Analysis of interaction and co-editing patterns amongst OpenStreetMap contributors

Peter Mooney and Padraig Corcoran

June 9, 2013

Abstract

OpenStreetMap (OSM) is a very well known and popular Volunteered Geographic Information (VGI) project on the Internet. In January 2013 OSM gained it's one millionth registered member. Several studies have shown that only a small percentage of these registered members carry out the large majority of the mapping and map editing work. In this paper we discuss results from a social-network based analysis of seven major cities in OSM in an effort to understand if there is quantitative evidence of interaction and collaboration between OSM members in these areas. Are OSM contributors working on their own to build OSM databases in these cities or is there evidence of collaboration between OSM contributors. We find that in many cases high frequent contributors ("senior mappers") perform very large amounts of mapping work on their own but do interact (edit/update) contributions from lower frequency contributors.

1 Introduction

OpenStreetMap (OSM) is currently one of the most popular Volunteered Geographic Information (VGI) projects on the Internet. VGI, the term coined by Goodchild [16, 17], is the collection of spatial data captured by "citizen sensors" where this data is then edited and managed within a collaborative web environment. OSM has global coverage, is multilingual, is constantly changing and updating, and contains spatial data and attribution representing almost every conceivable geographical feature [26]. It is edited in a collaborative web environment using several different types of web-based and desktop editor software tools. The successful progress of OSM since it's emergence in 2004 is very often attributed to it's status as a "crowdsourced" VGI project. Dodge and Kitchin [11] define crowdsourcing as the "collective generation of media, ideas, and data undertaken voluntarily by many people. The crowd metaphor signifies the power that can emerge from a mass of individuals converging to tackle a set of tasks". This "crowd" in OSM has grown significantly in recent years and in January 2013 the OSM project gained it's one millionth member [32]. Recently research works such as: Haklay et al. [19], Mooney and Corcoran [26], Neis and Zipf [30], and Budhathoki and Haythornthwaite [5] have shown that the idea that there is a large crowd of contributors working together to gather geographic data and build the OSM database is inaccurate. These authors have shown that it is actually a small percentage (in most cases not more than 10%) of these OSM members who produce and maintain the vast majority of the geographic data and associated metadata in the global OSM database. This is not a trivial task. The volume of data currently in OSM is significant. In February 2013 (from OpenStreetMap [31]) there were over three billion GPS track points and approximately 170 million ways (polygons or polylines).

In this paper we use seven major cities for a case-study: London, Berlin, Munich, Prague, Minneapolis, Paris, and Sydney. We extracted the entire editing and contribution history for these seven locations from the OpenStreetMap History database [28]. We generate statistics and supporting information to give an overview of the OSM database in each location. We then extract the contribution histories of the most frequent contributors to OSM in each city respectively. These contribution histories are then used to build a co-edit network model using a graph theoretic approach from the editing behaviour of OSM contributors. Our paper aims to answer the following research questions. From previous work we know that a very small percentage of OSM contributors perform the majority of the data gathering and editing work. We shall verify this for our case-study cities and then apply social network analysis techniques to measure and quantify the linkages between these high frequency contributors and the remainder of the OSM community in a given city. We will also give an interpretation, from our OSM experience, as to what our results from this social network analysis means and what additional insight it provides us with about the OSM community using a quantitative only approach.

The remainder of the paper is organised as follows. In section 2 we discuss the most relevant literature in the areas of OSM, VGI, and the contributors to VGI projects. We pay particular attention to research work investigating high frequency contributors to VGI and collaborative projects in section 2.2. In section 3 we discuss the experimental setup for our analysis and provide an overview of the OSM characteristics of the case study cities. This includes a discussion of the growth of the OSM databases in these cities and some descriptive statistics about the OSM contributors in these cities. Section 4 describes the development of a social network model for OSM contributors and their contributions in the case study cities. In particular section 4.1 introduces the concept of co-editing which forms the basis of our co-edit network graph model. Section 5 then provides analysis of some key social network measures which help us to quantitatively understand the properies of our co-edit network graph model of OSM contributions. Finally, in section 6 we close the paper with a discussion of the key results and research contributions of this paper. We also provide some ideas for immediate and longer term future work.

2 Related Research Work

In this section we shall describe some of the most closely related existing research. We will not be discussing the OpenStreetMap data model in detail. For a more detailed overview of the OpenStreetMap data model, tools for contribution, etc. the reader is directed to papers with excellent overviews including: Neis et al. [28], Neis and Zipf [30]. There is a steadily growing body of peer-reviewed research literature appearing which performs analysis of various aspects of OpenStreetMap or uses OpenStreetMap as a source of case-study data. In our literature review here we concentrate on research work which analyses the behaviour of the contributors to OpenStreetMap.

2.1 The growth of OpenStreetMap

OpenStreeMap has been successful in the reasonably short space of time since it's emergence in 2004. What has motivated it's contributors to work hard at collecting and managing geospatial data to support and develop OpenStreetMap? In the work of Budhathoki and Haythornthwaite [5], and the same author's earlier work in Budhathoki et al. [4], empirical studies of the motivations to contribute to OSM are performed. The potential motivation factors identified in Budhathoki et al. [4] include both *intrinsic* (self expression, altruism, learning, etc.) and *extrinsic* (social relations, community, networking, etc) factors. Over 400 OSM members contributed to this study, by survey, and Budhathoki and Haythornthwaite classify OSM contributors into "serious" and "casual" mappers. "Serious" mappers are those who have contributed several thousand nodes of geographic data to the OSM database. They find that "serious" mappers are more "directed to the OSM community" than casual mappers. They also found that for "serious mappers" the motivation to fix errors on the map of their local area is a particularly strong motivating factor. This mix of "serious" and "casual" mappers sees OpenStreetMap grow in a non-linear and sometimes unpredictable fashion. Neis et al. [29] outlines the development of the OpenStreetMap project from 2007 to 2011 in Germany by specifically considering the expansion of the total street network and the route network for car navigation. The authors predict that by 2013 OpenStreetMap for Germany should provide a route network for cars which is comparable to that provided by commercial companies such as TomTom. Corcoran and Mooney [7] attempt to characterise the metric and topological evolution of OpenStreetMap network representations by applying a graph theoretic approach. They use the historical databases for three city locations in Ireland. They found no uniform editing behaviour exhibited across the three regions. They conclude that despite the apparent lack of coordination of mapping effort their results for street density and coverage suggests that contributors initially map roads/streets predominantly of a greater length or prominence before subsequently mapping streets predominantly of a shorter length. In their related work Corcoran et al. [8] speculates that the manner in which networks in general grow could explain why contribution patterns in OSM are difficult to model. Many different networks (for example road, rail, etc) exhibit small-world and/or scale-free properties and contributors in OSM may tend to *grow* their OSM network representations in this fashion.

2.2 The influence of high frequency contributors

This type of growth and evolution of OpenStreetMap mentioned above relies on the interaction and collaboration of the contributors to the project. One of the exciting aspects of crowdsourcing is the possibility of hundreds and thousands of users generating and managing data or information to solve some problem or further the goals of a project [10]. However, many authors have chosen to concentrate on a smaller percentage of this crowd. Usually this subset of the crowd is compromised of the users or contributors who generate the most data or information, are involved in a certain crowdsourcing project for the longest time, or who display certain attributes which make them "stand out from the crowd". In OpenStreeMap anonymous changes to the database are no longer supported and any Internet user who registers for the project can add information to the map and change existing data. While this open approach to collaborative data collection creates questions about the quality of the spatial data [30] it provides researchers with an opportunity to analyse these users more closely. Neis and Zipf [30] perform analysis of the number of node contributions of all OSM members. The results gathered, in 2012 (when membership was approximately $700,000$), showed that there are approximately 5% of all the registered members, referred to as "Senior Mappers", who have contributed 1000 nodes or more. About 14% of the total number of members, created at least 10 and fewer than 1,000 Nodes, and these members may be referred to as "Junior Mappers". Crucially, 19% of members created less than 10 Nodes, which makes them the least active, but also the largest member group. Members falling into this class are referred to as "Nonrecurring Mappers". The largest group without any action in the OSM project is represented by 62%. These members have simply signed up to OSM but never performed editing of *any* kind. Rehrl et al. [33] proposes a conceptual model as a foundation for a uniform and standardized process for analysing user contributions in OpenStreetMap which could be extended to other types of VGI. They detailed a proposed action set for describing VGI contribution tasks which carefully lists the different types: create, edit, update, and delete actions that can be performed.

In other collaborative projects and social media there are similar patterns found. Achananuparp et al. [1] propose a novel framework to model information propagation behaviour of Twitter users. Specifically they introduce two propagation behaviours, namely originating and promoting behaviours and focus their analysis and model development on the "top 1000 Twitter users in Singapore". Liu and Ram [23] remark that crowdsourced projects such as Wikipedia make it easy to edit content but this does not mean that all contributors edit the same way, or with the same intensity. In a single Wikipedia edit, a contributor can insert a number of sentences or just change a single word. Collaboration on Wikipedia is not represented by a group of contributors making homogeneous contributions. Just as Neis and Zipf [30] found for OSM, in their statistical

Feature	**Berlin**	**London**	**Minn**	**Munich**	**Paris**	**Prague**	**Sydney**
No. POI	43521	49404	20397	22434	15033	6475	6548
No. Ways	362925	362720	168252	189146	317928	159431	85268
No. Relations	7466	11375	928	3780	13823	2039	2195
POI $1v$(%)	38	60	79	20	59	58	43
POI $\leq 5v$(%)	92	98	100	82	97	98	99
Ways $1v$(%)	57	56	74	56	55	74	45
Ways $\leq 5v$(%)	90	95	97	90	97	96	89
Rels $1v$(%)	48	20	58	46	86	59	44
Rels $\leq 5v$(%)	83	93	91	80	97	85	86

Table 1: Characteristics of the OSM databases for the 7 cities chosen for this study

analysis of Wikipedia Liu and Ram [23] found that contributors to Wikipedia could be clustered into six classes: All round contributors, watchdogs, starters, content justifiers, copy-editors, and cleaners. On $1,600$ featured articles 82.74% of contributors had less than 4 actions for a given article. Liu and Ram [23] call these "casual contributors" and do not include contributors with these characteristics in any futher analysis. Liu and Ram [23] concluded that articles developed using patterns where all-round editors played a dominant role are often of high quality, while patterns where starters and casual contributors dominate are often associated with low quality. Singh [36] reports on a study of the impact of community-level networksrelationships that exist among developers in an Open Source Software (OSS) community and then on the productivity of member developers. Singh finds that the OSS community networks are characterized by small-world properties that positively influence the productivity of the member developers by providing them with speedy and reliable access to more quantity and variety of information and knowledge resources and connectivity with other influential member developers. As in the case of the other studies of OSM and Wikipedia, Singh finds that the number of these influential member developers is rather small.

This is true for OSM also. Mooney and Corcoran [26] investigate heavily edited objects in OSM and find that 87% of contributions/edits to these objects are performed by 11% of the total 4128 contributors in their case-study. In 79% of these edits additional spatial data (nodes) are added to objects with the remainder of edits related to changing or updating tagging information. Lin [22] derived empirical data from interviewing a small number of OSM contributors. She draws the conclusion from these interviews that OSM "itself acts as a boundary object that enables actors from different social worlds to co-produce the OSM Map through interacting with each other and negotiating the meanings of mapping, the mapping data and the Map itself". In an application of VGI Comber et al. [6] evaluated the quality of land cover information provided by volunteers through the incorporation of a set of control locations where the land cover was known. The dataset contained $42,474$ records after filtering down to a smaller set of 47 volunteers who contributed more than 20 validation points and for whom robust reliabilities could be calculated. In the next section we shall outline the experimental setup for our analysis and some characteristics of the seven case-study cities in OpenStreetMap we have used.

3 Experimental Setup and Case-study area characteristics

In the OpenStreetMap database there are three primitive data types/objects: nodes (points), ways (polygons and polylines), and relations (logical groupings of nodes and ways). All of these objects can be annotated with tags which are key-value pairs where both the key and value are free format text fields, although in practice there are agreed conventions regarding how tags are used for most common purposes. The datasets for the 7 cities contain all of their edit histories from as far back as 2005 until April 2012. This was the last stable history release from OSM until late 2012. There have been changes made by redaction bots associated with the OSM License change in late 2012. This may cause some slight changes in the results associated with areas such as Sydney. However we feel that our time period of 2005 to April 2012 is adequate for the purposes of this study. The seven cities were extracted from the OSM history file which is available for download from `http://wiki.openstreetmap.org/wiki/Planet.osm/full`. These seven cities were chosen to ensure a good overview of the types of scenarios encountered in OSM including: cities which have had bulk import of free geodata (Paris and Minneapolis), cities which have had little or no bulk import (London, Berlin), and then finally cities which have very active OSM communities and have a mix of bulk import and manually collected spatial data (Prague, Munich, and Sydney). Data for each city was extracted using the administrative boundary relation/polygon for each city in OpenStreetMap. For consistency ways which overlapped these boundaries were cut to the boundary. The next section discusses the growth of the OSM database in each of these cities.

3.1 OSM database growth in the case-study cities

In this subsection we provide an overview of the growth of the OSM database for each of our case-study cities. This provides us with an opportunity to understand the key stages in the evolution of OSM in these cities and also the patterns of contribution from the OSM contributors over the past few years. Table 1 provides a tabulated summary of the characteristics of the geographic data stored in the OSM database for the seven cities. This includes a total count of all edit versions of all features or objects for each city. Rather than provide a count of all nodes in the OSM databases we have highlighted the number of Points of Interest (POI) for each city. There were some issues before 2011 in how node versions (within objects) were increased by certain OSM editor software [30]. To avoid any inaccuracies with node versions in our study we consider POI as an OSM node with a "name" attribute and is usually a single node object [27]. There are also POI in OSM which do not have a "name" attribute such as traffic signals, speed controls, etc. However for the purposes of our study we concentrate on POI with "name" tag attributes. POI are one of the easiest and most popular objects for mapping in OSM particularly amongst new and less experienced mappers. Usually after starting to map POI new contributors to the OSM project will move onto mapping ways. Minneapolis has the lowest number of relations of the seven cities. Relations are considered a complex relationship in OSM. They logically group collections of ways and nodes (including POI) together. A relation may represent a train station or an airport, for example. We can see that for all of our cities there are a small number of 'high edit' relations which have more than 5 versions of editing. In London, Prague, and Sydney there are 17, 15, and 14 percent of relations with more than 5 versions of editing. Mooney and Corcoran [26] studied objects (which they called "heavily edited objects") which are subject to a high number of revisions or versions but these usually only account for around 12% of all objects in a given city or region. Prague and Sydney both clearly show the presence of "heavily edited objects" in particular related to relations where almost 15% of these objects have greater than 5 versions.

It is also interesting to take a longitudinal view of how these cities developed in OpenStreetMap over the entire span of their edit histories. Figure 1 shows a timeseries plot of the total number of POI, Ways, and Relations created in Minneapolis and the number of subsequent edits to POI, Ways, and Relations in Minneapolis. An import of the 2005 TIGER/Line data was completed in 2007 which is represented by the very high 'create' spikes. There have been several subsequent efforts in recent years to perform a 'Tiger Fixup' to repair a wide range of data quality problems caused by the 2007 import. There is a similiar scenario for Paris in Figure 2 where there are very dramatic 'create' and 'edit' spikes in 2010. There was a bulk import of geographic data into the French OSM database during this time. The subsequent spike in 'edit' behaviour occurring in 2011 was the result of an effort by the French OSM community

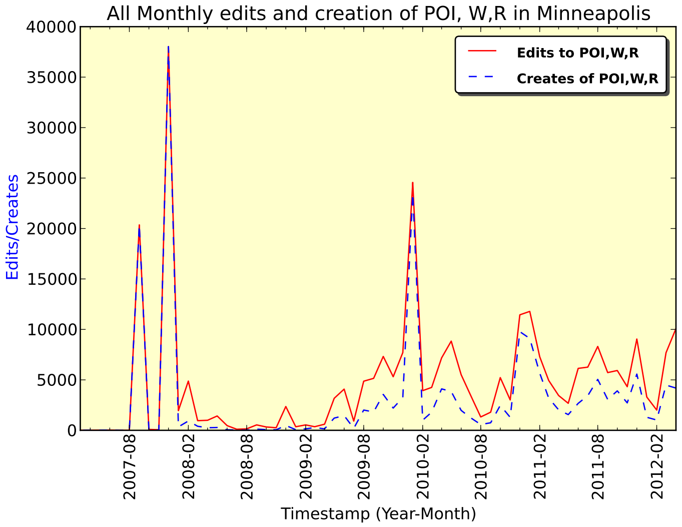

Figure 1: This plot shows the timeseries of the monthly totals for all OpenStreetMap objects created and edited in Minneapolis. An import of the 2005 TIGER/Line data was completed in 2007 which is represented by the very high 'create' spikes. There have been several subsequent efforts in recent years to perform a 'Tiger Fixup' to repair a wide range of data quality problems caused by the 2007 import.

to fix some of the data quality issues caused by the bulk import of 2010. The same timeseries is shown for London in Figure 3. For the most part the London OSM database has not been subject to very large scale bulk import of freely available geographic data as has been the case in Figure 2 for Paris or Figure 1 for Minneapolis. However, there are still several severe spikes in 'edit' behaviour and one significant spike in 'create' behaviour. These can all be traced back to significant events in OSM in London such as well coordinated mapping parties. Similar timeseries are seen for the other cities in this case-study. The key message from these timeseries is that mapping behaviour in these cities is not a process which is easily quantified. Events such as bulk import of geographic data or mapping parties can have a significant effect on the OSM database causing a sudden large number of edits or creation of new geographic objects.

3.2 Contributor Characteristics in the case-study cities

In the previous section we commented on the characteristics of the geographic data objects in the OSM databases for the seven case-study cities. In this section we shall summarise the characteristics of the contributors to the OSM databases in the case-study cities in order to begin understanding the structure of the contributor community. Table 2 summarises the contribution frequencies of all contributors in all seven cities. As outlined in the literature review in section 2 several authors [5, 23, 26, 30] have shown that a very large percentage of work in crowdsourced projects such as VGI is performed by a small percentage (around 10%) of participants. In our summary here we will specifically highlight the work of the top 10% contributors to OpenStreetMap as ranked on their total number of contributions. These top 10% contributors are those within the boundaries of the selected cities and are not necessarily amongst the top 10% contributors globally in OpenStreetMap. Where applicable we removed the edits performed by automated bot agents operating on the OpenStreetMap data in each city. In all cases these bots are easily detected and removed from our analysis. In the case of the top 1 contributor to Paris and Minneapolis these contributors were responsible for a bulk import of data to OpenStreetMap.

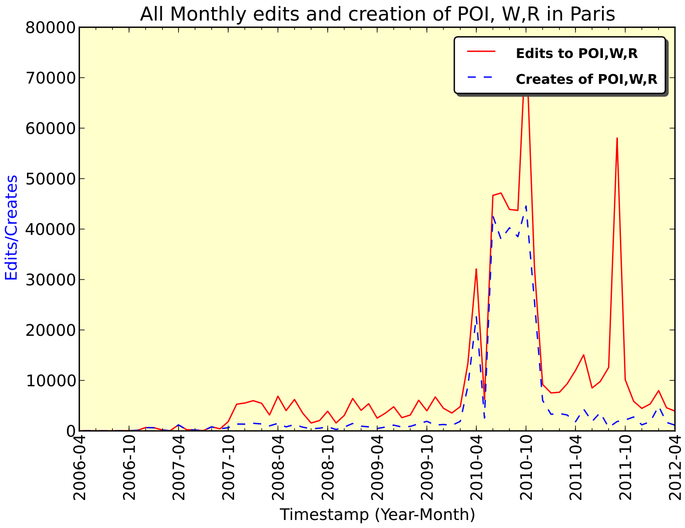

Figure 2: This plot shows the timeseries of the monthly totals for all OpenStreetMap objects created and edited in Paris. The major spikes in creation and editing of objects in 2010 and editing in 2011 can be directly linked to bulk import of national scale geographic data (land parcels, housing, etc.). Significant editing took place in 2011 to try to improve some of the data quality issues inherent in the bulk import in 2010.

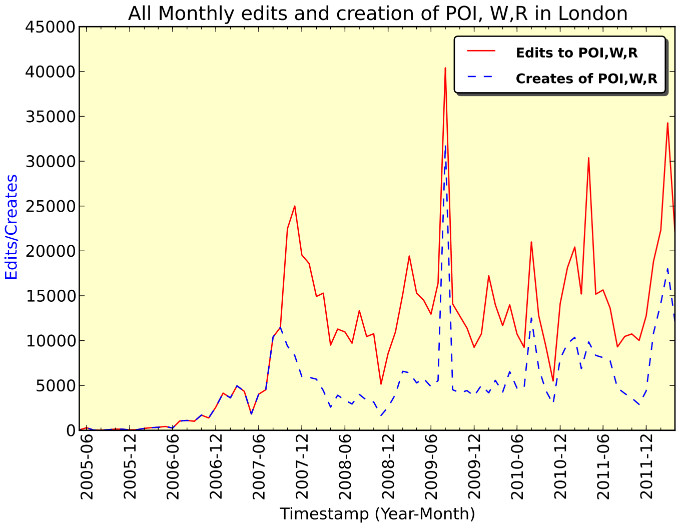

Figure 3: This plot shows a timeseries of the monthly totals for all OpenStreetMap objects created and edited in London. The spikes visible at 2007, 2009, 2011, 2012 can be directly attributed to increased participant in OpenStreetMap events such as mapping parties. There is little or no bulk import for London in OpenStreetMap

Feature	**Berlin**	**London**	**Minn**	**Munich**	**Paris**	**Prague**	**Sydney**
City Population	$3,501,872$	$8,278,251$	$385,378$	$1,378,176$	$2,211,297$	$1,257,158$	$4,575,532$
Area (km^2)	885	$1,510$	207	307	560	495	460
Contributors	5849	3928	589	3235	2403	1414	800
Senior [30]	409(7%)	236(6%)	77(13%)	292(9%)	120(5%)	113(8%)	56(7%)
Junior [30]	2749(47%)	1885(48%)	295(50%)	1423(44%)	1009(42%)	537(38%)	376(47%)
Non-recur [30]	2691(46%)	1807(46%)	218(37%)	1519(47%)	1274(53%)	764(54%)	368(46%)
Total Creates	$413,912$	$423,499$	$189,577$	$215,360$	$346,784$	$167,945$	$94,011$
Total Edits	$667,582$	$467,810$	$103,740$	$365,485$	$317,917$	$121,690$	$160,010$
Top 10% Create	96%	97%	97%	95%	98%	98%	96%
Top 10% Edit	94%	94%	91%	93%	96%	94%	96%
Top 5% Create	91%	94%	94%	88%	97%	96%	92%
Top 5% Edit	88%	89%	82%	86%	93%	90%	92%
Top 1% Create	68%	75%	74%	55%	89%	86%	67%
Top 1% Edit	67%	64%	34%	58%	79%	70%	68%
TOP 1: Create	13%	7%	31%	8%	22%	34%	29%
TOP 1: Edit	7%	4%	0%	9%	12%	28%	20%

Table 2: This table summarises the contribution activity of all contributions to OpenStreetMap in seven major cities. Population data is taken from the United Nations [38] World Demographic Yearbook. The area (km^2) of the region considered is calculated directly from the relations and polygons used from OpenStreetMap to determine the administrative boundaries of the selected cities

In the analysis by Neis and Zipf [30] the authors analyse the registrations to OSM and then check how many edits each registered contributor performed. In our analysis we are only concerned with registered users who have performed at least one edit to OSM. Berlin and London have the largest number of contributors with 5849 and 3928 respectively. The table summarises the editing and creation frequencies for the Top 10%, Top 5% and Top 1 of contributors for all of the cities. Neis and Zipf [30] classify OSM contributors as “Senior Mappers”, “Junior Mappers”, “Non-recurring Mappers”. Table 2 shows the distribution of these categories of contributors. The ‘non-recurring’ mappers/contributors identified by Neis and Zipf [30] are those who make a very small number of contributions and then effectively disappear from the project. These ‘non-recurring’ mappers/contributors make up a very significant percentage of all contributors in the seven cities studied. With the exception of Minnesota all cities have over 46% of non recurring mappers. The top 10% in all cities create 95% or more of all OSM objects and edit at least 91% of these objects. The main conclusion which can be drawn from the information in Table 2 is that those contributors in the top 10% of all contributors to OSM in these cities have carried out a very substantial body of voluntary work. We also find that the Neis and Zipf [30] classification is very consistent across all cities with the possible exception again of Minneapolis. In the next section we attempt to further classify the work performed by these top 10% by investigating their object creation and editing operations more closely.

3.3 Mapping behaviour of the top 10% of contributors

As we discussed in the previous section the top 10% of contributors perform over 90% of all object creation and object editing in OpenStreetMap in our case-study cities. In this section we will investigate the mapping behaviour of these subset of contributors more closely to investigate if we can understand what the most commonly occuring mapping task being carried out in each city is. Are these contributors now in a “map maintenance” phase of OSM in a given city? Is

City	Creators	Geometry	Tags & Geometry	Tags	Unclassified	Total
London (392)	36	85	140∗	109	22	392
Berlin (584)	38	159	60	300∗	27	584
Minneapolis (58)	4	10	4	38∗	2	58
Munich (323)	19	99	72	117∗	16	323
Paris (240)	17	83	21	115∗	4	240
Prague (141)	16	54∗	25	38	8	141
Sydney (80)	9	17	17	34∗	3	80

Table 3: This table summarises the individual mapping behaviour of Top 10% ('Senior Mappers') for each city by clustering their create and edit actions into 5 classes. The number of actual contributors selected is provided in the first column

there continued editing work on the geometry of objects by these contributors? As mentioned above, Liu and Ram [23], Yasseri et al. [40] and others have used a similar approach to study high frequency contributors in other collaborative projects such as Wikipedia. We performed $k-$means data clustering on the contribution history of each of the top 10% contributors in all cities in an attempt to ascertain if contributors can be classified based on the types of edit interactions they perform. Using a multivariate approach we selected $k = 4$ clusters and four variables for each contributor's history: c_1 the percentage of objects they created but did not edit any further (table column: 'creators'), c_2 the percentage of geometry only updates or edits (table column: 'Geometry'), c_3 percentage of tag only updates or edits on objects (table column: 'Tags'), and c_4 percentage of geometry and tagging updates committed during the same contribution edit or version (table column: 'Tags & Geometry'). These four variables were then used to form a contribution vector (c_1, c_2, c_3, c_4) which is provided as input to the clustering algorithm. Cluster centroids for $k-$means were chosen manually after a visual analysis of the input contribution vectors. The centroids for variable i assigned 80% to c_i and 10% to c_j where $i \neq j$.

These classifications are a more higher level than the more extensive set proposed by Rehrl et al. [33]. In their work a geometry operation could correspond to several different actions: create, split, merge, move, etc. Applying nearest neighbour cluster centroid selection and naive Bayes classification to our data we calculated that contributors formed clusters (which were then verified manually). The results are outlined in table 3. For each city we place an aestericks beside the mapping behaviour which forms the largest cluster. In London the largest cluster represents "geometry and tagging" actions indicating that 140 of the top 10% of contributors in London performed edits on objects which edited both the object geometry and the associated tags. For Munich and Prague there are slightly more "geometry only" contributors whilst all of the other cities have predominantly tagging ("Tags") as their predominant mapping behaviour. Editing or adding tags to objects in OSM is technically one of the simplest operations which contributors can perform as there is very good support in all of the software and web-based editors for this edit action. No special technical skills are required to perform tagging. However it must be stated that it is crucially important the tag editing is performed in alignment with the community agreed ontology in OSM [2, 3] despite it's technical simplicity. In-depth local knowledge is also crucial for tagging. Importantly this table shows us that the top 10% of contributors in all cities are not predominantly creating geographic objects (the "creator" column) but rather are deeply involved in editing geometries of existing objects and editing tag attributes of those objects or what Liu and Ram [23] refers to in Wikipedia as "all-rounders".

In this section we have given an indepth overview of the characteristics of the geographic objects in the OSM databases for our seven case-study cities (section 3.1. We also discussed the characteristics of the contributors to OSM in these cities (section 3.2) and the types of contribution behaviour which are most commonly encountered (section 3.3). In the next section we will introduce the development of a social network model for contribution behaviour to OSM and then apply it to the seven cities in our case study.

4 Co-edit Network Development

In this section we will use the edit history in OSM to build a co-edit network model of the contribution/editing behaviour of the top 10% of contributors in our case-study cities. To develop a co-edit network model we must quantify both the actors/nodes (contributors) in the network and then the interactions (edges) between these actors. It is often straightforward to define the network's nodes whereas defining the edges can be more challenging and requires additional computation or data mining [34]. In OSM there are thriving consultation and collaborative discussions on Wikis and mailing lists [4, 5] and at "mapping parties" [11] but these are not very easily quantifiable. In OSM there are no explicit mechanisms for contributors/registered members to 'follow' or 'like' each other as in the case of social networking applications such as Twitter or Facebook. There is a concept of 'friendship' amongst contributors which is outlined on personal information wiki pages which members can setup on the OSM Wiki site. However, this information is also difficult to harvest programmatically. In addition to this, after a visual inspection, we concluded that there are only a very small number of registered members using this feature anyway. Consequently, we have developed our own concept called "co-edits" as a means of generating our own synthetic model of collaboration amongst contributors. Extracting social interaction, from datasets where social interaction or attribute information is not explicitly stored, has been attempted in many other areas. In machine vision Cristani et al. [9] attempted to detect social interaction from photographic images. In web forums Gómez et al. [15] analyzed the structure and evolution of discussion cascades on discussion websites such as Slashdot as a means of extracting social interaction. Building a social network data structure involves the analysis of the attribute information of the actors involved, logs of interaction between actors, explicit links between actors, or collaboration outputs (such as articles or software libraries) Massa [25] warns that the collection procedure for data about a social network, and the collection assumptions, highly influence the collected network and hence the findings that can be inferred from it. We will now proceed to explain the concept of "co-edits" in the next section.

4.1 Co-editing explained

To develop a social interaction model for OpenStreetMap which can be generated quantitatively and automatically from the edit history it is necessary to define how the contributors in the OpenStreetMap network are related. This in turn allows us to build the network data structures necessary for our analysis. Actors in social networks tend to select partners that are socially or cognitively similar [37]. Wong et al. [39] shows that the degree of an actor is the number of social ties he/she has and concludes that in many social networks, a majority of actors have relatively small degrees, while a small number of actors may have very large degrees. Luthi et al. [24] show that some social networks do not specifically choose neighbours using locality. By actually giving up a strictly local geographical structure, cooperation often still emerges, provided that the co-edit patterns remain stable in time, which is a first step toward a social network structure. Co-edits occur when two contributors edit the same object in OSM. As we do not have biographical details about the contributors, we do not know if these contributors are from the locality where the object in question is located.

The concept of co-edits (CE) are shown in Figure 4. Suppose Bob creates a polygon in OSM at version 1. Mary then edits this polygon thereby creating version 2. There is a CE created between Mary and Bob such that $S(Mary, Bob, Obj) = 1$. Mary then edits the polygon by

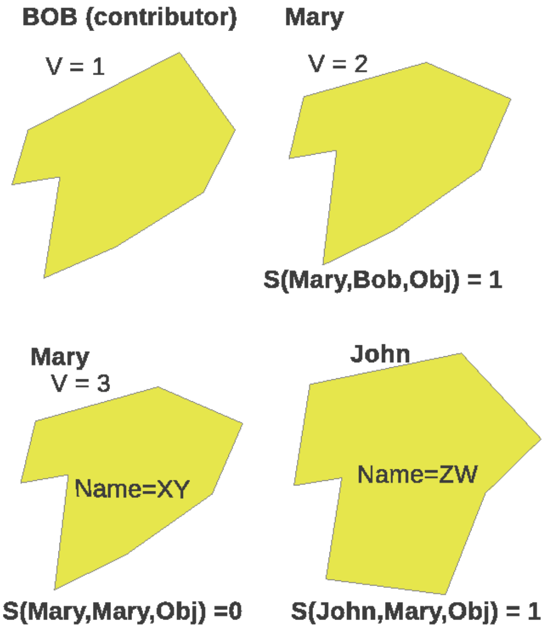

Figure 4: This figure illustrates the concept of Co-Edits (CE) on a single polygon where three contributors are involved in editing

attaching a name tag to the polygon. For consistency the CE here is $S(Mary, Mary, Obj) = 0$. In the final, and most current version, of the object John then edits the value assigned to the name tag. Then $S(John, Mary, Obj) = 1$. Table 4 provides an illustrative example of an actual series of CE from an object representing a tramline in Berlin, Germany. The object is created by contributor 13203 (ranked $6th$ overall contributor to Berlin OSM) on $2007-09-28$. The first CE happens when user 5453 edits the geometry of the object giving $S(5453, 13203, Obj) = 1$. The original creator 13203 makes some subsequent edits. As time passes other CE are generated with the three highest ranked contributors all making contributions to the development of this object in the Berlin OSM database. When contributors are editing OSM data they can very easily access the OSM 'user-id' and OSM 'user-name' which created or edited the current version of the object(s) they themselves are preparing to edit. Contributors may know the OSM 'user-id' and OSM 'user-name' of their friend who is also an OSM contributor or they may know the OSM 'user-id' and OSM 'user-name' of very well known OSM contributors ('FrederickRamm', 'SteveChilton', 'SteveC',etc). Indeed it is possible that a contributor has no knowledge of the other contributor(s) to the object they themsevles are preparing to edit.

4.2 Defining the co-edit network graph model

Now that we have defined the concept of co-edits (CE) we can define co-edit network model. We model the contributors to OSM in each of the case-study cities and their CEs as a graph. This directed graph G is defined as $G = (C, E)$. There is a set of nodes C where $|C|$ is the number of contributors in a given OSM history database for a given city/region. The set of geographical objects in this database is denoted by Ob. Each contributor c_i is a member of C: that is $c_i \in C \, \forall \, 1 \leq i \leq |C|$. If $S(c_i, c_j, O) = 1$ for some object $O \in Ob$ in the OSM database, with $i \neq j$ then these two contributors are said to have "co-edited" O. The time t which c_i and c_j edited O are given by $t(c_i, O)$ and $t(c_j, O)$ with $t(c_j, O) > t(c_i, O)$. The set of edges E is defined where $S(c_i, c_j, O) = 1$ for all $c \in C$ and all objects. The edge $e_{c_i,c_j} \in E$ can be assigned a weight which relates the CE between the nodes c_i and c_j. The weight on any $e_{c_i,c_j} \in E$ is defined as $\sum S(c_i, c_j, o) \, \forall \, o \in Ob$. We can also define an undirected graph of G

Version	**User ID**	**Date**	**Edit Type**
1	13203 (6^{th})	$2007-09-28$	1st Version
2	5453 (12^{th})	$2007-10-15$	Geometry
3	13203	$2007-10-22$	Geometry
19	13203	$2007-11-05$	Geometry and Tags
20	16267 (43^{rd})	$2007-12-04$	Geometry
25	10549 (1^{st})	$2008-01-02$	Geometry
26	13203	$2008-01-28$	Geometry
27	16267	$2008-10-30$	Geometry
28	13203	$2008-12-09$	Geometry and Tags
30	43566 (3^{rd})	$2009-01-28$	Geometry
31	13203	$2009-02-16$	Geometry
33	43566	$2009-03-14$	Geometry
34	6669 (2^{nd})	$2009-08-09$	Geometry
35	69922 (54^{th})	$2009-09-26$	Geometry
36	115651 (5^{th})	$2009-10-06$	Geometry
40	115651	$2010-03-18$	Geometry
48	13203	$2011-01-08$	Geometry and Tags
50	167417 (60^{th})	$2011-12-13$	Geometry and Tags

Table 4: This table shows a subset of edits to an OSM object representing a tram line in Berlin, Germany. Some edits have been removed for purposes of illustration. The overall ranking of each user from their total contribution to the OSM database in Berlin is provided. The edit type column indicates what changes/updates that user made to the object at a given time and version

as $G = (C, E)$ where $e_{c_i,c_j} \in E$ is defined as $\sum S(c_i, c_j, O) + S(c_j, c_i, O)$. The undirected case enumerates any co-editing between c_i and c_j regardless of which contributor edited each other's work. For analysis purposes we can extract various sub-graphs of G by imposing constraints on the definition of $S(c_j, c_i, O)$ or the attributes of each $c_i \in C \, \forall \, 1 \leq i \leq |C|$. For example we could impose a constraint that the subgraph only contain $S(c_j, c_i, O)$ where the period of time between $t(c_j, O) > t(c_i, O)$ is one week or one month.

5 Calculating Social Network Measures for our graph model

Now that we have defined the co-edit network graph model we shall apply some social network measures to understand the properties of this graph. Two important social network characteristics are Eigenvalue Centrality (EC) and Betweenness Centrality (BC). These network characteristics are not new but have gained a renewned prominence in recent years with the rise of Internet-based social networking and social media. One of the major concerns of network analysis is the definition of the concept of centrality. This concept measures the importance of a node's position in a network. In social, biological, communication, and transportation networks, among others, it is important to know the relative structural prominence of nodes to identify the key elements in the network [12, 14]. For our study we shall concentrate on Eigenvector Centrality and Betweeness Centrality.

5.1 Eigenvector and Betweeness Centrality

Eigenvector Centrality (EC) is one of the most powerful techniques in the social network analysis toolkit. EC takes into account not just the number of links that each actor has (as in degree centrality), but also the number of links of the connected actors, and their links too, and so on throughout the network. So if A is the very high frequency contributor to OSM, with lots of co-edits with many other contributors, then a person B connected directly to A (but only to A) still has a lot of importance, even though B may only have one connection. To prevent this we have set the value of $\alpha = 10$. Another contributor Z might be connected to three people, but if those individuals are not of high importance themselves, then Zs importance is similarly low. If we rank people by EC we can see who the key contributors are in the co-edit network for OpenStreetMap. At the top of the list these contributors may be obvious candidates but this could help us to identify those other contributors who have high eigenvector centrality even though they are not necessarily highly ranked in OSM in their city by their total contribution rates. Their appearance high in the ranking of contributors by EC suggests that we may need to investigate further to determine the reason for their high EC value. A node's eigenvector centrality is proportional to the sum of the eigenvector centralities of all nodes directly connected to it. In our OSM co-edit network the nodes representing contributors to OSM with high EC values are those nodes most likely to be involved in the most co-editing behaviour: that is either editing other contributors work or having their own work edited by other contributors. EC values are normalized and lie between 0 and 1.

Betweenness Centrality (BC) is a measure of the centrality of a node in a network, and is normally calculated as the number of shortest paths between node pairs that pass through the node of interest. Betweenness is, in some sense, a measure of the influence a node has over the spread of information through the network. Nodes with a high betweenness centrality score lie in the shortest path of information or work ow between a number of other nodes [36]. In our OSM context this means that such a node represents a contributor which can potentially exert control over the co-editing of information in the OSM database by other contributors or who has edited the work of many other contributors. BC values are normalized and lie between 0 and 1

Feature	Berlin	London	Minn	Munich	Paris	Prague	Sydney
N (Top 10%)	584	342	58	323	240	141	80
$\lvert C^*\rvert$	$2,143$	$1,988$	302	$1,816$	$1,938$	$1,043$	433
$\lvert E^*\rvert$	$38,765$	$28,787$	8900	$23,458$	$16,333$	$9,870$	$1,302$
$\alpha >= 10$	16%	15%	21%	15%	16%	12%	16%
Density	0.008	0.007	0.009	0.007	0.004	0.009	0.006
Mean BC	0.021	0.018	0.009	0.021	0.009	0.017	0.017
Mean EC	0.012	0.008	0.006	0.017	0.006	0.012	0.012

Table 5: Eigenvalue Centrality and Betweeness Centrality statistics from the co-edit network model involving the top 10% of contributors in each city

5.2 Analysis of Eigenvector and Betweenness centrality of the OSM co-edit network

In this section we shall discuss the results of the analysis of the measurement of Eigenvector Centrality (EC) and Betweeness Centrality (BC) for the OSM co-edit networks. Table 5 shows the CEs involving the top 10% of contributors in all cities. This is a subgraph of $G = (C, E)$ where we begin to build a directed CE graph $G*$ which includes $c_i \in C$ where c_i is one of the top 10% ranked contributors in a given city. For $G* = (C*, E*)$ the members of $C*$ is made up of $c^*_i \in C$ where c^*_i are in the top 10% ranked contributors. For some CE $S(c^*_i, c^*_j, O)$ or $S(c^*_j, c^*_i, O)$ the node c^*_i may not be in the top 10% ranked contributors. So $C* = c^*_i + c^*_j \forall i, j$. As before we can constrain the construction of edges connecting c^*_i and c^*_j. For Table 5 we imposed the constraint that there is a threshold α to control this construction. α is the number of CE existing between two contributors c_i and c_j. After an analysis of the distribution of CE between all contributors for all cities we found that $\alpha \geq 10$ was a very suitable value. This threshold is also reasonably consistent over all cities as outlined in Table 5. $\alpha \geq 10$ is then bounded from below by α CE to a single object O or α CE to α distinct objects in Ob. Using α then the CE are expressed as $\sum S(c^*_i, c^*_j, O) \geq \alpha$ or $\sum S(c^*_j, c^*_i, O) \geq \alpha$ for the directed version of $G*$ or $\sum S(c^*_i, c^*_j, O) + S(c^*_j, c^*_i, O) \geq \alpha \ \forall O \in Ob$ for the undirected version. Setting $\alpha \geq 10$ also prevented the social network graphs becoming unwieldy and difficult to analyse. Table 5 presents results when considering $G* = (C*, E*)$ as a directed graph. The construction of $G* = (C*, E*)$ using $\sum S(c^*_i, c^*_j, O) \geq \alpha$ or $\sum S(c^*_j, c^*_i, O) \geq \alpha$ means that $|C*|$ will almost certainly be larger than the number of contributors in the top 10% for a given city. In the case of Berlin, in table 5, there are 584 contributors in the top 10% but the CE graph $G* = (C*, E*)$ has $|C*| = 2143$. This is due to contributors in the top 10% editing, or being edited by, contributors ranked outside the top 10%. A visualisation of this social network is provided in Figure 5. For all graphs the density is calculated. Graph density is expressed in the range 0 to 1 for a maximum connected graph. The value of $\alpha \geq 10$ indicates the percentage of all CE which have been considered for this $G* = (C*, E*)$. As stated above we investigated setting α at different values. Over 50% of all CE have $\alpha \leq 2$. The problem with this model is that the graph is too large. Despite the imposition of a constraint on α there is in all cases at least three times as many nodes in C^* than in the top 10%. This makes it very difficult to interpret the significance of the BC and EC measurements for this graph.

To develop a more informative social network graph model we decided to constraint the selection of contributors in C^* more strictly. In Table 6 and Table 7 we provide results of extracting EC and BC measures from the co-edits involving only the top 10% of contributors in all cities. As above we constructed $G* = (C*, E*)$ as an undirected graph where each edge $e \in E*$ joining c^*_i and c^*_j has a weight of $\sum S(c^*_i, c^*_j, O) + S(c^*_j, c^*_i, O) \geq \alpha \ \forall O \in Ob$. The contributors c^*_i and c^*_j must both be ranked within the top 10% of contributors in the corresponding city. The tables contain the following information. The row “Top 10%” represents

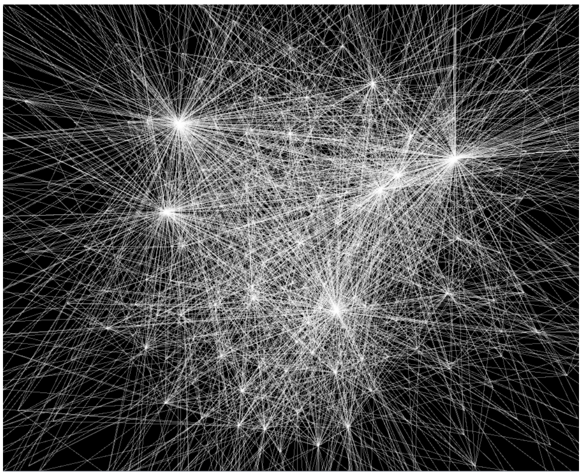

Figure 5: In this figure the co-edit network for Berlin from Table 5 is shown. The visualisation enhances the co-edit linkages between nodes (contributors). It is evident from the visualisation (white bursts) that there are several key contributors in this network who are co-editing with other contributors.

the Top 10% ranked contributors in the corresponding city. The row "Nodes" represents $|C*|$ in $G* = (C*, E*)$ while the row "Edges" represents $|E*|$ in $G* = (C*, E*)$. The row "Isolated" represents the difference in the number of nodes in the Top 10% and $C*$. The rows "$EC > 0.2$" and "$EC > 0.1$" represent the number of contributors in $C*$ with EC greater than 0.2 and 0.1 respectively. The rows "Mean EC", "STD EC", and "Max EC" represent the mean, standard deviation, and maximum EC value of all $c_k^* \in C*$. The rows for BC are defined in a similar fashion. The rows "EC rank" and "BC rank" are provided to give an idea of the $c_k^* \in C*$ who have the highest EC and BC values respectively. The data in these rows show the overall rank of c_k^* in the Top 10% of contributors and their corresponding EC or BC value.

In Table 6 there are some interesting features from our computation of various social network model characteristics of CE between the top 10% ranked contributors when editing ways only. There is a small but not insignificant group of contributors with $EC > 0.1$ which indicate that these contributors are co-editing with other contributors with high EC. The number of isolated nodes is high for Berlin, Munich, and Paris. It is difficult to understand the editing behaviour of these contributors as they have neither edited or being edited by other top 10% ranked contributors. We speculate that these contributors might be involved in mapping ways of very specific or niche geographic objects such as patches of shubbery, flowers in a park, etc. Alternatively these contributors have made contributions but have long since left the OSM project. The number of contributors with high BC values are very small indeed. We feel that this might be a facet of the construction of our co-edit network model. BC is based on shortest paths between nodes and this may not be a suitable measure to calcuate for this co-edit network. The interesting aspect of both $ECRank$ and $BCRank$ is that these are not necessarily dominated by the very highest ranked contributors. Lower ranked contributors (such as those ranked $28^{th}, 42^{nd}$, and 44^{th} in Berlin) have amongst the highest EC values for the corresponding co-edit network. There is a similar scenario for Prague in the $BCRank$. A visualisation of the CE network for Prague is shown in Figure 7. This particular network is worthy of further investigation as two contributors are completely dominant. These users "Bilbo" and "Petr Dlouhy" are two of the most frequent contributors to OSM in Europe with over $250,000$ and $146,000$ way creations and $644,663$ and $70,000$ way edits respectively (correct to June 2013). In Figure 6 an example of the CE network from Table 6 for Berlin is shown where larger nodes represent those contributors with the highest EC measurements.

In Table 7 there are some interesting features of our social network model of edit interac-

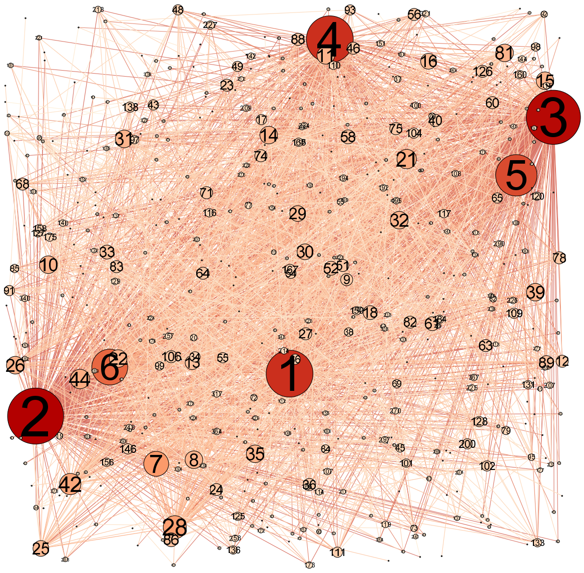

Figure 6: In this figure the co-edit network for Berlin from Table 6 is shown. The visualisation enhances the edit interaction linkages between nodes (contributors). The larger nodes have the largest EC values with the node label representing the overall rank of that contributor.

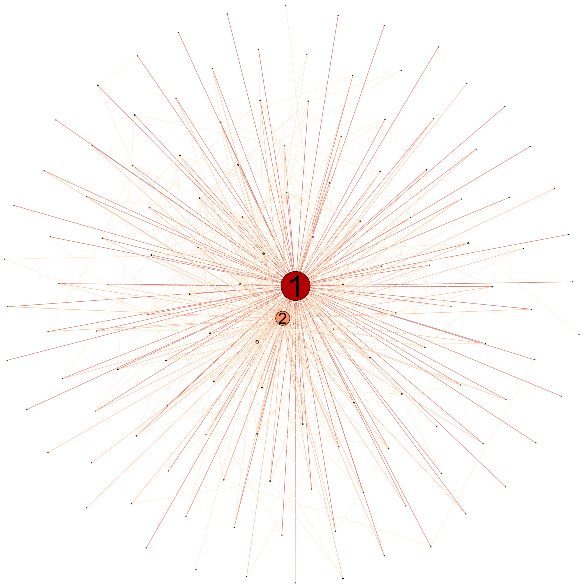

Figure 7: In this figure the co-edit network for Prague from Table 6 is shown. The visualisation enhances the CE linkages between nodes (contributors). The larger nodes have the largest BC values with the node label representing the overall rank of that contributor. Node 1 (BC = 0.6) and 2 (BC = 0.283) are completely dominant using this measure.

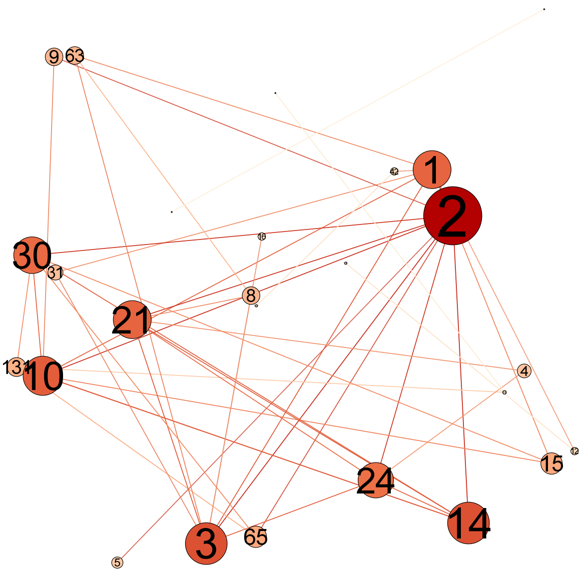

Figure 8: In this figure the co-edit network for Berlin from Table 7 is shown. The visualisation enhances the CE interaction linkages between nodes (contributors). The larger nodes have the largest EC values with the node label representing the overall rank of that contributor.

tions between the top 10% ranked contributors when editing relations only. As we mentioned previously relations are a logical collection of ways and nodes in OSM. They are more complex to represent in OSM either in the database itself and by creating/editing them using the various OSM editor software [20]. As in Table 6 there is a small but not insignificant group of contributors with $EC > 0.1$. This time the number of isolated nodes is much higher. This indicates that there is little co-editing of relations between these top 10% ranked contributors. When we look a the EC and BC rankings for co-editing of relation objects we find that there is an even greater spread of ranked contributors involved. Both Berlin and London have contributors who are ranked outside the top 200 contributors having high BC and EC values. We feel that this may potentially be a result of certain contributors concentrating on a specific area of OSM or applying their own expertise to a small set of geographic objects such as railway stations or airports as represented by relations. Figure 8 an example of the CE network for OSM relations from Table 7 for Paris is shown where larger nodes represent those contributors with the highest EC measurements. In this smaller network the nodes all have slightly greater EC but dominant contributors are still evident.

6 Discussion and Conclusions

To our current knowledge this is the first report which has discussed results of the application of a graph theory approach to the analysis of interaction and collaboration amongst contributors to the OpenStreetMap project. This paper has attempted to explore the interaction between contributors to the OpenStreetMap project. Our work focussed on those contributors in our case study cities who have contributed very frequently to OSM in that city. The characteristics of these contributors are outlined in Table 2. The characteristics of the OSM databases for the selected cities are summarised in Table 1. The results in this paper shall provide a solid platform for further studies into the study of the motivations contributors to OpenStreetMap, the social aspects of their contributions, and their editing behaviour.

This analysis of the co-editing amongst high frequency contributors to OSM has provided a number of important results:

- *Contributions to OSM are not linear or predictable*: This is discussed in Section 3.1. There are bursts of object creation and object editing after events such as mapping parties or after the import of freely available geodata.

- *The top* 10% *contributors (ranked by their quantity of contributions) perform over* 90% *of all object creations and edits.* As discussed in section 3.2 this has been shown in other work. However, our work here validates this for seven different cities with different OSM communities and OSM histories.

- *The top* 10% *contributors can be clustered into distinct classes of mapping behaviour.* While Rehrl et al. [33] outlines a very extensive set of actions which can be performed by a VGI contributor our analysis found that these top 10% contributors can be clustered into four distinct classes. In most cases "tagging" is the most popular action followed by "geometry only" actions. There is also a distinct set of contributors who only "create" objects in OSM but carry out little other editing. The aim of this clustering approach was to investigate what the dominant type of contribution was amongst this top 10%. Tagging (creation and update) is a very important contribution. While the geometry of objects in OSM may not change very frequently the values assigned to attributes in tags (such as the name of a store, open/closed status of a seasonal park, etc) can change quickly. Consequently it is necessary that contributors ensure tags are temporally and spatially correct.

- *We developed the concept of co-editing (CE) as a means of developing a social networking model of contributors to OSM.* As we discussed in section 4, OSM does not have any explicit "follow" or "friendship" structures between contributors. We developed the concept of co-editing to represent when two distinct OSM contributors edited the same object. The OSM contributors X and Y must co-edit such that X edits the contribution of Y or vice-versa. Table 5 shows the characteristics of the network of co-edits generated from the CE contribution history for each city.

- *We calculated Eigenvalue Centrality and Betweenness Centrality measurements for several different co-edit network configurations.* EC and BC are two very powerful techniques for quantitatively understanding social networks. In Table 6 and Table 7 we show the results of calculation of EC and BC for co-edit networks involving the top 10% of contributors based on their edit interactions on ways and relations respectively. We found that in both cases there are a very small number of contributors who have high EC and BC attributes based on their co-editing characteristics in the network. EC and BC helps us to identify these contributors as very important contributors in terms of the overall structure of the edit interaction network for OSM in the seven case-study cities. We also found that contributors who are not necessarily the very highest ranked in their respective city can have very high EC and BC measures. There are a number of possible explanations for this. These contributors are: editing popular, heavily edited, objects in OSM, they are interested in and consistently editing in the same geographical areas as other highly ranked contributors, or their editing work is being corrected or enhanced by other contributors. Very dominant contributors in some cities such as Prague (Figure 7) are easily identified in visualisations of EC and BC. However it is necessary to find out more information about these contributors in order to understand the reason for their dominant positions in the networks.

- *It is possible to control the level of co-editing using the α parameter.* As outlined in Section 4.1 it is possible to control the size of the co-edit network graph using the α parameter. We found this effective in allowing us to build graphs which were more suitable for analysis. In future work we will consider adding temporal conditions to the calculation of α. For example $S(c_i, c_j, O) = 1$ only if $t(c_i)$ and $t(c_j)$ are within the same month, week, or day. This will potentially allow us to automatically detect social events in the OSM contributor network where a higher than normal number of CE will occur during the time duration specified in the α constraints.

- *There are "isolated" contributors in all test-case cities.* In table 6 and table 7 we see that there is a small, but not insignificant, number of contributors in the top 10% which are isolated in the co-edit network graph. This means that the objects which they have created or edited are not co-edited by any other contributor in the top 10%. This result was not expected. It is beyond the scope of this current study to understand the circumstances for this isolation. It will be necessary to investigate the types of objects that these contributors are contributing and their location. Perhaps these objects are very rarely contributed objects in OSM and subsequently do not generate enough interest for other contributors to contribute to their editing and growth?

- *Co-editing occurs amongst the top contributors in all test-case cities.* Our results in table 6 and table 7 confirm anecdotal claims of contributor co-editing in OSM. However the quantitative evidence of this co-editing, using BC and EC, provides us with statistical confirmation of anecdotal evidence. What is unknown is how or why do these high frequency contributors perform co-edits on each other's contribution? Are they correcting errors or problems? Are they ensuring that a given object or set of objects are maintained to ensure geometrical, spatial, and semantic consistency (through tagging)? The results for BC and EC indicate that these top contributors are not just correcting the errors or problems generated by new contributors to OSM but rather are co-editing the work of other very experienced contributors. We tentatively speculate that this behaviour could form the basis of a self-regulating and self-administrating community like Wikipedia. If this could be quantified, through a more rigourous social analysis of co-editing, then this would be a very positive result for the sometimes maligned and misunderstood OSM community. Sepehri Rad et al. [35] remarks that predicting the positive or negative attitude of individuals towards each other in a social environment has long been of interest. If it were the case that top contributors reacted poorly to a co-edit by another contributor then tag-wars, disputes, etc would become rife and endanger the stability and quality of the OSM database. Our analysis, for the test-case cities, do not indicate this type of behaviour. Identification of *positive* co-editing between contributors in OSM could lead to better quality OSM data because as Liu and Ram [23] points out for their study on Wikipedia the identification of co-editing patterns which are preferable or detrimental for article quality can providing insights for designing tools and mechanisms to improve overall article quality.

Despite these very insightful results we feel that there are still some more interesting questions which merit further research work. Why do these "serious mappers" or "senior mappers" contribute so much effort to a project such as OSM? Dodge and Kitchin [11] remark that this apparent willingness to participate for free in crowdsourcing projects is "undoubtedly based on the fact that these projects provide genuniely effective platforms to connect socially, communicate meaningfully, and contribute collectively". As we have found in this paper, and in agreement with authors such as Dodge and Kitchin [11], that OSM is "crowdsourced by a few and not the many". In our test-case cities and beyond there are relatively small active groups of contributors who perform much of the mapping work, quality control, community development etc. Despite the rhetoric of mass involvement there are "small numbers of dedicated individuals in comparison with the large numbers consuming OSM and other VGI" [11]. The "few" referred to here could be what Haklay [18] calls a "a small technical elite" who posses significant technical knowledge to creating new geographic data collection tools in OSM to allow the production of free geographic information that is accessible to anyone and for any purpose. It will be necessary to perform further investigation into who these "few" are. Information could be gathered from web-based surveys, interviews, etc. similar to the work performed by Budhathoki and

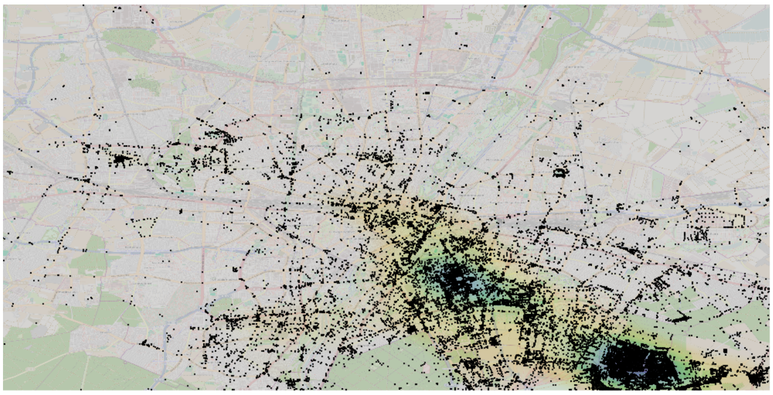

Figure 9: This graphic shows the density of edits by the top contributor to Munich OSM over the entire history of their edits. The map of Munich from OSM is used as the base-layer. It is very evident from the density of these contributions that this contribution is focussed on central Munich (south-east of the image) and very less so the north and west of the city

Haythornthwaite [5]. How active are the most high frequency contributors on a daily, monthly, or yearly basis? It will be necessary to assess the role that these important contributors are now playing to understand if their rates of contributions are sustainable into the future [21]? For other authors working on a similiar analysis we do feel that it will be necessary to obtain more biographical information about contributors. We feel that there is a limit to the extent of the insights which can be gained without this type of information and potential qualitative information related to the reasons by a contributor contributed to OSM in a particular way.

In this paper we have not considered the geographic location of the edits made by contributors. We shall be investigating, in our future work, if there are specific groups of contributors who are not only co-editing the same objects but the same objects in the same general regions of these cities. This will help to identify local cliques which have formed organically rather than in a pre-planned fashion. Figure 9 shows the density of contributions from the top ranked contributor to Munich OSM. There is an obvious higher density of contribution work to the south-east of the city which includes the downtown region. Using Rehrl et al. [33]'s contribution model for VGI we intend to investigate, at a more detailled, level when those seemingly lower ranked contributors have high EC and BC values. Is this because their work is being "corrected" or merely enhanced by other contributors? This could yield some useful insight into quality control in OSM by "senior mappers". While we concentrated on seven cities our approach is very flexible and is extendible to other cities and regions provided there is sufficient OSM history data. The continent of Africa or central America would provide a very interesting case-study. This would also provide an opportunity to investigate cities with different socio-economic characteristics and populations. Very often OSM and VGI networks are formed in these regions by aid agencies and local governments to support various humanitarian plans or programmes [41]. Contributors to local OSM and VGI projects do not have to be physically located in these areas [13]. It would be very interesting to investigate if these networks are more socially integrated than those outlined in this paper.

References

[1] Achananuparp, P., Lim, E.-P., Jiang, J., and Hoang, T.-A. (2012). Who is retweeting the tweeters? modeling, originating, and promoting behaviors in the twitter network. *ACM Trans. Manage. Inf. Syst.*, 3(3):13:1–13:30.

[2] Ballatore, A. and Bertolotto, M. (2011). Semantically enriching vgi in support of implicit feedback analysis. In Tanaka, K., Frhlich, P., and Kim, K.-S., editors, *Web and Wireless Geographical Information Systems*, volume 6574 of *Lecture Notes in Computer Science*, pages 78–93. Springer Berlin Heidelberg.

[3] Ballatore, A., Bertolotto, M., and Wilson, D. (2012). Geographic knowledge extraction and semantic similarity in openstreetmap. *Knowledge and Information Systems*, pages 1–21.

[4] Budhathoki, N., Nedovic-Budic, Z., and Bruce, B. (2010). An interdisciplinary frame for understanding volunteered geographic information. *Geomatica*, 64(1):11–26.

[5] Budhathoki, N. R. and Haythornthwaite, C. (2012). Motivation for open collaboration: Crowd and community models and the case of openstreetmap. *American Behavioral Scientist*.

[6] Comber, A., See, L., Fritz, S., der Velde, M. V., Perger, C., and Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, 23(0):37 – 48.

[7] Corcoran, P. and Mooney, P. (2013). Characterising the metric and topological evolution of openstreetmap network representations. *The European Physical Journal Special Topics*, 215:109–122.

[8] Corcoran, P., Mooney, P., and Bertolotto, M. (2013). Analysing the growth of openstreetmap networks. *Spatial Statistics*, 4(0):5–19.

[9] Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Bue, A. D., Menegaz, G., and Murino, V. (2011). Social interaction discovery by statistical analysis of f-formations. In *Proceedings of the British Machine Vision Conference*, pages 23.1–23.12. BMVA Press. http://dx.doi.org/10.5244/C.25.23.

[10] Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96.

[11] Dodge, M. and Kitchin, R. (2013). Crowdsourced cartography: mapping experience and knowledge. *Environment and Planning (A)*, 45:19–36.

[12] Donninger, C. (1986). The distribution of centrality in social networks. *Social Networks*, 8(2):191 – 203.

[13] Georgiadou, Y., Bana, B., Becht, R., Hoppe, R., Ikingura, J., Kraak, M.-J., Lance, K. T., Lemmens, R., Lungo, J. H., McCall, M., Miscione, G., and Verplanke, J. (2011). Sensors, empowerment, and accountability: a digital earth view from east africa. *Int. J. Digital Earth*, 4(4):285–304.

[14] Gomez, D., Figueira, J. R., and Eusebio, A. (2013). Modeling centrality measures in social network analysis using bi-criteria network flow optimization problems. *European Journal of Operational Research*, 226(2):354 – 365.

[15] Gómez, V., Kappen, H. J., and Kaltenbrunner, A. (2011). Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 181–190, New York, NY, USA. ACM.

Measure	Berlin	London	Minneapolis	Munich	Paris	Prague	Sydney
Top 10%	584	342	58	323	240	141	80
Nodes	515	323	57	303	198	130	77
Isolated	69	19	1	20	42	11	3
Edges	2592	2350	386	1980	914	441	447
$EC > 0.2$	5	3	9	3	7	3	7
$EC > 0.1$	11	27	22	26	27	21	23
Mean EC	0.027	0.035	0.112	0.038	0.046	0.064	0.089
STD EC	0.034	0.042	0.07	0.043	0.054	0.059	0.07
MAX EC	0.304	0.233	0.281	0.272	0.29	0.422	0.301
EC Rank	(2:0.304) (3:0.291) (4:0.252) (1:0.251) (5:0.228) (6:0.195) (7:0.137) (28:0.125) (42:0.118) (44:0.108)	(2:0.233) (1:0.212) (14:0.203) (9:0.195) (20:0.192) (7:0.187) (4:0.184) (8:0.172) (30:0.166) (18:0.162)	(1:0.281) (2:0.259) (3:0.251) (15:0.250) (4:0.240) (14:0.232) (19:0.229) (7:0.219) (16:0.207) (11:0.196)	(7:0.272) (1:0.258) (5:0.216) (14:0.187) (9:0.183) (6:0.181) (8:0.177) (3:0.162) (13:0.157) (15:0.153)	(10:0.290) (1:0.266) (2:0.252) (15:0.236) (9:0.230) (12:0.221) (4:0.200) (28:0.172) (5:0.164) (14:0.157)	(1:0.422) (2:0.381) (3:0.241) (5:0.198) (9:0.180) (6:0.178) (8:0.177) (7:0.155) (15:0.144) (10:0.132)	(1:0.301) (6:0.272) (2:0.270) (5:0.227) (7:0.220) (8:0.218) (3:0.203) (9:0.198) (10:0.197) (16:0.187)
$BC > 0.2$	2	0	1	2	1	2	1
$BC > 0.1$	4	3	2	3	3	2	2
Mean BC	0.002	0.004	0.0145	0.0044	0.006	0.008	0.012
STD BC	0.2	0.017	0.0374	0.022	0.025	0.057	0.049
MAX BC	0.317	0.1636	0.238	0.283	0.243	0.6	0.41
BC Rank	(2:0.317) (3:0.243) (1:0.142) (4:0.141) (5:0.060) (6:0.054) (42:0.032) (32:0.020) (81:0.018) (7:0.016)	(2:0.163) (1:0.153) (14:0.106) (8:0.097) (4:0.091) (9:0.074) (7:0.072) (20:0.0499) (16:0.0371) (25:0.0235)	(0.238) (2:0.100) (15:0.088) (3:0.068) (4:0.067) (14:0.051) (19:0.037) (34:0.036) (7:0.027) (16:0.023)	(7:0.283) (1:0.226) (5:0.102) (6:0.057) (3:0.050) (2:0.050) (9:0.0348) (8:0.0341) (14:0.033) (4:0.030)	(10:0.243) (2:0.155) (1:0.134) (9:0.099) (12:0.087) (4:0.066) (3:0.065) (15:0.06) (8:0.048) (21:0.038)	(1:0.600) (2:0.283) (3:0.042) (6:0.0215) (15:0.018) (64:0.015) (5:0.013) (51:0.010) (8:0.008) (28:0.007)	(1:0.410) (2:0.112) (6:0.094) (8:0.047) (7:0.042) (5:0.035) (3:0.029) (10:0.028) (4:0.027) (9:0.018)

Table 6: This table provides the result of analysis of the EC and BC for co-edits (CE) between contributors ranked in the top 10% of each city only. This analysis investigates co-edits of way objects only

Measure	**Berlin**	**London**	**Minneapolis**	**Munich**	**Paris**	**Prague**	**Sydney**
Top 10%	584	342	58	323	240	141	80
Nodes	64	52	7	51	26	24	14
Isolated	520	290	51	272	214	117	66
Edges	122	91	9	86	45	40	25
$EC > 0.2$	2	9	6	8	8	6	7
$EC > 0.1$	21	15	7	17	16	21	9
Mean EC	0.088	0.088	0.353	0.097	0.151	0.169	0.225
STD EC	0.088	0.107	0.134	0.101	0.124	0.114	0.142
MAX EC	0.496	0.414	0.547	0.495	0.452	0.584	0.449
EC Rankings	(3:0.496) (4:0.368) (2:0.295) (5:0.219) (92:0.195) (1:0.186) (6:0.176) (42:0.173) (175:0.172) (420:0.167)	(20:0.415) (55:0.348) (82:0.342) (4:0.320) (2:0.315) (15:0.286) (25:0.246) (3:0.231) (14:0.205) (137:0.133)	(15:0.548) (2:0.481) (16:0.442) (4:0.365) (3:0.254) (11:0.216) (47:0.167)	(1:0.496) (15:0.339) (29:0.300) (23:0.287) (13:0.267) (17:0.244) (40:0.226) (76:0.221) (5:0.181) (6:0.171)	(2:0.452) (3:0.323) (14:0.320) (10:0.301) (21:0.295) (1:0.293) (30:0.282) (24:0.274) (65:0.161) (15:0.160)	(2:0.584) (3:0.331) (1:0.288) (9:0.245) (5:0.224) (8:0.205) (27:0.188) (17:0.185) (67:0.157) (56:0.157)	(5:0.450) (2:0.404) (29:0.374) (3:0.374) (1:0.326) (9:0.324) (10:0.240) (17:0.178) (6:0.177) (32:0.086)
$BC > 0.2$	2	2	2	1	2	1	3
$BC > 0.1$	3	5	4	6	5	3	4
Mean BC	0.021	0.035	0.014	0.029	0.06	0.051	0.09
STD BC	0.071	0.064	0.115	0.084	0.086	0.164	0.142
MAX BC	0.487	0.359	0.366	0.549	0.334	0.815	0.461
BC Rankings	(3:0.488) (4:0.302) (42:0.107) (2:0.071) (92:0.056) (82:0.055) (98:0.042) (490:0.030) (146:0.030) (134:0.030)	(20:0.329) (14:0.266) (82:0.165) (15:0.150) (224:0.113) (11:0.074) (55:0.073) (36:0.071) (3:0.071) (166:0.071)	(15:0.367) (16:0.233) (3:0.167) (11:0.100) (47:0.067) (2:0.067) (4:0.001)	(1:0.549) (15:0.166) (23:0.163) (2:0.145) (76:0.111) (29:0.105) (6:0.075) (56:0.038) (10:0.038) (7:0.038)	(1:0.334) (2:0.288) (3:0.169) (131:0.141) (14:0.109) (30:0.075) (69:0.073) (42:0.073) (12:0.073) (21:0.073)	(2:0.816) (12:0.166) (1:0.100) (30:0.087) (3:0.055) (9:0.006) (8:0.006) (5:0.001) (67:0.000) (56:0.000)	(1:0.462) (6:0.295) (5:0.291) (29:0.109) (2:0.056) (9:0.034) (3:0.024) (32:0.001) (17:0.001) (14:0.001)

Table 7: This table provides the result of analysis of the EC and BC for co-edits (CE) between contributors ranked in the top 10% in each city. This analysis investigates co-edits of Relation objects only

[16] Goodchild, M. F. (2008). Whither vgi? *GeoJournal*, 6(72):239–244.

[17] Goodchild, M. F. (2009). Neogeography and the nature of geographic expertise. *Journal of Location Based Services*, 3(2):82–96.

[18] Haklay, M. (2013). Neogeography and the delusion of democratisation. *Environment and Planning (A)*, 45:55–69.

[19] Haklay, M., Basiouka, S., Antoniou, V., and 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.

[20] Jones, C. E. and Weber, P. (2012). Towards usability engineering for online editors of volunteered geographic information: A perspective on learnability. *Transactions in GIS*, 16(4):523–544.

[21] Kittur, A., Nickerson, J., Bernstein, M. S., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. J. (2013). The future of crowd work. In *ACM Conference on Computer Supported Cooperative Work (CSCW)*. Social Science Research Network (SSRN) ID: 2190946.

[22] Lin, Y.-W. (2011). A qualitative enquiry into openstreetmap making. *New Review of Hypermedia and Multimedia*, 17(1):53–71.

[23] Liu, J. and Ram, S. (2011). Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.*, 2(2):11:1–11:23.

[24] Luthi, L., Pestelacci, E., and Tomassini, M. (2008). Cooperation and community structure in social networks. *Physica A: Statistical Mechanics and its Applications*, 387(4):955 – 966.

[25] Massa, P. (2011). Social networks of wikipedia. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 221–230, New York, NY, USA. ACM.

[26] Mooney, P. and Corcoran, P. (2012). Characteristics of heavily edited objects in openstreetmap. *Future Internet*, 4(1):285–305.

[27] Mlligann, C., Janowicz, K., Ye, M., and Lee, W.-C. (2011). Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In Egenhofer, M., Giudice, N., Moratz, R., and Worboys, M., editors, *Spatial Information Theory*, volume 6899 of *Lecture Notes in Computer Science*, pages 350–370. Springer Berlin Heidelberg.

[28] Neis, P., Goetz, M., and Zipf, A. (2012). Towards automatic vandalism detection in openstreetmap. *ISPRS International Journal of Geo-Information*, 1(3):315–332.

[29] Neis, P., Zielstra, D., and Zipf, A. (2011). The street network evolution of crowdsourced maps: Openstreetmap in germany 20072011. *Future Internet*, 4(1):1–21.

[30] Neis, P. and Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project the case of openstreetmap. *ISPRS International Journal of Geo-Information*, 1(2):146–165.

[31] OpenStreetMap (2013a). OpenStreetMap: daily statistics for the osm global database. Online: `http://www.openstreetmap.org/stats/data_stats.html` (Last Checked: February 15th 2013.

[32] OpenStreetMap (2013b). OpenStreetMap gains it's one millionth member. Online: `http://opengeodata.org/1-million-openstreetmappers` (Last Checked: February 15th 2013).

[33] Rehrl, K., Grechenig, S., Hochmair, H., Leitinger, S., Steinmann, R., and Wagner, A. (2013). A conceptual model for analyzing contribution patterns in the context of vgi. In Krisp, J. M., editor, *Progress in Location-Based Services*, Lecture Notes in Geoinformation and Cartography, pages 373–388. Springer Berlin Heidelberg.

[34] Rhodes, C. J. and Keefe, E. M. J. (2007). Social network topology: A bayesian approach. *The Journal of the Operational Research Society*, 58(12):pp. 1605–1611.

[35] Sepehri Rad, H., Makazhanov, A., Rafiei, D., and Barbosa, D. (2012). Leveraging editor collaboration patterns in wikipedia. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, HT '12, pages 13–22, New York, NY, USA. ACM.

[36] Singh, P. V. (2010). The small-world effect: The influence of macro-level properties of developer collaboration networks on open-source project success. *ACM Trans. Softw. Eng. Methodol.*, 20(2):6:1–6:27.

[37] Ter Wal, A. and Boschma, R. (2009). Applying social network analysis in economic geography: framing some key analytic issues. *The Annals of Regional Science*, 43:739–756. 10.1007/s00168-008-0258-3.

[38] United Nations (2013). World demographic yearbook 2013: Population and vital statistics report. United Nations Statistics Division Demographic Statistics: Available Online `http://unstats.un.org/unsd/demographic/` Last Checked: June 2013.

[39] Wong, L. H., Pattison, P., and Robins, G. (2006). A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications*, 360(1):99 – 120.

[40] Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertesz, J. (2012). Dynamics of conflicts in wikipedia. *PLoS ONE*, 7(6):e38869.

[41] Zook, M., Graham, M., Shelton, T., and Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: A case study of the haitian earthquake. *World Medical and Health Policy*, 2(2):7–33.