Modelling Growth of Urban Crowd-Sourced Information

Giovanni Quattrone
University College London
London, UK
g.quattrone@cs.ucl.ac.uk

Afra Mashhadi
Bell Labs, Alcatel Lucent
Dublin, Ireland
afra.mashhadi@alcatel-lucent.com

Daniele Quercia
Yahoo Labs
Barcelona, Spain
dquercia@acm.org

Chris Smith-Clarke
University College London
London, UK
chris.smith@ucl.ac.uk

Licia Capra
University College London
London, UK
l.capra@cs.ucl.ac.uk

ABSTRACT

Urban crowd-sourcing has become a popular paradigm to harvest spatial information about our evolving cities directly from citizens. OpenStreetMap is a successful example of such paradigm, with an accuracy of its user-generated content comparable to that of curated databases (e.g., Ordnance Survey). Coverage is however low and most importantly non-uniformly distributed across the city. Being able to model the spontaneous growth of digital information in these domains is required, so to be able to plan interventions aimed at gathering content about areas that would otherwise be neglected. Inspired by models of physical urban growth developed by urban planners, we build a model of digital growth of crowd-sourced spatial information that is both easy to interpret and dynamic, so to be able to determine what factors impact growth and how these change over time. We build and test the model against five years of OpenStreetMap data for the city of London, UK. We then run the model against two other cities, chosen for their different physical and digital growth's characteristics, so to stress-test the model. We conclude with a discussion of the implications of this work on both developers and users of urban crowd-sourcing applications.

Categories and Subject Descriptors

[**Information Systems Applications**]: Spatial-temporal systems-Geographic information systems

General Terms

Measurement

Keywords

Crowd-sourcing, Cellular Automata, Modelling

WSDM'14, February 24–28, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2556195.2556244.

1. INTRODUCTION

Urban crowd-sourcing has become a popular mechanism to harvest knowledge about our evolving cities directly from their citizens. Equipped with powerful mobile devices, citizens have become surveyors, with council-monitoring applications like FixMyStreet;[1] reporters, with micro-blogging sites such as Twitter;[2] and cartographers, with geo-wikis like Cyclopath[3] and OpenStreetMap.[4]

OpenStreetMap (OSM) is perhaps one of the most successful examples of urban crowd-sourcing, with currently over 547,270 users, collectively building a free, openly accessible, editable map of the world. Research has shown that accuracy of the geographic information stored in OSM is very high and it sometimes supersedes the most reputable centrally-maintained geographic datasets, performing especially well in urban areas [19, 31]. As testimony of this success, businesses like Foursquare[5] are now switching from proprietary datasets to OSM.[6]

However, relying entirely on user-generated content for urban mapping raises concerns, not only in terms of accuracy of the collected information (which, for OSM, is presently high), but also in terms of its coverage. Recent studies of OSM have revealed that coverage of information is non-uniformly distributed across a city, with areas that are either further from the centre, or central but income deprived, being particularly neglected [30]. This finding raises concerns in terms of the long-term sustainability of urban crowd-sourcing: is coverage going to spontaneously grow across the city? Or are there going to be areas that will continue to be neglected? Studies so far have been limited in that they quantify coverage at one specific point in time, and in so doing they cannot disclose the factors (e.g., spatial, social, demographic) that contribute to crowd-sourcing participation over time. Being able to build a model that *explains* growth, and that accurately detects what areas are most likely to suffer from neglect, will allow location-based services that make use of crowd-sourced information to plan evidence-based targeted interventions. Such interventions may include, for example, *(i)* the use of a hybrid approach of crowd-sourced and proprietary data, whereby proprietary data is used for seeding the areas that are predicted to suffer (e.g., because far from the city center or because located in economically-deprived neighbourhoods); *(ii)* the design of incentives to call the OSM community to edit spe-

[1] http://www.fixmystreet.com
[2] http://twitter.com/
[3] http://cyclopath.org/
[4] http://www.openstreetmap.org/
[5] https://foursquare.com
[6] http://blog.foursquare.com/2012/02/29/

cific areas, for example, in the form of OSM mapping parties, that is, social events organised by OSMers to descend on an area and map it exhaustively.

We present a model inspired by Cellular Automata theory [11] that accurately captures the *digital* growth of crowd-sourced urban information. Cellular Automata are computational methods that simulate the growth of complex systems by means of a set of simple rules that afford easy interpretation. These models have been extensively used by urban planners and policy makers to understand the evolution of *physical* urban systems, to predict their natural growth, and to evaluate the impact of alternative policy interventions [22]. In this work, we consider a dynamic yet easy-to-interpret version of such models: we divide the city area into cells, we capture past crowd-sourcing activity information for each of them, then use linear regressions to model their growth. We build and test the model over five years of OSM data for the city of London, UK. We discover that the digital growth of crowd-sourced urban information depends on different factors at different times: during the bootstrapping phase of urban crowd-mapping, geodemographic factors, such as distance from the city centre and population density, play a key role (e.g., early adopters naturally start mapping central areas); physical proximity and social influence subsequently become more important; finally, for areas whose digital coverage has already become high, past activity is sufficient to correctly estimate future growth. To assess the validity of the model beyond the city of London, UK, we then stress-test it under rather different urban settings - New York City and Boston in the US. We find the model to suggest that current growth in such cities is mostly explained by geodemographic properties, indeed the same that explained growth in London in its early days of OSM mapping.

The remainder of this paper is structured as follows: after a brief overview of the state-of-the-art, we describe the urban crowd-sourcing dataset at hand and the growth metric we compute on it. The construction of our model follows, with emphasis on the properties that we have mapped onto simple cellular automata rules. We then validate the model: using OSM data for the city of London, UK, we first assess the suitability of the chosen rules individually (and thus of their underlying properties), then of the model as a whole. We then evaluate the model's accuracy and generalisability using it for two other cities in the US. Finally, we conclude the paper by stating the implications of these findings and elaborating on future directions of research.

2. RELATED WORK

Urban crowd-sourcing generally refers to the collective gathering of user-generated content pertaining to the urban environment. The most popular example of such content is volunteered geographic information (VGI), such as that maintained by OpenStreetMap. For years, the research community has studied the accuracy of such information [17], and compared it to traditional geographical datasets maintained by national mapping agencies, as well as proprietary datasets maintained by commercial companies such as Navteq.[7] The findings show very high positional accuracy, both in the UK [18, 19], as well as France, Germany and Switzerland [13, 29].

While accuracy is consistently high, coverage has been shown to vary considerably. Zielstra *et al.* [45] found that road coverage in Germany sharply decreases as we move away from city centres and Girres *et al.* [13] discovered a positive correlation between the number of OSM road 'objects' in an area and number of OSM contributors in that area. Shifting focus from the road network to points-of-interest (POIs), Mashhadi *et al.* [30] found that both socio-economic factors (e.g., income deprivation) and physical distance from the city centre are negatively correlated with coverage in London, UK. Another well known example of urban crowd-sourcing application is Cyclopath,[8] a geo-wiki that is being successfully used to digitally map route information for cyclists in Minneapolis. Rather than focusing on information accuracy, the researchers behind Cyclopath have investigated issues of users' motivation that lead to contributions [38], and conducted studies in terms of user's participation and behavioural analysis [37, 35, 36]. An interesting finding was that manually highlighting areas in the city which were under-covered was sufficient to trigger motivation among cyclists to go those places and map them.

A common limitation of the above studies is that they analyse coverage at one specific point in time. However, crowd-mapping is a process that takes place over time, with contributors both joining and leaving the system all the time, and with new contributions continuously being added. In order to quantify the long term success and sustainability of an urban crowd-sourcing system, we must understand how digital crowd-sourced urban information *grows* over time. Studies in this vein are limited within the area of crowd-mapping. An exception is Neis *et al.* [34], who have recently measured coverage of the OSM road network in Germany over a period of three years, and highlighted at what point the map can be considered complete with respect to commercial datasets. However, this is a retrospective study, with no attempt to model future growth.

Shifting our attention from urban crowd-sourcing to online social networks, we find several studies that have attempted to model their dynamics instead. A large stream of research has developed models of the ways in which individuals join communities. Structural properties of the users' social networks have been found to have a strong influence on subsequent tie formation [5]. Geographical properties of the social network have also been linked to tie formation [21]. For example, Kleinberg [23, 24] modelled the probability of two individuals being friends based on the intuition that friendship probability increases with geographic proximity. In 2005, Liben-Nowell et al. [28] experimentally tested this model upon a half million profiles on the blogging site LiveJournal, on which people reported their locations within the USA and lists of friends. The authors incorporated population density into the initial model and determined that 66% of LiveJournal friendships form through geographic processes. Another stream of research has looked into how information diffuses through these existing social networks. Various models have been proposed based on contagion processes (e.g., [27, 9, 15]); these models have also been expanded, to account for homophily (e.g., [2, 3]), external influences (e.g., [33]), and the language used to describe topics (e.g., [40]).

These models share our goal of modelling 'growth', interpreted in its broader sense (e.g., of ties within a social network, of nodes reached by a topic). However, none of the above models is directly applicable to the area of urban crowd-sourcing (and OpenStreetMap in particular), as such scenario lacks the existence of an explicit social network, so we cannot reason about social network structure to capture growth dynamics. We hypothesise that factors reflecting engagement, influence, and geography will still be predictors of growth in our context though. In the next section, we present a spatio-temporal model of growth that uses properties such as self-reinforcement, physical proximity, social influence and geodemographic conditioning, to model digital growth for urban crowd-sourcing scenarios.

[7] http://www.navteq.com/

[8] http://cyclopath.org/

3. RESEARCH METHODOLOGY

We have built and subsequently tested a growth model of urban crowd-sourced information in the context of OSM. In this section, we thus begin with a brief overview of this dataset, followed by a definition of the growth metric we used. We then delve into a detailed description of the model we built, emphasising the properties that have informed the modelling process.

3.1 Dataset Description

OpenStreetMap is freely available to download[9] and contains the history of all edits since 2006 on all spatial objects performed by all users. In OSM jargon, spatial objects can be one of three types: *nodes*, *ways*, and *relations*. Nodes are single geospatial points, defined using latitude/longitude coordinates, and they typically represent POIs (e.g., cafes, restaurants, hospitals, schools); ways consist of ordered sequences of nodes, and mostly represent roads (as well as streams, railway lines, and the like); finally, relations are used for grouping other objects together, based on logical (and usually local) relationships (e.g., administrative boundaries, bus routes).

To reduce the dataset to a manageable size, we restrict our analysis to the area of Greater London, United Kingdom, where OSM was originally created and launched, and for which the history of contributions is longest. Furthermore, we consider OSM nodes only, thus focusing on POIs rather than the road network. We chose to sample POIs instead of roads as the contribution processes differ greatly between the two categories: road mapping is typically done by users who have high expertise in both the geography of an area and the editing tools required to digitally represent it, while POI mapping can be performed by any city dweller, with no specific cartographic skills required. The latter category is thus more representative of the broad urban crowd-sourcing setting. Finally, to consider only genuine users' contributions, we have looked into contributions that most likely correspond to bulk imports. Two bulk imports were detected in the whole dataset, with tens of thousands of edits done in a single day by a single user, spread throughout Greater London (e.g., more than 20,000 post boxes spread across all Greater London appeared in OSM in only one day in 2009 from the same user). We chose to discard such data for two main reasons: on one hand, we intend to model genuine 'bottom-up', user-generated contributions, of which massive imports are not representative; on the other hand, these bulk 'donations' of data were not geographically concentrated in an area, but rather spread throughout the whole of Greater London. As a result, even if we kept the data (and our model failed to capture it), the result would be a constant error added almost uniformly to all areas under predictions. We should note that, apart from these two extreme cases, we have identified many other peaks of edits by single users, which may or may not be smaller imports; in these cases, we kept and modelled all the data, to avoid falsifying results.

Figure 1 illustrates the cumulative number of editors and edits for each year in the sampled dataset, with 96,357 edits done by a total of 2,240 users overall. As shown, in the very first year of OSM being launched (2006), contributors and contributions were extremely low; to avoid our modelling being skewed by this cold-start phase, during which the technology was being tested but not widely available to the public, we did not try to model 2006 activity, and in the remaining of the paper we focus on the period 2007 (year of the launch of the OSM "State of the Map" conference[10]) onwards.

[9] http://www.geofabrik.de/data/download.html

[10] http://www.stateofthemap.org

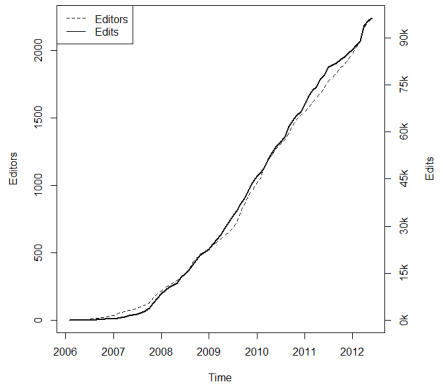


Figure 1: Cumulative number of OSM editors and edits over time

3.2 Metrics

In order to measure growth of crowd-sourced information across an urban area over time, the first step was to choose a *spatial* and *temporal* unit of analysis. In terms of spatial unit of analysis, previous studies have shown that factors such as population density and deprivation are correlated with coverage of crowd-sourced urban information [30]. We have thus chosen to operate at a level of granularity in London for which such information is available. This is the level of *wards*, of which there are about 600, as defined by London Local Authorities.[11]

In terms of temporal unit of analysis, we have tried different time units, from finer (3 months) to coarser (18 months) granularity. In the end, we chose to report the results for the smallest unit of granularity (12 months) that still afforded statistically significant results across *all areas* of Greater London. In particular, although in the very center of the city we could have chosen finer temporal granularity, doing so in the many wards falling within Greater London would not have provided us with sufficient data to have statistically significant results. This has forced us to consider 1 year as the unit of measurement for the time series.

Having defined this spatio-temporal unit of analysis, we then needed to define a metric that reflected which areas had been digitally mapped and which had been neglected instead. To this purpose, it is worth pointing out that not all areas naturally require the same amount of OSM edits to be mapped. For example, areas containing many services and attractions will require many OSM edits to be mapped (e.g., Soho in London); however, sparse areas like parks and industrial estates will require significantly less. To capture this reasoning, we chose as metric *OSM activity*, defined as the number of OSM edits *relative to* the number of physical POIs in each ward at that time:

$$\text{OSM activity} = \frac{\#\text{OSM edits}}{\#\text{POIs}} \quad (1)$$

#OSM edits is readily available from our OSM dataset. To estimate #POIs, that is, the actual number of POIs present in each area, we use Navteq, the leading global provider of maps and location data, covering millions of POIs of varying nature, from restaurants to hospitals and gas stations. Being a commercial service, Navteq's primary objective is to ensure the highest level of accuracy of its

[11] http://data.london.gov.uk/datastore/package/ward-profiles-2011

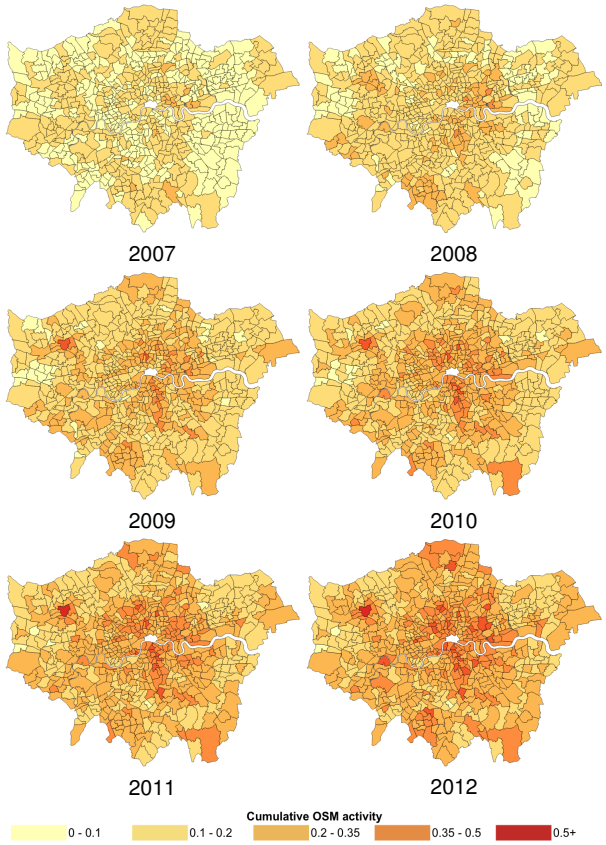


Figure 2: Cumulative OSM activity from 2007 until 2012

data (the information contained there is factually correct and up-to-date).

Figure 2 illustrates the cumulative temporal evolution of OSM activity (Equation 1) in London from 2007 to 2012. As shown, the vast majority of areas have low cumulative activity (with only a few wards slightly above 0.5); furthermore, complex dynamics are at play, with no clear pattern emerging (e.g., no core-to-periphery spreading). To then measure growth, in the next section we propose a model that captures how this metric changes over time.

3.3 Modelling Growth

Modelling the *physical* growth of urban areas has been extensively studied in the domain of urban planning [12, 1, 39]. These models offer an analysis and extrapolation of city dynamics that planners and policy makers use to forecast both natural growth and the effect of policy interventions. Cellular Automata models [11] are a well known example of such models, capable of reproducing complex spatial and temporal dynamics at a global scale, using simple local transition rules. More precisely, the urban area under investigation is first divided into a grid of cells (where each cell may correspond, for example, to a ward) and assigned an initial state t_0. An iterative computational process is then started, whereby the system moves from state t (with $t \geq t_0$) to state $t+1$ using a set of *transition rules* that are applied at each cell in the grid. Typically, transition rules are the same for each cell and follow this general mathematical function:

$$S(c_i, t+1) = f(S(nbh_1(c_i), t), \ldots, S(nbh_n(c_i), t), \\ S(nbh_1(c_i), t-1), \ldots, S(nbh_n(c_i), t-1), \\ \ldots, \\ S(nbh_1(c_i), t_0), \ldots, S(nbh_n(c_i), t_0)) \quad (2)$$

where c_i is the i-th cell of the grid, $S(\cdot,\cdot)$ is the state of a cell at a given discretized time, $nbh_1(\cdot), \ldots, nbh_n(\cdot)$ are the neighbourhood cells relative to the input cell whose transition is being computed (usually including the cell itself), and f is any mathematical function, receiving as input a list of states and returning the new state of the cell under transformation. As an example, a very simple Cellular Automata model, modelling evolution of the physical urban landscape, could comprise the following transformation rules: if a ward is non-urban and it is surrounded by three or more urban wards, then change it to urban; if the ward is urban, then keep it urban. More complex rules are used in practice, but generally all Cellular Automata models follow the basic idea that the state of urbanization of a part of the city depends on the previous state of urbanization of that part (or a sequence of its previous states), together with the previous states of urbanization of its neighbours.

In this work, we borrow from the theories of Cellular Automata to model the growth of *urban crowd-sourced information* in the *digital* domain. To this purpose, we considered each ward of London as a cell of a Cellular Automata model, and discretized time at one-year intervals. Finally, we defined the *state of a ward* w_i of London at time t as the OSM activity that occurred in that ward during the one year preceding t. In the following, we refer to such state as $act(w_i, t)$.

We built our model of crowd-sourced information growth based on processes and properties that capture factors of attractiveness, influence and geodemography.

PROPERTY 1 (SELF-REINFORCEMENT). *If a ward has been edited in the recent past, the ward will continue to be edited in the future.* □

This property paraphrases the preferential attachment process [44], and captures the intuition that, if a geographical area has attracted many contributions from OSM editors in the past (e.g., it contains 'attractions' of interest to such crowd), it will continue to do so in the future. We distinguish between recent and full history of past, so to capture a possible saturation effect that may arise when a ward has already been massively edited in the full past history, to an extent that there is not much else to edit. We then build a model where the predicted outcome $act(w_i, t+1)$ is computed as a linear function over these two quantities:

$$act(w_i, t+1) = \beta_1 \cdot act(w_i, t) + \beta_2 \cdot \hat{act}(w_i, t) + \Omega \quad (3)$$

where $act(w_i, t)$ is OSM activity measured in the last year, while $\hat{act}(w_i, t)$ is defined as $act(w_i, t) + \ldots + act(w_i, t_0)$ and condenses a whole history of activity for that ward; the parameters β_1 and β_2 are the weights we assign to $act(\cdot,\cdot)$ and $\hat{act}(\cdot,\cdot)$ respectively (in the next section, we will show how to determine them). In this and the following models, we use Ω to concisely represent linear pairwise parameter interactions.

As an OSM contributor maps an element in a ward, s/he may also do so for a neighbouring ward s/he passes through. In other words, a contagious process may take place, reminiscent of [20], leading to a spatial diffusion of OSM activity. The second property we consider is thus the following:

PROPERTY 2 (SPATIAL CORRELATION). *If the cells surrounding a ward have been heavily edited in the past, the ward is likely to be edited in the future.* □

To quantitatively capture this property, we first define two *spatial correlation scores* for each ward w_i at each yearly snapshot t: the maximum spatial correlation score $sc_M(w_i,t)$ and the average spatial correlation score $sc_A(w_i,t)$, computed as the maximum and the average activity value of the neighbouring wards respectively:

$$\begin{aligned} sc_M(w_i,t) &= \text{Max}_{w_j \in nbh(w_i)}(act(w_j,t)) \\ sc_A(w_i,t) &= \text{Avg}_{w_j \in nbh(w_i)}(act(w_j,t)) \end{aligned}$$

where $nbh(w_i)$ indicates the list of neighbouring wards of w_i.

We then build a more complex model which predicts activity in a ward w_i at time $t+1$, based on its own past state and that of its adjacent wards. As before, we use a linear function of the form:

$$\begin{aligned} act(w_i,t+1) &= \beta_1 \cdot act(w_i,t) + \beta_2 \cdot \hat{act}(w_i,t) + \\ &+\beta_3 \cdot sc_M(w_i,t) + \\ &+\beta_4 \cdot sc_A(w_i,t) + \Omega \end{aligned} \quad (4)$$

The previous property captures spatial contagion processes based on geographic adjacency between wards. A complementary definition can be given based on wards that, regardless of their spatial positioning within a city, have attracted the same OSM contributors (because they might, for example, offer related attractions/urban functions [10]). We thus formulate the third property as:

PROPERTY 3 (EDITING CORRELATION). *If a ward has been edited by contributors who have heavily edited other wards in the past year, the ward is likely to be edited in the future.* □

In other words, the activity of a ward depends on the past activity of its 'co-edited' areas, where two areas are defined as 'co-edited' if they are edited by some shared editors. To capture this property quantitatively, we define two *editing correlation scores* for each ward w_i at each yearly snapshot t:

$$\begin{aligned} ec_M(w_i,t) &= \text{Max}_{w_j \in cdt(w_i)}(\zeta(w_i,w_j,t) \cdot act(w_j,t)) \\ ec_A(w_i,t) &= \text{Avg}_{w_j \in cdt(w_i)}(\zeta(w_i,w_j,t) \cdot act(w_j,t)) \end{aligned}$$

where $cdt(\cdot)$ returns the list of co-edited wards of the input ward (that is, the list of wards that share at least one editor with the input ward); $\zeta(\cdot,\cdot,\cdot)$ receives as input two wards w_i and w_j, and a time window t, and returns a number belonging to [0,1] indicating how strongly w_i and w_j have been 'co-edited' in the time window t. Specifically, ζ returns the fraction of the edits of w_i which, during year t, has been made by editors who edited both w_i and w_j. Intuitively, the editing correlation score ec_M of w_i is high when there exists at least one ward w_j that has been heavily edited in the last year, by contributors who have also edited w_i. For the second score ec_A to be high, there must exist not just one but *several* heavily edited wards, sharing contributors with w_i.

The resulting model predicts the activity in a ward w_i at time $t+1$, based on its past state, the past state of its physically adjacent wards, and that of its co-edited wards. Once again, we use a linear function of the form:

$$\begin{aligned} act(w_i,t+1) &= \beta_1 \cdot act(w_i,t) + \beta_2 \cdot \hat{act}(w_i,t) + \\ &+\beta_3 \cdot sc_M(w_i,t) + \beta_4 \cdot sc_A(w_i,t) + \\ &+\beta_5 \cdot ec_M(w_i,t) + \beta_6 \cdot ec_A(w_i,t) + \\ &+\Omega \end{aligned} \quad (5)$$

All previous models rely on historical OSM activity information about a ward to be available to compute its next state. This means that, if a ward has not been edited at all in the past and neither have its adjacent ones, this ward is consistently predicted to not grow. However, Figure 2 shows this is not the case, and over time new, previously isolated areas start to be mapped. We thus consider the following factors that may determine the activation of a new ward:

Population. The more densely populated an area is, the more likely one of its residents will join the crowd of OSM editors and start to map the areas s/he knows (i.e., where they live). UK Census data published by the National Statistics Office[12] offers population density information at ward level; we can thus use this information within our model. On top of population density (that is, population divided by ward size), we also consider population per POIs (that is, population divided by number of physical POIs in an area), to differentiate between areas with different POI concentration.

Deprivation. Higher population density does not directly translate into higher number of contributors. Indeed, studies have revealed that contributors of crowd-sourced systems are a group of predominantly young, educated and wealthy males [14]. It is thus more likely that areas of low deprivation will have new OSM contributors. Furthermore, a recent study has revealed that Londoners living in well-off areas do not tend to visit areas that are deprived [26], thus increasing the probability that well-off wards will start to be mapped, as opposed to deprived ones. To quantify deprivation, we use the Index of Multiple Deprivation (IMD), a single parameter that aggregates indicators including income, employment, and education deprivation, as well as crime rates.

Distance from the Nearest Poly-centre. Social and economic activities tend to cluster around a small number of poly-centres in metropolitan areas [8]; a recent study [41] has found that London has 10 different polis. We expect that the further a ward is from its nearest poly-centre, the less-likely this ward will be activated, based on previous OSM studies that have shown road coverage to dramatically decrease as one moves away from the city centre [45].

Based on these considerations we formulate our final property:

PROPERTY 4 (GEODEMOGRAPHY). *The geodemographic factors of a ward (e.g., its population, deprivation, and distance from the centre) influence the probability of such ward to be mapped in the future.* □

Our final model combines all four properties above, and predicts the activity in a ward w_i at time $t+1$, based on its past state, the past state of its physically adjacent wards, that of its co-edited wards, as well as its own geodemographic factors:

$$\begin{aligned} act(w_i,t+1) &= \beta_1 \cdot act(w_i,t) + \beta_2 \cdot \hat{act}(w_i,t) + \\ &+\beta_3 \cdot sc_M(w_i,t) + \beta_4 \cdot sc_A(w_i,t) + \\ &+\beta_5 \cdot ec_M(w_i,t) + \beta_6 \cdot ec_A(w_i,t) + \\ &+\beta_7 \cdot pop\_dens(w_i) + \\ &+\beta_8 \cdot pop\_poi(w_i) + \beta_9 \cdot imd(w_i) + \\ &+\beta_{10} \cdot dist(w_i) + \Omega \end{aligned} \quad (6)$$

In the above formulation, $pop\_dens(\cdot)$, $pop\_poi(\cdot)$, $imd(\cdot)$, and $dist(\cdot)$ are, respectively, the population density, population per POI, IMD, and the Euclidean distance from the nearest poly-centre of a

[12] http://www.ons.gov.uk/

ward w_i. We have measured population density every year, and the changes are minimal in the five year span we consider. IMD is computed every three years, so for this study we have used the version released in 2010 (whose results are based on data collected mainly in 2008). The geodemographic variables are thus not time-dependent in the course of this study.

Table 1 summarises the models above, from its simplest (Equation 3, based on Property 1 only), to the most complex (Equation 6, based on all four properties):

Eq. Number	*Properties*
Eq. (3)	Self-reinforcement
Eq. (4)	Self-reinforcement + Spatial correlation
Eq. (5)	Self-reinforcement + Spatial correlation + Editing correlation
Eq. (6)	Self-reinforcement + Spatial correlation + Editing correlation + Geodemographic

Table 1: Summary of Growth Models

4. RESEARCH RESULTS

We structure the evaluation of the previously constructed growth model in four parts: first, we use correlation analysis to support the validity each of our properties separately; second, we use multiple linear regression to validate the models constructed from such properties, from the simplest to the more complete; third, we turn the models into a classification tool and measure their accuracy in modelling OSM activity growth; finally, we apply the models to other cities to understand their generalisability.

4.1 Testing the Properties

The four properties put forward in the previous section assume that the variables listed in Table 2 (e.g., OSM activity of the ward itself in the last year, OSM activity of the ward itself in all previous years, maximum and average OSM activity of the spatially correlated wards in the last year, etc.) correlate with OSM activity at time $t + 1$.

As a first step in our analysis, we have thus computed the Pearson's Correlation Coefficient r between the value of each such variables at year t and the actual value of OSM activity registered in the following year $t+1$. As some variables exhibited a moderately skewed distribution, we first computed their square root, so to obtain a normal distribution. Correlation values computed on a yearly basis are presented in Table 3.

Property 1 - Self-reinforcement. According to Property 1, OSM activity in a ward at time $t + 1$ depends both on its past activity at time t, and its cumulative past activity. Table 3 confirms this property: the correlation between future OSM activity and its past-year OSM activity goes from a minimum of $r = 0.31$ in 2007/08 to a maximum of $r = 0.37$ in 2009/10. Similarly, the correlation between future OSM activity and the cumulative past OSM activity is positive, and goes from a minimum of $r = 0.11$ in 2007/08 and a maximum of $r = 0.43$ in 2010/11. Note that there exists a positive correlation value between future OSM activity and cumulative past activity in all years analysed (i.e., the more OSM activity there has been in the aggregate past history, the more OSM activity growth will be experienced next). This suggests that OSM in London is yet far from experiencing a saturation effect.

Property 2 - Spatial correlation. The intuition behind Property 2 is that OSM contributors that are very active in one area are likely to spread some of their attention to adjacent areas, thus affecting their growth. Table 3 confirms this property, with OSM activity at time $t + 1$ being positively correlated to both parameters we defined to capture such spatial correlation (maximum spatial correlations goes from $r = 0.27$ in 2010/11 to $r = 0.34$ in 2007/08, and average spatial correlations goes from $r = 0.28$ in 2010/11 to $r = 0.36$ in 2009/10).

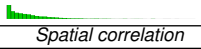
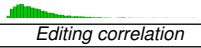
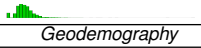
Variable	*Freq. Distribution*	*Max*
Self-reinforcing		
Activity		0.56 (edit/POI)
Cumulative activity		0.92 (edit/POI)
Spatial correlation		
Maximum		0.57 (edit/POI)
Average		0.28 (edit/POI)
Editing correlation		
Maximum		0.34 (edit/POI)
Average		0.03 (edit/POI)
Geodemography		
Population density		23,746.80 (ppl/ha)
Population per POI		14,310.00 (ppl/POI)
IMD		54.90
Distance		22.90 (km)

Table 2: Variables used to model OSM activity growth, together with frequency distribution (starting at zero) and maximum value

Variable	2007/08	2008/09	2009/10	2010/11	2011/12
Self-Reinforcing					
Activity	**0.31**	**0.35**	**0.37**	**0.31**	**0.35**
Cumulative activity	**0.11**	**0.22**	**0.30**	**0.43**	**0.40**
Spatial Correlation					
Maximum	**0.34**	**0.29**	**0.29**	**0.27**	**0.28**
Average	**0.33**	**0.31**	**0.36**	**0.28**	**0.30**
Editing Correlation					
Maximum	**0.22**	**0.21**	**0.17**	**0.18**	**0.20**
Average	**0.21**	**0.13**	0.00	0.06	0.04
Geodemography					
Population density	**0.18**	**0.18**	**0.25**	**0.19**	**0.18**
Population per POI	0.08	0.08	**0.11**	0.09	0.05
IMD	-0.05	-0.09	**-0.15**	-0.06	-0.08
Distance	**-0.26**	**-0.34**	**-0.36**	**-0.32**	**-0.30**

Table 3: Pearson correlation r between individual variables and growth (in bold are those results that are statistically significant – p-value < 0.01)

Property 3 - Editing correlation. Next we consider Property 3, which posits that OSM activity at time $t + 1$ depends on the activity occurred in the past year in co-edited wards, no matter where physically located. We considered two variables: one that expects growth to happen if *at least one* co-edited ward has seen high activity in the past ('maximum' editing correlation score), and a more stringent one, that requires *many* co-edited wards to have exhibited high activity for the ward to be affected ('average' editing correlation score). The former has weak positive correlation with OSM activity growth (from a minimum of $r = 0.17$ in 2009/10 to $r = 0.22$ in 2007/08); the latter has similar correlation in the first two years; however, in the latter two the correlation is not statistically significant (this is possibly because, in our data, the cases where the more stringent condition held true were very few).

Property 4 - Geodemographic factors. Finally, we turn our attention to the correlation between geodemographic factors and a ward's activity at time $t+1$. As Table 3 illustrates, distance from the closest poly-centre is the geodemographic factor with the strongest (negative) correlation (from a minimum of $r = -0.26$ in 2007/08 to $r = -0.36$ in 2009/10), supporting the intuition that, the further a cell is from the centre, the less likely it will be mapped. Weaker correlations exist for all other geodemographic factors.

4.2 Testing the Models

The previous analysis suggests that all four properties put forward are grounded. We now proceed to validate the growth models

Model	Whole London	Top 75%	Top 50%	Top 25%
		2007/08		
Eq. (3):	0.10	0.11	0.12	0.17
Eq. (4):	0.15 (+50%)	0.19 (+73%)	0.16 (+33%)	0.20 (+18%)
Eq. (5):	0.17 (+13%)	0.23 (+21%)	0.20 (+25%)	0.25 (+25%)
Eq. (6):	0.28 (+65%)	0.41 (+78%)	0.44 (+120%)	0.54 (+116%)
		2008/09		
Eq. (3):	0.12	0.10	0.17	0.17
Eq. (4):	0.15 (+25%)	0.15 (+50%)	0.22 (+29%)	0.24 (+41%)
Eq. (5):	0.17 (+13%)	0.20 (+33%)	0.31 (+41%)	0.33 (+38%)
Eq. (6):	0.23 (+35%)	0.34 (+70%)	0.51 (+65%)	0.55 (+67%)
		2009/10		
Eq. (3):	0.15	0.25	0.30	0.34
Eq. (4):	0.18 (+20%)	0.30 (+20%)	0.35 (+17%)	0.40 (+18%)
Eq. (5):	0.21 (+17%)	0.33 (+10%)	0.37 (+6%)	0.42 (+5%)
Eq. (6):	0.29 (+38%)	0.45 (+36%)	0.52 (+41%)	0.58 (+38%)
		2010/11		
Eq. (3):	0.20	0.23	0.31	0.33
Eq. (4):	0.21 (+5%)	0.25 (+9%)	0.34 (+10%)	0.36 (+9%)
Eq. (5):	0.24 (+14%)	0.30 (+20%)	0.38 (+12%)	0.41 (+14%)
Eq. (6):	0.31 (+29%)	0.43 (+43%)	0.53 (+39%)	0.57 (+39%)
		2011/12		
Eq. (3):	0.17	0.24	0.30	0.30
Eq. (4):	0.20 (+18%)	0.25 (+4%)	0.32 (+7%)	0.32 (+7%)
Eq. (5):	0.24 (+20%)	0.28 (+12%)	0.36 (+13%)	0.38 (+19%)
Eq. (6):	0.28 (+17%)	0.41 (+46%)	0.50 (+39%)	0.54 (+42%)

Table 4: Adjusted multiple R^2 of our models, together with relative improvement w.r.t. the previous model (every model showed in this table is statistical significant – p-value < 0.001)

we constructed based on these properties, from its simplest to its final composite one (Table 1). To this purpose, we used multiple linear regression to determine the unknown β parameters of Equations 3–6; furthermore, we compute the adjusted R^2 of each such model, as a way to measure its fit with respect to modelling growth. To highlight the role that different properties play at different times in the evolution of OSM, we break down the analysis on a per-year basis. Furthermore, we decided to report results both for the whole 600 wards in London, as well as the 75%, 50% and 25% most dense wards (in terms of POIs) over the whole 600 wards in London. In so doing, we can analyse the validity of our models for areas of different growth *potential* (i.e., areas that contain only a handful of physical POIs are less amenable to growth modelling, contrary to POI-dense areas). Table 4 shows the results.

Year-based analysis. We observe that the complete model that combines all four properties (Equation 6) has the best fit, significantly improving over all partial models that consider only a subset of them. More interestingly, the gap between the adjusted R^2 of the complete model and of the partial model that leaves out the geodemographic parameters (Equation 5 – Self-reinforcing + Spatial + Editing) significantly decreases as time progresses. For example, in whole London in 2007/08, the complete model (Equation 6) improves the adjusted R^2 of 65% w.r.t. the second best model (Equation 5), whereas in 2010/11 the improvement is 29% only. This confirms the intuition that, in the early stages of OSM (years 2007/08 – 2008/09) all models that do not consider the geodemographic factors are not able to accurately explain OSM activity growth, possibly because of little historical OSM data available. In a second step, that is after 2009, geodemographic factors are not so important anymore, in the sense that we can have a good model of OSM activity growth even without them (possibly because the model has now plenty of historical data to be trained on).

Density analysis. Finally, we observe that the adjusted R^2 increases significantly, when considering models comprising the top 75%, top 50% and top 25% POI dense wards in London. That is, the denser the areas in terms of physical POIs, the better the models to explain their growth.

Model	TN Rate	TP Rate	Accuracy	Senitivity
		Predicted Year: 2009		
Eq. (3)	0.59 (+18%)	0.56 (+12%)	0.58 (+15%)	0.58 (+15%)
Eq. (4)	0.63 (+7%)	0.61 (+9%)	0.62 (+8%)	0.62 (+8%)
Eq. (5)	0.67 (+6%)	0.65 (+7%)	0.66 (+6%)	0.66 (+7%)
Eq. (6)	0.78 (+16%)	0.75 (+15%)	0.77 (+16%)	0.77 (+17%)
		Predicted Year: 2010		
Eq. (3)	0.64 (+28%)	0.60 (+20%)	0.62 (+24%)	0.63 (+25%)
Eq. (4)	0.66 (+3%)	0.61 (+2%)	0.64 (+2%)	0.64 (+3%)
Eq. (5)	0.71 (+8%)	0.68 (+11%)	0.70 (+9%)	0.70 (+9%)
Eq. (6)	0.81 (+14%)	0.78 (+15%)	0.80 (+14%)	0.80 (+15%)
		Predicted Year: 2011		
Eq. (3)	0.66 (+32%)	0.59 (+18%)	0.63 (+25%)	0.63 (+27%)
Eq. (4)	0.70 (+6%)	0.64 (+8%)	0.67 (+7%)	0.68 (+7%)
Eq. (5)	0.75 (+7%)	0.70 (+9%)	0.73 (+8%)	0.74 (+8%)
Eq. (6)	0.82 (+9%)	0.79 (+13%)	0.81 (+11%)	0.81 (+11%)
		Predicted Year: 2012		
Eq. (3)	0.62 (+24%)	0.59 (+18%)	0.61 (+21%)	0.61 (+22%)
Eq. (4)	0.68 (+10%)	0.65 (+10%)	0.67 (+10%)	0.67 (+10%)
Eq. (5)	0.75 (+10%)	0.73 (+12%)	0.74 (+11%)	0.74 (+11%)
Eq. (6)	0.79 (+5%)	0.75 (+3%)	0.77 (+4%)	0.78 (+5%)

Table 5: True Negative Rate (slow growth), True Positive Rate (fast growth), Accuracy and Sensitivity of the classification models. Relative improvement of each model w.r.t. the preceding model is also reported in parentheses. For the baseline model of Equation 3, the improvement is computed w.r.t. a random classifier.

4.3 Predicting Growth

The previous regression results afforded us the ability to understand what properties affect growth of OSM activity in different years, and to what extent. To quantify the *predictive* accuracy of our models, we have then conducted a classification experiment, whereby we first determined the unknown β parameters of Equations 3–6 using multiple linear regression, as done before; we then used these models to classify activity for the upcoming year. For example, we used 2007/08 to estimate the parameters, then made predictions for 2009. In this case, we divided the outcome of our models in two distinct categories: 'slow future OSM activity growth' (when $act(w_i, t+1) < 0.3$) and fast future OSM activity growth' (when $act(w_i, t+1) \geq 0.3$), with 0.3 being the median value of OSM activity growth for the time windows under consideration. Finally, we considered the top 75% wards in London only, as predicting OSM activity growth of very sparse areas (e.g., parks) has little significance. Table 5 presents the results of the classification. As shown, the accuracy of our models is quite high: the simplest model based on Equation 3 already gains up to 32% for slow growth (TN rate) and up to 20% for fast growth (TP rate) over a random classifier. In all predicted years, the full model based on Equation 6 offers the highest accuracy, with up to 82% for slow growth and 79% for fast growth.

4.4 Beyond London

Our modelling and evaluation has so far focused on a single city, and this raises concerns in terms of the validity of our properties and models elsewhere. Indeed, London is in itself quite a peculiar city: from a geodemographic perspective, London is by far the largest city in the EU (the most densely populated within city limits), with a diverse range of people and cultures spread all over its area. In terms of OSM-related characteristics, London was the city were OSM was born, it is the one with the longest-living and largest user base, and it is considered a prime example or organic, bottom-up crowd-sourcing growth with the vast majority of OSM edits being genuine users' contributions, rather than bulk imports (i.e., data donations to the OSM foundation by third-party providers [32]).

To address this concern, we present in this section a brief study of two cities, Boston and New York City, chosen for their very different characteristics with respect to London. From a geodemographic perspective, we scale both population density and area size of a factor up (for New York City) and down (for Boston) with respect to London. From an OSM perspective, both cities are much less 'mature' than London: as shown in Table 6, they have a much smaller user base, smaller number of edits, and lower level of engagement from OSMers (edits per user), especially in New York City. This city is also fundamentally different from London in terms of growth pattern: while in London only 21% of OSM edits can be attributed to bulk imports, in New York City these represent 65% of total OSM edits. New York City thus resembles a sort of worst case scenario when it comes to predicting growth using our model, as only 35% of its contributions are actually genuine users' contributions (i.e., those that the properties in our models aim to describe). In repeating the previous analysis on Boston and New York City, we have used US census data, which is available at the level of tracts (closest US-equivalent of wards);[13] as we did not have Navteq data for the US, we have used POIs information from Yelp instead[14] (to gain confidence that Yelp was indeed a reliable source to estimate the number of physical POI in a certain area, we computed the Pearson correlation between the number of POIs in Yelp and that in Navteq for the city of London, and we found it to be above 0,9).

City	*Edits*	*Users*	*Avg(edits/user)*	*Import ratio*
London	96,357	2,240	43.0	21%
New York	12,275	543	22.6	65%
Boston	4,808	149	32.3	41%

Table 6: Comparisons between London and US Cities on OSM usage

Table 7 shows the average adjusted R^2 values for the models, over the period 2007 to 2012. As observed in London, the complete model (Equation 6) is the most accurate, with an adjusted R^2 value as high as 0.61 for Boston, suggesting that the same properties we previously tested for London are now able to explain a large portion of this city data's variability, despite its many differences (e.g., in terms of size, population density, import ratio). The adjusted R^2 value for New York City is much lower (0.22); however, what is most important to observe here is that such value is in line with what observed for the whole of London in the year 2007/08 (see Table 4 - value 0.28). Furthermore, the gain in adjusted R^2 with respect to the model that does not account for geodemographic factors is enormous ($+144\%$ for New York City and $+53\%$ for Boston). Based on these results, we speculate that adoption of OSM in the US cities under examination is still in its infancy, and as such the geodemographic properties are those that account the most for variability in the data at this stage, as it also happened in London back in the early days of OSM adoption.

5. DISCUSSION AND CONCLUSION

In this paper, we have presented a model inspired by Cellular Automata theory that leverages characteristics of the urban crowd-sourcing domain to accurately describe its digital evolution over time. The model considers properties of self-reinforcement, spatial and editing correlation, and geodemographic influences. We have used correlation analysis to support each of our properties, multiple linear regression to validate the model, and classification to measure its accuracy. We now discuss the theoretical and practical implications of this work.

[13] http://www.census.gov/geo/reference/gtc/gtc_ct.html

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

Model	*New York*	*Boston*
Eq. (3):	0.07	0.23
Eq. (4):	0.07 (+0%)	0.24 (+4%)
Eq. (5):	0.09 (+29%)	0.40 (+67%)
Eq. (6):	0.22 (+144%)	0.61 (+53%)

Table 7: Adjusted multiple R^2 of our models, together with relative improvement w.r.t. the previous model in New York and Boston (every model showed in this table is statistical significant – p-value < 0.001)

Theoretical Implications. The study reported in this paper has shown that digital mapping of spatial urban information is governed by complex dynamics, with geodemographic factors being particularly important to determine area seeding in the very early stages of technology deployment; once the bootstrapping phase is over, contagion and self-reinforcement processes explain most of the digital growth instead. This two-step process for adoption is in line with recent findings in the area of social networks, that suggest there exist two different classes of individuals who contribute to two parallel adoption processes: early participants, who start contributing and create random seeding, followed by low threshold individuals who are responsible for the subsequent contributions' spreading [4, 16, 7].

Practical Implications. Understanding '*what* factors are most important *when*' can be leveraged by designers of urban crowd-sourcing applications, so to better engineer them: from choosing who to target as early adopters of the technology, so to influence area seeding, to embedding gamification elements, so to drive contagion and self-reinforcement processes (e.g., via leader-boards, competitions, virtual rewards). Once the urban crowd-sourcing application is deployed, the model described in this paper can be used to monitor its growth. Being able to *predict* what areas that will not be digitally mapped gives businesses and public agencies time to plan and execute interventions. Such interventions may span a wide spectrum: from allocating financial resources to cover neglected areas, to gamification (e.g., in the form of competitions or mapping parties) to location-based social network features [43, 25] so to direct the crowd towards specific mapping goals. Having an accurate growth model at hand implies that these limited resources (human and/or financial) can be best allocated so to maximise return on investment. For example, using influence maximization schemes (e.g., [42]), one could decide how many resources to allocate to areas predicted to be severely neglected, so to maximise expected growth in the following year(s), both as an immediate result of investment *and* thanks to the contagion and self-reinforcement processes that our model estimates to follow.

5.1 Limitations & Future Work

Our work suffers from a number of limitations. First, we studied one specific example of urban crowd-sourcing application (OSM), on a number of cities that, even though they exhibit different characteristics, they all belong to the modern urbanised part of the world. If we were to model OSM growth in different parts of the world, different properties might have to be considered within the model. For example, our model works under the assumption that the city under examination is already urbanised and thus the physical space constraints prevents the formation of many new physical urban elements in a short period of time. This assumption does not hold for cities in developing countries, where the actual physical changes are of much higher magnitude and at a faster pace. Also, our model has been tested in a scenario where *digital* POI cov-

erage is far from complete (indeed, there are very few areas with OSM cumulative activity above 50%, as shown in Figure 2). This means there is as yet no saturation effect to be accounted for. This is also evidenced by the fact that the positive correlation of the self-reinforcement property (for both distant and recent past) increases over the years – Table 4 (i.e., the more an area has been edited in the past, the more it will be in the future). Once saturation starts to appear, segmentation analysis should be performed, so to assess our model separately in areas under-covered and areas near-complete, as the same factors (e.g., past growth) might yield opposite effects.

Second, the 'pace' at which mapping takes place in OSM is fundamentally different to that taking place during disaster recovery efforts [46], and we do not know whether the processes of self-reinforcement and contagion that our model leverages would be able to describe growth in these settings. A direction of future work is to study the suitability of the model presented in this paper in different urban crowd-sourcing settings (both in terms of different world regions, and in terms of different applications).

We presented a model of growth that focuses on 'spatial' factors (e.g., distance to the centre) and processes (e.g., spatial contagion). In so doing, we have not investigated characteristics of the adopters of the technology themselves: their motives for taking part in digital mapping, their interests (what types of POIs they map – e.g., services vs leisure POIs), and how they respond to incentives, such as the frequently organised OSM mapping parties.[15] Modelling these factors and the processes behind adopters' dynamics is a direction of future research.

Acknowledgment

This work was partly supported by the Knowledge and Innovation Community EIT ICT Labs, and by the Intel Collaborative Research Institute on Sustainable Connected Cities.

6. REFERENCES

[1] S. Alkheder and J. Shan. Cellular Automata Urban Growth Simulation and Evaluation-A Case Study of Indianapolis. In *Proceedings of the 8th International Conference on GeoComputation*, pages 1–19, 2005.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 7–15, 2008.

[3] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, December 2009.

[4] Aral, S. and Walker, D. Identifying Influential and Susceptible Members of Social Networks. *Science*, 337, 2012.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54. ACM, 2006.

[6] P. Bibby. Land use change in Britain. *Elsevier Journal on Land Use Policy*, 2009.

[7] A. Brodersen, S Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. *WWW 2012*, 2012.

[15] `http://wiki.openstreetmap.org/wiki/Mapping_Weekend_Howto`

[8] S.D. Brunn, J.F. Williams, and D.J. Zeigler. *Cities Of The World: World Regional Urban Development*. Rowman & Littlefield Publishers, 2003.

[9] D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri X. Lan. Sequential influence models in social networks. In *Proc. 4th International AAAI Conference on Weblogs and Social Media*, 2010.

[10] J. Cranshaw, R. Schwartz, J.I. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

[11] N. Ganguly, B.K. Sikdar, A. Deutsch, G. Canright, and P. Pal Chaudhuri. A Survey on Cellular Automata. Technical report, 2003.

[12] A.M. García, I. Santé, M. Boullón, and R. Crecente. A comparative analysis of cellular automata models for simulation of small urban areas in Galicia, NW Spain. *Computers, Environment and Urban Systems*, 36(4):291–301, 2012.

[13] J.F. Girres and G. Touya. Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4):435–459, 2010.

[14] R. Glott, P. Schmidt, and R. Ghosh. Wikipedia Survey–overview of results. *United Nations University: Collaborative Creativity Group*, 2010.

[15] M. Gomez-Rodriguez, D. Balduzzi, and B. SchÃűlkopf. Uncovering the temporal dynamics of diffusion networks. In *Proc. of the 28th International Conference on Machine Learning*, 2011.

[16] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The Dynamics of Protest Recruitment through an Online Network. *Scientific reports*, 1, 2011.

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

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

[19] M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. 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, 2010.

[20] P. Hedström. Contagious Collectivities: On the Spatial Diffusion of Swedish Trade Unions, 1890-1940. *American Journal of Sociology*, 99(5):1157–1179, March 1994.

[21] K. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of global social media. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[22] D. Kim and M. Batty. Calibrating Cellular Automata Models for Simulating Urban Growth. Technical Report 176, UCL Centre for Advanced Spatial Analysis, December 2011.

[23] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[24] J. Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51(11):66–72, 2008.

[25] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the wikipedians talk: network and tree structure of

wikipedia discussion pages. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*. AAAI, 2011.

[26] N. Lathia, D. Quercia, and J. Crowcroft. The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility. In *Proc. of Pervasive*, June 2012.

[27] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506, 2009.

[28] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[29] I. Ludwig, A. Voss, and M. Krause-Traudes. A comparison of the street networks of navteq and osm in germany. *Advancing Geoinformation Science for a Changing World*, 1(2):65–84, 2011.

[30] A. Mashhadi, G. Quattrone, and L. Capra. Putting Ubiquitous Crowd-sourcing into Context. In *Proceedings of ACM CSCW*, 2013.

[31] A. Mashhadi, G. Quattrone, L. Capra, and P. Mooney. On the Accuracy of Urban Crowd-Sourcing for Maintaining Large-Scale Geospatial Databases. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym'12)*. ACM, 2012.

[32] P. Mooney and P. Corcoran. Who are the contributors to openstreetmap and what do they do? In *Proc. of GISRUK*, 2012.

[33] S.A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 33–41, 2012.

[34] P. Neis, D. Zielstra, and A. Zipf. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany. *Future Internet*, 4(1):1–21, 2011.

[35] K. Panciera, R. Priedhorsky, T. Erickson, and L. Terveen. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the 28th International Conference on Human factors in computing systems (CHI)*, pages 1917–1926. ACM, 2010.

[36] K.A. Panciera, M. Masli, and L. Terveen. "how should i go from _ to_ without getting killed?": motivation and benefits in open collaboration. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration (WikiSym'11)*, pages 183–192, 2011.

[37] R. Priedhorsky, M. Masli, and L. Terveen. Eliciting and focusing geographic volunteer work. In *Proceedings of the 13th International Conference on Computer Supported Cooperative Work (CSCW'10)*, pages 61–70. ACM, 2010.

[38] R. Priedhorsky, B. Jordan, and L. Terveen. How a personalized geowiki can help bicyclists share information more effectively. In *Proceedings of the 3rd International Symposium on Wikis and Open Collaboration (WikiSym'07)*, pages 93–98, 2007.

[39] D. Pullar and C. Pettit. Improving Urban Growth Forecasting with Cellular Automata: A Case Study for Hervey Bay. In *International Congress on Modelling and Simulation*, pages 14–17, 2003.

[40] D.M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 695–704, 2011.

[41] C. Roth, S.M. Kang, M. Batty, and M. BarthÃl'lemy. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1), 01 2011.

[42] Y. Singer. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proc.of the 5th ACM International Conference on Web Search and Data Mining*, pages 733–742, 2012.

[43] Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner. The length of bridge ties: structural and geographic properties of online social interactions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*. AAAI, 2012.

[44] G. U. Yule. A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London*, 213:21–87, 1925.

[45] D. Zielstra and A. Zipf. A comparative study of proprietary geodata and volunteered geographic information for germany. In *Proceedings of the 13th International Conference on Geographic Information Science*, 2010.

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