A Bayesian Framework for Reputation in Citizen Science

Joan Garriga[1,2], Jaume Piera[1,3], and Frederic Bartumeus[1,2,4]

[1] Centre de Recerca Ecològica i Aplicacions Forestals (CREAF),
Carrer de les Cases Sert 54, 08193, Cerdanyola del Vallès, Barcelona
[2] Centre d'Estudis Avançats de Blanes (CEAB-CSIC),
Carrer Accés Cala Sant Francesc 14, 17300, Girona
[3] Institut de Ciències del Mar (ICM-CSIC),
Passeig Marítim de la Barceloneta 37-49, 08003, Barcelona
[4] Institució Catalana de Recerca i Estudis Avançats (ICREA),
Passeig de Lluís Companys 23, 08010, Barcelona
jgarriga@ceab.csic.es

Abstract. The viability of any Citizen Science (CS) research program is absolutely conditioned to the engagement of the citizen. In a CS framework in which participants are expected to perform actions that can be later on validated, the incorporation of a reputation system can be a successful strategy to increase the overall data quality and the likelihood of engagement, and also to evaluate how close citizens fulfill the goals of the CS research program. Under the assumption that participant actions are validated using a simple discrete rating system, current reputation models, thoroughly applied in e-platform services, can be easily adapted to be used in CS frameworks. However, current reputation models implicitly assume that rated items and scored agents are the same entity, and this does not necessarily hold in a CS framework, where one may want to rate actions but score the participants generating it. We present a simple approach based on a Bayesian network representing the flow described above (user, action, validation), where participants are aggregated in a discrete set of user classes and we use the global evidence in the data base to estimate both the prior and the posterior distribution of the user classes. Afterwards, we evaluate the expertise of each participant by computing the user-class likelihood of the sequence of actions/validations observed for that user. As a proof of concept we implement our model in a real CS case, namely the Mosquito Alert project.

Keywords: citizen science, reputation system, Bayesian network

1 Introduction

Since its origins, back in the mid-90's, *citizen science* (CS) has been questioned by the scientific community as an adequate scientific methodology [7]. Pros and cons aside, a basic principle to bring citizens and scientists into a productive relationship is to match the public understanding of science with the science's

understanding of the public [7]. To this end, modern citizen science is rethinking methods for citizen engagement [3, 1]. Key concepts in participants engagement are connection and reward. Connection refers to connecting the scientific goals of the CS research program with the citizen perception of a social worry or interest (the basic motivation to start cooperating). Reward refers to providing feedback that can be neatly perceived as a reward (the basic motivation to keep cooperating). Nevertheless, it is well known by psychologists [1] that the effect of a reward resides in its expectation and vanishes as soon as it is achieved. Thus, in order to increase the likelihood of participation in the long run, it is necessary to generate continuous reward expectations. A successful strategy to achieve sustained participation requires the implementation of a reputation system as a core component of any CS research program. In addition, well-grounded reputation systems provide back-end information of participants that is valuable to augment data quality and to increase the *fitness for use* [12] of CS research programs.

Reputation is a broad concept not only suitable to people but also to many kinds of things or services [2]. Extending the notion given in [11], reputation is *the perception that an agent (or item) creates through past actions about its intentions, norms, knowledge, expertise or value*. Reputation can be seen as an asset, not only to promote oneself, but also to help us to make sound judgments in the absence of any better information. However, reputation is highly contextual and what works well in a specific context may inevitably fail in many others. As a consequence details about reputation systems are profusely treated in the literature [2, 8, 6, 11, 14]. The simplest reputation systems scale down to a ranking/voting system where information is aggregated into a single score used to qualify and sort items (*e.g.* songs in *iTunes*, users in *Stackoverflow*). Systems for collecting and representing reputation information are usually based on simple rating mechanisms such as thumbs-up/down or a five star rating. The difficulties arise at the time of aggregating this information.

Many rating aggregation systems recently proposed (*e.g.* Amazon, iTunes, YouTube) are different forms of *Bayesian Rating* (BR), a pseudo-average computed as a weighted combination of the average rating of a particular item and the average rating for all items. In a k-way rating system, (*i.e* with k discrete rating levels $r \in \{1, \ldots, k\}$), with a total of m rates and an overall rating $\bar{r}\,(all) = \sum_{j=1}^{m} r\,(y)_j\,/m$, the BR of an item y with n ratings, and an average rating $\bar{r}\,(y) = \sum_{j=1}^{n} r\,(y)_j\,/n$ is given by,

$$BR\,(y) = \frac{n\;\bar{r}\,(y) + m\;\bar{r}\,(all)}{m+n} = w\;\bar{r}\,(y) + (1-w)\;\;\bar{r}\,(all) \tag{1}$$

with $w = n/\,(n+m)$. A clear benefit of BR is that an item with only a few ratings (*i.e* $w \to 0$) will approach the overall mean rating, hence, does not receive the lowest (unfair and discouraging) rate but the average rate, while the more the item is rated (*i.e* $n \gg 0$) the largest the weight of its own average rating. In any case $m \gg n$, hence the scoring is focused on the quality of the ratings rather than on the quantity of ratings.

The Beta reputation system [9] (binomial) and the Dirichlet reputation system [10] (DR), the multinomial generalization of the former, are reputation models based on a sound statistical machinery that explains away the *Bayesian rating* concept and frames it in a real Bayesian perspective. Consider a k-way rating system and let the rating level be indexed by i (*i.e.* $1 \leq i \leq k$). Let $n(y)_i$ be the rating counts for item y (the observed evidence), and let a_i be a base rate expressing the biases in our prior belief about each rating level. A Dirichlet (or Beta for $k = 2$) rating yields a multinomial probability distribution $S(y)_i$ over the k rating levels, where the expectation value for each rating level is computed as,

$$S(y)_i = \frac{n(y)_i + C\,a_i}{C + n(y)} \tag{2}$$

where $n(y) = \sum_{i=1}^{k} n(y)_i$ and C is the *a priori* constant that can be set to $C = k$ if we consider a uniform prior (*i.e.* $a_i = 1/k$). In this case Equation 2 defaults to the classical Laplace smoothing [9]. The larger the value of C with respect to k the less the influence of the observed ratings and the more $S(y)_i$ will approach the base rate a_i. Assuming the k rating levels evenly distributed in the range [0, 1], a point estimate reputation score is computed as,

$$DR(y) = \sum_{i=1}^{k} \frac{i}{k} S(y)_i \tag{3}$$

Multinomial form aside, the similarity with BR (Eq. 1) is clear. But the difference can not be obviated: while the weighting in Equation 1 emerges from a pure frequentist persepective, in a DR the factor $C\,a_i$ can convey specific *a priori* information provided by domain experts or any other external source [9].

Agents (and in particular human agents) may change their behaviour over time. This issue is usually approached either by incorporating a *cutoff factor* that limits the series of ratings to the most recent ones, up to a given period or a given number, or by introducing a *longevity factor* that assigns a time relative weight to ratings [9]. An additional concern in reputation systems for e-service platforms is its resistance against typical strategies for reputation cheating (*e.g.* whitewashing, sybil attacks, fraudulent ranking) reviewed in *e.g.* [4].

CS research programs constitute a different scenario where the aim is not to promote user interaction but to collect useful data for their scientific goals. Hence, reputation issues do not arise from peer-to-peer interaction but from the need to increase citizen engagement and data quality. However, a systematic review of 234 CS research programs presented in [5] reveals that despite of this general concern on data quality, very few has been made in terms of participants' reputation. Data validation is usually performed by a core of domain experts or project coordinators, eventually assited by authomated methods or with some level of intra-community interaction (*e.g. eBird*, *Galaxy Zoo*, *iSpot*) or more broadly via crowd-sourcing (*e.g.* www.crowdcrafting.org). In a few cases, local coordinators take into account the participants' experience for validating data

(*e.g Common Bird Monitoring*, *Weather Observations Website*), and just in a handful of them it is the community of participants itself who directly validate data (*e.g Galaxy Zoo*, *iSpot*, *oldWeather*). Among the later, *Notes from Nature* and *iSpot* [15] are the ones going further in terms of community-based validation and participants' reputation, implementing simple agreement based algorithms to rank participants and assign digital badges in recognition of specific achievements. Community-based validation explicitly requires a core reputation system integrated with the CS research program. However, there is not a general approach and each research program implements reputation in a functional way to fit its needs, neither framing its system in general conceptual frameworks, nor making it available to the scientific community.

Notably, an implicit assumption in any of the reputation models above is that the agent (or item) being rated is the one that is scored and, more explicitely, that the rating system used to collect ratings for an agent is the rating system used to score that agent (*e.g.* Equation2). This apparently obvious and irrelevant assumption might become more subtle in a CS framework. CS research programs expect participants to perform a set of actions (basically, reporting information in specific formats) and these actions are later on validated. In this case, the rated items are the actions, but the scored agents are the participants. Importantly, each kind of action might require its own discrete rating system (not necessarily coincident in the number of levels). Yet the expertise of participants might be expressed based on a specific set of user classes (with its own number of levels), and scored based on the ratings of all their possibly different actions. A straightforward way to overcome this problem is to compute separate scores for each type of action and get an overall score using a weighted combination of the former. Alternatively, we propose a novel model for user reputation based on a Bayesian network describing the characteristic flow of CS research programs, (*i.e.* user, action, validation). The proposed method (i) decouples action rating from participant scoring, (ii) provides a unified framework to process validation information regardless of the rating levels used for each kind of action, (iii) accounts for a good balance of both quality and quantity of evidence, and (iv) is more responsive to participants' actions, which may augment engagement dynamics.

2 Mosquito Alert: CS for public health

Mosquito Alert (MA) is a CS research program initially devised to monitor the expansion in Spain of the Asian tiger mosquito (*Aedes albopictus*), a disease carrying mosquito. Since the expansion of the Zika virus threat in 2016, MA included the monitoring of the yellow fever mosquito (*Aedes aegypti*). Both species are world wide distributed, living in urban environments, and being specially invasive and agressive vectors of tropical infectious diseases such as Dengue, Chikungunya, Yellow Fever, and Zika.

Aside from its scientific goals (*e.g.* unveiling the population expansion mechanisms, forecasting vector and disease threats), a particular challenge for MA

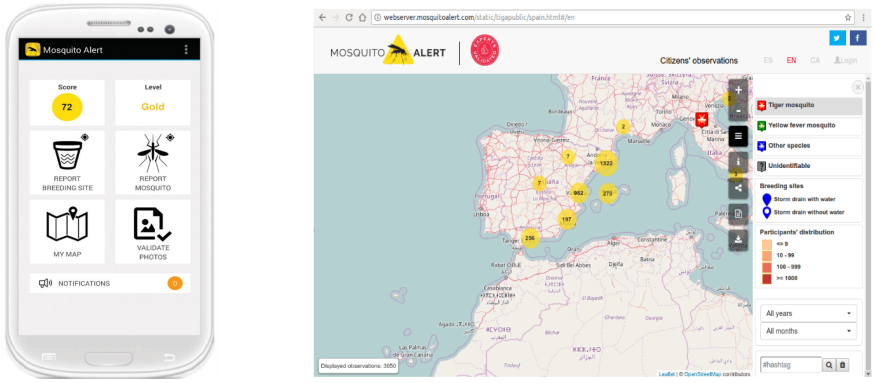

Fig. 1. Mosquito Alert (left) home screen of the app showing the rank and category; (right) web map showing the validated reports.

arises from its impact on the public health domain. MA is aimed to provide reliable early warning information (in recently colonised areas), and real-time management information (in areas where it resides) to public health administrators. Public health administrations at different organizational levels in the territory, use MA to improve their surveillance and control programs with the goal of decreasing mosquito populations, specially in urban areas. Because of all this, MA is designed as a multi-platform system structured as follows:

1. The MA smartphone app (freely available for Android and iOS), by means of which citizen can send reports of observations of mosquitoes (and their breeding sites) potentially belonging to disease vector species (namely the Asian Tiger and the Yellow Fever mosquito).
2. The corresponding server-side functionality (Django, SQL) managing the reception and storage of data, along with an ever evolving set of tools for the management and analysis of the data, including machine-learning algorithms to help automating the validation of information [13].
3. A private platform called *Ento Lab*. This is a restricted access service through which a set of experts can make a previous filtering of inappropriate reports and classify the rest as either positive or negative ones. Only classified reports are afterwards made visible to the rest of the services.
4. Another private platform named *Managers Portal* which grants on-demand-access to stakeholders (e.g. public health administrations, mosquito control services, private mosquito control companies), open GIS tools to visualize all the available data in the portal (including their own imported management data), and the possibility to directly comunicate control actions through the app.
5. A public platform http://www.mosquitoalert.com, providing data and visualization tools to all the public via interactive maps, where participants can find their individual contributions validated by the experts (Figure 1, right).

The direct involvement of public health institutions make citizen truly conscious of the usefulness of their contributions (much beyond science). Triggering mosquito control actions in the territory through MA participation constitutes the necessary reward to keep citizens engaged in the research program.

Table 1. Summary of Mosquit Alert's data-base (2015-2016).

	total	NC	hd	-2	-1	0	+1	+2
adult	4177	19	188	429	128	655	1317	1441
%		0.00	0.05	0.10	0.03	0.16	0.32	0.34

	total	NC	hd	-1	0	+1
bSite	1172	564	160	90	129	229
%		0.48	0.14	0.08	0.11	0.20

The work described in this paper is based on data corresponding to the last two years (2015-2016) of MA, summarized in Table 1, with more than 30000 app downloads, 2993 active users and 5349 reports submited. Reports are of type *adult* (4177) correponding to observations of adult mosquitoes (either Asian tiger or Yellow Fever) or *bSite* (1172) corresponding to potential mosquitoes' breeding sites. Reports of type *adult* and *bSite* are reviewed by experts who manually label them. Reports of type *adult* are labelled as: -2, definitely not corresponding to the species of interest; -1, probably not corresponding to the species of interest; 0, can not tell; 1, probably corresponding to the species of interest; and 2, definitely corresponding to the species of interest. Reports of type *bSite* are labelled as: -1, does not correspond to a breeding site; 0, can not tell; and $+1$, does correspond to a breeding site. *NC* stands for not-classified reports which either do not provide an image or have not yet been reviewed by the experts. The *hd* stands for hidden reports which correspond to reports with improper images that are hidden by the experts (not shown in the map).

3 Methods

Let's consider a Bayesian network describing the characteristic flow in a CS research program, what we call the User-Action-Validation (UAV) network (Figure 2, left, lightgrey nodes). In our particular case, the nodes of the UAV network represent the following:

- Users are participants of the CS research program aggregated in a variable $U = \{1, \ldots, k\}$ specifying an ordinal set of user-classes with increasing levels of expertise (*e.g.* beginner, advanced, expert). Beyond the intuition that U should be a discrete ordinal variable, we have no prior idea of the optimal number of user classes we should define nor about their prior distribution.
- Actions consist in submitting reports either of type *adult* or *bSite*. Thus, we define $A = \{adult, bSite\}$ such that $P(A|U)$ specifies the probability that a

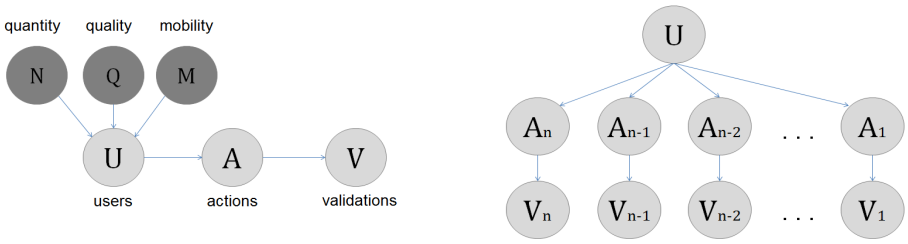

Fig. 2. The User-Action-Validation Bayesian network to model users' expertise on the Mosquito Alert's platform.

user of a given class submits a report of a particular type. We assume that each submitted report is an independent event.

– Validations are expressed in terms of ratings and each kind of action can be rated with a different number of discrete rating levels. Thus, we define a variable V such that $P(V|A)$ specifies the probability that an action of a given type gets the corresponding rating (or, transitorily, no rating), namely $(V|A = adult) = \{None, hd, -2, -1, 0, 1, 2\}$ and $(V|A = bSite) = \{None, hd, -1, 0, 1\}$.

From the joint distribution $P(U, A, V)$ and by simple Bayes' rule we have,

$$P(U|A, V) = \frac{1}{P(A, V)} P(V|A)\ P(A|U)\ P(U)$$

where $P(A, V)$ is just a normalization constant. So, if we have some means to estimate a prior distribution $P(U)$, and the conditional distributions $P(A|U)$ and $P(V|A)$, we can evaluate the posterior distribution of user-class expertise for a user y given the observed sequence of actions/validations $S(y)$ (Figure 2, right),

$$P(U|S(y)) = \prod_{A,V \in S(y)} P(U|A, V)$$

3.1 Guessing a prior $P(U)$

We start by guessing a number of user expertise levels. For each user we consider three features regarding to the sequence $S(y)$: (i) the quantity of reports, (ii) the quality of the reports, and (iii) a user's mobility index mI describing the average area covered by the user defined as the variance of the pairwise geolocation distances between the reports,

$$mI(y) = \frac{1}{2\,|S(y)|^2} \sum_{(p,q) \in S(y)} [\,(p_x - q_x)^2 + (p_y - q_y)^2\,] \tag{4}$$

where (p_x, p_y), (q_x, q_y) are the geolocation coordinates.

Based on these features we define the following proxy variables of the user-class U (Figure 2, left, darkgrey nodes): (i) a quantitative proxy aggregating users sending less or more than a given number θ_1 of reports ($N = \{less, more\}$); (ii) a mobility proxy aggregating users with a mobility index lower/higher than a given value θ_2 ($M = \{lower, higher\}$); (iii) a quality proxy aggregating reports in four categories: hidden, low quality (those labeled as $(-2, -1)$), medium quality (those labeled as 0), and high quality (those labeled as $(1, 2)$), ($Q = \{hidden, low, medium, high\}$, we do not count here not-classified reports). Note that (i) and (ii) account for the attitude of the participants, while (iii) accounts for their skills, and both aspects are deemed important. The joint combination of the above three proxys results in a primary partition of the users' expertise space into 16 categories summarized in Table 2. The threshold values where selected by looking at the corresponding histograms and taking the values that yield the most possible balanced distribution.

Table 2. Join distribution of the proxy variables, $P(N, M, Q)$, resulting in 16 expertise categories. Threshold values: $\theta_1 = 2$, $\theta_2 = 10^{-9}$.

	$N \leq \theta_1$, $M \leq \theta_2$	$N \leq \theta_1$, $M > \theta_2$	$N > \theta_1$, $M \leq \theta_2$	$N > \theta_1$, $M > \theta_2$
hiddden	0.0427	0.0220	0.0004	0.0086
low	0.0974	0.0132	0.0027	0.0228
medium	0.0853	0.0211	0.0065	0.0519
high	0.2685	0.0707	0.0322	0.2541

By looking at this table, we should now infer a set of user classes with increasing levels of expertise. We prioritize as following: (i) the quality of the reports before the quantity (low quality reports just result in a waste of experts' time); (ii) the quantity of reports before the mobility index of the users (we give the lowest priority to the mobility index because the meaning of this variable is double folded: for surveillance purposes it is important that participants send reports covering the broadest geographical area possible, but for control purposes it is also important that participants keep sending reports within their neighbourhoods). Also, we are not looking here for a fine grain discretization of the expertise space. Taking into account the unbalances present in Table 2 it looks reasonable to impose an ordering of the 16 expertise categories into a set of $k = 6$ user-classes, *i.e* $U = \{1, \ldots, 6\}$, as shown in Table 3. Tables 2 and 3 together express a joint distribution $P(N, Q, M, U)$ from which the prior $P(U)$ follows straightforwardly by marginalization,

$$P(U) = \sum_{N,Q,M} P(N, Q, M, U) \tag{5}$$

and we get a temptative prior for the user-class variable (Table 4).

Having defined the user classes we know the user-class value of each report, and we can make estimations (*maximum a posteriori*, MAP) for the action con-

Table 3. Definition of user classes

	$N \leq \theta_1,\ M \leq \theta_2$	$N \leq \theta_1,\ M > \theta_2$	$N > \theta_1,\ M \leq \theta_2$	$N > \theta_1,\ M > \theta_2$
hiddden	1	1	1	2
low	2	2	3	3
medium	3	4	4	4
high	5	6	6	6

Table 4. Expertise-class prior distribution, $P(U)$

U	1	2	3	4	5	6
	0.0648	0.1190	0.1106	0.0792	0.2691	0.3573

ditional distribution $P(A|U)$ (Table 5) and also for the validation conditional distributions $P(V|A = adult)$ and $P(V|A = bSite)$ (Table 6).

3.2 Computing the posterior $P(U|A, V)$

Applying Bayes' rule we compute the posterior distribution $P(U|A, V)$,

$$P(U|A, V) = \frac{1}{W}\ P(V|A)\ P(A|U)\ P(U) \tag{6}$$

where $W = \sum_{i=1}^{k} p(u_i|A, V)$. Equation 6 evaluates the probability that an action (report) of a given type, with a given validation (rating), belongs to a particular user-class (Table 7).

Note that, in our Bayesian approach ratings become a fuzzy qualification of the user-class. Also note that this is indeed a two parameter model (θ_1, θ_2) allowing a degree of control over the user-class prior and, ultimately, over the user-class posterior distributions, (*i.e.* we can push the classification of reports to a lower/upper user-class by increasing/decreasing either one or both of the parameters (this is shown later in Figure 4).

Table 5. Action conditional distribution, $P(A|U)$

U	1	2	3	4	5	6
adult	.3454	0.7994	0.8456	0.4234	0.9339	0.9144
bSite	0.6546	0.2006	0.1544	0.5766	0.0661	0.0856
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

3.3 Computing users' scores.

Note that, so far, our UAV-network (Figure 2 , left) just yields the user-class distribution of a single action, not the users' scoring that we aim. To compute

Table 6. Validation conditional distribution, $P(V|A,U)$

adult reports

U	1	2	3	4	5	6
NC	0.0059	0.0059	0.0059	0.0059	0.0059	0.0059
hd	0.9643	0.0559	0.0022	0.0033	0.0008	0.0006
-2	0.0060	0.7326	0.1401	0.0033	0.0008	0.0006
-1	0.0060	0.1996	0.0657	0.0033	0.0008	0.0006
0	0.0060	0.0020	0.7817	0.9778	0.0008	0.0006
+1	0.0060	0.0020	0.0022	0.0033	0.5511	0.4135
+2	0.0060	0.0020	0.0022	0.0033	0.4397	0.5780
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

bSite reports

U	1	2	3	4	5	6
NC	0.4742	0.4742	0.4742	0.4742	0.4742	0.4742
hd	0.5153	0.0932	0.0063	0.0064	0.0060	0.0035
-1	0.0035	0.4193	0.1837	0.0064	0.0060	0.0035
0	0.0035	0.0067	0.3294	0.5066	0.0060	0.0035
+1	0.0035	0.0067	0.0063	0.0064	0.5079	0.5152
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 7. User-class posterior distributions

adult reports

L	NC	hd	-2	-1	0	+1	+2
$U=1$	0.0272	0.7810	0.0016	0.0052	0.0012	0.0005	0.0004
$U=2$	0.1157	0.1926	0.8351	0.7363	0.0018	0.0007	0.0006
$U=3$	0.1137	0.0074	0.1571	0.2382	0.6856	0.0007	0.0007
$U=4$	0.0408	0.0040	0.0013	0.0042	0.3074	0.0004	0.0004
$U=5$	0.3055	0.0075	0.0025	0.0080	0.0019	0.5050	0.3683
$U=6$	0.3972	0.0075	0.0025	0.0080	0.0019	0.4927	0.6296
	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

bSite reports

L	NC	hd	-1	0	+1
$U=1$	0.2389	0.8849	0.0107	0.0050	0.0058
$U=2$	0.1346	0.0901	0.7252	0.0054	0.0062
$U=3$	0.0963	0.0044	0.2273	0.1922	0.0042
$U=4$	0.2575	0.0119	0.0212	0.7900	0.0115
$U=5$	0.1003	0.0043	0.0077	0.0036	0.3545
$U=6$	0.1724	0.0044	0.0078	0.0037	0.6178
	1.0000	1.0000	1.0000	1.0000	1.0000

the scores we consider the report sequences $S(y)$ (Figure 2, right). To add some dynamics to the model we consider a third parameter θ_3 (a *cutoff factor*) that limits the sequences to the last θ_3 reports. Assuming an *iid* sequence of reports, the corresponding user-class posterior distribution is given by,

$$P\left(U|S\left(y\right)\right) = \frac{P_0}{P_w} \prod_{j=1}^{\theta_3} P\left(U|A_j, V_j\right) \tag{7}$$

where $P_0 = (1/k, \ldots, 1/k)$ sets a starting uniform user-class distribution and $P_w = \sum_{i=1}^{k} P\left(u_i|S\right)$ is a normalization factor. Afterwards, an expertise score can be computed as the user's expected user-class,

$$X\left(y\right) = \frac{1}{k} E[U]_{P(U|S)} = \frac{1}{k} \sum_{i=1}^{k} u_i \, p\left(u_i|S\right) \tag{8}$$

Equation 8 yields a normalized score with a lower bound given by $\frac{1}{k} E[U]_{P_0}$ which avoids a discouraging zero-score for new comers. Usually, the computation of Equation 7 is subject to numerical precision problems and therefore we implement a log computation as,

$$\log P\left(U|S\left(y\right)\right) = \log P_0 + \sum_{j=1}^{\theta_3} \log \, P\left(U|A_j, V_j\right) - \log P_w \tag{9}$$

For gamification purposes, users are ranked based on their scores. Ties are solved by mobility index. The rank position, not the score, is notified to the users via the smartphone app, together with a quantile based category label as either *gold*, *silver* or *bronze* (Figure 1, left).

In summary, we use the global evidence in the data base to guess the joint distribution $P\left(U, Q, N, M\right)$ and estimate a prior $P\left(U\right)$, from which we can derive the posteriors $P\left(U|A, V\right)$. Afterwards, we use the evidence observed for each particular user $S\left(y\right)$, to evaluate the posterior distribution $P\left(U|S\left(y\right)\right)$ and compute a score for that user. Essentially, our scoring model is a naïve Bayes classifier where the number of features varies with the number of reports used to qualify the user. The larger the number of reports, the better the profiling of the user. Because we use the global evidence to estimate the user-class distribution, the scores change dynamically as the contents of the data base grow and all individual expertise scores are in part dependent on the overall average performance.

4 Results

Based on the user-class posterior distributions (Table 7) and applying Equations 9 and 8 we get the results shown in Figure 3. In the x-axis, users are ranked by score (blue line). Scores are plotted together with scaled versions of the *mobility index* (yellow), the number of *breedingSite* reports (cyan) and the number of *adultMosquito* reports (magenta). The scoring yields several plateaus corresponding to typical number of submitted reports. We highlight (darkgrey rows in Table 8) the position of users who only submitted one report, classified as +1 (positions 1494:2053, 560 users), or classified as +2 (positions 2244:2684,

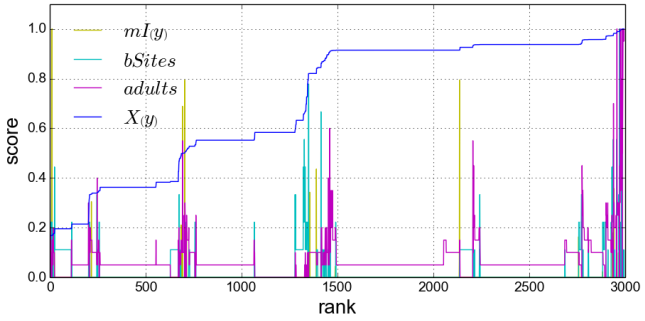

Fig. 3. Ranked score plot

Table 8. Ranked score table

				adult					*bSite*		
rank	score	hd	-2	-1	0	+1	+2	hd	-1	0	+1
1	.16668463							4			
2	.16684409							3			
3	.16728058	4									
26:112	.19559913							1			
116:201	.21448389	1									
221:245	.33933867		2								
263:553	.36258186		1								
555:628	.38297512			1							
629:670	.38557248								1		
730:756	.52790362				2						
764:1064	.55214793				1						
1069:1278	.58333333										
1284:1321	.63213868									1	
1356:1390	.82074814				1	1					
1494:2053	.91446382					1					
2139:2205	.92599170										1
2244:2684	.93735892						1				
2823:2882	.95749959						2				
2892:2899	.95809921										2
2911:2926	.97219402						3				
2936:2937	.97351476										3
2944:2952	.98252239						4				
2962:2965	.98930969						5				
2978	.99617983						7				
2984	.99888390										9
2991	.99999356					1	19				
2992	.99999384						18				1
2993	.99999632						20				

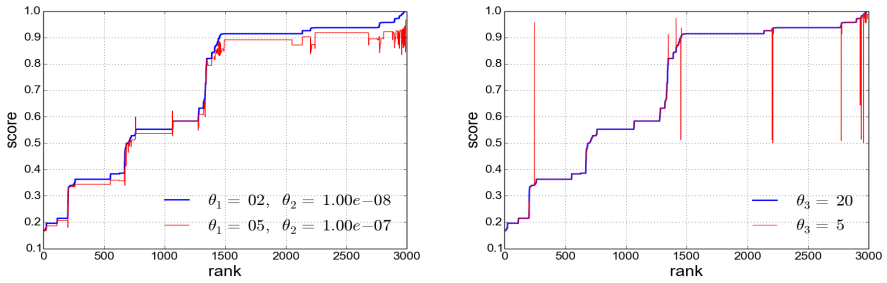

Fig. 4. Effect of the 3 threshold parameters: (left) effect of the number of reports, θ_1, and the mobility index, θ_2; (right) effect of the cutoff factor, θ_3

441 users), which correspond to the largest plateaus in Figure 3. As expected, the larger the number of positive reports the higher the rank (Table 8 bottom rows), and the larger the number of negative reports the lower the rank (Table 8 top rows), with hidden reports being strongly punished.

We also analyze the effect of the threshold parameters $(\theta_1, \theta_2, \theta_3)$ (Figure 4). In the x-axis, participants are ranked by score (blue line). The red line depicts the score coresponding to a different value of the threshold parameters. In general, the scores do not change much in terms of value though sudden breaks in the increasing trend of the red line reveal users whose position in the ranking has been affected by the change of the parameter value. The plateaus remain almost invariant and we only appreciate some changes of position at the borders of the plateaus. Increasing (decreasing) θ_1 and θ_2 (Figure 4, left) together, move reports to the left (right) columns of Table 2 and consequently force a change in the prior distribution. As a result, the plateaus are globally pushed lower (higher). This change is propagated to the posteriors and originates also the rank changes that can be observed at the borders of the plateaus (Figure 4, left). In the case of θ_3 we observe that by not looking so far in the past (Figure 4, right), some low rank users are upgraded (users who clearly improved their performance over time) while some high rank users are downgraded (users who worsened their performance over time).

The dynamics of the scoring are also shown. In Figure 5 (top) we simulate the evolution of the score (blue line) and rank (red line) for a particular user in a static situation where nothing is changing, no new users are coming and no reports are submitted by third users. Each submitted report is shown as a coloured dot, where the color indicates the validation value of the report. It is apparent that good/bad reports push the score/rank up/down. Figure 5 (bottom) shows the evolution of the score in a realistic situation to show the effect of third users' actions or new comers to MA. Note the double effect of the overall dynamics, inducing soft fluctuations in the score but really significant changes in the rank. The stronger dynamics of the rank makes it much more effective for gamification purposes.

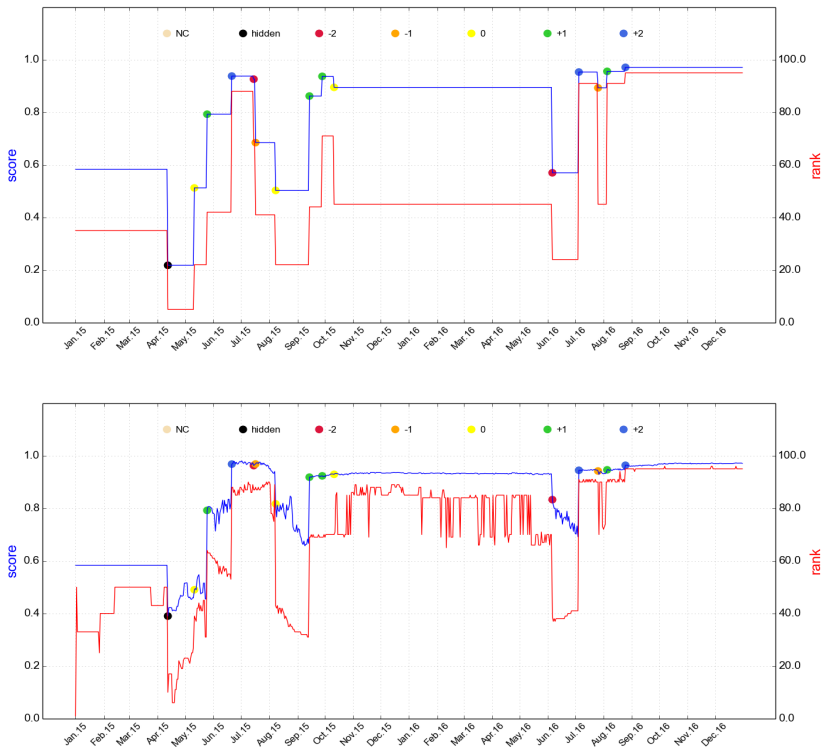

Fig. 5. Score dynamics: (top) simulating a static situation in which the rest of participants do not perform any action; (bottom) real situation with new reports submitted by third participants and new participants joining the Mosquito Alert research program.

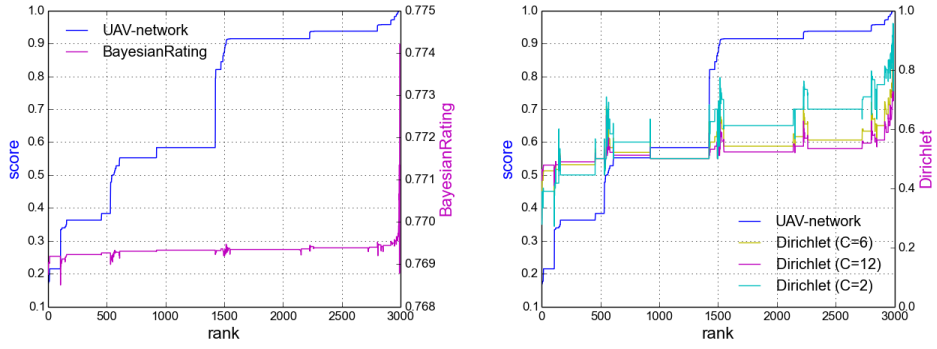

Fig. 6. Comparison of scores: (left) UAV-network *vs.* Bayesian Rating; (right) UAV-network *vs.* Dirichlet reputation system with different values of C.

We also show a comparison of our scores with Bayesian Rating and Dirichlet reputation scores (Figure 6). The scores used in the comparison have been computed taking into account only the reports of type *adult*. In this way we avoid to analyze second order effects due to weight-averaging of *adult* and *bSite* reports, given that in BR or DR these must be independently computed and combined later on. In the x-axis, participants are ranked by our score (blue line). BR scores (Figure 6, left, magenta line) are clearly affected by the weight of the overall average rating (note the scale of the right y-axis). However, BR still yields the same plateaus and we only observe slight ranking changes at the borders of the plateaus. These changes are due to the differences in the leverage of the rate values (*i.e.* the values $P(U|A,V)$ in Tables 7 *versus* the rating levels $r = \{1,\ldots,k\}$ used in Equation 1). DR scores (Figure 6, right) are computed for different values of the C constant. It is clear that C is playing the role of the overall average factor in BR, but DR gives us some control over it. The most important plateaus are also found, but the differences at the borders of the plateaus are more significant. Notably, there is a great difference in the sensibility of our model in comparison to BR and DR. In this context, sensibility represents a better responsiveness of the scoring in relation to participants actions, which we consider it to be a good property to improve participants' engagement in CS research programs.

5 Discussion

The model we propose is similar to a naïve-bayes classifier where the number of feature nodes varies with the number of actions performed (*i.e.* reports submitted) by the user. Starting from a uniform user-class distribution, each validated report contributes with new evidence to refine the profiling of the user.

The key issue of our approach is to estimate a user-class prior that suffices for our scoring purpose. We suggest to select a set of proxy variables of the user-class, with a clear semantics in terms of user expertise, to make a guess about this prior. Nonetheless, any alternative to compute the prior can be considered and applied as well. Anyway, it is crucial to make a guess of the prior that leads to a well balanced (as much as possible) prior and to well behaved (as much as possible) posteriors (*i.e.* good ratings favouring higher user classes and bad ratings favouring lower user classes). If the probability mass distribution of the posteriors is not in clear correlation with the user classes the behaviour of the algorithm can become non-monotonic with respect to increasing evidence about a certain class. Thus, this step must be carefully considered.

In the same way that it is not good to score new users excesively low, it is also not good to score them excessively high. The reason to initialize the score with a unifom instead of the prior user-class distribution is that the later will usually be unbalanced, in our case, clearly unbalanced towards the high expertise classes (Table 4), and consequently users with no validated ativity would be ranked either excesively high or excessively low. Using the prior to initialize the score, not-active users get a score of 0.74011666 (*i.e.* $P(U|S) = P(U)$ in Equation 8),

while using a uniform distribution their score is 0.58333333 (arround 0.5) and they are positioned by the middle-low part of the rank (rank positions 1069:1278 in Table 8) which is fairly reasonable.

As scores are relative to the performance of the whole community, scores are quite dynamic. As participants increase their expertise, all good scores are globally pushed higher. Nevertheless, it is the rank what is ultimately notified to the participants, thus along a period of no activity, a participant might be downgraded with time. Furthermore, in periods of no activity the score can indeed increase if better positioned participants suddenly start sending reports of low quality. These unexpected dynamics could easily generate some confusion or disappointment among the participants. We avoid this situation by giving the basic hints of our scoring system in the project's web page [5] where, indeed, we promote the gamification side of these features in order to use them in our favour. Alternatively, unexpected dynamics as described above could be controlled by implementing an age weighted rating as proposed in [9]. In our case, this solution should be implemented with special care because of the seasonality of mosquito population and, consequently, minimal report activity during winter and spring. This long periods of minimal activity would uniform all scores and many experienced participants might feel disappointed. At the moment, our decision is to keep participants' scores from one season to another.

With respect to BR and DR, while essentially capturing the same concept of rating-based reputation, our model shows a much higher sensibility to the observed evidence, and a good balance of both, evidence of quality (the rates themselves) and quantity of evidence (the number of ratings). The reason lies in the way that evidence is cumulated, *i.e.* by multiplication (Equation 8) instead of by addition as in BR (Equation 1) and in DR (Equation 2). A larger sensibility results in a stronger responsiveness to specific participant actions. Augmenting engagement dynamics with more sensible reputation systems may probably bind better the participants to the long term goals of CS research programs. Furthermore, our model decouples action validation from participant scoring by means of an integrative and unified treatment of any action under consideration, independently of the rating system used for each type of action.

By summer 2017, MA is going to collect extra data from participants with a recently added tool designed to reinforce citizen participation in the research program, whilst easing the experts' validation task. This new tool, natively incorporated to the app, allows citizens to validate mosquito and breeding site images sent by third users and challenge their expertise in identifying mosquito species. This new action will provide valuable information in terms of participants' expertise, based on a binary rating system, *i.e.* right/wrong. Given the structure of our scoring model, this information can be readily translated into a new user-class posterior distribution and incorporated to the scoring algorithm.

[5] http://www.mosquitoalert.com/en/project/send-data/

6 Conclusion

We propose a novel reputation system based on a Bayesian network representing the characteristic flow typically present in CS research programs where participants are expected to perform actions that are validated later on (*i.e.* user, action, validation), what we call the UAV network. In this network, the users node represents an aggregation of participants into expertise classes. The key issue of our approach is to estimate a prior for the user-class distribution that suffices for our scoring pourpose. We suggest to select a set of proxy variables of the user-class, with a clear semantics in terms of user expertise, to make a guess about this prior. However, any other means to get a valid estimate of the prior can readily be used. With respect to Bayesian rating and the Dirichlet reputation models, our approach presents some advantages: (i) is more responsive to the observed evidence, and thus, it bridges better participants with their actions, (ii) it decouples action rating from user scoring, providing a unified processing of any action under consideration, no matter the number of rating levels defined for each, and (iii) it yields a better balance of both, evidence of quality (the rates themselves) and quantity of evidence (the number of ratings). As a proof of concept this model is implemented as part of the Mosquito Alert CS research program.

7 Acknowledgments

We would like to thank the Mosquito Alert team for continuous effort and support and the Mosquito Alert community for its unvaluable cooperation. This work is part of Mosquito Alert CS program research funded by the Spanish Ministry of Economy and Competitiveness (MINECO, Plan Estatal I+D+I CGL2013-43139-R) and la Caixa Banking Foundation. Mosquito Alert is currently promoted by la Caixa Banking Foundation.

References

1. N. Eyal and R. Hoover. *Hooked: How to Build Habit-Forming Products.* Portfolio Penguin, 2014.
2. Randy Farmer and Bryce Glass. *Building Web Reputation Systems.* Yahoo! Press, USA, 1st edition, 2010.
3. BJ Fogg. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pages 40:1–40:7, New York, NY, USA, 2009. ACM.
4. Eric Friedman, Paul Resnick, and Rahul Sami. Manipulation-resistant reputation systems. In Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay Vazirani, editors, *Algorithmic Game Theory*, pages 677–698. Cambridge University Press, 2007.
5. Roy H.E., Preston M.J.O, C.D., Roy D.B., Savage J., Tweddle J.C., and Robinson L.D. *Understanding Citizen Science & Environtmental Monitoring. Final Report on behalf of UK-EOF.* NERC Centre for Ecology & Hidrology and Natural History Museum, November 2012.

6. Ferry Hendrikx, Kris Bubendorfer, and Ryan Chard. Reputation systems. *J. Parallel Distrib. Comput.*, 75(C):184–197, jan 2015.
7. A. Irwin. *Citizen Science: A Study of People, Expertise and Sustainable Development*. Routledge, 1995.
8. Audun Jøsang. Trust and reputation systems. In Alessandro Aldini and Roberto Gorrieri, editors, *Foundations of Security Analysis and Design IV: FOSAD 2006/2007 Tutorial Lectures*, pages 209–245, Berlin, Heidelberg, 2007. Springer-Verlag.
9. Audun Josang and Jochen Haller. Dirichlet reputation systems. In *Availability, Reliability and Security, 2007. ARES 2007. The Second International Conference on*, pages 112–119. IEEE, 2007.
10. Audun Jøsang and Roslan Ismail. The beta reputation system. *BLED 2002 Proceedings*, page 41, 2002.
11. L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation for e-businesses. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, volume 7 of *HICSS '02*, pages 188–, Washington, DC, USA, 2002. IEEE Computer Society.
12. Colin Robertson. Whitepaper on citizen science for environmental research. 2015.
13. Antonio Rodriguez, Frederic Bartumeus, and Ricard Gavaldà. Machine learning assists the classification of reports by citizens on disease-carrying mosquitoes. In Ricard Gavaldà, Indre Zliobaite, and João Gama, editors, *Proceedings of the First Workshop on Data Science for Social Good co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Dicovery in Databases, SoGood@ECML-PKDD 2016, Riva del Garda, Italy, September 19, 2016*, volume 1831 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.
14. Jordi Sabater and Carles Sierra. Review on computational trust and reputation models. *Artif. Intell. Rev.*, 24(1):33–60, September 2005.
15. J. Silvertown, M. Harvey, R. Greenwood, M. Dodd, J. Rosewell, T. Rebelo, J. Ansine, and K. McConway. Crowdsourcing the identification of organisms: A case-study of ispot. *ZooKeys*, 480:125–146, 2015.