Karin Hansson, Love Ekenberg (2016):
A framework for describing the social production of data in crowdsourcing.
In Karin Hansson, Tanja Aitamurto, Thomas Ludwig, Neha Gupta, Michael Muller (Eds.),
International Reports on Socio-Informatics (IRSI),
Proceedings of the CHI 2016 - Workshop: Crowd Dynamics: Exploring Conflicts and Contradictions in Crowdsourcing
(Vol. 13, Iss. 2, pp. 27-34)

A framework for describing the social production of data in crowdsourcing

Karin Hansson
Stockholm University, Stockholm, Sweden
khansson@dsv.su.se

Love Ekenberg
Stockholm University, Stockholm, Sweden
lovek@dsv.su.se

Abstract. This overview of the handling of user profile data and user activity in some crowdsourcing tools provides a framework for analyzing data production processes in terms of embodiment (the participants' social and cultural perspective) and gameplay (how the participants can interact through the tool). This can create a better understanding of the quality of the data in crowd produced environments, which can be particularly interesting in contexts were trustworthiness is aggregated in the network rather than provided by a single source (of unknown credibility), and as an alternative when normal sources cannot provide trustworthy information or information at all. By combining gameplay metrics with data indicating embodiment, the social production of data can become more transparent.

1 Introduction

Information production and consumption online often reinforce differences prevailing in other contexts. Participants represent specific groups, experiences and opinions of different modalities, social contexts and structuring norms, why crowd sourced information also reflects particular perspectives (Hansson, 2015). Furthermore, the mechanisms of excluding other groups and voices are definitely prevalent in the decision processes on so called open platforms, which reproduce and reinforce inequalities between different groups and identities. For instance, a geo-mapping tool such as OpenStreetMap is subject to a large demographic bias with significantly more men than women are contributing (Neis and Zielstra, 2014; Stephens, 2013). In the negotiation games that appear in such applications, there are always actors that are more successful than others in making their perspectives as the dominant ones. Unlike the gamificated environment of, for example, Waze, social games on Wikipedia's many discussion forums are trickier to master and are even not broadly accessible (Steinmann, Häusler, Klettner, Schmidt, & Lin, 2013).

There is nevertheless a promising potential on openly defined platforms such as Wikipedia and OpenStreetMap for developing elaborated forms of networked democracy that are not constrained to a predefined authority, and where trustworthiness is aggregated rather than provided by a single source. If we would be able to clarify the social production of data without restricting it and controlling it, the data produced within these types of peer-to-peer created contexts would probably provide a better basis for interpretation and management; also from a governing perspective.

In this position paper we investigate crowdsourcing tools of different types to better understand the social production of data, where after we suggest a framework for a systematic analysis of this phenomenon.

2 The social production of data

To understand the social production of data, we need to know the participants identities, in terms of age, gender, location, occupation, and education, as well as how they interact on the platforms. The latter can instrumentally be expressed in metrics, such as user activity, interaction, relations, and reputation. We also need to understand which types of data the interface asks for and which data they collect as well as whether this data is publicly available. In this study we investigated the interfaces in a representative subset of geo-mapping tools of the

abovementioned kinds: Wikipedia, Twitter, OpenStreetMap, Waze, Wikimapia, and Google MapMaker.

First we looked at if and how the interfaces are gathering information regarding the participants, thereafter we investigated how (if applicable) the interfaces gathered and displayed information about the participants' interaction and activity.

2.1 Embodiment: user data indicating social situation and literacy

The focus in these tools for crowd sourced information gathering is not so much on user identity as in other more social media contexts. Collecting user identity indicators such as *age* and *gender* are not very common. In general, it is difficult to find available data that can help the users to position contributors as belonging to a certain perspective or group, and none of the platforms make user statistics available in an easy accessible way. Even when there is a possibility to add a photo, image-based data is difficult to systematize and aggregate statistically. A profile page for a Wikipedia user is, for example, a blank page without predefined categories where it is up to the user to define themselves, complicating the data gathering and comparison.

Self-reported pre-categorized data might nevertheless be erroneous in any case and there are other ways to correct the information provided. For instance, in Twitter, gender, age, language and location is possible to understand through the API through a combination of looking at participants' language, IP-addresses, participants' relations to other relations, and usage (Malhotra, 2014; Underwood, 2012). There is a vast amount of research investigating the best ways for data verification, cf. e.g., (Culotta, Ravi, & Cutler, 2015), but the adequacy is debated (Adnan et al., 2014; Burger, Henderson, Kim, & Zarrella, 2011; Hargittai, 2015; Harris, n.d.; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011; Sloan et al., 2013). An interesting observation here is, for instance, that the tools generally do not describe any indications of significant qualities, such as the participants' possible *literacies*. The only tool that gathers information that might indicate users social and cultural capital, such as *education* and *occupation,* is Google MapMaker, with the help of data from Google's general profile management system. However, this information is not made available for all the users of the tool. *Geo-location* is, on the other hand, something that is gathered and made transparent in several of these tools, which is not surprising as most of these are about collaborative mapping. This is another way of determine where the users are living and working, which then can be compared to available demographic information on the area. Waze, a collaborative street-mapping tool, is even using location as part of the interface functionality. Here is the contributors' permission level restricted to areas where the users have been located, which means that the

user needs to prove their local expertise by showing that they actually have been in the area.

The user can provide a text description in some of the tools, but there are no categorizations of *interest, age* or *education*. User identity is foremost defined by user followers or friends and by those the user *follows* or are *friend* with, which is public in Twitter and OpenStreetMap, where *user relations* and *user activities* define the user. Thus, despite the lack of self-reported demographic data, investigating the *user relations* and *interests* could be used to determine group identity and educational level of the users.[3]

There are surprisingly little direct information collected in these tools when it comes to identifying participants and their attributes revealing something regarding their social and cultural capital - what we can call their *embodiment*, such as age, gender, work and education. However through location data, relations, and users' showed interest, group identity and the educational level of the users can nevertheless often be determined.

2.2 Gameplay: conditions for interaction, visualization of user activity, interaction, relations, and reputation.

Gathering, and sometimes also displaying, information on the users' activities and discussions was common in the systems we investigated. Sometimes open data from OpenStreetMap and Wikipedia is used to analyze and visualize user activities, but previous activities are usually not used by the various systems for e.g. providing special privileges. Instead the reputation mechanisms seem to be purely social in the sense that they are used as a "badge", rather than for being used in the reputation mechanisms. Often, the activities are differentiated into different categories. For example, OpenStreetMap, collects user activities divided into "Edits", "Map Notes", "Traces", "Send Message", "Diary", "Comments" and "Friends", which makes the information a bit difficult to overview and interpret. An exception is again Waze, where the user activity is simplified and translated to a score, making the system more game like. Users are earning points by leveling up from one role to another.

Some tools utilize forms of user rankings for example by giving "likes", or showing various metrics, such as visits to the users profile, mentions of user names, number of new followers/friends, and similar. The standard Twitter interface displays followers and the amount of tweets, and as an additional

[3] Regarding these types of methods, Mislove et al (2011), for example, shows a bias in the Twitter population relative to the U.S. population, indicating that entire regions of the U.S. are underrepresented on Twitter. The result indicates that social and cultural capital are important for Twitter activity, why a higher education and an urban type of social network can, as in other contexts, have significant advantages.

service, the users get access to more extended analytics. This information is also important when assigning credibility ranks. A trustworthy Twitter user might not be the one with the largest group of followers; quite on the contrary. A study of influence on Twitter showed that that credibility is more about showing sign of being in mutual relationships than having many followers (Bakshy, Hofman, Mason, & Watts, 2011).

Reputation generation can also be a function of more than just the number of followers or re-tweets. For example, user-sharing activities can be used as a way to position the users in social contexts. Similarly, the social neighborhood of users and topic popularity can be further properties used as measurements of reputation or influence (Han, Nakawatase, & Oyama, 2014). Castillo et al. (2013) suggest, based on twitter studies during crisis situations, a measure of credibility based on the characteristics of the users propagating the information in combination with the reactions that certain topics generate, and types of external sources utilized. These types of reputation mechanisms can be particularly useful in crisis situations when there is a limited access to other trustworthy information sources or when such are entirely absent (Castillo, Mendoza, & Poblete, 2011).

Despite that most of these systems gather, and sometimes also display, information on user activity and discussions, the functionality is seldom easy to overview and understand. Likewise, the process of gaining higher status is usually complex and unclear. Again, Waze is an exception utilizing gamification as a way to motivate contributions. On the other hand, Waze is not one of the more open systems and the game-rules are restricted and inflexible.

Relational metrics like profile visits, number of followers or citations are maybe most important when judging credibility. User sharing activity and characteristics as well as the reactions to actions, are other properties that can be used as measurements of reputation. These properties show how the tool is used and how the users are following the rules and opportunities within, what can be called, the *gameplay*.[4]

3 Conclusion

Obviously there are plenty of methods available that could make crowd-produced information easier to understand just using available data and without creating detailed control mechanisms. However, and not surprisingly, the accuracy depends on the combinations of different retrieval mechanisms for data and research methodology as well as the actual context.

[4] We take this concept from game studies, describing the specific ways in which players interact with a game. This is useful as a way to highlight the social engineering that takes place in these contexts, and how the design directs how participants interact with the tool.

By combining data indicating *embodiment* with data indicating *gameplay*, we provide a framework for describing data production that acknowledges the inequalities in these processes and uses this information as a knowledge base.

Embodiment: Describes how participants' bodies are structured in relation to social groups and material conditions*:* Does the tool collect and visualize belonging to categories - data about the user such as gender, race, location and age? What does it take in terms of social and cultural capital to participate? Are there means to measure social network or describing educational levels?

Gameplay: *Describes* how participants' interactions are structured in the tool: How are scores and rankings defined, accumulated, and exchanged? Is reputation counted for in the system or used as part of the interface? Is activity and "likes" measured and used in some sort of reputation mechanism? Are the algorithms for aggregating and analysing the data transparent and adequate?

Many platforms do not emphasize embodiment of the data production in terms of user profiles or by making user statistics easier available. But there are alternative ways of retrieving data on gender, age, language and location, by investigating user activity and in particular by combining it with, geo-location and demographic data as well as user relations and interests. This can be a way of making the gameplay more explicit and to create more transparent and equal information production and decision processes, especially when connecting it with mechanisms for visualizing reputation.

We will continue the investigation including a wider variety of tools. We hope this will lead to a better understanding of the available tools and contribute to the development of new tools designed for more transparent and equal information production and decision processes.

4 References

[1] Adnan, M., Lima, A., Rossi, L., Veluru, S., Longley, P., Musolesi, M., & Rajarajan, M. (2014). The uncertainty of identity toolset : Analysing digital traces for user profiling. In Proceedings of the 7th International Conference on Security of Information and Networks (p. 254). ACM.

[2] Bakshy, E., Hofman, J., Mason, W., & Watts, D. (2011). Everyone's an influencer: quantifying influence on twitter. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining SE - WSDM '11, 65–74. doi:doi: 10.1145/1935826.1935845

[3] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (Vol. 146, pp. 1301–1309). doi:10.1007/s00256-005-0933-8

[4] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web - WWW '11 (p. 675). doi:10.1145/1963405.1963500

[5] Castillo, C., Mendoza, M., Poblete, B., & Authors, F. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5), 560–588. doi:10.1108/IntR-05-2012-0095

[6] Culotta, A., Ravi, N. K., & Cutler, J. (2015). Predicting the Demographics of Twitter Users from Website Traffic Data. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (pp. 72–78).

[7] Han, H., Nakawatase, H., & Oyama, K. (2014). Evaluating credibility of interest reflection on Twitter. International Journal of Web Information Systems, 10(4), 343 – 362.

[8] Hansson, K. (2015). Accommodating differences: Power, belonging, and representation online. Stockholm University.

[9] Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. The ANNALS of the American Academy of Political and Social Science, 659(1), 63–76. doi:10.1177/0002716215570866

[10] Harris, J. (n.d.). Gender, Twitter, and the Value of Taking Things Apart - Learning - Source: An OpenNews project. Source: Journalism code and the people who make it. Retrieved July 7, 2015, from https://source.opennews.org/en-US/learning/gender-twitter-and-value-taking-things-apart/

[11] Malhotra, N. (2014). Introducing language targeting. Twitter. Retrieved June 9, 2015, from https://blog.twitter.com/2014/introducing-language-targeting

[12] Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. Artificial Intelligence, 554–557. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

[13] Neis, P., & 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

[14] Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. Sociological Research Online, 18(3). Retrieved from http://www.socresonline.org.uk/18/3/7.html

[15] Steinmann, R., Häusler, E., Klettner, S., Schmidt, M., & Lin, Y. (2013). Gender Dimensions in UGC and VGI: A Desk-Based Study. Proceeding of GI_Forum 2013 – Creating the GISociety, 355–364. doi:10.1553/giscience2013s355

[16] Stephens, M. (2013). Gender and the GeoWeb: Divisions in the production of user-generated cartographic information. GeoJournal, 78(6), 981–996. doi:10.1007/s10708-013-9492-z

[17] Underwood, A. (2012). Gender targeting for Promoted Products now available. Twitter. Retrieved June 9, 2015, from https://blog.twitter.com/2012/gender-targeting-for-promoted-products-now-available