Knowledge sharing on Q&A fora: challenges of automated interaction analysis

Thesis report

P. A. Ruben (4273818)

Thesis Science Communication 15 EC [SL3521]

TU Delft

Knowledge sharing on q&a fora: challenges of automated interaction analysis

Thesis report

by

P. A. Ruben (4273818)

In partial fulfillment of the degree of Master of Science
at the Delft University of Technology,

Project duration:	February 2020 – September 2020	
Supervision by:	Dr. Éva Kalmar,	TU Delft
	Dr. Caroline Wehrmann,	TU Delft
	Anna Labetski, TU Delft	
	Dr. Ken Arroyo Ohori, TU Delft	

TU Delft

Abstract

Q&A fora have developed into a precious tool for online knowledge exchange. Being community-driven, openly accessible and free, getting an answer to a complicated question can nowadays be a matter of minutes. This research investigates the specific case of *gis.stackexchange*, a technical Q&A forum used by *Geographic Information Systems* professionals. A central topic is the interaction between users. Based on existing research, an automated approach is elaborated and a big data analysis is performed. Also, a link between the outcomes and self-reflective behaviour found on the community meta forum is established. Factors positively influencing interaction as defined by the approach are identified (implication of the author(s), presence of images and code snippets). Furthermore, weaknesses such as the low share of interactions associated to an alteration (of the original post) are also discussed. These outcomes eventually lead to a series of recommendations for the forum itself and related formats.

P. A. Ruben (4273818)
Online, August 2020

Acknowledgements

In social research, being inside a phenomenon often does not make it easier to understand it. So-called heuristic research requires a lot of self-reflection, an aspect which is frequently hindered by biases. But this is not something limited to academia - we also encounter it in our daily life. And, unfortunately, the time in which the research of this thesis was conducted perfectly illustrates this. The SarS-COVID-19 pandemic which affects our lives since the start of the year is especially hard to understand as one lives inside it. Taking distance is necessary, but the deep implications the resulting phenomena have for nearly all aspects of our lives make it no easier to do so. The fate of social research suddenly became of importance far beyond specific academic research.

Although the pandemic was not directly part of the research, it would be wrong to claim that this thesis was not affected by it. As already stated, there were interesting times - thinking about the application of social research in a very critical context. And about the role of online communication, about the impact of online communication on studying. On the other hand, there were also moments of solitude. Moments of tiredness, of feeling far from a reality I was afraid of never finding back. But although we could not be as present for each other as we wanted, I still felt the support of many people who I would like to thank.

Let's start with the ones most directly involved in this research. Éva, Anna, Maarten, Caroline and Ken - thank you so much for accompanying me during this adventure. Sharing your personal and academic interest in the subject was a huge help to me. While I knew most of you from the respective masters, working on something slightly off your teaching comfort zone made this experience very human. Too often, ice is left to be broken at the end of classic study courses. This time it felt different!

I would also like to thank the numerous friends I made during these five years in Delft. Without you, my experience in Delft would have been completely different. Thanks for having made these times so special! Gijs, Dion, Stendert, Tess, Steven, Gitta, Meylin, Gabo, Davey, Leyden, Lucio, Laurens, Emiel, Camera, Ela, Josine, Rebecca, and to the many people I forgot to mention in this list: thank you!

Last but not least, a big thank you goes to the people closest to me. To my beloved Camille. To my father Mario, to my mother Maren and my brother Shura. As well as to my grandmothers Isa and Rosalinde, to my grandfathers Gerhard and Hans-Jürgen, to my uncles Tino and Tilo, to my aunt Egle and my cousins Laura and Miegle. Thank you for all the support in the best and worst times of this year-long learning. Thank you for having helped me achieve where I am now, the end of my formal studies and the start of an entirely new life chapter!

P. A. Ruben (4273818)
Online, August 2020

Contents

List of acronyms

SE fora managed by StackExchange
gis.SE The forum gis.stackexchange.com
GIS Geographic Information Systems
QGIS Quantum Geographic Information System
FOSS Free and Open Source Software
ESRI Environmental Systems Research Institute
HI High Interaction
Q&A Question and Answer
CMC Computer-Mediated Communication
LCPs Learning Centered Principles

1 Introduction

The education children experience at school has considerably evolved in the last decade. A core ingredient of which the positive effects are now recognised is interaction which is a central topic to elaborate school programs. As formulated by Duschl (2008) for the United States: “Since the first NSF-funded era of science education reform in the 1960s and 1970s, we see a shift in views [...] to science teaching focusing on the management of learners’ ideas, access to information, and interactions between learners”.

With the recent emergence of the internet and corresponding learning formats, one might ask whether this trend is continued using modern means. In fact, the internet has intrinsically always been a social endeavour. One can thus expect to find interaction. But is it similar to more traditional forms? And more importantly, how can it be defined and analysed? These are two central topics to which this master thesis relates.

1.1. Real-time knowledge sharing in the 21st century

In the past decade, the share of people using the internet has seen a steep increase. From the fraction of a per cent in 1990, it reached 53.6% in 2019 and the trend is still increasing[1]. With the ability to communicate and access knowledge nearly anywhere and at all times, information-sharing habits have also evolved. The concept of user-generated content emerged rather early in the history of the internet with the concept of the wiki being formulated as early as 1994 by Ward Cunningham[2]. Nowadays, such community-built knowledge resources, of which *Wikipedia.org* might be the most famous one, have proven to be popular. At the time of writing, *Wikipedia.org* ranked #13 in terms of global internet traffic and engagement on the Alexa Rank[3].

Even before the existence of wikis, user-generated content systems were already present in different forms. Bulletin Board Systems which appeared before the democratization of internet access might be regarded as an early form of these. The first internet fora appeared at a similar time, with, for instance, *Delphi*[4] in 1981. The *UseNet* system conceived in 1979 is an additional system with the particularity of being distributed.

Harper et al. (2008) performed a comparative study of several so-called Question and Answer (Q&A) sites, sometimes referred to as ’knowledge markets’[5]. Within these websites, it identified three types:

- Digital reference services: which are online but rather analogue services giving users the possibility to ask librarians.
- Ask an expert services: such websites provide a database of experts to which users can ask questions. A community thus exists, but interactions are very topic-oriented (e.g. an expert is assigned to a question based on the field).
- Community Q&A sites: as the most modern form, these services are typically more open, meaning they have a less predetermined structure or role-based organization. Some sites do nevertheless rely on moderators and statuses one can earn using a gamification system.

[1] https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2019.pdf
[2] https://en.wikipedia.org/wiki/History_of_wikis
[3] https://www.alexa.com/siteinfo/wikipedia.org
[4] https://www.delphiforums.com/index.ptt
[5] https://www.forbes.com/sites/jenniferhicks/2011/06/27/the-rise-of-the-knowledge-market/

Furthermore, a difference between free and paid services is mentioned (also referred to as market-based and non-market-based). In fact, some websites are driven by volunteers only, while others provide financial incentives (askers pay answerers). The comparative study itself was performed on a number of different platforms (Library service, Google answers, AllExperts as well as Live QnA). The main findings are that paying does increase the reliability of answer quality - but that the community (volunteer efforts) contributes considerably too (Harper et al., 2008).

Within Q&A fora, Gazan (2006) identified two forms of users: synthesists and specialists. Investigating a system called *Answerbag*, it was observed that some users are not specialists in a matter but gather information from several sources to answer a question in that matter. For that specific community, it was furthermore noted that answers from such synthesists achieved more popularity - showing the potential of such roles. This suggests that Q&A fora might not necessarily be about connecting a question to the right specialists, but also about getting help from people with more ability in *finding* solutions or pointing in the right direction.

Additionally, it should be noted that Q&A processes can also be found outside of services specifically designed for it, such as on social media platforms. As described by Panahi et al. (2012), a key characteristic of social media is that the same users which animate peer to peer contact also generate the content themselves. Further characteristics include the 'networking' aspects such as statuses ('friendships', point rewards, etc.), user-friendliness and finally, the possibility to share multimedia content beyond the mere text. Q&A processes occurring in this less formal context (sometimes referred to as *social* Q&A) might cover a bigger range of topics but have varying success (Wang et al., 2017). The Alexa ranks (for global internet traffic and engagement) of widespread examples such as *YouTube.com* (ranking #2) and *facebook.com* (#4) or *reddit.com* (#20 and has some traits of traditional fora) nevertheless stress their relevance.

In-between these fora and social networks, there are forms which build on both principles but intend to create a knowledge base. The most famous of these is probably *stackoverflow.com* which ranks #45 (for global internet traffic and engagement) and has a rather specific user community. Part of a bigger group of websites, the network of SE, it is a Q&A website on which programmers can share and solve the challenges they face. Since its creation, it has collected no less than 19 million questions in 11 years.

Based on previous research, Dondio and Shaheen (2019) describe Stackoverflow.com as being a new "computer science department where people go to learn" in which learning fulfils needs in real-time. This is linked to a new learning paradigm associated to "digital natives" which solves problems by finding solutions online and thus learns "on the go".

1.1.1. Introducing the StackExchange network

Created in 2008, Stackoverflow.com is one of the most popular question & answer sites for programmers. Being open in a similar fashion to Wikipedia and Reddit, it is more result-focused and became a central reference for many IT-professionals. From 2009 on, creators Jeff Atwood and Joel Spolsky expanded the concept by creating a network with the name SE. By doing so, the successful concept was expanded, allowing other communities to create similar sites. While the critical mass is still made of computer programming ones, other scientific (e.g. mathematics, physics, geographical information sciences) and less scientific topics (e.g. cooking, travelling, religions, etc.) are covered too.

One should be aware that the SE 'system' goes well beyond content regulation as it also englobes social aspects. The gamification process is identical for all fora and is based on recognition by other members (up- and downvotes) which allows gaining badges and privileges (right to comment, to vote, to moderate, etc.). The creation of new SE fora is subject to a similarly clear process: ideas can be submitted on the *Area51.stackexchange* forum. If a critical number of members with sufficient endorsement (reputation scores) are convinced, the page goes into the beta stage. From there on, any forum consists of a meta-discussion page (for the organisation of the forum itself) and the actual Q&A forum. Once several key metrics have been achieved, the site leaves the beta stage and becomes a full SE site (a process which gis.SE successfully fulfilled). More details of the structure of the forums are provided in part 3.2.

Among the websites of this network, the one this research focuses on is gis.SE. This forum which was initiated in 2010 and, with 46,000 daily visits, is the 29th most visited of the 175 sites of the network (at the time of writing). Focusing on Geographic Information Systems (GIS) technology, it is rather technical and built a knowledge base of 127,000 questions and 145,000 answers.

Content-wise, the forum states “This site is all about getting answers. It’s not a discussion forum. There’s no chit-chat.” and defines[6] its scope as follows:

- Specific questions concerning geographic information systems and science
- Real problems or questions that you’ve encountered

Also, the following are explicitly excluded:

- Anything not directly related to geographic information systems
- Questions that are primarily opinion-based
- Questions with too many possible answers or that would require an extremely long answer

To position the gis.SE forum within its context, frameworks as provided by Stanoevska-Slabeva (2002) or Bos et al. (2007) can be utilised (for more details, see part 4.1). These ones highlight three characteristics of the forum. First, that it is a *(virtual) community of practice*, thus based on a professional community. Second, that it is an *interest community* which is built around a shared interest in and expertise of a specific technology. Finally, this is also related to the *open community contribution systems* trait: anyone can get involved and the core product, the knowledge base, is freely shared.

This positioning is also valid for other technical fora within the SE ecosystem. Fora such as StackOverflow, SuperUser, Ask Ubuntu, Unix & Linux (and many more) are also built around both an *interest community* and a *community of practice*. For some other fora such as English Language Learners, SeasonedAdvice (on cooking) and Motor Vehicle Maintenance & Repair, the *community of practice* are less present as the users have a less professional but hobby profile (although the level of expertise of some hobby practitioners might doubtlessly be high too). Some other fora such as Role-playing games, Anime & Manga or Arqade are also *interest communities* but are additionally related to ’virtual worlds’ in the entertainment industry.

1.2. The state of existing research on StackExchange and identified gap

Considerable research has already been conducted on some SE forums (mainly *StackOverflow*). However, it generally focuses on question-answering from a general, mainly quantitative point of view rather than analyzing the interactions and their content. Often, the user votes are taken as an indication of quality while this is actually questionable. While some studies take the thread as the unit of analysis, a study of the literature did not deliver any result regarding studies investigating interactions/discussions on SE fora. This obviously raises the question about whether interaction is actually a critical topic. Before continuing, a distinction between the knowledge types a question addresses should therefore be made[7].

- *Declarative knowledge* which is relatively easy to share but does not connect to prior/context knowledge. This might, for instance, be a list of commands to execute to reach a specific goal. A typical example in the case of gis.SE, are situations in which users ask for help in debugging their code and their errors are directly identified. Such situations do not aim at discussing the general approach but mainly at identifying glitches in the reasoning. An extract of such a situation can be found in Figure 1.2.
- Procedural knowledge which, in opposition to the previous one, provides information on the system to which one is confronted. This is generally harder to share and is linked to the development of specific and reusable insights in the systems one uses. Typically, these are situations in which *understanding* the method used and suggesting a better one is required to answer the question. An example, namely an answer given to an issue with interpolation can be found in Figure 1.1. Note that the solution is further discussed in the comments shown in Figure 1.3.

[6] https://gis.stackexchange.com/tour

[7] terminology used by Ummelen (1997); knowledge typology being a topic of research itself, more information is provided in part 2.3

1 Answer

Active | Oldest | Votes

0

If you reduce the cell size it seems to fill the missing values, and a way to guesstimate roughly 100km * 100km is to check the sqrt of the area of pixels, here it varies from 80km to 128km. (They are always taller than they are wide so might depend if balancing area or dimension is more important)

I don't know why it gets missing values with lower res, however.

```
res(parGrid) <-  1.2 ## ~about 100km?? Not sure if this is correct.
sqrt(area(parGrid))
class      : RasterLayer
dimensions : 39, 103, 4017  (nrow, ncol, ncell)
resolution : 1.2, 1.2  (x, y)
extent     : -72, 51.6, -67.8, -21  (xmin, xmax, ymin, ymax)
crs        : +proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0
source     : memory
names      : layer
values     : 82.99557, 128.4049  (min, max)
```

Then I get what seems reasonable:

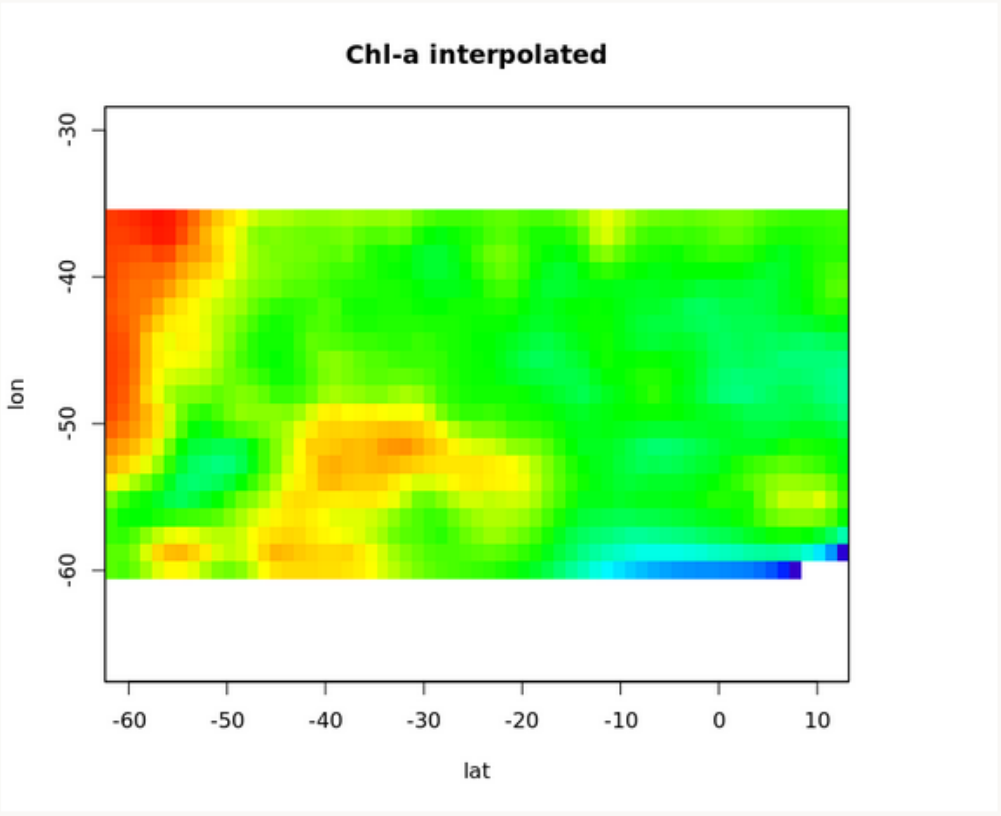


Figure 1.1: Extract of a thread in which rather procedural knowledge is shared[8].

One might note that in the case of declarative knowledge, a satisfying result might be reached without interaction beyond the question-answer pair. This is at least the case if the user who posted the question is looking for a viable solution without further wanting to understand the implications. If, however, the matter is to find a more subtle solution for a specific task, such as in the example for procedural knowledge, mindset(s) of different users might be relevant and need to be expressed (and

[8]https://gis.stackexchange.com/questions/327619/r-kriging-raster-returns-na-for-some-prediction-points

ideally combined) to formulate a complete answer. An example of such a situation and linked to the first example (Figure 1.1 on procedural knowledge) can be found in Figure 1.3. One might also find situations in which the mindset(s) of different users are shared without interactions occurring. This does imply that the choice of which answer to use stays with the reader of the knowledge. This process of choosing is, however, a form of procedural knowledge too. A situation in which the mindsets are combined (or at least a tentative to do so is done) and the process behind it is shared is thus providing the reader with more procedural knowledge. Furthermore, the potential overlap between mindsets (and thus partial duplicity) might also be avoided by co-construction and thus interaction.

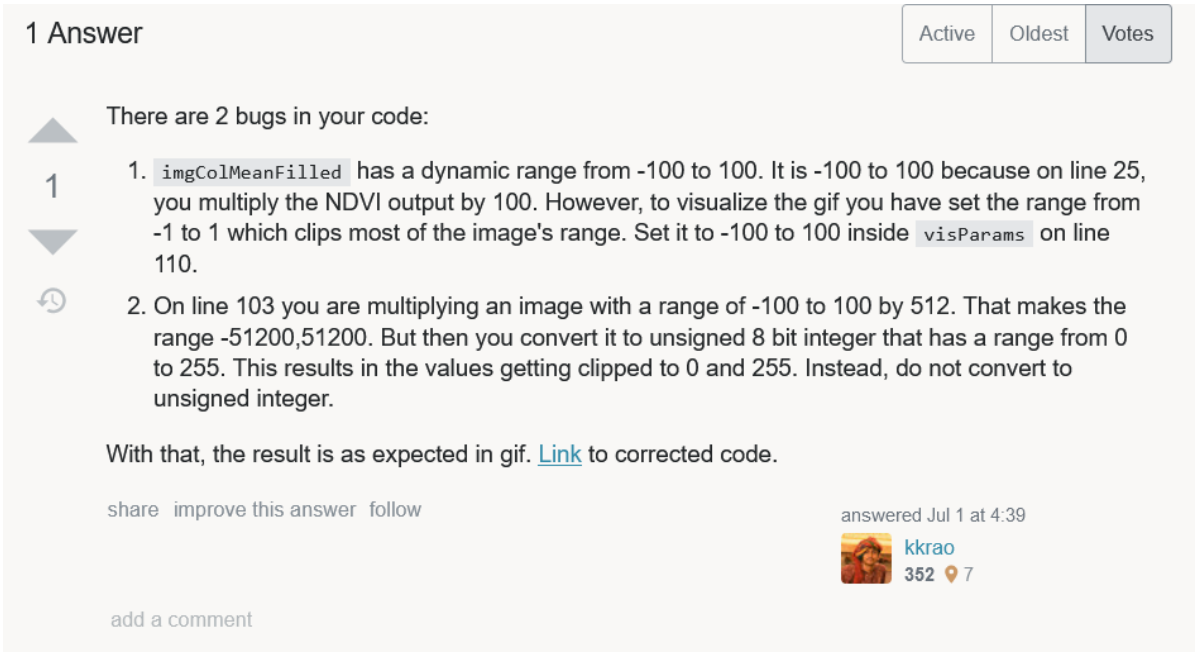


Figure 1.2: Extract of a thread in which rather declarative knowledge is shared[9].

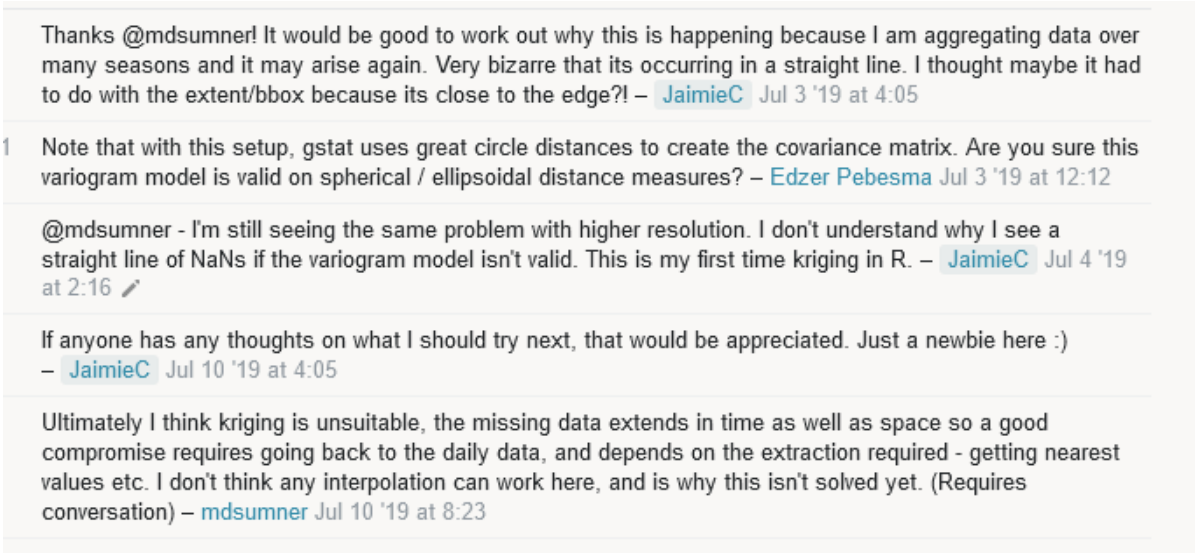


Figure 1.3: Comments on the answer shown in 1.1.

To put it in a nutshell, the role of interactions and their content is thus twofold:

[9]https://gis.stackexchange.com/questions/366451/sentinel-ndvi-timelapse

- on the one hand, interactions can be an indicator of shared understanding, of co-construction processes.
- on the other hand, interactions also produce valuable additions to the knowledge base on the website. Due to the openness of the SE fora, they produce outputs which other users might later retrieve. Therefore, understanding is shared beyond the immediate participants of the interaction.

The investigation of interactions and their roles within fora such as gis.SE is the gap which this thesis aims to address. Before addressing it, the relevant topics in the academic literature are addressed in more detail in the theoretical framework formulated in chapter 2.

Beyond the Q&A websites themselves, one might note that ideas to reuse the openness and user-generation of content in other fields has arisen. A good example is an approach formulated by Tennant et al. (2017) which tackles the weaknesses of traditional peer-review processes in academia. It sees the open review system associated with the knowledge base of SE fora as an approach from which academia could benefit. Rather than being journal based, they argue that peer review should be community-based, making it an intrinsically open and more transparent process. To establish such a system, three major aspects are highlighted: quality control, certification and incentivization. This illustrates that research into the dynamics occurring *around* the actual knowledge base might be of interest beyond the use case of the gis.SE forum itself.

1.3. Problem statement and scope of the research

Based on the existing gap in the scientific literature on interaction within SE fora and the potential reuse of the review system for other fields, the following problem statement has been established: “For effective learning, Q&A fora should not only share solutions but also facilitate constructive interactions sharing the mindset(s) of the solution(s) and encourage thread improvement. Based on cues of the existing situation, how can fora further facilitate such interactions?”. The word “mindset(s)” is used here as a term for the exchange of different views (and the development of shared understanding) in the cases of procedural knowledge (as mentioned earlier, see Figure 1.3).

Rather than focusing on the gis.SE forum and community from a general point of view, the research in this report thus focuses on the nature and the occurrence of interactions within which the solution development process is shared. These interactions are expected to occur within the comments which thus play a central role. Furthermore, this study focuses on the outputs produced, thus the content which website visitors can eventually access (more specifically, the 1st of June 2020 has been taken as a date of reference). From a general paradigm, this study will thus perform a *data analysis*.

To perform this analysis, mixed methods will be employed. On the one hand, quantitative data analysis in order to identify high scale phenomena and to develop an automated approach. On the other hand, qualitative analysis is used to provide a human coder perspective on the content and the research choices. By combining these ones, the study is expected to confront itself with two challenges: first, combining algorithms with human coding in communication sciences and second, getting insight through a computer-mediated environment (through the forum gis.SE rather than by interacting/observing the people involved directly).

1.4. Research questions

The main research question of this research is therefore formulated as follows: **In which cases do threads on gis.stackexchange share the ‘mindset(s)’ necessary to solve the question by facilitating constructive interaction?**.

To address it, several sub-questions have been formulated. Each of them are mainly but not exclusively addressed in respectively chapter 4, 5 and 6:

1. What are the characteristics of the gis.SE forum/community?
2. *a.* How can threads in which constructive interaction (1) occurs and (2) discloses the ‘mindset(s)’ linked to the answers be identified in an automated way? *b.* Which factors influence the occurrence of such threads?
3. *a.* To which extent does the community appreciate threads sharing such ’mindset(s)’? *b.* Are these threads in conflict with the enforcement of forum rules?

Before the chapters addressing these problems, further insights into the scientific background is shared in chapter 2 and the approach is highlighted in chapter 3.

2
Background

This chapter introduces three scientific fields of relevance for this study and upon which the research was eventually built. Online knowledge sharing is a topic for which several theoretical approaches exist. In scientific literature, links between these appear to be rather weak. This chapter will therefore consist of three parts which have been linked by the author.

A first part establishes a link between online knowledge sharing and social learning. After discussing the role of interaction in traditional offline education, existing literature on new forms of learning using websites such as SE fora is introduced.

A second part takes a closer look at the human-machine duo formed by the users and the Q&A forum on which they interact. The multi-dimensional impact of knowledge exchange occurring with computer-mediated communication (CMC) is discussed there. In fact, the presence of CMC does not only impact interaction itself but also poses challenges to methods analysing it.

A final part is dedicated to the content of communication, namely the knowledge exchange itself. Several approaches to classify knowledge into types are introduced and the challenges of applying these are highlighted too.

2.1. Social Learning and online interaction

2.1.1. Social learning in online environments

Within the field of social learning, the term 'learning in the wild' was introduced by Kumar et al. (2018) to describe a form of decentralised learning as occurs on platforms such as Reddit, Twitter and StackOverflow. Decentralised because, in contrast to formal courses and degrees: "It is crowdsourced learning, but not of curricula or courses, but in conversation-sized pieces, based on crowdsourcing interest in answering just-in-time questions." Therefore, there is no direct goal beyond the knowledge itself, there is no test to pass or certificate to achieve[1]. Similarly, there is not a central planning and content coordination as within university curricula - rather, the contributors (as SE is an open system, thus anyone) decide which questions are asked and if/how they are answered.

2.1.2. Social Learning and the importance of interaction

Social learning itself is a field established in the 1970s of which Bandura and Walters (1977) were two pioneers. It states that direct experience in a social context positively impacts learning processes:

"In the social learning system, new patterns of behaviour can be acquired through direct experience or by observing the behaviour of others. The more rudimentary form of learning, rooted in direct experience, is largely governed by the rewarding and punishing consequences that follow any action." (Bandura and Walters, 1977)

In sum, the real setting of the learning exposes the 'learner' to the behaviour of others and provides a rather direct feedback loop facilitating direct observation of one's actions. Moreover, the authors state: "It is commonly believed that responses are automatically and unconsciously strengthened by their immediate consequences." This puts social learning in contrast to other, less interactive forms such as exercise books where feedback might be obtained (if the solutions to the exercise are provided) but stays at the individual level.

Mulder et al. (2002) states that *shared understanding* is key to the group learning process. In order to cooperate, participants must, for instance, understand each other and their concepts must overlap.

[1] NB: SE fora have a reputation system - but unless this one serves outside of the SE ecosystem (e.g. during job applications), its advantages are limited to the system itself

Implicitly, this means that throughout the process the understanding is updated. Furthermore, conceptual learning (the exchange, reflection and refinement of concepts), feedback and motivation are stated as key ingredients to reach this. The fact that conceptual learning can be split in four steps (accretion, tuning, restructuring and co-construction) further illustrates the need for interaction to achieve *shared understanding*.

It should further be noted that these approaches are not mere theories but were also implemented in formal educational programs. As Duschl (2008) states, social learning goals played a major role in the science education reforms since the 1980s. While science curricula traditionally focused on what one needs *to know* to do science, they gradually shifted focus to what one needs *to do* to learn science. The goal is not anymore to have insights in the stage of science, but to be able to actively participate in its development. In practice, this means that student learning became *active productive* (not just accumulating knowledge but applying it), *integrated* (i.e. in several domains, e.g. social and cognitive processes) and that student thinking should be *made visible* and *monitored* by teachers (Duschl, 2008).

2.1.3. Social Learning Analytics

A key method related to 'learning in the wild' is *Social Learning Analytics* which the focus is on "processes in which learners are not solitary, and are not necessarily doing work to be marked, but are engaged in social activity". By using online platforms facilitating crowdsourced learning (and thus being open), learners do not only consume content but also create new one. In several cases including SE fora, Twitter and Reddit, their interactions (asking, reacting, editing, voting, etc.) with others leave traces which others can later experience.

A key theory to performing *Social Learning Analytics* is earlier research conducted by Mercer (2007). These scholars established and developed the distinction between three types of discourse, originally in formal education settings:

- *Disputational Talk* in which disagreement takes place and few attempts are made to find common ground or a constructive outcome (e.g. short criticism 'I don't agree').
- *Cumulative Talk* in which participants build upon each other in an uncritical way. Examples are repetitions and confirmations (e.g. 'I agree with all you said').
- finally *Exploratory Talk* which "represents a joint, coordinated form of coreasoning in language, with speakers sharing knowledge, challenging ideas, evaluating evidence and considering options in a reasoned and equitable way. The children present their ideas as clearly and as explicitly as necessary for them to become shared and jointly analysed and evaluated. Possible explanations are compared and joint decisions reached. By incorporating both constructive conflict and the open sharing of ideas, *exploratory talk* constitutes the more visible pursuit of rational consensus through conversation." (Mercer and Littleton, 2007)

In this study, the most comprehensive angle on social learning is *Exploratory Talk* as it is the form in which most interaction, most social exchange occurs. While the first two forms might also be seen as "rewarding and punishing consequences" (Bandura and Walters, 1977), only the third form makes sure that participants fully understand consequences such as criticism. This work was extended in Ferguson (2009) where evidence of similar talk in asynchronous dialogue was identified. Subsequently, Ferguson et al. (2013) also developed a method for automatic classification of discussion into *Exploratory* and *Non-exploratory talk*. Applied to the discussion logs of several online conferences (thus synchronous dialogue), it builds upon a coding scheme which achieved a Cohen's Kappa of 60% on the inter-annotator agreement (see Figure 3.4. The method itself uses self-training from labelled features: the classifier is first trained with labelled data and then expands into unlabeled data, prioritising data on which is similar and on which it is thus more likely to produce correct results. Overall, this approach demonstrated an accuracy of 79%, but it must be noted that the discriminating features obtained (the training results of the classifier) are highly specific to the nature of the data used.

Originally developed for educational and academic settings, these approaches were first adapted for asynchronous online fora in a study by (Kumar et al., 2018) on four 'Ask' communities of Reddit ('subreddits'). The coding scheme of Ferguson et al. (2013) was adapted in several iterative cycles. Subsequently, it was also applied to Twitterstorians, a part of Twitter (Kumar and Gruzd, 2019). More recently, Sengupta and Haythornthwaite (2020) developed a coding scheme to classify SO comments.

Interestingly, this one builds upon the previous research but focuses on general categories, rather than on identification of *Exploratory Talk*.

The Social Learning approach clearly shows the benefits of interaction, for which a number of requirements must be met. On the one hand, learning must take place in a social environment, on the other hand, the learning approach must facilitate the creation of shared understanding and eventually lead to co-creation. Existing research has further established the term *Social Learning Analytics*, showing that the type of interaction can further be refined and applied in both manual and automated analysis of online platform contents.

While this approach is rather strong from a qualitative point of view, it falls short from a quantitative one. Clues to classify content are provided, but little indicators are given beyond the unit of analysis. The study conducted by Sengupta and Haythornthwaite (2020) on SO, for instance, classifies single comments but does address the discussion layer. Therefore, it is complementary to the second approach, which looks more specifically at the impact of CMC platforms on social interaction and how they build networks.

2.2. The impact of computer-mediated communication (CMC)

2.2.1. Openness and role of participants

To clarify the changes induced by exchanging knowledge in online rather than traditional distance learning environments, a framework introduced by Moore (1989) proves particularly handy. This one distinguishes between three types of interaction: *learner-learner*, *learner-instructor*, *learner-content*. In the case of Q&A websites such as gis.SE, such frameworks are, however, not applicable. As mentioned in parts 1.1.1 and 2.1.1, these are a new form of more open, decentralised and less formal learning. The mediation is done to a higher degree by a system (e.g. the software on which all SE fora rely) and to a lesser degree by humans (although human moderators do still exist and play a critical role). The line between traditional learner and instructor roles is thus blurred as the content, the knowledge base, is created both by people asking and answering questions (as well as by the ones posting comments). As

2.2.2. Approaches to online interaction using CMC

While the applicability of the framework introduced by Moore (1989) is limited, Hillman et al. (1994) added a fourth component which is highly relevant: learner-interface interaction. In fact, users of a system such as a Q&A website must necessarily interact with the technology, the tool as it is a prerequisite to participate in the system. If one feels uncomfortable with the interface, for instance, then interaction using that interface is rather unlikely to happen. According to Chou (2004), two relevant theoretical principles to increase learner-learner interaction in computer-mediated interaction are Learning Centered Principles (LCPs) and constructivism. Although these originally apply to a more traditional form of (formal) distance education, these ones provide clues on the differences induced by computer-mediation.

LCPs are a framework englobing a number of areas of learning (cognitive/metacognitive, motivational/affective, developmental/social and individual factors). They base on the consideration that learners operate holistically (thus based on their individual characteristics), behave based on their perceptions and the interpretation thereof and that they follow a process of dynamic change and growth (Chou, 2004).

In order to apply these, Chou (2004) suggests using Jonassen et al. (1995)'s four constructivist attributes for building learning systems. The first attribute is *context* and refers to the real world in which learners carry out their tasks. *Construction* emphasises the fact that knowledge is best acquired by referring to their own experiences. The two last attributes are probably the most important ones with regard to social learning and interaction introduced earlier in part 2.1.1: *collaboration* is about learners being exposed to multiple perspectives of other learners, and finally *conversation* is about learners engaging with each other to facilitate the actual learning activities (scheduling, informal discussions, etc.).

Although the work of Chou (2004) does not actually describe the impact of CMC on learning, this approach shows us that a number of aspects which can be deemed fulfilled in face-to-face education (as the learners share the same geographical location and thus to some extent the same context) require special attention. This is in line with the observations of Kreijns et al. (2002) who describe taking social

interaction for granted and the lack of attention to the social psychological dimension (especially for non-task-oriented discussions) as two major pitfalls. To tackle these, Kreijns et al. (2002) suggest the usage of *social affordances* on the one hand and of a *group awareness widget* on the other. For the first one, five levels based on the concept of *teleproximity* were developed (see Figure 2.1). This one states that group members perceiving the presence of each other is a prerequsite for interaction, and should thus be achieved at each level. The second suggestion, the *group awareness widget* resemble the approaches described for *Social Learning Analytics* earlier (see part 2.1. However, it goes further as it should also include user's activities (for instance, the task one works on - e.g. writing a paper or visiting a digital library). By sharing this information with all users, interaction of participants is expected to be stimulated (Kreijns et al., 2002).

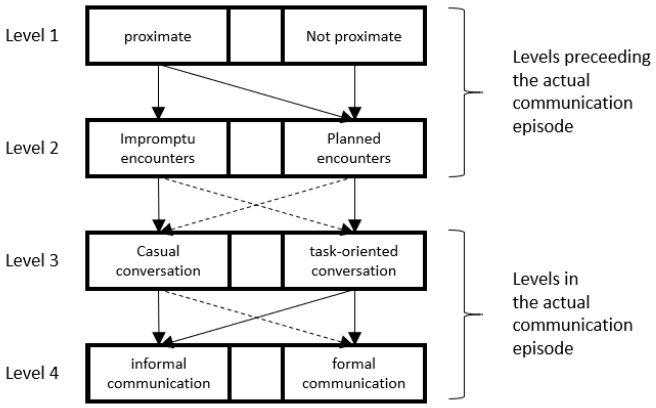


Figure 2.1: Five levels on which the concept of *teleproximity* should be applied (Kreijns et al., 2002). The lines represent the impact of different factors and their aspect the strength of it (dashed for weak, full for strong).

In a more recent study, Ioannou et al. (2015) conducted an experiment to compare learner interaction using two formats related to the SE format. The first one is the wiki format (the output document has a more prominent position than the communication tools) for which results suggested a condensing process. This is in contrast to the threaded discussion format (the communication tool has a more prominent role than the output document) for which results suggest an expanding nature. In other words, the final output is similar but the threaded discussion resulted in more verbose interaction.

In a study focusing on the content rather than the interaction, Dondio and Shaheen (2019) replaced traditional higher education exercises by *StackOverflow* questions. By exposing a group to the new and another to the old situation, the effectiveness of the forum as a learning complement was investigated. The results were slightly better for the new approach using the forum, but not sufficient to identify a clear trend.

Overall, these studies show that (online) learning and interaction using CMC can be fruitful but is highly influenced by the format. While the addressed topics (e.g. education in a specific technology) and the presence of human moderators (e.g. teachers, forum moderators) might be similar, the social interaction is highly influenced by the user interface and format used. By having a mixed format, SE fora position themselves in between wiki and fora, giving the user the opportunity for relatively verbose interaction (comments and chatroom) while keeping the output, the knowledge base and the contributions to it (the Q&A core), central. This balance illustrates the challenges to produce valuable content (which requires interaction but also condensation).

2.2.3. CMC and inherent structure: Social Network Analysis

The usage of CMC has the inherent advantage that user actions can be logged and used for analysis of the social network. Such activities can be performed in near real-time, creating an inherent link to *Social Learning Analytics* (see part 2.1.3). A concept which which is related to this (and thus to some extent to social learning) is the one of the *reflective practitioner* as introduced by Schön (1987). This one represents the action of reflection on one's actions, which eventually allows to engage in continuous learning. In the approach of Schön (1987), there are two types of reflection, one is *knowing in action*

and the other one is *research-based theory*. The usage of data to monitor a social network activity can be seen as a form of the second type. However, the possibility to use it in a nearly instantaneous way does also link it to *knowing in action*. Beyond the stimulation of formal and informal activities suggested by Kreijns et al. (2002) (see part 2.2.2) which might lead to reflective interaction, so-called *Social Network Analysis* thus offers an additional tool for relective practice.

Among the methods found in previous research, *Social Network Analysis* represent persons/participants as nodes and the relationships between those as edges (Helms et al., 2016). Hereby, a network graph is created and mathematical methods can thus be used for analysis. An example applying such network graphs to SO is the research led by Menshikova (2018) in which overlapping user expertise (tags which do not belong to the same cluster) is identified and its impact is analysed. This focus on thematic expertise does, however, limit the scope of the research to the tags of the questions answered by users.

Nevertheless, substantial research has also been conducted on interactivity using *Social Network Analysis*. A recurring concept, of which Henri (1992) has been one of the pioneers in computer-mediated communication. The basic pattern consists of three steps: (1) someone formulates and sends a message; (2) another one replies to this message (3) the author of the first message replies to the second message. This implies that interaction consists of at least three actions; the process can be represented schematically in the following manner:

A==>B==>A

While this might hold for conversations with only two participants, the situation of online (and open) fora such as gis.SE is much more complex. This is related to the fact that there are more participants. Messages are rarely 'sent' to only one person (as anyone can read them). Also, they are usually read and responded by more than one other person. The patterns that can arise from interactivity can thus take many more shapes. An approach to build graphs taking this into account is proposed by Manca et al. (2009).

However, it provides only limited clues on how to identify the actual patterns (to whom a message is addressed, whether it replies to a previous one). This is especially critical in situations where input data is unstructured (in opposition to threaded). Threaded fora often make the relationship between messages and thus participants explicit as its messages must be posted in a given hierarchy (e.g. message trees). In contrast, it is much harder to deduce the relationships when using unstructured data which consists of a list, where posters do not necessarily reply to a specific message. Generally, timestamps are still available and it is possible to order messages by creation date. Several methods to deduce relationships do then exist and must be selected depending on the specific situation. Helms et al. (2016) conducted a literature review which results in an overview of several such methods (also see Figure 2.2):

- everyone interacts with everyone (method 1)
- every participant interacts with the first and previous participant (method 2)
- every participant interacts with the previous participant only (method 3)

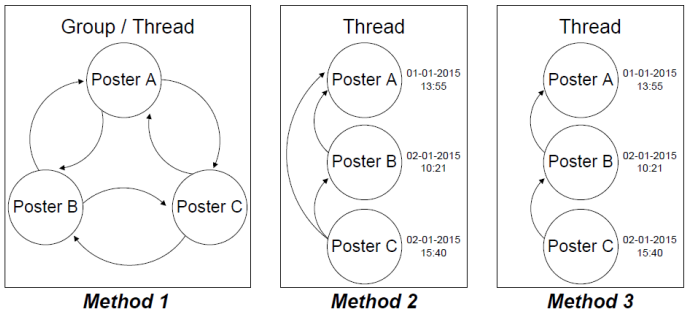


Figure 2.2: Methods to structure unstructured data identified in the literature by Helms et al. (2016).

Beyond these approaches, Helms et al. (2016) suggests a further one developed within the research presented in the same paper, namely the analysis of the *Hallo!* forum of the Dutch Chamber of Commerce. This one is established upon three rules (an illustration can be found in Figure 2.3):

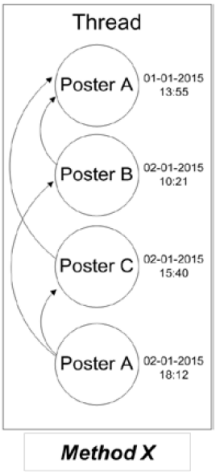


Figure 2.3: Method developed by Helms et al. (2016) to structure unstructured data.

- The first message does not build any relation to other users as it has only a broadcasting nature.
- Any reply posted by another user builds a relationship to this first message.
- If the thread starter replies in his/her turn, then relations are built to all messages who had previously replied.

In a similar study conducted by Petrovčič et al. (2012), several options in between method 1 (a message replies to all earlier messages) and 3 (a message replies only to the previous message) were explored. These are formulated as *replying to k earlier messages*, where *k* can either be a number or a time frame. In both cases, a constant or a variable (e.g. derived from the time passed or the number of messages since the first post) can be used. Furthermore, explicit mentions of previous users in the message (e.g. quotations) were also mentioned. Despite their unambiguous nature, these have the disadvantage to not always be 'common practice' and to be used inconsistently (some users use them, others do not), thus potentially creating a bias. Nevertheless, the research successfully used them to estimate the accuracy of different settings and eventually choose the most performing one. The highest correlation achieved is 0.71 and shows the challenge to transpose manual approaches (as performed by Hara et al. (2000)) in algorithmic ones.

Beyond the creation of the network itself, *Social Network Analysis* also confronts itself to the evaluation of this one. The first set of metrics focuses on the way users cluster by using indicators of the following nature: clustering coefficient, degree correlation/mixing coefficient, reciprocity/mutual links. A second set focuses on the community structure by using degree distribution, user scores distribution as well as agglomerative clustering (distribution of the strengths of the edges between participants). Finally, the third set analyses the discussion structures. This is done by observing evolution over time, branching numbers and comments per nesting level. A key metric introduced is the h-index which is defined as *given a radial tree corresponding to a discussion thread and its comments organized in nesting levels, the h-index h of a post is then the maximum nesting level i which has at least h > i comments* (Gómez et al., 2008). This indicator does thus capture the distribution of the comments, rather than just its number. In order to identify controversial posts, the following formula was then used:

ranking score = h-index + 1/num comments

An overview of additional metrics (e.g. quotation based ones) which might be used to characterise messages can be found in a paper by Gómez et al. (2008) which focuses on automatic scoring of posts. This research was applied to the Slashdot forum too.

2.3. Types of Knowledge: Tacit vs. Explicit

2.3.1. Knowledge typologies

There is a substantial literature corpus with regard to knowledge typologies. Courtney (2001) provides an overview, according to which most authors do the underlying distinction between data (raw facts), information (data in a given context, with some interpretation) and knowledge (which facilitates action). Within knowledge itself, several typologies such as the already introduced (see part 1.2) distinction between declarative and procedural knowledge (Ummelen, 1997) exist. Another major approach is the one of tacit vs. explicit knowledge developed by Nonaka (1991). Tacit knowledge is more subtle, harder to express. In the original publication (Nonaka, 1991), the skills a baker learns from a master when making bread are taken as an example. It stands in contrast to explicit knowledge which can readily be articulated and shared.

2.3.2. Challenges in operationalization

While the concepts on knowledge typologies introduced here are rather popular within the social sciences (at least, according to the citation scores), there are also shortcomings. The concepts appear to be difficult to operationalize, especially when applying them to knowledge directly. During the literature search, the only empiric set-ups which were encountered measured the knowledge indirectly (i.e. the usage of knowledge by humans). This observation is further confirmed by a taxonomy which Mitchell and Boyle (2010) established upon literature review and does not include examples of direct measurement methods either.

This is in line with criticism expressed by Gourlay (2006a) on Nonaka (1991)'s approach. Closely analysing the original publication, lack of substantial evidence, such as the transfer of tacit knowledge by socialization, is demonstrated (cognitive mapping is suggested as a method). Similarly, the examples used (e.g. prototyping of a bread baking machine) are criticized for lack of detail and the success of the interpretation is attributed to its power concerning organisational priorities, rather than to its actual validity. In a later publication, Gourlay (2006b) proposes to limit tacit knowledge to inarticulable forms. It then also makes sense that previous research used to develop this approach focuses on manifested behaviour and usage, rather than on the (inarticulable) knowledge itself.

As the scope of this study is to perform an analysis on the website contents that users can access (see part 1.3), knowledge typologies such as the distinction between tacit and explicit knowledge are considered out of the scope. In fact, substantial research has been conducted in this field but all cases encountered required interaction with the user. The operationalization always takes into account the pair of knowledge and human behaviour (thus analysing more than the output produced). Within the scope of this study, human behaviour can only be measured indirectly (using the output) and thus to a limited extent. The methods related to this theory have thus been dismissed.

2.4. Summary of the theoretical framework

This chapter further specified and studied a number of theories related to interaction in which 'mindset(s) necessary to solve a question' are shared (see parts 1.2). The first topic covered is the one of social learning, which states that learning best takes place in a social environment. Litterature shows that shared understanding is key to achieve interaction and co-construtions. Shared understanding furthermore relates to *exploratory talk*, in contrast to *cumulative* and *disputational talk* as described by Mercer (2007). This framework has also been used by previous research in a method named *Social Learning Analytics* of which both manual and automated coding approaches have been developed.

Despite its high relevance, this first part lacks details on the impact of online environments, of CMC which were thus addressed in a second part. On the one hand, it was shown that social interaction is a major challenge which must be addressed when designing online knowledge sharing. This can for instance be done by proceeding in a learner-centered way. A concrete example to do this is to stimulate the perception of each other's presence among users and to monitor the outcome. On the other hand, CMC does also impact the research methods involved in this analysis and monitoring. In the case of unthreaded or partially threaded data such as found in gis.SE, lists of messages must be split into interactions chains. Previous research confronted to this matter relate this to the term *Social Network Analysis*.

Furthermore, theories relating to the usage of CMC allow us to position this study in the context of online learning. As elaborated in part 4.1, the gis.SE forum is a rather technical one focusing on the

creation and maintenance of a knowledge base. Within the framework provided by Kreijns et al. (2002) and shown in Figure 2.1, it can be noted that this study does only cover a part of the levels that can be involved in CMC. gis.SE tends by definition towards weak proximity between users. This is related to their global distribution and to the limited options for informal communication (the only notable ones are chatrooms, as even the meta forum is formal to some degree). Furthermore, most encounters can be expected to be formal as they rely on the thematic classification of the content by using tags (for more information see part 3.2). The type of interaction investigated by this study are therefore rather task-oriented conversations and thus belong to formal communication.

Finally, knowledge typologies of which the distinction tacit vs. explicit is a popular example have also been tackled. However, literature study has not found previous research conducted without direct interaction with the knowledge users. This is in line with conceptual criticism expressed by some authors about this theory, which also explains difficulties in operationalizing it for analysis of the outputs of interaction only. Therefore, the methods related to this topic are deemed incompatible with this study.

3
Methodology

3.1. Approach chosen by the study
3.1.1. General Paradigm

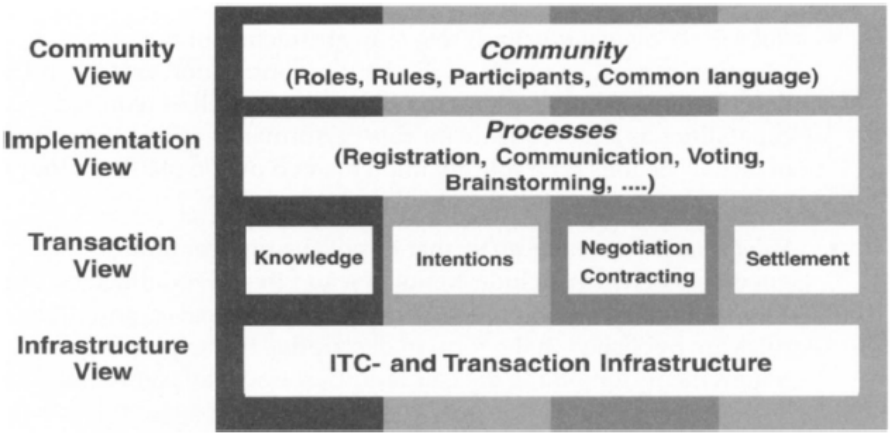


Figure 3.1: Within the media reference model introduced by Schmid (1999), this research takes an implementation view by analysing the interaction that arises from the intertwined system of users and technology.Image by (Stanoevska-Slabeva, 2002).

Within the media reference model introduced by Schmid (1999) (see Figure 3.1), this research positions itself in-between social and technical aspects. The phenomena which are studied are rather social (interaction being a social act) but are intrinsically linked to the technology, to the platform that facilitates them. Therefore, the phenomena studied can not be dissociated from the 'system' in which it takes places (the SE fora - e.g. types of posts, interface, gamification system). The same goes for the community aspects such as the moderation policy which undoubtedly influence the content and thus the interactions too.

3.2. Data terminology
For the sake of clarity, this section defines the data terminology used in this report. An illustration is also provided in Figure 3.2.

The forum gis.SE consists of **threads** which contain at least a **question**. The question and thus the thread are associated to a number of **tags** which can be found below the **body** (or text) of the question. Any **question** does also have an **author** and might have been edited, in which case the **last editor** is displayed. Also, the **question vote score** is shown on the left. Additionally, the comment might also have a **comment list related to the question**.

Next to the question, the thread might also contain one or several **answers**. Each answer consists of a **body** (or text) which was written by the **answer author**. On the left again, the **answer vote score** is displayed. Just as questions, answers might have a **comment list** related to them. Furthermore, this comment list might contain one or several **comment chains** which are composed of comments related to the same discussion.

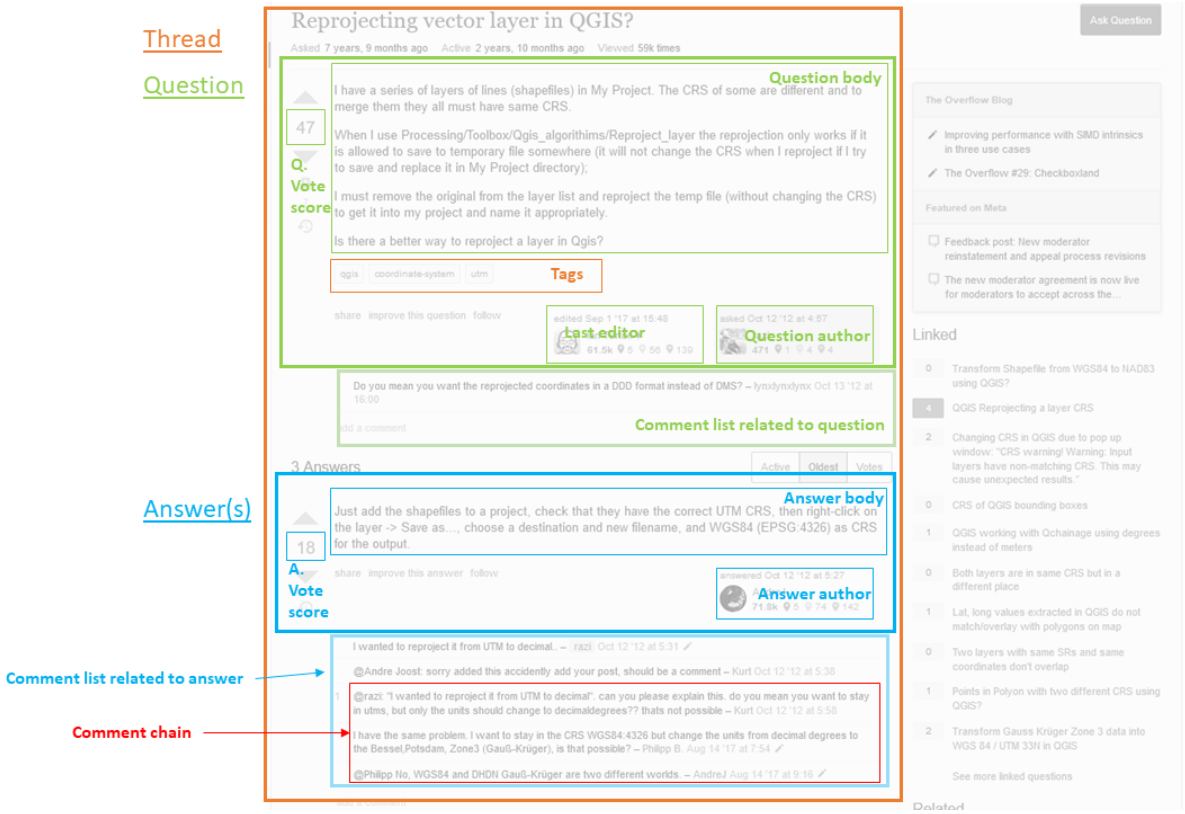


Figure 3.2: Data terminology used in this report.

3.2.1. Steps of the research

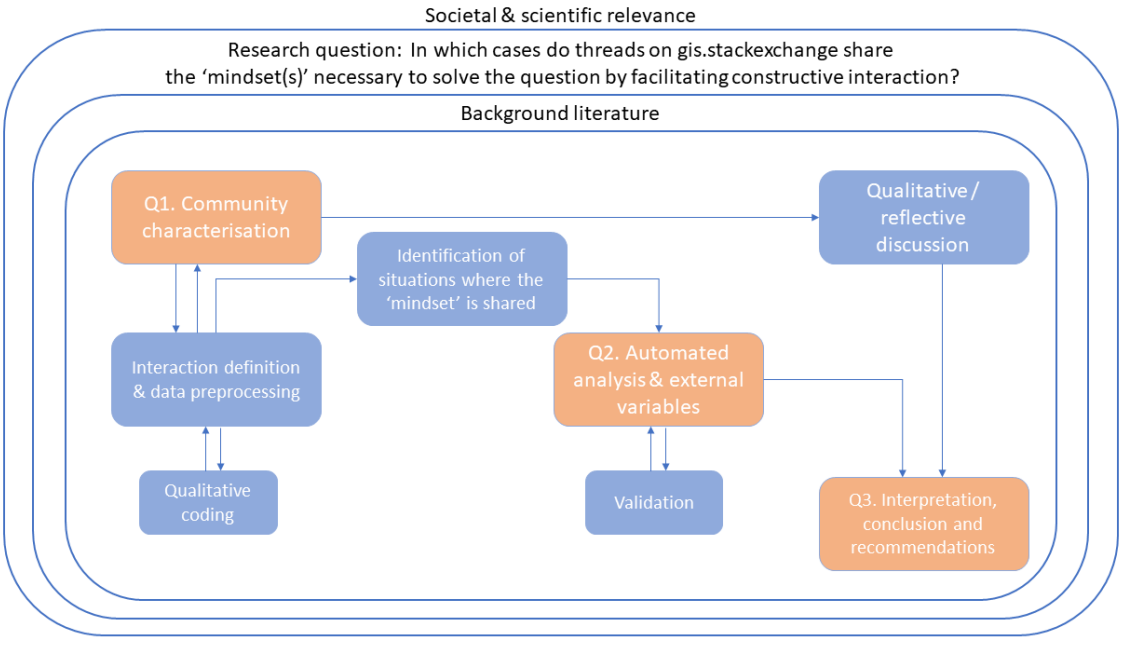


Figure 3.3: Methodological steps of this research. The steps addressing specific questions are shown in yellow. Substantial steps to be performed in between are shown in blue.

The research presented in this report follows the methodology which can be found in Figure 3.3. In a first step which is covered in chapter 4, the community is characterised. This includes data pre-processing and the definition of interaction. Here, a first loop is performed by a qualitative analysis (coding) to characterize the nature of forum interactions. The outcomes of this first part then serve as a basis for the development of an automated method to identify situations in which the 'mindset' to solve a question is shared. In chapter 5, the performance of the automated method is validated. Subsequently, data analysis concerning the cases identified by the method is performed. In chapter 6, the qualitative reflection expressed by forum users themselves are analysed and taken into account during the interpretation of the outcomes. A more detailed description of the methods involved in these steps can be found below.

3.2.2. Research methods

Forum characterisation

Within chapter 4, a qualitative characterisation of the gis.SE forum is performed. On the one hand, this characterisation is performed by positioning the forum as an online community, by using typologies from literature. On the other hand, a quantitative characterisation is performed by analysing the evolution of user and moderator activity in the relevant period (2017-2019, which is related to moderator elections in 2016 - see the end of this part). Beyond the monthly evolution, indicators of weekday activity are also calculated and analysed at a yearly level.

A final part of the characterisation is a qualitative analysis of the content of the forum. For this one, a representative sample of 50 comment lists related to questions and 50 comment lists related to answers was taken (for more details on the sampling see part 3.2.2 below. This one was performed by manual coding based on previous research in which a similar question was addressed: for the development and validation of an automated way to classify online conference logs, Ferguson et al. (2013) developed a coding scheme consisting of four characteristics which allow the identification of *exploratory talk* which can be found in Figure 3.4.

1.1	Category	1.2	Description	1.3	Examples
1.4	**Challenge**	1.5	A challenge identifies that something may be wrong and in need of correction	1.6 1.7 1.8 1.9 1.10 1.11 1.12 1.13	calling into question calling to account contradicting disputing finding fault with proposing revision putting forward an opposing view raising an objection
1.14	**Evaluation**	1.15	An evaluation has a descriptive quality	1.16 1.17 1.18 1.19	appraising assessing expressing in terms of something already known judging
1.20	**Extension**	1.21	An extension builds on, or provides resources that support, discussion	1.22 1.23 1.24 1.25 1.26	applying idea to a new area increasing the range of an idea linking to, developing or providing related resources requesting additional resources to support understanding taking the same line of argument further
1.27	**Reasoning**	1.28	Reasoning is the process of thinking an idea through.	1.29 1.30 1.31 *1.32* 1.33 1.34 1.35	asking questions *about content* changing position in the light of arguments presented explaining inferring justifying your position reaching a conclusion working ideas out in a logical manner

Table 1: Coding scheme for sub-categories of exploratory dialogue. Dialogue turns coded in any of these categories were also coded as exploratory. All other turns were coded non-exploratory

Figure 3.4: Coding scheme developed by Ferguson et al. (2013) to identify *Exploratory Talk* in conversations of online conferences.

As the coding scheme of Ferguson et al. (2013) is rather detailed, a reduced and more practicable

and time-efficient version was made. This allows dismissing nuances such as, for instance, the difference between calling into question, putting forward an opposing view and contradicting. Each of the 100 comment lists was read and checked for the following traits, the results were stored in an excel table (see appendix B):

- **Challenge:** the traits which were considered here are expressions of criticism, objection and concern.
- **Evaluation:** approval/disapproval, confirmations and judging were considered here.
- **Extension:** here, the extension of an existing idea beyond its original scope was considered. This might include cases that were a higher degree of insight/detail was chosen to discuss the matter.
- **Reasoning:** this includes all cases in which more precise information was asked and/or given. Furthermore, messages which explain the matter, justify positions and conclude discussions are also taken into account.

Quantitative automated identification of chains

In chapter 5, an automated approach to split unstructured lists of comments into meaningful chains is introduced. This one bases on the research of Petrovčič et al. (2012) introduced in part 2.2.3. Among the possible methods to quantify the 'distance' between messages, the time elapsed has been selected (another possibility would have been the similarity of messages). Therefore, the main rule which the automated approach follows is:

- Any comment separated by more than [...] hours from the previous comment does not belong to the same interaction 'chain'.

Moreover, several exceptions to this rule are considered and can be summarised as follows:

- If a comment has sufficiently explicit references to a previous message, this comment can not be the start of a new chain.

Similarly to the chapter 4, a validation on a sample of comment lists referring to 50 questions and 50 answers is conducted in this part too. For each of these lists, the differences (number of identified chains) between the automated approach and human interpretation are identified. Beyond discrepancies in the number of chains, the number of comments which are attributed to a wrong chain (wrongly added or removed) is quantified too.

In the next step, these results provide an estimation of the accuracy of the automated approach. Using this one, a practicable definition to discriminate between regular and 'high-interaction' comment chains is established. This is done in a way to minimize the impact of the automated method's weaknesses.

Eventually, these steps facilitate the most conclusive part of the chapter, namely a data analysis. This one is performed on all content of the period 2017-2019 and thus works at a scale which can only be handled by an automated analysis. The analysis is performed at three levels (see part 3.2 for the terminology): comment lists, question/answer as well as thread. Thematically, the analysis focuses on three axes: the popularity metrics, question typology and content as well as the users and their interactions.

Validations and representative sampling

An important aspect for the validation performed in both chapters 4 and 5 is the sampling method. To create a representative sample, so-called *tags* are used. Introduced in part 3.2, *tags* are an existing thematic classification tool which is used by the community to cluster users around shared interests. Each thread has typically one or multiple *Tags*. For the representative sampling, a list of *tags* (after potential edits) is obtained for all questions/answers which had at least one comment in the period 2017-2019 or were created before 2020. The 50 most frequent *tags* (see appendix A) are then selected and used for sampling. Out of a total of 77662 questions/answers, this approach allows covering 87.6% of the cases.

It should be noted that while doing the sampling and the selection, combinations of *tags* are not considered (the tags other than the one for which a thread was selected were deleted). This process is done once for comment lists associated with questions and once for comment lists associated with answers (one answer among the thread was chosen). While doing this, the distribution of the number of comments is calculated and applied to the sampling as well(see Figures 3.5). Cases with less than two comments are not taken into account, and cases with exactly two comments are only taken into account if the second comment was posted by the author of the parent answer/question.

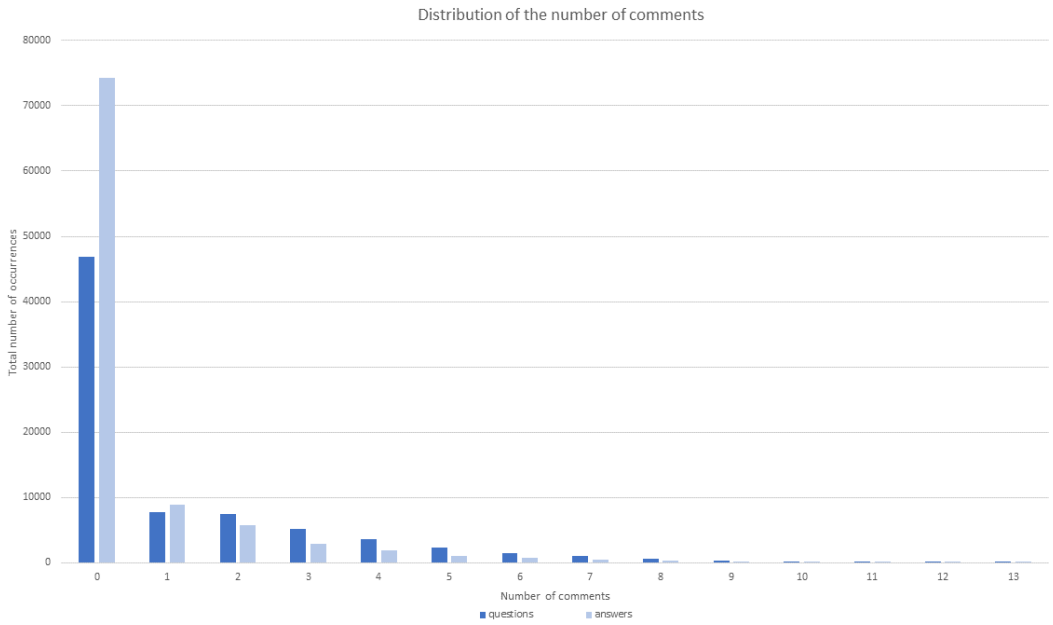


Figure 3.5: Distribution of the number of comments per question/answer.

3.2.3. Qualitative analysis of discussions in the meta forum

The last chapter uses a qualitative method, namely a content analysis to identify important topics in the forum community discussions. This allows linking the dynamics identified in the data analysis to the perceptions of the users involved in these very same dynamics.

To select the topics, the ten most voted posts (with the highest vote activity) on the gis.SE meta forum were consulted. Out of these, three topics were selected (moderation policy/forum rules, guidelines provided to users, feedback style, limited opinion plurality and unanswered questions) as chapter axes. For each of these, relevant user contributions are grouped, quoted and discussed. Additionally, a topic which does not appear in the top ten but is highly relevant for this research and the SE system (the status of comments) was covered similarly.

Sources outside of gis.SE's meta forum were considered as well. However, the usage of two search engines (*google.com*, *duckduckgo.com*) did not deliver meaningful results with four keyword combinations ('gis.stackexchange', 'moderation', 'criticism', 'overmoderation'). The only potentially interesting occurrence identified is a tweet which was not followed by insightful discussions[1].

Overview of methods

An overview of the methods employed in this research is shown in Table 3.1.

[1] https://twitter.com/scw/status/679001514690613248

chapter	method used	related concept(s)
4	positioning with regard to typologies in literature	Online communities (Stanoevska-Slabeva, 2002), Academic research collaboratories (Bos et al., 2007) see part
4	manual analysis of comment list contents	Exploratory talk (Ferguson et al., 2013)
4 & 5	Sampling using list lengths and tags	Stratified and probability proportionate to size sampling (Babbie, 2015)
5	Automated splitting of comment lists into chains	Social network interactivity analysis (Petrovčič et al., 2012)
5	Definition of 'high-interaction chains' and data analysis along different axes	Automated scoring of posts (Gómez et al., 2008)
5	Qualitative analysis of discussions in the meta forum	Reflective practitioner, see part 2.2.3

Table 3.1: Table showing the methods used in this research and their related concepts in literature.

3.3. Tools and code availability

The source code used for the data analysis performed within this project is available at https://github.com/Flyalbatros/gis.SE.thesis. The main programming languages used were *Python* and *PostGreSQL*.

3.4. Data availability, cleaning and gaps

The data of the gis.SE forum was downloaded from the internet archive where SE regularly releases data dumps[2]. Additionally, the full data schema is also available[3]. The obtained dataset was downloaded in July 2020 and covers the period from 22 July 2010 until 01 July 2020. To perform the data analysis, it was further decided to choose a period of relative stability community-wise. Therefore, the dataset was reduced to the period between 01 January 2017 and 31 December 2020, which corresponds to the period between the most recent moderator changes (the current ones were elected in September 2016) and the Covid-19 pandemic. By doing so, the number of questions was reduced from 125 167 to 61 638.

It should be noted that these data dumps do not provide complete information from a technical point of view. This is due to some limitations concerning the availability of deleted data. On the one hand, automatics deletions are performed by all SE sites following clear rules[4]. As this process mainly applies to questions with low activity (e.g. no answers within 30 days and score lower than -2; less than 2 comments within 365 days), its impact is not expected to be major for this study. An ill-defined problem (which is addressed in chapter 6) is, however, that several such questions might have no activity due to deliberate closures by moderators.

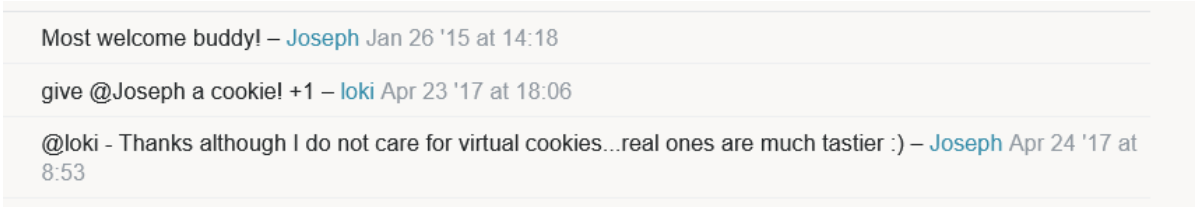


Figure 3.6: Example of evidence showing that a comment was deleted[5]. NB: Joseph is the author of the answer this comment lists refers too - the list shows all comments visible on 23/08/20.

[2]https://archive.org/download/stackexchange
[3]https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede
[4]https://meta.stackexchange.com/questions/78048/enable-automatic-deletion-of-old-unanswered-zero-score-questions-after-a-year
[5]https://gis.stackexchange.com/questions/131788/adjust-fill-style-behavior-qgis/131789

A more critical data gap results from the deletion of comments by the users themselves. Contrasting with the rules for questions, there are no constraints for deleting one's own comments. These deletions are not included in the data dumps - the original comments are thus missing. The comments deleted by the community or moderators are acceptable as they are, in fact, a policy filtering content which should not be shared. The comments deleted by the users themselves (which cannot be stopped by others) are, however, more problematic. According to information provided by SE upon request, these ones represent 10.05% of all comments posted (incl. the deleted ones) in 2019 and 9.48% of all comments posted in 2018. In some situations, gaps in interactions patterns due to subsequent messages not being deleted can be found, as shown in the example in Figure 3.6.

A final consideration during the data cleaning is the high degree of moderation on the gis.SE forum. As mentioned in the introduction, gis.SE is a forum which is very content-oriented and aims at building a knowledge database. Therefore, discussions with other aims often referred to as 'chit-chat', are explicitly not welcome. As a consequence, it is rather rare to find comments which do not focus on the content of the thread (this observation is confirmed by a study on the SO forum Sengupta and Haythornthwaite (2020) and by the sample validations discussed in part 4.3.1). Content filtering was therefore not considered.

3.4.1. Case of edits

In a study by Li et al. (2015) edits of the SE websites were studied in depth. In line with Stack Overflow's guidelines[6], edits were divided in five types:

1. Fixing grammar and spelling mistakes
2. Clarifying the meaning of the post
3. Including information found only in comments
4. Correcting mistakes or adding updates
5. Adding related resources

To filter out comments which are irrelevant for this study, their impact on the question/answer's length was considered. Comments not altering the length of a post can not belong the category 5. Category 1 is typical for edits not altering the length as it represents the very small scale corrections. For edits belonging to category 2, 3 or 4, and unchanged length usually indicates small scale modifications too. There might also be cases where a major addition and deletion with similar lengths occurred in the same edit, but such coincidences are left out of scope. Category 5 is the only category in which an edit not altering the length is impossible. There might however be edits which increase the length only minimally.

Based on the reasoning that length alteration is representative for the degree of editing (and leaving cases with considerable deletions and additions of similar length out of scope), it was decided to filter the edits by the degree to which they alter the length. Edits adding or reducing the length by less than 10% were thus filtered out and not considered for the research. Practical examples can be found in Figures 3.7 and 3.8.

Beyond this length filter, it should be noted that edits can be rolled back by post owners. As the downloaded data is raw, thus containing both refused edits and rollback logs, it is worth noting here that edits subject to such rollbacks were still taken into account as they show user interaction. Analysis of rollback was considered but dismissed due to their low quantity within the samples used in this research.

[6] https://stackoverflow.com/help/editing

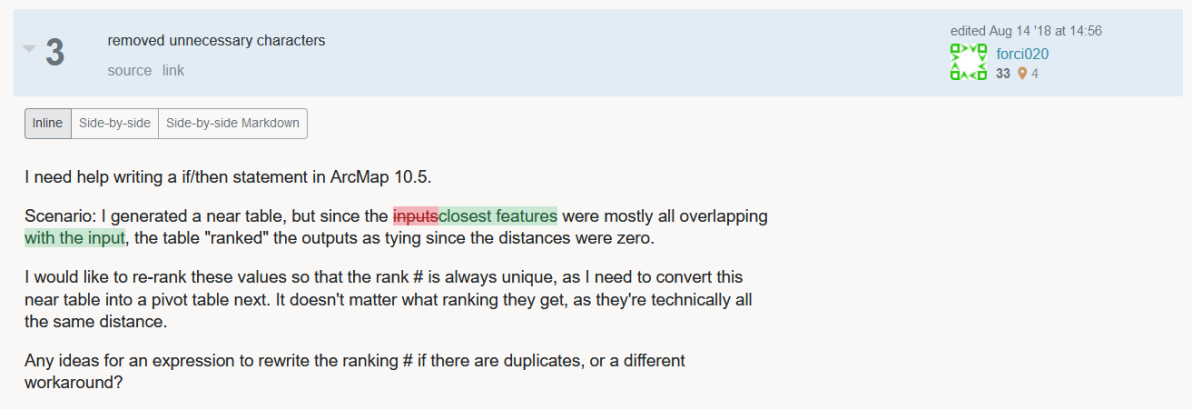


Figure 3.7: Example of an edit altering the question length by less than 10% [7]. Here, only a few words are modified and added. Such edits are not considered in this research.

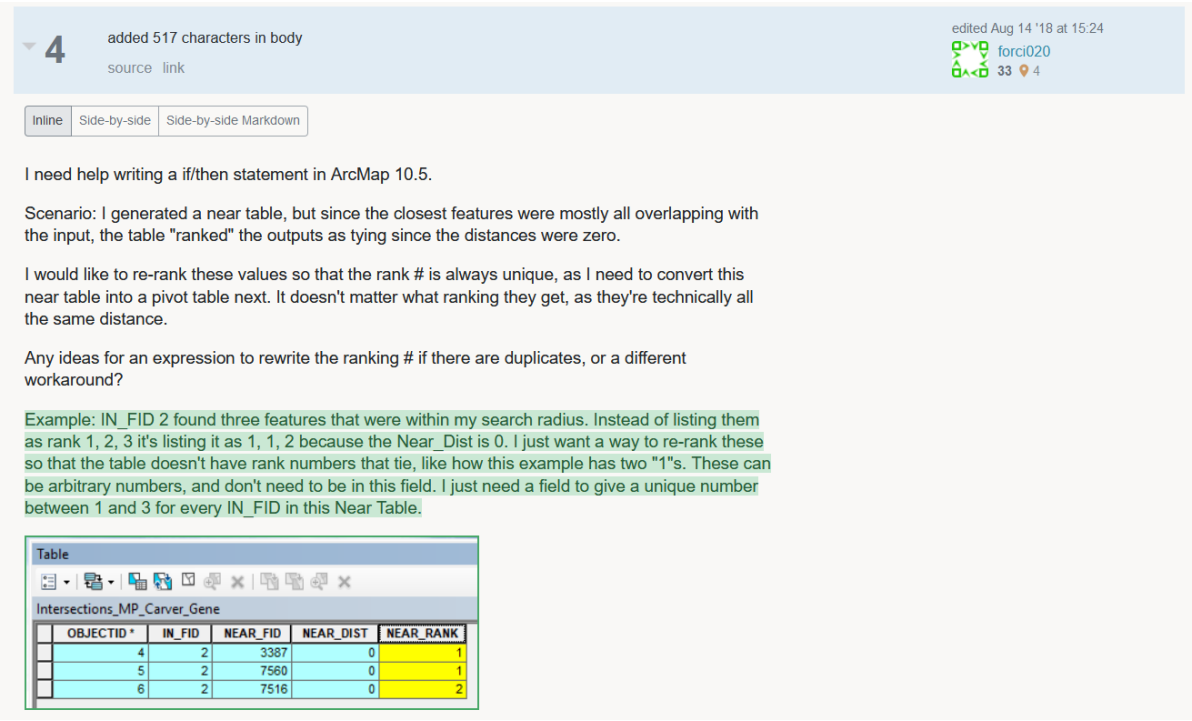


Figure 3.8: Example of an edit altering the question length by more than 10% [8]. Here, a substantial addition takes place adding an example illustrating the question. Such edits are considered in this research.

3.5. Ethical considerations

Ethical considerations for this study are rather limited as content posted on the SE fora is freely accessible by anyone and users are very well aware of this when they discover the system (as you can use SE passively without ever registering).

From a copyright point of view, SE publishes all data under the CC-BY-SA-3.0/4.0 license[9]. Therefore, the strictest license among 3.0 and 4.0 should apply. The data can thus be freely reused, as is

[7] https://gis.stackexchange.com/posts/292730/revisions
[8] https://gis.stackexchange.com/posts/292730/revisions
[9] There have been some issues as SE decided to move to 4.0 without permission: https://meta.stackexchange.com/questions/333089/stack-exchange-and-stack-overflow-have-moved-to-cc-by-sa-4-0

regularly done for academic research.

Although there is no formal requirement to do so, special care has been taken for the case of users that deleted their account. These users have made the explicit choice to remove themselves from the SE system as accounts can only be deleted upon request (no auto-deletion). In practice, this means that instead of being linked to a user id, all posts/activities of the deleted user obtain the attribute *UsedDisplayName*. In theory, it would thus be possible to link the activities of these users, reducing the limitations to the absence of metrics such as the registration date, the reputation score etc. However, this was not done to respect the choice of (former) users who deleted their account.

4

Characterisation of the gis.SE community

This chapter aims to specifically characterise the gis.SE community. While it is part of a bigger group, the SE fora, it has its specificity which impact the communication(s) taking place in it. To identify the specificity, three approaches were selected.

First, a social perspective is elaborated. Within this one, the forum is linked to existing theories and taxonomies on (online) communities. Second, a data science perspective has been chosen. A general data analysis is performed to identify key indicators and potential time variations. Finally, a qualitative analysis is performed as well to identify the types of interactions present among the ones defined by Ferguson (2009) (see section 2.1). The chapter is closed by an overview of the main findings and a discussion on how they relate to the assumptions made.

4.1. Social network perspective: gis.SE as a community of practice

The name of the gis.SE forum explicitly refers to *Geographic Information Systems* (GIS), a term which was first used in 1968[1]. The acronym itself is a well-established one in the field of geography, especially as it is often used for another term, namely *Geographic Information Science*. The latter is the related academic and professional field which was first defined around 1990 by Goodchild (2010). The professional orientation of the forum is thus already present in the name.

From a social point of view, the community formed by the gis.SE forum is a very open one. Beyond a mere internet connection, there are absolutely no prerequisites for registering an account and participating. Even stronger, the content of the forum can be consulted by any internet user. gis.SE applies the principle of Open Data to the knowledge repository of GIS it represents. This is also coherent with the official support role[2] the forum plays concerning the widespread software named *Quantum Geographic Information System (QGIS)*[3]. This is one of the two major tools present in the field of GIS, the other one being Environmental Systems Research Institute (ESRI)'s *ArcGIS* software. In opposition to the first one, the latter is a proprietary software and does thus not share the principles of Free and Open Source Software (FOSS). Nevertheless, both software are widely present as the tags among the forum's top 50 show (see appendix D).

In this sense, the gis.SE community connects people sharing a common field of expertise, in this case built around a set of software technologies. By being exclusively online, the community aims at being as open as possible and thus connects people working in many different contexts (e.g. countries, organisations, etc.) but sharing a common expertise. This can be linked to the vision of Barley (1996) who studied such professions and their position within organisations. As phrased by Brown and Duguid (2001), technicians tend to more easily build communities with people of the same expertise in other organisations than with other professions inside the same organisation. In their work, the authors see this phenomena as the key to the term *community of practice*. They see this one as a central unit of analysis to understand knowledge in companies, putting particular emphasis on the implications of practice. Furthermore, the authors develop on the presence of a shared know *how* which is required to effectively exchange know *that*, more explicit knowledge. In the specific situation here, this is associated to a foundation concerning using computers generally (e.g. programming languages) and the

[1] https://www.esri.com/news/arcnews/fall12articles/the-fiftieth-anniversary-of-gis.html

[2] https://qgis.org/en/site/forusers/support.html

[3] Additionally, one of the eight moderators (https://gis.stackexchange.com/users/187/underdark) is a member of the OSGeo charter, the organisation managing, among others, QGIS

community specific software. The concept of *community of practice* is thus of relevance for the study of online communities such as gis.SE.

To position the gis.SE forum within its context, the framework provided by Stanoevska-Slabeva (2002) is useful. In it, online communities are defined "through their features as associations of participants who share a common language, world, values, and interests, obey a commonly defined organizational structure, and communicate and cooperate ubiquitously connected by electronic media". Based on existing typologies, Stanoevska-Slabeva (2002) then establishes four main types of online communities, with each several subtypes:

- Discussion or conversation communities dedicated to the exchange of information about a defined topic. The following subtypes exist:
    1. *Relationship communities* targeted at the creation of social bonds between members sharing a situation or passion.
    2. *Interest communities* emerging around a defined topic and attracting interested participants.
    3. *Communities of practice* which are discussion communities around a domain of knowledge. This category builds upon a concept established by Wenger (2010). According to this one, *communities of practice* are a special form of groups which build a common stock of knowledge, accumulate expertise and develop shared practice.
    4. *Implicit discussion communities* that are also known as recommendation and reputation communities. These ones often arise around e-commerce activity.
- Task and Goal-oriented communities which have the aim to achieve a common goal by cooperation efforts.
    1. *Transaction communities* which focus on the exchange of economic products or services.
    2. *Design communities* where the cooperative design of products and goods occurs, this includes less material products such as software.
    3. *On-line learning communities* which emerge in online education settings.
- Virtual worlds which create parallel 'worlds' to facilitate fantasy and gamification.
- Hybrid communities which are combinations of any of the previous ones.

Within this first taxonomy, gis.SE can be seen as a hybrid of two forms of discussion or conversation communities. On the one hand, it can be seen as an *interest community* as the forum is built around a common interest, around expertise in a specific technology. On the other hand, the forum is also built around a professional community, which makes them qualify as *community of practice*. This is especially true as the main aim is to build a knowledge base of answered Q&A threads, thus to share common expertise.

Beyond the categories by Stanoevska-Slabeva (2002) (introduced in part 1.1.1), one can use the approach provided by Bos et al. (2007). This one takes the angle of academic research 'collaboratories'. By performing a literature review, the author identified seven types of these:

- *Shared instruments*: initiatives aiming at increasing the usage of scientific instruments (e.g. telescopes) by giving easier (remote) access.
- *Community data system*: sharing information resources, for instance by easy online access to existing datasets
- *Open community contribution system*: remote cooperation tools allowing to aggregate research efforts towards a problem
- *Virtual community of practice*: tools allowing online exchange about a shared research area
- *Virtual learning communities*: mainly occur in on-lin education settings and facilitate remote courses
- *Distributed research centre*: by analogy an 'on-line University', allowing to coordinate activities similarly but on-line

- *Community infrastructure project*: cooperation seeking to develop of infrastructure (tools facilitating further work) in a particular domain.

In this second approach, SE fits mainly two categories: first, the *Virtual communities of practice* as there is a shared expertise. Although not strictly academic, one might note that gis.SE and more generally SE fora are rather goal driven communities structured around one field of practice. Second, the category of *Open community contribution systems* also applies as the fora have each an own but common goal, namely the creation of an open and shared (as mentioned earlier), knowledge base. There is no coordination of 'research' efforts in a classical way - but rather an aggregation of existing initiative/projects and the issues to which users are confronted. To achieve this, cooperation such as the one with QGIS mentioned earlier are key as a knowledge base should preferably be centralised at one location to avoid duplicate efforts (i.e. avoid several knowledge bases with the same content).

4.2. Data science perspective: gis.SE in numbers

This second part takes a data analysis perspective and provides insights into the relative quantities of the activities taking place within the gis.SE forum. The aim is to give an overview of how the general SE system applies to the specific forum (as introduced in part 1.1.1), including the degree of moderation. A first part focuses on raw data and ratios thereof (parts 4.2.1) while a second one uses first computations to calculate reaction times (part 4.2.2).

4.2.1. Seasonal evolution, data by month or week

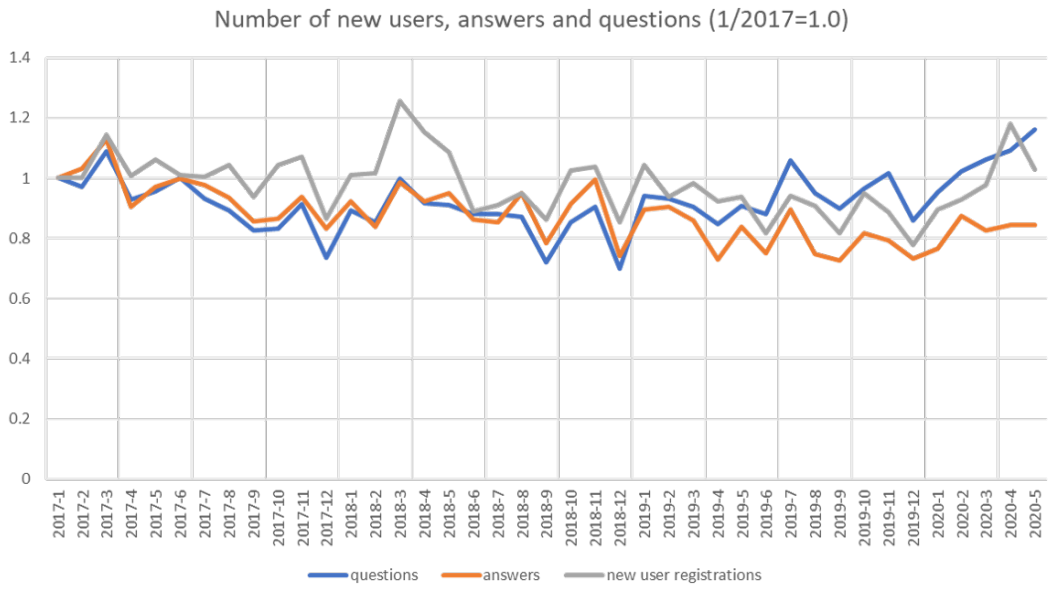


Figure 4.1: Monthly number of new users (1.0=1768), answers (1.0=1447) and questions (1.0=1398)

Questions vs. Answers Over the period 2017-2019 (see Figure 4.1), one can see that the number of answers is globally stable, fluctuating between 1.0 and 0.7. Among the variations, no particular seasonality is to be noted. The curve of the number of answers follows the one of the number of questions until January 2019 when a gap arises and gradually increases to about 0.2. From 2020 on, the gap becomes even bigger. The number of new user registrations is globally stable, fluctuating between 0.8 and 1.1. It mostly follows the number of answers - except in March 2018 and in April 2020 when numbers outside of this range can be observed.

Questions without activity As a question can have more than one answer, it is necessary to have a closer look at the number of unanswered questions to understand the increasing gap (see Figure

D.1 in annex D). This observation confirms that it is not the number of answers per questions but the monthly number of threads/questions which never received an answer (orange line) which increases from January 2019 on. While the share of questions having neither an answer nor a comment (grey line) is lower, a similar increase can be observed there. This increase might be linked to an increase in the absolute number of questions but only from January 2020 on.

Edits The fact that questions are being asked but not answered might point to a potential mismatch. One way of correcting such a mismatch is to edit the questions, which would increase the total number of edits (considering that the question answering activity does stay normal). However, this is not the case, as Figure D.2 (in annex D) shows: the number of edits is rather stable, fluctuating between 0.5 and 0.7. Also, the moderator share is stable between 0.3 and 0.5, except for August 2018 and the year 2020.

Closures Another option to deal with mismatching questions might be to close them. Closures are typically applied when a question does not fit the rules or scope of the forum. However, Figure D.3 (in appendix D) shows that the number of closures decreased until August 2018 when it nearly reached 0.2. After this, it never fully recovered as it fluctuated between 0.3 and 0.7 until it dropped again in 2020. Closures are thus no reaction to the increasing number of questions left unanswered. The moderator share was also calculated and stayed rather stable (between 0.7 and 1) until it dropped in 2020. Here, one must note that this is the share of the decisions, not the number of closure votes themselves. A question can either be closed by a moderator action (immediate effect) or by a critical number of 'regular' users who voted for a closure. Nevertheless, these votes can also incite a moderator to look at it and act. The moderator share is thus the share of closures where the *effective* decision was made by a moderator, thus not meaning that no 'regular' users were involved.

Comments per thread Focusing more on the topic which will be addressed in chapter 5, the number of comments per thread was calculated (see Figure D.4 in annex D). This one shows a slight decrease from January 2019 on, moving from fluctuations between 3 and 3.5 to fluctuations between 2.5 and 3. Here, the decrease in 2020 makes less of a kink and seems to be more in line with the previous trend. Second, the moderator share (see Figure D.5 in D) was calculated. This one is rather stable, fluctuating between 0 and 10% except for the first half of 2017 when it fluctuated between 10 and 15%.

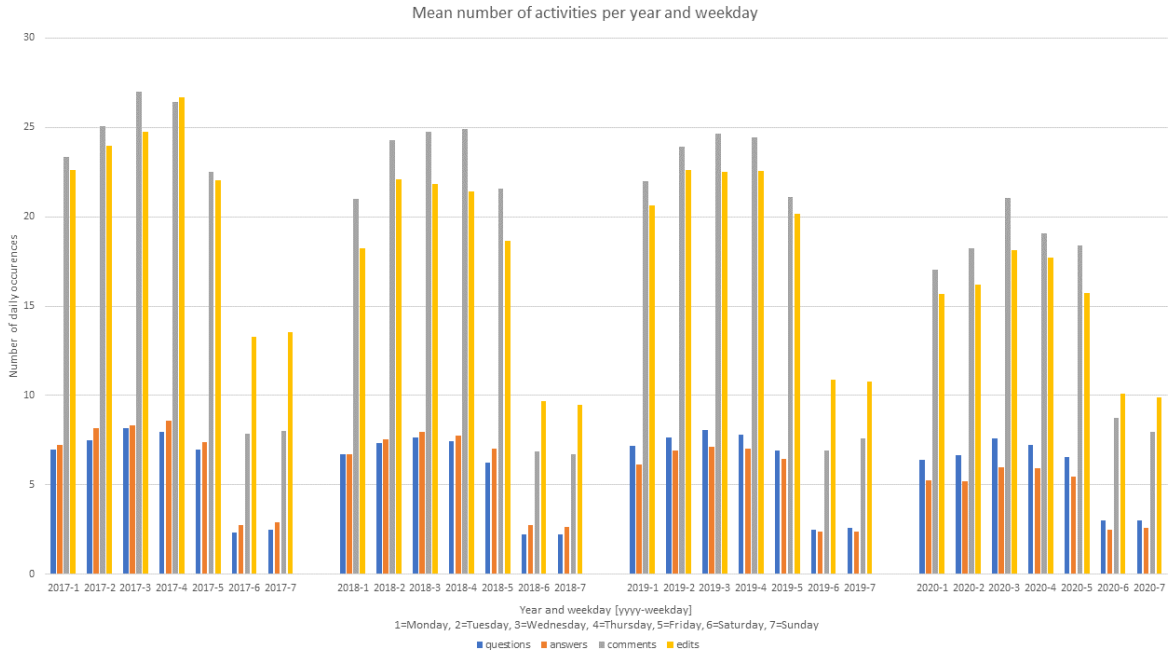


Figure 4.2: Average number of questions, answers, edits and comments, per year and weekday.

Overview of activities per weekday The daily average number for each of the activities is provided in Figure 4.2. Coherently with the previous results, one can note that on all weekdays, the question/answer activity is lower than the edit/comment activity. Also, a drop of activity on Saturdays/Sundays can be noted for all activities. For the years 2017 - 2019 and for all categories, the highest activity can be observed on the days Tuesday, Wednesday and Thursday.

4.2.2. Reactivity per month and weekday

In this second part, the reaction time between two users interacting with each other was investigated. To do so, two approaches were chosen and compared.

The first approach bases on the informal habit of some users to use handles in the comments they post. Handles are an explicit reference to another user's message consisting of a combination of the symbol '@' and the username of the other user, thus resulting in a word of the structure '@username'. By identifying the combination and handles (up to three messages before the comment containing the handle), reaction times are calculated.

A second approach assumes that interaction occurs when a comment is both preceded and followed by a comment of the same other user. This results in a so-called 'ABA' scheme were A and B are the authors of the comments. By identifying the presence of such schemes in messages, reaction times are calculated. Here, the analysis goes even one step further as the reaction times for the first part (referred to as 'AB') and for the second part (referred to as 'BA') are calculated separately to avoid duplicates in the case of interactions with more than three messages.

This analysis was performed for the entire existence of the forum and interestingly, the suitability of the methods contrasts. While the 'ABA' scheme identification was able to identify 124 863 different references, the handle approach only identified 22 941. The overlap between the two methods is limited to 1108 references (respectively 0.8 and 4.8%), which highlights that both approaches address different subsets. Additionally, the reaction time between a question and the first answer was also calculated and analysed. Due to the high spread of the results, averages turned out to be unpractical to get insights into the data. Therefore, 10th and 90th percentiles were chosen instead. Moreover, it should be noted that the date and time of the earliest message are taken as a reference for plotting the data (i.e. by weekday).

Based on usage of handles

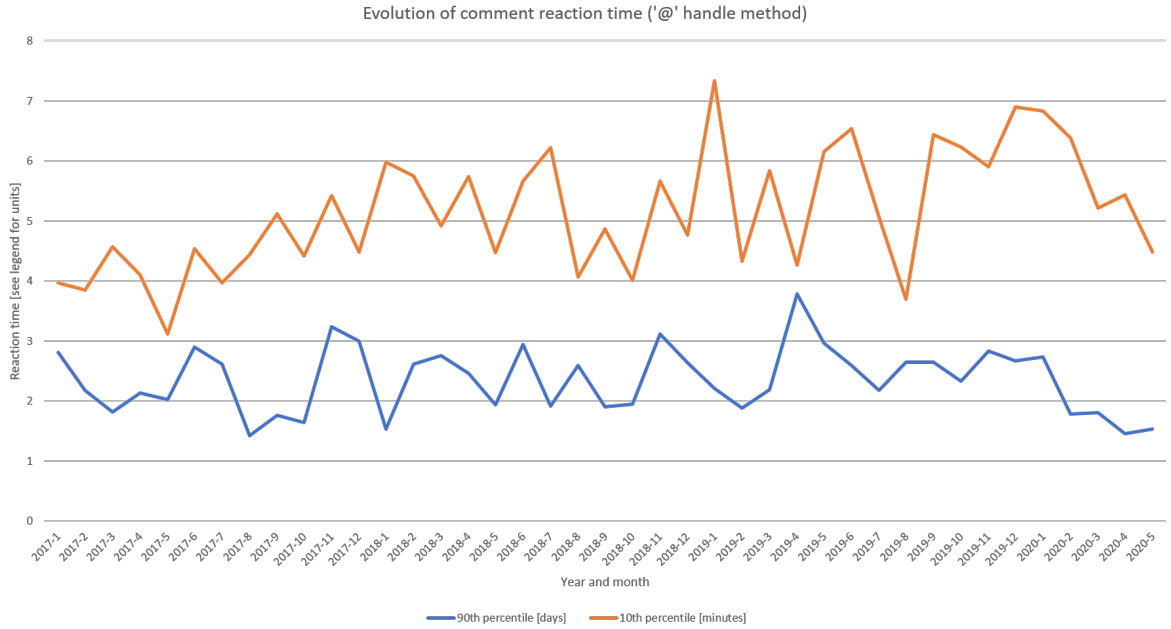


Figure 4.3: Monthly average reaction time, obtained by the handle approach.

Monthly evolution As can be seen in Figure 4.3. the monthly reaction time based on the handles is of 4-7 minutes for the 10th percentile and 1-4 days for the 90th percentile. Globally, the values are stable but a decreasing trend can be observed from January 2020 on.

Yearly weekday evolution Figure D.6 shows that for the 10th percentile (red), an increase of the reaction time can be found during weekends. This increase is less pronounced in 2018. For the 90th percentile reaction time, however, peaks reaching nearly 4 days on Friday (2017, 2018 and 2020) or Thursday (2019) are observed.

Based on identification of ABA schemes

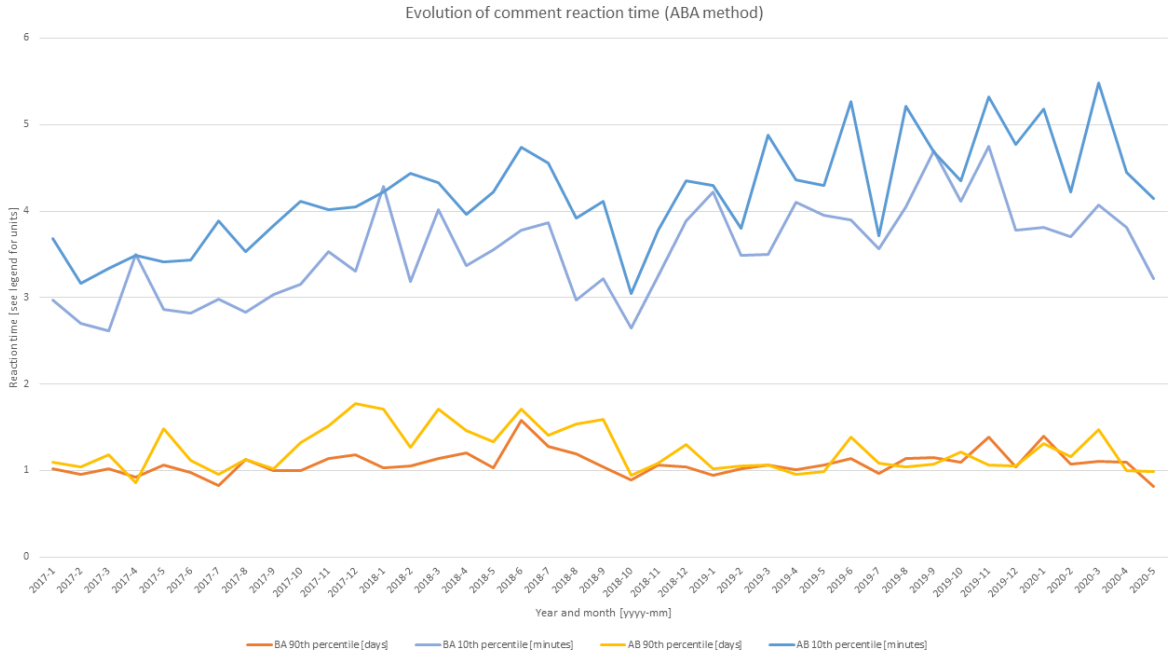


Figure 4.4: Monthly average reaction time, obtained by the ABA approach.

Monthly evolution The evolution of 90th and 10th percentile of the BA and AB reaction times that can be seen in Figure 4.4 (in annex D) follow a globally similar evolution but show differences on a small scale. Generally, the BA part (thus the second part of the reaction) has slightly lower values than the AB (thus the first) one. The 10th percentile is typically located within 3-6 minutes with an increasing trend from November 2018 on. The 90th percentile fluctuates between 1 and 2 days.

Yearly weekday evolution For the evolution of the ABA approach results per weekday (see Figure D.6, differences in the peaks of the different sets can be found too. For the 10th percentile (both AB and BA), higher reaction times are obtained on Sundays and Saturdays, reaching up to about 8 minutes in 2019, while being around 4 minutes during the week. For the 90th percentile, the BA part peaks at about 3 days in 2017-2019. For the AB part, the 90th percentile peaks at a similar value on Sunday and Monday. This suggests that for starting a discussion, long reaction times tend to be longer on Sunday and Monday and that for an already started discussion this is the case on Friday. However, this observation is different for relatively short reaction times. Also, it should be noted that, here too, the situation is considerably altered in 2020.

Based on the question- first answer pairs

Monthly evolution For the time between a question and its first answer (Figure 4.5), an interesting phenomenon can be observed in the period 2017-2019. While there is a decreasing trend for the 90th

percentile (moving from about 10 to about 5 days - except for two peaks in January and July 2017), the 10th percentile shows a slight increase since January 2019, moving from 15 towards 20 minutes.

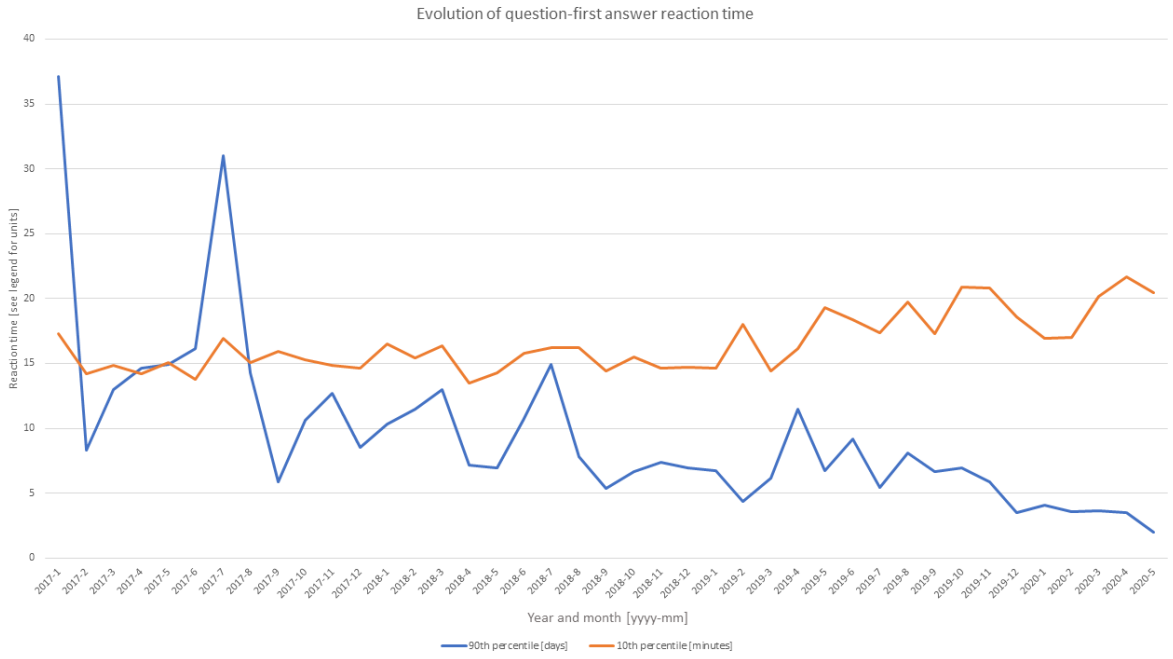


Figure 4.5: Monthly average reaction time, for question - first answer pairs.

Yearly weekday evolution For the yearly weekday evolution, similar observations as for the handle and ABA reaction times can be made (Figure D.8 in annex D). The 10th percentile peaks on Saturday/Sunday at 30 to 40 minutes. In 2018 and 2019, the 90th percentile peaks on respectively Thursday and Friday at about 10 days. For 2017, 90th percentile reaction times are higher (in coherence with Figure 4.5) and peak on Thursday and Sunday at about 18 days.

4.3. Distribution of the different types of interaction: exploratory, cumulative and disputational

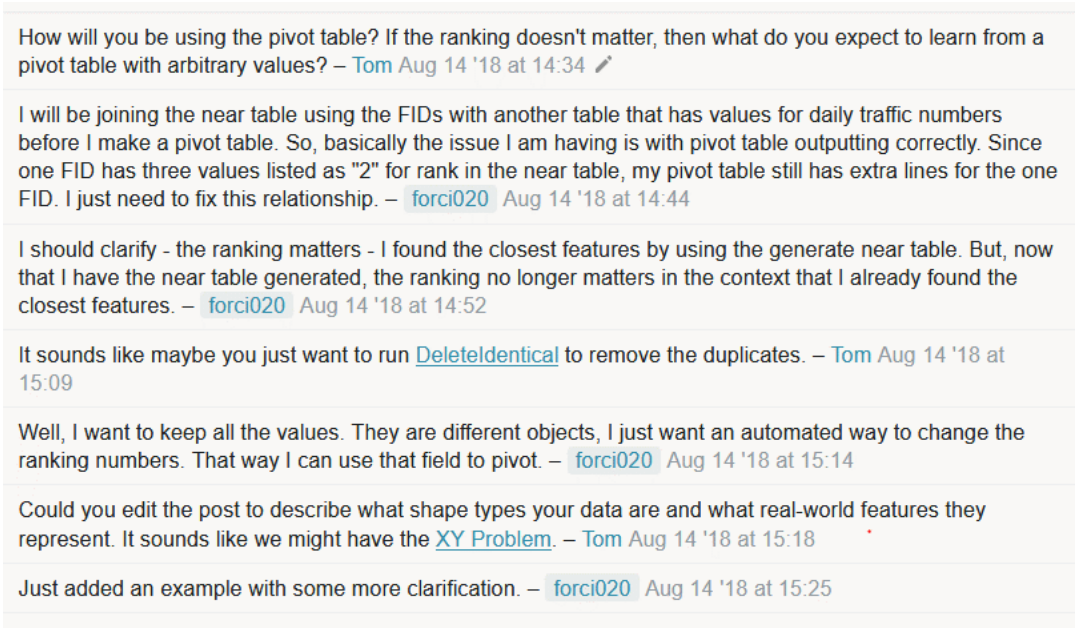
This last section of the chapter focuses on a qualitative analysis performed on a sampled number of discussion lists. The aim is to determine the nature of the interactions occurring on the forum. Details on the approach taken were already introduced and can be found in part 3.2.2.

Overall, all three types of *talk* (*exploratory, cumulative and negative*) were encountered during the validation. The cumulative one was, however, the most frequent one, followed by the *exploratory* and the *negative* one (which is practically non-existent and was only encountered in the validation discussed in part 5.1). An example of each of the three types of talk can be found in Figures 4.6, 4.7, 4.8.

> -1. You write degrees in a column that expect meters, so your point is 52 meters away from 0;0, in the ocean near the equator – JGH Nov 28 '18 at 14:04
>
> The -1 is a downvote for my ignorance?! Anyway, what makes this confusing for me is that in my mapbox-gl.js the points show exactly where they belong, so probably I accidently 'correct' things within my mapserver. – musicformellons Nov 28 '18 at 14:15

Figure 4.6: Example of *negative talk*[4].

[4]https://gis.stackexchange.com/questions/304139/what-is-the-metric-of-the-radius-postgis-query



How will you be using the pivot table? If the ranking doesn't matter, then what do you expect to learn from a pivot table with arbitrary values? – Tom Aug 14 '18 at 14:34

I will be joining the near table using the FIDs with another table that has values for daily traffic numbers before I make a pivot table. So, basically the issue I am having is with pivot table outputting correctly. Since one FID has three values listed as "2" for rank in the near table, my pivot table still has extra lines for the one FID. I just need to fix this relationship. – forci020 Aug 14 '18 at 14:44

I should clarify - the ranking matters - I found the closest features by using the generate near table. But, now that I have the near table generated, the ranking no longer matters in the context that I already found the closest features. – forci020 Aug 14 '18 at 14:52

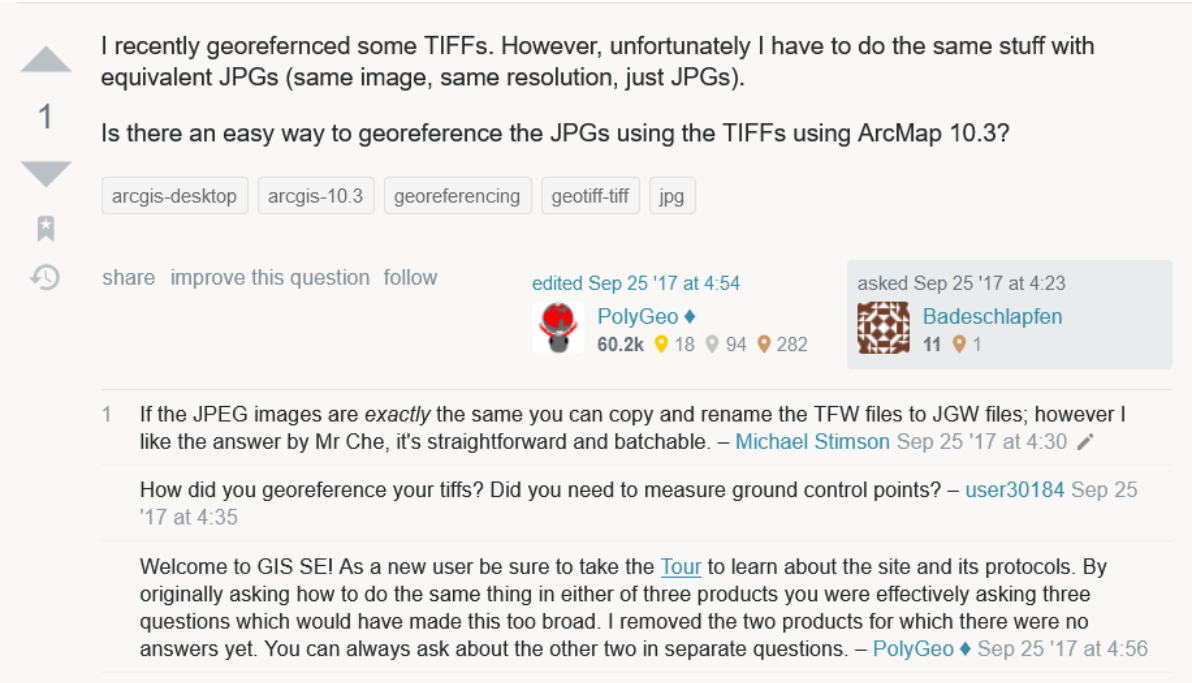
It sounds like maybe you just want to run DeleteIdentical to remove the duplicates. – Tom Aug 14 '18 at 15:09

Well, I want to keep all the values. They are different objects, I just want an automated way to change the ranking numbers. That way I can use that field to pivot. – forci020 Aug 14 '18 at 15:14

Could you edit the post to describe what shape types your data are and what real-world features they represent. It sounds like we might have the XY Problem. – Tom Aug 14 '18 at 15:18

Just added an example with some more clarification. – forci020 Aug 14 '18 at 15:25

Figure 4.7: Example of *exploratory talk*[5].



1

I recently georefernced some TIFFs. However, unfortunately I have to do the same stuff with equivalent JPGs (same image, same resolution, just JPGs).

Is there an easy way to georeference the JPGs using the TIFFs using ArcMap 10.3?

arcgis-desktop arcgis-10.3 georeferencing geotiff-tiff jpg

share improve this question follow

edited Sep 25 '17 at 4:54 PolyGeo ♦ 60.2k 18 94 282

asked Sep 25 '17 at 4:23 Badeschlapfen 11 1

1 If the JPEG images are *exactly* the same you can copy and rename the TFW files to JGW files; however I like the answer by Mr Che, it's straightforward and batchable. – Michael Stimson Sep 25 '17 at 4:30

How did you georeference your tiffs? Did you need to measure ground control points? – user30184 Sep 25 '17 at 4:35

Welcome to GIS SE! As a new user be sure to take the Tour to learn about the site and its protocols. By originally asking how to do the same thing in either of three products you were effectively asking three questions which would have made this too broad. I removed the two products for which there were no answers yet. You can always ask about the other two in separate questions. – PolyGeo ♦ Sep 25 '17 at 4:56

Figure 4.8: Example of *cumulative talk*[6].

[5]https://gis.stackexchange.com/questions/292730/if-then-statement-in-arcmap-to-remove-duplicates
[6]https://gis.stackexchange.com/questions/256469/georeferencing-jpeg-from-tiff-using-arcmap

4.3.1. Findings

The validation performed by manual coding (which can be found in annex B[7]) of a representative sample of 100 comment lists (half referring to questions and half referring to answers) led to three findings:

- First, the vast majority (about 90 %) of comment lists considered contain *exploratory talk* and thus constructive interactions. For the questions, only two cases of non-*exploratory* discussions were found among the set of 50 answer, and four among the questions. In combination with the next points, this indicates that less than 10% of the discussion lists are subject to non-*exploratory* talk.
- Second, no offensive comments were seen in the sample, suggesting that *negative talk* might be uncommon on gis.SE. In a validation performed in the next chapter (see part 5.1) two offensive comments were identified among the 100 lists of comments inspected. This suggests that the share of offensive comments is well below 1%.
- Third, in a total of 15 cases, the interaction was limited to two or three comments reacting on each other. These cases lead to limited interaction only (i.e. one question and one answer) and tend to occur more often among short lists (two or three comments). In some cases, aborted interactions were observed too (see Figure 4.9).

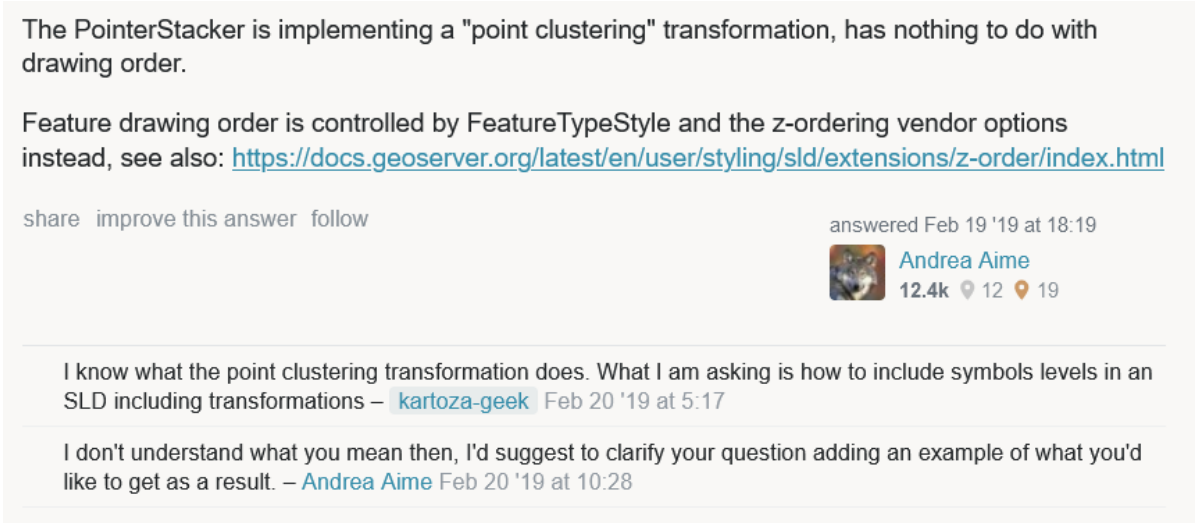


Figure 4.9: Example of an aborted, minimal *exploratory talk* interaction[8].

4.4. Chapter summary and link to assumptions made

Overall, this chapter has introduced the nature of gis.SE beyond the intentional focus on a tool and the openness while building and sharing knowledge. The professional nature of the forum was further confirmed by the fact that activities mainly take place in the period Monday-Friday, which matches the working week in a majority of countries in the world[9].

For the period 2017-2019, most indicators of the forum are stable. This includes, the number of new subscriptions, answers posted and comments and edits. For the questions, however, a concerning phenomena is that the number of questions without activities increases since January 2019. For the number of closures, a decrease was observed in the period January 2017 - August 2018, followed by a more stable period. This decrease did, however, not affect the moderator share which stayed stable. The moderator share for the edits and comments is similarly stable.

[7] the question/answers ids indicate the link with which the post can be found. Simply replace ID in the following link: https://gis.stackexchange.com/questions/ID in a browser

[8] https://gis.stackexchange.com/questions/312785/ordering-symbols-in-geoserver-using-sld-with-rendering-transformations

[9] https://en.wikipedia.org/wiki/Workweek_and_weekend

Beyond the scope of this study, one might note that there seems to be a peak (or a drop) for some activities around January 2017 and that the first half of 2020 shows substantial differences concerning previous years.

For the reaction time, two approaches were tested and compared. On the one hand, the handle method identified a small set of explicit reactions. On the other hand, the ABA method identified different, more implicit reactions and resulted in a bigger set. The trends observed for the 10th and 90th percentile are globally coherent, albeit that the reaction times are slightly higher for the handle method and that an increase since November 2018 is clearer with the ABA method. These differences might be attributed to the fact that handles are an informal habit and thus more likely to be found with experienced users. Such users might require a higher reaction time as they regularly visit the forum (in opposition to one-off engagement) or because the topics they address are more specialised. For both methods, weekday peaks are located in proximity of the weekend (Thursday-Monday). It should also be noted that the question-first answer reaction time is higher than the reaction time between comments, which is coherent with the efforts one is expected to make for the respective actions.

Eventually, manual coding was applied and *exploratory talk* identified as dominating in the representative sample studied. These results are in line with the forum's professional nature introduced by the social science perspective and confirmed by the data analysis. The results confirm the professionalism and seriousness of this content oriented nature of the forum. One should note that this analysis does not mean that no *negative talk* exists. As mentioned in part 3.4, comments can also be deleted - a process of which it is hard to find any traces.

When considering the results of this chapter, it should be kept in mind that they are limited to the period 2017-2019 and see the forum as a homogeneous whole. It is rather likely that several communities exist within the forum and that they follow different dynamics. As this analysis only considered data at forum level, such lower scale dynamics are not addressed. Similarly, the time-based disaggregation of data was only performed at monthly and weekday (yearly) level. This means that phenomena taking place at another level (e.g. the time of the day) are not covered by this analysis.

5
Interaction analysis

While chapter 4 addressed the characteristics of the forum and its different elements, this chapter will focus more specifically on the *exploratory talk*, on the occurrence of constructive interaction. Such interactions are facilitated by the ability to post comments on questions or answers. As, however, the output one can download is a mere list of comments posted at different moments throughout the lifetime of the answer or questions, pre-processing the data is required. The mere number of comments is not representative for the quantity of interaction and has thus only limited meaning. The challenge of transforming comment lists into 'chains' is tackled in section 5.1. In section 5.2, the results are then linked to external factors and an overview is finally provided in section 5.3.

5.1. The problematic of identifying conversation chains within lists

5.1.1. Threaded vs. unthreaded data

The problematic of unthreaded data which was introduced earlier in part 2.1 is definitevely of application for the analysis of comments on SE fora. These belong to the unthreaded category as comments are simply added to a list of other comments referring to the same question. In opposition to other fora (an example of a study on such is Weninger et al. (2013)), the position of comments is no explicit indicator for the relationship to other comments. To address this challenge, several methods studied by Petrovčič et al. (2012) and Henri (1992) served as a theoretical basis.

5.1.2. Chosen approach

The overall approach that a comment replies to a given number of previous comments (introduced by Petrovčič et al. (2012), see part 2.2.3) was taken as a starting scheme. It was decided to use the time between messages as a measure of distance, resulting in the following rule (based on the 90th percentile of the reaction time observation which can be found in Figure D.7):

- Any comment separated by more than 72h from the previous comment does not belong to the same interaction 'chain'.

The choice of the 72h was based on numeric analysis which indicates that, using the ABA approach (see part 4.2.2) for this value, respectively 95.8% (AB) and 94.2% (BA) of the reactions are covered. For the handle method, 72h corresponds to 91.7% of reaction times.

As the 90th percentile of the handle reaction time observed is however higher than 72h on some weekdays (see Figure D.6), the following exception is applied:

- Any comment containing the symbol '@' belongs to the same interaction chain as the previous comment, independently of the time distance between the two comments.

Moreover, the likeliness that a user already involved earlier in the comment list starts a new discussion is considered rather low. Therefore, a final exception is formulated as follows:

- Any comment written by a user who has already contributed to the comment list before belongs to the same chain as the previous comment, independently of the time distance between the two comments.

For GeoJSON - CSS styles are used to modify your points, line, polygons with thickness & color

21

```
{
    "type": "Feature",
    "geometry": {
    "type": "Polygon",
    "coordinates": [[
        [-180.0, 10.0], [20.0, 90.0], [180.0, -5.0], [-30.0, -90.0]
        ]]
    },
    "style": {
        "__comment": "all SVG styles allowed",
        "fill":"red",
        "stroke-width":"3",
        "fill-opacity":0.6
    },
    "className": {
        "baseVal":"A class name"
    }
}
```

http://wiki.openstreetmap.org/wiki/Geojson_CSS

share improve this answer follow

edited Apr 2 '19 at 17:44 Community ♦ 1

answered Mar 29 '12 at 13:25 Mapperz ♦ 46.3k 7 62 123

1 This doesn't appear to be part of the GeoJSON spec. Is this a common implementation? – Mr_Chimp Mar 29 '12 at 13:45

yes common common implementation, that works - GeoJOSN is a 'geospatial data interchange format' – Mapperz ♦ Mar 29 '12 at 14:17

a bit of topic, but is this geoson_css related to carto mapbox.com/carto – Francisco Puga Mar 29 '12 at 15:11

6 That isn't a standard thing and each implementation is going to do this differently. – Calvin May 17 '13 at 10:11

3 QGis (which uses GDAL under the hood) and geojsonlint.com, to name 2 examples, throw errors when using the "style" attribute. – Marian Oct 17 '13 at 14:11

@Mapperz `"style"` is missing a quotation mark. – ustroetz Apr 9 '14 at 12:25

Unfortunately this does not seem to work on Google Maps :(– DaSh Oct 6 '14 at 22:18

@Mr_Chimp See this answer, according to which there is the simplestyle spec by MapBox. – Abbafei Nov 6 '14 at 9:25

The reason this throws errors is because it's missing the required `properties: {}` attribute. Extra attributes such as `style` are allowed by the GeoJSON spec. – Steve Bennett Jun 14 '19 at 6:10

add a comment

Figure 5.1: Example of an ambiguous comment list referring to an answer[1].

[1] https://gis.stackexchange.com/questions/22474/geojson-styling-information/22492

5.1.3. Validation

Sampling approach To evaluate the reliability of this method, a validation was performed using the same sampling approach as used earlier in part 3.2.2. Here again, tags were used to obtain a representative sample of 50 comment lists applying respectively to questions and answers. Additionally, each sample of 50 was randomly divided into two parts:

- A first part with cases where the algorithm only identified a single chain. Here, the number of comments in the original list was maximized while respecting the tag sampling (see appendix A).
- A second part with cases where the algorithm identified at least three interaction chains. Here, no particular number of comments was favored.

The validation performed here is thus optimized to evaluate the behaviour of the automated approach under rather extreme circumstances. This builds upon the assumption that the algorithm is more likely to make errors when it has either a lot of comments to process or when it results in a lot of chains.

Results The results (see annex C[2]) show that the automated approach performs less well than a human. Discrepancies with manual interpretation were obtained in ten out of 50 question comment lists and in twelve out of 50 answer comment lists. One might note that in a few cases the order of the comments is of very limited meaning as several discussions are intertwined (see Figure 5.1 for an example). A total of nine comments were found to be ambiguous in the answer and twenty-two in the question set. In such cases, the angle of the manual approach was to maximize the number of chains, only grouping messages with explicit hints to another one in a same chain.

As discrepancies in ten comment lists out of 50 refering to questions and twelve out of 50 referring to answers are not deemed acceptable, an additional consideration was taken in the analysis. When comparing the results of the manual and automated approach, it was noted that in some cases (see Figure 5.2 for an example), no more than one message per chain was wrongly added/missed. When allowing a tolerance of one wrong message per chain, the number of errors decreases to three out of 50 comment lists with a wrong result (for both question and answer sets; see Figure 5.3 for an example). Given that the validation is set up to test extreme situations, this error is deemed acceptable.

5.1.4. Selection of additional criteria

Given the results of the validation, it was decided to further implement the tolerance of one wrong message in the analysis. Also, one of the findings of the first validation (see part 4.3) is that a limited number of comments tends to limit the interaction between users. Therefore, a 'high interaction' definition was chosen to scope the analysis of *exploratory talk* within this research:

- A High Interaction (HI) chain is an comment chain in which at least two different users posted each two comments. Moreover, a comment following a previous comment by the same user is not taken into account.

In the case where a comment is matched to the wrong chain, the likeliness that the identification of high interaction chains will be affected is reduced. On the one hand, it can not be completely excluded that the first or last of several comments from a user involved in the chain is missed (or that a comment of the same user but from another discussion is wrongly added before or after the first comment belonging to the chain). On the other hand, the second exception to the 72h rule (that no new chain can be started if the author was already active earlier in the comment list) reduces the likeliness of such cases.

Furthermore, the one comment tolerance is only problematic in cases where the user posted *exactly* one comment in the chain and another one is added by error, or where the user posted *exactly* two comments in the chain and one of these is missed by error. Beyond this, there might also be cases where more than one comment is wrongly added or missed. However, the corresponding error rate identified during validation is deemed acceptable. For other situations, however, such as when two users each posted more than two comments, or more than two users each posted two comments, the one comment tolerance is not problematic.

[2] the question/answers ids indicate the link with which the post can be found. Simply replace ID in the following link: https://gis.stackexchange.com/questions/ID

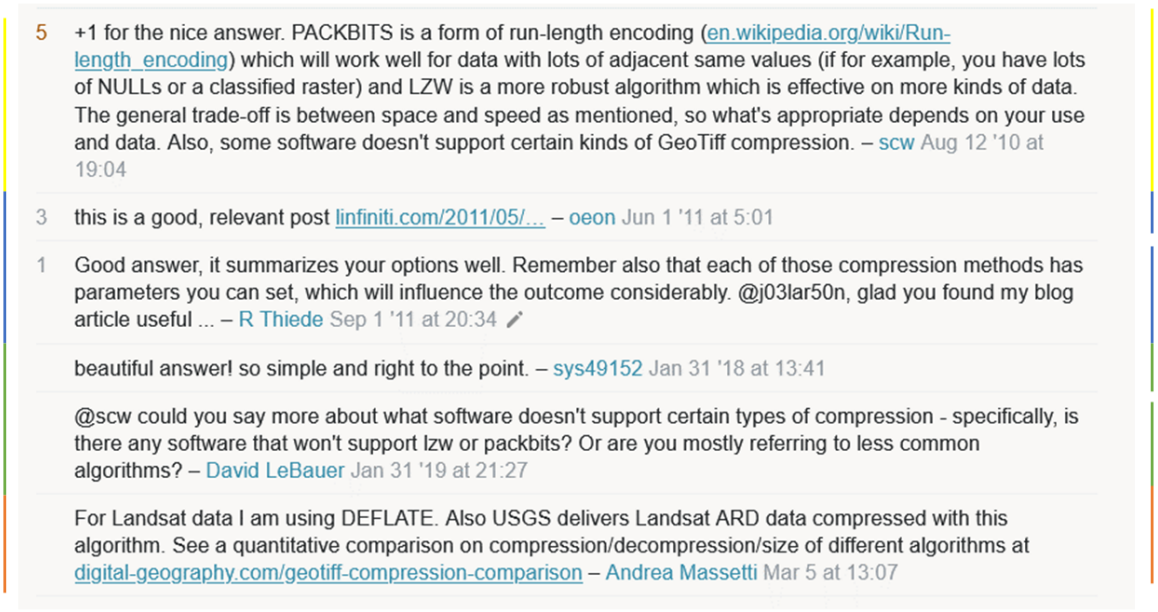


Figure 5.2: Example of a comment list referring to an answer in which two messages were wrongly added[3], but for two different chains. Here, no more than one wrong message per chain can thus be found. The line on the left shows the chains automatically identified and the line on the right the result of the manual analysis.

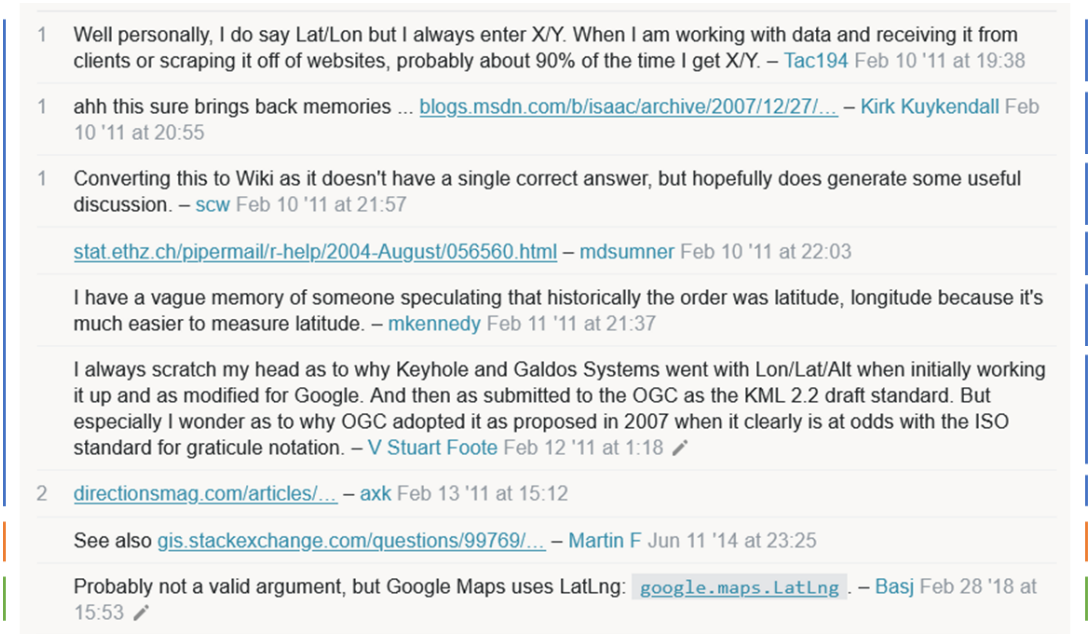


Figure 5.3: Example of a comment list referring to a question in which seven messages were wrongly grouped to a chain[4]. This is a typical case in which wrong results are obtained even with a tolerance of one wrongly added/missed message per chain. The line on the left shows the chains automatically identified and the line on the right the result of the manual analysis.

[3]https://gis.stackexchange.com/questions/1104/should-gdal-be-set-to-produce-geotiff-files-with-compression-which-algorithm-sh
[4]https://gis.stackexchange.com/questions/6037/displaying-coordinates-and-inputs-as-latlon-or-lonlat

5.2. Identification of external and internal factors related to constructive interaction

Following up on the definition of high interaction, this second part performs several analyses on the circumstances in which such high interaction chains occur. These analysis follow three thematic axes. The first one looks at popularity metrics at thread and question/answer level. The second one looks at the types of questions and at their content. A third axis investigates the interactions around them and the users involved in the threads.

5.2.1. Units of analysis

The analysis performed in this section links HI chains to three different levels of analysis: the thread, the question/answer and the chains themselves. Before presenting the different calculations performed by leading themes, an overview of the data at the different levels is provided.

It should be noted that the numbers are not fully coherent. This is because at each level of analysis, the units with at least one activity (i.e. answer, question or comment) in the period 2017-2019 are taken into account. As the three levels do not only refer to units of different sizes but also linked at different scales, it was tried to reach maximum coherence by applying the same time frame rule at all levels. Any other option would have been hardly applicable coherently across the levels (e.g. how to handle threads with one HI chain which took place inside and another one outside the years 2017-2019).

At thread level, the dataset (presented in Table 5.1) contains 61638 threads (with at least one question/answer/comment in the period 2017-2019) of which the vast majority (94.5%) contains no HI. The most relevant set to compare these with are the threads with at least one HI per thread (thus within the comments of either a question or an answer) which represent 5.4%. The set with threads containing two HIs is also used as an indicator, although its size of 109 is very limited. The set with threads containing three HIs however, contains only 6 threads and is therefore dismissed in the analysis.

number of HI chains	number of threads [%]	share [%]
0	58272	94.5
1	3251	5.4
2	109	0.2
3	6	0.001
total	61638	

Table 5.1: Table showing the overview of the data at thread level.

At question/answer level, the data contains 61,638 questions and 143,756 answers (see Tables 5.2 and 5.3). The share of questions containing at least one HI is slightly higher for answers than for questions (10.0 vs. 6.0%).

number of HIs	number of questions [%]	share [%]
0	135303	94.1
1	8050	5.8
2	403	0.2
total	143756	

Table 5.2: Table showing the overview of the data at answer level.

number of HIs	number of answers [%]	share [%]
0	56082	90.1
1	4931	9.0
2	625	1.0
total	61638	

Table 5.3: Table showing the overview of the data at answer level.

Finally, at the lowest level - the HI interaction chains themselves - the *exploratory talk* is not compared with its context anymore. Instead, the characteristics of the HI interaction themselves are anal-

ysed. In total, 2951 chains identified within lists relating to questions and 5004 chains in lists relating to questions are analysed.

5.2.2. Popularity metrics

Number of views (thread level)

To analyse the impact of HIs on popularity metrics, two approaches were taken: on the one hand, the number of views was considered. This one takes into account visits from people external to the forum that were, for instance, redirected from a search engine (for which SE has optimized its sites[5]). On the other hand, as the number of views is an indicator which keeps growing as long as the knowledge of the thread is being retrieved (and thus relevant), not only the total number as of 1st of June 2020 but also the number of views since the creation of the thread was calculated.

Absolute number of views The analysis of the absolute number of views (table 5.4) is not conclusive when looking at the average and standard deviation. While the average is higher, the standard deviation also increases, indicating a bigger spread of the data. The 10th and 90th percentile analysis allows to further specify this spread, clearly showing that the increase of the number of views is mainly taking place among the highly viewed threads, as shown by the 90th percentile. The increase for threads with two HIs should, however, be taken with care considering the small sample size in that category.

number of \ac{HI}s	average	standard deviation	10th percentile	90th percentile
0	1135.54	6049.67	24	1771
1	2750.28	10872.15	89	5436
2	7074.55	14422.93	131	16036
total				

Table 5.4: Table showing the statistics for the thread's absolute number of views, grouped by number of HIs.

Views per day Looking at the number of views per day (table 5.5), similar observations as for the absolute number of views can be made. Here again, the increase of views with the number of HIs is mostly found in the 90th percentile.

number of \ac{HI}s	average	standard deviation	10th percentile	90th percentile
0	0.81	2.61	0.04	1.61
1	1.70	4.25	0.15	3.76
2	3.38	5.77	0.29	7.11
total				

Table 5.5: Table showing the statistics for the thread's daily average number of views, grouped by number of HIs.

Votes (question/answer level)

Votes For the questions, the difference in terms of the total number of votes and in terms of net vote score is less clear than the one of the number of views. As shown in Table 5.6, the only indicator clearly changing between questions is the standard deviation. As the 10th and 90th percentiles show, this is however not related to higher extremes. For the answers, the situation is a bit different (see Table 5.7). For both the number of votes and the net score, the average and the 90th percentile do increase with the number of HI chains. The increase in these two indicators is of similar proportion. As the standard deviation and the 10th percentile stay stable, this indicates that answers with HI chains are likely to achieve higher net scores (which requires more votes). This does not necessarily indicate that the answers are more appreciated but does at least indicate that they are more popular among registered users (i.e. more users tend to upvote them).

[5] https://meta.stackexchange.com/questions/14056/seo-in-stack-overflow

number of HI chains	average	standard deviation	10th percentile	90th percentile
	total of votes up & downvotes (questions)			
0	4.895545	10.42752	2	8
1 or more	4.95092	7.526138	2	8
	net score (questions)			
0	2.570331	10.41059	0	5
1 or more	2.601227	7.433506	0	6

Table 5.6: Table showing the statistics for the question's absolute number of votes and net result scores, grouped by number of HIs.

number of HI chains	average	standard deviation	10th percentile	90th percentile
	total of votes up & downvotes (answers)			
0	5.983287	14.35707	2	11
1 or more	7.820588	14.39826	2	14
	net score (answers)			
0	3.630919	14.34366	0	8
1 or more	5.314706	14.21468	0	11

Table 5.7: Table showing the statistics for the answer's absolute number of votes and net result scores, grouped by number of HIs.

5.2.3. Question types and content

Question typology: number of answers (thread level)

A second factor which was studied is the correlation between the number of answers and the number of HI chains in the same thread. As can be seen in Table 5.8, there is no notable difference between the sets with 0 and 1 HI threads. The set with 2 HI chains tends to indicate that threads with more HI chains tend to be threads with more answers. However, care when interpreting these results is necessary once more due to the lower sample size.

Number of answers The number of answers (table 5.8) appears to increase with the number of HI chains. However, this increase is considerably less strong than for the number of views. Also, lower standard deviations indicate less spread of the data. Overall, the indicators do not suggest notable differences linked to the number of HI chains in the thread.

number of HI chains	average	standard deviation	10th percentile	90th percentile
0	1.47	1.00	1	2
1	1.74	1.36	1	3
2	3.18	1.77	2	6
total				

Table 5.8: Table showing the statistics for the average number of answers, grouped by number of HIs.

Differences in frequency of tags (thread level)

In this step, the share of HI chains among all chains was calculated for each tag (taking into account the tags as of the 1st of June 2020, thus after edits). Table 5.9 offers an overview of the results, showing the 10 tags with the highest and the lowest share. The total number of occurrences (for all tags) is 162,175 for the regular chains and 10,204 for the HI chains. Only tags appearing in at least 0.1% of either of these sets were taken into account.

Interestingly, the tags starting with 'arc' and thus referring to proprietary software by ESRI appear more often in the threads without HI chains. Additionally, the tags '3D-analyst' and 'network-analyst' refer to extensions to ESRI software. Beyond this, one might note that 'google-maps-api', 'mapbox' and 'mapbox-gl-js' also refer to proprietary software. The tag 'vector-tiles' refers to a technology closely linked to mapbox and google. Eventually, only the tag 'layouts' is not connected to proprietary software here.

tag name	number of regular chains (out of 162 175)	number of HI chains (out of 10 204)	share of HI chains [%]
highest shares of regular chains			
arcgis-10.6	183	3	1.6
layouts	175	4	2.2
mapbox	334	8	2.3
vector-tiles	180	5	2.7
network-analyst	251	7	2.7
mapbox-gl-js	173	5	2.8
3d-analyst	170	5	2.9
google-maps-api	200	6	2.9
arcgis-server	669	22	3.2
arcgis-online	576	19	3.2
highest shares of HI chains			
null	65	12	15.6
fiona	105	15	12.5
cursor	328	44	11.8
iteration	91	12	11.7
overpass-api	136	17	11.1
extract	123	15	10.9
linestring	157	19	10.8
select-by-attribute	208	25	10.7
union	111	13	10.5
iterator	94	11	10.5

Table 5.9: Table showing the tags with the highest and lowest share of HI chains (only tags with at least 0.1% of total occurrences among HI or regular chains were taken into account).

Among the tags with high shares of HI chains, one might note that 'null' which refers to empty data is on the top of the list. The tags 'cursor', 'iteration', 'extract', 'linestring', 'select-by-attribute', 'union', 'iterator' belong to the technical terminology of the geometry and programming fields. Additionally, one can find two tags referring to FOSS software, namely 'overpass-API' and 'fiona'. There is thus a clear contrast with regard to the proprietary software mainly encountered among non-HI chains. A plausible explanation to these results is that other fora might be more popular for commercial software. In the case of ESRI, one might note the official GeoNet forum[6] on which users are thus more likely to get answers to their questions.

One should however note that the meaning of this tag analysis is limited. First, one should note that no highly frequent tag is present among the highest share of HI and regular chains. For regular chains, the highest value is 669 which is roughly equivalent to 0.4% and for HI chains, this one is 44, roughly equivalent to 0.4% too. Second, the tags only give hints about the theme addressed but not necessarily about the form and the exact content of the question (e.g. presence of technical content). The next analysis, namely the Q/A level, however, addresses this level more specifically.

Most distinguishing words in posts the comments lists result from (question/answer level)

To extend the analysis of the question and answer content beyond the tags, the words composing the actual content of the questions and answers were analysed too. Here again, the frequency of the word occurrence within questions and answers of which the related list of comments does or does not contain an HI chain was calculated (using the content as of 1st of June 2020). Here, it is important to note that code sections have been skipped (when correctly marked in the raw data) and will be addressed in the next step (see Table 5.1). In total, 582,0728 words in answers and 11,227,925 words in questions without HI chains in comments were processed. For the cases where an HI chain is present in comments, 1,012,591 words of answers and 789,452 words of questions were processed.

[6] https://community.esri.com/welcome

top 20 words in regular answers	**share of HI chains in occurrences [%]**	**regular occurrences (total: 11 227 925)**	**HI occurrences (total: 1 012 591)**
was	4.6	19407	931
ArcGIS	5.7	12394	744
my	6.0	17707	1122
map	6.5	17727	1233
I	6.6	76498	5445
out	6.7	13506	971
also	6.7	19395	1400
up	7.0	12817	958
may	7.0	14734	1111
but	7.0	41552	3138
other	7.1	15827	1206
tool	7.1	16086	1229
could	7.2	20432	1574
no	7.2	11718	904
data	7.2	37886	2936
way	7.2	12802	994
there	7.2	17298	1350
used	7.3	14312	1126
has	7.3	20052	1587
some	7.4	19886	1579
top 20 words in HI answers			
+	12.5	24403	3495
*	11.6	9724	1274
=	11.2	160750	20319
import	11.2	9778	1230
var	11.0	11219	1388
return	10.5	11566	1364
description	10.0	22563	2495
values	9.8	13857	1506
value	9.8	13250	1439
field	9.7	18328	1978
if	9.6	51821	5520
output	9.6	10183	1084
code	9.6	16017	1697
points	9.5	15158	1592
each	9.4	19971	2074
polygon	9.4	11086	1149
image	9.3	31078	3203
point	9.3	18171	1860
line	9.2	13572	1383
first	9.2	11732	1193

Table 5.10: Table showing the top words contained in regular and HI chains referring to answers (only taking into account words appearing in at least 0.1% of either regular or HI chains).

top 20 words in regular questions	**share of HI chains in occurrences [%]**	**regular occurrences (total: 5 820 728)**	**HI occurrences (total: 789 452)**
add	9.2	6615	673
var	9.8	10253	1113
how	9.9	13480	1487
function	10.0	7041	778
know	10.1	9182	1027
ArcGIS	10.2	7661	866
map	10.2	13945	1585
can	10.2	25448	2894
way	10.4	11701	1354
there	10.4	18280	2120
want	10.5	16789	1978
layers	10.5	6660	785
like	10.6	18729	2211
create	10.6	10161	1204
do	10.6	19930	2374
use	10.7	16678	1997
any	10.7	11074	1332
would	10.8	15942	1926
a	10.8	143449	17338
need	10.8	11699	1420
top 20 words in HI questions			
:	18.6	8185	1865
script	15.4	4762	867
1	14.9	6035	1054
description	14.8	13275	2309
#	14.1	6812	1118
0	14.1	5681	930
error	13.9	7350	1189
output	13.9	4942	795
shapefile	13.8	5809	927
+	13.7	12573	1993
run	13.6	5059	799
problem	13.6	5792	910
tried	13.6	10007	1571
same	13.5	10521	1644
no	13.3	7987	1226
|	13.3	5474	840
at	13.3	21345	3268
-	13.3	10517	1609
image	13.2	20445	3120
=	13.0	96245	14429

Table 5.11: Table showing the top words contained in regular and HI chains referring to questions (only taking into account words appearing in at least 0.1% of either regular or HI chains).

Subsequently, the word-specific share of HI chains as well as the 20 words most characterising for their presence were computed. The results for questions and answers are shown in Tables 5.10 and 5.11. One can note that rather technical words and symbols tend to produce the highest share of HI chains (e.g. 'error', 'script', 'shapefile' for questions and 'import', 'value', 'polygon', 'point' for answers). In contrast, the words with the lowest share of HI chains tend to be rather typical words belonging to the general language. Some exceptions do exist, for the answers these are: 'ArcGIS', '|' symbol, as well as 'data'; and for the questions: 'var', 'function', 'ArcGIS', 'map' as well as 'layers'. Overall, this is coherent with the observations made for the tags in two ways. First, this confirms that the term 'ArcGIS' is negatively impacting the occurrence of HI chains (share of only 5.7% for answers and 10.2 % for questions). Second, the technicality of terms with a high share of HI chains is in line with the technicality suggested by the tags previously shown in Table 5.9.

Care is however recommended when interpreting the results as the gap between the regular and HI occurrences is smaller than for the tags (see Table 5.9). Taking only into account the top 10, the gap for the tags was 7.3% (10.5 - 3.2%), while it is only 2.7 % for the answers (9.7 - 7.0%) and 3.3% for the questions (13.7 - 10.4%). This indicates that the groups of words identified here are less qualifying for regular or HI chains than the tags were. Concerning the share of the occurrences with regard to the total, better scores than for the tags (maximum of 0.4 %) are achieved: for the questions, the word 'a' represents 2.4 % of regular and 2.1 % of HI occurrences, followed by '=' which represents 1.6 % of regular and 1.8 % of HI occurrences. For the answers, the maximum is reached by the symbol '=' which represents 1.4 % of regular and and 2.0 % of HI occurrences, followed by 'if' (about 0.5 % of both regular and HI occurrences). To put it in a nutshell: while the terms obtained by this analysis at question/answer content level are more representative, they are less characterising than the ones obtained at thread/tag level in Table 5.9.

Presence of code snippets, images and external links in the posts (question/answer level)

The last analysis performed with regard to the content of questions and answers is the presence of code snippets, images and external links in the the posts to which the comment lists refer. These can be reliably identified as the text content downloaded is in raw format. Codes snippets, for instance, are thus marked with the indicator '<code&gt'.

As can be seen in Table 5.12 for the situation of questions, the presence of an HI chain increases in all studied scenario. The increase is strongest for the presence of images (4%), followed by code snippets (2.2%) and external references (2.1%). Similar observations can be made for the answers, as shown in Table 5.13. Here too, images show the strongest effect (an increase of about 3.3%), followed by code snippets (2.2%) and external references. For the latter, the increase is minimal with only 0.4%. The strongest increase is observed for the combination of all three factors. When a code snippet, image and external reference are all three present, the increase is of respectively 6.5% (questions) and 4.8% (answers). Furthermore, these trends are globally valid for the situations with two HI chains too.

questions:	**all**	**with code snippet**	**with image**	**with ext ref**	**all three**
total number	61638	26170	15404	25881	5517
2 high interaction chains [%]	1.004	1.299	1.623	1.325	1.976
1 high interaction chains [%]	9.014	11.303	13.100	11.170	15.516
low interaction [%]	90.986	88.697	86.900	88.830	84.484

Table 5.12: Table showing the relation between the presence of code snippets, images, external references in answers and the presence of HI chains in the comment lists.

	all answers	**with code snippet**	**with image**	**with ext ref**	**all three**
total number	143756	65059	20205	77155	8313
2 high interaction chains [%]	0.280	0.383	0.460	0.315	0.650
1 high interaction chains [%]	5.880	8.110	9.201	6.213	11.681
low interaction [%]	94.120	91.890	90.799	93.787	88.319

Table 5.13: Table showing the relation between the presence of code snippets, images, external references in answers and the presence of HI chains in the comment lists.

	probability [%]
HI among all questions	9.9
HI given code snippet in question	12.4
HI given image in question	14.5
HI given ext. ref in question	12.3
HI given code, image and ext. ref	17.1

Table 5.14: Results of the bayesian inference for the presence of code, images and external references in questions

	probability [%]
HI among all answers	6.1
HI given code snippet in answer	8.5
HI given image in answer	9.6
HI given ext. ref in answer	6.5
HI given code, image and ext. ref	12.3

Table 5.15: Results of the bayesian inference for the presence of code, images and external references in answers

The bayesian inference calculations shown in Tables 5.15 and 5.14 further confirms these observations. Here, no distinction between the presence of one and multiple HI chains were made. Similarly to the raw results, the presence of images increases the probability of encountering a HI chain more than other factors. The highest increases are, here again, achieved with the combination of all three factors. It should however be noted that such situations are only encountered in 5.8% of the answers and in 9.0% of the questions, while external references or code snippets can be encountered in nearly half of the cases.

Furthermore, one might note that for both questions and answers, the number of messages with images is lower than the one with code snippets or external references. This suggests that, despite the clearly positive effect in the case of answers, this mean is not applied as often as others. This might also be related to less efforts required to include code snippets (and external references) than images.

5.2.4. User types and interactions

A last axis of analysis is the users involved in the threads and HI chains. First, the closure and reopenings of threads are studied. Second, to characterize their role, a check with regard to their status in the forum (moderator) or earlier in the thread (question or answer author) is performed. To study the impact, a check with regard to the number of words in the HI discussion chain is performed. The same approach is used to explore the impact of edits on the length of the HI chains. Finally, their time since registration on the forum and the distribution of their activities before participating in a HI chain is analysed.

Closure/reopening votes (thread level)

For this part, the relation between the presence of HI chains and the application of moderation policies was analysed (see Table 5.16). This was done in two steps. First, the percentage of closures and reopenings was analysed at the thread level. The percentage of threads locked was also calculated but applies to a sample too small to give conclusive results. Subsequently, the impact of closure decisions taken by moderators vs. 'normal' users was analysed too.

Number of HI chains	% of threads closed	% of threads reopened	% of threads locked
1	9.7	3.5	0.1
0	12.0	3.1	0.4

Table 5.16: Table showing the relation between closures/reopening and the presence of an HI chain.

Relation between thread closures and presence of HI chains Overall, it seems that the presence of HI chains is negatively linked to the number of closures (see Table 5.16). While the 58,272 threads

without an HI chain end up closed in 12.0% of the cases, this metric is reduced to 9.7% (out of 3,251 cases) in the presence of one HI chain. Furthermore, the number of thread reopenings is slightly higher in the presence of one HI chain (3.5% vs. 3.1%).

Closure decisions by moderators To get a better insight in the role of moderators within the closing process, the share of closures/reopenings in which moderators took the decision was calculated (see Table 5.17). If a sufficient number of non-moderator users vote for closing/reopening, such decisions can be taken without the action of a moderator. Overall, the distribution is coherent with the shares observed in Figure D.3. There is no notable difference related to the presence of HI chains, except that the reopening share is slightly higher for the cases where HI chains are present.

Number of HI chains	% of moderator closures	% of moderator reopenings
1	70.3	85.0
0	70.6	80.2

Table 5.17: Table showing the moderator share among closures and reopenings.

Number of words per HI chain and impact of moderator presence (chain level)

To cover the impact of moderator interventions at the chain level, the number of words in the HI chains was computed. As the downloaded text is of a raw form and might thus include links or code snippets, the number of (single) space characters was taken as a proxy. For each chain, it was also established whether a moderator is among the high-intensity users (thus the users which contributed two or more messages to the chain). Figure 5.4 shows the results for the questions and 5.5 for the answers. In both cases, the distribution are similar for the HI chains in which a moderator is involved and for the HI chains in which only regular users are. The participation of moderators does therefore not seem to have an impact on the number of words in the HI chain.

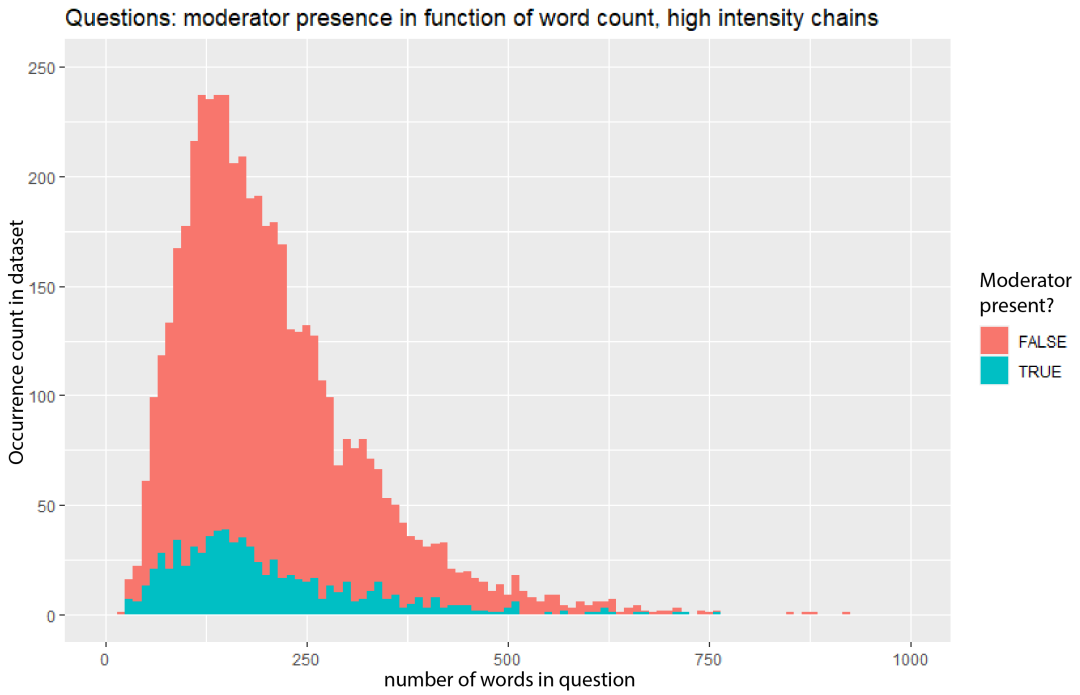


Figure 5.4: Histogram showing the number of words per HI chain referring to a question and the cases in which a moderator was one of the high intensity users of the same chain.

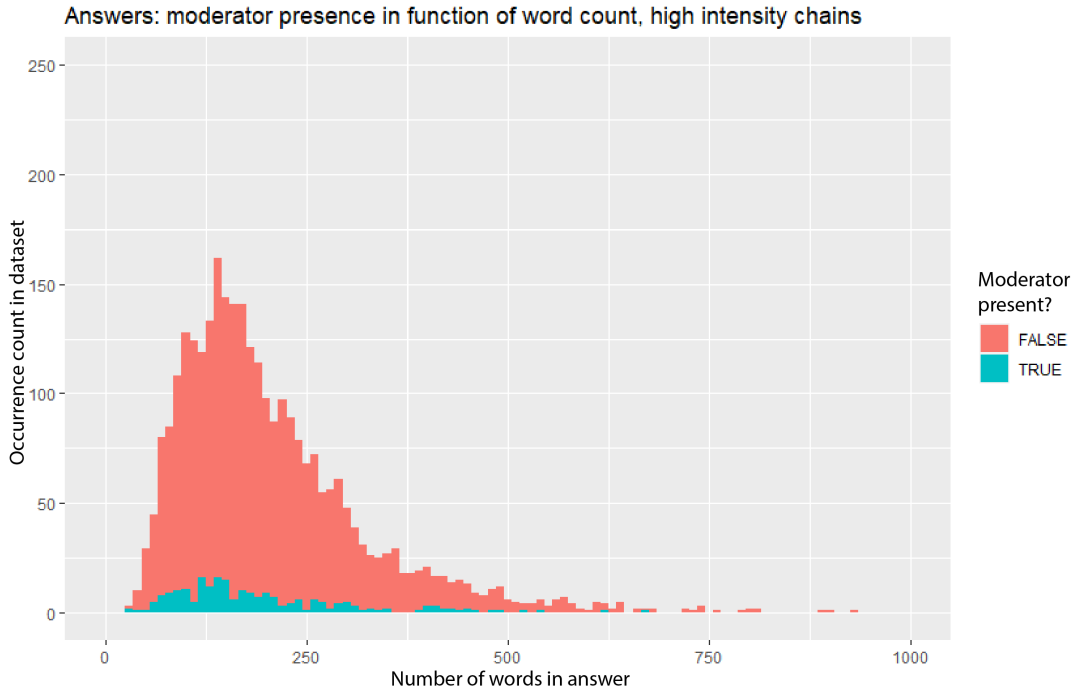


Figure 5.5: Histogram showing the number of words per HI chain referring to an answer and the cases in which a moderator was one of the high intensity users of the same chain.

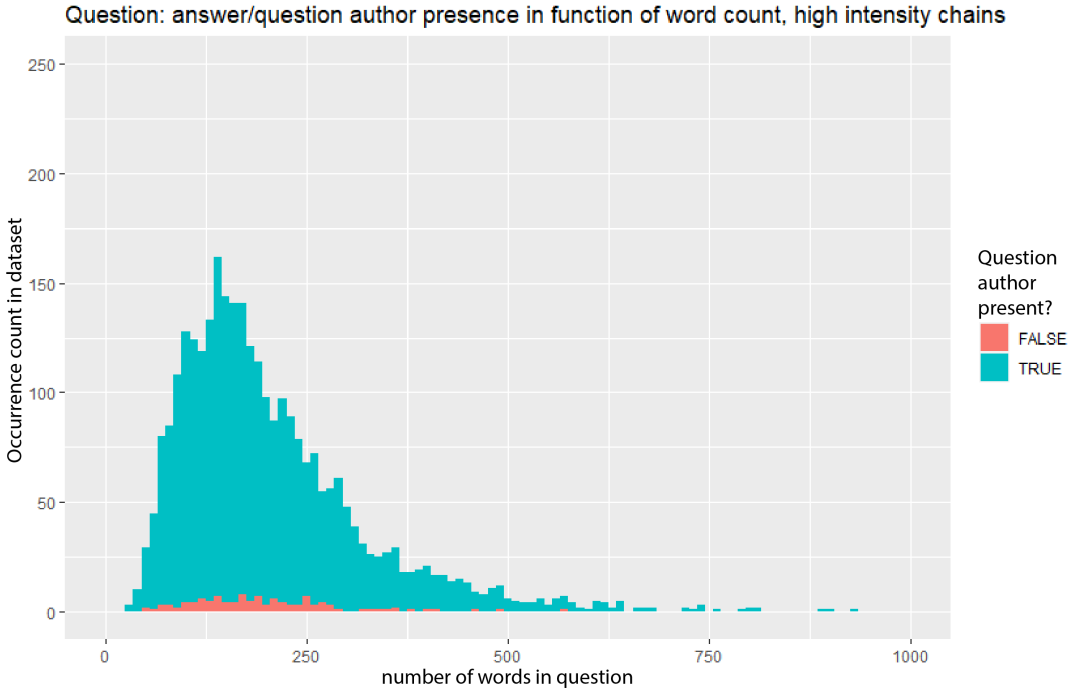


Figure 5.6: Histogram showing the number of words per HI chain referring to a question and the cases in which the original question author was one the high intensity users of the same chain.

Presence of question/answer author(s) in HI chains and impact on number of words.

Extending on the analysis performed in part 5.2.4, the impact of the presence of the question/answer author (as a chain participant with at least two comments) on the number of words of the HI chain was

analysed.

Similarly to the previous analysis regarding the participation of moderators in discussions, the distribution of all subsets is very similar. The participation of the authors does therefore not have an impact on the word number of the high-intensity chains. Nevertheless, it is worth noting the share of their participation. Out of 2951 HI chains linked to answers, the answer author was present in 2837 cases (96%) and in 2512 cases (85.1%) even the author of the thread question (see Figure 5.6). For the HI chains referring to questions (see Figure 5.7), the question author was involved in 4831 of all the 5004 chains (96%).

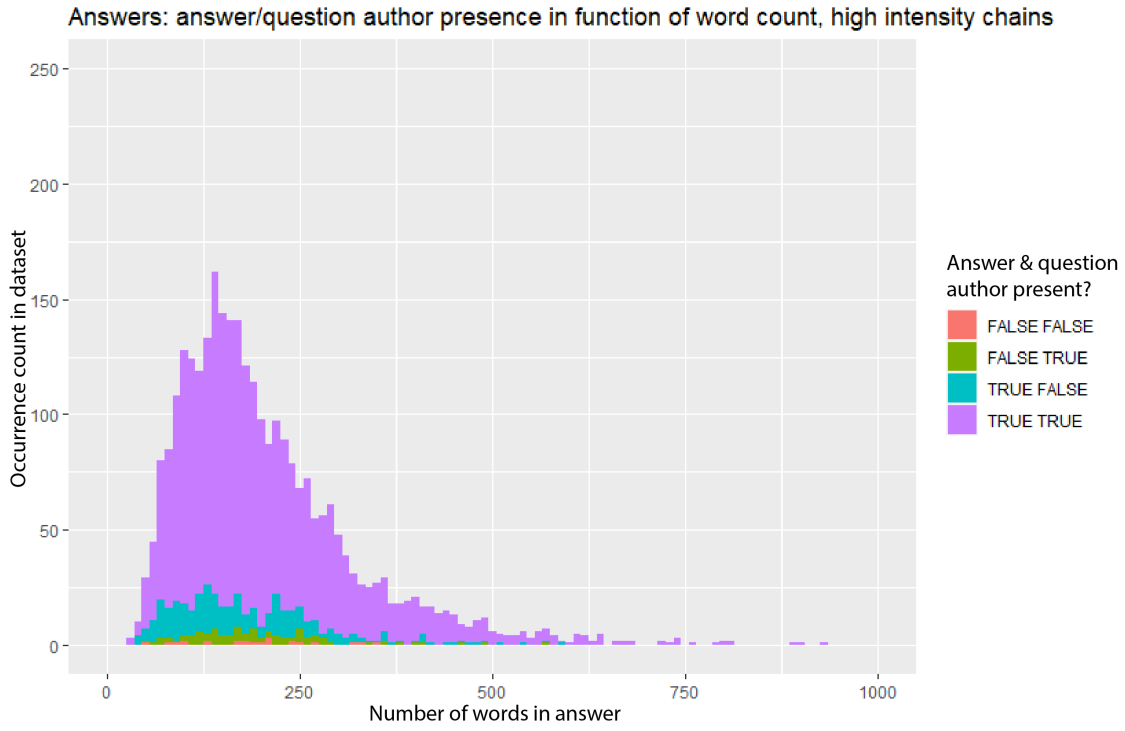


Figure 5.7: Histogram showing the number of words per HI chain referring to an answer and the cases in which a question or answer author was one of the high intensity users of the same chain.

HI chains resulting in edits and impact on number of words

Out of the 5004 HI chains related to questions, 1697 (34%) have an edit that can directly be linked to the HI chain (thus performed by a user that has participated with at least two messages in the HI chain)[7]. Out of these 5004 cases, only 44 are edits by moderators. For the 1697 HI chains related to answers, 969 edits have been identified - resulting in a higher share of 57%. Moderators were, here again, only involved in few cases - namely 52. The impact of such edits on the number of words in the HI chain was plotted in a histogram (see Figures 5.8 and 5.9). However, the presence of edits does not notably affect the distribution of the number of words in the HI chain.

[7] To avoid edits which only address spelling and grammar, only cases which have an impact of more than 10% on the word count were considered - see part 3.4.1. Also, a 10-minute tolerance was implemented to allow edits posted shortly before or after comments to be included

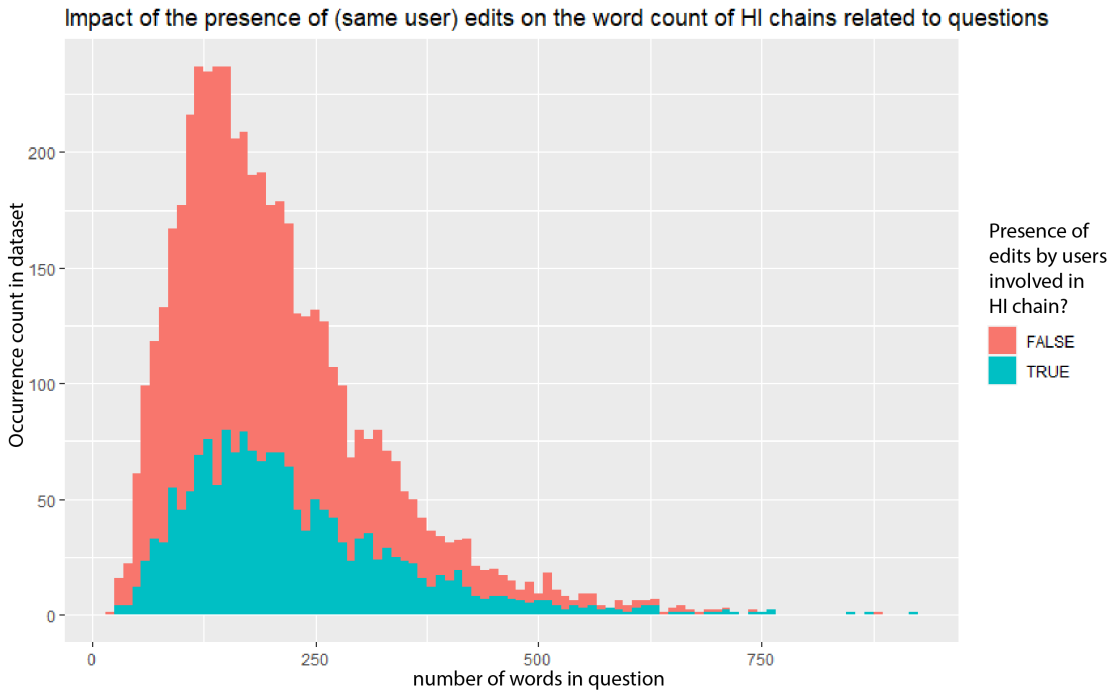


Figure 5.8: Impact of the presence of edits linked to a question-related HI chain on the number of words of the chain.

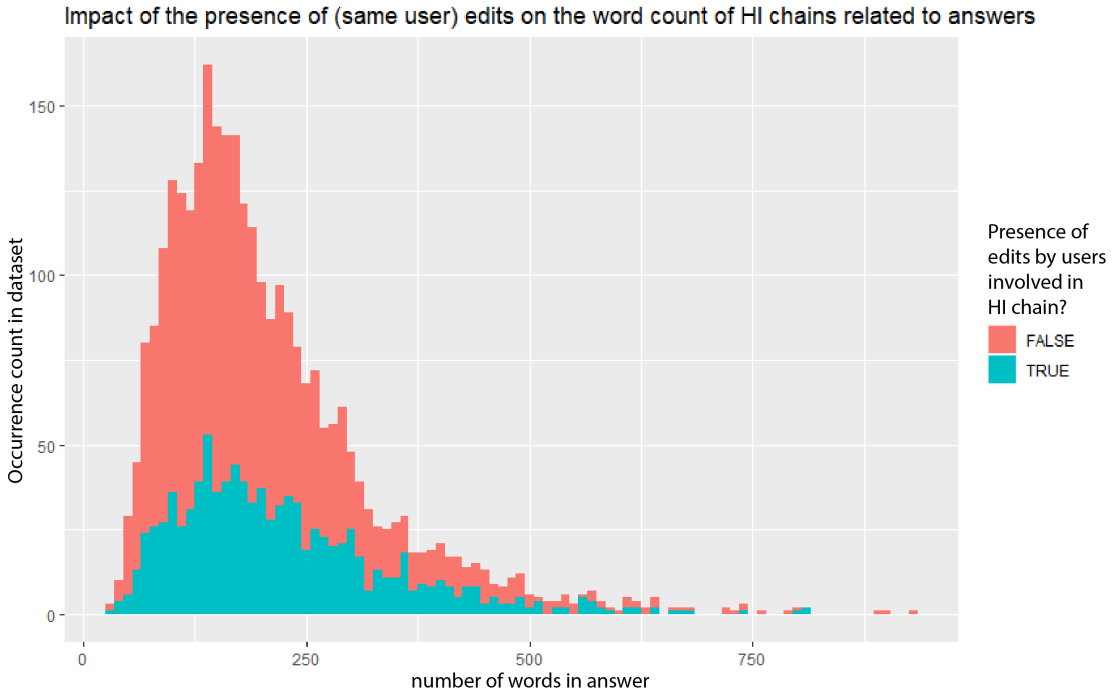


Figure 5.9: Impact of the presence of edits linked to an answer-related HI chain on the number of words of the chain.

Time since registration (chain level)

The time elapsed between the registration of the user and the start time of the HI chain in which at least two comments were contributed was computed for this step. The results can be found in Figure 5.10. There is a strong tendency of users who registered shortly before participating, reflected in the highest peak which is close to the y-axis of the plot (bin located between 19 and 55 days). Nevertheless, later peaks do also exist, namely at approximately 1.5, 3, 4.5 and 6 years[8].

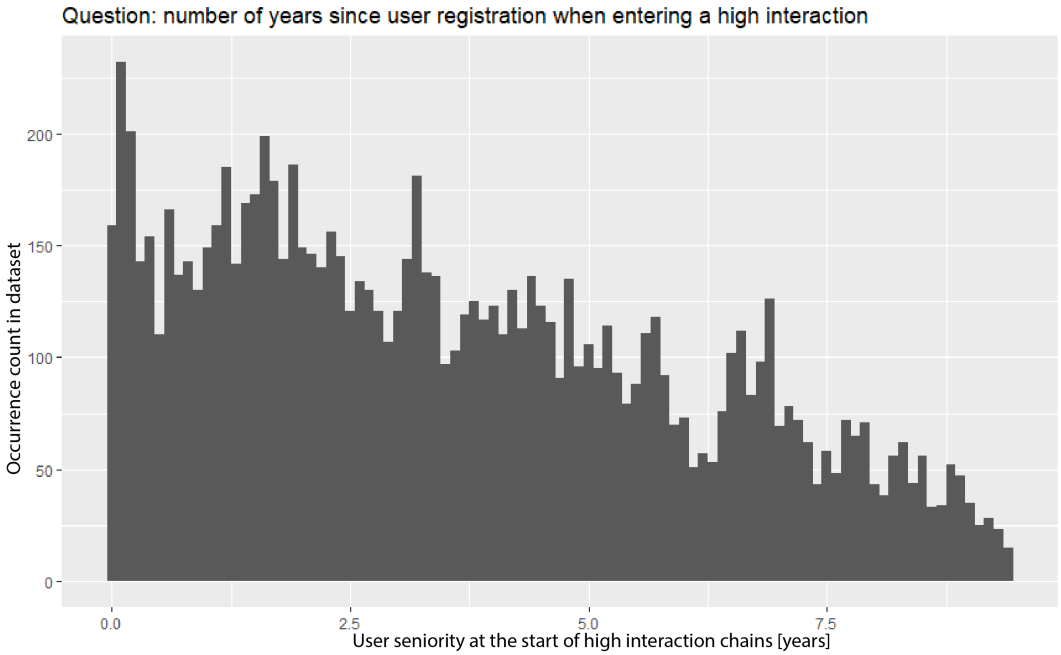


Figure 5.10: Histogram of the time elapsed between the registration of the user and the start time of the HI chain in which at least two comments were contributed by the same user.

Types of activities performed at the time of contribution (chain level) Next to the time since registration, the previous activities of each user contributing at least two messages to a HI chain were also computed for the time of the chain started. For this analysis, new users, defined as having less than 11 previous activities, were not taken into account. Interestingly, there is no single case in which the number of previous answers or questions was higher than any other activity. Therefore, the number of comments is highest in 73.3% of the cases and the number of edits in 26.7% of the cases. If only considering the questions and answers, it appears that 72.6% tend to answer more questions than they ask. This indicates that most users contributing with at least two messages to HI chains had comments as their main previous activity. In some cases, users appear to edit more than they comment. Also, they rather tend to answer than to ask questions.

role	**frequency [%]**
commenter	73.3
editer	26.7
—	
answerer	72.6
asker	27.4

Table 5.18: Table showing the distribution of the dominant activity at the moment of a HI chain start, for each user involved with at least two messages in that same chain. The editer/commenter roles are dominant in the absolute sense - meaning that in no single case the number of questions/answers previously posted exceeded the number of answers and comments.

[8]this might indicate a periodicity in the forum usage but might also be a mere coincidence

5.3. Chapter overview and link to assumptions made

This chapter has conducted a series of analyses to identify situations in which interaction with *exploratory talk* occurs (which is used as a proxy for the sharing of the 'mindset' required to solve a question). To identify such situations in an automated way, a strict definition tackling the challenge of identifying interaction in unstructured comment lists was necessary. After developing and implementing an algorithmic approach, validation was performed. As the identified error margin exceeded the tolerance for reliable analysis, the definition was further refined. Eventually, so-called HI interaction in which at least two users posted each at least two comments were used as a central indicator for dialogue with *exploratory talk*.

Analysis performed The analysis itself was conducted at several levels corresponding to the functioning of the SE forum system. Three different units of analysis were used: the thread, the question/answer and the HI chain itself. Beyond this, four major axes were followed in the research, with in each case both conclusive and non-conclusive results.

Popularity metrics For the popularity metrics, one might note that the number of absolute and daily views is higher for threads with a HI chain. More specifically, this increase is strongest for threads with high popularity, located in the upper percentiles of the study set. For the score resulting from votes, similar observations were made in the case of answers.

Content typology The analysis of the content typology started by looking at the number of answers present in the thread was, which did, however, not show any relevant differences. In the next step, the likeliness of HI chains to occur was investigated among post tags. The results are of limited statistical relevance, but do indicate that proprietary software might negatively impact the appearance of HI chains. The outcome of the analysis of words present in questions and answers is globally coherent, and indicates that technical topics positively impact the appearance of HI chains. While the results analysing the words are more representative, the trends observed are less clear. Confirming these differences would thus require additional research. More conclusive results are, however, obtained for the impact of the presence of code snippets, images and external references. For both questions and answers, the presence of such elements is positively linked to the presence of HI chains and is strongest when all three elements are simultaneously present. For images, the effect is strongest, but for external references alone, the effect is, however, minimal.

User actions and types For the analysis of user actions and types, the relation between thread closures and the presence of HI chains was investigated first. Interestingly, in the presence of HI chains in a thread, the number of closures decreases and the number of reopenings increases. The share of such decisions being taken by moderators does, however, stay the same. Several analyses concerning the number of words of the HI chain were subsequently performed. Differences in the distribution of HI chains with or without the involvement of question/answer author(s) and moderators in them were investigated but none found. It is, however, worth noting that in 96% of the HI chains related to answers, the author of that answer was present with at least two comments. In 85.1% this is even the case for both the answer and the thread's question author. For the HI chains related to questions, the same metric for the question author is 96%. Beyond this, the presence of edits linked to the users involved (with at least two comments) in the HI chain was also investigated. Such an edit is present in 57% of the HI chains referring to answers and in 36% of the HI chains referring to questions. Plotting the distributions of word count was, however, not conclusive. Eventually, the time since registration and the dominant type of previous activity were computed to characterise the involved users themselves. It appears that the distribution of time since registration peaks around a few dozens of days and decreases, reaching values nearly as high as the existence of the forum. The dominant activity type is commenting in 73.3% of the cases and editing in 26.7% of the cases. Answering and asking activities appear to never exceed these two activities. Among them answering is, however, dominant with 72.6%.

Assumptions made in the process Beyond the assumption made in chapter 4 that the means used sufficiently take into account potential subcommunities; and that interaction with *exploratory talk* is a

way of exchanging mindsets linked to the solutions of technical questions, the research process of this chapter involved a few additional assumptions:

- By performing an analysis with a limited time frame at three different levels (thread, question/answer and HI chain), differences between the respective sets could not be avoided. The analysis of this chapter does, therefore, assume that the sets are sufficiently overlapping to provide insights valid across the levels.
- A second limitation is the stricter definition of HI chains concerning *exploratory talk*. Instances of *exploratory* interaction are likely to occur outside of HI chains too, but can not be reliably identified in an automated way (which is needed to process the amounts of data involved in this project). Therefore, this chapter assumes that HI chains are representative of a certain degree of interaction with *exploratory talk*.
- A third and final limitation is that the investigation conducted in this chapter did not look at the chronological order of events. It, therefore, established mainly correlations and care must be exerted for identifying causalities.

6

gis.SE, a reflective community?

Beyond the core activity of creating an open knowledge base, gis.SE also shows traits of the *reflective practitioner* (see part 2.2.3) at community level. Just as any other of the SE fora, there is a so-called 'meta' forum on which community rather than content-related issues are discussed [1]. These have the same format as the regular fora and can serve as a qualitative source to see which issues the community is struggling with and how the discussions evolve. This chapter aims at linking the analysis of the previous chapters to these discussions occurring on the meta forum. The main focus is the enforcement of forum rules which were introduced in part 1.1.1. This includes a review of criticism expressed by users on the behaviour of moderators. Where relevant, data insights from chapter 5 are used to complement the views. In the second step, the long term increase of unanswered questions observed in chapter 1 is covered. In this step too, explanations provided by members are discussed and, where possible, linked to data insights. Finally, the non-permanent view which the overarching SE system has on comments is discussed. Comparing it with the situation in the gis.SE forum, a review of it is provided. A the end of the chapter, an overview of the main findings of these steps is provided. Also, the challenges concerning the usage of data to tackle real-life community discussions are addressed.

6.1. Forum rules and criticism on moderation policy

A major input to analyse the current discussions of the forum is a post with the title 'Is GIS StackExchange ill?' which was created in February 2020 and is the most recent of the top ten highest-score questions. In it, a user who started contributing in 2013 shares the feeling that activity has decreased. Data provided points to an increase in unanswered questions. This is put in contrast with the forum website traffic (thus monthly number of views) which was claimed to be rather stable at the time of the post (about 57000 visits a day at the time of the thread creation[2]).

While the problem of the increasing number of questions will be addressed later in this part (see 6.1.4), the question raised addresses the decrease of activity from a general point of view. Some of the answers, criticize the behaviour of some moderators and will be discussed here.

In fact, no less than four of the seven answers mention the enforcement of forum rules, sometimes referring to it as *curation*. The comments contain also considerable discussions about this topic. A number of major aspects were identified: difficulties in accessing (official and unofficial) forum rules, criticism on the role taken by moderators and the fact that some frustrated users might stop contributing. For each of these aspects, an overview of the content, quoting original contributions is first shared and then discussed. Where relevant links with the findings of chapter 5 are established.

NB: unless state otherwise, all quotes are from the thread 'Is GIS.stackexchange ill?'[3]

6.1.1. The resources with which people get about the guidelines for questions:

"One of the problems is that the tour (*NB: what one sees when registering to the forum*) doesn't actually describe the standards of the most active moderator. It doesn't tell you that you must restrict the solution to a single piece of software, or that you must not ask questions about PostGIS without providing a block of SQL, whether or not it is relevant, or that you must include only one question mark symbol in your question; you may not phrase the same question multiple ways, etc, etc." — user dbaston

[1] https://gis.meta.stackexchange.com/

[2] NB: this number had however dropped to 46000 visits a day as of July 11 2020

[3] https://gis.meta.stackexchange.com/questions/5142/is-gis-stackexchange-ill

According to the user dbaston, the fact that the forum applies a number of rules which are not described inside the welcome tour [4] displayed when registering is problematic. In a subsequent comment, user PolyGeo (one of the moderators) mentions that there are two additional resources providing additional rules:

"The Tour is necessarily very concise but if you follow this link sequence from within it Visit the Help Center[5] - What topics can I ask about here?[6] - What makes a good question?[7] (or go via any other path to that Meta Q&A FAQ) you'll see the suggestions I offer for what I think makes clear and focused questions."

Given that newly registered users are only forced to look at the tour, this raises an interesting point. In fact, one might want to take the shortest path when having a question that was not asked earlier. However, in order to correctly ask a question, one should also have a look at the other sites which have more content than the tour. At the time of writing, only the help center was mentioned on the tour page itself. It is thus likely that a new user posting a question only discovers the rules beyond the tour once a moderator acts on the question asked, which can create frustrations.

This is especially important as, among subjective questions, there is a fine line between *constructive* and *non-constructive* questions. As the page 'What kind of questions should I avoid asking?'[8] states:

"If your motivation for asking the question is "I would like to participate in a discussion about ...", then you should not be asking here [...] Avoid asking subjective questions where:

- every answer is equally valid: "What's your favorite ...?"
- your answer is provided along with the question, and you expect more
- answers: "I use ... for ..., what do you use?"
- there is no actual problem to be solved: "I'm curious if other people feel like I do."
- you are asking an open-ended, hypothetical question: "What if ... happened?"
- your question is just a rant in disguise: "... sucks, am I right?"

[...] All subjective questions are expected to be constructive. What does that mean? Constructive subjective questions:

- inspire answers that explain "why" and "how"
- tend to have long, not short, answers
- have a constructive, fair, and impartial tone
- invite sharing experiences over opinions
- insist that opinion be backed up with facts and references
- are more than just mindless social fun"

While some guidelines are provided for the discrimination between questions that fit the forum and the one's that don't, they are are not applicable in a strict way. For instance, hypothetical questions might still be backed up by facts. For instance, one might ask what the disadvantages of writing a given coding assignment in a programming language not used anymore would be. Overall, it seems that the rules established here care mostly about the actual usefulness of the questions (e.g. excluding 'mindless social fun' and 'sharing experiences over opinions')[9]. The forum rules are therefore not necessarily opposed to discussions, as long as these discussions serve a defined direction and are not solely for the sake of interacting.

[4] https://gis.stackexchange.com/tour
[5] https://gis.stackexchange.com/help
[6] https://gis.stackexchange.com/help/on-topic
[7] https://gis.meta.stackexchange.com/questions/3349/asking-good-questions-for-gis-stack-exchange
[8] https://gis.stackexchange.com/help/dont-ask
[9] also see https://stackoverflow.blog/2010/09/29/good-subjective-bad-subjective/

6.1.2. The style with which feedback is given by moderators

“Finally, there’s the rise of what I’d call meta-moderation. Moderation has slowly gone from loosely enforcing some minimum standards in an effort to improve the community to a highly structured rule based enforcement that can be downright oppressive at times.” — by user ’Evil Genius’

“There is more than one moderator on GIS-SE who over-zealously penalises people who are new to GIS for asking questions that are obvious to most readers, but do not necessarily fit the format they wish the questions to be asked in. At least one of those same moderators expends a great amount of time investing in minor edits on every other post. When they do reply, the replies are often condescending, overly negative or disparaging, or do not make clear what they would like to be fixed about the post / how to do it.” — by user ’anakaine’

“It turns out that the way you curate is a serious discouragement to my willingness to participate here (and apparently others). I don’t expect you will change. But the question was asked, why has activity decreased, and this is my part of the answer to that question: I don’t participate much, because I don’t enjoy experiencing the heavy-handed actions of this particular moderator” — by user ’Steve Bennett’ (in a comment)

“This is how some of us are feeling. We feel invaded. You have more edits than the top 60 editors of the site COMBINED. You say you do this to bring quality to the site but here is the thing: GIS Stackexchange is not your backyard.” — by user Albert (in a comment)

“I think you can look at the pattern “question was asked, then closed, then edited, and reopened” as support for many things. In my case, step 3 of that pattern has often been a begrudging and irritated edit to conform to what I see as an arbitrary rule, adding little value in the process. But for a moderator it could look like “see! curation works! by closing the question, it got improved!” — by user Steve Bennett

When analysing the criticism expressed here, one should keep in mind that absolutely all contributors to the gis.SE forum, including moderators are volunteers (the only employees are working for the overarching SE organisation but do not participate in the topics content-wise). It is therefore rather interesting that the criticism volunteer moderators face goes up to reproaching them to take some degree of ownership of the forum, to abuse of their status. To nuance these critics, it should nevertheless be noted that positive feedback is also provided:

“I have written about 1000 answers now, many of them has been moderated and all the edits so far have made my answers more understandable. [...] I believe that most original posters benefit from the moderation because they have better chance for getting an answer. Opposite effect is rather unlikely IMHO.” —user30184 (in a comment)

Overall, it appears that moderator interventions and especially thread closures tend to be controversial. The privilege of moderators to close or reopen threads by a single decision contrasts with the ability of normal users. To do so without the intervention of a moderator, a critical number of users needs to vote for closure/reopening. On the one hand, this leaves moderators considerable discretion in how to apply the forum rules. On the other hand, this liberty is required to react promptly to filter out questions which do indeed not match the forum scope (e.g. duplicates) so that no energy of users is lost on these.

Interestingly, the analysis of the presence of HI chains performed in part 5.2.2 indicates that posts which create more discussions (as HI chains) are treated more gently in terms of both closures and reopening. Threads with a HI chain tend to get closed less often and reopened more often than threads without. This is rather surprising as one might expect HI chains to appear, at least for a part, around threads of which the question is unclear and would thus be more likely to be closed. Also, the share of moderators in these actions is identical to regular threads[10].

Assuming that the statement of the comments that edits by moderators are common (such as expressed by user Albert), it is worth noting that there are only very few cases (less than 5%) where such moderator edits lead to HI discussions (see part 5.2.4). This suggests that edits by moderators are usually applied with no or only limited discussion (either before or after). Eventually, this might create an impersonal feeling and even frustration, especially when the original author does not fully agree with the edit.

[10] A limitation which might still be explored is whether the HI chain appears before or after the closing and potentially reopening

In contrast, assuming that HI discussions chains are a positive process in terms of problem-solving, it appears that the presence of question/answer authors is more critical than the intervention of moderators (as mentioned in the analysis of figure 5.6 and 5.7, the author of the question/answer to which the HI chain refers is nearly always present). In 34% of questions and 57% of answers, such HI chains lead to the improvement of the question/answer by editing and thus modify the content as well, but in a more interactive and social way.

Nevertheless, this statement needs to be put in context as it appears that HI chains are not necessarily more popular/successful (see part 5.2.2) - or at least this is not reflected by the number of answers and the votes on them. Only in the case of threads achieving relatively high views and of votes regarding answers, a critically higher number than for regular threads can be observed. This suggests that not necessarily all threads with a HI chain should deserve higher attention either.

6.1.3. Some users with opinion differences get voiced out as they loose motivation in participating in the site

"NB: this is a comment that starts by citing another user: *I encourage people to continue voice their opinions, with reflection and thought, with enough personal flavour that it's real, and enough consideration-of-other that it transcends flame-war rhetoric.* it seems like there is a number of people who have been voicing these concerns over the years and nothing has changed in terms of "curation" or reflection among some of these readers. At what point do we reach a critical threshold?" —user GISKid

The last aspect tackled in the answers is the fact that the criticism on moderation policies leads to frustrated contributors dropping out which in turn reduces the diversity of opinions. This is an interesting statement as it shows the problems in analysing the impact of moderation. Beyond mere content actions, moderation policy also impacts the user that actively involved in the forum. Analysing this impact has some aspects of wicked problems (Rittel and Webber, 1973) as it is very hard (if not impossible) to find a test group acting in sufficiently similar circumstances. A possible approach might nevertheless be a 'research by design' approach (Zimmerman et al., 2007).

At a bigger scale, it should also be noted that SE underwent considerable evolution in the past years. As of 2011, the initiators founded a commercial company with activity in the recruitment sector, operating as a side branch to the fora. From 2016 on, a degradation of the relation between the management of SE and its users (including moderators) also occurred. This resulted mainly from bad decision making (not ensuring community support before acting) and miscommunication[11]. Such circumstances which even led to controversies on other platforms[12] doubtlessly impact the users involved in SE's fora.

6.1.4. Awareness about an increasing amount of unanswered questions

A second topic which is also addressed by the meta-question post "Is GIS stackexchange ill?" is the concern for the increasing amount of unanswered questions that were also identified in 4 and more specifically figure D.1. It is important to note that a similar issue was already addressed in January 2016[13]. In January 2018, another post addressing this issue too[14] managed to enter the top ten (vote-wise) of the meta forum. This shows that even before there was an increase in the monthly count (January 2019) of unanswered question, the gis.SE concerns were already raised. At the time, there were 20,000 unanswered questions. With the pace of about 300 unanswered questions (per month) observed in figure D.1 for the period January 2017-2019, this would be roughly equivalent to 65 months, thus more than five years (although the metrics since the creation of the forum would have to be taken into account for a precise number)[15]. Despite the consciousness expressed on the meta forum, the number of unanswered questions has thus been increasing again afterwards. Among the reasons mentioned for this by users, three can be found: that expanding a knowledge base changes with the

[11] for more information, see https://meta.stackexchange.com/questions/316934/revisiting-the-hot-network-questions-feature-what-are-our-shared-go and https://meta.stackexchange.com/questions/334551/an-apology-to-our-community-and-next-steps and https://cellio.dreamwidth.org/2019/10/05/stack-overflow-fiasco-timeline.html

[12] see https://www.theregister.com/2019/10/01/stack_exchange_controversy/ and https://www.reddit.com/r/programming/comments/de2has/stack_exchange_chose_persecution_over/

[13] https://gis.meta.stackexchange.com/questions/5142/is-gis-stackexchange-ill

[14] https://gis.meta.stackexchange.com/questions/4771/reaching-80-000-answered-questions-before-questions-on-our-site-reaches-100-000

[15] additionally one must note that a number of these questions have been answered in a challenge started as a result of the post and do thus not appear in the metrics used for figure D.1

progression, the change of the user profiles with the maturity of the site and answer being located in comments. Each of these explanations will be explained by quoting original contributions and analysed by linking it to the data of this research.

6.1.5. Expanding a knowledge base changes as it grows

“The longer the site has been around, the more likely it is that any given question has already been asked. Or to put it another way, the more likely it is that the question a user wants an answer to has already been asked and answered.” — user Son of a Beach[16]

A first explanation is that as the number of threads on the forum increases, the less likely it is that a new question will be a relevant addition. However, this implies that such questions would be duplicates. Another option is that the question does not match the scope of the gis.SE forum but would better fit another one (e.g. *https://earthscience.stackexchange.com/*). In both cases, such a question should be closed - a phenomenon which should thus lead to an increase in such closures. However, this is not the case, as shown in figure D.3.

Another question this statement poses is the evolution of the field with time. One might expect knowledge to be something dynamic: new scientific developments take place, software companies release new versions, etc. As one of the comments on the post states, the tag distribution among the unanswered questions should be checked to see whether the questions address newer technologies:

“For me it is a good sign that relatively more complex question are emerging. Maybe the stats should compare tags of “new” softwares (e.g. ArcGIS Pro, GEE) to see if the rate is different in there.” — user radouxju[17]

Although only indirectly, this remark suggests another interesting aspect: namely that the types of additions to the knowledge base change over time. In the beginning, the knowledge base still lacks some general, fundamental topics. Therefore, questions forming additions to it can be answered by a relatively big number of people. Later on, however, such questions have already been answered by the knowledge base and only more specialised ones can still form additions to the knowledge base. To answer these, it takes specific specialists (i.e. people who are able to understand and explain the matter) and more effort (i.e. investigating the topic). This evolution might be another explanation for an increase of unanswered questions.

6.1.6. Change of the user profiles as site reaches higher maturity

“*This*[18] seems to confirm something that I’ve suspected has been going on, which is that there are a whole lot more people asking questions than answering questions than there used to be.” — user blah238[19]

A second explanation is a change in maturity with the growth of the gis.SE forum. As user blah238 states, it was first the project of several passionate specialists with the willingness to share their knowledge and help each other openly. As the forum gained popularity, however, the dominating activity among user switched from answerer to user. This is still the case nowadays (as the results when running the query in the quoted comment show). Interestingly, the user profile of users associated with HI comment chains is the opposite: 72.6% tend to comment and answer more than they ask (as shown in table 5.18). This suggests that the persons involved in these chains are a different community than the ones asking questions.

6.1.7. Comments containing comment answers or partial answers

“Some users don’t understand the difference between an answer and a comment. The comment text box is the first thing you see after the question, so it’s a nice convenient place to simply start typing. Comments even look somewhat like a traditional forum [...] But more often than not, those users simply

[16] https://gis.meta.stackexchange.com/questions/5142/is-gis-stackexchange-ill
[17] https://gis.meta.stackexchange.com/questions/5142/is-gis-stackexchange-ill
[18] data query at following link:https://data.stackexchange.com/gis/query/76996/number-of-users-answering-or-questioning
[19] https://gis.meta.stackexchange.com/questions/4117/improving-on-19-000-unanswered-questions?cb=1

don't have the time, the complete story, or simply do not have the inclination to post a full answer [...]" — user Robert Cartaino [20]

A final explanation is that instead of being answered formally, sufficient hints are provided in the comments on the question. This would be indicated by an increasing comment activity on these questions, which is however partially refuted by the data shown in figure D.3. Not only the number of questions without answers but also the number of questions without both comments and answers increase. While this explanation might be true for a part of the unanswered questions, it can thus not explain the increasing trend.

6.2. The status of comments

While the previous point is unlikely to apply for the period 2017-2019 (as the number of questions without comments increases too), it tackles an important topic which is the role, the permanence of comments. The official position of the overarching SE system is that comments are not permanent, which does, however, differ from the opinion of some gis.SE users.

6.2.1. Position of SE vs. gis.SE users

"Comments are temporary "Post-It" notes left on a question or answer. You should not expect them to be around forever. Once a clarification has been made, an edit added to the post to include new information, or the issue in the comment is otherwise resolved, it can be deleted. Additionally, any comment that violates the comment guidelines listed above is subject to deletion." — SE community wiki[21]

This vision provided by the people who designed and contributed the overarching SE system stands in contrast with the statement of some users for the gis.SE forum:

"I don't usually bother deleting comments. When I do, it's usually for one of these reasons:

- The comment includes inaccurate or misleading information.[...]
- The tone of the comment was overly harsh, judgmental, or mean.[...]
- I misunderstood something in the original question.[...]

Basically, I delete comments when I think they might confuse/mislead people, hurt someone's feelings, or make me look bad. Otherwise, I see no reason to bother. I don't see any downside to having old comments still exist. I don't really see any upside either." — user csk [22]

Here, the results of the analysis of HI chains tend to confirm the statement of the gis.SE user. Assuming that HI chains result in the exchange of relevant information which was not yet included in the question/answer they refer to, the share of HI chains linked to an edit indicates us the extent to which discussed content is eventually transferred (added to the question/answer the discussion is about). This is the case for 57% of chains relating to answers and for 34% of chains relating to questions. Therefore, a considerable number of HI chains of which no content can be found back in the posts exists. Although some cases might cover unfruitful discussions (e.g. inconclusive debugging), these numbers illustrate the problem of such content being officially temporary. A reader who is interested in all details will, at some point, read the comments and is likely to find additional, relevant information in them.

6.3. Chapter overview and link to assumptions made

The issues addressed by the community on the meta forum show that the gis.SE community is reflective, addressing issues to which it is confronted in an explicit and structured way. By using the same format as all regular SE fora, the community can build a knowledge base about itself. This is not only useful to incite users to reflect on matters and express their opinion, but also valuable input for understanding the historical evolution of topics. The meta forum is similarly open as the main forum as only five reputation points (which only requires one or two answers/questions) are needed to participate.

[20] https://gis.meta.stackexchange.com/questions/580/answering-questions-with-a-comment
[21] https://meta.stackexchange.com/questions/19756/how-do-comments-work
[22] https://gis.meta.stackexchange.com/questions/5025/what-are-the-downsides-of-not-deleting-our-own-comments

By identifying critical discussion topics on the meta forum and linking content to the data analysis, user insights are taken into account during the interpretation of the results. This was done along with three topics of which two result from the top 10 meta-threads with highest scores and one specifically related to comments, which are central to this research.

Role of moderators The first topic is the role of moderators who are enforcing the forum rules and appear to be strongly criticized about the way they do this. A first issue is that the forum rules are not necessarily clear enough, and hard to apply in some cases such as subjective questions.

The criticism on moderators does, however, go further than that. Statements go as far as reproaching them to take ownership, which is linked to their frequent edits. The analysis provides a clue here as the share of HI chains related to edits performed by moderators is only minimal. This indicates that the editing process by moderators is leading to extended discussion in exceptional cases only. This is in contrast to the presence of question and answers authors which are much more critical to the occurrence of HI chains. This is also true for the content of the questions and answers themselves as a substantial part of HI does indeed produce edits (about a third of the ones relating to questions and more than half of the ones relating to answers).

The fact that threads with HI chains are less likely to be closed and more likely to be reopened (if closed in first instance) is rather encouraging with regard to the moderation. Building on the assumption that HI chains indicate exploratory talk, this would mean that moderators do not necessarily see questions leading to constructive comment discussions negatively or as out of scope.

A potential source of frustration which was discussed in this part is the presence of forum rules and how to access them. While there are some guidelines which users are actively invited to use, a contribution fulfilling the requirements needs to fulfil guidelines explained on other pages too.

Increasing number of unanswered questions The phenomena which was observed in chapter 2 appears to be well known to the community. While a major increase was observed from January 2019 on, the problem appears to have been discussed in 2016 and 2018 already. Among the possible explanations provided, only one appears to be in line with data. This one is the change of user profiles dominating on the forum. As the popularity of the forum increased, the share of askers has gradually increased too, reducing, in turn, the share of answerers. The data to prove so is provided by the user blah238 advancing data that still holds as of March 2020. This is interestingly contrasting with the users involved in HI chains who post more answers than they ask questions. This suggests that the community involved in HI is a community who might be stimulated in order to increase the answerer share of the overall forum.

Another explanation which can only be partially valid is that after having built a knowledge base, expanding it continues at a slower pace. This would mean that a substantial share of unanswered questions would be duplicated and, normally, closed. However, no increasing trend in the closure actions has been observed as the number of unanswered questions increased. Another explanation, namely that answers are 'hidden' in comments does not fully stand either as the number of the question without answers or comments did also increase.

The temporary nature of comments A final topic which was addressed concerns a central element of this research, namely the comments. Although the data analysis indicates that HI chains and thus comments are definitively a relevant phenomena, it appears that the official SE policy still defines comments as temporary 'post-its'. This policy is however questioned by a user who explicitly states that there is no need to delete all comments that have been processed (i.e. lead to a conclusive decision) - unless they are confusing or harm someone. The fact that the data analysis shows that only some HI chains lead to an edit of the question/answer they refer to is an additional argument in that direction.

The wicked problematic of the gis.SE community A major limitation of this analysis which is also tackled by user opinions is that investigating a forum such as gis.SE is to some extent a wicked problem. In fact, there are users who would probably be interested in contributing to the community but renounce from doing so (or choose to resign after a given time). Such phenomena are very hard to distinguish from the 'regular' evolution of a forum as a control group subject to sufficiently similar circumstances is rather hard to find.

Assumptions made Among the assumptions of the research, only a few were added in this chapter. The major one is that the community is indeed able to reflect on itself in a fact-based manner. Controversial discussions such as the one about the moderation policy would indeed be of limited meaning if they were only intended to harm people. However, users regularly provide numeric analysis and share them for discussion. Criticism about moderation is for instance strengthened by the fact that the number one editor (who is a moderator) accumulates more edits than the sixty next together. Overall, this chapter assumes that such kind of statements were already fact-checked by the community and are thus valid (unless refuted in their discussion).

7
Conclusion

7.0.1. Answers to the research questions

1. What are the characteristics of the gis.SE forum/community? The gis.SE and the community belonging to it appear to be built around a tool, a technology, namely the Geographic Information Systems. On the one hand, the community qualifies as a *community of practice* (Barley, 1996) with a content-oriented and professional behaviour. From the data side, this is confirmed by higher activity in the period Monday-Friday which is equivalent to the working week in many countries. On the other hand, the community is also a very open one, qualifying as an *open community contribution system* within the framework of Stanoevska-Slabeva (2002).

Content-wise, a qualitative analysis of comment lists was performed to determine the types of *talk* present within the framework of Ferguson et al. (2013). In order to cover a potential heterogeneity of the community, representative sampling was applied in two ways. On the one hand, the community's own thematic organizing tool, namely the tags, was used to maximize thematic diversity. On the other hand, the distribution of the comment list length was extracted and applied during the sampling as well. The results show that a strong majority of lists qualify as *exploratory talk*. A few cases (<10%) of *cumulative talk* exist, and the share of *negative talk* (i.e. offensive comments) is very low (<1%). This further confirms the content-oriented nature of the forum and suggests that due care is taken to filter out comments not fitting in the scope (e.g. spam, clashes, etc.). For further characterisation, a monthly analysis of the forum activity during the period 2017-2019 was conducted. The main findings are:

- The forum is globally stable, as shown by the number of new user subscriptions, answers posted, comments and edits.
- Some trends do nevertheless exist, the most important ones being an increase of questions without answers (and without either answers or comments) from January 2019. In the period January 2017-August 2018, a decrease in the number of closure and deletion votes was furthermore observed.
- The reaction time to comments was analysed using two methods that address different subsets. In both cases, the 10th percentile is below 10 minutes and the 90th percentile below 5 days. For one of the methods, an increase in reaction the 10th percentile was observed from January 2019 on. The comment reaction times are globally lower than the reaction times observed between questions and their first answers (10the percentile: 15 minutes and 90th percentile 4-15 days). Globally, this shows that gis.SE allows not only easy but also relatively rapid access to support by other specialists.

2.a. How can threads in which constructive interaction (1) occurs and (2) discloses the 'mindset(s)' linked to the answers be identified in an automated way? In order to identify situations in which the 'mindset(s)' linked to answering a question is shared, the presence of interaction with *exploratory talk* was taken as a central indicator. While the forum characterisation has shown that *exploratory talk* is present in a majority of cases, the presence of interaction (thus at least two users discussing a matter) is an important requirement. As the SE format stores such discussions in an unthreaded way, namely as a chronological list of raw comments, an automated approach to identify interaction chains was developed. This automated approach follows one rule and two exceptions:

- **Rule:** Any comment separated by more than 72h from the previous comment does not belong to the same interaction 'chain'.

- **Exception one:** Any comment containing the symbol '@' always belongs to the same interaction chain as the previous comment.
- **Exception two:** Any comment written by a user who has already contributed to the comment list before belongs to the same chain as the previous comment.

The validation of this approach does, however, show that single comments are wrongly added to or removed from chains. Therefore, a more robust definition with regard to these errors was chosen for the analysis. So-called high-intensity (HI) chains (which consist of at least two users contributing each with at least two messages) were therefore used as an indicator of *exploratory talk* in the analysis. By adding these criteria, the definition is narrowed down and the identified errors are addressed but the sample size relevant for the analysis is reduced. In total, 2951 chains relating directly to questions and 5004 chains relating to answers were identified, which represents about 5-10% of the set, depending on the unit of analysis used.

2.b. Which factors influence the occurrence of such threads? An analysis of factors potentially related to the appearance of such HI chains was conducted using different levels of analysis. The most important conclusive outcomes are:

- Threads containing a HI chain tend to achieve more absolute and daily views, especially for the upper percentiles, thus the threads with relatively high popularity. This indicates a link with the popularity of threads.
- The analysis of the voting score of answers containing a HI chain shows similar links, again for the high scoring cases. This indicates a link with the appreciation of such answers by the forum.
- Tag (threads) and content-wise (words of questions/answers), the differences between the presence and absence of HI chains are statistically limited, indicating trends with limited reliability.
- A conclusive observation is, however, that the presence of images, code snippets and to some extent external references (after edits) is higher among questions which have a HI chain in their comments. These positive links reinforce each other when combined.
- Closures and reopenings are negatively linked to the presence of HI chains in a thread. In their presence, the share of closures decreases and the one of reopenings increases. This indicates that moderators do not see threads in which HI chains occur more negatively.
- The author of a question is present in 96% of the HI discussions chains. In the case of a chains related to an answer, the author is present in 96% of the cases and in 85%, the author of the thread question is present too. This illustrates that authors have a major role in the occurrence of HI discussions.
- Only 36% of HI chains referring to questions and 57% of chains referring to answers are related to an edit performed on the respective question or answer (by a user involved in the HI chain). While some HI conversations are thus related to content changes, this is far from being the case for all.
- In opposition to the general forum trend, the contributors to HI chains tend to answer more questions than they ask. Nevertheless, their comment and edit activities are even higher, indicating that their contribution to HI chains are linked to high activities among the forum comments and content changes.

3.a.To which extent does the community appreciate threads sharing such 'mindset(s)'? The results of the data analysis indicate that threads in which HI chains are shared reach a higher number of daily and total views in some situations. While the values might be similar to other threads in some cases, they clearly increase when measured at the 90th percentile, indicating that the presence of HI chains is linked to an increase of views among popular threads (in terms of views).

For the number of votes, a similar trend can also be identified but only for the score of answers. In the case of relatively high scoring answers, the presence of HI chains in answers is thus positively

linked to higher scores of the answer. In the case of questions, no positive link exists, but no negative one either.

These two indicators indicate that the community (including non registered) users appreciate threads, questions and answers just as much as in cases where a HI chain is present in the comments as in the cases where it is not. In some situations, the appreciation is actually higher for the cases with a HI chain than without.

3.b. Are these threads in conflict with the enforcement of forum rules? While the official forum rules are rather clear on some points, it appears this is not the case for subjective questions (despite the fact that they try). Such questions might lead to more HI chains in the comments as they tend to be subject to more discussions. In the case of such a link, forum policy measures, namely thread closures would be expected to be higher among threads with a HI chain (and thread reopenings lower). Data does, however, indicate the opposite as the number of closures is actually lower and the number of reopenings higher among threads with a HI chain. This indicates that in the eyes of moderators (who are taking a majority of closing/reopening decisions), the presence of HI is not particularly conflicting with the forum rules. It should, however, be noted that this finding is only partial. In fact, identifying threads which would *potentially* have a high number of HI chains, but do not due to moderation is a challenge not addressed in the research.

Overarching research question: In which cases do threads on gis.stackexchange share the 'mindset(s)' necessary to solve the question by facilitating constructive interaction? Overall, this research has developed an automated approach to identify a limited set of situations in which the 'mindset(s)' which accompanies a technical question resolution is shared. In opposition to other threads, these are instances in which an interaction linked to either the question itself or to a proposed answer took place. The minimum criteria for these interactions were set to two different users posting each two comments (not taking into account subsequent comments of the same user).

Although this definition is rather strict, validations performed have confirmed that they represent cases of *exploratory* interaction. It is, however, not excluded that interactions in which a mindset is shared also take place outside this definition. For the analysis, it is therefore assumed that the identified cases are representative of the instance in which 'mindset(s)' necessary to solve questions are being shared.

Among the circumstances which are beneficial for the occurence of the sharing of such 'mindset(s)', a major one is the involvement of the author(s). In a large majority of identified cases, the author of the thread and, in case the comments refer to an answer, also of this answer is involved. Another trend among the involved actors is that the users are mostly commenting and editing questions/answers. They tend to answer more questions than they ask.

Content-wise, the major circumstance identified is that the sharing of such mindset(s) mostly occurs when the post (question or answer) to which the comments refer includes particular content forms. In the case of both questions and answer, these are code snippets and images. One should note that these circumstances do not necessarily have to appear before the interaction in which 'mindset(s)' to answer the question are shared.

8
Discussions

8.1. Assumptions made

During the different steps of this research, a number of assumptions were made in order to develop a model matching automated data analysis. In fact, humans usually outperform algorithms when it comes to interpreting the meaning of natural language (at least if the interpretation of human codes is taken as a reference, which is the case in communication sciences). However, they have the disadvantage of being rather costly, which limits the scope of projects. In order to analyse a high data quantity, as done in this research, automated approaches must, therefore, be developed. This does however not go without difficulties, as qualitative concepts need to be translated into more objective forms. To perform this translation, a number of assumptions had to be made, of which an overview is provided here:

- A first assumption, highlighted in the introduction, is that interaction with *exploratory talk* is a reliable indicator of the occurrence of 'mindset(s)' involved in solving a question.
- The qualitative analysis performed in chapter 4 indicates that *exploratory talk* is dominating on the forum (about 90% of the cases). The research assumes that the sampling performed is representative (that the method sufficiently accounts for potential heterogeneity) and the high share of *exploratory talk* is sufficient to assume that when interaction occurs on the forum, it is of *exploratory* nature.
- Furthermore, a number of comments posted on the forum are deleted. This decision is either taken by the community or by the author. The analyses performed in this research assume these deletions (for the ones done by authors themselves, about 10%) not to be problematic.
- While humans have relative ease in reconstructing conversations from an unthreaded list of comments, this is not the case for machines. Based on the second validation performed in chapter 5, it is assumed that the algorithm developed to perform this task automatically is sufficiently reliable. In this assumption, the exact definition developed for HI chains (which is more robust to errors on single comments) was taken into account.
- As the definition is relatively strict, it leads to limited sample size. Depending on the unit of analysis used, its size is of about 5-10%. It is assumed that such a size is sufficient to perform comparisons with the bigger set.
- The timeframe of the period 2017-2019 was used for all analysis and thus at three different layers/units of analysis. As layers, questions/answers and chain are linked in different ways, this does unavoidably lead to different data sets. It is assumed that the overlap of the respective data sets and their overall homogeneity is sufficient to allow for common analysis.
- A final limitation is that within the analysis, the chronological order of events (e.g. edits) was not taken into account. The outcomes can thus indicate correlations, but not necessarily causality.

8.2. Recommendations and their specificity

8.2.1. gis.SE forum

A first series of recommendations is specific to the gis.SE forum. In order to include interaction in the forum design, a number of aspects identified by this study should be considered:

- **Clarify forum rules:** as the forum rules seem to be a source of frustration, a new approach to explain these ones should be considered. The forum tour page as it is now appears to be concise but sometimes insufficient. To keep an appealing layout while communicating a clearer message, more attractive formats such as videos or websites with moving elements should be considered.
- **Community surveys/reports:** this research shows that there is potential for the investigation of communities using both quantitative and qualitative methods. While the discussions on the gis.SE meta forum do to some extent resemble such investigation, they are not conducted in a systematic way. Most discussions and thus reflections on the forum are started by users voicing their opinion on a given matter. While keeping in mind that volunteers mean limited means, a more systematic approach should be considered. This one might consist of an annual report produced to stimulate discussion (building on methods of *Social Network Analysis* and *Social Learning Analytics* mentioned in chapter 2), which might include a survey (an idea that was mentioned in the past but not realised[1]).
- **Use KPIs to facilitate structured decisions:** Next to a more systematic approach, the usage of KPIs within the discussions is encouraged. Some examples of data evidence used in discussions on the gis.SE forum were identified in chapter 5. In order to integrate this reflective behaviour with the previous point, a number of KPIs should be selected and monitored to stimulate regular discussions.

A second series of recommendations applies to the SE system from a relatively general point of view. To some extent, these points apply to other Q&A fora too.

- **Add the possibility to give comments the status 'outdated' rather than deleting them:** as discussed in part 6.2, the SE system explicitly allows authors to delete their own comments. The main use for such action seems to be when the context changes (e.g. the question was edited) and the comments thus appear inappropriate or confusing. The deletion might, however, also lead to the removal of valuable information. Even if the information might be confusing at first hand, it might still be of value for users digging deeply into the subject (and thus also into the edits of the question/answer the comment refers too). Therefore, an alternative should be provided - allowing users to flag a comment as potentially confusing, as 'outdated'. The term users here should include more persons than the comment author (e.g. all users involved in the thread so far) as the correct flagging otherwise depends on the long term implication of the comment author. Graphically, such comments could be shown only upon a click by the user (similarly to long comment lists which need to be expanded by the user before being fully displayed). Additionally, such comments could be shown with some transparency.
- **Distribute badge rewards for combined interaction and editing:** As mentioned in the conclusions, interactions do not always lead to an edit of the question/answer they refer to. While there might be situations in which no edit might make sense (e.g. the interaction was mainly the discussion of an idea but was eventually refused), there might also be situations in which edits are required but no such effort was taken (see part 6.1.4). The behaviour of producing edits after interactions could therefore be stimulated by the creation of a specific badge (e.g. 'discussion master' badge) within the already well developed SE gamification system.
- **Draw attention on situations with interaction resulting in no edits:** An additional measure might be to stimulate other users rather than the interaction participants to produce the edits based on interaction. In order to do so, the SE would need to draw users attention on discussions without edits, for instance by altering the selection for the 'top questions' page[2].

Finally, a number of recommendations for use cases beyond Q&A websites can be formulated. For open academic peer-review processes for instance, these are:

- **Facilitate *exploratory talk*:** as the goal is to improve draft papers before their publication, discussions with a fruitful outcome should be encouraged. Implicitly, this also requires to formulate and enforce rules which serve as a basis to discourage other types of talk.

[1] https://gis.meta.stackexchange.com/questions/5082/gis-survey-similar-to-stack-overflows-developer-survey
[2] https://gis.stackexchange.com/?tab=hot

- **Combine both expanding and condensing processes:** on the one hand, reaching fruitful outcomes requires room for interaction, for discussions expanding the matter. On the other hand, gis.SE shows that condensing processes (e.g. edits to the matter being discussed) are essential too. Therefore, both parts should be facilitated.
- **Implicate the authors:** while the situation is slightly different to the one of Q&A fora (quality insurance rather than solving a challenge people face), the implication of the original authors of the paper is likely to be important. The case of gis.SE shows that a high majority of interactions occur when the original author is involved. This would be a substantial addition to more traditional peer review processes in which authors rarely get an opportunity to answer reviewers or ask for clarification (except when several review rounds take place with the same reviewers).

It should, however, be noted that substantial differences are to be expected outside the Q&A cases. For the peer-review process, the main function is to enforce quality standards before publication. Making such process open is challenging as it would require to fit processes similar to the Q&A one in a limited time (although authors that share preprints on online platforsm are an example). Another option, which would change the academic publication paradigm, is to lift the limits of the review and encourage the improvement of a publication over a longer period of time. It should, however, be kept in mind that keeping the authors implied also become harder over a longer period of time (but, as already discussed, is positively impacting interaction).

On the other hand, the question of openness must also be raised. Even if throwing the peer-review process fully open, there will always be a bias to the participants in such process (similarly to the users involved gis.SE). An important question is thus how to overcome this, how to convince a maximum number of contributors to participate in the process. Additionally, the issue of how to formulate and enforce a peer review scope and rules is a point which is likely to require attention. In fact, who would be supervising such an open peer review process and what would be the meaning thereof? Depending on the implicated community, mindsets might have a limited compatibility with the processes of open review. As Tennant et al. (2017) formulates it, open peer review would imply decoupling editorial selectivity from the peer review process: it assumes that all research deserves to be published. The role of editors might thus move from establishing strategies and organising the quality insurance for papers fitting in it to formulating thematic guides to navigate through a bigger amount of published guides.

8.3. Reflection on the usage of algorithms for social research

Overall, this research has also shown that the usage of algorithms in social research on communication can be challenging. The often ambiguous and unthreaded nature of the raw data (the comment lists) is, even more, a challenge to machines than it is already to human coders. In order to translate concepts in a way that machines can use them, a number of assumptions are necessary. To avoid the propagation of limitations or even errors, thorough quality insurance is needed. Each choice needs to be correctly reflected and regular validations must be performed to confirm them. To some extent, this is also a topic in human coding where steps such as iterative coding approaches and inter-coder agreement are not uncommon. The same does thus apply, if not to an even stronger degree, for the usage of automated approaches in communication sciences.

This distinguishes communication sciences from other, more mathematical sciences. While they are also confronted with the problematic of input sampling, instrumentation with rather well-known specification and reliability is often used. In social sciences, algorithms transforming the raw input into numbers (such as the algorithms used in this research) might be analogous to such 'instrumentation'. Their specificity to the context they work in is, however, very high. At each step, at each decision taken in their development, errors are unavoidable and their meaning in the context must be thoroughly analysed. Moreover, this also means that the approach developed in this research can not be applied as-is to other fora. In fact, the outcome of the design choices made and the validations performed might be very different in other contexts.

8.4. Link to theoretical framework

Looking back at the theoretical framework formulated in chapter 2 and more specifically in section 2.4, a number of observations are worth mentioning:

A first one is related to the scope of the study which is limited to the outputs, to the content that can be found on the gis.SE site. It is challenging to draw conclusions upon this as it has the consequence of making the outputs very specific. While interaction as such can be defined and the nature of it defined basing on the output, it is much harder to specifically analyse the impact it has on (social) learning. This is true for both active participants and (more) passive website users. In this study, a specific type of interaction was taken as an indicator for situations in which the 'mindset(s) to solve a question' is/are being shared. However, this does not address what is going on between the moment the message is read and memorised by the user. Intrinsically, this also addresses informal learning as Q&A sites are usually not incorporated in formal curricula. In order to more precisely understand the learning process involved, it would be necessary to address *how* users are using the website content. Such research would necessarily require interaction with the users on top of the analysis performed here.

This challenge of interpretation and understanding is even more critical when relating it to the definition of interactions in CMC with unthreaded data. In fact, the approach taken this study assumes that one human coder is sufficient to correctly split an unstructured list into interaction chains. While some ambiguities might be filtered out by the definition which tolerates an error margin of the automated approach too, it is unclear whether a consensus would be possible at all in some cases. To push it to the extreme, CMC could even lead to situations in which one of the discussion participants believes being involved in an interaction chain, while the other one does not share this observation (e.g. the person was answering another message than the other believes). In practice it is thus challenging to define interaction chains without the direct involvement (i.e. interviewing) of the participants in the study.

As already mentioned earlier, gis.SE is a technology-oriented platform on which rather formal communication takes place. This is in line with several quantitative indicators (e.g. activity per weekday) and with the intentions of the moderators as described in the meta forum. The guideline of avoiding opiniated discussions can be seen as a will to avoid dynamics only expanding knowledge. Instead, condensing dynamics are encouraged. As discussed in the theoretical framework, expanding dynamics tend to be associated with fora while condensing dynamics tend to be associated with wikis. As gis.SE positions itself in between these two forms, it is not unsurprising that the balance between these two dynamics is critical. This is in line with the theoretical framework's statement that shared understanding and thus interaction is key to co-creation. Interestingly, while some users criticize the moderators for putting this balance at danger, the numbers do not suggest moderators being interaction-averse. The numbers do however not cover the potential interactions that did not occur. Hereby, the study shows that the definition and monitoring of this balance is very challenging in practice.

Finally, it is worth further specifying gis.SE as a human-machine hybrid (or duo) here. This nature can be found on multiple levels, starting with the content, the technology which is central to the forum: the one of GIS. The participant of the forum are already human-machine hybrids in the activity that links them to each other, that makes them form a *community of practice*. At a second level, they use another technology to create and maintain a knowledge base - the SE system. By doing so, a second human-machine hybrid is present, not at the level of the expertise but at the level of knowledge exchange. Finally, a third level is the one of reflective practice within which community members reflect on the knowledge exchange. Interestingly, the technology employed here is identical to the second level. The content does however differ as much less moderation applies to the third than to the second level: opinionated discussions are no exception there. This contrast highlights that not only the technology but also the humans play an important role in such system. On one hand, communication is computer-mediated, but on the other hand it is also human-mediated - and not only because users participate in the communication but also because they participate in the system (i.e. by moderation, voting, etc.). In sum, gis.SE is a complex phenomenon of human-machine integration.

8.5. Axes for future research

This last part of the conclusion looks at potential improvements and extensions to the research done. A part of them addressed the research and the choices made while another part rather addresses the context.

The meaning of 'mindet(s)' The second research question was formulated with regard to the sharing of the 'mindset(s)' required to solve a question asked on the forum. While the presence of *exploratory* interaction was deemed sufficient within this research, a more precise definition might be useful too.

One might note that the coding approach developed by Ferguson et al. (2013) used to identify *exploratory* interaction itself already contains four categories. A more specific approach with regard to which degree *exploratory* interaction facilitates the exchange of such 'mindet(s)' might thus be possible.

The wicked nature of the analysis of forum users A second topic which was already addressed in part 6.1.3 is the wicked trait of the problematic around forum users. The circumstances in the forum tend to favour users who agree with these ones but it is very hard to measure such an effect. Chapter 4 has, however, shown that one the impact of the Covid-19 pandemic is that the moderator share among closure/deletion actions and edits has suddenly dropped. As these posts have a limited age (which is one of the reasons why the timeframe was limited to 2017-2019), they were not covered in the analysis. These exceptional circumstances are nevertheless of interest and future research might investigate if this drop of moderation had any impact on the user activity.

Other channels of CMC Another aspect out of the scope of this research but nevertheless of interest are the chatrooms associated to each SE forum. In opposition to the forum, the chat is a real-time discussion platform with less formal guidelines. Its intention is to facilitate more interactive discussions around the topics addressed in the forum[3]. At the time of writing, the gis.SE chatroom had only four chatrooms[4] with the most recent activity two days before checking. Despite the fact that chat room information is not as publicly highlighted as comments, it might be worth investigating the role of this tool within knowledge exchange.

[3] https://chat.stackexchange.com/faq
[4] https://chat.stackexchange.com/?tab=site&host=gis.stackexchange.com

Appendices

A

50 most frequent tags, used for the sampling of the validations

arcobjects	geometry	pyqgis
javascript	arcgis-10.0	arcgis-desktop
shapefile	arcpy	arcgis-10.2
python	geoserver	geotiff-tiff
postgis	geojson	openlayers
gdal	symbology	arcgis-10.3
raster	modelbuilder	arcgis-pro
arcmap	postgresql	openlayers-2
sql	coordinate-system	google-earth-engine
wms	r	carto
openstreetmap	arcgis-javascript-api	
labeling	arcgis-10.1	
dem	field-calculator	
spatial-analyst	arcgis-online	
polygon	leaflet	
qgis	attribute-table	
grass	line	
point	layers	
arcgis-server	qgis-plugins	
vector	qgis-3	

Table A.1: 50 most frequent tags, used for the sampling of the validations.

B

Validation results for the identification of type of talk present in comment lists

B.1. Questions

question id	challenge	evaluation	extension	reasoning	comments
284585	x		x	x	
311768	x			x	
259210	x		x	x	
283823	x			x	
292730	x		x	x	on question formulation
239445				x	not conclusive
303701				x	not conclusive
325702				x	not conclusive
305358		x		x	
279550	x			x	
345823				x	
244804					cumulative
284450				x	
266350				x	
285144		x		x	
101674				x	
155096				x	
252801				x	
249668					cumulative
239305	x			x	
239417				x	
247989	x		x	x	
308161	x			x	
339066		x		x	
313012				x	not conclusive
332499				x	
240686				x	
315715					cumulative
262538	x				
334385				x	
228671				x	
302114		x			
256469					cumulative
136143				x	
252094				x	
333274	x				
294971	x				
284183	x			x	
332236	x			x	
318647				x	minimal
229550				x	minimal
257868				x	minimal
337731				x	
224918				x	minimal
306528				x	
247242		x		x	
297943				x	
343840		x		x	
295785				x	minimal
276290				x	minimal

cumulative total 4

B.2. Answers

answer id	challenge	evaluation	extension	reasoning	comments
334040		x	x	x	
219170	x			x	
297321		x		x	
233028	x	x		x	
284410	x	x	x		no improvements, mainly confirmation and how to spread it
27458	x		x	x	
241517	x	x			
89344	x	x		x	
40548					cumulative
276697	x	x		x	
91685					cumulative
26355	x	x	x		
179247	x	x		x	
255420	x	x			
336708		x	x	x	
152866	x		x		
52952			x	x	
310887	x			x	
27104	x	x			
309163	x	x			
281790				x	
341292		x			minimal
242686	x				
37584			x		
133563				x	minimal
272030	x		x	x	
264225	x		x		
131789		x			full discussion probably deleted
52888	x	x	x		
252			x		partially exploratory
146396	x		x		
282165	x		x		
84992		x	x		
279713	x			x	
298077	x			x	
318886	x			x	
249725				x	
226906				x	
250017				x	
312807				x	minimal
310452	x	x			
228761	x		x		
244485	x			x	
240235				x	minimal
339700	x			x	
268502	x			x	
337598	x		x		
91963				x	minimal
329270	x			x	
290304		x			

cumulative total	2
no discussion total	5

C

Validation results for the comment list splitting algorithm

C.1. Questions

question_id	count	validation	comments	chat	closed	offensive/defensive	wrongly added	missed	#affected chains
346059	0	0	lof of explaining forum rules	no	yes				
282089	0	0	quest for solution	no	yes				
304516	0	0	doubts on fitness of q for gis.SE (	no	yes				
254232	0	0	investigation to find error in data	no	no				
313975	0	0	users trying to help	no	no				
230605	0	0	question on computing speed	no	yes				
233785	0	0	guessing of a CRS	no	no				
273946	0	0	gdal installation pb	no	no				
253475	0	0	CRS file setting problems	no	no				
270327	0	0	geometrical problem discussions	no	no				
277603	0	0	coding speed discussion	no	no				
284652	0	0	two people discussion	no	no				
291751	0	0	discussion on finding CRS	no	no				
304139	0	0	CRS importing chit chat	no	no	1			
314664	0	0	q understanding and CRS discuss	no	no				
332587	0	1	displaying data problem	no	no		1		1
223498	0	0	database bug	no	no				
224668	0	0	two people discussion	no	no				
234630	0	0	need for different dbs/schemas	no	no				
242337	0	0	displaying layers	no	no				
258651	0	0	pixels and polygon intersection	no	no				
292368	0	0	cumulative, on valide polygons	no	no				
293799	0	0	software speed	no	no				
306342	0	0	software bug slope computing	no	no				
317620	0	0	unclear question on moving poin	no	no				
6037	2	8	extreme cumulative	no	no		6		**1**
32119	3	2	discussion on software	no	no		2		2
36795	2	5	extreme cumulative	no	no		4		**2**
310347	2	1	library versions saga	no	no		1		2
137025	2	2	admin rights	no	no				
137025	-		duplicate						
144077	2	2	uploading data	no	no				
165950	2	2	gdal data export	no	no				
18814	2	3	tool for arcpy	no	no		1		1
57338	3	3	generating points on line	no	no				
64383	2	1	arcgis wfs compliance	no	no			1	2
94469	2	0	clip raster speed	no	no			3	3
126380	3	3	qgis tools	no	no				
182670	3	3	software in R, limit cumulative	no	no				
182670	3	-	duplicate						
211000	2	2	bug of qgis	no	no				
300860	2	0	bug in qgis	no	no			3	**2**
57651	2	1	settings in esri	no	no			1	2
216871	2	2	bug qgis psql	no	no				
256849	2	2	cumulative, json	no	no				
93159	2	2	cumulative, json	no	no				
172504	2	2	debug discussion	no	no				
227384	3	3	ge engine	no	no				
263629	2	1	psql esri database	no	no			1	

total with wrong result	10
cases with >1 wrong messa	**3**

C.2. Answers

answer_id	count	validation	uncertain	comments	chat	closed	offensive/	wrongly added	missed	affected c
230749	0	0	0	gdal debug conversation	no	no				
102122	0	0	0	debugging conversation p	no	no				
7650	0	0	0		no	no				
280958	0	0	0		yes	no				
318205	0	1	0		yes	no		1		1
199685	0	0	0	2 person debug conv	no	no				
312501	0	0	0		yes	no				
245492	0	0	0	3 person discussion	no	no				
86185	0	0	0	2 person debug conv	no	no				
250432	0	0	0	2 person debug conv	no	no				
244296	0	1	0	mostly 2 person conv	no	no		2		**1**
306981	0	0	0	2 person debug conv	no	no				
5787	0	0	0	2 person debug conv	no	no				
34387	0	0	0	file format conv	no	no	1			
52231	0	0	1	2 person debug conv	yes	no				
267985	0	0	0	2 person debug conv	no	no				
272392	0	0	0	2 person debug conv	no	no				
284414	0	0	0	2 person debug conv, abo	no	no				
288655	0	0	0	2 person debug conv	no	no				
176225	0	0	0	2 person debug conv	no	no				
220274	0	0	0	2 person debug conv	yes	no				
252544	0	0	0	2 person debug conv	yes	no				
267486	0	0	0	2 person debug conv	no	no				
307271	0	0	0	2 person debug conv	no	no				
322956	0	0	0	2 person debug conv	no	no				
21661	2	2	0	debug conv + 2 comments	no	no				
33070	2	2	0		no	no				
126550	5	5	0	debug conv	no	no				
209131	3	5	1	development of plugin	no	no		2		**1**
274600	2	3	1	python environment	no	no		1		1
109969	4	3	1		no	no		0	3	**1**
5642	2	2	0		no	no				
22492	4	4	2	cumulative	no	no				
54792	3	3	0	cumulative cos + conversa	no	no				
92227	2	3	0	several convs	no	no			1	1
191446	3	4	0	several convs					1	1
31415	5	4	2					2		2
81176	2	2	0							
304598	2	2	1							
1105	3	5	0						2	2
104528	3	3	0							
83027	2	2	0							
104380	2	1	0					1		1
155448	2	1	0					1		1
273718	2	1	0					1		1
			9							

cases with wrong result	12
cases with >1 wrong message per chain	**3**

D

Additional graphs

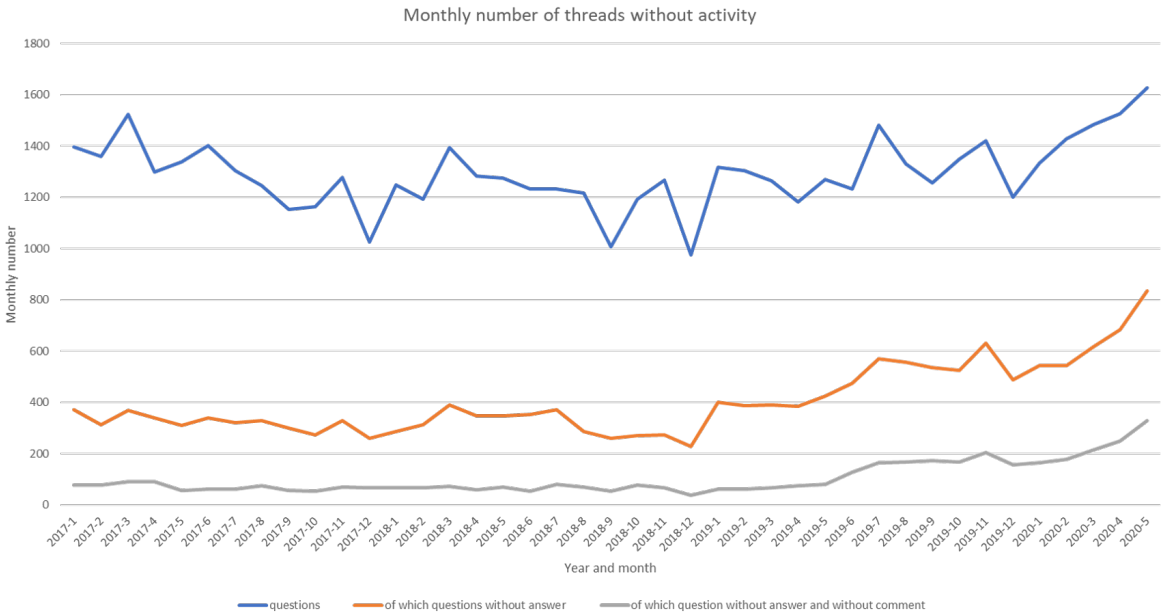


Figure D.1: Monthly number of threads without activity

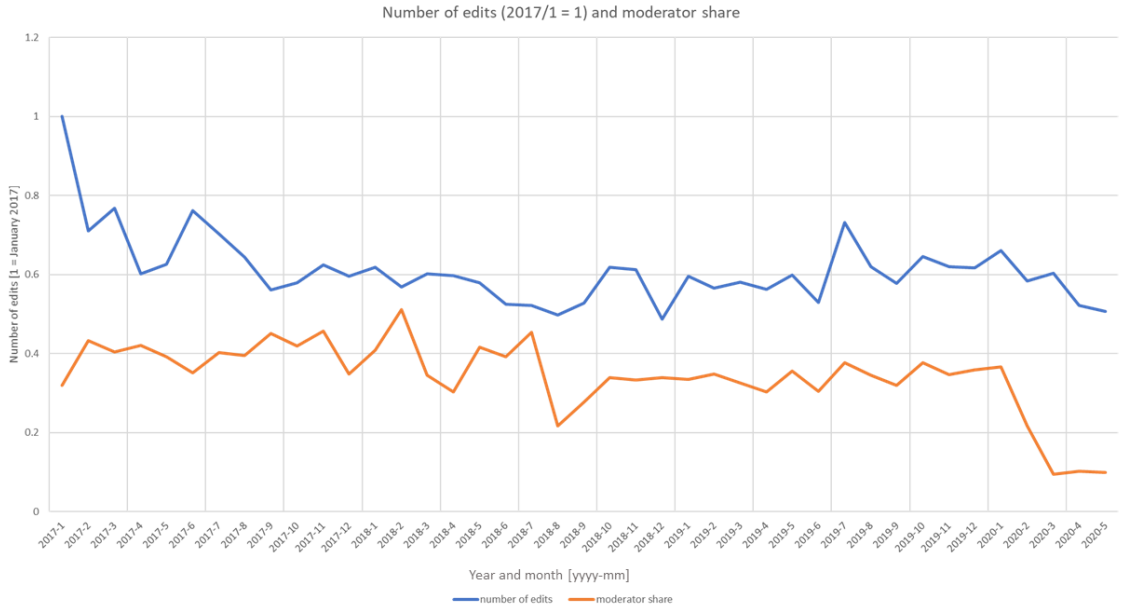


Figure D.2: Monthly number of edits (1.0=6565) and moderator share

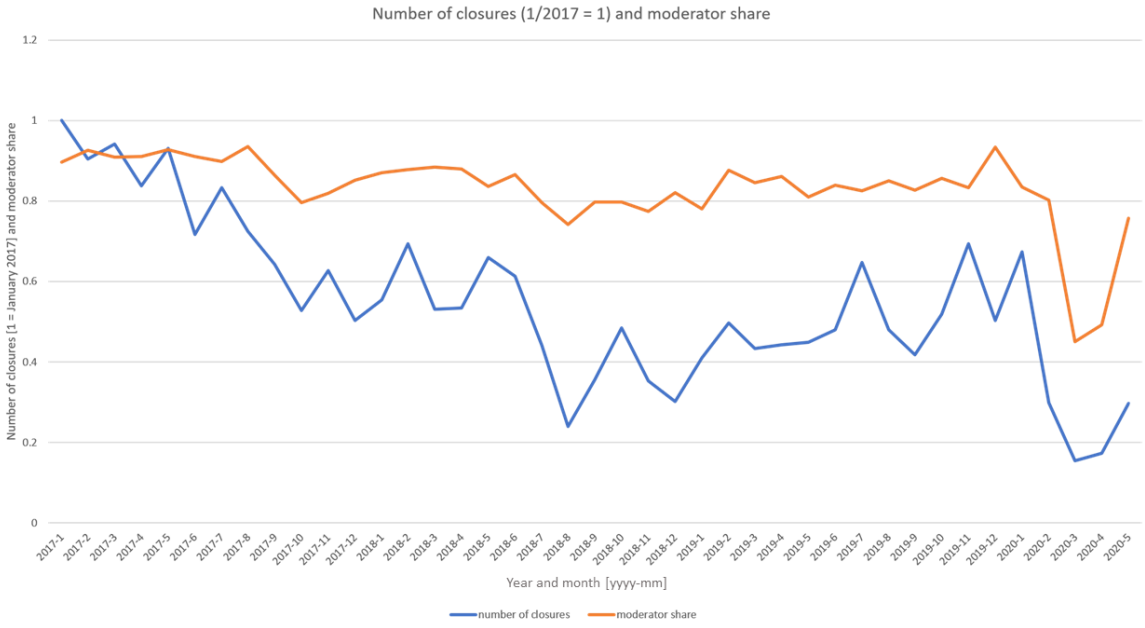


Figure D.3: Monthly number of closures (1.0=388) and moderator share

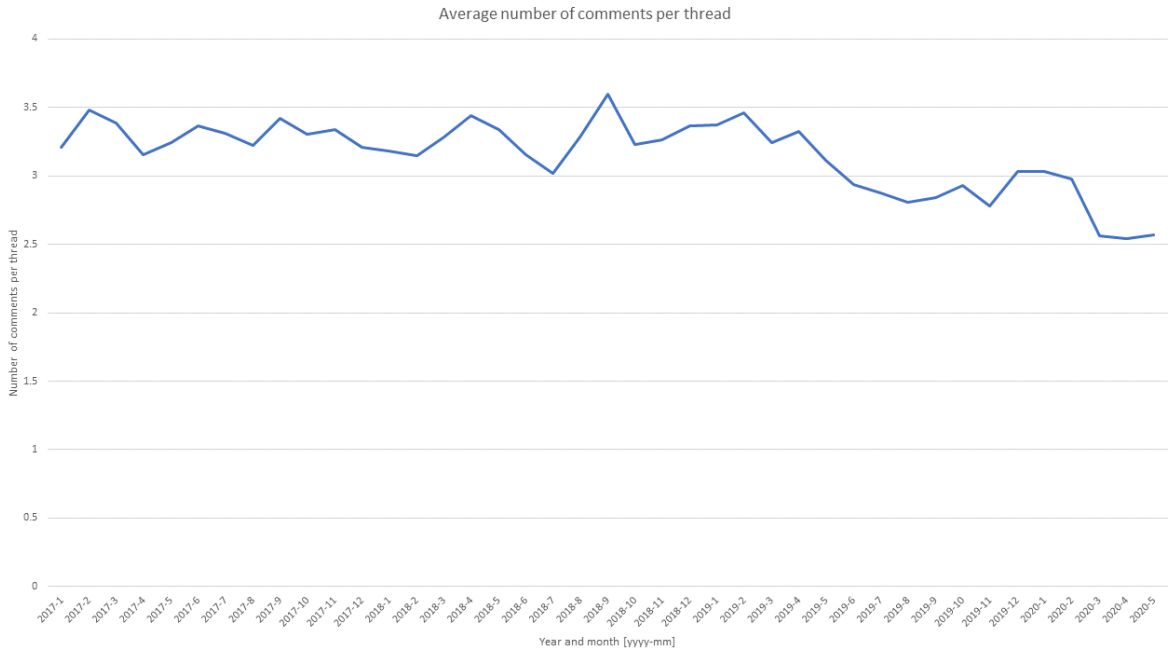


Figure D.4: Average number of comments per thread

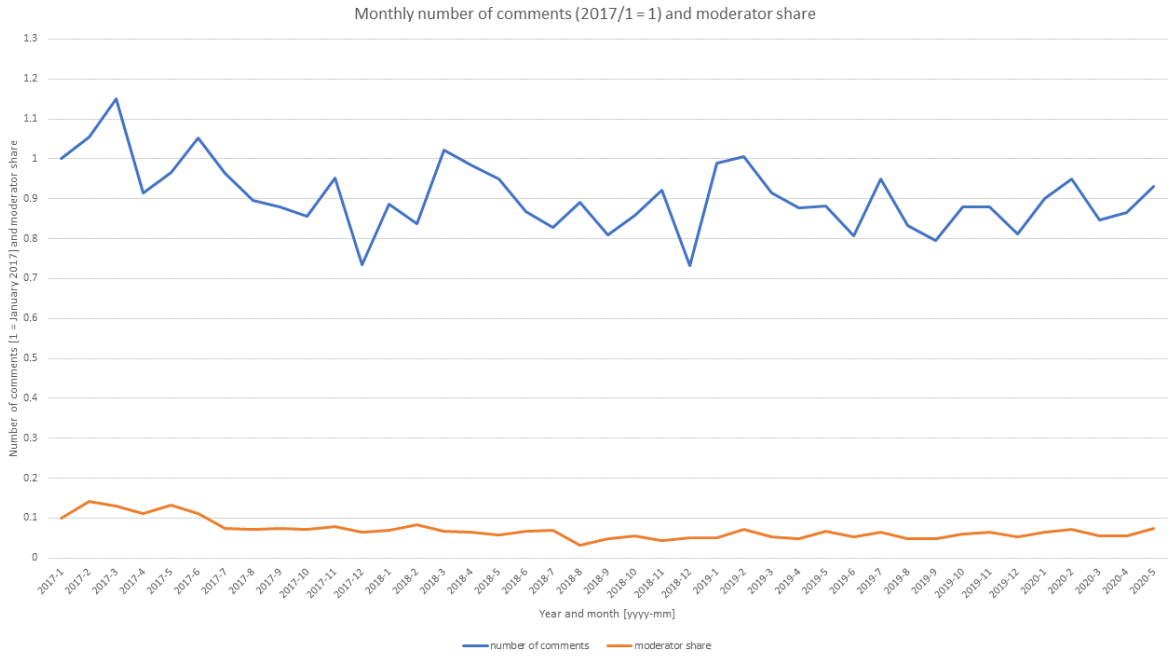


Figure D.5: Monthly number of comments (1.0=4485) and moderator share

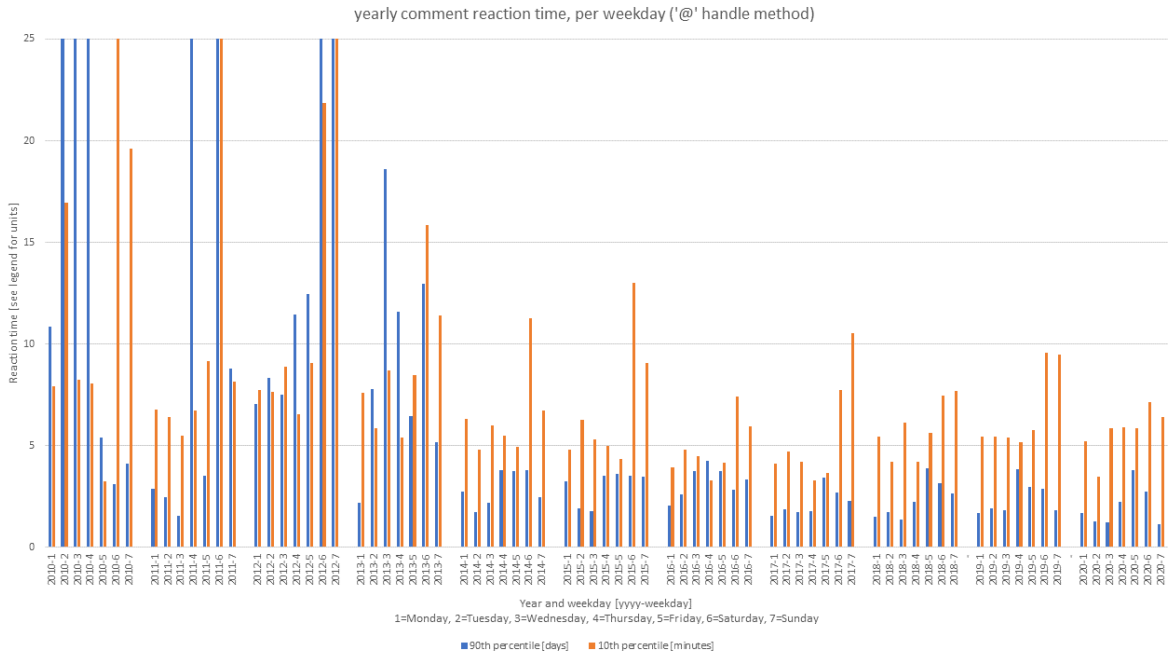


Figure D.6: Average reaction times per weekday and year, obtained by the handle approach.

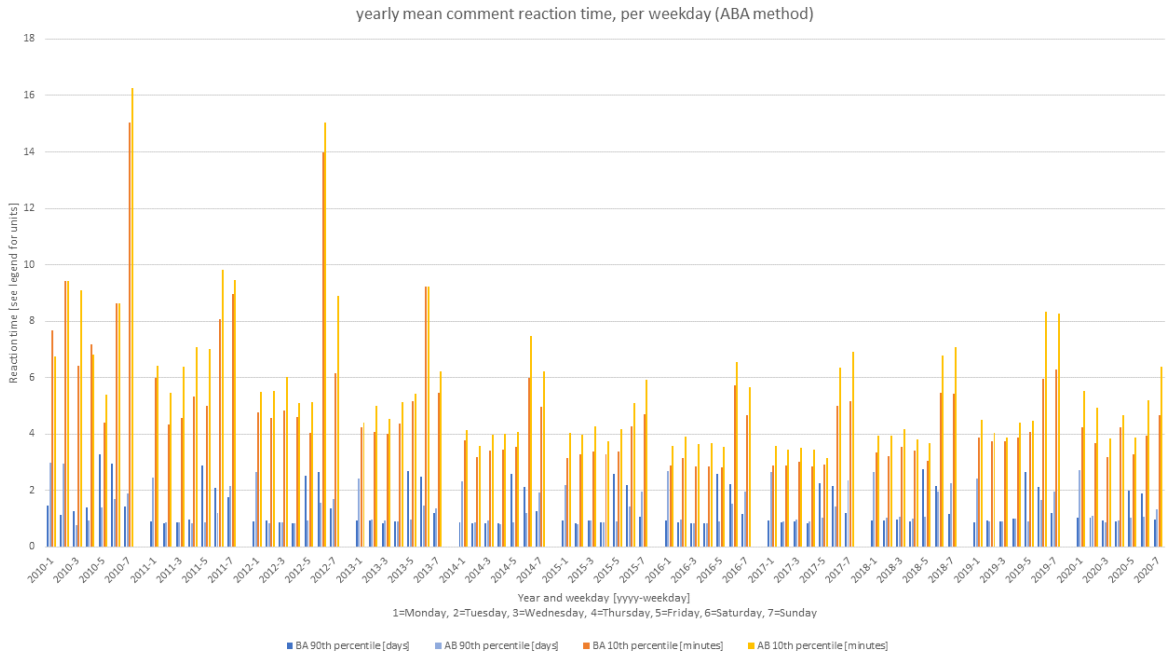


Figure D.7: Average reaction times per weekday and year, obtained by the ABA approach.

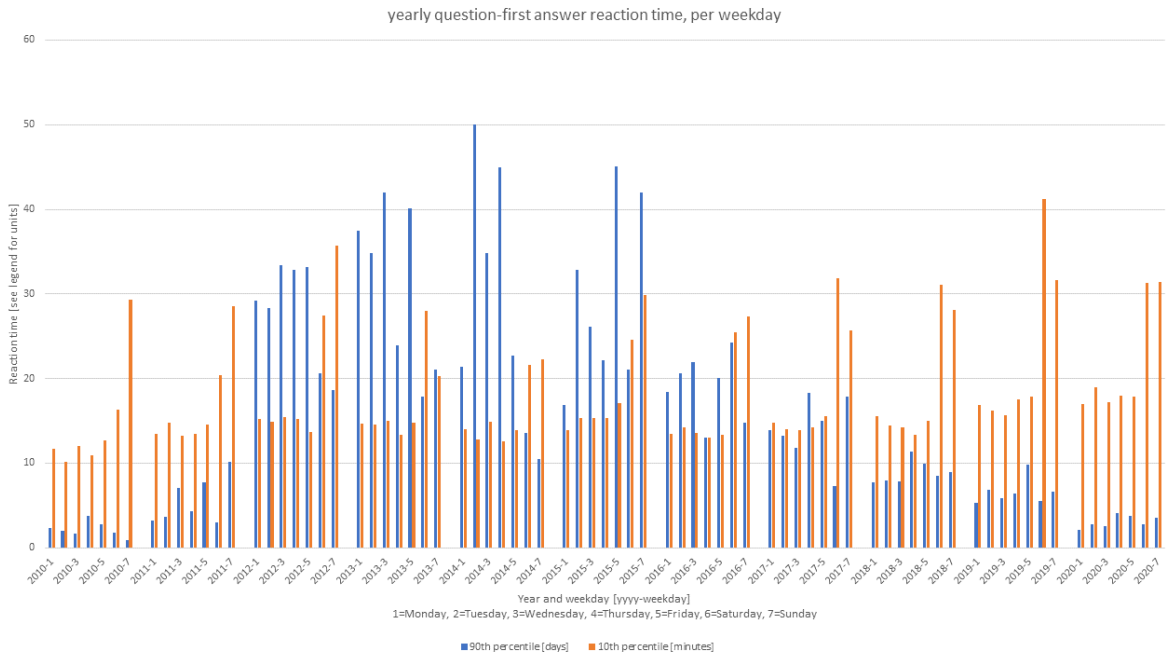


Figure D.8: Average reaction times per weekday and year, for question - first answer pairs.

Bibliography

Earl R Babbie. *The practice of social research*. Nelson Education, 2015.

Albert Bandura and Richard H Walters. *Social learning theory*, volume 1. Prentice-hall Englewood Cliffs, NJ, 1977.

Stephen R Barley. Technicians in the workplace: Ethnographic evidence for bringing work into organizational studies. *Administrative Science Quarterly*, pages 404–441, 1996.

Nathan Bos, Ann Zimmerman, Judith Olson, Jude Yew, Jason Yerkie, Erik Dahl, and Gary Olson. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, 12(2):652–672, 2007.

John Seely Brown and Paul Duguid. Knowledge and organization: A social-practice perspective. *Organization science*, 12(2):198–213, 2001.

C Candace Chou. A model of learner-centered computer-mediated interaction for collaborative distance learning. *International Journal on E-learning*, 3(1):11–18, 2004.

James F Courtney. Decision making and knowledge management in inquiring organizations: toward a new decision-making paradigm for dss. *Decision support systems*, 31(1):17–38, 2001.

Pierpaolo Dondio and Suha Shaheen. Is stackoverflow an effective complement to gaining practical knowledge compared to traditional computer science learning? In *Proceedings of the 2019 11th International Conference on Education Technology and Computers*, pages 132–138, 2019.

Richard Duschl. Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of research in education*, 32(1):268–291, 2008.

Rebecca Ferguson, Zhongyu Wei, Yulan He, and Simon Buckingham Shum. An evaluation of learning analytics to identify exploratory dialogue in online discussions. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 85–93, 2013.

Rebecca M Ferguson. *The construction of shared knowledge through asynchronous dialogue*. PhD thesis, The Open University, 2009.

Rich Gazan. Specialists and synthesists in a question answering community. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–10, 2006.

Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645–654, 2008.

Michael F Goodchild. Twenty years of progress: Giscience in 2010. *Journal of spatial information science*, 2010(1):3–20, 2010.

Stephen Gourlay. Conceptualizing knowledge creation: A critique of nonaka's theory. *Journal of management studies*, 43(7):1415–1436, 2006a.

Stephen Gourlay. Towards conceptual clarity for 'tacit knowledge': a review of empirical studies. *Knowledge Management Research & Practice*, 4(1):60–69, 2006b.

Noriko Hara, Curtis Jay Bonk, and Charoula Angeli. Content analysis of online discussion in an applied educational psychology course. *Instructional science*, 28(2):115–152, 2000.

F Maxwell Harper, Daphne Raban, Sheizaf Rafaeli, and Joseph A Konstan. Predictors of answer quality in online q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 865–874, 2008.

Remko Helms, W Ai, and Jocelyn Cranefield. TALKING TO ME? CREATING NETWORKS FROM ONLINE COMMUNITY LOGS. *ECIS 2016 Proceedings*, 2016.

France Henri. Computer conferencing and content analysis. In *Collaborative learning through computer conferencing*, pages 117–136. Springer, 1992.

Daniel CA Hillman, Deborah J Willis, and Charlotte N Gunawardena. Learner-interface interaction in distance education: An extension of contemporary models and strategies for practitioners. *American Journal of Distance Education*, 8(2):30–42, 1994.

Andri Ioannou, Scott W Brown, and Anthony R Artino. Wikis and forums for collaborative problem-based activity: A systematic comparison of learners' interactions. *The Internet and Higher Education*, 24: 35–45, 2015.

David Jonassen, Mark Davidson, Mauri Collins, John Campbell, and Brenda Bannan Haag. Constructivism and computer-mediated communication in distance education. *American journal of distance education*, 9(2):7–26, 1995.

Karel Kreijns, Paul A Kirschner, and Wim Jochems. The sociability of computer-supported collaborative learning environments. *Educational technology & society*, 5(1):8–22, 2002.

Priya Kumar and Anatoliy Gruzd. Social media for informal learning: a case of# twitterstorians. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

Priya Kumar, Anatoliy Gruzd, Caroline Haythornthwaite, Sarah Gilbert, Marc Esteve del Valle, and Drew Paulin. Learning in the wild: coding reddit for learning and practice. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

Guo Li, Haiyi Zhu, Tun Lu, Xianghua Ding, and Ning Gu. Is it good to be like wikipedia? exploring the trade-offs of introducing collaborative editing model to q&a sites. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1080–1091, 2015.

Stefania Manca, Manuela Delfino, and Elvis Mazzoni. Coding procedures to analyse interaction patterns in educational web forums. *Journal of computer assisted learning*, 25(2):189–200, 2009.

Anastasiia Menshikova. Evaluation of expertise in a virtual community of practice: The case of stack overflow. In *International Conference on Digital Transformation and Global Society*, pages 483–491. Springer, 2018.

Neil Mercer. Sociocultural discourse analysis: Analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics and Professional Practice*, 1(2):137–168, 2007.

Neil Mercer and Karen Littleton. *Dialogue and the development of children's thinking: A sociocultural approach*. Routledge, 2007.

Rebecca Mitchell and Brendan Boyle. Knowledge creation measurement methods. *Journal of Knowledge Management*, 14(1):67–82, 2010.

M Moore. Three types of interaction; the american journal of distance education, 1989.

Ingrid Mulder, Janine Swaak, and Joseph Kessels. Assessing group learning and shared understanding in technology-mediated interaction. *Journal of Educational Technology & Society*, 5(1):35–47, 2002.

Ikujiro Nonaka. The knowledge-creating company harvard business review november-december. *Google Scholar*, 1991.

Sirous Panahi, Jason Watson, and Helen Partridge. Social media and tacit knowledge sharing: developing a conceptual model. *World academy of science, engineering and technology*, (64):1095–1102, 2012.

Andraž Petrovčič, Vasja Vehovar, and Aleš Žiberna. Posting, quoting, and replying: a comparison of methodological approaches to measure communication ties in web forums. *Quality & quantity*, 46 (3):829–854, 2012.

Horst WJ Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4 (2):155–169, 1973.

Beat Schmid. Elektronische märkte-merkmale, organisation und potentiale. 1999.

Donald A Schön. Educating the reflective practitioner. 1987.

Subhasree Sengupta and Caroline Haythornthwaite. Learning with comments: An analysis of comments and community on stack overflow. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

Katarina Stanoevska-Slabeva. Toward a community-oriented design of internet platforms. *International Journal of Electronic Commerce*, 6(3):71–95, 2002.

Jonathan P Tennant, Jonathan M Dugan, Daniel Graziotin, Damien C Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, Lauren B Collister, Christina K Pikas, Tom Crick, et al. A multidisciplinary perspective on emergent and future innovations in peer review. *F1000Research*, 6, 2017.

Nicole Ummelen. *Procedural and declarative information in software manuals: Effects on information use, task performance and knowledge*, volume 7. Rodopi, 1997.

Huangxin Wang, Zhonghua Xi, Jean X Zhang, and Fei Li. An empirical study of financial incentivized question answering in social websites. In *Proceedings of the fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies*, pages 1–6, 2017.

Etienne Wenger. Communities of practice and social learning systems: the career of a concept. In *Social learning systems and communities of practice*, pages 179–198. Springer, 2010.

Tim Weninger, Xihao Avi Zhu, and Jiawei Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 579–583. IEEE, 2013.

John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502, 2007.