Internet Collaboration on Extremely Difficult Problems: Research versus Olympiad Questions on the Polymath Site

Isabel Kloumann Chenhao Tan Jon Kleinberg Lillian Lee
Cornell University
imk36@cornell.edu,{chenhao|kleinber|llee}@cs.cornell.edu

ABSTRACT

Despite the existence of highly successful Internet collaborations on complex projects, including open-source software, little is known about how Internet collaborations work for solving "extremely" difficult problems, such as open-ended research questions We quantitatively investigate a series of efforts known as the *Polymath* projects, which tackle mathematical research problems through open online discussion. A key analytical insight is that we can contrast the polymath projects with *mini-polymaths* — spinoffs that were conducted in the same manner as the polymaths but aimed at addressing math Olympiad questions, which, while quite difficult, are known to be feasible.

Our comparative analysis shifts between three elements of the projects: the roles and relationships of the authors, the temporal dynamics of how the projects evolved, and the linguistic properties of the discussions themselves. We find interesting differences between the two domains through each of these analyses, and present these analyses as a template to facilitate comparison between Polymath and other domains for collaboration and communication. We also develop models that have strong performance in distinguishing research-level comments based on any of our groups of features. Finally, we examine whether comments representing research breakthroughs can be recognized more effectively based on their intrinsic features, or by the (re-)actions of others, and find good predictive power in linguistic features.

1. INTRODUCTION

Groups interacting on the Internet have produced a wide range of important collaborative products, including encyclopedias, annotated scientific datasets, and large pieces of open-source software. These successes led the Fields Medalist Timothy Gowers to ask whether a similar style of collaboration could be used to approach open research questions. In particular, his focus was on his own domain of expertise, mathematics, and in early 2009 [7] he famously asked, "Is massively collaborative mathematics possible?"

Shortly after posing this question, he and a group of colleagues set out to test the proposition by attempting it. They began the first in a series of so-called *Polymath projects*; in each Polymath project, an open, evolving group of mathematicians communicate via a shared blog attempt to solve an open research problem in mathematics. The groups have been quite diverse in background; they have included active participation from Gowers and a second Fields Medalist, Terence Tao, along with a large set of both professional and amateur mathematicians. To date there have been nine Polymath projects; three of them have led to published papers and one to notable partial results preceding the subsequent resolution of its central question, thus demonstrating that this approach can lead to new mathematical research contributions with some regularity.

The Polymath projects have an explicitly articulated set of guidelines that strongly encourage participants to share all of their ideas via online comments in very small increments as they happen, rather than thinking off-line and waiting to contribute a larger idea in a single chunk. We can thus see, through the comments made on the site during the project, almost all the ideas, experiments, mistakes, and coordination mechanisms that participants contributed.

Attempts to think about the nature of the collaboration underpinning Polymath lead naturally to analogies in several different directions. One analogy is to the online collaborations one finds in other settings, such as Wikipedia [11] and open-source software projects [19]. A second analogy is to large decentralized collaborations that take place in "traditional" scientific research [10].

But both analogies are limited. The first does not quite fit because our existing models of collaborative work on the Internet involve domains where the task is inherently "doable": the feasibility of the task — authoring an encyclopedia article or writing an open-source computer program to match a known specification — is not in doubt, and the primary challenge is to achieve the requisite level of scale and robustness. In Polymath, on the other hand, we see people who are the best in the world at what they do struggling with a task that might be beyond them or impossible as they work on open problems in their field.

The second analogy also does not quite fit: as noted by Gowers [7], decentralized scientific collaborations have typically focused on problems that are inherently decomposable into separate pieces. With Polymath, on the other hand, we see problems that present themselves initially as a unified whole, and any decomposition needs to arise from the collaboration itself. Anyone with Internet access can participate for any period of time that they wish.

For all these reasons, Polymath provides a glimpse into a novel kind of activity — the use of Internet collaboration to undertake world-class research — in a way that is not only open but completely chronicled. In the same way that co-authorship networks have provided a glimpse into the fine-grained structure of scientific partnerships [9, 6], the contents of Polymath offer a look at the minute-by-minute communication leading to the research that these partnerships enable.

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883023.

With a growing number of sites where people congregate to discuss solutions to hard problems, it is useful to also appreciate the basic similarities between Polymath and other Web-based communication and collaboration platforms. Even if the specific findings about Polymath do not generalize to all other contexts, the questions themselves can often be generalized. With this in mind, an additional goal of the paper, beyond the investigation of Polymath as a domain, is to present a template for questions that we believe can be productively asked in general about the type of data that sites like Polymath generate. We hope that this template will help facilitate direct comparisons and contrasts with future studies of collaborative Web-based problem-solving.

1.1 Summary of contributions

Data from Polymath 1 was analyzed in an interesting paper by Cranshaw and Kittur [2]; in their own words, they provide "an in-depth descriptive analysis of data gathered from [Polymath 1]," focusing on the role of leadership in the progress of the project, and the interaction between established members and newcomers as the projects proceeded. With the inception of eight new Polymath projects, and rich variation in their evolution and success, a new set of opportunities arises in the type of questions we can explore with Polymath data. We organize our analysis around Here we attempt to address *two central questions* regarding Polymath.

(1) Research or hard problem-solving?

At a general level, our first question is to analyze some of the distinctions between online discussion about open research questions versus online discussion about tasks where the outcome is more attainable.

To address this question, and to make the comparison as sharp as possible, we use a source of discussion data that comes from Polymath itself: the *mini-polymath projects*. Shortly after Polymath was successfully underway, Terence Tao assembled a group to solve something hard but more manageable than a research question; each mini-polymath problem is a question from a past International Mathematical Olympiad (IMO). The existence of the mini-polymaths provides us with a very natural contrast between the two types of activities. Specifically, we can understand the differences between tackling an open-ended research problem, where current techniques may be completely inadequate for finding a solution, vs. solving a problem that, while difficult, is known to be feasible, in a setting where, to a large degree, there is control for topic (in both cases, difficult mathematics) and for participants (there are dozens of people who participated in *both* Polymath and the mini-polymaths). We study and contrast the polymath and mini-polymath projects with three lenses: the roles and relationships of the authors, the temporal evolution of the projects, and the linguistic properties of the comments.

Roles of authors and leadership. First, we analyze the role of the authors, the role of leadership, and differences in patterns of conversation networks in the two domains. In particular, in the research domain we observe that there is a substantially higher concentration of activity in the hands of fewer people, indicating that there was a more distinct notion of contribution leadership in the research domain than the somewhat easier mini-polymath domain. We further observe that there is significantly more symmetry in the global conversation network than what would be initially expected, which is not the case in the mini-polymath projects.

Temporal dynamics. Second, we consider how progress in the two domains evolved over time, and observe interesting patterns both in differences and similarities between the two domains. The two types of projects differ in the temporal properties of the discussion: overall, comments come more quickly in mini-polymath projects, befitting their smaller-scale format, but, interestingly and unexpectedly, on the shortest time scales comments actually come more quickly in Polymath, indicating that the research discussions have the potential to reach the most rapid-fire rate.

Linguistic properties. Third, we study the use of language in the two domains, in both content and high-level linguistic features such as politeness, relevance, and specificity, again finding interesting differences between the two domains. Strong signals in the text distinguish comments in Polymath projects from those in mini-polymath projects. At the most naive level, using bag-of-words classification achieves an accuracy above 90%, since problem-specific terms and time differences (as expressed by words such as "primes" or "July") can be prominent in these two kinds of discussions. But surprisingly, and more importantly, restricting attention to just words that are *not* topic-focused still achieves 90% accuracy, suggesting stylistic differences in Polymath comments and mini-polymath comments. Additionally, high-level linguistic features beyond just individual words display significant differences between the two domains: research discussions in Polymath projects have higher average word distinctiveness, higher relevance to the original post for the topic, greater politeness, and greater usage of the past tense.

(2) General contribution or research highlight?

Our second question is based on a key aspect of research collaborations — they they pass through "milestones" when important progress is made. Can we characterize such milestones as the collaboration unfolds? With the ability to do this, one may be able to set up If we are able to do this, we may set up mechanisms that help researchers focus on promising directions, which can potentially result in more productive research collaboration. Alternately, a more pessimistic hypothesis is that these milestones may only be realized in retrospect. To characterize these milestones, we formulate a prediction problem that asks whether it is possible to identify comments that were marked "highlights" by participants.

The task of identifying highlights turns out to be more challenging than our first task, distinguishing Polymath comments from mini-polymath ones. Nevertheless, we still obtain prediction performance significantly above the baselines for the task. To help understand whether the challenge is inherently in the task or in the shortcomings of our prediction algorithms, we compared to the performance of applied mathematics graduate students in recognizing highlights from Polymath discussions. Algorithms using the strongest feature sets achieve comparable performance to these human judges. We also find that features based on the individual comments themselves outperform features that try to capture reactions or the run-ups to the comments in question.

2. DATA

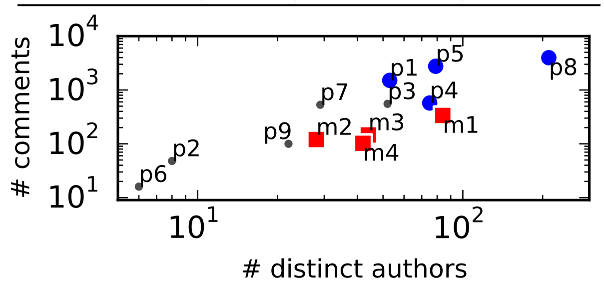
The Polymath and mini-polymath projects share their common roots in a gateway wiki hosted by Michael Nielsen[1]. Starting from that site, we parsed all discussion comments, and for each comment retained its text, its author's WordPress username, its timestamp (with minute-level granularity), and its permalink.

For portions of our analysis we use all the Polymath projects, but in other parts we focus on the most active and successful. As Table 1 indicates, there is a relatively wide variation in the amount of content produced as part of each Polymath project, as well as variation in their levels of success. The mini-polymaths, on the other hand, are more uniform and each solved the Olympiad problem that

[1] `http://bit.ly/1Spixks`

Table 1: Activity summary for each of the polymath and mini-polymath projects. Focal polymath projects of the present study are highlighted in blue, other polymath projects are shown in black, and mini-polymath projects in red. *Tag*: label used in subsequent figures. *Papers*: number of papers written by the corresponding project. *See Footnote 2 regarding partial results from Polymath 5. *Active days*: number of days on which at least one comment was made. The figure shows the number of comments and distinct authors in each project.

Project (tag)	*Papers*	*# of comments*	*Active days*
Polymath 1 (p1)	2	1509	112
Polymath 4 (p4)	1	573	103
Polymath 5 (p5)	0*	2757	238
Polymath 8 (p8)	2	3975	413
Polymath 2 (p2)	0	48	10
Polymath 3 (p3)	0	553	110
Polymath 6 (p6)	0	16	4
Polymath 7 (p7)	0	531	81
Polymath 9 (p9)	0	100	28
Mini 1 (m1)	n/a	336	15
Mini 2 (m2)	n/a	120	7
Mini 3 (m3)	n/a	146	16
Mini 4 (m4)	n/a	102	10



they focused on. When comparing Polymath to mini-polymath, we often focus on the subset of Polymath projects whose successful outcomes are analogous to the successes of the mini-polymaths; these are the Polymaths that led to publications (Polymaths 1, 4, and 8) as well as Polymath 5 which was also highly active and led to important partial results on the Erdos Discrepancy Problem (EDP).[2] Unless otherwise stated when we refer to *quantitative* results or observations about the Polymath projects, we are referring to this subset.

In addition, we collected data about which comments in the Polymath 1 project were identified as research highlights, which was recorded on a subpage of the Polymath projects wiki page.

The data studied in this paper has been made publicly available online at https://bitbucket.org/isabelmette/polymath-data.

3. ROLES AND LEADERSHIP

3.1 Leadership and inequality in research discussions

What is the role of the top contributors in the Polymath research setting compared to mini-polymath's simpler domain? Similarly, what role do authors who contribute less frequently play in the two settings? And how does the interaction structure of the authors vary across the projects? We find striking differences between the two domains; contrasts in the leadership structures are present by design, but the differences in the organic structure of participation stand out equally strongly.

There is an initial superficial difference between the Polymath and mini-polymath projects: in the Polymath projects, the leaders were also among the main contributors, while the mini-polymath projects were designed so that the leaders did not contribute extensively.[3] In a bit more detail, there is a clear definition of "the leadership" in the Polymath projects, as Tao and Gowers were both the project hosts (they collaboratively hosted Polymath 1 on their two blogs) and its two most prolific authors. Table 2 lists the hosts for each project alongside each project's two top contributors. In the Polymath projects the hosts are almost always among the top contributors.

Table 2: Overview of leadership in the Polymath projects and mini-polymath projects.

Project	Host(s)	Top two contributors
Polymath 1	Tao, Gowers	Gowers, Tao
Polymath 4	Tao	Tao, Croot
Polymath 5	Gowers	Gowers, Edgington
Polymath 8	Tao, Morrison	Tao, Paldi
Mini 1	Tao	Bennet, Speyer
Mini 2	Tao	Bennet, Hill
Mini 3	Tao	Thomas H, Narayanan
Mini 4	Tao	Gagika, Olli

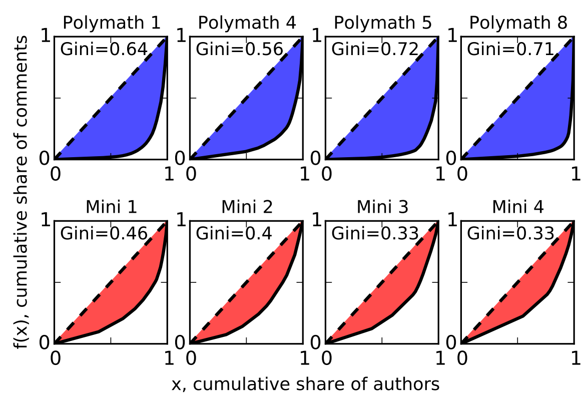


Figure 1: The Gini coefficient — the area between the solid and dashed lines — indicates that there is more equality in the mini-polymath author-comment distributions than in Polymath's. The vertical axis $f(x)$ is the cumulative fraction of comments that have been contributed by the corresponding cumulative share of authors x, where the authors are sorted by increasing number of comments written. Dashed line: $f(x)$ for a hypothetical uniform distribution. Solid line: observed distribution in the given project.

[2]Polymath 5 was very active (see Table 1) and led to partial though unpublished results, which were then cited by Terence Tao when he published his resolution of the EDP in 2015 [18].

[3]As Tao noted in setting up the mini-polymath projects, he hosted them (either via his own blog or as the moderator on the polymath-projects.org blog), but he refrained from contributing to the collaborative effort, stating, "I myself worked it out ... in order not to spoil the experiment, I would ask that those of you who have already found a solution not to give any hint of the solution here until after the collaborative effort has found its solution. ... I will not be participating in the project except as a moderator."

Moving beyond this straightforward distinction between moderators and contributors, we explore to what extent contributions in the successful Polymath and mini-polymath projects were made by a small group of active authors versus shared across a larger group.

On one hand we have the hypothesis that the easier mini-polymath projects could more easily be dominated and solved by just a handful of people, while the more difficult projects would require contributions from a greater number of people. On the other hand, it may be that work in the mini-polymaths would be distributed more evenly among many people because their lower difficulty level made them accessible to a larger group, whereas in Polymath the problems are so difficult that very few people are able to make a substantial number of contributions.

We explore this question and find clear differences in the role of leadership and heterogeneity using the Gini coefficient, a well-known measure of a system's inequality, as shown in Figure 1. In this domain we apply the Gini coefficient to the fraction of authors who contribute a given fraction of the total number of comments in a system. The Gini coefficient is computed via the Lorenz curve, the fraction of comments $f(x)$ made by the x fraction of people who provided the least number of comments. Larger Gini coefficients indicate more inequality.

From Figure 1, we find that the mini-polymath projects possess a notably greater degree of commenting equality (a lower Gini coefficient) than the research projects. This means that in the research domain a larger fraction of comment contributions was made by a smaller fraction of authors. But while research discussions tend to be dominated by fewer people, do the less dominant people still make meaningful contributions? We find that the answer is yes. Recall from the introduction that a subset of the comments in Polymath 1 were labeled as "highlights" by participants. We can thus measure the Gini coefficient on two separate sub-populations defined by these labels: the highlights and the complement of the highlights. We find that the two sub-populations have nearly identical distributions, and thus to the extent that lower frequency contributors participated in Polymath 1, they were making contributions that were indeed classified as highlights in the overall success of the project.

3.2 Symmetry and Sticky Conversations

What does the sequence of participants in a conversation tell us about the domain? How does the reply structure of a conversation aimed at solving an extremely hard problem compare to the reply structure in an easier problem-solving domain? To investigate these questions, we pinpoint two closely related metrics: *reply symmetry* and *stickiness*.

Setup and baseline. Both metrics, reply symmetry and stickiness, are computed using the sequence of authors who comment on the project.

In particular, for each project we have the set of authors who comment, denoted $\mathcal{A} = \{a^i\}_{i=1}^n$, and the sequence S in which their m comments were made: $S = \{a_1^i, a_2^j, \dots\}$. The random baseline for these metrics will be based on a time-zone-controlled random sequence. That is, to create a random sequence S^{rand}, for position S_i^{rand}, we select a random author from the set of authors who have commented in that hour of the day, proportional to how frequently they have commented during that hour.

Definition: reply symmetry. To define reply symmetry we consider the reply matrix A: A_{ij} is the number of times author j follows author i in the sequence S. We then define symmetry in the matrix as $sym(A) = 1 - \dfrac{|A - A^T|_1}{|A|_1}$. The 1-norm of the matrix A, $|A|_1$, is the total number of comments, and $|A - A^T|_1$ is the number of alterations that would be made to the sequence of comments such that the reply matrix is completely symmetric.

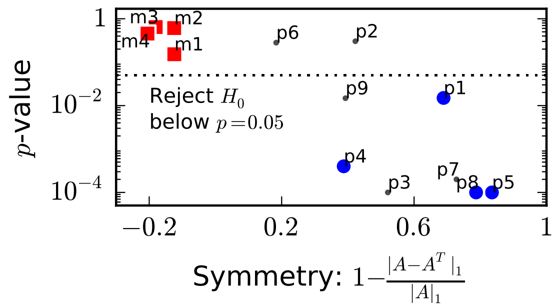


Figure 2: Symmetry of conversations in Polymath and mini-polymath projects. The horizontal axis is the amount of symmetry observed in comment threads: higher symmetry indicates that authors follow up comments from the same authors who follow up their comments, and the p-value on the vertical axis is the bootstrapped estimate of the level of significance at which we can reject the null hypothesis that the symmetry is due to random variations.

Table 3: The increased likelihood of the motif in the Polymath projects and mini-polymath projects. *, **, and * indicate that the result was significant when measured against a time-zone-controlled random baseline (as defined in the text) at the 95%, 99%, and 99.9% significant levels respectively. Otherwise, the number in parentheses indicates the $p-$value. nan indicates that there were no examples of this motif in the temporally-controlled random baseline.**

Project	aaa	aba
Polymath 1	5.15***,	1.41 (0.25)
Polymath 4	2.34***,	1.42**
Polymath 5	3.51***,	1.54*
Polymath 8	3.65***,	1.91***,
Mini 1	3.55***,	1.5 (0.14),
Mini 2	5.14***,	0.86 (0.82)
Mini 3	1.9***,	0.68 (0.92)
Mini 4	nan***,	0.82 (0.79)

This definition captures the extent to which people respond to the same people who respond to them, regardless of whether they respond immediately in real time, or at a later time.

Definition: stickiness. Next we define the notion of *stickiness*, which captures the local author symmetry in comment sequences. In the author sequence, we first count the number of times we observe the sequence motif aba — an author a is followed by another author b, who is then followed again by a. Similarly, the motif abc corresponds to comments by three distinct authors in succession, while the motif aaa corresponds to three comments in a row by the same author. We define *stickiness* of the interaction to be the extent to which the aba motif is overrepresented; it is the probability of observing the motif aba in the real sequence relative to the probability of observing it in a time-zone-controlled random baseline (the likelihood ratio).

Results: symmetry and stickiness in research domains. In Figure 2 we test the hypothesis that the amount of symmetry observed is as much as would be observed by a random, asymmetric graph. We find that in each case the bootstrapped $p-$value for the Polymath projects is less than 0.05, indicating that we can reject this hypothesis and that the symmetry we observe is more than one would expect from random fluctuations (the exceptions are Polymath projects 2 and 6, which both have fewer than 10 authors and 100 comments total, which is too little data to compute a meaningful estimate). On the other hand, for each of the mini-polymath projects the estimated $p-$value is above 0.05, indicating that the observed symmetry may be due to random variations.

Similarly, in Table 3, we observe that in three of the four polymath projects under question there is significantly more stickiness than in the random baseline, whereas in three of four mini-polymath projects, there is less.

What we find surprising about these phenomena of increased symmetry and stickiness is not that it occurs at all, but that we observe it in the Polymath projects while not observing it to the same extent in the mini-polymath projects, which was hosted on the same platform and involved a similar group of people.

We expect that in the Polymath projects it is at least in part thanks to a norm that emerged from the collaboration: as conversation in each project developed, there were a large number of subproblems that needed to be completed (everything from running simulations, to reviewing related work, to building information sharing web-apps), and subgroups of people would work on them together. These subgroups of people would tend to communicate with each other more frequently than with other people, leading to the asymmetry we have observed.

The apparent lack of stickiness in the mini-polymath projects compared to the polymath projects may indicate that the role of smaller groups discussing subproblems was less important in this easier problem domain.

4. TEMPORAL LEVEL FEATURES

4.1 Response time dynamics

The time scale on which mini-polymath projects play out is quite different from that of the Polymath projects, with the latter taking place over the course of several months to a year and the former being concluded in a matter of days. This difference in overall time scales suggests that we consider contrasts in the responsiveness dynamics for Polymath versus mini-polymath projects: when an author posts a comment, how quickly do people follow up after them and how do those dynamics compare in the two types of collaborations? We find that the answer is subtle and depends on the temporal scale of analysis itself.

First, we define the *response time* of a comment to be the amount of time that has elapsed since the comment immediately preceding it was posted.[4] We then consider the mean response time in Polymath and mini-polymath, conditioned on those response times being less than some upper threshold. That is, for some value t, what is the mean response time of all comments whose response time is less then t? We denote this quantity by $\bar{r}^t_{main}$ and $\bar{r}^t_{mini}$ for Polymath and mini-polymath, respectively.

Given that mini-polymath projects played out much more quickly overall than Polymath projects, it would be natural to expect that response times on mini-polymath should be less than those on Polymath for all values of the threshold t; that is, one would expect a positive difference $\bar{r}^t_{main} - \bar{r}^t_{mini}$.

What we find is more subtle, in that it depends on the threshold t; we get a different answer if we condition on comments made within a few minutes of each other. In Figure 3(top), we observe that when

[4]The blog data includes comment timestamps with one-minute granularity.

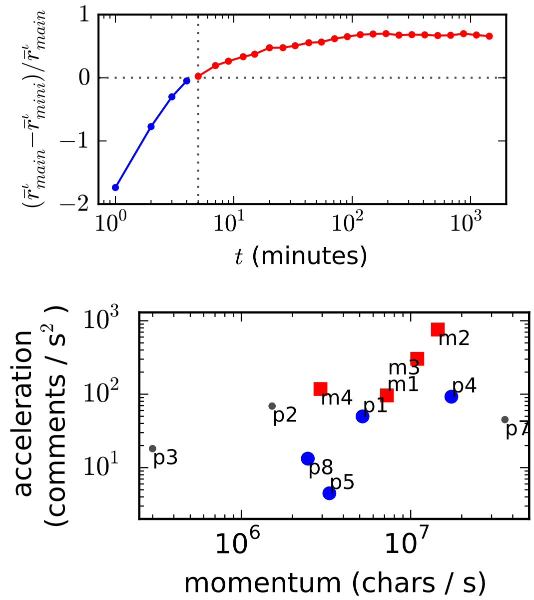


Figure 3: (Blue vs. red indicates temporal regime where Polymath vs. mini-polymath has faster response time). Top figure: Vertical axis: fractional difference in response times between the Polymaths and mini-polymaths, conditioned on the response time being less than what is indicated by the horizontal axis. Hence, negative values indicate that responses in Polymath are faster than in mini-polymath. Bottom figure: Average commenting acceleration vs. average commenting momentum. Velocity units are #comments per minute, and acceleration units are #comments per minute, per minute.

we focus in on very small time scales of less than five minutes, commenting in Polymath is actually faster than in mini-polymath. This is reflected in the negative difference $\bar{r}^t_{main} - \bar{r}^t_{mini}$ for $t < 5$ minutes. And then (as expected) as we allow for comments with larger and larger response times, the mean response time in Polymath becomes larger than in mini-polymath. In the figure we report the mean difference, and consider the $p-$value corresponding to the significance with which we reject the null hypothesis that the means are the same, estimated using Welch's t-test (for comparing population means between populations with unequal variance). For all thresholds except 4 and 5 minutes, at which the transition between mean signs is observed, $p < 0.001$.

4.2 Momentum and acceleration: comment dynamics

Next we consider the question of how commenting rates evolve over time in the Polymath and mini-polymath projects. To explore this process we draw on two measures from physics for quantifying motion: acceleration and momentum. We define them formally below, but broadly speaking, *acceleration* captures whether authors are commenting on the project at a constant rate, an increasing rate, or a decreasing rate; *momentum* captures the overall rate at which the progress is advancing, considering both the rate at which authors are creating new comments, and also the amount of content that they are producing in those comments.

Definitions. Let us refer to the current "position" of the project as $x(t_i)$, where $x(t_i)$ is the number of comments that have been made up to time t_i. Then the project's instantaneous velocity and acceleration are the first and second time derivatives of $x(t)$, which can be measured using the central difference formula: $v(t_i) = x'(t_i) \approx \frac{x(t_{i+1}) - x(t_{i-1})}{t_{i+1} - t_{i-1}}$, and similarly for $a(t_i) = v'(t_i)$. We compute the average velocity with units of comments per minute, providing a summary measure of how rapidly each project progressed. The average acceleration then has units of comments per minute per minute, and tells us whether or not the speed of the project was picking up (positive acceleration) or slowing down (negative acceleration).

Finally, we introduce the notion of a comment's *momentum*: borrowing from physics, the momentum of an object is the product of its mass and its velocity. We interpret the number of characters in a comment as its mass and so compute the momentum as the product of a comment's length and its velocity. This notion of momentum enables us to distinguish between projects with, for example, the same commenting *rate* but with different average comment *lengths*.

High-momentum projects pick up more speed. Surprisingly, in Figure 3B we find that all Polymath and mini-polymath projects have a positive average acceleration. Earlier we observed that comment response times were on average faster in mini-polymath than in Polymath; we also observe that they tend to have higher acceleration

Perhaps most strikingly, in Figure 3 (bottom figure), we see that the average acceleration and momentum in this case have an approximately monotonic relationship with each other, meaning that the projects with the highest momentum were also the projects that were picking up the most speed. This monotonic relationship is not something to be expected a priori: for example, a project that started off with long, rapid comments and slowly decayed would have high average velocity and negative acceleration; but all of the examples observed here have the opposite pattern, with the higher momentum projects accelerating more rapidly.

5. LINGUISTIC FEATURES

Following the plan outlined in the introduction, we continue by studying the distinctions between Polymath projects — representing research on open problems — and mini-polymath projects, which are efforts to solve Math Olympiad problems. This investigation offers the opportunity to understand the contrasts between these related but qualitatively different types of collaborative activities. In this section, we introduce the high-level linguistic features that we consider and the differences observed in how they manifest in the two domains.

5.1 Exploring high-level linguistic features

Our set of high-level linguistic features draws on recent innovations in natural language processing that have been used for applications including the memorability of movie quotes [3], the effects of wording on message propagation [17] and the popularity of online posts [13]. We supplement these features with several more basic ones as well.

We divide the features into four groups: relevance, distinctiveness, politeness and generality. To get an initial understanding of how these features differ between Polymath and mini-polymath projects, for each one we conduct a t-test between feature values extracted from Polymath comments and mini-polymath comments (Table 4). We find that Polymath comments are indeed significantly different in many of these features compared to mini-polymath comments. Later in §6, we will see how they perform in a prediction setting in comparison to topic-based linguistic features, as well as the role- and temporal-based features discussed in §3 and §4.

Table 4: T-test results for high-level linguistic features. For each feature, we conduct a t-test from two independent samples, extracted from Polymath comments and mini-polymath comments respectively, where the null hypothesis is that the two kinds of comments come from the same distribution. The number of arrows in the table visually indicates the p-value magnitude: $p < 0.05$: 1 arrow, $p < 0.01$: 2 arrows, $p < 0.001$: 3 arrows, $p < 0.0001$: 4 arrows. ↑ indicates that Polymath comments have larger values; ↓ indicates that mini-polymath comments have larger values.

Feature	test results
Relevance	
similarity to original post	↑↑↑↑
similarity to current post	↑↑↑↑
Distinctiveness	
average log POS unigram prob	↑↑↑↑
average log POS bigram prob	-
average log POS trigram prob	↑
average log lexical unigram prob	-
average log lexical bigram prob	↑↑↑↑
average log lexical trigram prob	-
Politeness	
politeness [4]	↑↑↑↑
number of hedges	↑↑↑↑
fraction of words that are hedges	↓
Generality	
frac. indefinite articles	↓↓↓↓
frac. past tense	↑↑↑↑
frac. present tense	-

We begin by describing the feature-level differences between Polymath and mini-polymath comments. For each category of differences, we summarize it first in a bold-faced sentence and then elaborate in the subsequent paragraph.

Research discussions match the original problems more closely. We first ask how much the language used in the discussion drifts away from the language used at the outset of the project to describe the problem. We do this by computing Jaccard similarity between each post and the original post for the project. Since the discussions are segmented up into *threads* of roughly 100 consecutive posts each, we also compute a related measure — the Jaccard similarity between each post and the initial post in the thread it belongs to.

One might expect that since research discussions are open-ended, the language might drift quickly away from the description of the initial problem. In fact, we find that posts are significantly more similar to the original posts for Polymath projects, both in the discussion and in the thread.

Research discussions have *less* distinctive language. One might expect the language in tackling hard research problems to be more "distinct" from daily language compared to that in solving problems with known solutions. We formalize distinctiveness using language model scores, defined as the average logarithm of word probabilities [3, 13, 17]. Our language model, based on frequencies of one, two, and three word sequences (unigrams, bigrams, and trigrams) of words and part-of-speech tags, is developed from the Brown corpus [12].

Perhaps somewhat surprisingly, research discussions resemble daily language more in terms of part-of-speech tag patterns. When it comes to actual words, research discussions also employ more common word patterns, although it is not statistically significant for unigrams and trigrams. The greater robustness of the part-of-speech analysis, in comparison to the word-level analysis, may reflect the fact that both projects contain a large amount of language infrequently used outside of mathematical discussions.

Research discussions are more polite. As participants are discussing harder problems for a longer period of time in Polymath projects, a natural hypothesis is that they are more polite to one another. We test this using a recently developed method for estimating the politeness of pieces of text [4], and we find that indeed there is significantly more politeness in the text of the Polymath projects.

We obtain an inconclusive comparison when we study the related phenomenon of *hedging* in the language use of the posts — a term coined by Lakoff [14] to describe the expression of uncertainty, which would be natural to have in posts discussing hard technical problems. Although Polymath comments have significantly more hedges, mini-polymath comments have a larger fraction of hedges.

Research discussions are more "specific". One hypothesis may be that we can see a difference in how general the arguments are, i.e., mini-polymath may be more specific due to the limited scope of the problem. Previous work has used the occurrences of indefinite articles and tense-related expressions to capture generality [3, 17]. Somewhat surprisingly, Polymath comments are less general, with significantly more past tense and fewer indefinite articles.

6. PREDICTING DOMAIN: RESEARCH VS. HARD PROBLEM SOLVING

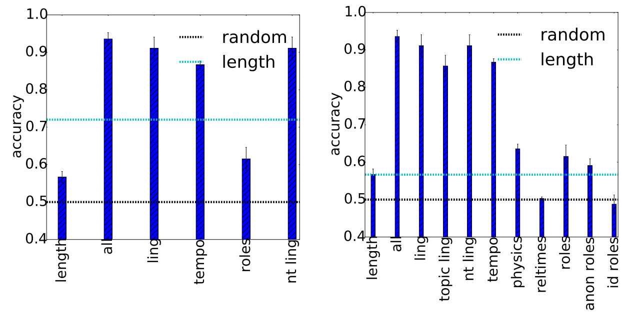
We now have a broad set of features characterizing the posts and can leverage them to use in our basic prediction problem. Our model uses these features to determine whether a given post comes from a Polymath project or a mini-polymath project.

The features discussed above fall into three categories: author roles, temporal dynamics, and linguistics. We focus on the performance of each set in turn. The author roles can be further distinguished by whether they are being used anonymously (omitting author identities) or non-anonymously; the temporal by whether they are simple elapsed time differences or more nuanced dynamics metrics such as acceleration and momentum; and the linguistic properties by whether they have topic information or non-topic information.

Surprisingly, we will find that in a controlled setting, prediction using these anonymous structural and non-topical features can actually outperform topic-based and identity-based features. We also find that the dynamics metrics (drawn from physics) offer better prediction performance than the simpler, elapsed-time metrics.

Prediction setup. We set up balanced prediction tasks for distinguishing Polymath comments from mini-polymath comments. Specifically, as there are fewer comments in the mini-polymath projects, we sample a Polymath comment for each mini-polymath comment. Thus we have a pair of comments in each instance of our data, with one comment from each of Polymath and mini-polymath respectively. (We randomly order these two comments when presenting them to the algorithm.) We use two different ways of sampling pairs from the overall data.

- Random (704 pairs). For each mini-polymath comment, we randomly sample a comment from the Polymath projects.
- Controlled (203 pairs). For each mini-polymath comment, we find the Polymath comment from the same author with the minimum length difference in terms of the number of words. (We only use mini-polymath comments for which the same author has written at least one Polymath comment.) This constructs a much more difficult prediction task.



(a) Overview with no additional controls

(b) Breakdown with author and length controls

Figure 4: Results for predicting polymath comments vs. minipolymath comments. x-axis: different feature sets, each group is defined in §6.1. y-axis: accuracy. Error-bars represents the standard error of the performance across 5 folds. The black line shows the performance of random guessing; the cyan line shows the performance of the length baseline. (a) and (b) respectively show the performance without any control and when we control author and length. (b) shows the more detailed performance breakdown for each of the three elements (roles: anon. vs. non-anon, timing: elapsed times vs. physics, linguistics: topic-based vs. non-topic based).

6.1 Feature definitions and motivations

We now discuss the features we use for the prediction task, drawing on the features defined above. Our plan is to compare the prediction performance using different sets of these features.

The features can be categorized as follows; the keyword in parentheses preceding each definition indicates the feature category as labeled in the performance results plots (Figures 4 and 5).

- Length. Given that comments in mini-polymath projects are generally shorter than the comments in Polymath projects, the length of a comment already provides a non-trivial baseline for prediction. Our notion of length actually includes three quantities for each post: the number of words, the number of characters, and the number of MathJax characters as features.
- Roles
  - (id roles) Author and surrounding authors: numeric id of comment author and those authors of the ten comments leading up to it and the ten succeeding it;
  - (anon roles) Anonymous structural: same as *id roles* but with generic structural representation of the author sequence;
- Temporal
  - (reltimes) Elapsed times: hours, days, and minutes elapsed since project inception; number of comments and number of threads since project inception;
  - (physics) Dynamic properties: instantaneous velocity, acceleration, and momentum of comment, where position is defined as comment id, and mass of a comment is defined as the number of characters in it. These features are defined formally in §4.2.
- Linguistic features. The linguistic features consist of *non-topical features* (denoted "nt-ling") listed in the first four bullet points, and *topical features* (denoted "topic ling") listed in the latter two bullet points.

– (nt ling) High-level linguistic features, as discussed in §5.1: politeness, generality, specificity, hedging, fraction of *novel* words with respect to the entire preceding conversation or to a fixed-size window of previous comments.

– (nt ling) LIWC. Linguistic Inquiry and Word Count (LIWC) includes a dictionary of words classified into different categories, along dimensions that include affective and cognitive properties [16]. We use the frequency of each LIWC category in a comment as features.

– (nt ling) Part-of-speech tags (POS). Part-of-speech tags can provide us with stylistic information for a comment. All possible part-of-speech tags are considered as features.[5].

– (nt ling) Stopwords from the NLTK[6]; most frequent 50 words from the training data; most frequent 100 words from training data.

– (topic ling) Bag-of-words (BOW). This is a very strong method typically used in natural language processing tasks. We include all the unigrams that occur at least 5 times in our training data as features. We use the tokenizer from the NLTK package after replacing urls and MathJax scripts with special tokens.

– (topic ling) Bag-of-words for the preceding and succeeding comments. The same definition as the feature above, but now for each of the five comments before the comment in question, and each of the five after.

Computational evaluation of prediction. We use 5-fold cross-validation in our computations to measure prediction performance. Since the task is balanced, we use accuracy as our evaluation metric. In the computations, for each feature set, we extract the values from each comment in a pair, and then take the differences between the first comment and the second comment in this pair. For BOW and POS based features, we normalize the feature vectors using L2-norms, while for the other features, the values are linearly scaled to $[0, 1]$ based on training data. We use scikit-learn in all prediction computations.

Prediction: Roles, Temporal. In Figure 4 we observe that using the anonymized roles (author motifs as discussed in §3.2) offers good performance as model features. This positive performance may be due to the distinctions we observed above. In particular, the Polymath projects tend to have larger and significant correlations in the reply structure of the comment threads.

We also observe that the temporal features offer significant improvements over the random baseline. As with the role features, this performance increase can potentially be understood as thanks to the substantial differences in temporal dynamics in the two projects that we discussed in §4.

Linguistic prediction performance: topical vs. non-topical. We make several observations about the prediction results based on linguistic-only features. First, all the feature sets improve on the length baseline for both the uncontrolled task (when we form a pair for each mini-polymath comment) and the controlled task (when we match the author and approximately match the length within each pair).

Second, the bag-of-words feature set slightly outperforms the non-topic feature set on the uncontrolled task, but when we add length and author controls, in fact the non-topic feature set significantly outperforms the bag-of-words features, achieving close to

Table 5: Top 20 features in main- vs. mini-polymath prediction. Features are separated by spaces. High-level linguistic features are in quotes. Other non-topical features are named by concatenating the category name and feature name; for instance, "POS-adj" means the feature "adjectives" from the part-of-speech category.

Top bag-of-word features	
Polymath	sequences " is sequence primes prime - now values at " in different of by 3 also latex paper x
mini-polymath	m then can points ... mine number mines point n coins proof moves comments added all any partial thread 2
Top non-topical features	
Polymath	"similarity to original post" "similarity to current post" POS-adjective POS-adverb "POS-verb (past)" POS-" "frac. past tense" POS-preposition liwc-work POS-noun numchars liwc-adverb liwc-auxverb nummathchars liwc-preps "POS-verb (non-3rd present)" liwc-they POS-: liwc-time "average log unigram prob (lexical)"
mini-polymath	liwc-motion liwc-assent liwc-we liwc-certain liwc-cause liwc-negemo liwc-achieve "frac. indefinite articles" liwc-filler liwc-conj liwc-nonfl liwc-quant liwc-number POS-NONE "POS-adjective (superlative)" "POS-verb (base form)" POS-$ "POS-proper noun (singular)" POS-determiner POS-particle

90% accuracy. It is interesting that the non-topic feature set should achieve this, since it is not attuned to the content of the posts themselves. Moreover, the non-topic features actually give better performance on the controlled task than on the uncontrolled task, despite the fact that the controlled task was set up to limit the effectiveness of various features; meanwhile, the performance of the bag-of-words feature set in the controlled task (along with stopwords and POS) drops significantly.

As for individual categories, high-level linguistic features actually outperform all other non-topical categories despite the small number of features in this category, including commonly used LIWC features. This observation is robust across both tasks. It is worth noting that there are fewer high-level linguistic features than POS or LIWC features.

In terms of top features (Table 5), similarity to the original problem statement is the most prominent signal for Polymath comments, followed by part-of-speech tags including adjectives; in contrast, LIWC categories and part-of-speech tags tend to be top indicators of mini-polymath comments. Table 5 also shows the top word-level features that emerged for the bag-of-words feature set, including topical words such as "sequence", "prime" and "mine"[7].

7. IDENTIFYING RESEARCH HIGHLIGHTS: INTRINSIC VS. CONTEXTUAL EVIDENCE

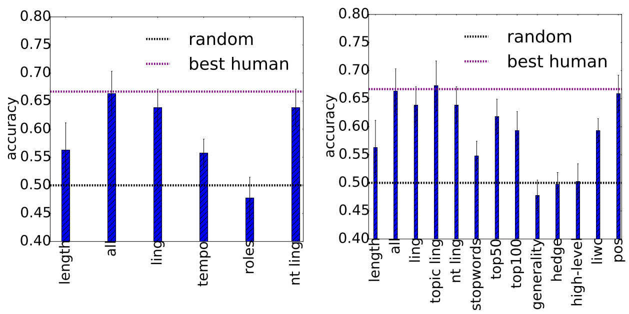
We now investigate the second main question we posed in the introduction: Are research breakthroughs identifiable in a string of comments? If they are, can one best recognize them solely from their content, a finding that could indicate that authors know the eventual importance of their statements? Or are breakthroughs best recognized by the (re-)actions of others, suggesting that it can be hard to know in the heat of the moment which results are key ones?

Polymath 1 serves as a particularly nice setting for investigating this question because, fortunately, breakthroughs have already been identified by a domain expert: Terence Tao set up a wiki timeline of

[5]Througout we use the NLTK maximum entropy tagger with default parameters, which is based on the Penn Treebank Dataset (`http://www.cis.upenn.edu/~treebank/home.html`)

[6]`http://www.nltk.org/`

[7]"Mine," in the sense of an explosive device, occurred in one problem in IMO.



(a) Overview; author and length controls

(b) Linguistic breakdown; author and length controls

Figure 5: Highlight-prediction results for different feature sets. Error bars: standard error for 5-fold cross-validation. Black line: random guessing. Purple line: best human performance on the author-control-only (easier) task.

Polymath 1 highlights.[8] While Cranshaw and Kittur [2] employed this highlights list to study *whether less active users* had impact, we use the list to constitute instances for the task of classifying *which comments* have impact, and to identify the most helpful intrinsic vs. extrinsic features for this task.

Prediction setup. In setting up the prediction task, we employed two paradigms: (A) classifying individual instances as being either a highlight or not, or (B) choosing one comment from a pair where it is known that exactly one was a highlight, and the other is the non-highlight written by the *same* author that is closest in length to the highlight. Due to space constraints, we only describe (B) in this paper, for three reasons. First, author- and length-controlled findings are more likely to generalize to other settings. Second, we believe that for judges (human or algorithmic) that are not domain experts, it is more reasonable to be asked to pick the more important-looking comment in a pair than to judge a single text in isolation. Third, describing (B) allows us to be more concise than describing (A), where mechanisms for handling class imbalance would need to be explained. We note, though, that (B) gives us less data to work with (since we can only construct as many pairs as there are highlights), and doesn't directly map onto the application of classifying individual comments as they naturally appear.

Feature sets and cross validation. For this task, we use the same feature sets that we employed for our first task of distingishing Polymath comments from mini-polymath comments. These features are described in §6. Further, in all experiments, we employ the same experiment protocol as in the first task. Random guessing yields a baseline accuracy of 50%.

7.1 Prediction Performance

To assess the difficulty of the task for humans as a point of comparison for our algorithmic performance, we asked three Applied Mathematics graduate students to attempt the classification task on 30 author-controlled pairs in an approximately 30-minute session.[9] They got accuracies of 66.7%, 63% and 46% (agreeing 60% of the time); these results, together with post-hoc feedback from the students, indicates that our task is fairly difficult.

Prediction performance: roles, temporal. In Figure 5 we observe that the features based on the authors' roles in the project and those based on temporal properties do not offer additional prediction performance beyond what can be achieved by a random baseline.

Prediction performance: topical vs. non-topical linguistic. Meanwhile, the linguistic features perform 15% above the random baseline, and in fact achieve the same performance as that of the human evaluators. It is interesting to note that the text was all that was available to the human judges, just as it was precisely the words and high-level linguistic features and parts of speech that were available to the model we trained. We continue by further exploring the performance of these varying groups of linguistic features.

Figure 5 shows that the best-performing classifier uses bag-of-words features and yields accuracy comparable to that of our best human subject (on a slightly simpler task). The second best performing feature set is part-of-speech tags. Adding other non-topical features actual hurts the performance slightly in this task. Neither preceding comments or reactions outperforms comment-internal features. All in all, the evidence suggests that authors often write texts that eventually turn out to be highlights in a fashion that indicates they may be aware of the importance of their remarks at the time.

8. RELATED WORK

Shortly after Polymath 1's success Tim Gowers and Michael Nielsen wrote a retrospective opinion piece on open collaboration in Nature [8], in which they took the opportunity to share their vision for the incredible potential that the Web offers to the future of science, as a collaborative tool that is ideal for facilitating communication and information sharing.

Michael Barany [1] wrote about Polymath from a qualitative sociological perspective, focusing on the interaction of the participants with the technological system that supported the collaboration. In particular, he considers the mutual adaptation of that technological system, the participants, and the overall collaboration as the project advanced from its uncertain beginning to a successful conclusion.

In addition to Barany's piece, Cranshaw and Kittur [2] provide a quantitative overview of Polymath 1. They find that activity tends to spur other activity, and that activity by either of the two leaders, Terence Tao and Tim Gowers, tends to spur even more activity. They observed that the numbering-threading convention was successful in allowing multiple threads to develop simultaneously, but that cross-references were limited. By constructing the comment mention-graph and cross-referencing authors' Wordpress profiles with Google Scholar accounts they were able to show that, while the top two contributors were the Fields Medalists, there were much "smaller names" close behind – indicating that Gowers's vision of the project being accessible to a broad audience was achieved at least in part.

Finally, mini-polymath has been studied by Pease and Martin [15]; they show how the approaches there follow well-studied frameworks for problem-solving.

9. CONCLUSION

Polymath is an interesting experiment in promoting Internet collaboration on a type of activity — working on open mathematical research problems — that is otherwise not really represented in large open online collaborative efforts. Using this site as a lens, we have sought to contrast Internet collaborations on open research

[8] `http://michaelnielsen.org/polymath1/index.php?title=Timeline`. The page's revision history reveals that Tao's intent was indeed to list "highlights", but he switched to the milder term "events" to alleviate Gowers' apparent embarrassment at one of his contributions being deemed "highlight-worthy".
[9] At the time, we had not installed the length controls, but the task is strictly easier when length is a potential clue.

problems with Internet collaborations on "merely" difficult problems.

Limitations While Polymath is the most visible effort at open Internet collaboration on mathematical research problems, one should be careful about generalizing too far from a single domain. Moreover, we can ask whether there are specific aspects of Polymath that played a role in the findings. Perhaps most importantly, the participation guidelines of the main Polymaths promoted rapid, incremental posting over the arguably more typical research mode wherein one engages in longer periods of off-line reflection and independent thought. The (laudable) intent was to make the project more accessible, but it is possible that the collaboration was less natural as a result. Regardless of these concerns, of course, it is clear that several projects had sucessful outcomes, resulting in publications and/or important partial progress toward the stated goal.

Future Directions Many of our findings open up promising future directions. First, the reply-time properties are interesting, with the intriguing fact that Polymath, which is significantly slower than Mini-Polymath overall, becomes faster at the shortest time scales. We would like to understand the reason for this fast pace; it is also natural to ask whether this "organically" developed fast pace is good for collaborations, or whether it is more effective to proceed more slowly at the shortest time scales. It is also interesting to ask whether we can trace any potential effects that the high-level linguistic properties have on the trajectory of the discussion or the quality of the outcome.

Finally, our second prediction task, on identifying highlights in real time, raises potential questions for the design of future iterations of Polymath-style sites. If it were possible to flag predicted highlights as they happen, is this a useful thing to make explicit for a group engaged in research? And if so, is it more productive to call attention to these predicted highlights as they happen, or at a later point? Questions in this style point to the potential opportunities for algorithms trained on this type of data to assist in guiding future discussions, when on-line groups assemble to work on hard problems together.

Acknowledgments. We thank the students that took the course CS6742 (Fall 2014) at Cornell and the anonymous reviewers for helpful comments. We also particularly thank John Chavis, Zachary Clawson, Stephen Cowpar, David Eriksson, Hyung Joo Park, and Evan Randles for serving as judges in the prediction tasks. This work was supported in part by NSF grant IIS-0910664, a Simons Investigator Award, a Google Research Grant, a Facebook fellowship, and a NSF Graduate Research Fellowship.

References

[1] M. J. Barany. '[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 *Proc. Symp. on Wikis and Open Collaboration*, pages 10:1–10:9, 2010.

[2] J. Cranshaw and A. Kittur. The Polymath project: Lessons from a successful online collaboration in mathematics. In *Proc. CHI*, pages 1865–1874, 2011.

[3] C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee. You Had Me at Hello: How Phrasing Affects Memorability. In *Proc. ACL*, 2012.

[4] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A Computational Approach to Politeness with Application to Social Factors. In *ACL*, 2013.

[5] J. DiNardo. Natural experiments and quasi-natural experiments. In *The New Palgrave Dictionary of Econ.* Palgrave Macmillan, 2008.

[6] W. Glänzel and A. Schubert. Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research*, pages 257–276. 2005. ISBN 978-1-4020-2702-4. doi: 10.1007/1-4020-2755-9_12.

[7] T. Gowers. Is massively collaborative mathematics possible?, Jan. 2009. URL `https://gowers.wordpress.com/2009/01/27/is-massively-collaborative-mathematics-possible/`.

[8] T. Gowers and M. Nielsen. Massively collaborative mathematics. *Nature*, 461(7266):879–881, 2009.

[9] D. B. Horn, T. A. Finholt, J. P. Birnholtz, D. Motwani, and S. Jayaraman. Six Degrees of Jonathan Grudin: A Social Network Analysis of the Evolution and Impact of CSCW Research. In *Proc. CSCW*, 2004.

[10] B. F. Jones, S. Wuchty, and B. Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[11] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proc. CSCW*, pages 37–46, 2008.

[12] H. Kučera and N. Francis. *Computational analysis of present-day American English*. Brown University Press, 1967.

[13] H. Lakkaraju, J. McAuley, and J. Leskovec. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *ICWSM*, 2013.

[14] G. Lakoff. Hedges: a study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2, 1975.

[15] A. Pease and U. Martin. Seventy four minutes of mathematics: An analysis of the third Mini-Polymath project. In *Proc. Artificial Intell. Simulation of Behavior*, 2012.

[16] J. W. Pennebaker, M. E. Francis, and R. J. Booth. *Linguistic inquiry and word count: LIWC 2007*, 2007.

[17] C. Tan, L. Lee, and B. Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proc. ACL*, 2014.

[18] T. Tao. The erdos discrepancy problem. *arXiv preprint arXiv:1509.05363*, 2015.

[19] Y. Yamauchi, M. Yokozawa, T. Shinohara, and T. Ishida. Collaboration with lean media: How open-source software succeeds. In *Proc. CSCW*, pages 329–338, New York, NY, USA, 2000. ACM. ISBN 1-58113-222-0.