordingly, although missing that date is not ideal, according to their social level (number of friends) and prolificness (number of photos) no major loss of data would happen to our final results.

[30] It is important to bear in mind that the Flickr database is constantly being updated. As a result, photos taken during the sampling period and uploaded afterwards are not included in this data set.

poster, other than their GB photos, has the potential to demonstrate the geographic footprint of the user and brings the possibility of assessing the relation between home location of the users and places in which they take photos and the location where their friends live.

4.3.3 Summary

This chapter reviews the attributes and specifications of two sets of Flickr data collected for this study. The first data set was collected during a pilot study for assessment and evaluation of Flickr data. It contains a random sample of all photos uploaded to Flickr. The second set of data was collected according to more restricted criteria. It contains all photos uploaded to GB borders during the five-year sampling period. The next chapter explains the methods developed and applied to the finalized data. A significant amount of cleaning, mining, filtering and interaction has been done with the aim of bringing out the meaning, patterns and trends in the large examined data set.

5 Classification, Disambiguation and Visualization of Spatio-Social Data

5.1 Introduction

This chapter covers the methods developed and applied for analysis, classification, disambiguation and visualization of the spatio-social Flickr data. The first section explains the initial attempts made to assess the general characteristics of Flickr data through analysis of the Flickr sample world data set (chapter 4, section 4.3.1). The second section explains the amendments applied, according to the findings of the previous assessments, in order to make the final modified methods appropriate for GB data analysis. Moreover, the existing visualization packages have been assessed against the research requirements (section 5.4.1). Accordingly, an appropriate Java based visualization package is selected. The final section explains the step-by-step stages of design decision in the development of a visualization application for study and analysis of the complex spatio-social relations of Flickr users.

5.2 Flickr Sample World Data

As described in the previous chapter (chapter 4: section 4.3.2.1), the sample world data set has the following spatial information:

- Geo-tagged photo collections
- FHLI

The geo-tagged photo collections come with formally defined coordinates, latitude and longitude, and therefore with a straightforward conversion can be mapped onto the screen coordinates (section 5.4.2.1). However, the FHLI, as expected and explained in the context chapter (chapter 2: section 2.5), needs to be disambiguated before analysis and visualization.

Referring to chapter 4 (section 4.2), Flickr does not apply any restriction on how to define home location on profiles. Therefore, FHLI comes in any format with no restriction on size, length, format, accuracy, etc. Consequently, it is inherently imprecise and fuzzy. It varies from fuzzy vernacular geographic terms that people use in their everyday lives, to precise coordinates.

This section gives a brief overview of the existing disambiguation methods. It assesses each solution against the attributes of the FHLI. Finally it justifies the need for a new customized disambiguation application. The application has been designed, developed, implemented, applied and evaluated with the collected FHLI in the following sections of this chapter.

5.2.1 Disambiguation in Literature

FHLI of users varies from detailed information, including house number and coordinates, to geographic terms people use in their every day lives. This information comes in any format without any restriction. It is also acceptable to leave the home location field as blank or enter several locations. Consequently, this kind of data can reflect the natural behavior and spatial mental maps of people and has been named as “fuzzy psycho-geographical” data by Waters and Evans (2003). They argue that the large amount of locational data on the web is defined in people’s natural day-to-day language, and therefore there is a need for a reliable system for capturing and normalizing that “fuzzy vernacular geographic" information. As discussed and can be expected from Purves et al. (2005), the unstructured web documents are usually a combination of two types of location information:

- Locations with well defined footprints
- Locations that may not map directly onto an ontology or gazetteer.

FHLI collected for this study contains a combination of the two Purves et al. categories. Geo-photo collections can easily be mapped according to their coordinates. However, FHLI comes in different levels of detail, from the informal natural language of people (vernacular geography) to scientific geographical vocabulary (latitude, longitude, postcode). People can leave their home location as blank or name as many locations as they want, or even write a description of a place in which they live. The data set thus contains variable kinds of locational data that make the grounding (localization) and normalization a challenging task (Li et al., 2003). Therefore, the nature of the spatial data set of this study comes with considerable uncertainties and vagueness that requires an appropriate disambiguation method.

The best-known solution for disambiguation of place names in literature is based on the assumption that if a word is repeated more than twice in a document it is highly likely that all occurrences have the same meaning (Gale et al. 1992). In addition to discourse constraints, another suggested method of disambiguation is to assign a default location to the ambiguous name (Montello et al., 2003; Jones et al., 2008). This can be done according to several assumptions:

- Most frequently occurring place (Smith and Mann, 2003)
- Higher population of the alternative places (Rauch et al., 2003)
- Web mining (Li, 2003)
- Referring to the hierarchy associated to the given name in gazetteers (Jones et al., 2008).

The above-mentioned methods, although introduced and applied to ambiguous terms in different situations and concepts, are all limited to at least one of the following:

- Well-edited text document (Amitay et al., 2004), discourse analysis and textual content (Smith, 2002) or semi-automatic extraction from the web (Li et al., 2003)
- Referring to an attribute available in gazetteers: spatial hierarchy (Arampatzis et al., 2006) or population (Rauch et al., 2003)
- Adapting the Prim's and Kruskal's algorithms (implemented in an application called InfoXtract by Li et al., 2003) to measure the minimum spanning tree for each location. This performs relatively well for short documents but will fail for large documents (which is the case for FHLI of this study).

Although all the above methods and applications deal with ambiguity and uncertainty in place names, Flickr data has the following attributes that make the existing solutions to be of limited use:

- FHLI comes in different formats. It varies from scientific geographical vocabulary to fuzzy vernacular place names.

- FHLI is not part of a text document or discourse. The geographic locations are the only information that are available, and therefore the text/web mining methods do not sound promising.
- FHLI can come in any levels of precision and accuracy. The information varies from blank, to one specific place, several places, and description of a place to precise coordinates in some cases.
- The locations are not specific to a limited area or country.
- Not all FHLI have an entry in gazetteers and also not all locations in the gazetteers have population registered.
- The size of the data is large; therefore efficiency of the application is essential.

As a result of the above points, it can be concluded that applying the existing disambiguation methods and looking into the existing gazetteers cannot disambiguate the FHLI (vague terms come in VGI). Therefore, the grounding and normalization in that respect require new methods to be developed and new classification decisions to be made.

5.2.2 *Classification of the Ambiguous* FHLI

Initial attempts to classify the retrieved FHLI revealed six classes of vague terms that require new disambiguation methods to be developed and new precision measurements to be applied (Table 5.1). These classes are ordered according to different types of vague terms occurring in the data set.

Vague Classes	**Examples**
Doesn't exist	• Never: 'Outer space', 'L???' • Current: 'Standel, Kent County'
Multiple Alternatives	• Same name for different places o Multiple scale: '*Netherland* (country, city, town, hamlet) o Multi-site place: '*Nanyang Technological University*' • Different names for same places: '*Germany, Allemand, Deutschland*'
Multiple Entities	'*UK, Paris one day Italy*'
Abbreviation	• Single scale: '*Philly' (Philadelphia)* • Multiple scale: *'PC, US' (Pacific Coast, Panama City, Park City, Penn Central)*
Misspelling	• 'Toru?, Poland'
Descriptive	• 'I live somewhere with lots of sunshine'

Table 5.1. Classes of vague terms in FHLI.

5.2.3 Precision Classification Model

Initial attempts were made to apply Flickr's geo-photos classification model to measure and classify the FHLI. As mentioned in section (chapter 4: section 4.2.1), Flickr uses a 16 level accuracy classification:

- World level = 1
- Country ~3
- Region ~6
- City ~11
- Street ~16

The above model is based on the zoom level that the user selects for uploading their geo-photos. Accordingly, photos that are uploaded on highly zoomed maps are considered more accurate than others. Since the zoom level of the map is not applicable to FHLI and considering the fuzzy vernacular geographic terms that are frequently found in FHLI (Table 5.1), the above model was found to be inadequate for classification of the FHLI. Consequently, the Flickr model has undergone essential changes to include more detailed hierarchical spatial units. Accordingly, fourteen distinct precision levels were identified (Table 5.2)[31].

Precision Level	***Spatial Unit***	***Precision Level***	***Spatial Unit***	***Precision Level***	***Spatial Unit***
0	Blank	5	Region	10	Village
1	Unknown	6	State	11	Street
2	World	7	City	12	Postcode
3	Continent	8	Town	13	House No.
4	Country	9	Borough	14	Coordinates

Table 5.2. Precision classification model for unstructured locational information of Flickr users.

The above model is based on governmental divisions that have sharp geographical boundaries. A particularly important goal of this model was to produce a classification according to standard and flexible administrative structures, so as to enable the comparison between FHLIs. Applying the six cases that appear frequently in Flickr (Table 5.1) made it clear that the governmental divisions are not reliable scales for classification and measuring the precision of FHLI. This conclusion is supported by the following:

- Different internal administrative names for land units in each country (Census designated place and unincorporated communities in US, Parroquia in Spain, Civil parish in UK, Commune in France and Italy, Ward and prefecture in Japan)
- Different internal organizations exclusive to each country. The form and structure in which the land is divided can vary significantly from one country to another. For example provinces in China are very different from Canadian provinces[32].
- Inconsistency between size and population and the hierarchy of governmental divisions. Since there is no universally accepted size and population for governmental divisions the assumption of applying hierarchical precision to the governmental divisions is not valid. For example not every city is larger in area and more populated than every town. Ipswich is a larger and more populated town than the city of St. Davids (Both in UK). This argument can be made for any of the two selected divisions in the model.

Therefore, has become clear that the governmental units, although having sharp boundaries, are not consistent and comparable amongst different countries. This shortcoming stems from what Fisher (2007) referred to as “lack of definition and not lack of data”. There is no universally accepted definition for administrative units that can be applied for all units in the world. Therefore, the administrative divisions selected and modeled in Table 5.2 found insufficient scales for measuring the precision of FHLI. Table 5.3 summarizes the identified

[31] It is worth referring to the section (4.2.1) in chapter 4 on explanation of accuracy and precision terms used in Flickr and in this study. As mentioned the accuracy and precision definitions are applied differently to FHLI in comparison to geo photos.

[32] (http://www.wisegeek.com, accessed 01/11/09)

inconsistencies in the accuracy classification model based on administrative divisions.

Description	Example
Different internal administrative names for land units in each country	*Parroquia* in Spain, *Ward* in Japan.
Different internal organizations (land divisions) exclusive to each country	*Province* in China and Canada.
Inconsistency between size and population and the hierarchy of administrative divisions.	Ipswich (town) larger than St Davids (city)

Table 5.3. Inconsistencies in spatial units amongst different countries.

Accordingly, in order to minimize the mentioned inconsistencies a suitable uncertainty number (from 1 to 5) is used to reflect the confidence in this classification (Table 5.4).

Uncertainty Classification	Description
1	Less uncertain than the following uncertainties *('London, UK')*
2	Nested spatial units e.g. city and county *('Denver', 'New York')*
3	Different places in one country *('Portland, US', 'Cangas, Spain')*
4	Different places in different countries *('Netherland')* or several places for a single user ('Anchorage, *Los Angeles, Someday New York, may be Paris')*.
5	Blank or information that cannot be associated with any place in the world ('*Outer Space, L????')*.

Table 5.4. Uncertainty classification for FHLI

5.2.4 Disambiguation Methods

As described in the previous section none of the existing disambiguation methods can successfully disambiguate the FHLI. As a result, a customized disambiguation method is required to disambiguate the FHLI before further analysis. Since the existing gazetteers have avoided the vernacular place names, geo coding the available VGI according to gazetteers would remain a subject of debate and, as predicted by Twaroch et al. (2008), will cause problems and failures. In order to overcome the reference and referent ambiguity (section 2.6) that is present in the majority of FHLIs, a 'notion of default sense' (Arampatzis et al. 2006) has been used in this study. Accordingly, attempts were made to assign a best possible default location to each of the examined FHLI. Rauch et al. (2003) introduced a solution that, although it has been applied to place names in one single document, could be adapted for this study as well. They argue that by keeping the history of the vague term in a whole document, the probability that the given place means the given name will improve over time.

P (Place, name): Probability of the given place means the given name

For example if a hometown is mentioned as 'London', the disambiguation process is as follows: Following the suggestion of Rauch et al. 2003, the probability of London to be London UK is directly related to the number of occurrences of the potential alternatives for London in the document.

Occurrence (placeName): Number of times that the given place name has occurred in the examined document

Therefore,

P (London, London UK) > P (London, London Ontario)

If

Occurrence (London UK) > Occurrence (London Ontario)

In geo coding the vague locations, Purves et al. (2005) suggested two alternative methods:

- Assigning multiple locations to a single reference or
- Assigning a default location

Following the above methods FHLIs are assigned a default location according to the occurrence frequency of the possible references. In order to be more precise an uncertainty ranking is also added to the location metadata (adapted from Purves et al. 2005). In order to reflect the confidence in selecting a default value for each FHLI, an uncertainty number is also assigned (Table 5.4).

C (place, name): The confidence that the given name refers to the given place (adapted from Rauch et al. 2003. Overall, the process of classifying and disambiguating the FHLI consists of three consecutive steps:

1. Home location retrieval
2. Disambiguation
3. Uncertainty classification

This method is adapted from the process of assigning geographic weight to documents (web pages or digital libraries) which originally has three steps (Smith and Crane 2002):

1. Name identification
2. Categorization and
3. Disambiguation

In the modified version for this study, 'name identification' is replaced by location retrieval. This decision is based on the fact that unlike web mining process the locational information of Flickr users is not part of a document that needs to be separated from other words. All the words are considered as a place name. Therefore, the first step is to retrieve the information from the Flickr database. In the second step, disambiguation replaces the categorization step. This is based on the fact that in text mining, after identifying the names they need to be categorized as person, place or date, etc., while in Flickr data, it has been assumed that all the home locations are in the place category. Finally, an uncertainty number is used to reflect the confidence in selecting a default location for each of the FHLI. In developing and applying the above method to the Flickr sample world data, it should be mentioned that the existing disambiguation algorithms do consider the discourse constraints either by probability rules or by following the one sense per discourse analysis. However, FHLI is not in a body of text and is not accompanied by any other information. Therefore, in order to adapt the existing solutions (Jones et al. 2008, Smith and Crane 2002), in the disambiguation process the following two assumptions were made:

- Instead of looking for other accompanied more precise places within the extent of a place name in the document, all retrieved FHLI of the sample world data set is considered as one document (which can be replaced with body of text in the text mining process). According to Li et al. (2003) default values play important roles in disambiguation of vague location terms. Their research found that people refer to 'Los Angeles' as the city in *California* more than the city in the *Philippines, Chile, Puerto Rico,* or the city in *Texas* in the *USA*. Therefore, in the Flickr data set the probability of a place to be chosen as a default value is directly related to the number of GB posters who live in that location. Therefore, the locations that have more GB posters have a higher chance of being selecting as a default location than other alternative places.
- Moreover, additional spatial information has been introduced to replace the role of other accompanied names in the document. In FHLI, geographic distribution of friends and photos for each poster is considered as extra spatial information contributing to the disambiguation process or hypotheses testing.

The above disambiguation method is applied manually to the Flickr sample world data (chapter 4: section 4.3.1) and the results are shown in chapter 6 (section 6.2.6). This experiment has been conducted in order to identify a series of requirements and challenges during the disambiguation and selecting a default location for each FHLI. In other words, manual disambiguation of random samples of FHLI helped in developing an automated method that implements the human knowledge and experience in disambiguating the vague locational terms. The disambiguation method is explained in the next section.

5.3 Disambiguation Methods for Flickr Sample GB Data set

As realized during the analysis of the sample world data set, a disambiguation method is required before analysis and visualization of FHLI. This section describes the disambiguation process developed according to the above disambiguation method and in line with the findings of the above experiment (chapter 6: section 6.2.6). The method developed here is slightly different from that applied to the world data set in the previous section. These changes stem from the shortcomings identified in the previous model (section 5.2.2-3) and the results of analysis provided in chapter 6 (section 6.2.6). Efforts have been made to minimize those drawbacks while successfully disambiguating the FHLI automatically.

5.3.1 Disambiguation Modification

According to the results and findings of the application of the introduced method to the Flickr Sample World Data in the previous section, the following modification was conducted in the classification and disambiguation of the Flickr GB Data. Attempts were made to minimize the shortcomings and weaknesses identified in analysis and disambiguation of the world data set (section 5.2). As a result of the uncertainty measurements (Figure 6.15), it has been identified that uncertainty and ambiguity is higher in national scope. Therefore, in order to get a rich and indicative sample for visualization purposes and to answer the research questions, data were limited to GB borders only. As found in Figure 6.15, two out of 12 precision classes have contained the majority of the locations. The rest of the classes have either none or a few locations. Posters either left their home locations as blank or provided information in city level. In other words, there is only one precision level that encompasses the majority of the spatial information. Therefore, precision measures (Table 5.2) that might lead to several identified inconsistencies (Table 5.3) were excluded from the analysis of the Flickr sample GB Data.

As a result of the above finding the uncertainty assignment (Table 5.4) has also been modified. A new uncertainty measurement based on city level (as a unit of measurement) has been developed and applied.

- Unc2: one place in one country
- Unc3: several places in one country
- Unc4: several places in several countries
- Unc5: unknown

According to the inconsistencies recognized in comparing the administrative units in different countries, Geo-Names standard feature code (spatial hierarchical unit, Appendix 2) has been used for assigning precedence over places with the same name (Table 5.12).

5.3.2 Disambiguation Process for Flickr GB Data

As also described in previous section FHLI contains the natural expression of people from the place they live. The data of this kind is ambiguous and prone to wrong interpretation. It includes vernacular geographical terms that people use in their everyday lives and, conventionally, they are not included in the official gazetteers. As described in section 5.2.1 none of the existing algorithms can complete the disambiguation of FHLI successfully. As the nature of the Flickr GB data is by no means different from the Flickr sample world data that has been collected randomly and examined in the previous section, (chapter 4: section 4.3.1) the existing disambiguation algorithms at the time of this research (summer, 2010) were not applicable and suitable for disambiguation of FHLI (section 5.2).

5.3.2.1 Modified Disambiguation Algorithm

A customized method was required to read FHLI for each poster, disambiguate the string and select a default location for the examined FHLI. Hence efforts were made to develop an appropriate algorithm to find a comprehensive list of place names around the world. This was achieved by looking into geo-databases. GeoNames[33] geographical database was found as a suitable source of information that provides list of place names that covers all countries. All the data in GeoNames can be downloaded free of charge. It contains several files ready to download in different categories. It includes place names for each country, city names of all countries with several restrictions on the population (e.g. greater than 5000, etc.), country information and GeoNames specific feature codes (that can be replaced by admin codes of different countries). It also includes information about unhabitable places, i.e. mountains, lakes, forests, etc. After study and assessing the available data files in GeoNames, four files were downloaded from the server:

- AllCountries: All Countries contains information about all countries combined in one file. This file includes names of all places around the world. The additional metadata about places follows the main GeoNames table format (appendix 1). The original file also includes information about unhabitable places (i.e. sea, hills, lakes and forest). In order to improve the efficiency of the algorithm and reduce the size of the file, the unhabitable places were removed from the AllCountries file. The remaining information includes only the habitable places and is called PopulatedPlace file in this study.
- CountriesInfo: CountriesInfo contains countries' names plus additional information, i.e. bounding box, language, ISO code, FIPS code, and capital.
- Admin1Codes: Admin1Codes file contains first administrative order in each country. In other words it is the repeat of the entries in the populated places file that have their feature code as ADM1. This information is provided with country code.
- AlternateNames: AlternateNames is a list of alternative names for each populated place listed in the AllCountries file. The field 'alternatenames' in (Appendix 1) is a short version of this file. In order to minimize the risk of missing a place that has been referred to with its alternative names in FHLI, the complete version of the AlternateNames file was also downloaded and used in the disambiguation process. This file will be used for dealing with FHLI cases with 'referent ambiguity' (chapter 2: section 2.5, Jones et al. 2008).

The above files were used in finding alternative locations found around the world for the given place names. All the GeoNames data files are formatted to include the fields and attributes described in the 'Standard GeoNames Table' (Appendix 1). They were used for finding potential candidate places for each location name and also find alternate names for each location. The following section explains in detail the steps followed in order to develop a disambiguation algorithm for Flickr sample GB data.

[33] http://geonames.org

5.3.2.2 FHLI Disambiguation Algorithm

Finally a comprehensive list of all place names in the world was downloaded from GeoNames' geographic database (http://www.geonames.org/)[34]. As explained in the previous section the file has undergone essential changes to be formatted, filtered and saved in an appropriate, ready to use, format. Each FHLI was then separated according to delimiters within the text. In doing that a Java application was developed for cleaning the FHLI text file. It reads each string and replaces all symbols and delimiters with commas, then separates the string according to the commas. For each separated text, all places with the same name in the database are saved. Each saved location also is assigned one of the following types (according to the files downloaded from Geo-Names):

- Country
- Admin1
- PopulatedPlace (PPL)

In the second step the algorithm searches for nested locations amongst the alternate locations found in the previous step. The following explains the necessary and sufficient condition for a populated place to be nested in admin1:

A populated place is nested in Admin1 if: PPL.Admin1Code = Admin1 Code

The above line should read as 'code of the admin1 of the examined populated place equals to this admin 1 code'.

Following the same strategy necessary and sufficient condition for an admin1 to be nested in a country is as under:

An Admin1 is nested in a country if: Admin1.CountryCode = Country Code

Likewise the above line says as ' country code of the examined admin1 equals to this country code'.

For each occurrence of nested locations, the remaining alternatives will be excluded from further analysis. Two examples of the FHLI and how they have been disambiguated according to this algorithm can be found in Appendix 3. Depending on the combination of the location types, the FHLI is classified into seven different categories (Tables 5.5-11). An appropriate uncertainty number is also assigned to reflect the confidence in selecting a default location (Table 5.12).

[34] http://download.geonames.org/export/dump/

PPL, Admin1, Country	Description		Location	Unc
Nested	One country	PPL nested in Admin1 and Admin1 nested in Country	Most nested PPL	2
	Two or more countries	PPL and Admin1 are nested but not in mentioned country	Mot nested PPL	4
Not nested	One country	If more than one PPL is found, the one with higher number of occurrences in the data set	Most occurred PPL	3
	Two or more countries	PPL and Admin1 are not nested inside the mentioned countries	Most occurred PPL	4

Table 5.5. FHLI contains three different types of location (populated place, admin1, country)

PPL, Admin1	Description		Location	Unc
Nested	One country	One occurrence	Most nested PPL	2
	Two or more countries	More occurrences in more than one country	Most occurred PPL	4
Not nested	One country	PPL not inside the mentioned Admin1, both in one country	Most occurred PPL	3
	Two or more countries	PPL and Admin1 have occurrences in several countries	Most occurred PPL	4

Table 5.6. FHLI contains two types of location (populated place, admin1). No country is mention

PPL, Country	Description		Location	Unc
Nested	One country	One occurrence in mentioned country	Mentioned PPL	2
		More than one occurrence in the mentioned country	Most occurred PPL in the data set	3
	Two or more countries	Ignore other places with same name as the mentioned PPL in other countries	Most occurred location in the data set in the mentioned country	3
Not nested	Two or more countries	PPL does not have any occurrence in the mentioned country	Most occurring alternative of the PPL in the data set	4

Table 5.7. FHLI contains two types of location (populated place, country). No admin1 is mentioned.

Admin1, Country	Description		Location	Unc
Nested	One country	Admin1 is nested in the mentioned country	Mentioned admin1	2
	Two or more countries	Admin1 has alternative places in other countries, excluded	Admin1 in the mentioned country	2
Not Nested	Two or more countries	Admin1 has several occurrences in several countries, not the one mentioned	Most occurred admin1 in the data set	4

Table 5.8. FHLI contains two types of location (admin1, country). No populated place is mentioned.

PPL(s)	Description		Location	Unc
Nested	One country	One PPL with one occurrence in one country	Mentioned PPL	2
		Several PPLs all nested within one country	The PPL that encompasses the others	3
Not nested	One country	Several PPLs all in one country	Most occurring PPL	3
	Two or more countries	Several PPLs in different countries	Most occurred PPL	4

Table 5.9. FHLI contains one type of location (populated place) only.

Admin1(s)	Description	Location	Unc
One country	One location in one country	Mentioned admin1	2
	Several locations in one country	Most occurred admin1 in data set	3
Two or more countries	Several admin1s in several countries	Most occurred admin1 in data set	4

Table 5.10. FHLI contains one type of location (admin1) only.

Country(s)	Description	Location	Unc
One country	One country is mentioned	Mentioned country	2
Two or more countries	Two or more countries are mentioned	Most occurred countries in data set	4

Table 5.11. FHLI contains one type of location (country).

In cases where there is insufficient information for identification of the nested locations, an alternative with the highest GeoNames hierarchy code is selected (Appendix 2). Examples of this are highlighted in Table 5.12.

FHLI Example	Alternatives	GeoNames Code	Default location
'London'	UK	PPLC	London in UK with Code PPLC (Capital of a country)
	17 places in US	PPL	
	Canada	PPL	
	Brazil	PPL	
'Italy'	Italy	Country	Italy with Country code
	China	PPL	
'IN'	India	Country	India with Country code
	Indiana	Admin1	

Table 5.12. Examples of GeoNames hierarchy precedence.

An uncertainty number is assigned to each FHLI. The uncertainty is quantified according to the number of candidate cities for each FHLI. It also reflects the confidence in selecting a default location as a representative of the examined FHLI. As briefly mentioned in (section 5.3.2) the uncertainty here has undergone some changes in response to the findings in chapter 6 (section

6.2.6). Subsequently, the uncertainty numbers are based on candidate cities that can be associated to the FHLI. According to the analysis of classification and uncertainty measurements (chapter 6: section 6.2.6 and 6.5.5) 'City' has been used as a unit of measurement. Examples of the FHLI and their uncertainties are summarized in the following table.

FHLI Examples	Example	Uncertainty
One city	Islington, London, UK	2
Several cities within same country	London, Brighton, UK	3
Several cities in several countries	Tehran, London, UK	4
Unknown	Damned Village, Sunny Place	5

Table 5.13. Uncertainty numbers reflect confidence in assigning default location for each FHLI.

A Flickr GB poster with a hypothetical name of 'Helen' with FHLI as 'Sydney, New South Wales, Australia' can be used as an example to demonstrate how the developed algorithm works. Since 'Helen' has her home location available in her profile she is categorized as have-home posters (section 4.2.2). Her FHLI is being disambiguated according to the developed disambiguation algorithm in this section. Her FHLI is in the category of feature code combinations demonstrated in table (5.5). 'Sydney' is PPL (populated place feature code) nested in New South Wales (Admin1) that is nested in Australia (Country). Accordingly, the first combination of nested locations in one country is best matched with Helen's FHLI. Referring to the table (5.5) the algorithm assigns the most nested PPL as a default location for the examined FHLI. In that case the disambiguated home location is 'Sydney' with uncertainty 2 (one place within one country, section 5.3.1).

5.3.2.3 Evaluation of the Disambiguation Algorithm

The artificial intelligence still has a long way to go in order to automate the process of understanding the natural language of humans (Erik Rauch, 2003). Human beings have a powerful ability to react in vague and ambiguous situations by applying the fuzzy rules combined with real world experience and knowledge in the right time and place (Fisher, 2007; Rauch, 2003). In developing the disambiguation algorithm in this study attempts were made to create an automated application by implementing the steps that the author followed to disambiguate the Flickr sample world data manually (sections 5.2.2-3). The implemented algorithm tries to imitate the power of the human mind in decision-making according to common sense and real world knowledge. As mentioned in section (5.2.1) specifically, and discussed in general in (section 2.5), predetermining a default location for each ambiguous place name has been done in regards to different metadata:

- The most commonly occurring place (Smith and Mann, 2003)
- Population of the place name (Rauch et al. 2003)
- Semi-automatic extraction from the web (Li et al. 2003)

The disambiguation algorithm in this study is unique in that it does not predetermine a default location for each ambiguous place name. It deals with each FHLI individually without any fixed default location selected beforehand. For each FHLI, it disambiguates the phrase according to the combination of the spatial information provided (Tables 5.6-11). According to the combination types of the provided spatial information, number of potential candidates and the hierarchy depth of the Geo-Names' feature codes, it assigns the best possible location as an indicative of the studied FHLI. In addition to that it also assigns an uncertainty number (Table 5.13) to the final disambiguated location depending on the levels of uncertainty involved (number of potential candidates and combination types). The developed disambiguation algorithm has been applied to the Flickr sample GB data. Its performance was assessed against vague classes indentified in FHLI (Table 5.1). The results are summarized in Table 5.14.

Vague classes	Successfully Disambiguated	Vague Classes	Successfully Disambiguated
Does not exist	Yes	Abbreviations	Yes/No
Multiple Alternatives	Yes	Misspelling	No
Multiple Entities	Yes	Descriptive	No

Table 5.14. Assessment result of the disambiguated algorithm against the 6 classes of vague terms in Flickr (Table 5.1).

The three classes of vague terms that failed to be disambiguated successfully by the developed algorithm are highlighted in Table 5.15.

Vague Classes	Example	GeoNames	Disambiguated Successfully
Abbreviations	Formal 'USA' -> United States of America	Yes	Yes
	Informal 'U of A' -> United States of America	No	No
Misspelling	Typo 'Dubliln' 'United Lingdom'	No	No
	Intentional typo 'En ger land' 'Brizzzzzzle'	No	No
	Other languages 'Turkiye' 'Spania'	No	No
Descriptive	Absence of any delimiters/separateors to split the FHLI 'Brighton Sussex UK' or ' I live in dirty dirty Leeds'	No	No

Table 5.15. Vague classes of FHLI that were not disambiguated successfully with the disambiguation algorithm. The third column indicates if the GeoNames database contains occurrences of the mentioned types.

All occurrences of the above classes in FHLI were disambiguated manually. These only equate to 4.3 % of the FHLI that are categorized in the above three classes in Table 5.15. The cases that the developed algorithm failed to disambiguate successfully are described in detail in the next section.

5.3.2.4 Failure Cases and Recommendations

To the best of the author's knowledge the disambiguation application developed and applied in this study is the first disambiguation process that successfully deals with the ambiguous, vernacular geographic terms in an unedited and un-formatted text with no single context. Moreover, its performance is not restricted to place names of a specific city or country. The following table summarizes the cases in which the algorithm fails to detect the expected default location.

Failure Case		Examples
Description	Spatial	'When not travelling London', 'Dirt dirty Leeds,
	Non-Spatial	'Travelling widely', 'Where the stars come out at nights,
Latin/Local Names		'Italia', 'Turkiye', 'Spania', Oztralia'
Misspelling	Typo	'United Lingdon', 'Dubliln'
	Emphasis	'Brizzzzzle', 'En ger land'
Slang		'Bortsm'f', 'U of A'
No Delimites[35]		'Brighton East Sussex UK'
Numbers		'31', '514414'
Inconsistent Titles for Countries		'Republic of', 'Federal District of', 'Federal Territory'
Post Codes		'SE5', '9011'

Table 5.16. Failure cases of the disambiguation application.

The following three failure cases can be improved by modifying the disambiguation algorithm:

- **Description with spatial information (e.g. 'When not travelling, London')**
  Splitting the FHI can be done with two different separators (other than the delimiters only). It can also be split by space. The rest of the algorithm can remain the same and should be run once for each separator. By doing this, at the end of the process there will be two default locations according to the results of each separator. If the resultant locations are not the same then the one with less uncertainty is selected. This solution, although it can improve the performance of the application, reduces the efficiency of the code and doubles the processing time for dealing with cases that might occupy less than 5% of the locations.

- **No delimiters (e.g. 'Brighton East Sussex UK')**
  As per the solution recommended above, here again each phrase can be split with a space (in addition to delimiters) and then it follows the same steps of the developed algorithm. This also puts extra burden on the processing procedure, especially in cases where there is no location

[35] Since the developed algorithm replaces any delimiters with commas, those FHLI that do not use any symbol for separating locations names have failed to disambiguate successfully.

information in the phrase (e.g. 'All around the world', 'Travelling every where').

- **Misspelling and Emphasis (e.g. 'Dubliln')**
  Making a less strict comparison for finding places with the same name can disambiguate this case relatively well. The potential candidate for a name can be found by finding the best match in the database and not necessarily an exact match. However, it increases the potential places that can have relatively the same name, therefore making the process of selecting a final default location more complicated than the existing algorithm.

  The following cases (table 5.16) might be improved to some extent by using some other geo databases[36] in addition to Geo-Names' files:

  - Latin/local spellings
  - Slang
  - Numbers
  - Titles
  - Postcodes

  Connecting to more than one database for place names and searching for potential places with the same name in several files will reduce the efficiency of the application. Therefore, this should be applied only in cases where ambiguous places occupy a high percentage of the data, which is not the case for the GB data of this study.

5.3.2.5 Justification

Referring back to the tables (5.5-13) indicates the logic and rules that applied in the disambiguation algorithm. Here briefly, reviews the algorithm and justifies the decisions made.

Overall, according to the possible combination of the three location types a decision is made to assign a disambiguated location to the examined FHLI. In general, the most nested locations (based on Geo Names' hierarchy) or the most occurred locations in the data set have been given higher probability of being classified as disambiguated home location. The justification for this decision comes from a simple fact. The most nested location in FHLI is the most precise place that can be associated to home location of a person. For those of places that there is not sufficient information to identify a nested location among several alternatives, the most occurred place in the data set is selected (for example several places within the same country). This is also based on the assumption that the most occurred locations come from higher Geo Names' hierarchy (Table 5.12). This seems to accord with the assumption that most occurred locations within the data set are those with higher population. In FHLIs where there are several place names in different countries, the one that comes with its hierarchy is selected for default location. For example, FHLI as 'Jerusalem, London, UK' is disambiguated to

[36] http://www.mapsofworld.com/
http://www.worldatlas.com/
http://www.travelmath.com/places,
http://www.infoplease.com/countries.html

London capital of UK. An uncertainty number is used to reflect the uncertainty in making this assignment. This rule also can be explained by the assumption that the location that comes with its next feature code (London, UK) is given precedence over other place names in the same FHLI that have insufficient information for the identification of their locations. The disambiguation algorithm and its performance have been assessed against its ability to disambiguate the FHLI of GB data set (described and collected in section 4.3.2). In the first step, the algorithm failure cases were compared to the identified classification of vague terms in Flickr (Table 5.1). The results are summarized in Tables (5.14-15). Referring back to the data indicates that only 4.6% of the FHLI of GB posters come from classes of vague terms that the algorithm failed to disambiguate successfully. In addition to that, as demonstrated in Figure (6.33) the majority of the FHLI are either blank or within uncertainty 2. The blank category is exempt from disambiguation and those places that have been disambiguated with uncertainty 2 are the most successfully disambiguated ones. Studying the precision classification (Figure 6.15) together with uncertainty categories (Figure 6.33) implies that the FHLIs at city level occupy the majority of the disambiguated FHLIs with Uncertainty 2. The remaining FHLI (14%) have been disambiguated with uncertainty 3 or 4. Theses are FHLIs that can be associated to several places in one (UNC3) or several countries (UNC4). 4.6% of this category are FHLIs that have been failed to be disambiguated successfully. Overall, considering the percentage of the FHLI that come in the combination types that fail to be disambiguated successfully in line with the uncertainty categories (Figure 6.33) indicates that only 14% of the data are assigned with higher uncertainty (UNC2-3). Considering the percentage of the data that this category occupies in the whole data set one can conclude that the developed algorithm can successfully disambiguate the FHLI of GB posters within reasonable level of certainty.

However, the method applied here can be criticized on the basis that the author is the only user who evaluated the disambiguated FHLIs. As a result here comes the subject of ‘reliability of agreements’. ‘Cohen Kappa’ and ‘Fleiss Kappa’ have been introduced in literature as statistical measures for assessing the agreement among group of participants who assign ranks to or classify items (Sim and Wright, 2005; Gwet, 2010). Fleiss Kappa is recommended here to enable the experiment to include several participants[37] (probably 3-4). An indicative sample of FHLIs in regards to classes of vague terms (Table 5.1) and combination of feature codes (Tables 5.5-11) can be given to the algorithm for disambiguation. The output is a test data that can be assessed by several participants (including the author). The results of this test will show the degree to which the amount of agreements noticed would exceed what would be expected if all participants were to do their ratings completely randomly. This result can be useful in evaluation and studying the effects of assessing the disambiguated locations by the author only.

[37] Cohen Kappa is designed to assess the agreements between two participants only.

5.4 Visualization

This section describes the methods that have been set out for achieving the visualization of the formatted and disambiguated data. The whole process is broken down into the following two major tasks:

- Identify appropriate visualization package
- Visualization methods and design decisions

5.4.1 Visualization Packages

Since this research aims to study the spatio-social relations in a spatially structured social group, it involves the consideration of the social network analysis from the perspective of visualization and geography (chapter 2: sections 2.3 and 2.4). Therefore, potential kinds of available packages suitable for this study were limited to the following, with capabilities of visualization of spatial social network data:

- Social network analysis packages
- General visualization packages

Alternative sites from each of the above categories were selected as representative subsets of the options available. Attempts have been devoted to make the number of alternatives selected from each category to be proportional to the total amount of available options. Consequently, three social network packages (*Pajek, Ucinet, Many Eyes*), and three general-purpose visualization toolkits (*prefuse, processing, and improvise*), were selected for further study and analysis.

Assessment criteria are selected precisely based on the research requirements and are essential for the suitability of the toolkits. All the findings and conclusions are based on demos, tutorials, existing applications and manuals available for each package. Where necessary the developers were contacted to ensure the validity of the assessments. The findings also benefited from the comments, discussions and feedbacks received from students, researchers and those who are also involved in using the packages.

5.4.1.1 Assessment Criteria

The following eight criteria are developed that sound necessary according to research requirements. Each criterion is assessed qualitatively and rated in Table 5.17 according to the guide summarised in table 5.18.

- ***Criterion 1(C1) Visualizing Relational Data***: Since social networks are built up with relational data it is essential for the package to support visualization of that kind of data (e.g. adjacency matrix).
- ***Criterion 2(C2) Visualizing Spatial Data***: This criterion assesses the ability of the package to visualise geographical data, e.g. (lat, lon). Since this study will work with VGI, assigning location to nodes in the networks is one of the fundamental tasks that need to be conducted successfully.
- ***Criterion 3 (C3) Calculate Social Network Properties***: This criterion examines the capability of the package in measuring social properties of

the networks. This might be useful in study and analysis of the spatio-social data in later steps of this research.

- ***Criterion 4 (C4) Calculate Statistical Properties of Data****:* This criterion tests the ability of the package to calculate and conduct some statistical analysis, e.g. standard deviation. This feature sounds useful for testing possible hypotheses and comparing results in different networks.
- ***Criterion 5 (C5) Visualizing Large Networks***: This criterion examines the maximum number of nodes that the software can process. This feature is necessary since this study will involve visualizing large spatio-social data sets.
- ***Criterion 6 (C6) Interactivity****:* This criterion is defined as the ability of the network to develop interactive visualization in order to enable exploratory analysis of VGI.
- ***Criterion 7 (C7) Flexibility****:* This criterion is set to test the ability of the package to provide the developer with flexibility in design. Since large spatio-social data sets will be the focus of this study, possible future design decisions and possibilities are heavily reliant upon the flexibility of the application.
- ***Criterion 8 (C8) Learning Support****:* The last but not the least feature that will be considered for the package selection is the availability of learning support (e.g. books, tutorials, user manuals or API).

5.4.1.2 Alternatives Assessed

1. Many Eyes

Authors: Wattenberg, Viégas, Ham, Kriss, McKeon, IBM's Visual Communication Lab

Project Webpage: http://services.alphaworks.ibm.com/manyeyes/home

Evaluated Version: 2004

Requirements: Internet explorer

Overview: Democratize Visualization, Makes Visualization Technologies Accessible to Everyone.

According to the developers the aim of *Many Eyes* is to standardize visualization, enabling anyone on the Internet to create strong interactive visualizations and start communicating about their own visualizations with others. Accordingly, it has also been referred to as a social visualization package[38]. The developers believe that sharing and discussing visualizations could investigate unexpected and hidden patterns. They provide users with a new kind of data analysis, which is achieved through sharing and discussing visual patterns with others. In other words the package develops a data analysis method in social settings. *Many* Eyes is an online social environment where anyone can share their visualizations with others. The site contains more than 6,000 visualization samples (May 2008) each with a separate discussion forum. Therefore members can discuss share and exchange ideas in order to gain common insight about the existing patterns.

[38] Social visualization package is a visualization toolkit that enables a new social kind of data analysis. A discussion forum is assigned to each visualization created by the package. The aim is to understand and examine data sets through online interactions and discussions. http://manyeyes.alphaworks.ibm.com/manyeyes/

C1: The software supports visualization of relational data by drawing graphs. It shows the relationships among data points by either scatter plot or network diagram. It can read only one type of data. *Many Eyes* data sets should only be transferred to table format. The first row in the table should be headers describing columns and the rest should be filled with values separated by tabs.

C2: Although it supports network and graph drawing it does not have the functionality to locate the nodes exactly on the spot according to X, Y coordinates intended by users. Therefore, it is unable to visualize spatial data. But in a high level fashion it is possible to overlay data values on geographic regions. It only supports a limited number of countries according to their regions or territories. Although it can fill regions with different colors or different sized bubbles, there is no chance to depict the connectivity among entities in the same or different regions.

C3 and C4: The software does not possess any functions or routines for social measures or statistical analysis. Therefore, it will not be appropriate in future analysis and hypotheses testing of the study.

C5: It is very time consuming to develop *Many Eyes* applications with data sets embracing more than a thousand vertices. The optimized layout could not work efficiently in networks with a high density. In addition the software could not process data sets bigger than 5 megabytes.

C6: The package provides some fixed options for the user to choose from. Therefore, users can see the same data set in different views with basic changes, e.g. shape or colour of nodes. The developed applications have a limited number of interactive options:

- Zoom and pan in Networks and Graphs
- Search options in Word Tree and Tag Cloud and
- Colour or bubble in geographic maps

All these options are automatically allocated to applications according to their categories and users are not in the position to delete or edit them. Accordingly, the developer is very much limited to these options and it is not possible to add new features to the viewing options.

C7: The extendibility and flexibility of the developed applications rely on the future development of the package. All *Many Eyes* visualizations come from limited views with single formatted data sets. In other words it is to a great extent inflexible and incapable of accepting changes. Therefore the fixed applications are vulnerable against future possible extensions and necessary changes of the research. It requires the user to upload the data set and it does the visualization job itself without giving the developer any chance to change any feature of the representation.

C8: The package is easy to use and there is an extensive tutorial which helps beginners to do the visualizations.

2. Prefuse

Author: Heer, UC Berkeley
Project Webpage: http://prefuse.org
Evaluated Version: beta, July 15, 2006
Requirements: Java 1.4

Overview: Data Modeling, Visualization and Interaction, Animation and Rendering Support

Prefuse is a user interface program for creating interactive information visualization applications. It is a flexible visualization toolkit for developing customized data visualization applications using Java programming language. It is specifically designed for development of sophisticated, highly interactive and flexible information visualization. The package provides a large number of different classes each with a few functionalities. Those separate classes are joined together with object-oriented techniques. Consequently developers can customize Java codes to implement their needed functionality.

C1: Apparently *Prefuse* API documentation contains classes and methods for visualization of graphs and networks (e.g. Graph, GraphDistanceFilter, GraphicsLib, GraphLib, GraphListener and GraphMLReader). Accordingly, the package supports network and graph drawing and is capable of handling relational data.

C2: *Prefuse* provides classes and methods to assign and retrieve location from nodes on the page (e.g. AxisLayout, AxisLableLayout). It also contains a specified layout algorithm that allows developers to define how to draw the networks. Zone manager[39], an example application developed by *Prefuse*, proves the ability of the package to demonstrate the spatial information.

C3: As a visualization toolkit with no emphasis on visualizing relational data, there are no classes for calculating social measures.

C4: Statistical analysis methods are not included in the software library. But since it is a Java programming language statistical routines developed in standard Java API[40] can be used in the package.

C5: According to the software manual it supports visualization of data sets even if too large to fit in the memory.

C6: As a Java based visualization toolkit the user can take the advantage of Java programming functionalities in order to develop applications with a high level of interactivity.

C7: Since each feature and function of the visualization has been developed by Java codes with *Prefuse* API and Java standard API, the applications can be changed and extended according to future requirements of the research. *Prefuse* creates visualizations with high flexibility and extendibility.

C8: Although the user manual of the package is incomplete, it has a discussion forum for user support.

3. Improvise

Author: Weaver, Penn State University
Project Webpage:
http://www.personal.psu.edu/faculty/c/e/cew15/improvise/introduction.html
Evaluated Version: Version 2

Requirements: Java 2 (1.4.2).

Overview: Highly Coordinated, Multi View and Open Ended Visualization

[39] http://goosebumps4all.net/profusians/wiki/Zonemanager

[40] http://java.sun.com/j2se/1.4.2/docs/api/

Improvise is about modeling, creating and analysing multiple data sets in highly interactive environments. *Improvise* visualizations contain an interactive combination of data, views and coordinates. The main goal of the software is to improve coordination flexibility both in simple and complex data exploration and analysis. The package is capable of visualizing different number of related data sets in different linked views all contained in a single visualization.

The main difference of *Improvise* to the other examined packages is that it produces dynamically linked and coordinated views in a single visualization. Users can build and browse relational data by pre-defined coordination. Therefore, a subset of a large database or a group of related data sets can be visualized in different views. As a result of this architecture comparison between different sets of data would be possible without switching between different views.

According to the developer *Improvise* is a research grade software without any specific direction to fit into a particular purpose. Although some papers have been written on the software architecture it still lacks a manual, tutorial or even a brief description of how to use the software. As such, the assessment of the following criteria is merely based on the examples of the software application available on the package webpage.

C1: The software supports basic elements of graph drawing based on imported relational data. But the developed networks lack some of the essential functionalities e.g. different sizes for nodes, assigning weight to edges and applying a layout algorithm, such as force directed algorithm, etc.

C2: The software lacks any functions to attribute (X, Y, Z) coordinates to nodes. It is possible to assign data to specific areas on the map but connectivity between nodes could not be depicted on the geographical maps through the existing views.

C3 and C4: The software does not have any functions to measure social attributes of networks. Since it provides multi view visualization of data in different level of details it encourages comparison and the developing of possible hypotheses. However, it lacks statistical routines for testing possible developed hypotheses.

C5: While the software is capable of visualization of large amounts of data simultaneously, no emphasis has been put on the maximum number of the nodes that the software can handle in a single graph. This might be due to the fact that *Improvise* is not specifically designed for network and graph visualization.

C6: In terms of interactivity the software is mainly focused on interactive building and browsing the highly coordinated visualizations. Users normally start from a view and they are in the position to see the effects of the changes they made in multiple linked coordinated views.

C7: As an undocumented package with the codes available under GPL (General Public License), assessing the flexibility and extendibility of the *Improvise* applications sounds to be a demanding and unfeasible task. As can be inferred from the existing examples, the package requires some developing codes, while providing ready visualization views.

C8: There is no established learning support for *Improvise* users.

4. Processing

Authors: Fry and Reas, 2001, ACG (Aesthetic and Computation Group)
Project Webpage: http://processing.org

Evaluated Version: Processing 1.0 (Beta)

Requirements: Java 1.4 and earlier (Included in the package)

Overview: Simplified Java Based Programming Package for Complicated Visual and Conceptual Structures.

Processing is a programming language and environment built for the electronic arts and visual design communities, e.g. artists and designers. It is an open source software package mainly focused on graphical issues of drawing a variety of shapes, images and animation. The *Processing* environment is written in Java. It provides users with a graphics library and a special programming approach that simplifies most of the advanced concepts of object-oriented programming. Overall it is a simple programming language environment while still providing sophisticated graphical functionality and a reasonable amount of flexibility for advanced users.

C1: There is a "social network library" in the core library of the package, contributed to the software by Todd Holloway[41]. It includes network library element stressing methods for drawing and exploring networks based on the relational data provided.

C2: Coordinate class in the extended library of the package enables the user to locate the points to specific location. Therefore, the package is capable of visualizing spatial data.

C3: While it does not include any method and function for calculating social measures it possesses a "PajekReader" class which accepts I/O files in Pajek network format. Therefore, Pajek could conduct all the social measure calculations. This brings the opportunity to take the advantage of both programs.

C4: The package doses not have any function and method for calculating statistical analysis in its library. But as a Java based programming visualization toolkit, it has the advantage of applying the statistical classes and methods available in Java standard API.

C5: While there are some concerns regarding the time efficiency and visualization options as the number of nodes increases, the package does not include any limitation on the amount of data that it could process.

C6: Since it is mainly designed for artistic image programming and animation, most of the applications are highly interactive.

C7: As for the flexibility and future extendibility of the developed applications, *Processing* takes the advantage of Java programming and Java API which simplifies the complicated concept of object-oriented programming. Therefore, users are provided with ready and pre-written classes and methods in the software API documentation, and also in comprehensive Java API specification

[41] PhD Student, Computer Science Department, Indiana University. Available at http://ella.slis.indiana.edu/~tohollow/SocialNetworksLibrary/ Accessed January 2009.

available on the web. Therefore *Processing* applications are flexible in terms of necessary future changes in the research requirements

C8: The package is supported by discussion forum, tutorials, code examples and number of books elaborating on how to use the package, from basic applications to advanced visualization.

5. Pajek

Authors: Batagelj, Mrvar, University of Ljubljana.

Project Webpage: http://vlado.fmf.unilj.si/pub/networks/pajek/

Evaluated Version: Pajek 1.21

Requirements: Windows 2000/XP or Linux

Overview: Visualization of Large Social Networks

Pajek (spider in Slovenian language) is a program for visualization, exploration and analysis of large networks (more than one million nodes). The main goal of the software is to develop specifically designed programs to handle very large data sets. In doing so the software provides users with powerful visualization tools. *Pajek* has been under development since 1996. It is a menu driven program. The algorithms used in the package are specifically designed to handle large data sets. *Pajek*'s visualizations are based on six data structures: networks, partitions, permutations, clusters, hierarchies and vectors. It provides manipulation options for all the above-mentioned data structures.

C1: *Pajek* is specially designed for network analysis; therefore the relational data is the only acceptable data format for visualization.

C2: The package itself is not capable of visualizing spatial data. Recently a new feature called “ESRI GIS/Pajek Interface” has been proposed to be added to *Pajek* by Douglas White. It combines *Pajek* SVG networks with bitmap background pictures in GIS. Therefore, nodes could be laid on the page according to their assigned X, Y coordinates in a new interactive user interface. It is an ongoing project and, as yet, is by no means complete (May 2008).

C3: The package includes a comprehensive menu for calculating different social network attributes (e.g. centrality, betweenness, geodesic path, etc.) straightforwardly.

C4: Although *Pajek* itself only supports a few simple statistical routines (i.e. Median and Standard deviation) the statistical package R could also be used within *Pajek* data structures.

C5: The package is capable of processing more than one million nodes, as stated earlier.

C6: The software does not support creation of highly interactive visualizations.

C7: Since it is a high level menu driven program the future extendibility of the developed applications is limited to the development of the whole package.

C8. The *Pajek* developers have a book that includes comprehensive tutorials and discussions regarding the software functionalities.

6. Ucinet

Authors: Borgatti, Everett and Freeman

Project Webpage: http://www.analytictech.com/ucinet/ucinet.htm

Evaluated Version: Ucinet 6.

Requirements: Windows 95/98/NT/2000/XP

Overview: Analysis of Social Network Data

Ucinet is an extensive program for the analysis of relational data in a social context. It includes a number of network analytic routines for exploration of proximity data. It contains comprehensive explanatory network analysis procedures including social measures and statistical routines. Integrated with *Ucinet* is another program called *NetDraw* that provides visualization. It also has export options to *Mage* and *Pajek* programs which could also provide the visualization. Applying one of the above mentioned programs enhances the software capabilities to handle multiple relations at the same time, e.g. selecting subsets, merging data sets, assigning colours, shapes and sizes according to attributes of the nodes, etc.

C1: Although it is specially designed for network analysis it does not include any methods for visualizing relational data in graph format. It uses an external program *(NetDraw)* to visualize the relational data.

C2: Since the main goal of the program is to provide users with comprehensive network analytic routines, the software is not powerful in graphical depictions of data. Therefore, the ability of the package to visualise spatial data is apparently limited to the packages that could be selected for the visualization part of the network analysis. Normally *NetDraw* is used to draw the networks' diagrams. It can locate nodes on the page in accordance with the specified X, Y coordinates. Longitude and latitude attributes can be saved in the nodes' attributes files. And thereafter nodes could be laid out on the page according to their geographic coordinates.

C3: *Ucinet* provides comprehensive network analytic routines. Therefore, social measures of the networks could be calculated in a straightforward process. Detecting cliques, clans, components, cores, centrality analysis, structural holes, betweenness, geodesic path, etc. are automatically calculated and could be accessed through a log file.

C4: Various statistical routines for one or more variables are available in the package, such as p1 model, autocorrelation methods, QAP (Quadratic Assignment Procedure) correlations and permutation tests and the like. In addition, the package has strong matrix analysis routines, such as matrix algebra.

C5: *Ucinet* can handle maximum of 32,767 nodes in a graph, although some of the procedures get too slow for networks with more than 5,000 nodes. Considering the fact that this study will involve the visualization of large VGI, this limitation stands out significantly.

C6: *Ucinet* is not designed for developing interactive visualization applications.

C7: Since it is a menu driven program the extendibility of the application is limited to the functionality of the package. The user cannot extend the existing applications beyond the available options in the package menu. It is a high level package with emphasis on providing users with a straightforward process of statistical analysis of network data.

C8: The package has a comprehensive user manual but it lacks online support.

5.4.1.3 Conclusion

As can be inferred from the evaluation conducted above and Table 5.18, two low level programming languages, *Prefuse* and *Processing*, are ranked positively for most of the examined attributes. The only minor difference is that *Prefuse* does not measure the social properties and is not capable of accepting files from a social network package, while *Processing* can do this successfully. According to the results found by this report and also the existing demos and available applications, *Processing* could be ranked first and *Prefuse* second for meeting the research requirements. The two social network analysis packages, *Ucinet* and *Pajek*, stand nearly at the same level. *Pajek* scores are slightly more promising than *Ucinet*. Consequently they could be ranked as third and fourth suitable packages respectively.

Improvise, with positive scores just for relational and spatial data set and interactivity features, is ranked fifth. And finally the least applicable toolkit to satisfy the research requirements is *Many Eyes*. It can hardly meet two of the essential required features.

The above assessments left the researcher with two options:

- *Prefuse*
- *Processing*

In order to justify the decision it should be mentioned that a productive, easy to use package that is prone to failure in satisfying the future unexpected requirements of this study is considered as inappropriate. Therefore, the decision is made to assess the productivity (user friendliness) against interactivity and flexibility. Consequently in the following evaluation, interactivity and flexibility criteria are prioritized over the other assessed criteria. According to the nature of the research that deals with visualization of large spatio-social data accessible through VGI in addition to the flexibility and interactivity (6th and 7 criteria respectively) visualization of large networks (4th criterion) is also given precedence in assessments.

Prefuse and *Processing,* although they have shortcomings, provide wide and flexible capabilities and attributes that lend themselves to fulfill the requirements of the proposed research. They both provide a high number of different classes, each with specific functionality, and all joined together with object-oriented techniques. As a result of this architecture, both packages exploit the power of object-oriented designs to share, reuse and extend the existing classes in libraries. Consequently users can customise Java codes to implement the desired functionalities only. The packages are not easy to use and they require some adapting work to develop a well thought out, clear, flexible and elaborate application. They give flexibility to application functionality and minimise the risk of possible limitations in later steps of the research. Therefore, the future unexpected requirements of the proposed research can be satisfied with minimum effort. In Table 5.18 the results of the above evaluation are summarized by the followings symbols:

++

+

-

+/-

?

The above symbols, although used in their conventional meaning, also convey specific meanings regarding the assessed criteria that are elaborated in the following table.

Criterion	++	+	-	+/-	?
C1: Relational Data	NA	Capable	Incapable	Incapable, can accept Output from other packages	NA
C2: Spatial Data	NA	Capable	Incapable	Incapable, can accept Output from other packages	NA
C3: Social Networks Properties	NA	Capable	Incapable	NA	NA
C4: Statistical Properties	NA	Capable	Incapable	NA	NA
C5: Large Network	NA	Capable	Incapable	NA	Not documented
C6: Interactivity	NA	Fully interactive	No interactive options	Limited interactivity options	NA
C7: Flexibility	NA	Low level	High level	High level environment providing features of Low level functionalities	NA
C8: Learning Support	More than one kind of support	One kind of support	No support	NA	NA

Table 5.17. A guide to qualitative results of the assessment.

Sites	C1	C2	C3	C4	C5	C6	C7	C8
Many Eyes	+	-	-	-	-	+/-	-	+
Prefuse	+	+	-	+	+	+	+	+
Processing	+	+	+	+	+	+	+	+
Improvise	+	+	-	-	?	+	-	-
Pajek	+	+/-	+	+	+	-	-	+
Ucinet	+/-	+/-	+	+	-	-	-	+

Table 5.18. Assessment results of the visualization packages.

As is summarised in Table 5.18, processing has been found as the most appropriate visualization package for study and analysis of the spatio-social data set of this study.

5.4.2 Visualization Methods

After disambiguating the FHLI successfully, and selecting a suitable visualization tool, this section describes the methods developed, modified and applied for visual analysis and synthesis of the large unstructured Flickr data that has been collected, formatted and disambiguated during the previous steps of this study. Visualization and visual analytics solutions (Thomas and Cook 2005) are applied in the design and redesigning of an interactive application. The aim is to develop an application that provides systematic analysis of Flickr spatio-social data in line with research questions (chapter 1: section 1.4). The ultimate goal is to be able to generalize the inferences and findings that might reveal useful patterns about spatio-social attributes of social entities and their associated VGI. According to the data attributes and considering the research questions, for each GB poster there are four different kinds of attributes for visualization and analysis:

- Poster's home locations
- Friends' home locations
- Poster's geo-tagged photos
- Friends' geo-tagged photos

All the above information is to be visualised in an interactive application that can result in confirming the expected and exploring the unexpected patterns and relations in spatio-social Flickr data. The application is expected to help in answering the research questions and final conclusions are to be generalised to broader sets of spatio-social data.

In the first step, before writing any code to put the data into visual presentation, a pen and paper sketch of the initial design (Figures 5.1 and 5.2) revealed that regardless of the interactivity level of the application, due to size of the data and variety of the spatio-social attributes to be examined, incorporating all data in one single map, although feasible, would be inefficient and would end in a congested layout.

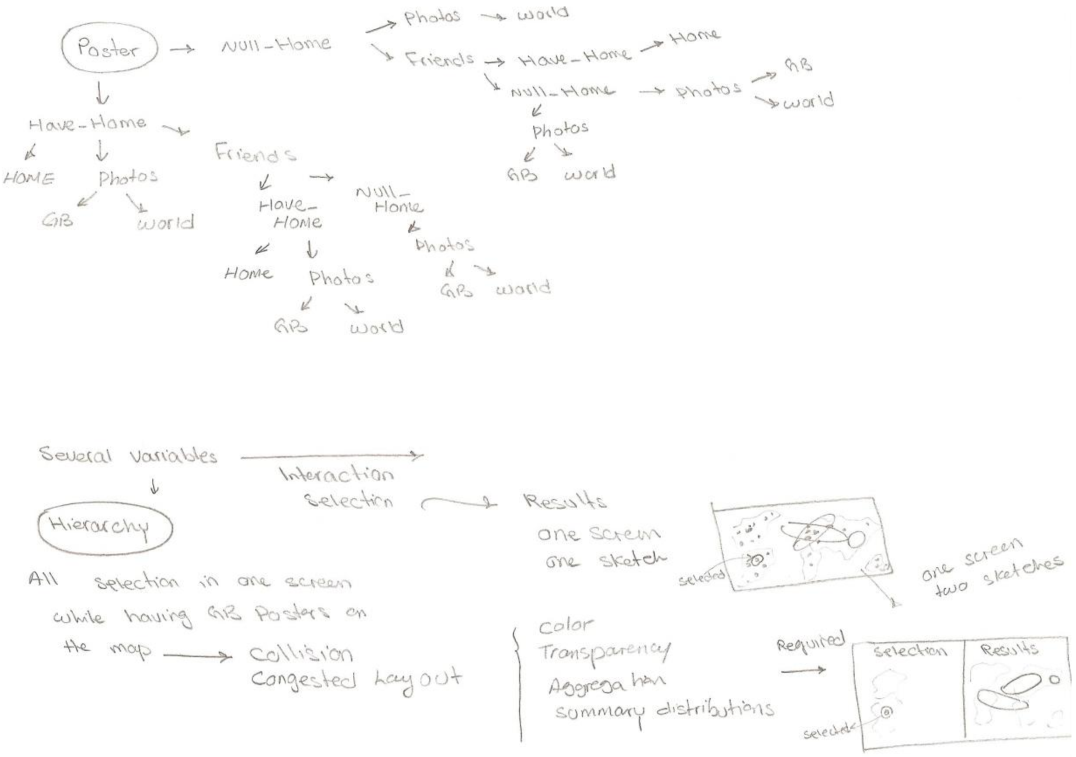


Figure 5.1. Spatio-social attributes of the data that were considered on paper before visualization.

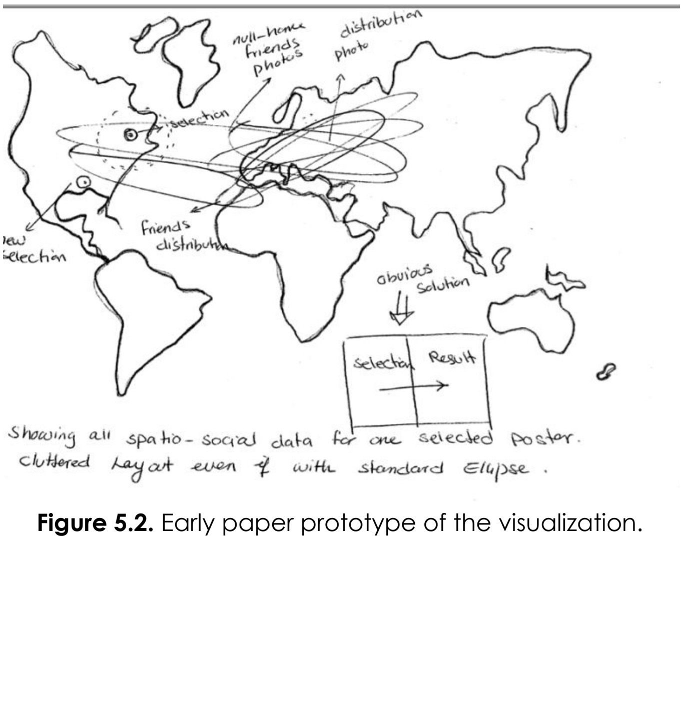


Figure 5.2. Early paper prototype of the visualization.

Figures (5.3-4) are the examples of the early attempts of visualizing the Flickr sample GB data. They demonstrate the volume of the data and level of complexity in efficient visualization for only one of the variables (geo-tagged photos). Consequently, having an application that is capable of visualizing several different spatio-social data in an interactive environment that can update the results according to users' interactions and selection proved to be a challenging task that requires several design decisions to be made.

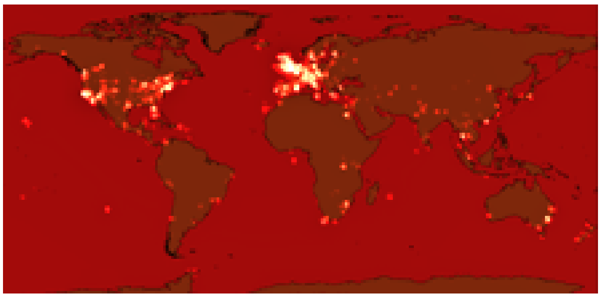


Figure 5.3. Early attempts for visualization of geo-tagged photos collected in Flickr sample GB data

Figure 5.3 shows the density of the uploaded photos in each area according to the brightness of the colour. As can be seen Europe has attracted the majority of the examined photos. Figure 5.4 visualizes the same sets of data but with different format. The points distribution are demonstrated with purple dots in each area. As can be seen in both visualizations Europe and US have the highest number of photos respectively. According to the size and volume of the data shown in Figures (5.3-4) study and analysis of the patterns of the required spatio-social data in this research needs more advanced method that can result in simpler layout for the data with this much density.

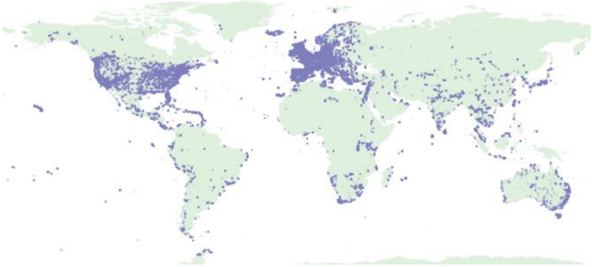


Figure 5.4. Early attempts for visualization of geo-tagged photos collected in Flickr sample GB data

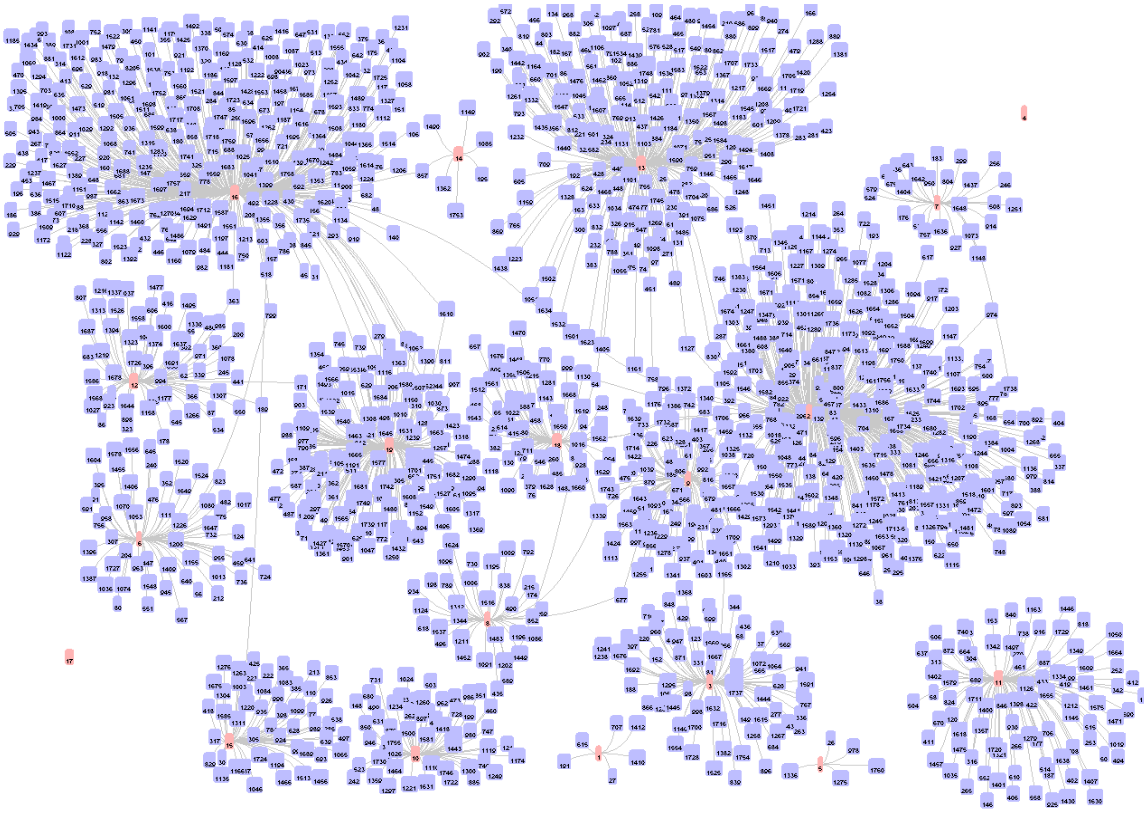


Figure 5.5 Graph layout of friendship network of 20 randomly selected GB posters.

Using the same data set the social network of 20 randomly selected Flickr members is shown in this picture as an indicative of typical patterns of Flickr friendship network. As can be seen the network is neither small world (most nodes are not neighbours but can be reached via short steps) no scale free (distribution of connectivity is extremely uneven. Some nodes act as very connected using a power law distribution) and even for only 20 members without any spatial information included and with one of the most effective force directed algorithms for social networks, the layout is cluttered. This confirms the fact that modelling and exploring the complex geography of large social network of this type need novel methods to be developed and new design decision to be made. Consequently, from the beginning of developing a suitable visual design, two maps were considered. The design of the desired application has been based on an original map that allows the user to interact with the data and a second map that reflects the result of user interactions with the application.

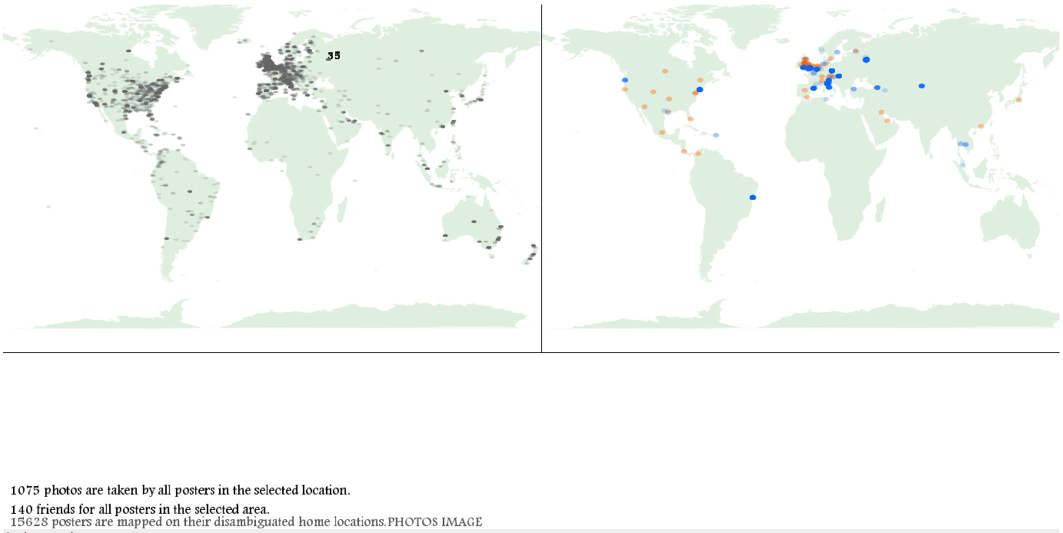
5.4.2.1 Visualization of Spatial Data

Accordingly, the first attempt to visualize the data is by plotting the have-home posters[42] according to their disambiguated home locations on the map. Since the aim of this application is to answer the research questions, and the research questions' emphasis is not on individual posters but on behaviour and attributes of posters who live in the same location, the prime focus of the application is set to allow the exploration and analysis of a group of posters who live in the same location. Therefore, each have-home poster is plotted on the map according to

[42] Those posters who have their home locations available in their profiles

the associated disambiguated FHLI. Considering the size of the data, number of posters and the fact that majority of the posters live in few specific areas, transparency is used to demonstrate the density of posters in each location (dark to light grey shades in left map, Figure 5.6).

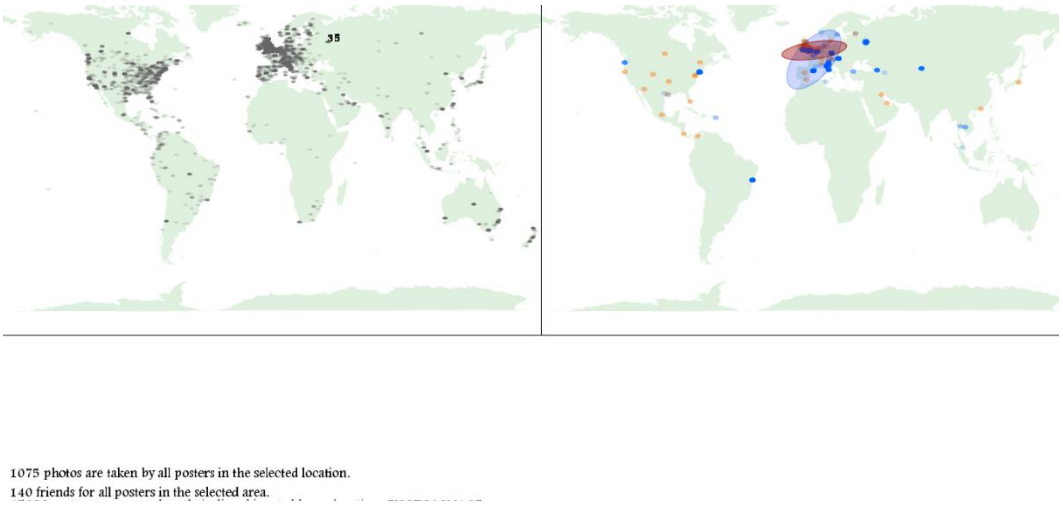
In the second step, in order to see the relation between the posters' home locations and the places that they take photos (section 8.2, RQ2), the photos taken by posters of the selected area are depicted in a separate map in the existing sketch (right map on Figure 5.6). Therefore, the geographic distribution of photos of the selected posters on the left map can be depicted on the right map. The same can also be done to show the distribution of have-home friends of the posters in the selected location. Different colours (blue, orange) are used to differentiate the photos from friends respectively. By having both distributions (photos and friends) on the same map with different colours, it is possible to compare the places that posters take photos and the places that their friends live (section 8.2, RQ2-3). Figure 5.6 demonstrates a screenshot of the application.



1075 photos are taken by all posters in the selected location.
140 friends for all posters in the selected area.
15628 posters are mapped on their disambiguated home locations.PHOTOS IMAGE

Figure 5.6. Screenshot of the first design of the application. It shows the distribution of the 1075 photos (blue) and 140 friends in (orange) on the right map for the 35 posters who live in the selected area (Moscow, Russia). Both maps are in one single sketch.

As shown in Figure 5.6, in addition to transparency, the actual number of posters (35) who live in the selected area is also written on the original map (left map). The places that the posters of the selected location take photos, and where their friends live, are highlighted in blue and orange respectively in the result map (right map). Transparency is also applied to show, to a certain extent, the density of the numbers in each location. As can be seen in Figure 5.6, plotting large number of photos and friends on the map is not useful for demonstrating structure or trends of their distributions. Although the transparency can help, it is very difficult to see the direction, centre and most favourite places that posters have friends or take photos. Accordingly, a method has been sought to produce a visual summary of distributions of the spatial points that can help in getting overall insight about the point distributions. 'Standard Ellipse', or as referred to in spatial statistics 'Standard Deviational Ellipse' (Mitchell, 2005), has been found as an efficient way of summarizing the directional distribution of set of spatial points (photos and friends). In doing so, Standard Ellipse class of the ‘giCentre Utilities’[43] library is applied in the exiting visualization with processing (more description is provided in section 5.4.2.3).



1075 photos are taken by all posters in the selected location.
140 friends for all posters in the selected area.

Figure 5.7. Summary of the geographic distribution of 1,075 photos and 140 friends (blue and orange respectively) for 35 posters in the selected area (Moscow, Russia) are shown with standard ellipses. An area of interest can be interactively selected and details are provided on the right hand side map.

Summarizing the distributions (Figure 5.7) has improved the application considerably. The distributions centre and direction are now easily recognizable and can be compared against other attributes. The existing design at this stage works relatively well in not busy areas, like Africa or Asia, but selecting a specific location in busy areas, like GB in Europe and several cities in the USA, is not feasible due to the restricted space on the sketch and high number of dots on the area. Zooming option can help the situation to some extent. By zooming in on the busy areas, detailed specific locations can be selected. However, the results of the selection (friends and photos distributions) cannot be seen while the left map has been zoomed in. Switching between the zoomed in map for selection, and then seeing the result in the un-zoomed map, is inefficient and confusing. Keeping the relation between the two maps by zooming in/out is not straightforward. It can result in confusion, misinterpretation and the wrong conclusion. Therefore, the zooming functionality of the application should be independent of the sketch (window). Therefore, attempts were made to enable the application to apply the zooming to the maps independent of each other. In that case the left map can be zoomed for selection purposes, while the results can be seen on the second (un-zoomed) map.

Consequently, in redesigning the application this time two separate sketches were developed. In doing so, another class of the giCentre Utilities called multiple sketches[44] has been used. Accordingly, two sketches were used to accommodate the original (left map on Figures 5.6 and 5.7) and results map (right map on Figures 5.6 and 5.7) in two separate windows. In this case the maps can be zoomed in independently while still being able to communicate with each other. The new design allows zooming in busy areas of the original map to

[43] Developed by giCentre.org →http://www.gicentre.org/utils/

[44] http://www.gicentre.org/utils/multiWindow

select a location, and the result can be seen in the second map in a separate window. Although the right sketch is dependent upon the interactive selection of an area in the left sketch, the two sketches are separate and can be run independent of each other (Figure 5.8).

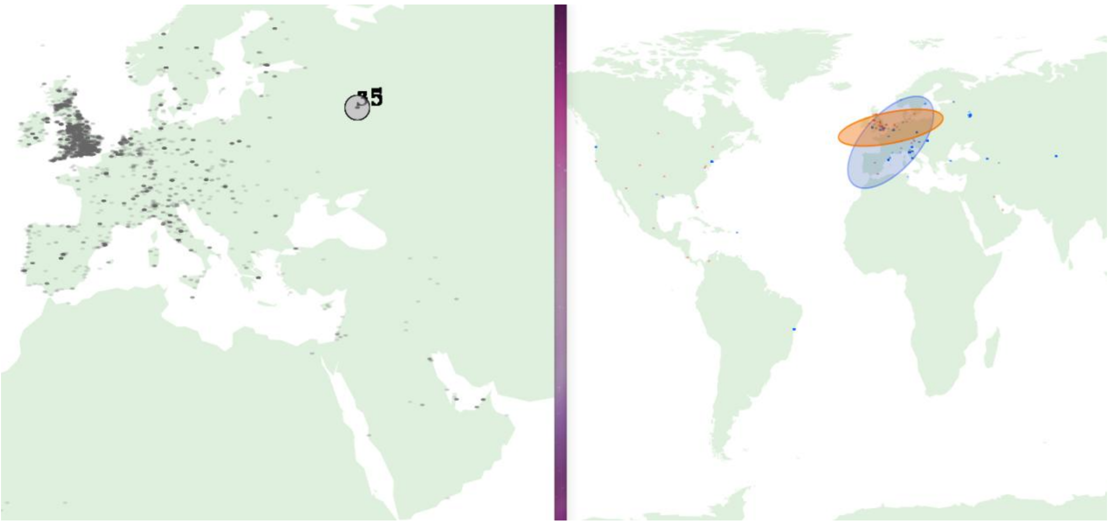


Figure 5.8. Standard Ellipse for the photos (Blue) and friends (Orange) of the 35 posters who live in the selected location (Moscow, Russia) in two separate sketches with independent zoom/pan functionalities.

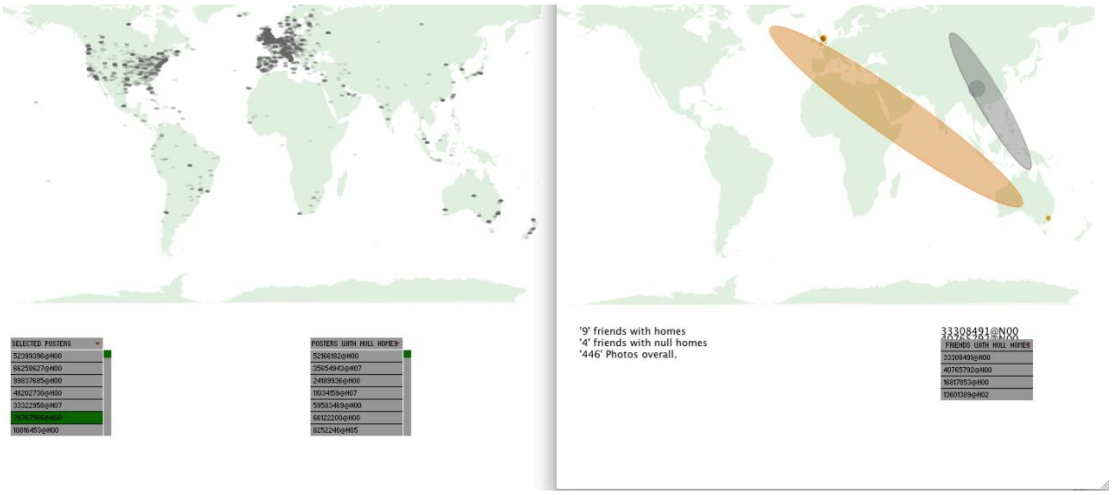
Up to this stage the spatial and relational information for the posters of a selected area can be successfully visualised in a separate sketch. Plotting the points on the map shows distributions of the photos and friends and summary of the distributions are depicted through standard ellipses (Figure 5.8).

Here, it should be noted that in the existing application the distribution of the photos and friends are demonstrated in aggregated view in which the results of all posters of the same location are summarized together. The next attempt is to analyse the geographic distribution of photos and friends for each poster of the selected area individually. The aim is to allow analysis and exploration of the social role and contribution of each poster individually, in addition to seeing the summary of all photos and friends for all posters of a selected area. In doing so a list box is added to the bottom left of the original sketch. For adding the GUI (Graphical User Interface) features to the existing processing application an additional library called 'ControlP5'[45] has been applied. The library has custom GUI features that can be interactively appear, hide and move while the processing sketch is running. The list box for each selected area includes the poster id of all posters in that location. By selecting each item of the List Box the geographic distribution of photos and friends for that poster will be shown on the map. By clicking a button the aggregated view can switch to individual view and vice versa.

As the Figure (5.8) shows the standard ellipses of photos and friends of the selected poster from the list box (highlighted in green) are drawn on the second sketch. The circles are sized according to number of photos and friends they are demonstrating (more explanation in section 5.4.2.3). As before, standard ellipses are used to show the summary of the points

[45] http://www.sojamo.de/libraries/controlP5/

distribution in each category. In this application one can easily compare how prolific and social the poster is compared with other users who live in the same location. Switching between the overall layout and the individual layout can be achieved through a click. At this stage, distributions of friends and photos for each selected location can be visualised in aggregated view as well as individual lay out. Switching between the layouts is easy and allows the user to study and compare the contribution of each poster towards achieving the aggregated view of photos and friends for each location.



SELECTED POSTERS
52399396@N00
66258627@N00
99837685@N00
48202730@N00
33322958@N07
70767566@N00
10816453@N00

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00
8252240@N05

'9' friends with homes
'4' friends with null homes
'446' Photos overall.

33308491@N00

FRIENDS WITH NULL HOMES
33308491@N00
40765792@N00
16817853@N00
13601389@N02

Figure 5.9. Photos and friends distributions of the selected individual from the list box (highlighted green) are demonstrated in grey and orange circles respectively

Figure 5.7 shows the ability of the application to navigate through the individual posters who live in the same location. The summary of the photos and friends of all posters of the selected location is shown in Figure 5.6. Here, the Figure 5.9 shows how individuals can be navigated in addition to the aggregated view. The selected poster from the list box is highlighted in green (left sketch). The photos and friends distributions are demonstrated in grey and orange respectively (right sketch). Switching between the aggregated view (Figure 5.11) and individual view (Figure 5.9) can be done by clicking a button. Consequently, the contribution of each poster towards the aggregated view can be studied easily through interaction with the application.

Step by step analysis of the application can be demonstrated through a Flickr member with Id number (7006877@N00, highlighted in green in the list box in left sketch Figure 5.9) and a hypothetical name as 'Helen'. As the application shows Helen has 9 have-home friends and 4 null-home friends (written in right sketch). The identified have-home friends are stretched across Australia and UK (orange ellipse). The grey ellipse shows the geographic distribution of geo-tagged photos uploaded by Helen. The photos are spread between Asia and Australia.

5.4.2.2 Visualization of Non-spatial Data

With this stage the spatial information of have-home posters can be visualised individually or in aggregated view. In other words, the existing application produces the geographic distribution of photos and friends for have-home posters only. However, there are considerably large amounts of data that are potentially spatial but do not come with enough information to be associated with a specific location and therefore have been excluded from the visualization. As can be seen in Figure (5.9) the information about the null-home friends of 'Helen' has been wasted in the visualization. Moreover, as is discussed in chapter 6 (section 6.3: table 6.3) nearly half of the GB posters are in the null-home category and excluding them from the analysis has a considerable effect on the quality of the final results and conclusions. The overall photo taking behaviour of posters needs to be analysed by visualization of all posters, regardless of the availability of their FHLI. Mapping the have-home posters only is not adequate for answering the research questions. Moreover, the null-home posters[46] are also part of the friendship network of GB posters, and without taking them into account the sub-network of GB friends is not complete and vulnerable against deriving any final conclusions. Therefore, integrating these posters into a visual analysis of the data is essential. As a result the next section covers the steps followed to integrate the non-spatial and potentially spatial data into the existing visualization, enabling further analysis and more robust conclusions about the studied spatio-social data.

5.4.2.2.1 Null-Home GB Posters

In addition to those posters who can be mapped according to their home locations, the null-home posters are demonstrated with their posters' id in a list box on the bottom right of the first sketch (Figure 5.10). By clicking on each poster id in the list, the geographic distribution of photos and friends for that poster are drawn on the second (or right) sketch. The circles are sized according to the number of friends or photos at each location (more explanation in section 5.4.2.3). Standard ellipses show the summary of the direction and distribution of the points. Therefore, regardless of the availability of FHLI, photos and friends distributions can be visualised on the results map (Figure 5.10). Have-home posters can be navigated according to their home locations while the null-home posters can be selected in a list box based on their poster ids. The added list box allows the exploration and analysis of null-home posters and comparison of their behaviours against have-home posters.

[46] Those posters who have their home locations 'null' in their profiles

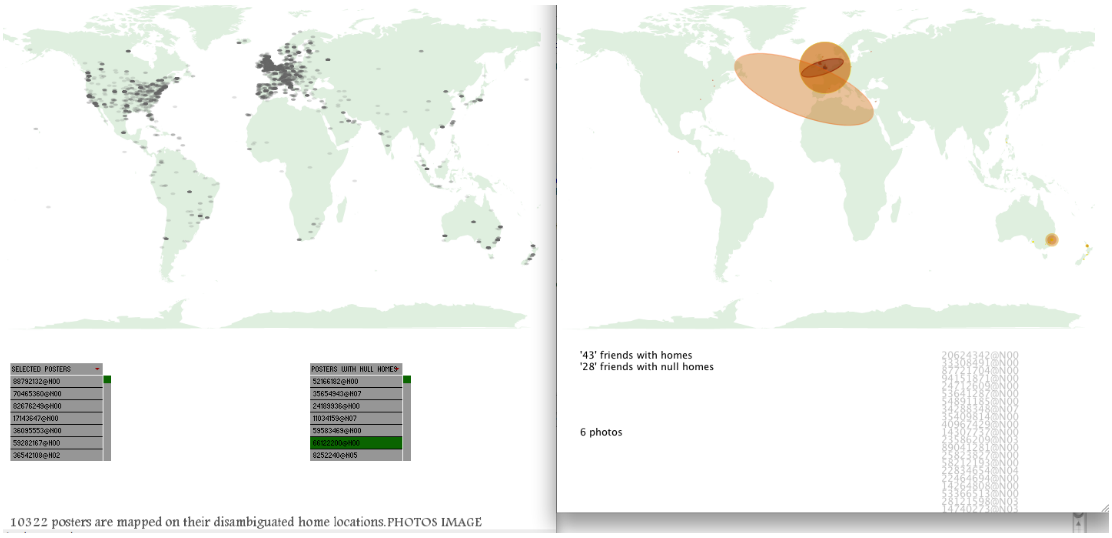


Figure 5.10. Spatial attributes of the selected null-home poster from the list box (highlighted in green) are demonstrated in the right sketch.

Orange and red ellipses are visual summary of the friends and photos respectively. Size of the circles is proportional to number of photos or friends at each location. In addition to visual factors on the map (points, colours, ellipses) and in order to add more information on the sketch, the number of have-home and null-home friends are counted and written on the bottom left of the sketch. The have-home friends are shown on the map according to their FHLI and the null-home friends are listed on the bottom right of the sketch according to their poster id (Figure 5.10).

5.4.2.2.2 Null-Home GB Friends

The GB Friendship network of this study for each poster is made up of two kinds of friends:

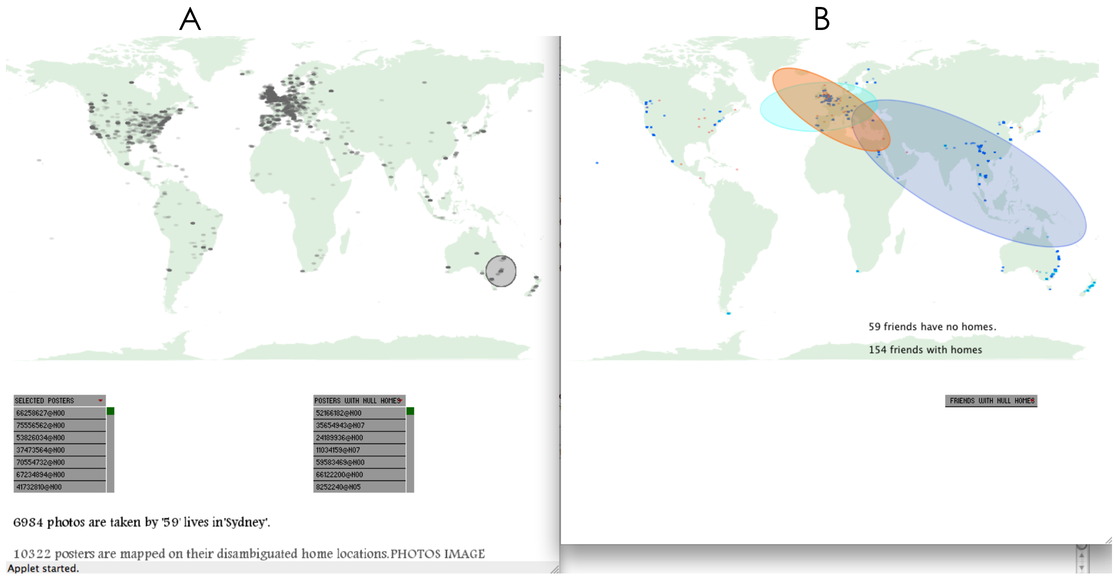
- Have-home friends
- Null-home friends

The selected worked example, 'Helen', demonstrates the typical spatio-social properties of Flickr GB posters of this study. 'Helen' has home location as 'Sydney, New South Wales, Australia' and has a collection of geo-photos stretched between Asia and Australia (Figure 5.9). Her have-home friends in majority lives in Australia and UK and her 4 null-home friends have potentially spatial collections of geo-photos in their profiles. The have-home friends can be visualised according to their disambiguated FHLI. However, including the null-home friends is not straightforward and requires a design decision to be made.

As explained above and already mentioned in the data set chapter, each GB poster in this study has the following spatial information:

- Home location
- Geo-photos collections
- GB friends' home location
- GB friends' photos collection

As mentioned in chapter 4 (section 4.3), the data set of this study is based on the posters who have at least one GB photo in their collections. The rest of the above attributes can be null, unknown or ambiguous. Therefore, in the absence of the home location of null-home friends, distribution of their photo collections can be used as an indication of their whereabouts. This decision is based on the following two points. Firstly, each poster has at least one GB photo in their profile. Secondly, each GB photo of this study has an associated metadata in the form of latitude and longitude with the highest available accuracy (accuracy 16, chapter 3). Therefore, summary distribution of the photos is the most reliable attribute of the GB posters that can be used as indicative of their home locations. In summary, in visual presentation of the GB friendship sub-network, have-home friends are mapped according to their disambiguated home locations and standard ellipses of geo-photos are used as an indicative of null-home friends' locations (light blue ellipse in Figure 5.11).



66258627@N00
75556562@N00
53826034@N00
37473564@N00
70554732@N00
67234894@N00
41732810@N00

52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00
8252240@N05

6984 photos are taken by '59' lives in'Sydney'.

10322 posters are mapped on their disambiguated home locations.PHOTOS IMAGE

Applet started.

59 friends have no homes.

154 friends with homes

Figure 5.11. Visualization of photo distributions and friendship sub-network for 59 posters who live in Sydney. The GB friendship sub-network is demonstrated with two distributions. Disambiguated FHLI of have-home friends (orange) and geo-tagged photo collections of null-home friends (light blue).

The above Figure shows the distribution of the three examined attributes in this study. Therefore, in order to clarify exactly how to interpret the rest of the Figures in this chapter and especially Figures in chapter 7 and 8, right and left sketches are referred to with A and B symbols respectively. Accordingly, Figure 5.11 shows the distribution of 59 posters who live in Sydney (sketch A). The size of the circle on the selected location (Sydney) is proportionate to the number of posters who live in that location (in this example 59). The blue ellipse as before shows the distribution of 6,984 photos taken by the 59 posters (sketch B). The orange ellipse also shows the distribution of have-home friends (sketch B). Here, one new variable is added to the visualization and that is the light blue ellipse indicating the photo distributions of null-home friends. As can be seen in sketch B, the three distributions can easily be compared. By having the above three distributions summarized with emphasis on direction and centre it is possible to make hypotheses and derive some useful conclusions in the next stage of this research. For example, here can summarise that GB posters of Sydney have photos distributed from home location towards Europe. Their have-home friends live in majority in close proximity of Europe so does the places that null-home friends take photos. Referring to the worked example in section (5.4.2.1, Figure 5.9) can identify that Helen is one of the 59 GB posters from Sydney. Her poster Id is in the left list box in the sketch A. Her contribution towards the patterns identified here (Figure 5.11) can be seen by switching between the aggregated view and individual view (as shown in Figures 5.8-9).

5.4.2.3 Summary of the methods used to uncover Patterns in Data

This section summarises the visualization methods applied, adapted and in some cases modified, in developing the visualization application. Overall, attempts were made to make the application an appropriate tool for answering the research questions.

1. Plotting points:

Geo-tagged photo collections and disambiguated home locations in this study were associated with sets of latitude and longitude. Since (lat, lon) values show a point on a spherical globe, a projection was required to convert the points to positions applicable to screen coordinates. While two maps were drawn in two separate sketches, instead of screen coordinates the points were converted to coordinates appropriate to the width (x axis for longitude) and height (y axis for latitude) of the sketches (equations 5.1 and 5.2).

$$x = \frac{width \times (longitude + 180)}{360}$$

Equation 5.1 Conversion of longitude values to proportional points in x-axis

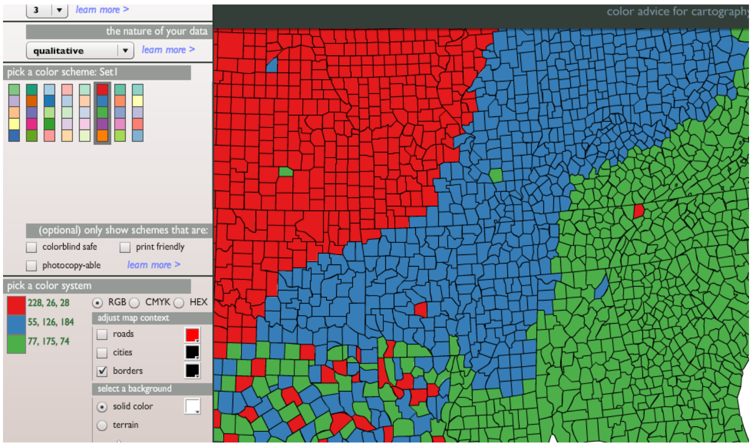
$$y = \frac{height \times (latitude - 90)}{-180}$$

Equation 5.2 Conversion of latitude values to proportional points in y-axis

The process of drawing the spherical globe on the flat surface in section (5.4.2) involved in scale alteration and system transformation. There are several alternative map projections in cartography to choose from. In different fields of research from geographers to navigators, politicians and historians they all use flat maps instead of globes and that necessitate a map projection. Selecting a projection depends on the nature of the research. Each projection can be useful for certain purposes and can be unfit for another. Each projection retains some properties and distorts others (i.e. Angular relationship, relative size of Figures, distance, direction, areas etc). According to the variables under analysis a proper projection need to be selected. In this thesis, a flat map is used for visual demonstration of geographic distribution of photos and friends. There is no specific hypothesis regarding angles, areas or directions. The only important feature of the maps here is location of posters and geographic distributions of their photos and friends. Accordingly, a simple projection called ‘Equidistant Cylindrical’ has been applied. It is a form of geographic projection that distorts shape and area but preserves relative distance (Longley, P.A. et al. 2005). According, to the developed research questions in section (1.5) keeping distance relatively accurate is of more importance than shape and area of places.

2. Colour:

Colour is used to distinguish several different categorical attributes (posters/friends home location and their photos, collections and distributions). The examined attributes are demonstrated with circle or ellipse symbols and are coloured differently (Figure 5.11) to reflect the attributes they represent. The choice of colours has been supported according to the colourBrewer application[47] (Figure 5.12) and is reflected in chapter 7 & 8.



3

the nature of your data: qualitative

pick a color scheme: Set1

(optional) only show schemes that are:
□ colorblind safe □ print friendly
□ photocopy-able

pick a color system
🟥 228, 26, 28
🟦 55, 126, 184
🟩 77, 175, 74

● RGB ○ CMYK ○ HEX

adjust map context
□ roads
□ cities
☑ borders

select a background
● solid color
○ terrain

Figure 5.12. ColourBrewer colour scheme that used for the viz application of this study.

The colourBrewer application provides guidance on how to choose appropriate sets of colour for maps. Sequential, diverging and qualitative data categories have different sets of colours to explore and choose from. The colour scheme used in the viz application of this study has followed the scheme recommended in the colourBrewer application under the qualitative category (Figure 5.12).

3. Transparency:

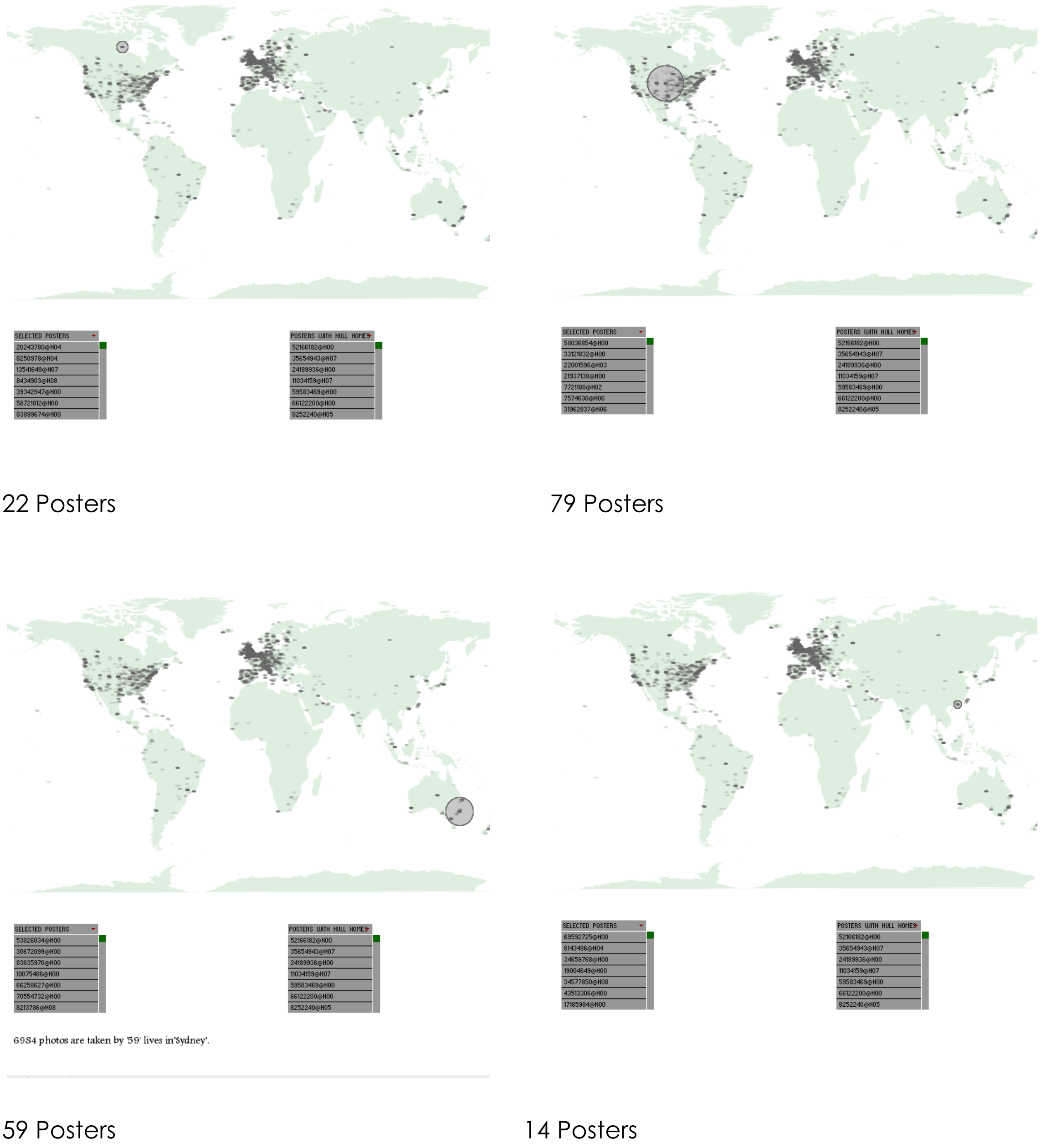
Since in majority of cases more than one item (poster or photo) is referenced to a specific point on the map, transparency is used for depiction of the areas with high density. The transparency to some extent improves the overall layout of the points’ density. In addition, different sizes for circles were added for visualization of the number of points on each location on the map.

4. Size:

Size has been applied in some cases to reflect the number of points in each location on the map. In the original map for each selected location by user (mouse rollover), a circle highlights that location. The size of that circle is

[47] http://colourbrewer2.org/

proportional to number of posters living on that location. The circle has been used as a feature, additional to transparency, in presenting the density of each location (Figure 5.13).



22 Posters

79 Posters

6984 photos are taken by '59' lives in'Sydney'.

59 Posters

14 Posters

Figure 5.13. Circles are sized according to number of posters in each location

Moreover, in individual layouts of the application, distributions of photos and friends of a selected poster from the list box are visualized with circles that have been sized according to the magnitude of the data they represent (Figures 5.9 and 5.10).

5. List Box:

GUI features have been added to the application in order to achieve higher interactivity and also to create individual layouts in addition to aggregated views only. List Box has been used in two places in the application. Firstly it lists the null-home posters and allows the study and analysis of their photo and friends distributions. Secondly, the second list box lists the poster ids of posters who live in the location selected by the user. It allows the analysis and visualization of the individual behaviors, as well as their aggregated view, with other posters who live in the same location. It also enabled the analysis of posters' photos and friends individually. In both cases the posters listed in the list box can be selected and their friends and photos distributions are visualized automatically in the results map (Figure 5.9 and 5.10).

6. Data Summary:

In order to get broader view of patterns and in line with the research questions efforts have been made to visualize the summary of data (photos and friends of all posters living in the same location - Figure 5.7-8). Therefore, the results map shows the distributions of friends and photos of all posters who live in the same selected location (aggregated view). Before getting into any details regarding the standard ellipses a summary of alternatives are given here and the choice of ellipses are justified.

The first obvious choice is to study the data in table format or excel files. The numerical analysis could have been conducted but no pattern can be visualised. Point distribution on the map is also another obvious choice. As demonstrated in the section (5.4.2), although it is useful in identifying the general distributions, in areas with high density (lots of points) the distribution becomes too cluttered to discern any pattern. Another option worth mentioning here is a spatial summary statistic called 'Density Surface'. This technique has been used by geographers to explore different kinds of spatial patterns. It uses a graded colour scheme to represent the density in each area. It can also demonstrate the density by topographic landscape. It uses the number of points in each area as if they were elevation values and it demonstrates mountains, peaks and valleys (Longley et al. 2005). High-density areas are shown as mountains and flat areas are spars and empty places. This method although useful for exploring spatial patterns in data, does not sound appropriate for visualization of the spatio-social data of this study:

1. There are more than one spatial attribute to be studied
2. There are non-spatial and potentially spatial data as well
3. An appropriate summary visualization for this study should allow the visualization of the distributions by plotting the points as well as any other summarized pattern
4. The summarizing technique should allow the user to compare three different distributions at the same time in the same map

Considering the above points, another methods of summary distribution of points have been taken into consideration. The visual summary of photo and friend distributions is demonstrated with Standard Ellipse[48]. This method summarises the distribution of collections of points by an ellipse. The size of the ellipse, its centre, major and minor axes are respectively proportional to the scope in which points are scattered around their centre, the mean centre of values, and the maximum/minimum directional pattern of points. Analysis of the standard ellipses can reveal any particular orientation among points. While it is possible to get a sense of patterns and trends by plotting the points on the map (Figure 5.6), measuring the standard ellipse highlights the trend and orientation of the examined points (photos and friends). The standard ellipses have been selected for the visualization part of this study for several reasons. Firstly, they can

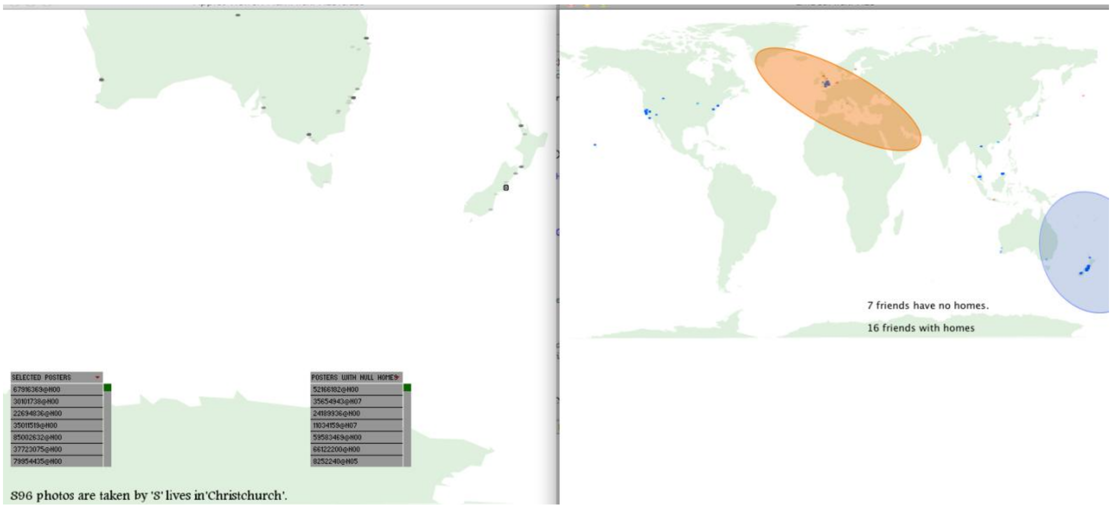
[48] http://www.gicentre.org/utils/ellipse/

successfully summarise a collection of points by emphasizing on the centre and direction of the distributions. The summarized layout of the ellipses is very efficient. It occupies as much space on the map as is required to demonstrate the direction of min/max dispersion of the points. By summarising the data in standard ellipses, it is possible to keep the original distribution (plotting points) on the map as well without making the layout congested.

Secondly, potentially spatial data (null-home posters) can be visualized by summarising the desired distributions. Several different points distributions can be summarized by standard ellipses on the same map. Applying different colour for each distribution adequately differentiate the distributions' centre and orientation.

Moreover, according to the nature of this thesis illustration of the orientation and direction of the distributions plays important role in answering the research questions. Standard ellipses are very well capable of doing that. Overall, producing the ellipses as well as their understanding, interpretation and comparison, is efficient and easy. However, in applying the standard ellipses to geographic distribution of spatial variables, it has been identified a drawback that needed to be rectified before analysis.

In measuring the standard ellipses for collection of points spread all around the world (photos and FHLI), the points that come on the edges of the map (close to longitude 180 and -180) were found to be problematic. This can happen in directional as well as circular data. For example, in the directional case, angles of -179.9 and 180 are very close while in circular cases angles of 359.9 are very close to 0. However, not all measurements and calculations with the angles and degrees consider the cyclic and directional nature of data of that kind. The Standard Ellipse class of the giCentre library also has failed to consider the directional nature of the points on the map. For instance, the photos distribution for a collection of photos heavily focused in Australia, New Zealand and the USA was calculated as a long ellipse drawn all over the map stretching from New Zealand to the USA. The resultant ellipses were measured according to the longest distance between the points on the edges and therefore cannot be a correct indication of the points' distribution. Consequently, in order to reflect the cyclic nature of the data and visualize a standard ellipse of points on the edges of the 2 dimensional coordinates, a modification was required for the giCentre utilities library.



SELECTED POSTERS
67916363@N00
30101738@N00
22694836@N00
35011519@N00
85002632@N00
37723075@N00
79954435@N00

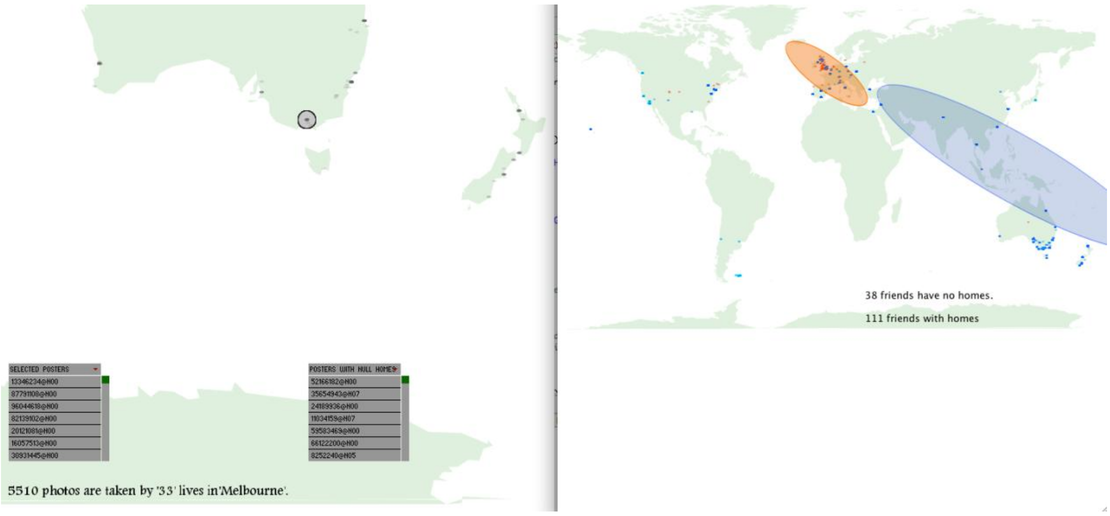
POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00
8252240@N05

896 photos are taken by 'S' lives in'Christchurch'.

7 friends have no homes.

16 friends with homes

Figure 5.14. An Example of the modification made to draw broken ellipses for photos taken by posters in Christchurch, New Zealand in order to reflect the circular nature of the global coordinate system.



SELECTED POSTERS
13346234@N00
87791108@N00
96044618@N00
82139102@N00
20121081@N00
16057513@N00
30931445@N00

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00
8252240@N05

5510 photos are taken by '33' lives in'Melbourne'.

38 friends have no homes.

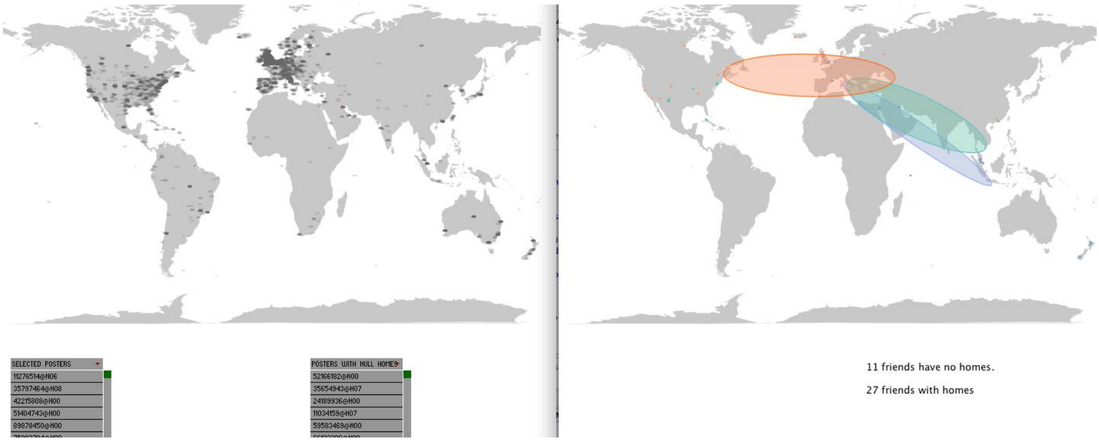
111 friends with homes

Figure 5.15. An Example of the modification made to draw broken ellipses for photos taken by posters in Melbourne, Australia in order to reflect the circular nature of the global coordinate system.

In the literature there are some techniques for dealing with directional and orientational data (Fisher, 1993; Jammalamadaka and SenGupta, 2001 and Mardia and Jupp, 1999). Circular statistics have been suggested for angular data analysis (Brundson, 2006). Accordingly, in the Standard Ellipse class of the giCentre utilities library a new constructor was developed and added to the library by the developer. The new constructor allows creating of ellipses by a collection of latitude and longitude instead of converted (X, Y) values. Mean and standard deviation calculations were modified accordingly to reflect the spherical nature of the earth that is hidden in latitude and longitude degrees. As a result of the modification, the new standard ellipse drawn in the developed application considers the directional nature of the data. Therefore, the ellipses are broken in the right direction instead of stretching wrongly all over the map (Figures 5.14 and 5.15). As can be seen in the mentioned two Figures, the points that happen to be on the edges of the map are resulted in the broken, ellipses indicating the circular nature of the global coordinate system.

7. Aggregation of Home Locations:

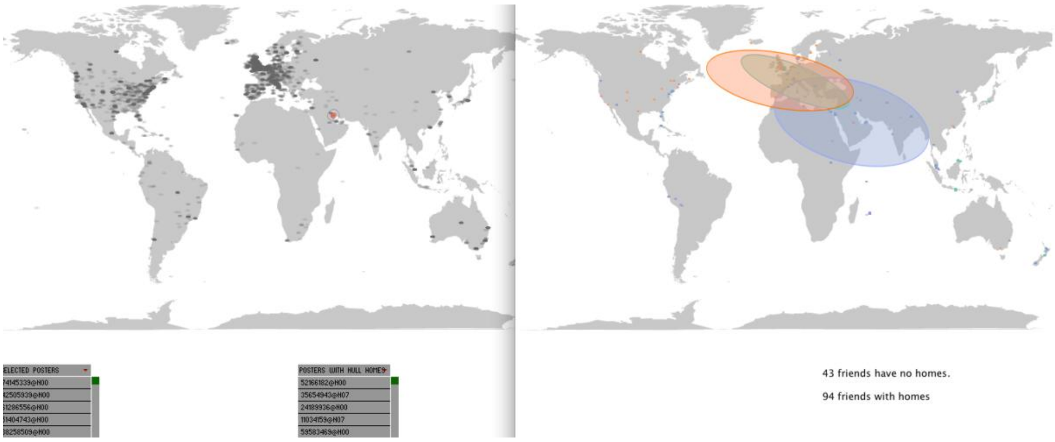
In order to see more robust patterns based on a larger sample of posters than those who live in exactly the same location, the home locations of the posters can be aggregated together according to the threshold distance defined by the user. By touching the arrow keys on the keyboards the threshold distance increases or decreases by 10 pixels and therefore, a greater or smaller number of posters and their spatial and non-spatial data can be visualized in the second sketch. This modification is applied to the application in order to avoid deriving conclusions based on small samples of posters who live in exactly the same selected location. By doing that the conclusions and findings are also less vulnerable to bias or the inevitable consequence of sampling strategies. Consequently, the application can demonstrate how patterns change by increasing or decreasing the number of posters through expanding the examined area. The following Figures (5.16 - 5.22) show an example of a sample location (Kuwait) and how the application can be used in the study and analysis of how spatio-social patterns could change by expanding the examined location. In the first step a location is selected and the application visualizes the requested variables for posters who live exactly on that location only (Figure 5.16). Afterwards, the consecutive Figures (5.17 – 5.22) demonstrate how the patterns of the examined variables change while expanding the selected area to nearby proximities (10 pixels at a time).



11 friends have no homes.

27 friends with homes

Figure 5.16. Photos distribution, have-home friends' homes and null-home friends' photos for 8 GB posters in Kuwait (threshold: 10)


ELECTED POSTERS

74145339@N00
42505939@N00
61286556@N00
51404743@N00
38258509@N00

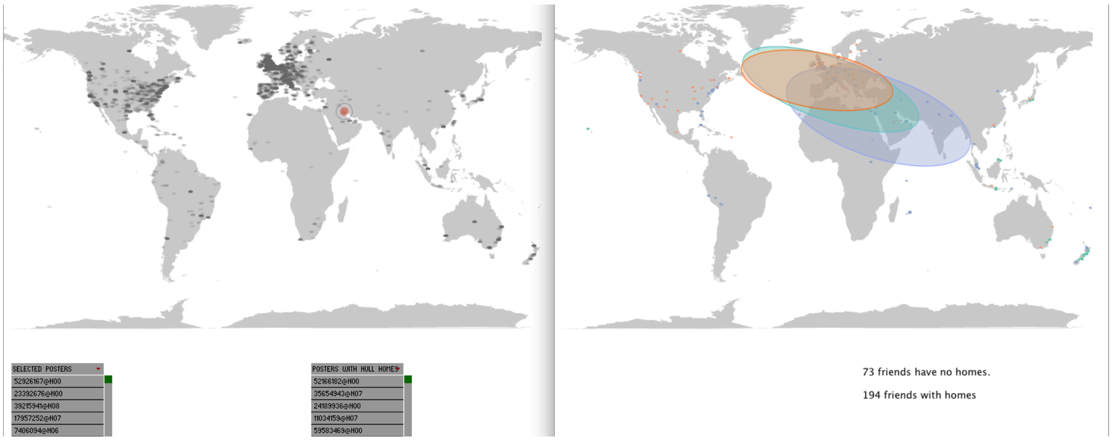
POSTERS WITH NULL HOMES

52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00

43 friends have no homes.

94 friends with homes

Figure 5.17. The above same variables for 30 GB posters who live in close proximity of Kuwait (threshold: 20)


SELECTED POSTERS

52926167@N00
23392676@N00
39215941@N08
17957252@N07
7406094@N06

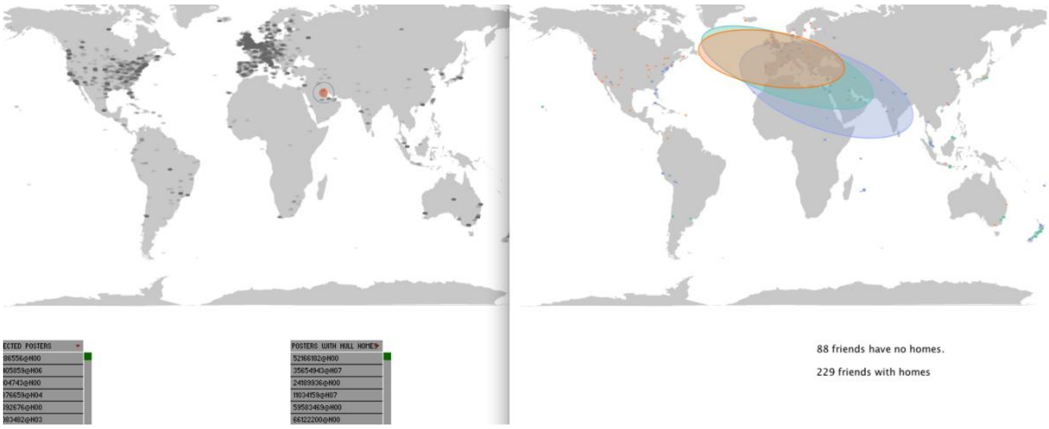
POSTERS WITH NULL HOMES

52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00

73 friends have no homes.

194 friends with homes

Figure 5.18. The above same variables for 48 GB posters who live in close proximity of Kuwait (threshold: 30)


ECTED POSTERS

86556@N00
105859@N06
04743@N00
376659@N04
392676@N00
083482@N03

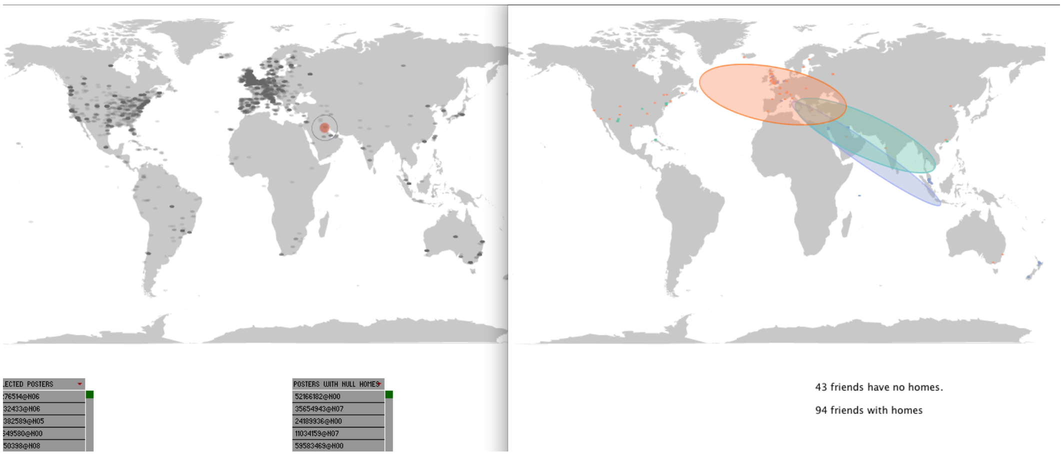
POSTERS WITH NULL HOMES

52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00

88 friends have no homes.

229 friends with homes

Figure 5.19. The above same variables for 53 GB posters who live in close proximity of Kuwait (threshold: 40)



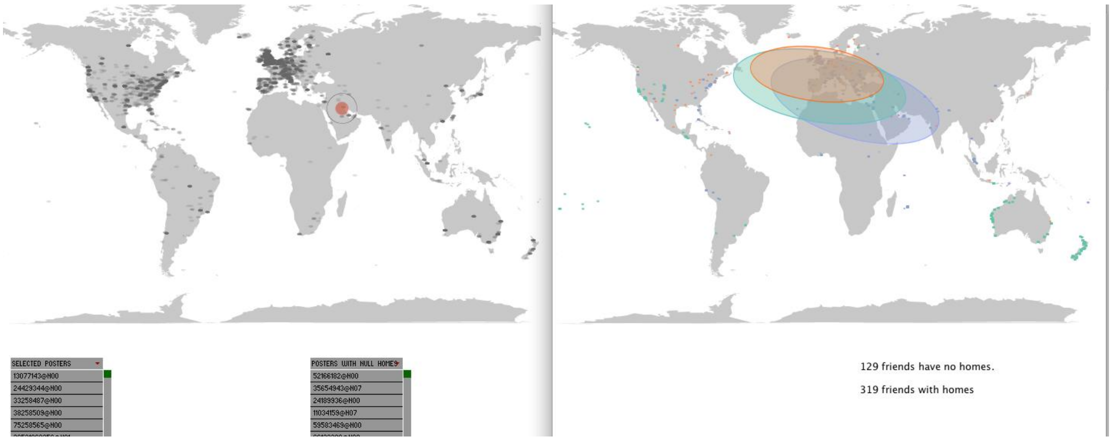
LECTED POSTERS
76514@N06
32433@N06
382589@N05
49580@N00
50398@N08

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00

43 friends have no homes.

94 friends with homes

Figure 5.20. The above same variables for 63 posters who live in close proximity of Kuwait (threshold: 50)



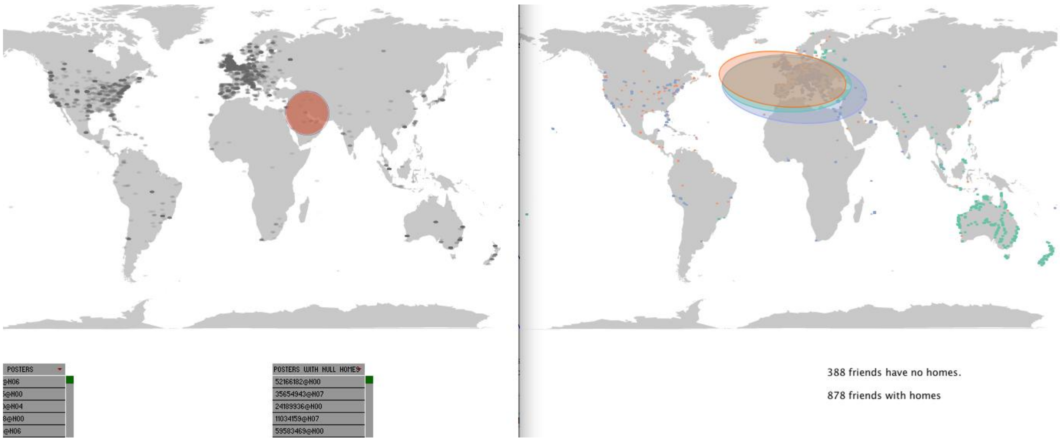
SELECTED POSTERS
13077143@N00
24429344@N00
33258487@N00
38258509@N00
75258565@N00

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00

129 friends have no homes.

319 friends with homes

Figure 5.21. The above same variables for 104 GB posters who live in close proximity of Kuwait (threshold: 60)



POSTERS
@N06
5@N00
@N04
8@N00
@N06

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00

388 friends have no homes.

878 friends with homes

Figure 5.22. The above same variables for 781 GB posters who live in close proximity of Kuwait (Threshold 90).

8. Appearance Compensation (Flannery[49]):

In the developed visualization application in this study, size has been applied to reflect the number of points (posters or photos) that are associated with one specific location. In doing so, attempts were made to visualize the number of photos/friends through an area of circles. In other words, the areas of circles are used to reflect the magnitude of data in each location. However, as research in psychology and psychophysics suggest human minds underestimate the area of larger circles (Montello, 2002). While we have ability to differentiate the length accurately, the area and volumes are very much subject to underestimation (Ihaka, 2003). Accordingly, the areas of larger circles need to be increased by the radius to the power of 0.6 instead of conventional square root of the radius (radius to the power of 0.5). This amendment has been applied to minimise the risk of perceptual problems and optical illusions that might lead to wrong interpretation of the results through visual analysis of the data (Lecture Note, City University, Jo Wood 2011).

5.5 Summary

This chapter described the methods applied for the study, classification, disambiguation and uncertainty assignment of sample Flickr world data. It has been found that in addition to photo collection of GB posters, their home locations also bear important spatial information. This kind of spatial information varies from formal geographical coordinates to fuzzy vernacular geographic terms that people use in their everyday lives. Accordingly, the existing disambiguation methods were inadequate in disambiguating the FHLI successfully. Therefore, a method was developed and applied to Flickr sample world data manually. Based on that experiment and according to the final results (chapter 6: section 6.2.6) amendments were applied to the method and a customized algorithm was developed and implemented in an automated Java application (Appendix 6). The application performance was assessed against the GB data and failure cases were classified.

The final section explains the design stages of the development of the visualization application. It covers the visualization methods studied and applied in designing an appropriate application for visualization of the Flickr spatio-social data and answering the research questions. Several visualization packages were assessed against the research requirements. A suitable visualization package was selected. The step-by-step design decisions and redesigning process of the application development were described. The last section gave a summary of all methods applied in developing the application. Next chapter covers the study and analysis of the Flickr world data and Flickr GB data.

[49] James Flannery is Arthur Robinson's student and one of the founders of American academic Cartography. He is amongst first geographers who applied psychophysical methods in geography and on human perception of circles (Making Maps: DIY Cartography, John Krygier, 2007 available at http://makingmaps.net/2007/08/28/perceptual-scaling-of-map-symbols/

6 Numerical Analysis of Flickr Sample World Data Against Flickr GB Data

6.1 Introduction

This chapter contains a numerical analysis of the data. The first section covers the analysis of the sample world data set. It studies the patterns in the number of photos and friends in the sampling time periods. The results of the proposed FHLI classification and uncertainty measurements (chapter 5: section 5.2) are also demonstrated. Accordingly, the fundamental findings that play important roles in selecting a rich and indicative Flickr sample GB data and modifications in the disambiguation algorithm are summarized. The second section reviews the Flickr sample GB data. It contains a brief numerical analysis of the data in general and according to null-home and have-home categories.

6.2 Flickr Sample World Data

This section covers the analysis of the sample world data set collected from Flickr (chapter 4: section 4.3.1). As described in chapter 4 (section 4.3.2) 3,245,866 photos were randomly selected on a daily basis. This section covers the study and analysis of the general attributes of this Flickr sample world data. The overall numbers, geo-tagged photos' collections, friendship network and FHLI have all been examined here. The results of the proposed FHLI classification and uncertainty measurements (section 5.2) are also summarized and discussed.

The results and findings of the analysis have been saved and used as a guide and a measurement scale representing the general attributes of Flickr data. Accordingly, certain criteria have been defined for selecting rich and indicative data set for this study (section 4.3) and achieving the defined aims and objectives (1.3-5). Comparison between the findings of the two assessments can reveal how close or far the GB data is from the general attributes of Flickr data. The assessment reveals the validity, level of robustness and consistency of final findings and conclusions.

6.2.1 World Level vs. Street Level Precision

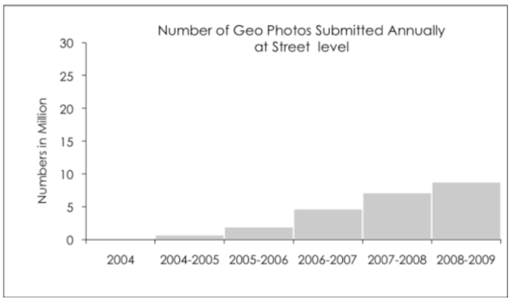


Figure 6.1. Number of geo-photos submitted to Flickr annually at street level during the sampling intervals

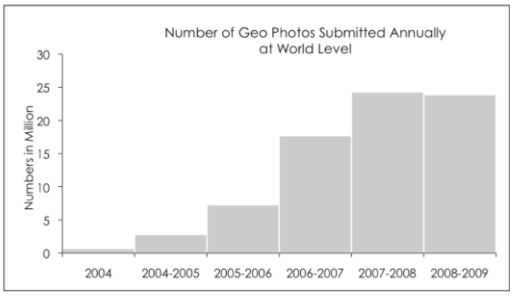


Figure 6.2. Number of all geo-photos submitted to Flickr annually during the sampling interval

As Figures 6.1 and 6.2 show, the number of Flickr geo-tagged photos in both examined categories has increased gradually each year, with the most significant rise in the fourth period (2006-2007). As can be expected the annual GB photos (demonstrating photos with any kind of geo-spatial information attached) are considerably more than the photos with the highest locational precision (street level). However, despite the small drop in 2008-2009 for all geo-tagged photos, the street level photos have continued the increasing trend (Figure 6.1).

6.2.2 Geo-Tagged Photos per Poster

The following Figures demonstrate the number of geo-tagged photos submitted by posters during each sampling period.

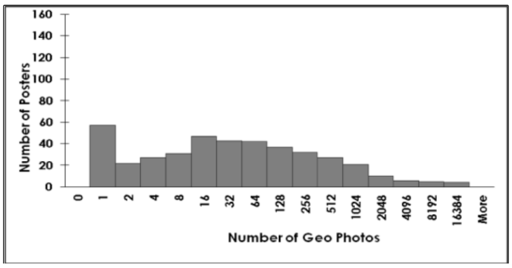


Figure 6.3. 01/02/04 to 20/07/04. 2,670 photos uploaded by 411 posters

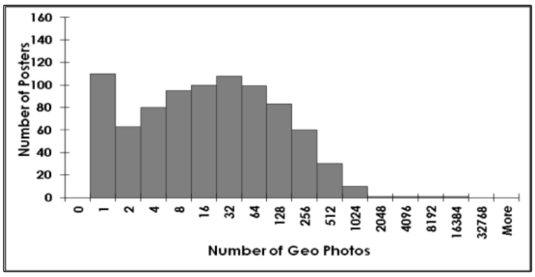


Figure 6.4. 2004-2005. 5,715 photos uploaded by 947 posters

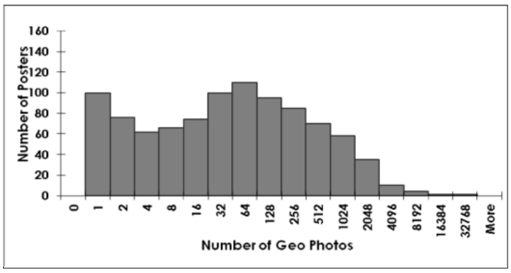


Figure 6.5 2005-2006. 5,730 photos uploaded by 947 posters

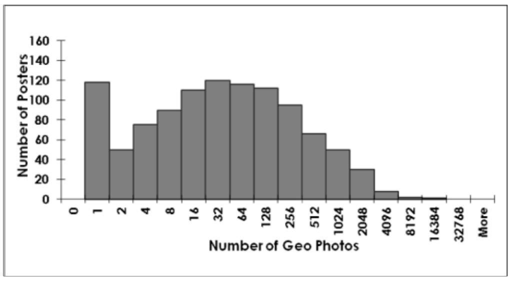


Figure 6.6. 2006-2007. 5,730 photos uploaded by 1,043 posters

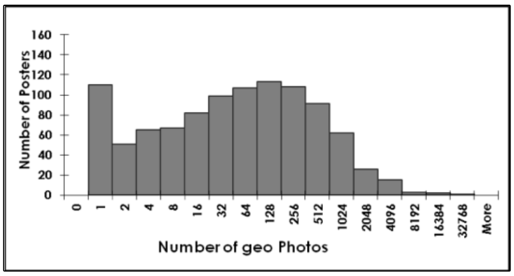


Figure 6.7.2007-2008. 5,730 photos uploaded by 991 posters

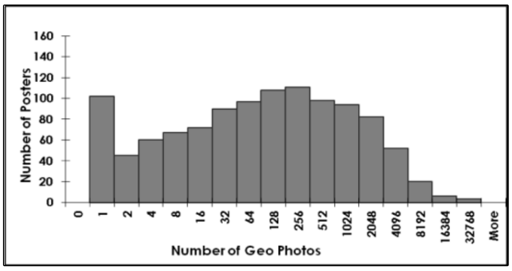


Figure 6.8. 2008-2009. 5,729 photos uploaded by 1,142 poster

As the results show, the frequency histograms of the geo-tagged photos sampled in all twelve-month periods follow a bimodal log-normal distribution with a population with one photo only, and the second population with log-normal distribution. The smaller number of geo-photos and the different behavior of users in the first six-month period (Figure 6.3) might be the result of the early days of Flickr when there were some experts and a few novice users affecting the results by their extreme contradicting behaviors.

6.2.3 Number of Friends per Poster

The following Figures demonstrate the number of public friends for the sampled posters during each sampling period.

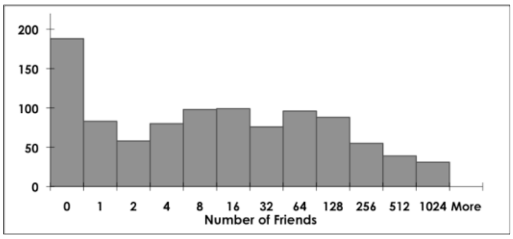


Figure 6.9. 411 Flickr members (six month period, 2004)

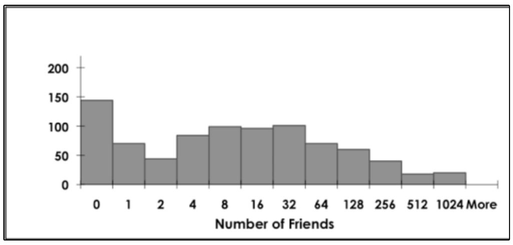


Figure 6.10. 846 Flickr members (2004-2005)

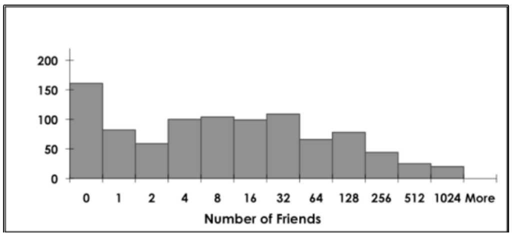


Figure 6.11. 947 Flickr members (2005-2006)

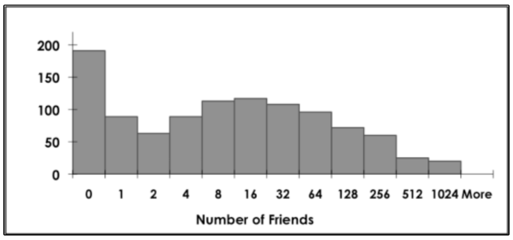


Figure 6.12. 1,043 Flickr members (2006-2007)

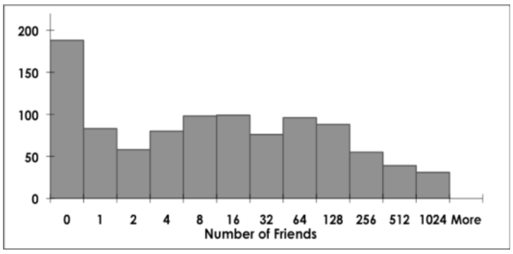


Figure 6.13. 991 Flickr members (2007-2008)

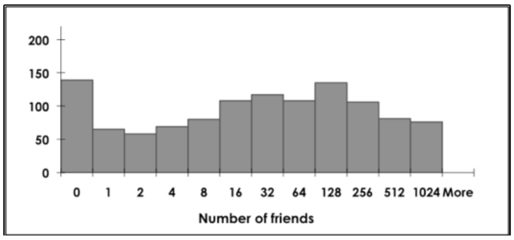


Figure 6.14. 1,142 Flickr members (2008-2009)

According to the above Figures the friendship trend in Flickr since 2004 has had three populations with different behaviors:

- A large population of users with no friend at all
- Second population with only one friend and
- Third population with some irregularities compared to the bell shaped curve of the log-normal distribution.

Overall, the most recent period is the most sociable period of the Flickr users with the lowest number of people with no friend and the highest average, median and eight times higher favourite number of friends than the previous two year intervals. This might indicate that Flickr users have started to change their online behaviors by developing larger networks of friends.

6.2.4 Analysis

As Figures 6.1 to 6.2 demonstrate, the continuous increasing popularity of the street level geo-tagging photos among Flickr users, especially in the most recent examined period when the total number of geo-tagged photos has dropped, might be the result of the availability and popularity of spatially aware devices (e.g. iPhone, iPad etc.) among the general public. It might also indicate that people have gradually built up trust and confidence in providing more accurate geo-spatial information in online environments.

Analysis of Figures demonstrating the geo-tagged photos for the sampled world data (Figures 6.3 to 6.8) reveals that, interestingly, the highest numbers of photos uploaded by the majority of users are in the last and most recent examined periods (Figures 6.7 and 6.8). This trend might be defined by the fact that in the early days Flickr had two different categories of users. There might be groups of posters who were absolutely new to the geo social networking sites and were experiencing the environment, while the second category might be those posters who were interested in sharing the spatial information in online environments. In other words, the second category of posters in early days of Flickr probably comprised those people who had a good perception of sharing geo-referenced photos in online environments. These two considerably different kinds of users ended up with the highest average, median and mode. As Flickr became more popular among online socializers, and also due to the huge improvement in devices that uploaded locations, more users started submitting photos, therefore decreasing the most favorite number of submitted photos. Although there were still people submitting a large amount of photos, the number of them did not outweigh the average of the photos uploaded by the second category of posters in the early days of Flickr. After three years experiencing and practicing, and with the noticeable convenience in using geo- spatial technologies that are built into the everyday use of accessories of people, more users started submitting a large amount of geo-tagged photos.

Regarding the number of friends for the sampled world data (Figures 6.9 to 6.14) it can be inferred that interestingly there is no one with more than 1,024 friends in the all of the sampling periods and the numbers of people with no friends stand out in all of the examined intervals. Flickr users have tended to make more friends in the first two years (Figures 6.9 and 6.10) than the following two consecutive years (Figures 6.11 and 6.12). The most recent examined period (2008-2009) has the least number of users with no friends and an eight times higher second favourite number of friends (128 friends) with the highest average and median.

In addition, the smallest gaps between the number of people with no friends and the highest favourite number of friends are in the most recent periods (Figures 6.13 to 6.14). Considering the different behavior of posters in the first six months of the Flickr existence, this finding can also be attributed to the possible noticeable difference between the behaviors of the two considerably different categories of posters in Flickr in the early days. Whereas, in the most recent period the highest favorite number of friends can be seen as increasing the tendency of posters in developing social connections on Flickr rather than using the site merely for uploading geo-photos.

6.2.5 Findings

In conclusion and according to all the analysis done up until now, we can conclude that the earliest and the most recent examined periods have had the most significant attributes.

The first six months of Flickr (01/02/04 to 20/07/04) have:

- Highest number of the favorite number of geo-tagged photos
- Smallest gap between the number of users with no friends and users with the favorite number of friends

The last twelve-month period (19/07/08 to 20/07/09) have:

- Highest number of the favorite number of geo-tagged photos
- Highest number of the most popular number of friends
- Increasing trend in street level photos despite the decrease in world level photos

Consequently, one might hypothesize that gradually, and especially in the most recent examined periods, Flickr users are more likely to develop larger networks of friends (Figure 6.14) in Flickr, are more interested in photos with highest levels of precision (Figure 6.1) and are more willing to upload geo-tagged photos (Figure 6.8). The last, but not the least interesting finding here, is that the bi-modal log-normal distributions have emerged consistently over the examined period for both the number of posted photos (Figures 6.4 to 6.8) and the number of friends (Figures 6.9 to 6.14). This indicates a repeatable and interesting pattern.

Referring to the literature there are two general classes of statistical distributions. The random variables are usually expected to follow the bell-shaped distribution where all items are clustered around a single mean value (Casella and Berger, 2001). The dependent variables that are not random (e.g. citation network) can usually be approximated by a Zipfian distribution. Zipfian distribution comes from the power law probability family and is named after the person who first proposed it (George Kingsley Zipf, 1949). In Zipfian distributions small occurrences happen considerably more than large instances. In other words, frequency of occurrences is inversely proportionate to their frequency range (Adamic and Hubeman, 2002). However, in this study a third statistical distribution has been identified. Figures 6.3 to 6.14 demonstrate that the bi-modal log-normal distribution was consistently repeated for friendship and photo distributions. Here the examined variables (number of GB friends and number of geo-tagged photos) cannot be directly recognized as dependent or random. The number of photos might seem to be a random variable in the first instance, but considering other factors like accessibility of the spatially aware devices, internet access, possible limitations on the maximum number of photos that can be uploaded at the same time for each poster, etc., these make the random category vulnerable. Likewise, the number of friends might be random or dependent on one or more factors (e.g. GB friendship sub-network, limitation on maximum number of friends on social sites etc.). In the same line, the identified pattern explains that the examined data are neither random nor dependent. They are random numbers that might be affected by possible factors, but these factors have not been influential enough to make the data merely dependent on them. Overall, and as discussed here and also demonstrated in Figures 6.3 to 6.14, the identified patterns can be approximated with bi-modal log-normal distributions at all periods. This consistency in the distributions reveals interesting and unexpected patterns about the examined variables.

The result of applying the first proposed disambiguation algorithm and classification method (chapter 5: section 5.2) to the 'Flickr sample world data' is demonstrated and discussed in the next section.

6.2.6 *Results of Applying the Pilot Disambiguation and Classification Methods to Flickr Sample World Data*

Classifying the FHLI of the sample world data according to the proposed method (section 5.2) can demonstrate how precisely people refer to their home locations on the web. In Figure 6.15 the five examined periods (2004-2008) in this study are demonstrated with a slightly different colour. The vertical bars on each precision class are indicative of the number of posters who have their home locations in that precision category. Each colour shows different time period.

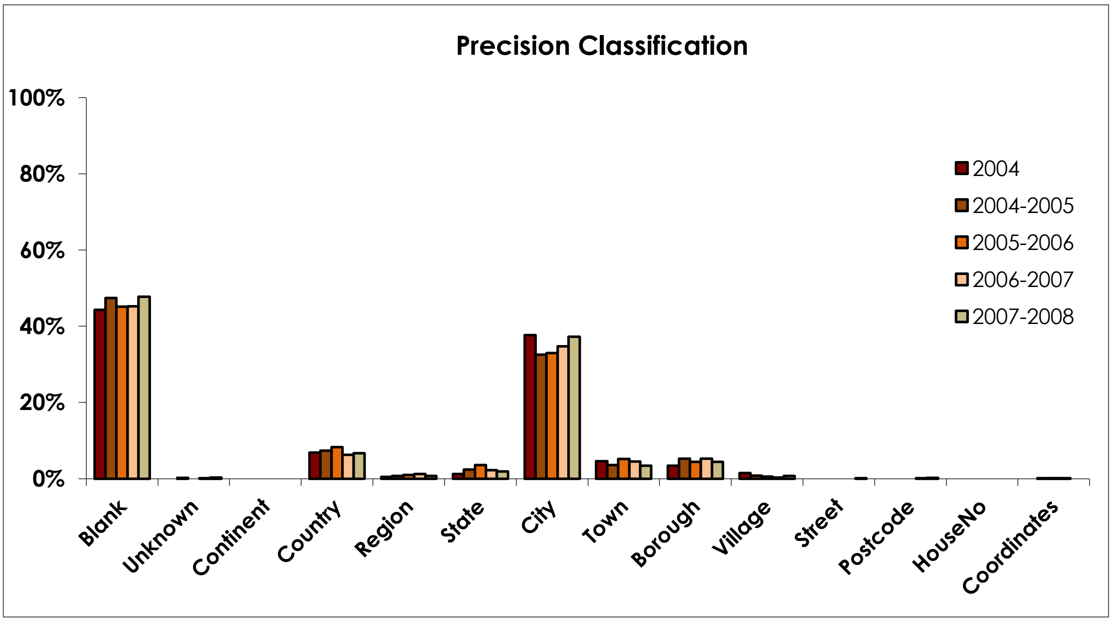


Figure 6.15. Precision classification for Flickr sample world data. For each precision category the number of FHLI are drawn from left to right in chronological order (2004-2008)

As the Figure 6.15 illustrates for each of the 5 examined periods in this study (2004-2008) posters have higher tendency in leaving their home location as blank or associate themselves to a city. Studying the blank category in Figure (6.15) indicates that: in 2004, 44% of the randomly selected posters have left their home location as blank (dark red bar on the left side of blank category).

Likewise, in 2004-2005 47% left their home locations blank

45% in (2005-2006) and (2006-2007) periods and

47% in (2007-2008) period provided no information about their homes.

The rest of the chart can be interpreted as above for the remaining categories. Accordingly Figure (6.15) indicates that there are remarkably consistent patterns in all the examined time periods (2004-2008) with the most significant number for blank and city level. In simpler terms posters have higher tendency in leaving their home locations as blank or associate themselves to a city. This is an important finding in modifying and customizing the proposed algorithms for disambiguation and classification of Flickr sample GB data (section 5.3).

Figure (6.16) demonstrates how many percent of FHLI in each precision category (Table 5.2) is occupied by each of the five types of examined uncertainties (section 5.2.3).

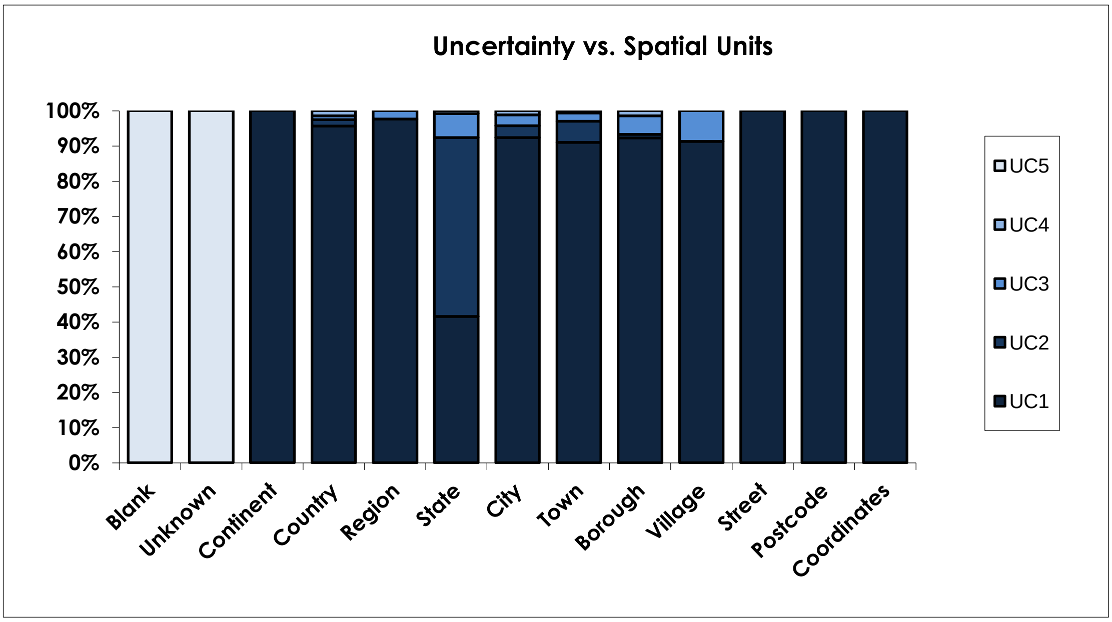


Figure 6.16. Uncertainty assigned to spatial units for Flickr sample world data

Understandably, since no information has been associated to home locations in blank and unknown categories, FHLI in this class have the highest uncertainties (UC 5, lightest Blue). The home locations that can be associated to continent, street, post code and coordinates categories are assigned the least uncertainty (UC1, darkest blue). Since each can be associated to one location although in different precision levels. The rest of the precision categories contained combination of uncertainties according to the FHLI that posters provide on their Flickr profiles. For example around 90% of the FHLI in city category can be associated to exactly one city in the world (UC1, darkest blue, example: 'London, UK'). About 4% of FHLI in this precision class can be referred to more than one city but in the same country (UC2, dark blue, example: 'London, Brighton, UK). Slightly more than 3% of the home locations can be referred to several cities in different countries (UC3, medium blue, example: 'Tehran, London, UK'). Interestingly, as can be seen here (Figure 6.16), in all examined categories uncertainty 3 is higher than uncertainty 4. This indicates the fact that uncertainty in place names is higher at a national level than across countries. This finding also plays an important role in defining criteria for data selection of this study (chapter 5: section 5.3).

Referring back to the example mentioned in section (5.4.2.2.2) a GB poster called Helen is used as a hypothetical member who has FHLI as 'Sydney, New South Wales, Australia'. Since she has mentioned spatial information for her home location she is categorized as have-home posters (section 4.2.2). Her FHLI is being disambiguated according to the developed disambiguation algorithm in section (5.3.2). Her FHLI is in the category of feature code combinations demonstrated in table (5.5). 'Sydney' is PPL (populated place feature code) nested in New South Wales (Admin1) that is nested in Australia (Country). Accordingly, the first combination of nested locations in one country is best matched with Helen's FHLI. Referring to the table (5.5) the algorithm assigns the most nested PPL as a default location for the examined FHLI. In that case the disambiguated home location is 'Sydney' with uncertainty 2 (one place within one country, section 5.3.1). Therefore, Helen's home location falls in the 'city' category of the Figure (6.16) with uncertainty 2 (dark blue).

6.2.7 Summary

Overall, the analysis of the 'Flickr sample world data' revealed that:

- Ambiguity and uncertainty in FHLI is higher in national scope
- Analysis of the distribution of geo-tagged photo collections in line with analysis of FHLI might improve the confidence in classification and disambiguation process.

According to the study and evaluation of the Flickr sample world data in this section, an appropriate data selection strategy was developed in selecting a rich, unbiased and indicative data (chapter 5: section 5.3) from the Flickr database. The next section examines the Flickr sample GB data from different numerical perspectives.

6.3 Flickr Sample GB Data

This section summarises the statistics of the sample Flickr GB data set. As mentioned in chapter 4 (section 4.3.2), 3,245,866 photos were collected out of which 1,482,170 were taken in Britain (45.7%). Accordingly, 19, 780 unique posters were found who have at least one photo of GB.

	World Photos	GB Photos
Total Numbers	3,245,866	1,482,170
Mean	162.4	92.38
Median	10	10

Table 6.1. Mean and Median summary of GB photos and World Photos

	World Friends	GB Friends
Mean	67.38	4.65
Median	19	1
Max	982	231

Table 6.2 Mean, Median and Maximum numbers for World Friends and GB Friends

Out of the overall 19,780 GB posters, 52.19% (10,332) of them have accessible home locations and 47.91% (9,458) have no home locations information at all. Classifying the posters according to the availability of home locations reveals that:

Out of the 52.19% of posters with home locations, 7,839 (75.87%) have at least one GB friend in their friendship network. 64.42% of the have-home posters have at least one null-home friend and 84.30% have at least one have-home friend. The same analysis for the remaining 47.91% of posters with no home location reveals that:

57% (2,912) have at least one friend with no home location available and 76.27% (3,889) have at least one friend with available home location. 10.82 % have at least one GB friend in their GB friendship sub network. Overall, 65.41% (12,938) posters have one GB friend. 60.59% (7,839) of them have their home locations available and 39.41% (5,099) have null value for their homes. All the above numbers can be summarized in the following table.

		Have at least one GB friend	Have at least one Have-home Friend	Have at least one Null-home Friend
19,780 GB posters with 1,480,170 GB photos and 3,245,866 world photos	52% have-home posters	75%	84%	64%
	47% null-home posters	10%	76%	57%

Table 6.3. Summary of the basic numerical analysis of the GB data set

6.4 Flickr Sample World Data vs. Flickr Sample GB Data

This section examines the attributes and statistics of the GB data set collected according to the 19,780 GB posters. The aim here is to study and compare the Flickr GB data to the Flikcr world data (used for completing a pilot study of Flickr - see chapter 5: sections 5.2 and 5.3).

6.4.1 Photos

As expected the frequency distribution of GB photos (Figure 6.17) has relatively the same trend as the the sample world photos collected and analysed (Figures 6.3-8). It follows a bi-modal log-normal distribution with large population with one photo and the second population which follows the log-normal distribution. The fact that the GB posters follow the same trend as a randomly selected Flickr posters (sample world data) reconfirms that the selected Flickr sample GB data can be an indicative sample and the findings can be generalized to other sample data in same situation and concept.

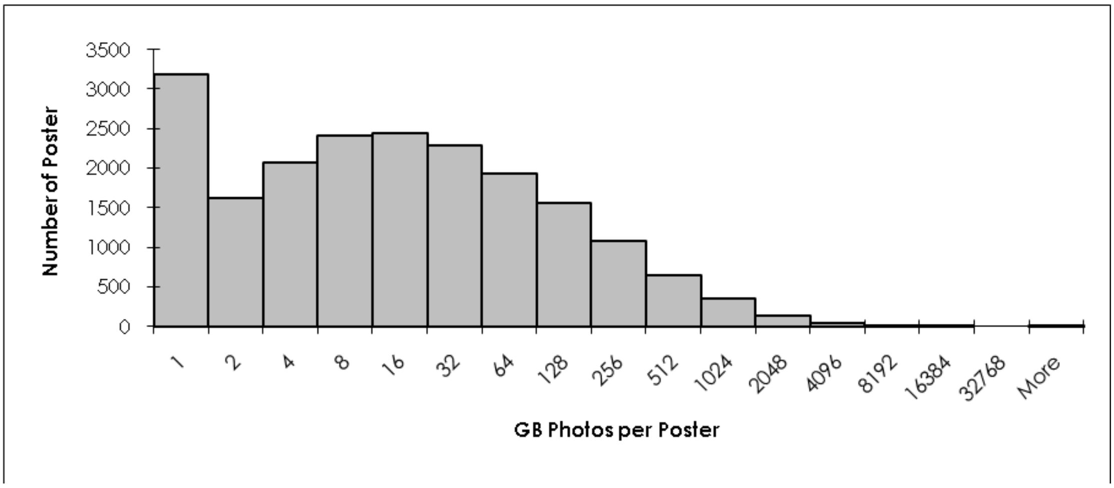


Figure 6.17. Frequency Histogram of GB photos per poster

6.4.2 Friends

Figures 6.18 and 6.19 compare the number of GB friends (friends who uploaded at least one photo of GB) with the number of world friends (total number of friends with no restriction for each GB poster).

As can be inferred there is a large population with no GB friend and then a considerably high number of posters with one friend only. The second population follows the log-normal distribution. The population of the posters with no GB friend significantly outweighs those with one world friend. This might be explained by the restriction applied to get the GB friendship sub-network for each poster. Those with no GB friend might have several other friends who do not have GB photos uploaded in the examined period, and therefore have been excluded from analysis. Overall, the two examined variables have relatively similar patterns with slightly higher peak for world friends and larger population of posters with no GB friends. This trend can be very much expected while considering the 19,780 GB posters that make the GB friendship sub-network against all users of Flickr that make the world friendship network. Having a friend in Flickr has a higher probability than having a friend from one specific group of posters (GB posters) in Flickr.

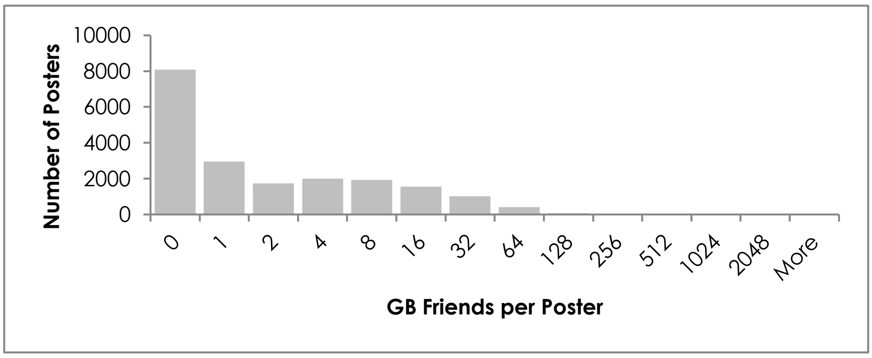


Figure 6.18. Frequency histogram for number of GB friends

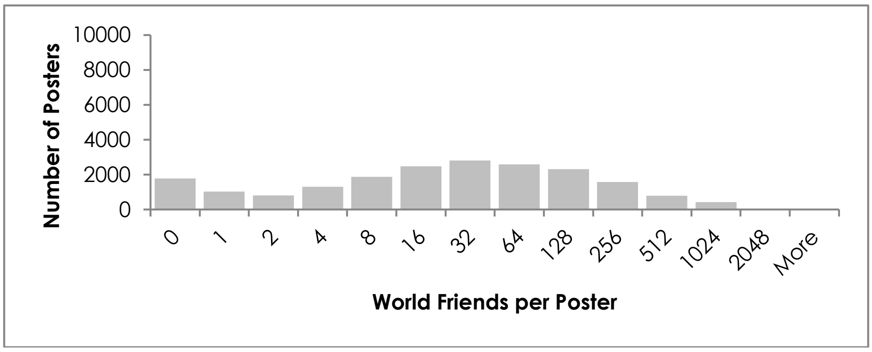


Figure 6.19. Frequency histogram for number of world friend.

6.4.3 Friendship Network Size vs. Prolificness

Figure 6.20 demonstrates the relation between the number of photos and number of friends for each GB poster. It shows there is no specific relation. The poster with the highest number of photos does not have the most or least number of friends, and the posters with zero friends have variety number of photos. In order to get a better understanding of the relation, the most prolific poster is removed from Figure 6.20. Consequently, more posters with a high number of friends are revealed. Interestingly the poster with highest number of photos does not have the highest number of friends, while the posters with a large number of friends lying on the x axis have small number of photos. The majority of the most prolific posters have a small number of friends and the majority of posters with a high number of friends have a small number of photos (Figure 6.21).

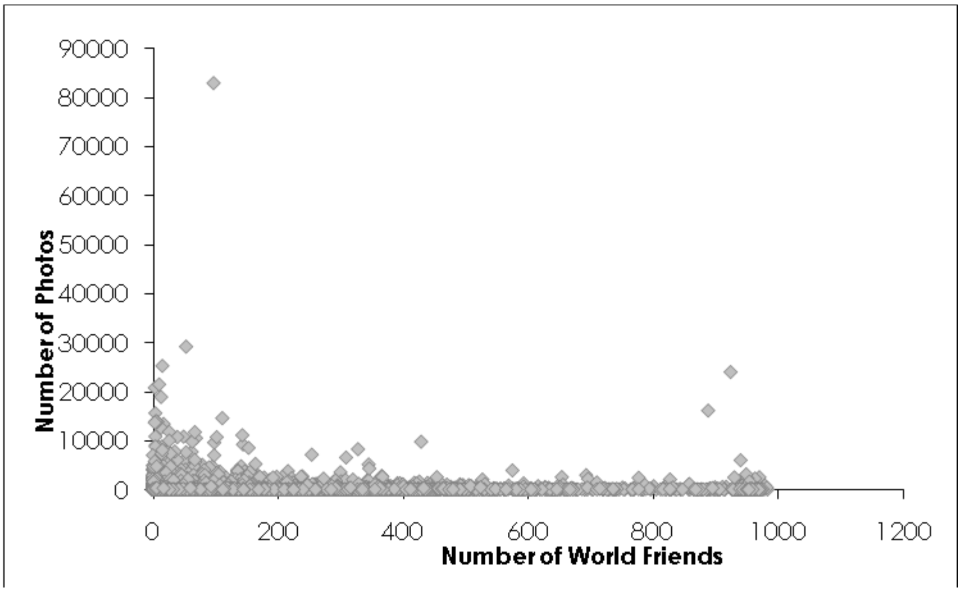


Figure 6.20. Friendship network size vs. prolificness for GB posters

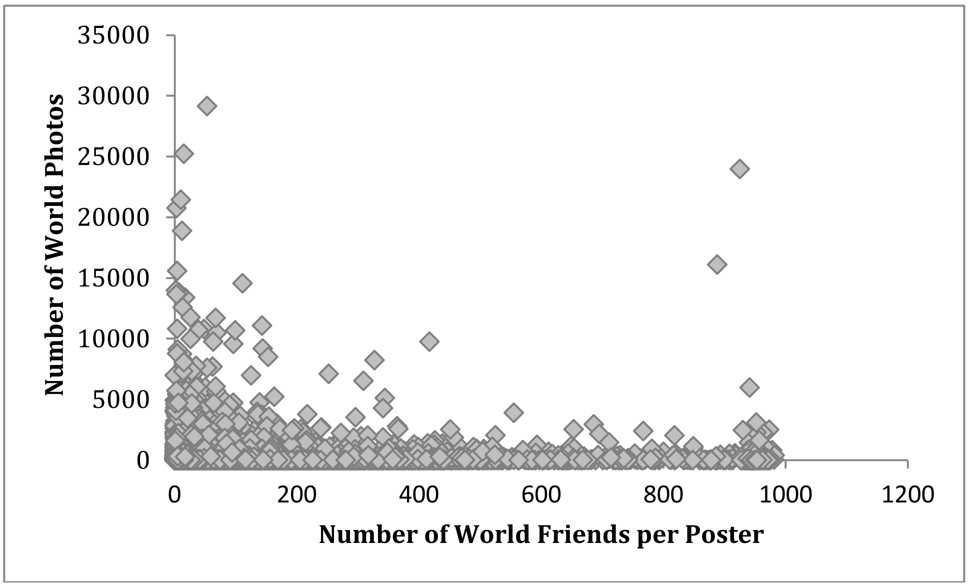


Figure 6.21. Friendship network size vs. prolificness after removing one poster with extremely high number of photos

Figures 6.22 and 6.23 illustrate the relation between number of GB friends and prolificness (number of GB photos). As the Figure shows, a poster with an unusual large number of photos (65,283) has a small number of GB friends (8). In order to get better view of the relation for other posters, in Figure 6.22 the extremely prolific poster is removed from analysis. In addition to general posters with a random relation between number of photos and friends, still the posters with the largest GB friendship sub-networks (150- 200- 250 friends) have no or a very small number of GB photos. This can lead to conclude that overall there is no specific relation between size of the GB friendship sub-network and prolificness of GB posters. However there is an inverse relation between the two variables for the most prolific posters and the posters with the largest GB friendship sub-network. The patterns here are not different from those identified for the world data set Figures 6.20 and 6.21. This can indicate that the sub-network of GB friends and GB photos does not have specifically different attributes from the sample world data set in respect to the examined variables in this section.

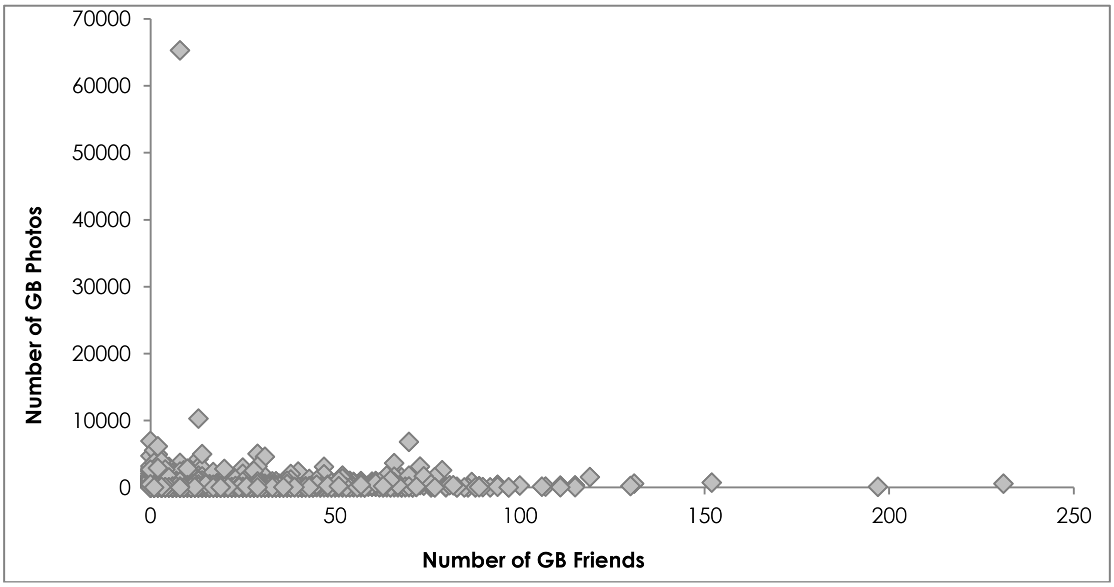


Figure 6.22. Friendship network size vs. prolificness for GB posters

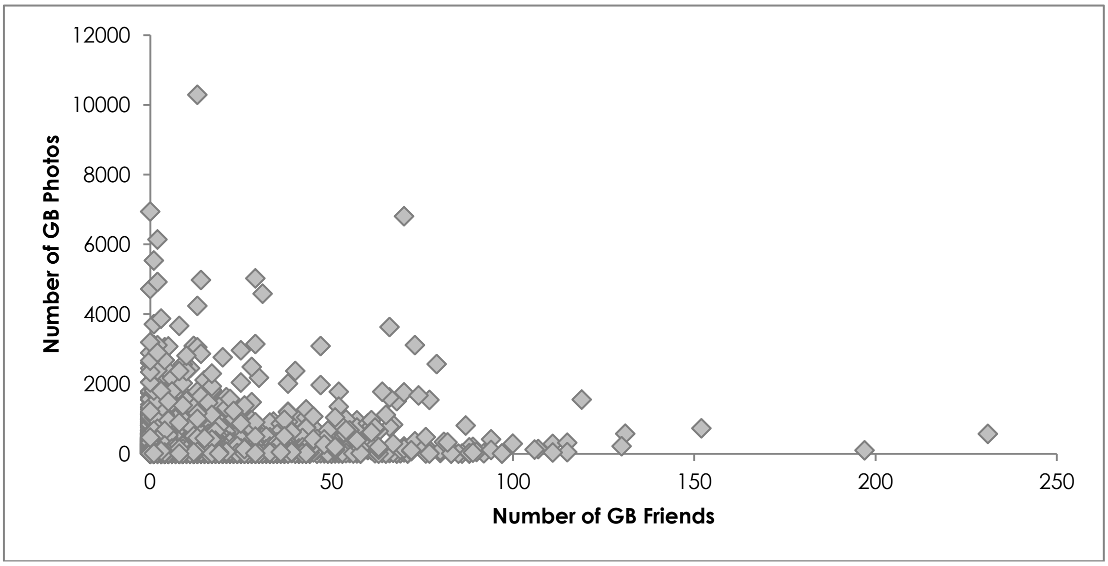


Figure 6.23. Friendship Network vs. prolificness for filtered GB posters

6.5 Have-Home Posters vs. Null-Home Posters

This section studies the posters according to two different categories:

- Posters with home locations accessible in their profiles (Have-Home posters)
- Posters without any home locations (Null-Home posters)

6.5.1 Photos

The following two Figures show the distribution of GB photos per poster in two different poster categories. They compare the behaviour of posters according to their availability of FHLI. Interestingly, both categories have relatively similar patterns. There is a large population of posters with one photo only and then trend is followed by normal distribution. Overall, the have-home posters are slightly more prolific than null-home posters.

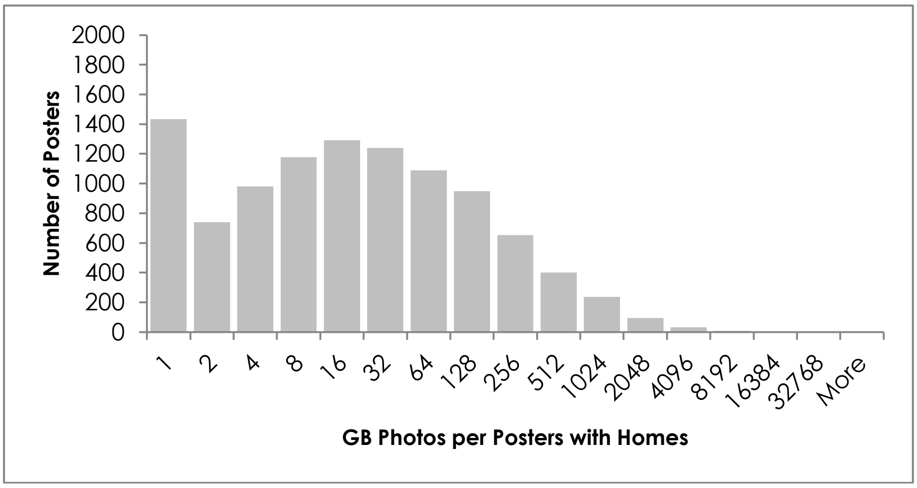


Figure 6.24. Distribution of GB photos per have-home poster

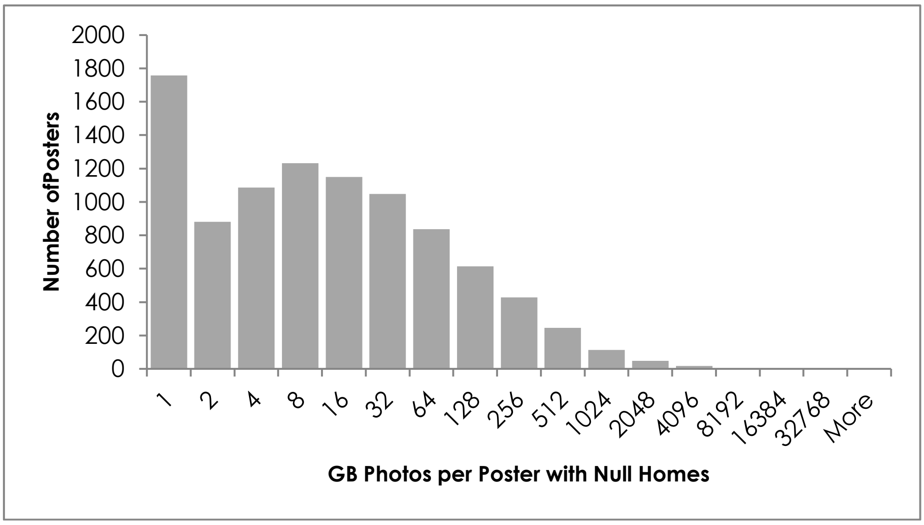


Figure 6.25. Distribution of GB photos per null-home poster

6.5.2 GB Friends

The following two Figures compare the number of GB friends for posters with home locations available and those with null homes. As can be seen in Figures 6.26 and 6.27 the distributions in both categories have a large population with no GB friend followed by those with one friend only. Although the identified patterns are relatively similar, more than half of the null-home posters have no or one friends only. The remaing 33% have two or more friends. Therefore, one can hypothesize that have-home posters have slightly higher probabilty of developing larger GB friendship networks than those posters who are not willing to reveal their home locations.

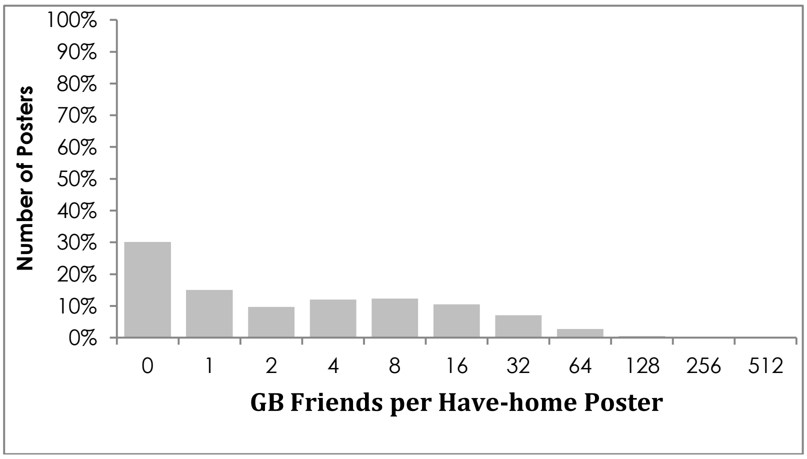


Figure 6.26. Number of GB friends for have-home posters

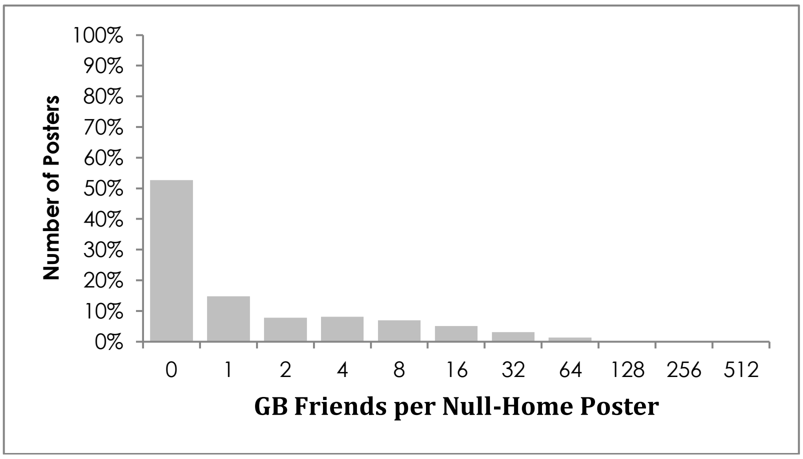


Figure 6.27. Number of GB friends for null-home posters

This finding here is totally consistent with the numbers and percentages calculated in section 6.3.

6.5.3 World Friends

Figures 6.28 and 6.29 demonstrate the frequency distribution of world friends for have-home and null-home posters. Here the difference among the two population is slightly more than the previous section (GB friends). The majority of have-home posters have more than 8 friends, while the null-home posters has a significantly higher number of posters with no or one friends only. The existing pattern here reconfirms the slightly less populaity of null-home friends in friendship network. As has been hypotheised (section 6.5.2) and can be recognized here, have-home posters are slightly more prone to have larger networks of friends (world and GB). This pattern can be attributed to the limited number of GB posters compared to all Flickr posters. Consequently, the restricted number of GB posters in GB friendsip sub-networks has a little effect on the patterns of GB friends compared to world friends for have-home and null-home posters.

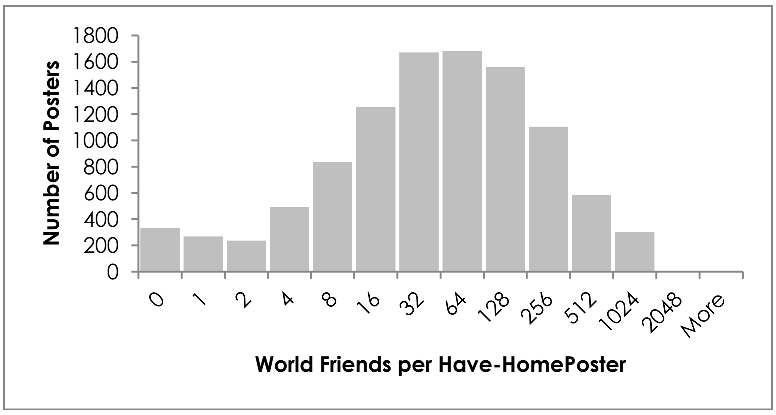


Figure 6.28. Number of world friends for have-home posters

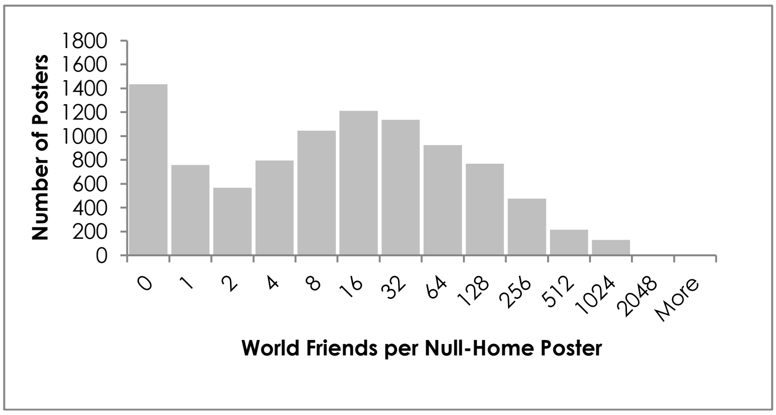


Figure 6.29. Number of world friends for null-home posters

6.5.4 Friendship Network Size vs. Prolificness

The last thing to consider for the two categories is to test if there is any relation between number of GB photos and GB friends. Figure 6.30 demonstrates the number of GB friends (number of friends) vs. Prolificness (number of photos) of have-home posters. There is an extremely prolific poster with 65,283 photos and 13 friends only. Accordingly, in order to see the result for average posters, the extreme poster is removed from analysis and the results are depicted in Figure 6.31. The second graph better demonstrates the relation of the two examined variables. As can be seen there is no obvious relation between the number of photos and the number of friends. The most extreme posters with the highest number of photos are not necessarily most prone to develop large networks of friends, and the posters with the higher tendency of having large networks of friends are not the most prolific ones. There are no specific recognisable patterns between the number of friends and photos for have-home and null-home posters (Figures 6.30 and 6.32).

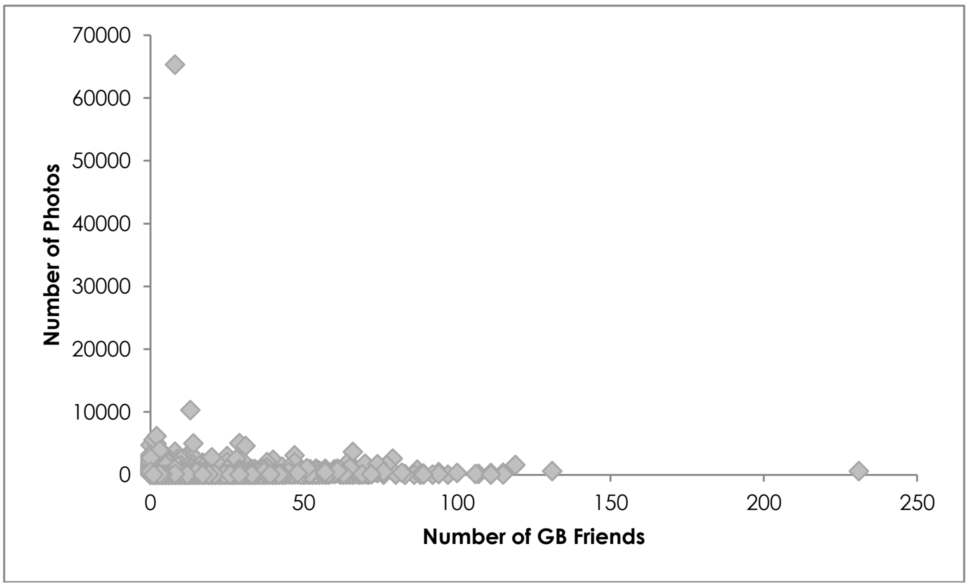


Figure 6.30. Friendship network size vs. prolificness for have-home GB posters.

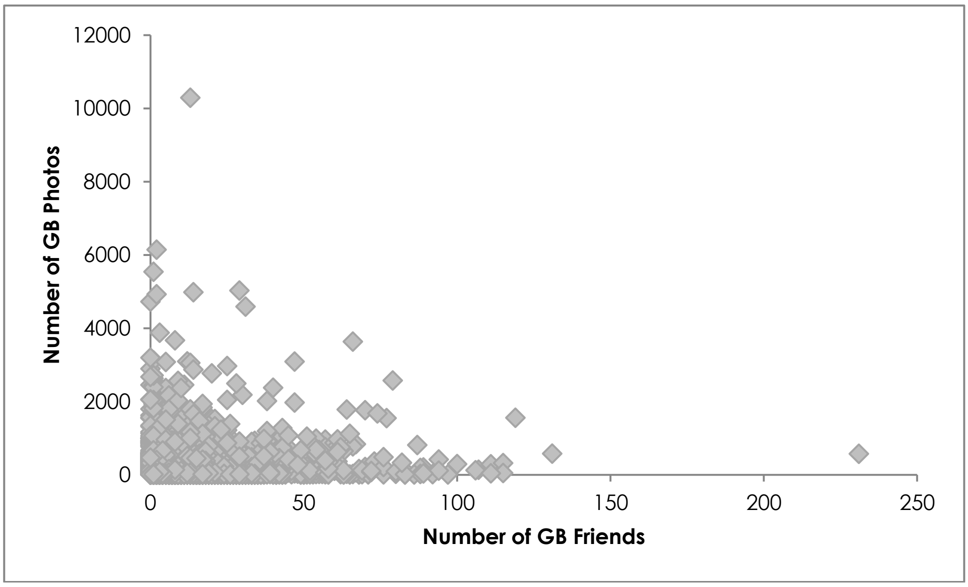


Figure 6.31. Friendship network size vs. prolificness for filtered have-home GB posters

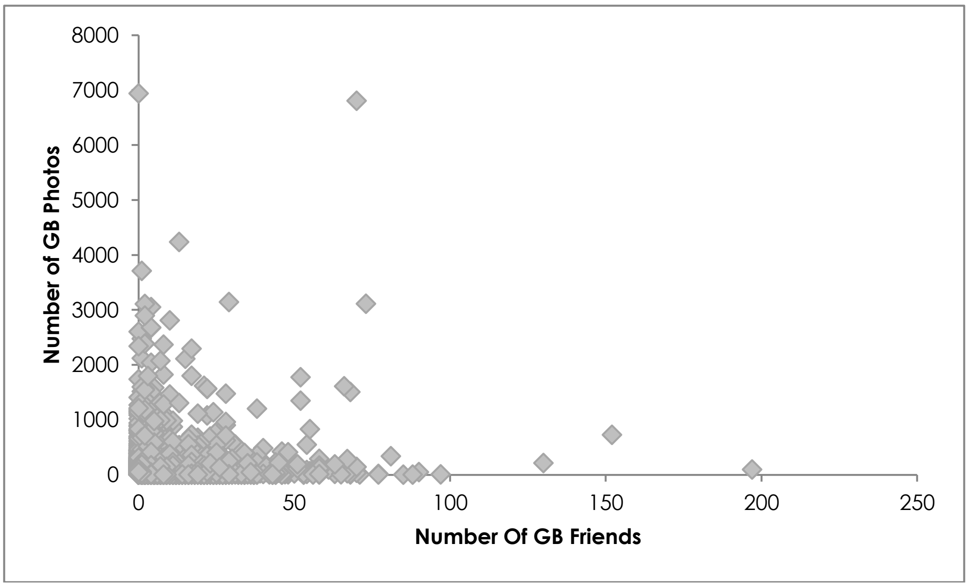


Figure 6.32. Friendship network size vs. prolificness for null-home GB posters

6.5.5 Uncertainty Classification

This section demonstrates the result of the uncertainty assignment to FHLI of the examined GB posters. The disambiguated home locations of GB posters are assessed against the four uncertainty classifications developed in chapter 5 (section 5.3). As shown in Figure 6.33 the null home (uncertainty 5) and uncertainty 2 (one location in one country) have the highest percentage of the data respectively with a small difference (7%). In other words as has been identified before, Flickr posters in majority leave their home location as blank or associate themselves with one specific location within a country (example: 'London, UK'). However, the remaining FHLI are assigned either Uncertainty 3 (several places within a county, example: 'London, Brighton, UK') or Uncertainty 4 (several places in several countries, example: 'Tehran, London, UK'). This finding is totally consistent with what have been found in the analysis of the sample world data set (section 6.2.6: Figures 6.15-16). This consistency reconfirms the validity of the amendment applied to the uncertainty assignment of sample world data set before applying it to the GB data set (chapter 5: section 5.3). Moreover, the fact that the UC3 is considerably higher than UC4 also vindicates the decision made in restricting the data to a national scope (chapter 4: section 4.3.2 and chapter 5: section 5.3).

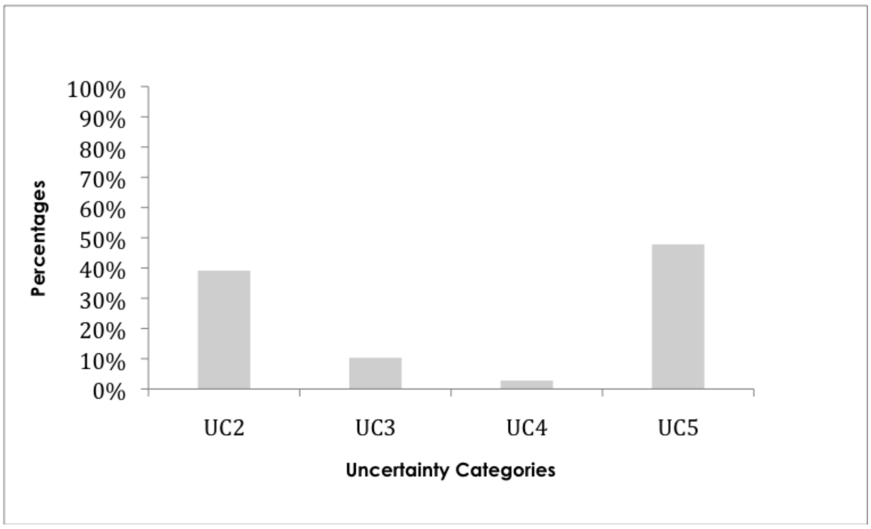


Figure 6.33. Percentage of the examined FHLI in each of the developed uncertainty classifications

6.5.6 Friendship and Home Location Availability

The numbers in Table 6.4 reveal that have-home posters and null-home posters both are more likely to befriend with have-home posters. This finding is consistent with what has been discussed in sections (6.5.2-3), which indicates that null-home posters are slightly less likely to develop large networks of friends.

	Have-Home Friends	Null-Home Friends
Have-Home Posters	%84	%64
Null-Home Posters	%76	%57

Table 6.4. Percentage of have-home and null-home friends for each of the examined posters category

6.6 Summary

This chapter covered the basic numerical analysis of the two sampled data sets of this study. The first section studied the sample world data set. It included the general analysis of the sample world data set. The photos and friendship patterns and trends were demonstrated and examined during the five year examined period. The results of the uncertainty classification and measurements (proposed in chapter 5: section 5.2) are demonstrated in section 6.2.6.

The second section studied the Flickr sample GB data set. It contained a more detailed numerical analysis from a variety of perspectives (i.e. GB friends vs. world friends, friendship network size vs. prolificness, have-home posters vs. null-home posters and uncertainty). Where appropriate the non-parametric Mann-Whitney U test[50] has been conducted to measure the probability if the examined two sampled come from different populations. By interpretation of probability (p value) that exceeds the null hypotheses[51] declaration (p<= 0.05) one can conclude that there is no significant difference between the two sampled population (have-home and null-home posters, 6.5.1-3).

[50] Introduced by Frank Wilcoxon, 1945

[51] Null hypothesis assumes that the two sampled groups do come from the same distribution.

The next chapter applies the developed visualization application (chapter 5: section 5.4) to visually explore the data and find appropriate answers to the research questions (chapter 1: section 1.5).

7 Visual Analysis

7.1 Introduction

In this chapter attempts are made to demonstrate the capability and usefulness of the developed visualization application in data analysis and better understanding of the spatio-social patterns and interactions among social entities. In line with research questions the data here have been studied visually in regards to

1. Geographic distribution of photos
2. Geographic distribution of GB friends and
3. FHLI

For each of the above attribute several indicative examples are provided. As explained in the methods chapter, the standard ellipse was selected to summarize point distributions. Therefore, size, direction and centre of the ellipses have been used for assessing and comparing the distribution of the photos and friends in different locations. In order to follow a systematic strategy, the identified populated places with GB posters, were broken into several smaller categories:

- Australia and New Zealand
- Asia
- Africa
- South America
- North America and Canada
- Europe

These categories were selected according to the distance, similarity, culture, language and number of GB posters. According to the visualization application's capabilities, the exploration starts from the selected areas with the mouse and then the area will enhance to larger vicinity until robust and consistent patterns are found.

7.2 Geographic Distribution of Photos

In order to study the geographic distribution of photos for each of the above selected areas, data have been studied visually through the developed visualization application. As described in chapter (5) a geographic area can be selected by the user in the left sketch (named as A in Figure 5.11). The boundary of the selected location can be expanded or shrunk. Accordingly, the total number of posters living in the chosen area can be demonstrated along with the geographic distribution of their photos in standard ellipse format. For example Figure (7.1) shows a selection that starts from Wellington and expands by threshold 100 until the whole New Zealand is covered. This is shown in sketch A with empty grey circle all around the chosen area. The orange ellipse inside the threshold circle is proportionate to the number of GB posters who live in New Zealand (194 posters). According to the selection in sketch A, the photos uploaded by the posters of the selected area are demonstrated in sketch B. As is shown in sketch B of the Figure (7.1) the geographic distribution of GB posters from New Zealand are stretched from home location toward Europe and US (blue ellipse).

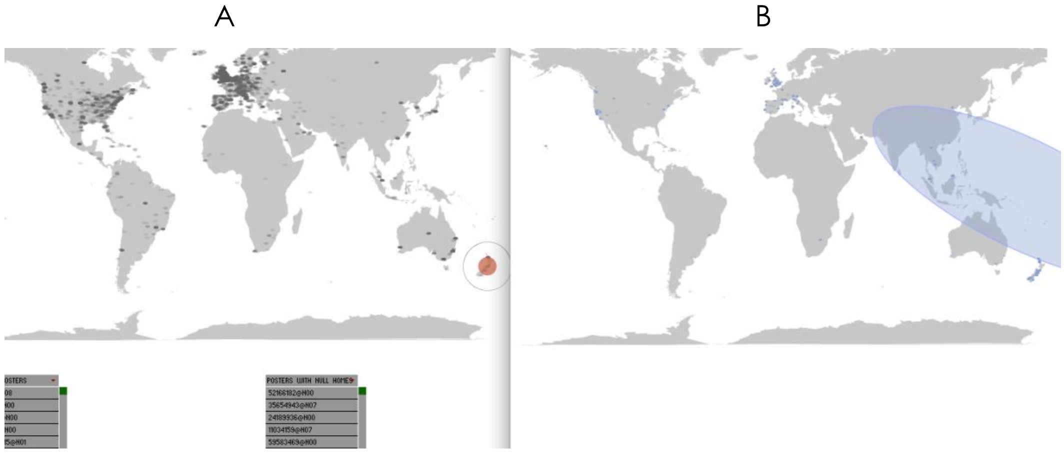


Figure 7.1 33,936 photos for 194 GB posters in close proximity of Wellington (threshold: 100 pixels)

Figure (7.2) is another example of how Flickr spatio-social data can be studied. Unlike the previous example, here the number of posters who live in the selected area have resulted in a bigger circle than the threshold circle. Therefore, the threshold circle is inside the orange circle. While in Figure 7.1 the small number of posters in comparison to the selected area (New Zealand) successfully drawn with an orange circle inside the threshold circle. Figure 7.2 shows the most popular places that the posters who live in Calgary and close vicinity (grey circle inside the orange circle) have taken photos (sketch A). The blue ellipse in sketch B shows the summary distribution of the photos.

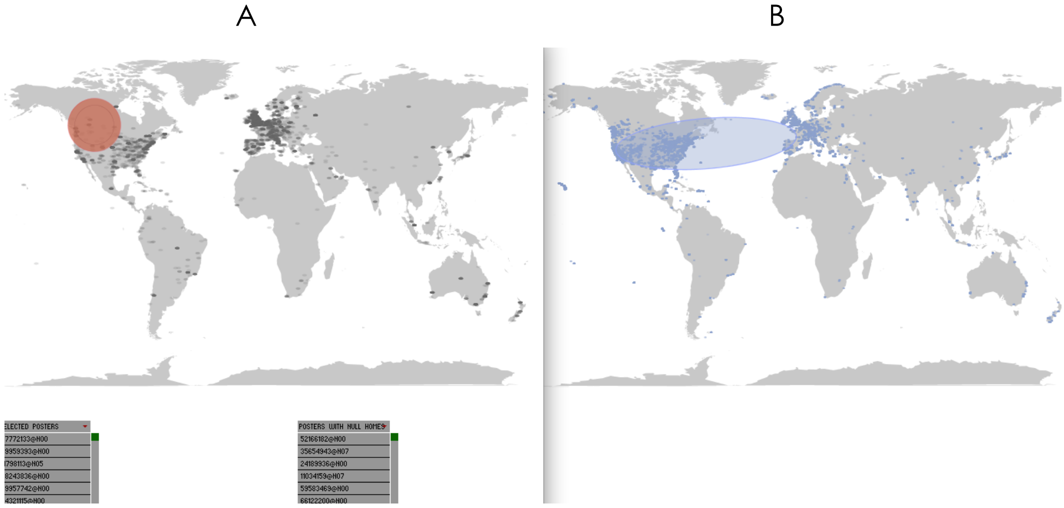


Figure 7.2 388,406 photos for 1,154 GB posters living in Calgary and close vicinity expanded by threshold 80 (grey circle inside the orange circle).

Analysis and study of the rest of the places by repeating the same methods have revealed interesting patterns. In the first place as the two example Figures demonstrate it has been identified that there is strong relation between the FHLI and geographic distribution of photos. In other words, GB posters are more likely to take photos of their nearby places more than farther away. The worked example already introduced in this study with a hypothetical name of 'Helen' also demonstrates this pattern in an individual view (Figure 5.9). Overall, the geographic distributions of photos in the majority of the examined places (Australia, New Zealand, Asia, South

America, Canada, North America and Europe) are lending towards a close proximity to where the posters live and Europe. In Africa, less evidence of taking photos of close proximity has been found (Figure 7.3).

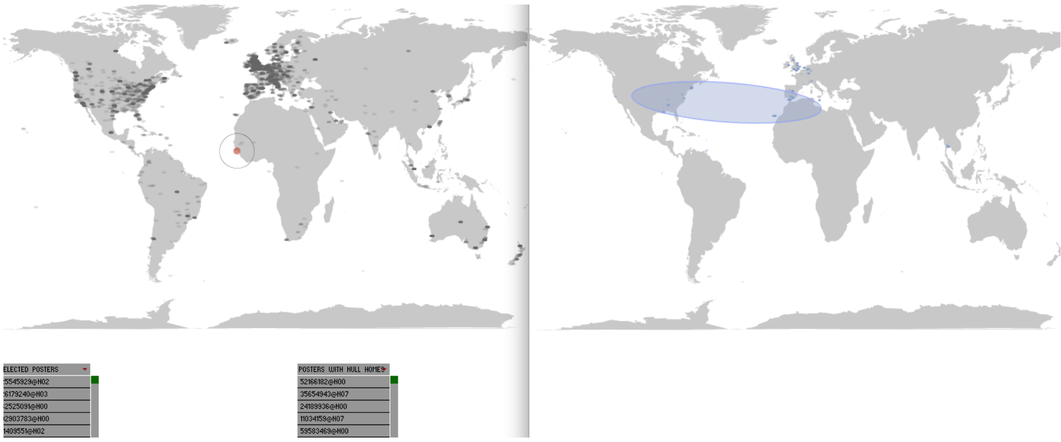


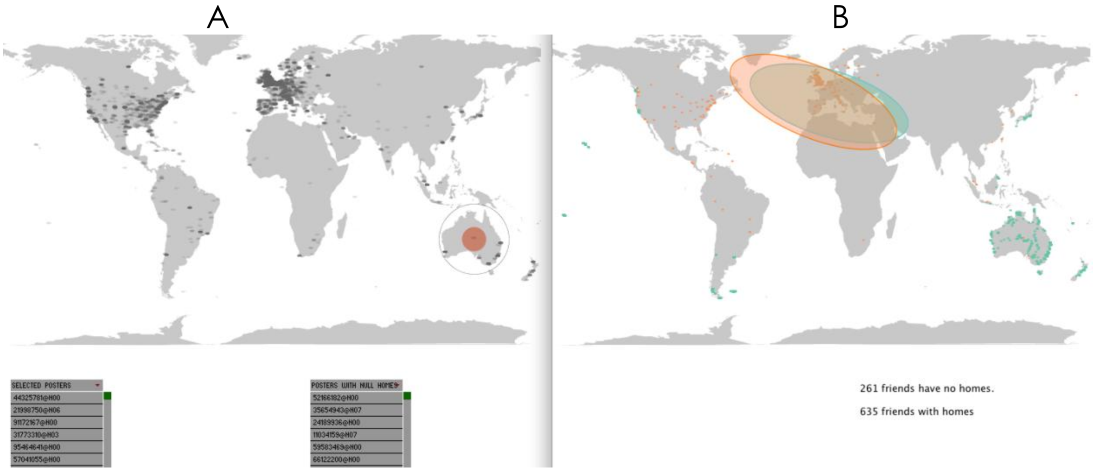
Figure 7.3 5,451 photos for 36 GB posters in Tombali expanded by threshold 70.

Unlike other examined locations, African GB posters have found to be least interested in taking photos of their nearby places. Figure 7.3 shows the typical patterns in photo taking behavior of African GB posters.

7.3 Geographic Distribution of Friends

Before using the visualization application for analysis of the friendship network for the sampled GB posters, it is necessary to emphasize here that the friendship network of GB posters in this study (chapter 4: section 4.2.2) is limited to those posters who also have photos of GB. In other words, friendship is studied amongst the selected GB posters. Those friends of GB posters who do not have a GB photo are excluded from analysis. Therefore, the examined friendship network here is called 'GB friendship sub-network'.

In the early stages of the analysis of the friendship network it has been realized that since there are two different categories of friends an appropriate strategy is needed to examine the friendship network comprehensively. The first category of friends is GB posters who have accessible home location information in their profiles. This category of friends has been visualized on the map according to their disambiguated FHLI. Therefore, their whereabouts and distributions can be easily explored and navigated interactively in the application. However, the null-home friends need to be studied according to another variable that can be indicative of their geographic footprints (chapter 5, section 5.4.2.2). In doing so and according to the data set collected (section 4.3.2), geographic distribution of photos has been used as an indicative of the null-home friends' footprints. This decision is supported by the fact that as explained in section (5.4.2) each sample poster in this study has at least one geo-tagged photo. In addition, the geo-tagged photos bear some reference to the physical places that the poster actually has been. Therefore, a collection of geo-tagged photos implicitly demonstrates the geographic footprint of its poster. Accordingly, exploring spatial relation between posters of a selected area and their friends' locations can be done through interaction with the visualization application. As discussed and demonstrated in section (5.4.2) have-home and null-home posters are visualized according to their home location or photos distributions respectively. For example Figure (7.4) shows the geographic distribution of GB posters who live in Australia (the area demonstrated by empty grey circle in sketch A). The orange ellipses in the other sketch show where the 261 have-home friends live and the green ellipse shows where the 635 null-home friends have taken photos. Reviewing from the chapter 5, the orange circle on the sketch A is proportionate to the 301 Australian GB posters. The grey circle around Australia shows the threshold of the area selected for analysis (sketch A).



SELECTED POSTERS
44325781@N00
21998750@N06
91172167@N00
31773310@N03
95464641@N00
57041055@N00

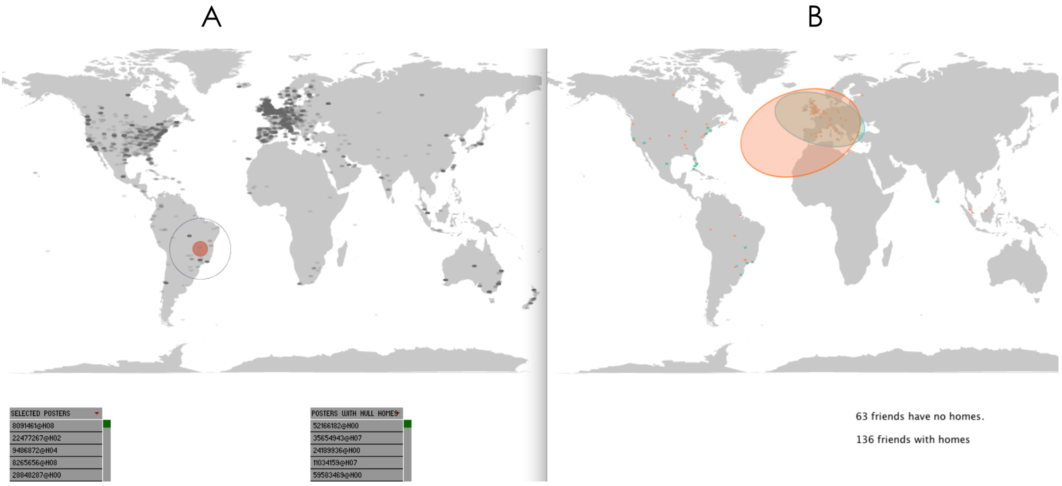
POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00

261 friends have no homes.

635 friends with homes

Figure 7.4 Spatial pattern of friendship for 301 GB posters in Australia.

Spatial patterns of friendship for the rest of the populated places with the sampled GB posters were explored visually in the developed visualization application (as demonstrated above). Interestingly, a consistent trend has been found in the spatial pattern of friendship. In majority of the cases the geographic footprints of have-home friends share a relatively similar distribution with the geographic distribution of the places that null-home posters have taken photos. Another example of this pattern is illustrated in Figure (7.5) with 147 GB posters who live in Brazil and its vicinity with threshold 130 (sketch A). As can be seen on the picture 136 friends are mapped according to their home locations (orange circles) in sketch B while the remaining 63 null-home friends are visualised according to their geographic footprints of their geo-tagged photos. As can be seen and was expected the summary distribution of points have close centre, size and direction focused around Europe.



63 friends have no homes.

136 friends with homes

Figure 7.5 303 friends for 147 GB posters in Brazil expanded by threshold 130.

This identified pattern might not be correct for individual views (section 5.4.2) but in aggregated view, that covers more posters and larger area, the two examined distributions tend to be quite close to each other.

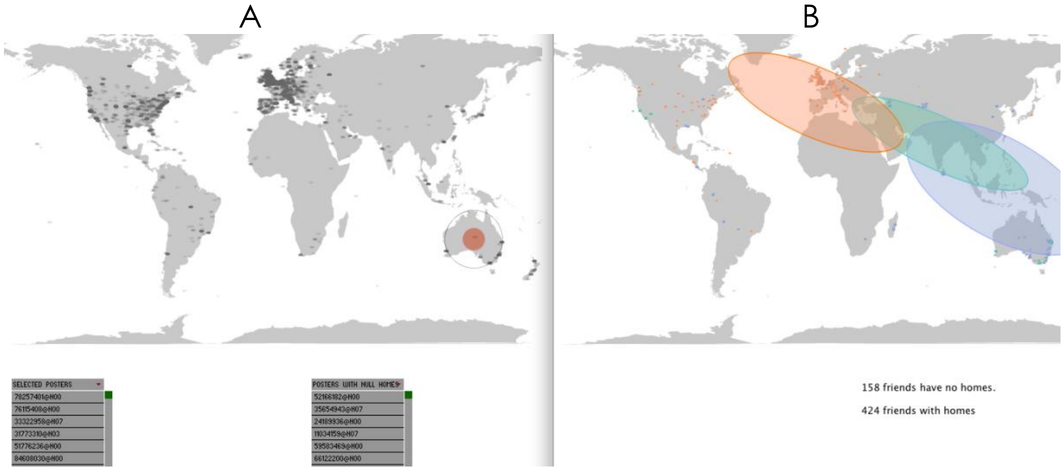
7.4 FHLI, photos and friends

Up to this stage the visual analysis of friends and photos have been conducted successfully through interaction with the developed application (chapter 5). Here, efforts have been done to examine all the three collected variables for sampled GB posters visually. In other words for posters of each selected area the following three variables are calculated and shown accordingly:

- Geographic distribution of photos (blue)
- Geographic distribution of home location of have-home friends (orange)
- Geographic distribution of photos of null-home friends (green)

From each of the location categories populated with the sampled GB posters, several locations (proportional to the density of each area) are selected for study and analysis.

Figure (7.6) has been selected as a typical patterns identified in majority of the locations populated by GB posters. This particular example shows the three examined attributes of the Australian GB posters. As described before, sketch A (left side) shows the selected area for analysis and number of posters who live in that area. The orange circle size is proportional to the number of posters who live in the selected area. Sketch B interactively updates itself according to the area selected on sketch A. Accordingly, this picture demonstrates that the Australian GB posters tend to take more photos in their close vicinity (blue ellipse). Their have-home friends are focused around Europe (orange ellipse) while their null-home friends have photos collection stretched between Australia and Europe (green ellipse). Interestingly all the three examined distributions share relatively the same direction (stretched from Australia towards Europe).

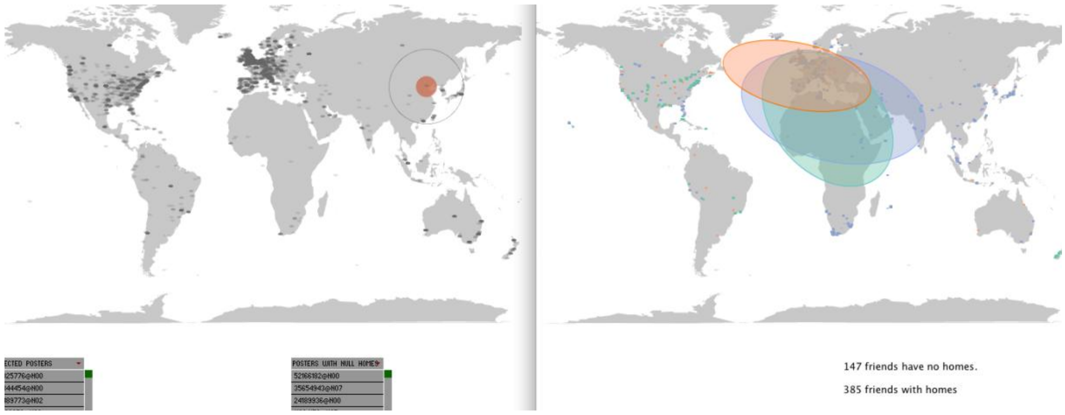


158 friends have no homes.

424 friends with homes

Figure 7.6 Distribution of photos, have-home friends' homes and null-home friends' photos for posters of Australia

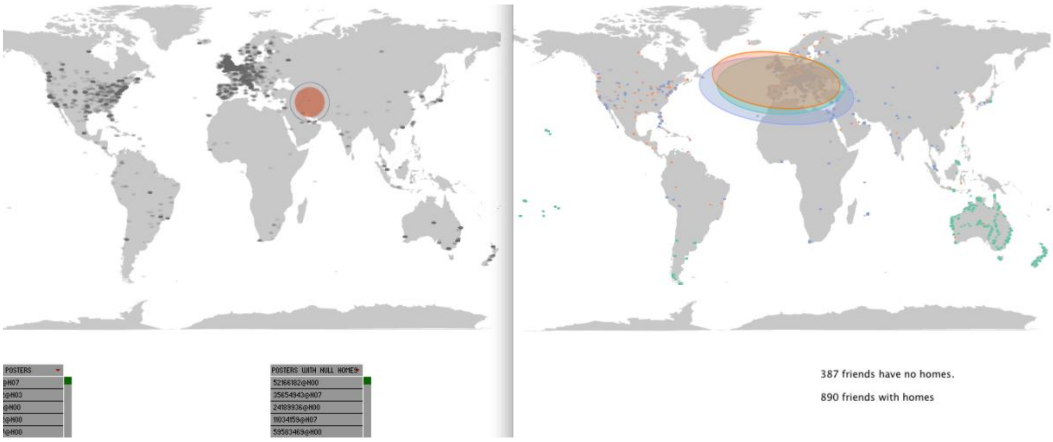
Exploring the data visually by comparing the distribution of the mentioned three variables for the locations populated by sampled GB posters of this study revealed that in most of the cases the distributions are very close to each other or at least share the same direction and or centre. The following two Figures are selected to demonstrate the two extreme cases in the distribution of the examined three variables. Figure (7.7) shows the distribution of photos, have-home friends' homes and null-home friends' photos for Asian GB posters. As can be seen although the examined variables are relatively focused in close proximities, their distributions have slightly different directions and their centres are located in slightly different locations. On the other hand, Figure (7.8) is indicative of those locations that the distributions are extremely close to each other (centre and direction).



147 friends have no homes.

385 friends with homes

Figure 7.7 Distribution of photos, have-home friends' homes and null-home friends' photos for posters of Asia (Beijing, threshold 160)



387 friends have no homes.

890 friends with homes

Figure 7.8 Distribution of photos, have-home friends' homes and null-home friends' photos for posters of Asia (Tehran, threshold 80)

Study and analysis of the three spatio-social attributes of the sampled GB posters have been conducted by following the same strategy through interaction with the developed visualization application.

7.5 Summary

This chapter demonstrates how the developed visualization application enables users to explore the data visually. Attempts were made to explore the data from three different perspectives. That includes geographic distribution of photos, geographic distribution of friends and FHLI. For each of the examined variable an indicative Figure has been selected and a snapshot of the application is provided for clarity. The next chapter answers the research questions through exploring the data interaction with the application.

8 Discussion

8.1 Introduction

This chapter applies the findings of the visual analysis of the data that have been conducted in the previous chapter to discuss and answer the research questions (section 1.5). For each research question the identified patterns of the examined distributions are discussed and possible assumptions and hypotheses are made accordingly.

8.2 Research Questions

RQ1. What limitations are involved in applying VGI in spatial social network presentation?

In applying VGI to spatial social network presentations the identified limitations can be defined in three different categories:

Servers and service providers

- Bug: as mentioned in the data chapter (chapter 4, section 4.3) during the course of the data collection of this research it has been identified that there is a bug in the Flickr service. In response to requests that return large amounts of data, the server returns duplications. In the sample world data, breaking the queries into smaller time periods resolved the problem (chapter 4, section 4.3). However, in the GB data collection it was aimed to collect all photos in GB borders during the 5-year sampling period. In that situation the bug resulted in a more significant problem. As a result the first (15 pages with 250 photos) 3,750 photos of each day were collected and the rest, that were duplicates, were removed from the data.
- Flickr service: Considering the large volume of data that needs to be collected, the internet connection and speed of the server play essential roles in the time and efficiency of the data collection process.

VGI

Ambiguity and uncertainty: since VGI is provided by people voluntarily without any particular format and accuracy, there are considerable levels of ambiguity and uncertainty that need to be dealt with before analysis. Although FHLI disambiguation has been completed in the most appropriate format for this study (chapter 5, section 5.3), there is an unavoidable level of uncertainty in the data that affects the generalization of the inferences, assumptions and conclusions.

Visualization: Packages and Techniques

- **Viz packages:** The pre-packed ready to use visualization packages are inadequate in dealing with unique characteristics of different data sets (Fry, 2007). Therefore, a customized visualization application was developed (chapter 5, section 5.4.2). Although the application was useful for detecting the patterns and exploring the data, considering the time constraint there is always place for improvement of the application for getting better results (specially for null-home posters and the uncertainty involved).
- **Data analysis through graphics**: There are some concerns about using graphics in data analysis (Montello, 2002). Visual analysis of the data can be affected by perceptual failings and optical illusions and therefore can result in deriving wrong conclusions from the analysis. This problem is more severe when it comes to judging the value of circles on the map (Flannery

perceptual scaling). Extra efforts are required while developing a visual application to make up for underestimation of areas and volumes through human minds. In order to follow Tufte's rule of 'telling the truth about the data', and as suggested by Psychology and Psychophysical research in geography and visualization (Montello, 2002; Dent, 2008), perceptual scaling has been applied to compensate for the underestimation of human's perception of areas (Krygier and Wood, 2002). Although attempts were made to reduce the risk of illusions while analysing the data and answering the research questions (chapter 5, section 5.4.2.3) through visual analysis, there is always a possibility of unavoidable human perceptual problems and optical illusions (Ihaka, 2003)[52].

Overall, the user generated data, although freely available and having potential in revealing patterns and trends in many aspects of human interactions and behaviours, will inevitably come with inherent ambiguity, uncertainty, inconsistency and conflict that leave the quality of the data as a subject of debate and make the analysis and visualization a challenging task.

RQ2. Can geographic distribution of photos contribute to prediction of FHLI?

This question requires the analysis of the photo distribution of GB posters in regards to their home locations. Referring back to chapter 7 (section 7.2) photo distributions of GB posters have been examined visually through interaction with the developed visualization application. Based on the visual analysis of the data (Figures 7.1-3) a general pattern has been identified for photo distribution of the sampled GB posters. According to the consistent identified patterns, one can hypothesize that the majority of the GB posters have higher tendency in taking photos of their nearby places than farther distances. In other words, the summary of the distributions indicates that the centre and direction of the photo taking behavior of GB posters are stretched between home locations and Europe (mandatory place for having at least one photo).

Here for each of the selected five categories of populated places with GB posters (chapter 7, section 7.1) the distribution patterns are studied against the above general pattern. GB posters of Australia, New Zealand and South America have their photos distribution stretched from their close vicinity towards Europe. This assumption might not work for each individual poster, but looking with a wider perspective reveals that it is the case for aggregated view of posters. Figure (8.1) demonstrates the typical patterns in the geographic distribution of geo-tagged photos of the mentioned place categories.

[52] Statistics 120, Information Visualization lecture notes, chapter 5, Available athttp://www.stat.auckland.ac.nz/~ihaka/120/Notes/ch05.pdf Accessed 7th Oct 2011

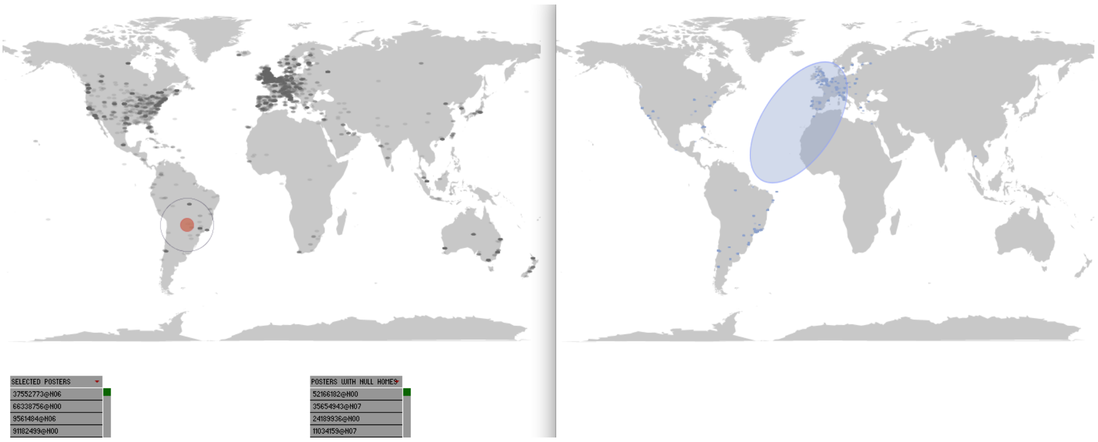


Figure 8.1 23,876 photos for 121 GB posters in Fortaleza expanded by threshold 110

Exploring the Asian GB posters' photo taking behavior in some cases shows a slightly more bias towards US than expected. Although they have their photos distributions spread between home locations and Europe, there is considerable number of photos taken in US as well. An example of this pattern is shown in Figures 8.2-3.

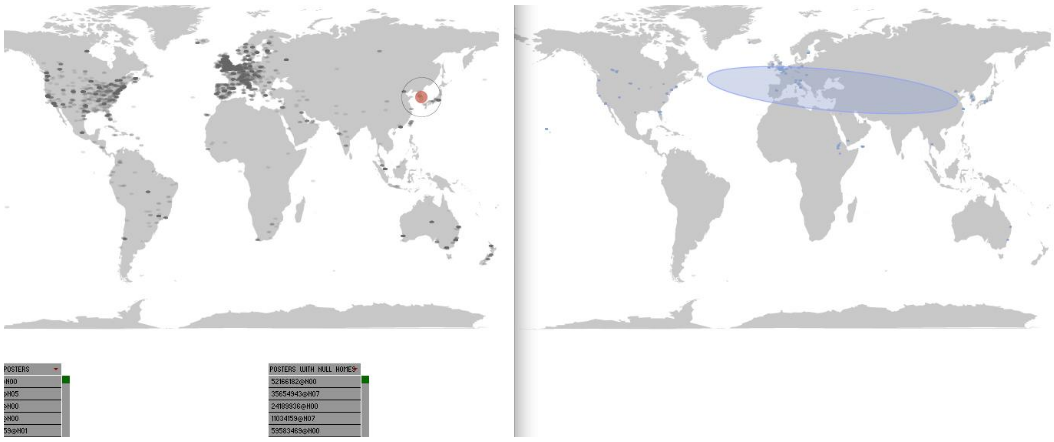


Figure 8.2 36,736 photos 99 GB posters in close proximity of Seoul (threshold 80)

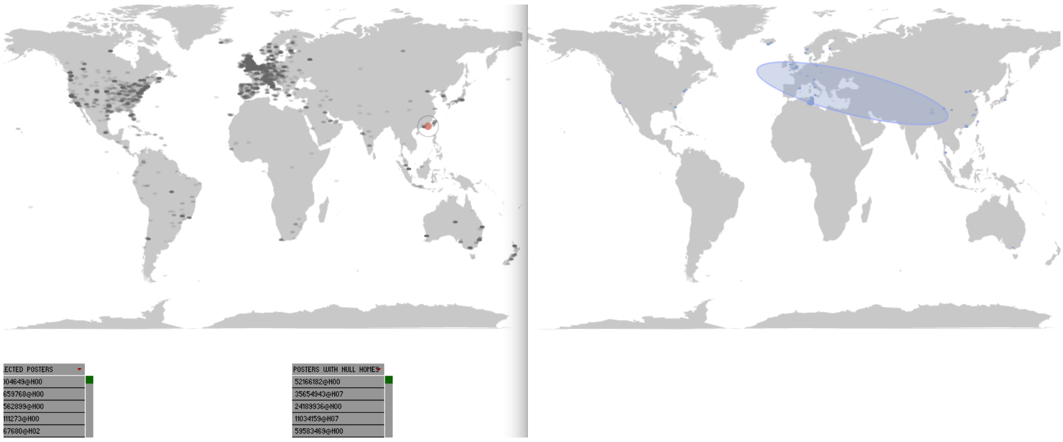


Figure 8.3 10,528 photos for 40 posters in Hong Kong expanded by 40

For the three most populated areas of Africa with GB posters, three relatively different photo distributions have been found (Figures 8.4-6). As Figure (8.4) shows, for posters of Cape Town the photos are distributed between Europe and home with a higher density around Europe. However, the 36 posters in Tombali's vicinity have no photos of their nearby places at all (Figure 8.5). Bujumbura whereabouts with the highest number of GB posters in Africa, has the distribution centered close to Europe and home with a slight tendency towards other places. Overall, there are no consistent photo distribution patterns among the African GB posters. In other words, the effect of home location on photo taking behavior of African GB posters is not immediately recognizable through the analysis of their photos' distribution.

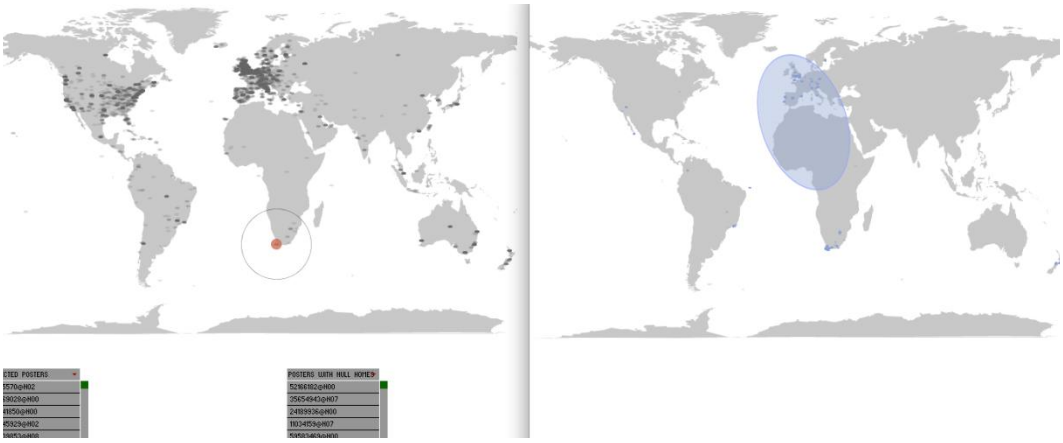


Figure 8.4 19,989 photos for 80 GB posters in Cape Town expanded by threshold 180

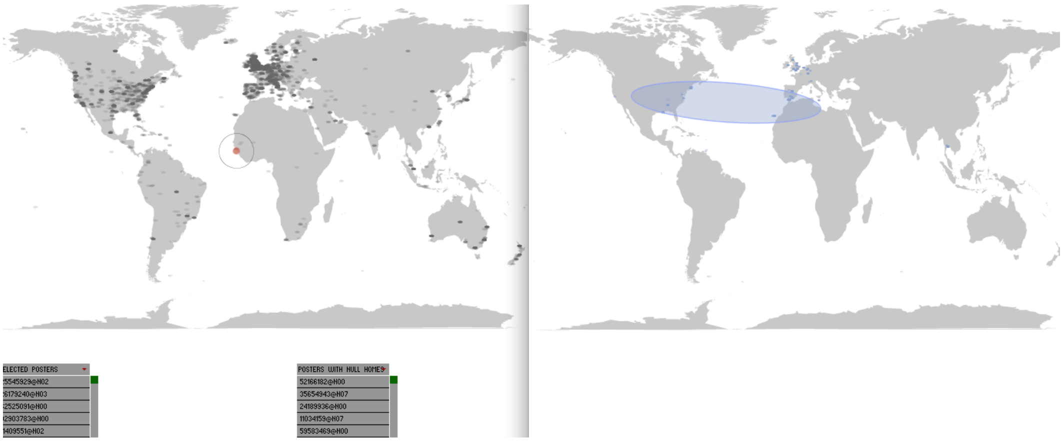


Figure 8.5 5,451 photos for 36 GB posters in Tombali expanded by threshold 70

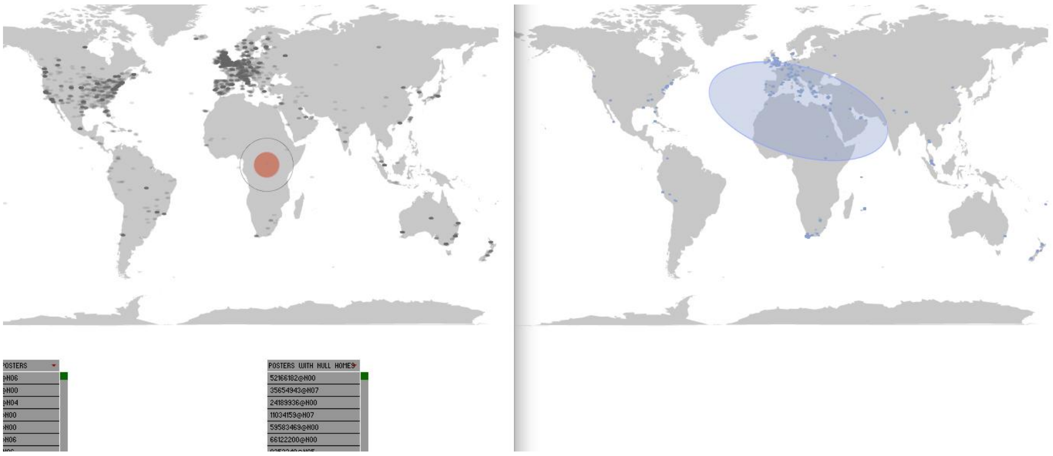
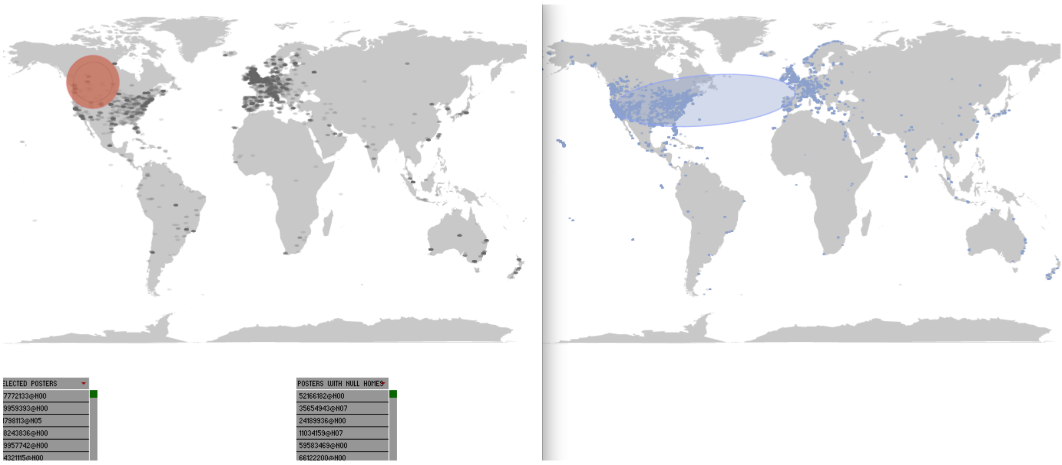


Figure 8.6 35,526 photos for 327 GB posters in Bujumbura expanded by threshold 110

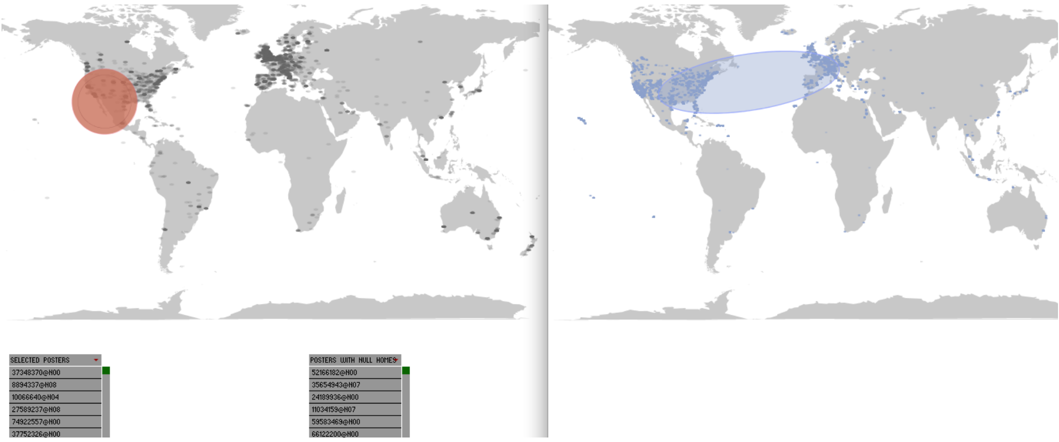
The GB posters from US and Canada follow the expected pattern of spreading photos distribution between home and Europe. The majority of photos are either in the US or Europe, with no particular popularity in either of them. In some cases the number of photos taken in nearby places of home location outweighs (Figure 8.7), while in other cases the distribution is relatively equally stretched between home location and Europe as expected (Figure 8.8).



ELECTED POSTERS
7772133@N00
9959393@N00
1798113@N05
8243836@N00
9957742@N00
4321115@N00

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00

Figure 8.7 388,406 photos for 1154 GB posters in Calgary expanded by threshold 80



SELECTED POSTERS
37348370@N00
8894337@N08
10066640@N04
27589237@N08
74922557@N00
37752326@N00

POSTERS WITH NULL HOMES
52166182@N00
35654943@N07
24189936@N00
11034159@N07
59583469@N00
66122200@N00

Figure 8.8 573,645 photos and 1,628 GB posters in Las Cruces expanded by threshold 110

Photo distributions for European GB posters are considerably focused in Europe. Examples of this pattern are shown in Figures (8.9-10). Despite the large number of photos taken from all around the world, Europe and USA have attracted the majority of the photos uploaded by European GB posters. In other words, European GB posters tend to take more photos of their nearby places than anywhere else. This might be based on the fact that since all posters have at least one GB photo in their profiles, so the nearer the home location to GB, the more likely are the photo distributions to be local.

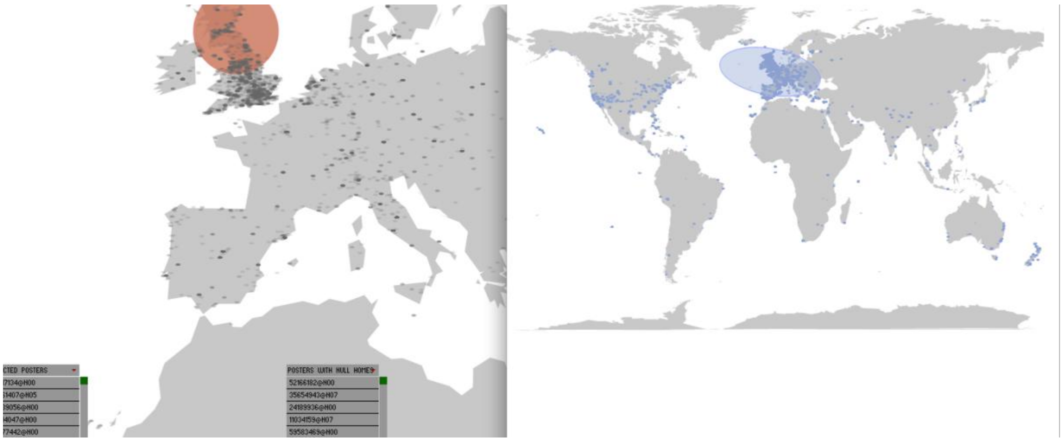


Figure 8.9 378,009 photos for 2,565 GB posters in Edinburgh (threshold 10)

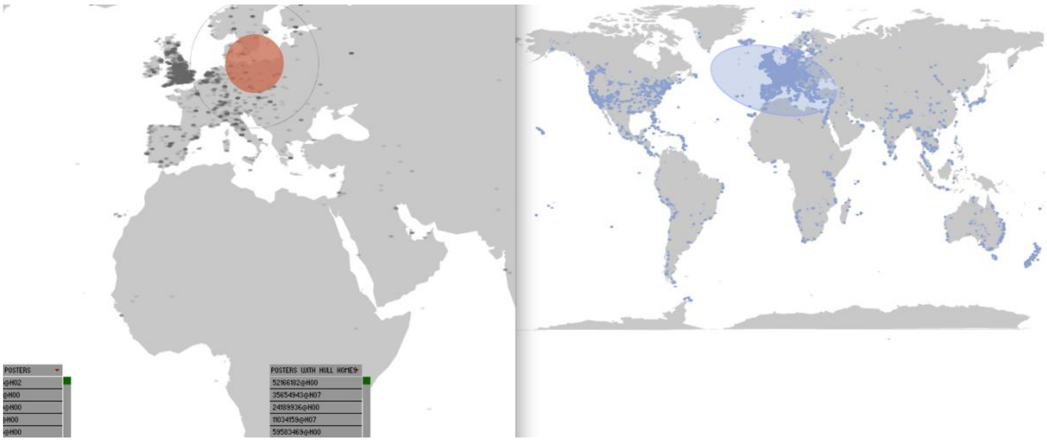


Figure 8.10 1,333,859 photos and 8,152 posters in Poland expanded by threshold 90

Overall, according to the analysis of the geographic distribution of photos in line with home location of posters, it can be concluded that:

Yes - geographic distribution of photos can contribute

to the prediction of FHLI of GB posters 'to some extent'.

The results are slightly different for different places that can be attributed to the nature of the countries, culture, language, economical situations and tourist destinations. Geographic distributions of photos in the majority of the examined places (Australia New Zealand, Asia, South America, Canada, North America and Europe) are lending towards a close proximity to where the posters live and Europe. In Africa, less evidence of taking photos of close proximity has been found. Interestingly, in the less economically challenging area of Africa (Cape Town) that is also known as a tourist destination, a relatively similar pattern, to the general pattern identified for GB posters of the rest of the world, has been found (Figure 8.4). Accordingly, one can hypothesize that:

For a null-home poster, the home location can be estimated according to the two most popular places where the majority of their photos were taken. If one of them is GB (or close proximity in Europe) the other place has significantly higher chance of being the poster's home location than other alternate places. If the whole photo collection is focused around Europe and very close proximity, then the chance of the GB poster living in Europe is higher than anywhere else. For any other distributions that do not match the above theory, the home location has the higher chance of being close to Africa.

RQ3. Can geographic distribution of friends contribute to prediction of FHLI?

As described in the data set chapter, there are two different categories of posters. In the first place the friendship network is made of GB posters who have accessible home location information. In the second place attempts were made to include null-home friends - who do not have any home location available in their profiles – into the friendship network as well. Visualizing the null-home friends on the map (chapter 5, section 5.4.2.2.2) allows the analysis and exploration of the complete GB friendship sub-network. Consequently, a decision is made and justified for analysis of null-home friends (chapter 5, section 5.4.2.2) as a part of a GB friendship sub-network. Accordingly, the developed visualization application allows the study and analysis of the friendship distribution in regard to have-home friends as well as null-home friends (chapter 5, section 5.4.2.2). Based on the identified pattern in the comprehensive friendship network of GB posters, has been concluded that the FHLI of have-home friends are relatively close to places where the null-home friends take photos (chapter 7, section 7.3). It can also be inferred that the GB friendship-sub network of GB posters (considering where have-home friends live and where null-home friends take photos) in majority of cases are centered in Europe (Figures 7.4-5). This suggests two hypotheses:

- GB posters are more likely to make friends from places where they take photos or
- The majority of sampled GB posters live in GB and nearby places (Europe). Consequently, the GB friendship sub-network has a higher chance of being comprised of posters who live in Europe than anywhere else.

Since this study examines the GB friendship sub-network, it does not cover those friends of GB posters who do not have any GB photos. Therefore, the first hypothesis cannot be investigated further at this stage. The second hypothesis implies that the majority of GB posters live in close proximity to GB. In other words, in a social network of posters with one mandatory common place in their photo collections, the effects of that location on other examined variables (friendship network, photo distributions) are quite significant. This finding is consistent with what has been concluded in RQ2, indicating that people tend to take more photos of their nearby places. Consequently, considering the fact that the studied posters have at least one GB photo would result in hypothesizing that the examined GB posters have a higher probability of being in Europe (and specifically GB) than anywhere else.

In addition, another interesting finding here is that since the photo distribution of null-home friends and the home distribution of have-home friends in the majority of cases are relatively close to each other, and referring to the aforementioned assumption in RQ2, one can hypothesize that the GB friendship sub-network of GB posters has a noticeable tendency towards Europe, regardless of the availability of home location of the friends. In other words, there is no significant different pattern in home location of have-home and null-home friends. That means availability of home location does not affect where GB friends live. In either case they live in close proximity to each other and it is not possible to derive any conclusion as to where null-home friends have more tendency to live compared to have-home friends. As the observations here indicate, the majority of the GB friends in general have a higher probability of being in Europe.

All in all, the examined sub-network of GB friends on its own does not contribute towards prediction of FHLI of GB posters directly. However, as two different variables (home locations of have-home friends and photos' distribution of null-home friends) were studied visually in answering this research question, it is possible to conclude that GB posters have more friends, who live in GB and nearby places (i.e. Europe), who take more photos of GB and nearby places (i.e. Europe) than anywhere else.

RQ4. To what extent locational information dominates a pattern in online communities?

In order to study the effects of spatial data on social interactions of social entities the relation among the FHLI, geographic distribution of photos as well as geographic distribution of friends have been examined for the sampled GB posters. As explored visually and have been discussed in chapter 7 (section 7.1) from each category of the selected populated locations with GB poster, several sample locations, proportional to the density of each area, are selected. In the visualization application for each examined variable an individual view is demonstrated in the form of transparent dots as well as an aggregated (summary) view in the form of standard ellipses. As has been identified through the visual analysis (Figures 7.6-8) it is clear that all the three examined variables in all sampled locations are closely distributed with relatively close centers and/or direction. This leads to the general conclusions that:

- GB posters have their photos' distribution spread between their home locations and Europe.
- GB posters share close photo distributions with their GB friends and
- Null-home friends take photos in close proximity to where have-home friends live.

Although this trend might not be recognized for any selected individual, the aggregated view in all cases confirms the validity of this finding to a certain extent. After all, despite the difference in the direction and size of the three ellipses in some cases, the distributions overall are not very far from each other and that indicates the role of space and location in the online behavior of GB posters. As the analysis and visualization application have demonstrated (chapter 7, section 7.4), the three examined ellipses in all cases have a tendency towards Europe. In other words, the social network of GB posters have higher chance of living in GB and nearby places (Europe), have their photo distributions stretched towards Europe and have more friends around GB and close proximity than anywhere else. This indicates the significant role of location in online behavior of social entities. Consequently, in answering the above question it is possible to argue that:

Locational information dominates a significant pattern in the social network of posters who all have common spatial information in their profiles (i.e. in this study one GB photo in their photo collection). The role of that common location can be seen in photo distributions, home location of posters, friendship network and photo taking behavior of friends.

RQ5. Are spatial social network maps useful for better understanding of the geographic distribution of individuals and their online behaviors?

In this study, a visualized map is considered to be useful if it enables the user to answer the questions at hand by interactively exploring the data. Since three of the research questions of this study, which aimed to be answered by visual analysis of the data, have been successfully answered via the interaction with the developed visualization application, one can argue that the developed spatio-social maps have achieved the desired objectives. Hence,

Although there is always room for improvement of any application, the developed spatial social network maps have been found useful for a better understanding of the geographic distribution of individuals.

The useful information gained through interaction with the spatio-social maps are reviewed here as a proof of usefulness of the visualization application in understanding the examined data. The following findings were very difficult to achieve (and in some cases quite impossible to be inferred) by analysing the data in its original format.

- GB posters tend to take more photos of their nearby places than farther locations (RQ2, chapter 7, section 7.2).
- The GB friendship sub-network of GB posters in the majority of cases is centered in Europe (RQ3, chapter 7, section 7.3).
- Availability of home location does not affect where posters live or take photos (chapter 7, section 7.4). In other words null-home posters, while being less prone to develop large networks of GB friendships and being less popular in the examined GB friendship sub-network (chapter 6: section 6.5.2), have no idiosyncratic behavior in their photo distributions (chapter 6: section 6.5.1 and Figures 7.6-8). Although they are less likely to have a large GB friendship sub-network and are less popular in friendship sub-networks (section 6.5.2), that does not affect where they live or take photos (RQ3).
- In the majority of the examined locations, distribution of FHLI of have-home friends and photos of the null-home friends are close to each other. Assuming that posters take more photos of their nearby places would lead to hypothesizing that the friends of selected posters live in close proximity to each other. In other words, have-home friends and null-home friends live in relatively close proximity to each other (chapter 7, section 7.4)
- In cases where the distributions of the three examined variables in RQ4 (photos, have-home friends' homes and null-home friends' photos) are close to each other, one can conclude that GB posters have more friends who live in close proximity to GB and are more likely to befriend with those who have photos of the same places (Figures 7.6-8).
- Photo distributions and the GB friendship sub-network are spatially related to the examined spatio-social variables. However, photo distributions more significantly depend on home location of GB posters than the GB friendship sub-network.

All in all, according to the above findings, one can hypothesize that GB posters tend to take more photos of places close to where they live, while the friendship sub-network does not reveal such an obvious spatial relation to FHLI. In other words, friendship distribution has stronger geographic footprints compared to photo distributions. Hence, in a social network of posters where they all have at least one photo of a specific location, the friendship sub-network is significantly biased towards that location.

RQ6. Is there any uncertainty involved in interpretation of spatio-social relations?

This work has faced several levels of uncertainty in different concepts. The data availability and quality have been affected by the inherent uncertainties in the nature of the spatio-social data. The identified uncertainties are summarized here in the following three categories:

Ambiguous FHLI

- Unavoidable level of uncertainty in people's everyday vocabulary i.e. near, close, not very far, downtown (chapter 2: section 2.5 and chapter 5: section 5.2.1)
- Uncertainty in human perception of location names compared to the official boundaries defined in gazetteers (chapter 5: section 5.2.1).
- In addition to the uncertainty in the nature of the vernacular and fuzzy geographic names, depending on the sensitivity and awareness of the GB posters the level of uncertainty in FHLI varies. For example, ambiguity in 'London, UK' is considerably less than 'London'. This kind of uncertainty was measured and quantified in chapter 5 (section 5.2) and chapter 6 (section 6.2.6).

Disambiguated FHLI

Although attempts were made to develop an appropriate customized algorithm for disambiguation of the examined FHLI (chapter 5: sections 5.2.4 and 5.3), the final disambiguated home locations bear some level of uncertainty.

- Depending on which category of place names the disambiguated home locations belong to (section 5.2.2, Table 5.1), they are likely to suffer from a level of uncertainty that can be attributed to the accuracy in performance of the application and the probability of errors in the disambiguation process (section 5.3.2.2, Table 5.14).
- There is also another potential uncertainty regarding the level of accuracy and closeness of the assigned default location to a real place where the examined poster spends the majority of their time. The measurement of that accuracy is not possible unless conducting a qualitative analysis (interview, observation, questionnaire etc.) and also overcoming the challenge of online and spatial privacy (chapter 2: section 2.7).

Geo-tagged Photos

In addition to the FHLI that bears certain levels of ambiguity and uncertainty, there are some concerns regarding the spatial information attached to the examined geo-tagged photos. Although the data collection application (chapter 4: section 4.3.2) was limited to the most precise geo-photos (in Flickr terms 'accuracy 16' that is equivalent to street level (chapter 4: section 4.2.1), couple of uncertainties has been identified in that respect.

- The precision of the geo-tagged photos in Flickr depends on the zoom level of the map. The user can adjust the zooming level of the map relatively proportionate to the desired precision for each photo (country, area, or in a very zoomed in scale photos can be dragged onto the appropriate street). Therefore, spatial information attached to the photos bears particular levels of uncertainty that can be attributed to the user's attention, accuracy and sensitivity in placing photos on the correct locations on the map.
- There is another kind of ambiguity regarding locational information attached to the photos. The latitude and longitude of each photo can simply reflect

the location of the camera (or photographer) rather than the actual place in the photo.

Overall, analysis of spatio-social data is involved with different levels of uncertainty in different concepts. The identified uncertainties in the spatio-social data set of this study are inherent and inevitable due to the nature of the data. Where possible, efforts have been made to reduce the effects of uncertainties to a particular level on data collection, analysis and conclusions. In doing so, the most accurate geo-photos were collected for study and analysis. In addition to that a classification and uncertainty measurements were developed and applied to FHLI. Accordingly, it has been found that the Flickr posters in majority have higher tendency in associating themselves to a place within boundary of a city (UC2-3). In other words, in the worst-case scenario the majority of findings and assumptions of this thesis in regards to ambiguous FHLI are vulnerable against different locations within the boundary of a city. Location uncertainty of a poster with the variation scope of different places with boundary of a city is a reasonable level of uncertainty in global geographical analysis. The remaining identified uncertainties in regards to FHLI and geo-tagged photos are inevitable part of analysis of spatio-social relations. This can suggest that the analysis and identification of the uncertainties involved in spatio-social data are far from straightforward. Accordingly, one can conclude that spatio-social data present significant challenges with respect to vagueness, ambiguity and accuracy.

8.3 Summary

This chapter demonstrated how the visualization application developed in chapter 5 is applied here in answering the research questions. Each question has been answered with appropriate evidence from the previous steps of the research, or by referring to the maps developed through the visual interaction with the data. The next chapter reviews the shortcomings of this research and proposes potential solutions for overcoming some of them. In addition, it reviews the aims and objectives and ensures that they have been achieved successfully.

9 Conclusion and Future Work

9.1 Introduction

This chapter reviews the aims and objectives of this study and summarises how they have been achieved. It also highlights the contributions and important findings. Limitations and shortcomings identified during the course of this research are described in the final section. Potential solutions on how to avoid or overcome the identified problems in the proposed methods and applications are discussed and potential amendments for future direction of this work have been proposed.

9.2 Revisiting Research Aims and Objectives

In this section a brief review is given on how attempts were made to achieve the aims and objectives that were set out in chapter 1 (sections 1.3-4).

1. **Identify concerns regarding locational uncertainties of online social entities and their implications.**
   Three different locational uncertainties were identified in this study (RQ6, section 8.2). During the classification, disambiguation and uncertainty measurements attempts were made to reduce the unavoidable effects of the uncertainty on the global geographical analysis to improve the robustness and reliability of the results. However, the locational uncertainties regarding the accuracy of home location of posters and geo-tagged photos were identified as an inherent complexity of spatio-social data that is unavoidable in analysis of a data set of this kind. Considering the volume and nature of the data that has been widely spread around the world, a qualitative analysis (i.e. interviews, observations, questionnaires etc.) for examining the ambiguous behavior of social entities in online environments does not sound efficient, feasible or promising. Overall, it has been concluded that the inherent and unavoidable level of uncertainty has effects on the quality of both data sets and the results of analysis. However, as discussed (section 5.3.2.3), where possible attempts were made to systematically classify and measure the uncertainties in the data set.

2. **Develop new forms of spatial social network presentations that support the visual synthesis of locational data and online relations.**
   This aim has been successfully achieved during completion of the following objectives (chapter 1: section 1.4):

   1) Develop techniques to discover and analyze spatio-social relations within a large spatially structured social group.
   2) Assess the strengths and weaknesses of the developed visualizations in identifying the role of geography in online interactions and friendship patterns.

   The first step in achieving the above aim and objectives was to select an appropriate data set (chapter 4: section 4.3) and to identify the spatio-social attributes for analysing spatio-social relations among social entities (chapter 4: sections 4.3.1 and 4.3.2). Accordingly, a Java application (appendices A-B) was developed that can send queries to the Flickr database to collect required spatio-social data during any selected time period. Several methods and techniques were developed and applied in cleaning, filtering, disambiguating and dealing with uncertainty in the collected data (chapter 5). Having clean and disambiguated data, some extra techniques were applied in order to complete several numerical analyses (chapter 6) before embarking on visual analysis of the data. In the final step, an appropriate visualization application was developed (chapter 5: section 5.4) in Processing in line with the research questions (chapter 1: section 1.5). As has been explained in the method chapter (chapter 5: section 5.4), the developed visualization application produces maps that demonstrate all the selected attributes of examined Flickr posters:

- Spatial (FHLI, photos distribution, friends home location and friends' photos distribution)
- Social (GB friendship sub-network)
- Potentially spatial (null-home posters)
- Non-spatial (number of posters in each location, number of friends, have home friends ID and null-home friends ID, number of GB photos, number of GB have-home friends and null-home friends)

The application, in addition to the spatial information, has also visualized the potentially spatial and non-spatial data. It allows the user to interact with the three variables at the same time. Therefore, the spatial social networks of this study are made in regards to social connections as well as associated spatial information (i.e. home location of have-home posters). In other words, as has been aimed originally in this study, each poster has been studied according to multi geographies associated to them. The developed disambiguation algorithm has been assessed in regards to its performance against disambiguation of 'Flickr GB Data' (section 5.3.2). Failure cases have been identified and potential solutions have been proposed (section 5.3.2.4-5).

3. **Identify the challenges of visualizing large spatio-social information in a spatially structured social group.**

Every single step followed and completed in the data collection, disambiguation and visualization sections of this study has faced challenges regarding the characteristics of spatio-social data set of this study. Visualization of large spatio-social data in this study involved the following five consecutive steps:

1. Collecting and storing a large volume of spatio-social data (chapter 4: section 4.3) from Flickr
2. Cleaning, filtering, formatting, classifying and disambiguating the spatio-social data (chapter 5: sections 5.2 and 5.3)
3. Identifying and measuring the associated uncertainty (section 5.3 and section 8.2, RQ4)
4. Design decision on visualization of spatial-social data (chapter 5: section 5.4)
5. Design decision on including potentially spatial and non-spatial data in the visualization results (chapter 5: section 5.4.2.2)
6. Reducing the risk of inaccurate perceptions of values through the appearance compensation (perceptual, apparent scaling) in analysis of the data visually (chapter 5: section 5.4.2.3)

A complete overview of the identified uncertainties in analysis of spatio-social data is provided in answering RQ6 (chapter 8, section 8.2).

9.3 Contributions

The main contributions of this work are noted and discussed in relevant chapters. However, a summary of what have been achieved during the course of this study are given as under:

1. Efficient application for collecting spatio-social data from Flickr (section 4.3)
2. Uncertainty classification model for quantifying the uncertainty in FHLI (sections 5.2-3)
3. Appropriate algorithm implemented in a Java application for disambiguation of fuzzy vernacular geographic terms that frequently happened in FHLI (section 5.3.2)
4. Visualization application that provides opportunity to augment bounding box geography and explores spatio-social relations and associated multi geographies of social entities in an interactive environment (section 5.4.2).

Here, in addition to what have been mentioned above, key achievements of this thesis are highlighted in the areas of 'home location disambiguation' and 'social entities and their spatio-social relations'.

9.3.1 Home Locations Disambiguation

The most significant contribution of this work is the placeName disambiguation in chapter 5, where a disambiguation algorithm is developed for disambiguation of fuzzy vernacular geographic terms that people use to refer to their home locations in online environments. The algorithm has been assessed against the random sample data from Flickr. Appropriate modifications have been applied in section (5.3.2) to improve the results for 'Flickr GB Data'. The disambiguation here, does not depend on the other similar documents and context and is not limited to a specific area or country. Although there are some failure cases in the disambiguation process (section 5.3.2.4-5), according to the developed classification it has been realized that the majority of the FHLI can be successfully assigned to a specific location on the earth. Overall, the author suggests that in order to gain the best possible accuracy for disambiguation of place names one can take into consideration the following modifications to the proposed method:

- Context in which the locations are mentioned
- Usage of several place name databases
- Comprehensive text analysis (i.e. capital letter, delimiters, spaces etc.)

It should also bear in mind that the above amendments should be conducted while maintaining a balance between the accuracy of the results and efficiency of the application (discussed in section 5.3.2.5).

9.3.2 Spatio-Social Relations and online Social Entities

Visual analysis of the data through interaction with the visualization application developed in chapter 5, demonstrates convincingly that locational data dominates a pattern in online environments. Developed hypotheses and their consequent discussions also confirm the effects of spatial data on online behavior of social entities (RQ2-5, section 2.8). Similar hypotheses have been put forward by other researches in the field in the past (Liben-Nowell et al. 2005; Escher, 2007), however to the best of author's knowledge, ambiguity, uncertainty and effects of spatial data in online behavior of social entities have never been studied and quantified in the scale conducted in this thesis. As a result of analysis of the spatio-social relations among a group of social entities who interact together in online environments two hypotheses have been made.

In answering the RQ2 (section 8.2) a hypothesis has been made to estimate the FHLI of null-home posters according to the geographic distribution of their photos. In addition, another hypothesis has been made in RQ3 (section 8.2). It states that posters who live in close proximity of each other share relatively close online behaviors (i.e. geographic distribution of photos and friends). These findings have interesting and useful implications on spatio-social analysis for businesses. Estimating an individual's whereabouts has considerable benefit for data aggregators, site developers and businesses that plan their marketing strategies according to the user participations (e.g. collaborative knowledge, reviews, feedback etc.). Effects of spatial data in online behavior of social entities can be exploited in customizing the product or making intelligent recommendation or strategy for targeting other users with similar spatial information. For example, if a survey concludes that the majority of consumers of a specific product are Londoners, the advertisement strategy can target the Londoners' friends who have higher chance of sharing similar online behaviors. Within the same scenario, can also mention that, for those of member of friendship network that do not have any spatial information available, those who have similar usage pattern to Londoners, have higher probability to live in London or nearby places than those members who have totally different usage pattern.

Overall, the study and analysis of spatio-social relation on the web has considerable potential for providing insight into hidden patterns in online behavior of social entities. However, as has been identified in this research, the existing tools and techniques for collecting data of this kind is far more advanced than our ability to analyse and make sense of it. Therefore, the author here predicts that study and analysis of spatio-social relations among social entities on the web will attract considerable attention in the very near future. More efficient methods for collecting large volume of data and better classification methods are expected to be developed and applied by researchers in the field.

9.4 Limitations

This section gives a brief overview of the shortcomings identified during the course of this research.

9.4.1 Flickr as a corpus

The body of the work of this thesis relies on two assumptions:

1. Firstly, Flickr users are representative sample of online social entities and
2. Secondly, Flickr is a representative corpus of spatio-social sites

On the whole, attempts were made to reduce the amount of bias in data collection and analysis (section 3.4, 4.3.1 and 4.3.2) as much as practically possible. However, there are some concerns regarding the validity and application of hypotheses derived from one sets of data to different corpora (Mika et al. 2008). For example Overell (2009) demonstrates that the two terms 'race' and 'party' have been used in different context and meaning by Wikipedians in comparison to Flickr members. Mike et al. (2008) argued that in language processing tasks models that are developed on one sets of data poorly perform on other sources of data. A rule of thumb in language processing task has been introduced as to put into account the context of the subject before choosing an appropriate model (Mike et al, 2008). However, since the disambiguation algorithms in this research only deals with location names and it has been assumed that all the FHLI are referring to a place on the earth therefore, the developed algorithms are not limited to one specific context and/or situation. Consequently, the classification methods (chapter 5, section 5.2.2) and disambiguation algorithm (Chapter 5, sections 5.2.4. and 5.3.2) developed and applied in this study are expected to perform reasonably well for place names people use to refer to their homes in other corpora.

Notwithstanding, in data collection and classification of this study, efforts have been made to select suitably diverse and general data (section 3.4-6 and section 4.3). However, the inevitable bias introduced during data collection (chapter 4, section 4.3) and inherited uncertainties in spatial data (chapter 5, section 5.3.2.3, RQ6 section 2.8) will have to be considered when applying this work in other data sets.

9.4.2 Disambiguation Algorithm

It has been identified that 35.6% of place names in English is ambiguous while in the other 12 examined languages ambiguity in place names were less than 20% (Overell, 2009). According to the developed classification scheme for this study (Chapter 5, section 5.2.2) majority of the FHLI are within city level (regardless of the ambiguities attached). It has also been recognised that 78% of the existing websites and their consequent spatio-social data is also in English (Myers and Dyke, 2000). Consequently, although the developed algorithm here can disambiguate the English place names only, the fact that 78% of the websites are in English and ambiguity in English names are the highest (Overell, 2009), one can justify the algorithm generality and its application in other situations. However, like any other applications there are always rooms for improvements. As mentioned in chapter 5 (section 5.3.2.3) the developed disambiguation algorithm failed in nine cases of location types that were summarized in table 5.16. Chapter 5 (section 5.3.2.4-5) provides potential solutions for improvement of the performance of the algorithm in the mentioned failure cases.

In general, despite recent advances in the field of word sense disambiguation (Li et al. 2002; Bilhaut et al. 2003) and geographic information retrieval (Jones et al, 2008) there are still impeding limitations. Simulation of how human brain functions in ambiguous situations in automatic systems is not fully feasible yet (Fisher, 2007). However, by the advent of Web 2.0 and proliferation of spatially aware devices the need for analysis, understanding and disambiguation of data of that kind is becoming increasingly important. As a result, the author here predicts that place name disambiguation will attract extensive usage in very near future. In the medium terms, perhaps 2 to 3 years there will be improved least supervised disambiguation algorithms that can improve many different angles of spatio-social field of research from accuracy of search results, to more precise geographically augmenting search engines, better marketing strategies for businesses and more accessible data for site owners and developers. The presented work here contributes to this realm of predicted research.

9.4.3 Limitations on Methods

The above two limitations have been found as inevitable part of research of this kind. However, the results have also suffered from some limitations as a consequence of the methods developed and applied in this thesis. Therefore, in the following sections the identified limitations are discussed and some solutions are proposed for improving the results.

9.4.3.1 Data collection

As mentioned in chapter 4 (section 4.3) there was a bug in Flickr services that resulted in some corrupted duplicate data that have been removed from the data set of this study before analysis. Being aware of the bug before data collection could have made it possible to retrieve the complete desired data set with minimum corruption. In order not to be affected by the bug, a slight modification in the algorithm could break the time periods into smaller slots until the returned results are no more than 3,750 photos. The solution, although recognised soon after the duplicates were found in the data, was not applied due to time constraint and the small percentage of the corrupted data.

9.4.3.2 Uncertainty

It is also useful to apply the uncertainty of the examined FHLI (quantified in chapter 5: section 5.3) to visualization and develop hypotheses regarding the visual analysis of spatial uncertainty of posters and their online behaviours. However, bearing in mind that the majority of the examined FHLI were categorized in City level (chapter 6: sections 6.5.5 and 6.2.6) the existing hypotheses and conclusions in the worst case might be vulnerable to ambiguity within the boundary of a city (or different locations within a country UC3 > UC4). Accordingly, one can argue that the findings and conclusions of global geographical analysis in this study are robust and valid within a reasonable level of certainty.

9.5 Important Finding

Overall, this thesis concludes that even in places where geography does not seem important directly (e.g. photographic archive) location plays a dominant role. This is known as first law of geography:

'Everything is related to everything else but closer things are more closely related[53]'. Accordingly, one can conclude that first law of geography is valid in spatio-social relations of social entities who interact together in online environments. This research proved that spatial information has significant pattern in online behavior of social entities in regard to their photo distributions and their friends. This finding has important implication on every aspect of human life from business, tourism, proximity sensing and customized products to field of image and information retrieval. The

[53] Waldo Tobler's first law of geography, 1970

hypotheses made in regards to relation between home location and other spatio-social attributes of members (section 8.2) can play vital information for personalization and customization of products and services.

9.6 Further Research

Like any other application there is always room for further improvement and adding new functionalities for better understanding of the data. Evaluations of the proposed application in visualization and analysis of spatio-social data have been conducted regarding the performance of the application in answering the research questions (section 8.2). However, the developed application can be improved in the following three areas:

- Incorporating the ability of comparing have-home posters with null-home posters in the visualization application can considerably improve the ability of the application to analyze the data and develop hypotheses regarding the have-home and null-home posters' online behaviors.
- Another improvement can be done by visualizing the estimated home location of null-home poster (according to the RQ2 hypothesis) on the map. Doing that could have facilitated the understanding of the overall view of the patterns in posters' behaviors.
- Incorporating the uncertainty classifications and measurements into the visualization application can improve the capability and usefulness of the application in revealing the patterns and extraction of relevant information from large volumes of data.

Apart from putting all the developed methods into the visualization application functionalities, it is also interesting to apply the developed methods and techniques in this thesis to other sets of spatio-social data. By doing so, more comprehensive evaluation of the methods can be conducted. In addition, the findings and hypotheses of this study can be tested against different data set and any distinctive behavior of Flickr users can be eliminated from generalizations of findings.

9.7 Summary

This chapter has provided a brief overview on how the aims and objectives (chapter 1: sections 1.3-4) of this study have been achieved successfully. It also highlighted the important findings of this research and the contributions that it has made. The last section proposed some potential places that the application can be improved and that could lead the future direction of this research.

10 Glossary

In these Glossary acronyms, abbreviations and notations referred to in this thesis are defined.

Geo-Social Network: a type of social network that provides geographic services (i.e. geo-tagging, geo-coding).
Online Social entity: a person who socializes with others via Internet
SNA: Social Network Analysis
FHLI: Flickr Home Location Information
E-Social Science: is about technological developments and approaches in social science field. It helps the social scientists to conduct new research or do their existing research quicker.
Fuzzy: is attribute of any variable that is approximate rather than fixed and exact.
Place Names: Refers to a location on the earth
Threshold Circle: shows the selected area by the user on the map
Vernacular Geographic Terms: are place names that people describe in their ordinary every day language.
RQ: Research Questions
REQ: Requirements
Geo-photo: photos that have attached spatial information (Lat, Lon)
Web 2.0: allows users to participate and share information on the web
UGC: User Generated Content
VGI: Volunteered Geographic Information
Poster: anyone who uploads photos on Flickr
NLP: Natural Language Processing
Disambiguation: is the act of identifying the intended use and/or meaning of a word that has or can be associated to several meanings or locations.
Spatio-Social: spatial (locational) and social (relational)
Spatial: any variable that carries locational data
Social: any variable that relates individuals together (i.e. friendship)
IR: Information Retrieval
GIS: Geographic Information Systems
UNC: uncertainty

11 Bibliography: Cited References

Abbasi, R., Chernov, S., Nejdl, W., Paiu, R., & Staab, S., 2009, Exploiting Flickr Tags and Groups for Finding Landmark Photos. Available at: http://www.l3s.de/~chernov/SergeyChernovECIR2009.pdf.

AbdelMalik, P., Kamel Boulos, M.N., & Jones, R, 2008, The perceived impact of location privacy: A web-based survey of public health perspectives and requirements in the UK and Canada, Bio Med Central (BMC) Public Health, 8(156). Available at: http://www.biomedcentral.com/content/pdf/1471-2458-8-156.pdf.

Abril, P.S., 2007, A (my) space of one's own: On privacy and online social networks, Northwestern journal of technology and intellectual property, 6(1). Available at: https://www.law.northwestern.edu/journals/njtip/v6/n1/4/.

Adamic, L.A. & Huberman, B.A., 2002. Zipf's Law and the Internet. Glottometrics, 3:143-150.

Adams, A., 2000. Multimedia Information Changes the Whole Privacy Ballgame. Computers, Freedom and Privacy. ACM Press: 25-32.

Adamic, L.A., Buyukkokten, O., & Adar, E., 2007, A social network caught in the web, First Monday, 8(6). Available at: http://www.firstmonday.org/issues/issue8_6/adamic/index.html.

Adamic, L. & Adar, E., How to Search a Social Network? Available at: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VD1-4FPDR53-1&_user=910131&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000047819&_version=1&_urlVersion=0&_userid=910131&md5=22c84564dfd9d040e53a7ef5a598b783.

Ahern, S., Naaman, M., Nair, R., & Yang, J., 2007. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections, in Proceedings of Joint Conference on Digital Libraries (JCDL), Vancouver, Canada. Available at: http://delivery.acm.org/10.1145/1260000/1255177/p1ahern.pdf?key1=1255177&key2=6197595521&coll=GUIDE&dl=GUIDE&CFID=57203968&CFTOKEN=31411227.

Albert, R., Jeong, H., & Barabasi, A.L., 1999, Internet: Diameter of the World Wide Web, Nature 401:130-131.

Almeida, R.B., Mozafari, B. & Cho, J., on the Evolution of Wikipedia. Available at: http://www.cs.ucla.edu/~barzan/papers/icwsm_2007.pdf.

Ames, M., & Naaman, M., 2007, Why we tag: motivations for annotation in mobile and online media. CHI'07: Proceedings of the SIGCHI conference on human factors in computing systems, ACM Press: 971-980.

Amitay, E. et al., 2004, Web-a-Where: Geotagging Web Content, 27th annual international ACM SIGIR conference, Sheffield, UK: 273-280.

Andrienko, G., Andrienko, N., Jankowski ,P., Keim, D., Kraak, M.J., MacEachren, A., & Wrobe, S., 200, Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science, 21(8).

Andris, C., Using Prefuse Software Toolbox to Visualize Research Interests and Specializations of Faculty, Staff and Graduate Students at the University of South Carolina Department of Geography.

Anthony, D., & Williamson, T., 2005, Explaining Quality in Internet Collective Goods: Zealots and Good Samaritans in the Case of Wikipedia, Available at: http://web.mit.edu/iandeseminar/Papers/Fall2005/anthony.pdf [Accessed January 28, 2008].

Arampatzis, A., Kreveld, M.V., Jones, C.B., Vaid, S., Clough, P., Joho, H., Sanderson, M., Benkertt, M., & Wolf, A, 2006, Web-based delineation of imprecise regions, Computers, Environment and Urban Systems (CEUS), 30(4): 436-459.

Arguello,, B.B.J., Elisabeth Joyce, Kimberly S. L.,Kraut, R., & Wang, X., 2006, TaLk to Me: Foundation of Sucessful Indivudual Group Interactions in Online Communities. CHI 06, ACM conference on human factors in online communities. Available at: http://www.cs.cmu.edu/~kraut/RKraut.site.files/articles/arguello06-foundations%20Success%20_final.pdf [Accessed January 29, 2008].

Armstong, M.P. & Ruggles, A.J., 2005, Geographic information technologies and personal privacy, Cartographica, 40(4): 63-75.

Banaszak-Holl, J., Rundall, T. & Wholey, D., 2004. social network analysis in health services research: theory, methods and examples. Academy Health.

Barabasi, A.L., & Albert, R., 1999, Emergence of Scaling in Random Networks, science, 286(509). Available at: http://arxiv.org/PS_cache/cond-mat/pdf/9910/9910332v1.pdf.

Barabasi, A. L., & Bonabeau, E., 2003, Scale-Free Networks, scientific American, 288(5): 60-69.

Battista, G.D., Tollis, I.A., Eades, P., & Tamassia, R., 1998, Graph Drawing: Algorithms for the Visualisation of Graphs, PTR Upper Saddle River, NJ, USA: Prentice Hall.

Bausch, P., & Bumgardner, J., 2006, Flickr Hacks: Tips & Tools for Sharing Photos Online, Published by O'Reilly. Available at: http://books.google.co.uk/books?id=WLx95OWUMYkC.

Beck, U., 2002, The Cosmpolitan Society and its Enemies, Theory, Culture and Society, 19(1-2):17-44.

Beigl, M., Zimmer, T., & Decker, C., 2002, A location Modle for Communication and Processing Context, Personal and Ubiquitous Computing, 6(5-6): 341-357.

Bellomi, F., & Bonato, R., 2005, Network Analysis for Wikipedia, Proceedings of Wikimania 2005, Frankfurt, Germany. Available at: http://www.fran.it/articles/wikimania_bellomi_bonato.pdf.

Bender-deMoll, S., & McFarland, D.A., 2006, The art and science of dynamic network visualization, Journal of Social Structure, 7, Available at: http://www.stanford.edu/group/sonia/papers/DNV_JOSS.pdf.

Bergs, A., 2006, Analyzing online communication from a social network point of view: questions, problems, perspectives, language at internet, 3(371). Available at: http://www.languageatinternet.de/articles/2006/371.

Bernard J., 2009, Understanding User-Web Interactions via Web Analytics, Jansen Synthesis Lectures on Information Concepts, Retrieval, and Services, 2009, Vol. 1, No. 1 , Pages 1-102

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., & Weitzner, D.J., 2006, Web Science: an Interdisciplinary Approach to Understanding the Web, Science 313, 769.

Berry, M.J., & Linoff, G., 1997, Data Mining Techniques: for Marketing, Sales and Customer Support, New York, USA: John Wiley & Sons.

Bilhaut, F., Charnois, P., Enjalbert, P., & Mathet, Y., 2003, Geographic reference analysis for geographic references, Edmonton, Alberta, Canada, NAACL_HLT (North American Chapter of the Association for Computational Linguistics - Human Language Technologies).

Borgatta, E.F. & Montgomery, R.J.V., 2000, Encyclopedia Of Sociology - Volume I, Macmillan Reference USA, 2nd edition.

Borgatti, S., 2002, Basic social network concepts. Available at: http://www.analytictech.com/networks/basic%20concepts%202002.pdf.

Borgatti, S., & Foster, P.C., 2003, The Network Paradigm in Organizational Research: A Review and Typology, Journal of Management, 29(6): 991-1013.

Borgman, C.L., 2003, From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World MIT Press., United States of America.

Borner, K., 2007a. Mapping Science. Available at: http://colab.cim3.net/file/work/Expedition_Workshop/2007-08-14_MappingPublicInformation_ConnectingToScienceAndScholarlyKnowledge/KatyBorner--Mapping-Science_2007_08_14.pdf.

Borner, K., 2007b, Wikipedia Activity Visualization. Available at: http://abeautifulwww.com/NewWikipediaActivityVisualizations_AB91/07WikipediaPS3150DPI.png.

Boyd, D., 2007, Friends, Friendsters and top 8: Writing community into being on social network sites, First Monday, 11(12). Available at: http://www.firstmonday.org/issues/issue11_12/boyd/.

Boyd, D. M., 2007, Social network sites: definition, history and scholarship, Journal of Computer-Mediated Communication, 13(1). Available at: http://www.danah.org/papers/JCMCIntro.pdf.

Brandes, U., & Lerner, J., 2007, Visual Analysis of Controversy in User-generated Encyclopedias, IEEE Symposium on Visual Analytics Science and Technology, Available at: http://www.inf.uni-konstanz.de/algo/publications/bl-vacuge-07.pdf [Accessed January 15, 2008].

Brandes, U., 2003, Visual unrolling of Network Evolution and the analysis of Dynamic Discourse, Information Visualization 2: 40–50, Available at: http://www.inf.uni-konstanz.de/algo/publications/bc-vunea-03.pdf.

Buriol, S.C., Donato, D. & Millozzi, S., 2006, Temporal Analysis of Wikigraph, Web Intelligence Conference, IEEE CS Press. Available at: http://www.dcc.uchile.cl/%7Eccastill/papers/buriol_2006_temporal_analysis_wikigraph.pdf [Accessed January 28, 2008].

Brunsdon, C., & Corcoran, J., 2006, Using Circular statistics to analyse time patterns in crime incidence. Computers, Environment and Urban Systems, 30: 300-309.

Brunsdon, C., & Corcoran, J., 2005, Using Circular Statistics to analyse time patterns in crime incidence, Computers, Environmnets and Urban Systems, pp: 300-319. www. Elsevier.com/locate/compenvurbsys

Brusilovsky, P., 1996, Methods and Techniques of Adaptive Hypermedia, User Modeling and User-Adapted Interaction, 6(2-3):87-129.

Bryant, S.L., Forte, A., & Bruckman, A., 2005, Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. GROUP International Conference on Supporting Group Works, Sanible Island, FL, 1-10.

Cairncross, F., 1997, The Death of Distance, London: Orion Business Books.

Caldwell, D.R., 2008, Unlocking the Mysteries of the Bounding Box, MAGERT (Ala Map and Geography Round Table), Available at: http://purl.oclc.org/coordinates/a2.pdf.

Capocci, A., Servedio, V.D., Buriol, L.S., Donato, D., Leonardi, S., & Caldarelli, G., 2006, Preferential attachment in the growth of social networks: the case of Wikipedia. Physical Review Series E, PubMed, 74(3).

Caroline H., 2005, Social Network and Internet Connectivity Effects, Information, Communication & Society, 8(2):125-147.

Casella, G., & Berger, R.L., 2001, Statistical Inference 2nd edition.

Caslon, 2004. Caslon Analytics privacy guide. Available at: http://www.caslon.com.au/privacyguide20.htm.

Cattuto, C., Loreto, V., & Servedio, V.D.P, 2006, A Yule-Simon Process with Memory, Europhys, Lett., 76 (208).

Cattuto, C., Loreto, V., & Pietronero, L., 2005, Semiotic Dynamics and Collaborative Tagging, In Proceedings of National Academy of Sciences, 104: 1461-1464.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. 2007, I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System, In proceedings of the 5th ACM Internet Measurement Conference (IMC), San Diego, CA, Available at: http://www.imconf.net/imc-2007/papers/imc131.pdf.

Chen, H., Smith, T.R., Larsgaard, M.L., Hill, L.L. & Ramsey, M., 2008, A geographic knowledge representation system for multimedia geospatial retrieval and analysis. Available at: http://www.springerlink.com/content/qkh8hyecby6glww5/ [Accessed December 2, 2008].

Chen, H., Smith, T.R., Larsgaard, M.L., Hill, L.L., Ramsey, M., 1997a, A geographic knowledge representation system for multimedia geospatial retrieval and analysis, International Journal on Digital Libraries, 1(2):132-152.

Chen, H., Smith, T.R., Larsgaard, M.L., Hill, L.L., Ramsey, M., 1997b. SpringerLink - Journal Article. Available at: http://www.springerlink.com/content/qkh8hyecby6glww5/ [Accessed November 26, 2008].

Chen, J.,MacEachren, A.M., & Guo, D., 2008, Supporting the Process of Exploring and Interpreting Space-Time Multivariate Patterns: The Visual Inquiry Toolkit, Cartography and Geographic Information Science, 35(1): 33.

Chesney, T., 2006, An empirical examination of Wikipedia's credibility, First Monday, 11(11).

Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C., 2000, Developing a Context-Aware Electronic Tour Guide: some Issues and Experiences. Conference on Human Factors in Computing Systems (CHI), The Hague, The Netherlands, Association for Computer Machinery (ACM) Press, 17-24.

Cohen, J.D., 1997, Drawing Graphs to Convey Proximity: An Incremental Arrangement Method, ACM Transactions on Computer-Human Interaction, 4(3): 197-229.

Comber A., Fisher, P., & Wadsworth, R., 2005, What is land cover? Environment and Planning B: Planning and Design, 32(1):199-209.

Conniss, L.R., Ashford, A.J. & Graham, M.E., 2000, Information Seeking Behavior in Image Retrieval: Visor I Final Report. Library and Information Commission Research Report 95. Institute for Image Data Research, Newcastle upon Tyne.

Consolvo, S., Smith, I.E., Mathews, T., LaMarca, A., Tabert, J., & Powledge, P., 2005, Location disclosure to social relations: why, when and what people want to share. Proceedings of ACM Conference on Human Factors and Computing Systems: CHI, 81-90, Available at: http://seattleweb.intel-research.net/people/lamarca/pubs/chi05-locDisSocRel-proceedings.pdf.

Cooper, M., Foote, J., Girgensohn, A., & Wilcox, L., 2003, Temporal Event Clustering for Digital Photo Collections, In Proceedings of 11th ACM international conference on Multimedia, ACM Press, 364-373.

Crandall, Backstorm, L., Huttenlocher, D., & Kleinberg, J., 2009, Mapping the World's Photos, Proceedings of the 18th International Conference on World Wide Web, ACM Press, 761-770.

Craswell, N. & Hawking, D., 2009, Web Information Retrieval, in Information Retrieval: Searching in the 21st Century, edited by Göker, A., & Davies, J., John Wiley & Sons, Ltd, Chichester.

Crucitti, P., Latora, V., & Porta, S., 2006, Centrality measures in spatial networks of urban streets. Phys. Rev. E, 73(3): 036125-036130.

Cui, W., Zhou, H., Qu, H., Wong, P. C., & Li, X., 2008, Geometry based clustering for Graph visualization, IEEE Transaction on Visualization and Computer Graphics, 14(6):12771284.

Cummings, J.N., Lee, J.B. & Kraut, R., 2006, Communication Technology and Friendship during the Transition from High School to College, Edited by Kraut, R., Brynin, N., & Kiesler, S., Computers, phones and the internet: Domesticating information technology, Oxford University Press, 809-851.

Davies, C., Holt, I., Green, J., Harding, J., & Diamond, L., 2009, User Needs and Implications for Modeling Vague Named Places, Spatial Cognition and Computation, 9(3):174-194.

Davis, M., King, S., Good, N., & Sarvas, R., 2004, From context to content: leveraging context to infer media metadata, In Proceedings of 12th International Conference on Multimedia, ACM Press, 188-195.

Davis, M., Smith, M., Stentiford, F., Bamidele, A., Canny, J., Good, N., King, S., & Janakiraman, R., 2006, Using context and similarity for face and location identification, In Proceedings of the IS&T/SPIE 18th Annual Symposium on Electronic Imaging Science and Technology.

Daz, L., Granell, C., Gould, M., & Huerta, L., 2011, Managing UGC in geospatial cyber-infrastructures, Future Generation Computer Systems, 27(3): 304 314.

Dent, B. Torguson, J., & Thomas, H., 2008, Cartography: Thematic Map Design, 6th Edition, ISBN 9780072943825.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Wenzhong, S., 2010, Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. Transactions in Geographic Information Systems, 14(4): 387-400.

Dijck, J., 2008, Users like you? Theorizing agency in user-generated content, Media, Culture & Society, 31(1).

Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D., & Tomkins, A., 2002, Self-Similarity in the Web, ACM Transactions on Internet Technology 2, 205-223.

Dilo, A., De By, R.A. & Stein, A., 2007, A System of types and operators for handling vague spatial objects, International Journal of Geographical Information Science, 21(4): 397-426.

Dorogovtsev, S.N., Mendes, J.F.F., & Samukhin, A.N., 2001, Anomalous Percolation Properties of Growing Networks, Physics Review, 63.

Dorogovtsev, S & Mendes, J.F.F, 2003, Evolution of Net- works: From Biological nets to the Internet and WWW, Oxford University Press.

Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., & Tomkins, A., 2006, Visualizing Tags Over Time, In Proceedings of World Wide Web, 193-202.

Durand, N. & Lancieri, L., 2002, Study of the Regularity of the Users' Internet Accesses, in Proceedings of the 3rd International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 02, 2412: 173-178, Manchester, UK,

Dykes, J., 1998, Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/TK and cdv, the statistician, 47(3):485-497.

Eades, P. & Feng, Q.W., 1996, Multilevel visualization of clustered graphs, Proceedings of the Symposium on Graph Drawing, Springer-Verlag,101-112.

Egenhofer, M.J. & Mark, D.M., 1995, Naive Geography, Proceedings of Conference on Spatial Information Theory (COSIT), Springer-Verlag,1-15.

Elizabeth Goodman, A.M., 2006, Community in Mashups: The Case of Personal Geodata. Available at: http://mashworks.net/images/5/59/Goodman_Moed_2006.pdf.

Ellison, N. B., Steinfield, C., & Lampe, C, 2006, Spatially bounded online social networks and social capital: the role of facebook, Annual Conference of the International Communication Association (ICA), Montreal, Available at: https://www.msu.edu/~nellison/Facebook_ICA_2006.pdf.

Ellison, N. B., Steinfield, C., & Lampe, C, 2007, The benefits of Facebook friends: Social capital and college students' use of online social network sites, Journal of Computer-Mediated Communication, 12(4). Available at: http://jcmc.indiana.edu/vol12/issue4/ellison.html.

Emigh, W., & Herring, S.C., 2005, Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias, International Conference on System Sciences, Available at: http://csdl2.computer.org/comp/proceedings/hicss/2005/2268/04/22680099a.pdf.

Escher, T., 2007., Researching the Geography of Social Relations, Analysis of the spatial distribution of friendship networks on social network sites, third international conference on e-social science.

Elwood, S., 2009, Geographic information science: new geovisualization technologies–emerging questions and linkages with GIScience research, Progress in Human Geography, 33:2, 256-263.

Faloutsos, M., Faloutsos, P. & Faloutsos, C., 1999, On Power-Law Relationships of the Internet Topology, ACM Special Interest group on Data Communication (SIGCOMM) 251-262.

Fabrikant, S.I., 2001, Visualizing Region and Scale in Information Spaces, The 20th International Cartographic Conference, ICC, 2522-2529.

Fekete, J.D. & Plaisant, C., 1999, Excentric Labeling: Dynamic Neighborhood Labeling for data Visualization. Human factos in Computer Systems, ACM, New York, pp.512-519.

Fellbaum, C., 1998, WordNet: An Electronic Lexical Database and some of its Applications, Massachusetts Institute of Technology (MIT) press, Cambridge, MA.

Forte, A.B., 2005, Why do people write for Wikipedia? GROUP, Available at: http://jellis.org/work/group2005/papers/forteBruckmanIncentivesGroup.pdf.

Fisher, D., 2007, Hotmap: Looking at Geographic Attention, IEEE transaction and computer graphics, TVCG, 13(6). Available at: http://research.microsoft.com/pubs/69446/fisher_infovis_hotmap.pdf.

Fisher, P., 2007, Sorties paradox and vague geographies, Fuzzy Sets and Systems, 113(1), 7-18.

Fisher, N.I., 1993, Statistical analysis of circular data, New York: Cambridge University Press.

Fisher, P., 1999, Models of uncertainty in spatial data, in Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., & Fisher, P., Geographical Information Systems, Principles and Technical Issues, Chichester: John Wiley and Sons, 191-205.

Flanagin, A.J. & Metzger, M.J., 2008, The credibility of volunteered geographic information, GeoJournal, 72:137-148.

Freeman, L.C., 1988, Computer programs and social network analysis, Connections, 11(2): 26-32.

Freeman, L.C., 2000, Visualizing social network, Journal of Social Structure, 1(1). Available at: http://www.cmu.edu/joss/content/articles/volume1/Freeman.html.

Fruchterman, T. & Reingold, E.M., 1991, Graph Drawing by Force directed Placement, Software- Practice and Experience, 21(11):1129-1164.

Fu, G., Jones, C.B. & Abdelmoty, A.I., 2005, Building a Geographical Ontology for Intelligent Spatial Search on the Web, in Proceedings of IASTED International Conference on Databases and Applications,167-172.

Gale, W.A., Church, K.W. & Yarowsky, D., 1992, One Sense Per Discourse, Proceedings of the 4th DARPA Speach and Natural Language Workshop, 233-237.

Gargi, U., 2003, Consumer media capture: Time-based analysis and event clustering, Technical Report HPL,165, HP Laboratories.

Giles. J., 2005, Internet encyclopedias go head to head, Nature , 438:900–901.

Girardin, F., & Blat, J., 2007, Place this Photoon a Map: a study of Explicit Disclosure of Location Information, Late Breaking Result at Ubicomp. Available at: http://www.girardin.org/fabien/publications/girardin_ubicomp2007_lbr.pdf.

Girardin, F., Blat, J., et al., 2008, Digital Footprinting: Uncovering Tourists with User-Generated Content, IEEE Pervasive Computing, 7(4): 36-43.

Girardin, F., Calabrese, F., 2008, Digital Footprinting: Uncovering Tourists with User-Generated Content, Pervasive Computing IEEE, 7(4):36-43.

Girardin, F., Fiore, F.D., Ratti, C., & Blat, J., 2008, Leveraging Explicitly Disclosed Location Information to Understand Tourist Dynamics: A Case Study, Journal of Location Based Services, 2(1):41-56.

Girardin, F., Fiore, F.D., Blat, J., & Ratti, C., 2007, Understanding of Tourist Dynamics from Explicitly Disclosed Location Information, 4th International Symposium on LBS and Telecartography, Available at: http://www.girardin.org/fabien/publications/girardin_dalfiore_blat_ratti_lbs2007_final.pdf.

Golder, S.A., & Huberman, B., 2006, Usage patterns of collaborative tagging systems, Journal of Information Science, 32(2):198-208.

Goodchild, M.F., 2007a, Citizens as sensors: the world of volunteered geography, GeoJournal, 69: 211-221.

Goodchild, M.F., 2007b, Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0, International Journal of Spatial Data Infrastructures Research, 2: 24-32.

Goodchild, M.F., 1999, The Future of the Gazetteer. Presented at the Digital Gazetteer Information Exchange Workshop. Available at: http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/Perspectives.

Goker, A., Myrhaug, H., & Bierig, R., 2009, Context and Information Retrieval, in Information Retrieval: Searching in the 21st Century, edited by Goker, A., & Davies, J., John Wiley & Sons, Ltd.

Gow, G.A., 2004, Pinpointing Consent: Location Privacy, Public Safety, and Mobile Phones, The Global and the Local in Mobile Communication' conference, Institute for Philosophical Research of the Hungarian Academy of Sciences; Budapest. Available at: http://www.fil.hu/mobil/2004/Gow_webversion.pdf.

Grafarend, E.W., & Krumm, F.W., 2010, Map Projections, Cartographic Information Systems, Springer, ISBN 978-3-642-07178-2.

Graham, A. et al., 2002, Time as Essence for Photo Browsing Through Personal Digital Libraries, in Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries. Available at: http://ilpubs.stanford.edu:8090/740/1/2002-34.pdf.

Grothe, C. & Schaab, J., 2009, Automated Footprint Generation from Geotags with Kernel Density Estimation and Support Vector Machines, Spatial Cognition & Computation, 9: 195-211.

Gunther, R., Levitin, L., Shapiro,B., & Wagner, P., 1996, Zips's law and effect of ranking on probability distributions, International Journal of Physics, 35(2): 395-417.

Guo, D. et al., 2006, A Visualization System for Space-Time and Multivariate Patterns, IEEE transactions on visualization and computer graphics, 12(6):1461- 1474.

Gupta, M., Li, R., Yin, Z., & Han, J., 2010, Survey on social tagging techniques, SIGKDD Explor. News1, 12: 58-72, November.

Gutmann, M. & Stern, P.C., 2007, Putting people on the map: Protecting confidentiality with linked social spatial data, US: The national academies press, editors committee on the human dimensions of global change (HDGC). Available at: http://books.nap.edu/openbook.php?record_id=11865&page=1.

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A., 2010, How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information, The Cartographic Journal, 47(4): 315-322.

Haklay, M., & Weber, P., 2008, OpenStreetMap – User Generated Street Map, IEEE Pervasive Computing, October-December, 12-18.

Hanneman, R A & Riddle, M., 2005, Introduction to social network methods, Riverside, CA: University of California. Available at: http://faculty.ucr.edu/~hanneman/.

Heer, J., & Boyd, D., 2005, vizster: visualizaing online social networks, Infovis, IEEE Symposium on Information Visualization, Available at: http://www.danah.org/papers/InfoViz2005.pdf.

Heer, J., 2004, Prefuse a Software Framework for Interactive Information Visualization, Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of Master of Science, Plan II. Available at: http://jheer.org/publications/2004-Heer-prefuse-MastersApp.pdf.

Henriksson, R., Kauppinen, T. & Hyvönen, E., 2008, Core Geographical Concepts: Case Finnish Geo-ontology, Proceedings of the 1st International Workshop on Location and the Web, 17th International World Wide Web Conference, 57-60.

Henry, N. & Fekete, J.,D., 2007, MatLink: Enhanced Matrix Visualization for Analyzing Social Networks, Proceedings of the 11th IFIP TC13 International Conference on Human Computer Interactions, 4663: 288-302.

Henry, N. & Fekete, J.D., 2006, MatrixExplorer: a Dual-Representation System to Explore Social Networks, IEEE Transactions on Visualization and Computer Graphics, 12(5):677-684.

Henry, N., Bezerianos, A.,& Fekete, J.D., 2008, Improving the Readability of Clustered Social Networks using Node duplication, IEEE transactions on visualization and computer graphics, 14(6):1317-1324.

Henry, N., Fekete, J.D., & McGuffin, M.J., 2007, NodeTrix: A Hybrid Visualization of Social Networks, in Proceedings of IEEE InfoVis, 13(6):1302-1309.

Herman, I., Melançon, G., & Marshall, M.S., 2000, Graph Visualization and Navigation in Information Visualization: A Survey, IEEE Transactions on Visualization and Computer Graphics, 6(1):24-43.

Hewitt, A., & Forte, A., 2006, Crossing boundaries: identity management and student/faculty relationships on the facebook. Available at: http://www.cc.gt.atl.ga.us/grads/f/Andrea.Forte/HewittForteCSCWPoster2006.pdf.

Hightower, J., & Borriello, G., 2001, A Survey and Taxonomy of Location Systems for Ubiquitous Computing, IEEE Computer, 34(8):57-66.

Hill, L.L., 2000a, Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints, in proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries. Available at: http://citeseer.ist.psu.edu/cache/papers/cs/23366/http:zSzzSzwww.alexandria.ucsb.eduzSzzCz7ElhillzSzpaper_draftszSzECDL2000_paperdraft7.pdf/hill00core.pdf.

Hill, L.L., 2000b.,Core Elements of Digital Gazetteers: Placenames,Categories, and Footprints. In Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, Edited by Borbinha, J.L., & Baker, T., Lecture Notes in Computer Science, 1923: 280-290.

Hill, W.C., Hollan, J.D., Wroblewski, D., & McCandless, T., 1992, Edit wear and read wear, Proceedings of CHI'92, the SIGCHI Conference on Human Factors in Computing Systems, Monterey, CA, May 3-7, 3-9.

Hill, L.L., Frew, J., & Zheng, Q., 1999, Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library, D-Lib Magazine, 5(1). Available at: http://www.dlib.org/dlib/january99/hill/01hill.html.

Hirst, G., 1987, Sematic Interpretation and the Resolution of Ambiguity, Cambridge University Press, Cambridge.

Hollenstein, L., & Purves, R.S., 2009, Exploring place through user-generated content: using Flickr to describe city cores, Journal of Spatial Information Science (JOSIS), 14. Available at: http://www.josis.org/index.php/josis/article/view/13.

Holloway, T., Boievi, M. & Börner, K., 2005, Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors, Available at: http://arxiv.org/ftp/cs/papers/0512/0512085.pdf [Accessed January 16, 2008].

Holten, D., 2006, Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data, IEEE Transactions on Visualization and Computer Graphics,12(5): 741-748.

Holton, D., 2006, Hierarchical Edge Bundles: Visualization of Adjacency relations in Hierarchical Data, IEEE Transaction on Visualization and Computer Graphics, 12(5). Available at: http://www.win.tue.nl/~dholten/papers/bundles_infovis.pdf.

Huberman, B.A., Pirolli, P.L, Pitkow, J.E., & Lukose, R.M., 1998, Strong Regularities in World Wide Web Surfing, Science 280, 95.

Hull, D., 1993, Using statistical testing in the evaluation of retrieval experiments, In the Association for Computer Machinery, Special Interest Group in Information Retrieval (SIGIR) conference on Research and Development in Information Retrieval, 329–338.

Hwee Tou, N., & Lee, H. B., 1996, Integrating Multiple Knowledge Sources to Disambiguate Word sense: an exemplar-based Approach, Association for Computational Linguistics (ACL), 96: 40-47, California.

Humphreys, L., 2007, Mobile ocial network and social practice: a case study of Dodgeball. Available at: http://jcmc.indiana.edu/vol13/issue1/humphreys.html.

Izquierdo, L.R., & Hanneman, Robert, A., 2004, Introduction to the formal analysis of social networks using mathematics.

Izquierdo, L.R., & Hanneman, Robert A., 2006, Introduction to the formal analysis of social networks using mathematica. , (version 2). Available at: http://luis.izqui.org/papers/Izquierdo_Hanneman_2006-version.

Jacquez, G.M., Maruca, S., & Fortin, M.J., 2000, From Fields to Objects: A Review of Geographic Boundary Analysis, Journal of Geographical System, 2(3): 221-241.

Jaffe, A., Tassa, T., & Davis, M., Generating Summaries and Visualization for Large Collections of Geo-Referenced Photographs, in proceedings of Multimedia Information Retrieval (ACM), 89-98.

Jaffe,A. Naaman, M., Tassa, T., & Davis, M., 2006, Generating Summaries and Visualization for Large Collections of GeoReferenced Photographs, Multimedia Information Retrieval (MIR).

Jammalamadaka, S.R., & SenGupta, A., 2001, Topics in circular Statistics, Singapore: world Press.

Jain, A.K., Murty, M.N. & Flynn, P.J., 1999, Data Clustering: A Review. ACM Computing Surveys, 31(3): 264-323.

Jakob, V., 2005, Measuring Wikipedia, International Conference of the International Society for Scientometrics and Informetrics: Stockholm (Sweden), 10th edition. Available at: http://eprints.rclis.org/archive/00003610/01/MeasuringWikipedia2005.pdf [Accessed January 16, 2008].

Jia, Y., Hoberock, J., Garland, M., & Hart, J., 2008, On the Visualization of Social and other Scale-Free Networks, IEEE Transaction on Visualization and Computer Graphics, 14(6): 1285-1292.

Jones, C.B., & Purves, R.S., 2008, Geographical Information Retrieval, International Journal of Geographical Information Science, 15(22): 219-228.

Jones, C.B., Purves, R., Clough, P., & Joho, H., 2008, Modeling Vague Places with Knowledge from the Web, International Journal of Geographical Information Science, 22(10): 1045-1065.

Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Kreveld, M., & Weibel, R., 2002, Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT project. SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 387-388.

Jones, Q., Sukeshini, A., Grandhi, A., Terveen, L., & Whittaker, S., 2004, People to people to geographical places: the P3 framework for location based community systems, Computer supported cooperative work, Kluwer academic publishers., 13: 249-282.

Kamada, T. & Kawai, S., 1989, An algorithm for drawing general undirected graphs, Information Processing Letters, 31: 7-15.

Karimi, A., 2008, Handbook of Research on Geoinformatics, University of Pittsburgh, USA: Information Science Reference. Available at: http://www.igiglobal.com/reference/details.asp?ID=9706&v=preface.

Katz, J.S., 1993, Geographical proximity and scientific collaboration, scientometrics, 31(1): 31-43.

Kawachi, I., & Berkman, L.F., 2001, Social ties and mental health, Journal of urban health The New York Academy of Medicine, 78(3): 458-468.

Keim, D. A., 2002, Information Visualization and Visual Data Mining, IEEE Transaction on Visualization and Computer Graphics, 7(1), Available at: http://fusion.cs.uni-magdeburg.de/pubs/TVCG02.pdf.

Keim,, D. A., Andrienko, G., Fekete, J.D., Gorg, C., Kohlhammer, J., & Melancon, G., 2008, Visual analytics: Definition, Process, and Challenges, Information Visualization - Human centered Issues and Perspectives, Springer,154-175.

Keim,D.A., Hao, M.C., Dayal, U., Hsu, M., 2001, Pixel bar charts: a visualization techniques for very large multi attributes data sets, Information Visualization, in Symposium on Information Visualization (San Diego), IEEE Computer Society: Silver Spring, 113-122.

Kennedy, L. et al., 2007, How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections, In Proceedings of ACM Multimedia, New York, NY, USA, 631-640.

Kilkenny, M. & Nalbarte, L., 2000, Keystone sector identification: A graph theory-social network analysis approach The web book of regional science., West Virginia university: Regional research institute, Available at: http://ideas.repec.org/p/isu/genres/10308.html.

Kim, T., Jeong, H., Chew, Y., Bonner, M., Stasko, J., 2009, Social Visualization for Micro-Blogging Analysis, posters at VisWeek, Atlantic City, New Jersey, USA.

Klir, G.J., Yuan, B.,1995, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, Englewood Clis, NJ.

Kossinets, G. 2006, Effects of missing data in social networks, Social Networks, 28(3), Available at: http://arxiv.org/PS_cache/cond-mat/pdf/0306/0306335v2.pdf.

Kraak, M.J., 1999, Visualization for exploration of spatial data, geographical information science, 13(4): 285-287.

Kittur,A., Pendleton, B., & Mytkowicz, T, 2007, Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie, 25th ACM Conference on Human Factors in Computing Systems, CHI. Available at: http://www.viktoria.se/altchi/submissions/submission_edchi_1.pdf [Accessed January 28, 2008].

Kittur, A., Suh, B., Pendleton, B.A.,& Chi, Ed. H., 2007, He Says, She Says: Conflict and Coordination in Wikipedia, Human Factors in Computing Systems, San Jose, CA, USA. Available at: http://kittur.org/files/Kittur_2007_Wikipedia_CHI.pdf [Accessed January 16, 2008].

Kleinberg, J, 1999, Authoritative sources in a hyperlinked environment, Journal of the ACM, 46: 604-632.

kleinberg, J.M, 2000, Navigation in a Small World, Nature, 406(845). Available at: http://www.cs.cornell.edu/Info/People/kleinber/nat00.pdf.

kleinberg, J., 20001, small world phenomena and the dynamics of information, ACM Symposium on theory of computing,163-170.

Knoke, D. & Kuklinski, J.H., 1982, Network Analysis, Sage publications.

Krishnamoorthy, K. 2001, Handbook of Statistical Distributions with Applications, Bocaraton, Chapman and Hall, ISBN 1-584-88635-8 of.

Koskinen, I., 2003, User-generated content in mobile multimedia: empirical evidence from user studies, IEEE Xplore, Proceedings of International Multimedia and Expo.

Kossinets, G., 2008, Effects of missing data in social networks, social networks, 28(3): 247-268.

Kraak, M.J., 2003, The space-time cube revisited from a geovisualization perspective, In Proceedings of the 21st International cartographic Conference.

Krygier, J. & Wood, D., 2002, Making Maps: A Visual Guide for Map Design in GIS, Guilford Publications, ISBN 1593852002.

Kwan, M.P., Casa, I., & Schmitz, B.C., 2004, Protection of Geoprivacy and Accuracy of Spatial Information: How Effective are Geographical Masks? Cartographica, The International Journal for Geographic Information and Geovisualization, 39(2): 5-28.

Lampe, C., Ellison, N., & Steinfield, C., 2006, A Face(book) in the crowd:(Social searching vs. social browsing, Proceedings of CSCW, New York: ACM Press, 167-170.

Larson, R.R., 1995, Geographic Information Retrieval and Spatial Browsing, Edited by Linda C. Smith and Myke Gluck, editors, Geographic Information Systems and Libraries: Patrons, Maps, and spatial Information, 81-123.

Laudon, K.C., 1996, Markets and privacy, Communication of the ACM, 39(9), Available at: http://delivery.acm.org/10.1145/240000/234476/p92laudon.pdf?key1=234476&key2=3463075121&coll=GUIDE&dl=GUIDE&CFID=77552941&CFTOKEN=94861573.

Lee, B., Parr, C., Plaisant, C., Bederson, B.B., Veksler, V.D.,Gray, W.D., & Kotfila, C., 2006, TreePlus: Interactive Exploration of Networks with Enhanced Tree Layouts, IEEE Transactions on Visualization and computer Graphics,12(6):1414-1426.

Lee, J., & McKendree, J., 1999, Learning vicariously in a distributed environment, Active Learning, 10: 4-9.

Lee, B.,Plaisant, C., & Fekete, J.D., 2006, Task Taxonomy for Graph Visualization, in proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization (BELIV 06), ACM Press: 82-86.

Lemmens, R. & Deng, D., 2008, Web2.0 and semantic web: clarifying the meaning of spatial features, Workshop "Semantic Web meets Geospatial Applications", held in conjunction with AGILE 2008, short paper. Available at: http://musil.uni-muenster.de/wp-content/files/agileworkshop/Lemmens_DengClarifyingTheMeaningOfSpatialFeatures.pdf.

Lengerich, A.C.R., Chen, J., Eugene J., & Meyer, H., and MacEachren, A.M., 2005, Combining usability techniques to design geovisualization tools for epidemiology, Cartography and Geographic Information Science, 32(4): 243-255.

Lew, M.S., Sebe, N., Djeraba, C., & Jain, R., 2006, Content-based multimedia information retrieval: State of the art and challenges, ACM Transaction on Multimedia Computing, Communications and Applications. 2(1): 1-19.

Li, H., Srihari, R., Niu, C., & Li, W., 2003, InfoXtract location normalization: a hybrid approach to geographic references in information extraction, in Workshop on the Analysis of Geographic References, Edmonton, Canada. Available at: http://acl.ldc.upenn.edu/W/W03/W03-0106.pdf.

Li, H., Srihari, R., Niu, C., & Li, W., 2002, Location Normalization for Information Extraction, in proceedings of the 19th Conference on Computational Linguistics, Taipei, Taiwan.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A., 2005, Geographic Routing in Social Networks, National Academy of Science, 102(33):11623-11628.

Liberman, J, 2006, Upper and Lower Case on the Geosemantic Web, Available at: http://www.ordnancesurvey.co.uk/oswebsite/partnerships/research/research/TerraCognita_Papers_Presentations/Lieberman.pdf.

Liu, H., Maes, P., & Davenport, G., 2006, Unraveling the taste fabric of social networks. International Journal on semantic web and information systems, 2(1):42-71.

Longley, P.A., Goodchild, M.,F., Maguire, D., & Rhind, D.W., 2005, Geographic Information Systems and Science, Wiley, 2nd Edition, ISBN: 0-470-87000-1.

Longueville, B., Ostländer, N. & Keskitalo, C., 2009, Addressing vagueness in Volunteered Geographic Information (VGI) - A case study, International Journal of Spatial Data

Infrastructures Research, Special Issue GSD-11, Available at: http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/viewFile/132/137.

Lynch, C., 1998, The evolving internet: Applications and network service infrastructure, Journal of the American Society for Information Science, 49(11): 961-972.

MacEachren, M. A., 1992, Visualizing Uncertain Information, Cartographic Perspective, 13: 10-19.

MacEachren, A.M, & Kraak, J.M., 2001, Research Challenges in Geovisualization, Cartography and Geographic Information Science, 28(1). Available at: http://www.geovista.psu.edu/sites/icavis/pdf/visagenda.pdf.

MacEachren, A., Kraak, M.J., & Dykes, J., Exploring geovisualization, 143-158.

Maluf, R., 2002, Markov Models for Language-Idependent Namded Entitiy Recognition, in proceedings of Conference on Natural Language Learning (CoNLL), 187-190, Taipei, Taiwan.

Man, P.D., 1983, Blindness and Insight: Essays in the Rhetoric of Contemporary Criticism Second Edition, London: Routledge.

Mansbridge, L., 2005, Perceptions of Imprecise Regions in Relation to Geographical Information retrieval.

Marczak, B., Lat/Long Tutorial. Available at: http://209.85.229.132/search?q=cache:FyMghlXOOsoJ:www.al911.org/wireless/XYTUTORP.DOC+WHAT+IS+THE+LAT,+LONG+FOR+lONDON&cd=5&hl=en&ct=clnk&gl=uk&client=firefox-a.

Mardia, K.V. & Jupp, P.E. 1999, Directional Statistics, Germany: Willey, Print ISBN: 9780471953333.

Marlow, C., Naaman, M., Boyd, D., Davis, M., 2006, Tagging Paper, Taxonomy, Flickr, academic Article, ToRead, in Proceedings of Hypertext, ACM Press, 31-40.

Martin, D., Cockings, S., Leung, S., 2006, An Introduction to Geographical Referencing for Social Scientist, Available at: http://www2.geog.soton.ac.uk/georefer/files/Workshop1%20Pres.pdf.

McCurley, S.K., 2001, Geospatial mapping and navigation of the web, In Proceedings of the 10th International WWW Conference HongKong, 221-229.

McDonald, J., 2003, Let's get more positive about the term 'lurker' – CPSquare Class Project. Available at: http://www.cpsquare.org/edu/News/archives/LurkerProjectCoPWorkshopSPring03a.doc.

McGrath, C., Blythe, J. & Krackhardt, D., 1997, The effect of spatial arrangement on judgments and errors in interpreting graphs, Social Networks, 19(3), 223-224.

McRoy, S., 1992, Using Multiple Knowledge Sources for Word Sense Discrimination, Computational Linguistics,18(1): 1-30.

Meiss, M.R., Menczer, F. & Vespignani, A., 2005, On the lack of typical behavior in the global Web traffic network, In proceedings of International World Wide Web Conference, 510-518.

Meiss, M.R., Menczer, F., Fortunato, S., Flammini, A., & and Vespignani, A., 2008, Ranking Web sites with real user traffic, In proceedings of the International Conference on web Search and Web Data Mining.

Mika, P., Ciaramita, M., Zaragoza, H., & Atserias, J., 2008, Learning to tag and tagging to learn: A case study on Wikipedia, EEE Intelligent Systems , 23(5):26–33.

Miller, P., 2005, Web 2.0: Building the New Library, Ariadne, Web Magazine for Information professionals, issue:45.

Mitchell, A., 2005,The ESRI Guide to GIS Analysis, Volume 2, ESRI Press.

Menczer, F., 2004, Lexical and Semantic Clustering by Web Links, Journal of American Society for Information Science and Technology, 55(14): 1261-1269.

Mok, D., Wellman, B. & Basu, R., 2007, Did Distance Matter Before the Internet? Interpersonal Contact and Support in the 1970s, Social Networks, 29(3): 340-461.

Monge, P.R. & Contractor, N.S., 2003, Theories of communication networks, US: Oxford University Press.

Montello, D.R., Goodchild, F., Gottsegen, J., & Fohl, P., 2003, Where's Downtown?: Behavioral Mehtods for Determining Referents of Vague Spatial Queries, Spatial Cognition and Computation, 3(2&3):185-204.

Montello, D.R., 2002, Cognitive Map Design Research in the Twentieth Century: Theoretical and Empirical Approaches, Cartography and Geographic Information Science. 29(3): 283-304.

Mooney, P., Sun, H. Corcoran, P., & Yan, L., 2011, Citizen Generated Spatial Data and Information: Risks and Opportunities, American Journal of Engineering and Technology Research, 11(9).

Naaman, M., 2006, Eyes on the world. Yahoo! Research Berkeley, IEEE computer society, 39(10):108-111.

Morgan, L.J., 2004, New dimension in privacy: spatial privacy in the geographic information age, Available at: http://www.spatial.maine.edu/~nittel/lp/joe_morgan_abstract.pdf.

Motulsky, H., 1995, Intuitive biostatistics, Oxford University Press.

Mountain, D. & Raper, J., 2001, Modeling human spatio-temporal behavior: a challenge for location based services, GeoComputation, University of Queensland, Brisbane, Australia.

Myers, W., & Dyke. C., 2000, State of the Internet, United States Internet Council and International Technology and Trade Associations (ITTA) Inc.

Naaman, M., Harada, S., Wang, Q., & Paepcke, A., 2004, Adventures in Space and Time: Browsing Personal Collections of Geo-Referenced Digital Photographs, Technical Report, Stanford University.

Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., & Paepcke, A., 2004, Context Data in Geo-Referenced Digital Photo Collections, Proceedings of the 12th Annual ACM Conference, Multimedia, ACM Press,196-203.

Naaman, M., Song, Y., Paepcke, A., Garcia-Molina, H., 2004, Automatic Organization for Digital Photographs with Geographic Coordinates, Proceedings of 4th ACM/IEEE-CS Joint Conference on Digital Libraries ACM Press, 53-62.

Naaman, M., Paepcke, A. & Garcia-Molina, H., 2003, From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates, in 10th International Conference on Cooperative Information Systems, Available at: http://ilpubs.stanford.edu8090/598/1/2003-37.pdf.

Nair, R., Reid, N. & Davis, M., 2005, Photo LOI: Browsing multi user photo collections, In Proceedings of the 13th International Conference on Multimedia, ACM Press, Available at: http://fusion.sims.berkeley.edu/GarageCinema/pubs/pdf/pdf_138B148E-B6F5-4685-AAC94AA2B9EE3183.pdf

Natsui, T., 2002, FTC workshop report, location based service for the commercial purpose and protection.

Newman, M.E.J., 2001a, The structure of scientific collaboration networks, Proceedings of National Academic Sciences (PNAS), 98(2): 404-409.

Newman. M.E.J., 2005b, Power laws, Pareto distributions and Zipfs law, Cotemporary Physics, 46:323.

Nielsen, J., 2006, Participation Inequality: Lurkers vs. Contributors in Internet Communities. Accessible at http://www.useit.com/alertbox/participation_inequality.html Retrieved on 13th January 2012.

Nonnecke, B. & Preece, J., 2001, Why Lurkers Lurk, American Conference on Information Systems (AMCIS), Boston, June 2001, Available at: http://snowhite.cis.uoguelph.ca/~nonnecke/research/whylurk.pdf

Nonnecke, B., Preece, J. & Andrews, D., 2003, The top five reasons for lurking: improving community experiences for everyone, Available at: www.elsevier.com/locate/comphumbeh.

Nonnecke, B., Preece, J., & Andrews, D., 2004a, What lurkers and posters think of each other. Paper presented as part of the proceedings of the 37th Hawaii International Conference on System Sciences , Hawaii, USA.

Nonnecke, B., Preece, J., Andrews, D. & Voutour, R., 2004b, Online Lurkers Tell Why, Paper presented to the Tenth American Conference on Information Systems, New York, New York, USA.

Nooy, W., Mrvar, A. & Batagelj, V., 2005, Exploratory Social Network Analysis with Pajek, Cambridge UK: Cambridge University Press.

Ochoa, X., & Duval, E., 2008, Quantitative analysis of user generated content on the web, In Proceedings of Webevolve2008: Web Science Workshop at WWW2008, (April 2008).

O'Hare, N., Gurrin, C., Jones, J.F., & Smeaton, A.F., 2005, Combination of Content Analysis and Context Features for Digital Photograph Retrieval, In 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies. Available at: http://doras.dcu.ie/391/1/ewimt_2005.pdf.

Olivares, X., 2011, Large Scale Image Retrieval Based on User Generated Content, PhD thesis, PF University.

Onsrud, H.J., Johnson, J.P. & Lopez, X., 1994, Protecting personal privacy in using geographic information systems, Photogrammetric Engineering and Remote Sensing, 60(9): 1083-1095.

O'Reilly, T., 2005a, What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly Media, Inc. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20. Html.

O'Reilly, T., 2007b, What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software ,International Journal of Digital Economics, 65(4580): 17-37. Available at: http://mpra.ub.uni-muenchen.de/4580/1/MPRA_paper_4580.pdf.

Ortega , F.& Gonzalez-Barahona, J.M., 2007, Quantitative Analysis of the Wikipedia Community of Users, WikiSym, October 21–23, Montreal, Quebec, Canada. Available at: http://libresoft.es/downloads/wiki35f-ortega.pdf.

Ortega, F., Gonzalez-Barahona, J.M. & Robles, G. 2008, On The Inequality of Contributions to Wikipedia.

Overell, S.E., 2009, Geographic Information Retrieval: Classification, Disambiguation and Modeling', PhD Thesis, Department of Computing, Imperial College London, July.

Pasley, R.C., Clough, P. & Sanderson, M., 2007, Geo-tagging for imprecise regions of different sizes, Proceedings of the 4th ACM, Workshop On Geographic Information Retrieval, Lisbon, Portugal, 77-82.

Pachet, F., 2005, Knowledge Management and Musical Metadata. Encyclopedia of Knowledge Management. Idea Group.

Pastor-Satorras, R & Vespignani, A., 2004, Evolution and Structure of the Internet: A Statistical Physics Approach, Cambridge University Press.

Perer, A. & Shneiderman, B., 2006, Balancing Systematic and Flexible Exploration of Social Networks. IEEE Transaction on Visualization and Computer Graphics, 12(5). Available at: http://hcil.cs.umd.edu/trs/2006-25/2006-25.pdf.

Pigeau, A. & Gelgon, M., 2003, Organizing a personal image collection with statistical model-based ICL lustering on spatio-tempral camera phone meta-data. Journal of Visual Communication and Image Representation, 15(3), pp.425-445.

Pirolli, P.,Wollny,E., & Suh, B., 2009, So You Know You're Getting the Best Possible Information: A Tool That Increases Wikipedia Credibility, Proceedings of the 27th International Conference on Human Factors in Computing Systems, 1505-1508.

Popescu, A., Grefenstette, G. & Moëllic, P.A., 2008, Gazetiki: Automatic Creation of a Geographical Gazetteer, International Conference on Digital Libraries, 85-93.

Porta, S., Crucitti, P. & Latora, V., 2006, The Network Analysis of Urban Streets: a Primal Approach, Environemnet and Planning, 33: 705-725, Available at: http://arxiv.org/ftp/physics/papers/0506/0506009.pdf.

Priedhorsky, R., Chen, J., Lam, K., Panciera, K., Terveen, L. & Riedl, J., 2007, Creating, Destroying and restoring Value in Wikipedia, Proceedings of the 2007 international ACM conference on Supporting group work, 259-268.

Purchase, H, 1994, Which aesthetic has the greatest effect on human understanding, Edited by Battista, G.D., Graph Drawing, Rome, Italy, Springer 1998, 248-261.

Purves, R.S. & Jones, C.B., 2006, Geographic Information Retrieval (GIR), Computers Environment and Urban Systems, 30: 375-377.

Purves, R.S., Clough, P. & Joho, Hideo, 2005, Identifying imprecise regions for geographic information retrieval using the web, In Billen, R., Drummond, J., Forrest, D., and Joao, E., Proceeding of the GIS research UK 13th Annual Conference, Glasgow, UK, 313-318.

Purves, R., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., & Vaid, S., 2007, The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet, International Journal of Geographical Information Science, 21(7): 717-745.

Purves, R., Dykes, J., Edwardes, A., Hollenstein, L., Mueller, D., & Wood, J., 2009, Describing the space and place of digital cities through volunteered geographic information, GeoViz Hamburg.

Rafaeli, S., Ravid, G. & Soroka, V., 2004, De-lurking in virtual communities: a social communication network approach to measuring the effects of social and cultural capital. Paper presented to the 37th Hawaii International Conference on System Sciences, Hawaii, USA.

Ralph Gross, A.A., 2005, Information Revelation and privacy in online social networks, the facebook case, ACM Workshop on Privacy in the Electronic Society, Available at: http://www.heinz.cmu.edu/~acquisti/papers/privacy-facebook-gross-acquisti.pdf.

Raper, J., 2002, Location privacy: a new challenge for geographic information science, 5th AGILE Conference on GIScience, Palma, Mallorca.

Rapoport, A., 1958, Nets with Reciprocity Bias, Bulletin of Mathematical Biology, 20(3): 191-201.

Rattenbury, T. & Naaman, Mor, 2009, Methods for Extracting Place Semantics from Flickr Tags, ACM Transaction on the Web, 3(1). Available at: http://infolab.stanford.edu/~mor/research/RattenburyPlacesSemanticsTweb09.pdf.

Rattenbury, T., Good, N., & Naaman, M., 2007, Towards Automatic Extraction of Event and Place Semantics from Flickr Tags, In Proceeding of SIGIR, Amsterdam, The Netherlands. Available at: http://delivery.acm.org/10.1145/1280000/1277762/p103rattenbury.pdf?key1=1277762&key2=7896595521&coll=GUIDE&dl=GUIDE&CFID=57201841&CFTOKEN=39070420.

Rauch, E., Bukatin, M. & Baker, K., 2003, A confidence-based framework for disambiguating geographic terms, In Workshop on the Analysis of Geographic References, Edmonton, Alberta, Canada, Available at: http://delivery.acm.org/10.1145/1120000/1119402/p50-rauch.pdf?key1=1119402&key2=9283345521&coll=GUIDE&dl=GUIDE&CFID=56252354&CFTOKEN=23762071.

Reda, K., Tantipathananandh, C., Berger-Wolf, T., Leigh, J., & Johnson, A., 2009, SocioScape - a Tool for Interactive Exploration of Spatio-Temporal Group Dynamics in Social Networks, in Proceedings of the IEEE Information Visualization Conference (InfoVis '09), Atlantic City, New Jersey, 2009. Available at: http://vis.computer.org/VisWeek2009/infovis/sessions_posters.html.

Raymond, E. S., 2001, The Cathedral and the Bazaar, Musing on Linux and Open Source by an Accidental revolutionary, O'Reilly Media, California, US.

Rodgers, P., 2004, Graph Drawing Techniques for Geographic Visualization, Edited by

Rogowitz, B.E. & Treinish, L.A., 1996, How NOT to Lie with Visualization, Computers in Physics, 10(3): 268-273.

Salmon, G., 2002, E-Tivities the key to active online learning, London England: Kogan Page Limited.

Salmon, G., 2003, E-moderating the key to teaching and learning online, London England: Routledge Falmer Taylor & Francis Group.

Sanchez Abril, P., 2007, A (My)Space of One's Own: On Privacy and Online Social Networks. Northwestern Journal of Technology and Intellectual Property, 6(1), 72-89.

Santini, S., 2001, exploratory image databases: content-based retrieval, Academic Press, Inc., Duluth, MN, USA.

Schmitz, P., 2006, Inducing Ontology from Flickr Tags, Proceedings of Collaborative Web Tagging Workshop.

Schroeder, R., Huxor, A. & Smith, A., 2001, Activeworlds: geography and social interaction in virtual reality, futures, 33(7): 569-587.

Schultz, N., & Beach, B., 2004, From Lurkers to Posters, Australizan flexible Learning Framework, available at flexiblelearning.net.au

Scott, J., 1991, Social Network Analysis: A Handbook, Sage Publications, London, London: sage publication.

Scott, J., Tallia, A., Crosson, J.C., Orzano, A.J., Stroebel, C., DiCicco-Bloom, B., O'Malley, D., Shaw, E., & Crabtree, B., 2005, Social network analysis as an analytic tool for interaction patterns in primary care practices, Annals of family medicine, 3(5): 443-449.

Shardanand, U., Maes, 1995, 1995, Social Information Filtering: Algorithms for Automating 'Word of Mouth', in Proceedings of Computer Human Interactions (CHI), Human Factors in Computing Systems, 210-217.

Shankland, S., 2008, Webware news on geotagging, Available at: http://www.webware.com/8300-1_109-2.html?keyword=geotagging.

Sholtz, P., 2001, Transaction Costa and the Social Costs of Online Privacy, First Monday, 5(6). Available at: http://outreach.lib.uic.edu/www/issues/issue6_5/sholtz/index.html.

Sigurbjörnsson, B. & Zwol, R. van, 2008, Flickr Tag Recommendation based on Collective Knowledge, Proceedings of the 17th International conference on World Wide Web, ACM Press, 327-326.

Silva, M. J., Martins, B., Chaves, M., Cardoso, N., & Afonso, A.P., 2006, Computers Environment and Urban Systems, 378-399.

Sim, J., & Wright, C.C., 2005, The Kappa Statistic in Reliability Studies: Use, Interpretation and Sample Size Requirements, Physical Therapy, 85(3): 257-268.

Simon, N., 2010, Participatory Museum, Published by MUSEUM, Santa Cruz, California.

Skupin, A. & Fabrikant, S.I., 2003, Spatialization methods: a cartographic research agenda for non geographic information visualization, Cartography and Geographic Information Science, 30(2): 99-119.

Slingsby A, Dykes J., & Wood J., 2008, Using Treemaps for Variable Selection in Spatio-Temporal Visualization, Information Visualization, 7: 1473-8716.

Slingsby, A., Dykes, J. & Wood, J., 2009, Configuring Hierarchical Layouts to Address Research Questions, IEEE, Available at: http://gicentre.org/papers/slingsby_configuring_2009.pdf.

Slocum, T.A., 1999, Thematic Cartography and Visualization, New Jersey: Prentice Hall.

Smith, D.A., 2002, Detecting and Browsing Events in Unstructured Text, In proceedings of the 25th Annual ACM, Special Interest Group in Information Retrieval (SIGIR) Conference, Finland, 73-80.

Smith, D.A., & Crane, G., 2002, Disambiguating Geographic Names in a Historical Digital Library. ECDL'01: Proceedings of the 5th European conference on research and advanced technology for digital libraries, London, UK, Springer-Verlag, 127-136.

Smith, D.A., & Crane, G., 2001, Disambiguation geographic names in a historical digital library, In proceedings of the 5th European Conference on Research and Advances Technology for Digital Libraries (ECDL'01).

Smith, D.A. & Mann, G.S., 2003, Bootstrapping toponym classifiers, In Proceedings of theHuman Language Technology (HLT-NAACL), Workshop on analysis of geographic references, Association for Computational Linguistics, Morristown, NJ, USA, 45-49.

Smith, B., & Mark, D., 2001, Geographical categories: an ontological investigation. International Journal of Geographic Information Science, 15 (15): 591-612.

Stolte, C., Tang, D. & Hanrahan, P., 2002, Polaris: A System for Query, Analysis and Visualization of Multidimensional Relational Databases, IEEE Transaction on Visualization and Computer Graphics, 8(1). Available at: http://graphics.stanford.edu/papers/polaris_extended/polaris.pdf.

Stvilia, B., Twidale, M.B., Smith, L.C., & Gasser, L., 2005, Assessing Information Quality of a Community Based Encyclopedia, international conference on information quality. Available at: http://www.isrl.uiuc.edu/~stvilia/papers/quantWiki.pdf [Accessed January 29, 2008].

Suh, B., Kittur, A., Pendleton, B., 2007, Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations, IEEE Symposium on Visual Analytics Science and Technology.

Tantipathananandh, C., Berger-Wolf, T. & Kempe, D., 2007, A Framework for Community Identification in Dynamic Social Networks, In Proceedings of 13th ACM SIGKDD international Conference on Knowledge Discover and Data Mining, 717-726.

Taylor, J. R.,1999, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, University Science Books,128-129, ISBN 0-935702-X

Thomas, J., &Cook, K., 2006, Visual Viewpoints: A visual analytics agenda, IEEE Computer Graphics and Applications, 26(1): 10-13

Thomas, J. & Cook, K., 2005, Illuminating the path: the research and development agenda for visual analytics, US: National Visualization and Analytic Centre, Available at: http://nvac.pnl.gov/agenda.stm.

Thompson, G.F., 2003, Between hierarchies and markets: the logic and limits of network forms of organization, Oxford university press.

Tichy, M., Tushman, L.,& Fombrun, C., 1979, Social Network Analysis for Organizations, Academy of Management Review, 4(4): 507-519.

Toyama, K., Logan, R., Roseway, A., & anandan, P., 2003, Geographic Location Tags on Digital images, In Proceedings of the 11th ACM International Conference on Multimedia, 156-166.

Troll, G & Graben P.B., 1998, Zipf's law is not a consequence of the central limit theorem, Physical Review.E, 57(2): 1347-1355.

Tufte, E., 2001, The Visual Display of Quantitative Information, 2nd Edition.

Twaroch, F.A., Jones, C.B. & Abdelmoty, A.I., 2008, Acquisition of a Vernacular Gazetteer from Web Sources, Proceedings of the First International Workshop on Location and the Web, 17th International World Wide Web Conference, 61-64.

Twaroch, F.A., Smart, P.D. & Jones, C.B., 2008a, Mining the web to detect place names. Workshop On Geographic Information Retrieval, Proceeding of the 2nd international workshop on Geographic information retrieval, Napa Valley, California, USA, Available at: http://delivery.acm.org/10.1145/1470000/1460017/p43-twaroch.pdf?key1=1460017&key2=6472194521&coll=&dl=&CFID=15151515&CFTOKEN=6184618.

Twaroch, F.A., Smart, P.D. & Jones, C.B., 2008b, Modeling the Web to Detect Place Names, 5th Workshop on Geographic Information Retrieval, 43-44.

Van House, N.A., 2007, Flickr and Public Image Sharing: Distant Closeness and Photo Exhibition, Computer Human Interaction (CHI), April 28-May 3, San Jose, California, USA.

Vanwey, L.K., Rindfuss, R., Myron, Gutmann, M.P., Entwisle, B., & Balk, D.L., 2005, Confidentiality and Spatially Explicit Data: Concerns and Challenges, Proceedings of the national academy of sciences of the united states of America: Spatial Demography special feature., 102(43):15337-15342.

Vickers, D. & Rees, P., 2006, Introducing the National Classification of Census Output Areas, Population Trends, 125.

Viegas, F.B. & Donath, J., 2004, Social Network Visualization: Can We Go Beyond the Graph? Workshop on Social Networks for Design and Analysis: Using Network Information in Computer Supported Cooperative Work (CSCW), 4: 6-10.

Viégas, F.B., Wattenberg, M., Ham, F., Kriss, J., McKeon, M., et al., 2007, Many Eyes: A Site for Visualization at Internet Scale, IEEE transactions on visualization and computer graphics, 13(6). Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4376131&isnumber=4376125.

Viégas, F.B., Wattenberg, M., Kriss, J. & Ham, F. 2007, Talk Before You Type: Coordination in Wikipedia, Hawai International Conference on System Sciences (HICSS) 40., Available at: http://www.research.ibm.com/visual/papers/wikipedia_coordination_final.pdf [Accessed January 16, 2008].

Viégas, F.B., Wattenberg, M., & Dave, K., 2004, Studying Cooperation and Conflict between Authors with History Flow Visualization. CHI, Vienna, Australia. Available at: http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf [Accessed January 16, 2008].

Volkel, M., Krotzsch, M., Vrandecic, D., Haller, H., & Studer, R., 2006, Semantic Wikipedia, In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26), Available at: http://citeseer.ist.psu.edu/735238.html.

Ware, C., Purchase, H., Colpoys, L., & McGill, M., 2002, Cognitive measurements of graph aesthetics, Information Visualization, 1(2): 103-110.

Wasserman, F. & Faust, K., 1994, Social Network Analysis, Cambridge UK: Cambridge University Press.

Waters, T. & Evans, A.J., 2003, Tools for web-based GIS mapping of a "fuzzy" vernacular geography, Geographic Information Research UK (GISRUK), pp.9-11.

Waters, N., 2007, Why you can't cite Wikipedia in my class, Communications of the Association for Computer Machinery (ACM), 50(9): 15–17, 2007.

Watts, D.J. & Strogatz, S.H., 1998, Collective Dynamics of Small World Networks, Nature, 393: 440-442.

Watts, D.J., 2007, Connections A twenty-first century science, Nature 445:489, [PubMed: 17268455].

Wellman, B, & Berkowitz, S.D., 1988, Structural Analysis in the Social Sciences: a Network Approach, UK: Cambridge University Press.

Wellman, B., 1996, Are personal communities local? A Dumptarian reconsideration, Social Networks, 18(4): 347-354.

Wellman, B., & Hampton, K., 1999, Living Networked in a Wired World, Contemporary Sociology, 28(6), Available at: http://chass.utoronto.ca/~wellman/publications/onandoffline/onandoff.pdf.

Welty Lefever, D., 1926, Measuring geographic concentration by means of the standard deviational ellipse, The American Journal of Sociology, 32(1): 88-94.

Westman, S., 2009, In Information Retrieval: Searching in the 21st Century, Edited by Goker, A., & Davies, J., John Wiley &Sons, Ltd, 62-83.

Wetherell, C., Plakans, A. & Wellman, B., 1994, Social networks, kinship and community in Eastern Europe, Journal of Interdisciplinary History, 24(4): 639-663.

White, D.R., & Harary, F., 2001, The Cohesiveness of Blocks in Social Networks: Node Connectivity and Conditional Density, Sociological Methodology, 31(1): 305.

Wijffelaars, M., Vliegen, R., Wijk, J., & Linden, E, 2008, Generating Colour Palettes using Intuitive Parameters, Computer Graphics Forum, 27(3): 743-750.

Wijk, J. & Selow, E., 1999, Cluster and Calendar based Visualization of Time Series Data, IEEE Symposium on Information Visualization, 4-9.

Wilkinson, D.M., & Huberman, B.A., 2007, Assessing the Value of Cooperation in Wikipedia, First Monday, 12(4), Available at: http://www.firstmonday.org/issues/issue12_4/wilkinson/.

Williams, B., 2004, Participation in on-line courses – how essential is it? International Forum of Educational Technology and Society Formal Discussion Initiative. http://ifets.ieee.org/discussions/discuss_september2004.html

Wong, J., 2008, What Do We "Mashup" When We Make Mashups, in Proceedings of the 4th international workshop on End-user software engineering, 35-39.

Wood, J., Dykes, J., Slingsby, A., & Clarke, K., 2007, Interactive Visual Exploration of a Large Spatio-Temporal Data set: Reflections on a Geovisualization Mashup, IEEE Transactions on Visualization and Computer Graphics, 13(6): 1176-1183.

Wood, J., Dykes, J., Slingsby, a., & Clarke, K., 2007, Interactive Visual Exploration of a Large Spatio-Temporal Data set: Reflections on a Geovisualization mashup, IEEE Transaction on Visualization and Computer Graphics, 13(6):1176-1183.

Wood, J. Dykes, J., Slingsby, A, & Radburn, R., 2009, Flow Trees for Exploring Spatial Trajectories. Proceedings of the GIS Research UK 17th Annual Conference, edited by Fairbairn, D., Durham, UK, 229-234.

Wood, J., Slingsby, A. & Dykes, J., 2010, Layout and Colour Transformations for Visualizing OAC Data. GISRUK 2010, UCL, London. Available at: http://gicentre.org/papers/gisruk10/wood_layout_2010.pdf.

Wood, J., Slingsby, A., & Dykes, J., 2010., Edited by MacEachren, A., & Miksch, S., Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 285-286.

Woodruff, A.G., & Plaunt, C., GIPSY: Automated Geographic Indexing of Text Documents. Journal of the American Society for Information Science, 45(9): 645–655.

Yamaguichi, K., & Buskens, B., 1999, A New Model for Information Diffusion in Heterogeneous Social Networks, Sociological Methodology, 29(1): 281-325.

Young, J., 2006, Wikipedia founder discourages academic use of his creation, The Chronicle of Higher Education: The Wired Campus , June 2006, Available at http://tinyurl.com/lxhxo.

Zachary, J., & Iyengar, S., 2002, Content Based Image Retrieval Systems, Technical Report TR, Department of Computing Science.

Zadeh, L.A., 2008, Is there a Need for Fuzzy logic?, an International Journal Information Sciences, o178: 2751-2779. Available at: http://www.eecs.berkeley.edu/~zadeh/papers/Is%20there%20a%20need%20for%20fuzzy%20logic.pdf

Zips, G.K., 1949, Human Behavior and the Principle of Least Effort, Addison-Wesley, Hafner publication company.

Zlatic, V., Bozicevic, M., Stefancic, H., & Domazet, M., 2006, Wikipedias: Collaborative web-based encyclopedias as complex networks, arXiv:physics, 3. Available at: http://arxiv.org/PS_cache/physics/pdf/0602/0602149v3.pdf.

Zwol, R., Murdock, V., Liuis, P., & Georgina, R., 2008, Diversifying image search with user generated content, Proceeding of the 1st ACM international conference on multimedia information retrieval, October 30-31, Vancouver, British Colombia, Canada.

Zwol, R., Murdock, V., Garcia Pueyo, L., & Ramirez G, 2008, Diversifying Image Search by user Generated Content, Proceeding of the 1st ACM international conference on Multimedia information retrieval.

12 Appendices

Appendix 1

This appendix demonstrates the features and format of the Geo-Names Feature Codes in the ‘Geo-Names Standard Table’.

The main 'geoname' table has the following fields:

----

geonameid : integer id of record in geonames database

name : name of geographical point (utf8) varchar(200)

asciiname : name of geographical point in plain ascii characters, varchar(200)

alternatenames : alternatenames, comma separated varchar(5000)

latitude : latitude in decimal degrees (wgs84)

longitude : longitude in decimal degrees (wgs84)

feature class : see http://www.geonames.org/export/codes.html, char(1)

feature code : see http://www.geonames.org/export/codes.html, varchar(10)

country code : ISO-3166 2-letter country code, 2 characters

cc2 : alternate country codes, comma separated, ISO-3166 2-letter country code, 60 characters

admin1 code : fipscode (subject to change to iso code), see exceptions below, see file admin1Codes.txt for display names of this code; varchar(20)

admin2 code : code for the second administrative division, a county in the US, see file admin2Codes.txt; varchar(80)

admin3 code : code for third level administrative division, varchar(20)

admin4 code : code for fourth level administrative division, varchar(20)

population : bigint (8 byte int)

elevation : in meters, integer

gtopo30 : average elevation of 30'x30' (ca 900mx900m) area in meters, integer

timezone : the timezone id (see file timeZone.txt)

modification date : date of last modification in yyyy-MM-dd format

Appendix 2:

This appendix demonstrates the hierarchy in Geo-Names feature codes.

Accuracy/precision classification (GeoNames hierarchical levels)

“

PPLC capital of a Political entity

PPLG seat of government of a political entity PPLG seat of Government of a Political entity

PPLA seat of a first-order administrative division (PPLC takes precedence over PPLA)

PPLA2 seat of a second-order administrative division PPLA2 seat of a second-order administrati division

PPLA3 seat of a third-order administrative division PPLA3 seat of a third-order administrati division

PPLA4 seat of a fourth-order administrative division PPLA4 seat of a Fourth-Order Administrati Division

PPL populated place a city, town, village, or other agglomeration of buildings where people li and work

PPLX section of populated place

PPLS populated places cities, towns, villages, or oth It is a plural feature if a couple of villag
agglomerations of buildings where people live and wc are merged into one GeoNames entry

PPLF farm village a populated place where the population is largely engaged in agricultu activities

PPLL populated locality an area similar to a locality but with a small group of dwellings or oth buildings

PPLR religious populated place a populated place whose population is largely engaged religious occupations

PPLQ abandoned populated place

PPLW destroyed populated place a village, town or city destroyed by a natural disaster, or war

STLMT Israeli settlement.

Appendix 3:

This appendix includes the step-by-step example of how the disambiguated algorithm disambiguates the given FHLI.

Example 1: FHLI (Austin, TX, US)

```
Populated place saved in arrayList successfully.2647529
 Countries read successfully.247
 Admin1 read successfully.3763
 old admin list read successfully: 0
 reading line: '333    austin, TX, US'
 an object of type DisambigFHLI is created successfully
about to seperate the fhli by ,
 locations to be disambiguated: austin
 locations to be disambiguated: TX
 locations to be disambiguated: US
 reading Location: austin
 spliting: austin
 size of the splitted array: 1
 reading Location:  TX
 spliting:  TX
 size of the splitted array: 2
 reading Location:  US
 spliting:  US
 size of the splitted array: 2
 locations to be disambiguated after resolving south, east, noerth, west3
austin
TX
US
30 populated places are found with the name austin
 ---------------------Content of admin1 list ----------------------:
 -----------------Content of populated Place list--------------------
Austin in BR -22.71667 -43.53333
Austin in CA 45.18339 -72.2824
Austin in HT 18.80639 -72.52333
Austin in US 35.44202 -92.52655
Austin in US 34.99842 -91.98376
Austin in US 35.53063 -92.79072
Austin in US 38.75839 -85.80803
Austin in US 36.82533 -86.01915
Austin in US 38.50279 -94.29995
Austin in US 36.1609 -90.15926
Austin in US 34.64066 -90.44927
Austin in US 36.32374 -80.97702
Austin in US 39.43673 -83.22241
Austin in US 35.20175 -88.2392
Austin in US 30.26715 -97.74306
Austin in US 41.88753 -87.76478
Austin in US 41.73727 -84.69412
Austin in US 46.28439 -87.45708
Austin in US 43.66663 -92.97464
Austin in US 42.8109 -76.47077
Austin in US 41.63118 -78.09139
Austin in US 41.59066 -71.65507
Austin in US 37.24244 -121.99829
Austin in US 38.78109 -107.9509
Austin in US 39.49326 -117.06953
Austin in US 38.67219 -112.12159
Austin in US 46.6391 -112.24529
Austin in US 44.60266 -118.49661
Austin in US 47.99148 -122.53987
Austin in US 47.6849 -117.16465
 alternatePP content after removing those with country name
Austin
Austin
Austin
Austin
Austin
```

```
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
0 alternate countries have been found for austin
0 populated places are found with the name TX
Texas is added to the alternate states list
 ---------------------Content of admin1 list ----------------------:
Texas in US
 -----------------Content of populated Place list--------------------
 alternatePP content after removing those with country name
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
0 alternate countries have been found for TX
3 populated places are found with the name US
 ---------------------Content of admin1 list ----------------------:
Texas in US
 -----------------Content of populated Place list--------------------
Us in FR 49.1 1.96667
Us in GT 14.95 -91.21667
Us in SN 16.09233 -16.37397
US is added to the alternate countries list................
 alternatePP content after removing those with country name
Austin
Austin
Austin
Austin
Austin
```

```
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
Austin
1 alternate countries have been found for US
................... contents of the alternate countries:
United States
 number of countries mentioned in FHLI is: 1
 comparing: Austin
Texas
IL for populated place is compared with: TX
IL with TX
 comparing: Austin
Texas
NY for populated place is compared with: TX
NY with TX
 comparing: Austin
Texas
NC for populated place is compared with: TX
NC with TX
 comparing: Austin
Texas
CO for populated place is compared with: TX
CO with TX
 comparing: Austin
Texas
RI for populated place is compared with: TX
RI with TX
 comparing: Austin
Texas
WA for populated place is compared with: TX
WA with TX
 comparing: Austin
Texas
MO for populated place is compared with: TX
MO with TX
 comparing: Austin
Texas
PA for populated place is compared with: TX
PA with TX
 comparing: Austin
Texas
TN for populated place is compared with: TX
TN with TX
 comparing: Austin
Texas
21 for populated place is compared with: TX
21 with TX
 comparing: Austin
Texas
UT for populated place is compared with: TX
UT with TX
 comparing: Austin
Texas
AR for populated place is compared with: TX
AR with TX
```

```
 comparing: Austin
Texas
MI for populated place is compared with: TX
MI with TX
 comparing: Austin
Texas
IN for populated place is compared with: TX
IN with TX
 comparing: Austin
Texas
CA for populated place is compared with: TX
CA with TX
 comparing: Austin
Texas
WA for populated place is compared with: TX
WA with TX
 comparing: Austin
Texas
TX for populated place is compared with: TX
TX with TX
 comparing Austin country code: US with Country@47f70bc1 admin1 code : US
 doNested method incoming parameters:
        PopulatedPlace@563e4d49 and Admin1@7b3d2b1c and [PopulatedPlace@14441fdb,
PopulatedPlace@31ce069f, PopulatedPlace@698c10b8, PopulatedPlace@1f5f5727,
PopulatedPlace@feb5de5, PopulatedPlace@87bd5a0, PopulatedPlace@4144e4e0,
PopulatedPlace@4d14ca44, PopulatedPlace@20579615, PopulatedPlace@7d5e8834,
PopulatedPlace@713a4e73, PopulatedPlace@48b30ae, PopulatedPlace@4876d42,
PopulatedPlace@6e821522, PopulatedPlace@25616d8d, PopulatedPlace@30e4f894,
PopulatedPlace@563e4d49, PopulatedPlace@120e4f9a, PopulatedPlace@5a34d004,
PopulatedPlace@1ccfa5c1, PopulatedPlace@57160a60, PopulatedPlace@2c52d188,
PopulatedPlace@20b0b3b0, PopulatedPlace@2af49a18, PopulatedPlace@6803514a,
PopulatedPlace@4e28f1d6, PopulatedPlace@6a9dd62a, PopulatedPlace@1e39a3dc,
PopulatedPlace@49404e39, PopulatedPlace@621d40b0] and [Admin1@7b3d2b1c]
30 size of the incoming parameter of alternatePPObj.
30 is the size of the alternate pps that needs to be compared.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 ONE with the same name as Austin is found.
 Uncertainty 2 is selected.Austin is nested in Texas inUS
 inside one country: Austin and Texas in United States
IL compared with TX
 comparing: IL with Estado de Texas
 comparing: NY with Estado de Texas
 comparing: NC with Estado de Texas
 comparing: CO with Estado de Texas
 comparing: RI with Estado de Texas
 comparing: WA with Estado de Texas
 comparing: MO with Estado de Texas
 comparing: PA with Estado de Texas
 comparing: TN with Estado de Texas
```

```
comparing: 21 with Estado de Texas
comparing: UT with Estado de Texas
comparing: AR with Estado de Texas
comparing: MI with Estado de Texas
comparing: IN with Estado de Texas
comparing: CA with Estado de Texas
comparing: WA with Estado de Texas
comparing: TX with Estado de Texas
comparing: MS with Estado de Texas
comparing: MO with Estado de Texas
comparing: 11 with Estado de Texas
comparing: MI with Estado de Texas
comparing: NV with Estado de Texas
comparing: OR with Estado de Texas
comparing: MT with Estado de Texas
comparing: MN with Estado de Texas
comparing: AR with Estado de Texas
comparing: OH with Estado de Texas
comparing: KY with Estado de Texas
comparing: 10 with Estado de Texas
comparing: AR with Estado de Texas
comparing: IL with Estado de la Estrella Solitaria
comparing: NY with Estado de la Estrella Solitaria
comparing: NC with Estado de la Estrella Solitaria
comparing: CO with Estado de la Estrella Solitaria
comparing: RI with Estado de la Estrella Solitaria
comparing: WA with Estado de la Estrella Solitaria
comparing: MO with Estado de la Estrella Solitaria
comparing: PA with Estado de la Estrella Solitaria
comparing: TN with Estado de la Estrella Solitaria
comparing: 21 with Estado de la Estrella Solitaria
comparing: UT with Estado de la Estrella Solitaria
comparing: AR with Estado de la Estrella Solitaria
comparing: MI with Estado de la Estrella Solitaria
comparing: IN with Estado de la Estrella Solitaria
comparing: CA with Estado de la Estrella Solitaria
comparing: WA with Estado de la Estrella Solitaria
comparing: TX with Estado de la Estrella Solitaria
comparing: MS with Estado de la Estrella Solitaria
comparing: MO with Estado de la Estrella Solitaria
comparing: 11 with Estado de la Estrella Solitaria
comparing: MI with Estado de la Estrella Solitaria
comparing: NV with Estado de la Estrella Solitaria
comparing: OR with Estado de la Estrella Solitaria
comparing: MT with Estado de la Estrella Solitaria
comparing: MN with Estado de la Estrella Solitaria
comparing: AR with Estado de la Estrella Solitaria
comparing: OH with Estado de la Estrella Solitaria
comparing: KY with Estado de la Estrella Solitaria
comparing: 10 with Estado de la Estrella Solitaria
comparing: AR with Estado de la Estrella Solitaria
comparing: IL with Lone Star State
comparing: NY with Lone Star State
comparing: NC with Lone Star State
comparing: CO with Lone Star State
comparing: RI with Lone Star State
comparing: WA with Lone Star State
comparing: MO with Lone Star State
comparing: PA with Lone Star State
comparing: TN with Lone Star State
comparing: 21 with Lone Star State
comparing: UT with Lone Star State
comparing: AR with Lone Star State
comparing: MI with Lone Star State
comparing: IN with Lone Star State
comparing: CA with Lone Star State
comparing: WA with Lone Star State
comparing: TX with Lone Star State
comparing: MS with Lone Star State
comparing: MO with Lone Star State
comparing: 11 with Lone Star State
comparing: MI with Lone Star State
comparing: NV with Lone Star State
comparing: OR with Lone Star State
comparing: MT with Lone Star State
comparing: MN with Lone Star State
comparing: AR with Lone Star State
```

```
 comparing: OH with Lone Star State
 comparing: KY with Lone Star State
 comparing: 10 with Lone Star State
 comparing: AR with Lone Star State
 comparing: IL with State of Texas
 comparing: NY with State of Texas
 comparing: NC with State of Texas
 comparing: CO with State of Texas
 comparing: RI with State of Texas
 comparing: WA with State of Texas
 comparing: MO with State of Texas
 comparing: PA with State of Texas
 comparing: TN with State of Texas
 comparing: 21 with State of Texas
 comparing: UT with State of Texas
 comparing: AR with State of Texas
 comparing: MI with State of Texas
 comparing: IN with State of Texas
 comparing: CA with State of Texas
 comparing: WA with State of Texas
 comparing: TX with State of Texas
 comparing: MS with State of Texas
 comparing: MO with State of Texas
 comparing: 11 with State of Texas
 comparing: MI with State of Texas
 comparing: NV with State of Texas
 comparing: OR with State of Texas
 comparing: MT with State of Texas
 comparing: MN with State of Texas
 comparing: AR with State of Texas
 comparing: OH with State of Texas
 comparing: KY with State of Texas
 comparing: 10 with State of Texas
 comparing: AR with State of Texas
 comparing: IL with TX
 comparing: NY with TX
 comparing: NC with TX
 comparing: CO with TX
 comparing: RI with TX
 comparing: WA with TX
 comparing: MO with TX
 comparing: PA with TX
 comparing: TN with TX
 comparing: 21 with TX
 comparing: UT with TX
 comparing: AR with TX
 comparing: MI with TX
 comparing: IN with TX
 comparing: CA with TX
 comparing: WA with TX
 comparing: TX with TX
30.26715,-97.74306for Austin in USand state: Texas
Austin in Texas in United States
333     austin, TX, US Austin  30.26715       -97.74306      2
```

Example 2: FHLI (Tehran, London, Uk)

```
Populated place saved in arrayList successfully.2647529
Countries read successfully.247
Admin1 read successfully.3763
old admin list read successfully: 0
 reading line: '333    tehran, London, UK'
 an object of type DisambigFHLI is created successfully
about to seperate the fhli by ,
 locations to be disambiguated: tehran
 locations to be disambiguated: London
 locations to be disambiguated: UK
 reading Location: tehran
 spliting: tehran
 size of the splitted array: 1
 reading Location:  London
 spliting:  London
 size of the splitted array: 2
 reading Location:  UK
 spliting:  UK
 size of the splitted array: 2
 locations to be disambiguated after resolving south, east, noerth, west3
tehran
London
UK
6 populated places are found with the name tehran
 ---------------------Content of admin1 list ----------------------:
Ostƒ?n-e Tehrƒ?n in IR
 -----------------Content of populated Place list--------------------
Tehrƒ?n in IR 35.69439 51.42151
Tehrƒ?n in IR 27.45861 55.47556
Tƒ´rƒ?n in IR 32.7026 51.1537
Tehrƒ?n in IR 35.6373 51.3516
Tehrƒ?n in IR 35.7261 51.3304
Tehrƒ?n in IR 35.75 51.5148
 alternatePP content after removing those with country name
Tehrƒ?n
Tehrƒ?n
Tehrƒ?n
Tehrƒ?n
Tehrƒ?n
Tƒ´rƒ?n
0 alternate countries have been found for tehran
35 populated places are found with the name London
 ---------------------Content of admin1 list ----------------------:
Ostƒ?n-e Tehrƒ?n in IR
 -----------------Content of populated Place list--------------------
London in BZ 17.98333 -88.43333
London in CA 42.98339 -81.23304
London in GB 51.50051 -0.12883
Vauxhall in GB 51.48582 -0.12205
Pimlico in GB 51.48897 -0.13699
Victoria in GB 51.49624 -0.14402
London in GQ 2.28333 9.8
London in KI 1.98333 -157.46667
Lonton in MM 25.1 96.28333
London in NG 5.72257 5.78787
London in PH 6.00972 125.12944
London in US 31.29767 -87.08775
London in US 32.2732 -86.05024
London in US 35.32897 -93.25296
London in US 39.6256 -85.92026
London in US 37.12898 -84.08326
London in US 39.88645 -83.44825
London in US 35.86982 -83.00209
London in US 30.67685 -99.57645
London in US 32.23099 -94.94438
London in US 38.19455 -81.36872
London in US 42.02004 -83.61327
London in US 43.52607 -93.0627
London in US 47.2027 -91.56962
London in US 40.445 -95.23498
London in US 40.91033 -82.62934
London in US 41.14367 -80.14867
London in US 43.04778 -89.01289
London in US 36.47606 -119.44318
London in US 43.63457 -123.09285
London in US 36.48091 -119.44401
London in VE 10.36389 -66.73333
London in ZA -24.81667 31.05
London in ZA -24.76667 30.86667
London in ZA -24.3 30.58333
 alternatePP content after removing those with country name
London
London
London
London
London
Victoria
Tehrƒ?n
London
London
Tehrƒ?n
London
Pimlico
London
Tehrƒ?n
London
```

```
London
London
London
Tehrƒ?n
London
London
London
London
Tehrƒ?n
London
London
Lonton
London
London
London
London
London
London
London
London
London
London
Vauxhall
Tƒ´rƒ?n
London
London
0 alternate countries have been found for London
2 populated places are found with the name UK
 ---------------------Content of admin1 list ----------------------:
Ostan-e Tehran in IR
 -----------------Content of populated Place list--------------------
Uk in RU 54.94722 57.26167
Uk in RU 55.07694 98.85389
 alternatePP content after removing those with country name
London
London
London
London
London
Victoria
Tehrƒ?n
London
London
Tehrƒ?n
London
Pimlico
London
Tehrƒ?n
London
London
London
London
Tehrƒ?n
London
London
London
London
Tehrƒ?n
London
London
Lonton
London
London
London
London
London
London
London
London
London
London
Vauxhall
Tƒ´rƒ?n
London
London
1 alternate countries have been found for UK
.................... contents of the alternate countries:
```

```
United Kingdom
 number of countries mentioned in FHLI is: 1
 comparing: London
Ostƒ?n-e Tehrƒ?n
KY for populated place is compared with: 26
KY with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
OR for populated place is compared with: 26
OR with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
09 for populated place is compared with: 26
09 with 26
 comparing: Tehrƒ?n
Ostƒ?n-e Tehrƒ?n
26 for populated place is compared with: 26
26 with 26
 comparing Tehrƒ?n country code: IR with Country@23610f1f admin1 code : GB
PopulatedPlace@64650ddb and PopulatedPlace@64650ddb testing namesTehrƒ?n with Tehrƒ?n
 comparing: Victoria
Ostƒ?n-e Tehrƒ?n
ENG for populated place is compared with: 26
ENG with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
MN for populated place is compared with: 26
MN with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
OH for populated place is compared with: 26
OH with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
IN for populated place is compared with: 26
IN with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
WV for populated place is compared with: 26
WV with 26
 comparing: Tehrƒ?n
Ostƒ?n-e Tehrƒ?n
11 for populated place is compared with: 26
11 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
AR for populated place is compared with: 26
AR with 26
 comparing: Pimlico
Ostƒ?n-e Tehrƒ?n
ENG for populated place is compared with: 26
ENG with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
ENG for populated place is compared with: 26
ENG with 26
 comparing: Tehrƒ?n
Ostƒ?n-e Tehrƒ?n
26 for populated place is compared with: 26
26 with 26
 comparing Tehrƒ?n country code: IR with Country@23610f1f admin1 code : GB
PopulatedPlace@4d560eeb and PopulatedPlace@4d560eeb testing namesTehrƒ?n with Tehrƒ?n
 comparing: London
Ostƒ?n-e Tehrƒ?n
MO for populated place is compared with: 26
MO with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
OH for populated place is compared with: 26
OH with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
36 for populated place is compared with: 26
36 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
TX for populated place is compared with: 26
TX with 26
 comparing: Tehrƒ?n
Ostƒ?n-e Tehrƒ?n
26 for populated place is compared with: 26
26 with 26
 comparing Tehrƒ?n country code: IR with Country@23610f1f admin1 code : GB
PopulatedPlace@2670d85b and PopulatedPlace@2670d85b testing namesTehrƒ?n with Tehrƒ?n
 comparing: London
Ostƒ?n-e Tehrƒ?n
15 for populated place is compared with: 26
15 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
C6 for populated place is compared with: 26
C6 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
04 for populated place is compared with: 26
04 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
CA for populated place is compared with: 26
CA with 26
 comparing: Tehrƒ?n
Ostƒ?n-e Tehrƒ?n
26 for populated place is compared with: 26
26 with 26
 comparing Tehrƒ?n country code: IR with Country@23610f1f admin1 code : GB
PopulatedPlace@3a8c5214 and PopulatedPlace@3a8c5214 testing namesTehrƒ?n with Tehrƒ?n
 comparing: London
Ostƒ?n-e Tehrƒ?n
00 for populated place is compared with: 26
00 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
WI for populated place is compared with: 26
WI with 26
 comparing: Lonton
Ostƒ?n-e Tehrƒ?n
04 for populated place is compared with: 26
04 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
TX for populated place is compared with: 26
TX with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
AL for populated place is compared with: 26
AL with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
PA for populated place is compared with: 26
PA with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
CA for populated place is compared with: 26
CA with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
08 for populated place is compared with: 26
08 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
MN for populated place is compared with: 26
MN with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
07 for populated place is compared with: 26
07 with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
AL for populated place is compared with: 26
AL with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
07 for populated place is compared with: 26
07 with 26
```

```
 comparing: London
Ostƒ?n-e Tehrƒ?n
00 for populated place is compared with: 26
00 with 26
 comparing: Tƒ´rƒ?n
Ostƒ?n-e Tehrƒ?n
28 for populated place is compared with: 26
28 with 26
 comparing: Vauxhall
Ostƒ?n-e Tehrƒ?n
ENG for populated place is compared with: 26
ENG with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
TN for populated place is compared with: 26
TN with 26
 comparing: London
Ostƒ?n-e Tehrƒ?n
MI for populated place is compared with: 26
MI with 26
        US with ........GB
        US with ........GB
        ZA with ........GB
        IR with ........GB
        GB with ........GB
        US with ........GB
        US with ........GB
        US with ........GB
        US with ........GB
        IR with ........GB
        US with ........GB
        GB with ........GB
        GB with ........GB
        IR with ........GB
        US with ........GB
        US with ........GB
        NG with ........GB
        US with ........GB
        IR with ........GB
        VE with ........GB
        PH with ........GB
        BZ with ........GB
        US with ........GB
        IR with ........GB
        GQ with ........GB
        US with ........GB
        MM with ........GB
        US with ........GB
        US with ........GB
        US with ........GB
        US with ........GB
        CA with ........GB
        US with ........GB
        ZA with ........GB
        US with ........GB
        ZA with ........GB
        KI with ........GB
        IR with ........GB
        GB with ........GB
        US with ........GB
        US with ........GB
 highest place is: London with 1 inGB
 adding: Victoria 7 GB
 adding: Pimlico 8 GB
 adding: London 1 GB
 adding: Vauxhall 11 GB
London encloses all with the same names in the same country.
         alternatePP content..................................
London
London
London
London
London
Victoria
Tehrƒ?n
London
London
```

```
Tehrƒ?n
London
Pimlico
London
Tehrƒ?n
London
London
London
London
Tehrƒ?n
London
London
London
London
Tehrƒ?n
London
London
Lonton
London
London
London
London
London
London
London
London
London
London
Vauxhall
Tƒ´rƒ?n
London
London
        Adding to the notNestedTehrƒ?n in IR
        Adding to the notNestedTehrƒ?n in IR
        Adding to the notNestedTehrƒ?n in IR
        Adding to the notNestedTehrƒ?n in IR
        Adding to the notNestedTehrƒ?n in IR
        Adding to the notNestedTƒ´rƒ?n in IR
Tehrƒ?n inIR is not the same as the nested ones.
Tehrƒ?n inIR is not the same as the nested ones.
Tehrƒ?n inIR is not the same as the nested ones.
Tehrƒ?n inIR is not the same as the nested ones.
Tehrƒ?n inIR is not the same as the nested ones.
Tƒ´rƒ?n inIR is not the same as the nested ones.
uncertainty 4 is selected. there are places in different countries.
 highest place is: London with 1
 there are the followings within the same country:
Victoria with7
London with1
 highest places are:
London with 1
51.50051,-0.12883 for London although there are other places mentioned in other countries.
333     tehran, London, UK    London  51.50051       -0.12883       4
```

Appendix 4

This appendix includes the source codes of the data collection application developed for collecting Flickr data for the pilot study (chapter 4, section 4.3.1). The aforementioned codes can be found on the attached CD.

Appendix 5

This appendix includes the source codes of the data collection application developed and used for collecting photos uploaded within boundary of GB (chapter 4: section 4.3.2). The aforementioned codes can be found on the attached CD.

Appendix 6

This appendix includes the source code of the disambiguation application developed and used for disambiguating the FHLI (chapter 5: section 5.2-3). The aforementioned codes can be found on the attached CD.

Appendix 7

This appendix includes the source code of the visualization application developed and used for visual analysis of the data (chapter 5: section 5.4 and chapter 7: section 7.2). The Aforementioned codes can be found on the attached CD.

Appendix 8

Twenty four LBS sites and blogs

A selection of recent LBS sites I've visited.

Useful Networks
M Spatial
Cloud Made
Zodigo
Yotta
230 miles of love
POI Friend
Walking Hotspot
London Calling
My Loki
Jentro
Upingme
Gypsii
Enkin
Seero
Socialight
Lightpole
Brightkite
Zkout
Wildknowledge
Whrrl
Ipoki
Placebase
TurfTag