Characterizing a collaboration by its communication structure

Adrian Wüthrich[1]

Abstract

I present first results of my analysis of a collection of about 24,000 email messages from internal mailing lists of the ATLAS collaboration, at CERN, the particle physics laboratory, during the years 2010–2013. I represent the communication on these mailing lists as a network in which the members of the collaboration are connected if they reply to each other's messages. Such a network allows me to characterize the collaboration from a bird's eye view of its communication structure in epistemically relevant terms. I propose to interpret established measures such as the density of the network as indicators for the degree of "collaborativeness" of the collaboration and the presence of "communities" as a sign of cognitive division of labor. Similar methods have been used in philosophical and historical studies of collective knowledge generation but mostly at the level of information exchange, cooperation and competition *between* individual researchers or small groups. The present article aims to take initial steps towards a transfer of these methods and bring them to bear on the processes of collaboration *inside* a "collective author."

1 Knowledge generation in modern scientific collaborations

The large-scale experiments at CERN, the particle physics laboratory in Geneva, are among the paradigm examples of successful scientific collaborations. Among other achievements, two of them, ATLAS and CMS, have discovered the Higgs particle in 2012 (ATLAS Collaboration, 2012; CMS Collaboration, 2012), the existence of which was predicted long ago by the by now Nobel Prize winners François Englert and Peter

✉ Adrian Wüthrich
adrian.wuethrich@tu-berlin.de

1 Institute of History and Philosophy of Science, Technology, and Literature, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

Higgs and others (Higgs, 1964; Englert & Brout, 1964). But how are these large scientific collaborations organized? How do they accomplish a successful "division of cognitive labor" (Kitcher, 1990)? Are they particularly collaborative in ways that are conducive for epistemic success? And how can we determine how collaborative a collaboration is at all?

One of the most important empirical studies on the structures of scientific collaboration found that collaborations in particle physics are "atypical" because they feature "close interdependencies, low bureaucracy, and fluid organization" (Shrum et al., 2007). Karin Knorr-Cetina's seminal studies of knowledge generation in communities, such as that of high-energy physicists (Knorr-Cetina, 1999, 1995) brought to the fore the importance of the composition of the experimental group and choices of technology, which often grow out of previous generations of similar experiments. She also noticed the importance of the internal communication inside these communities which almost constantly articulate and reflect the ongoing research process by means of this communication. In such a "humming beehive", to use one of the metaphors from Knorr-Cetina, no single individual is able to have knowledge in the sense of justified belief. Rather, the collaboration is a genuinely "collective author" (Galison, 2003). Kent Staley (2004) has shown how the CDF collaboration at Fermilab, another "collective author", determined the exact nature of its results in discussions while writing the results up for publication. A more recent study on ATLAS identified the details of such excellently adapted internal communication structures that allowed a collaboration to share knowledge among its members, to validate it and to form a consensus on the results that will eventually be published (Graßhoff & Wüthrich, 2012).

But these studies have not, and could not have, attempted to capture such a collaboration's internal communication at the level of day-to-day messages, let alone in a nearly complete fashion. Although Harry Collins (2017) based his detailed study mostly on e-mail exchange among members of the LIGO collaboration, the discoverers of gravitational waves, he did so in a largely unsystematic manner, focusing on messages that "caught his eye" (cf. Collins, 2017, p. 1) thanks to his long involvement in the collaboration. Collins thus also did not attempt to come to terms with the seemingly intractable abundance of internal, almost exclusively born-digital communication (Kirschenbaum, 2013; Rosenzweig, 2003) such as e-mails, messages on chat platforms, and documents and comments on internal servers. It is only with the help of recently developed tools in computational approaches to the history of science (Laubichler et al., 2019) that a comprehensive study, which is neccessary for a bird's eye view analysis of the internal communication of a collaboration, has finally come within reach.

At the same time, the philosophy of science is also witnessing a trend to increasing use of digital tools and has yielded several intriguing results, especially from computer simulations, as to which structures for sharing information among researchers are epistemically most successful (see, for instance, Zollman, 2007). Although, at this stage, these models may not be directly applicable to the case of an internal communication structure, they provide growing evidence that characteristic measures of epistemic networks, such as the density or the extent of clustering, could be important relevant factors in facilitating the epistemic success of a collaboration or collective of

researchers. The present article is meant as an initial step towards bringing the abstract philosophical models in touch with a case study of actual scientific practice.

For my analysis, I can build on the previous study, *MetaATLAS*, mentioned above (Graßhoff & Wüthrich, 2012), in which I was involved. Among other things, with the support of the ATLAS spokesperson at the time, we collected a large amount of e-mail messages from ATLAS internal mailing lists. The present article is, however, also meant as a methodological suggestion for the analysis of other scientific collaborations provided there is a similar collection of internal documents at one's disposal.

In Sect. 2 I will describe our collection of e-mail messages in more detail and how I represent the ATLAS internal e-mail communication by a network. Section 3 suggests several ways of how to associate characteristic network measures with the "collaborativeness" of the collaboration. Section 4 motivates the analysis of the "community" structure of the network, which I interpret as an indicator of how cognitive labor is divided in the collaboration. Section 5 presents and discusses the results my student assistants and I obtained of the particular case of the ATLAS collaboration. Section 6 summarizes the main findings and takes stock of open questions.

## 2 Data and methods

2.1 The MetaATLAS collection

The central part of my analysis will rely on the collection of 24,000 e-mail messages obtained in the framework of "MetaATLAS" (Graßhoff & Wüthrich, 2012) between 2010 and 2013. The messages were stored using common e-mail-client software in a common format.[1] The messages came either from standard mailing list setups or from "HyperNews", "one of the main ATLAS collaborative tools."[2] In private communication, several members of the ATLAS collaboration have also confirmed that the mailing lists (standalone or as part of the HyperNews platform) of the MetaATLAS collection were important for the communication inside ATLAS at the time under study here. Messages from HyperNews were received just as regular messages over a mailing list. The communication platforms of ATLAS at the time are described in Marti and Schefer (2012). A detailed analysis of some of these messages is part of Graßhoff and Wüthrich (2012) and Wüthrich (2017).

The mailing lists and HyperNews groups served various purposes, ranging from general messages such as weekly agendas proposed by the ATLAS spokesperson and addressed to all ATLAS physicists ("current-physicists") to specialized discussions of issues in dedicated working groups ("jet-etmiss-wg").[3] I could not find much explicit information about the character of the standalone mailing lists. However, in a document

---

[1] The format used was "mbox." For a specification, see https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml (accessed 2022-01-19).

[2] From a document which is part of the MetaATLAS collection and cited in (Schefer, 2012, fn. 12, p. 74).

[3] For reasons of data protection, I do not display the complete addresses of the lists but only give a short label extracted from the part before the "@"-sign.
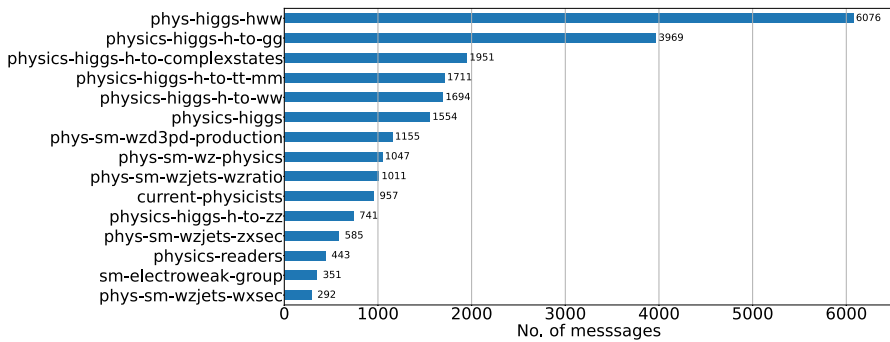
Fig. 1 The 15 lists with the most number of messages from November 2010 to March 2013. For reasons of data protection, I do not display the complete addresses of the lists but only give a short label extracted from the part before the "@"-sign

contained in the MetaATLAS collection, the HyperNews platform is described as a "discussion management system combining a forum and an email list."[4]

From the document just mentioned one can also learn that the HyperNews mailing lists mirror the discussions in the forum. The ATLAS members can participate in the dicussions over a web-interface or via the associated mailing lists. The document also states that the HyperNews forums are open to every ATLAS members; a restriction to subgroups is not possible. However, the ATLAS members can choose to which forum and associated mailing list they want to subscribe. I could not find any indication, either in the email messages or in the document describing the HyperNews platform, to the effect that the mailing lists and forums are moderated. There was the possibility that new forums (and associated mailing lists) are created on the HyperNews platform. Those who are subscribed to the "systems announcements" forum would be informed about such changes. Judging from the analysis provided here and from my close reading of selected messages, the standalone mailing lists and the HyperNews mailing lists seem completely similar in character. So, for the present analysis, I treat the messages on the standalone mailing lists and the HyperNews mailing lists as the same type of source.

2.2 Quantitative overview

In total, the MetaATLAS collection contains 23,826 messages from 175 mailing lists or lists associated with a HyperNews group.[5] The messages are, however, unevenly distributed over the different lists. The 15 lists with the most messages over the analysis period (see Fig. 1) make up more than 90 per cent of all the messages sent over that period.[6]

---

[4] See Footnote 2.

[5] Not counting 205 messages which do not have a properly formatted date field.

[6] Note that we are counting a single message several times if it is sent to more than one list. Thus the "sum of sums" is not equal to the total number of messages.

Also, only 36 of the lists have received, on average, more than one message per month over the analysis period (29 months).

The two most active lists (phys-higgs-hww and physics-higgs-h-to-gg) received over 6000 or close to 4000 messages over the analysis period. These two lists were, perhaps not surprisingly, concerned with two important decay channels in which the Higgs boson was first observed: Higgs decay into two W bosons or into two photons.[7] The communication concerning the decay into two W's was even more active than the number of messages on "phys-higgs-hww" indicates because, around March 2012, there was a switch in activity to this list with its 6,076 messages from a HyperNews group called "physics-higgs-h-to-ww" with 1,694 messages over the analysis period. Also, other lists, such as "phys-higgs-hww-winter2012" deal with the decay of the Higgs boson into two W bosons.

Another important decay channel, the Higgs decay into two Z bosons, had a less active mailing list associated with it ("physics-higgs-to-zz") but the list is still among the 15 most active ones. Also many of the other most active lists are concerned with the analysis of a possible Higgs decay (for instance into two tau leptons or two mu leptons, corresponding to "physics-higgs-h-to-tt-mm", or into "complex states", corresponding to "physics-higgs-h-to-complexstates").

Others among the top 15 lists are concerned with physics of the bosons of the electroweak interaction, W and Z, or with messages to all "current physicists" or "readers." The list for the "current physicists" contains different kinds of messages such as the announcements of weekly agendas or concerning the approval of plots showing preliminary results. The messages to the "readers" are almost exclusively announcements of draft papers for which comments are invited.

In the collection, the first email message is from January 4, 2010, and the last one from April 1st, 2013. Figure 2 shows the number of messages over the whole period of our collection.

A peak in the month before the announcement of the discovery of the Higgs-boson (July 4, 2012) is clearly visible. It is, indeed, very plausible that close to the announcement of the discovery of the Higgs boson, there was a particularly high intensity of internal communication inside the ATLAS collaboration. The fact that the MetaATLAS e-mail collection shows such an increase in communication intensity speaks to the representativeness of the collection.

On average, nearly 600 messages per month were sent (mean: 596, median: 584.5). In the first 9 months (until September 2010), however, as well as in the last month (April 2013) much fewer messages were sent than on average. In these months, the MetaAT-LAS team received messages only from a couple of lists mostly of general purpose such as "current-physicists." From October 2010, and even more so from November 2010, they received at least some messages from 175 mailing lists or HyperNews groups. Also, on all lists with no messages before October 2010, the number of messages increase almost immediately to roughly the average value over the whole period

---

[7] This consideration and the following are based on the information on https://espace.cern.ch/atlas-forums/Lists/Atlas%20forums/By%20category.aspx (retrieved 2021-09-15), close reading of selected messages and the names of the mailing lists, where "h" is an abbreviation for the Higgs boson, "w" for the W boson, "g" for photons ("gamma" particles) etc. For a description of the decay channels and their relevance to the Higgs discovery, see ATLAS Collaboration (2012).
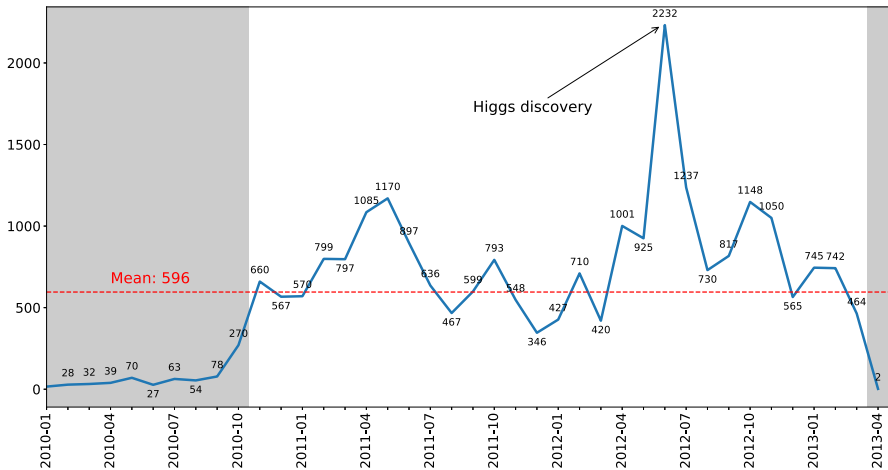
Fig. 2 Number of messages per months during the MetaATLAS-collection

in October or November 2010 and drops again to zero in April 2013. I conclude from this that the small number of messages in the initial and the last months is because the MetaATLAS team didn't have access to these additional lists rather than that the lists were not yet active. The initial months and the final one are, therefore, not likely to be a part of a representative collection and we will, accordingly, exclude them in our further analysis.

2.3 Comparison to other mailing lists

I know of no study of the e-mail communication inside a scientific collaboration, which we could use to calibrate our quantitative results. We will, therefore, use a sample from the "Linux Kernel Mailing List" as one of our reference points.[8] The messages are freely available on the internet, and the list is described as "a rather high-volume list, where (technical) discussions on the design of, and bugs in the Linux kernel take place."[9] The collaborative character of the mailing list may be similar, to some extent, to that of the ATLAS mailing lists. We collected the messages using a python script that downloaded all the messages from 2012 without manual intervention but with some time lapse between the queries in order not to overcharge the server.[10] Another reference point will be a study of general e-mail traffic on the server of the University of Kiel (Ebel et al., 2002).

---

[8] I thank Dirk Wintergrün for bringing up this idea.

[9] https://lkml.org (accessed 2021-08-03).

[10] This was almost exclusively the work of Michael Zichert, student assistant in two of my current projects.

2.4 Reconstructing the communication network

The core of my analysis consists in the analysis of a network representing the e-mail communication given by our collection. I built up the network as follows: Each ATLAS member who sent a message to one of the mailing lists at least once is represented by a node in the network. Two nodes A and B are connected by (directed) edges if and only if the person A replied to the message sent to the list by person B. This is similar to grouping e-mails by a discussion "thread." However, we use "direct" edges only, i.e. a reply from A to a message from B which is a reply to a message from C does not imply an edge from A to C. Several replies from one person to another are represented by several edges or, equivalently, by a single edge with "weight" equal to the number of replies.

We identified the ATLAS members by their full name.[11] These turned out to be unique in our collection. To determine the full names, we used the author list of ATLAS' "Higgs discovery paper" (ATLAS Collaboration, 2012) as a pool of potential last names and initials.

The procedure is described in more detail in Appendix A, and it may seem cumbersome. But the best possible reduction of the number of cases with nodes in the network that represent more than one person is important in order to produce a reliable characterization of the network and its nodes by centrality measures or the degree distribution etc. (see below). For instance, if a person is erroneously represented by two distinct nodes, then her or his importance as measured by the number of its edges (the network "degree") will often be significantly underestimated. The importance of the person as per this measure should be the sum of the "degrees" of the nodes whereas with two distinct nodes one will conclude that there are two persons with lower degrees (or one with equal degree and the other with degree zero).

Most of these tasks did not need manual intervention such that we could perform them with our own python scripts and readily available modules such as "mailbox" or "networkx."[12] For the visualizations we used the python libraries matplotlib and also gephi, a software specialized for network analysis.[13]

3 Measuring collaborativeness

3.1 Density and clustering

One way to quantitatively characterize a network such as ours is by determining some characteristic global measures. Some of them are, I will argue, good candidates for providing an idea of how collaborative the collaboration is. One such candidate is the *density* of the network. The density of a network is defined as the fraction of edges actually present in the network relative to all possible edges given the number of nodes

---

[11] This part of the analysis has mostly been conducted by Beate Federau, student assistant at TU Berlin in 2019.

[12] https://docs.python.org/3.5/library/mailbox.html, https://networkx.org/ (accessed 2021-04-10). See also Hagberg et al. (2008).

[13] See https://matplotlib.org/ and www.gephi.org (accessed 2021-08-03). See also Bastian et al. (2009).

of the network (Newman, 2018, p. 128). This is equivalent to the mean degree of a node in the network divided by the number of nodes. Informally, a higher density means that the nodes in the network are "better connected." As a representation of the communication over the mailing lists the density thus measures the number of replies to messages on the list relative to the number of persons who replied at least once to any of the messages sent to one of the lists. If we share the premise that more communication is an indicator for more "collaborativeness," this suggests a first definition:

Measure 1 The collaboration is more collaborative the denser the network representing its communication structure is.[14]

In a collaborative collaboration, according to this measure, the persons engaged in the communication in our network tend to reply often to the messages sent over the lists.

A similar candidate for a measure of collaborativeness is the *clustering coefficient* of the network. Contrary to what the name may suggest, the coefficient does not measure the number of clusters or the amount of group structure in the network.[15] Rather, it measures the density of "triangles" or, in other words, the *transitivity* of the connections between nodes. The clustering coefficient of a network is higher the more often a given node A, which is connected to node B, which in turn is connected to node C, is also connected to node C. To state it more informally: If you are in a highly clustered network of friends, the friend of your friend is likely also your friend. Applied to the e-mail communication in ATLAS this means that in a highly clustered network the persons tend to reply also to those to which the recipient of their reply has replied. This also means, that the information flow does not tend to be canalized through only a few individuals but can rather take many different routes and spread out in the network—there are no "structural holes" (cf. Burt, 1995; Newman, 2018). Accordingly, I propose a second sense of collaborativeness:

Measure 2 The collaboration is more collaborative the higher the clustering coefficient is of the network representing its communication structure.

The focus on triangles that comes with the clustering coefficient is a special case for the search of *motifs* in the network (Milo et al., 2002, Newman, 2018, p. 334). Motifs are subgraphs (triangles, boxes, chains etc.) that occur more often in the given network than one would expect from random networks. They often can be associated with particular functions of the network such as certain kinds of filtering in information processing. It would be interesting to see whether such motifs exist in the network under study here, too. But this would go beyond the scope of the present article. Also, the measurement of the frequency of triangles, i.e. the clustering coefficient is by far the most common instance of motif analysis. There are far fewer studies on the more general motifs in real-world networks. Also the Kiel study (Ebel et al., 2002) does not feature any study of motifs.

---

[14] Strictly speaking, in network theory, "dense" means that the density of the network does not tend to zero in the limit of an infinite number of nodes (Newman, 2018, p. 128). For empirical networks, such as the present case, "dense" rather just means that a considerable fraction of the edges that are possible combinatorially are also there actually.

[15] Cf. Sect. 4.2 and Newman, 2018, p. 183, fn. 16)

3.2 Degree distribution and mixing

In addition to these rather straightforward global measures, I propose two other approaches to quantify the collaborativeness of ATLAS. These other approaches use measures related to the *degrees* of the nodes in the network and I will propose to associate these measures with the presence of hierarchical structures in the collaboration represented by the network. The degree of a node is simply the number of edges it has to other nodes. In a directed network, i.e. in a network where the edges from a node *A* to a node *B* are distinguished from the edges from *B* to *A*, the number of edges pointing towards a node is its *in-degree*; its *out-degree* is the number of edges pointing from the node to other nodes.

In many real-worlds networks such as citation networks and the world wide web, the degrees of the nodes are distributed following a power-law (Newman, 2018, Fig. 10.1, p. 305):

$$n_k = ck^{-\alpha}, \tag{1}$$

where n_k is the number of nodes that have degree k, α is a parameter that can be tuned such that n_k fits the empirical values in an optimal way. The parameter c is determined by the condition that sum of all n_k's equals the total number of nodes. Also the study on the email traffic at Kiel University (Ebel et al., 2002) found such a degree distribution.

Perhaps the most remarkable instance of these findings are citation networks where the in-degrees often follow a power-law distribution (de Solla Price, 1976; Newman, 2018, p. 435). Taken as idealized models of the empirical networks, it has been shown that these networks can be generated by "preferential attachment" (Barabási & Albert, 1999). Thus, much of the empirically observed citation patterns can be explained by the fact that papers that have already received a great deal of citations up to a certain point in time will continue to receive more citations than papers which have not been cited very often.

So whether an empirical degree distribution follows a power-law or not may suggest some or other mechanisms of how a network featuring such a distribution may have come into existence. Power-law distributions can be explained by preferential attachment; other distributions don't—at least not without significant modification. If our empirical data does not fit well a power-law distribution but rather other kinds of degree distributions, we will have to come up with other possible explanations of how such a network has formed. And also if we do indeed find a power-law degree distribution of our network we can, of course, not be sure that it indeed has formed according to preferential attachment. Alternative explanations are not excluded. Rather, we are dealing here with an instance of an inference to the best explanation (Lipton, 2004). I will come back to such issues for our ATLAS network below (Sect. 5.3). Here I will first deal with the question of what we can learn from the degree distribution about the "collaborativeness" of the collaboration.

Theoretical network studies have found that the exponent of the power-law that best fits the degree distribution of a network also determines how "top-heavy" the distribution is (Newman, 2018, p. 328). Among the best known examples of top-heavy distributions is monetary wealth in the world or in a country (cf. Newman, 2005, p. 11).

Most of the wealth is held by only a small minority of persons. More precisely, mathematical considerations show that in a network with a degree distribution following a power-law with exponent $\alpha > 2$, the fraction W of wealth that is held by the fraction P of the population is given by (Newman, 2018, p. 328)

$$W = P^{(\alpha-2)/(\alpha-1)} \quad (\alpha > 2). \tag{2}$$

If the exponent α of the power-law is less than or equal to 2, W tends to 1 (or 100%) for any fraction P, such that "essentially all of the wealth (or other commodity) lies in the tail of the distribution" (Newman, 2005, p. 11). Thus, the value 2 of α separates two regimes: one in which we can calculate the wealth that a certain fraction of the population accumultes and one in which the mathematical expressions become ill-defined and all the wealth is concentrated in an infinitely small fraction of the population. As far as I can see, α does not have any meaning beyond that such that it would refer to certain characteristic features of a mechanism such as preferential attachment.

Applied to scientific collaboration networks, more precisely for networks of co-authorship in biomedical research, physics and computer science, these considerations have led to the conclusion that if the exponent of a power-law degree distribution is less than 2, the network is "dominated by the few individuals with a large number of collaborators" (Newman, 2001b, p. 407).

Below, I will investigate whether the degree distribution in the ATLAS-e-mail network is reasonably fitted by a power-law and, if yes, with what exponent. After all, also the Kiel study (Ebel et al., 2002) found such a pattern in the email communication. Also, because networks with power-law degree distributions are well investigated we could build our further analysis on well-established theoretical results. In particular, we might be able to tell whether the e-mail communication is rather a matter of a small number of persons who reply to a lot of messages on the lists or whether the activity on the lists is more equally distributed, which I would take as a sign of "collaborativeness." Accordingly, I propose a third measure of collaborativeness:

Measure 3 A collaboration is collaborative if the network representing its communication displays a degree distribution following a power-law with exponent larger than 2.

One further approach to quantifying collaborativeness also derives, as mentioned, from the degree distribution in the network. The relevant criterion here is whether nodes of a certain degree (in the above sense of number of edges) in the network are more frequently connected to nodes with a similar or equal degree. I propose to take such "assortative mixing by degree" (Newman, 2018, p. 209) to suggest that the network is rather hierarchical and, therefore, less collaborative than networks with nodes that are equally well connected to "unlike" nodes, i.e. nodes with significantly different degrees than their own. A fourth definition of collaborativeness thus reads:

Measure 4 A collaboration is collaborative if the degrees of the nodes in the network representing its communication are not assortatively mixed.

Applied to the ATLAS e-mail network this means that the collaboration should be seen as rather collaborative if ATLAS members reply to messages with a certain

probability that is independent of whether the message comes from members who receive, on average, a similar number of replies as their own messages to the list. I will give a mathematical formulation below when we will calculate the mixing for the ATLAS case.

4 Distributive cognition and division of cognitive labor

4.1 Collaborations vs. collectives

The philosophy of science and cognate fields have, since some time, taken up the challenge of finding out how, especially in large scientific collaborations, the research about complex objects of investigation can be split up in optimal ways and eventually put together to yield a justified over-all result (Knorr-Cetina, 1999; Giere, 2007). Such research is often an instance of "distributed cognition." In the framework of the project "MetaATLAS", and in particular in the work by Maya Schefer (Schefer, 2012), a significant part of the process has been identified in which small bits of knowledge and preliminary results make their way through the ATLAS collaboration and thus become justified findings of the whole collaboration. Schefer's analysis was the result of close reading selected internal electronic documents collected in MetaATLAS, among other things the e-mails that are the focus of the present analysis.

A related strand of philosophical research has also dealt with collective knowledge generation but with regards to a collective of individual authors or small research teams (Kitcher, 1990; Zollman, 2007).[16] In this context, the most relevant process is not the splitting up of complex problems into more manageable ones. Rather, several individuals or small groups try to solve the problem completely on their own with possibly different methods and working hypotheses. The important question then is what the optimal procedure is such that at least one researcher or team, and thus the collective as a whole, will eventually solve the problem. The main result of this strand of research is that the probabilites of success are often highest if there is a "division of cognitive labor" such that "cognitive diversity" is maintained sufficiently long in order for certain methods or hypotheses not to be discarded prematurely.

My study cross-cuts these two strands. It is mostly concerned with the distributed cognition inside a scientific collaboration (as is the first strand) but it proposes to use methods and concepts mostly from the second strand. Like the second strand, I use network analysis to characterize the communication structure of a group of researchers in potentially epistemically relevant terms. The details of such a methodological transfer have yet to be spelled out but roughly I suggest the following mapping.

In both cases, the "epistemic success" consists in things like solving a scientific problem, making an empirical discovery or developing a valid theory. However, in the case of a collaboration (the first strand) the success is achieved by the collaboration as a whole through coordinated work on parts of the overarching problem. In the case of a collective of researchers (the second strand), success is achieved by individuals or small

---

[16] I use the term "collective" instead of "community" here because, in this article, "community" refers to the groups of network nodes detected by computer algorithms, see below.

groups in the first place. The epistemic success of the collective can then be understood in the sense that at least one individual in the collective was successful or that a large proportion of individuals was (such that there is a consensus in the collective). Yet, information flow, and thus the communication structures under investigation with the second strand's (and my) methodological approach, presumably plays similar roles in both cases.

For instance, something like the "Zollman effect," a remarkable result of the second strand, may occur also in the context of the first strand. In the case of the original Zollman effect (Zollman, 2007) a reduction of the information flow makes the agents in the collective pursue eventually successful methods that they would have abandoned had they had information about the other agents' findings. But one can imagine similar scenarios for the communication inside a collaboration. In this case, the reduction of information flow may, for instance, make the members of the collaboration maintain correct partial results that they might have discarded had they known about temporarily erroneous partial results of other members.

4.2 Communities and central persons

These considerations motivate the following attempt to characterize, through a network analyis, the communication structure of ATLAS in a way that allows some conclusions about how the research is distributed and eventually combined. A first step towards such a characterization is the detection of "communities" inside the collaboration. This is a fairly standard procedure in network analysis of all sorts and some procedures to determine such communities have established themselves. The most widely used definition of a "community" in a network is based on *modularity*, yet another global measure (besides density and clustering, which we encountered above). Communities in this sense are those partitions of the network which maximize modularity (Newman, 2018, Sect. 14.2). This means that when you break up the entire network into communities as per this definition the nodes inside these communities have more connections to each other than they have to the nodes in other communities. It is, in fact, another case of assortative mixing, as mentioned above in discussing the degrees of the nodes. In order to find, at least approximately, the modularity-maximizing partitions of the ATLAS-e-mail network we used the popular "Louvain algorithm" (Blondel et al., 2008) as implemented in networkx. The communities so obtained will then be interpreted as a sign for division of cognitive labor, i.e. it is probable that inside these groups the communication focuses on a particular aspect of the over-all results produced by the collaboration as a whole.

In order to gain, from this bird's eye view, also some insights into how the results are put together, we identified "central" persons in the network according to varying definitions (see, for instance, Düring, 2016). The two most important definitions are "degree centrality" and "betweenness centrality." Degree central nodes simply have a lot of connections to other nodes. Betweenness central nodes are those through which many shortest paths between two other nodes pass. In other words, betweenness central nodes (in our case representing persons) are those who may have only few connections but those connections are to people who are highly connected in special ways. Such

persons often function as "brokers" between communities in the network and can indeed be the most important persons in the sense that they hold the communities together and might be crucial when it comes to combine the partial results obtained in the communities.

Below, we will also investigate how much "cross-activity" there is over the lists, i.e. how often ATLAS members tend to be active on various lists, which are dedictated to different topics, or rather on the lists that are the most relevant for the community (as defined above) they belong to. This should also give us a sense of to what extent the cognitive labor is divided and how much activity is needed to bring the partial results together.

It goes without saying that such a bird's eye view will not be able to tell the exact procedures that take place inside ATLAS to collectively obtain scientific results. Close reading analyses, such as Schefer's (2012), are necessary to identify, for instance, the exact nature of the tasks required for the combination of partial results. However, among other things, the bird's eye view analysis may bring to our attention the importance of persons that a close reading analysis of a selection of documents may have overlooked. Moreover, finding algorithms or measures for determining persons with characteristic roles in a network will also help to spell out more explictly and precisely the function of these roles (cf. Herfeld & Doehne, 2019). The digital tools may thus serve not only as heuristics for more in depth analysis but also as "explicators" for certain otherwise vague or incomplete concepts.

5 Results

5.1 General structure

Figure 3 shows a visual representation of the network generated by the above mentioned rules, i.e. nodes represent persons and edges between two nodes the fact that a person replied to the other person's message. The visualization has been produced using the "Force Atlas" layout of gephi. The size of the nodes corresponds to the "betweenness centrality" of the person in the network. The colors represent "communities" as detected by the "Louvain algorithm" (for the undirected variant of the network). I will discuss possible interpretations of the communities below.

The network has 749 nodes and 5591 directed edges, not counting 255 self-loops, i.e. edges from a node to itself. A self-edge would represent the fact that a person sends a message to one of the lists and later the person responds or adds some information to the initial message by replying to it. We exclude self-loops from our present analysis for various reasons: First, we focus here on interactions inside the collaboration; second, they make up less than 5% of all edges; third, most network measures (cf. Sect. 5.2) are usually calculated without self-loops. So in order to make our study as comparable as possible to other studies we do exclude them, too.

By the construction rules of our network, a message to the list to which nobody replies is not represented as an edge but only by the node representing the sender. If, for instance, a person only sends announcements that may need no reaction and thus get no replies will remain an isolated node in the network. The author list of ATLAS'

Fig. 3 Visualization of of the network representing replies to messages on mailing lists used inside the ATLAS collaboration (November 2010–March 2013). Only the largest component is shown. It is a two-dimensional layout produced by gephi's "Force Atlas" algorithm. The colors represent "communities." The size of the nodes varies with their "betweenness centrality" (see text for further explanation)

"Higgs discovery paper" (ATLAS Collaboration, 2012) lists close to 3000 authors. I.e. about a fraction of 25% of the ATLAS authors are also represented as a node in the network.

The *weight* of the edges represents the number of replies between the persons represented by the nodes connected by the edges. That is, if a person (represented by node A) replies 10 times to a message sent to one of the lists by another person (represented by node B), there will be a directed edge of weight 10 from node A to node B. The mean value of the weight of the edges is 2.22 (with a median of weight 1). The maximum weight is 64.

As most networks of interest which are generated from empirical data the network representing the ATLAS communication over mailing lists features a distinct *largest component*. In a directed network a (weakly connected) component is a subset of all nodes in which there exists at least one path between any two nodes in that subset.

A path is a set of edges by which a node is connected to another node, possibly through intermediate nodes and not going through a node twice. In a weakly connected component a path is allowed to contain edges in any direction. In a strongly connected component the path must follow the direction of the edges. Different networks can best be compared if one considers weakly connected components. The largest such component in the ATLAS network includes 92% of all nodes. I.e., almost all nodes are in the largest component, which means that almost all nodes can be reached from any other node through a path (disregarding the direction of the edges). Figure 3 only shows the largest component of our network.

The ATLAS network also displays a "small world" effect according to which the average shortest path connecting two nodes traverses only around 6 edges in many social networks (Milgram, 1967; Travers & Milgram, 1969; Newman, 2001a). In our case, the average shortest path in the largest component is just about 2.62, meaning that most pairs of nodes are connected by paths of either length two or length three. If we also take into account the smaller components of the network, and follow a common definition of mean distance in such cases (Newman, 2018, Eq. 10.3), we obtain an average shortest path of length 3.62, which still qualifies as an indicator of a "small world", given that the network contains around 700 nodes. If these nodes formed a chain, the average shortest path (or mean distance) would be around 230.[17] As with most real-world networks, the particular structure of our network dramatically reduces the average shortest paths.

5.2 Global measures

Let us start our more detailed analysis of this network with the two global measures mentioned above, density and clustering. These quantities can be readily determined by the corresponding algorithms provided by the networkx library for python.

For the calculation of the density of the ATLAS e-mail network, we disregard the weight of the edges. That is, we represent even several replies from one person to another as just one edge between the corresponding nodes. Otherwise, the density would not be bounded from above and could not be used as a measure of how well the nodes in the network are connected. Similarly, we continue to exclude self-loops, i.e. edges from a node to itself, because this would also lead to no standard upper bound of the density. Thus, the maximum density ($d = 1$) would obtain in the network if all nodes were connected to all nodes, where "connected" means that the represented persons replied at least once to the other person's messages, and vice versa. In other words, the network would be a *complete* (directed) graph.

In a simple, unweighted, directed network, the density d is given by

$$d = \frac{m}{n(n-1)}, \tag{3}$$

---

[17] This result can be obtained, for instance, by using the methods "path_graph" and "average_shortest_path_length" of the networkx library for python.

where m is the number of edges and n the number of nodes in the network.[18] For our case of the ATLAS e-mails we obtain $d = 0.010$, i.e. quite precisely one percent of all the possible edges are indeed present. This may come across as very little but compared to the Kiel study of e-mail messages (Ebel et al., 2002) and also compared to our own rough and ready analysis of e-mail messages of the Linux kernel developers such a density is still two to four orders of magnitude higher. In this respect, we may thus conclude that the ATLAS collaboration is indeed rather collaborative (according to Measure 1).

As mentioned, some prominent results from the philosophy of science, in particular by Kevin Zollman (2007), suggest that as far as epistemic success is concerned, dense communication networks fare rather badly.[19] That is, while I think a dense network should often qualify as "collaborative" it may well be the case that collaborative in this sense is not necessarily conducive to epistemic success. However, in quantitative terms, the ATLAS network is still not dense at all compared to the range of densities where Zollman observed a decline in epistemic success. These seem to be values above 0.3 (Zollman, 2007, Fig. 4). A more thorough comparison of my emprical study with the theoretical models by Zollman and others would go beyond the scope of this article. Let me just mention that such a comparison would also have to include a detailed assessment of how applicable Zollman's and other's models are to real world cases such as ATLAS. For instance, Zollman considered only networks with up to 10 nodes whereas in ATLAS (and many other real world networks) we are dealing with much more (749 nodes in case of ATLAS). Also, Zollman's models are intended to represent the interaction of individual mostly independent researchers rather than researcher inside a formal collaboration (cf. Sect. 4.1).[20]

For the calculation of the clustering coefficient (cf. Measure 2), it is custom to not only disregard the weights of the edges but also their direction (Newman, 2018, p. 186). There are some definitions of clustering and transitivity for *directed* networks as well but these are not often used and thus difficult to compare between networks. For our case of the ATLAS e-mails we obtained a value $C = 0.239$ for the (undirected) clustering as defined by (Newman, 2018, p. 184)

$$C = \frac{\text{(number of triangles)} \times 3}{\text{(number of connected triples)}}. \tag{4}$$

"Triangles" are groups of three nodes which are all connected to each other (by three edges in total). "Triples" are groups of three nodes which form a path of length two, i.e., for instance, X is connected to Y, and Y is connected to Z. Such a group is counted as a triple even if it is also a triangle, i.e. even if, in the example, X is also connected to Z. The factor 3 is there for combinatoric reasons: The existence of a triangle means that three triples are also a triangle. Thus, for instance, in the simple case of a network with three nodes forming a triangle, C should be 1, because every triple is also a triangle. But there are three triples and just one triangle. The factor 3 accounts for that.

---

[18] In an *undirected* network, the density would be twice this value because there can only be one edge between two nodes whereas in a directed network there can be two (one in each direction).

[19] Cf., however, critical discussions of these effects such as the one by Rosenstock et al. (2017).

[20] I thank Dunja Šešelja for reminding me of this difference.

For the ATLAS case we thus observe that in about 24% of the cases situations like the following occur: When a member X replies to the messages of another member Y and a member Z also replies to the messages of Y then X also replies to the messages of Z. Since we are dealing with an undirected network situations in which the direction of the reply is reversed should also be considered.

Unfortunately, in the Kiel study, the clustering coefficient has been determined using another definition following Watts and Strogatz (1998), denoted here as C_{WS}. This alternative definition derives from the concept of *local* clustering in the network. The local clustering coefficient of a node is defined by the number of pairs of nodes, connected to that particular node, which are also connected themselves (Newman, 2018, p. 186). The clustering of the entire network can then be defined as the mean value of the local clustering coefficients. In case of ATLAS, again represented as a simple undirected network, we obtain $C_{WS} = 0.319$, meaning that on average about 30% of a node's pairs of neighbours are connected, i.e. in about 30% of the cases two persons who have replied to the messages of another person A, or have received replies from A, have also sent at least one reply between them. Again, replies in reverse order should also be considered.

Both the Kiel study and our own preliminary study of the communication of the Linux kernel developers show values of the average local clustering (C_{WS}) of half of the amount that we find in ATLAS. Also the global clustering (C) in ATLAS is twice as high as the value we find in the Linux kernel mailing lists. Again, also on this measure, ATLAS is rather collaborative.

5.3 Degree distribution

In this section, we will determine and analyze how the "degrees" are distributed among the nodes of our ATLAS network. As mentioned above, this can be indicative of certain kinds of generative mechanisms of the network but also of whether nodes with high degrees "dominate" the network (cf. Newman, 2001b, p. 407). Most of such conclusions, however, hinge on the fact that the degree distribution can be fitted reasonably well by a *power-law*

$$n_k = ck^{-\alpha}, \tag{5}$$

where n_k is the number of nodes that have degree k, α is the parameter to be fitted, and c can be determined by normalizing the n_k's such that their sum equals the total number of nodes.

Like in many empirical networks, the in-degrees as well as the out-degrees in the ATLAS network are distributed non-uniformly. The distributions are proncouncedly "top-heavy" (or, synonymously, "heavy-tailed"), i.e. many nodes have rather low (in or out) degrees whereas only a few nodes have high (in or out) degrees. And, indeed, the degree distribution in the ATLAS network follows, at least roughly, what one would expect for a quantity that is theoretically distributed following a power-law.

However, to determine whether such a degree distribution indeed follows a power-law, and if so with what coefficient, is notoriously more difficult than it

may appear at first sight (Clauset et al., 2009; Newman, 2018, Sect. 10.4.1). Roughly, this is because there are often very few nodes with high degrees. In particular, this should indeed be the case, if the distribution of degrees follows a power-law. But it is a standard result of statistics, the statistical values about rare events tend to fluctuate significantly. Therefore, the standard fit procedures may be led astray when applied to potentially power-law distributed degrees in a network.

We performed several variants of fitting our empirical degree distribution and obtained significantly different results with each variant. The procedures are described in more detail in Appendix 1. With some of them we obtained values of α of almost exactly 1, both for the in- as well as the out-degrees. With other procedures we obtained values of α of 2.7 or 1.5 for both types of degrees. Yet another variant yielded $\alpha = 2.58$ for the in-degrees and $\alpha = 2.22$ for the out-degrees. In fact, an important variant supports the hypothesis that the degree distribution in our network does not follow a power-law at all but is better fitted by a "log-normal" distribution.

The first conclusion I draw from these considerations is that, given the current data, the degree distribution does not follow a power-law but rather a log-normal distribution. This then suggests that the network has formed by other mechanisms than a simple preferential attachment. Similar log-normal distributions have been found in citation analyses of physics papers. In these cases, it has been suggested that preferential attachment may still be at work in the formation of the network but that the preferential attachment is *non-linear* (Sheridan & Onodera, 2018).

The second conclusion is this: The data is relatively sparse, in particular for high degrees. This is also visible in the relatively large fluctuations in the histograms and scatter plots in the range of high degrees. Therefore, with more data (i.e. a larger sample of the communication over the ATLAS mailing lists) a more conclusive assessement of whether the degree distribution follows a power-law may be possible. If, for the moment, we assume that the distributions indeed follow a power-law, it is not so clear what the characteristic exponents are. They depend much on the fitting procedures. In particular, with some fitting procedures, the coefficients lie below 2 while with others they lie above 2. Remember that an exponent of 2 is a dividing line between networks where almost all edges are among the high degree nodes and networks where the edges are more equally distributed.

Also for our preliminary study of the Linux kernel mailing list we have not yet obtained sufficiently robust results. The Kiel study (Ebel et al., 2002) found $\alpha = 1.5$ for the in-degrees and $\alpha = 2.0$ for the out-degrees. So in this general purpose e-mail network, the degree distributions follow indeed a power-law. The exponent for the in-degrees is such that according to my Measure 3, the network would qualify as rather not collaborative and "dominated" by the nodes with high network degrees. With respect to the out-degrees of the network a conclusion is difficult to draw because the value of α lies precisely on the dividing line of 2 between more collaborative and less collaborative networks in the sense of my Measure 3.

5.4 Assortative mixing by degree

As mentioned before, there is another approach to determine the relative relevance of high or low degree nodes in the network. What we've just done in the previous section was asking whether the high degree nodes "dominate" the network in the sense that they accumulate most of the connections in the network. Now we will discuss whether the high degree nodes are mostly connected to other high degree nodes, and similarly for the low degree nodes. This would amount to an *assortative mixing by degree*. If, instead, we find *disassortative* mixing by degree, high degree nodes are often connected to low degree nodes and vice versa—or, in Newman's (2018, p. 210) words, "the gregarious people were hanging out with the hermits and vice versa."

For directed but unweighted networks, there exists an established definition for the correlation coefficient that measures assortativity by degree (Newman, 2003, Eq. 26):

$$r = \frac{\sum_i j_i k_i - M^{-1} \sum_i j_i \sum_{i'} k_{i'}}{\sqrt{\left[\sum_i j_i^2 - M^{-1} \left(\sum_i j_i\right)^2\right] \left[\sum_i k_i^2 - M^{-1} \left(\sum_i k_i\right)^2\right]}}, \tag{6}$$

where the sums are over all (directed) edges of the network (labeled by i or i'). j_i is the in-degree of the node at the start of the edge i; k_i is the out-degree at the end of the edge i.[21] For undirected (and unweighted) networks, one can use the same formula if one first replaces every (undirected) edge in the network by two directed ones in opposite directions to each other. For weighted networks, no definition seems to have established itself yet but proposals for it exist (Yuan et al., 2021). Below I will use the definition of the networkx library, which implements Newman's definition (or almost so, see below) and extends it to weighted networks and to different handling of in- and out-degrees. I will also propose an alternative definition which seems more suitable for weighted networks than Newman's and networkx' definition. For better comparison to known data of other networks, I will also calculate the assortativity coefficient for the representation of the ATLAS communication as an *undirected* network. For this, I will take our normal ATLAS network, which is directed and weighted, and just replace (weighted) directed edges by (weighted) undirected ones. If, in the original network, two nodes are connected by one (directed) edge the weight of the new (undirected) edge is equal to the weight of the original edge. If two nodes were connected by two edges (in opposite directions) the weights of the two original edges are added up to give the weight of the new (undirected) edge.

For directed networks, Newman, with his definition (Eq. 6), proposes to calculate the correlation coefficient of the out-degrees of the source node of a directed edge and the in-degrees of the target node. In the case of the ATLAS e-mail network, this would mean that we check wether a person A is more likely to reply to messages on the list by a person B if A *sends* a similar number of replies to the list as B *receives*. For the purposes of assessing how collaborative or hierarchical the collaboration is, this does not seem to be the most appropriate definition. Receiving and sending a similar

---

[21]   In fact, according to Newman (2003), one should evaluate Eq. 6 with "excess" or "remaining" degrees, instead of just the degrees. I will come back to this issue briefly below.

number of messages is not likely to indicate a similar status of two persons in the collaboration. Rather it is the fact that two persons *receive* a similar number of replies. Although there may certainly be exceptions, I imagine that in an idealized hierarchical or uncollaborative collaboration an "important" person *A* receives a lot of replies to her or his messages and that she or he would tend to reply only to messages of another person *B* which receives about the same amount of replies to her or his messages such that *A* judges her or him of equal importance or status. This means that we should adapt Newman's own reading of his equation and take k_i to be the *in*-degree of the source node rather than the out-degree (cf. Newman, 2003, p. 6).[22]

An appropriate definition of a degree correlation coefficient for our case thus seems to be a weighted "in-in" version of Newman's definition for a directed network. It turns out that the networkx library provides a function (by the name of "degrees_assortativity_coefficient") which can be applied to just such a case. The value we obtained is $r = -0.071$. Taken as an undirected network (as described above), we obtained $r = -0.073$. The network representing the ATLAS mailing list communication is thus slightly *disassortative* but very close to neutral: The information about the degree of a node (representing a person) does hardly imply any information about the degree of the nodes connected to it by an edge in the network. If anything at all, it implies that the degrees of these nodes tend to be different from the first node. For the behavior of the ATLAS members in their communication over mailing lists this means that the frequency with which they reply to each other's messages does not depend much on their similarity with respect to how many replies each other gets from others. If at all, they tend to reply to those who receive different amounts of replies. As indicated, I would consider such a behavior rather collaborative.

However, disassortativity is unusual for a "social" network (in a broad sense). For a physics coauthorship network, for instance, (Newman, 2002) found a rather strong *assortative* mixing ($r = 0.363$). In this line of research, it has been proposed that the reason for the assortative mixing of social networks can be explained, to a large extent, by the presence of sub-groups, or "communities," which are often absent in other "non-social" networks such as technological or biological ones (Newman & Park , 2003). The ATLAS communication network seems thus quite special. It definitely displays a community structure (see below) and also a considerable amount of clustering (see above), which both confirm its "social network" character. Nevertheless the network shows a slightly disassortative, rather than assortative mixing. Thus, in view of our attempt to assess the collaborativeness of the ATLAS collaboration, I would interpret this finding as indicating *collaborativeness in spite of group structures*.

At this stage, I can only conjecture what the reasons for this could be.[23] Newman and Park (2003) have found a similar exception to the expected value but in the opposite direction. The "board of directors" which they studied (cf. Newman, 2001) was

---

[22] To be precise, remember that Newman uses "excess" degrees in his definition of r.

[23] Given that also the Linux Kernel mailinglist is disassortative, a possible explanation for the disassortativity of the communication in ATLAS is the fact that the communication takes place on mailinglists rather than through direct messages. I thank Dirk Wintergrün for this observation. I plan to investigate this possibility further. Unfortunately, the Kiel study does not provide values for assortativity. In any case, the ATLAS communication still seems special because it is denser and tends to be more clustered than the communication at Kiel University and on the Linux Kernel mailinglist.

even more assortative (by degree) than can be explained by the group structure of the network. Accordingly, one is led to infer "true sociological or psychological effects" (Newman, 2003, p. 1). Similarly, we have good reasons to infer the presence of such effects—or rather mechanisms that bring about these effects—in the ATLAS collaboration. In contrast to the case of the board of directors, however, these mechanisms should lead to less rather than more assortativity. A charitable, but to my mind also plausible, interpretation would be that the ATLAS collaboration has a tradition, or actively encourages, their members to communicate with other members irrespective of their supposed position in some hierarchies or group structures. One could further ask where such a tradition could come from and why the ATLAS management, for instance, may encourage such a collaborative spirit. Again, I can only conjecture, but it seems again plausible to me that the kind of problems the ATLAS collaboration is trying to solve is best achieved by a neutral or slightly disassortative network.

There is, however, a worry one might have in calling neutral or even disassortative networks "collaborative." Results from network theory suggest that *assortative* networks, as opposed to neutral and disassortative ones, are more stable under the removal of nodes. As Newman (2003, p. 10) explains, in an assortative network, even "targeted" removal, i.e. the removal of high degree nodes, is not likely to affect the existence of paths across almost the whole network. This is so because, in an assortative network, the high degree nodes tend to be connected to each other and form a dense core of the network such that if one of them is removed, there still exist alternative paths connecting the rest of them, and other parts of the network are not affected by the removal. By contrast, in a disassortative network, the removal of a few high degree nodes can often cut off significant parts of the network. This is so because, in a disassortative network, high-degree nodes tend to be connected to "bundles" of lower degree nodes (Newman, 2018, Fig. 7.14b, p. 210). In this sense, assortativity implies more stability, and disassortativity implies instability. Thus, the removal of nodes affects the functioning of an *assortative* network (for instance for communication) much less than the functioning of a neutral or *disassortative* network.

This also means that, the ATLAS e-mail network, slightly disassortative as it is, tends to be unstable, in the above sense, which may not be optimal for communication purposes. A few persons (in particular those represented by high-degree nodes in the network) are essential to the functioning of the network. If they became absent, the information might not flow any more through all the relevant parts of the network, i.e. some persons might not get the information they need. In a sense this may mean that, on the one hand, ATLAS is not so collaborative after all, since it depends to some extent on the presence of some particular individuals. On the other hand, the instability under node removal might be explained by an efficient division of labor with very few redundant processes.

For completeness, Table 1 shows the values of alternative definitions of the assortativity coefficient as discussed above. Similar discussions as the one just made for the "in-in" coefficient will apply to these variants, but I will not do this in the present article. Of particular interest may be the "out-in" type of r because that is what corresponds to Newman's definition of r in a directed network (see above). I also add a column with the values we obtained with a further variant which takes into account the weights when calculating the "excess" degrees (see Footnote 21). Newman (2003)

Table 1 Various types of degree correlation coefficients (r) and their values

Type	Value	Value (weighted)
in–in	−0.071053	−0.085620
out–in	−0.071585	−0.087534
in–out	−0.070313	−0.085728
out–out	−0.071357	−0.088190
undir	−0.073226	−0.085946

The type column indicates which type of degree is considered at the sources and at the targets of the edges. For instance, "out-in" means that the out-degrees are considered at the sources and the in-degrees at the targets of the edges. The last line ("undir") shows the value of r if the e-mail communication is represented as an undirected (weighted) network. The third column lists the value we obtained when "weighted excess degrees" are taken into account

calculates the excess degrees only by subtracting 1 from the usual degree of the nodes and not the weight of the edge connecting two nodes in the calculation of r.

How does the ATLAS collaboration fare on this measure compared to the Linux kernel developers and the e-mail users at the University of Kiel? Unfortunately, the Kiel study does not contain a calculation of the assortativity coefficient. Our own preliminary study of the Linux kernel mailing list yields a value of $r = -0.126$ (of the in-in variant), i.e. slightly more negative but of the same order of magnitude.

5.5 Community structure

So far, we analyzed global measures of the network such as the density and clustering coefficient. As a further component of our analysis of ATLAS' communication structure we now turn to the question of what substructures the network shows. Are there subgroups of the collaboration that communicate more among themselves than they do with members of other subgroups?

This is a standard problem in network theory and the most widely used solution is the "Louvain" algorithm of "community detection" (Blondel et al., 2008). The algorithm is only applicable to undirected networks but the weights of the edges can be taken into account and we will do so. That is, for the following analysis we will convert our directed network in an undirected (but weighted) one. This just means ignoring the directions of the existing edges. Also, the algorithm makes most sense when applied to the largest connected component of the network. The largest connected component almost always contains over 90% of all nodes of a real-world network (Newman, 2018, p. 306), which is also our case. If we applied the algorithm to the whole network, we would just get additional small unconnected communities, which are not of much interest with regards to the major communication activities of the collaboration.

If we apply this algorithm to the largest component of our network we get 6 communities (numbered arbitrarily from 0 to 5). The smallest one contains about 50 nodes;

the largest about 200 nodes.[24] Each community features about 1 to 5 nodes that stand out by their betweenness-centrality, see Fig. 3.

5.5.1 The mailing list profile vector

What are these "communities"? In what respects are the members inside a community similar to each other? What is the characteristic feature of each community? To answer these questions, we introduce a *mailing list profile vector* (m.p.v., for short) of a person. The m.p.v. has as many components as there are mailing lists (or HyperNews groups) in our collection of e-mails. In its unnormalized form, each component of the vector is the number of messages the person under consideration has sent to the list plus the number of replies the person has received to posts to that list. It's basically a kind of "degree", i.e. the total weight of edges that a node (representing the person under consideration) would have if we only took into account messages of the single list under consideration.

By means of the m.p.v.'s, we can measure how similar the activity of two persons is with respect to which mailing list they are most active on. If two persons write to the same lists in similar proportions, or receive messages in similar proportions over the same lists, their m.p.v.'s will point in the same direction in the vector space. If we normalize the m.p.v.'s to one, they will also lie close together, i.e. the distance between them will be small. If two persons write to or get replies over mostly different lists, their normalized m.p.v.'s will be separated by a considerable distance. We thus created a matrix of distances between the normalized m.p.v.'s of every pair of persons. If one groups the persons by the community the Louvain algorithm has assigned to it a distinct pattern emerges, see Fig. 4.

The figure shows that the m.p.v.'s tend to be similar inside the communities. The distances between their m.p.v.'s tend to be small (represented as shades of orange in the figure). In contrast, between persons who are not assigned the same community, the distances between their normalized m.p.v.'s tend to be large (represented as shades of purple in the figure).

This quite distinct visual pattern also bears out in some quantitative measures albeit perhaps not as distinctly as one might expect. The mean distance of the m.p.v.'s of all possible pairs of nodes is given by

$$\mu = \frac{1}{M} \sum_{ij} d_{ij}, \qquad (7)$$

---

[24] The Louvain algorithm does not yield unique results because its output depends on the order in which the nodes are processed in the search for a modularity optimizing partition. In our calculations, we used the default behavior of the "python-louvain" package, which chooses the node order at random each time the algorithm is called. In about 98% of calls of the algorithm, we obtain 6 modularity classes; in the other cases 5 or 7. In the relevant case of 6 classes, the sizes vary with an approximate standard deviation of 4% relative to the mean size of a given class. Also, the members of the classes remain mostly constant across different calls of the algorithm. About 95% of the members of a given class according to the first run were in the same class in 90% of the subsequent runs (checked with 10,000 runs).
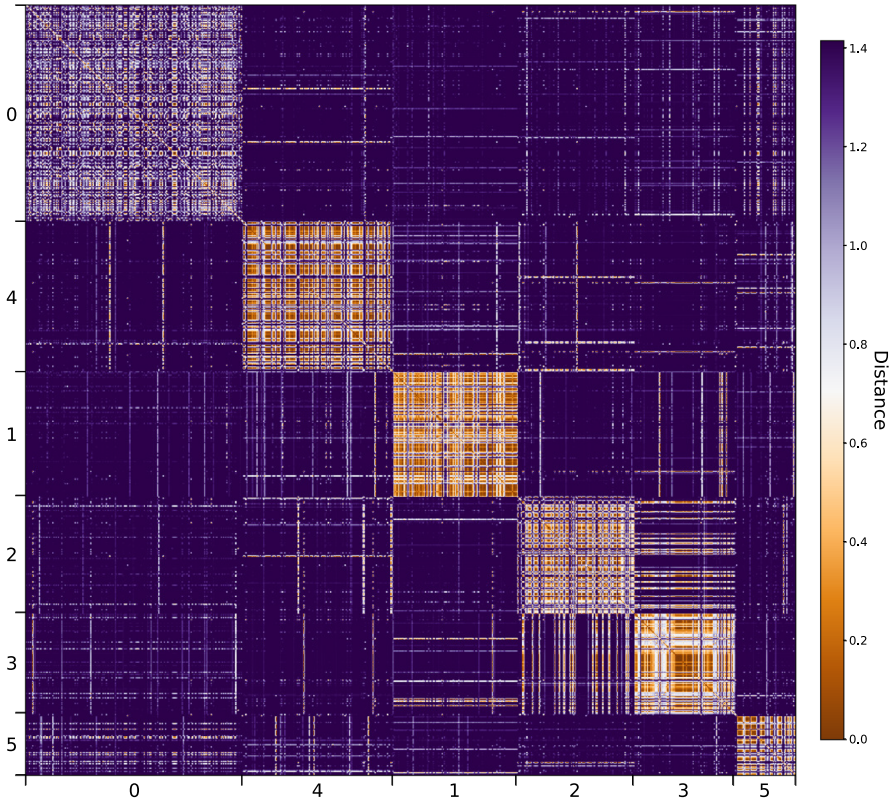
Fig. 4 "Heatmap" of distances between every pair of nodes in the largest component of the network. The figure is symmetric with respect to reflection on its diagonal. Each pair of nodes appears twice in the figure, once above and once below the diagonal according to the symmetry. The nodes are grouped by "communities." Inside the communities, the order of the nodes is random. The communities (labeled arbitrarily from 0 to 5) are sorted by size

where $M = \frac{1}{2}n(n-1)$, n being the number of nodes, is the number of all possible ordered pairs of (distinct) nodes. d_{ij} is the distance between the ith and the jth node in the list of nodes of the network. The sum should actually run over only unequal values of i and j, but the distance of m.p.v.'s of a nodes to itself is zero. Also, in M as well as in d each pair is considered twice (with the orders of the nodes exchanged). A factor of $1/2$ that one might want to add would therefore cancel out. For our network we get $\mu = 1.28$.

Inside the communities, i.e. taking into account only pairs of nodes that lie in the same community, the mean distance can be calculated thus:

$$\mu_c = \frac{1}{M_c} \sum_{ij} d_{ij} \delta_{g(i)g(j)}, \tag{8}$$

where M_c is the number of pairs of nodes which lie in the same community, and the function g assigns each node (indexed by i or j) its modularity class (= "community").

δ is the Kronecker delta. For our network we get $\mu_c = 0.92$, which is considerably lower than the mean over all.

The mean distance is even lower among all the nodes that are connected by an edge. This is determined by

$$\mu_e = \frac{1}{2m} \sum_{ij} d_{ij} A_{ij}, \tag{9}$$

where m is the total number of edges in the largest component of the network and A is the adjacency matrix of the largest component.[25] A factor of $1/2$ is in order here because the distances associated with one of the edges appears twice in the matrix d. For our network we get $\mu_e = 0.68$.[26]

By contrast, the mean value is again considerably higher if we distributed the edges at random. The idea and method of calculation for this scenario resembles closely the definition of *modularity* (Newman, 2018, p. 204), which is at the basis of the Louvain algorithm of community detection. The modularity measures how much more edges there are between nodes of the same type in the actual network compared to the number of edges between nodes of the same type if we distributed the edges randomly but consistent with the degrees of the nodes given by the empirical data. Here we ask about the mean distance among nodes connected randomly in this sense. The relevant quantity is given by

$$\mu_r = \frac{1}{4m^2} \sum_{ij} d_{ij} k_i k_j, \tag{10}$$

where k_i is the degree of the ith node. The additional factor of $1/2m$ comes from the double counting of edges when one calculates their probabilities from the degrees of the nodes attached to it. For our network we get $\mu_r = 1.20$.[27]

The mean distance of the m.p.v.'s in the network with edges distributed at random is thus almost twice as high as with the actual edges. This amounts to a kind of assortative mixing of properties that are represented by a vector. This is an understudied topic in network theory (Newman, 2018, p. 209). The present case thus provides a rare instance of such a mixing pattern and a proposal of how to quantitatively deal with it.

Moreover, we can see that the modularity maximizing partition (the "communities" according to the Louvain algorithm) also tends to minimize the distances between the m.p.v.'s of the nodes. Although this effect is less pronounced as with considering only connected nodes (cf. μ_e), we may still conclude that the modularity maximizing partition partitions the nodes into classes with similar m.p.v.'s. Each modularity class

---

[25] The element A_{ij} of the adjacency matrix of an undirected network is equal to the weight of the edge connecting node i to node j. If there is no edge between node i and node j A_{ij} is zero. In an unweighted network, $A_{ij} = 1$ if there is an edge between node i and node j, and $A_{ij} = 0$ if there is none.

[26] In this calculation, we have not taken into account the weights of the edges. If we do, we get $\mu_e = 0.56$.

[27] Again, in this calculation, we have not taken into account the weights of the edges. If we do, we get $\mu_r = 1.12$. The following conclusion, that μ_r is about twice as high as μ_e thus remains valid also when taking into account the weights of the edges, cf. Footnote 26.
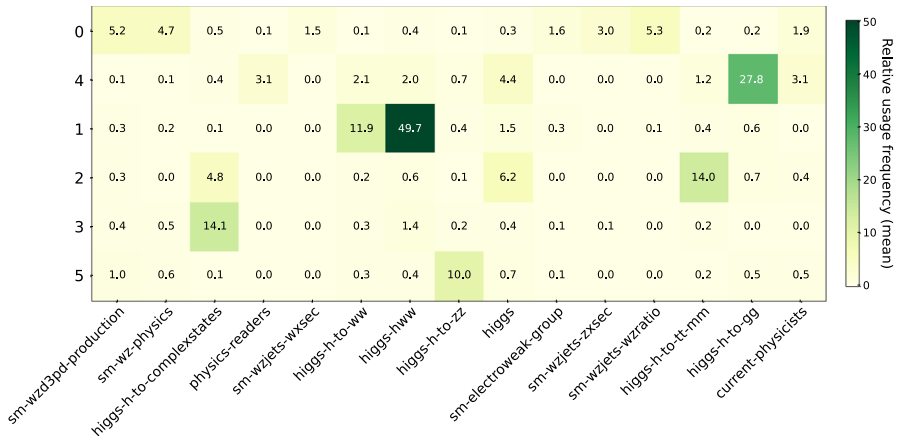
Fig. 5 Heat map of mean mailing list profile vector per community (corresponding to the rows labeled from 0 to 5). Only the mailing lists with a mean value of at least one message over the whole analysis period are taken into account. The communities are sorted by size, i.e. number of nodes in the class. Only the distinct parts of the e-mail addresses are used as labels

seems to have, roughly, a unique typical m.p.v. The ensuing question then is what that typical m.p.v. looks like.

5.5.2 The typical profile of the communities

To determine the typical mailing list profile vector for a particular modularity class (or "community") we simply take the mean vector, that is the sum of all m.p.v.'s of nodes in the same community, divided by the number of nodes in that community. For this calculation, we use the absolute m.p.v.'s, not the normalized ones as we did for the calcuation of the distances between them. As mentioned (see Sect. 2), quite a number of the 175 mailing lists (or HyperNews groups) in our collection have very little traffic. So most of them can be neglected; they would only clutter the following analysis.

More concretely, we will reduce the mean m.p.v. of a community to only those of its components which in at least one of the communities is greater than one. In other words, in the following, we take into account only those mailing lists over which (on average in a given community over the whole analysis period) at least one message was sent. This leaves us again with the 15 top mailing lists of Table 1. In Fig. 5, the components of the m.p.v. corresponding to these 15 mailing lists are represented as a "heatmap" for each of the six modularity classes.

One can see, in the figure, that in 5 of the six modularity classes, there is a relatively distinct maximal component of the mean m.p.v. For instance, the mean m.p.v. of community 1 gets its most important contribution from the mailing list phys-higgs-hww. This is, in fact, the "forum" of the analysis subgroup dedicated to the decay of the Higgs boson into two W boson, abbreviated as HSG3.[28] Also the remaining four

---

[28]   For a list of mailing lists serving as "forums" for the Higgs analysis subgroups, see https://espace.cern.ch/atlas-forums/Lists/Atlas%20forums/By%20category.aspx (retrieved 2021-09-15).

Table 2 Most important mailing lists per modularity class

Class	Mailing List	Subgroup
0	wzjets-wzratio, wzd3pd-production, wz-physics	WZ derived physics data
4	physics-higgs-h-to-gg	H to Gamma-Gamma (HSG1)
1	phys-higgs-hww	H to WW (HSG3)
2	physics-higgs-h-to-tt-mm	H to Tau-Tau and Mu-Mu (HSG4)
3	physics-higgs-h-to-complexstates	H to Complex Final States (HSG5)
5	physics-higgs-h-to-zz	H to ZZ (HSG2)

For Class 0, which has no significantly distinct maximum, abbreviations of the three most important lists are shown. These lists are also not listed as a "forum" of a dedicated subgroup. Only a short characterization is therefore given. The classes are sorted by decreasing size, i.e. number of nodes (largest class first)

classes with a distinct maximum component can thus be associated with one of the "Higgs subgroups", see Table 2.

The components of the mean m.p.v. of community 0 show much less variation. Still, three major contributions can be distinguished coming from the mailing lists phys-sm-wzjets-wzratio, phys-sm-wz-physics and phys-sm-wzd3pd-production. These mailing lists are not listed as a "forum" in the list mentioned in Footnote 28. But from their names, one can guess that they are dedicated to discussions of physics analysis and data production related to the W and Z boson. This community does thus not seem immediately involved in the analysis of Higgs decays but rather provide the necessary background data.

An exact characterization of the tasks of each of these groups would require a more detailed analysis—by close reading the e-mail messages on the corresponding lists, by consulting other types of documents, interviewing involved physicists etc.—but it seems fair to say that communities detected by maximizing the modularity of the network (using the Louvain algorithm) correspond to some of the most important subgroups of the ATLAS collaboration in the period covered by our analysis.

We also saw that the m.p.v.'s of members of different modularity classes are often very different (see Fig. 4) as are the mean m.p.v.'s of the communities (see Fig. 5). So there is, over all, very little e-mail communication *across* the various "communities." In view of our main question of how to characterize the type of collaboration inside ATLAS, this suggests that the collaboration occurs mostly by a clear division of labor rather than through an attempt to involve a maximum number of researchers in a maximum number of different tasks.

However, in order to be successful, the collaboration must also pull together the results from the different subactivities and aggregate them to general findings, endorsed by the whole collaboration, such as the "Observation of a New Particle" (the Higgs boson) in 2012. Parts of that process have been analyzed, through close reading documents from our collection of digital sources, by (Graßhoff, 2012; Schefer, 2012) and characterized as a collaborative process of "quality management" of scientific results. Here we take another approach and attempt to identify signs of such a process in the collaboration's communication structure.

5.5.3 Octopus nodes

Standard "centrality" measures of nodes in a network can give valuable insights into the role a person (or whatever is represented by the node) has in the process or structure represented by the network. As regards the process of pulling partial results together in a scientific collaboration represented by its communication network, a good candidate measure is the "betweenness" centrality of a node (see also Sect. 4). As mentioned, nodes with high betweenness represent "brokers" through which many shortest paths pass. They are therefore also likely to be the links between communities. However, this need not be so. In principle, a node can also get a high betweenness because of its broker role *inside* a community.

We are, therefore, looking for further measures which could capture the role of a connector *between* the communities (as detected, for instance, by the Louvain algorithm as in our case). Moreover, in view of the process of aggregating partial results of the collaboration, we attempt to identify in particular those nodes which connect not just two communities but many. Such a node should have a "tentacle" into each of the communities; let's call them *octopus nodes*.

The best measure we could come up with is constructed as follows. For a given node n_i we categorized the replies of the person represented by it according to the community of the node representing the person that sent the original message. This yields a list of (in our case) six integers, each indicating how often a person replied to a message of a person in community 0, community 1 etc. up to community 5. The sum of the values in this list is the total number of replies the person has sent. It corresponds to the out-degree (number of attached edges) of the corresponding node in the network. This is a measure of how "active" a person was in the e-mail communication. The main characteristics of an octopus node, however, is not just that the corresponding person writes a lot of messages but that it tends to write *equally* into all communities. We are thus looking for a way to measure the variation in the number of replies to the different communities.

Statistical measures for such a variation of "nominal data" are not very prominent but a couple of them have been proposed and discussed in the literature (Gibbs & Poston, 1975; Mueller et al., 1977; Kvålseth, 1988). One such measure is the most natural to choose because it is, to my mind, largely continuous with measures of variation of better known cases such as the distribution of a random variable; let's call it *qualitative variation* following Mueller et al.. Our version of qualitative variation v of the values x_i in a category i is given by

$$v = 1 - \frac{\sum_{i=1}^{I} \left(x_i - \frac{N}{I} \right)^2}{\frac{N^2}{I}(I-1)}, \tag{11}$$

where I is the number of categories and $N = \sum_i x_i$. In our case, $I = 6$ and N is also the degree of the node for which we are calculating v. The basic idea behind the measure v is to take the mean of the x_i, which is equal to N/I, and then sum up the deviations of the x_i from that mean. All the other factors and the difference to 1 are there to have $0 \le v \le 1$, and $v = 1$ if all x_i are equal to the mean.

In our case, if we applied this measure in isolation, a person who writes just 6 messages but each in a different community would get the maximal value of v, which is 1. In some sense, this is an "octopus" but not a very active one. For people pulling together partial results in a scientific collaboration we rather expect that they have a large value of v but also a large degree in the network sense of a high number of edges attached to the node representing it. In this combination, we will identify persons who write a lot of replies but also tend to write them in equal proportions to all communities.

A systematic discussion of which persons appear to be "central" on different network measures such as degree, betweenness and our tailor-made "octopus" measure would go beyond the scope of the present article. But a combination of degree of a node and its value of v turns out to be very appropriate to identify octopus nodes in our case. We first sorted all the nodes of our network by their degree. To exclude "lazy octopuses" we selected only the top 20 nodes of this ranking, 20 being a pragmatic choice to get the "few" most active ones. These top 20, we sorted by their value of v and the best score is achieved by one of the two convenors of the ATLAS Higgs group at the time of the Higgs discovery.[29]

The two Higgs convenors were the main responsibles for coordinating the work of the different subgroups dedicated to the analysis of the various decays of the Higgs boson (cf. Graßhoff, 2012, p. 34; Marti & Schefer, 2012, p. 63). Our analysis shows that one way to fulfill this role is to be an "active octopus" as defined by our network-theoretical measure, i.e. by communicating a great deal and in roughly equal amounts with all the relevant subgroups.

This result as well as the match of the communities as detected by the Louvain algorithm with the most important subgroups of ATLAS (see Sect. 5.5.2) gives us additional reasons to believe in the representativeness of the MetaATLAS collection. It would be unlikely that an unrepresentative sample reproduces these features (as well as the peak in communication activity mentioned in Sect. 2.2).

6 Summary and outlook

In the study presented here, I subjected a collection of about 24,000 e-mail messages from the internal communication of the ATLAS Collaboration at CERN to a network analysis by digital means. The digital tools employed, and to some extent further developed here, should also be applicable to similar collections of born-digital sources. As a test case we ran our algorithms on a collection of freely available messages of the Linux kernel developers and obtained, in most cases, sensible and interpretable results.

In order to characterize the ATLAS communication through an analysis of its communication structure, I proposed four network measures to assess, in a quantifiable way, how collaborative the collaboration really is. On most of these measures, the ATLAS collaboration turns out to be collaborative, indeed. The communcation network of ATLAS is relatively dense, clustered and not assortative by degree.

---

[29] Our data contains the full names of most of the persons represented by a node in our network. We then found the official role of the person under consideration here by a websearch.

Through the use of mathematical vector, representing a person's usage profile of the different ATLAS mailing lists, and standard procedures of community detection in networks, we were able to identify 6 major subgroups of the ATLAS collaboration at the time of the discovery of the Higgs boson. A tailor-made "octopus" measure, combined with the standard measure of network degree, identified one of the persons which, according to their official role, was in charge of coordinating the "divided" labor of the several Higgs analysis subgroups.

The presence of a pronounced group structure, relatively high density and considerable clustering and disassortivity by degree, as we have found them in ATLAS, is an unusual combination of network characteristics. In particular, group structure and clustering usually leads to an assortativity (rather than disassortativity) by degree, which are all typical for "social" networks. There must be some mechanisms at work, in the ATLAS communication, that make this special combination possible.

The MetaATLAS collection, which forms the basis of the present study, includes, among other things, a considerable amount of messages from "one of the main ATLAS collaborative tools" (see Sect. 2.1). To assess in more detail how representative the collection is and how robust the results we obtained are data over a longer period of time and from additional communication channels would be desirable. Also more selective close reading of the communication, and interviews with ATLAS members would be helpful. But with its methodological approach and its first results about the communication inside a major scientific collaboration, the present article takes initial steps towards making empirical studies of scientific collaborations comparable to abstract models in theoretical social epistemology.

Appendix A Identification of e-mail addresses

We identified the ATLAS members by their full name.[30] These turned out to be unique in our collection. To determine the full names, we used the author list of ATLAS' "Higgs discovery paper" (ATLAS Collaboration, 2012) as a pool of potential last names and initials.

In case of identical last name and first initial we determined, by searching the internet, the most probable full name (using the CERN website, institutional websites of potential ATLAS authors, authorship names in other publications etc.). We then checked, in our collection of messages, whether parts of the e-mail addresses (before the "@"-sign) matched one of the last names of the author list. We then took the most complete name as a potential name associated with the e-mail address in question. For some special cases, which did not seem to contain first and last names, we used the complete "From"-field of the e-mail message.

Another special case is the e-mail address "atlas.secretariat." It is used by several persons but we registered "ATLAS Secretariat" as the single most complete name for it. There is thus certainly some error in the exact identification of the persons active on the lists. However, in our collection only about 1% of all messages come from this e-mail address. We therefore take potential errors due to a lack of discrimination resulting from this joint e-mail address to be negligible for our further analysis. For the e-mail addresses that ended up with more than one potential most complete name, we eliminated all but one alternative, using the second initial or other information, mostly available on the internet. From this list of pairs of e-mail addresses and most complete names, we then created a list of unique most complete names with their associated e-mail addresses.
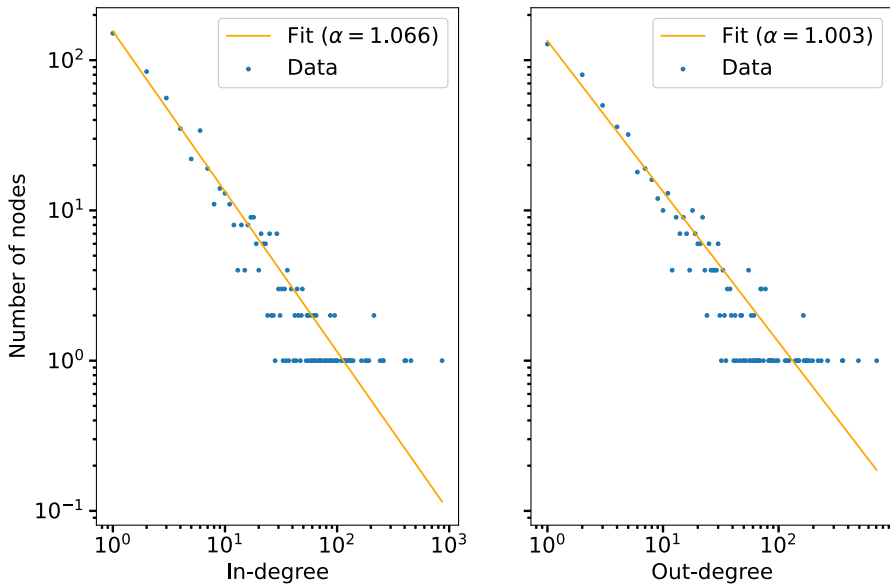
Fitting the degree distribution

Like in many empirical networks, the in-degrees as well as the out-degrees of the ATLAS network are distributed non-uniformly. The distributions are proncouncedly "heavy-tailed", i.e. many nodes have rather low (in or out) degrees whereas only a few nodes have high (in or out) degrees. And, indeed, the distribution follows, at least roughly, what one would expect for a quantity that is theoretically distributed following a power-law, see Fig. 6.

The figure also shows the result of a straightforward least-squares fit to the empirical data by a power function of the form

$$n_k = ck^{-\alpha}, \tag{12}$$

---

[30] This part of the analysis has mostly been conducted by Beate Federau, student assistant at TU Berlin in 2019.

Fig. 6 Scatter plot and fit of in- and out-degrees

where n_k is the number of nodes that have degree k, and α and c are the parameters to be fitted.[31] Of particular interest for the interpretation of the degree distribution as an indicator for "collaborativeness" is the parameter α. For the distribution of in-degrees, the fit by this method yields $\alpha = 1.066$; for the out-degrees we get $\alpha = 1.003$.[32]

However, to determine whether such a degree distribution indeed follows a power-law, and if so with what coefficient, is notoriously more difficult than it may appear at first sight (Clauset et al., 2009; Newman, 2018, Sect. 10.4.1). Roughly, this is because the empirical data for high degrees are very sparse. There are only a few nodes with high degrees. Therefore, the standard fit procedures may be led astray by the more numerous data points for low degrees.

In addition to the straightworward fit just performed, we will therefore also apply two alternative methods, which have been argued to be more adequate in cases such as the present one. We will first apply the method of *logarithmic binning* (Newman, 2018, p. 320). For a distribution that follows a power-law, the resulting histogram should visually display a "straight-line behavior" (Newman, 2018, p. 321). While we're at it, we will also perform a linear least-squares fit to the tips of the logarithmic histogram. Although this is not considered "good practice" (Newman, 2018, p. 324) it will show to what extent the methodological difficulties in determining power-law behavior manifest themselves in our case or not. For a second attempt to detect a power-law and its characteristic exponent we will use a dedicated python library, called "powerlaw" (Alstott et al., 2014), which implements most of the recommendations of

---

[31] One could also fit only α and determine c by requiring $\sum_k n_k$ to equal the total number of nodes.

[32] We performed the fit using the "curve fit" function of the "scipy optimize" library for python, see https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html (accessed 2021-08-12).
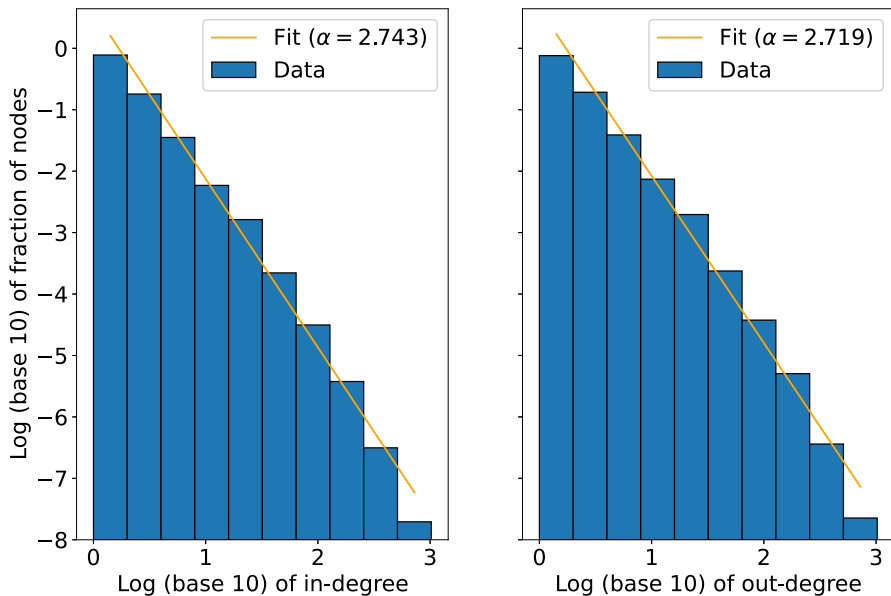
Fig. 7 Histogram with logarithmic bins of in- and out-degrees and a linear fit to the tips of the bin centers

Clauset et al. (2009), which, to my knowledge, still represents the state-of-the-art for such a task.

A fairly good approach to detect a possible power-law behavior of the degree distribution is by creating a histogram of degrees and their relative frequency. In order to deal with the fact that only a few nodes have a high degree in the empirical sample, which leads to large statistical uncertainties, it is best to bin the data logarithmically. That is, instead of bin of equal lengths (on a linear scale), one should uses bins of equal lengths on a *logarithmic* scale. In order not to overestimate the relative frequencies of nodes with high degrees with such a binning, one should devide the relative frequency by the respective widths of the bins.

The resulting histograms are shown as "Data" in Fig. 7.

The histogram for the in-degrees as well as that for the out-degrees shows indeed a rather distinct "straight-line behavior" as does Newman's example of the Internet (Newman, 2018, pp. 320–321). This is an indication that the empirical distribution could result from a theoretical power-law distribution, but it is by no means a conclusive argument. Other theoretical heavy-tailed distributions could manifest themselves in similar ways empirically. Assuming, for a moment, that the distribution indeed comes from a power-law we fitted the center of the histograms (on the logarithmic scale) with a polynomial function of degree one (in other words, a "straight line").[33] This procedure yields an exponent α of about 2.7 for the in- as well as for the out-degree distribution. We will discuss possible interpretation of these quantities with regard to

---

[33] To implement this, we used the "fit" method of the "Polynomial" class of the numpy libary for python, see https://numpy.org/doc/stable/reference/generated/numpy.polynomial.polynomial.Polynomial.html (accessed 2021-08-12).

the "collaborativeness" of the collaboration below, keeping in mind that such a fit is usually not considered "good practice."

Before trying to interpret these quantities, however, let us investigate a little bit more whether the degree distributions follow a power-law at all. We delegate this task to a large extent to the capabilities of the aforementioned "powerlaw" library in order to apply the current state-of-the-art fitting procedures.

The best fits according to these procedures yield exponents $\alpha = 2.58$ for the in-degrees and $\alpha = 2.22$ for the out-degrees. However, the optimal fit interval as determined by this method only covers about 4 per cent of the nodes in the case of the in-degrees and only about 9 per cent of the nodes in the case of the out-degrees. Thus, the power-law package could determine a power-law behavior only in the extreme tails of the distributions.

In order to assess how well a power-law can fit the degree distribution for the majority of nodes we set the degree from which the distribution should be fitted to 10. Like this, about 30 per cent of the nodes are covered by the fit. With these modifications, we get $\alpha = 1.56$ for the in-degrees as well as for the out-degrees. However, using the powerlaw library, it was also relatively straightforward to see that the *log-normal distribution* is a better fit to this range of the degree distribution than a power-law.

References

Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: A Python package for analysis of heavy-tailed distributions. *PLoS ONE, 9*(1), e85777. https://doi.org/10.1371/journal.pone.0085777

ATLAS Collaboration. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Physics Letters B, 716 (1), 1–29. https://doi.org/10.1016/j.physletb.2012.08.020

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Burt, R. S. (1995). *Structural holes: The social structure of competition (1. paperback ed.).* Cambridge: Harvard UnivPress.

Clauset, A., Shalizi, C. R., & Newman, M. (2009). Power-law distributions in empirical data. *SIAM Review, 51*(4), 661–703. https://doi.org/10.1137/070710111

CMS Collaboration (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Physics Letters B, 716 (1), 30–61. https://doi.org/10.1016/j.physletb.2012.08.021

Collins, H. (2017). *Gravity's kiss-the detection of gravitational waves.* Cambridge: The MIT Press.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*, 292–306. https://doi.org/10.1002/asi.4630270505

Düring, M. (2016). How reliable are centrality measures for data collected from fragmentary and heterogeneous historical sources? A case study. *The Connected Past: Challenges to Network Studies in Archaeology and History, 8*, 85–102.

Ebel, H., Mielsch, L. I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E, 66*(3), 035103. https://doi.org/10.1103/PhysRevE.66.035103

Englert, F., & Brout, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters, 13*(9), 321–323. https://doi.org/10.1103/PhysRevLett.13.321

Galison, P. (2003). The collective author. In M. Biagioli & P. Galison (Eds.), *Scientific authorship: Credit and intellectual property in science* (pp. 325–355). New York: Routledge.

Gibbs, J. P., & Poston, D. L. (1975). The division of labor: Conceptualization and related measures. *Social Forces, 53*(3), 468–476. https://doi.org/10.2307/2576589

Giere, R. N. (2007). Distributed cognition without distributed knowing. *Social Epistemology, 21*(3), 313–320. https://doi.org/10.1080/02691720701674197

Graßhoff, G. (2012). Die ATLAS-Collaboration als wissenschaftlicher Autor. In G. Graßhoff & A. Wüthrich (Eds.), *MetaATLAS: Studien zur Generierung, Validierung und Kommunikation von Wissen in einer modernen Forschungskollaboration. Bern: Bern Studies in the History and Philosophy of Science*.

Graßhoff, G., & Wüthrich, A. (Eds.). (2012). *MetaATLAS: Studien zur Generierung, Validierung und Kommunikation von Wissen in einer modernen Forschungskollaboration*. Bern: Bern Studies in the History and Philosophy of Science.

Hagberg, A.A., Schult, D.A., & Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. G. Varoquaux, T. Vaught, & J. Millman (Eds.), Proceedings of the 7th python in science conference (pp. 11-15). Pasadena.

Herfeld, C., & Doehne, M. (2019). The diffusion of scientific innovations: A role typology. *Studies in History and Philosophy of Science Part A*, 77, 64–80. https://doi.org/10.1016/j.shpsa.2017.12.001

Higgs, P. W. (1964). Broken symmetries, massless particles and gauge fields. *Physics Letters, 12*, 132–133. https://doi.org/10.1016/0031-9163(64)91136-9

Kirschenbaum, M. (2013). The .txtual condition: Digital humanities, born-digital archives, and the future literary. *Digital Humanities Quarterly, 7*(1).

Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy, 87*(1), 5–22. https://doi.org/10.2307/2026796

Knorr-Cetina, K. (1995). How superorganisms change: Consensus formation and the social ontology of high-energy physics experiments. *Social Studies of Science, 25*(1), 119–147.

Knorr-Cetina, K. (1999). *Epistemic cultures. How the sciences make knowledge*. Cambridge: Harvard University Press.

Kvålseth, T. O. (1988). Measuring variation for nominal data. *Bulletin of the Psychonomic Society, 26*(5), 433–436. https://doi.org/10.3758/BF03334906

Laubichler, M. D., Maienschein, J., & Renn, J. (2019). Computational History of Knowledge: Challenges and Opportunities. *Isis, 110*(3), 502–512. https://doi.org/10.1086/705544

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.

Marti, B., & Schefer, M. (2012). Das ATLAS-Experiment. In G. Graßhoff & A. Wüthrich (Eds.), *MetaATLAS: Studien zur Generierung, Validierung und Kommunikation von Wissen in einer modernen Forschungskollaboration* (pp. 41–82). Bern: Bern Studies in the History and Philosophy of Science.

Milgram, S. (1967). The small world problem. *Psychology Today, 2*(1), 60–67.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science, 298*, 824–827. https://doi.org/10.1126/science.298.5594.824

Mueller, J. H., Schuessler, K. F., & Castner, H. L. (1977). *Statistical reasoning in sociology* (3rd ed.). Boston: Mifflin.

Newman, M. (2001a). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*(1), 016132. https://doi.org/10.1103/PhysRevE.64.016132

Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, 98*(2), 404–409. https://doi.org/10.1073/pnas.98.2.404

Newman, M. (2002). Assortative mixing in networks. *Physical Review Letters, 89*(20), 208701. https://doi.org/10.1103/PhysRevLett.89.208701

Newman, M. (2003). Mixing patterns in networks. *Physical Review E, 67*(2), 026126. https://doi.org/10.1103/PhysRevE.67.026126

Newman, M. (2005). Power laws, pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Newman, M. (2018). *Networks: An introduction* (2nd ed.). Oxford: Oxford University Press.

Newman, M., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E, 68*(3), 036122. https://doi.org/10.1103/PhysRevE.68.036122

Newman, M., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E, 64*(2), 026118. https://doi.org/10.1103/PhysRevE.64.026118

Rosenstock, S., Bruner, J., & O'Connor, C. (2017). In epistemic networks, is less really more? *Philosophy of Science, 84*(2), 234–252. https://doi.org/10.1086/690717

Rosenzweig, R. (2003). Scarcity or abundance? Preserving the Past in the Digital Era. *Amercian Historical Review, 108*, 735–762.

Schefer, M. (2012). Der Wissensbegriff am Limit? Kollaborative Wissensgenerierung im Grossforschungsprojekt ATLAS am CERN. In G. Graßhoff & A. Wührich (Eds.), *MetaATLAS: Studien zur Generierung, Validierung und Kommunikation von Wissen in einer modernen Forschungskollaboration* (pp. 83–212). Bern: Bern Studies in the History and Philosophy of Science.

Sheridan, P., & Onodera, T. (2018). A preferential attachment paradox: how preferential attachment combines with growth to produce networks with log-normal in-degree distributions. *Scientific Reports, 8*, 2811. https://doi.org/10.1038/s41598-018-21133-2

Shrum, W., Genuth, J., & Chompalov, I. (2007). *Structures of scientific collaboration*. Cambridge: Massachusetts Institute of Technology.

Staley, K. (2004). *The evidence for the top quark: Objectivity and bias in collaborative experimentation*. Cambridge: Cambridge University Press.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry, 32*(4), 425–443. https://doi.org/10.2307/2786545

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442. https://doi.org/10.1038/30918

Wüthrich, A. (2017). Discovering and "rediscovering" the W boso. *Annalen der Physik, 5*, 529. https://doi.org/10.1002/andp.201700101

Yuan, Y., Yan, J., Yan, J., & Zhang, P. (2021). Assortativity measures for weighted and directed networks. *Journal of Complex Networks, 2*, 9. https://doi.org/10.1093/comnet/cnab017

Zollman, K. (2007). The communication structure of epistemic communities. *Philosophy of Science, 74*(5), 574–587.