The Social Lives of Literary Characters: Combining citizen science and language models to understand narrative social networks

Andrew Piper † **Michael Xu** † **Derek Ruths** †

†Department of Languages, Literatures, and Cultures ‡School of Computer Science

McGill University

Abstract

Characters and their interactions are central to the fabric of narratives, playing a crucial role in developing readers' social cognition. In this paper, we introduce a novel annotation framework that distinguishes between five types of character interactions, including bilateral and unilateral classifications. Leveraging the crowd-sourcing framework of citizen science, we collect a large dataset of manual annotations (N=13,395). Using this data, we explore how genre and audience factors influence social network structures in a sample of contemporary books. Our findings demonstrate that fictional narratives tend to favor more embodied interactions and exhibit denser and less modular social networks. Our work not only enhances the understanding of narrative social networks but also showcases the potential of integrating citizen science with NLP methodologies for large-scale narrative analysis.

1 Introduction

Characters and their interactions are a fundamental feature of storytelling. As a prominent dimension of cognitive literary theory has argued, fictional characters provide readers with the opportunity to identify with other imaginary human beings and model social relationships (Zunshine, 2006; Mar et al., 2006, 2009, 2010; Palmer, 2004). According to this theory, the interactions among characters and the resulting social networks provide an important training ground for the development of social cognition (a.k.a. Theory of Mind) (Kidd et al., 2016).

Well over a decade ago, a robust body of work began to emerge in NLP to address the extraction of social networks from narrative texts (Elson et al., 2010; Lee and Yeung, 2012; Agarwal et al., 2013, 2014). That work established important methodological foundations for the study of literary social networks, which were understood to be a key component of research in the Digital Humanities (Moretti, 2011).

Nevertheless, this work was faced with two significant challenges: 1) automated methods never exhibited robust levels of accuracy to be applicable to real-world cases; 2) the costliness of deriving high-quality training data made it difficult to build more accurate models.

In this paper, we aim to address these challenges to further the study of narrative social networks by focusing on the following methodological contributions:

1. Establishing a novel annotation framework that includes five distinct interaction types, including a second level classification of *bilateral* versus *unilateral* types (i.e. whether one of the characters involved is aware of the interaction).

2. Illustrating the value citizen science can have for research in NLP by releasing and validating a large-scale dataset of manual annotations of character interactions (N=13,395).

3. Training, validating, and publicly sharing a performant, open-weight small language model or SLM (Phi3-7B) for the task of interaction labeling (F1 = 0.70).

4. Testing the effects that different genres and audience types have on social network structure within a sample of contemporary books (N=390). Here we demonstrate a proof-of-concept analysis of the ways in which social networks in stories are shaped by genre and audience factors such as fictionality, cultural context, prestige, and expected reader age.

We conclude with a discussion of areas for future work.

2 Prior Work

Some of the earliest theoretical work concerning the value of social networks for literature was undertaken by Moretti (2011) and Woloch (2009). Woloch (2009)'s theory of *character-space* has

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 472–482

been particularly influential. This theory highlights the skewed distribution of attention around a primary character, referred to as 'the one and the many' structure by Woloch. This work established an important theoretical framework for studying character *relations* in addition to individual characters (Propp, 1968; Frow, 2014).

Early methodological work on the extraction of character networks was undertaken by several researchers (Sudhahar and Cristianini, 2013; Agarwal et al., 2014; Trovati and Brady, 2014; Nijila and Kala, 2018). Much of this work focused on the use of sentence-level co-occurrence or subject-verb-object triplets as the foundation of building character interactions. Labatut and Bost (2019) provide an extensive survey of methods of social network extraction applied to cultural texts (i.e. stories and screenplays), ranging from interaction identification to network analysis methods.

In terms of applications, Mac Carron and Kenna (2012) analyzed character networks within three European epics (*The Iliad, Beowulf, Tain Bo Cuillange*) to understand their relation to contemporary real-world social networks. Volker and Smeets (2020) compared fictional networks in Dutch literature with real-world networks with respect to racial groups. And Ardanuy and Sporleder (2014) and Agarwal et al. (2021) used social networks as a mechanism to detect book genres.

Dialogue networks have also been studied as a subset of literary social networks (Elson et al., 2010; Waumans et al., 2015), with a similar principal applied to the study of drama (Algee-Hewitt, 2017; Lee and Lee, 2017; Fischer and Skorinkin). Finally, substantial work has focused on the detection of *relationship* types (instead of individual interactions), such as kinship ties (Iyyer et al., 2016; Chaturvedi et al., 2016; Massey et al., 2015; Makazhanov et al., 2014) and conflict groups (Smeets et al., 2021).

3 Methods

3.1 Defining Character Interactions

We define a character interaction as occurring when *a character / group of characters engage in an action that involves another character / group of characters within the story world of a narrative*. This definition allows for the inclusion of a single character or group at the level of the agent or patient (object). And following the work of Agarwal and Rambow (2010), it also supports a base-level distinction between "bilateral" and "unilateral" interactions, i.e. when both characters are aware of the interaction or only one of the characters is aware. While an interaction requires two characters / groups to be an interaction, it does not require cognizance of the action by the patient.

Accordingly, we identify five possible types of interactions: *communicating, thinking about, observing, touching* (physical contact), and *associating* (which we use as a catch-all). Only *observing* and *thinking about* can be unilateral. Table 1 provides example sentences of the different interaction types.

3.2 Using Citizen Science for Manual Annotation

"Citizen science" is a term used to describe the general public engagement in scientific research activities (Consortium et al., 2013). Citizen science projects have annotated over 250 million pieces of data over the past two decades, ranging from the identification of galaxies, bird species, to the location of marine-based trash. Research shows that data produced by citizen science projects can be of high quality and correlate strongly with expert opinion when best-practices are employed (Kosmala et al., 2016; Wiggins and He, 2016). It also provides a cost-efficient means of data collection (Sauermann and Franzoni, 2015).

Several citizen science projects have emerged in the humanities in recent years (Ridge, 2016; Terras, 2015; Dobreva and Azzopardi, 2014), although the quantity of projects is still small compared to the natural sciences which represent an estimated 80% of all projects (Hecker et al., 2018). To date, most citizen science initiatives in the humanities have focused on document transcription. Our project, called *Citizen Readers*, uses the popular platform Zooniverse.org and focuses on text annotation common to the NLP community, which has traditionally been undertaken through fee-based crowd-sourcing platforms. Our project thus seeks to illustrate the opportunities that await both the humanities and NLP through the use of volunteer citizen scientists.

Figure 1 provides an illustration of our task structure. Participants were presented with two-sentence passages, where the first sentence serves as the context and the second sentence includes two highlighted characters for interaction classification. Passages were randomly sampled from the CONLIT dataset of contemporary books (Piper,

Type	Sentence
Associating	When Admiral Bloch left the Dauntless, **he** placed **me** in command of the fleet.
Communicating	Then I saw **them** waving at **me** from the far end of the restaurant.
Observing-Uni	**She** peers out at **the sniper**, but the angle is enough to hide her from his sight.
Observing-Bi	**Ange** looked over at **me**, then sprung up and headed my way.
Thinking	**She** thinks about **Ned** in his brown Doc Martens.
Touching	**Dr. Fell** gently brought **Ethel Pusster** to her feet.

Table 1: Example sentences of our various interaction types. Boldface represents the highlighted characters.

2022), which contains twelve different genres of fiction and non-fiction books, and characters were automatically detected using bookNLP (Bamman, 2021).

Participants were then asked a series of conditional questions: 1. Are the highlighted characters interacting in the story? If yes: 2. Is the interaction unilateral or bilateral? Given their answer: 3. What kind of interaction is it? for which the relevant classes were presented.

In addition to the task itself, Zooniverse provides an area for a custom tutorial, a field-guide with more in-depth descriptions, an about page to inform participants about the goals and intentions of the project, and a talk area where moderators can respond to participant questions. For this project we employed four student moderators who were indispensable in responding to the volume of questions.

A total of 1,915 citizen scientists participated in our project completing 73,648 unique annotations. The project took three months to complete. Out of the initial 19,006 passages posted to the platform, 15,641 were annotated by three or more annotators. The total number of passages where we observed a majority consensus on the label was 13,395. We found that 1,189 participants (or 62%) annotated five or more passages, and only 249 annotated a single passage. We also observe the Pareto principle at work, with 20% of our participants completing 72% of our annotations.

In order to assess the quality of annotations by citizen scientists, we hired three trained students to annotate a small subset of passages (N=261). We then compared agreement scores for three cases: inter-student annotations, inter-citizen scientist annotations, and student-citizen scientist annotations. We calculate Fleiss's Kappa scores for two conditions: all annotations and only those with majority votes. As we can see in Table 2, student annotators exhibit slightly higher agreement for all annotations but when conditioning on those with majority agreement the scores converge. We also show very high levels of agreement between student and citizen-scientist majority annotations, suggesting the high quality of our final annotations.

Condition	student	citizen scientist	student-citizen
All annotations	0.48	0.41	0.49
Majority votes	0.50	0.51	0.79

Table 2: Comparison of agreement scores using Fleiss's Kappa between students and citizen scientists.

We present the distribution of interaction types in our 13,395 majority-labeled passages in Fig. 2. The most common label is "no interaction," followed by "communicating" and "associating." While the other three types are far rarer, we will see in later sections their significance. We release our data set, known as the "Citizen Readers for Character Interactions" dataset (*CR4Interact*), which is available in our project's long-term repository.[1]

3.3 Finetuning an SLM for Interaction Detection

3.3.1 Training and Test Data

We then use our labeled data to fine-tune and test the performance of a small-language model (SLM) for the task of interaction type detection. For training and testing purposes, we extracted an equal amount of data from each class and a confidence score to partition the data, understood as the average agreement percentage per passage. We use passages with the highest confidence to build the test dataset, moderate confidence for the validation dataset, and the remaining data for the training dataset, as shown in Table 3. This approach guarantees the most accurate evaluation results possible, although it introduces some difficulties for the SLM because it will be trained on the lowest quality data.

[1] https://doi.org/10.5683/SP3/QMIARS

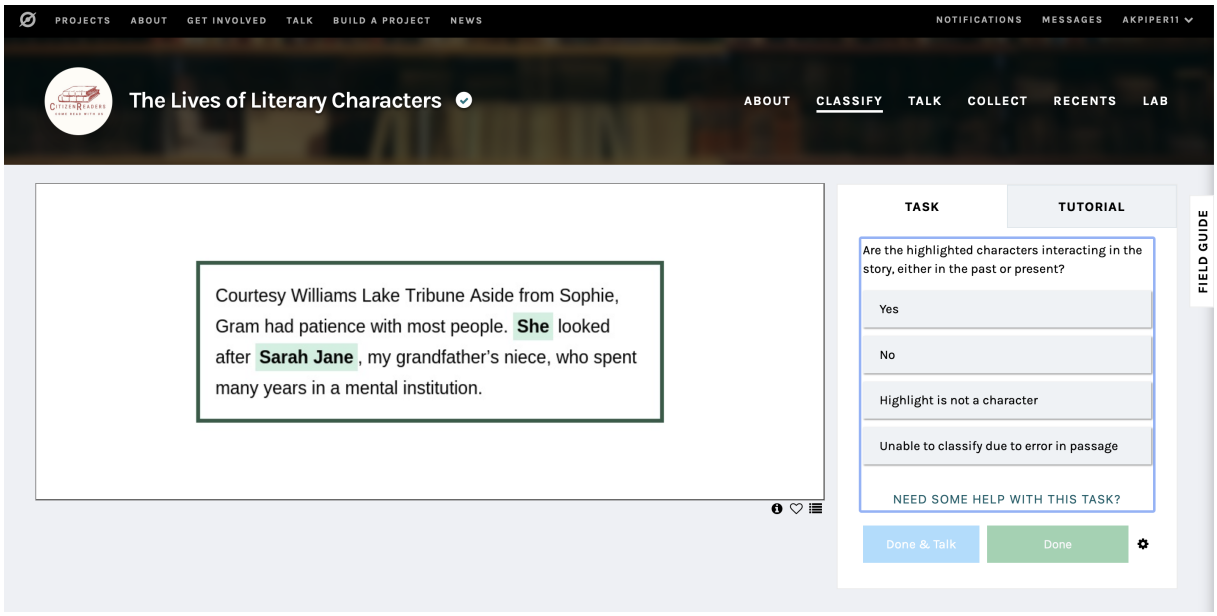

Figure 1: Image of our annotation task on Zooniverse.

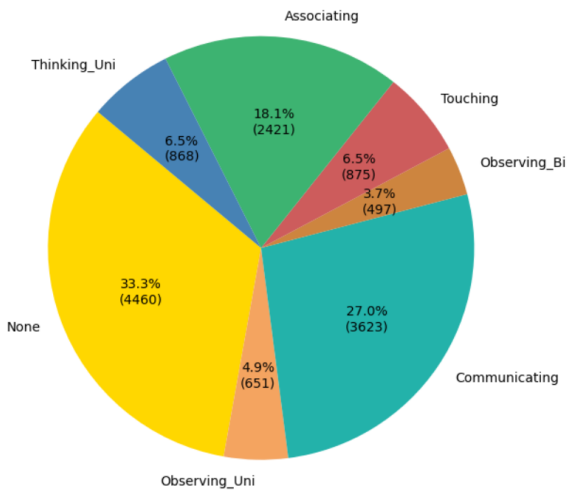

Figure 2: Label distribution in the annotated data.

Class	Training	Validation	Test
Associating	2032	225	225
No	4085	225	225
Thinking	443	225	225
Communicating	3223	225	225
Observing	721	225	225
Touching	437	225	225
Total Amount	**10941**	**1350**	**1350**

Table 3: Class distribution and total amounts for the training, validation, and test datasets.

3.3.2 Model Selection

To select the optimal combination of models and prompting strategies, we compare Phi-3-Mini-4K-Instruct, Phi-3-Small-8K-Instruct, Mistral-7B-Instruct-v0.3, Llama-3-8B-Instruct with GPT-4o (version 2024-05-13). We use the different prompting strategies listed in Table 4 and fully described in the appendix A and use a small balanced subset of our test data consisting of 10 passages per category for initial testing.

Among the open-source models, the detailed prompt on the Phi-3 (7B) achieves the highest accuracy of 0.73, as shown in Table 4. Phi-3 (7B) and GPT-4 are the only models that surpass 0.70 accuracy, and we anticipate an additional performance improvement after the finetuning stage. Note these performance numbers are based on only a small subsample of our test data.

Model	Base	Detailed	One-shot	Many-shot
Phi-3 (3.8B)	0.30	0.62	0.48	0.30 (8-shot)
Phi-3 (7B)	0.55	0.73	0.56	0.55 (8-shot)
Mistral (7B)	0.43	0.58	0.22	0.17 (8-shot)
Llama-3 (8B)	0.52	0.47	0.58	0.45 (5-shot)
GPT-4	0.65	0.70	0.73	0.72 (10-shot)
Tuned Phi-3	0.68	0.80	0.78	0.77 (8-shot)

Table 4: Performance comparison across different models and prompt strategies using the initial small, downsampled test dataset. Tuned Phi-3 (7B)'s performance is included for reference.

3.3.3 Our model: Phi-3-interact

We use a single A100 PCIE 80GB to finetune the Phi-3 (7B) on the training dataset and monitor the loss on the validation dataset. We set the learning

rate to `2.5e-5` and utilize the `paged_adamw_8bit` optimizer to reduce memory usage and accelerate tuning speed.[2]

After finetuning on the 10K training data, the model's accuracy increases from 0.58 to 0.71 on the validation dataset. Subsequently, tuning the model on the smaller, but higher-quality 1K validation dataset further improves accuracy to 0.80 on the small test dataset (as shown in Table 4) and from 0.727 to 0.735 using the full test dataset, surpassing GPT-4's 0.70 accuracy on the same data.

Table 5 indicates that both models struggle to distinguish the "No interaction" class, which is most often confused with the "Associating" class. Otherwise, categories range from a low of 0.70 accuracy (Thinking) to a high of 0.96 (Touching). A closer examination of the classification errors reveals that both models tend to assign positive interactions for hypothetical scenarios such as "If I could talk with her," which annotators were instructed not to consider as interactions because they do not actually take place in the storyworld. Our finetuned model, Phi-3-interact, demonstrates significantly higher precision in the "No interaction" category compared to GPT-4, indicating that Phi-3-interact is more reliable for predicting the absence of interaction.

Class	Precision	Recall	F1	Acc
Associating	0.75	0.56	0.64	0.56
	0.59	0.75	0.66	0.75
Communication	0.75	0.89	0.81	0.89
	0.64	0.85	0.73	0.85
No	0.89	0.21	0.34	0.21
	0.54	0.20	0.29	0.20
Observing	0.82	0.94	0.88	0.94
	0.89	0.79	0.84	0.79
Thinking	0.60	0.85	0.70	0.85
	0.61	0.71	0.65	0.71
Touching	0.76	0.96	0.85	0.96
	0.92	0.90	0.91	0.90
Mean	0.76	0.74	0.70	0.74
	0.70	0.70	0.68	0.70

Table 5: Performance metrics using our fine-tuned model (Phi-3-interact, upper row) and GPT-4 (lower row) on the full test dataset.

3.4 Constructing social networks from our data

In order to analyze social networks at the book level, we first sub-sample the CONLIT dataset down to 390 books to represent the genre and audience categories described more fully in Section 4 and shown in Table 6. We use bookNLP (Bamman, 2021) to perform sentence tokenization, entity recognition (NER tag "PER"), and co-reference resolution on the book level data. From there we extract all possible candidate pairs of characters for every sentence in each book for a total of 3,928,602 possible interactions. We then use our fine-tuned Phi-3 model to label all interactions.

To construct the nodes of our networks, we use the master character IDs provided by bookNLP that are derived from the co-reference resolution step. This gives us a list of unique characters per book. We then construct weighted edge lists for each book, where an edge represents the sum of all interactions between two characters. We then construct network graphs for each book for all interaction types and one aggregate network per book.

Finally, given our edge lists we then extract the following set of network statistics for each book according to two conditions: all characters and only characters whose degree (number of relationships) is five or greater, in order to focus on more significant characters.

Protagonist Centrality. The degree of the most connected character, normalized by dividing by the total number of edges. Equivalent to the percentage of all relationships consumed by the most central character.

Density. The ratio of the number of actual edges in a graph to the potential number of edges. Ranges from 0 to 1.

Transitivity. The global transitivity of the graph also known as the clustering coefficient. This measures the ratio of the number of closed triplets (or triangles) to the total number of triplets (both open and closed) in the network.

Average Shortest Path. The average length of the shortest path between all pairs of nodes in the network.

Modularity. Measures the strength of division of a network into communities, quantifying the degree to which nodes within the same community are more densely connected to each other than to nodes in different communities. Higher modularity values indicate stronger sub-community structure. Here we use the Fast and Greedy algorithm.

[2]https://huggingface.co/ChunB1/Phi-3-interact

4 Analysis

In this section, we aim to illustrate the potential utility of our data for the large-scale study of cultures of storytelling. We measure the effects of the following four stylistic and audience categories on the distribution of character interaction types and the resulting book-level social networks:

Fictionality. Here we test for the effects that fiction has on social interactions. For interactions we look at all adult genres. For social networks, we sample books from the Prizewinners and Bestsellers categories for fiction and Biography for non-fiction.

World. We test for the effects that books published in non-Western cultures have on social interactions. Specifically we look at books published in English in India, South Africa, and Nigeria that were reviewed in major literary reviews in their respective regions and compare them to our Western Prizewinner category.

Prestige. Prior work has identified strong stylistic differences between best-selling fiction and fiction written to appeal to literary elites on prize committees (Piper and Portelance, 2016). Here we test whether these findings extend to social interactions and their resulting networks.

Youth. For this category we compare middle-school fiction with adult fiction as represented by Prizewinners and Bestsellers. We expect to observe strong effects that are designed to make narratives more accessible to younger readers.

4.1 Book type effects on character interactions

In order to study the effects of our book categories on interaction types, we utilize count data to compute the log odds ratios through Fisher's Exact Test, focusing on the rate of each interaction type relative to the overall interaction rate for each category. Our findings indicate that fiction uniquely exhibits statistically significant effects (Fig. 3).

Specifically, non-fiction surpasses expectations in rates of communication and association, whereas fiction emphasizes observation and physical contact. These results corroborate existing theories that highlight the importance of embodied behavior in fictional narratives (Caracciolo and Kukkonen, 2021; Piper, 2024). Interestingly, there is no meaningful difference in the rate of unilateral versus bilateral actions in either corpus. Fiction does not indicate a preference for unilateral interactions as might be hypothesized by the strong emphasis on social cognition theories of reading fiction (Zunshine, 2006).

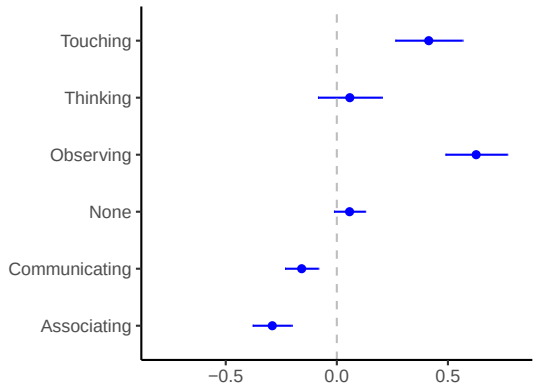

Figure 3: Log odds of interaction type appearing in fiction compared to non-fiction.

4.2 Book type effects on social networks

The first notable fact about our book-level social networks is the long tail of degree distribution (i.e. the number of relationships per character). For both fiction and non-fiction more than 86% of characters per book have fewer than five relationships in their respective social networks, suggesting a long tail of minor characters and a small, central core of main characters (Woloch, 2009).

Table 6 shows the results of our regression analysis for the aggregate social network structures in our sampled books by category. For our analysis, we require five or greater connections for inclusion in the network given the long tail of very minor characters (though we note that the overall results do not change in any meaningful way if we include all characters).

Our analysis reveals that fictional narratives exhibit significantly increased density and transitivity, coupled with lower modularity, average shortest path, and time to completion. To quantify the effect sizes, we converted these findings into Cohen's d values. This translation demonstrates substantial effects, with values ranging from d=1.3 for density to d=-2.2 for modularity, indicating large to very large effects. We also find that non-fictional narratives take 50% longer to complete their social networks when compared to fictional narratives.

These results indicate how strongly fictional narratives tend to create denser, more connected networks than biographical narratives. Although biographies condition explicitly on a single life, they

exhibit on average more modular narrative structures (strongly connected components that are less connected overall). Fiction by contrast maintains a denser overall relational network, one in which the introduction of significant new characters is completed much earlier than in biographical narratives.

Category	Fictionality	World	Prestige	Youth
Density	+ (***)	.	.	+ (***)
Centrality	.	.	.	+ (*)
Transitivity	+ (***)	.	.	+ (***)
Completion	- (***)	.	.	.
Modularity	- (***)	.	.	- (***)
Shortest Path	- (***)	.	.	- (***)

Table 6: Results of the regression analysis. +/- refer to positive or negative effects and the number of asterisks refer to p-value magnitude (* < 0.01, ** < 0.001, *** < 0.0001). A period denotes no meaningful effect.

These results lend further support to prior work suggesting a "small world effect" of fictional narratives. Prior work has shown that fictional characters cover smaller geographic distances (Matthew Wilkens, forthcoming) and that fictional narratives exhibit considerably lower overall informational surprise, favoring narrative 'exploitation' (covering familiar characters and situations) over narrative 'exploration' (introducing new characters, themes, and situations) (Piper et al., 2023). Here we can add the denser social network structures as a further index of this small world effect of fictional narratives.

Within fiction, we observe no meaningful effects for either social prestige (books receiving literary prizes) or books published in non-Western cultures. Youth books on the other hand exhibit very clear signals of simplified social networks with lower modularity, shorter paths, more centralized protagonists and greater relationship density. Youth books in other words tend to amplify the effects of fictionality.

Our results suggest two important points: the first is that the expected values we are observing with respect to major distinctions like fiction/nonfiction and adult/youth indicate that our social networks are capturing important information about the underlying social structures of the sampled books. While we do not yet have a way to validate the accuracy of the constructed social networks from local character interactions these results give us confidence that broadly speaking we are capturing meaningful differences in narrative construction. That said, the more subtle differences we observe between different cultural contexts or levels of social prestige may yet be due to measurement error. Future work will want to investigate this more fully.

One further question we investigated was whether interaction-type sub-networks differ significantly from the larger networks in which they are imbedded. Do observational or communicative or other types of interactions lead to structurally different properties that might initiate new theories about the relationship between social interactions and social networks within narratives?

To measure structural equivalence between social networks, we utilized cosine similarity as our primary metric. Structural equivalence traditionally involves assessing the commonality of neighbors between pairs of vertices; however, a simple count of common neighbors does not account for variations in vertex degrees or the broader distribution of common neighbors among other vertex pairs. Cosine similarity addresses these limitations by considering the degree of the vertices and their neighbors.

In our method, we treat the rows/columns of the adjacency matrix as vectors. The cosine similarity between two vertices i and j is calculated as the cosine of the angle between their corresponding vectors. Mathematically, the cosine similarity of vertices i and j is defined as the number of common neighbors divided by the geometric mean of their degrees. This measure produces a value ranging from 0 to 1, where 1 indicates that the two vertices share exactly the same neighbors, and 0 indicates no common neighbors. For vertices with a zero degree, we conventionally set their cosine similarity to 0.

When doing so, interestingly we find no meaningful distinctions between the structural similarity of different types of sub-graphs when compared to the main graphs to which they belonged (Fig. 4). While the rate of different interaction types differed strongly between fictional and non-fictional narratives, for example, the underlying network structures to which they contribute do not.

5 Conclusion

Understanding narratives at large scale is a core concern of the Digital Humanities (Underwood, 2019; Piper, 2018). The social interactions of characters (Zunshine, 2006) and the resulting social networks (Moretti, 2011; Woloch, 2009) have long

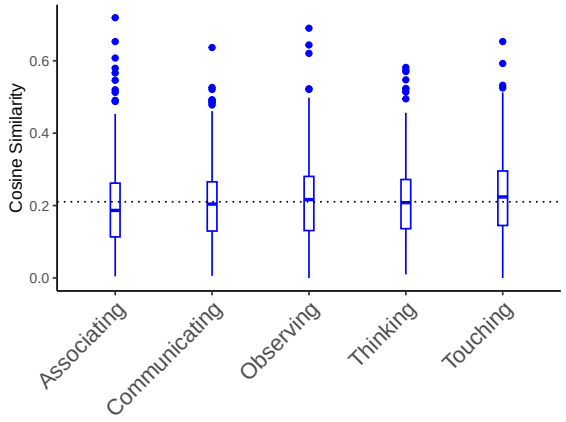

Figure 4: Comparison of structural similarity of subgraphs to main graphs by type. The dotted line represents the global mean.

been theorized as important dimensions of human storytelling. In this paper we have endeavored to illustrate the potential that citizen science has as a means of generating data for training and testing language models towards the goal of understanding the social lives of characters.

With respect to Citizen Science as a mechanism of crowd-sourced text annotation, we find annotation quality on par with trained student annotators. As prior work has suggested, Citizen Science projects achieve the same quality standards as other approaches and bring with them the affordances of a volunteer, community-based approach to scientific discovery (Kosmala et al., 2016; Wiggins and He, 2016). We hope more projects in NLP and DH will utilize this significant resource.

With respect to narrative understanding, we have identified two notable findings in our work. First, fictional narratives strongly favor embodied forms of interactions such as physical contact and sensory perception in support of prior work (Piper, 2024). Second, fictional narratives also strongly favor far denser and less modular social networks. The physicality of relations between characters is amplified by the overall *connectedness* of characters (Mar and Oatley, 2008).

We could add here one additional negative finding: social networks built around individual interaction types do not appear to differ from the overall narrative social networks to which they belong. Social networks built around contact, communication, observation, etc. follow the same patterns as the full network. So while it appears interaction types are useful for distinguishing fictional narratives, they do not contribute much to our understanding of the larger social network structures.

We also highlight a number of areas for future work: both our SLM and GPT indicate significant limitations with respect to the detection of non-interactions, which has also been demonstrated with respect to grammatical models (Agarwal et al., 2013). Much of this can be attributed to the fuzzy boundary around the concept of "interaction" – when two characters are grammatically proximate the rejection of their interaction depends on a number of subtle factors (hypotheticality, co-presence versus interaction, etc.). Future work will want to further explore this boundary in particular.

A second key area is the validation of the social networks themselves. Book-level data on narrative social networks remains a costly endeavor. To date, the field still lacks reasonably sized ground truth when it comes to validating book-level social networks. While we show that our constructed social networks from local interactions align well with theoretical expectations, further validation of their accuracy awaits.

Finally, while we introduce a novel interaction framework in our work, future work will want to think about further nuance with respect to labeling interactions. Our work does not address the valence of interactions, an important property of narrative relations (Smeets et al., 2021), nor does it address overall relationship types (such as kinship or narrative properties such as antagonist, etc.). These too can be valuable frameworks for understanding the structural properties of narratives.

We hope that our publicly shared training data and SLM can be useful tools for researchers to further study the nature of narrative social networks.

Limitations

As we mention above, our work is subject to a number of limitations. First, we note that despite the relatively large size of our training data particular interaction types are significantly less well represented (e.g. observing, thinking, touching). Future work will want to concentrate on expanding our understanding and coverage of those categories. As we also note, while our ability to identify specific types of interactions is high, our ability to distinguish between non-interactions and associations is weak. Future work will want to explore this boundary more fully.

We also highlight that future work will want to

provide book-level annotations of social networks to validate the accuracy of moving from local interaction prediction to global social network modeling.

Another important limitation is the cultural specificity of our data. While our data is drawn from a broad array of genres and a few world cultures, they are limited to the English language. Future work will want to assess cultural differences more deeply with respect to interaction types and social networks.

Ethics Statement

Relying on crowd-sourced labor brings with it important ethical considerations, specifically around fair labor practices and the representativeness of the participating community (Harmon and Silberman, 2019). Citizen Science makes two important contributions to these issues: first, it relies on volunteer rather than paid labor and thus depends on the project-specific interest of participants. Platforms like Zooniverse further contribute to this through the use of About pages, team descriptions, and talk pages where participants can interact with researchers. Participants are far more aware of research goals of a project when it comes to Citizen Science.

In addition to promoting greater project transparency, Citizen Science projects also promote greater researcher-citizen connections, which can help support the democratization of scientific knowledge and facilitate participant learning (Bonney et al., 2016) without sacrificing quality.

We note that while Citizen Science projects can lower the cost of large-scale annotations they do require far more planning and design investment. The initial adaptation of tasks to a particular platform can take time, but we have found that after initial learning projects can take about 2-3 months to prepare for launch. Additionally, because Talk pages are actively used by participants it is essential to have moderators available to handle the volume of queries from users. Nonetheless, all of this can contribute to more transparency and involvement by citizens which is a decidedly positive contribution.

Acknowledgements

The authors wish to thank Robert Budac and Geoffrey Rockwell for their management of the Zooniverse Citizen Readers platform and the Social Sciences and Humanities Research Council of Canada for their generous support of this research.

References

Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014. Frame semantic tree kernels for social network extraction from text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–219, Gothenburg, Sweden. Association for Computational Linguistics.

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1202–1208.

Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1024–1034.

Divya Agarwal, Devika Vijay, et al. 2021. Genre classification using character networks. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 216–222. IEEE.

Mark Algee-Hewitt. 2017. Distributed character: Quantitative models of the english stage, 1550–1900. *New Literary History*, 48(4):751–782.

Mariona Coll Ardanuy and Caroline Sporleder. 2014. Structure-based clustering of novels. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 31–39.

David Bamman. 2021. Booknlp. a natural language processing pipeline for books. `https://github.com/booknlp/booknlp`. Accessed: 2022-01-30.

Rick Bonney, Tina B Phillips, Heidi L Ballard, and Jody W Enck. 2016. Can citizen science enhance public understanding of science? *Public understanding of science*, 25(1):2–16.

Marco Caracciolo and Karin Kukkonen. 2021. *With bodies: Narrative theory and embodied cognition*. Ohio State University Press.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Socientize Consortium et al. 2013. Green paper on citizen science. *Citizen Science for Europe. Towards a better society of empowered citizens and enhanced research. Brussels*.

Milena Dobreva and Daniela Azzopardi. 2014. Citizen science in the humanities: a promise for creativity. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Frank Fischer and Daniil Skorinkin. Social network analysis in russian literary studies. *The Palgrave handbook of digital Russia studies*, page 517.

John Frow. 2014. *Character and person*. Oxford University Press.

Ellie Harmon and M Six Silberman. 2019. Rating working conditions on digital labor platforms. *Computer Supported Cooperative Work (CSCW)*, 28(5):911–960.

Susanne Hecker, Lisa Garbe, and Aletta Bonn. 2018. The european citizen science landscape–a snapshot. UCL Press.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

David Kidd, Martino Ongis, and Emanuele Castano. 2016. On literary fiction and its effects on theory of mind. *Scientific Study of Literature*, 6(1):42–58.

Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

James Lee and Jason Lee. 2017. Shakespeare's tragic social network; or why all the world'sa stage. *Digital Humanities Quarterly*, 11(2).

John Lee and Chak Yan Yeung. 2012. Extracting networks of people and places from literary texts. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pages 209–218. Waseda University.

Pádraig Mac Carron and Ralph Kenna. 2012. Universal properties of mythological networks. *Europhysics Letters*, 99(2):28002.

Aibek Makazhanov, Denilson Barbosa, and Grzegorz Kondrak. 2014. Extracting family relationship networks from novels. *arXiv preprint arXiv:1405.0603*.

Raymond A Mar and Keith Oatley. 2008. The function of fiction is the abstraction and simulation of social experience. *Perspectives on psychological science*, 3(3):173–192.

Raymond A Mar, Keith Oatley, Jacob Hirsh, Jennifer Dela Paz, and Jordan B Peterson. 2006. Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of research in personality*, 40(5):694–712.

Raymond A Mar, Keith Oatley, and Jordan B Peterson. 2009. Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes.

Raymond A Mar, Jennifer L Tackett, and Chris Moore. 2010. Exposure to media and theory-of-mind development in preschoolers. *Cognitive Development*, 25(1):69–78.

Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.

Sandeep Soni David Bamman Andrew Piper Matthew Wilkens, Elizabeth F. Evans. forthcoming. Small worlds: Measuring the mobility of characters in english-language fiction. *Journal of Computational Literary Studies*, 3(1).

Franco Moretti. 2011. Network theory, plot analysis. Technical report, Stanford Literary Lab.

M Nijila and MT Kala. 2018. Extraction of relationship between characters in narrative summaries. In *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR)*, pages 1–5. IEEE.

Alan Palmer. 2004. *Fictional minds*. U of Nebraska Press.

Andrew Piper. 2018. *Enumerations: Data and Literary Study*. University of Chicago Press.

Andrew Piper. 2022. The conlit dataset of contemporary literature. *Open Humanities Data*, 8.

Andrew Piper. 2024. What do characters do? the embodied agency of fictional characters. *Journal of Computational Literary Studies*, 2(1).

Andrew Piper and Eva Portelance. 2016. How cultural capital works: Prizewinning novels, bestsellers, and the time of reading. *Post45*, 10.

Andrew Piper, Hao Xu, and Eric D Kolaczyk. 2023. Modeling narrative revelation. In *CHR*, pages 500–511.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas press.

Mia Ridge. 2016. *Making digital history: The impact of digitality on public participation and scholarly practices in historical research*. Open University (United Kingdom).

Henry Sauermann and Chiara Franzoni. 2015. Crowd science user contribution patterns and their implications. *Proceedings of the national academy of sciences*, 112(3):679–684.

Roel Smeets, Maarten De Pourcq, and Antal van den Bosch. 2021. Modeling conflict: Representations of social groups in present-day dutch literature. *Journal of Cultural Analytics*, 6:1–31.

Saatviga Sudhahar and Nello Cristianini. 2013. Automated analysis of narrative content for digital humanities. *International Journal of Advanced Computer Science*, 3(9):440–447.

Melissa Terras. 2015. Crowdsourcing in the digital humanities. *A new companion to digital humanities*, pages 420–438.

Marcello Trovati and James Brady. 2014. Towards an automated approach to extract and compare fictional networks: An initial evaluation. In *2014 25th International Workshop on Database and Expert Systems Applications*, pages 246–250. IEEE.

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Beate Volker and Roel Smeets. 2020. Imagined social structures: Mirrors or alternatives? a comparison between networks of characters in contemporary dutch literature and networks of the population in the netherlands. *Poetics*, 79:101379.

Michaël C Waumans, Thibaut Nicodème, and Hugues Bersini. 2015. Topology analysis of social networks extracted from literature. *PloS one*, 10(6):e0126470.

Andrea Wiggins and Yurong He. 2016. Community-based data validation practices in citizen science. In *Proceedings of the 19th ACM Conference on computer-supported cooperative work & social computing*, pages 1548–1559.

Alex Woloch. 2009. *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel*. Princeton University Press.

Lisa Zunshine. 2006. *Why we read fiction: Theory of mind and the novel*. Ohio State University Press.

A Appendix: LLM Prompts

A.1 Base Prompt

> **What kind of interaction between char1 and char2?** Choose one of six options: No, Associating, Thinking, Touching, Observing, Communicating.

A.2 Detailed Prompt

> **Task Description:** Classify the type of interaction between char1 and char2 in a given passage. There are six categories of interaction:
>
> **No interaction:** Direct or indirect interaction does not occur between char1 and char2. Any imagination or assumption of interaction also counts as No.
> **Communicating:** char1 and char2 are engaged in some form of communication, such as speaking, writing, or signaling.
> **Associating:** char1 and char2 are linked by a social or relational context, such as friendship, teamwork, or other associative bonds.
> **Observing:** at least one character is observing or watching another one, without direct interaction.
> **Thinking:** at least one character is thinking about or recalling memories of another one, without direct interaction.
> **Touching:** char1 and char2 are engaged in physical touch or contact.
>
> **What kind of interaction between char1 and char2?** Choose one of six options: No, Associating, Thinking, Touching, Observing, Communicating.

A.3 One-shot and Many-shot Prompt

Append examples with passage, char1, char2, and label before the detailed prompt. In the many-shot setting, any shot contains one example from each class.