Statistical model of the association between collaboration dynamics and project performance

Marc Santolini, Camille Masselot, Rathin Jeyaram

Crowd4SDG

Crowd4SDG
Citizen Science for the Sustainable Development Goals

Deliverable 4.7

Report. Statistical model of the association between collaboration dynamics and project performance

Deliverable identifier: D4.7
Due date: 30/04/2023
Justification for delay: due to exceptional activity as agreed with Project Officer
Document release date: 14/06/2023
Nature: Report
Dissemination Level: Public
Work Package: 4
Lead Beneficiary: UP
Contributing Beneficiaries: UNIGE
Document status: Final

Abstract:

This report investigates the association between collaboration dynamics and project performance in the context of Crowd4SDG consortium's GEAR cycles 2 and 3, using self-reported data collected through the CoSo platform and digital traces from Slack. The analysis reveals the importance of team engagement, diverse compositions, activity span, and effective collaboration strategies in determining project outcomes. The findings also indicate that compositional and structural aspects at the Evaluate phase can serve as early predictors of teams' eventual performance. Based on these insights, the report recommends fostering robust team engagement, assembling diverse teams, and implementing efficient collaboration strategies to enhance the success of future GEAR cycles or similar programs.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 872944.

For more information on Crowd4SDG, please check: http://www.crowd4sdg.eu/

Document history

	Name	Partner	Date
Authored by	Marc Santolini, Camille Masselot, Rathin Jeyaram	UP	April 2023
Revised by	Oguz Mulayim Elena Proden Anais ZODEOUGAN-QUIST	CSIC UNITAR UNITAR	18/04/2023
Edited by			
Reviewed by			
Approved by			

Table of Contents

Project Partners

	Partner name	Acronym	Country
1	Université de Genève	UNIGE	CH
2	European Organization for Nuclear Research	CERN	CH
3	Agencia Estatal Consejo Superior de Investigaciones Científicas	CSIC	ES
4	Politecnico di Milano	POLIMI	IT
5	United Nations Institute for Training and Research	UNITAR	CH
6	Université de Paris	UP	FR

Crowd4SDG in brief

The 17 Sustainable Development Goals (SDGs), launched by the UN in 2015, are underpinned by 169 concrete targets and 231 unique measurable indicators. Some of these indicators initially had no established measurement methodology. For others, many countries do not have the data collection capacity. Measuring progress towards the SDGs is thus a challenge for most national statistical offices.

The goal of the Crowd4SDG project is to research the extent to which Citizen Science (CS) can provide an essential source of non-traditional data for tracking progress towards the SDGs, as well as the ability of CS to generate social innovations that enable such progress. Based on shared expertise in crowdsourcing for disaster response, the transdisciplinary Crowd4SDG consortium of six partners is focusing on SDG 13, Climate Action, to explore new ways of applying CS for monitoring the impacts of extreme climate events and strengthening the resilience of communities to climate related disasters.

To achieve this goal, Crowd4SDG is initiating research on the applications of artificial intelligence and machine learning to enhance CS and explore the use of social media and other non-traditional data sources for more effective monitoring of SDGs by citizens. Crowd4SDG is using direct channels through consortium partner UNITAR to provide National Statistical Offices (NSOs) with recommendations on best practices for generating and exploiting CS data for tracking the SDGs.

To this end, Crowd4SDG rigorously assesses the quality of the scientific knowledge and usefulness of practical innovations occurring when teams develop new CS projects focusing on climate action. This occurs through three annual challenge based innovation events, involving online and in-person coaching. A wide range of stakeholders, from the UN, governments, the private sector, NGOs, academia, innovation incubators and maker spaces are involved in advising the project and exploiting the scientific knowledge and technical innovations that it generates.

Crowd4SDG has six work packages. Besides Project Management (UNIGE) and Dissemination & Outreach (CERN), the project features work packages on: Enhancing CS Tools (CSIC, POLIMI) with AI and social media analysis features, to improve data quality and deliberation processes in CS; New Metrics for CS (UP), to track and improve innovation in CS project coaching events; Impact Assessment of CS (UNITAR) with a focus on the requirements of NSOs as end-users of CS data for SDG monitoring. At the core of the project

is Project Deployment (UNIGE) based on a novel innovation cycle called GEAR (Gather, Evaluate, Accelerate, Refine), which runs once a year.

The GEAR cycles involve online selection and coaching of citizen-generated ideas for climate action, using the UNIGE Open Seventeen Challenge (O17). The most promising projects are accelerated during a two-week in-person Challenge-Based Innovation (CBI) course. Top projects receive further support at annual SDG conferences hosted at partner sites. GEAR cycles focus on specific aspects of Climate Action connected with other SDGs like Gender Equality.

Grant Agreement description of the deliverable

The focus of Work Package 4 (WP4), led by the University of Paris, is to conduct research on Citizen Science. This encompasses the establishment of methods and the collection of data to inform the development of effective, high-quality citizen science projects. To that aim, this work package develops metrics and statistical models in order to assess the many-faceted outcomes of the citizen science projects developed within the Crowd4SDG consortium.

In tasks 4.2 and 4.3, we quantitatively monitored and analysed the activity of teams working within the GEAR cycle framework, and the activity and engagement patterns of citizen science participants, by leveraging the digital traces from online tools that document project progress and citizen engagement and the self-reported data collected via CoSo.

This deliverable presents the results of Task 4.4 detailed below:

T4.4: Build a predictive model of project quality from the collected multi-scale data (UPD, UNIGE, CSIC)

The tools and measurements described in Task 4.2 and Task 4.3 will be applied to the CS projects running within Crowd4SDG to provide a basis from which to **predict performance and impact of the various projects.** Using the collected data, we will investigate various organizational features of the teams. The previously defined measures of Team Energy (number and frequency of interactions), Team Engagement (the degree to which people close conversation loops, that can be computed using network clustering), and Team Exploration (going outside the core group for additional interactions and information) will be calculated in this context. Moreover, the existence and importance of leadership will be explored by looking at node centrality and triadic closure around that node. Beyond leadership, the existence of a core group can be revealed using information theoretical

metrics based on the non-uniformity of activity patterns. Since the teams can be relatively large, modularity may be an important success factor. Such modularity can be exhibited in the context of temporal interaction network using Non-Negative Tensor Factorization. The detected sub-teams may, for example, be specialized in certain sub-tasks, and the integrity of the team as a whole would then rely on them being properly connected through broker nodes that can be detected using Burt's constraint. We will also analyse the working dynamics in terms of burstiness by departure from a Poissonian activity pattern, revealing the importance of a teams' internal deadlines and the "sprints" that precede them. Finally, by collecting personal attributes, we will explore the role of diversity of skills/age/gender in sustained interactions (Energy) and team end performance. By reflecting the diversity and adaptivity of insights as well as the learning experience of the participants and overall societal impact, we believe such metrics can provide a refined and much needed picture of the complexity of collective knowledge production in the 21st century.

1. Introduction

WP4 focuses on creating and monitoring new metrics and statistical models of team engagement and collaboration, which contribute to the diverse outcomes of citizen science projects within the Crowd4SDG consortium over its 3-year duration. WP4 has two primary objectives: 1) Develop standardised metrics and descriptors for assessing the diversity, originality, effectiveness, sustainability/robustness, and adaptation/appropriateness of solutions and insights obtained from citizen science projects; and 2) Implement these metrics and descriptors as tools for analysing digital records of citizen science collaborations and their generated solutions and insights. As a result, WP4 supports Crowd4SDG's specific objectives of enhancing citizen science skills, producing high-quality scientific outcomes, and generating economic and social outputs relevant to achieving SDGs through challenge-based citizen science events, particularly focusing on climate change resilience.

In this report, we present a **statistical modelling framework for identifying predictors of performance and impact metrics for citizen science projects**. Prior research has identified key characteristics of high-performing teams (Pentland 2012), such as Team Energy (interaction quantity and frequency), Team Engagement (closing conversation loops, assessed using network clustering), and Team Exploration (seeking external interactions and information). By analysing digital traces from the Slack workspace, demographic data, and self-report surveys collected in GEAR 2 (D4.4) and GEAR 3 (D4.5), we extract various team

organisational features related to these characteristics. We then utilise social network analysis to investigate centrality measures in communication processes and informal advice networks. Ultimately, we evaluate their association with the success and quality of citizen science projects using regression analyses on the performance metrics defined in our initial report on the epistemology of citizen science in D4.1.

2. Methods

In this section, we first provide an overview of the data collected in GEAR 2 and GEAR 3, which were previously introduced in D4.4 and D4.5. We then dive into the methods used for the statistical modelling developed in this report.

2.1 Communication data

A Slack workspace was used by the teams during the GEAR cycle as a means to communicate with other teams and with the organising team. The data was extracted in JSON format using the export function available to the owners/admins of the Slack workspace. This allowed us to gather, across all public channels, a data frame containing the messages (post contents) and information on each message's timestamp, sender, and target channel. The raw data was then processed to obtain mentions. A mention occurs when a Slack user types in a message the Slack username of a target user prefixed by "@". Each recorded mention has information on the source (who wrote the message), target (who is being mentioned) and the timestamp (when the message was sent). Slack also allows users to broadcast messages by citing all users in a channel or a workspace by using specific commands (@all, @here, @channel_name). The messages containing these built-in commands were not included as mentions in order to focus on direct interactions only.

Using the available Slack data, we employed the number of posts and number of reactions of a user as a marker of individual engagement, or team engagement when aggregated over team members. Furthermore, for each GEAR cycle we built social interaction networks where a user is linked to another user if he/she mentions him/her, with a weight corresponding to the number of mentions. When aggregating at the team level, intra-team mentions are encoded as self-loops, and the weights of the intra-team links are summed to create a final team-level network on which to compute centralities such as weighted degree. This allows to represent the flow of information characterising this phase, in particular highlighting the interactions with the organisation team.

2.2 CoSo self-reported interaction data

During GEAR cycles, we conducted two types of surveys: those related to participant attributes (e.g. their background, country of origin, etc), and those related to participant interactions (e.g. who they collaborated with, sought advice from, etc).

The initial survey was related to attributes only and was disseminated using a Google Form at registration to the Evaluate phase. We then disseminated 4 weekly surveys related to social interactions and activities using the CoSo platform (Tackx et al., 2021). The CoSo platform is designed to collect self-reported interaction data with a simple, reactive interface, and an analysis-ready database (see D4.6). To document their interactions, the users select target users across all other participants and organisers. The interactions span prior ties in the first survey ("Which of these people did you know personally before?"), and on a weekly basis their advice seeking interactions ("Who did you seek advice from last week?") and work collaborations ("Who did you work with last week?"). To document their activity, they could also select across 26 activities encompassing routine activities within research teams inspired from the CRediT contribution taxonomy[1], as well as specific questions regarding Crowd4SDG, for example specific tool usage.

The surveys were advertised through Slack and the organising team dedicated 10 minutes for participants to fill them during weekly sessions, ensuring a high engagement (see D4.5 p12).

CoSo networks were directly inferred from the surveys. For each GEAR, we aggregated the networks over all time points collected, yielding weighted interaction networks where edge weights correspond to the number of times an interaction was reported. When considering team-level network centrality measures, that is, measures that indicate how strategic the position of the team is in the network of interactions, we further aggregated the individual networks at the team level. Network centrality measures were computed using the `igraph` library in R (Csardi 2006).

2.3 Team characteristics

The ability of teams to develop their project depends on compositional features such as who is in the team, as well as how the team operates, such as their collaboration activity and division of labour. Here we used the digital traces and survey data to derive and monitor

[1] https://credit.niso.org/

features related to team **composition**, **communication**, **collaboration,** and **activity** which we detail below.

For team composition, we built measures of size, diversity, education level, and prior experience with SDGs. **Team size** was assessed using the number of members of a team. **Background diversity** was assessed by computing the background span, that is the number of unique academic backgrounds in the team as declared in the registration form. The **education level** was computed by taking the average level of education in a team based on the response to the question "What is your current or highest level of education" to which we attributed the following score based on the answer: 0 for secondary school, 1 for high school, 2 for undergraduate and 3 for graduate. Finally, **prior experience with SDGs** was computed as the average answer to the question "Have you participated in data projects or contributed as citizen scientist to data production before?" (yes = 1 and no = 0) within each team.

For communication, we leveraged the activity and interactions on Slack public channels. The **Slack activity** was assessed as the total number of messages posted by team members. For interactivity, we measure **Slack interaction intra-team** as the number of mentions among members of a team, and **Slack interaction organising team** as the number of mentions between members of a team with the organising team. We counted mentions regardless of their directionality.

For collaborations, we focused on the amount of collaborations within the teams, as well as the centrality of the teams within the advice network. For the **intra-team collaborations** (CoSo interaction intra-team), we summed for each team the weights of the intra-team edges in the “work with” collaboration network. For the centrality in the advice network, we computed the **Burt constraint** (Burt, 2004), a measure of social capital that takes low values when a neighbourhood is diverse (existence of links with separate neighbourhoods), and higher values when the neighbourhood is constrained (dense connections to the same neighbourhood). Burt constraint is essentially a measure of the extent to which a team *v* is invested in teams who are invested in other of *v*'s alters (neighbours). The "constraint" is characterised by a lack of primary holes around each neighbour, i.e. the neighbourhood has a large number of triadic closures. Burt's constraint is higher if ego has less, or mutually stronger related (i.e. more redundant) contacts. **Network diversity** was computed by taking the negative of the Burt constraint, with higher values indicating higher levels of diversity (more structural holes). This quantifies the ability of a team to leverage diverse sources of information for advice seeking or collaborations.

Finally, for the activity, we focused on measures of diversity and engagement of activities performed. For diversity, we computed the **activity span** as the proportion of activities performed by a team among the 26 listed activities. For engagement, we considered the **activity regularity** by first computing the number of activities reported by a team each week, and then computing the negative of the Gini coefficient[2] on the resulting vector. The Gini index ranges from 0 (perfectly regular) to 1 (perfectly irregular). The (1 - Gini) value is higher if activities are regularly conducted across weeks. Finally, we quantified for each team the **survey engagement** as the proportion of survey responses per team across all CoSo surveys, a measure of engagement to the study.

2.4 Team performance data

To quantify team performance, we used the scores that teams obtained in their assessment by the jury and the Crowd4SDG organising team, which were co-constructed using the results from D4.1 (pp 32-33).

At the end of each phase, experts composing a jury scored each team from 0 to 5 on the following criteria. We indicate the weight of each score between squared brackets. The sum of these scores constitute the final jury score, with a maximum value of 50.

- Novelty: Is the pitch based on a new idea or concept or using existing concepts in a new context? [10]
- Relevance: Is the solution proposed relevant to the challenge or potentially impactful? [10]
- Feasibility: Is the project implementable with reasonable time and effort from the team? [10]
- Crowdsourcing: Is there an effective crowdsourcing component? [10]
- Overall: How would you rate this team's overall presentation skills during this pitch? [10]

[2] The Gini coefficient measures the inequality among values of a frequency distribution, such as levels of income. A Gini coefficient of 0 reflects perfect equality, where all income or wealth values are the same, while a Gini coefficient of 1 (or 100%) reflects maximal inequality among values. For example, if everyone has the same income, the Gini coefficient will be 0. In contrast, a Gini coefficient of 1 indicates that within a group of people, a single individual has all the income or consumption, while all others have none.

Between the Evaluate and Accelerate phases, additional criteria presented below were used by the organisation team. We indicate the weight of each score between squared brackets, summing to a maximum possible jury score of 40.

- Appropriateness of Methodology [5] (only for GEAR 2)
- Weekly Evaluation [10]
- Use of Toolkit [5]
- Data Collection and NSO [5]
- Commitment [5] (only for GEAR 2)
- Attendance [5]
- Deliverables [5]

The final score accounted for 60% of the jury score and 40% of the organisation team score:

Final Score = jury score*(60/50) + organisation team score.

More precisely, crowdsourcing was assessed using the mean score attributed by judges to the question “Is there an effective crowdsourcing component?” (yes = 1 and no = 0). We measured the feasibility, relevance, and novelty by computing the mean score attributed by the jury on a scale from 0 to 5 to the questions “Feasibility: Is the project implementable with reasonable time and effort from the team?”, “Novelty: Is the pitch based on a new idea or concept or using existing concepts in a new context?”, and “Relevance: Is the solution proposed relevant to the challenge or potentially impactful?”.

All variables were integer values with scores ranging from 0 to 5 for deliverables and attendance, 0 and 1 for commitment. For weekly evaluation, the score was a continuous value ranging from 0 to 10 scoring the overall quality of their weekly pitch sessions. Deliverable score was measured by the total number of deliverables submitted and documented on the platform Innprogress (https://innprogresstest.unige.ch/) among the expected ones. Attendance was estimated by the proportion of sessions attended by team members. Commitment was scored 1 if teams were willing to continue their project after the end of theEvaluate phase, or 0 otherwise.

2.5 LASSO regression and statistical model

Statistical and network analyses were conducted using the R software. We leveraged libraries `glmnet` (Friedman 2010), `MASS` (Venables 2002) and `jtools` (Long 2022) for statistical modelling, and `igraph` for network centralities.

Associations between team characteristics and performance measures were done as follows:

First, since the data originated from two different GEAR cycles, we considered the possible variation in overall values of both team features and performance by normalising the data. To do so, the features were centred (mean of 0) and rescaled (variance of 1) within each GEAR cycle using the `scale()` function in R, and concatenated into an overall dataframe.

Then, each performance variable was defined as a dependent (outcome) variable, and the data frame of team features was used as independent variables. Missing data was handled by imputation using the means of the nonmissing values (by the `makeX()` function of the `glm` package). We then conducted a LASSO regression (Tibshirani 1996) in order to eliminate team features that are not statistically contributing to the outcome, and select only the relevant features. For this, we first conducted a 10-fold cross-validation to find the optimal penalty value that minimises the Mean Squared Error of the regression to the outcome (by the `cv.glmnet()` function of the `glm` package). A final model was run for this optimal penalty value on the whole dataset to derive regression coefficients for all team features. Any feature with a coefficient equal to 0 was then discarded. A standard regression (by the `lm()` function) was then run using the remaining features to obtain standardised regression coefficients, 95% confidence intervals and p-values. Features with p-values less or equal to 10% were finally kept for the final figures shown in this report.

Overall, we considered for each outcome the features that i) are selected during a cross-validation step of the LASSO regression and ii) have less than 10% chance to be contributing to the outcome in a randomised setting. This stringent selection process ensures a significant reduction of the noise in the estimator considering the relatively small (N=26) number of data points.

2.6 Pseudo-anonymization and ethical approval

The data collection tools and research questions received the ethical approval of the Inserm committee attached to the University of Paris team (IRB00003888), in charge of collecting

the data. Participants gave their consent to the collection of data as they registered to the Evaluate phase (see D4.5). Data was pseudo-anonymized by our team before the analysis.

3. Results

We report the results of the statistical modelling of the association between collaboration dynamics and project performance. Because of the low number of data points (N=26 teams), we leverage a stringent analysis in order to i) combine both GEAR 2 and GEAR 3 (batch correction) and ii) select relevant features (LASSO regression) for regression analysis (see Methods). We consider two main outcomes: the team performance at the Evaluate phase, and the advancement in the GEAR cycle. The former is directly related to the team characteristics measured at the Evaluate phase, while the latter interrogates whether early monitoring at the Evaluate phase informs on the ultimate stage achieved in the GEAR cycle (the Accelerate or Refine stage). In addition, we explore several fine-grain performance measures that are aggregated to compute the Evaluate performance, such as the novelty, relevance, or feasibility of the projects.

3.1. Performance at Evaluate phase

We first focus on the performance at the Evaluate phase, which accounts for 60% of the jury score and 40% of the organisation team score (see Methods). Results of the LASSO feature selection and linear regression method are shown in Figure 1. Features are ordered by decreasing significance (i.e. higher p-values), with all features having $p<0.1$.

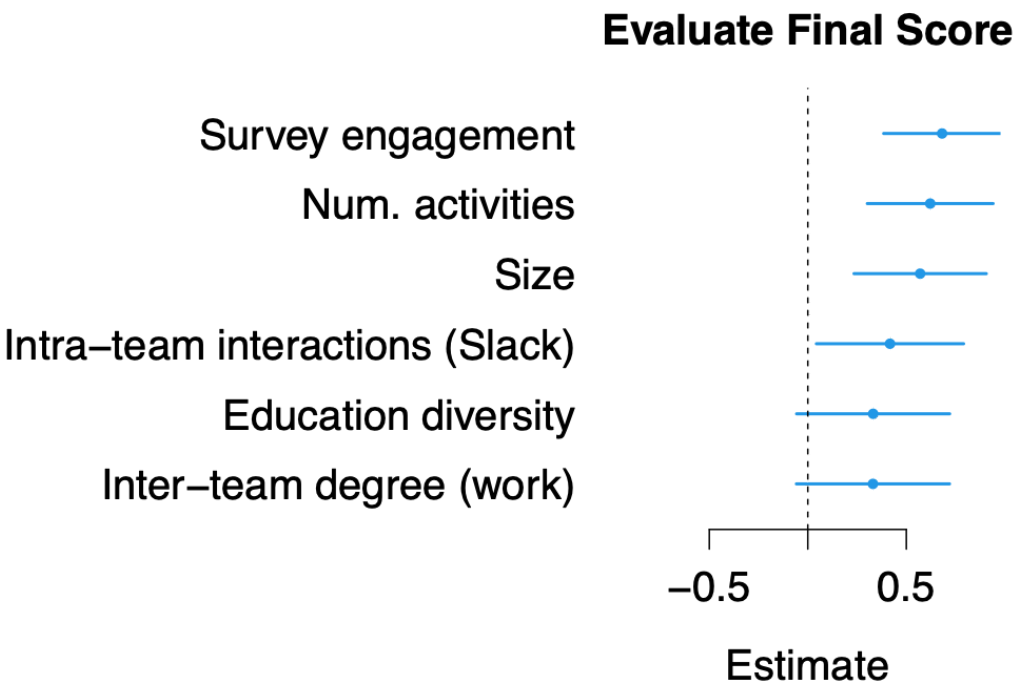

Figure 1: Standardised regression coefficients for the team characteristics associated with the Evaluate final score, selected through the LASSO regression (see Methods). Error bars denote 95% confidence intervals. Positive estimates denote a positive association between the feature and the outcome.

Firstly, our analysis reveals that a team's engagement in the CoSo survey (mean answers per week) is the most significant predictor. This finding suggests that, beyond its data collection function, the self-report activity serves as an indicator of the team's dedication to participating in the program, and that these efforts impact the quality of their project (jury score) and the engagement perceived by the organising team. This is supported by the subsequent feature, the total number of activities performed during the phase, which is positively linked to performance. These two measures demonstrate that engagement in Evaluate activities influences performance at the end of the phase.

We also discover that team composition plays a role in performance, with a positive correlation between the number of team members (size) and the diversity of education levels within the team (education Shannon index[3]). This implies that larger, more diverse teams have a performance advantage.

[3] The Shannon diversity index is a measure of the entropy of a distribution, in that case the number of team members at a given education level. Higher values mean a flatter distribution, that is a more diverse set of education levels in the team. The mathematical definition can be found at https://www.statology.org/shannon-diversity-index/.

Additionally, our findings show that a team's position within the interaction network is crucial. Teams that collaborate with a higher number of teams (degree of inter-team collaboration), and have members who communicate more frequently (intra-team Slack interactions) perform better.

In summary, these results indicate that both composition and structural features are important in determining the outcome at the Evaluate phase. However, these are aggregated outcomes, and we will now shift our focus to specific fine-grained outcomes to delve deeper into which features are crucial for their success.

3.2. Aspects of project quality

In the Crowd4SDG project, teams have to design and pitch early-stage citizen science projects. As such, these projects must hold certain properties: they have to be relevant for the topic of the GEAR cycle, feasible, innovative, and involve a crowdsourcing component. We used the fine-grained data from the jury scores to compute relevant performance variables and explore team features that underlie them. Results are shown in Figure 2.

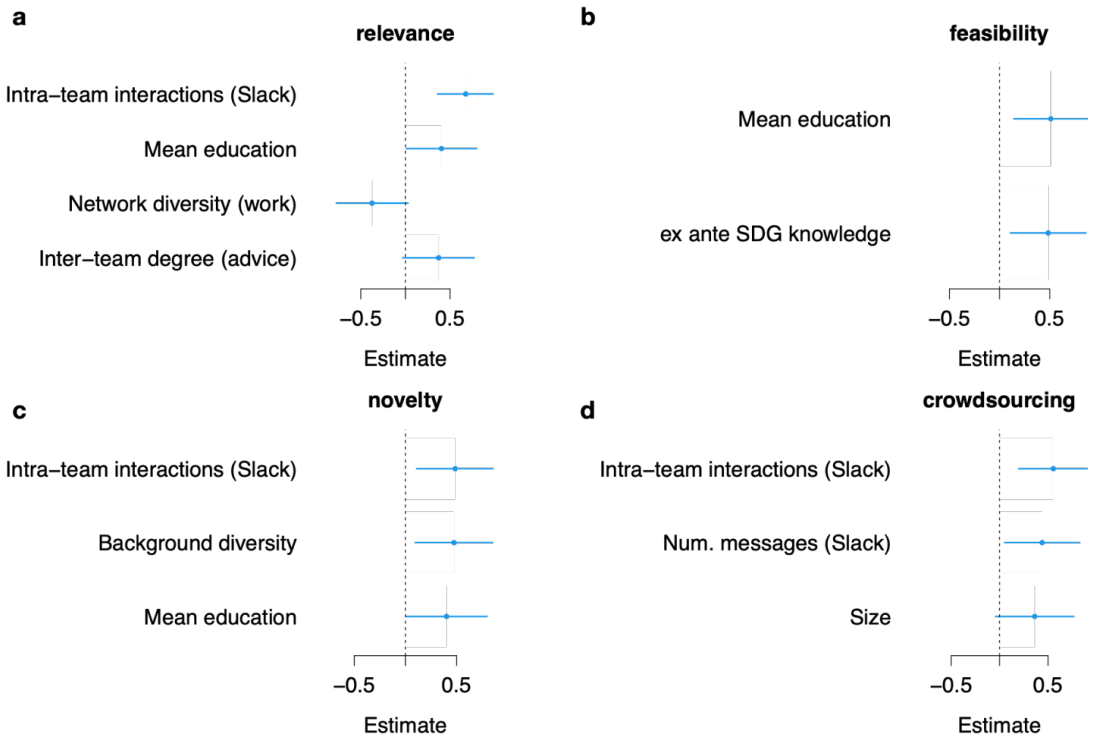

Figure 2: Same as Figure 1, for the outcomes shown in bold.

First, we find that internal team communication, as measured by Slack activity (Intra-team interactions and Number of messages), plays a crucial role in project relevance, novelty, and the incorporation of a crowdsourcing component. This suggests that teams use the Slack workspace for brainstorming and refining their projects, making it a central hub for team processes. As in-person meetings were not possible due to the Covid pandemic, teams effectively leveraged Slack for their interactions and collaboration.

Various aspects of team composition were found to be important. First, higher education levels (mean education) within teams correlated with more relevant, feasible, and novel projects, emphasising the importance of advanced academic skills for developing innovative yet realistic projects. Second, team size played a role in crafting crowdsourcing components, highlighting the benefits of a larger number of individuals to accomplish this task. Third, the diversity of backgrounds in the team, indicative of interdisciplinarity, was linked to novelty, a finding consistent with scientometrics research showing that interdisciplinarity fosters innovation (Singh 2022). Lastly, the average level of prior experience with SDGs (*ex ante* SDG knowledge) was associated with project feasibility, suggesting that participants draw on their SDG experience (possibly within the Goodwall platform, from which the majority of participants originated) to refine their ideas into viable projects.

Finally, team interactions proved to be crucial for project relevance. Teams that sought advice from a larger number of teams (Inter-team degree) and collaborated within a focused, tight network (low network diversity for "work with") were more likely to achieve high relevance scores. This reflects a balance between seeking advice (gathering information from the network) and exploiting advice (collaborating with a more limited set of actors).

In summary, these findings demonstrate that team composition features (size, education level, and diversity of backgrounds), internal communication (engagement on Slack), and collaboration strategy (advice seeking and work interactions) are associated with distinct aspects of project quality.

3.3. Advancement in the GEAR cycle

Beyond the results from the Evaluate phase, we asked whether the obtained data at the Evaluate phase, which encompasses the largest number of teams (compared with

Accelerate of Refine), could be used as an early predictor of the final stage achieved by teams during the GEAR cycle.

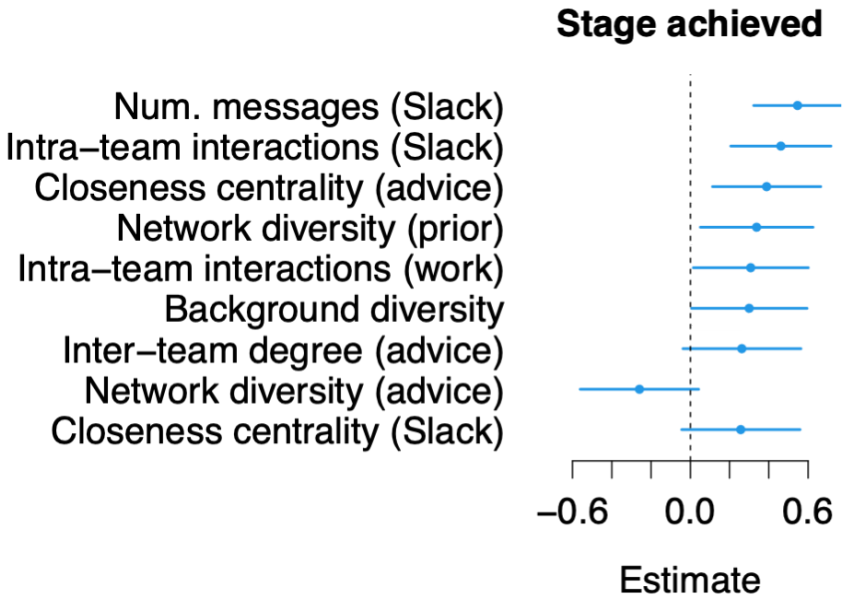

Figure 3: Same as Figure 1, for the final stage achieved in the GEAR cycle.

Figure 3 presents the results of feature selection and regression analysis for the stage achieved. These findings are consistent with previous insights and can be summarised as follows.

First, team activity in the Evaluate phase is associated with the final stage reached in several ways: the number of messages shared on Slack (both within the team and overall) and the number of self-reported work interactions within the team. In essence, hard work plays a significant role in ultimate success.

Second, we find that several diversity measures are associated with success: the diversity of backgrounds (background span), which suggests that team interdisciplinarity is essential to address the global challenges at hand, and network diversity of prior ties, indicating a broader reach within the informal network.

Lastly, advice-seeking behaviour is identified as important on multiple levels. In fact, we find that both local (inter-team degree, a measure of the number of immediate neighbours of a node) and global (closeness centrality, a measure of how close a node is to all other nodes in

the network through shortest paths[4]) centrality in the advice-seeking network are important, while maintaining strong connections with a focused, tightly knit neighbourhood (low network diversity).

In summary, these results demonstrate that compositional and structural aspects during the Evaluate phase serve as early indicators of the teams' eventual performance in the GEAR cycle.

4. Discussion of the results

The GEAR cycle analysis provides valuable insights into the factors that contribute to project performance and advancement in the context of citizen science. Specifically, aspects of team composition, internal communication, and collaboration strategy are crucial determinants of success, highlighting the interplay between individual and collective factors.

Engagement in the CoSo survey and the number of activities performed during the Evaluate phase significantly impact team performance, reflecting the importance of commitment and dedication. This finding echoes the social psychological concept of group cohesion, which is known to positively affect group performance (Carron et al., 1985).

Larger teams with diverse education levels and interdisciplinary backgrounds have a performance advantage, consistent with theories that emphasise the benefits of interdisciplinary collaboration for innovation (Singh, 2022). This resonates with research on the benefits of diverse teams in science, which shows that heterogeneous groups can bring different perspectives and expertise to bear on complex problems (Page, 2007).

Internal communication through Slack proves critical for project relevance, novelty, and crowdsourcing components, demonstrating that digital platforms can facilitate effective collaboration, especially during remote work scenarios. This aligns with prior studies examining the role of digital tools in fostering collaborative research networks (Lazer et al., 2009).

The study found that advanced academic skills, team size, and prior experience with SDGs correlate with distinct aspects of project quality. Teams that sought advice from a larger number of teams and collaborated within a focused network achieved higher relevance scores. This balance between information gathering and effective collaboration aligns with

[4] In a connected graph, closeness centrality (or closeness) of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph.

Granovetter's (1973) "strength of weak ties" theory, which posits that weak ties provide access to novel information, while strong ties foster trust and collaboration. This balance is also consistent with Burt's (2004) structural hole theory, which suggests that individuals who bridge gaps in networks can access diverse information and resources, leading to improved performance and innovation.

In summary, the GEAR cycle analysis offers valuable insights into the interplay between individual and collective factors that contribute to project performance in citizen science initiatives. These findings emphasise the importance of fostering interdisciplinary teams, effective communication, and strategic collaboration, which are supported by existing social theories on networks and collaboration in science. This research provides a foundation for further exploration into the dynamics of collaborative networks in citizen science and the development of strategies to optimise project outcomes.¶

5. Conclusion~~Discussion~~ and perspectives

Work Package 4 aims to develop and monitor new metrics and develop statistical models of team engagement and collaboration that contribute to the many-faceted outcomes of the citizen science projects developed within the Crowd4SDG consortium. In this report, we presented a data-driven approach to develop a statistical model of the association between collaboration dynamics and project performance during the GEAR cycles 2 and 3. We leveraged the CoSo platform for collecting self-reported data on collaborations and task allocation structure of participating teams, allowing us to measure characteristics of team composition, activity and interaction dynamics.

While our initial hope was to follow the in-situ dynamics of teams participating in the program, the unexpected shift to a fully online program led us to shift gears, focusing on digital traces from the team coordination tools, as well as building a smartphone application to facilitate the reporting of collaborative activities. By leveraging a single sign-in system and the ability to run all future surveys in one place using the CoSo app, we were able to collect comprehensive profile information about participants along with temporal insights on their collaborative activities, avoiding extensive manual curation from the organising team and ensuring analysis-ready data. We show how the CoSo allows us to monitor features that are associated with aspects of performance related to the process and not just the final outcome. Moreover, we show that beyond background diversity (important for project

novelty), the diversity of advice-seeking ties (a network diversity metric) matters for the final project performance.

In this report, we demonstrated how different data sources on teamwork, effort, communication and collaborations inform on various measures of performance of their project. Given the relatively small number of teams (N=26), we leveraged a LASSO regression analysis in order to perform feature selection. We then investigated the association between collaboration dynamics and project performance in the context of the GEAR cycles, focusing on team performance at the Evaluate phase and advancement in the GEAR cycle.

Results show that team composition and structural features are equally important in determining the outcome at the Evaluate phase. Key factors include team engagement, activity span, team size, diversity of education levels, and embeddedness in the interaction network. Further analysis of fine-grained outcomes reveals that team composition features (size, education level, and diversity of backgrounds), internal communication (engagement on Slack), and collaboration strategy (advice seeking and work interactions) are associated with different aspects of project quality.

We also examined whether data from the Evaluate phase can serve as an early predictor of the final stage achieved by teams during the GEAR cycle. Findings indicate that compositional and structural aspects at the Evaluate phase are indeed early predictors of the eventual performance of teams. Specifically, team activity in the Evaluate phase, diversity measures, and advice-seeking behaviour were found to be important for final success.

Overall, the study highlights the significance of team engagement, composition, and collaboration strategy for project performance in the GEAR cycle. The self-reported and surveyed data offer an opportunity to operationalise metrics and descriptors underlying the quality and novelty of citizen science projects. Our contribution extends beyond the Crowd4SDG project to the general evaluation of citizen science by informing project leaders, citizen scientists, and decision makers on what can be assessed online to perform high-quality citizen science based on the criteria provided in D4.2 and operationalised in this report.

In light of the findings presented in this report, we put forth the following recommendations to enhance the success of future GEAR cycles or comparable programs. Coordinators should prioritise cultivating robust team engagement, assembling teams with diverse compositions, and implementing efficient collaboration strategies. It is advisable to motivate participants to

actively partake in activities and maintain frequent communication via platforms like Slack, which has proven beneficial for idea generation and project refinement. Forming teams with a diverse mix of education levels, backgrounds, and experiences can foster innovation and improve project quality. Additionally, establishing a collaborative atmosphere in which teams can access advice from an extensive network of peers while sustaining strong connections with a select group of collaborators is essential. By emphasising these aspects, coordinators can contribute to a more favourable environment for achieving successful project outcomes in GEAR cycles or similar initiatives.

6. References

Burt, R.S., 2004. Structural Holes and Good Ideas. Am. J. Sociol. 110, 349–399. https://doi.org/10.1086/421787

Carron, A. V., Brawley, L. R., & Widmeyer, W. N. (1985). The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire. Journal of Sport Psychology, 7(3), 244-266.

Csardi G, Nepusz T (2006). "The igraph software package for complex network research." InterJournal, Complex Systems, 1695. https://igraph.org.

Friedman J, Tibshirani R, Hastie T (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software, 33(1), 1–22. https://doi.org/10.18637/jss.v033.i01.

Granovetter, M. S. (1973). The Strength of Weak Ties. American Journal of Sociology, 78(6), 1360–1380.

Lazer, D., Pentland, A., Adamic, L. A., Aral, S., Barabási, A. L., Brewer, D., ... & Van Alstyne, M. (2009). Life in the network: the coming age of computational social science. Science, 323(5915), 721-723.

Long JA (2022). jtools: Analysis and Presentation of Social Scientific Data. R package version 2.2.0, https://cran.r-project.org/package=jtools.

Page, S. E. (2007). The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press.

Pentland, A. (2012). The New Science of Building Great Teams. Harvard Business Review. https://hbr.org/2012/04/the-new-science-of-building-great-teams

Singh, C. K., Barme, E., Ward, R., Tupikina, L., & Santolini, M. (2022). Quantifying the rise and fall of scientific fields. PLOS ONE, 17(6), e0270131. https://doi.org/10.1371/journal.pone.0270131

Tackx, R., Blondel, L., Santolini, M., 2021. Quantified Us: a group-in-the-loop approach to team network reconstruction, in: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, UbiComp '21. Association for Computing Machinery, New York, NY, USA, pp. 502–507. https://doi.org/10.1145/3460418.3479363

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.