Citation and Attribution in Open Science: A Case Study

Yla R. Tausczik
iSchool
University of Maryland,
College Park
ylatau@umd.edu

ABSTRACT
Technology is changing the way in which scientists and mathematicians communicate. New platforms for scholarly communication enhance informal communication and imbue it with some of the functionalities of formal communication. A citation analysis was conducted to examine how content from one of these platforms, MathOverflow, has been cited and referenced within the mathematics literature. Citation patterns suggested that some authors were treating MathOverflow content as a legitimate source of scholarly knowledge. Some problems with references and attribution occurred, which can be addressed by changes in the design of these technologies. Developing better systems for scholarly communication can help advance open science.

Author Keywords
open science; bibliometrics; large-scale collaboration; scientific publishing; scholarly communication

ACM Classification Keywords
H.5.3. Information Interfaces and Presentation (e.g. HCI): Group and Organization Interfaces

INTRODUCTION
From as early as 1990, academics believed that the internet and web technologies would change scholarly communication [17]. Scholarly communication, the means by which academics disseminate knowledge and discuss ideas, can be categorized into two classes: formal and informal communication. Formal communication is exemplified by peer-reviewed journal publications while informal communication includes casual conversations and the exchange of early drafts and unpublished manuscripts.

Peer-reviewed journal publications[1] have been the standard accepted form of formal communication for the last 350 years [27]. Journal publications serve four main functions: archiving, registration, dissemination, and certification [31, 29]. Journals archive knowledge by preserving a permanent record of articles. These articles can in principle, be retrieved at any time. Journals also serve to register the order in which ideas are announced, providing canonical date stamps in the form of articles. Journals disseminate knowledge through publication to their subscribers. The peer review mechanism provides legitimacy by requiring the approval of a small number of scholars, whose approval is typically accepted by the larger community.

Through informal communication, scholars are able to share new results quickly and get feedback on ideas. Participating in informal communication is important for both disseminators and recipients. For recipients it is a way to access intermediate results, access new results more quickly and discuss implicit knowledge not included in formal communication. Scholars in many disciplines report seeking out information through informal communication to support their research activities [7]. Historically, the importance of informal communication varies among disciplines, and is particularly important in certain fields like economics and mathematics [7, 15].

In the quarter-century since 1990, the internet has indeed changed scholarly communication [21], particularly in the area of informal communication. Personal communication that might have taken place in person can take place on listservs, blogs and through social media [15, 29]. Manuscripts and preprints are being deposited and stored in repositories, such as arXiv [21]. These forms of communication reach a larger audience faster and with less effort than previous forms of informal communication. As a result there has been an expansion in scholarly informal communication, with more and richer information shared through a wider class of media [27].

The affordances of new communication technologies have led to an overlap in the functionality of informal and formal communication [29]. For example, pre-internet manuscripts were exchanged directly between scholars without a permanent record or collection. Now repositories, such as arXiv, collect manuscripts in one place and store manuscripts versions with unique ids. These repositories serve to archive, register, and disseminate knowledge by storing articles in a way that can be retrieved later, creating a primacy record for ideas, and distributing articles through mailing lists and search accessibility. Other communication technologies, including online discussion forums, serve similar functionality [15].

[1]For simplicity, by journal publications I am also including publications in journal-like venues, such as conference proceedings and electronic journals

CSCW'16, February 27-March 02, 2016, San Francisco, CA, USA
 ACM 978-1-4503-3592-8/16/02 $15.00
DOI: http://dx.doi.org/10.1145/2818048.2820070

The increasing functionality of informal communication and impediments in the journal system, such as slow publication rate, have led to an increase in legitimacy of informal communication. In economics, citations of preprints have risen to 14% as of 2007 [1]. Web sources, including preprint repositories and online discussion forums, have become a larger part of works cited in many disciplines [9]. Change in scholarly communication is slow and incremental; while informal communication has not usurped formal communication, it is gaining recognition.

Web technologies are also enabling radically new forms of informal communication. Academics are using web technologies to create tools for metascience to address widespread methodological problems through collective action [30]. Psychfiledrawer.org is a place to document and aggregate failed replication attempts which often fail to appear in the literature. Open Science Framework is a platform to document studies and to hold researchers accountable for scientific practices [32]. Academics are also using web technologies to engage in larger-scale collaborations. Polymath project 1 created collaboration on mathematics proof at a scale not seen to date [12]. On MathOverflow, a question and answer platform, mathematicians engage in *ad hoc* collaboration on micro-research questions every day [33].

These projects are a part of a larger movement toward open science. There is a widespread belief, with some empirical support, that openness in science will accelerate scientific progress [6]. Through disclosure of intermediate results, open science allows for the cumulative generation of knowledge through smaller, more rapid contributions. Use of web technologies in efforts like polymath and MathOverflow create functionality beyond what is typical for informal scholarly communication. These forms of communication are more generative, with new knowledge built through interaction [33]. Often, individual contributions are small and collaborations are large, allowing for division of labor in which individuals can take on specialized roles.

Addressing two broad questions would help to evaluate the impact of these new technologies on scholarly communication.

1. To what extent are new forms of communication being used in practice to replace important aspects of the journal system?

2. To what extent are new forms of communication being used to extend the functionality of the journal systems?

As scientists, theoreticians, and technologists grapple with the design of new technologies for scholarly communication, one important goal is to support greater openness and sharing in science. One avenue for such development leverages changes in information and communication technology (ICT) which enhance informal communication, imbuing it with some of the functions of formal communication. In this way, sharing of partial, unpublished ideas becomes easier and gains some of the benefits of sharing published ideas. In this paper I examine a community of mathematicians which has coopted a Q&A platform to study research-level mathematics questions, and we quantify the way that this community interacts with and extends more traditional forms of formal communication. This work contributes to an understanding of whether and how scholars make use of new informal communication technologies, and our results can inform the design of future platforms which support open science.

RELATED WORK

Communication and collaboration among scientists, and the technology which supports it, has been a major focus of CSCW research. Changes in science are driven by many factors, such as funding and an emphasis on interdisciplinary research, and these have led to systematic shifts in the nature of collaboration. For example, the importance of teams in science has increased over time [38], and this has been accompanied by other shifts such as collaborations which span larger geographic distances, more institutions, and a greater variety of disciplines [16, 11].

The availability of new forms of technology has also led to changes in scientific collaboration. Collaboratories leverage internet technologies to create virtual labs in which large groups of researchers share data and computing resources [14]. Citizen science projects use technology to engage the general public in science. Ubiquitous computing and mobile technology support virtual organizations which allow citizens to collect of data over wider geographic areas and longer time periods [37]. Crowdsourcing techniques allow both the general public and experts to pool resources in the interest of solving scientific problems. Several of these projects have led to scientific progress, including FoldIt [10] and Galaxy Zoo [8]. Notably, mathematics is one area in which large-scale collaboration through online technologies has led to major discoveries and may be especially fruitful [12].

Despite the increased emphasis on and potential for interaction, researchers have found that deep collaboration is hard to cultivate and support. Integrative collaboration is rare even among groups which are funded as a unit [2]. Moreover, collaboration is often impeded by social, cultural, geographic and institutional barriers (e.g. [25, 13]). Some authors have emphasized the importance of human aspects of organization within cyberinfrastructure [23]. The effective design of collaborative technology requires an understanding of the sociotechnical aspects of its use.

One particular research focus in this area has been the design of technology to encourage sharing and reuse within science. Most often this research has focused on sharing of data, but it may also include sharing of other scientific knowledge and resources, such as specimens, software, materials, protocols and unpublished ideas [36]. Velden [35] identifies an inherent tension between sharing and secrecy in which the former allows for greater cooperation while the latter ensures credit for individual ideas.

The trade-offs between these two motivations varies by field, and incentives play a major role in determining sharing behavior. A misaligned incentive structure can hamper effective collaboration, as was found in studies of scientific soft-

ware production [19]. Birnholtz and Bietz [5] suggest that careful engineering of systems to support data sharing may bypass some of the risks associated with sharing, for example by having scientists share data abstractions rather than datasets themselves. In mathematics, sharing is somewhat less problematic because, compared to other fields, mathematics research does not depend as heavily on supplemental resources such as data, software, or protocols. In addition, there is less of a financial motive for secrecy, as mathematical ideas cannot be patented. Therefore, in mathematics there is a greater emphasis on sharing knowledge and unpublished findings [12, 33].

CURRENT STUDY

MathOverflow (MO) is a question and answer website used for the discussion and generation of mathematical knowledge by academic mathematicians [34, 33]. MO is a member of the Stack Exchange community, which contains many other popular Q&A sites, including Stack Overflow.

Affordances of formal communication

Communication on MO shares many of the functions of formal communication in journals, including archiving, registration, dissemination, and certification. Question and answer pages are stored in a database, displayed on the website mathoverflow.net, and archived on the Internet Archive by the Stack Exchange company. Content can easily be retrieved by anyone using the unique ID given to a Q&A page or unique ID given to an individual contribution on a Q&A page, such as an answer or a comment. Every contribution is dated; this creates a record of the date and time when an individual first submitted an idea. Content is made available online, is returned by search engines, and is emailed to subscribers of topic areas. Content is lightly moderated and peer-reviewed by readers through comments and votes.

As informal communication gains many of the functions of formal communication it begins to take on some of the its roles as well. This is evident in the increasing prominence of preprints. Because preprint repositories have mimicked many of the functions of journals (i.e. archiving, registration, dissemination) preprints can be treated as journal articles. Some preprints have achieved a large impact without ever being published [22] and preprints in general are becoming a larger fraction of published citations in some disciplines [1]. However, academia is slow to adapt and change takes widespread community acceptance. A record of when an idea is first "published" is meaningless unless a community agrees on what counts as being published [21]. Any form of peer review is valid only to the extent that a community accepts it as a valid certification.

Of course, community acceptance of MO as legitimate scholarly discourse is not an all or nothing consideration. There will inevitably be some early adopters. If these early adopters are experienced mathematicians, more central to the professional community, this may help MO to gain legitimacy. Similarly, if early adopters cite MO in high-impact, mainstream journals, this will add to its credibility. On the other hand, if MO citations are associated with inexperienced mathematicians and/or lower-tier journals the path toward legitimacy may be longer and perhaps unsuccessful.

My first set of questions concerns the degree to which MO content is being treated as formal communication in practice. Each question addresses the practical adoption of the functions of journals. Legitimacy results from dissemination of content and acceptance of certification. Retrieval results from adoption of the archival system. Attribution results from recognition of MO as registering ideas to specific people.

Research Question Set 1: Is MathOverflow content being treated as formal communication?

a) Are authors citing MO contributions in the literature (Legitimacy)?

b) Are authors providing complete references for citations to MO (Retrieval)?

c) Do references credit individual MO authors for their contributions (Attribution)?

Affordances of informal communication

MathOverflow has affordances that create functionality not present in the journal system. On MO, individuals can solicit the content they want by posting questions. This functionality is more similar to traditional forms of informal communication, such as asking an expert in a field. The main difference is that, on MO, content can be solicited from many more people at once, with little social capital and little effort. These discussions can be more interactive and involve a larger number of people. Individuals may make very small contributions to these discussions, such as a single answer or a comment. These small individual contributions by many authors can build on each other to generate a larger contribution [33].

It is unclear whether journal authors are using MO content as a complement to their principal arguments or in a more central role. For example, authors may use MO content in the introduction, related work, and discussion sections to frame and discuss the problem at hand by providing motivation for the problem, background information, related problems, and/or examples. Alternatively, authors may use content as part of the main arguments of the journal article and/or to support specific claims made in these main arguments. The former use suggests MO content is strengthening the journal article, but not in an essential way, whereas the latter would suggest that MO content is essential to the main results and that MO could become an critical tool in the generation of publishable results.

My second set of questions concern the way in which MO content is being cited. If MO is being used to generate knowledge to be used in formal communication then this will be reflected in citations. If journal article authors recognize these small contributions as stand alone contributions then they will appear as individual citations in their papers.

Research Question Set 2: Do citations reflect the unique affordances of MathOverflow?

a) Are authors of the journal articles prompting the generation of the content being cited (Generativity)?

b) Are small contributions on MO being cited (Level of Granularity)?

Interaction between formal and informal communication

Career advancement in academia relies on receiving credit for one's work. One challenge to open science is incentivizing participation when there are no individual rewards for sharing intermediate results [24]. One way to address this is to give credit for work in aggregate. Large collaborations, like those on MathOverflow, could be credited as a unit (e.g. discussion by X, Y, Z). This is how authorship was handled for the first polymath project [26]. The work of all 39 contributors was credited under a single pseudonym DHJ Polymath. However, when equal credit is given to a large number of authors, it can be difficult for individuals to receive the recognition they want and difficult for outsiders to judge who did most of the work [4]. Alternatively, credit can be given to each small contribution. This has the advantage that it is clear who contributed what. To date there is no evidence that MO contributions are being used as part of hiring, tenure or promotion decisions. If one day these contributions do reflect on measures of scholarly productivity, such as impact factor, the distribution of credit in these collaborative endeavors will be an important consideration.

The third and final set of questions concerns the interaction between the traditional functions of formal communication and the new forms of communication occurring on MO. If authors are going to receive credit for small contributions on MathOverflow, then references must be at the appropriate level of granularity (i.e. a citation for an answer should provide archival information for an answer) and references ought to specifically name individuals who have made these small contributions.

Research Question Set 3: How is credit for MathOverflow content being assigned?

a) Are authors providing complete references for citations to MO at the appropriate level of granularity (Retrieval X Level of Granularity)?

b) Do references credit individual MO authors for their contributions even when those contributions are small (Attribution X Level of Granularity)?

These three sets of questions were addressed by collecting and analyzing all MO citations in published journal articles up to January 2015.

METHOD

Data Collection - Peer-reviewed articles

Google Scholar (scholar.google.com) was used to gather a collection of published articles that cited MathOverflow. A search for "MathOverflow" returned 1,670 results. Google displayed the first 1,000 entries ranked as the most relevant. Each entry was manually reviewed; the majority were not published articles (e.g. blog posts, MO pages). Only entries which were journal articles, articles in conference proceedings, preprints of journal articles, masters theses or dissertation theses were retained.

For each article the following inclusion criteria were applied to obtain a final set of published articles. An article had to be published in a peer-reviewed journal or conference proceedings (computer science only). Preprints which clearly stated that they had been accepted and were to appear in a specific peer-reviewed journal were also included. Articles had to be published in a journal related to mathematics or a related discipline (e.g. theoretical computer science). These rules meant that the following types of publications were excluded: conference presentations without a proceedings; books or book chapters; lecture notes or seminar notes; preprints which did not indicate that they had been accepted; and papers about MO.

All articles which met the inclusion and exclusion criteria were downloaded from the journals' websites or another source (e.g. arXiv). Two articles could not be downloaded due to licensing restrictions and are not included in this study. Finally, only articles in which the citation was included in the body of the journal article were included; this excluded articles in which MO was cited in the acknowledgements section only.

In addition to the collection of published papers, the study also includes theses, dissertations, and accepted preprints. Theses and dissertations citing MO were retained from the original collection of articles before the inclusion criteria had been applied. A count of the number of preprints was determined by searching arXiv.org, where authors upload their articles as PDFs. Because arXiv does not support full-text search of the articles on its site, Google's search engine was used to search the site "arXiv.org/pdf" for the term "MathOverflow". This allowed a complete search of preprints hosted by arXiv that cited MO.

Data Collection - Citations and References

Citations and references were identified within the collection of published articles. The term *reference* will be used to indicate the text of the reference section, which provides information about the source of the content being cited, e.g.:

"[8] MathOverflow, http://mathoverflow.net/questions/54851."

The term *citation* will be used to refer to text in the body of the article which concerns the given reference, e.g.:

"It was discovered by F. Brunault [8]."

MO mentions were segmented into a collection of distinct citation-reference pairs using a few basic rules. A reference was linked to every place in the text where it was cited. A citation was considered on its own if no reference was provided. Any citations to the same content which shared a reference and appeared in the same article were treated as a single citation (e.g. two different mentions of reference "[8]" in the same article). A single citation which mentioned two parts of a Q&A discussion (e.g. two answers) and which only provided one reference for the citation were treated as a single citation. Citation-reference pairs were matched to particular

Q&A page using the information provided in the published article, if it could be determined.

Variables and Measures

Information about the citations and references was gathered directly from the journal articles and MO. Citation-reference pairs were coded according to the following eight variables:

Self- vs. Other- Citations: In order to determine whether a citation was a self-citation, the names of the journal authors were compared to the names of the authors of the content being cited. If any journal author was an author of the original content it was considered a self-citation. The majority of users on MO use their real names [34]. When the name of the original author(s) could not be determined (e.g., a pseudonym was used) the citation was not coded.

Author experience: For each author, their number of mathematics papers was gathered from MathSciNet, the main database for mathematics literature. These authors were compared to a random sample of contributors to MO. The random sample of authors came from the 150 Q&A posts collected in [33]. I chose this comparison sample because, while it was randomly selected, it had already been screened to eliminate off-topic Q&A posts and was from a similar time period.

Journal Reputation: A measure similar to Impact Factor was collected from MathSciNet for each journal. MathSciNet reports the mathematical citation quotient (MCQ), which is the number of citations of articles in the journal divided by the number of items published in that journal calculated on data from the last 5 years. The reputation of these journals was compared to the average MCQ across all of MathSciNet and to the MCQ scores for a sample of other journal articles published by the same authors (specifically, the most recent paper by that author which did not cite MO).

Complete Reference: A reference was considered complete if it included information that uniquely identified the Q&A page being cited, such as the title of the page, the url of the page, or the question's unique ID. The data was also coded as to the inclusion of a date (month and year).

Individual Attribution: Journal authors were considered to have provided attribution if at least one name was provided in the reference entry.

Author Asked the Question: Whether or not one of the journal authors asked the MO question that resulted in the cited content.

Level of Granularity: Content can be cited at various levels of granularity. Four levels of granularity were coded, presented in order of increasing granularity: Q&A discussion, multiple individual contributions, individual contribution (question, answer, or comment), and edit to an individual contribution. The text surrounding a citation provided information about the level of granularity being cited. For example, the citation "Micciancio [15] showed that ..." was coded as a citation to an individual contribution whereas "...see the discussion in [10] and [11]" was coded as a citation of a coarser Q&A discussion.

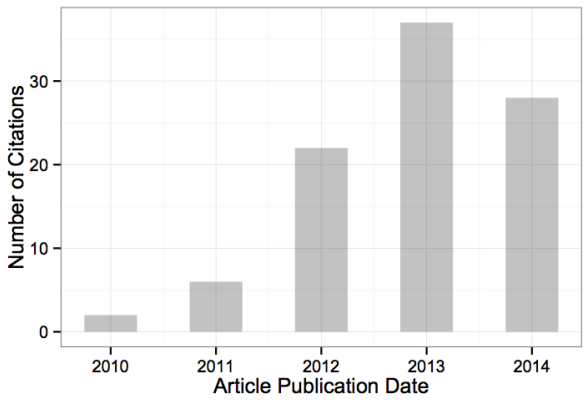

Figure 1. Number of journal articles citing MO per year.

Location of Citation: Citations were classified based on whether they were cited in the main body of the journal article or only appeared in the introduction, related work, discussion section or appendix. Classification was based on standard headers used by authors (e.g. introduction, preliminaries, general discussion, concluding remarks). If no standard headers were used then classification was based on location relative to the first and last proofs given for theorems, corollaries or propositions.

Specific Reference: A reference may or may not provided sufficient information to identify the specific content mentioned in a citation. For example, a reference-citation pair would be considered specific if its reference indicated in some way the answer being cited, such as by providing the answer's unique ID.

Specific Attribution: A reference was considered to provide a specific attribution if it named all the original author(s) of the content being cited (e.g. when citing an answer, giving the name of the answer's author in the reference).

Shared Attribution: Journal authors were considered to have provided shared attribution if they named two or more people in the reference entry (e.g. named both the question-asker and a respondent).

RESULTS

Research Question Set 1: Is MathOverflow content being treated as formal communication?

93 published articles in mathematics and related fields cited MO contributions. A larger body of unpublished work has also cited or acknowledged MO, including 1 Master's thesis, 21 Ph.D. dissertations, and 778 preprints [2]. Some fraction of these articles may eventually end up in the published literature. Figure 1 shows the number of published articles with citations over time. There was early growth in the number of citations per year, which tracks the growing popularity of MO during these early years. The number of citations to MO may be stabilizing or slightly declining to around 30 per year.

[2]Counts of theses and preprints may partially overlap with the number of published articles (e.g. a portion of a dissertation has already been published).

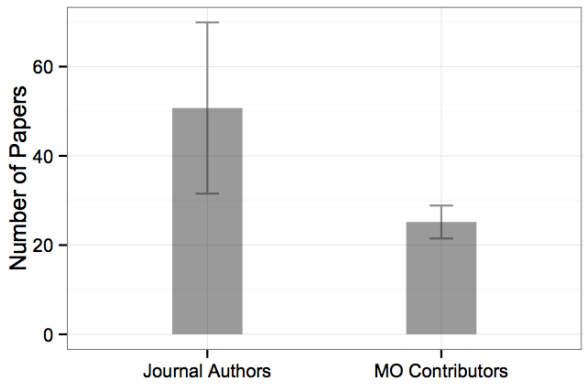

Figure 2. Number of papers with 95% confidence intervals for journal authors citing MO and published contributors to MO.

There were a total of 109 citations within published articles. There were 18 self-citations; these citations represent authors formalizing contributions first made in informal communication channels by publishing them as part of a journal article. The remaining 90 citations[3] are examples of informal communication being treated as stand-alone artifacts. These citations suggest that MO is being treated as a legitimate source of information.

On average, MO content was typically cited by experienced authors in central journals. The journal authors had more papers on average than a typical contributor to MO, even after limiting consideration to published authors (Negative binomial regression, Coef. = 0.70, z = 5.46, df = 380, p < 0.001)[4]. The journal authors had, on average, 25 more papers than a typical MO contributor (Journal authors: M = 51, $Med.$ = 17, SD = 108; Contributions to MO: M = 25, $Med.$ = 16, SD = 30). The impact factor of journals citing MO was higher than that of the average paper on MathSciNet (one tailed t-test: t(65) = 4.4, p < 0.001): 0.62 for MO citations compared to 0.38 on average (SD = 0.44). The journals in which authors published MO citations were equivalent to those in which their other publications appeared; there was no statistical difference in impact factor between the two (paired t-test: t(105) = -0.37, p = 0.71; Other Journal M = 0.63 Journal Citing MO = 0.65).

In addition, the nature of these citations suggest that individuals are using MO for other functions of formal communication. Authors included a reference entry in the reference section for the majority of MO citations, 79% ($\chi^2(1) = 43.9, p < 0.001$). Every reference included either the title of the Q&A page, a unique URL to the Q&A page, or the unique ID for the Q&A page. This information allows the MO content to be retrieved by future readers. Journal authors named at least one of the original authors of the MO content in the reference entry for the majority of citations, 67% ($\chi^2(1) = 20, p < 0.001$). This provides original authors with attribution for their work.

[3]One citation could not be classified due to missing information.

[4]Published authors meant any MO contributor with at least one paper on MathSciNet.

Level of Granularity		% of Citations
Q&A Discussion		9%
Multiple Contributions		10%
Individual Contributions		81%
	Question	27%
	Answer	73%
	Comment	0%
Edits to Contributions		0%

Table 1. Percentage of citations broken down by level of granularity and type of individual contribution.

The fact that MO content was treated more like a journal article and less like a web resource led to one unexpected consequence: the majority of references did not include the date when the content was retrieved ($\chi^2(1) = 4.4, p = 0.04$). Only, 39% of references included the retrieval date (or the date of the content creation). For archival purposes, dates are especially important in this context because questions and answers on MO are often edited and many of these edits are substantial [33].

Research Question Set 2: Do MathOverflow citations reflect the unique affordances of MathOverflow?

MO provides a platform in which the authors can quickly and easily seek answers from a large audience. The importance of this capability is demonstrated by the fact that 58% of citations involved a question asked by one of the article's authors. In the remaining 42% of the instances the authors of the journal article were not involved in generating the cited content[5]. Presumably these authors discovered the content through browsing or searching. There was no statistical difference in the frequency of these two different practices ($\chi^2(1) = 2.2, p = 0.13$).

The majority of citations were at the level of individual contributions, such as a question or an answer (see Table 1; $\chi^2(1) = 44.8, p < 0.001$). No citations cited content at the level of a single edit to a question or answer. Some citations were at higher levels of granularity and acknowledged multiple contributions as a unit. Among citations to individual contributions, citing answers was more common than citing questions, and no comments were cited on their own.

Slightly more than half of the citations, 56%, were included as part of the main body of the journal article; the remaining 44% were cited in the introduction, related work, discussion or appendices only. Inexperienced journal authors were more likely to use MO content in the main body of the article. Authors who had prompted the results by asking the question were also more likely to the cite content in the main body of the article. Logistic regression showed that both these variables, journal authors' number of papers (odds = 0.73, z = -2.1, p = 0.04) and whether an author had asked the original question on MO (odds = 4.2, z = 2.5, p = 0.01), were significantly related to use of MO content in the main body ($\chi^2(2) = 11.5$, p = 0.003). Only some authors may be using MO to

[5]Measured to the best of my knowledge. Using real names is strongly encouraged, but authors may have used a pseudonym when posting

Level of Granularity	Reference	Attribution		
	Specific	Present	Specific	Shared
Q&A Discussion	100%	0%	0%	0%
Multiple Contributions	29%	71%	43%	43%
Individual Contributions	74%	84%	82%	5%

Table 2. Citations providing a specific reference or a form of named attribution broken down by level of granularity, conditional on the presence of a reference.

support essential elements of publishable results and in only some circumstances. In other cases, this communication may be helpful, but not necessary.

Research Question Set 3: How is credit for MathOverflow content being assigned?

For typical citations, that is citations to journal articles, there is only one unit of granularity, a single article. Credit and attribution are determined by the original authors and citing authors follow their lead. Citations of MO content, on the other hand, cite content at various levels of granularity from individual contributions to entire Q&A discussions. Multiple levels of granularity for citations make credit and attribution more complex. References were examined by the level of granularity.

Citations were equally likely to be included with a reference regardless of level of granularity ($\chi^2(2) = 1.0, p = 0.60$; Q&A Discussions: 75% had a reference; Multiple Contributions: 88%; Individual Contributions: 89%). However, there were significant differences in whether a reference contained enough information to uniquely identify the specific contribution being cited (see Table 2; $\chi^2(2) = 9.0, p = 0.01$). References to Q&A discussions were the most likely to include identifying information. References to multiple contributions were the least likely to identify specific contributions in the reference entry, followed by individual contributions. These differences are best explained by differences in the accessibility of identifying information for components of a Q&A page. In order to cite specific parts of a Q&A page, such as an answer, a reader must sift through the source page html or find the cite button for a specific contribution. The cite button is hidden and not easy to find. In addition some types of content, such as comments, do not have a cite button.

There were significant differences in whether and how journal authors gave credit to original authors based on the type of citation. Journal authors credited MO authors more often when the content being cited was more granular ($\chi^2(2) = 9.0, p = 0.01$). Journal authors who referenced individual or multiple contributions usually gave credit to at least one original author, whereas no journal authors gave credit to original authors when citing an entire Q&A discussions (see Table 2). When credit was given to original authors it was typically given to the author most directly responsible for the content being cited. For example, 100% authors of questions were referenced by name and 98% of authors of answers were referenced by name when the content being cited was a question or an answer respectively and credit was given to someone. This was not true to the same extent for citations of multiple contributions together, only 60% of these references mentioned all the direct authors of the content being cited.

Journal authors did not make much use of shared attribution (i.e. including multiple people's names in a reference). There were significant differences based on the granularity of the citations ($\chi^2(2) = 12.1, p = 0.002$). The most frequent use of shared attribution was when citing multiple contributions together (see Table 2). Rarely were shared attributions given when citing individual contributions. Unexpectedly, when citing an entire Q&A discussion shared attribution was not used. No attributions were given for these citations.

GENERAL DISCUSSION

One strategy to support greater openness in science and mathematics is to blend informal and formal scholarly communication, mitigating the risks associated with revealing prepublication ideas (e.g. theft of ideas) and instead inheriting the benefits associated with publication (e.g. recognition and credit). This blurring of the boundaries between formal and informal communication has led to optimism regarding the potential for web technologies to supplant traditional forms of scholarly communication [29], in particular improving the ease, speed, and reach of informal communication [21]. MathOverflow is one example of an emerging Q&A platform that has been coopted for scholarly communication in mathematics and exhibits these design features. At present, contributions to MO unquestionably fall into the category of informal communication. Indeed, several mathematicians think of these contributions as analogous to personal communication [20]. While these conversations might once have happened face-to-face, now they occur online, using the new functionality provided by modern technology. Nonetheless, these informal communications are now supplemented by the four affordances of formal communication: archiving, registration, dissemination and certification.

In this paper we find empirically that, given technology for informal communication which supports formal affordances, some individuals will cite informal communication in ways similar to published articles. We found evidence that informal communication from MO was cited in published articles, by experienced mathematicians, in standard, well-regarded mathematics journals. In addition, these contributions were cited in ways which allow for retrieval and attribution, although these did not always reflect the increased granularity available in the new technology. This is consistent with prior theoretical work, which has suggested that imbuing informal communication with the four aspects of formal communication would be sufficient to mimic publication in journals (e.g. [29]). These results add to a growing body of empirical work showing that contributions via informal communication channels are being cited at substantial rates within published journal articles [1, 9] and suggest that informal communication is

becoming a more legitimate source of scholarly knowledge. The legitimization of new forms of informal communication provide scholars with better tools to seek out the unpublished knowledge they need to pursue their own research. We also found that journal authors often use MO to elicit knowledge from a large audience and that the knowledge generated and cited was typically small in scope (e.g. a single answer).

Compared to other types of informal communication supported by web technology, such as email listservs, blogs, preprint servers, and social media, Stack Exchange Q&A platform may be particularly well-designed to support the requisite aspects of both formal and informal communication. The Stack Exchange voting system provides a form of peer review (certification) which most other forms of informal communication lack; it also serves to promote the most valuable ideas to a wider audience (dissemination). Like arXiv and unlike, email listervs and blogs, the fact that so much content is on a single platform makes archiving and registration easier. All content is stored using consistent, unique identifiers and is currently being archived using a public repository. In contrast, links to emails and blogs in published articles are often subject to link rot and may be unavailable only a few years after a journal article is published.

Unlike other forms of informal communication which have been integrated into scholarly citations such as blogs and preprints [9], affordances of Q&A technology also allow greater use of the unique affordances of informal communication. On MO it is easy for journal authors to ask for the information they need to bolster their own work, by asking a question directed to their interests. It is less straightforward to elicit content from blogs or preprint servers, although this is possible on email listervs and social media. Different forms of informal communication also cater to different quantities of knowledge transmission: social media encourages small, off-the-cuff exchanges whereas preprints are limited to article-size contributions. Others, such as blogs and MO allow for contributions of varying sizes.

For these reasons, the Q&A platform may be particularly advantageous as an informal communication platform; it allows authors to elicit content and exhibits flexibility in the character of their responses. In addition to advantageous design features, historical uses and community attitudes may also dictate whether valuable prepublication content is shared on a particular platform. While social media could be used to share prepublication ideas, and indeed shares many similarities with Q&A platform conversations, in-depth discussion are less common than simply providing links to published work [28, 18].

The results of this paper suggest that technology design which supports informal communication with the functions of formal communication can foster greater openness and sharing in science and mathematics, goals which present an ongoing problem faced by researchers in the CSCW community (e.g. [5, 35]). Successful sharing depends on incentivizing sharing and openness [24]. Here we make the argument that by blending informal communication with the affordances of formal communication we can lower the risks of sharing and increase its benefits to individual researchers. Crucially, this requires that individuals receive credit for the ideas they share.

Another major finding from this study is that problems with references and attribution arise from the integration of formal and informal ways of communicating. Although the majority of MO citations were provided with references that linked to a unique Q&A page, fewer citations of individual and multiple small contributions provided enough information to identify specific contributions within the Q&A page. Without such pointers, specific content (and its authors) are not easily retrievable. Without a clear indication in the reference section of the specific content being cited there is no link between reference and content, which could in the future contribute to the impact factor of a particular contribution. While static references to specific content were problematic for small contributions, attribution to specific individuals was problematic for larger collaborative contributions. Specific individuals were not credited in the reference section for citations of the entire discussions and fewer specific individuals were credited when multiple contributions were cited collectively.

Design Recommendations

The use of new media for scholarly communication disrupts traditional patterns of citation and reference. New community guidelines are needed to integrate new forms of communication into formal references. Citations of this content can vary in their specificity from short comments to lengthy, multi-author discussions, and references should reflect this while preserving traditional elements of attribution and archiving. With the emergence and rapid change of new technologies supporting scholarly communication, new guidelines must develop through an interaction between the community and the developers of its tools. Potential features to support this development could include the provision of governance tools for the community to decide on citation and reference guidelines, allowing authors to give attributions of their own, and providing tools to automatically generate appropriate references for citations.

Community governance tools: Reference style and attribution are fundamentally a discipline-specific, community decision. Traditionally, these sorts of decisions have been made by scholarly organizations or journals. Such organizations often react slowly and, as novel platforms like MO appear, decisions on these matters may also become more informal. Platform design should allow communities to settle on new reference styles and promote their decisions. Stack Exchange sites already make use of a meta Q&A forum to discuss moderation and other community issues; there are several discussions of proper citation and reference styles for MO. These platforms could also use governance tools, such as voting systems for community norms, so that group decisions like the selection of an appropriate reference style can be made with technological support. The more easily these decisions can be made, the easier they will be to modify as technology changes.

Attribution tools: Journal authors did not reference individuals as often when they cited multi-author content. One way to address this would be to give MO authors tools to declare

who should get credit for a particular contribution. Authors often revise their answers, sometimes heavily, in response to comments left by others. They usually acknowledge these influences in the text, but there is no systematic way to assign shared credit. If authors could specify attributions themselves, this would help readers to accurately assign credit for ideas. Similarly the question-asker, who can already select a preferred answer, could decide who should receive credit for citations of the entire Q&A discussion (e.g. everyone equally, the top answerer, the top answerers). Defaults could be set based on community guidelines, but such a tool would leave the ultimate decision with the individuals involved.

Reference tools: Currently MO supports a citation widget that generates a reference as a BibTeX or plain text snippet. This is an important feature and should be more thoroughly developed. The widget is hidden on the site and seems to be rarely used, as only a minority of references matched the widget's style. It should be displayed more prominently. While the citation widget allows a reader to cite a particular question or answer, it does not support references at multiple levels of granularity. It could easily be extended to create appropriate references for citing an entire Q&A discussion, multiple contributions together, or even individual edits. This tool would be helpful because it would create a *de facto* standard for reference styles. It could also handle the tedious retrieval of author, unique id for contribution, date, and archival information which journal authors often omit. The specific way in which this reference information is instantiated is not as important as the fact that it should be done in a way that is 1) consistent, 2) complete, and 3) made easily available to users to encourage high usage rates. For example, instead of a citation widget, URLs could be created for each contribution (now only the Q&A as a whole has a url, not each post) and designed to have all necessary attribution and archival information.

Limitations & Future Work

By design, this study only focused on positive examples of MathOverflow citation. MO content that was included in journal articles without any citation could not be collected and was not studied. There may be a substantial number of contributions that do go uncited. For example, one author asked for advice on meta.mathoverflow (a site discussing MO) to help address a situation in which his MO work was included in a preprint without a citation (e.g. [3]). In addition, only citations that used the word "MathOverflow" were collected. Citations that referred to the source of the content as personal communication without saying that the communication occurred on MO were not collected.

A qualitative study, such as an interview study with mathematicians familiar with MO who either do or do not cite MO content in their journal articles would help validate these quantitative results and provide deeper context to understand usage. This quantitative study shows that some individuals are citing MO, but it cannot answer deeper questions about how MO content is viewed and being used. For example, this study cannot tell us the degree to which individuals feel that MO is legitimate, credible source of material (maybe only some content is seen as credible or it is only seen as partially credible). This study also cannot tell us whether MO is actually being used to generate new content (maybe it is only being used to make work developed by the journal authors seem more credible and grounded).

Future work should gather qualitative data on perceptions and attitudes directly from users to complement these quantitative results. In particular, it would valuable to know how perceptions of MO's credibility varies across different subdisciplines of math, how it varies across different levels of mathematical experience, and in what contexts mathematicians do or do not view MO as an appropriate (citable) resource.

CONCLUSION

As new technologies emerge for scholarly communication, there will be a continued blurring between informal and formal communication. The design of these technologies can have a large impact on the success of these systems for open science, and the long-term value of these citations within the scientific literature. Citations of MathOveflow content reveal that journal authors make use of the platform's features to use informal communication in traditionally formal ways, drawing on the platform's archival features. They also reflect the traditional values of informal communication in providing tailored, interactive interactions. Problems with citations and references emerge when these formal and informal functions mesh. Design of future systems should focus on the development of community citation guidelines and the creation of citation tools to help readers follow these guidelines.

REFERENCES

1. Ofer H. Azar. 2007. The slowdown in first-response times of economics journals: Can it be beneficial? *Economic Inquiry* 45 (2007), 179–187.
2. Aruna D. Balakrishnan, Sara Kiesler, Jonathon N. Cummings, and Reza Zadeh. 2011. Research team integration: What is it and why it matters. In *Proceedings of the ACM conference on Computer Supported Cooperative Work (CSCW '11)*. ACM Press, 523–532.
3. Andras Batkai. 2014. Reference to Mathoverflow. (2014). http://meta.mathoverflow.net/questions/1629/reference-to-mathoverflow
4. Jeremy P. Birnholtz. 2006. What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science* 57 (2006), 1758–1770. DOI:http://dx.doi.org/10.1002/asi
5. Jeremy P. Birnholtz and Matthew J. Bietz. 2003. Data at work: Supporting sharing in science and engineering. In *Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work (GROUP '03)*. ACM Press, 339–348. DOI: http://dx.doi.org/10.1145/958160.958215
6. Kevin Boudreau and Karim Lakhani. 2013. Cumulative innovation & open disclosure of intermediate results:

Evidence from a policy experiment in bioinformatics. (2013).

7. Cecelia M. Brown. 1999. Information seeking behavior of scienctists in the electronic information age: Astronomers, chemists, mathematicians and physicists. *Journal of the American Society for Information Science* 50, 10 (1999), 929–943. http://www.trans.uma.es/numeros.html

8. Carolin Cardamone, Kevin Schawinski, Marc Sarzi, Steven P. Bamford, Nicola Bennert, C. M. Urry, Chris Lintott, William C. Keel, John Parejko, Robert C. Nichol, Daniel Thomas, Dan Andreescu, Phil Murray, M. Jordan Raddick, AnŽe Slosar, Alex Szalay, and Jan Vandenberg. 2009. Galaxy Zoo Green Peas: Discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society* 399, 3 (2009), 1191–1205. DOI:http://dx.doi.org/10.1111/j.1365-2966.2009.15383.x

9. Chuanfu Chen, Kai Sun, Gang Wu, Qiong Tang, Jian Qin, Kuei Chiu, Yushuang Fu, Xiaofang Wang, and Jing Liu. 2009. The impact of internet resources on scholarly communication: A citation analysis. *Scientometrics* 81, 2 (2009), 459–474. DOI:http://dx.doi.org/10.1007/s11192-008-2180-y

10. Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466 (Aug. 2010), 756–60. DOI:http://dx.doi.org/10.1038/nature09304

11. Elizabeth A. Corley, P. Craig Boardman, and Barry Bozeman. 2006. Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research Policy* 35, 7 (2006), 975–993. DOI:http://dx.doi.org/10.1016/j.respol.2006.05.003

12. Justin Cranshaw and Aniket Kittur. 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)*. ACM Press, Vancouver, Canada, 1865–1874.

13. Jonathon Cummings and Sara Kiesler. 2007. Coordination costs and project outcomes in multi-university collaborations. *Research Policy* 36 (2007), 1620–1634.

14. Thomas A. Finholt and Gary M. Olson. 1997. From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science* 8, 1 (1997), 28–36. DOI:http://dx.doi.org/10.1111/j.1467-9280.1997.tb00540.x

15. Jeremy Fox. 2012. Can blogging change how ecologists share ideas? In economics, it already has. *Ideas in Ecology and Evolution* 5 (2012), 74–77. DOI:http://dx.doi.org/10.4033/iee.2012.5b.15.f

16. Oliver Gassmann and Maximilian von Zedtwitz. 1999. New concepts and trends in international R&D organization. *Research Policy* 28, 2-3 (1999), 231–250. DOI:http://dx.doi.org/10.1016/S0048-7333(98)00114-0

17. Stevan Harnad. 1990. Scholarly skywriting and the prepublication continuum of scientific inquiry. *Psychological Science* 45, 1990 (1990), 342–343. DOI:http://dx.doi.org/10.1111/j.1467-9280.1990.tb00234.x

18. Kim Holmberg and Mike Thelwall. 2014. Disciplinary differences in Twitter scholarly communication. *Scientometrics* 101 (2014), 1027–1042. DOI:http://dx.doi.org/10.1007/s11192-014-1229-3

19. James Howison and James D. Herbsleb. 2011. Scientific software production: Incentives and collaboration. In *Proceedings of the ACM conference on Computer Supported Cooperative Work (CSCW '11)*. ACM Press, 513–522.

20. Steve Huntsman. 2010. No Title. (2010). http://tea.mathoverflow.net/discussion/64/where-to-keep-track-of-math-overflow-success-stories/

21. Rob Kling, Geoffrey Mckim, and Adam King. 2000. A bit more to it: Scientific multiple media communication forums as socio-technical interaction networks. *Journal of the American Society for Information Science* 54, 2001 (2000), 47–67.

22. Greg Kuperberg. 2002. Scholarly mathematical communication at a crossroads. *Nieuw Archief Wiskunde* 5 (2002), 262–264. http://arxiv.org/abs/math/0210144

23. Charlotte P. Lee, Paul Dourish, and Gloria Mark. 2006. The human infrastructure of cyberinfrastructure. In *Proceedings of the ACM conference on Computer Supported Cooperative Work (CSCW '06)*. ACM Press, 483–492. DOI:http://dx.doi.org/10.1145/1180875.1180950

24. Arijit Mukherjee and Scott Stern. 2009. Disclosure or secrecy? The dynamics of Open Science. *International Journal of Industrial Organization* 27, 3 (2009), 449–462. DOI:http://dx.doi.org/10.1016/j.ijindorg.2008.11.005

25. Gary M. Olson and Judith S. Olson. 2000. Distance matters. *Human-Computer Interaction* (2000).

26. DHJ Polymath. 2012. A new proof of the density Hales-Hewett theorem. *Annals of Mathematics* 175 (2012), 1283–1327.

27. Jason Priem. 2013. Scholarship: Beyond the paper. *Nature* 495 (2013), 437–40. DOI:http://dx.doi.org/10.1038/495437a

28. Jason Priem and Kaitlin Light Costello. 2010. How and why scholars cite on Twitter. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem (ASIS&T '10)*, Vol. 47. 1–4. DOI: http://dx.doi.org/10.1002/meet.14504701201

29. Jason Priem and Bradley M. Hemminger. 2012. Decoupling the scholarly journal. *Frontiers in Computational Neuroscience* 6 (2012), 1–19. DOI: http://dx.doi.org/10.3389/fncom.2012.00019

30. David De Roure, Carole Goble, Sergejs Aleksejevs, Sean Bechhofer, Jiten Bhagat, Don Cruickshank, Paul Fisher, Duncan Hull, Danius Michaelides, David Newman, Rob Procter, Yuwei Lin, and Meik Pschen. 2010. HPCTOOLKIT: Tools for performance analysis of optimized parallel programs. *Concurrency Computation Practice and Experience* 22 (2010), 685–701. DOI: http://dx.doi.org/10.1002/cpe

31. Fytton Rowland. 2002. The peer review process. *Learned Publishing* 15 (2002), 247–258.

32. Gordon B. Schmidt and Richard N. Landers. 2013. Solving the replication problem in psychology requires much more than a website. *Industrial and Organizational Psychology* 6 (2013), 305–309. DOI: http://dx.doi.org/10.1111/iops.12056

33. Yla R. Tausczik, Aniket Kittur, and Robert E. Kraut. 2014. Collaborative problem solving: A study of MathOverflow. In *Proceedings of the ACM conference on Computer Supported Cooperative Work (CSCW '14)*. ACM Press, Baltimore, Maryland, 355–367. DOI: http://dx.doi.org/10.1145/2531602.2531690

34. Yla R. Tausczik and James W. Pennebaker. 2011. Predicting the perceived quality of online mathematics contributions from users' reputations. In *Proc. Human Factors in Computing Systems*. ACM Press, Vancouver, Canada, 1885–1888. DOI: http://dx.doi.org/10.1145/1978942.1979215

35. Theresa Velden. 2013. Explaining field differences in openness and sharing in scientific communities. In *Proceedings of the ACM conference on Computer Supported Cooperative Work (CSCW '13)*. ACM Press, 445–457. DOI: http://dx.doi.org/10.1145/2441776.2441827

36. Theresa Velden, Matthew J. Bietz, E. Ilana Diamant, James D. Herbsleb, James Howison, David Ribes, and Stephanie B. Steinhardt. 2014. Sharing, Re-use and Circulation of Resources in Cooperative Scientific Work. In *Companion CSCW 2014*. ACM Press, 347–350. DOI: http://dx.doi.org/10.1145/2556420.2558853

37. Andrea Wiggins and Kevin Crowston. 2010. Distributed scientific collaboration: Research opportunities in citizen science. In *CSCW 2010 Workshop*. 1–4. http://www.sci.utah.edu/images/docs/cscw2010/wiggins_2.pdf

38. Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science* 316 (May 2007), 1036–1039. DOI: http://dx.doi.org/10.1126/science.1136099