AAAI Technical Report WS-12-04
When the City Meets the Citizen

Mapping Community Engagement with Urban Crowd-Sourcing

Desislava Hristova, Afra Mashhadi, Giovanni Quattrone, Licia Capra
Dept. of Computer Science, University College London
Gower Street, London, WC1E 6BT, UK

Abstract

Cities are highly dynamic entities, with urban elements such as businesses, cultural and social Points-of-Interests (POIs), housing, transportation and the like, continuously changing. In order to maintain accurate spatial information in these settings, crowd-sourcing models of data collection, such as in OpenStreetMap (OSM), have come under investigation. Like many crowd-sourcing platforms (e.g., Wikipedia), these geo-wikis exhibit tailing-off activity, bringing into question their long-term viability. In this paper, we begin an investigation into the sustainability of urban crowd-sourcing, by studying the network structure and geographical mapping of implicit communities of contributors in OSM. We observe that spatially clustered crowd-sourcing communities produce higher coverage than those with looser geographic affinity. We discuss the positive implications that this has on the future of urban crowd-sourcing.

Introduction

The world's population has grown sevenfold in the past two centuries and now half of us live in cities, with the rate of urbanisation still approaching its peak. While economic growth is welcomed in urban hubs (Bettencourt and West 2010), high dynamicity increases the cost of centrally maintaining up-to-date spatial information such as maps, rendering some public datasets obsolete (Masser 1998). A solution made possible with the advent of Web 2.0 is crowd-sourcing, where user-generated content can be cultivated into meaningful and informative collections, as exemplified by sites like Wikipedia (Voss 2005). This form of citizen science has been amplified by the rise of location-based services and the wide adoption of powerful mobile devices. Equipped in this manner, citizens can become surveyors, with council-monitoring applications like FixMyStreet[1]; reporters, with micro-blogging sites such as Twitter[2], and cartographers, with geo-wikis like OpenStreetMap[3].

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]http://www.fixmystreet.com/

[2]https://twitter.com/#!/iranprotestnews/

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

OpenStreetMap is a successful example of a crowd-sourcing platform, where people with basic geographic skills and an affinity for digital mapping can contribute to the free wiki map of the world. There are currently 547,270 users registered on the OSM platform[4]. The geographic information they collectively provide has been demonstrated to be of high quality, where quality has been measured in terms of positional accuracy, attribute completeness and consistency. Indeed, OSM's quality has shown to sometime supersede the most reputable geographic datasets, performing especially well in urban areas (Haklay 2010).

Relying entirely on user-generated content for urban mapping raises concerns, not only in terms of quality of the collected information (which, for OSM, is presently high), but also in terms of its long-term sustainability, on account of the driving force behind it (Panciera et al. 2010). Will citizens *continue* to remain engaged with the crowd-sourcing process, and will they do so *accurately*?

As a first step towards assessing the viability of urban crowd-sourcing, we build a *spatial interest network*, where an edge exists between two users if they have been editing in the same areas of the city of London, UK. The higher the number of edits in the same areas, the higher the similarity weight on the edge connecting such users (i.e., the stronger the 'interest' of such citizens in the same parts of the city). We then observe whether geographic clustering affects the quality of coverage of an area.

The reminder of the paper is structured as follows: we begin with a brief review of the state of the art in urban crowd-sourcing and community analysis. We then present the urban crowd-sourcing dataset we have used in our investigation, the process we have followed to construct a virtual network of contributors from it, and the method used to extract communities from this network. We analyse the spatial and structural characteristics of the detected communities and their relation to quality of coverage, allowing for a better understanding and, in the future, prediction of quality in urban crowd-sourcing settings. Finally, we conclude the paper with a discussion of our research agenda.

[4]http://www.openstreetmap.org/stats/data_stats.html

Related Work

Quality of crowd-sourced information has been extensively researched in the domain of Volunteered Geographical Information (VGI) (Goodchild 2007). In this domain, quality of OpenStreetMap data has been assessed in comparison to traditional geographical datasets maintained by national mapping agencies, as well as proprietary datasets maintained by commercial companies such as Navteq. For example, Haklay et al. (Haklay 2010; Haklay et al. 2010) measured the positional accuracy of OSM road networks in the UK and found it to be very accurate (i.e., on average within 6 meters of the position recorded by Ordnance Survey). Overall, the attention of the VGI community has focused on road networks only; however, the contribution process associated with editing roads and that associated with editing POIs differ greatly: indeed, the former is typically done by users who have some expertise in both the geography of an area and the editing tools required to digitally represent it, whilst the latter can be performed by any city dweller with local knowledge. It is the latter that is representative of citizen engagement, and it is thus the focus of this paper.

The link between physical and online communities has been observed in the prediction of social ties from space and time co-occurrence (Crandall et al. 2010), in the adoption of social roles (Welser et al. 2011) and vastly in the observation of offline characteristics within online social networking sites (Gilbert and Karahalios 2009; Pennacchiotti and Popescu 2011). In Social Network Analysis (SNA), research has shown that users cluster around moods, personality traits, beliefs as well as geographically. This analysis is derived from the richness of personal information projected online through social networking sites. Research shows that online communities do in fact highly resemble real social communities in their dynamics and structure, particularly in their scale-free and small-world characteristics (Fu, Liu, and Wang 2008). This mirror effect that the web has can be used to analyse not only online social networks but also implicit networks such as in Flickr, the photo-sharing website. For example, (Crandall et al. 2010) derived accurate social ties between the users of the site by observing the geo-tag and timestamp of photographs. Ties need not be personal to be meaningful as demonstrated by experiments with Twitter and Wikipedia, where people form interest networks which have beneficial implications for the content produced.

The relationship between quality and community has been observed in sociological theory, where quality is shown to be localised within tightly-knit homophilic communities. Homophily, a similarity bond that can generate social networks is produced by homogeneity and common interests, through which people cluster together. The principle of homophily as the glue of social networks was originally discussed in Lazarsfeld and Merton's 1954 publication (Lazarsfeld and Merton 1954), where a distinction was made between *status* and *value* homophily. The former type of relationship takes into consideration the social status of an individual, while the second operates around ideas, regardless of factors such as wealth and education of an individual. More recently, the phenomenon has been thoroughly discussed in McPherson, Smith-Lovin and Cook's 2001 work *Birds of a feather flock together: Homophily in social networks*. Their work discusses gender and age in homophily, among many other factors, demonstrating that males tend to segregate into larger homogeneous groups in work establishment networks. Men are also more likely to create voluntary associations with other men roughly their age, than women who tend to have more heterophilic relationships in general. Most importantly, the authors show that quality becomes localised in sub-networks and that it is primarily bred by geographic factors (McPherson, Smith-Lovin, and Cook 2001). Like many other crowd-sourcing applications, OSM editors form a highly homogeneous group of predominantly young and educated male contributors (Lam et al. 2011).

Collectivism and collaboration have been shown to shape the vibrant community of OSM's users. Despite lacking formal social networking facilities, the contributors to OSM join forces in "mapping parties" and actively participate in forums and wikis (Perkins and Dodge 2008). This behaviour generates implicit communities of interest, as is also the case with other crowd-sourcing platforms such as Wikipedia, whose contributors form interest networks around topics and ideas. This community formation property of wikis is also the main reason for their success, creating inner workings of quality control and a feeling of attachment for its users (Christakis and Fowler 2009).

Urban Crowd-Sourcing Dataset

In this section, we describe the urban crowd-sourcing dataset at hand. We then discuss how we have extracted community information from the user-contributed data.

OpenStreetMap Dataset

OpenStreetMap (OSM) is the most famous example of VGI publicly available today. Registered users can contribute spatial content describing map features to the global OSM database, thus collectively building a free, openly accessible, editable map of the world. The OSM dataset contains the history of all edits (since 2006) on all spatial objects performed by all users. Spatial objects can be one of three types: *nodes*, *ways*, or *relations*. Nodes broadly refer to Points of Interest (POIs), ways are representative of roads, and relations are used for grouping other objects together.

For the present analysis, we have restricted our attention to a subset of the openly available OSM dataset. In particular, we have selected edits done to POIs of the city of London, UK in the one year period between 19-06-2010 and 19-06-2011. We selected London as the context of our investigation because of its large number of users and contributions, which is partly due to the fact that OSM originated there. Our study is confined to the above one year period in order to capture a sufficiently representative static snapshot of the community. To ensure we are considering real citizens, and not bots for example, users with unnaturally high numbers of edits (over 40 edits in the same 'changeset' – i.e., same session in OSM) were filtered out. Our focus is only on POIs (and not roads), capturing edits requiring less specialised skill from the citizens contributing them.

The characteristics of our filtered dataset are summarised in Table 1. A preliminary analysis of the number of edits per

#Users	#POIs	#Edits	Power Law Coefficients
819	9, 718	10, 623	$\alpha = 1.328$, $R^2 = 0.9681$

Table 1: Characteristics of OSM Filtered Dataset

user shows a power-law distribution, with an alpha value exponent of 1.33 and a fitting coefficient of $R^2 = 0.97$. Given a set of pairs $\langle x_i, y_i \rangle$, where $x_i \in \mathbb{N}$ is a user and $y_i \in \mathbb{N}$ is the number of edits performed by x_i, α is chosen such that $\sum (a x_i^{-\alpha} - y_i)^2$ is minimized. Our value signifies that the majority of contributions are made by a small subset of so-called "power users", that is, users who contribute heavily.

Interest Network and Communities

Our *virtual* network was crafted from the raw dataset described above, where nodes are OSM editors, and an edge exists between any two editors if they have made contributions in the same geographical area. When studying the implicit community structure of this network, we make the underlying assumption that citizens predominantly edit in urban areas of *relevance* to them. That is, areas where they spend time (e.g., near where they live, or work), as they must have visited a location physically to edit it, and they care enough to place it on the map. POIs do not include houses or private residences but only businesses and public locations. The following analysis therefore focuses specifically on the urban elements which represent the user's locus of edits and from which we can confidently build a representation of the user's spatial interests.

We divided London into a 20x20 grid, with each cell covering an area of 2Km x 2Km. For each user, a vector of 400 elements was created (one per geographic cell), counting how many edits that user made in each such cell. After normalising these counters, we computed pair-wise user similarities, as the cosine similarity of their corresponding vectors. Following manual inspection of the distribution of these weights, we constructed a network with an edge between two users if their similarity was at least 0.5. The resulting network contains 714 (out of the original 819) nodes and 8,839 edges (the top 40%); it has a diameter of 13 and an average clustering coefficient of 0.8, representative of strong interconnectedness. To study community structure within this network, we ran the Louvain modularity optimisation algorithm for community detection (Blondel et al. 2008), as implemented in Gephi (http://gephi.org). The algorithm operates iteratively, beginning with one community per node in the network, and then repeatedly aggregating communities together so to optimise, at each aggregation step, the division of network modules. The process stops when further iterations fail to increase the modularity and a hierarchy of communities is then produced. The Louvain algorithm detected 98 communities in our network, with a final modularity value of 0.63, showing that the network is naturally highly divided into non-overlapping communities, although the vast majority of these communities are singletons. In the next section, we focus our attention on the top 6 communities in terms of size, and analyse their properties.

Mapping Citizen Engagement

For each of the six communities detected above, we now investigate their properties in terms of spatial affinity, network characteristics and quality of coverage.

To investigate spatial affinity (or geographic clustering), we constructed a map mosaic based on our grid division of London, where each grid is assigned to the community that has most significantly contributed to it over the 12 month period under investigation (Figure 1). In the case of a tie, the cell is assigned to more than one community. Intuitively, this shows what communities are most responsible for the data coverage of a given cell in the year-long period. However, such mapping does not reveal the intensity of contributions in different areas; also, it hides the geographic spread of each community's contributions, highlighting only the area they edited the most w.r.t. any other community. We thus also show heatmaps for each community in Figure 2.

In terms of network characteristics, Table 2 reports, for each community, the number of users it contains, the number of edits performed by these users, and the number of cells in the grid that such community 'owns', as per mapping above. We also report the average clustering coefficient for each community, which gives us an insight into the density of the community, and therefore how strong the forces of similarity are within it. The alpha value exponent determines the proportion of heavy editors with respect to moderate and light editors in each community. The higher the exponent, the more occasional editors it contains and the more similar to the original distribution it is. The fitting coefficient of the power-law distribution of edits per user, determines how closely it follows the power-law curve. Finally, we compute the average quality of coverage per cell of each area where a community resides (as per Figure 1). In order to measure this, we compared the OSM POI dataset to the Navteq proprietary POI dataset as a benchmark. We calculated coverage in every cell as the ratio of POI matches between the two datasets to the total number of POIs appearing in the Navteq dataset. We considered two POIs of the two datasets a match when their Euclidean distance was no more than 0.1Km and their names had less than 0.35 in lexographic distance.

We now go back to our goal of investigating the presence of spatial clustering in OSM. Evidence of it suggests that the community dedicates its mapping effort to a well-defined area and according to our hypothesis does so meticulously. As shown in Figure 2 (c)(e)(f), communities 3, 5 and 6 exhibit very high concentration, with all their contributions happening in geographically small areas. For example, Community 3 is active in the city centre and just South of it; Community 5 is situated mainly in the North and centre, while Community 6 is concentrated in the area of Kilburn, in the North-West. Interestingly, these communities have much higher clustering coefficients than the network of OSM contributors as a whole (0.91, 0.96 and 0.94 respectively, as opposed to 0.8), suggesting that they are very tightly-knit virtual communities. Furthermore, if we look at the distribution of edits per user, we note that Community 6 does not follow a power-law distribution (low fitting coefficient); Communities 3 and 5 do follow a power-law distribution, but with a much lower alpha exponent w.r.t. that of the whole network.

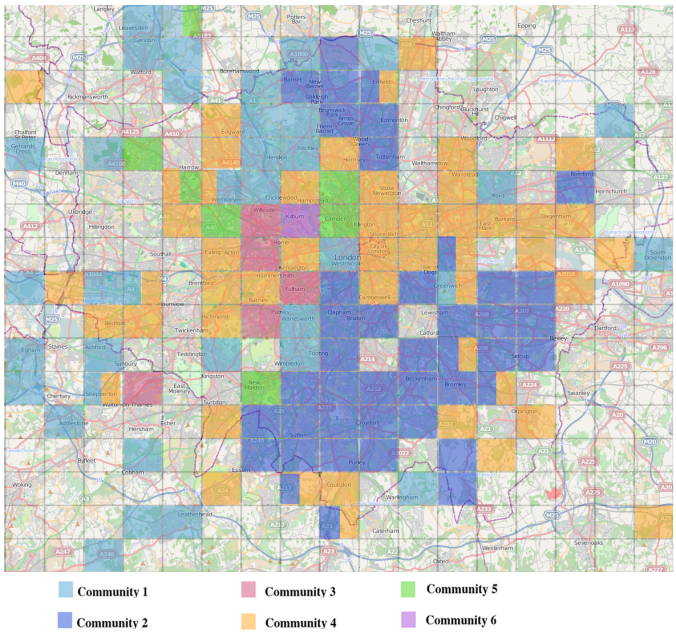


Figure 1: Map of the Spatial Distribution of the Six Communities on the London Grid

	#Users	#Edits	#Cells	Clustering Coefficient	α	R^2	Coverage
1	87	1,374	34	0.93	1	0.96	0.22
2	63	856	39	0.87	0.95	0.95	0.2
3	54	749	5	0.91	0.97	0.96	0.24
4	53	1,260	43.5	0.9	1.28	0.92	0.13
5	38	218	6	0.96	0.84	0.95	0.17
6	27	427	1	0.94	1.47	0.87	0.39

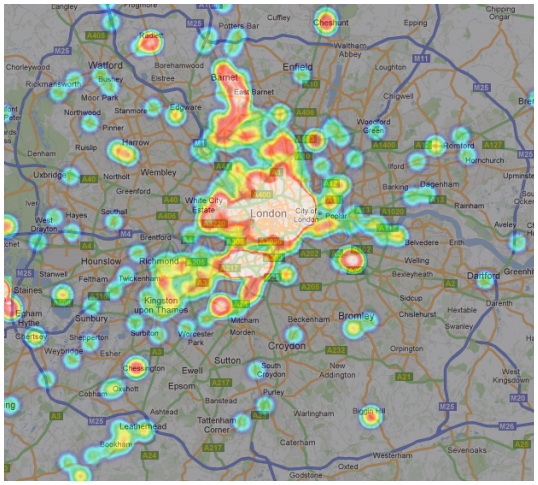
Table 2: Community Properties

These findings suggest that members of these communities are actively and more equally engaged with the OSM mapping effort, as opposed to what the network as a whole would suggest (where a small group of power users is responsible for the vast majority of contributions). It is also the case that the coverage recorded in the cells of the two communities is the highest, suggesting that citizen engagement in OSM does indeed depend on community structure and spatial affinity. Community 6, which is the most closely knit community, is geographically associated with a single cell having a coverage value of 0.39, significantly higher than all the others.

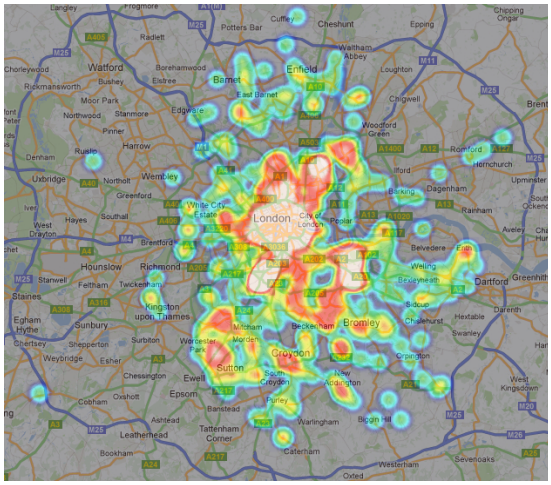
Although less spatially clustered than the previous three, Communities 1 and 2 also exhibit geographic affinity, albeit with a more polycentric behaviour. Indeed, whilst they are both active in the city centre (which is not surprising as that is where POIs are most concentrated), we can observe Community 1 (the biggest in terms of number of users) is predominantly engaged with the North of London, while Community 2 (second biggest) is very active in the South. The map mosaic shows this North-South divide quite clearly indeed (Figure 1). It is interesting to note that even the two largest communities have alpha values significantly lower than the original network: once again, contributions are more evenly distributed among members of these communities, and this seems to be a good indicator of 'healthy' communities. This is confirmed by the coverage metric, which is quite high for both communities, with values of 0.22 and 0.2 respectively.

Community 4 is the only one with no spatial association (i.e., highest geographic spread), as shown both by the map mosaic (Figure 1) and the community heatmap (Figure 2 (d)). This is also the only community with both a high α exponent and a high fitting coefficient R^2 (aligned with those of the original overall network), indicating that contributions are not evenly spread amongst its members: the high edits per user ratio suggest the presence of a few power users, responsible for the majority of edits. Note also that this is the community with the lowest coverage, possibly a consequence of its geographically spread edits.

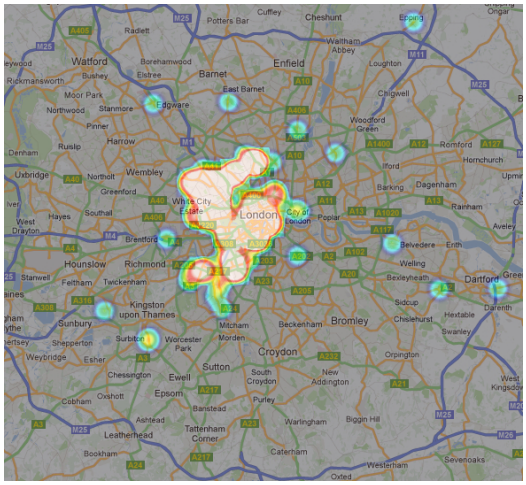
Based on the above, there is an apparent relationship between the presence of geographic clustering within a community, and (a) the even spread of contributions within such community, and (b) the high quality of coverage in the areas it occupies. Spatial clustering is present in five of the six communities detected and they all exhibited higher levels of coverage. These communities also show an active engage-



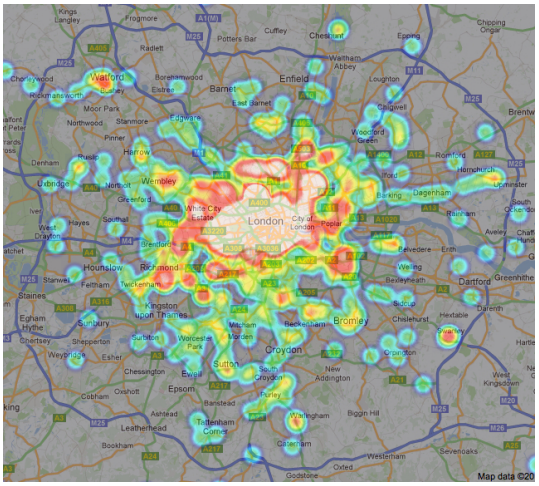
(a) Community 1



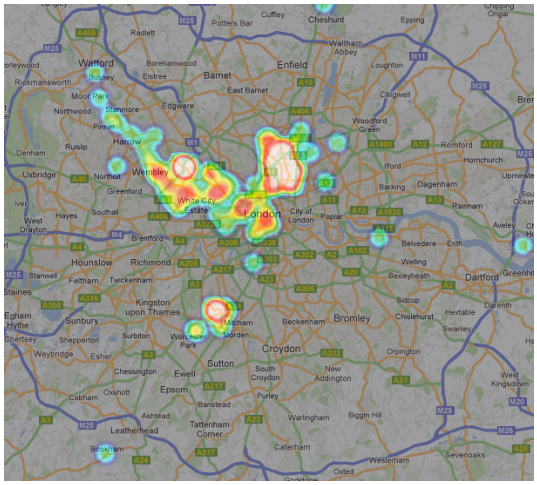
(b) Community 2



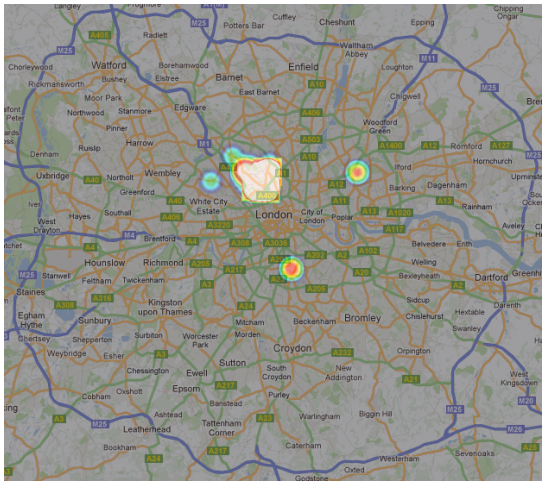
(c) Community 3



(d) Community 4



(e) Community 5



(f) Community 6

Figure 2: Heatmap of Contributions across Top 6 Communities

ment with the mapping process that is evenly distributed among its members. This preliminary analysis inspires further inquiry because it shows a clear correlation between spatial affiliation, the internal community structure and the community's engagement in terms of coverage.

Future Work

The work presented in this position paper is an initial step in a line of enquiry that aims to understand the factors determining the viability of urban crowd-sourced geographic data. The ultimate goal is to build predictive models of information quality and sustainability in these settings.

The hypothesis that we have started to verify in this paper is that spatially attached virtual communities produce consistently high levels of coverage. So far, we have studied the presence of geographically clustered communities in OSM, where the overall quality has been measured and observed to be generally high. Indeed, we have observed that OSM editors do cluster together in geographically well-defined areas of London and that the quality of coverage in those areas is greater. The next step in this line of enquiry is to uncover what specific *properties* of these communities determine what levels of *quality*. Properties that we intend to look at go beyond topological characteristics of these communities, and include socio-cultural aspects of the geographical areas these communities belong to (e.g., wealth, average education level, income, employment). As for quality, we intend to measure both accuracy and coverage of the edited information: by accuracy, we refer to both lexicographic correctness in the spelling of the POI names, and geographic accuracy, in terms of the spatial positioning of POIs (compared to proprietary mapping datasets such as Navteq). By coverage, we refer to the proportion of POIs that have been crowd-mapped, w.r.t. those that exist in the real world (once again, as recorded in proprietary mapping datasets).

In order to refine our methodology, we will fine-tune our analysis of central areas in accordance with the density of edits. We intend to do this using ranked distance in lieu of normalised counts (Liben-Nowell et al. 2005). In order to classify a user's interest in an area we will use distinctive edits which are less common rather than popular edits which most users will have. This will be done using a tf*idf-like weighing technique (Salton and McGill 1983), commonly used to calculate the relevance of documents, so to create a more meaningful network of editors.

So far, we have focused on a single temporal snapshot of OSM, thus disregarding its dynamic nature, which is an essential aspect of its sustainability over time. Our final step is to study the evolution of OSM over the years, both in terms of its communities and the measured quality. We also intend to repeat this study across a different cities, so to observe the formation and evolution of communities in various urban contexts. In so doing, we hope to be able to build accurate predictive models of sustainability of the urban crowd-sourcing paradigm in the future.

Acknowledgements. The research leading to these results has received funding from the European Community (FP7-SST-2008-RTD-1) under Grant Agreement n. 234239.

References

Bettencourt, L., and West, G. 2010. A unified theory of urban living. *Nature* 467(7318):912–913.

Blondel, V.; Guillaume, J.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10(10008):12–18.

Christakis, N., and Fowler, J. 2009. *Connected. The Amazing Power of Social Networks and How They Shape Our Lives*. London: Harper Press.

Crandall, D.; Backstrom, L.; Cosley, D.; Siddharth, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *PNAS* 107(52):22436–22441.

Fu, F.; Liu, L.; and Wang, L. 2008. Empirical analysis of online social networks in the age of web 2.0. *Physica A: Statistical Mechanics and its Applications* 387:675–684.

Gilbert, E., and Karahalios, K. 2009. Predicting tie strength with social media. In *Proceedings of CHI*, 211–220.

Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221.

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. *Cartographic Journal, The* 47(4):315–322.

Haklay, M. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B* 37(4):682–703.

Lam, S. T. K.; Uduwage, A.; Dong, Z.; Sen, S.; Musicant, D. R.; Terveen, L.; and Riedl, J. 2011. Wp:clubhouse?: an exploration of wikipedia's gender imbalance. In *Proc. of WikiSym*, 1–10.

Lazarsfeld, P., and Merton, R. 1954. Friendship as a social process: a substantive and methodological analysis. *Freedom and Control in Modern Society* 18–66.

Liben-Nowell, D.; Novak, J.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2005. Geographic Routing in Social Networks. *Journal of the National Academy of Sciences* 102(33):11623–11628.

Masser, I. 1998. *Governments and Geographic Information*. London: Taylor and Francis.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–444.

Panciera, K.; Priedhorsky, R.; Erickson, T.; and Terveen, L. 2010. Lurking? cyclopaths? a quantitative lifestyle analysis of usr behaviour in a geowiki. In *Proc. of CHI*, 1917–1926.

Pennacchiotti, M., and Popescu, A.-M. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proc. of 17th ACM SIGKDD* , 430–438.

Perkins, C., and Dodge, M. 2008. The potential of user-generated cartography: a case study of the openstreetmap project and mapchester. *North West Geography* 8:19–32.

Salton, G., and McGill, M. J. 1983. *Introduction to modern information retrieval*. McGraw-Hill.

Voss, J. 2005. Measuring wikipedia. In *Proc. of ISSI*, 24–28.

Welser, H.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; and Smith, M. 2011. Finding social roles in wikipedia. In *In Proceedings of the 2011 iConference*, 122–129.