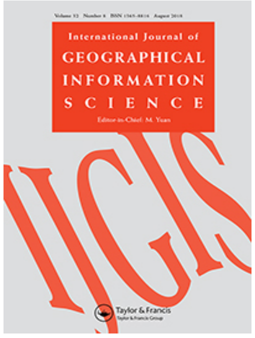

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: https://www.tandfonline.com/loi/tgis20

The life cycle of contributors in collaborative online communities -the case of OpenStreetMap

Daniel Bégin, Rodolphe Devillers & Stéphane Roche

To cite this article: Daniel Bégin, Rodolphe Devillers & Stéphane Roche (2018) The life cycle of contributors in collaborative online communities -the case of OpenStreetMap, International Journal of Geographical Information Science, 32:8, 1611-1630, DOI: 10.1080/13658816.2018.1458312

To link to this article: https://doi.org/10.1080/13658816.2018.1458312

Published online: 16 Apr 2018.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2018, VOL. 32, NO. 8, 1611–1630
https://doi.org/10.1080/13658816.2018.1458312

RESEARCH ARTICLE

The life cycle of contributors in collaborative online communities -the case of OpenStreetMap

Daniel Bégin[a], Rodolphe Devillers[a,b] and Stéphane Roche[b]

[a]Department of Geography, Memorial University of Newfoundland, St. John's, NL, Canada; [b]Centre de Recherche en Géomatique, Université Laval, Québec, QC, Canada

ABSTRACT

Over the last two decades, online communities have become ubiquitous, with millions of people accessing collaborative project websites every day. Among them, the OpenStreetMap project (OSM) has been very successful in collecting/offering volunteered geographic information (VGI). Very different behaviours are observed among OSM participants, which translate into large differences of lifespan, contribution levels (e.g. Nielsen's 90–9-1 rule) and attitudes towards innovations (e.g. Diffusion of innovation theory or DoIT). So far, the literature has defined phases in the life cycle of contributors only based on the nature of their contributions (e.g. role of participants and edits characteristics). Our study identifies the different phases of their life cycle from a temporal perspective and assesses how these phases relate to the volume and the frequency of the contributions from participants. Survival analyses were performed using both a complementary cumulative distribution function and a Kaplan-Meier estimator to plot survival and hazard curves. The analyses were broken down according to Nielsen and DoIT contributors' categories to highlight potential explanatory variables. This paper shows that two contribution processes combine with three major participation stages to form six phases in contributors' life cycle. The volume of edits provided on each active day is driven by the two contribution processes, illustrating the evolution of contributors' motivation over time. Since contributors' lifespan is a universal metric, our results may also apply to other collaborative online communities.

ARTICLE HISTORY
Received 30 November 2017
Accepted 24 March 2018

KEYWORDS
OSM history; behaviour; lifespan; survival analysis; diffusion of innovation theory

1. Introduction

Online communities have become ubiquitous features of today's life. Well-known communities have developed around social networking (e.g. Facebook, Instagram and LinkedIn), while others have focussed on knowledge sharing projects such as Wikipedia, birdwatching (e.g. Audubon) or mapping (e.g. OpenStreetMap). Millions of people visit those websites every day and the scientific community is increasingly referring to these communities as being both an important source of data and a valuable work force (Kimura and Kinchy 2016, Michelucci and Dickinson 2016). These collaborative projects require a continuous flow of new participants to compensate for those who withdraw after being active for a period of time.

CONTACT Daniel Bégin d.begin@mun.ca

In a previous study on volunteered geographic information (VGI) contributors (Bégin *et al*. 2017a), we observed that the rate at which people enrol in an online community depends on two main factors: interest in, and awareness of, a project. A project triggers the interest of people because of its appealing objectives (Chacon *et al*. 2007, Budhathoki *et al*. 2010, Nov *et al*. 2011), the nature of the tasks (Hemetsberger and Pieters 2003, Houle *et al*. 2005, Borst 2010) or because people foresee their participation as being potentially enjoyable (Budhathoki *et al*. 2010, Nov *et al*. 2011). Awareness about a project, whether online or not, usually comes from credible acquaintances, colleagues or friends (Rogers 1983, Brown and Reingen 1987, Nov *et al*. 2014, Rotman *et al*. 2014). When a project succeeds, awareness may also come from mass media, blogs or conferences. We also demonstrated that the many events that dot the history of a project have an influence on the number of people that register and contribute to it (Bégin *et al*. 2017a).

In another study (Bégin *et al*. 2017b), we found that the rate at which VGI contributors withdraw from an online community depends on project's ability to fulfil contributors needs, desires or aspirations (Clary *et al*. 1998, Penner 2002, Nov 2007, Budhathoki 2010), and the time it takes them to lose interest in the project. Factors such as required knowledge and skills, community norms and rules and other participants' behaviours may discourage most new contributors from pursuing their participation beyond the first few days. The same study (Bégin *et al*. 2017b) proposed three overarching stages in the life cycle of contributors adapted from the life cycle of complex systems in reliability engineering (Wang *et al*. 2002). First, an 'assessment' stage (i.e. early defects) over which a majority of participants withdraw after having estimated the costs and benefits of contributing to the project. Second, an 'engagement' stage (i.e. useful life), in which participants often contribute for years. Finally, a 'detachment' stage (i.e. wear out), being a period over which the rate of withdrawal increases exponentially. These stages are spread over two periods that are characterized by distinct contribution processes (Wang *et al*. 2002, Danescu-Niculescu-Mizil *et al*. 2013). These life cycle stages reflect a learning-adaptation dominated process in which active contributors seek to adhere to evolving norms and tools, followed by a cumulative-damage dominated process in which contributors adopt a more conservative attitude and eventually withdraw (Bégin *et al*. 2017b).

The literature has so far characterized contributors' life cycle based solely on the nature of their contributions. For instance, according to contributors' motivations, knowledge and skills (Coleman *et al*. 2009), their roles in the community (Bryant *et al*. 2005, Cheung *et al*. 2005, Preece and Shneiderman 2009) or the volume of their edits (Neis and Zipf 2012).

Very different levels of participation are observed between participants. This participation inequality is expressed by the 90–9-1 rule proposed by Nielsen (2006), in which 90% of the members do not contribute much data (Schneider *et al*. 2013, Sun *et al*. 2014), 9% contribute sporadically, and the remaining 1% produces most of the content (Ochoa and Duval 2008, Neis and Zielstra 2014, Ma *et al*. 2015). The history of a project has also an influence on the level of participation. Most collaborative online projects have been created over the last two decades, offering new ways of sharing information. In this context, the profile of participants might evolve over time as predicted by the Diffusion of Innovation Theory (DoIT). This theory (Rogers 1983) describes how an innovation diffuses through a population and characterizes participants according to the time at which they adopt the innovation during its diffusion process.

Unravelling the phases of contribution in an online project could help determine at what point in time the properties of contributions are likely to change (e.g. volume, content and quality). To the extent that both are related, this could shed a new light on the structure of VGI contributions, so valuable to GIScience. In order to better understand contributors' life cycle, its phases, and the potential relationships these phases may have with the data they provide, we analysed both the lifespan and the contributions of the OpenStreetMap (OSM) participants. The distribution of contributors' lifespan was examined over years and survival analyses were performed to identify the different phases of their life cycle. Potential changes in the nature of their contributions were assessed for both the frequency and the volume of contributions at each phase. In addition, the analyses were broken down according to both the Nielsen (90–9-1) and DoIT categories to understand the effects they may have on our results. Finally, we examined whether contributors' life cycle phases are defined by the time the participants spend in the project or the number of days they are actually active.

This paper analyses the life cycle of online contributors from a temporal perspective. Section 2 describes the methods used to study contributors' lifespan and to identify both the phases of their life cycle and the nature of their contributions during each phase. Section 3 presents the evolution of contributors' lifespan over years and identifies specific events that seem to have had an impact on the project. The results from survival analyses are presented and broken down according to both DoIT and Nielsen categories. The different phases of contributors' life cycle are presented in detail and the impacts they have on the nature of the contributions are described. Finally, Section 4 discusses the evolution of contributors' lifespan and the nature of their contributions according to the different phases of their life cycle.

2. Materials and methods

OSM is a large collaborative project that aims to build a comprehensive map of the world. The OSM community uses a Wiki approach to create and improve the map by collecting local knowledge from its members (Mooney and Corcoran 2012, Napolitano and Mooney 2012, Bright *et al*. 2017). OSM has been widely studied by the GIScience community to understand key questions about VGI, both because of OSM's success and also because the documentation of the project is easily accessible (Sui *et al*. 2013, Arsanjani *et al*. 2015, Capineri *et al*. 2016).

A history dump file released by OSM on 1 September 2014, was downloaded for the purpose of this study (OpenStreetMap contributors 2014a). The file contained all the contributions made to the project over years. In addition to the contributions, the file included the virtual containers (i.e. changesets) in which the edits were supplied, identifying both the temporal and the geographical extents of each editing session, as well as the identifier of the contributor who made them. A detailed chronicle of the project history is maintained by OSM contributors (OpenStreetMap contributors 2014b) and was consulted when required. FME workbenches (Safe software 2015.0) were developed to obtain contributors' registration timestamps from OSM website and to extract and load the 1 TB history dump file to a PostgreSQL v.9.3 database. Statistical analyses and visualizations were performed using the R software v.3.4.1.

First, the Nielsen 90–9-1 rule was used to categorize participants as being either 'prolific' (i.e. having contributed 90% of the edits), 'casual' (i.e. the following 9%) or 'inactive' (i.e. remaining 1%). A cumulative sum of edits was then assigned to each contributor after having ordered them based on the volume of their respective edits, from largest to smallest. Second, DoIT categories were assigned to contributors based on the results from previous studies (Bégin *et al*. 2017a, 2017b). Contributors that registered prior to 2007 were identified as 'innovators', 'early adopters' were those who registered from 2007 to 2009 and 'early majority' was assigned to those who registered after 2009. The number of days between the history dump file creation and contributors' registration dates was computed to use the same reference system as their lifespan (i.e. days since the first contribution).

Some survival analyses require differentiation between withdrawn and active contributors. In order to make this distinction, we used a systematic method that identifies contributors' status from a statistical analysis of individuals' contributions (Bégin *et al*. 2017b). Withdrawn contributors were identified by comparing the time passed since their last contribution with the longest period of inactivity expected from their contribution history. The Chebyshev theorem was applied to the time spent between contributions estimating maximum duration with a 95% probability. Periods of inactivity were computed in days after having removed biases induced by contributors' location and time zone.

2.1. *Contributions' span over time*

A scatterplot of contributors' first and last edits was created to visualize general trends in the contribution spans over years. A contribution span is defined as the time interval between the first and last edits made by a contributor. The opening timestamp of the first changeset was used for the time of the first edit, and the closing timestamp of the last changeset was used as last edit. Previous studies linked variations in the number of new and withdrawn contributors (i.e. variations of lifespan) with specific events that dotted the history of the project (Bégin *et al*. 2017a, 2017b). Consequently, a plot of contributions' span over time should help highlight the impact that specific events may have had on the life cycle of contributors. Due to the large number of contributors, we used R's 'smoothScatter' procedure that plots kernel density estimates instead of actual data points (R Core Team 2016). Contrary to standard scatterplots, the density scatterplot shows the relative number of contributors represented by each point.

2.2. *Survival analysis*

Two complementary survival analyses were used to identify phases in contributors' life cycle.

First, a complementary cumulative distribution function (CCDF) measured for each contributor the proportion of participants whose lifespan was greater and plotted the results by lifespan. This function, also called 'Survival function', was used to globally assess contributors' lifespan, without discriminating between active and withdrawn participants. Inflection points on the CCDF graph were expected to show changes in contributors' engagement in the project. The analysis was run using an empirical cumulative distribution function from R software (2016). The CCDF distribution was plotted using different scales on each axis to support a manual identification of inflection points.

Second, a Kaplan-Meier estimator modelled contributors' survival rates by considering the status of each contributor (i.e. withdrawn or not). The Kaplan-Meier estimator was computed using the 'survfit' procedure from R's 'survival' package (Therneau and Lumley 2017). Participants that were considered as being active at the time the history dump was created were 'censored' for the procedure to consider their lifespan as incomplete. Both survival and hazard function curves were derived from the analysis. Survival curves show the proportion of participants that are still active, while the hazard function curves show the daily rates of withdrawal. Hazard function curves are of particular interest since they often illustrate the different phases in the life cycle of the studied phenomena (Wang *et al*. 2002, Weon 2016). Those curves were filtered using a moving average over a 30-day window and inflection points were identified manually on the resulting curves. The effects Nielsen and DoIT categories may have had on the life cycle were examined using strata analyses, breaking down the Kaplan-Meier analysis using these categories.

2.3. *Identification of contributors' life cycle phases*

The life cycle phases of OSM contributors were identified by comparing the inflection points found on the curves resulting from both survival analyses. Since our data are empirical and no theoretical model could have located these inflection points, the points were identified manually.

Different metrics were defined to characterize each phase. First, the proportion of contributors that completed a given phase was established using Kaplan-Meier survival rates. The number of members belonging to each phase resulted from classifying all contributors based on the time they spent in the project. Estimating the number of active participants used the same process, but counting only those who were still considered as being active. The formal evaluation process used to differentiate withdrawn from active contributors may require long delays before confirming a contributor has left a project with 95% probability, particularly when they contributed only on few occasions (Bégin *et al*. 2017b). The proportion of active contributors (PAC) was then expected to be overestimated for early phases; to address this, an alternative estimation was used, as described in Equation (1).

$$\mathrm{ACp} = \mathrm{Nc} * \mathrm{Ep} * \mathrm{Sp}, \tag{1}$$

where p is the phase considered, ACp is the number of active contributors at the end of phase p, Nc the number of new contributors per day, Ep is the number of days since first edits at the end of phase p, and Sp is the survival rate at the end of phase p. The smallest estimates of active contributors derived from both methods (i.e. formal and alternative) were applied to each phase of the OSM project.

2.4. *Volume and frequency of contributions*

In order to assess the volume and the frequency of contributions at each phase, the number of edits and the number of days since previous contribution were registered for each users and each active day. However, assessing contributions from the number of edits or the number of active days would introduce biases since both are correlated to contributors' lifespan and the duration of each phase. Instead, a volume ratio (edits per active day) and a frequency ratio (days between edits) were calculated. With this method, a specific behaviour produces the same results, regardless of contributors' lifespan or duration of the phase. Boxplot procedures were used to analyse the nature of contributions in order to consider their long-tail distributions. Finally, since 'prolific' participants generate 90% of the data, their contributions were also evaluated separately from the overall population.

3. Results

The history dump file retrieved from the OSM website spanned over 3433 days. It contained 25.1M changesets related to 464,857 accounts, considered herein as distinct contributors. Within these changesets, 8381 had no associated contributors and were not used in the analysis. Overall, 58% of OSM contributors have withdrawn over years. The breakdown of both DoIT and Nielsen categories is presented in Table 1.

Regarding DoIT, the proportions are expected to evolve since other categories of participants should join the project over years. For Nielsen's categories, the proportions we obtained for OSM participants were surprisingly similar to the expected 90–9-1 values.

3.1. *Contributions' span over time*

The scatterplot of OSM contributors' lifespan (Figure 1) was used to better understand general patterns of contributors' life cycle, as well as the impact of specific events on the recruitment and the withdrawal of contributors.

Figure 1 is characterized by a dark diagonal and variations of density over the vertical and horizontal axes. The diagonal highlights the fact that a large number of participants contributed only for a very short period of time after enrolling in OSM. Shading density generally increases from left to right, showing a cumulative growth in the number of OSM contributors over time. The darker line at the top of the graph represents active contributors at the time of the history dump. Vertical lines show specific peaks in recruitment, followed by a gradual withdrawal of corresponding participants. Horizontal lines show bursts of participants withdrawing from the project at a specific time. Some noteworthy variations of density (i.e. labels A-H) were linked to specific events in the project's history (Table 2).

The span of some of horizontal lines (Figure 1) indicates that only the older contributors were affected by the exclusion of users who did not agree to the new ODbL license (E) and the subsequent collective withdrawals that followed, up to 2013. The later contributors, who agreed to the new licence when joining the project (D), did not seem concerned by these events (i.e. horizontal darker lines ending on D).

Table 1. OSM contributors according to DoIT and Nielsen categories, the number of participants (N) and their proportion (%).

DoIT classification	N	%	Nielsen classification [a]	N	%
Innovators	1453	0.31%	Prolific (90% of data)	8189	1.76%
Early adopters	49,866	10.73%	Casual (9% of data)	41,722	8.98%
Early majority	413,538	88.96%	Inactive (1% of data)	414,946	89.26%

[a] The names provided here illustrate the nature of contributions according to Nielsen.

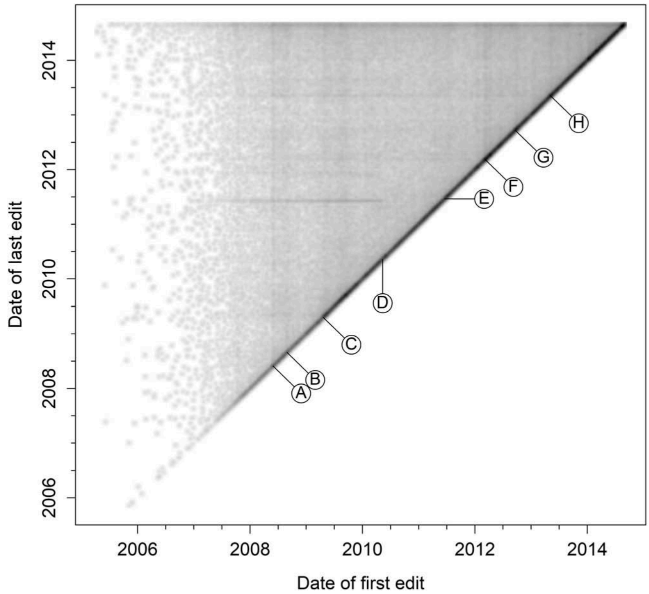

Figure 1. Density of OSM contributors' first and last edits over time. Each point represents about a week of contribution. The larger the number of contributors, the darker the colour is. Some noteworthy density variations are identified using labels (see Table 2 for details).

Table 2. Events related with notable variations of graph density. The 'Effect' column characterizes the effects these events may have had on the number of OSM contributors as being either positive (+), negative (-) or both (*).

Id	Date	Effect	Event description
A	2008–05-30	+	The German journal *Der Spiegel* compares OSM to Wikipedia
B	2008–08-29	+	The BBC[a] quotes the president of the BCS[a] being positive about OSM
C	2009–04-21	*	API v. 0.6 brings changesets and drops anonymous edits
D	2010–05-12	+[b]	New users must now agree to the new ODbL Licence [c] to register
E	2011–06-19	-	Established users who declined the Licence are excluded from OSM
F	2012–03-08	*	ArcGIS Editor for OpenStreetMap is made available
G	2012–09-20	*	Import guidelines require dedicated accounts
H	2013–05-08	*	New ID editor is made available on the OSM website

[a] BBC: British Broadcasting Corporation; BCS: British Cartographic Society
[b] This event did not increase enrolment but has limited subsequent withdrawals after some users were excluded from OSM (E).
[c] OSM switched to an Open Database Licence (ODbL) after a lengthy process that lasted almost four years.

3.2. *Survival analysis*

The results of the CCDF analysis are presented in Figure 2.

Six inflection points were identified on the CCDF, bounding seven periods when contributors showed similar patterns of withdrawals (i.e. relatively constant slopes on the graphs). Figure 2(a) shows an abrupt drop of contributors starting exactly 1 h after enrolment in the project, and continuing for a few hours. The proportion of remaining contributors declined from 80% to 50% during this very short period. After 24 h, the proportion of contributors that remained active declined to 40% and the slope becomes constant until it reaches 6 months. Figure 2(b) shows a constant slope from 6 months to about 4.5 years,

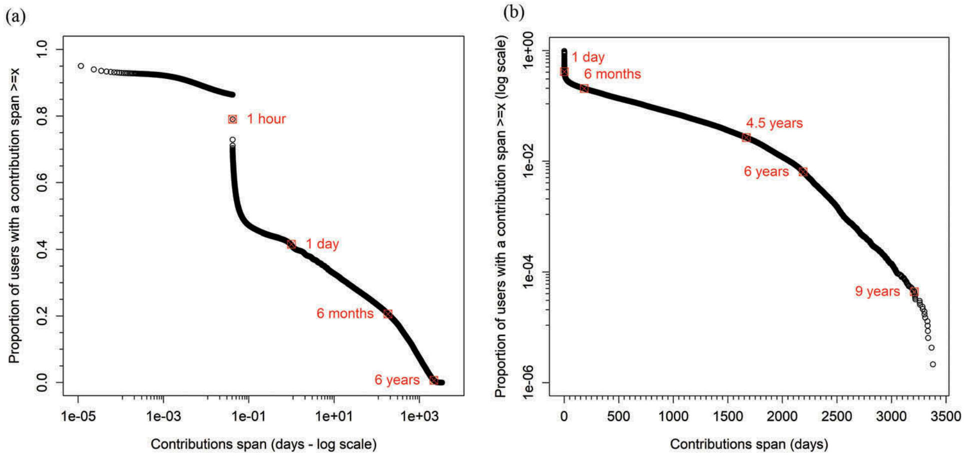

Figure 2. Complementary cumulative distribution function (CCDF) of contributions' span. Short contributions were enhanced by applying a logarithmic scale on the X axis (a) while a logarithmic scale on the Y axis enhanced longer contributions (b).

after which the slope increases until it reached 6 years. From this point, the slope remains relatively constant until it starts increasing again after 9 years.

CCDF inflection points of Figure 2 were reported on Kaplan-Meier analysis results (Figure 3) to combine both results in order to identify phases in contributors' life cycle.

The hazard curve (Figure 3(b)) shows a bathtub shape typical of the stages in the life cycle of complex systems. These three stages, described earlier, are identified on the figure. The 'assessment' stage includes both the first (1 h) and second (1 day) inflection points from CCDF (merged in the first symbol). The boundary between the 'assessment' and 'engagement' matches the third inflection point (6 months). During the 'engagement' stage, withdrawal

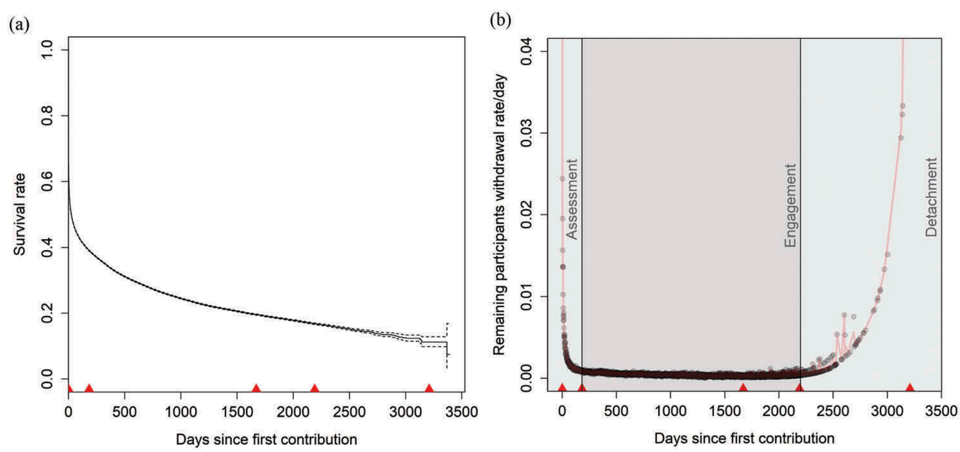

Figure 3. Survival curves from Kaplan-Meier estimators on the entire OSM population, with (a) Survival rates over time with confidence intervals, and (b) the daily withdrawal rates (Hazard curve) with the three stages from Bégin *et al.* (2017b). Locations of CCDF inflection points are reported on both X axes (red triangles).

rates are low and almost constant (i.e. about 15% per year). During this stage, the lowest rate of withdrawal matches the inflection point found at 4.5 years, where contributors potentially switch from a learning-adaptation to a cumulative-damage dominated behaviour (Bégin *et al*. 2017b). This switch seems to occur after years of irritation and annoyance start affecting participants' motivation to keep contributing, which leads to the 'detachment' phase. The boundary between 'engagement' and 'detachment' stages fits the location of the next CCDF inflection point (6 years). Finally, the 'detachment' stage contains the last CCDF inflection point (9 years), a point after which most long-term contributors quit the project. The curves were stratified according to Nielsen (Figure 4) and DoIT categories (Figure 5).

Figure 4(a) (DoIT) and Figure 5(a) (Nielsen) show very distinct survival rates, although the shape of their curves remains similar, except for 'innovators' and 'prolific' contributors. While the survival curve of 'innovators' is relatively linear, the 'prolific' contributors are characterized by a convex survival curve rather than a concave one. The curve eventually converges

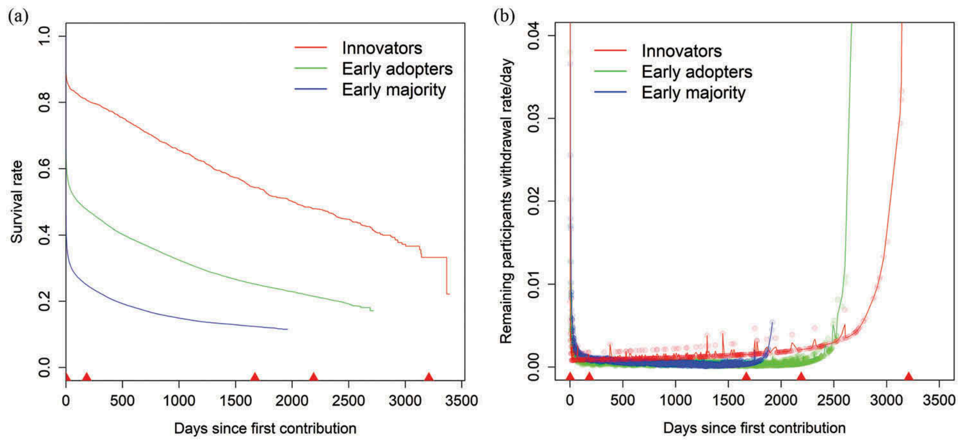

Figure 4. Survival curves from Kaplan-Meier estimators stratified by DoIT categories, where (a) illustrates the survival rates, and (b) shows daily withdrawal rates.

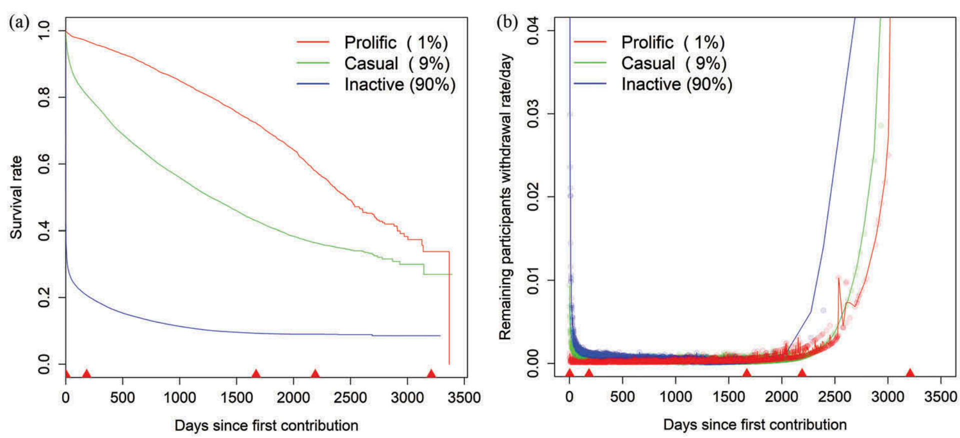

Figure 5. Survival curves from Kaplan-Meier estimators stratified by Nielsen's categories, where (a) illustrates the survival rates, and (b) shows daily withdrawal rates.

towards 'casual' mappers around the last inflection point (9 years). Interestingly, the curves of both 'inactive' and 'early majority' contributors stabilize around the fourth CCDF inflection points (4.5 years). The stair case effect, visible on both graphs, results from a smaller number of contributors as we get closer to the last inflection point.

The daily rates of withdrawal from both DoIT (Figure 4(b)) and Nielsen (Figure 5(b)) categories reproduce a pattern that is very similar to the one of Figure 3(b), with only slight variations between categories. The various extents of the curves from Figure 4(b) (DoIT) were expected by definition. However the truncated 'assessment' stages of 'innovators' (Figure 4(b)) and 'prolific' contributors (Figure 5(b)) were not anticipated.

3.3. *Identification of contributors' life cycle phases*

Both the CCDF (Figure 2) and the hazard curves (Figure 3(b)) were expected to be complementary, but they rather mutually confirmed each other. The CCDF provided clear inflection points identification while the hazard curve provided meaningful context to interpret each phase. Six phases covering the life cycle of OSM contributors were identified from these points. The phases were identified without ambiguity since both methods agreed on their approximate location. The curves from Kaplan-Meier analysis on DoIT and Nielsen classifications provided clues about the underlying structure of contributors' life cycle. The resulting phases are presented below (Table 3) with the volume and the frequency of contributions made at each phase.

The first phase (visitors) results from combining the first two segments from the CCDF analysis (Figure 2(a)). The lifespan of OSM participants was measured from the changesets they provided, which in turn depends on the OSM application programming interface (API). The OSM API applies constraints regarding the time over which a changeset has been opened, by automatically closing it either after being inactive for 1 h, or after being active for 24 h. Since the first inflection point of Figure 2(a) is found at exactly 1 h, it is most probably a consequence from API operations, and the point was excluded from the analysis.

Since the boundaries of each phase were determined manually, their locations are approximate, particularly regarding the later phases. In these cases, contributors' withdrawal could

Table 3. Detailed description of the phases in the life cycle of the OSM contributors. Phases' name aims at characterizing contributors' lifespan and/or behaviour. Definitions of each column are provided as footnotes.

Phase	End	Span	Rate	Members	Active	PAC	MV	MF
Visitors	1 day	1	65%	263,848	281 [a]	<1%	4	NA
Explorers	6 months	182	39%	105,262	20,088 [a]	25%	30	4
Adopters	4.5 years	1487	20%	83,357	48,604	61%	56	77
Veterans	6 years	520	17%	9411	8338	10%	125	47
Elders	9 years	1019	11%	2957	2751	3%	189	20
Founders	NA	NA	NA	22	21	<1%	242	14

End: estimated termination of the phase since contributors' first edits.
Span: duration of the phase (days).
Rate: survival rates at the end of the phase according to Figure 3(a).
Members: number of contributors belonging to the phase according to their lifespan.
Active: number of active members at the time of the history dump.
PAC: proportion of all active members belonging to the phase.
MV: median volume of edits over members' whole lifespan (edits per active day).
MF: median frequency of edits over members' whole lifespan (days between contributions).
[a] Value adjusted for withdrawal uncertainty over first 591 days (see explanations in the text).

occur several months before or after the observed dates without significantly changing our results. In Table 3, the number of active participants at each phase (Active) is derived from the sum of participants and the number of those who withdrew. These withdrawn participants were identified with a 95% probability by using a time threshold since their last contribution. As a result, one-time contributors were considered active until they reached 591 days without contributing, even if 70% of them will have withdrawn at that time (Figure 3(a)). Consequently, the numbers of active 'visitors', 'explorers' and 'adopters' were potentially overestimated since their phases extend beyond that threshold. Equation (1) was considered to provide more realistic estimations of active participants in these categories, using the latest trend in new contributors' enrolment (281 participants/day). 'Visitors' and 'explorers' phases were found to be overestimated using this evaluation and their values were replaced. The PAC refers to the sum of all active contributors at the time of the history dump.

3.4. *Volume and frequency of contributions*

A boxplot analysis looked at the contributions made by participants at each phase of their lifespan in the project. The contributions of each participant were distributed over each corresponding phase. The first analysis looked at the contributions made by all OSM participants and the results are presented in Figure 6.

Figure 6(a) shows that the average number of edits per active day is relatively constant over all phases (approximately 52 edits), with the exception of the first day (visitors; characterized by 12 edits). The range of outlier values decreased over time. Figure 6(b) shows that the average time spent between contributions increased up to the 'adopters' phase, before

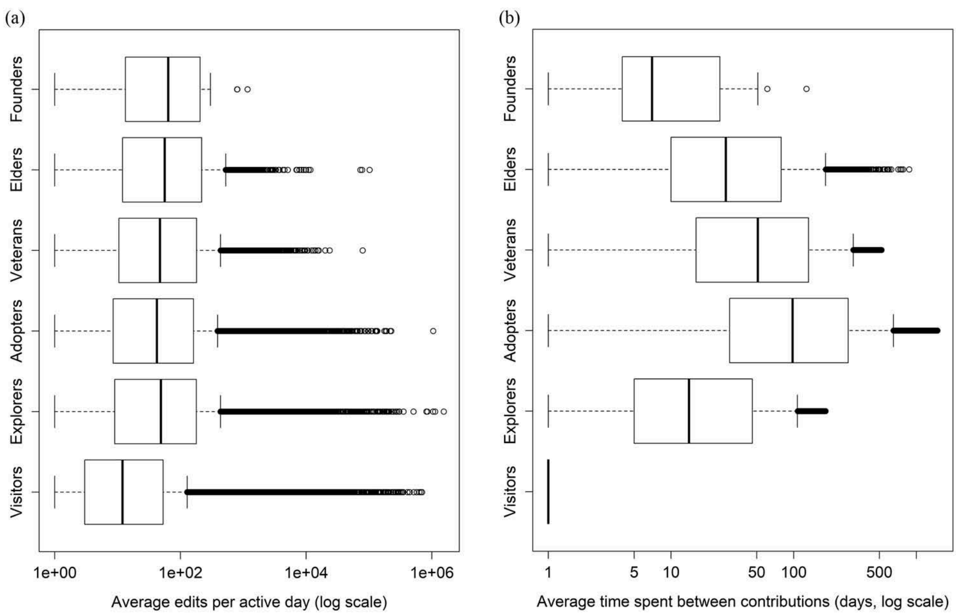

Figure 6. Volume and frequency of contributions made by all OSM participants according to the phase they belong to at the time of contributions, where (a) shows the volume of edits (edits per active day) and (b) the frequency of edits (time spent between active days).

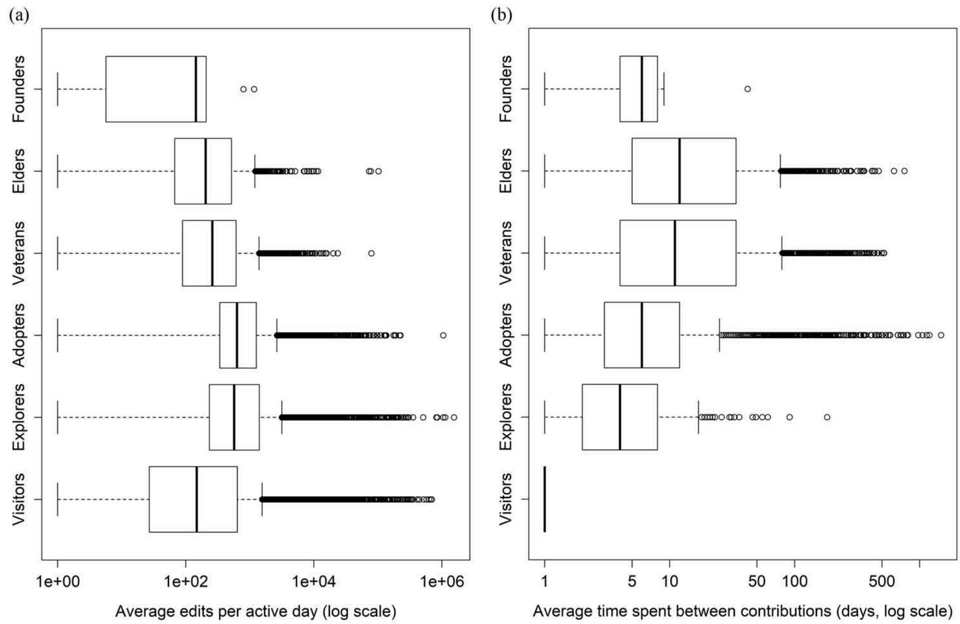

Figure 7. Contributions made by 'prolific' participants according to the phase they belong to at the time of contributions, where (a) shows the volume of edits (edits per active day) and (b) indicates the frequency of edits (time spent between active days).

decreasing in later phases. A detailed analysis has shown that the maximum daily rate of edits occurred after about 20 active days, regardless of the contributors category.

The same analyses were conducted on a subset made up of 'prolific' contributors (Figure 7). The volume of edits per active day (Figure 7(a)) increased and reached its maximum at the 'adopters' phase before decreasing in the following phases. As expected from prolific contributors, the median values were much higher than for other contributors (Figure 6(a)), but the spans of outliers were similar. Figure 7(b) shows that the frequency of edits generally decreased (i.e. the time spent between contributions increases), except over the last phase ('founders'). The spans of outliers were much larger than in Figure 6(b), particularly over the 'adopters' phase.

In order to better understand the nature of contributions from the participants, the results were also broken down for each phase using Nielsen's categories (Table 4).

These results illustrate that sustained contributions are not only the fact of highly productive contributors since the other participants also contribute several times a year.

4. Discussion

Earlier studies have proposed different classifications to describe phases in the life cycle of online contributors, based on contributors' motivations, knowledge and skills (Coleman *et al.* 2009), or their roles in the community (Bryant *et al.* 2005, Cheung *et al.* 2005, Preece and Shneiderman 2009). However, none of those classifications

Table 4. Contributions from participants at each phase of their life cycle. The volume of edits (average edits per active day) and the frequency of edits (average time spent between edits) are broken down using Nielsen's classification.

	Volume of edits			Frequency of edits (days[a])		
Phase	Prolific	Casual	Inactive	Prolific	Casual	Inactive
Visitors	149	140	9	1	1	1
Explorers	569	233	20	4	10	18
Adopters	630	136	14	6	42	188
Veterans	259	45	11	11	57	110
Elders	206	38	10	12	36	72
Founders	144	51	12	6	8	24

[a] Frequency of edits is expressed as the number of days spent between edits (active days) where the larger the number is, the less often the participants contributed.

clearly linked contributors' behaviours to the time they spent in a project or the number of days they actually contributed. In a previous study we suggested that the life cycle of OSM contributors exhibited three important stages (Bégin *et al*. 2017b). In this study, we confirmed these three stages and further subdivide them into six distinct phases, providing the first detailed analysis of temporal patterns in OSM contributors' lifespan. Stages and phases of OSM contributors' life cycle are summarized in Figure 8.

Furthermore, we found that contributors' lifespan and contributions are driven by two processes (i.e. earning-adaptation and cumulative-damage) very similar to those described by Danescu-Niculescu-Mizil *et al*. (2013). These authors were able to identify the transition from first to second process to estimate contributors' remaining lifespan. Therefore, it might be possible to identify this transition for OSM contributors and then predict their remaining lifespan.

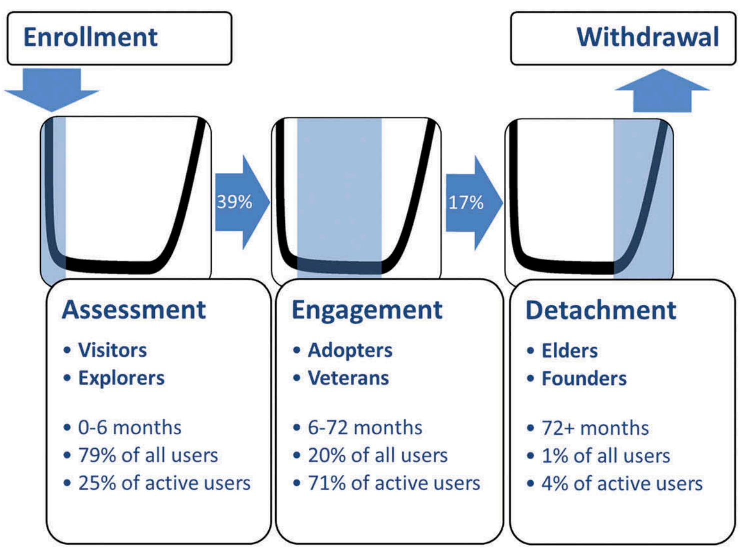

Figure 8. The life cycle of OSM contributors, from enrolment to withdrawal. The three stages of Figure 3(b) are presented with associated phases, stage duration and the proportion of contributors associated to the stage. Arrows show the proportion of contributors reaching the next stage.

4.1. Phases description

In this study, two distinct analyses have corroborated the different phases of OSM contributors' life cycle. The results show strong evidence that contributors' lifespan follows six distinct phases that reflect the evolution of contributors' motivation and interests in the project. However, some results also suggest that later phases' boundaries may continue to evolve over the years.

Metaphorically, the 'visitors' phase could be seen as dipping one's toe in a project (Preece and Shneiderman 2009), a first impression that drives about two third of participants to come back to the project. However, using DoIT categories as temporal stratification, we see that this proportion ranged from 90% for 'innovators' to 46% with the 'early majority' (Figure 4(a)). Paradoxically, 'innovators' enrolled over a period of time when contributing to the project was much more complex than it is for current 'early majority' participants. We previously suggested that while learning to contribute, participants may be less inclined to withdraw from the project (Bégin *et al*. 2017b). This has been seen in communities of practice (Lave and Wenger 1991, Wenger 1998) where new participants slowly move from the periphery to the core of an activity by learning from others. The improvements made to APIs and editing applications over years may have enabled many 'early majority' participants to be autonomous and to reach the core of the activity much faster than previously possible. Consequently, this retention mechanism may no longer apply, resulting in an immediate decision to disengage from the project when expectations are not met.

During the 'explorers' phase, participants assess the fragile equilibrium between engagement and withdrawal (Nov *et al*. 2011, Aknouche and Shoan 2013), balancing the costs (e.g. time invested and learning efforts) with benefits (e.g. pleasure and outcome usefulness) of contributing to the project. This phase is crucial in determining contributors' engagement in a project since most of those who go through the phase will stay in the project for years (this forms the 'assessment'/'engagement' boundary).

For most participants who withdraw at the explorer phase, the decision to quit the project is made over the first few weeks even though the rate of withdrawals stabilizes only after 6 months (Figure 3(b)). Looking at DoIT categories (Figure 4(b)), we found that the duration of the 'explorers' phase has changed over time. During the early years of OSM ('innovators') this phase barely existed, while its duration increases for 'early adopters' (2007–2009) and for 'early majority' participants (2009–at least 2014). Several reasons can be invoked to explain the apparent absence of 'explorers' and 'adopters' phases for early 'innovators'. Among others, a recent study shows that early OSM participants ('innovators') took on average 2 years to contribute after having registered to OSM project and suggests they may have experienced their 'assessment' stage otherwise (Bégin *et al*. 2017a, Figure 5). This long delay may also have excluded de facto most of those whose expected lifespan was shorter according to our results. Globally, the proportion of withdrawal at the end of the phase is about 60%, which includes only 20% of 'innovators' but almost 75% of 'early majority' participants.

The next two phases (i.e. 'Adopters' and 'veterans') capture the long-term 'engagement' of OSM contributors (Figure 8). Over almost 6 years, the daily rates of withdrawal remained low, with less than 17% of remaining contributors quitting the project each year. The boundary between both phases is illustrated in Figure 3(b) where the slope of the curve switches from negative to positive after 4.5 years, when the contribution processes change. Again, phase duration could lengthen over time as illustrated by Figure 4(b).

The 'elders' phase coincides with the beginning of the 'detachment' stage when the odds that a contributor will withdraw from the project increases exponentially. The upper limit of the phase is expected to increase in the future, as the maximum lifespan of OSM participants has not been reached yet (i.e. Figure 3(a) does not end at 0%).

Finally, the 'founders' phase seems to be an artefact of the recent history of the project and concerns only about 20 contributors. The small appendix at the end of the CCDF curve (Figure 2(b)) may not show the last 'survivors' of the project but rather its initiators: people that have a special attachment to the project. As they eventually withdraw from the project, we expect the corresponding segment to disappear from the CCDF curve and from OSM contributors' life cycle phases.

In summary, our results suggest that the different phases of the life cycle apply regardless of the volume of edits the contributors provide (Nielsen's classification) or the phase of the diffusion of the project at the time they enrol (DoIT). The next DoIT category of participants to enter the project should not impact the phases except for their duration. According to DoIT, the next type of contributors that should be interested in the project is the 'late majority'. The different personality traits proposed by Rogers (1983) describe late majority participants as conservative people that believe far more in tradition than in progress (Moore 2001, p. 34). In a context where OSM is not really considered as a conventional map provider, it suggests that this 'late majority' should not constitute the mainstream of OSM contributors yet. However, and paradoxically, this 'open' project may eventually show less openness over time as more conservative participants will join the community.

4.2. *Nature of contributions over time*

The assessment of the contributions made at each phase revealed some interesting findings. Figure 6(a) displays an apparent stability of the volume of edits over each phase. This is the result of a complex combination of edit rates and proportion of contributors from Nielsen's categories at each phase. The same phenomenon has affected the apparent frequency of contributions presented in Figure 6(b). The actual variations of volume and frequency show no such stability over the different phases (Table 4).

In this context, the profile of contributions from prolific participants is of a particular interest since they provided 90% of OSM data. The average volume of edits (Figure 7(a)) seems to follow the contribution processes described earlier. The volume of edits increases over the first three phases (i.e. over learning-adaptation process), before declining over the last three (i.e. during cumulative-damage process). Such dichotomous behaviour has also been observed in other online communities (Danescu-Niculescu-Mizil *et al*. 2013).

The volume of edits provided over the first phases is characterized by a large participation inequality (outliers). For instance, while most contributors provided a few edits and withdraw over the 'visitor' phase, some participants provided hundreds of thousands of edits on that same day. Similar inequality also applies to the 'explorers' and 'adopters' phase. The profiles of the hundred most prolific OSM contributors indicate that approximately half of them were dedicated import or bot accounts. However, the participants from the remaining half remained active for longer periods of time and often show mixed content (i.e. imports, personal edits and GPS tracks), at least until 2012 when OSM guidelines on import operations were updated.

The frequency of contributions from 'prolific' participants was expected to match trends observed in the volume of edits. However the analysis revealed that the frequency dropped over time (Figure 7(b)). The other Nielsen categories (Table 4) show that 'casual' and 'inactive' mappers rather increase the frequency of contributions over latest phases. Such behaviour is counter-intuitive considering it happens over the 'detachment' stage. A potential explanation is that many of these 'casual'/'inactive' participants, who did not withdraw after so many years, may not have had the opportunity to contribute at will, throughout their lifespan. They may also have changed their objectives over time, bringing new motivations to contribute. The convex/concave shape of survival curves (Figure 5(a)) may be related to this 'incomplete' experience, further distinguishing 'prolific' participants from other contributors.

4.3. *History of contributions to OSM at a glance*

Survival analyses and the analysis of contributions at each phase provided an in-depth understanding of participants' life cycle. Our first analysis (Figure 1) proved to be a simple yet powerful approach to better understand both contributors' lifespan over years, and the history of a project. Trends identified by the literature regarding the behaviour of contributors in online communities are revealed by this simple graph (Figure 1).

The absence of diagonal patterns in the upper left of the graph illustrates findings from our survival analyses. First, it shows that the life cycle of contributors is not affected by sudden changes after initial withdrawals, which would have created diagonal fading of density towards the upper left corner. Second, it demonstrates that the maximum lifespan of contributors has not been reached yet since there is no definite blank triangle on the top left corner.

The slight variations in density over horizontal and vertical axes tell intimate stories about the project and its participants. These patterns reveal how people are brought to the project following media coverage and illustrate how conflicts between participants' personal values and beliefs can be expressed by collective withdrawals. These conflicts can be openly shared with the community as with the ODbL licence change, for example. Opponents took a public stance and drew a significant number of contributors to their arguments. When actions are taken in relation to the conflict (e.g. blocking opponents and implementing a disputed solution), a significant number of supporters may withdraw from the project in response, even long after the opponents have left the project. Such conflicts may not be public, nevertheless resulting in similar collective withdrawals. For instance, when Esri, a prominent player of the GIS industry, proposed an interface to the project, contributor withdrawal patterns suggest that a portion of the community was offended or experienced reduced motivation to participate in the project. Although this may not be surprising in an environment dominated by free and open-source software (FOSS) enthusiasts (Perkins 2011, Elwood *et al*. 2012, Budhathoki and Haythornthwaite 2013), such an assumption cannot be confirmed based on the available data.

Finally, the graph also reveals an unexpected relationship between withdrawals and application improvements. Among other reasons, a change in familiar habits may appear as too difficult to cope with, particularly for contributors who are in the cumulative-damage-dominated phase. These participants may have withdrawn from the project because the perceived cost of contributing exceeded the derived benefits. However, since the graph interpretation is purely qualitative, further analyses are needed to confirm these relationships.

5. Conclusions

This paper has shown that two contribution processes and three participation stages combined into six phases to define OSM contributors' life cycle. We found that the first two phases are critical to engage OSM contributors in the long term, particularly for those who contribute only little to the project. While their survival rates greatly differ from more productive contributors during this period (i.e. 'assessment'), they show very similar patterns of withdrawal after that stage. Actions to engage contributors in the long-term must then focus on this 6-month period. At the same time, using a density scatterplot of contributors' first and last edits could be an easy and effective way to monitor how the different interventions affect the community.

Our results have shown that OSM contributors behave according to Nielsen's 90–9-1 rule. They illustrated that sustained contributions are not only the fact of highly productive contributors, but also come from many other participants, even if these participants contribute only a few times a year. It has revealed that the volume of edits provided on each active day is driven by two contribution processes illustrating the evolution of contributors' motivation over time. Surprising increases of the frequency of contributions among non-prolific contributors over the later phases of their life cycle will require further study.

Analyses confirmed that the phases of the life cycle derived from contributors' lifespan, not the number of active days they experienced as considered at the beginning of the study. Although it could not be verified within the scope of this study, the number of active days may have an influence on contributors' behaviour following project adoption.

Presuming that experienced contributors have a deeper knowledge of OSM features and more skills regarding data capture, we can expect the diversity and quality of OSM data to improve as the proportion of experienced contributors should increase until the maximum lifespan in OSM is reached.

The temporal approaches described in this paper offer novel methods for determining the phases of contributors' life cycle and shed a new light on the nature of VGI contributions that are increasingly valuable to GIScience. An objective metric, available in most online projects (i.e. contributors' lifespan), was used in conjunction with a rigorous approach to identify the phases in contributors' life cycle. The phases identified have revealed very natural behaviours that may apply to a broad range of collaborative online communities. Finally, according to the Diffusion of Innovation Theory, most of these new 'open' communities (e.g. OSM and Wikipedia) may eventually show less openness as more conservative participants will join them.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [Discovery Grant] awarded to Rodolphe Devillers and by Memorial University of Newfoundland;

References

Aknouche, L. and Shoan, G., 2013, *Motivations for Open Source Project Entrance and Continued Participation*. Thesis (Master). Lund University.

Arsanjani, J.J., *et al*. 2015. *OpenStreetMap in GIScience*. Cham (CHE): Springer.

Bégin, D., Devillers, R., and Roche, S., 2017a. Contributors' enrollment in collaborative online communities: the case of OpenStreetMap. *Geo-Spatial Information Science*, 19 (3), 282–295. doi:10.1080/10095020.2017.1370177

Bégin, D., Devillers, R., and Roche, S., 2017b. Contributors' withdrawal from online collaborative communities, the case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 6 (11), 340.1–340.20. doi:10.3390/ijgi6110340

Borst, W.A.M., 2010. *Understanding crowdsourcing - Effects of motivation and rewards on participation and performance in voluntary online activities*. 1st ed. Rotterdam (NLD): Erasmus University of Rotterdam.

Bright, J., De Sabbata, S., and Lee, S., 2017. Geodemographic biases in crowdsourced knowledge websites: do neighbours fill in the blanks?. *GeoJournal* [Online], 1–14. Available from: https://doi.org/10.1007/s10708-017-9778-7 [Accessed 21 Mar 2018].

Brown, J.J. and Reingen, P.H., 1987. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14 (3), 350–362. doi:10.1086/jcr.1987.14.issue-3

Bryant, S.L., Forte, A., and Bruckman, A., 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, November 6-9 Sanibel Island (USA). New York (USA): ACM, 1–10.

Budhathoki, N.R., 2010, *Participants' motivations to contribute geographic information in an online community*. Thesis (PhD). Graduate College of the University of Illinois.

Budhathoki, N.R. and Haythornthwaite, C., 2013. Motivation for open collaboration crowd and community models and the case of OpenStreetMap. *American Behavioral Scientist*, 57 (5), 548–575. doi:10.1177/0002764212469364

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

Capineri, C., *et al*., 2016. *European handbook of crowdsourced geographic information*. 1st ed. London (GBR): Ubiquity Press.

Chacon, F., Vecina, M.L., and Davila, M.C., 2007. The three-stage model of volunteers' duration of service. *Social Behavior and Personality*, 35 (5), 627–642. doi:10.2224/sbp.2007.35.5.627

Cheung, K.S., *et al*., 2005. The development of successful on-line communities. *International Journal of the Computer, the Internet and Management*, 13 (1), 71–89.

Clary, E.G., *et al*., 1998. Understanding and assessing the motivations of volunteers: a functional approach. *Journal of Personality and Social Psychology*, 74 (6), 1516–1530. doi:10.1037/0022-3514.74.6.1516

Coleman, D.J., Georgiadou, Y., and Labonté, J., 2009. Volunteered geographic information: the nature and motivation of producers. *International Journal of Spatial Data Infrastructures Research*, 4, 332–358.

Danescu-Niculescu-Mizil, C., *et al*., 2013. No country for old members: user lifecycle and linguistic change in online communities. *Proceedings of the 22nd international conference on World Wide Web*, ACM, 307–318.

Elwood, S., Goodchild, M.F., and Sui, D.Z., 2012. Researching volunteered geographic information: spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, 102 (3), 571–590. doi:10.1080/00045608.2011.595657

Hemetsberger, A. and Pieters, R., 2003. *When consumers produce on the internet: the relationship between cognitive-affective, socially-based, and behavioral involvement of prosumers* [online]. CiteSeerX. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9299&rep=rep1&type=pdf.

Houle, B.B.J., Sagarin, B.J., and Kaplan, M.F., 2005. A functional approach to volunteerism: do volunteer motives predict task preference? *Basic and Applied Social Psychology*, 27 (4), 337–344. doi:10.1207/s15324834basp2704_6

Kimura, A.H. and Kinchy, A., 2016. Citizen science: probing the virtues and contexts of participatory research. *Engaging Science, Technology, and Society*, 2, 331–361. doi:10.17351/ests2016.99

Lave, J. and Wenger, E., 1991. *Situated learning: legitimate peripheral participation*. 1[st]ed. Cambridge (GBR): Cambridge university press.

Ma, D., Sandberg, M., and Jiang, B., 2015. Characterizing the heterogeneity of the OpenStreetMap data and community. *ISPRS International Journal of Geo-Information*, 4, 535–550. doi:10.3390/ijgi4020535

Michelucci, P. and Dickinson, J.L., 2016. The power of crowds. *Science*, 351 (6268), 32–33. doi:10.1126/science.aad6499

Mooney, P. and Corcoran, P., 2012. Who are the contributors to OpenStreetMap and what do they do? *Proceedings of the GIS Research UK 20th Annual Conference*, April 11-13 Lancaster (GBR): Lancaster University, 355–360.

Moore, G.A., 2001. *Crossing the chasm - marketing and selling high-tech products to mainstream customers*. Reviseded. New York (USA): HarperCollins.

Napolitano, M. and Mooney, P., 2012. MVP OSM: a tool to identify areas of high quality contributor activity in OpenStreetMap. *The Bulletin of the Society of Cartographers*, 45 (1), 10–18.

Neis, P. and Zielstra, D., 2014. Recent developments and future trends in volunteered geographic information research: the case of OpenStreetMap. *Future Internet*, 6 (1), 76–106. doi:10.3390/fi6010076

Neis, P. and Zipf, A., 2012. Analyzing the contributor activity of a volunteered geographic information project—the case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1 (2), 146–165. doi:10.3390/ijgi1020146

Nielsen, J., 2006. *The 90-9-1 Rule for Participation Inequality in Social* Media *and* Online Communities [online]. Nielsen Norman group. Available from: http://www.useit.com/alertbox/participation_inequality.html [Accessed 26 Oct 2012].

Nov, O., 2007. What motivates Wikipedians? *Communications of the ACM*, 50 (11), 60–64. doi:10.1145/1297797

Nov, O., Arazy, O., and Anderson, D., 2011. Technology-mediated citizen science participation: a motivational model. *Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media*, July 17-21 Barcelona (ESP). Menlo Park (USA): The AAAI Press, 249–256.

Nov, O., *et al*., 2014. Scientists@ Home: what drives the quantity and quality of online citizen science participation? *PloS One*, 9 (4), e90375. doi:10.1371/journal.pone.0090375

Ochoa, X. and Duval, E., 2008. Quantitative analysis of user-generated content on the web. *In*: D. De Roure and W. Hall, eds. *Proceedings of the first international workshop on understanding web evolution (WebEvolve2008): a prerequisite for web science*. April 22. Beijing (CHN). New York (USA): ACM, 19–26.

OpenStreetMap contributors, 2014a. *Complete OSM Data History* [online]. OpenStreetMap Wiki. Available from: http://planet.openstreetmap.org/planet/full-history/ [Accessed -03 Jul 2014].

OpenStreetMap contributors, 2014b. *Main Page* [online]. OpenStreetMap Wiki. Available from: http://wiki.openstreetmap.org/wiki/Main_Page [Accessed 19 Jun 2017].

Penner, L.A., 2002. Dispositional and organizational influences on sustained volunteerism: an interactionist perspective. *Journal of Social Issues*, 58 (3), 447–467. doi:10.1111/josi.2002.58.issue-3

Perkins, C., 2011. Researching mapping: methods, modes and moments in the (im) mutability of OpenStreetMap. *Global Media Journal-Australian Edition*, 5 (2), 1–12.

Preece, J. and Shneiderman, B., 2009. The reader-to-leader framework: motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, 1 (1), 13–32. doi:10.17705/1thci

R Core Team, 2016. *R: a language and environment for statistical computing*. Vienna (AUT): R Core Team.

Rogers, E.M., 1983. *Diffusion of innovations*. 3rd. New-York (USA): The Free Press.

Rotman, D., *et al*., 2014. Motivations affecting initial and long-term participation in citizen science projects in three countries. *iConference 2014 Proceedings*, 4–7 March Berlin (GER). ACM, 110–124.

Schneider, A., Von Krogh, G., and JäGer, P., 2013. "What's coming next?" Epistemic curiosity and lurking behavior in online communities. *Computers in Human Behavior*, 29 (1), 293–303. doi:10.1016/j.chb.2012.09.008

Sui, D.Z., Elwood, S., and Goodchild, M.F., 2013. *Crowdsourcing geographic knowledge*. 1sted. New York (USA): Springer.

Sun, N., Rau, P.P., and Ma, L., 2014. Understanding lurkers in online communities: a literature review. *Computers in Human Behavior*, 38, 110–117. doi:10.1016/j.chb.2014.05.022

Therneau, T.M. and Lumley, T., 2017. *R survival package - survival analysis*. Fermanagh (IRL): CRAN.

Wang, K.S., Hsu, F., and Liu, P., 2002. Modeling the bathtub shape hazard rate function in terms of reliability. *Reliability Engineering & System Safety*, 75 (3), 397–406. doi:10.1016/S0951-8320(01)00124-7

Wenger, E., 1998. *Communities of practice: learning, meaning, and identity*. 1sted. Cambridge (GBR): Cambridge university press.

Weon, B.M., 2016. Tyrannosaurs as long-lived species. *Scientific Reports*, 6 (srep19554), 1–5. doi:10.1038/srep19554