CHAPTER 3

Why is participation inequality important?

Mordechai (Muki) Haklay
Department of Civil, Environmental and Geomatic Engineering,
University College London, m.haklay@ucl.ac.uk

Abstract

Participation inequality – the phenomenon that a very small percentage of participants contribute a very significant proportion of information to the total output – is persistent across Volunteered Geographic Information (VGI) and citizen science projects. It has been identified in both online and offline projects that rely on volunteers' effort over the past 20 years and, therefore, can be expected to appear in new projects. This chapter looks at participation inequality (also known as the 1% rule or the 90-9-1 rule), its origins and some of its characteristics. The chapter also explains how participation inequality emerges in a project at both temporal and spatial scales, and also evaluates its implication on the use of VGI and citizen science data. The chapter suggests a generic rule for analysts of VGI and citizen science datasets, in the form: '*When using and analysing crowdsourced information, consider the implications of participation inequality on the data and take them into account in the analysis.*'

Keywords

Participation inequality, patterns of contribution, citizen science, online and offline communities, 1% rule, 90-9-1 rule

How to cite this book chapter:
Haklay, M. 2016. Why is participation inequality important?. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 35–44. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.c. License: CC-BY 4.0.

Introduction

One of the most persistent aspects that can be noted in systems which facilitate user-generated content (among them volunteered geographic information and citizen science data) is the inequality in the level of participation that they exhibit. According to Jakob Nielsen (2006), participation inequality was first recognised by Hill and his team (1992) while studying the development of digital documents and analysing the contributions by different people to the final product. It manifests itself in online forums such as mailing lists, discussion forums, games and ecological observations (e.g. Hill et al. 1992; Mooney & Corcoran 2012; Lund et al. 2011; van Mierlo 2014; Silvertown et al. 2015). In each of these cases, the overwhelming majority of people who use the information or are registered to the service do not contribute any information to it. The proportion of registered people who do not contribute can reach 90% or even more of the total number of users. Of the remaining participants, the vast majority contribute infrequently or fairly little – these account for 9% or more of the users. Finally, the last 1% contribute most of the information. This has led to framing the phenomenon as the 90-9-1 rule (Nielsen 2006). However, participation can be very skewed. As Nielsen demonstrates, in Wikipedia, 0.003% of users contribute two-thirds of the content, with a further 0.2% contributing infrequently, making the relationship 99.8-0.2-0.003% (with the increased use of Wikipedia since 2006, the situation has worsened). There is some evidence to suggest that the proportion can be different – for example, Budhathoki (2010) suggests that in OpenStreetMap the proportions are 70-29.9-0.01%. Recent analysis by Harry Wood (2014) provides an indication of this relationships (Figure 1), with the contribution of the first ranked 1,000 participants dwarfing the effort of all other contributors, and only about 300,000 participants contributing more than 10 points of data - although at the time there were 2 million registered users.

Participation inequality has been observed in VGI and citizen science projects such as OpenStreetMap (Budhathoki 2010; Mooney & Corcoran 2012; Neis & Zipf 2012), Galaxy Zoo (Ponciano & Brasileiro 2014) and bird watching (Cooper & Smith 2010). It is especially noteworthy that participation inequality is not only appearing in online projects, but also can be observed in projects that mainly happen offline, such as participation in environmental volunteering or when analysing the levels of contribution of different volunteers in biological observations across London.

In this chapter, we look at the implications of participation inequality and argue that it is among the most significant aspects of VGI and citizen science. We start by noticing what we already know about participation inequality and its manifestations. This is followed by suggesting possible explanations for how it occurs and evolves over time. The fourth section discusses the potential implications on project development and the use of information that emerges from it. We conclude with open research questions and future directions for investigation that are of specific interest to researchers of VGI.

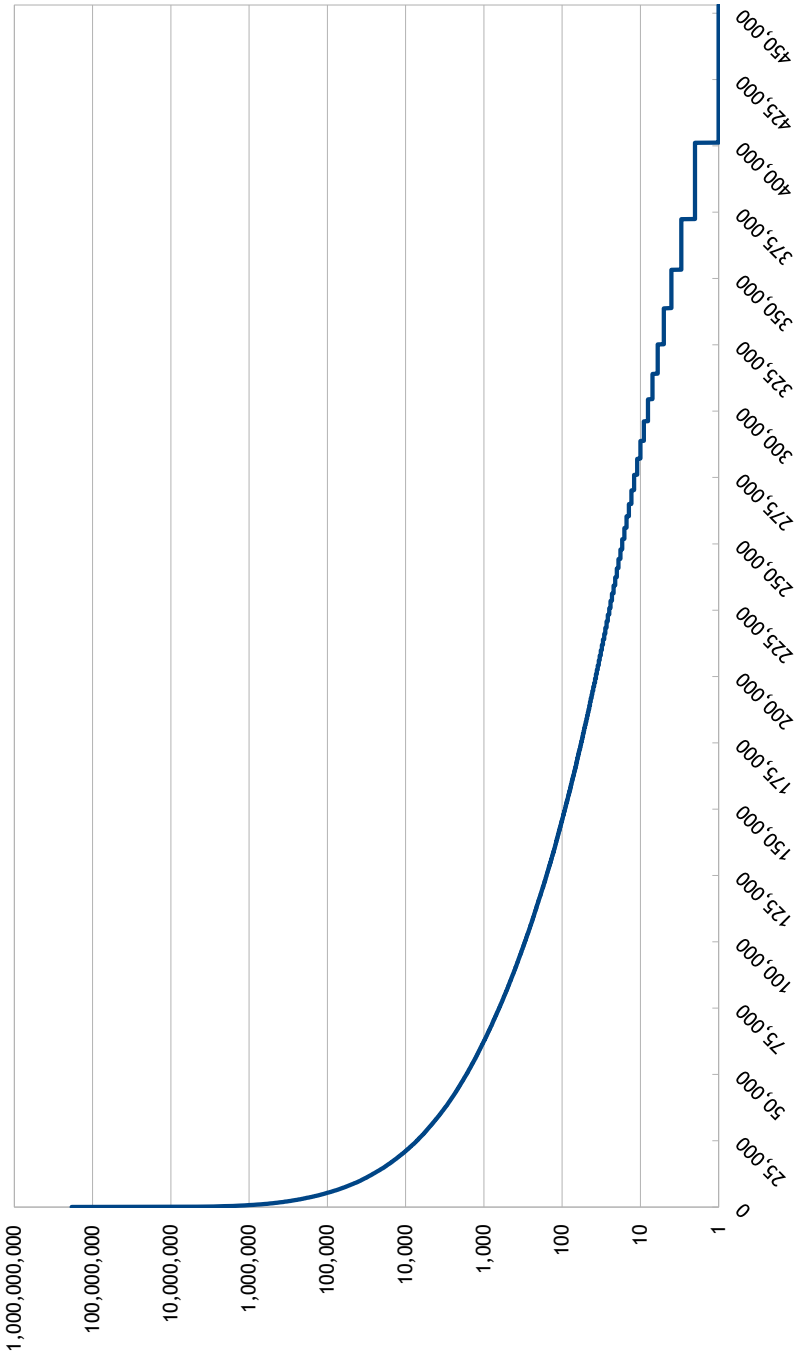

Figure 1: OpenStreetMap contributions (Wood 2014).

Throughout the chapter, OpenStreetMap is being used to demonstrate the nature and implications of participation inequality. While OpenStreetMap have specific characteristics in terms of participants' profiles and social dynamics (Budhathoki 2010; Haklay 2010), it can be used to illustrate the general aspects of the phenomena. Other projects are being used to augment the picture.

Participation inequality – what do we know?

Unlike command and control processes that are common in industrial information creation, VGI and citizen science are produced through a distributed, less coordinated system. Within industrial processes, there is scope for planning of coverage and allocation of resources. For example, when planning the surveying of a city in an industrial process, it is possible to divide the efforts of the surveyors to ensure uniform level of coverage and time allocation to different parts in proportion to the amount of work that is required. Of course, the abilities of the different surveyors will have an impact on the final results but, in general, these can be minimised through quality assurance so the final product is uniform.

Within a system that relies on 'crowdsourcing' – the use of a large group of people with whom there are no direct employment relationships – there is far less ability to dictate to the participants where, when and how they should contribute information. For example, in a system that provides traffic information on the basis of users' satellite navigation devices, there is a co-dependence between the number of users in a given location and the ability to provide information about this place. Moreover, because the devices are used within the context of daily activities, such as the school run or a trip to the local supermarket, there will be more information about places in which many people travel daily (e.g. city centre) and especially during rush hour. While both industrial and crowdsourced systems are socio-technical systems, in the latter the 'socio' requires special attention, particularly to the way it influences the resulting information that emerges from the system.

In the case of participation inequality, since it has been so persistent over the years, it is highly likely to appear in any crowdsourcing project. It has been observed from the pre-Web internet messaging system Usenet (Whittaker et al. 1998) to current large-scale online citizen science (Ponciano & Brasileiro 2014). It is, therefore, part and parcel of VGI and citizen science.

Just as interesting is that the phenomenon repeats itself at various scales (something akin to Power Laws), so analysing the level of participation in OpenStreetMap for the area of London, Europe or across the world will show participation inequality (Haklay 2010; Mooney & Corcoran 2012; Neis & Zipf 2012). Participation inequality also occurs at different temporal scales of weeks, months or years (Neis & Zipf 2012). As can be expected with statistical analysis of this sort, the larger the area or the longer the time frame, the clearer the pattern and the position of various participants.

Another important aspect known about participation inequality is that lowering the barrier for participation does help, but to a limited extent. Even in volunteer computing projects, in which participants download software to their computers that utilises unused processing resources for scientific research, participation inequality persists. IBM World Community Grid serves as an example. This project is an aggregator of volunteer computing projects, and yet few members contributed most of the processing. Of the 350,000 participants, the top contributor has contributed 325 times more than the 250th contributor, and 875 times more than the 1,000th contributor.

The use of a leader board and providing credits to emphasise the position of participants has been shown to encourage competition among contributors, but with a potential to alienate some participants and reduce their motivation (Massung et al. 2013). The assumption that it is always valuable to encourage competition among participants to yield more information should be questioned, and there are alternative, such as the mechanism that encourage collaboration that Silvertwon et al. (2015) offer.

Participation inequality also manifests itself through geographic and temporal patterns. Thus, places that are within the coverage area of highly active participants will have more contributions than areas that do not have many participants. More generally, the geographic distribution of information shows that some places are more popular and receive much more attention than others. Similarly, the temporal pattern of highly active contributors has a disproportionate impact on the temporal patterns of data collection activities as a whole. Thus, the sleeping and working patterns that can be observed within the contributed information will be influenced by the practices of high contributors (Yasseri et al. 2013).

Finally, while high contributors receive a lot of attention, in comparison to the very large group of people who contribute very little both individually and to the overall size of the dataset, we should not forget that they are, statistically, outliers. They are not representative of the overall population, nor should we expect them to be so. There is a need to have the majority of people as consumers of information, as otherwise the producers would lose the *raison d'ĕtre* to create and share information.

How participation inequality evolves over time and space

One of the puzzling questions regarding participation inequality is how it evolves. After all, at first look the participants are acting as volunteers and therefore there is no limitation on the number of people who can join a specific activity in citizen science or VGI or how much each of them contributes. Second, arguably, the actions of one participant do not stop another, for example when viewing the same bird or taking a geotagged picture of Big Ben (see Jayaraman 2012). Furthermore, the participants are only loosely coordinated

and therefore not necessarily aware of the actions of other participants, and there is no reason for one to compete with another or even be aware of their contribution. However, some of these observations are inaccurate, and a further analysis of the process that created participation inequality can explain the source of the observed patterns.

Firstly, we can start by noticing that, in many VGI and citizen science projects, some resource is finite. For example, in OpenStreetMap or Wikimapia, once a participant has tagged a location and mapped it, this specific place is no longer available to other users to carry out the mapping. This is also true in volunteer thinking projects in which participants help scientists in classifying information online. In such projects, the system allocates the images to participants and, after the image has been viewed by a given number of participants, it is not shown anymore. Therefore, if one participant becomes highly active, they reduce the amount of work that is left to other participants to carry out.

Secondly, the temporal aspects of the project also play their part in generating participation inequality. For example, participants who joined OpenStreetMap early on were facing an empty map, in which it was relatively easy to identify and digitise objects such as motorways. Over time, the ability to digitise objects rapidly diminished as the map became complete. For a volunteer who joins the mapping process today, in many places the effort that is left requires adding more intricate details of building or address information. This is also true in citizen science, for example in the British Trust of Ornithology (BTO) breeding bird survey which started over twenty years ago. A volunteer that joined the project in the early stages will have collected many more records over the years than a person that will join the project today, who will not be able to 'catch up' to such levels of recording.

Thirdly, another side to the temporal aspect is demonstrating the link between participation inequality and other social inequalities. The contributions of participants can be translated into time – for example, one of top contributor to OpenStreetMap in June 2015 (Władysław Komorek) edited over 4.94 million objects in 966 active mapping days over 3 years, contributing on average about 5,100 points in an active day. With an assumption that it is possible to record 2 objects per second, this represents an average investment of about hour in digitising only (without any breaks). This is, of course, a low estimation, since such a participant spends time on mailing lists, meetings and going out mapping. When considering that, across advanced economies, people have about 36.5 hours of leisure a week (OECD 2009), it is clear that, for this participant, OpenStreetMap is the most important leisure activity during that period. However, since leisure time is more available to men, and is reduced in people with major caring responsibilities, it is more likely that men with a well-paid job will be able to become major contributors of VGI and in many citizen science activities. Indeed, many projects have people with such profiles as their top contributors (e.g. Cooper & Smith 2010). However, one should be careful of sweeping generalisations about the profile of top contributors, as they are

specific to projects and research area – for example in the EyeWire project, in which participants help brain research by analysing the structure of neurons, 65% of top contributors are women (Kim et al. 2014), while bird watching is dominated by men (Cooper & Smith 2010).

Fourthly, access to financial resources can have an impact on the ability of people to become high contributors. For example, an Australian study of bird-watchers concluded that some of them travel 300 to 1,900 km from home to record an observation (Tulloch & Szabo 2012). Such extensive travel, apart from dedication, also requires financial resources. Other VGI and citizen science activities also involve purchasing specialised equipment and dedicating time to learn how to use it.

Finally, there is a need to consider internal and external motivations of high contributors. The various studies that were mentioned above, and others, demonstrate clearly that the top contributors represent a different demographic group – for example, in EyeWire they are older than the average participant (Kim et al. 2014). Studies show that their internal and external motivations play an important part in maintaining their engagement with a project. For some participants, competition is a significant motivation (Massung et al. 2013) while for others the joint contribution to science is a major one (Nov et al. 2011).

The implications of participation inequality

Based on the analysis above, we can formulate a general rule for crowdsourced geographic information: *'When using and analysing crowdsourced information, consider the implications of participation inequality on the data and take them into account in the analysis.'*

As we have seen, crowdsourced information, either VGI or citizen science, is created through a socio-technical process, which, by necessity, will have impacts on the final outputs. Yet, all too often it is easy to forget the social side – especially when using the information without paying due attention to the metadata of who collected it and when. Even though analysts who use the information are aware that the data source is expected to be heterogeneous because of the nature of the crowdsourced process, it is easy to forget participation inequality and treat each observation as similar to other observations and assume they were all produced in a similar way.

Yet, data is not only heterogeneous in terms of consistency and coverage; it is also highly heterogeneous in terms of contribution, which can have far-reaching implications on quality, coverage and content. As we have explored, the outcome is dependent on the expertise of heavy contributors, their spatial and temporal engagement, and even on their social interactions and conduct.

For example, some of the top contributors of OpenStreetMap naturally concentrate their effort in the city where they live. Knowing where these individuals are active can help in quality assurance processes by comparing novice

practices to their actions, potentially changing the number of people that are required to map an area well (Haklay et al. 2010). In some projects, such as iSpot (Silvertown et al. 2015), in which participants help in the identification of a range of species, there are mechanisms to reward high contributors with trust marks and to give their opinions more weight during the identification process.

Another aspect of the impact of high contributors is the social evolution of the project. In some projects, high contributors might exhibit abrasive behaviour towards other participants or protect 'their patch' (the area in which they operate) by aggressively editing any new information to fit their standards. Such conduct is not welcoming to new participants, and can impact on the growth of the project and even its resilience in cases where the high contributor leaves the project.

The specific background and interests of high contributors will, by necessity, impact on the type of data that is recorded. This is especially important in VGI projects where the details of what to record are left to the participants. For example, lack of interest in a class of facilities (e.g. wheelchair accessible toilets) will mean that such information will be lacking from the resulting dataset and might shape the activities of other participants (Stephens 2013).

Interestingly, while some research analysed the biases that are created by high contributors (Haklay 2010; Bégin et al. 2013; Mooney 2013), there is relative lack of attention within the VGI literature to the wider impact that they have on the information and on other participants.

Conclusion

In this chapter, we looked at participation inequality and its implications on VGI and citizen science datasets. We have seen that participation inequality – the phenomenon in which a very small percentage of participants contributes a very significant proportion of information to the total outcome – is persistent. It occurs across spatial and temporal scales and is driven by multiple factors.

Participation inequality impacts on the social and technical outcomes of a project and, because of that, it is critical to remember the impact and implications of participation inequality during the analysis and use of the information. There will be some analysis to which it will have less impact and some where it will have major impact. In either case, it needs to be taken into account. This can be done by including an analysis of participation patterns early on in the analysis of a dataset, and examining the biases that are caused by it.

While we can expect it, we do need to understand more about the process that created it and its impact on the resulting datasets. There is plenty of scope for spatio-temporal analysis to identify the actions of high contributors from their early actions, and evaluate to what degree they impact on other contributors. There is also value in more detailed analysis of how people at different levels of contribution add to the project and whether there are ways to encourage

people to move between contribution groups. Finally, the ethical and practical implications of high contributors should be assessed, especially in commercial VGI projects.

Acknowledgements

The author would like to thank Prof Tanya Berger-Wolf for her suggestions on explaining participation inequality and the work of Vyron Antoniou, Valentine Seymour and Gianfranco Gliozzo at UCL on various datasets. Many thanks for Cristina Capineri and Stephen Winter for their comments on an earlier version of this paper. The research was supported by EPSRC (EP/I025278/1, EP/K022377/1) and the EU FP7-ICT (317705).

References

Bégin, D., Devillers, R., & Roche, S. 2013 (May). Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours. In *Proceedings of the 8th international symposium on spatial data quality ISSDQ*: pp. 149–154.

Budhathoki, N. R. 2010. *Participants' motivations to contribute geographic information in an online community* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Cooper, C. B., & Smith, J. A. 2010. Gender patterns in bird-related recreation in the USA and UK. *Ecology and Society, 15*(4): 4. Available at: http://www.ecologyandsociety.org/vol15/iss4/art4/.

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

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *Cartographic Journal* 47 (4) 315–322.

Hill, W. C., Hollan, J. D., Wroblewski, D., & McCandless, T. 1992. Edit wear and read wear. In: *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*. ACM: pp. 3–9.

Jayaraman, K. 2012. Tragedy of the Commons in the Production of Digital Artifacts. *International Journal of Innovation, Management and Technology, 3*(5): 625–627

Kim, J. S., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., Purcaro, M., Balkam, M., Robinson, A., Behabadi, B. F., Campos, M., Denk, W., Seung, H. S., & the EyeWirers. 2014. Space-time wiring specificity supports direction selectivity in the retina. *Nature, 509*(7500): 331–336.

Lund, K., Coulton, P., & Wilson, A. 2011 (November). Participation inequality in mobile location games. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. ACM: p. 27.

Massung, E., Coyle, D., Cater, K. F., Jay, M., & Preist, C. 2013. Using crowdsourcing to support pro-environmental community activism. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM: pp. 371–380.

van Mierlo, T. 2014. The 1% rule in four digital health social networks: an observational study. *Journal of medical Internet research, 16*(2).

Mooney, P. 2013. *Understanding the activity of contributors to Volunteered Geographic Information projects. How, why, where, and when do they contribute geographic information?* 2nd Meeting of the EU COST Action TD1202, Dresden, Germany.

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

Nielsen, J. 2006. Participation inequality: Encouraging more users to contribute. *Jakob Nielsen's alertbox*, *9*: 2006.

Neis, P., & 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.

Nov, O., Arazy, O., & Anderson, D. 2011 (February). Dusting for science: motivation and participation of digital citizen science volunteers. In: *Proceedings of the 2011 iConference*. ACM: pp. 68–74.

OECD. 2009. *Society at a glance 2009: OECD social indicators*, OECD.

Ponciano, L., & Brasileiro, F. 2014. Finding volunteers' engagement profiles in human computation for citizen science projects. *Human Computation, 1*: 5–28.

Silvertown, J., Harvey, M., Greenwood, R., Dodd, M., Rosewell, J., Rebelo, T., Ansine, J., & McConway, K. 2015. Crowdsourcing the identification of organisms: A case-study of iSpot. *ZooKeys*, (480): 125.

Stephens, M. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, *78*(6): 981–996.

Tulloch, A. I., & Szabo, J. K. 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets. *Emu*, *112*(4): 313–325.

Whittaker, S., Terveen, L., Hill, W., & Cherny, L. 1998. The dynamics of mass interaction. In: *CSCW '98 Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*: pp. 257–264.

Wood, H. 2014. *The long tail of OpenStreetMap*. Available at: http://harrywood.co.uk/blog/2014/11/17/the-long-tail-of-openstreetmap/#slide18 (accessed July 2015).

Yasseri, T., Quattrone, G., & Mashhadi, A. 2013. Temporal analysis of activity patterns of editors in collaborative mapping project of OpenStreetMap. In: *Proceedings of the 9th International Symposium on Open Collaboration*. ACM: p. 13.