ASSESSING THE RELIABILITY OF CONTRIBUTORS IN VGI USING IMPLICIT FACTORS

R. Ghasemi Nejad [1, *], R. Ali Abbaspour [1], A. Chehreghan [2]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
- (r.ghasemi, abaspour)@ut.ac.ir
[2] Faculty of Mining Engineering, Sahand University of Technology, Tabriz, Iran
- chehreghan@sut.ac.ir

KEYWORDS: Volunteered Geographic Information (VGI), Reliability, OpenStreetMap (OSM), Intrinsic Measures, Contributors, Implicit Interactions

Commission IV, WG IV/3

ABSTRACT:

VGI projects include geographic information, which are the product of many unorganized volunteers, making it a challenge to ensure the quality of their information. In this field of study, several researchers have suggested using intrinsic factors to evaluate the quality of VGI instead of using explicit methods such as comparing with real or reference datasets. In addition, the measurement of the reliability of VGI contributors as an essential intrinsic factor in determining the credibility of their contributions remains an open question. Various types of contributors' activities and interactions are introduced and discussed in detail in this study at first. Then a comprehensive spatio-temporal contributor reliability model is proposed to assess their performance based on multiple implicit interactions between volunteers in their contribution process. Finally, several cities with different contribution rate (based on their population, number of users and area extent) are chosen and the proposed model is applied to the VGI data of selected regions, finally the results are compared and discussed.

1. INTRODUCTION

Volunteered geographic information (VGI) (Goodchild, 2007) is recently served as a growing source that refers to geographic data to be created, modified, or removed within a map by volunteers. However, the quality of VGI is questionable because it is heavily user-based and relies on community efforts. The community consists of non-professional users contributing to VGI with no pre-defined standards (Mohammadi and Sedaghat, 2021). Thus, contributor analysis provides a better overview of user-generated data and cannot be ignored (Rehrl et al., 2013).

There are a lot of activities or interactions between users in VGI that indirectly reflect their peers' feedback and evaluation relations. These implicit relations can be achieve for determining the user's reliability score, which changes over the contribution process time. This score is related to some factors, for example the user's previous performance, personal information; and feedbacks that receives from others (Yang et al., 2013).

Several papers suggest various models determine the reliability of VGI users from a diverse point of view (Bishr and Janowicz, 2010; D'Antonio et al., 2014; Lodigiani and Melchiori, 2016; Zhou and Zhao, 2016; Fogliaroni et al., 2018; Muttaqien et al., 2018; Zhang et al., 2021). However, they have gaps and do not take into account all effective parameters on the contributor's reliability level.

This research attempts to fill the gaps where computing the contributors' reliability level has been ignored by the researches of the community in VGI projects like OSM. To achieve this aim, a contributor's reliability model is suggested based on effective parameters that can be used to intrinsically assess the VGI quality. The proposed model is implemented to the regions with different numbers of users and populations to illustrate its scope of application.

The remaining of this document is organized in the following way. Section 2 introduces the different types of contribution processes in detail, and then a comprehensive reliability model for VGI users is presented. The first part of section 3 introduces the selected study areas and their datasets for experimental aims. The suggested model is implemented to the desired data in section 3.2. After that, in section 3.3, the results of the suggested model are evaluated and discussed in detail and finally, the conclusion is presented.

[*] Corresponding author

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W1-2022-215-2023 | © Author(s) 2023. CC BY 4.0 License.

2. PROPOSED MODEL

2.1 Various types of OSM activities

OSM is among the most popular VGI projects and has a vast community with over 8.3 million registered volunteers worldwide in 2022 (OpenStreetMap Wiki stats website). It enables its users to collaborate and modify the available data. The OSM dataset is the result of several interactions among its users, which implicitly reflect some feedback about them.

Each feature can be added to the OSM dataset with three shapes: node (for point features), way (for linear and areal features), or relation. In addition, some tags can be attached to them to express the thematic information.

When an individual registers in the OSM as a user. He/she observes a real world feature and its neighbourhood and wants to contribute about that feature to the OSM database. The different types of activities which the contributor can do are:

Creation: Adding a new object to the OSM dataset.

Modification: Editing an existing object which is created or modified before.

Deletion: Removing an existing object from the OSM dataset. It happens for some reasons: that object no longer exists in the reality, it is false, or its user is vicious (Fogliaroni et al., 2018).

Confirmation: making some contribution to the neighbourhood of an object without changing it (Keßler and De Groot, 2013).

2.2 VGI User's Reliability Model

In VGI and other crowdsourcing projects, most customers do not know contributors and how they produce content. Therefore they must rely on contributors' reliability level, representing their trust. In other words, a higher reliability level can mean a higher degree of confidence (Hendrikx et al., 2015). As Tavakolifard and Almeroth (2012) mentioned, reliability level is related to a user's behaviour based on available information about their past operations. Therefore, the various interactions and rating relations between contributors in the VGI community can determine contributors' reliability levels. Generally, the measures and interactions which have effects on the reliability level of OSM contributors are:

2.2.1 User's Personal information (Yang et al., 2013)

The personal information of contributors is available when they complete the profile at the registration time. For example, gender, age, cell phone number, address, education level, field of study, and e-mail. Such information may assist in understanding the user's behaviour and skill level (Zhang et al., 2021). Further, fake profiles can be recognized by evaluating this information with machine learning techniques (Xiao et al., 2015). Furthermore, the contributor's profile can be used as the initial value for the reliability level. However, The OSM does not disclose this information such as education level and e-mail addresses for privacy reasons. Such information may assist in understanding the behaviour and skill level of the contributor (Zhang et al., 2021). Also, fake profiles can be recognized by evaluating this information with machine learning techniques (Xiao et al., 2015). Furthermore, the contributor's profile can be used as the initial value for the reliability level. However, The OSM does not disclose this information for privacy reasons.

2.2.2 Type of contribution (Fogliaroni et al., 2018)

As mentioned in section 2.1, users make various types of contributions. Each type of activity affects the contributor's reliability level. For example, in creation, the contributor tries to complete the dataset by adding new objects. Consequently, this contribution positively impacts user reliability and should be rewarded. When the next contributor modifies a version, it means that this version is not completely correct and requires some edition. As a result, modification harms the reliability level score of the user.

2.2.3 The similarity of an object versions (Zhou and Zhao, 2016)

The effect of modification on the reliability level score of the user is based on the amounts of the changes (e.g. spatial or thematic similarity) done by the following user. For example, a major change shows a high reduction in reliability level.

2.2.4 Stability or time duration of each version (Severinsen et al., 2019)

Time between two consecutive versions of a feature also affects the previous user's reliability level. A longer time duration can show that a number of contributors see the latest version in this period and don't do any modification on it, so they implicitly confirm it (Severinsen et al., 2019). Indeed, this sentences is true in a specific extent of time (e.g., two years) as the content of information will be outdated after a time duration (Gusmini et al., 2017).

2.2.5 The number of modifications after each version of a feature (Keßler and De Groot, 2013)

The number of edits and modifications on a version after it is created affects the reliability level of its user. So, the high number of changes has negative impacts. (Keßler and De Groot, 2013).

2.2.6 The number of vicinity confirmations (Keßler and De Groot, 2013)

Implicit confirmations from the contributors who collaborate in the neighbourhood of an object positively influence its contributor's reliability level (Keßler and De Groot, 2013).

The suggested algorithm of how these measures are combined to compute the users' reliability levels is illustrated in **Figure 1**. Whenever each activity is done in the dataset of OSM, the associated user's reliability level is updated as this algorithm. In this algorithm, L is a time-ordered list of interactions. R_0, R_u, and R'_u are initial, current, and new reliability levels of the user u, respectively. The coefficients w_c, w_v, and w_n are the reward of creating, the negative effect of later modifications, and confirmation of adjacent users, respectively.

```
Algorithm: VGI Contributor Reputation System
Input: A time-ordered list of interactions (L), initial reputation value ($R_0$), Reward of creating a
new object ($w_c$), Negative effect of later modifications ($w_v$), Adjacent user's confirmation ($w_n$)
Result: reputation scores of users

While True do
    get next interaction from L
    Switch interaction do
        Case new user u is registered
            $R'_u \longleftarrow R_0$
            break;

        Case the first version of an object is created by u
            $R'_u \longleftarrow R_u + w_c$
            break;

        Case an existing version $v_i$ of object O, generated by u, is modified with another user and $v_j$
          is generated
            $R'_u \longleftarrow R_u - \left(\frac{1-Sim_{ij}}{\Delta t_{ij}}\right)$
            break;
            If i>1 (there are some previous versions)
                For k=(1:i)
                    $R'_{u_k} \longleftarrow R_{u_k} - w_v$
            If there exist some other objects (N numbers) whose last versions are generated before
              by some users in the vicinity of O
                For m=(1:N)
                    $R'_{u_m} \longleftarrow R_{u_m} + w_n$
            break;
```

Figure 1. The algorithm of the suggested contributor's reliability model

3. IMPLEMENTATION AND DISCUSSION

3.1 Study Regions and Dataset

All of the activities in the OSM database associated with each user affect that user's reliability level. In addition, the users are allowed to collaborate in OSM globally. Thus, the contributors' reliability model must consider all changes made to OSM data globally. However, as such data is massive, in this study the effectiveness of the suggested model is evaluated in several regions with different rates of contribution. Therefore, ten cities are chosen from various areas of Iran. It is tried to select cities with diverse populations and built-up area extents at the time of the selection procedure. Figure 2 illustrates the location of the ten chosen regions in Iran. Table 1 summarises the cities' names, area extent, the population of each region, and their population density (i.e. population/area) information.

From the Planet web page (https://planet.openstreetmap.org), the complete OSM history data in .pbf format with a size of 180 GB is downloaded in January 2022. OSM objects of each selected cities have been extracted. Each element has some properties, e.g., OSMID, username, user ID, timestamp, version, and some user-defined tags.

To obtain an overview of the state of the OSM dataset in the target cities, the number of contributions and unique OSM users in each city are shown in

Figure 3 and Figure 4, respectively.

City name	Area (km^2)	Population in 2016	Population Density (/km^2)
Tehran	751	8,693,706	11576.17
Esfahan	268	1,961,260	7318.13
Tabriz	244	1,558,693	6388.09
Zahedan	78	587,730	7535.00
Sari	35	309,820	8852.00
Bushehr	50	223,504	4470.08
Birjand	35	203,636	5818.17
Ilam	30	194,030	6467.67
Minab	35	73,170	2090.57
Bardsir	17	25,152	1479.53

Table 1. Population and area information of selected regions extracted from https://www.amar.org.ir/english

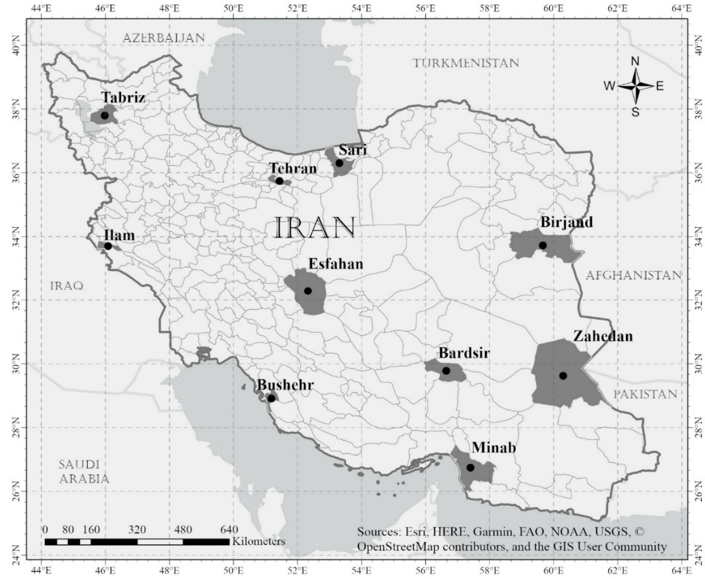
AZERBAIJAN
TURKMENISTAN
IRAN
IRAQ
AFGHANISTAN
PAKISTAN
SAUDI ARABIA
Tabriz
Sari
Tehran
Birjand
Ilam
Esfahan
Zahedan
Bardsir
Bushehr
Minab
N
W
E
S
0 80 160 320 480 640 Kilometers
Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community
44°E 46°E 48°E 50°E 52°E 54°E 56°E 58°E 60°E 62°E 64°E
24°N 26°N 28°N 30°N 32°N 34°N 36°N 38°N 40°N

Figure 2. Location of ten chosen areas in Iran

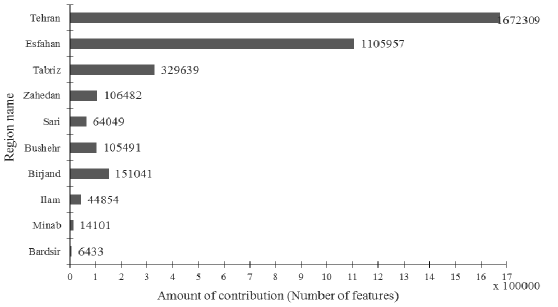

Figure 3. Amount of OSM contribution in each city.

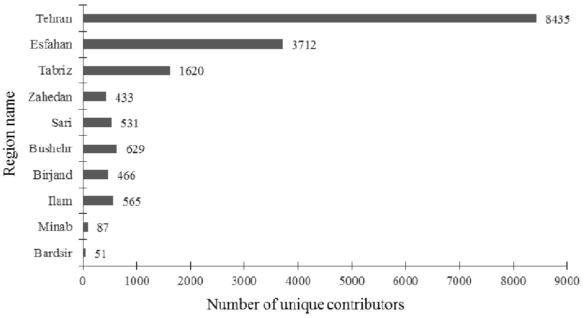

Figure 4. The number of unique users in each city.

Although the amount of population in a city may represent the number of contributors, a larger population is not a direct reflection of a greater amounts of activities (Mashhadi et al., 2015). Another parameters like extent of region and active users' number in addition to the population indicate the city's contribution rate. Therefore the number of users per population density can reflect the level of activity of a community within a region (Neis Pascal et al., 2013). In addition, the population density reduces the effect of the area size of the city on the results. Figure 5 shows users' number per population density in each city. This figure indicates that the highest value is related to the Tehran. In general, three different groups may exist:

- The first group involves the cities with the most significant values of contribution rates, like Tehran in this study.
- The second group consists of cities with average rate values, such as Esfahan and Tabriz.
- The last group covers all other cities with the lowest rate values.

In order to avoid repetition and to evaluate the suggested model, three different regions are chosen from the above groups (Tehran, Tabriz, and Zahedan) as experimental data in this study.

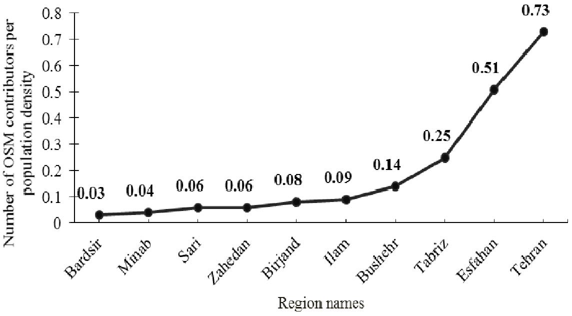

Figure 5. The users' number per population density in cities.

3.2 Implementation details and results

The model suggested can be employed as an online contributors' reliability model for VGI projects. In this study, to determine the reliability level of each user, the parameters w_c (the reward of creation), w_v (punishment of versioning), and w_n (the reward of confirmation from vicinity) are set respectively to the values 0.1, 0.01, and 0.15.

Measuring the similarity of two objects is also a common problem in GIS. There have been several studies on this issue that suggested a variety of methods to calculate the spatial or thematic similarity. (e.g. (Arkin et al., 1991; Masuyama, 2006; Fan et al., 2014)). However, in this paper, some existing methods are used to compute the similarity of two consecutive versions to complement this model.

The reliability levels of each contributor in the three cities of Tehran, Tabriz, and Zahedan are computed and updated in each iteration based on the proposed model. All contributors receive their reliability level according to the associated interactions. Finally, to make it easier to understand, the reliability level values are normalized within the range [0,1], and the contributors are divided into ten groups based on their normalized values at identical intervals. Table 2 shows ten groups and their intervals, the proportion of contributions, and users in each category in three study regions. From Table 2, some information can be extracted. Most of the users (more than 95%) belong to the first group with a reliability level of 0-0.1. It demonstrates that most users are not too active and make few contributions to the OSM dataset or are judged negatively by others. On the other hand, only a few users take high-reliability levels in the study regions.

Group	Reliability Interval	Tehran		Tabriz		Zahedan	
		users	contributions	users	contributions	users	contributions
1	0-0.1	98.79	39.43	95.27	36.43	96.26	36.96
2	0.1-0.2	0.52	8.84	2.03	14.84	1.87	22.65
3	0.2-0.3	0.29	7.62	0.68	8.07	0.0	0.0
4	0.3-0.4	0.06	2.08	0.34	3.72	0.93	18.76
5	0.4-0.5	0.12	9.67	0.0	0	0.0	0.0
6	0.5-0.6	0.06	7.05	0.68	5.37	0.0	0.0
7	0.6-0.7	0.0	0.0	0.68	20.17	0.0	0.0
8	0.7-0.8	0.0	0.0	0.0	0.0	0.0	`0.0
9	0.8-0.9	0.12	10.50	0.0	0.0	0.0	0.0
10	0.9-1.0	0.06	14.80	0.34	11.40	0.93	14.80

Table 2. The percentage of the selected cities users and their contributions in each group

3.3 Discussion

It can be concluded from the results shown in Table 2 that only a little number of OSM users take a high value of reliability level. For example:

Only one user in Tehran with username “kiaraSh-Q” and userID 2693232 takes the maximum reliability level value and belongs to group 10. Furthermore, two contributors with usernames, Ali Behzadian Nejad (ID: 671793) and Kesler (ID: 13908), are in group 9 with 0.8-0.9 intervals. These three volunteers contribute nearly 25% of the OSM features in this city.

In Tabriz, group 10 with the highest reliability level, consists of one user with username “Khalil_hz” and 1729935 ID. He contributes nearly 11.4% of OSM features of Tabriz.

The contributor Kesler (ID: 13908) is the only one with a high-reliability level with a contribution near 21.8% of all features in Zahedan.

Neis (Neis P, 2015) provides the HDYC webpage (https://hdyc.neis-one.org), which presents detailed and comprehensive information on how a volunteer contributes to the OSM on a global scale. It means that the active users in a specific region may also be active and make some contributions in other areas worldwide. The properties of the four users who are very active in the three study areas are captured from the HDYC webpage and mentioned in the Table 3. Furthermore, randomly three users are chosen from the group one (with the lowest reliability level values) to present their properties in Table 3 from this website. These table contents show the consistency between the results of the suggested model and the extracted information from the HDYC webpage.

Username	Registered	Map changes per Mapping days	Type
KiaraSh-Q	2015/02/22	856.61	Super Mapper
Kesler	2007/09/19	1566.33	Great Mapper
Khalil_hz	2013/08/30	390.28	Heavy Mapper
Ali Behzadian Nejad	2012/05/01	740.47	Casual Mapper
PATMAP	2011/04/01	102.16	Newbie
Iri98	2013/12/11	92.44	Newbie
Haddadian	2014/11/16	67	Newbie

Table 3. Summary of users’ properties extracted from the HDYC webpage.

The suggested model is implemented to the selected cities Tehran, Tabriz, and Zahedan OSM data with various populations, numbers of users, and areas extent. A comparison of the results in these three regions shows that in the region with a small value for the number of users per density of population, the probability of the existence of contributors in the various groups with higher reliability levels is little. It may be because of the fewer activities of contributors. For instance, in Zahedan, the contributors are classified into only four groups that three of them have a low-reliability level. Generally, this leads to the conclusion that the higher contribution rate in an area shows a larger the variety of contributors in different categories.

4. CONCLUSION

VGI projects, through the collection and publication of geographical information worldwide, can be a source of valuable and helpful content for various services and applications. However, estimating the quality of this content remains a questionable and open research problem. Because of the nature of this data and some restrictions for actual datasets, using the intrinsic indicators is suggested. Assessing the reliability of contributors and identifying their levels can be a critical intrinsic factor for verifying the quality. Thus, it is essential to understand how each contributor behaves and gives feedback from others in VGI. At first, the contribution process types are explained in detail in the present study, and effective user interactions are described as implicit and indirect evaluation relationships. These relationships can show each user's performance as negative or positive feedback and can calculate their reliability levels via the suggested model. The proposed reliability model may lead to new insights into the quality assessment. The reliability level of the users in Tehran, Tabriz, and Zahedan are computed using the suggested model, and the obtained results are discussed. They show that the reliability level of most of the users (more than 95%) is a small value. This means that most users are less active or achieve negative judge from others. Furthermore, few users contribute most of the features and take higher reliability levels. Anyway, in this study, the authors attempt to propose a comprehensive contributor reliability model, consider most of the temporal and spatial relations in the database, and eliminate the other research gaps.

REFERENCES

Arkin EM, Chew LP, Huttenlocher DP, Kedem K, Mitchell JS, 1991. An efficiently computable metric for comparing polygonal shapes. *Cornell Univ Ithaca NY.*

Bishr M, Janowicz K, 2010. Can we trust information?-the case of volunteered geographic information. *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume*.

D'Antonio F, Fogliaroni P, Kauppinen T, 2014. VGI edit history reveals data trustworthiness and user reputation. *17th AGILE Conference on Geographic Information Science* 3-6 june; Castello de la Plana.

Fan H, Zipf A, Fu Q, Neis P, 2014. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4),700-719.

Fogliaroni P, D'Antonio F, Clementini E, 2018. Data trustworthiness and user reputation as indicators of VGI quality. *Geo-spatial Information Science*, 21(3),213-233.

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

Gusmini M, Jabeur N, Karam R, Melchiori M, Renso C, 2017. Reputation evaluation of georeferenced data for crowd-sensed applications. *Procedia Computer Science*, 109,656-663.

Hendrikx F, Bubendorfer K, Chard R, 2015. Reputation systems: A survey and taxonomy. *Journal of Parallel and Distributed Computing*, 75,184-197.

Keßler C, De Groot RTA. 2013. Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. Geographic information science at the heart of Europe. Springer; p. 21-37.

Lodigiani C, Melchiori M, 2016. A pagerank-based reputation model for VGI data. *Procedia Computer Science*, 98,566-571.

Mashhadi A, Quattrone G, Capra L. 2015. The impact of society on volunteered geographic information: The case of OpenStreetMap. OpenStreetMap in GIScience. Springer; p. 125-141.

Masuyama A, 2006. Methods for detecting apparent differences between spatial tessellations at different time points. *International Journal of Geographical Information Science*, 20(6),633-648.

Mohammadi N, Sedaghat A, 2021. A framework for classification of volunteered geographic data based on user's need. *Geocarto International*, 36(11),1276-1291.

Muttaqien BI, Ostermann FO, Lemmens RL, 2018. Modeling aggregated expertise of user contributions to assess the credibility of OpenStreetMap features. *Transactions in GIS*, 22(3),823-841.

Neis P, 2015. How did you contribute to OpenStreetMap. [accessed]. https://hdyc.neis-one.org.

Neis P, Zielstra D, Zipf A, 2013. Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future internet*, 5(2),282-300.

OpenStreetMap Wiki stats website. [accessed]. https://wiki.openstreetmap.org/wiki/Stats.

Rehrl K, Gröechenig S, Hochmair H, Leitinger S, Steinmann R, Wagner A. 2013. A conceptual model for analyzing contribution patterns in the context of VGI. Progress in location-based services. Springer; p. 373-388.

Severinsen J, de Roiste M, Reitsma F, Hartato E, 2019. VGTrust: measuring trust for volunteered geographic information. *International Journal of Geographical Information Science*, 33(8),1683-1701.

Tavakolifard M, Almeroth KC, 2012. Social computing: an intersection of recommender systems, trust/reputation systems, and social networks. *IEEE Network*, 26(4),53-58.

Xiao C, Freeman DM, Hwa T, 2015. Detecting clusters of fake accounts in online social networks. *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*.

Yang H, Zhang J, Roe P, 2013. Reputation modelling in Citizen Science for environmental acoustic data analysis. *Social Network Analysis and Mining*, 3(3),419-435.

Zhang D, Ge Y, Stein A, Zhang WB, 2021. Ranking of VGI contributor reputation using an evaluation-based weighted pagerank. *Transactions in GIS*.

Zhou X, Zhao Y, 2016. A version-similarity based trust degree computation model for crowdsourcing geographic data *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 41.