Applied Network Science

RESEARCH

Open Access

Large language models recover scientific collaboration networks from text

Rathin Jeyaram[1,2], Robert N Ward[2,3,4] and Marc Santolini[1,2,3*]

*Correspondence:
Marc Santolini
marc.santolini@cri-paris.org
[1]Université Paris Cité, Inserm, System Engineering and Evolution Dynamics, Paris F-75004, France
[2]Learning Planet Institute, Research Unit Learning Transitions (UR LT, joint unit with CY Cergy Paris University), Paris F-75004, France
[3]School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA
[4]Department of Bioengineering, Stanford University, Stanford, CA, USA

Abstract

Science is a collaborative endeavor. Yet, unlike co-authorship, interactions within and across teams are seldom reported in a structured way, making them hard to study at scale. We show that Large Language Models (LLMs) can solve this problem, vastly improving the efficiency and quality of network data collection. Our approach iteratively applies filtering with few-shot learning, allowing us to identify and categorize different types of relationships from text. We compare this approach to manual annotation and fuzzy matching using a corpus of digital laboratory notebooks, examining inference quality at the level of edges (recovering a single link), labels (recovering the relationship context) and at the whole-network level (recovering local and global network properties). Large Language Models perform impressively well at each of these tasks, with edge recall rate ranging from 0.8 for the highly contextual case of recovering the task allocation structure of teams from their unstructured attribution page to 0.9 for the more explicit case of retrieving the collaboration with other teams from direct mentions, showing a 32% improvement over a fuzzy matching approach. Beyond science, the flexibility of LLMs means that our approach can be extended broadly through minor prompt revision.

Keywords Social networks, Network reconstruction, Large language models, Collaboration networks, Task allocation structures

Introduction

A large amount of social network data is recorded in unstructured text: nodes are mentioned with non-standard labels, and the edges between them, which may constitute one or more kinds of interaction, are described with idiosyncratic, contextually dependent phrases. This makes it difficult to accurately reconstruct networks in an automated way. While manual annotation can be performed for small datasets, this approach becomes prohibitively costly at scale. In particular, leading scholars to either rely on low quality data, or to direct attention elsewhere.

Research on scientific collaboration is a good example. Identifying who works with whom, and what they contribute is essential for the allocation of resources and credit, as well as for improving the organization of scientific work (Fortunato et al. 2018). Existing research focuses almost exclusively on co-authorship, in part because it can be easily

observed in an article's metadata. However, sociologists have long described the invisible college that underlies research production: helpful scientists who share resources and advice without being included as authors (Solla Price and Beaver 1966; Oettl 2012). These informal collaborations are described in the acknowledgements that accompany tens-of-millions of articles. They are accessible to researchers, but unused because they would be prohibitively difficult to extract with existing methods. Similarly, the division of labor amongst co-authors is central to the quality and reliability of their work. Journals are increasingly adopting contribution statements to describe the division of labor. But only a fraction of them are reported in a structured format, and even those that do typically lack fine-grained descriptions of the tasks (Larivière et al. 2020; Sauermann and Haeussler 2017). Other extensive qualitative approaches such as surveys (Walsh and Lee 2015), interviews and manual annotation (Lazega et al. 2008) or self-reported statements (Masselot et al. 2023) have been used, and despite ensuring high quality in the curated data, are severely limited by time, granularity and resource constraints. As a result, we have a rather narrow and uncertain view of how scientific collaborations are formed, structured, and affect knowledge production (Hall et al. 2018).

In this study, we focus on a context-rich, large-scale text dataset of wiki-based digital laboratory notebooks from 2,000+scientific teams participating in the international Genetically Engineered Machines (iGEM) synthetic biology competition (Santolini et al. 2023). In their wiki page, teams document their scientific project and outcomes, along with team member attributions – who did what – and collaborations with other teams. Prior work using the iGEM dataset reconstructs a proxy of the task allocation network from how members of a team co-edit different sections of the wiki. However, moving beyond digital traces, the rich textual information provides a more elaborated view on task allocation structure in scientific teams. As such, obtaining these inter-team collaboration networks and intra-team organizational structures across this large number of teams requires extracting specific information from heterogeneous and unstructured text.

To address these challenges in measurement, we present a semi-supervised approach leveraging on Large Language Models (LLMs). We evaluate the performance of using LLMs on two network retrieval tasks from unstructured text of increasing difficulty. First, we aim to identify the inter-team collaboration network, where direct interactions between teams are encoded within the wiki text. In this task, we have some prior knowledge about team names, and are interested in retrieving team collaborations along with the type of interaction. Second, we identify the task contribution structure of iGEM teams from the self-reported attribution statements of team members. Here, we retrieve a member-to-task bipartite network where we have prior knowledge about team members, and are inferring task allocation from contextual data. For both cases, we use a manual labeling test set to evaluate the accuracy of LLMs in retrieving relationships and their contexts, and investigate their ability to accurately reproduce local and global network properties. Beyond the quantitative results, we aim to provide the reader with a guide on the good practices and pitfalls we experienced with using LLMs for these information retrieval tasks.

The paper is organized into 4 further sections. We talk about recent approaches in using LLMs in information retrieval and present some background for the iGEM competition and dataset. We then describe in the Methods section the pipeline to curate and

validate the intra-team contribution structure and the inter-team collaboration network. Finally, we present the results and discuss perspectives for future work in leveraging the iGEM dataset and using LLMs for network reconstruction in computational social science research.

Related work

Text data encodes relationships between heterogeneous entities. Recognizing these relations is crucial across various scientific disciplines, and has benefitted from the advent of natural language processing and machine learning methods (Detroja et al. 2023; Pawar et al. 2017). For example, biologists might be interested in retrieving interactions between proteins from the published literature to build a comprehensive protein interaction network that can serve medical insights (Rolland et al. 2014). For this, researchers will have to navigate the different ways these proteins can be referred to, and the different phrase structures that can evoke a similar relation. In the humanities, scholars might be interested in retrieving character networks from fiction works to study narrative patterns across stories (Labatut and Bost 2020). In social sciences, the extraction of relationships from online data, free text surveys or unstructured self-reports can recover multimodal social relationships at scale (Deri et al. 2018; Irfan et al. 2015).

Ensuring a high quality of information retrieval from text is challenging and requires tedious manual annotation and validation with experts. The use of information coding software, such as MAXQDA, has provided a strong support for the analysis of qualitative and mixed data, such as surveys with both standardized and open-ended questions (Kuckartz and Rädiker 2021). Yet, with the increasing availability of large-scale digital text data collections, the complexity of extracting and annotating these relationships increases further. To cope with scale, a popular method has been the use of Amazon's Mechanical Turk (MTurk) service. By leveraging non-expert crowd workers, the MTurk platform presents a low-cost and scalable method for text annotation tasks, showcasing significant correlations with expert annotation (Snow et al. 2008). However, recent studies have shown a significant decrease in data quality and reliability, requiring additional response validity indicators and data screening (Chmielewski and Kucker 2020).

Against that background, Large Language Models (LLMs) present an opportunity for assisting mixed methods research at scale (Karjus 2023). Leveraging their ability to understand linguistic contexts and address low to medium complexity tasks, studies have quantified the efficacy of using LLMs for text annotation, and showed they improve over manual curation using Amazon Mechanical Turk (Alizadeh et al. 2023; Gilardi et al. 2023; Törnberg 2023). By refining prompts using manually curated validation data (Reiss 2023), annotation tasks can be customized for each use case to yield more relevant results (Goel et al. 2023). These encouraging findings have led to the development of open-source tools such as GraphGPT (Shenoy, 2023/2024) that extract heterogeneous information networks out of text data, paving the way for the application of LLMs in the context of network reconstruction. Preliminary insights using the textbook example "*Les Misérables*" character network reconstruction from Hugo's eponym fictional work show a promising increase in richness of information retrieval compared to the smaller manually obtained version, including non-plot characters discussed by Hugo in tangential sections of the book (Karjus 2023).

As such, LLMs are a promising tool for reconstructing social networks from unstructured text data. Focusing on the digital lab notebooks from teams participating in the iGEM competition, we aim to address the gap in evaluating the efficacy in using them as an adaptive solution for curating social networks and present the framework as a low-effort, yet reliable solution to address similarly structured tasks in the information and the social sciences.

The iGEM dataset

Launched in 2003, the iGEM competition has become a cornerstone in the Synthetic Biology field, promoting an open, collaborative approach to solving real-world problems using standardized DNA elements, or "BioBricks." It encourages a bottom-up, community-based learning method, fostering dialogue and transparency around ethical and safety concerns. From its inception with just 5 teams, iGEM has expanded globally, involving over 4,300 university and high school teams from 40 countries. Annually, teams use BioBricks to create synthetic biology solutions, culminating in a Jamboree where their projects are judged. The competition emphasizes collaboration and high-quality documentation, with each team's wiki serving as a key evaluative tool.

The team wiki website serves as an extended scientific article that includes technical specifications of the project, experimental and modeling methods, project results, and outreach practices the team members undertook. In addition to describing the content of the team project, the website contains context-rich information on the team members, their background and contribution to the project, and the collaborations the team maintained with other iGEM teams. This rich data source enables analysis of collaboration networks and project outcomes, offering insights into the practice and impact of collaborative science in interdisciplinary contexts (Santolini et al. 2023).

While wikis have an overall topical structure dictated by the needs of the competition – i.e. they must showcase collaborations with other teams and describe the contributions of team members–, the presentation of the content within each page varies significantly across teams. With over 2,200 participating teams in 2008 to 2018 and an average 17 members per team, manually curating and labeling individual contributions within teams and collaborations across teams is a time and resource consuming process. Here we use LLMs and specifically the GPT family of models to tackle this challenge and evaluate their performance.

To reconstruct the team collaboration and task allocation networks, we first extract the raw text from all teams public wiki pages. The text content of the pages was extracted using the "KeepEverythingExtractor" option in the boilerpipe.extract library for processing and removing boilerplate content after web scraping. This amounts to 106,207 pages from 2,265 teams over 11 years, from 2008 to 2018. The raw html and processed text of the wikis is available in Zenodo, along with team metadata from the iGEM website (Blondel et al. 2024).

Methods

Network reconstruction pipeline

We show in Fig. 1 an overview of the workflow to reconstruct the collaboration and task allocation networks in iGEM teams using GPT models. The workflow is similar for both the intra- and inter-team networks, with specific changes in the implementation.

Pipeline to extract the team contribution and inter-team collaboration structure from the text data of wiki pages from iGEM teams. The text from the wiki page is processed and passed to the language model along with the prompt. The output is a table with the decorated relationships, i.e. the user identifier, the activity they performed and the matched category for the contribution network and the team, target and the context of collaboration for the inter-team collaboration network. Schematic networks are shown in the bottom left, where circular blue nodes denote teams in the inter-team collaboration network, green pentagons denote team members and yellow rectangles the categories of tasks they performed in the intra-team contribution structure.

The first step is to process the text from the wiki pages into chunks. This is to primarily remove pages that do not contain information about either inter-team collaborations or about intra-team contributions. This step also ensures that input size limits (tokens) for the GPT models are preserved.

The curation using the GPT models is split into two steps. The first step extracts all relationships between entities present in the provided input text along with their described context. The objective of this step is to maximize recall and to ensure that there are minimal false negatives. The second step aims to match the extracted entities with their corresponding identifiers. In the inter-team case - it is to match team names to their official team IDs and the context of their collaboration to one of five standard collaboration categories in iGEM - "Work", "Material Transfer", "Meetup", "Advice" or "Other". For the intra-team case, it is to match team members to their member IDs and the tasks they undertook to a standard list of tasks that teams are expected to perform during the iGEM competition. This list is inspired by the CRediT Contributor Roles Taxonomy (Larivière et al. 2020), and constructed to encompass all tasks and deliverables that teams work on and are evaluated on in iGEM. **The categories are: "Design", "Experiments", "Documentation", "Interlab", "Modeling", "Analysis", "Parts", "Safety", "Entrepreneurship", "Hardware", "Software", "Human Practices", "Public Engagement", "Collaboration", "Fundraising", "Creative Contributions", "Administration", "Material Supply", "Supervision", "Training" and "Other"**. In this step, we eliminate the false positives and increase the precision of the final curated network.

To evaluate the curated networks, we create a validation set by manually curating relationships from teams and text chunks selected at random. For the inter-team case, we also use fuzzy matching to construct an alternate reference set - although this approach does not identify the contexts of inter-team collaborations. As a part of their iGEM project in 2016, team Waterloo constructed a database of collaborations between teams in 2015 - which we use as a further benchmark to evaluate the quality of the approach.

Specifications of the intra- and inter-team curation - such as the text processing, the prompts used, constructing the validation sets and the implementation details using the GPT models are detailed in the Supplementary Information.

Statistical analyses

Standard errors for the precision and recall scores of the inter-team collaboration networks in Fig. 2a are computed using the jackknife method over chunks. In the case of the intra-team contribution networks, precision and recall are computed at the team level and we show in Fig. 3a the average and standard error across the manually curated

teams. The confusion matrices between relationship contexts for the intra- and inter-team networks in Figs. 2b and 3b are row normalised using Z-scores: each observation x is centred and standardised using $Z = (x - \mu)/\sigma$, where μ is the row average and σ its standard deviation.

We use the pROC library in R for computing the ROC curves and Area under the ROC curve (AUC) of Fig. 2c.

Hierarchical clustering in Supplementary Fig. 5 is performed by computing the Euclidean distances between predicted categories from the confusion matrix (i.e. the rows) with the R dist function, and then using the "hclust" function in R.

All network analyses were performed with the igraph library in R. The betweenness and closeness centrality measures are unweighted and computed using their namesake functions. The coreness of the network was computed using the "coreness" function and the local clustering coefficient using the "transitivity" function setting the type argument to be "local". The network assortativity measures were computed for the categorical vertex attributes region, country and section using the "assortativity_nominal" function in igraph. The degree assortativity is computed using the "assortativity_degree" function. Standard errors for the network properties are computed using the jackknife method by removing one team from the network in each iteration, recomputing the corresponding network property, and calculating the standard deviation across the obtained values.

The degree distributions of the intra-team bipartite networks are computed using the in-built function degree_distribution from the igraph library with the "cumulative" option set to true. To quantify nestedness, we compute the standardized NODFc nestedness metric based on overlap and decreasing fill, described in (Song et al. 2017), using the maxnodf library in R (Hoeppke and Simmons 2021).

Results

Team collaboration networks

Accuracy of inferred relations

We first investigate the accuracy of GPT models and a fuzzy matching approach in retrieving the collaboration relations between iGEM teams. Our approach is described in Methods. We manually curate 200 text chunks from the *Collaboration* wiki pages across teams to extract the relationships present in the text. The same text chunks are then passed through the GPT and fuzzy matching pipelines to retrieve predicted

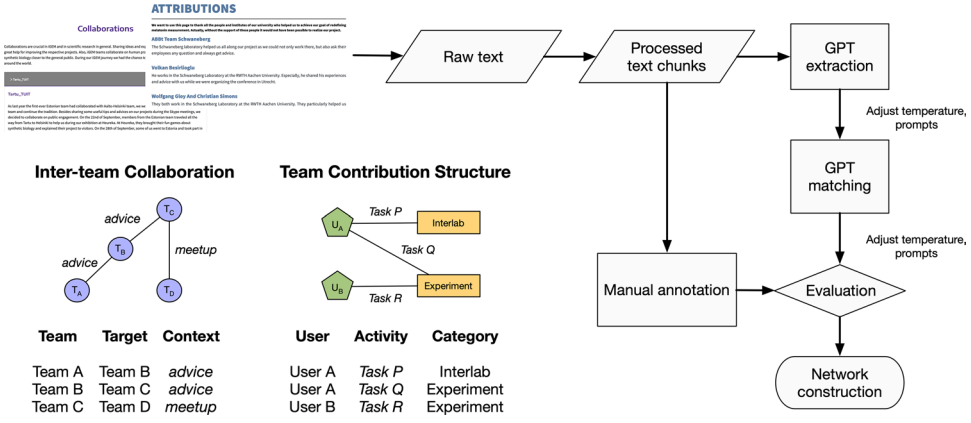

Fig. 1 Overview of the pipeline to reconstruct social networks from text

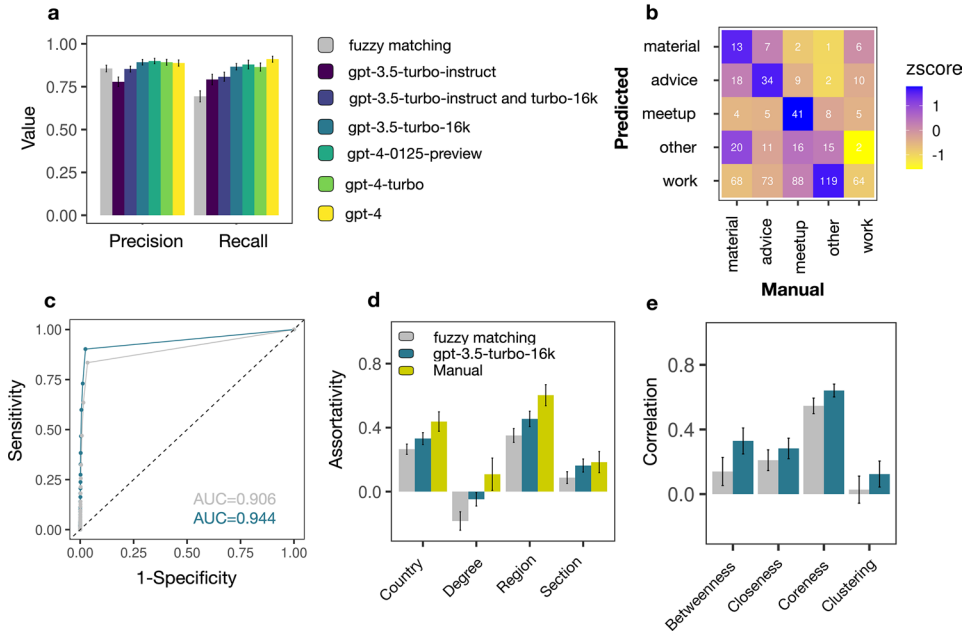

Fig. 2 Reconstruction of inter-team collaboration networks **(a)** Precision and recall of GPT models and fuzzy matching approach against manually curated chunks. **(b)** Confusion matrix showing the accuracy of category prediction for relationship contexts. The color coding indicates row-standardized Z scores values. **(c)** Receiver Operating Curves (ROC curves) for weighted interactions from 2015. Dashed line indicates a line of slope 1 and intercept 0. **(d)** Comparison between network assortativity values for reconstructed and manually obtained 2015 inter-team network. **(e)** Correlations of network properties between the 2015 reconstructed and manual network. In all plots, the error bars are computed using the jackknife method

relationships. We compare the performance of the GPT and fuzzy models against manually curated data in Fig. 2a. We focus on precision and recall, quantifying respectively the proportion of predicted relations that are in the manually curated set, and the proportion of manually curated relations that are retrieved. We find that newer GPT models consistently increase in precision and recall. Amongst the various models tested, gpt-4 has a higher recall while gpt-3.5-turbo and gpt-4-1025-preview models have a comparable performance, with the former being 10x cheaper. The turbo instruct models allow for inbuilt parallel request calls, but their performance is a detrimental tradeoff for information extraction and curation tasks. While precision flattens after GPT3.5-Turbo-16k, recall shows a significant improvement with GPT4, with a recall of 0.91. Interestingly, while fuzzy matching does not yield too many false positives – it has competing precision with GPT3.5 models–, it does not provide a good coverage of manual edges, with a recall of 0.69, giving GPT4 a 32% improvement over the fuzzy approach.

Inference of the collaboration context

Beyond edge retrieval, we evaluate the ability of the GPT models to infer the collaboration context from the surrounding text. In the following we focus on gpt-3.5-turbo-16k as it has the higher performance-to-cost ratio. This harder problem, that the fuzzy approach does not tackle, is evaluated by comparing the results to manual labeling. We show in Fig. 2b the comparison between the categories inferred with gpt-3.5-turbo-16k and the manually labeled ones. We find an overall accuracy of 0.26 – i.e. the predicted category matches the manually labeled category 26% of the time. However, there is variation across categories. 'Meetups' are identified with a high accuracy of 0.65, while

categories 'Advice' and 'Material transfer' have accuracies of 0.46 and 0.45 respectively, and are sometimes confused with one another. On the other end of the spectrum, most of the relationships identified as a 'Work' relationship by GPT are manually labeled as 'Other', leading to a poor accuracy of 0.16 for this category. This may be due to a lot of non-work relationships still using phrases like 'we worked with', 'we collaborated with' in their descriptions. Overall, these results show that while there is some degree of confidence in extracting categories from context, there is a strong variation between categories, with some being very accurate and others very hard to predict. We note that this exercise depends on the taxonomy chosen for the categorization, and the poor results in some categories might as well be seen as a poor definition of the category itself. Supplementary Fig. 2 shows the confusion matrix between categories for the other GPT models, indicating a similar pattern with the other models.

Accuracy of the reconstructed network

While previous measures focus on the ability to infer an edge and its context, here we investigate the ability of GPT models to infer accurate network properties in the reconstructed network. To do so, we evaluate the gpt-3.5-turbo model and fuzzy approach against a manually curated network of the significant inter-team collaborations in 2015, extracted by the 2016 Waterloo iGEM team (see Supplementary Information). This curated network does not comprise all team mentions that can be retrieved from the pages, but only the ones deemed significant by the curators. For both inference methods, the edges are weighted by the number of unique text chunks where the corresponding collaboration is reported. These weights can then be used to predict the occurrence of a significant edge in the manually curated network. We show the resulting ROC curve in Fig. 2c, along with corresponding Areas under the ROC curve (AUCs). We find that GPT outperforms fuzzy matching with an AUC of 0.944 (edge recall of 0.90) compared to 0.906 (recall of 0.83) for the fuzzy method. Finally, we investigate the extent to which the reconstructed networks, though still somehow noisy at the edge level, recover some key network properties of the manually reconstructed network. We show in Fig. 2d that GPT3.5 consistently improves over fuzzy matching to recover assortativities of the team attributes. We finally investigate if GPT can recover the relative importance of nodes in the network. For this we use Spearman correlation as a measure of the similarity of ranks under various network centrality measures between the inferred networks and the manually curated one. We find larger Spearman correlations across several local (clustering) and global (betweenness, closeness, and coreness) centrality measures for GPT compared to fuzzy matching, showing the ability of the former to build a more accurate overall representation of the network. We note poor results for clustering, which might be due to curation decisions – e.g. meetups between several teams can create large cliques that might not be considered significant by the Waterloo team.

Team contribution structure

Accuracy of inferred relations

In this second part, we focus on the question of identifying the tasks that team members have done in their project from the attribution page. Identifying the contribution structure of a team is a more complex task as only the name of the member is encoded in the text, while the category of the relationship they were involved in is inferred from

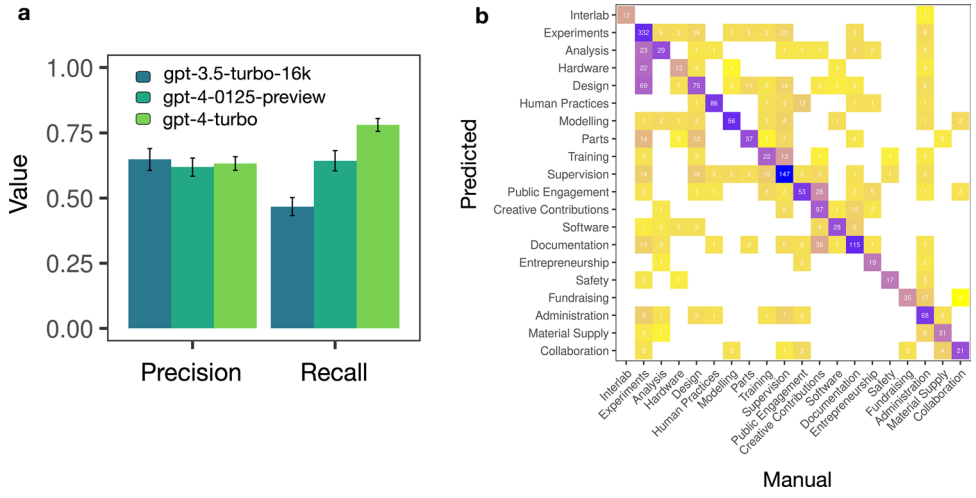

Fig. 3 Reconstruction of team contribution structure **(a)** Precision and recall of GPT models against manually curated contribution networks. **(b)** Confusion matrix of inferred category labels

the description of their activities. We first compare the precision and recall scores for the task categories retrieved using GPT models against the manually curated teams. Figure 3a shows a significant performance improvement in using the newer gpt-4-turbo class of models, with an edge-level recall across the manually labeled teams of 0.79, and a precision of 0.62. The lower precision means that descriptive task contexts are often matched to both the manually assigned category (high recall), but also to other categories, decreasing the precision.

Task-level accuracy

To investigate this mismatch, we compare the manual and GPT matching of the description of the individual members' task attributions to the defined task categories. Figure 3b shows the confusion matrix between the predicted and manual labels. Overall, we find an accuracy of 0.66, with variation between categories (see Supplementary Fig. 3 for precision and recall values across categories). We observe higher accuracy for categories with specific descriptions - such as being involved in interlab (accuracy=0.92), performing human practices (0.80) or modeling work (0.79). However, categories with high similarities in their descriptions are either misidentified, or identified to all similar alternatives, especially amongst experiments, performing analyses, design or between supervision and training. The clustering of categories based on the similarity of their labeling profile in the confusion matrix (Supplementary Fig. 5) identifies groups of strongly interrelated tasks (e.g. supervision and training), corresponding to similar tasks when working within an iGEM team. Information from these clusters could be used to increase the specificity of each subcategory description and reduce their similarity, or merge similar categories to improve the overall accuracy.

Accuracy of inferred network properties

FInally, we go beyond the edge level to investigate the ability of GPT to reproduce key properties of the team bipartite networks. First, we show in FIgure 4a the degree distributions for each layer in the inferred and manual networks. We find similar distributions, with large Pearson correlations between (the log10 of the) degrees of $r=0.76$

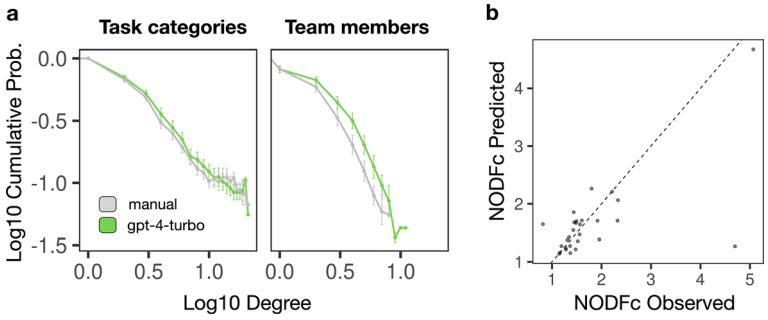

Fig. 4 Recovering bipartite network properties. **(a)** Cumulative degree distributions for the task categories and team members layers of the bipartite network, comparing manually curated networks (grey) with gpt-4 inferred networks (green). **(b)** Nestedness values, computed using NODFc (Song et al. 2017), are shown between the predicted and manual networks. Dashed line indicates a line of slope 1 and intercept 0

($p<$2e-16) for users and $r=0.85$ ($p<$2e-16) for tasks. In addition to degree distributions, bipartite networks are often characterized by a correlation between the degrees of the layers, leading to nested structures that can be quantified with the NODFc value (Mariani et al. 2019). We show in Fig. 4b the predicted and observed values of nestedness, with a high Pearson correlation of $r=0.66$ ($p=$1e-4). We find the presence of an outlier team with a high manually curated NODFc and a low predicted one. Under inspection it is due to the specificity format of the team page, involving images of team members rather than their name. Such outlier events can be taken care of using preprocessing of high-confidence pages that can offer a good inference, for example by evaluating the proportion of names of team members present in the page.

Discussion

This study aimed at showcasing the use of GPT models to infer team collaboration networks from text, providing a quantified account of what had been scoped in previous work (Karjus 2023). Overall, we found that the use of GPT models has clear benefits for extracting such information at scale. In extracting direct relationships between teams encoded within the text, it performs significantly better than a fuzzy approach, and can recover key aspects of the network structure. For example, fuzzy matching fails in cases where teams mention the names of members using their first name or nicknames, or in cases where they describe teams by the names of the universities they are based in rather than their team name. Similarly, the partial substring matching of the fuzzy approach has difficulty disambiguating team names with similar acronyms, such as UCL and UCLA. GPT models overcome these challenges through their larger inbuilt contexts which allows better disambiguation. Extracting the contexts of the relationships is another advantage that GPT models have over fuzzy matching methods, which would require additional language processing methods such as topic modeling to extend to other similar settings.

The similar performance between gpt-3.5-turbo and the more expensive gpt-4 class of models, unlike the lower performance observed in the second, more complex task, suggests a higher feasibility in reconstructing these direct relationships with simple few shot prompts and minimal post-processing. We also note that further pre-processing of the dataset aiming to provide higher quality text through filtering out pages with low abundance of certain words (such as team names or team member names), would probably

increase the overall precision and recall, while decreasing the amount of data that can be treated with the method.

Matching the contexts of the relationships to a list of categories has certain caveats. With the pipeline first extracting the relationship contexts and then performing the matching, some information beyond the sentence describing the relationship is lost. Attempting to perform both the extraction and the matching in a single step preserves most contextual information from the text, but performs poorly in precision and recall, especially when tested with the cheaper gpt-3.5 class of models. This could potentially be caused by an individual request now performing a more complex task, which current advances in model complexity might resolve.

Evaluating the performance of reconstructing the indirect relationships of the intra-team contribution structure shows promising initial results, with drastic improvements observed with newer GPT models. This might reflect the task complexity that heavily relies on context-awareness for the task allocation to team members. The similarity between the properties of the manual and inferred networks – degree distributions and nestedness values – indicates that the inferred networks can be leveraged for future studies comparing their structure across the large number of teams of the iGEM competition.

We also note that the process of manual curation is itself prone to inter-individual variation. Decisions on assigning a description to a specific category relies on personal judgment, and quantifying this variation would require a more extensive investigation using platforms like Mechanical Turk. As such, it is yet not clear whether GPT models already outperforms human annotators or not in the context of this study. The high recall rates and comparatively low precision have been interpreted as a larger richness of GPT models (Karjus 2023), capturing more peripheral relations than immediately catches the eye. Future work could investigate whether this richness helps capture more subtle phenomena that can be useful in downstream analyses.

Finally, the ability to extract and label relationships from text in bulk saves on human effort, but does not eliminate it. There are several drawbacks in using proprietary LLMs - including but not limited to reproducibility, sensitivity to temperature, prompt adjustment and model updates. This creates the necessity to carefully evaluate model performance, ideally with high quality manually annotated data (Ollion et al. 2023, 2024; Reiss 2023). An additional concern is working with surveys and other confidential data sources. A potential workaround is by leveraging open source LLMs for network reconstruction. Open source LLMs have shown to outperform crowd workers, but lag behind the GPT models in text annotation (Alizadeh et al. 2023). However, this gap can potentially be reduced by fine-tuning open source LLMs with contextual data to improve on performance. The approach requires further human work in curating a training set, but allows customizability and executing requests in-house. This is a potential extension of the study.

Conclusion

We assessed the performance of using OpenAIs GPT family of models in curating and annotating team collaboration networks from unstructured text. We leveraged digital laboratory notebooks from scientific teams participating in the iGEM competition to infer inter-team collaborations and the team contribution structure from heterogeneous

self-reported data. We show that despite wide differences in page structures, the networks can be extracted with Large Language Models by using prompts that include minimally supervised team-specific information (list of users, team names, and types of contributions), resulting in structured graph data showing precision in par with manually curated data. We find that recall rates consistently increase with each new model, indicating that future improvements are expected with upcoming releases. We also showcase that LLMs can partially retrieve contextual information to infer edge types, with future improvements possible with new models and improved prompt design and category definition. This work has implications for the study of division of labor in scientific teams that often relies on CREDIT contribution reports (Xu et al. 2022), providing both a Method and a fine-grained dataset (Blondel et al. 2024) for studying fine-grained task allocation structure. More generally, this study suggests that LLMs can be a useful, scalable and efficient approach to network reconstruction for assisting manual curation work in computational social science and digital humanities.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-024-00658-8.

Supplementary Material 1

Acknowledgements
We thank Lionel Deveaux for assisting us with the API setup.

Author contributions
RJ: Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. RNW: Conceptualization, Formal Analysis, Methodology, Software, Writing – review & editing. MS: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Funding
This work was supported by the French Agence Nationale de la Recherche (ANR), under grant agreement ANR-21-CE38-0002-01.

Data availability
The datasets generated and/or analyzed during the current study are available in a Github repository (https://github.com/InteractionDataLab/Reconstructing-Social-Networks-from-Unstructured-Text-using-Large-Language-Models) and the wiki content is available in Zenodo - https://doi.org/10.5281/zenodo.11072818 (Blondel et al., 2024)

Declarations

Competing interests
The authors declare no competing interests.

Received: 30 April 2024 / Accepted: 8 August 2024
Published online: 09 October 2024

References

Alizadeh M, Kubli M, Samei Z, Dehghani S, Bermeo JD, Korobeynikova M, Gilardi F (2023) *Open-Source Large Language Models Outperform Crowd Workers and Approach ChatGPT in Text-Annotation Tasks* (arXiv:2307.02179). arXiv. http://arxiv.org/abs/2307.02179

Blondel L, Jeyaram R, Krishna A, Santolini M (2024) iGEM: a model system for team science and innovation. Zenodo. https://doi.org/10.5281/zenodo.11072818

Chmielewski M, Kucker SC (2020) An MTurk Crisis? Shifts in Data Quality and the impact on study results. Social Psychol Personality Sci 11(4):464–473. https://doi.org/10.1177/1948550619875149

De Solla Price DJ, Beaver D (1966) Collaboration in an invisible college. Am Psychol 21(11):1011–1018. https://doi.org/10.1037/h0024051

Deri S, Rappaz J, Aiello M L, Quercia D (2018) Coloring in the links: capturing Social ties as they are perceived. Proc ACM Hum Comput Interact 2(CSCW):1–18. https://doi.org/10.1145/3274312

Detroja K, Bhensdadia CK, Bhatt BS (2023) A survey on relation extraction. Intell Syst Appl 19:200244. https://doi.org/10.1016/j.iswa.2023.200244

Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B, Vespignani A, Waltman L, Wang D, Barabási A-L (2018) Science of science. Science 359(6379):eaao0185. https://doi.org/10.1126/science.aao0185

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd-workers for text-annotation tasks. Proc Natl Acad Sci 120(30):e2305016120. https://doi.org/10.1073/pnas.2305016120

Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH, Hao X, Jaber B, Reddy S, Kartha R, Steiner J, Laish I, Feder A (2023) LLMs accelerate annotation for medical information extraction. Proc 3rd Mach Learn Health Symp 82:100. https://proceedings.mlr.press/v225/goel23a.html

Hall KL, Vogel AL, Huang GC, Serrano KJ, Rice EL, Tsakraklides SP, Fiore SM (2018) The science of team science: a review of the empirical evidence and research gaps on collaboration in science. Am Psychol 73(4):532–548. https://doi.org/10.1037/amp0000319

Hoeppke C, Simmons BI (2021) Maxnodf: an R package for fair and fast comparisons of nestedness between networks. Methods Ecol Evol 12(4):580–585. https://doi.org/10.1111/2041-210X.13545

Irfan R, King C, Grages D, Ewen S, Khan S, Madani S, Kołodziej J, Wang L, Chen D, Rayes A, Tziritas N, Xu C-Z, Zomaya A, Alzahrani A, Li H (2015) A survey on text mining in social networks. Knowl Eng Rev 30:157–170. https://doi.org/10.1017/S0269888914000277

Karjus A (2023) *Machine-assisted mixed methods: Augmenting humanities and social sciences with artificial intelligence* (arXiv:2309.14379). arXiv. https://doi.org/10.48550/arXiv.2309.14379

Kuckartz U, Rädiker S (2021) Using MAXQDA for Mixed Methods Research. In the routledge reviewer's guide to mixed methods analysis, Routledge, pp 305–318 https://doi.org/10.4324/9780203729434-26

Labatut V, Bost X (2020) Extraction and analysis of fictional character networks: a Survey. ACM-CSUR 52(5):1–40. https://doi.org/10.1145/3344548

Larivière V, Pontille D, Sugimoto CR (2020) Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT). Quant Sci Stud 2(1):111–128. https://doi.org/10.1162/qss_a_00097

Lazega E, Jourda M-T, Mounier L, Stofer R (2008) Catching up with big fish in the big pond? Multi-level network analysis through linked design. Social Networks 30(2):159–176. https://doi.org/10.1016/j.socnet.2008.02.001

Mariani MS, Ren Z-M, Bascompte J, Tessone CJ (2019) Nestedness in complex networks: Observation, emergence, and implications. Phys Rep 813:1–90. https://doi.org/10.1016/j.physrep.2019.04.001

Masselot C, Jeyaram R, Tackx R, Fernandez-Marquez JL, Grey F, Santolini M (2023) Collaboration and performance of Citizen Science projects addressing the Sustainable Development Goals. Citiz Science: Theory Pract 8(1):1. https://doi.org/10.5334/cstp.565

Oettl A (2012) Reconceptualizing stars: scientist helpfulness and peer performance. Manage Sci 58(6):1122–1140. https://doi.org/10.1287/mnsc.1110.1470

Ollion E, Shen R, Macanovic A, Chatelain A (2023) Chatgpt for Text Annotation? Mind the Hype! *SocArXiv. October*, *4*. https://files.osf.io/v1/resources/x58kn/providers/osfstorage/651d60731bc8650a79f376cf?action=download&direct&version=1

Ollion É, Shen R, Macanovic A, Chatelain A (2024) The dangers of using proprietary LLMs for research. Nat Mach Intell 6(1):1. https://doi.org/10.1038/s42256-023-00783-6

Pawar S, Palshikar GK, Bhattacharyya P (2017) *Relation Extraction: A Survey* (arXiv:1712.05191). arXiv. http://arxiv.org/abs/1712.05191

Reiss MV (2023) *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark* (arXiv:2304.11085). arXiv. http://arxiv.org/abs/2304.11085

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Vidal M (2014) A proteome-scale map of the human interactome network. Cell 159(5):1212–1226. https://doi.org/10.1016/j.cell.2014.10.050

Santolini M, Blondel L, Palmer MJ, Ward RN, Jeyaram R, Brink KR, Krishna A, Barabasi A-L (2023) *iGEM: A model system for team science and innovation* (arXiv:2310.19858). arXiv. http://arxiv.org/abs/2310.19858

Sauermann H, Haeussler C (2017) Authorship and contribution disclosures. Sci Advances 3(11):e1700404 https://doi.org/10.1126/sciadv.1700404

Shenoy V (2024) *Varunshenoy/GraphGPT* [JavaScript]. https://github.com/varunshenoy/GraphGPT (2023)

Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In M. Lapata & H. T. Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 254–263). Association for Computational Linguistics. https://aclanthology.org/D08–1027

Song C, Rohr RP, Saavedra S (2017) Why are some plant–pollinator networks more nested than others? J Anim Ecol 86(6):1417–1424. https://doi.org/10.1111/1365-2656.12749

Törnberg P (2023) ChatGPT–4 outperforms experts and crowd workers in Annotating Political Twitter messages with zero-shot learning. arXiv. arXiv:2304.06588. http://arxiv.org/abs/2304.06588

Walsh JP, Lee Y-N (2015) The bureaucratization of science. Res Policy 44(8):1584-1600. https://doi.org/10.1016/j.respol.2015.04.010

Xu F, Wu L, Evans J (2022) Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences*, *119*(23), e2200927119. https://doi.org/10.1073/pnas.2200927119

Publisher's Note