Gargiulo *et al. Applied Network Science* (2022) 7:19
https://doi.org/10.1007/s41109-022-00456-0

Applied Network Science

RESEARCH Open Access

Distribution of labor, productivity and innovation in collaborative science

Floriana Gargiulo[1*], Maria Castaldo[2], Tommaso Venturini[3] and Paolo Frasca[2]

*Correspondence:
floriana.gargiulo@cnrs.fr
[1] CNRS, Université Paris-Sorbonne - Paris IV, GEMASS, Paris, France
Full list of author information is available at the end of the article

Abstract

In this paper, we investigate the process of scientific discovery using an under-exploited source of information: the Polymath projects. Polymath projects are an original attempt to solve a series mathematical problems collectively and in a collaborative online environment. To investigate the Polymath experiment, we analyze all the posts related to the projects that have resulted in a peer-reviewed publication. We focus in particular on the organization of the scientific labor and on the innovations that result from the contributions of the different authors. We find that a high presence of occasional contributors increases the productivity of the most active users and the overall productivity of the forums (i.e., the number of posts grows super-linearly with the number of contributors). We argue that, in large-scale collaborations, the serendipitous interaction between occasional contributors can be crucial to the scientific process, and individual contributions from occasional participants can open new directions of research.

Keywords: Polymath, Collaborative science, Innovation

Introduction

With the advent of Internet technologies, we have witnessed the emergence of large-scale collaborative platforms that enable the creation of open-source content through the voluntary participation of large numbers of users. These platforms have fundamentally changed the practices of knowledge production and consumption, turning into public goods a number of cultural and scientific resources. The best-known example of these collective cultural platforms is the online encyclopedia Wikipedia, which has become one of the most important and trusted sources of information worldwide (Anthony et al. 2009; Fallis 2008). Other examples of online collaboration can be found in software production, from the Linux movement, to GitHub, to Q&A systems, like Stack Overflow, and in scientific task-sharing projects, like Galaxy Zoo.

Platforms such as the ones mentioned above foster interactions and allow a multiplicity of participants with different expertise to come together without a rigid and hierarchical role structure and collectively performed together simple and complex tasks. These collaborative systems have proved that spontaneous interactions between users can produce a shared knowledge that goes beyond the mere aggregation of individual

contributions, a process referred to as "collective intelligence" (Straub et al. 2021; Grasso and Convertino 2012).

Despite the success of these platforms and even though academic institutions have long insisted on the idea of open and participatory science, there are few actual examples of large-scale collective production in science.

The conceptual framework of online collaborative structures raises several important questions when applied to scientific production: Is science the craft of many or of few? Can research be conducted in a large-scale open collaborative environment? Can science be based on collaboration rather than competition? These questions are discussed in the book "Reinventing discovery" by Nielsen (2011), which presents several examples of collective problem solving. Among the cases described by Nielsen, one—the Polymath project—has attracted our attention, because it can be studied not only ethnographically but also computationally, since all its contributions are available as digital records.

The first Polymath project was proposed in 2009 by mathematician Tim Gowers, who, with a post on his blog, invited mathematicians to find a combinatorial proof of the density version of the Hales-Jewett Theorem, using a dedicated thread of discussion. Since then, fifteen other Polymath projects have been launched, six of which have resulted in one or more peer-reviewed publications signed with the collective name "Polymath Collaboration". The Polymath blogs not only enable the study of an important project in collaborative science, but also provide an unprecedented playground for the in-depth study of discovery processes.

In our work, we present a comprehensive statistical analysis of the Polymath ecosystem, looking specifically at the activities of participants and the content that they produce. Our findings are based on all projects that have resulted in a peer-reviewed publication (projects 1,4,5,8,15). We did not examine the other projects because they were abandoned by their contributors at an early stage and their data are insufficient to support a robust analysis.

After discussing some related work in Sects. , 1 we present the data and methods used in our analysis. In Sect. 1 we present our results. In Sects. 3.1 and 3.2 we analyze the internal structure of collaboration and its role in productivity patterns. We identify a clear hierarchy in participation patterns with a hyperactive elite responsible for 80% of the work. At the same time, we show that collaborative architecture plays an important role in promoting individual production: Indeed, we observe a dynamic of super-production in which the presence of occasional participants helps to increase the productivity of the elite.

After analyzing the organization of work in open science, we focus on the mechanisms of scientific discovery. A mathematical discovery is the rigorous verification of a formal statement, realized by bringing together a set of pre-existing theorems, conjectures, axioms, and so on. It is therefore part of a larger category of innovation processes in which the introduction of new ideas and concepts is crucial to intellectual progress. Innovation processes can be described by the notion of "adjacent possible expansion" introduced by Kauffman (2000). This term refers to the expansion or restructuring of the possible knowledge space, triggered by the introduction of novel concepts. This type of process has been shown to leave distinctive traces in the statistical properties of the knowledge produced, expressed by two key laws first observed in linguistics: Zipf's law and Heaps'

law (Tria et al. 2018). In Sect. 3.3 we show that Polymath's discovery dynamics exhibit the markers of adjacent possible expansion processes, similar to literary production and musical innovation. Finally, in Sect. 3.4 we examine the triggering factors for innovation and show that no rule determines a priori who the key innovators will be, as peripheral users can sometimes steer collective work in new directions.

Related work

There is an extensive and multidisciplinary literature on online collaborative systems, and on interaction patterns in large-scale collaborations: the case of Linux is, for example, analyzed in Maillart et al. (2008), Wikipedia in Voß (2005), Yasseri and Kertész (2013), Ciampaglia and Vancheri (2010), Yasseri et al. (2012), Kittur and Kraut (2008) and GitHub in Thung et al. (2013), Guzman et al. (2014), Sornette et al. (2014).

However, few works focus on large-scale scientific collaboration and even less on Polymath projects. In addition to the reflections by Gowers himself (Gowers and Nielsen 2009), a descriptive analysis of Polymath 1 project can be found in Barany (2010), where the authors provide a qualitative discussion of the rules that Polymath contributors developed to organize their work. For a more quantitative analysis of the initiative, but limited to the first project, see Cranshaw and Kittur (2011). Kloumann et al. (2016) are—to our knowledge—the only authors that have presented a statistical analysis of multiple Polymath projects. Their analysis compares full Polymath projects with the side initiatives of "Mini-Polymath projects", which are smaller collaborations concerning Math Olympics questions that, while quite difcult, have known solutions. A detailed description of the collective problem-solving approach in the third Mini-Polymath has also been provided by Pease and Martin (2012).

Drawing on this literature, our paper develops in three directions. First, we confirm and extend the results on the distribution of labor obtained by Cranshaw and Kittur (2011) on Polymath 1, for the five projects that achieved a final peer-review publication. We also extend this research by considering interaction patterns among contributors. Second, we investigate the productivity of collective intelligence in collaborative systems. Following Sornette's work on GitHub (Sornette et al. 2014), we study the superlinearity of production as a function of the number of users. Finally, based on innovation studies in online systems, such as the one presented in Tria et al. (2018), we introduce an innovation measure for the mathematical production process and identify the actors responsible for introducing innovations in the Polymath ecosystem.

Data and methods

We collected all the posts from Polymath projects 1, 4, 5, 8, and 15, starting from the links listed on the Polymath project wiki page (The Polymath Wiki 2021). The corpus for each project consists of a collection of posts identified by publication date, author, text, and parent post (for posts written in response to another contribution). The posts were published primarily on three blogs: Timothy Gowers's blog (2021), Terence Tao's blog (2021) and The Polymath blog (2021). Each of these blogs entails different technical restrictions on author interaction. On Gowers's blog, comments can only be posted in the main threads, limiting the depth of discussion and preventing authors from responding to comments in sub-threads. In contrast, Tao's blog allows

Project	Number of Posts	Number of Users	First Post	Last post
1	1867	118	2009-02-01	2009-07-04
4	545	63	2009-07-27	2010-07-06
5	2586	107	2009-12-17	2012-09-02
8	3363	199	2013-06-03	2014-08-14
15	1021	57	2018-01-24	2019-04-11

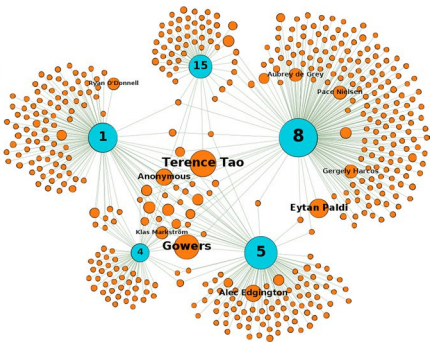

Fig. 1 The table on the left shows the number of posts, the number of contributors and the periods of development for each project. The network on the right represents the participation of each author to the projects

comments up to a depth of 4. The Polymath blog does not appear to limit nested comments at any level and shows comments up to a depth of 10.

Demography of the projects

The five projects we analyzed contain between 545 and 3363 posts and the number of contributors varies from 57 to 199. Detailed information about each project can be found in Fig. 1. The network in Fig. 1 represents the bipartite network of contributors and projects: In the graph, every edge represents an author's participation in a project. The size of the contributors' nodes represents the number of their contributions. The graph shows that there is a small core of very active authors who have participated in almost all projects, and a periphery of occasional contributors working on a single project.

Contents' identification

Since we are interested in reconstructing the collaborative processes that led to the discovery of a mathematical a solution, we need to identify the mathematical objects used in the posts. Natural language processing techniques performed poorly on this task and tended to identify non-mathematical terms, such as features of each participant's personal language patterns. Therefore, we built a mathematical vocabulary by means of a two-steps protocol. First, we collected the titles of all Wikipedia pages labeled as "mathematics" (Lists of mathematics topics 2021). Second, we added to the list all the expressions "Theorem of *", "*'s conjecture", etc. extracted from the corpora. The dictionary we obtained contains 25.035 mathematical concepts. Table 1 shows the number of independent mathematical concepts retrieved in the projects: in expressions like "theorem of *", the sub-string "theorem" is not considered as an independent concept. Through this mathematical dictionary, the content of each post can be qualified by the set of mathematical concepts that it contains: $K_i = \{kw_1, \ldots, kw_m\}$. A post will thus be generally characterized by its time (t), its author (α) and its content (K): $\Pi_i = (t_i, \alpha_i, K_i)$.

Table 1 Number of independent mathematical concepts in the corpora

Project	Concepts
1	1040
4	768
5	1089
8	1116
15	661

Table 2 Number of topics per project

Project	Topics	Modularity
1	9	0.70
4	9	0.75
5	10	0.72
8	12	0.72
15	9	0.69

Topic extraction

We then aggregate different mathematical concepts into topics. This aggregation allows us to study the collaboration between authors and the structure of the collective labor. To define topics, we create a co-occurrence network for each Polymath project, where the nodes represent different mathematical concepts. In this network, two concepts are connected if there is at least one post in which they were discussed together. The network is weighted according to the number of co-occurrences in different posts. Since this network is highly connected and extremely complex, we filter the edges to highlight relevant structures. In order to do so, we compute the Planar Maximally Filtered Graph[1] (PMFG) proposed in Tumminello et al. (2005). We then define our topics as the clusters of mathematical concepts identified by the Louvain community detection algorithm (Blondel et al. 2008) over the PMFG graph. Table 2 shows the number of topics extracted for each project and the modularity of the partition of keywords in the filtered co-occurrences network. Using this definition of topics, we label each post with the topic whose keywords appear most frequently in the text. In case of a tie, no label is assigned to the post. Therefore, in addition to its publication time (t), author (α), and content (K), each post is also characterized by a topic label (T): $\Pi_i = (t_i, \alpha_i, K_i, T_i)$.

Similarity and innovation

We first define the semantic similarity between two posts using the Jaccard measure between their contents: $J_{ij} = (K_i \cap K_j)/(K_i \cup K_j)$. We tweak this similarity by considering the temporal distance among the posts, thus introducing the *semantic-temporal* similarity:

[1] A Planar Maximally Filtered Graph is a filtered graph obtained by subsequently adding edges to a graph that was originally deprived of all its edges. Edges are added in a descending order according to their weights if and only if the resulting graph can still be embedded in a planar surface. A Planar Maximally Filtered Graph preserves the hierarchical organization of the Minimum Spanning Tree but contains a larger amount of information in its structure and proves to be efficient in filtering relevant information about the clustering of the system.

$$\Theta_{i,j} = J_{i,j} e^{-|t_i - t_j|/\tau_0}, \tag{1}$$

where τ_0 is the average time distance among all the pairs of posts (within each project). According to this measure, two posts that are similar in content but distant in time will be less similar than according to the standard Jaccard measure.

We use the semantic-temporal similarity measure to define an innovation index for each post. First we define two separate indicators for each post:

- The *in-debate index* measures the similarity between a post and the contents published before it. It is calculated as the average of the semantic-temporal similarity from the previous posts:

$$\nu_i = \frac{\sum_{j|t_j<t_i} \Theta_{ij}}{\sum_{j|t_j<t_i} 1}. \tag{2}$$

- The *impact index* measures how much a post content is reproduced in the posts following it. It is calculated as the average Jaccard similarity with the following posts:

$$\xi_i = \frac{\sum_{j|t_j>t_i} J_{ij}}{\sum_{j|t_j>t_i} 1}. \tag{3}$$

An innovative post is characterized by a low in-debate index (i.e., it is different from the earlier content) and a high impact (i.e., it influences the following contents that are therefore similar to it). For this reason we define the *innovation index* for each post as:

$$I_i = -\xi_i log(\nu_i). \tag{4}$$

Results

Organization of labor

As usual in collaborative systems, only a few contributors do most of the work (Barabasi 2003). When we analyze the number of contributions made by each author, we find a power-law distribution (Fig. 2B) and high Gini indices (Fig. 2C). In Fig. 2A we represented this distribution in the form of the Lorenz curve: authors are ordered by the number of contributions and curves represent the cumulative fraction of posts produced by the corresponding fraction of ranked authors. From the figure, we can see that the most active 10% produce the 80% of the posts (with the exception of project 4, which is characterized by a lower Gini index, where the 20% of contributors produce the 80% of the posts).

Following the procedure described in Bassolas et al. (2019), we use the Lorenz curve to categorize authors hierarchically: We take the derivative of the Lorenz curve at the point (1,1) and set an initial threshold at the point where the derivative crosses the horizontal axis (as you can see in Fig. 2A). The authors after this threshold represent the most productive *elite* of the project. We remove these *elite* contributors and we repeat the procedure recursively, identifying a group we define as the *first shell* (highly active authors but outside the hyperactive elite) at the first iteration, and the *peripheral shells* (namely shells E3, E4, E5, E6 in Fig. 2D) at subsequent iterations. In Fig. 2C, we display the number of contributors

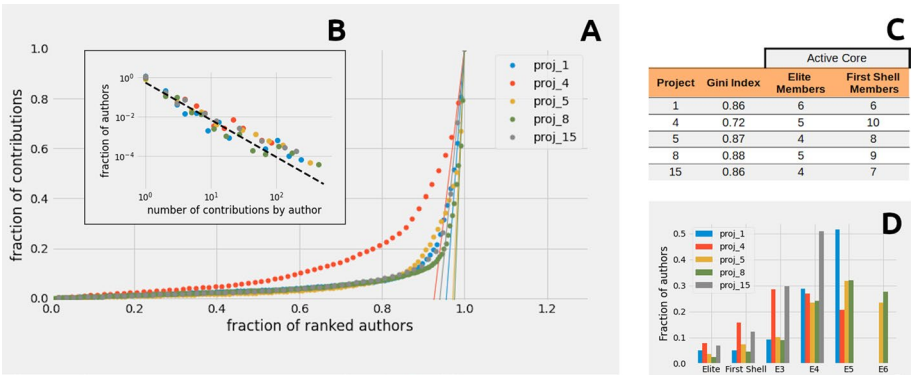

Fig. 2 **A** Lorenz curve for authors' contributions to the projects. **B** Empirical probability distribution of the number of contributions by author. **C** Statistics about contributors activity. **D** Elite–peripheries structure of the projects

in the elite group and in the first shell, while Fig. 2D shows the percentage of authors in each hierarchical category. We can see that, according to this classification, the elite group contains less than 10% of the authors while the peripheral shells are consistently the most represented. In the following we will refer to authors belonging to the *elite* and the *first shell* as the *active core*.

Interactions between the authors

To better understand the division of labor in Polymath, we investigated the distribution of interactions between authors. In particular, we focused on how the *active core* authors, as defined in Sect. 3.1, interact with the *peripheral shells*.

In order to do so, based on the dependencies between posts, we defined a *comments interaction network* $\mathrm{CIN} = (\mathcal{V}, \mathcal{E}, W)$ with the following properties: each node $i \in \mathcal{V}$ represents an author, an edge $(i,j) \in \mathcal{E}$ represents the existence of at least one comment by author i to a post of author j, and the weight W_{ij} associated to the edge (i,j) represents the number of times author i replied to a post of author j. To understand whether such interactions are highly concentrated in the active core of elite authors or more spread towards peripheral contributors, we compared the obtained graphs with a stochastic network model preserving, on average, the activity level of each node. Similarly to Roth et al. (2013), we hence simulate K networks $\{\Gamma^k = (\mathcal{V}, \mathcal{E}^k, Q^k)\}_{k\in\{1,\dots,K\}}$ with an expected degree for each node equal to the one of the authors of our dataset, keeping the same number of nodes $n = |\mathcal{V}|$ and edges $m = |\mathcal{E}| = |\mathcal{E}^k|$ for all $k \in \{1,\dots,K\}$. To do so we draw the weights Q^k_{ij} from a multinomial distribution with parameters m and $p = \{p_{ij}\}_{i,j\in\mathcal{V}}$ such that

$$p_{ij} = \frac{d_i^{\mathrm{out}} \cdot d_j^{\mathrm{in}}}{m^2}$$

where d_i^{out} is the out-degree of node i and d_j^{in} is the in-degree of node j in the comments interaction network. Figure 3A shows the distribution of the *fraction of in-core links* (i.e., the fraction of messages from elite contributors to other elite contributors) in our $K = 100$ simulations and compares these distributions with the actual fraction of

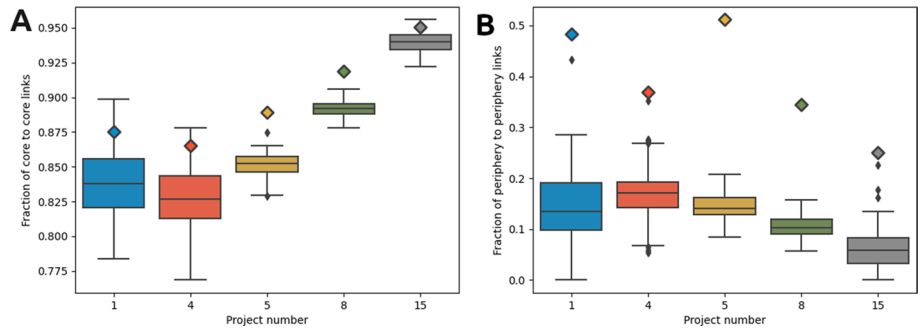

Fig. 3 Interactions through comments: **A** Boxplots represent the distribution of the fraction of in-core links after 100 simulations. ⋄ represents the fraction of in-core links in the Polymath dataset. **B** Boxplots represent the distribution of the fraction of in-periphery links after 100 simulations. ⋄ represents the fraction of in-periphery links in the Polymath dataset

in-core links in our dataset. Figure 3B shows the same comparison for the in-periphery links, i.e., the fraction of messages written by peripheral contributors directed to other peripheral contributors. Both plots show a peculiar division of labor in the Polymath project: both core-to-core and periphery-to-periphery links are more represented than in random simulations, underlining that authors are more likely to reply to contributors who participate in the discovery process to a similar extent.

As mentioned in the Data and Methods section, some of the blogs we studied limit the depth of response structures. We qualitatively observed a shift from a non-hierarchical structure in the very first project (i.e., only the presence of second-level comments and no deeper structures) to a more structured organization of posts in later projects. To evaluate the robustness of the results presented in Fig. 3, we compared them with the results obtained with a different definition of network interactions. We define a *topic interaction network* $\mathrm{TIN(T)} = (\mathcal{V}, \widetilde{\mathcal{E}}(T), \widetilde{W}(T))$ with the following properties: the node set $\mathcal{V}$ still represents the set of authors, an edge $(i,j) \in \widetilde{\mathcal{E}}$ represents the fact that authors i and j published a post on the same topic at a distance no bigger than T posts (when posts are ordered chronologically). The weight $\widetilde{W}_{ij}(T)$, associated to the edge (i, j), represents the number of times author i and j published a post on the same topic in the time window defined by parameter T. Notice that, by definition, such a network is undirected. Once again, in order to study authors interactions, we need to compare them with a set of simulated networks $\{\widetilde{\Gamma}^k = (\mathcal{V}, \widetilde{\mathcal{E}}^k, \widetilde{Q}^k)\}_{k\in\{1,\dots,K\}}$ where $\widetilde{m} = |\widetilde{\mathcal{E}}| = |\widetilde{\mathcal{E}}^k|,\ k \in \{1,\dots,K\}$. To do so, it is now sufficient to draw the solely values $\{\widetilde{Q}_{ij}\}_{i,j\in\mathcal{V},j\geq i}$ from a multinomial random distribution, as we want the network to be undirected and $\widetilde{Q}^k_{ij} = \widetilde{Q}^k_{ji}$ for all the K simulations. Therefore, we draw the values $\{\widetilde{Q}_{ij}\}_{i,j\in\mathcal{V},j\geq i}$ from a multinomial distribution of parameters $\widetilde{m}$ and $\widetilde{p} = \{\widetilde{p}_{ij}\}_{i,j\in\mathcal{V},j\geq i}$ such that

$$\widetilde{p}_{ij} = 2\frac{d_i \cdot d_j}{\widetilde{m}^2} \qquad \text{if } i \neq j$$
$$\widetilde{p}_{ii} = \frac{d_i \cdot d_i}{\widetilde{m}^2} \qquad \text{otherwise,}$$

where d_i is the degree of node i in the topic interaction network. The resulting distribution of in-core and in-periphery interactions is shown in Fig. 4. We notice that,

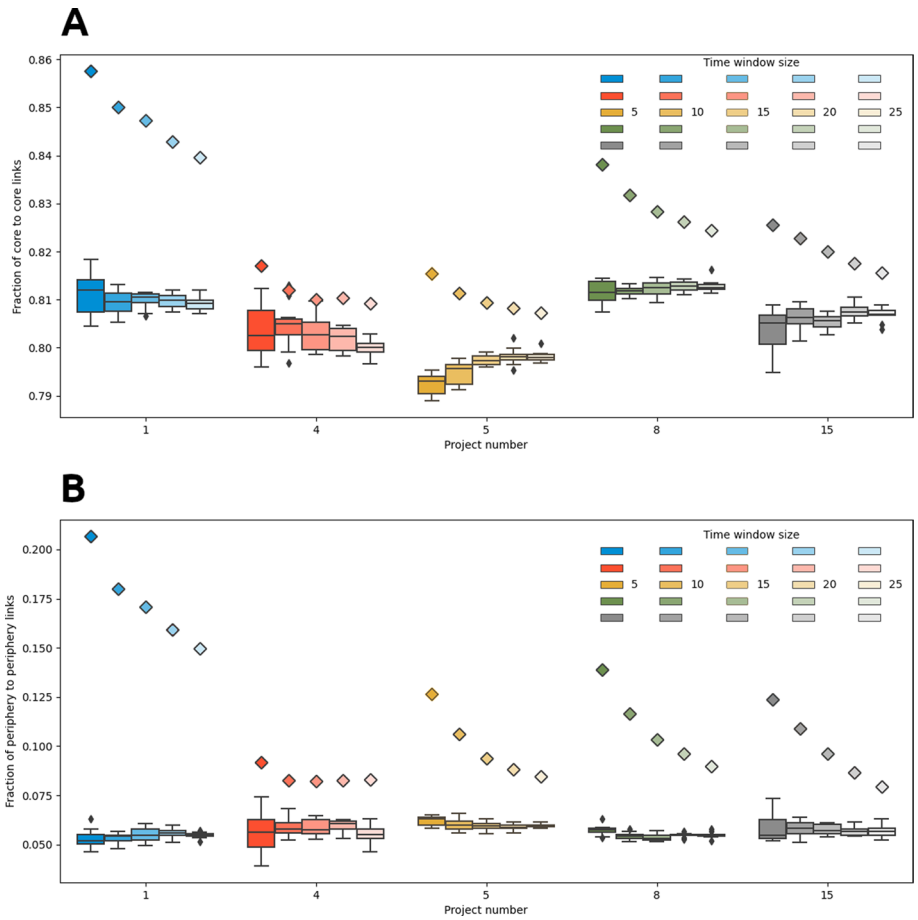

Fig. 4 Interactions on the same topic: boxplots represent the outcome of simulations while ⋄ markers represents real values coming from data. **A** Percentage of links stemming from a core node and ending in a core node over the total number of edges stemming from core nodes. **B** Percentage of links stemming from peripheral nodes and ending in peripheral nodes over the total number of edges stemming from peripheral nodes. The increasing transparency corresponds to the increase of the time window

regardless the definition of the window, in-core and in-periphery interactions are higher in the actual Polymath projects than in simulated networks. Moreover, results obtained on the *topic interaction network* confirm the ones obtained on the *comment interaction network*. We can thus conclude that in the Polymath collaborations, elite actors interact more with other elite actors, while peripheral actors preferentially respond to other peripheral actors. This result is consistent with the forms of status-based homophily observed by McPherson in social networks (McPherson and Smith-Lovin 1987). This is however surprising in a scientific context where interactions are generally assumed to be based on cumulative advantage processes (Merton 1968).

Collective intelligence at work

Several studies on collaborative systems have shown a super-linear effect of collaboration: The very expression "collective intelligence" suggests that the collective productivity (in our case the number of posts) is higher than the sum of the individual productions.

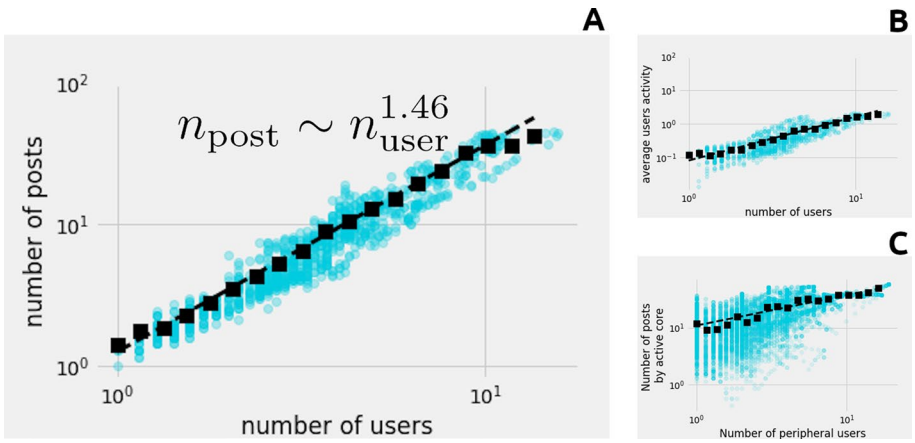

Fig. 5 **A** Superlinear growth of the number of posts with the number of users. **B** Average daily individual productivity as a function of the number of users. **C** Average daily individual productivity of core authors as a function of the number of peripheral contributors

To test this feature dynamically, we count the daily number of posts and the daily number of participants for all projects:

$$n_{post}(t) = [n_{post}(t_0), n_{post}(t_1), \ldots]$$
$$n_{user}(t) = [n_{user}(t_0), n_{user}(t_1), \ldots],$$

where $t_0, t_1, \ldots$ represent different days. To reduce noise, we smooth these time series with a 7-days rolling window. By plotting the pairs $(n_{user}(t), n_{post}(t))$, we obtain the curves representing the relationship between the number of users and the number of posts. Figure 5A shows a pronounced superlinear growth of the number of posts with the number of users, aggregated for all projects: $n_{post} = n_{user}^{\gamma}$ (with exponent $\gamma = 1,46$). Our results are similar to those of Sornette et al. (2014) for GitHub.

Figure 5B, C suggest that contributions have positive super-linear effects, even when they are relatively marginal. In Fig. 5B, we show that the average individual daily production (for all contributors with more than 10 posts in all the projects) grows with the number of users active on that day. Figure 5C displays the average daily productivity of the active core as a function of the number of users in the peripheral shells. We observe that an important presence of peripheral users boosts the productivity of the most active users.

In Fig. 5, we show the results obtained by aggregating all the projects. The individual analysis of each blog shows similar trends with very small variations in the growth exponents (blog1: $\gamma = 1.30$, blog4: $\gamma = 1.22$, blog5: $\gamma = 1.46$, blog8: $\gamma = 1.65$, blog15: $\gamma = 1.50$). Since the blog platforms are diverse, the robustness of these results suggests super-productivity to be an intrinsic characteristic of collaborative science, regardless the communication medium.

Statistical properties of scientific discoveries

While in the previous sections we analyzed collaborative patterns in open science, we now focus on the analysis of the scientific discovery process itself.

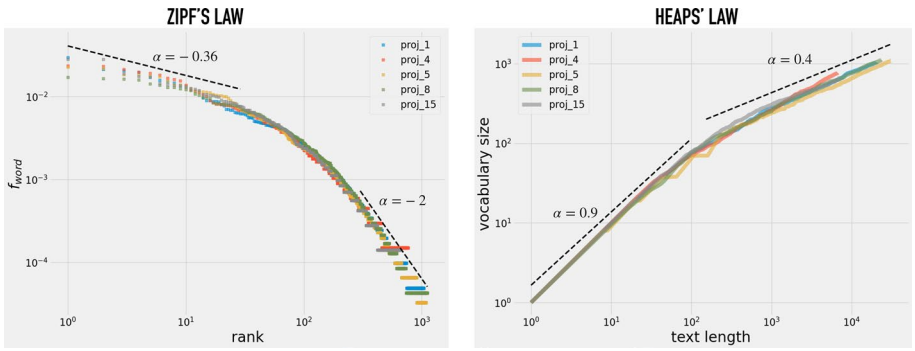

Fig. 6 Left plot: the rank-frequency relationship respects the Zipf's law with the exponents indicated in the plot. Right plot: the relationship between the text length and the vocabulary size respect the Heaps' law with the exponents indicated in the plot

First, we analyze the statistical properties of the mathematical concepts used in the projects. As described in the Methods Section, we have assigned a set of mathematical concepts to each post. We first test whether our corpus follows the basic laws of linguistic patterns: Zipf's Law and Heaps' Law. Zipf's law expresses the relationship between the frequency and the ranking of words. It states that the frequency of a word is inversely correlated with its rank, $f \sim r^{-\alpha}$. For example, looking at the Gutenberg Project corpus (a large sample of English literature), one can observe a value $\alpha \sim -1$ for low values of r and $\alpha \sim -2$ for high values of r. Heaps' law concerns the entry of innovative concepts into a text and expresses the relationship between the number of different words (i.e., the vocabulary size) and the total number of words used (i.e., the length of the text). It describes an initial linear growth followed by an asymptotic behavior according to the power law $l = v^{\alpha}$: in the Gutenberg corpus $\alpha \sim 1$ for low values of l and $\alpha \sim 0.44$ for high values of l have been observed. In Fig. 6, we see that not only are these laws respected in our corpus, but also all projects have the same behavior and exponents, Zipf's exponents being $\alpha = -0.36$ and $\alpha = -2$ and Heaps' exponents being $\alpha = 0.9$ and $\alpha = 0.4$. These values are consistent with those from Gutenberg corpus (Tria et al. 2018), although the first exponent of Zipf's law in our corpora is lower, due to the fact that we removed the non-mathematical expressions and stop words. This consistency means that, statistically, the creative process of scientific discovery follows the same basic rules that characterize literary production.

Second, we focused on the typical timing of the discovery process, based on the the hypothesis that posts that are close in time would also tend to be similar in terms of content. In Fig. 7 we show the average Jaccard similarity between all pairs of posts published within a given time delay. We observe a power law decay of similarity with time, $J \sim \Delta t^{-\gamma}$ (with $\gamma = 0.2$), once again similar for all projects. Thus, for all projects, there exists a typical time window in which the debate remains focused on the same topic before switching to new one.

Innovation patterns

Finally, we analyze how innovations affects the discovery mechanism, by using the innovation measure we defined in Sect. 2.4. As observed in Fig. 8A, the innovation

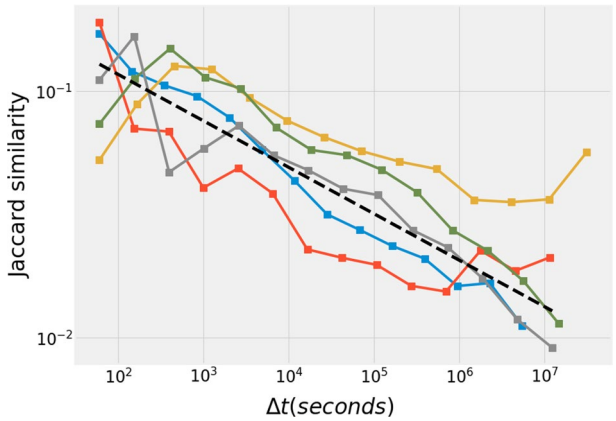

Fig. 7 Average distance between posts at a certain temporal distance

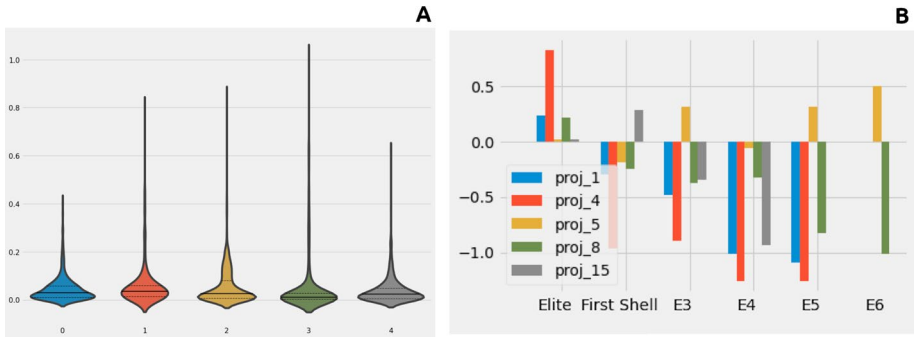

Fig. 8 Plot **A** Distributions of innovation for each project. Plot **B** Innovation z-score for each activity shell

values' distributions are long-tailed, meaning that few posts have a much larger innovative content compared to the others: high innovation is rare, but statistically significant.

We define posts in the top quartile of the innovation distribution as innovative. Then, referring to the definition of activity shells introduced in Sect. 3.1, we examine which actors lead innovation. Since the groups vary in size, we compare the number of innovations observed in each class with their multinomial expectation, namely the probability that a post is innovative (25%) multiplied by the number of posts produced by the group. We calculate the z-score between the observed and expected values. While the previous results showed a fairly homogeneous behavior between the different projects, here we observe significant differences. In projects 1,4,8, the elite produces more innovation than expected. In project 15, the first shell is the main driver of innovation. Finally, in project 5, the peripheral shells are the largest producers of innovation. This result highlights that in large-scale collaborations no rule determines a priori who will be the main innovators. An innovator can be a member of the hyper-active elite, but sometimes serendipitous interactions of peripheral participants can also have a large impact on the discovery process: an isolated contribution of an occasional participant can be responsible for opening a large adjacent possible and giving a new direction to the work.

Conclusion

Over the past few decades, we have witnessed the rise of large online collaborations such as Linux, GitHub, and Wikipedia. In 2009, the first Polymath project was launched with the goal of exploiting online collaborative environments to solve mathematical problems.

In this work, we have investigated how the path to scientific discovery develops in this collaborative environment, how the labor is organized between authors, and which actors are the main innovators. Our results, which are consistent with previous works, show that productivity is highly skewed between contributors and that there is a small hyper-productive elite that publishes the bulk of the contributions. Nonetheless, peripheral contributors also play a significant role, as content production grows super-linearly with the number of discussants.

Our analysis shows that, in Polymath projects, peripheral contributions boost the activity of other authors in a rather indirect way. Although interactions between the elite and the rest of the participants are relatively limited (as both peripheral and hyper-productive authors tend to interact mainly with authors with similar levels of activity), we have demonstrated that peripheral authors often play a crucial role in bringing new and innovative ideas to the debate. Our analysis has also shown that innovators cannot be defined only by their productivity level. Sometimes, occasional contributors can play a key role in innovation and be responsible for steering the research in new directions.

In this exploration of the Polymath ecosystem, we focused on four main directions: classifying contributors by their involvement in the projects, analyzing interactions among contributors, examining the impact of large-scale collaboration on productivity, and finally identifying the actors responsible for innovation. We conducted only a limited analysis of the content of the posts and of the semantic relationships between them. In a follow-up study, it would be interesting to examine the relationships between contributions based on the similarity of the content they produced rather than considering only their direct interactions in the response network. This structure would allow us to analyze the thematic cooperation patterns between contributors. This similarity network would also allow us to characterize the internal composition of users' "opinions" on solution techniques and their complex dynamics. Finally, it would be interesting in future works to compare the results obtained from the Polymath dataset with other online collaborative environments, in particular, to analyze the relationship between the level of participation and innovation, which remains largely unexplored in the literature.

Abbreviations

PMFG: Planar maximally filtered graph; CIN: Comments interaction network; TIN: Topic interaction network.

Authors' contributions

F.G. conceived the research, collected and analyzed the data, discussed the results, wrote the manuscript. M.C. collected and analyzed the data, discussed the results, wrote the manuscript. T.V. discussed the results, wrote the manuscript. P.F. discussed the results, wrote the manuscript. All authors have read and approved the manuscript.

Funding

This research has been supported by CNRS through the 80 PRIME MITI project "Disorders of Online Media" (DOOM). This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003) and by ANR grant HANDY (ANR-18-CE40-0010).

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author upon request.

Declaration

Competing interests
The authors declare to have no competing interests.

Author details
[1]CNRS, Université Paris-Sorbonne - Paris IV, GEMASS, Paris, France. [2]CNRS, Inria, Grenoble INP, GIPSA-Lab, Univ. Grenoble Alpes, Grenoble, France. [3]CNRS, CIS-lab, Paris, France.

Received: 20 September 2021 Accepted: 6 March 2022
Published online: 28 March 2022

References

Anthony D, Smith SW, Williamson T (2009) Reputation and reliability in collective goods: the case of the online encyclopedia Wikipedia. Ration Soc 21(3):283–306
Barabasi A (2003) Linked: how everything is connected to everything else and what it means for business, science, and everyday life. Plume Books, New York
Barany MJ (2010) '[b]ut this is blog maths and we're free to make up conventions as we go along': Polymath1 and the modalities of 'massively collaborative mathematics'. In: Proceedings of the 6th international symposium on wikis and open collaboration. WikiSym '10. Association for Computing Machinery, New York, NY, USA
Bassolas A, Barbosa-Filho H, Dickinson B, Dotiwalla X, Eastham P, Gallotti R, Ghoshal G, Gipson B, Hazarie SA, Kautz H, Kucuktunc O, Lieber A, Sadilek A, Ramasco JJ (2019) Hierarchical organization of urban mobility and its connection with city livability. Nat Commun 10(1):4817
Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008(10):10008
Ciampaglia GL, Vancheri A (2010) Empirical analysis of user participation in online communities: the case of wikipedia. ICWSM 4(1)
Cranshaw J, Kittur A (2011) The polymath project: Lessons from a successful online collaboration in mathematics. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '11, pp 1865–1874. Association for Computing Machinery, New York, NY, USA
Fallis D (2008) Toward an epistemology of Wikipedia. J Am Soc Inform Sci Technol 59(10):1662–1674
Gowers T, Nielsen M (2009) Massively collaborative mathematics. Nature 461(7266):879–881
Grasso A, Convertino G (2012) Collective intelligence in organizations: tools and studies. Comput Supp Cooper Work (CSCW) 21(4):357–369
Guzman E, Azócar D, Li Y (2014) Sentiment analysis of commit comments in GitHub: an empirical study. In: Proceedings of the 11th working conference on mining software repositories. MSR 2014, pp 352–355. Association for Computing Machinery, New York, NY, USA
Kauffman SA (2000) Investigations. Oxford University Press, Oxford
Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in wikipedia: quality through coordination. In: Proceedings of the 2008 ACM conference on computer supported cooperative work. CSCW '08, pp 37–46. Association for Computing Machinery, New York, NY, USA
Kloumann IM, Tan C, Kleinberg J, Lee L(2016) Internet collaboration on extremely difficult problems: research versus olympiad questions on the polymath site. In: Proceedings of the 25th international conference on world wide web, pp 1283–1292
Lists of mathematics topics on Wikipedia. https://en.wikipedia.org/wiki/Lists_of_mathematics_topics, visited on 2021-03-31
Maillart T, Sornette D, Spaeth S, Krogh G (2008) Empirical tests of Zipf's law mechanism in open source linux distribution. Phys Rev Lett 101:218701
McPherson JM, Smith-Lovin L (1987) Homophily in voluntary organizations: status distance and the composition of face-to-face groups. Am Sociol Rev 52(3):370–379. https://doi.org/10.2307/2095356
Merton RK (1968) The Matthew effect in science. The reward and communication systems of science are considered. Science 159(3810):56–63
Nielsen M (2011) Reinventing discovery. Princeton University Press, Princeton
Pease A, Martin U (2012) Seventy four minutes of mathematics: an analysis of the third Mini-Polymath project. In: AISB/IACAP World Congress 2012: symposium on mathematical practice and cognition II, part of Alan Turing Year 2012, pp 19–29
Roth C, Gargiulo F, Bringé A, Hamberger K (2013) Random alliance networks. Soc Netw 35(3):394–405
Sornette D, Maillart T, Ghezzi G (2014) How much is the whole really more than the sum of its parts? 1 ⊞ 1 = 2.5: superlinear productivity in collective group actions. PLoS ONE 9(8):103023
Straub VJ, Tsvetkova M, Yasseri T (2021) The cost of coordination can exceed the benefit of collaboration in performing complex tasks. arXiv: 2009.11038
Terence Tao's blog. https://terrytao.wordpress.com, visited on 2021-03-31
The Polymath blog. https://polymathprojects.org, visited on 2021-03-31
The Polymath Wiki. https://asone.ai/polymath/index.php?title=Main_Page, visited on 2021-03-31
Thung F, Bissyande TF, Lo D, Jiang L (2013) Network structure of social coding in GitHub. In: Proceedings of the 17th European conference on software maintenance and reengineering. CSMR '13, pp 323–326. IEEE Computer Society, USA
Timothy Gowers's blog. https://gowers.wordpress.com, visited on 2021-03-31
Tria F, Loreto V, Servedio VDP (2018) Zipf's, Heaps' and Taylor's laws are determined by the expansion into the adjacent possible. Entropy 20(10):752
Tumminello M, Aste T, Matteo TD, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci USA 102(30):10421–10426
Voß J (2005) Measuring wikipedia. In: Proceedings of ISSI, pp 221–231

Yasseri T, Kertész J (2013) Value production in a collaborative environment. J Stat Phys 151(3–4):414–439
Yasseri T, Sumi R, Kertesz J (2012) Circadian patterns of wikipedia editorial activity: a demographic analysis. PLoS ONE 7(1):30091–30091

Publisher's Note