Towards mathematical AI via a model of the content and process of mathematical question and answer dialogues

Joseph Corneli[1], Ursula Martin[2], Dave Murray-Rust[1], and Alison Pease[3]

[1] University of Edinburgh
jcorneli@staffmail.ed.ac.uk
[2] University of Oxford
[3] University of Dundee

Abstract. This paper outlines a strategy for building semantically meaningful representations and carrying out effective reasoning in technical knowledge domains such as mathematics. Our central assertion is that the semi-structured Q&A format, as used on the popular Stack Exchange network of websites, exposes domain knowledge in a form that is already reasonably close to the structured knowledge formats that computers can reason about. The knowledge in question is not only facts – but discursive, dialectical, argument for purposes of proof and pedagogy. We therefore assert that modelling the Q&A process computationally provides a route to domain understanding that is compatible with the day-to-day practices of mathematicians and students. This position is supported by a small case study that analyses one question from Mathoverflow in detail, using concepts from argumentation theory. A programme of future work, including a rigorous evaluation strategy, is then advanced.

Keywords: Q&A, argumentation, mathematics

1 Introduction

In this paper, we outline a computational approach to modelling mathematical dialogues, and show how it can be used to model *Q&A dialogues* in particular. We argue that a strongly empirical approach – along these lines – can support the development of robust, knowledge-rich, mathematical artificial intelligence. Mathematical dialogues convey the processes through which new mathematics is created and existing mathematics is taught. We claim that mathematical question and answer (Q&A) dialogues are a practically and theoretically important subclass. In particular, there is now a large corpus of mathematical questions, answers, and accompanying discussion available in online Q&A forums.

The Q&A forums that we consider here are "social machines" – defined by Tim Berners-Lee to be a class of systems "in which the people do the creative work and the machine does the administration" [4]. Question-and-answer websites differ from other popular social machine formats, like general purpose forums, wikis, and mailing lists – with which they nevertheless share some features – in the relatively explicit semantics that they support (and require).

This has to do with both content and form. The Stack Exchange network of Q&A sites has a network-wide norm of focusing on questions whose answers are not primarily opinion-based: in other words, questions which have answers that can be considered "right" or "wrong," or that can otherwise be compared with each other in objective (as opposed to purely subjective) terms. Stack Exchange sites treat a wide range of technical and non-technical subjects, ranging from computer programming, to travel, to advice on academic careers, the internal logic of science fiction universes, and beyond. It has two specialist sites devoted to mathematics: Mathoverflow, for research-level Q&A (often dealing with new, open, conjectures, and for which background approximately equivalent to a strong mathematics degree is a minimum barrier to entry), and math.stackexchange.com, which focuses on non-research mathematics (e.g., at school, university, or postgraduate level). Whereas a mailing list, for example, would permit more open-ended discussions, the questions discussed on these websites tend to have right and wrong answers, and the discussion focuses on exposition of the correct answer (or answers).

Technical Q&A often embody knowledge about "how to" as well as "what is." In this regard, Q&A is similar to computer programming [38] – and it is no coincidence that the most popular site on the Stack Exchange subject is devoted to programming concepts. As a commentary on the medium and its affordances, it is useful to note that, by number of questions, the math.stackexchange.com website is the second-most popular site in the Stack Exchange network.[4]

Mathematical Q&A has some interesting things in common with other kinds of mathematical discussions, such as the discussions that take place among professional mathematicians working on a paper together, or among novice problem-solvers leading to personally-new insights. Both these types of dialogues have been studied extensively – in some cases using online discussions as a ready source of data [3, 31, 35, 37].

Proof dialogues often contain Q&A sub-dialogues: for example, a discussant may state "I'm sorry, I don't understand what you did in this part of the proof" or "I don't understand why this works, but it seems to." Similarly, Q&A dialogues contain elements of *mathematical argumentation*, typically sufficient to convince the querent and subsequent readers. Following Walton [42], Aberdein has outlined a range of purposes that "proof" may serve: inquiry, persuasion, information seeking, deliberation, negotiation, and debate [1]. Q&A dialogues seem to serve many similar purposes. Martin and Pease [26] developed an empirically-founded typology of questions on Mathoverflow (expanding on Mendes and Milic-Frayling's earlier study [28]). The three popular types of questions observed were "Is it true that…?", "What is this?", and "Could I have an example please?" (we summarise this earlier work in more detail below).

We support our case for a strongly empirical approach to mathematical AI by building a proof-of-concept model of a Q&A dialogue, using a technique from argumentation theory [5] with suitable adaptations for the mathematics domain. We have stated the motivations for this approach elsewhere [27], and demon-

[4] `http://stackexchange.com/sites#questions`

strated its overall salience [32]. In brief, we show in [32] that it is possible to build meaningful computational models of proof dialogue. However, this earlier work dealt with the high-level logic of argument structure. Here we demonstrate extensions that provide a much more detailed model. We will summarise the relevant background below, and in our small proof-of-concept case study, introduce enough formalism to walk through an example in detail. The outline of the paper is as follows.

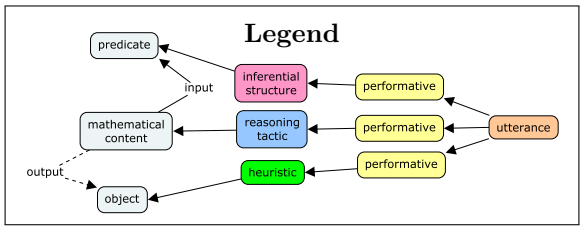
- We present a brief overview of relevant background in Section 2.
- Section 3 develops the main case study.
- Section 4 offers discussion and an outline of future work.

2 Background

2.1 Overview

The approach we take in this paper is grounded in a part of argumentation research called *Inference Anchoring Theory* (IAT), which was designed to model the inferential structure of dialogues, by connecting statements with their logical import [5]. In a recent paper we have described a broadly IAT-inspired theory that uses a constrained set of rules called a *dialogue game* to model mathematical discourse [32]. The specific set of rules were adapted from Lakatos's *Proofs and Refutations* [21]. The "Lakatos game" developed in [32] is a formalised approximation of the "informal logic" that people use when arguing about concepts on the way to a shared proof. The dialogue game shows which assertions are being used to support or to argue against a given conjecture, and it shows when assertions are in conflict with each other. This leads to a sociologically interesting, but rather atypical, idea of a proof as a developing set of mutually coherent statements that support a given conjecture, and that have yet to be successfully refuted. This conception of proof is at odds with the way mathematicians (and, especially, formal mathematicians) would describe proofs, i.e., that proofs are derivations from axioms by way of valid inference rules. Philosophers of mathematics have expressed doubt as to whether mathematical worldviews as different as these can be brought into alignment with each other in a routine way [39]. Nevertheless, mathematics has a logical structure, embodied in its theories and objects, which is typically not subject to debate. For instance, two numbers are either co-prime, or not. Accordingly, we have been developing a new strategy for representing mathematical discussions, in which 'The Cayley graph of group G', for example, is modelled as an object; the relationship between the proposition P and the proposition 'P is difficult to prove' is made explicit; and in which Lakatos-style conjectures, refutations, and repair are modelled. We call this framework *IAT+Content* or IATC. *Inferential structure*, like `implies`, describes statements about pieces of *mathematical content*; meta-level *reasoning tactics*, like `goal` or `auxiliary`, are used to strategise proof development; *heuristics* guide the proof or manipulate content. New nodes are brought

into being in connection with IAT-style *performatives* labelled `agree`, `assert`, `challenge`, `define`, `query`, `retract`, and `suggest`. The Legend at right provides a schematic summary of the features of IATC.



While the proto-language is not complete (and not yet implemented, like the work in [32]), we hope to show that it can be used to model real-world mathematical dialogues. In the current paper, we focus on Q&A dialogues. These were summarised above; some further detail on this specific domain follows.

2.2 Q&A

The strategies used by Mathoverflow "contributors [to] communicate and collaborate to solve new mathematical 'micro-problems' online" have been studied previously [40], and a typology of collaborative acts was proposed: 1. *provide information*; 2. *clarify the question*; 3. *critique an answer*; 4. *revise an answer*; 5. *extend an answer*. The study focused on understanding, quantitatively, how these different activities contributed to answer quality. Some more specific activities within this framework are noted in an example,[5] e.g., 1→ referencing a related Q&A post, which IATC might model as an '`auxiliary`' problem; 2→ stating that the question is harder than the related post, which IATC would model with a (reversed) '`easy`' heuristic value judgement. This framework is useful as a high-level check on the completeness of IATC.

The difficulty of questions in math.stackexchange.com has been studied [24]. However, this was done, not primarily by examining the content of questions, but by devising a "competition-based" score that estimates a given question's difficulty using the estimated expertise of discussion participants (learned via a Bayesian model). In particular: "the expertise score of the best answerer is higher than that of the asker as well as all other answerers." Then, the difficulty of a question is estimated to be "higher than the expertise score of asker u_a, but lower than that of the best answerer u_b." These authors defer a detailed analysis of question content to future work.

Mathoverflow. Related work by some of us examined the production of mathematics on Mathoverflow [26]. A typology of questions was developed, as follows:

- **Conjecture 36%** — asks if a mathematical statement is true. May ask directly "Is it true that" or ask under what circumstances a statement is true. (This corresponds to the purpose of 'inquiry' noted by Walton.)
- **What is this 28%** — describes a mathematical object or phenomenon and asks what is known about it. (This also corresponds to 'inquiry'.)

[5] `http://mathoverflow.net/q/12732`

– **Example 14%** — asks for examples of a phenomenon or an object with particular properties. (This may be 'inquiry' or 'information seeking'.)
– **Formula 5%** — ask for an explicit formula or computation technique.
– **Different proof 5%** — asks if there is an alternative to a known proof. In particular, since our sample concerns the field of group theory, a number of questions concern whether a certain result can be proved without recourse to the classification of finite simple groups.
– **Reference 4%** — asks for a reference for something the questioner believes to be already documented in the literature
– **Perplexed 3%** — ask for help in understanding a phenomenon or difficulty. A typical question in this area might concern why accounts from two different sources (for example Wikipedia and a published paper) seem to contradict each other.
– **Motivation 3%** — asks for motivation or background. A typical question might ask why something is true or interesting, or has been approached historically in a particular way.
– **Other 2%** — closed by moderators as out of scope, duplicates, etc.

Answers are also examined, although in less detail. Responses typically present information known to the respondent, and readily checked by other users, but not necessarily assumed to be known by them. Some specific findings:

– **Existing research literature 56%** — over half of the questions in the sample refer to existing literature
– **Errors 37%** — many questions (and answers) contain errors; these are acknowledged politely, and corrected when pointed out.
– **Examples 34%** — the use of specific examples gives some evidence of broadly "Lakatosian" reasoning (i.e., per [21]).

3 Case study

In this section we will use the IATC formalism to analyse one example Q&A dialogue in detail.[6] We quote the text of this dialogue verbatim, and present the analysis graphically. We selected one of the questions from Mathoverflow that was part of the sample described above: the example was classed as an "Is it true that…?" question; examples and references are supplied. As they appear on the site, both the specific question and the top-rated answer are quite succinct. Below, we will replay the conversation in order. As we will see, most of the interesting argumentation takes place in comments.

> (Original question, 18:05) I have seen this problem, that if G is a finite group and H is a proper subgroup of G with finite index then $G \neq \bigcup_{g \in G} ghg^{-1}$. Does this remain true for the infinite case also. [→ Figure 1]

The first follow-up comment to be submitted observes that this question doesn't quite make sense as written, and suggests a correction.

[6] `http://mathoverflow.net/q/34044`

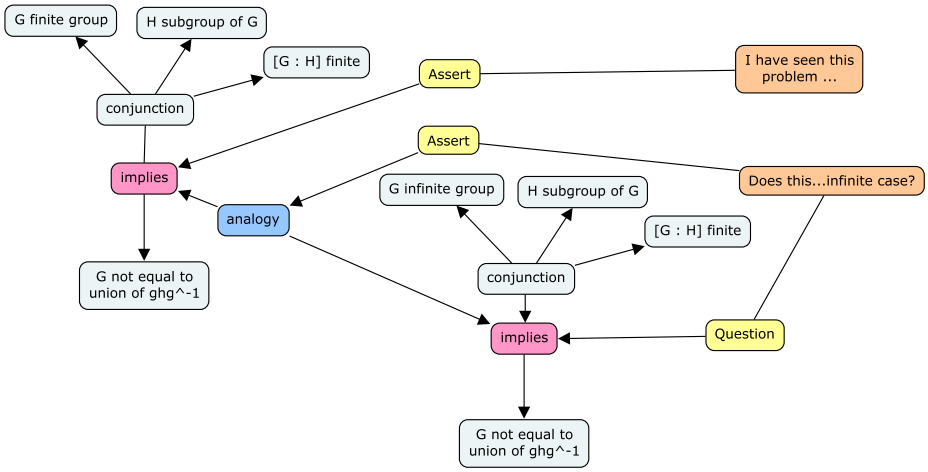


Fig. 1. Original question, diagrammed

> (First comment on question, 18:15) There's something I don't understand here: do you perhaps mean gHg^{-1} instead of ghg^{-1}? [→ Figure 2]

Meanwhile, it seems an answer was already being composed, since the following text appeared on the site one minute after the above clarification.[7]

> (Answer, 18:16) Not in general. Every matrix in $\mathrm{GL}_2(\mathbf{C})$ is conjugate to an invertible upper triangular matrix (use eigenvectors), and the invertible upper triangular matrices are a proper subgroup. [→ Figure 3]

Even though an answer has been given, suggestions for fine-tuning the question continue in the comments.

> (Second comment on question, 18:20) Yes, the statement is out of focus: gHg^{-1} is intended (and "infinite index case"). The natural starting point is to ask whether the proof for finite index breaks down. [→ Figure 4]

Quite a lot happens in the foregoing short comment. A change in the problem's set of hypotheses is suggested. The analogy to the known "inspiring" theorem for finite groups is deemed not particularly relevant, and its hypotheses are changed as well. This then suggests a strategy for proving the (revised) problem. (*NB.*, to save room, in the diagram that follows, we have elided some of the structure that accumulated earlier: the earlier nodes and links are still assumed.)

[7] At this point, the discussion becomes multi-threaded, since comments can now attach to the answer as well.

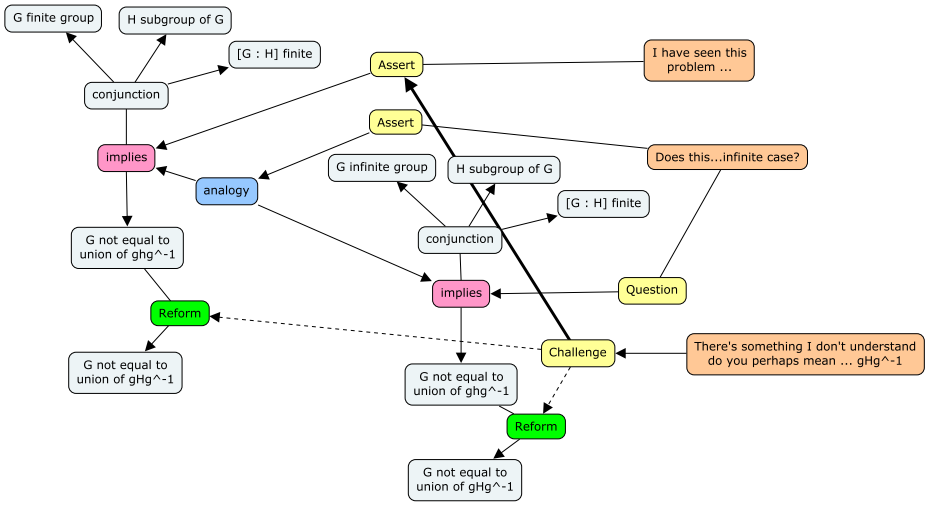


Fig. 2. First comment on question, diagrammed

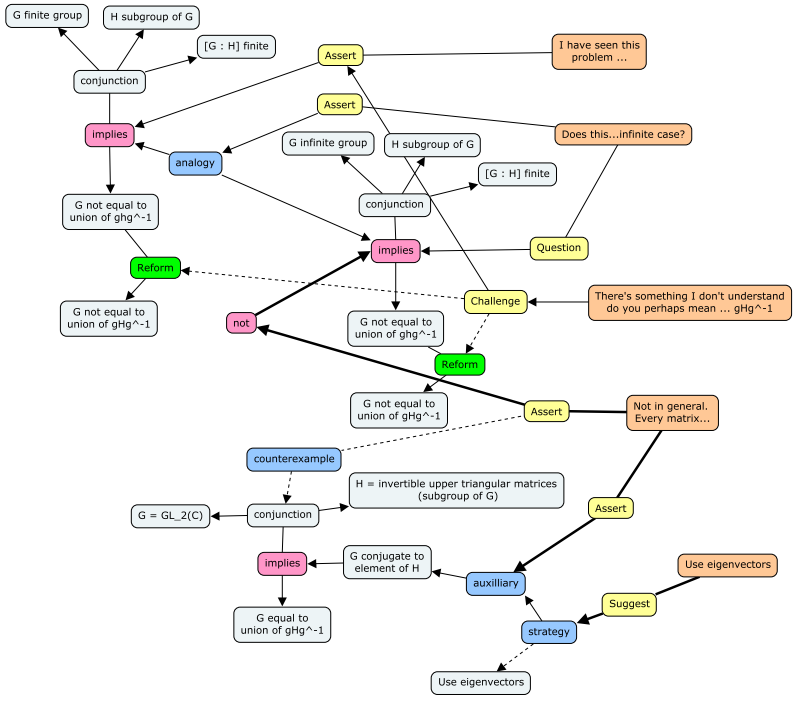


Fig. 3. Answer, diagrammed

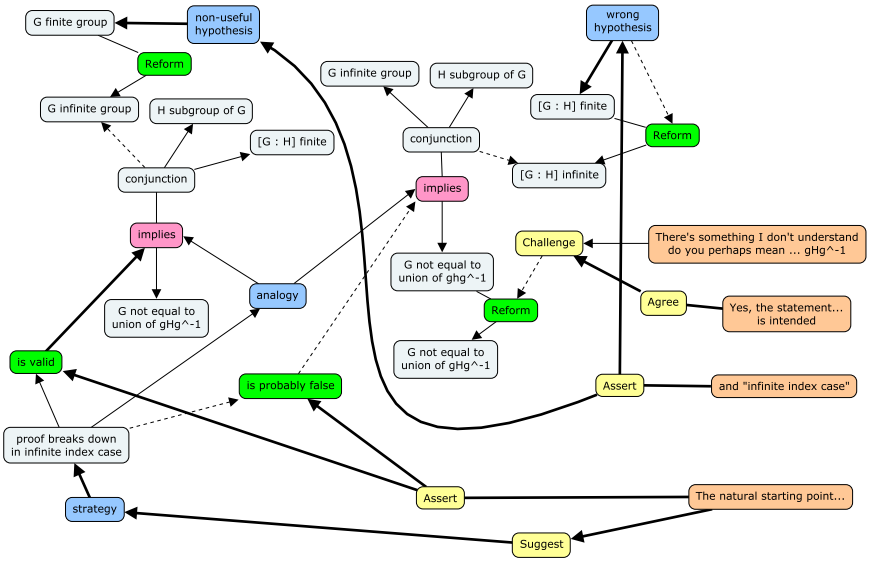


Fig. 4. Second comment on question, diagrammed

In the following comment, both the foregoing refinements and the earlier proposed answer are taken into account.

> (Third comment on question, 18:24) If G is a finite group then all its subgroups have finite index. What the statement should say is that if H is a proper finite index subset of G then $G \neq \cup_{g \in G} gHg^{-1}$ (the case of infinite G readily reduces to the case of finite G). As Keith shows, this is not always true for subgroups of infinite index. [→ Figure 5]

Here, the earlier assertion that $[G : H]$ is the "wrong hypothesis" is challenged. Essentially, this comment is looking for the most precise way to phrase the problem statement. "$[G : H]$ finite" is not wrong, but necessary if we want the implication to hold. The counterexample of invertible upper triangular matrices – "Keith's counterexample", diagrammed in Figure 3 – does indeed have infinite index in $\mathrm{GL}_2(\mathbf{C})$.[8] It is not, therefore, precisely a counterexample to the three-part conjunction in that diagram that it appears to refute; it does serve as a counterexample to the revised three-part conjunction in Figure 4. One should keep in mind that the statement "not in general" that prefaces the answer was addressed not to the specific conjunction in Figure 3, but instead to the OP's considerably more vague question "Does this remain true for the infinite case

[8] Intuitively, upper-triangular 2×2 matrices have one element that is zero, so the subgroup has one codimension in $\mathrm{GL}_2(\mathbf{C})$, namely, a copy of $\mathbf{C}$. The fact that the index is infinite follows (but an algebraic proof is also straightforward). For an example of an infinite group and a subgroup with finite index, consider $\mathbf{Z}$ and $2\mathbf{Z}$: in this case, the group is equal to the union of cosets.

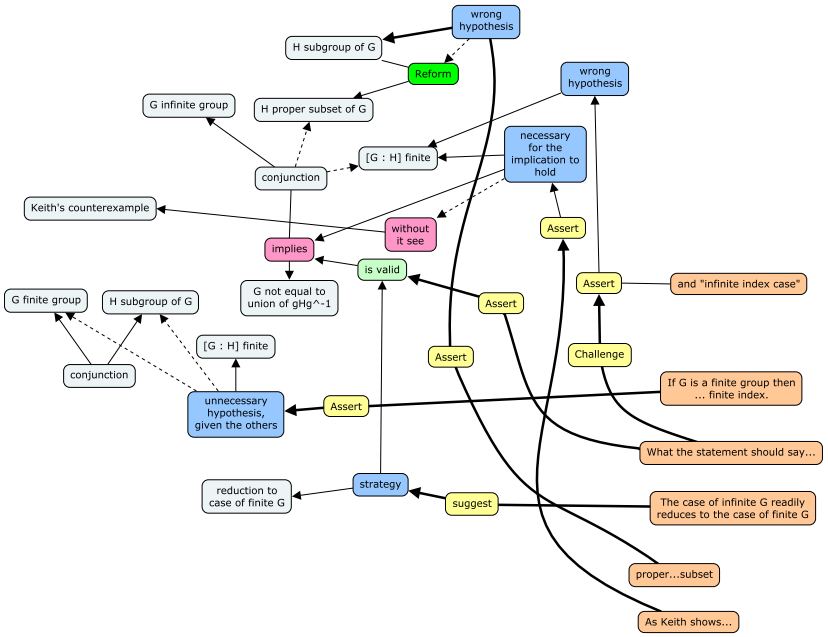


Fig. 5. Third comment on question, diagrammed

also?" Once the question has been clarified by other contributors, the logic of the answer works.

Subsequent to this, further terminological issues arise in comments on the answer, which again have to do with understanding exactly what is being asked, e.g., adding another condition of "discreteness" to the conjunction, for an answer with an even more "finite feel." New examples are proposed, both with reference to the literature and by straightforward adaptation of the answer already presented. In addition, four alternative answers are supplied, without attracting further comment (but with argumentation provided "in advance", so to speak). We will not draw diagrams for this material, because the illustrative presentation above is sufficient for our current purposes.

4 Discussion and Future Work

The case study that was examined above shows both the basic promise of the argument-theoretic approach to modelling mathematics, and some of the difficulties that would have to be overcome in 'scaling up' this approach. Bundy argues that the right representation is the key to successful reasoning [6, p. 16]. While both representation and reasoning are crucial (and, for 'functional inference', it is important to be strategic about representations *of* reasoning) in a material sense, the logically-prior step is that of building the representation. One

alternative to hand-coding is to search for catalogues of existing data, and more specifically, of reasoning, that have already been represented in some preliminary and readily digestible form. Learning from text was a key part of IBM's approach in building the version of Watson that competed in *Jeopardy* [11]. The Austin-based company, Cycorp, known for its large hand-coded AI system, also emphasises learning from documents in more recent work [34]. More broadly, "Knowledge Extraction from Text" is a well-known domain of computing research (with workshops running since 2013). The requirements for knowledge extraction from mathematical text, in particular, are being worked out in specialist efforts in linguistics and NLP [13, 10, 8] (and see in particular the survey in [14]).

Issues associated with mining Wikipedia and scientific literature have been explored (e.g., [7], [16], [17]) with successful scientific and technical proofs-of-concept at various scales. The technical issues associated with mining mathematical literature are increasingly well-understood [29] – and implementation work is ongoing.[9] *Argument mining* is, however, a relatively new area [33, 22]. In the current state of the art, discourse structure may have to be laboriously hand-coded.[10] We may be aided in our specific pursuit by the fact that much of mathematics is relatively "formulaic." For example, the first 100 most frequent schematic constructions (like "let X be a Y") were shown to cover half of the sentences on the detailed but mathematically informal ProofWiki website [18].

Even so, as we've seen in our case study, modelling mathematical arguments is far from a trivial task. The initially quite simple language of '`implies`' and '`analogy`' from Figure 1 is complemented by a range of much more complex relations in Figure 5, like '`necessary for the implication to hold`'. The '`Reform`' relation has been applied to individual pieces of mathematical content, but in some cases the relevant transformations would have to happen on the graph of statements, for example, if we were to specify the '`analogy`' between two problem statements; reasoning from analogy can be complicated [25, 36]. The precise representations need to be worked out; and, as indicated above, we also want our representations to be capable of simulating the reasoning evolved in an effective manner.

Looking to the future, given a suitably-represented knowledge base of mathematical facts from whatever source, one interesting line for research (with a Q&A feel) would build on the work of Nuamah et al [30], who describe a strategy for automatically assembling the answers to queries when the answers are not directly stored in the database. As a simple example, it may be the case that there is no prerecorded answer to the following question:

> *Does the country x that has the highest GDP per capita (GDP/c) out of all countries in South America have a higher GDP/c than the country y that has the highest GDP/c out of all countries in Africa?*

[9] `https://www.authorea.com/users/5713/articles/51708-understanding-a-dataset-arxiv-org/_show_article`

[10] `https://research.googleblog.com/2017/05/coarse-discourse-dataset-for.html`

Nevertheless the answer may be computed from information that is available in the database (namely, GDP/c for all countries, together with the association of countries to continents). We assert that at a high level the logic of 'functional inference from heterogenous data' would be similar if the data was not CIA Fact Book-style, but, instead, mathematical facts – or, indeed, facts about mathematical arguments.

Broadly, the proposal that we suggest pursues the slow refinement of mathematical knowledge from knowledge about mathematical arguments. We can compare and contrast this with a famous proposal in the computer mathematics community, the QED manifesto [2]. This proposed "a project to build a computer system that effectively represents all important mathematical knowledge and techniques." We agree with this general aim. However, the QED manifesto relied on the idea of "the use of strict formality in the internal representation of knowledge and the use of mechanical methods to check proofs of the correctness of all entries in the system." If relied on as the only method, formal mathematics may be a premature optimisation. More specifically, before achieving full formality, it may be necessary to be 'capable of being in uncertainties' [20]. Later reassessments of the QED proposal [43, 15] have had to deal with the fact that, by in large, it has not worked out as hoped.

In this paper, we have been inspired by the observation that Q&A is *both* a useful source *of* explicitly represented reasoning, *and* a useful (i.e., 'effective') modality *for* developing additional explicit reasonings. Q&A is a popular way for humans to learn from each other – including, in particular, about technical subjects like mathematics and computer programming. There are large existing 'traces' of Q&A dialogues available online, and many users actively engaging with these dialogues on a daily basis (both as querents and respondents). Q&A dialogues are, accordingly, a likely source of explicit knowledge – and they are, also, if our ansatz is correct, a potential modality for automated knowledge-building that could realise Turing's vision of machines that 'converse with each other to sharpen their wits' [41]. Naturally, dialogues between humans and automated agents/agencies would be a potential application. Systems like this could support activities ranging from automatic tutoring to mixed initiative proof and program-construction.

Argumentation techniques for agent research are surveyed in [9]. This work should be compared with Ganesalingam and Gowers's natural-language generating automated problem solver [12]. The kinds of natural language that are employed in Q&A dialogues is quite different from that employed in the textbook problems that are the focus of [12]. Nevertheless, at the level our proof-of-concept illustration above, we seem to have a good handle on the semantics of mathematical dialogues, and we should be able to support textbook reasoning as a special case. It should also be possible to map dialogue semantics to a theorem proving system (an LCF-inspired system like the one used by Ganesalingam and Gowers, would be one starting point but others should be considered as well).

Lastly, to evaluate systems working in this area – whether driven by argumentation, agents, theorem proving, simple machine learning heuristics, some com-

bination, or by some other approach entirely – we propose the simple knowledge-based computational benchmarking task SEMATCH, which is defined as follows. This challenge problem requires a system to match existing *questions* and *answers* selected from the Stack Exchange network. This can be solved in suitably-blinded forward and backward directions (Q↦A or A↦Q); that is, the program that is being tested will be presented with a sample of questions and answers, and will not be told which question matches which answer: it will then try to recover that information by reasoning about the sample's content. Accordingly, SEMATCH can be applied to routinely evaluate heuristics, without necessarily requiring a system that is capable of generating new answers or new questions. As a benchmarking problem for programs that reason about natural language, one could compare SEMATCH with the Winograd Schema Challenge [23], which is similarly open-ended as to methods of solution. However, the Winograd Schema Challenge currently requires expert intervention to generate new test questions, which follow a certain prescribed form. In the case of SEMATCH, a large corpus of ground-truthed data exists. Additional benchmarks related to the same ground-truthed data can be straightforwardly devised, e.g., to correctly sort the answers to a given question in order of empirical user ratings, based on their contents. Ultimately, after workouts with evaluation metrics like this, along with generative experiments, a system may be devised that can answer specific sets of domain problems correctly, and have its answers validated by human experts.

Acknowledgements

Martin and Corneli were support by Martin's EPSRC fellowship award "The Social Machine of Mathematics" (EP/K040251/1); Murray-Rust was supported by "SOCIAM - the theory and practice of Social Machines" (EP/J017728/2).

References

1. Aberdein, A.: The informal logic of mathematical proof. In van Kerkhove, B., van Bendegem, J.P., eds.: Perspectives on Mathematical Practices: Bringing Together Philosophy of Mathematics, Sociology of Mathematics, and Mathematics Education. Springer. Logic, Epistemology, and the Unity of Science, Vol. 5 (2007) 135–151
2. Anonymous: The QED Manifesto. Lecture Notes in Artificial Intelligence **814** (1994) 238–251
3. Barany, M.: '[B]ut this is blog maths and we're free to make up conventions as we go along': Polymath1 and the modalities of 'massively collaborative mathematics'. In: Proceedings of the 6th International Symposium on Wikis and Open Collaboration, ACM (2010)
4. Berners-Lee, T., Fischetti, M.: Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. Harper Information (2000)
5. Budzynska, K., Janier, M., Reed, C., Saint-Dizier, P., Stede, M., Yaskorska, O.: A model for processing illocutionary structures and argumentation in debates. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. (2014) 917–924

6. Bundy, A.: The interaction of representation and reasoning. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **469**(2157) (2013)
7. Buscaldi, D., Rosso, P.: Mining knowledge from Wikipedia for the question answering task. In: Proceedings of the International Conference on Language Resources and Evaluation. (2006) 727–730
8. Caprotti, O., Saludes, J.: The gf mathematical grammar library: from openmath to natural languages. In: Joint Proceedings of the 24th Workshop on OpenMath and the 7th Workshop on Mathematical User Interfaces (MathUI), Citeseer (2012) 49
9. Carrera, Á., Iglesias, C.A.: A systematic review of argumentation techniques for multi-agent systems research. Artificial Intelligence Review **44**(4) (2015) 509–535
10. Cramer, M., Koepke, P., Schröder, B.: Parsing and disambiguation of symbolic mathematics in the naproche system. In: International Conference on Intelligent Computer Mathematics, Springer (2011) 180–195
11. Ferrucci, D., et al.: Building Watson: An overview of the DeepQA project. AI magazine **31**(3) (2010) 59–79
12. Ganesalingam, M., Gowers, W.: A fully automatic theorem prover with human-style output. Journal of Automated Reasoning (2016) 1–39
13. Ganesalingam, M.: The language of mathematics. Springer (2013)
14. Ginev, D.: The structure of mathematical expressions. Master's thesis, Jacobs University, Bremen, Germany (2011)
15. Harrison, J., Urban, J., Wiedijk, F.: Preface: Twenty years of the QED manifesto. Journal of Formalized Reasoning **9**(1) (2016) 1–2
16. Hu, Y., Wan, X.: Mining and Analyzing the Future Works in Scientific Articles. CoRR **abs/1507.02140** (2015)
17. Juršič, M., Cestnik, B., Urbančič, T., Lavrač, N.: Cross-domain literature mining: Finding bridging concepts with crossbee. In: Proceedings of the 3rd International Conference on Computational Creativity. (2012) 33–40
18. Kaliszyk, C., Urban, J., Vyskočil, J., Geuvers, H.: Developing corpus-based translation methods between informal and formal mathematics [Poster of [19]] `http://cl-informatik.uibk.ac.at/cek/docs/14/ckjujvhg-cicm14-poster.pdf`.
19. Kaliszyk, C., Urban, J., Vyskočil, J., Geuvers, H.: Developing corpus-based translation methods between informal and formal mathematics. In: International Conference on Intelligent Computer Mathematics, Springer (2014) 435–439
20. Keats, J. In Rollins, H.E., ed.: The Letters of John Keats. Volume 1. Cambridge University Press, Cambridge (1958)
21. Lakatos, I.: Proofs and refutations: The logic of mathematical discovery. Cambridge University Press ([1976] 2015)
22. Lawrence, J., Reed, C., Allen, C., McAlister, S., Ravenscroft, A., Bourget, D.: Mining arguments from 19th century philosophical texts using topic based modelling. In: Proceedings of the First Workshop on Argumentation Mining, Citeseer (2014) 79–87
23. Levesque, H., Davis, E., Morgenstern, L.: The Winograd Schema Challenge. In Brewka, G., Eiter, T., McIlraith, S.A., eds.: Proceedings of the Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning, AAAI Press (2012)
24. Liu, J., Wang, Q., Lin, C.Y., Hon, H.W.: Question difficulty estimation in community question answering services. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2013) 85–90

25. Macagno, F., Walton, D.: Argument from analogy in law, the classical tradition, and recent theories. Philosophy and rhetoric **42**(2) (2009) 154–182
26. Martin, U., Pease, A.: What does mathoverflow tell us about the production of mathematics? In: SOHUMAN, 2nd International Workshop on Social Media for Crowdsourcing and Human Computation, at ACM Web Science 2013, May 1, 2013, Paris. (2013)
27. Martin, U., Pease, A., Corneli, J.: Bootstrapping the next generation of mathematical social machines. In Kuper, L., Atkey, B., eds.: Off the Beaten Track workshop at POPL, UPMC Paris, January 21, 2017, ACM (2017)
28. Mendes Rodrigues, E., Milic-Frayling, N.: Socializing or knowledge sharing?: characterizing social intent in community question answering. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 1127–1136
29. Miller, B.R.: Strategies for Parallel Markup. CoRR **abs/1507.00524** (2015)
30. Nuamah, K., Bundy, A., Lucas, C.: Functional inferences over heterogeneous data. In: International Conference on Web Reasoning and Rule Systems, Springer (2016) 159–166
31. Pease, A., Martin, U.: Seventy four minutes of mathematics: An analysis of the third Mini-Polymath project. In: Proceedings of AISB/IACAP 2012, Symposium on Mathematical Practice and Cognition II. (2012)
32. Pease, A., Lawrence, J., Budzynska, K., Corneli, J., Reed, C.: Lakatos-style collaborative mathematics through dialectical, structured and abstract argumentation. Artificial Intelligence **246** (2017, to appear) 181–219
33. Peldszus, A., Stede, M.: From argument diagrams to argumentation mining in texts: A survey. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) **7**(1) (2013) 1–31
34. Schneider, D., Witbrock, M.J.: Semantic Construction Grammar: Bridging the NL/Logic Divide. In: Proceedings of the 24th International Conference on World Wide Web, ACM (2015) 673–678
35. Schoenfeld, A.H.: Mathematical problem solving. Academic Press (1985)
36. Sowa, J.F., Majumdar, A.K.: Analogical reasoning. In Aldo, A., Lex, W., Ganter, B., eds.: International Conference on Conceptual Structures, Springer (2003) 16–36
37. Stahl, G.: Group cognition: Computer support for building collaborative knowledge. MIT Press Cambridge, MA (2006)
38. Sussman, G.J.: Why programming is a good medium for expressing poorly understood and sloppily formulated ideas. In: OOPSLA '05: Companion to the 20th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, ACM (2005) 6–6
39. Tanswell, F.: A problem with the dependence of informal proofs on formal proofs. Philosophia Mathematica **23**(3) (2015) 295
40. Tausczik, Y.R., Kittur, A., Kraut, R.E.: Collaborative Problem Solving: A Study of MathOverflow. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. CSCW '14, New York, NY, USA, ACM (2014) 355–367
41. Turing, A.M.: Intelligent Machinery, A heretical theory. Philosophia Mathematica **4**(3) ([1951] 1996) 256–260
42. Walton, D.: How can logic best be applied to arguments? Logic Journal of IGPL **5**(4) (1997) 603–614
43. Wiedijk, F.: The QED manifesto revisited. Studies in Logic, Grammar and Rhetoric **10**(23) (2007) 121–133