Title
Automated Conspiracy Theory Detection and Narrative Consensus Tracking in Social Media

Author
Shahsavari, Shadi

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Automated Conspiracy Theory Detection and Narrative Consensus Tracking in Social Media

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Shadi Shahsavari

2022

ABSTRACT OF THE DISSERTATION

Automated Conspiracy Theory Detection and Narrative Consensus Tracking in Social Media

by

Shadi Shahsavari
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2022
Professor Vwani P. Roychowdhury, Co-Chair
Professor Timothy Tangherlini, Co-Chair

A longstanding grand-challenge problem in AI is to build machines that are able to think and interact like humans do. A specific embodiment of this problem is a generalization of the cocktail party problem encountered in signal processing and blind signal separation: If an AI agent were to drop in at a crowded cocktail party then can it separate out and reconstruct the different underlying stories and narratives being discussed from a mixture of fragments of all the on-going conversations. Such a problem has taken on a renewed urgency: Narratives play a defining role in influencing critical decisions and worldviews of both the society at large and individuals, but the continual emergence of a multitude of conflicting narratives –enabled by large-scale adoption of social media– has created a global emergency, where the basic tenets of civil society and governance are being increasingly compromised. These narratives, some of which can be labeled as conspiracy theories, are composed of numerous characters connected by semantically diverse relationships situated in multiple and overlapping contexts. Injecting false facts happens in the context of such discussions, and solving such a misinformation problem is beyond a supervised classification task in natural language processing (NLP). In this dissertation, we develop a pipeline of interlocking computational and statistical modules - based on NLP tools and complex network theories– to extract meaningful narrative networks by distilling millions of social media posts. We develop a framework for semantic parsing of such narrative graphs (e.g. who are outsiders, their motivations and threats, and strategies

of insiders) and evaluate the quality of these automatically derived communities in different ways. We evaluate the quality of these automatically derived communities against domain-expert generated networks, with an average high recall confirming that these communities capture relevant contexts. In the event that such ground truths are absent, we track these communities in reliable news articles. With our close attention to context, the conspiracy theory structures extracted by our pipeline can improve systems for identifying fake news and support near-real-time analysis of these emerging narratives. We show that conspiracy theory narratives tend to glue different domains together with fragile connectivity that are based on some hidden knowledge not validated by civil institutions.

Our next attempt to better understand narratives evolves around reader perceptions of different novels. The rise of social reading sites offers an opportunity to capture a segment of readers' candid reactions to literature, giving automated text analysis tools the opportunity to mine critical insights into how people "read." Posts discussing an individual book on *Goodreads*, a social media platform that hosts user discussions of popular literature, are referred to as "reviews." They consist of plot summaries, opinions, quotes, hypotheticals, analyses, author credits, or a mixture of them about a particular novel. We model the reviewers' shared knowledge of the underlying story that allows us to discover the overall non-professional discussion space about the novel, including an aggregated summary of its plot and the readers' impressions of the main characters. Our entity mentions grouping and interactant relationship clustering methods help us obtain a more accurate reviews-based narrative network and compare it with the expert-generated original plots. We report metrics such as edge detection rate and accuracy by comparing original plots and the reviews'-based networks for four different novels. We further expand the reviews narrative networks and study the non-novel characters discussed on these online forums. We show the complexity of certain characters via our impression extraction method. The impressions are discovered based on descriptive extractions from the reviews and expand beyond sentiments.

To conclude, this dissertation takes a step towards enabling human-smart AI, where it can view the world as a human would. We attempt to understand the narrative and the context behind conspiracy theories from thousands of social media posts. Next, to develop a more

refined understanding of narratives, we study online book readers' reviews and track the book's original plots in the collective narratives extracted from these posts.

The dissertation of Shadi Shahsavari is approved.

Lieven Vandenberghe

Cho-Jui Hsieh

Timothy Tangherlini, Committee Co-Chair

Vwani P. Roychowdhury, Committee Co-Chair

University of California, Los Angeles

2022

To my mom, Shahin, whom I owe my education and life.

To my sisters: Soma, Gashin and Nina who always support me.

And to my teachers and mentors, who believed in me and taught me the joy of learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincerest gratitude and appreciation to my advisor, Prof. Vwani Roychowdhury. I am grateful for his support and patience throughout these years. He always looks for exciting and impactful ideas; I was inspired by his approach to formulating research problems and identifying steps to solve them. I will always cherish our thoughtful and intellectual conversations on various topics. I particularly love my Ph.D. research journey because of its multi-disciplinary subject. My Ph.D. studies would not have been as joyful without fruitful collaborations with Prof. Tangherlini. I was lucky to have the extraordinary opportunity of working with a brilliant expert with a folklore, narrative, and digital humanities background and learn a great deal about research. I sincerely appreciate his patience and invaluable insights that brought more meaning to my research projects. Through our weekly writing workshops, I learned how to present and write my ideas.

I would like to extend my gratitude to my other committee members: Prof. Lieven Vandenberghe and Prof. Cho-Jui Hseih, for taking the time to serve on this committee and supporting my doctoral training.

I would like to especially thank the ECE staff for always helping me and always being there for graduate students. I couldn't get here without Deeona's help and patience. It came as a shock to lose Ryo. I will never forget his warm smile and kindness; may he rest in peace.

At UCLA, I was lucky to work with amazing and talented friends, Pavan, Tianyi, Ehsan, Misagh, Arash, Tonmoy, Yipeng, Qiujing, Ramin, Behnam and many more. I thank them for their support. In particular I am grateful to collaborate with Pavan whom I learned a lot from. It was truly a pleasure to work with him.

I am fortunate to have a support group of friends. Zhowan, Rozhin, Priyanka, Vicky and many more are always there for me, no matter how far they are! I am indebted to my cool Venice Barry neighbors and friends: Shadi, Arezoo, Ali and Alireza, in the past couple of years. Our Wurstkuche hangouts will be a part of cheerful memories of my Ph.D. journey.

Finally, I would like to thank my family and loved ones who brightened my days and sup-

ported me unconditionally. Mom, thank you for putting your daughters' education in your first priority. I owe this to you and your hard work.

VITA

2016 B.S. (Electrical Engineering and Computer Science), Sharif University of Technology, Iran.

2019 M.S. (Electrical Engineering), University of California Los Angeles.

2017-2022 PhD Student, Electrical and Computer Engineering Department, University of California, Los Angeles.

PUBLICATIONS

Tangherlini, Timothy R., Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. "An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web." PloS one 15, no. 6 (2020): e0233879.

Shahsavari, Shadi, Pavan Holur, Tianyi Wang, Timothy R. Tangherlini, and Vwani Roychowdhury. "Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news." Journal of computational social science 3, no. 2 (2020): 279-317.

Shahsavari, Shadi, Ehsan Ebrahimzadeh, Behnam Shahbazi, Misagh Falahi, Pavan Holur, Roja Bandari, Timothy R. Tangherlini, and Vwani Roychowdhury. "An automated pipeline

for character and relationship extraction from readers literary book reviews on Goodreads.com." In 12th ACM Conference on Web Science, pp. 277-286. 2020.

Holur, Pavan, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R. Tangherlini, and Vwani Roychowdhury. "Modelling social readers: novel tools for addressing reception from online book reviews." Royal Society open science 8, no. 12 (2021): 210797.

Chong, David, Erl Lee, Matthew Fan, Pavan Holur, Shadi Shahsavari, Timothy Tangherlini, and Vwani Roychowdhury. "A real-time platform for contextualized conspiracy theory analysis." In 2021 International Conference on Data Mining Workshops (ICDMW), pp. 118-127. IEEE, 2021.

Holur, Pavan, Tianyi Wang, Shadi Shahsavari, Timothy R. Tangherlini, and Vwani Roychowdhury. "Which side are you on? Insider-Outsider classification in conspiracy-theoretic social media." Accepted to ACL 2022

CHAPTER 1

Introduction

Social media have become an inevitable part of modern societies. Every day we receive and propagate information via our interactions on social media outlets. A massive increase in investments in social communication mobile applications and their revenue shows the growth in the number of active daily users. This increase has led to a rise in the amount of information exchanged on these platforms and left us wondering about its effects on our daily life decisions, perceptions, and lifestyles. Processing information on social media has many challenges due to its fragmented nature. A single post from a forum does not reveal the underlying narrative or the whole related debated story. In this dissertation, we seek to bring narrative knowledge to understanding thousands of unstructured social media posts. We provide tools and devise computational methods to overcome this challenge. The contributions can be seen from two different perspectives. In the first part, we investigate narrative network graphs. We describe a generative model and a joint estimation method to estimate its parameters. Conspiracy theory is among the compelling story types circulating on social media. The rise of social media has made online users contribute in small incremental amounts in most cases to such theories. Conspiracy theories are at the core of misinformation and create serious harm. Since the early 1800s, these theories have played a vital role in shaping societies. Conspiracy theories have primarily existed on the fringes of popular culture in the past; however, the advent of social media has prompted a resurgence in their popularity, an outcome that has dire consequences for societies. During the Covid-19 pandemic, we observed the powerful effects of conspiracy theories on the real world that prompted people to engage in dangerous radical medical or political decision-making based on unsubstantiated information. In order to mitigate the social, economic, and polit-

ical consequences of conspiracy theories, it is crucial to understand the behaviors of these theories, primarily how they spread through sources such as social media. As a result, conspiracy theory is an impactful case study for narrative theory in general and our AI models. In section,1.1 we expand upon automated conspiracy theory detection. In the second part, we study literary reviews posted on social media. Narrative theorists have modeled story plots as networks, and following this viewpoint, we look into social media book reviews. We compare their narrative networks to the original book's story plots. We investigate online readers' impressions of various characters in different novels. In section 1.2 we explain our motivation and the challenges behind literary review narrative analysis.

1.1 Conspiracy Theory Detection

Devising computational methods for disentangling misleading stories from the actual facts is a pressing need. Such methods could be used to support fact-checking organizations, and help identify and deter the spread of misleading stories. Ultimately, they may help prevent people from making potentially catastrophic decisions, such as resisting efforts at containment that require participation by an entire citizenry or self-medicating with chloroquine phosphate, bleach, or alcohol. As decades of research into folklore has shown, stories such as those circulating on social media, however anecdotal, are not created from whole cloth, but rely on existing stories, story structures, and conceptual frameworks that inform the world view of individuals and their broader cultural groups [2] [3] [4] [5]. Taken together, these three features (a shared world view, a reservoir of existing stories, and a shared understanding of story structure) allow people to easily generate stories acceptable to their group, for those stories to gain a foothold in the narrative exchanges of people in those groups, and for individuals to try to convince others to see the world as they do by telling and retelling those stories. Inspired by the narratological work of Algirdas Greimas [6], and the social discourse work of Joshua Waletzky and William Labov [7], we devise an automated pipeline that determines the frameworks that form the narrative bedrock of diverse knowledge domains. We also borrow from George Boole's famous definition of a domain of discourse, recognizing

that in any such domain there are informal and constantly negotiated limits on what can be said:"In every discourse, whether of the mind conversing with its own thoughts or of the individual in his intercourse with others, there is an assumed or expressed limit within which the subjects of its operation are confined" [8]. We conceptualize a narrative framework as a network comprising the actants (people, organizations, places, and things) and the interactant relationships that are expressed in any storytelling, be it a journalistic account or an informal anecdote [9] [10]. In our model of storytelling, individuals usually activate only a small subset of the available actants and interactant relationships that exist in a discourse domain, thereby recognizing that individual storytelling events are often incomplete. This story's incompleteness presupposes knowledge of the broader narrative framework on the part of the storyteller's interlocutors.

Building on folkloric work in rumor and legend, we further recognize that a large number of the stories circulating on and across social networks have a fairly straightforward "threat narrative" structure, comprised of *orientation* (the who, what, where, and when), a *complicating action: threat* (identifying who or what is threatening or disrupting the in-group identified in the orientation), a *complicating action: strategy* (a proposed solution for averting the threat), and a *result* (the outcome of applying that strategy to the threat) [5]. To determine the extent of narrative material available—the actants and their complex, content-dependent interactant relationships—we aggregate all the posts or reports from a social media platform. For social media, in particular, we recognize that participants in an online conversation rarely recount a complete story, choosing instead to tell parts of it [11]. Yet even partial stories activate some small group of actants and relationships available in the broader discourse. We conceptualize this as a weighting of a subgraph of the larger narrative framework network.

1.2 Online Readers Perception Analysis

In our search to evaluate our extracted narrative graphs and expand our consensus tracking developments, we study underlying narrative graphs from book reviews posted on social media platforms. Online reader comments about works of literary fiction offer an intriguing

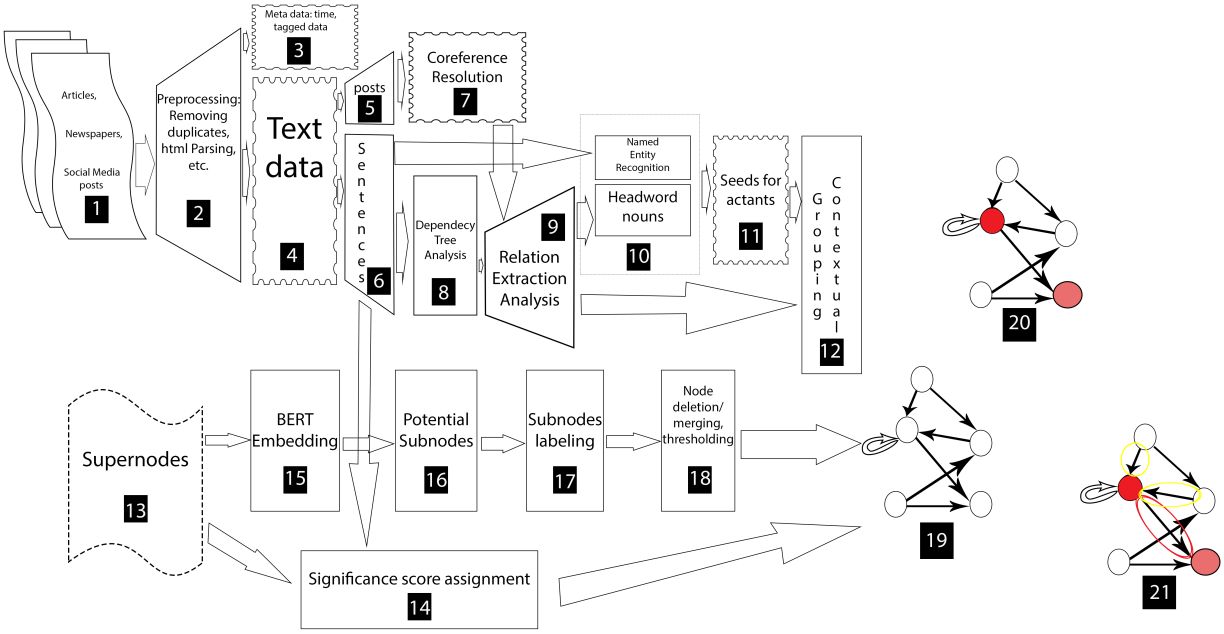

Figure 1.1: **Representation of the narrative framework discovery pipeline.** The numbered blocks are described in the chapter 3.

window into how people read. Although previously ignored in the realm of computational literary studies, recent studies have shown the value of these resources in understanding reader reception [12–14]. These comments can provide useful insight into how readers imagine the main storylines of a novel, how they understand the fictional struggles of characters, and how they develop varying impressions of the work. Taken together, the reviews of a single novel provide a view onto the collective imagining of what is important in the novel, including aspects of the plot, the interactions between various characters, and even the metadiscursive space of authors, genre, critics, film adaptations, and movie stars. These reviews thus provide the impetus for a data-driven analysis of readers' responses to a work of literary fiction [15]. They also help us understand how readers create an "imagined community" of readers, an extension of Fish's notion of "communities of interpretation", engaged in the collective enterprise of literary analysis [16, 17].

"Reader response theory" experienced a brief and productive heyday in literary theory during the 1960s. Despite the considerable attention this theoretical premise received, the focus of

much of this work centered on the hypothetical and highly theorized "individual reader". This theoretical orientation was expanded to include groups of readers, and led in part to Stanley Fish's important contributions concerning "communities of interpretation"–groups of readers who, through their shared experiences, converged on similar readings of texts [17, 18]. The consideration of broad-scale responses of readers to works of fiction, however, remains understudied, not because of a lack of interest on the part of literary historians and theorists, but because of a lack of access to those readers' responses, and a lack of methods to address this at times noisy data. While there is considerable investigation into how groups of individuals are likely to read (or to have read), investigations of how large groups of people outside of an experimental setting respond to the same work of fiction have only recently become possible [19]. The advent of social reading sites on the internet that allow individual readers to join in wide-ranging discussions of individual works of fiction through reader-generated reviews and comment threads on those reviews has enabled a revisiting of fundamental questions of how people respond to literary fiction [12, 15, 18, 19].Perhaps best known among these sites in the United States is *Goodreads* that, along with sites like it, represents an online attempt to reproduce the face-to-face space of book clubs and library groups, where there is no "right" answer to reading the work (as there might be, at least implicitly, in a classroom), nor any hierarchy of critical insight (as there might be in a forum where professional reviewers or literary critics might dominate the conversation) [18, 20]. Because these sites archive the reader reviews and the ensuing comment threads on those reviews, they offer an opportunity to explore computationally how people respond to individual works of fiction, and how they explore such a work as communities of interpretation emerge [15]. Since these reviews are unguided explorations of fiction, and since many of the readers read and review purely for entertainment, it is unlikely that the readings encode the types of literary-theoretical engagement found in academic work. Instead, the reviews encode a popular engagement with literature, focusing on aspects of plot, character, and the struggles of the characters in their fictional worlds. Our goal in this part is to model the collective expressions of thousands of readers as they review the same work of fiction. Can we discover what they see as important? Can we discover divergences in their readings? Do

their reviews provide us with information about reading, remembering, and retelling? And do they tell us anything about the process of writing a review itself?

In our work, we assume that we are given thousands of user reviews of a particular novel from a social cataloging/review website such as Goodreads.com. Given such a corpus, we ask the following questions: (i) Can one *automatically discover all the primary actants* as well as meta-actants (authors, actors and actresses from film adaptations, etc.) that are mentioned across all of the book reviews for a given novel? (ii) Can one also *discover and meaningfully cluster all the inter-actant relationships* that these reviews include? The results of goals (i) and (ii) provide, when properly thresholded and weighted, a representation of the consensus model of the novel as perceived by those readers who review the book. Inspired by the actantial narrative model noted above, we represent these results as an automatically generated narrative network, where nodes are actants and edges are directed multi-edges annotated with the extracted relationships. (iii) Finally, *given an expert generated ground truth narrative network*, can one *automatically compare that ground truth network with the auto-generated summary narrative framework network* and compute meaningful metrics such as recall and precision?

Solving the above problems is tantamount to developing a view of the reviewers' consensus about a target novel, as readers recollect and review the actual cast of actants and their inter-actant relationships.

Following the approach developed for our conspiracy theory work, we further attempt to study the underlying narrative graphs extracted from this corpus. Reader reviews, in the aggregate, constitute a collective process that converges on an underlying, yet broad, narrative framework [18]. This framework is represented as a narrative network where the nodes are actants and the edges are interactant relationships. The actants in the network are expanded beyond the canonical census of a novel's characters to include the metadiscursive space, populated by actants such as the author, filmatizations, film directors, actors and other extra-diegetic features of the work, since reader reviews often spill into discussions of other works by the same author, genre, and filmic interpretations of the novel.

Importantly, book reviews also encode the varying impressions that readers form of the novel's characters. By reading reviews at "internet scale", one can consider the extent to which partial reviews contribute to a representation of the character-interaction network of the target novel and how readers build a collective understanding of complex characters, even if their individual views of the characters may not capture that same complexity. These problems can be seen as part of a formal, computational assessment of readers' response to even quite long and complex novels, such as *The Hobbit* or *To Kill a Mockingbird*, to name but two of our target works.

The more often that an actant or relationship appears in the corpus, the more heavily it is weighted in the network graph. Importantly, the related methodologies presented here can be extended well beyond the realm of literary fiction to derive narrative frameworks undergirding nearly any collection of documents. We focus on literary fiction because of the unusual (for cultural datasets) presence of a ground truth against which to measure the accuracy of our results.

To address these challenges, we extract and aggregate a meaningful representation of the reader-generated shared narrative framework modeled as a network. This framework is based on a structured open-world infinite-vocabulary network of interconnected actants and their relationships. We introduce SENT2IMP, which presents a representation of the collective, at times differing, opinions of characters in a novel. In addition, we expand the narrative framework graph to include the important metadiscursive and extra-diegetic nodes noted above, thereby providing a fuller picture of how readers contextualize their engagement with a work of fiction.

1.3 Outline and Summary of Contribution

In chapter 2 we introduce our generative narrative model. We represent a graph that consists of characters and their relationships. We explain the intuition behind these elements and how they all form the narrative graph. We introduce a graph-based algorithm, Entity Mention Grouping (EMG) that helps us discover characters based on their role in narrative networks.

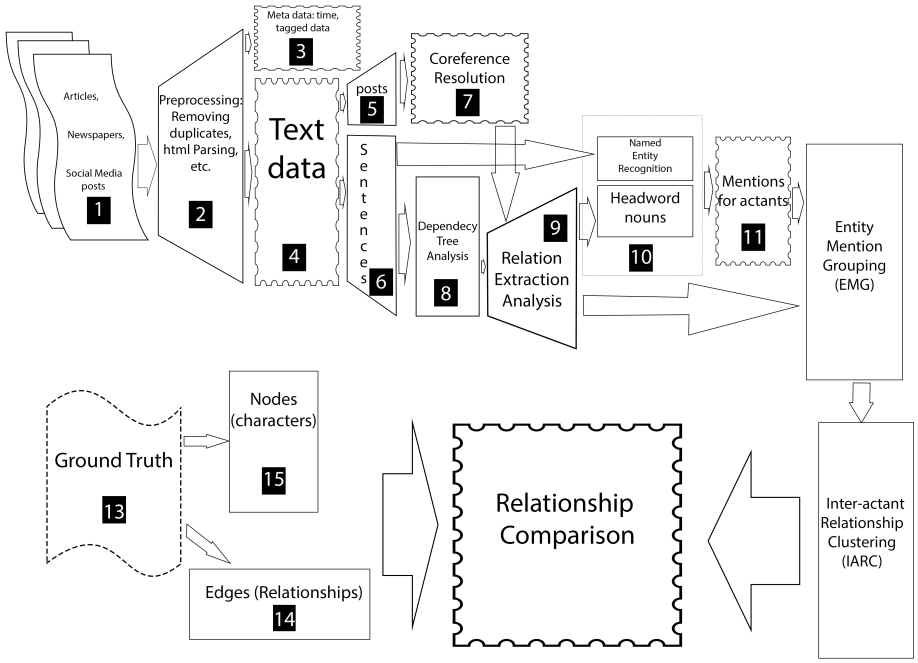

Figure 1.2: **A representation of the literary review analysis flow chart. Expert generated ground truth is compared with narratives extracted from unstructured social media posts.**

EMG runs after we extract the narrative graph. It looks into the nodes' roles in the narrative graph and discovers the characters called with different mentions. Given the expert-generated novel plots, we devise an unsupervised method to compare them with our extracted graphs. The increase in the accuracy after applying EMG proves its success in determining the characters. In chapter 3, we propose computational methods to extract the narrative graphs that include sentence-level relationship extraction, embedding-based supernode/subnodes derivation, graph formation, and a statistical method to mine multiple stories. We propose a threat detection method to find critical nodes in our graphs given the narrative graph. A threat node is one of the critical elements in the narrative model of a conspiracy theory. This classification problem opens the door to perform supervised machine learning based on our narrative graphs. Figure 1.1 represents a flow chart of our computational modules to derive narratives from social media. We provide the detailed implementation of the EMG algorithm and evaluation methods further in chapter 3. We demonstrate our computational results in three different chapters. In chapter 4, we study two datasets. The first dataset is about the Pizzagate conspiracy theory, and the second is the Bridgegate conspiracy. We derive the narrative networks based on the models earlier described in chapter 2. We demonstrate how a conspiracy theory consists of multiple loosely connected layers. In chapter 5 we perform the narrative extraction on a dataset on Covid-19 conspiracy theories, and we disentangle a narrative graph made from multiple conspiracy theories. We present our computational results comparing news-generated communities and our discovered stories. We report the homogeneity and completeness comparing two sets of communities. In chapter 6 we study the literary book reviews posted on social media. We develop EMG and interactant relationship clustering discussed earlier in chapter 2. We show improvements in the edge detection accuracies after applying EMG on four different novels. This algorithm relies on the semantic role of each mention in the extracted narrative graph. We compare the narrative graphs extracted from social media with the original plot. While individual reviews might not tell the whole story and may, on the individual level, fail to capture the complexity of characters, the collective impressions of thousands of readers provide essential insight into how people read, remember, retell and review. These methods allow us to do many things, including

reassembling a portrait of a tortured scientist and his monster.

CHAPTER 2

Narrative Extraction from Social Media

Our goal is to study underlying narratives existing on social media. To this end, we describe a generative story model that narrative theory studies have inspired. We devise methods to extract the model parameters from thousands of social media posts. In our next attempt to evaluate such story models, we use available ground truth story graphs to compare and study social media posts. The works done in this dissertation do not fall entirely under any of the standard natural language processing (NLP) tasks. Instead, we have used multiple NLP tools such as Named Entity Recognition, Stanford Corenlp, and Semantic Role Labeling to develop an interlocking pipeline to achieve our goal. In our attempt to study conspiracy theories, we used narrative graph models from narrative theory and developed a joint estimation approach to estimate such graphs. We start from a single conspiracy theory problem with a case study on the Pizzagate conspiracy theory and Bridgegate conspiracy. Next, we expand our work to study the multi-conspiracy theory problem where we extract multi-story narratives from a single dataset. Finally, our novel approach to study readers' responses contains new approaches to extract impressions toward different characters. We develop an algorithm to resolve multiple entities assigned to a single character, and later on, based on our relationships extractions, we examine each character in readers' eyes.

2.1 A Graphical Narrative Model for Generation of Posts and Story Fragments

We present a generative graph model that consists of a set of nodes and edges. The nodes in our model are called **actants** (people, places or objects), and the edges are **the relation-**

ships between pairs and groups of actants. These edges/hyper-edges are labeled with the nature of the observed relationships (for example, based on actions or attributes), the context of the relationships, and their likelihoods. In such networks, the actant categories or types are usually predefined, such as persons, organizations, and places. Similarly, different attributes and relationships among the actants are usually chosen from predefined attribute lists.

Our graphical models, by way of contrast, are primarily aimed at capturing actants and the interactant relationships which emerge under specific circumstances and situations, and that are driven by an underlying narrative framework. They are particularly suited for representing story and narrative dynamics where the overarching structure does not vary much, but the specific instances of the actants, their roles, and their relationships vary significantly based on the circumstances.

The narrative graph is characterized by a set of n nodes representing the actants, a set of r relationships $R = \{R_1, R_2, \ldots, R_r\}$ defining the edges, and k contexts $C = \{C_1, C_2, \ldots, C_k\}$ providing a hierarchical structure to the network. These parameters are either given *a priori* or estimated from the data. A context C_i is a hidden parameter, or the 'phase', of the underlying system that defines the particular environment in which the actants operate. It expresses itself in the distributions of the relationships among the actants, and is captured by a labeled and weighted network $G_{C_i}(V_{C_i}, E_{C_i})$. Here, $V_{C_i} \subseteq \{A_1, A_2, \ldots, A_n\}$, where each A_j is an actant. The edge set E_{C_i} consists of m_{C_i} ordered pairs $e_{C_i,j} = (A_{j_1}, A_{j_2})$, where each such pair is labeled with a distribution over the relationship set R.

Relationships are represented by categories of words (most often verbs) grouped together, where each category is comprised of verbs that imply a similar relationship.

Each post to a forum describes relationships among only a subset of actants (which are yet not known to our automated algorithms). To write a sentence, a reviewer first picks a context $C_i \in C$ and then samples an underlying context-dependent network $G_{C_i}(V_{C_i}, E_{C_i})$ (to be estimated by the algorithm) by drawing a pair of actants (A_k, A_j) according to a conditional actant recall distribution across all the actants, $p_{C_i}(A_j)$. A context could represent

a particular situation in the plot. For example, when someone wants to recount the scene in *Frankenstein* where Dr. Frankenstein creates the monster, then certain actants and relationships are described much more often than others. Following this, the reviewer draws a relationship for the pair (A_k, A_j) from a distribution associated with the context-dependent edges: $D_{(E_{C_i},(j,k))}(\mathcal{R})$. The writer then composes the review according to these outcomes by choosing the proper words and syntax. In particular, the reviewer chooses noun phrases (as mentions of the actants A_j and A_k) and the associated verb/relationship phrases (or other syntactical constructs) for the sampled relationship.

Problem Statement

In this section, our goal is to extract the graphical model $G_{C_i}(V_{C_i}, E_{C_i})$ from a set of social media posts following our generative model assumptions. We aim to understand different layers of narratives along with evaluating the networks by comparing them with available ground truth graphs. First, using sentence level extractions we transform a social media post, p, into a set of triplets,

$$\mathcal{M}_p = \{(m_l, r_{l,k}, m_k)\}$$

in which m_l and m_k are representations of actants and $r_{l,k}$ is relationship instant in the R. Given a set of social media posts along with their relationships we jointly estimate narrative parameters:

$$V = \{A_1, A_2, \ldots, A_n\},$$

$$R = \{R_1, R_2, \ldots, R_r\},$$

$$C = \{C_1, C_2, \ldots, C_k\}$$

and the underlying graphical structure:

$$\forall C_i \;\; \exists G_{C_i}(V_{C_i}, E_{C_i})$$

2.2 Context-Based Character Detection and Evaluation

In the second part of this dissertation; we study book reviews posted on social media platforms. Based on our generative model, we study the underlying narrative network. A network of characters (and other actants) interconnected by their relationships can serve as a useful representation of an aggregated model of readers' understanding of the narrative scope of a novel. This model has the advantage that it can show multiple, at times competing, claims to the underlying story line (or story lines) of the target work. We introduce a pipeline addressing two important tasks: Entity Mention Grouping (EMG) and Inter-Actant Relationship Clustering (IARC). Often we have multiple mentions/noun-phrases for the same actants, and multiple semantically equivalent relationship phrases to describe different contexts. In order to accurately estimate the different contexts C_i, actant frequency distributions $p_{C_i}(A_j)$, and the relationships $D_{(E_{C_i},(j,k))}(\mathcal{R})$, we must aggregate the different mentions of the same actant into a single group. In order to do that, we need to consider relationships: two mentions refer to the same actant only if the key relationships with other actants are semantically identical. Thus, the estimations of entity mention groups and relationships need to be done jointly. Our algorithm called Entity Mention Grouping (EMG) is used to recognize actants in a narrative graph. The resulting graph constitutes an end-state ranked consensus model of all actants and relationships.

Assuming we have a ground truth narrative graph, the evaluation of our results focuses on the similarity of the ground truth and learned narrative graph based on a matching of actants and their contextual relationships. The frequency distributions of the actants, p, and relationships, D, can be estimated based on the counts of the occurrences of the associated groups of phrases. We use a threshold to decide whether an actant or a relationship is included in the consensus narrative graph.

The EMG task is a labeling process that aggregates multiple entity mentions from the extracted relationship phrases (subject $\widehat{s}$ or object $\widehat{o}$) into a single character. This aggregation is accomplished through an evaluation of the similarity between a pair of entity mentions by observing their interactions with other entity mentions. For example, in *The Hobbit*, the

two entity mentions, "Bilbo" and "Baggins", frequently interact with the entity mention "Gandalf", and with semantically equivalent relationships; as a result, these two mentions, "Bilbo" and "Baggins", likely refer to the same character, here the hobbit, "Bilbo Baggins".

To formulate this task, let the set of entity mentions empirically observed in reviews be $\widehat{E}$. A smaller character set E refers to a finite vocabulary of distinct characters in a literary work. The EMG step is then defined as a surjective function $f : \widehat{E} \rightarrow E$ that maps entity mentions to characters. The resultant mapping of entity mentions to characters in the EMG task provides a semantically-informed aggregation tool for the original corpus of relationship tuples. Tuples sharing entity mentions mapped to the same character can now be aggregated to form larger relationship sets between a pair of characters (as opposed to a pair of entity mentions). For example, in *Of Mice and Men*, the entity mentions "George" and "Milton" are successfully mapped to the character "George" with the EMG task.

The IARC task, on the other hand, is designed to aggregate relationship phrases generated as output from a relationship extraction module. A relationship tuple consists of a subject mention, relationship phrase and object mention $(\widehat{s}, \widehat{r}, \widehat{o})$ that is directly extracted from a review and thus contains partial information about the structure of the underlying narrative model [21]. For every ordered pair of characters $\{e_i, e_j\}$, we obtain an aggregated set of relationship phrases from the reviews, $\widehat{R}_{ij}$, that connects e_i to e_j.

The EMG task implicitly aids in the aggregation of larger sets of relationship phrases since it aggregates entity mentions for each character. $\widehat{R}_{ij}$ is the union of all the relationship phrases between entity mentions that compose e_i and e_j. We seek to cluster these relationship phrases in $\widehat{R}_{ij}$ and assign each resulting cluster to a label in a set R_{ij}. This process of clustering, Inter-actant Relationship Clustering (IARC), can accordingly be defined by another surjective function $g_{ij} : \widehat{R}_{ij} \rightarrow R_{ij}$. Specific details about assembling the set of labels R_{ij} for the IARC task are provided in the relevant subsections below. For example, a few of the relationship phrases between "Atticus" Finch and "Tom" Robinson in reviews include: defends, is defending, protects, represents, supports. These phrases are semantically similar and appear in the same cluster; this cluster is subsequently labeled "defends".

After we extract the narrative networks, which we label "narrative frameworks", represent the broad consensus across all the reviews of the story network, with each node in the network representing a character and each directed edge representing a relationship between a pair of characters. Importantly, the narrative framework graph is derived entirely from the reader reviews.

In addition to estimating the narrative framework of a novel, which may reflect the readers' understanding of important characters and inter-character relationships, it is useful to estimate a larger graph that also includes key, often metadiscursive or extra-diegetic, actants related to the novel's *reception* such as the "author", a "reviewer", or even references to cross-media adaptions such as the "film director" and "actors" who played particular characters in the film. Similarly, while earlier work generated story network frameworks that were static, summarizing the entire story, estimating the temporal dynamics of the story underlying the reviews can greatly expand the usefulness of these graphs. This dynamic view of the novel can provide insight into how reviewers remember the unfolding of the narrative. Last, while the narrative graphs highlighted the interactions *between* actors, they did not model the reviewers' *impressions* of the actors, important information in the context of an analysis of reader response.

Problem Statement

We develop a character detection algorithm, EMG, to create a higher level understanding of narrative networks. We show that EMG improves the evaluation metrics by a significant increase in the accuracy. This accuracy is obtained through comparing the expert generated ground truth networks and our narrative graphs.

CHAPTER 3

Narrative Extraction Methods

3.1 Learning Narrative Structure from Large Scale and Unstructured Text Data

Our methodology relies on the underlying structure of the narrative framework that captures how a storytelling instance emerges via a collective negotiation process. Each post to a forum describes relationships among only a subset of actants (which are not yet known to our automated algorithms). We have described the posts generation process earlier.

From a machine learning perspective, given such a generative process, we need to estimate all the hidden parameters of the model, including the actants, the set of relationships, and the edges and their labels. This joint estimation provides us all of the parameters of the different layers of such a model.

3.1.1 Joint Estimation of Actants, Contexts, and Relationships

We assume that the given corpus is a sample syntactic output of our graphical generative model. The underlying sets of actants, their semantic relationships, and the contexts that determine different groups of relationships among the same actants are unknown. Thus, we need a formal data-driven function/measure to characterize these sets so that they can be estimated from the text corpus. In the sections below, we describe our steps to extract the actants or the nodes in our narrative graphs. Once these actants are recognized, we create our graphs using the extracted relationships from the corpus. We estimate the contexts as dense parts of the whole narrative graph. We deploy community detection algorithms to

infer the contexts or individual stories.

3.1.2 Actants

Actants in our narrative graph can in practice be described as follows: *a set of Noun Phrases or argument phrases (e.g., named entities and head words in a parse tree) that play almost the same semantic roles in the corpus.* The semantic role of a noun phrase is measured by the semantic similarity of the words and phrases around it in the parse tree. Our goal is to derive these actants from the thousands of social media posts. In the subsections below, we describe how we start from relationship extraction and use different techniques to estimate the actants.

3.1.3 Relationship Extraction

Each post comprises a set of a few sentences. Each sentence in the text corpus is processed to extract specific patterns of syntax relationship tuples in the form of (arg_1, rel, arg_2) where arg1 and arg2 are noun phrases, and rel is a verb or other type of phrase.

Our relation extraction utilizes dependency parsing tree and Semantic Role Labeling (SRL) [22] [23]. A similar, albeit more limited, approach to actant-relationship extraction is described by Samory and Mitra in their work on conspiracy theories [24]. In that work, their goal is to cluster semantically similar agent-action-target triplets, manually label the clusters, and align those labeled clusters with a manually curated topic model of a broader target corpus [24]. As opposed to limiting our extractions to agent-action-target triplets, we design a set of patterns (such as Subject-Verb-Object (SVO) and Subject-Verb-Preposition (SVP)) to mine extractions from dependency trees by using the NLTK package and various extensions [22, 25–31]; The patterns are based on extensions of Open Language Learning for Information Extraction (OLLIE) [32] and ClauseIE [33]. Next, we form extractions from the SENNA Semantic Role Labeling (SRL) model. We combine dependency-based extraction techniques with SRL to increase the recall of our system. A list of all the syntax relationship patterns, their definitions, and related examples are provided in the GitHub link for our

research. Following these steps, we apply cleaning and de-duplication techniques to select unique and high precision extractions. Relationship tuples scraped from reviews only include those entity mentions that match or exceed a frequency lower bound.

We process every post in our corpus and split it into sentences using the NLTK Python package. We extract relationships and aggregate all the noun phrases across the entire corpus to derive underlying actants. This aggregation process (based on the generative model of narratives) also takes into account contextual differences, where the relationships between actants change in different situations. Such corpus-level structure cannot be inferred by simply extracting relationships seen in sentences. In our approach, syntax-based relationships, such as SVO (subject, verb, object), are tuned to capture story-specific syntactic forms of expressions. For example, to fit our generative model, we often break up three-way relationships into multiple pairwise relationships: a sentence, such as "They're setting up these 5G towers that will control us." is broken up into two pair-wise relationships: : (They , 're setting up , these 5G towers) and (these 5G towers , will control , us); as illustrated in Figure 3.1. Because arg1 and arg2 could be pronouns, it is important that we determine to which nouns or noun phrases these pronouns refer. Since pronouns often refer to nouns in preceding sentences, we use groups of sentences belonging to the same post as input to a co-reference tool (Stanford corenlp package. We apply the output maps (pronouns resolved to nouns) to replace the resolved pronouns in the noun phrases, arg1 and arg2, with their corresponding nouns. As a result of this process, corresponding to block number 7, a major fraction of the pronouns are replaced by nouns. The input to this block is posts, and the output is used in block 9.

3.1.4 Supernode and Subnode Discovery

Every arg_i where $i \in \{1, 2\}$ is a semantic representation of A_j where $A_j \in \{A_1, A_2, \ldots, A_n\}$ and n is the number of actants (nodes) in the underlying narrative graph. Let us assume we have $\{m_1, m_2, \ldots, m_M\}$ where every m_i is a noun phrase and has appeared at least once in the extracted relationships as arg1 or agr2.

Our goal is to find a mapping function f where $f : m_i \rightarrow A_j$. This function associates each mention to an actant in the final underlying graph.

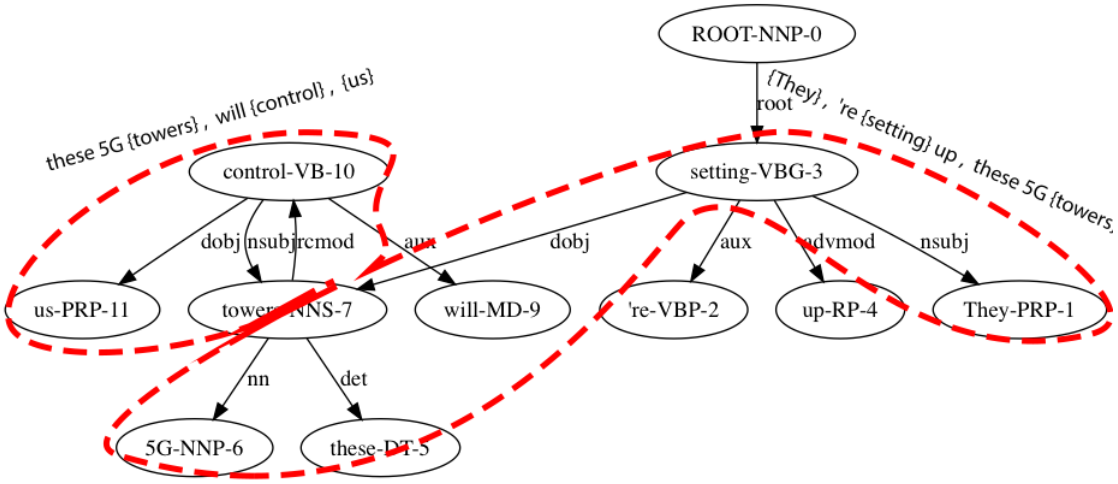

Figure 3.1: **An Example of syntax-based relationship extraction patterns:** The sentence, *"They're setting up these 5G towers that will control us."* is analyzed to extract two relationship triples. These relationships are then aggregated across the entire corpus to create the final narrative network.

For example, (i) phrases such as "Bill", "Gates", "Bill Gates" form one actant category because of their high frequency, both as individual "head" words, and as co-occurring words in noun-phrases. As per our intuitive definition of an actant, because they are part of the same arguments in syntactic relationships, they have similar semantic roles; (ii) phrases such as "Supporter of Clinton", "Clinton follower" and "Clinton insiders" form a distinct semantic context because of the close semantic similarity of the words, Supporter, Follower, and Insider; (iii) phrases such as "Clinton Foundation", "Clinton Foundation Fund raising", "Clinton Donor" and "Clinton Foundation Contributions" form yet another distinct actant context because of the semantic similarities of the words Foundation, Fund Raising, Donor, and Contributions. These examples guide not only the automatic determination of actants, but also reveal that the actants themselves have a hierarchical structure based on the different semantic and contextual roles they play. The phrases in (i) dealing with the different contexts for the actant Hillary Clinton can be considered a super-actant or a **supernode**, and the phrases in (ii) and (iii) dealing with different facets and distinct roles that are associated

with the actant, Hillary Clinton, can be considered sub-actants or **subnodes**. The subnodes are specific semantic contexts that directly relate to the supernode and are expected to have relationships that are semantically homogeneous with the rest of the actant groups.

Historically, the semantic and functional similarity of words has been difficult to compute. In the past, these similarities were manually cataloged in dictionaries, thesauruses, and manually created databases such as WordNet and VerbNet. Recent advances in data-driven methods of embedding words and phrases into a multidimensional vector space [34] [35] such that their Euclidean distances have correlations with their semantic similarity have made it possible to assign a quantitative measure to the similarity metric. The embeddings of syntactic argument phrases can be clustered with each cluster representing a separate actant. As we demonstrate in our results, this procedure of clustering embeddings of relationship phrases nearly automates the process of jointly estimating the actants and their attendant hierarchies.

Figure 1.1 provides a flowchart of the computational steps executed in our end-to-end pipeline. The salient computational steps are described below.

Actant Discovery (Blocks 10 **through** 18 **in Fig. 1.1):** Formally, let $\mathcal{P}$ be the set of all relationship and noun phrases (i.e. all phrases, arg1, arg2, and rel occurring in any syntactic extraction (arg1, rel, arg2)). We define an embedding mapping $\mathcal{E} : \mathcal{P} \rightarrow \mathcal{R}^n$, that maps the set of phrases to a real vector of dimension n. Given any phrase, $\mathcal{P}_i \in \mathcal{P}$, $\mathcal{E}(\mathcal{P}_i) = \mathbf{y_i} \in \mathcal{R}^n$ (and without loss of generality we assume $\|\mathbf{y_i}\| = 1$). Moreover, the mapping $\mathcal{E}$ is such that if $\mathcal{P}_i$ and $\mathcal{P}_j$ are semantically close phrases (i.e., they semantically mean almost the same even if they do not use the exact same words), then their corresponding embeddings must satisfy $\|\mathbf{y_i} = \mathbf{y_j}\| \approx 0$. This requirement enables an unsupervised approach to actant determination: One can cluster the embedding vectors to obtain semantically close actant groups.

A direct approach to this problem might be to take all the noun phrases (i.e., arg1 and arg2) and their contextual embeddings using BERT [36], and cluster them using an algorithm such as k-means to obtain actant candidates. These clusters could then be further processed to merge very similar clusters to form a combined larger actant group, or to delete clusters

that are not meaningful enough or are too heterogeneous (for example, as measured by the entropy of the word distributions in the phrases clustered together). This direct approach suffers from two major drawbacks: (i) The noun phrases, even after resolving pronouns/co-references, are dominated by high frequency pronouns, such as "they" "I" and "she", or not so meaningful online terminology, such as "URL". This ambiguity results in large clusters comprised of high-frequency but irrelevant actant groups, while more relevant actant groups get merged together to form heterogeneous clusters. (ii) The current embedding techniques tend to be flat (i.e., there is no inherent hierarchy in the vector space in which the words and phrases are embedded) and thus the example of the "Hillary Clinton" supernode and the subnodes related to "Clinton Foundation" and "Clinton Campaign" cannot be easily replicated. The above observations motivated us to adopt a two-step process: (i) Contextual grouping of high frequency entities and concepts to create **supernodes**: We first create a ranked list of named entities and concepts. Then we define a supernode as a context consisting of all the argument phrases that have a limited but unique and highly-correlated subset of the entities/concepts as substrings. In the Pizzagate corpus for example, we find all phrases with any of the following words {Clinton, Hillary, Hillary Clinton} as one single supernode. Similarly, we find {Pizza, Comet, Ping, Pong} as the seed words for another supernode. Thus a supernode defines a general context, which can be further divided into subactants or subnodes as described below. (ii) Embedding vectors to cluster arguments in a supernode to create **subnodes**: Once we have defined meaningful contexts, we cluster the embeddings of the phrases belonging to a supernode to create subnodes.

Determining Supernodes (Blocks 10 **through** 13 **in Fig. 3.3):** After retrieving syntax extractions from the corpus sentences, we generate and rank a list of entities, which is then used to form the seeds for potential actants. The ranking is based on the frequency of occurrences of the entities in the noun phrases arg1 and arg2. This ranking consists of both named entities as well as concepts such as "closures" and "email". For Named Entity Recognition (NER), we use the Flair framework [37], a character-level neural language model for contextualized string embeddings, along with the Flair pre-trained model. We limit the

candidate actants to nine main types. For concept discovery, we create a ranking of the frequent headwords in the noun phrases, arg1 and arg2. This method provides a second ranking of headwords including non-named entities. We then combine the two rankings, and rank each entity according to the summation of its frequency in the two lists. The list can be truncated to delete all nodes below a certain frequency threshold. The truncated list constitutes the original list of all entities/concepts to be considered for creating supernodes.

The subset of entities/concepts that define a supernode is computed in a hierarchical fashion: (**Step-0:**) The current entity/concept list is set equal to the original list. The maximum number of seed nodes in a supernode is set to k. (**Step-I:**) If the current list is empty, then Quit (supernode construction is complete). Otherwise, select the highest ranked entity/concept in the current list (in the first iteration, the entire original list is the current list). Let this entity be E_1. Add E_1 to the list of seed nodes for the new supernode, S. Remove E_1 from the current list. Set the seed-node list size, $|S| = 1$. (**Step-II:**) Find all phrases/arguments where any of the seed nodes in the set S (i.e. the set representing the supernode under construction) appears as a sub-string, and let this be called $\mathcal{P}$. (**Step-III:**) Compute the most frequent entity/concept in the original list (other than the seed nodes already extracted) in $\mathcal{P}$. Let this be E. (**Step-IV:**) If E has been processed before (i.e., it is no longer in the current list), then jump to **Step-VI**. (**Step-V:**) If E is in the current list, then add it to the list of seed nodes, S. Remove it from the current list of entities/concepts. Increase the size count, $|S| = |S| + 1$. If $|S| = k$ (where k is the maximum size of the supernode seed list S), then go to Step-VI. Otherwise jump to **Step-II**. (**Step-VI:**) The current list of seed nodes, S, is the new supernode. Return to **Step-I** to start creating a new supernode.

Subnode Creation and Labeling (Blocks 15 through 18 in Fig. 3.3): Each supernode represents a meaningful context, and is defined by its set of argument phrases. For each phrase we compute a BERT embedding [36] and cluster the embeddings of the phrases via k-means clustering, chosen for its simplicity, interpretability and suitability for

our data [38] [39]. Since supernodes have varying sizes (i.e. different supernodes have larger or smaller number of argument phrases), it is a computationally involved task to optimize k, the number of clusters for each supernode. In order to avoid such customization, we fix a single value of k (for both Pizzagate and Bridgegate, we picked $k = 20$) for all supernodes and then delete insignificant clusters or merge two very similar clusters as follows: (i) **Deletion of small size clusters**: For each supernode, we plot the size distribution of the k clusters, and we find that a certain percentage always has significantly smaller size than the average. Therefore, we define a threshold based on the ratio of the size of a cluster and the average size of the clusters for that supernode; all clusters with a ratio below this threshold are deleted. The rest of the larger clusters are processed as potential subnodes. (ii) **Merging of very similar clusters:** For each cluster, we generate a ranked list of the words that appear in the phrases that define the cluster. The ranking is based on a TF*IDF score, where TF is the frequency of the word/term in the phrases of the subnode, and IDF is the inverse of the number of paragraphs/posts that the word has appeared in the entire corpus. A list of n (corpus dependent, $n = 2$ for Bridgegate and $n = 5$ for Pizzagate) top significant words from this list is then used to create a *label* for the cluster. For the particular implementation in this work, we start with the first word in the ranked list, and then add the next word only if its score is greater than $\alpha *$ (score of its predecessor) for some corpus dependent $\alpha < 1$ (for Pizzagate we used $\alpha = 0.5$ and for Bridgegate $\alpha = 0.7$); if the next word is not significant then we stop. We also stop if we reach n top words in this list of significant words. Thus, for each cluster we determine a label of at most n representative words. Next we consider all the k clusters and merge all clusters with identical labels. *Each such merged cluster is now a subnode.*

3.1.5 Narrative Network Generation

Contexts and Context-dependent Relationships: For computational purposes, we define a particular context as the set of sentences where two actant categories as determined

by noun phrases belonging to the same supernodes appear together in the same sentence. A context is characterized by the set of relationship phrases that have already been computed from these sentences. To further distill this set of relationship phrases and create a ranked order among them, we consider only the verbs in the relationship phrases because verbs are known to capture binary relationships in large-scale corpora [40]. The contexts defined by verbs have discriminatory power since they capture the different roles played by the same actants in different contexts. In our conspiracy studies, in order to find the most significant relationship, we use the frequency (excluding stop words) to find the best describing verb. In an improved method we use the inverse document frequency (IDF) multipled by its frequency scores to create a weighted ranking relationships. Later in our novel reviews narrative networks, we define a context based clustering method to find the best set of relationships.

Multi-Scale Narrative Network Generation: The network defined by all the subnodes and their relationship edges, which are labeled by the most significant relationship phrases/verbs, is the final narrative framework or frameworks for a particular corpus. This network will tend to have a relatively large number of nodes and high edge density. The subnodes and supernodes play different roles with varying importance. Meaningful subnetworks can be extracted by projecting various facets of the narrative network such as power-relationship networks, ego networks, super-node level networks, and networks comprising a target set of entities or actants; these projections, in turn, can be used to support multi-scale analysis of a complex narrative.

Structural Centrality of Nodes and Edges: Various measures of centrality and importance can be computed for each of the nodes and edges in the network. Eigen-centrality or PageRank for nodes, and betweenness for edges are example measures. A set of cen-

tral nodes in a narrative network can be defined as a set of minimal size whose removal breaks up the network into disjoint connected components. For example, as illustrated in Figure 4.5, the removal of the Wikileaks supernode and its edges in the Pizzagate narrative network breaks it up into disjoint connected components that define different domains that the actants inhabit. For the Bridgegate narrative network, no small size set of central nodes exists because the rich set of connections among the main actants existed well before the conspiracy to close the lanes on the George Washington Bridge.

3.1.6 Community Detection

Intuitively, a *community* in a network is a set of nodes that are more "densely" connected within the set than with nodes outside of that set. Given the nature of the inter-actant relationship network, such communities correspond to the subdomains of interaction. Partitioning any given network into an optimal number of clusters or communities is a well-known problem in graph theory that is computationally intractable (i.e. the problem is NP-complete) [41]. Several approximate algorithms have been developed that generate both disjoint and overlapping community partitioning, depending on the precise definition of the "density" of any candidate community [42]. An algorithm based on the modularity index measure [41] has been shown to provide good partitioning and various implementations of the algorithm are widely used across numerous fields [43]. A potential limitation of this algorithm is that, for each run, it returns disjoint communities that can vary over different runs (based on the random choice of seeds used for each run). In practice, when averaged over many runs of the algorithm, one finds that: (i) nodes that are strongly connected appear together in the same community over a majority of the runs; these nodes can be said to form the *core nodes* that define a stable community; and (ii) nodes that are more loosely connected with the core nodes and therefore change their community assignments; these nodes can be considered as ones that are overlapping or shared among different core-nodes defined communities. In the context of narrative networks, both sets of nodes provide significant information about the different core actant groups, and how these core groups interact via shared actants.

In order to discover this nuanced community structure, we develop an algorithm described below. In the first step, given a network $G(V, E)$ (where V is the set of nodes, $N = |V|$ is the number of nodes, and E is the set of edges), our goal is to determine M (to be determined) core-defined disjoint communities, C_j ($j \in \{1, \ldots, M\}$), such that $C_j(i) = 1$ if node i belongs to community j, otherwise $C_j(i) = 0$, where $i \in \{1, \ldots, N\}$. Since the core nodes are not shared, $C_j^T C_k = 0$ for any two communities, $j \neq k$. To determine both M and the communities C_j's, we run the Louvain heuristic community detection algorithm (in NetworkX [44]) T_{max} times. Next, a co-occurrence matrix, A, is defined, such that its element $A(i, j) = k$, if nodes i and j co-occur in the same community k times over the T_{max} runs of the algorithm ($0 \leq k \leq T_{max}$). We normalize this co-occurrence matrix by dividing every entry by T_{max}, so that $A(i, j)$ is the probability that nodes i and j co-occur in any given run. We next create a graph by defining an adjacency matrix, $G_c(i, j)$, where $G_c(i, j) = 1$ if $A(i, j) \geq P_{th_1} = 1 - \epsilon$, where $\epsilon > 0$ is a small number. Every connected component with at least two nodes in this graph defines a core of a community, C_j. The number of non-trivial connected components (i.e., connected components with at least two nodes), M, is the number of communities. Note that, by construction, the cores are disjoint.

In the second step, we extend each core community C_j by bringing in nodes that co-occur sufficiently many times with any node in C_j. That is, for every $k \notin C_j$, if there exists an $i \in C_j$ such that $A(i, k) \geq P_{th_2}$ (where $0 < P_{th_2} < P_{th_1}$), then $C_j(k) = 1$. Thus, the core nodes have strong connectivity, and the extended nodes share sufficiently strong connectivity. Note that after extension, the communities can overlap.

Finally, each community network is formed by the subgraph of the original network, $G(V, E)$, defined by the nodes in each C_j. Community co-occurrence frequency counts are retained for each node, since nodes with lower co-occurrence can provide information on various components of the graph that are not part of the core communities. We disregard nodes that have co-occurrence probability less or equal than P_{th_2}.

Once a community structure is determined for the narrative network (defined over all the subodes), we do further processing to discover the most frequently activated communities, actants and relationships. In particular, we filter for subnodes with a corpus frequency $\geq$ the

Algorithm 1 Community detection for a network, $G(V, E)$, with overlapping nodes

Input: G, P_{th_1}, P_{th_2}

Output: $C_j(i)$, F

$A_{k,l} = 0$

for $i = 1 : T_{max}$ **do**

Run community detection algorithm on G

if nodes k,l in same community **then**

$A_{k,l} = A_{k,l} + 1$

end if

end for

Normalize by $A = A/T_{max}$

$A'_{k,l} = A_{k,l} \geq P_{th_1}$

Form Graph G_c defined by adjacency matrix A'

M = Number of Connected Components in G_c with at least two nodes.

for Connected component C_j ($|C_j| \geq 2$) $\in G_c$ **do**

$C_j(i) = 0$

if $i \in C_j$ **then**

$C_j(i) = 1$

end if

end for

for i, k and C_j **do**

if $(C_j(i) == 1)$ and $(C_j(k) == 0)$ and $(A_{i,k} \geq P_{th_2})$ **then**

$C_j(k) = 1$

end if

end for

For each C_j construct a subgraph of $G(V, E)$ with nodes in C_j. F is the union of all the community networks.

average frequency count of subnodes in the corpus. The surviving subnodes are then grouped by actant supernodes. This step allows us to identify the central narrative framework. It also helps us identify less frequently activated communities and their constituent nodes, which may include components representing meta-narratives, unrelated conversations, or the emergence–or to borrow a more fitting term from physics, *nucleations*–of other narrative frameworks.

3.1.7 Conspiracy Theory Evaluation

We evaluate our results by comparing the narrative graph we learn to an expert labeled "gold standard" graph (as opposed to a ground truth graph). The lack of ground truth for machine learning work based on data derived from social and news media is a well-known problem [45]. As with oral narrative traditions where there is no "correct" version of a story (or ground truth), there is no canonical version of Pizzagate against which one can compare. For news stories, while journalistic accounts attempt to represent a ground truth, that "truth" is often contested [46]. In many cases, it is not until long after the news event is over that canonical summaries of the event are published; even then, there can be considerable debate concerning whether the news has been reported accurately, and whether the canonical summary is an accurate one (i.e. ground truth). For Bridgegate, that canonical summary has yet to be written, in part because several of the indicted co-conspirators are appealing their convictions, and in part because additional information continues to be reported. Given this lack of ground truth data, we use high quality, expert labeled data for evaluation [47].

For both Pizzagate and Bridgegate, we use the NY Times illustrations as the basis of our gold standard evaluation data [48] [49]. It is generally accepted that the reporters and illustrators of the NY Times, as reflected by their status in the field of journalism, are capable of creating high quality, expert labeled data. Consequently, we consider their illustrations and accompanying explanatory articles as fulfilling reasonable criteria for external, expert generated validation data. Yet, while all of the nodes were labeled in these illustrations,

the relationships between nodes were either poorly labeled (Pizzagate) or labeled in an inconsistent manner (Bridgegate).

To generate the gold standard expert annotations for the Pizzagate relationships, we proceeded in steps. First, we kept the labels from the eight labeled edges in the original illustration. Then we employed a standard three-person annotator setup to generate labels for the remaining edges: two independent expert annotators, native speakers of English with experience in journalism and political science and trained in narrative mark-up, provided their own relationship labels based on a reading of the article accompanying the Pizzagate illustration. Once they had completed their annotations, whenever they were in agreement, that label was used as the label on that edge. Whenever they were in disagreement, an equally qualified arbitrator decided on which label to use as the label for that edge [50] [51] [52]. We did not ask the annotators to add additional nodes or edges, although both of them decided independently to annotate the Edgar Welch episode described in the article, adding two additional nodes: "Edgar Welch" and "Police". The annotators also added three additional edges: "investigates" and "shoots" from Welch to "Comet Ping Pong", and "arrest" from Police to Welch.

Unlike the Pizzagate illustration, the NY Times Bridgegate illustration included labeled inter-actant relationships. These labels were not consistent and, along with relationships, also included actant descriptors (e.g. "top Cuomo appointee"), evaluative statements of relationships (e.g. "They weren't."), and speculative questions (e.g. "what prompted Kelly to send this email?"). To address this problem, we used the same three-person annotation team as described above to derive clear inter-actant relationship labels from the illustration. As the speculative questions included in the illustration were issues raised by the illustrators and not a part of the inter-actant relationship graph, we did not include them in our revised gold standard graph.

To determine the accuracy of our narrative framework graphs, we performed two evaluations, one to measure the accuracy of our actant extractions and aggregations, and one to measure the accuracy of our interactant relationships.

For actants, we calculated, given a threshold, whether the nodes represented in the hand-drawn illustrations were present or not in our extractions, and then whether they were present or not without the threshold. We also counted the actants that we discovered that were not in the hand-drawn illustrations. This last measure is important since the hand-drawn illustrations do not represent a ground truth, but rather serve as an expert summary based on human assessment of the reports of an event. It is possible that even expert summaries such as the NY Times illustrations do not include key actants; this was the case for Pizzagate, where Bill Clinton and the Clinton Foundation, frequently mentioned actants in the narrative framework developed on the Pizzagate forum, were missing in both the illustration and the accompanying article. We report the accuracy of our extractions for actants in Table 4.3.

To evaluate the accuracy of our relationship extractions, we developed an automated algorithm comparing our relationship phrases to ground truth relationships. For a set of relationships between entities $J_{A_1A_2}$, we aim to find a mapping $h_{A_1A_2} : J_{A_1A_2} \rightarrow C_{A_1A_2}$. This process is described as follows: Use the scoring function $f_{cos}(a, b)$ to compute the cosine similarity between a, b. A gold standard relationship phrase is mapped to an automatically extracted relationship phrase only if its embedding is close enough to be considered a match, here cosine ≥ 0.85. This algorithm seeks to approximate a maximum likelihood estimation problem; $\mathcal{L}$ represents the cosine similarity f_{cos} implemented with thresholds:

$$h_{A_1A_2}(j) = \underset{C \in C_{A_1A_2}}{\operatorname{argmax}} \mathcal{L}(C, j), \ \forall\, j \in J_{A_1A_2}. \tag{3.1}$$

The evaluations of these interactant relationships are presented in Table 4.4.

3.1.7.1 News Based Evaluation

After we estimate narrative networks that represent the underlying structure of conspiracy theories in a large social media corpus (4Chan, Reddit) where they are most likely to originate, We compare communities with corresponding reporting about them in the news (GDELT). This approach allows us to analyze the interplay between the two corpora and to track the time-correlation and pervasive flow of information from one corpus to the other.

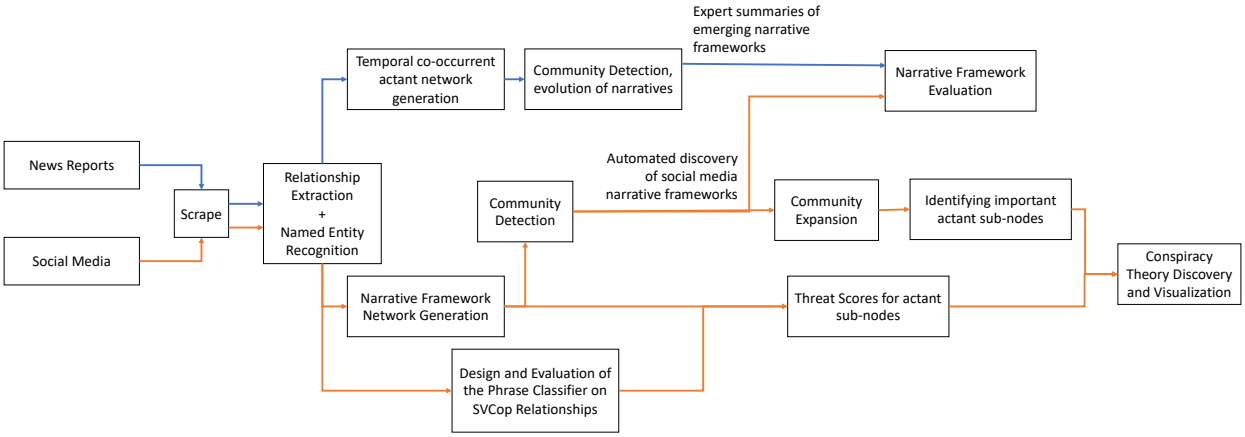

Figure 3.2: Automated pipeline of processing data and discovering narrative networks in social media and news reports.

The latent structure of the social media networks also provides features which enable the identification of key actants (people, places and things) in conspiracies and conspiracy theories, and the detection of threat elements in these narratives. The following subsections introduce the graphical narrative model for conspiracy theories in social media as well as the pipeline for processing news reports. The end-to-end automated pipeline is summarized in Figure 3.2.

3.2 Narrative network communities and their relationship to conspiracy theories

Because conspiracy theories connect preexisting domains of human activity through creative speculation, often presented as being based on a theorist's access to "hidden knowledge", we expect that the narrative frameworks that we construct will have clusters of nodes and edges corresponding to the different domains. Since these clusters are densely connected within themselves, with a sparser set of edges connecting them to other clusters, we can apply community detection algorithms to discover them. For example, the domain of *"public health"* will have dense connections between sub-nodes such as *"doctors"* and *"hospitals"*, with relatively few connections to the domain of *"telecommunications"*, which will in turn

have dense connections between sub-nodes such as *"5G"* and *"cell towers"*. Traversing these different communities mimics the conspiracy theorist's cross-domain exploration in the attempt to create a conspiracy theory.

Given the unsettled nature of discussions on social media concerning the Covid-19 pandemic, it seems likely that there are multiple, competing conspiracy theories in the corpus. Therefore one would expect to find a large number of communities in the overall network, some isolated from the rest and others with a limited number of shared sub-nodes. One would also expect that this network would have a hierarchical structure.

In order to capture any such hierarchical structure, we compute overlapping network communities, where each community is defined by (i) a core set of nodes that constitute its backbone, and (ii) a set of peripheral nodes of varying significance that are shared with other communities. Currently, to determine the communities in our network, we run the Louvain greedy community detection algorithm multiple (~ 1000) times using the default resolution parameter in NetworkX [53]. We define *two nodes as belonging to the same core if they co-occur* in the same community for almost all of the runs; here we use a threshold of 850. As in [10], the threshold is aligned with the precipitous drop in the size of the Giant Connected Component (GCC).

Next, a core that defines a community is a set of nodes that is closed under the co-occurrence transitive relationship: If nodes A and B belong to the same core, and nodes B and C also belong to that same core then, by transitivity, we say nodes (A, B, C) are all in the same core. The resulting disjoint sets of core nodes (i.e. equivalence classes under the co-occurrence transitive relationship), along with their edges in their original network, define non-overlapping communities that form the multitude of narrative frameworks in the corpus.[1] Overlapping nodes are then brought into the communities by relaxing the co-occurrence threshold [10].[2] These interactions among core communities, and hence, the respective narrative frameworks, capture the alignments among multiple knowledge domains that often underlie conspiracy

[1]See [10] for details on how to select an optimal co-occurrence threshold and how to efficiently determine the core community networks.

[2]Future work will focus on testing other network community detection methods.

theories.

3.2.0.1 Threats and special nodes in communities

Taken as a whole, the narrative framework comprising networks of actants and their inter-actant relationships (along with other metadata) reveals aspects of conspiracy theories including the threatening sub-nodes identified by the conspiracy theorists, and the possible strategies that they suggest for dealing with those threats. For instance, the network community consisting of sub-node [tower, 5g, danger] along with its associated SVCop relationships "[5g] is deadly", "[tower]s should be burned", imply a threat to human well-being posed by 5G, and a strategy for dealing with that threat: burn the cell towers (strategy) to protect people from the deadly 5G (threat). Because threats are often followed by strategies, we prioritize the classification of threats.

To classify threats, we look for sub-nodes in the network communities that, given their associated descriptions, might be considered threatening. For example, a descriptive reference to a sub-node "vaccines" that suggests that they "can kill", would allow us to code "vaccines" as a possible threat. We repeat this process for all the sub-nodes in the network communities, and find that strong negative opinions are associated with some subset of sub-nodes, which we identify as candidate threats. By applying a semi-supervised classification method to these candidate sub-nodes, we can confirm or reject our suspicions about their threatening nature.

The threat classifier is trained on the relationships extracted from social media posts. In particular, SVCop relationships (described in Section 4.1) play a special role in providing information about a particular sub-node: these relationships provide important information about the first argument and are generally descriptive in nature. In such relationships, the second argument is most often a descriptive phrase with an associated to-be verb phrase. For example (5g,is,dangerous/a threat/harmful) are SVCop relationships describing the [5g] argument. We consider these relationships as self-loops for their first arguments, which are aggregated into sub-nodes.

The most discussed sub-nodes tend to have a high number of such self-loop relationships, and the descriptive phrases often carry powerful characterizations of these entities. Sub-node-specific aggregation of these relationships can inform us about the role of a particular actant in its community. For example, we find $\sim$ 350 SVCop relationships describing the node "virus" as "harmful", "deadly", "dangerous", and "not real".

We aggregate the entire corpus of SVCop relationships ($\sim$ 10000) and then label them in a hierarchical fashion as follows: First, each such SVCop phrase is encoded using a 768 dimensional BERT embedding from a model fine-tuned for entailment detection between phrases [36]. Next, the vectors are clustered with HDBSCAN [54], resulting in a set of $\sim$ 1000 density-based clusters C, with an average cluster membership size of 7. Approximately 3000 of the phrase encoding vectors are grouped in a cluster labeled as -1, indicating that they are not close to others and are best left as singletons. For the rest, each cluster represents a semantically similar group, and can be assigned a group semantic label. Thus, the task of meaningfully labeling $\sim$ 10000 phrases as 'threat' or 'non-threat' is reduced by almost a factor of 10.

We define a binary label for each cluster. A *threat* is a phrase that is universally recognized as threatening: [5g] is *dangerous*, [a tower] is *a bioweapon*. Here, the phrases *dangerous* and *bioweapon* are clearly indicative of threats. The remaining phrases are labeled as *neutral/vague* comments.

For every cluster $c \in C$, we assign a label l_c to c such that every descriptive phrase $d \in c$ is also assigned label l_c. Clearly, label quality is contingent on the manual labeling of the clusters and the semantic similarity of descriptive phrases as aggregated by the BERT and density-based clustering. This is ensured by three independent reviewers labeling each cluster and, in the case of disagreement, choosing the label receiving the majority vote.

We measure the inter rater reliability with respect to the majority vote by the three different raters. Our results for a sample size of 100, are 0.745, 0.87 and 0.829.

The semantic similarity in each cluster is verified by a qualitative analysis of the clusters undertaken by domain experts. For example, most of the clusters have exact phrase matches

such as

- **Cluster 1**: [the ongoing trade war, the trade war]
- **Cluster 2**: [radiation, radiation, more EM radiation, a result of radiation, electromagnetic radiation, also radiation]

that support high-fidelity hierarchical labeling. Other clusters validate the usage of BERT embedding as a means for clustering semantically similar phrases. For example,

- **Cluster 1**: [SLAVERY, members of race enslaved, a slave]
- **Cluster 2**: [a liberal hoax, a liberal lie designed]
- **Cluster 3**: [rabid supporters of SCIENCE, rabid supporters of SCIENCE, scientists f***]

capture semantic similarity in addition to exact matches of phrases.

Since our BERT model is fine-tuned to detect entailment, the clustering is sensitive to negation in particular, which is important in classifying phrases as threats. For example, the following clusters are distinct and complete:

- **Cluster 1**: [not convenient, not beneficial, not fun, not helpful]
- **Cluster 2**: [useful, helpful]

These labeled phrases are used to train a k-nearest neighbor (kNN)-based phrase classifier to identify threatening descriptions. Once again we use the fine-tuned BERT embedding.

Many competing kNN models provide useful classification results for phrases. We found that setting $k = 4$ results in a model that reasonably classifies most phrases. The kNN classifier is *binary*: 0 represents the class of *non-threat* and 1 represents the class of "threat". The cross validation part is carried out at the level of the clusters: that is, when designing the training sets (for kNN, the set of phrases used in performing the kNN classification of a given

phrase) and validation sets, we partition the phrases based on their cluster assignments. All phrases belonging to the same cluster are assigned to the same set and are not split across the training and validation sets. Because the labeled phrases have duplicate second arguments and repeated phrases occur in the same cluster, this approach to cross-validation ensures against repeating phrases in both the training and validation set, which is achieved by partitioning data at the cluster level.

The primary purpose of designing the phrase classifier is to identify threatening sub-nodes, which appear as core nodes in the narrative framework communities. Aggregated second arguments of SVCop relationships corresponding to a particular sub-node are classified with the kNN phrase classifier. Based on a majority vote on these second arguments, we can classify a sub-node as a potential threat. An outline of this algorithm is provided in Algorithm 2.

Algorithm 2 Threat classification of sub-nodes using majority vote

Input: A sub-node p, descriptive phrases in SVCop relationships of sub-node p, D_p and the $(k = 4)$ kNN model f

Output: s_p: A score between 0 and 1 assigned to p as an proportional indicator of threat

```
count ← 0
samples ← |D_p|
for each descriptive phrase arg_2 in D_p do
    label ← f(arg_2)
    count ← count + label
end for
s_p ← count / samples
```

3.2.0.2 Conspiracy Theory Narrative framework discovery

A narrative framework for a conspiracy theory, which may initially take shape as a series of loosely connected statements, rumors and innuendo, is composed from a selection of subnodes from one or more of these communities and their intra- and inter-community

relationships. Each community represents a group of closely connected actant sub-nodes with those connections based on the context-dependent inter-actant relationships. Traversing paths based on these inter-actant relationships within and across communities highlights how members posting to the forums understand the overall discussion space, and provide insight into the negotiation process concerning main actants and inter-actant relationships.

This search across communities is guided by the extended overlapping communities (which connect the core communities) taking into consideration the sub-nodes that are classified as threat nodes. The inter-actant relationship paths connecting the dominant threat nodes, both within and across communities, are then pieced together to create the various conspiracy theories.

3.2.1 Searching conspiracy theories from social media in the news

Many conspiracy theories detected in social media are addressed in news reports. By temporally aligning the communities discovered from social media with the evolving communities emerging from news collected daily, we can situate the 4Chan commentary alongside mass media discussions in the news. Such a parallelism facilitates the analysis of information flow from smaller community threads in social media to the national news and from the news back to social media.

3.2.1.1 Extraction of inter-actant communities in the news

To aggregate the published news, we consider (1-day time-shifted) intervals of 5 days. This sliding window builds $s = 101$ segments from January 1, 2020 to April 14, 2020. We have discovered that a longer interval, such as the one chosen here, provides a richer backdrop of actants and their interactions than shorter intervals. In addition, narratives on consecutive days retain much of the larger context, highlighting the context-dependent emergence of new theories and key actants.

We use the major actants and their mentions discovered in the social media data to filter the named entities that occur in the news corpus. A co-occurrence network of key actants in news

reports (conditioned on those discovered from social media), provides a day-to-day dynamic view of the emergence of various conspiracy theories over time. In addition, we model the flow of information between social media and news reports by monitoring the frequency of occurrence of social media communities (as captured by a group of representative actants in each community) in the text of news reports (see Evaluation). With minimal supervision, a few actant mentions are grouped together including, [trump, donald] : **donald trump**, [coronavirus, covid19, virus] : **coronavirus** and [alex, jones] : **alex jones**. This actant-grouping enhances the co-occurrence graph by reducing the sparsity of the adjacency matrix representing subject-object interaction.

3.2.1.2 Co-occurrence actant network generation

For each 5-day segment of aggregated news reports, the corpus of extracted relationships R_i and the associated set of entities E_i are parsed with Algorithm 3 to yield a co-occurrence actant network. Day-to-day networks reveal the inter-actant dynamics in the news. While many metrics can be used for summarizing the dynamics within these networks, we considered the Number of Common Neighbors (NCN) between them. If the adjacent vertices of a_1 are S_{a_1} and of a_2 are S_{a_2}, the NCN score is defined as:

$$n_{a_1,a_2} = |S_{a_1} \cap S_{a_2}|. \tag{3.2}$$

3.2.2 Evaluation and Metrics

3.2.2.1 Temporal alignment of communities derived from news reports and social media

We temporally align the conspiracy theories discussed in social media and in news reports by first capturing a group of representative actants in each social media community. Let the set of keywords representing a particular community be V_i. The timestamps present in 4Chan and GDELT data make these corpora suitable for temporal analysis with respect to V_i (our Reddit corpus does not contain dates). In order to facilitate a comparison between

Algorithm 3 Co-occurrence Actant Network Generation for a Segment $i < s$ of News

Input: R_i relationship tuples, E_i entities

Output: $G_i(R_i, E_i)$

$M \leftarrow []$

for $(arg_1, rel, arg_2) \in R_i$ **do**

$s \leftarrow H(arg_1)$ {$H(arg)$ is the headword of arg}

$o \leftarrow H(arg_2)$

$r \leftarrow H(rel)$

if $(s, o \in E_i)$ AND $(s \neq o)$ AND (r NOT stop word) **then**

$M[s, o] \leftarrow M[s, o] + 1$

$M[o, s] \leftarrow M[o, s] + 1$

end if

end for

$M_{norm} = \texttt{normalize}(M)$ {along each row}

$G_i(R_i, E_i) \leftarrow M$ {Color-coded based on the labels of actants decided by the Entity Extractor}

the two corpora conditioned on V_i, let C_j denote the raw 4Chan data and D_j denote the raw GDELT news data in time-segment j. The time segments are 5-day intervals between March 28,2020 and April 14, 2020, which is the intersection of date ranges for which we have temporal 4Chan and GDELT data. We define a Coverage Score (m) that captures the presence of actants from V_i in C_j and D_j.

$$m_C(V_i, j) = \frac{\sum_{w_V \in V_i} \sum_{w_C \in C_j} \mathbb{1}(w_V = w_C)}{|V_i||C_j|}, \tag{3.3a}$$

$$m_D(V_i, j) = \frac{\sum_{w_V \in V_i} \sum_{w_D \in D_j} \mathbb{1}(w_V = w_D)}{|V_i||D_j|}. \tag{3.3b}$$

To normalize the coverage scores to a baseline, we compute a Relative Coverage Score (r) where V^* is a random set of actants (of size 500) as:

$$r_C(V_i, j) = \frac{m_C(V_i, j)}{m_C(V^*, j)}, \quad r_D(V_i, j) = \frac{m_D(V_i, j)}{m_D(V^*, j)}. \tag{3.4}$$

Computed across all time-segments, $r_C(V_i)$ and $r_D(V_i)$ represent a series of relative coverage scores for 4Chan and GDELT respectively, with one sample for every time segment. This metric now provides a normalized measure for coverage of a community derived from social media in the temporal corpora of 4Chan and GDELT data.

The cross-correlation function of these relative coverage scores $\mathcal{R}_{C,D}(\tau) = E[r_C(V_i, t) r_C(V_i, t+\tau)]$ can provide interesting insight into the co-existence of conspiracy theory communities in the two corpora where τ is the number of offset days between the news and 4Chan data (see Figures 5.3 and 5.4). This cross-correlation score peaks for the number of offset days that results in the maximum overlap of relative coverage scores. For example a τ of 10 days would imply that information about a specific set of representative actants occurred in the news and 4Chan data roughly 10 days apart. τ captures the latency or *periodicity* lag between communities mentioned in the news and in 4Chan data. The error bars are generated over 20 random communities used for normalizing the coverage scores before cross-correlation.

3.2.2.2 Other standard metrics to compare communities derived from the news and social media

We present standard metrics to further compare communities of actants derived from temporal news reports and social media. Our metrics are standard measurements used for clustering evaluations based on ground truth class labels [1]. Algorithm 4 describes this evaluation process.

Algorithm 4 Unsupervised evaluation of communities

Input: $C_{i,t}$ News community indexed i at time t, K_j Social media community indexed j

Output: Pr_t Percentage of coverage for time t, h_t Homogeneity at time t, c_t Completeness at time t, v_t V-Measure at time t

$Y_{gr} \leftarrow []$

$Y_{pred} \leftarrow []$

$\texttt{count} \leftarrow 0$

for each $C_{i,t}$ **do**

  for each actant a in $C_{i,t}$ and K_j **do**

    if a in K_j **then**

      $\texttt{count} \leftarrow \texttt{count} + 1$

      $Y_{gr}[a] \leftarrow j$

      $Y_{pred}[a] \leftarrow i$

    end if

  end for

end for

$Pr_t \leftarrow \texttt{count}/|K_j|$

$h_t \leftarrow \texttt{Homogeneity}(Y_{gr}, Y_{pred})$

$c_t \leftarrow \texttt{Completeness}(Y_{gr}, Y_{pred})$

$v_t \leftarrow \texttt{V-Measure}(Y_{gr}, Y_{pred})$

3.2.2.3 Evaluation of phrase-based threat detection

We use average recall and average accuracy to evaluate the performance of the phrase-based threat classifier. The average is computed across the 5-fold group-shuffled cross-validation of phrases. Here recall and accuracy are defined as,

$$\texttt{Recall} = \frac{\texttt{Detected Threats}}{\texttt{Ground truth threats}},$$
$$\texttt{Accuracy} = \frac{\texttt{Detected threats} + \texttt{Detected non-threats}}{\texttt{Size of the validation set}}.$$

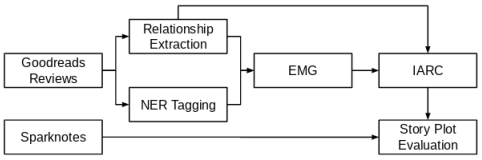

Figure 3.3: Pipeline to extract actant-relationship graphs. Our contributions introduce the Entity Grouping and the Inter-actant Relationship Clustering blocks

3.3 Context Based Character Detection

Our next chapter studies the underlying narrative extraction from social media posts on book reviews. In this chapter we introduce novel algorithms to further study characters or nodes in the narrative graphs. We group the nodes based on their interactions with other nodes.

3.3.1 Entity Mention Grouping (EMG)

As a semantically identifiable character in a book is expressed in reviews as diverse entity mentions, it is necessary to group these mentions and label them with the same character label.

Let the frequently-occurring set of entity mentions be M and let R_{ik} be the relationships between entity mention m_i and m_k, where m_i is the Subject and m_k be the Object. The set R_{ki} then denotes the relationships when the roles are reserved. First, we note that if there is a relationship triplet $(\text{Subject} = m_i, \text{Verb}, \text{Object} = m_j)$ then clearly m_i and m_j are mentions of different actants and are not to be grouped together. In order to avoid any noise-induced exclusion of such a pairing, we consider a pair m_i, m_j as incompatible if $|R_{ij}| + |Rji| \geq \gamma$. Based on our observation of the low frequency of noisy relationships, the hyperparameter γ is set to 3 in this work. In the following we assume that for each mention m_i we have removed all incompatible nodes m_j.

Intuitively, two compatible mentions m_i and m_j correspond to the same actant if, for every other mention m_k, the relationships between the pair (m_i, m_k) are semantically the same as the relationships between the pair (m_j, m_k). In practice, different mentions of the same

actant will share only a subset of the relationships when aggregated over all the extractions. In the following we provide an algorithm to quantify this intuitive idea that yields robust EMGs.

Let $T_{ik} = H(R_{ik})$ describe the set of headwords in R_{ik}. Also let G be the directed bipartite graph from the entity mentions M to M (see Fig. 6.1) with the edges representing the relationships between the entity mentions. We would like to find an Entity Mention Grouping (EMG) function $g : M \rightarrow [1, ..., N]$, $N \leq |M|$, where (i) if $g(m_i) = g(m_j) = k$ then entity mentions (m_i, m_j) are grouped together to form the k^{th} actant. Moreover, (ii) we want the groups to be complete: that is, for two groups $g^{-1}(k_1)$ and $g^{-1}(k_2)$ (with $k_1 \neq k_2$ and $k_1, k_2 \in [1, ..., N]$), the entity mentions are semantically similar within each set and are semantically differentiated across the sets. To measure semantic similarity between m_i and m_j, we consider the following measure involving another mention m_k:

$$\begin{aligned} s_{(ij)k} &= \Pr(T_{ik}|T_{jk}) + \Pr(T_{jk}|T_{ik}) \ , \\ \Pr(T_{ik}|T_{jk}) &= \frac{|H(R_{ik}) \cap H(R_{jk})|}{|H(R_{jk})|}. \end{aligned} \tag{3.5}$$

To understand why $s_{(ij)k}$ is an effective similarity measure, consider the following cases: (i) If $H(R_{ik}) = H(R_{jk})$, implying that m_i and m_j share the exact relationships with m_k and hence should be grouped together, then $s_{(ij)k}$ achieves the maximum value of 2, (ii) the m_j mention of an actant occurs less frequently then m_i and is reflected by $H(R_{ik}) \subset H(R_{jk})$, then $s_{(ij)k} \geq 1$. This captures the case where m_j shares all its relationships with m_i but not vice versa, (iii) m_i and m_j are indeed mentions of different actants, in which case $|H(R_{ik}) \cap H(R_{jk})|$ is expected to be a lot smaller than both $|H(R_{ik})|$ and $|H(R_{jk})|$ and $s_{(ij)k} << 1$.

To ensure that we compute similarity when m_k is the Subject, we define an analogous similarity score:

$$\begin{aligned} s_{k(ij)} &= \Pr(T_{ki}|T_{kj}) + \Pr(T_{kj}|T_{ki}) \ , \\ \Pr(T_{ki}|T_{kj}) &= \frac{|H(R_{ki}) \cap H(R_{kj})|}{|H(R_{kj})|}. \end{aligned} \tag{3.6}$$

Finally, the score matrix S is computed where the score S_{ij} between m_i and m_j aggregates

the measure on all feasible $m_k \in M - \{m_i, m_j\}$ and provides a metric for similarity across all entity mentions:

$$S_{ij} = \sum_{m_k \in M - \{m_i, m_j\}} s_{(ij)k} + s_{k(ij)}. \tag{3.7}$$

The grouping function g is now constructed as follows: For every entity mention m_i, the scores in the vector S_i are ranked in descending order. We next introduce two hyperparameters for each novel, $\alpha, \beta \geq 0$, such that an entity mention m_i is grouped with m_j only if the score S_{ij} satisfies: $S_{ij} \geq \alpha$ and $\frac{S_{i(j-1)}}{S_{ij}} \geq \beta$ (for $j \geq 2$).

We compute α from novel-specific distribution statistics. In particular, we compute the histogram of all non-zero S_{ij} and compute α as the 75th percentile (i.e. 25% of S_{ij}'s are $\geq \alpha$). For all considered books (except *To Kill a Mockingbird* where $\alpha = 2.6$), $\alpha = 2.0$. The hyperparameter β is set to 2.

The parameters α and β are similar to those in works such as the Elbow K-Means method [55], in which β correlates to inertia if the scores S_i correlate to the distortion, and α provides a means of resolution if the elbow is unreliable (common in our model for rarer entity mentions).

The entity mention groups, once found, are labeled with the most frequent mention in the respective groups. Empirically, these automatically computed labels match the ground truth entities as derived from SparkNotes.

3.3.2 Inter-actant Relationship Clustering (IARC)

The aggregated entity mentions captured in g are fed back into the standard relationship extraction task. Then, the relationships aggregated between any pair of actants, represented by their respective entity mention groups (e.g.: $A_1 = g^{-1}(k_1)$ and $A_2 = g^{-1}(k_2)$) is computed as:

$$R_{A_1 A_2} = \bigcup_{p \in A_1,\, q \in A_2} R_{pq}. \tag{3.8}$$

$R_{A_1 A_2}$ is a richer and potentially multi-modal set of relationships. This process enables a form of transfer learning, aiding relationship extractors in identifying connections at a higher

semantic level of characters and not merely at the level of entity mentions. The associated relationship clusters are found using the cosine similarity measure in the BERT embedding space (Algorithm 1).

Algorithm 5 Inter-actant Relationship Clustering

Input: $R_{A_1A_2}$

Output: $C_{A_1A_2}$

$\widehat{R}_{A_1A_2}, C_{A_1A_2} = \{\}$

for $r \in R_{A_1A_2}$ **do**

`append BERT`(r) `to` $\widehat{R}_{A_1A_2}$

end for

$C_{A_1A_2} =$ `Elbow K-Means Method on` $\widehat{R}_{A_1A_2}$

$C_{A_1A_2}$ is the set of clusters of relationships that describe the multi-modality in $R_{A_1A_2}$. For each cluster C we compute its dispersion (using the cosine similarity measure), β_C. We retain only those clusters with β_C greater than a threshold (here, we set it to 0.8) as a valid semantic relationship group.

3.3.3 Narrative Evaluation Methods:

We compare these relationship clusters to the ground truth relationships between characters (e.g.: $J_{A_1A_2}$). We aim to find a mapping $h_{A_1A_2} : J_{A_1A_2} \rightarrow C_{A_1A_2}$. This process is described in Algorithm 2, where $f_{cos}(a, b)$ is the function to compute the cosine similarity between a, b, and β_C is the dispersion of a cluster C using the cosine similarity measure. Thus, a ground truth relationship phrase is mapped to an automatically clustered semantic group only if its embedding is close enough to the centroid of the cluster.

Similar to the EMG task, the clusters are well differentiated, resulting in high-fidelity labels. Furthermore, Algorithm 2 seeks to approximate a maximum likelihood estimation problem, where $\mathcal{L}$ represents the cosine similarity f_{cos} implemented with thresholds:

$$h_{A_1A_2}(j) = \underset{C \in C_{A_1A_2}}{\operatorname{argmax}} \mathcal{L}(C, j), \ \forall\, j \in J_{A_1A_2}. \tag{3.9}$$

Algorithm 6 Evaluation: Mapping Relationship Clusters to Ground Truth

Output: $h_{A_1A_2}$

for $C \in C_{A_1A_2}$ **do**

  if $\beta_C \geq 0.8$ **then**

    if $\max_{r \in C,\, j \in J_{A_1A_2}} f_{cos}(r, \texttt{BERT}(j)) \geq 0.8$ **then**

      $h_{A_1A_2}(j) = C$

    end if

  end if

end for

3.3.4 Expanded Story Network Graph

While the nodes in a *regular* narrative framework graph for a novel are derived from an associated external resource such as a canonical character list from *SparkNotes* or the work itself, the discovered relationships from our pipeline (edges) are extracted from readers' reviews. These extractions reveal both readers' thoughts and their impressions of characters in the works under discussion. Not surprisingly, then, the reader-derived networks expand upon the regular narrative framework graphs, with additional nodes for film directors, authors, and other interlocutors, and additional edges for these additional actants' activities with the core story. To facilitate this expansion, we augment the regular story network for a novel with nodes for frequent entities *not* among the character mentions. First, we rank mentions based on their frequency and then we pick the top ranked entity mentions to add them to the story graph. For example, in *The Hobbit*, some of the extra candidate mentions are: ['tolkien', 'book', 'story', 'adventure', 'movie', 'jackson'].

Next, we find the edges that exist between these candidate nodes and the regular story network nodes. If a candidate node has significant edges with the existing characters in the story network, we augment the graph with this node. For example, the candidate node "Tolkien" is connected to the character "Bilbo" with the verb phrases ['to masterfully develop','provides','knocked','introduced']. There are other interesting but distant mentions of candidate nodes that do not have *direct* connections to nodes in the original story graph:

for example, "Jackson" (a reference to the director of *The Hobbit* film, Peter Jackson) only appears in the triplet (Jackson 's changes, distort, Tolkien 's original story). While this node represents the reviewers' acknowledgement of the novel's movie adaptation, we do not represent "Jackson" as a node in our expanded graph due to its sparse connection with the rest of the story network.

As a result, our algorithm discovers the available relationships between main novel characters and each metadiscursive candidate node mention. It draws an edge only if there are more than 5 relationships between the candidate and a story graph character. If that candidate node has a degree of 3 or greater, it is added to the expanded story network graph.

3.3.5 SENT2IMP: Character Impression Extraction

While reviews contain story synopses, they also include impressions: how did the reviewer feel about a particular character? In the expanded narrative network setting, we are not able to capture this information as the relationships represented on the graph are always between a pair of characters or meta-characters. To extract this additional structure, which has considerable importance in the evaluation by readers of the novel in this social reading space, we developed SENT2IMP, an unsupervised algorithm that aggregates user opinions in descriptive phrases of review text into distinct groups of semantically similar impressions.

The relationship extraction pipeline captures not only the relationships *between* characters, it also captures a set of relationships in which readers have expressed their impressions of characters. In our setting, an *impression* is formally defined as a cluster of semantically similar phrases from reviews that imply a single aspect of a character in the eyes of the readers. For example, while we obtain the classical relationships such as *Gandalf* "chooses" *Bilbo* for our expanded story networks, we also obtain relationships such as *Bilbo* "is" *unbelievably lucky*. These latter relationships, when aggregated, filtered and clustered, bring to light the different reader sentiments associated with each character. Our pipeline thus provides an unsupervised approach to model a character as a *mixture* of such impressions.

To extract these impressions, we start by selecting the relationship tuples labeled as "SVCop"

from the full set of extracted relationships. "SVCop" are those with the structure, *Subject Phrase* $\rightarrow$ *Verb Phrase* $\rightarrow$ *Copula*. These relationship tuples typically consist of a noun phrase and an associated adjective phrase that provide descriptive information about the noun phrase. We observe that a majority of these "copular" phrase relationships contain information about character impressions. For example, in the reviews of *Animal Farm*, the sentence "Snowball was humble and a good leader" yields the two SVCop relationships: (Snowball, was, humble), (Snowball, was, a good leader), where the phrases "humble" and "a good leader" describe Snowball. In the aggregate, these phrases contribute to *impressions* of Snowball being both humble and a good leader. In addition, we note that these relationships are frequent and comprise a significant portion of the extracted relationships per literary work: reviews contain a wide range of reader impressions when compared to the original work. Filtering and clustering these frequent relationships supports the creation of a *robust* pipeline for impression extraction.

Once we have extracted the SVCop relationships from the entire corpus, we group the relationships with respect to the character $e_i \in E$ present in the subject phrases of the extractions. Next, these phrases are transformed into vectors using a fine-tuned BERT [56] embedding. We embed these phrases in the BERT space to acquire a measure of semantic similarity. These vectors are then passed to the HDBSCAN clustering algorithm, which determines the optimal set of clusters, C'_{e_i}, per character e_i. This approach results in a qualitatively optimal clustering of phrases.

Due to the noisiness of the reader review data, as well as the inherent non-Gaussian distribution of BERT embeddings, some of the resulting clusters are not homogeneous. To mitigate the noisy clusters, we employ a modified TF-IDF [57] scoring function to weight the words in each cluster. The score of a word in a cluster is equal to its frequency in the same cluster times the TF-IDF score in the review corpus for that word, where each review is a document and each word is a term. This score for a distinct word w_m with frequency $f_{m,\widehat{C}_k}$ in a cluster of phrases $\widehat{C}_k \in C'_{e_i}$, $|\widehat{C}_k| = N$ is given as:

$$x[m, \widehat{C}_k] = \texttt{TF-IDF}(w_m) \times f_{m,\widehat{C}_k}.$$

After removing the stop words, we observe that the meaningful clusters have a skewed-tailed distribution over these scores. In some extreme cases, where a cluster only contains a few words or phrases, the distribution is centered on a high score with low variance. For example, for the character "Gandalf", we find numerous "Gandalf" clusters, with one that contains only the word "wizard". We select a cluster to be meaningful based on the variety of its highly scored words. An ideal cluster consists of a handful of high scoring words. The fewer low scoring words a cluster contains, the higher its quality and, consequently, we expect an ideal cluster to have less noise.

To quantify this cluster quality measure, we calculate the skewness g_1 [58]:

$$g_1 = \frac{m_3}{m_2^{3/2}},$$

where,

$$m_i = \frac{1}{N}\sum_{n=1}^{N}(x[n, \widehat{C}_k] - \bar{x})^i,$$

is the biased sample's i^{th} central moment, and $\bar{x}$ is the sample mean [59].

For a cluster $\widehat{C}_k$, the skewness of the distribution of $x[m, \widehat{C}_k]$, determines the quality of the cluster. High positive skewness shows that the cluster consists mostly of high-scored words with a low number of infrequent words. Occasionally, when a cluster has a small number of low-scored words, such as the "wizard" cluster for "Gandalf", it loses its skewness. Instead, we observe a high word score ($x(\cdot, \cdot)$) average. We consider a cluster to be valid if it has skewness above a certain threshold or if its words' scores have a high average with low variance.

As a result of these selection, clustering and filtering tasks, we obtain a filtered mixture of clusters of reader impressions per character, i, which we label, C_{e_i}. Each cluster presents a unique description of a character within a novel. For example, "George", in *Of Mice and Men*, has various clusters associated with him, such as ['basically Lennies protector', 'the guardian', 'in charge of Lennie ', 'Lennie guardian'] and ['clumsy', 'unhappy', 'sad', 'nervous', 'very rude', 'selfish', 'painfully lonely'].

In order to quantify the geometry of the obtained mixture, we define a distance metric between every pair of clusters of numerical embeddings (see Algorithm 7). The resultant

measure is in the range [-2,2] and is close to 2 if the pair is semantically similar and close to -2 if the pair of clusters are semantic opposites. We once again use BERT embeddings where semantic similarity is measured as the cosine distance between a pair of phrase embeddings. Before computing the cosine similarity, we reduce the dimension of the BERT embeddings to 4-principal components using Principal Component Analysis (PCA), having found that the resultant scores generalize well with this choice of principal components. It follows that for a *mixture* of clusters, we can represent the obtained inter-cluster distance measure between $\widehat{C}_i$ and $\widehat{C}_j$ (within a mixture C_{e_i}) on a heatmap that is symmetric – because our distance measure is symmetric – (see Algorithm 8) for a particular character. To ensure that the heatmaps we generate are rich, we limit our study to those characters with at least 4 clusters of descriptive phrases ($|C_{e_i}| \geq 4$).

3.3.5.1 Quantifying and Visualizing the Complexity of a Character

The heatmap for a single character shows an empirical measure of *character complexity*. In this task, we run Algorithm 8 such that both characters in the pairwise comparison are identical. A large and high-variance heatmap implies that readers consider the character to be *complex*. This complexity may (i) reflect disagreement in the reader discussions about that character, implying that impressions of the character are *controversial*, or (ii) capture contradictory and multi-modal character traits that authors develop – or that readers constitute through their reading – to portray remarkable characters in their novels. One finds instances of both in the impression clusters of Bilbo Baggins (See Table 6.6): The first two clusters that portray Bilbo as "unpleasant/boring" and "loveable/charismatic" are most likely instances of readers' contradictory perceptions; clusters 2 and 3, however, portray Bilbo simultaneously as a hero and a thief, and are most likely diverse dimensions inherent in the novel and picked up on by the readers. In other cases, when the heatmap is smaller and of low-variance, readers' perceptions of the profiled character are not as diverse, perhaps because of the character's secondary role in the novel and/or the character's narrow/singular purpose in the story. Such a character would be less *complex*.

This qualitative understanding of *complexity* can be quantitatively described by *entropy*. If one assumes that the heatmap entries of any character are samples from an underlying random variable, and the complexity of a character is its entropy, then the more spread out the underlying distribution is, the higher its entropy. Since we have only a few samples – coming only from the lower half of any given symmetric heatmap matrix – it is best to model the random variable as a discrete probability distribution in the range of scores in our heatmap, $[-2, +2]$.

First, we define the number of bins for the distribution, b. Then, from the impressions heatmap of a single character, we slot all the values below the major diagonal into these bins. Because most of our heatmaps do not have enough values to populate each bin, we adopt a standard numerical technique [60] of using a smoothing kernel – in our case uniform – (of width w bins) across the bins. In general, w and b are hyperparameters that change the sensitivity of calculating the *entropy*.

More formally, for a heatmap $S \in \mathbb{R}^{N \times N}$ where N is the number of impression clusters for the associated character, the entropy (complexity) $\mathcal{H}$ is defined:

$$\mathcal{H} = -\sum_{\{i \in [b]\}} P_i \log P_i,$$

where,

$$S_{ij} \xrightarrow[\forall i > j]{\text{aggregate}} \text{histogram bins}_{[-2,2]}^{b} \xrightarrow{\text{smoothing kernel (w), norm}} \text{Prob. Distribution } P_i.$$

3.3.5.2 Distinct Character Comparison

The observation that some characters are, in the minds of the readers as reflected in their reviews, more *complex* than others, motivates our use of the distance measure employed in Algorithm 7 for impression clusters derived from a pair of *distinct* characters. In this case, we project onto the heatmap the distance measures between clusters from two separate mixtures (the heatmap will not be symmetric and may not be square). Such a representation highlights the smaller number of *contexts* in which two characters are similar even across novels. For example, in *To Kill a Mockingbird*, one of Atticus's clusters consists of the phrases ['the

father of kids', 'the father of protagonist', 'the father of Jem', 'lenient father', 'the father of character', 'a father figure'] and can be compared to the first example of "George"'s attribute of being a "guardian". Although these two sets of phrases are not exactly the same, one can still recognize that "George" and "Atticus" are perceived similarly in their shared roles within their respective works.

Algorithm 7 Computing the Similarity Score between a Pair of Clusters of Numerical Embeddings

Input: $\widehat{E}_m$, $\widehat{E}_l$: Two Clusters of Numerical Embeddings

Output: $S_{l,m}$

********** From $\widehat{E}_l$ to $\widehat{E}_m$ **********

$s_1 = 0$

for every embedding u in $\widehat{E}_l$ **do**

$s_1 = s_1 + \max_{v \in \widehat{E}_m} \cos(v, u)$

end for

$S_1 = \frac{s_1}{|\widehat{E}_l|}$

********** From $\widehat{E}_m$ to $\widehat{E}_l$ **********

$s_2 = 0$

for every embedding v in $\widehat{E}_m$ **do**

$s_2 = s_2 + \max_{u \in \widehat{E}_l} \cos(v, u)$

end for

$S_2 = \frac{s_2}{|\widehat{E}_m|}$

$S_{l,m} = S_1 + S_2$

Algorithm 8 Computing the Heatmap between a Pair of Mixtures of Phrase Clusters

Input: C_{e_i}, C_{e_j} two mixtures of phrase clusters such that,

$C_{e_i} = [\widehat{C}_{1,e_i}, \widehat{C}_{2,e_i}, \ldots, \widehat{C}_{M,e_i}]$

$C_{e_j} = [\widehat{C}_{1,e_j}, \widehat{C}_{2,e_j}, \ldots, \widehat{C}_{N,e_j}]$

Output: $S \in \mathbb{R}^{M \times N}$: A heatmap S with S_{lm} equal to the similarity between the l^{th} cluster in one mixture and the m^{th} cluster in another mixture.

for iter in $[1, \ldots, M]$ **do**

$E_{\text{iter},e_i} = \text{BERT}\ [\widehat{C}_{\text{iter},e_i}]$

$\{E_{\text{iter},e_i} \in \mathbb{R}^{768 \times |\widehat{C}_{\text{iter},e_i}|}$: *our BERT embeddings are* 768*-dimensional vectors.*$\}$

end for

for iter in $[1, \ldots, N]$ **do**

$E_{\text{iter},e_j} = \text{BERT}\ [\widehat{C}_{\text{iter},e_j}]$

end for

$[\{\widehat{E}_{1,e_i}, \widehat{E}_{2,e_i}, \ldots, \widehat{E}_{M,e_i}\}, \{\widehat{E}_{1,e_j}, \widehat{E}_{2,e_j}, \ldots, \widehat{E}_{N,e_j}\}]$

$= \text{PCA}\,[E_{1,e_i}, E_{2,e_i}, \ldots, E_{M,e_i}, E_{1,e_j}, E_{2,e_j}, \ldots, E_{N,e_j}]$

$\{\widehat{E}_{\text{iter},e_i} \in \mathbb{R}^{4 \times |\widehat{C}_{\text{iter},e_i}|}$: 4 *principal components.*$\}$

for iterM in $[1, \ldots, M]$ **do**

for iterN in $[1, \ldots, N]$ **do**

Perform Algorithm 7 for the pair of clusters of numerical embeddings, $\{\widehat{E}_{\text{iterM},e_i}, \widehat{E}_{\text{iterN},e_j}\}$

end for

end for

CHAPTER 4

Single Conspiracy Theory Case Study

We base this work on two separate comprehensive repositories of blog posts and news articles describing the well-known conspiracy theory Pizzagate from 2016, and the New Jersey political conspiracy Bridgegate from 2013. We show how the Pizzagate framework relies on the conspiracy theorists' interpretation of "hidden knowledge" to link otherwise unlinked domains of human interaction, and hypothesize that this multi-domain focus is an important feature of conspiracy theories. We contrast this to the single domain focus of an actual conspiracy. While Pizzagate relies on the alignment of multiple domains, Bridgegate remains firmly rooted in the single domain of New Jersey politics. We hypothesize that the narrative framework of a conspiracy theory might stabilize quickly in contrast to the narrative framework of an actual conspiracy, which might develop more slowly as revelations come to light. By highlighting the structural differences between the two narrative frameworks, our approach could be used by private and public analysts to help distinguish between conspiracy theories and conspiracies.

4.1 Data

Data for this study were derived from two online repositories archived by the UCLA library. We received an exemption from UCLA's Institutional Review Board (UCLA IRB Exemption 19-001257) to make use of this data, as neither we nor the UCLA library had access to any personal identifying information (PII) nor any key to access such information. To ensure that we were in compliance with IRB approvals, prior to beginning our work, we confirmed that the datasets we accessed from the library contained no PII.

For the Pizzagate conspiracy theory, the library based their collection on the Reddit subreddit, r/pizzagate. As with many other conspiracy theories, the community discussing and negotiating the boundaries of Pizzagate archived their own discussions, particularly given their legitimate concern that Reddit was considering banning their subreddit [61]. The Pizzagate community moved their discussions to Voat in the aftermath of Reddit's decision, and continued their discussions on v/pizzagate. This data collection approach mirrors that of other research on conspiracy theories emerging and circulating on social media [62] [24]. As part of their initial collection process, the UCLA library confirmed that research use of the materials was in accordance with the terms of service of the sites. In addition, as part of their data preparation process, the library ensured that the collection was free from PII. After accessing this data through the UCLA library, we removed images, urls, videos, advertisements, and non-English text strings to create our research corpus, pizzagate.txt. To the best of our knowledge and to the best of the knowledge of the library, neither our corpus nor the library data contains data from private discussions, private chat rooms, or any other sources with restrictions on access for public use or that may violate the terms of our IRB exemption.

For Bridgegate, we relied on an archive of news reports developed by the UCLA library from a series of sources focusing on the northern part of New Jersey. This collection is also available through the UCLA library's archive site. The seed articles for the initial collection were either tagged or otherwise directly categorized as being about the closure of the lanes on the George Washington Bridge, and additional articles were indexed based on that initial seeding. We subsequently cleaned this collection to remove images, urls, videos, advertisements, and non-English text strings to create our research corpus, bridgegate.txt. Both of our data corpora can be accessed through a University of California Dash archive (DOI:10.5068/D1V665).

Pizzagate was "uncovered" by conspiracy theorists making use of the Wikileaks dump of emails hacked from the DNC servers, particularly those of John Podesta, who had served as the campaign manager for Hillary Clinton's unsuccessful run for the presidency in 2016. Through creative interpretations of these emails, conspiracy theorists alleged that they had

discovered Hillary Clinton's involvement in a child sex trafficking ring being run out of the basement of a Washington DC pizza parlor, "Comet Ping Pong". The conspiracy theory took root with a series of tweets in early November 2016, with the first appearance of the #Pizzagate Twitter hashtag on November 6, the day before the US presidential election [63]. Discussions of the conspiracy theory per measures from activity on Twitter lowered in December 2016, around the time that Welch was apprehended with his gun outside of the restaurant after surrendering to police [63]. Since then, Pizzagate has reappeared as part of the much larger QAnon conspiracy theory that began to develop in late October 2017.

The Bridgegate conspiracy, by contrast, was discovered by investigative reporters to be a political payback operation launched by the inner circle of New Jersey Governor Chris Christie, making use of their close alliances with highly placed officials in the Port Authority. The conspirators took aim at the Democratic mayor of Fort Lee, New Jersey, Mark Sokolich, who had refused to endorse the governor in his reelection bid. Christie's assistants conspired with members of the Port Authority to close several toll lanes to the George Washington Bridge, thereby causing catastrophic traffic jams that lasted for a week in early September 2013. When asked, these people said that the lane closures were part of a traffic study. A formal investigation into the decision to close the lanes was launched in 2014 and, during the ensuing five years, the overall contours of the conspiracy were revealed and various actors were indicted, tried and sentenced to prison. In late 2019, a petition filed by several of the conspirators was granted review by the U.S. Supreme Court, with initial oral arguments occurring in early 2020.

The Pizzagate data set consists of 17,498 posts yielding in 42,979 sentences, with an end date of February 2018. We used a similar end date for Bridgegate, and thus worked with an archive of 385 news reports comprising 20,433 sentences. Because of this end date, we missed the events of April and May 2019 based on the revelations of one of the main conspirators, Bridget Ann Kelley, subsequent to her sentencing for her role in the conspiracy. These revelations highlighted the role of an otherwise seemingly unimportant actant, Walter Timpone, and added several new relationship edges to the Bridgegate narrative framework. The fact that additional information related to an actual conspiracy emerged over a prolonged

period of time (here, five and a half years) might be one of the tell-tale signs distinguishing a conspiracy from a conspiracy theory. For Pizzagate, despite the three year scope of this study, the number of actants in the narrative remained stable beginning one month after the data collection period began.

Although Pizzagate was accessible through r/pizzagate and v/pizzagate, and the Bridgegate conspiracy was reported and archived by newspapers covering New Jersey politics, our approach does not require pre-established data sets. While access to comprehensive data collections eliminates an initial step in the narrative framework discovery pipeline, we have demonstrated methods for determining active domains of discussion in any collection of internet resources based on topic modeling [9] [64]. Although the selection of a target domain using this and similar approaches might result in overlooking posts related to a broader discussion, work on community formation suggests that people interested in a particular topic seek out forums where such topics are discussed and develop close knit communities [62] [65] [66] [67]. The first step in the pipeline can be tuned to capture actants that may be of interest; the extent of a domain can be discovered from there. In earlier work, we implemented this approach, and showed how a hierarchical topic-modeling method reveals broad topics of discussion in a large social media space that we identify as knowledge domains [9]. Posts, discussions and articles related to those knowledge domains can then be selected to constitute the study corpus. Cleaning the data results in a machine actionable corpus similar to those we developed for Pizzagate and Bridgegate. There are many other approaches that can be applied to the selection of target corpora from larger social media domains, although topic modeling has been to used to great effect in the context of social media [24] [68] [69].

4.2 Results

The joint estimation of the narrative framework network described in the Methods section relies initially on the relationship extractions. This process provides us with a ranked list of candidate entities used to seed the discovery of subnodes and supernodes, and a series of inter-

actant relationships (Table 4.1). For each of the two corpora, we find a very large number of relationships of various types and patterns (Fig 4.1). After tokenizing and stemming the extracted headword lists, the resulting unsorted grouping provides a seed for the subnode lists and supernode lists. Once we take the union of the arguments with each of these terms, and determine the BERT embedding for each argument, k-means clustering (k=20) results in a series of subnodes. After pruning and merging, we determine the supernodes and their corresponding subnodes for each narrative framework (Table 4.2).

	Supernodes	Subnodes	Rel Extractions	Labeled Rel	Avg Degree
Pizzagate	24	88	749	438	36
Bridgegate	134	144	5855	928	72

Table 4.1: **Summary statistics for the extracted graphs from the two corpora.**

To evaluate our actant discovery, we compare the actants discovered by our methods with those in the gold standard evaluation data. Even when we limit our actant list to those mentioned more than fifty times in the corpus, our methods provide complete coverage of the actants in the evaluation data. For Pizzagate, we provide comparisons with the illustration alone and with the expert labeled data, which includes the Edgar Welch meta-narrative (Table 4.3).

Our methods perform less well when actants are mentioned infrequently. The actant, "cannibalism", for instance, has a very low frequency mention in the Pizzagate corpus (4 mentions), and does not appear among the top ranked actants. Its inclusion in the NY Times illustration is visually arresting, however, which may in part explain why the illustrator chose to include it. By way of contrast, several highly ranked actants, including Bill Clinton and the Clinton foundation, do not appear in the NY Times illustration but are mentioned frequently in the Pizzagate discussions (Figure 4.2).

Similarly, some of the actants identified by the NY Times for Bridgegate are mentioned with relatively low frequency in our corpus. If, for example, we limit our actant list to only those mentioned more than 150 times (as opposed to 50 times), we miss five actants and

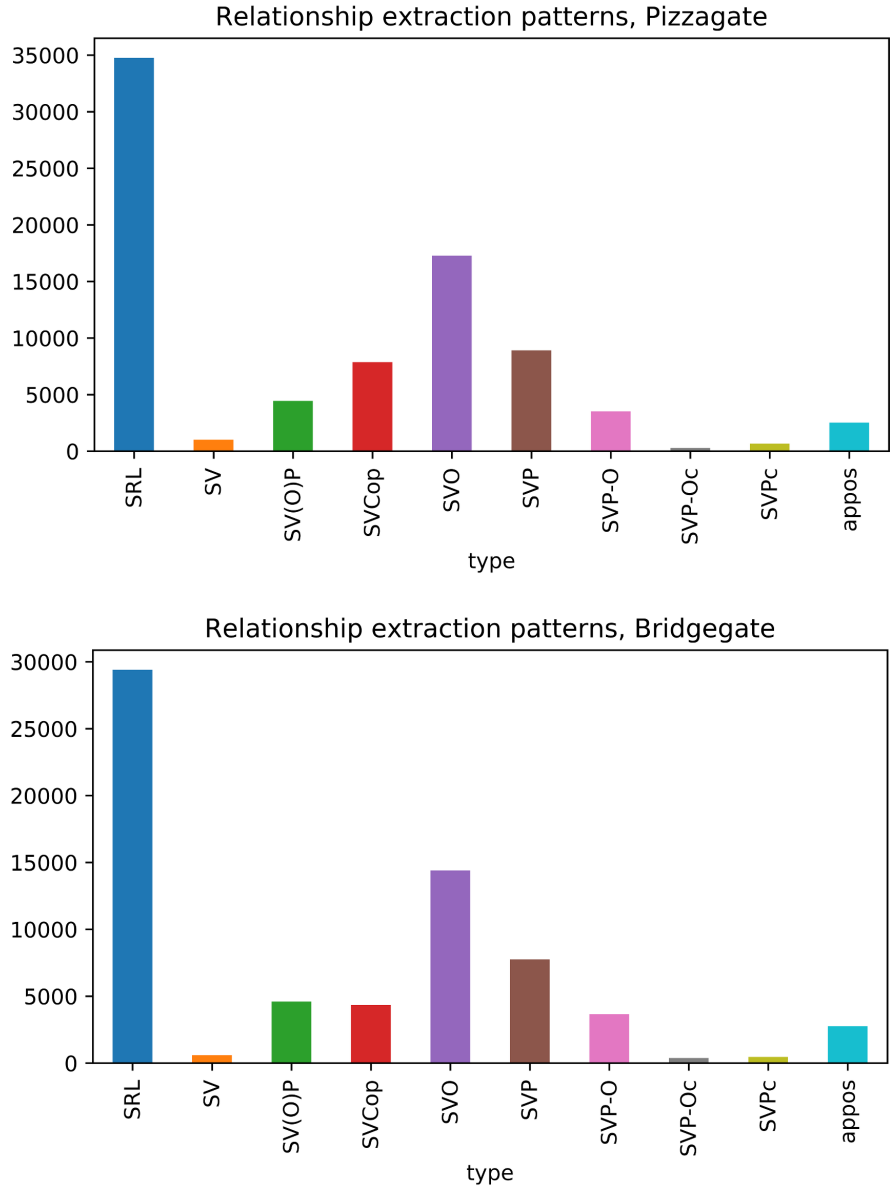

Figure 4.1: **Relationship extraction patterns.** Patterns by total number for A: Pizzagate (top) and for B: Bridgegate (bottom). For example, SVO is (nsubj, verb, obj), SRL is (A0, Verb, A1) and (A0, Verb, A2). A larger list can be found in supporting information (SI)

their various relationships (Lori Grifa, Evan Ridley, Phillip Kwon, Paul Nunziato and Nicole Crifo). These "misses", however, are replaced by actants such as Randy Mastro, a former federal prosecutor whose report exonerated Christie, Michael Critchlet, Bridget Anne Kelly's attorney, and Shawn Boburg, whose reporting broke the scandal, all of whom are central to

Table 4.2: **A sample of the top 5 supernodes and subnodes for Pizzagate and Bridgegate.**

Pizzagate		**Bridgegate**	
Supernodes	*Subnodes sample*	*Supernodes*	*Subnodes sample*
[Podesta]	John Podesta, Tony Podesta, leaked Podesta email, Podestas, Podesta	['christie', 'christi', 'christies', 'governor', 'chris', 'former']	christie governor, chris new jersey governor, christie
['pizza', 'comet', 'ping', 'pong']	comet pizza, comet pizza story, ping pong comet, comet, ping pong review facebook	['authority', 'author', 'authorizing', 'authorities', 'authors', 'authorization', 'port', 'executive']	['authority port', 'report authority port', 'executive director', 'baroni executive director', 'report', 'authority transportation']
[alefantis]	James alefantis, alefantis, james alefantis instagram, owner james alefantis	['wildstein', 'david']	['wildstein', 'wildstein david', 'wildstein david executive former']
[traffick]	child sex trafficking, ring trafficking, ring trafficking, human pedophilia trafficking	['lee', 'fort', 'mayor', 'sokolich']	['sokolich', 'fort lee', 'sokolich mark mayor', ' mayor effort sokolich', 'lee fort lane traffic']
[child]	child, child porn, child trafficking	['bridges', 'bridge', 'george', 'washington', 'lane']	['scandal bridge bridgegate', 'closure lane', 'george bridge washington closure lane', 'bridgegate', 'bridget kelly', 'closure gwb controversy lane', 'lane']

Table 4.3: Comparison of pipeline actant discovery with the gold standard evaluation data.

	New York Times	Pipeline Discovery	Matched > 50	Matched anywhere
Pizzagate (illustration)	21	88	20	21
Pizzagate (expert)	23	88	22	23
Bridgegate	36	144	36	36

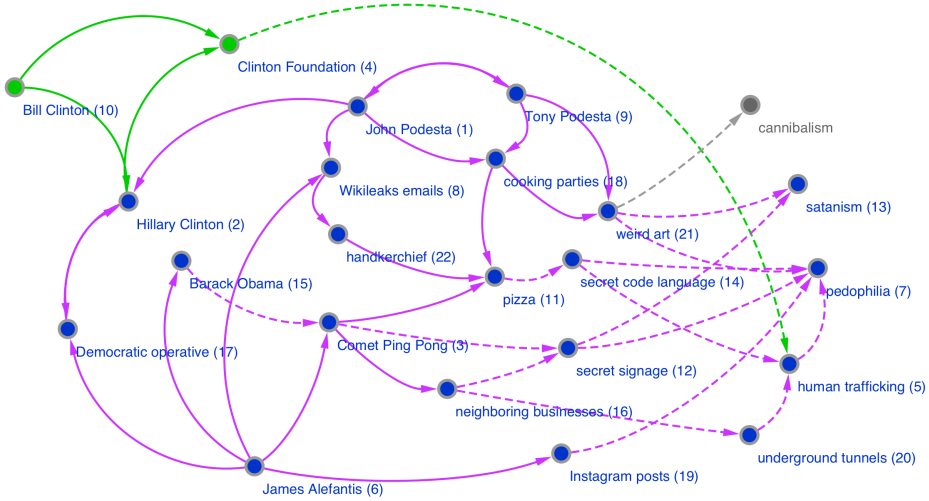

Figure 4.2: **Comparison of our results with the NY Times Pizzagate hand-drawn graph.** Edges and nodes that we do not discover in the top ranked actants through the pipeline are greyed out (cannibalism). Highly ranked edges and nodes that we discover not included in the NY Times illustration are in green (Bill Clinton and Clinton Foundation). We maintain the visual convention of dashed lines that the NY Times uses to identify relationships based on the interpretation by the conspiracy theorists of hidden knowledge. Immediately following the node label is the ranking of the actant as discovered by our pipeline.

the conspiracy and reporting on it.

Relationships between supernodes can be discovered by collapsing the subnode subgraphs, and labeling the edges between supernodes with the relationship with the highest relevance score over the subgraph edges (for example, Fig 4.3). A table summarizing the comparison of our relationship extractions and aggregations with the evaluation data lists the number of

edges in the NY Times illustrations, the number of expert labeled edges in the gold standard corpus, and the overall number of automated aggregated relationship extractions from our pipeline, as well as the recall of our extractions against the gold standard relationships, the average cosine similarity score for matched edges, and the standard deviation for this measurement (Table 4.4).

Table 4.4: Comparison of pipeline inter-actant relationship discovery with the NY Times and the gold standard corpora.

	NY Times illustration	Gold-standard corpus	Automated Extractions	Recall	Avg cos similarity	Std Dev
Pizzagate	35 (27 unlabeled)	38	749	83.7%	0.95	0.048
Bridgegate	46	122	5855	82.9%	0.89	0.0483

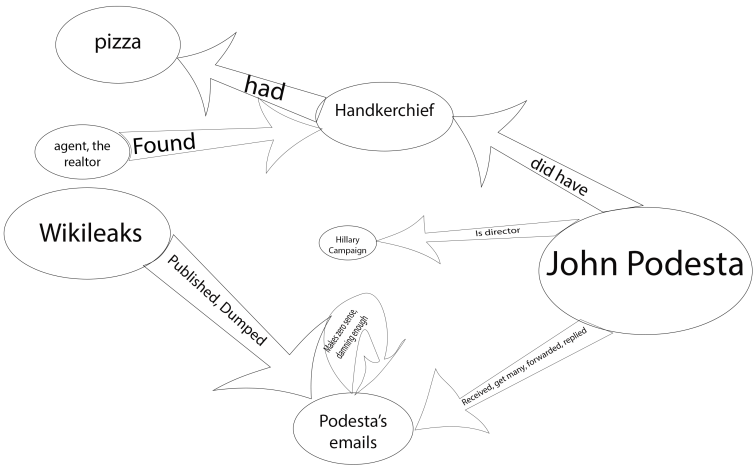

Figure 4.3: **A subnetwork demonstration of the Pizzagate narrative framework.** Some of the nodes are subnodes (e.g. "Podesta's emails"), and others are supernodes (e.g. "John Podesta"). Because we only pick the lead verbs for labeling edges, the contextual meaning of relationships becomes clearer when one considers the entire relationship phrase.

4.3 Discussion

Network representations of complex narratives are widely recognized as providing support for understanding narratives, including the actants, their various roles and the numerous inter-actant relationships of which they are comprised [70] [71]. Our methods allow us to derive these narrative frameworks automatically, and present them as a network graph. Visual representations of narrative networks such as those included in the NY Times and the ones we generate have also become a part of the reporting on complex events including conspiracy theories such as Pizzagate and conspiracies such as Bridgegate.

Running the community detection algorithm on Pizzagate reveals thirty-three core communities. After applying the thresholds on actant mention frequency and community co-occurrence as described in the Methods section, here $P_{th_1} = 0.7$ and $P_{th_2} = 0.4$, we discover a series of seven cores, corresponding to the main Pizzagate domains, as well as five nucleations of potentially emerging narrative frameworks, two meta-narrative cores, and two possibly unrelated cores. A visualization of those communities that only include subnodes with mention frequencies greater than or equal to the corpus average mention of 265 reveals the distinct narrative framework underlying the Pizzagate corpus, where central cores have a large number of edges connecting them (Fig. 4.4). Subsets of the supernodes define four domains that, outside of Pizzagate, would probably be less connected: (i) Democratic politics, where actants such as Hillary Clinton and Obama are dominant; (ii) the Podestas, with John Podesta as the major actant; (iii) casual dining, dominated by James Alefantis and Comet Ping Pong; and (iv) Child Sex Trafficking and Satanism, where actions such as child abuse and sex trafficking, and actants such as children and rituals are common. Nodes in the self-referential meta-narrative, "Pizzagate", have many edges connecting them to these core domains, while the narrative nucleations and unrelated discussions do not. This lack of connection suggests that they are not central to the Pizzagate narrative and their corresponding mention counts further reveal that they are not discussed frequently.

It is interesting to note that the Wikileaks domain, dominated by actants such as email and Wikileaks, provides the glue for the narrative framework. After eliminating the relationships

generated by the Wikileaks subnodes, the connections between the other domains disappear, leaving them as a disjoint series of smaller connected components (Fig 4.5). This disjuncture only occurs when the links generated by the Wikileaks subnodes are eliminated.

When we remove the mention frequency threshold of 265, the visualization is populated with the subnodes comprising the various communities (see Figure 4.6). Apart from the central Pizzagate narrative framework, an important meta-narrative component describing Edgar Welch's investigations of Comet Ping Pong appears. Although this meta-narrative is mentioned in the NY Times article on Pizzagate that accompanies the illustration, it was not included in the illustration [48]. Importantly, our detection of the Welch meta-narrative matches the annotations of the expert annotators as reported in the evaluation of our results above. A second meta-narrative focuses on Pizzagate as a topic of discussion, and includes references, for example, to the Pizzagate hashtag. Apart from these two meta-narrative communities, there are several other communities that we detect in the overall graph: (i) communities that provide background or support for the central Pizzagate narrative framework; (ii) communities that may represent nucleations of other narrative frameworks; and (iii) communities that are unrelated to the Pizzagate narrative framework.

Several of the communities providing background describe the internet itself and various social media platforms, thus presenting a general internet-based context for these discussions. Two additional background communities focus on general discussions about pedophilia, and various allegations against people such as the British entertainer and sexual abuser, Jimmy Savile. A final large background community focuses on American politics writ large, and provides the domain from which the various democratic operatives and Obama are drawn.

Other communities represent the beginnings of other narrative frameworks, which either represent indigenous nucleations of new narrative frameworks, or the intrusion of additional narrative frameworks from outside the target forum. One of these communities is redolent of anti-Semitic conspiracy theories related to Hollywood, which are common in other conspiracy theory forums, and may indicate the existence of conversations that propose links between Pizzagate and these broader conspiracy theories [24]. The rise of QAnon, which includes both anti-Semitic conspiracy theories and the Pizzagate conspiracy theory, suggests that this may

be the case [72]. Another small nucleation relates to suspicions that the Red Cross traffics in human organs. Other components that may represent emerging narrative frameworks include a community focused on Rabbi Nuchem Rosenberg and his efforts to reveal child abuse in certain Orthodox Jewish communities, and a community that includes narrative components related to secret orders within the Catholic Church, including the Knights of Malta. One final nucleation presents a possible narrative about George Soros, 9/11, Russia and Nazis.

There are two unrelated communities–one focused on discussions of aliens and alien films, including the film *Alien*, while the other is related to discussions about police and FBI investigations of Acorn and Katt Williams. These last two communities reveal how our methods work even with very noisy data that may include conversations not immediately relevant to the discovery of the narrative framework(s) in the target corpus. It is important to note that all of these non-central components are comprised of actants and their relationships that have lower than average frequency mentions in the corpus.

For Bridgegate, we discover a much simpler community structure, with a single giant connected component of 386 nodes. The community detection algorithm finds twenty-three communities, but only three of them have 20 or more nodes, with a mean size of 6.65 and a median of 3 for the remaining communities. This result is not surprising given that all of the actants in the Bridgegate conspiracy come from a single domain, namely that of New Jersey politics. Consequently, the narrative framework is not stitched together through the alignment of otherwise weakly connected domains, but rather is fully situated in a single domain. Similarly, there is no information source, such as Wikileaks, on which the framework depends to maintain its status as a single connected component. Even the deletion of a fairly important actant, such as Bridget Kelley along with her relationships, does not lead to a series of disjoint subgraphs as was the case in Pizzagate when the Wikileaks associated nodes were deleted. Indeed, even if all of the Bridgegate actants' conspiracy-related relationships were deleted—as if the conspiracy had never happened—New Jersey politics (for better or worse) would continue to exist as a giant connected component.

The automated pipeline for narrative framework discovery provides a clear pathway to de-

veloping a sophisticated, multi-scale representation of narratives. Not only does the pipeline capture the top level nodes and relationships such as the ones proposed by the NY Times in their hand-drawn illustrations, but it also captures additional nodes and relationships. For example, our extractions for Pizzagate include important actants such as Bill Clinton and contributions to the Clinton campaign and foundation, which are missing in the NY Times graph but were clearly central to the discussions among Pizzagate conspiracy theorists. Our approach also reveals certain details not captured by the hand-drawn illustrations, such as the central role played by Wikileaks in the Pizzagate conspiracy theory forums in gluing the otherwise disconnected domains of the narrative framework together. Indeed, these findings support our hypothesis that conspiracy theories are built by aligning otherwise unrelated domains of human interaction through the interpretation by the conspiracy theorists of discovered or hidden knowledge to which they claim either to have special access or a particularly astute interpretive ability.

An important aspect of our narrative framework discovery is its generative nature. Once the narrative framework is established, one can generate admissible stories or story parts (e.g. forum posts) that conform to the overarching framework by selecting already established actants and relationships. Although such a capacity might be used to create and perpetuate conspiracy theories, it might just as easily be deployed to interrupt narrative frameworks fueling anti-democratic behaviors or encouraging people to take destructive, real-world action. At the very least, our approach allows for deep and powerful insight into story generation, and the underlying factors that allow people to participate in the creation and circulation of these narratives. Similarly, understanding the significant structural differences in narrative frameworks between folkloric genres such as rumors, legends and conspiracy theories on the one hand, and factually reported conspiracies on the other hand, could be useful for testing the veracity of emerging narratives and might prove to be an important component of tools for private and public sector analysts.

4.4 Conclusion

We have observed the rise of the term known as "fake news" in the past few years and Trump presidency era in particular. Lazer et al propose that fake news be understood as "fabricated information that mimics news media content in form but not in organizational process or intent" [73]. Distinguishing fact from fiction given the sheer amount of generated stories and the speed of spreading is not easy. Accordingly, there is a pressing need, particularly in light of events such as the COVID-19 pandemic, for methods to understand not only how stories circulate on and across these media, but also the generative narrative frameworks on which these stories rest. Recognizing that a series of stories or story fragments align with a narrative framework that has the hallmarks of a fictional conspiracy theory might help counteract the degree to which people come to believe in—and subsequently act on—conspiracy theories.

We hypothesize that three features—a single domain of interaction, a robustness to deletions of nodes and relationships, and a proliferation of peripheral actants and relationships—are key characteristics of an actual conspiracy and may be helpful in distinguishing actual conspiracies from conspiracy theories. Reporting on actual conspiracies introduces new actants and relationships as part of the process of validating what has actually happened. This reporting feeds the core giant network with more evidence, resulting in a denser network over time. Conspiracy theories, by way of contrast, may form rapidly. Since the only evidence to support any of the actants and relationships comes from the storytellers themselves, we suggest that the network structure of a conspiracy theory stabilizes quickly. This stabilization is supported by studies in folklore, which reveal that an essentially constant and relatively small set of actants and relationships determines the boundaries of admissible stories (or story fragments) after the initial narrative burst finishes [74] [75] [76]. The addition of new domains through the process of alignment described above and symptomatic of the monological beliefs identified as a common feature of conspiracy theories may at times alter an otherwise stable framework, with sudden changes in the number of actants and relationships included in the network. In short, it seems likely that **a conspiracy theory is characterized by a comparatively small number of actants, multiple interconnected**

domains, and the fragility of the narrative framework graph, which can easily be disconnected into a series of disjoint subgraphs by the deletion of a small number of nodes or relationships. Our methods can help derive the narrative frameworks undergirding a corpus, and support the macroscopic analysis of these complex narratives.

Conspiracy theories have in the past been disregarded as the implausible fantasies of fringe members of society, not worthy of serious concern. An increasing awareness that people are making real-world, and at times violent or dangerous, decisions based on informal stories that circulate on and across their social networks, and that conspiracy theories are a significant part of that storytelling, countermands that idea. The rapid spread of conspiracy theories such as Pizzagate, COVID-19 conspiracies, and the capacious QAnon, coupled to the dangerous real world actions that people have taken based on a belief in these narratives, are no longer purely a fringe phenomenon. Consequently, knowledge derived from our methods can have clear and significant public safety impacts, as well as impacts on protecting democratic institutions.

Actual conspiracies and conspiracy theories threaten Democracy each in their own particular way. An actual conspiracy usually comes to light because of the investigative capacities of a free and independent press, and reveals corruption in government or industry; as such, the discovery of an actual conspiracy confirms the power of democratic institutions. Conspiracy theories, on the other hand, seek to undermine the very premise of democratic institutions. As Muirhead and Rosenblum note, "There is no punctilious demand for proofs, no exhausting amassing of evidence, no dots revealed to form a pattern, no close examination of the operators plotting in the shadows. The new conspiracism dispenses with the burden of explanation" [77]. Given the challenges that conspiracy theories present to democracy and a free and open society, we believe that the ability to automatically discover the underlying narrative frameworks for these accounts is of paramount importance. Such an awareness will, at the very least, provide insight into the type of muddled thinking promoted by propaganda campaigns [78] or other disinformation initiatives. It will also offer a clear overview of the domains of knowledge that conspiracy theorists link together through their imaginative interpretations of "hidden knowledge". Identification of the structural aspects of a conspir-

acy theory narrative framework fueling online conversations, such as the weak connection of multiple domains, can alert us to whether an emerging narrative has the hallmarks of a conspiracy theory. Importantly, these methods can provide insight into the potential strategies that adherents may be considering for dealing with the various threats identified in the narratives. Taken as a whole, the automated narrative framework discovery pipeline can provide us with a better understanding of how stories help influence decision making, and shape the contours of our shifting political environment.

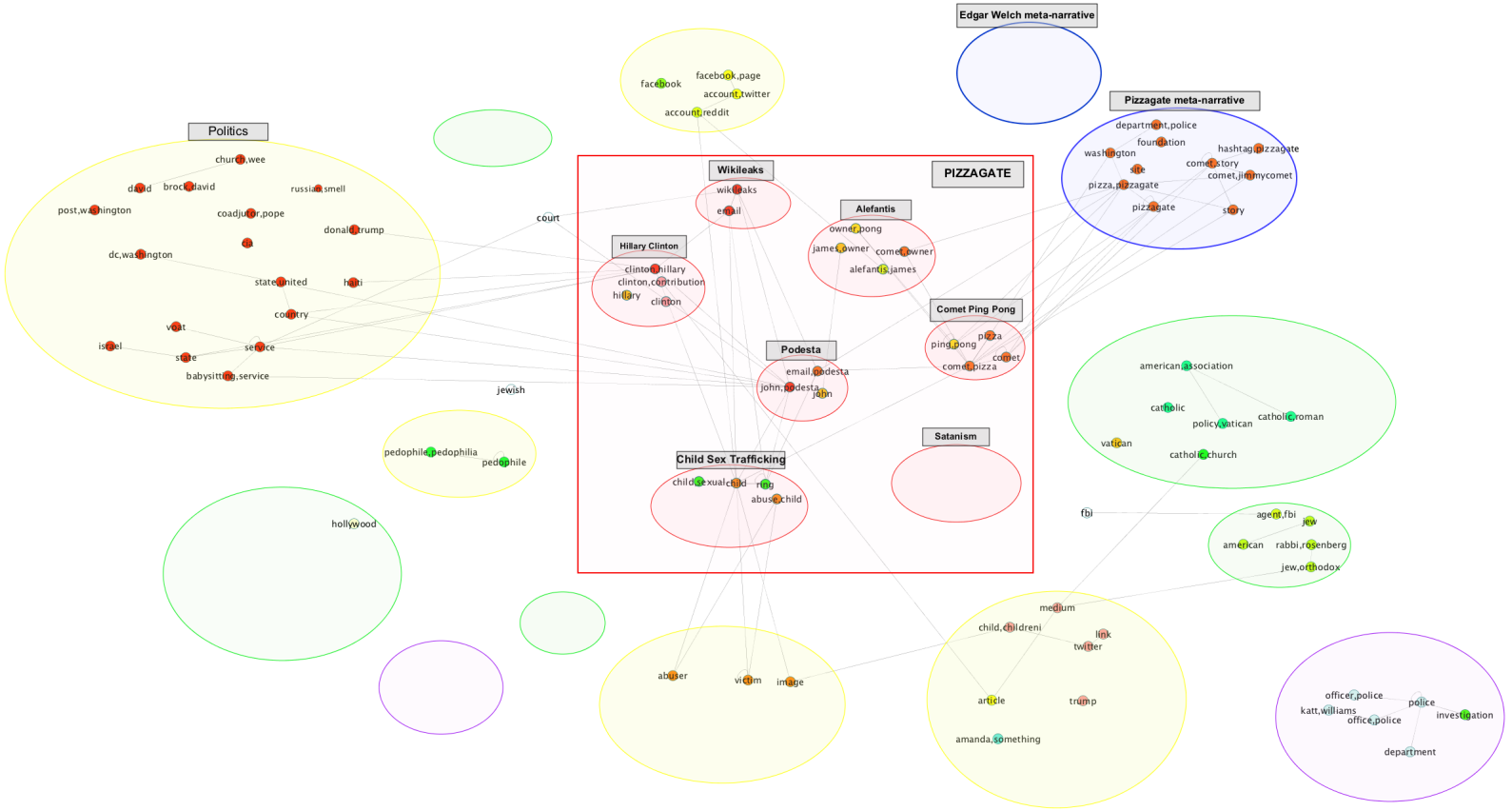

Figure 4.4: **Identification of the Pizzagate narrative framework from the Pizzagate corpus.** Subnodes with a mention frequency count < 265 and their edges are removed from the community-partitioned network obtained from Algorithm 1 (See Fig. 4.6 for the network before filtering). Solid nodes are core nodes, while nodes without color, such as "fbi", are non-core nodes. Colors are based on the core nodes' assigned community, while all relationships are collapsed to a single edge. These core nodes have an assignment based on the $P_{th_1} = 0.7$ threshold, while open shared nodes have an assignment based on $P_{th_2} = 0.4$ threshold (see Algorithm 1). Pizzagate subnodes are concatenated into their supernodes, and are outlined in red, while the subnodes retain their community coloring. Contextual communities are shaded with yellow, metanarrative with blue, nucleations with green, and unrelated discussions with purple.

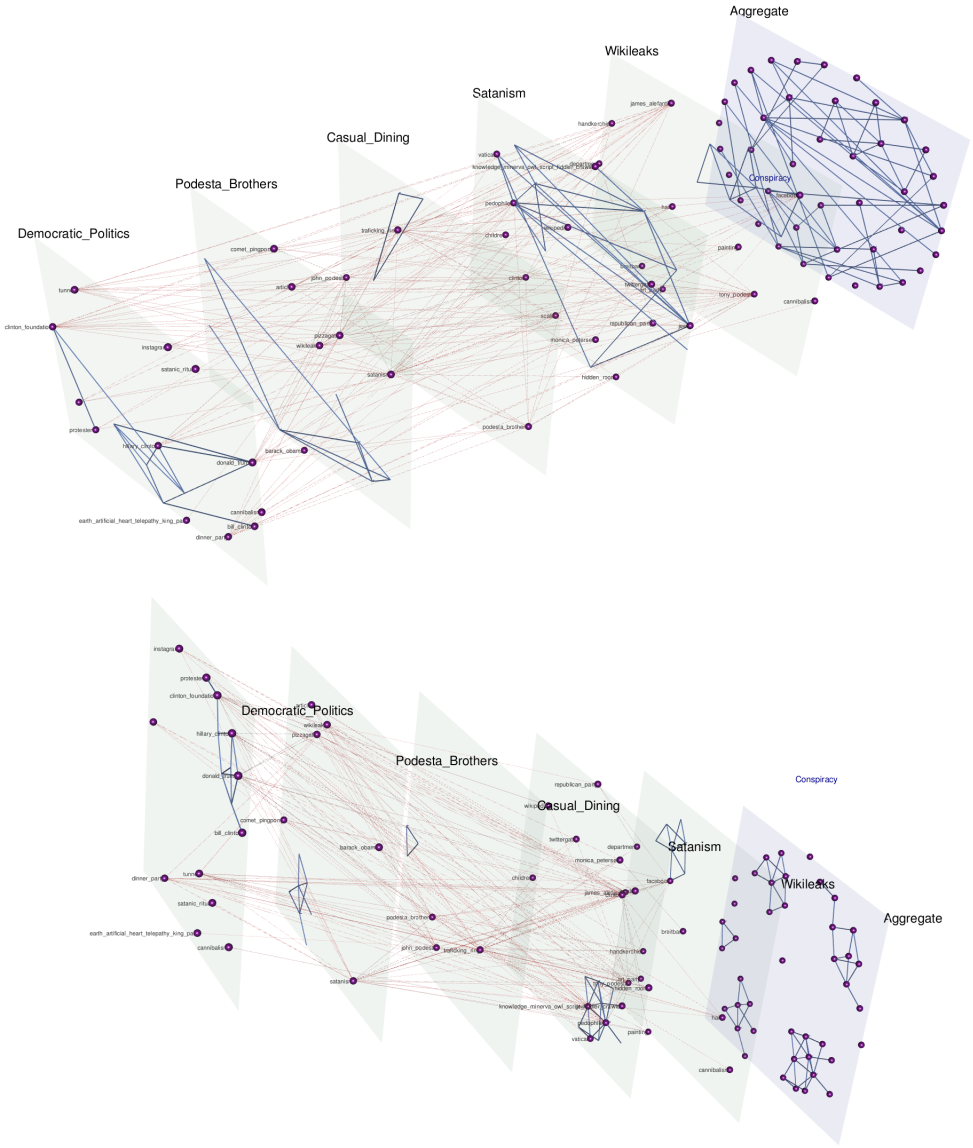

Figure 4.5: **A three dimensional visualization of the narrative framework for Pizzagate in terms of domains.** On the top, **A:** the graph with the inclusion of relationships generated by Wikileaks—the aggregate graph in blue shows a single large connected component. On the bottom, **B:** the graph with the Wikileaks relationships removed, shows on the aggregate level the remaining domains as disjoint components. In the Pizzagate conspiracy theory, the different domains have been causally linked via the single dubious source of the conspiracy theorists' interpretations of the leaked emails dumped by Wikileaks. No such keystone exists in the Bridgegate narrative Network.

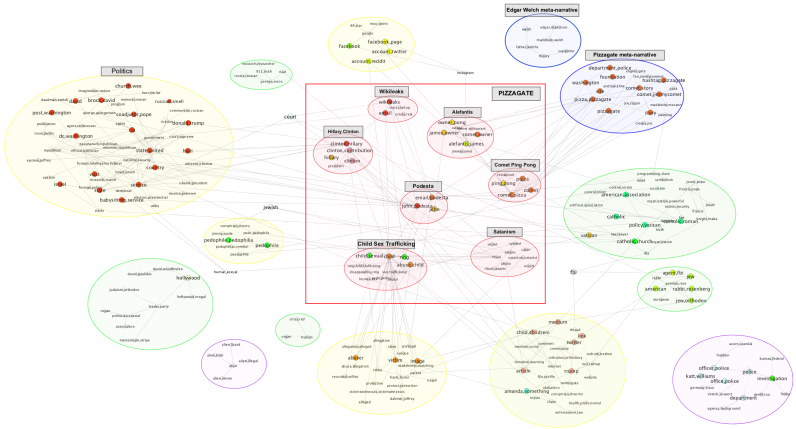

Figure 4.6: **Community detection on the overall Pizzagate corpus.** Subnodes are colored based on their assigned community, while all relationships between any two sub-node actant nodes are collapsed to a single edge. Solid core nodes have an assignment based on the $P_{th_1} = 0.7$ threshold, while open shared nodes have an assignment based on $P_{th_2} = 0.4$ threshold (see Algorithm 1). Main Pizzagate supernodes are outlined in red, and include their subnodes colored by community. Meta-narrative frameworks are shaded with blue. Context groupings are shaded with yellow, while narrative framework nucleations are shaded with green. Unrelated discussions are circled in purple. The entire Pizzagate narrative framework is highlighted with a red box (see Fig. 4.4 for a frequency-filtered version of this figure).

CHAPTER 5

Multi-Conspiracy Theory Study

Rumors and conspiracy theories thrive in environments of low confidence and low trust. Consequently, it is not surprising that ones related to the Covid-19 pandemic are proliferating given the lack of scientific consensus on the virus's spread and containment, or on the long term social and economic ramifications of the pandemic. Among the stories currently circulating in US-focused social media forums are ones suggesting that the 5G telecommunication network activates the virus, that the pandemic is a hoax perpetrated by a global cabal, that the virus is a bio-weapon released deliberately by the Chinese, or that Bill Gates is using it as cover to launch a broad vaccination program to facilitate a global surveillance regime. While some may be quick to dismiss these stories as having little impact on real-world behavior, recent events including the destruction of cell phone towers, racially fueled attacks against Asian Americans, demonstrations espousing resistance to public health orders, and wide-scale defiance of scientifically sound public mandates such as those to wear masks and practice social distancing, countermand such conclusions. Inspired by narrative theory, we crawl social media sites and news reports and, through the application of automated machine-learning methods, discover the underlying narrative frameworks supporting the generation of rumors and conspiracy theories. We show how the various narrative frameworks fueling these stories rely on the alignment of otherwise disparate domains of knowledge, and consider how they attach to the broader reporting on the pandemic. These alignments and attachments, which can be monitored in near real-time, may be useful for identifying areas in the news that are particularly vulnerable to reinterpretation by conspiracy theorists. Understanding the dynamics of storytelling on social media and the narrative frameworks that provide the generative basis for these stories may also be helpful for devising methods to disrupt their

spread.

5.1 Data

Data for this study were derived from two main sources, one a concatenation of social media resources composed largely of forum discussions, and the other a concatenation of Covid-19 related news reports from largely reputable journalistic sources.

We devised a scraper to collect publicly available data from Reddit subreddits and from 4Chan threads related to the pandemic. The subreddits and threads were evaluated for relevance by three independent evaluators, and selected only if there was consensus. All of the data are available in our Open Science Framework data repository [79].[1]

For 4Chan, we collected $\sim$ 200 links to threads for the term "coronavirus", resulting in a corpus of 14712 posts. The first post in our corpus was published on March 28, 2020 and the final post was published on April 17, 2020. For Reddit, we accessed $\sim$ 100 threads on various subreddits with 4377 posts scraped from the top comments. Because these top comments are not necessarily sorted by time but rather by the process of up-voting, we did not include these timestamps in our analysis. Specifically, we targeted r/coronavirus and r/covid19, along with threads from r/conspiracy concentrating on Corona virus. We removed images, URLs, advertisements, and non-English text strings from both sources to create our research corpus. After running our pipeline, we were able to extract 87079 relationships from these social media posts.

For news reports, we relied on the GDELT project, an Open Source platform that scrapes web news (in addition to print and broadcast) from around the world (https://www.gdeltproject.org/).[2] Our search constraints through this dynamic corpus of news reports included a first-order search for conspiracy theories. The corpus was subsequently filtered to only include articles

[1]We ensured that our data was free from personal identifying information (PII), and that our data collection was allowed by the terms of service of the two sites. To the best of our knowledge, neither our corpus nor the news data contains data from private discussions, private chat rooms, or any other sources with restrictions on access for public use.

[2]Research use of the platform is explicitly permitted on the GDELT "about" pages.

written in English (GDELT built-in feature) from U.S. news sources. The top 100 news articles (as sorted by the GDELT engine) were aggregated daily from January 1, 2020 to April 14, 2020 (prior to filtering), and the body of each filtered news report was scraped with Newspaper3K. These articles were then cleaned and staged for our pipeline to extract sentence-level relationships between key actors. We extracted ~ 60 relationships from each report, ~ 50 filtered news reports per day, and 324510 relationships.

5.2 Results and Evaluation

After running the pipeline and community detection, we find a total of two hundred and twenty-nine communities constituting the various knowledge domains in the social media corpus from which actants and interactant relationships are drawn to create narrative frameworks. Many of these communities consist of a very small number of nodes. It is worth noting that several of the communities are "meta-narrative" communities, and focus on aspects of communication in social media (e.g. communities 11 and 74), or platform specific discussions (e.g. communities 44 and 46 that focus on Facebook and 181 focusing on YouTube and Twitter). Other communities are "background" communities and focus on news coverage of the pandemic (e.g. communities 7 and 62), the background for the discussion itself (e.g. community 30 that connects the pandemic to death, and community 35 that focuses on hospitals, doctors, and medical equipment such as ventilators), or discussions of conspiracy theories in general (e.g. communities 108 and 109).

We find that these "meta-narrative" and "background" communities, after thresholding, tend to be quite small, with an average of 3.9 sub-nodes per community. Nevertheless, several of them include sub-nodes with very high NER scores, such as community 155, with only four nodes: "use", "microwave", "hybrid protein" and "cov", all with high NER scores. This community is likely to be included as part of more elaborated conspiracy theory narrative frameworks such as those related to 5G radiation.

The five largest communities, in contrast, range in size from 66 to 172 nodes. These five communities, along with several other large communities, form the main reservoir of actants

and inter-actant relationships for the creation of conspiracy theory narrative frameworks. We find thirty communities with a node count ≥ 14. (See Figure 5.1). Table 5.1 shows the temporary labels for these communities, which are based on an aggregation of the labels of the three nodes with the highest NER scores and node(s) with the highest-degree.

The relationship between the discussions occurring in social media and the reporting on conspiracy theories in the media changed over the course of our study period. In mid to late January, when the Corona virus outbreak appeared to be limited to the central Chinese city of Wuhan, and of little threat to the United States, news media reporting on conspiracy theories had very little connection to reporting on the Corona virus outbreak. As the outbreak continued through March 2020, the reporting on conspiracy theories gradually moved closer to the reporting on the broader outbreak. By the middle of April, reporting on the conspiracy theories being discussed in social media, such as those in our research corpus, had moved to a central position.

The connection between these two central concepts in the news—"coronavirus" and "conspiracy theory"—can also be seen in the rapid increase in the shared neighbors of these sub-nodes (defined in Equation (3.2)) in the overall news graph during the period of study (see Figure 5.2).

Since our dataset contains dated 4Chan *and* GDELT data from March 28, 2020 to April 14, 2020, communities from the social media corpus were explored within the subset of news media between the same dates using Relative Coverage Scores defined in Equation (3.4). The cross-correlation of the ratio of coverage scores for different fixed communities to a random community is provided in Figures 5.3 and 5.4.

The higher average scores for the "5G" community including words such as {"5g", "waves", "antenna", "radio", "towers", "radiation"}, suggests that this community was matched more frequently than other communities compared to a baseline random community. A peak at zero days offset within the time period from March 28, 2020 to April 14, 2020 implies that the news reports are correlated in time to 4Chan thread activity. In addition, these plots suggest that few communities dominate conspiracy theories more than others. The viability

Table 5.1: The largest thirty communities in the social media corpus in descending order of size. The labels are derived from the sub-node labels for the semantically meaningful nodes with the highest NER scores in each community (racially derogatory terms and swears have been skipped). The label of the highest degree node(s) not included in the community label is listed in the third column. Nodes with a threat score ≥ 0.5 are underlined.

ID	Core Size	Community label	High degree nodes
0	172	China, Government, End World	bioweapon
1	89	Chinese, lab, research	truth, animal(s)
5	88	Virus, 5G, cell	Bill Gates, vaccine
6	72	coronavirus, flu, test	SARS
35	66	chloroquine, doctor, patient	hospital
41	53	medium, fact, video	Chinese
21	39	question, trump, impeachment	Fauci
30	39	death, Connecticut, pandemic	CDC
51	32	bacterial, post, economy	YouTube, bot
32	27	medical, misinforming, life	quarantine, Italy
56	25	virus, 5G, vaccine	(radio) frequency
58	24	med, CIA, commie	evidence
74	24	physician, source, official	CNN
57	23	journal, conference, quantum tattoo	Bill Gates
40	22	Chinese, lab, bat	wet market
75	22	American, cognitive dissonance, question	lost cause
7	19	diagnosis, Fauci, wireless	coronavirus
18	19	Wuhan, advancement opportunity, medical company	flu
59	18	guy, repugnant organization, backstab	asymptomatic
82	17	virus, cell, vaccine	thermodynamic load
11	16	financial, community, coronavirus	biolab
42	16	nation, consequence, dedicated worker	attack
62	16	news, cancer, american	(un)reliable
73	16	accepted narrative, scientist, hand	cell level
8	15	corona, chan, 5g dumba**	friend
15	14	coronavirus, test, accurate	Spain
24	14	China, government, weapon	CCP, bioweapon

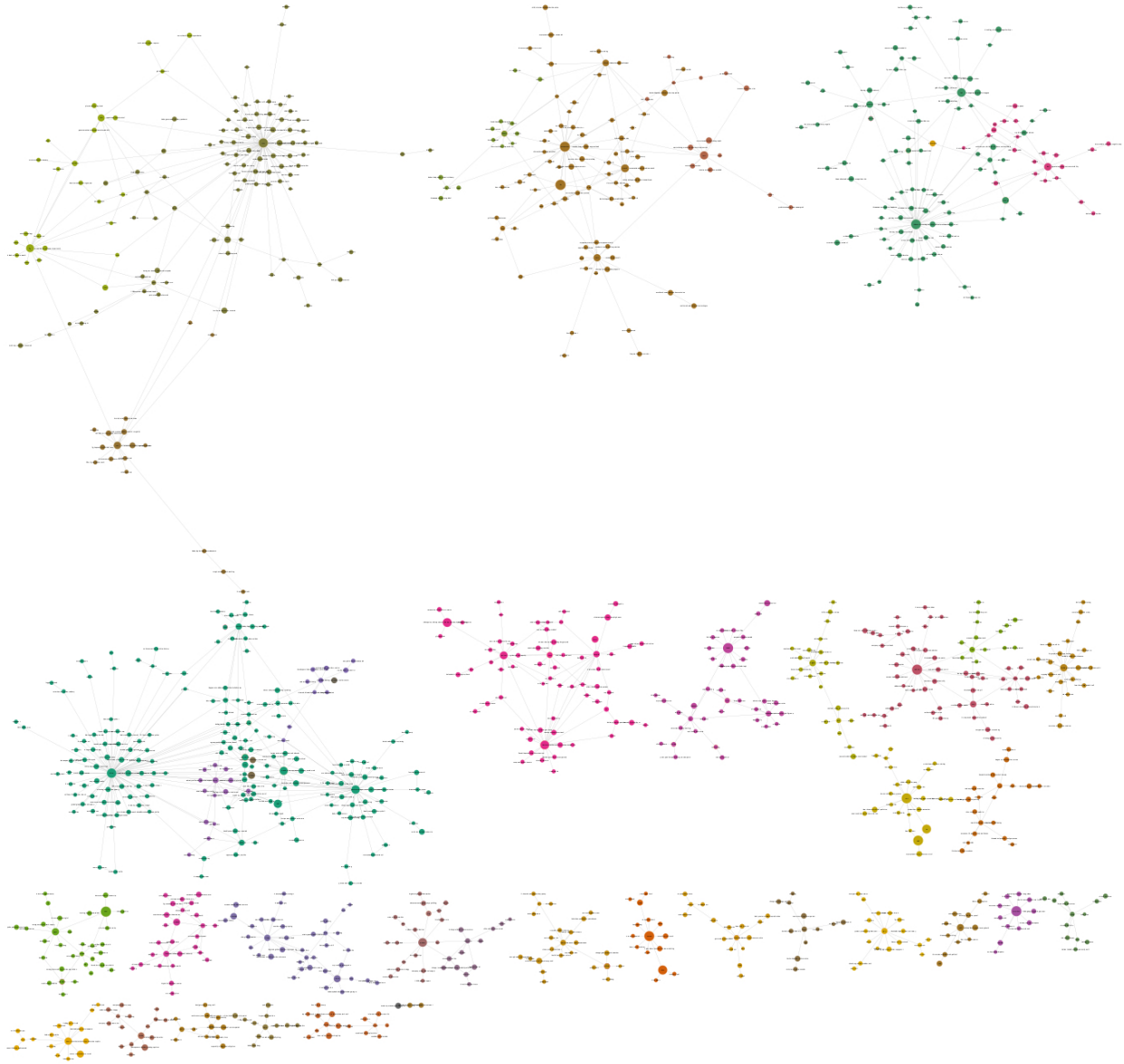

Figure 5.1: **Overview graph of the largest thirty communities in the social media corpus.** Nodes are colored by community, and sized by NER score. Narrative frameworks are drawn from these communities, each of which describes a knowledge domain in the conversation. Nodes with multiple community assignments are colored according to their highest ranked community. An overarching narrative framework for a conspiracy theory often aligns subnodes from numerous domains.

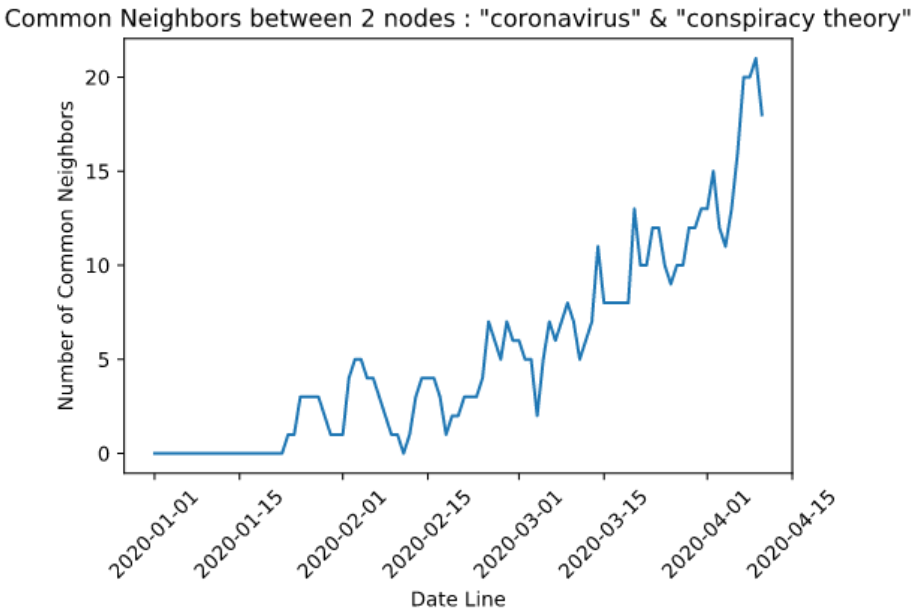

Figure 5.2: **Number of common neighbors between "coronavirus" and "conspiracy theory" over time in the news reports**: Across all 101 segments of 5-day intervals, the number of simple paths empirically increases rapidly, suggesting the closer ties between the two entities across time

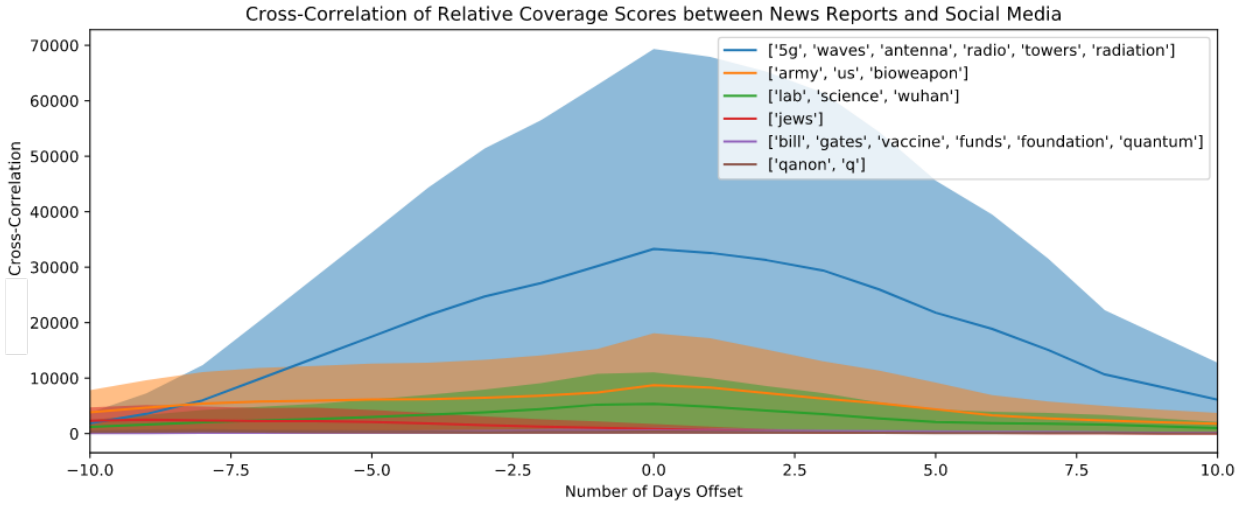

Figure 5.3: **Cross-Correlation of Relative Coverage Score for Word-Level Community Hits in social media against the news reports**: Words in a community are matched to words present in the news reports and social media. Both the news reports and social media are smoothed for 5-day intervals. The mean and standard deviation of the relative coverage score are computed per time stamp across 20 trials with 500 community members each. The peak at 0 days offset suggests that social media and the news are intertwined in a very responsive manner. Mean trajectories show the relative differentiation of each community.

of other communities such as {"army", "us", "bioweapon"} and {"lab", "science", "wuhan"} suggests the lack of a single dominant conspiracy theory *consensus* narrative. Instead, it appears that numerous conspiracy theories may be vying for attention.

We examine "Bill Gates", a key actor frequently found in the common neighbors set between "coronavirus" and "conspiracy theory". Key relationships extracted by our pipeline on the news reports provide a qualitative overview of the emergence of "Bill Gates" as a key actor (see Table 5.4).

Finally, the evaluations based on Algorithm 4 are shown in Figures 5.5, 5.6, 5.8 and 5.7. The plots indicate the saturation of completeness and homogeneity scores at $\sim 92\%$ and $\sim 82\%$ respectively across time. Similarly, the V-measure saturates at $\sim 86\%$. These scores per time sample, represent the fidelity of the process of cluster matching.

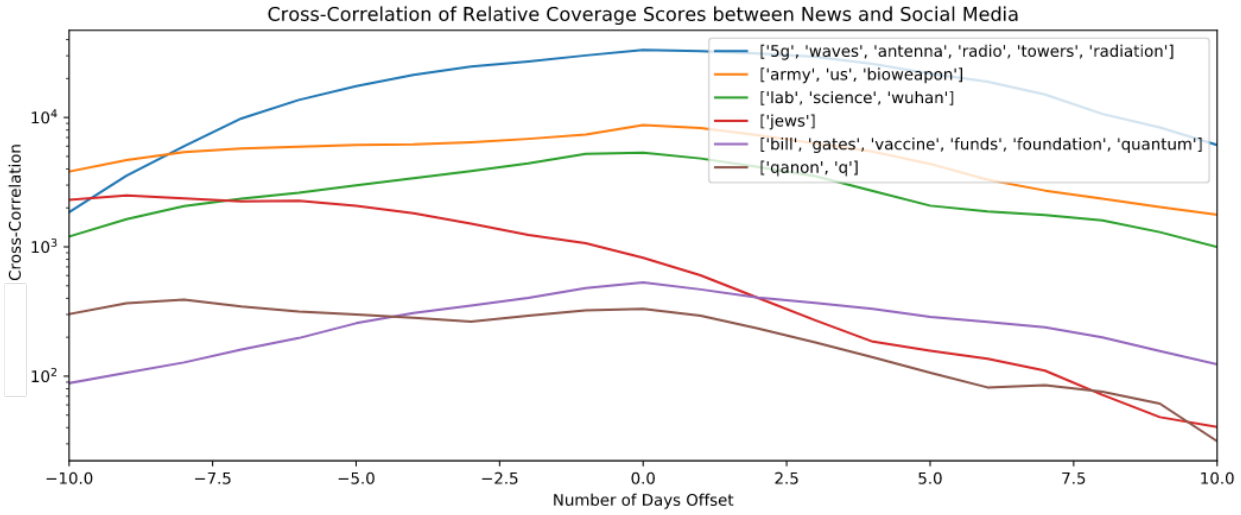

Figure 5.4: **Cross-Correlation of Relative Coverage Score for Word-Level Community Hits in social media against the news reports**: For better visualization we plot Figure 5.3 with y-axis in logarithmic scale.

5.2.1 Descriptive phrase classification of threats from SVCop relationships

The phrase classifier described in the methods was cross-validated and the recall and accuracy across the validation sets are provided in Table 5.2. Recall is used as the primary performance measure in the detection of threats, as the sensitivity to threatening phrases is the most important feature of the classifier.

Table 5.2: Cross-validation (5 fold) result of the phrase classifier

Hyperparameters	Recall	Accuracy
$k = 4$	$73.1\% \pm 3.8\%$	$84.9\% \pm 1.3\%$

5.2.2 Classification of sub-nodes as threats

The phrase classifiers applied to descriptive phrases of a particular sub-node provide insight into the context of the sub-node. For the phrase classifier, Figure 5.10 describes a histogram of the number of sub-nodes across the percentage of associated phrases classified as threats.

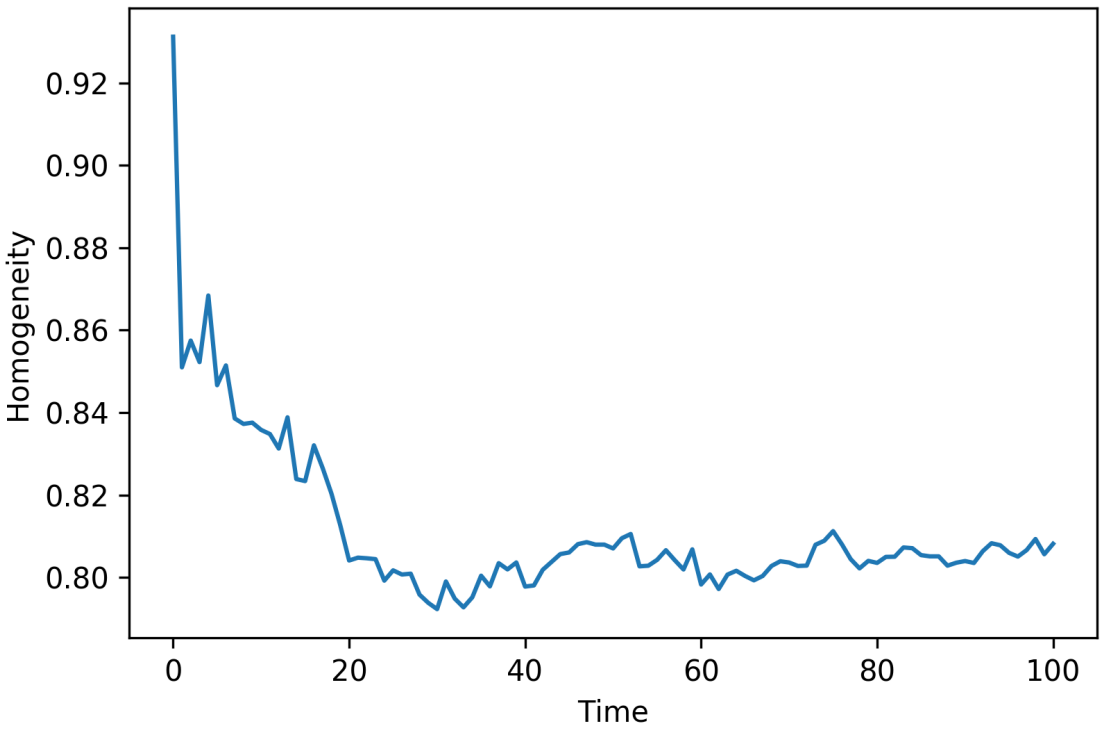

Figure 5.5: **Homogeneity of news based communities** is provided to compare News based communities with Social Media communities. We used Y_{pred} and Y_{gr} derived in algorithm 4 as our cluster label and classes. Homogeneity measures how each cluster contains only members of a single class.

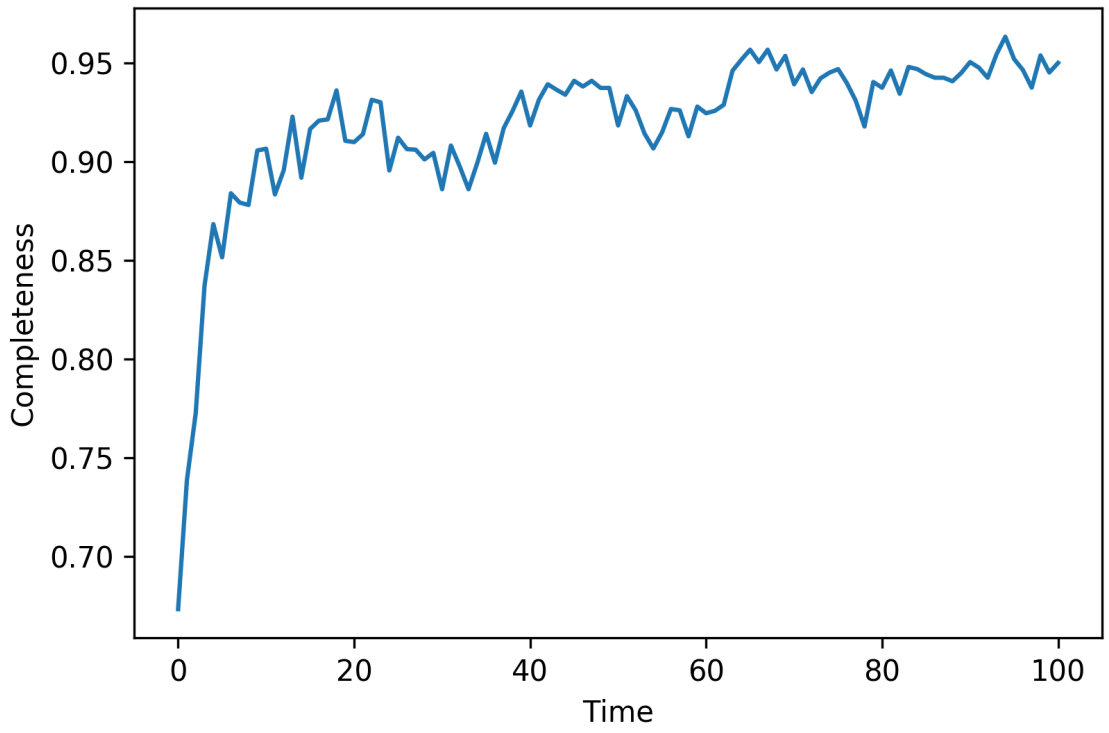

Figure 5.6: **Completeness of news based communities** is provided to compare News based communities with Social Media communities. We used Y_{pred} and Y_{gr} derived in algorithm 4 as our cluster label and classes. Completeness measures how members of a given class are assigned to the same cluster

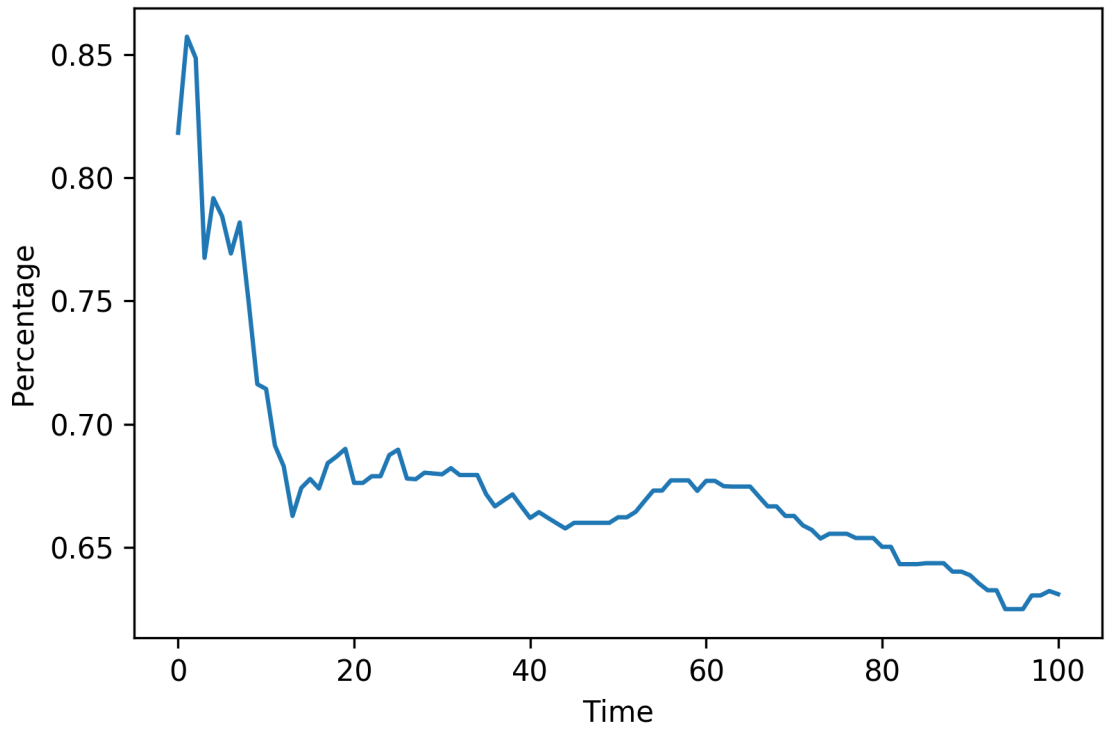

Figure 5.7: **Percentage of coverage** is provided to compare News based communities with Social Media communities. We used Y_{pred} and Y_{gr} derived in algorithm 4 as our cluster label and classes. Coverage percentage is the fraction of actants in news report communities that also are found in social media network communities.

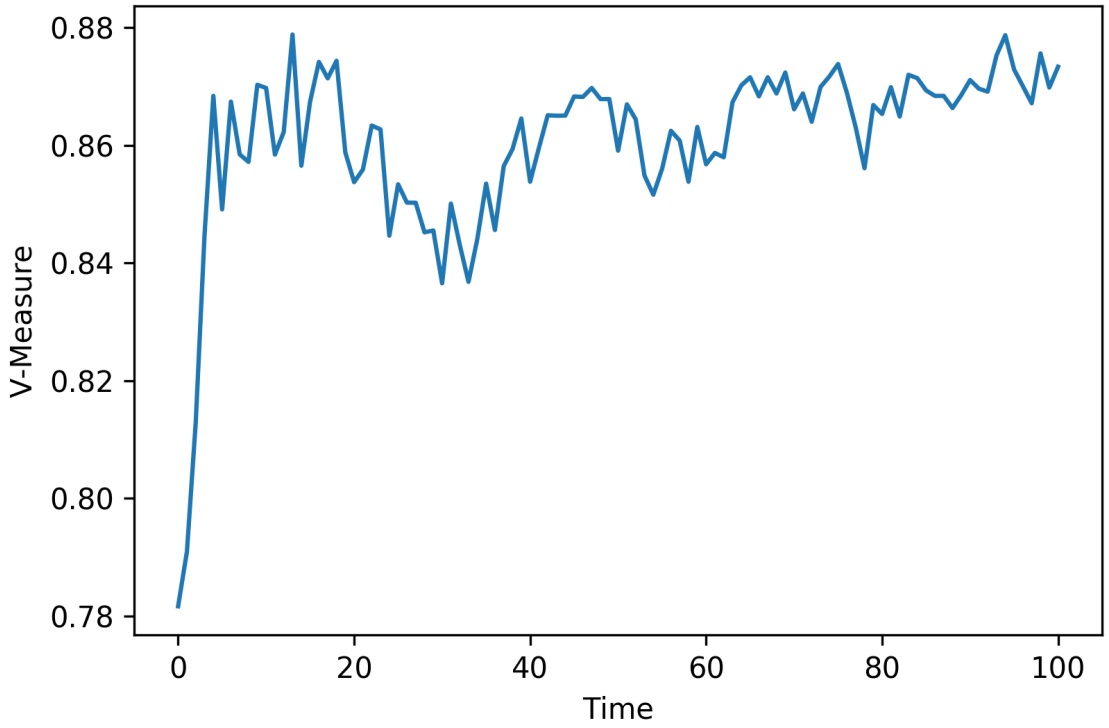

Figure 5.8: **V-Measure** is provided to compare News based communities with Social Media communities. We used Y_{pred} and Y_{gr} derived in algorithm 4 as our cluster label and classes. Completeness measures how members of a given class are assigned to the same cluster, while homogeneity measures how each cluster contains only members of a single class. Their harmonic mean is the V-Measure [1].

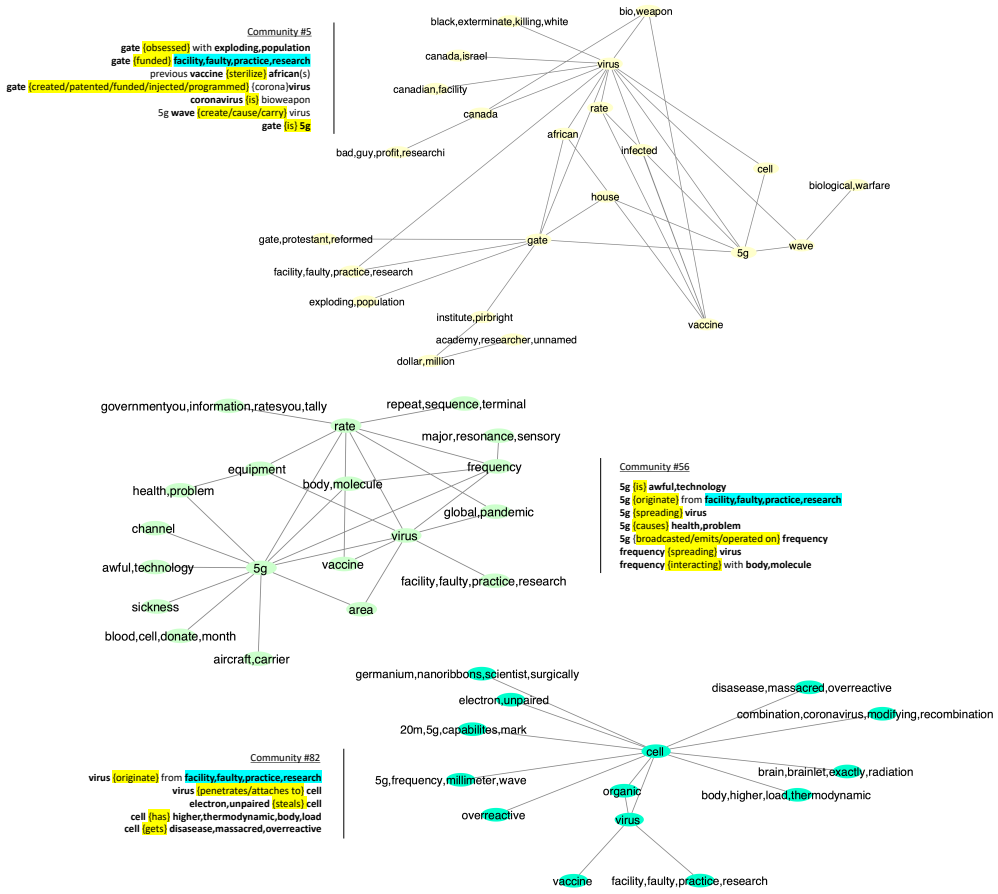

Figure 5.9: Communities with index 5, 56 and 82 sequentially describe the conspiracy theory surrounding "Bill Gates" and "5g". The words in **bold** are the sub-nodes present in the narrative network and the yellow-highlighted phrases are automatically extracted relationships between the sub-nodes. The blue-highlighted sub-node is a key actant that exists in all 3 communities and is one of the connecting components between "Bill Gates" and the conspiracy theory around "5g". Community 5 describes Gates's supposed *obsession* with population control along with his funding of faulty research. The same research is alleged to have created "5g" as a means of spreading the "virus" which is allegedly intended as a "bioweapon". Community 56 takes it a step further tying "5g" to its carrier frequency and the associated interactions of this frequency with the human body. Community 82 concludes the origin story of the virus (back to the "faulty" research conducted by "Gates") and mentions the cell-level interaction between the virus and the body.

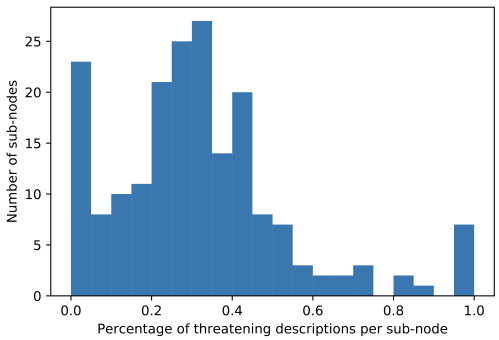

Figure 5.10: The histogram of threat scores across the sub-nodes from the phrase classifier. The bi-modality encourages binary classification thresholds around 0.2. In our networks, we use 0.25 which is at the 57^{th} percentile of sub-nodes classified as threats.

Sub-node	Score
china	0.50
chinese	0.53
china,government	0.57
ccp	0.53
5g	0.52
cia	0.64
shill	0.20
result	0.17
year	0.12

Table 5.3: Sample threat scores: Note the increasing threat score from the sub-nodes "china" to "chinese" to "chinese, government", which reflects the threat carried by more specific "china" contextualized actants

Table 5.3 provides a sample set of sub-nodes with their respective threat scores based on the majority vote. A sample sub-node "CCP" has 53% of its associated descriptive phrases classified as threats. The end-to-end classification pipeline, along with sample nearest neighbors during the phrase classification task, is shown in Figure 5.11.

5.3 Discussion

The lack of authoritative information about the Covid-19 pandemic has allowed people to provide numerous, varied explanations for its provenance, its pathology, and both medical and social responses to it. These conversations do not occur in isolation. They not only circulate on and across various social media platforms but also interact with news reporting on the pandemic as it unfolds. Similarly, journalists are keenly aware of the discussions

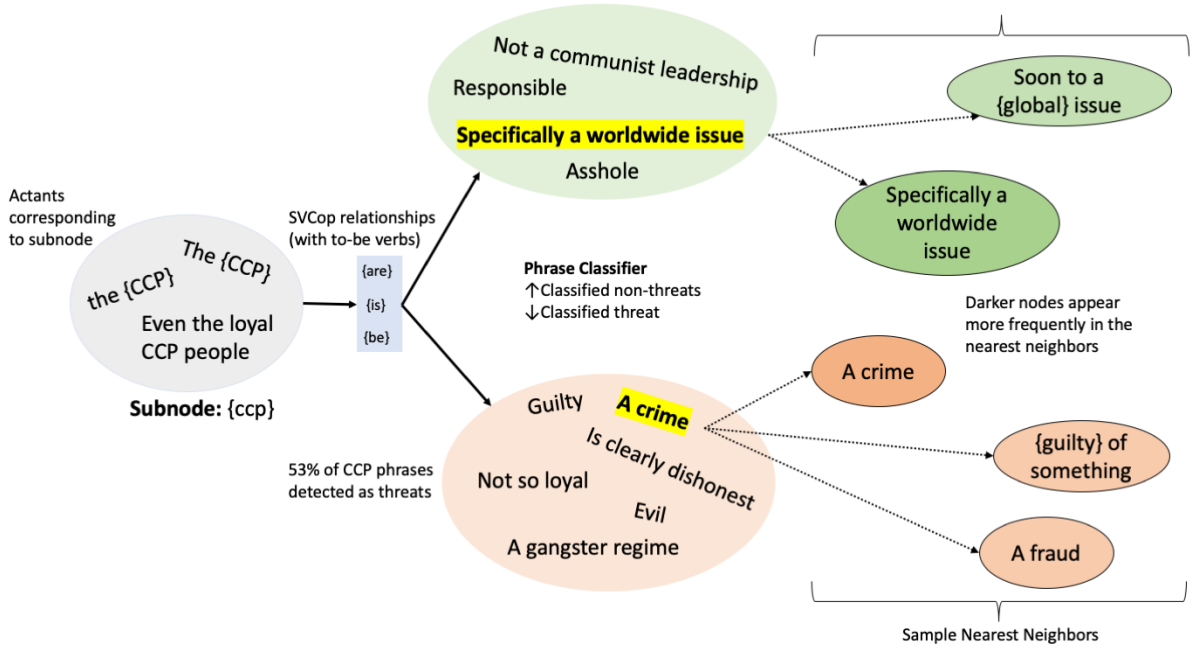

Figure 5.11: The sub-node "*CCP*" has associated noun phrases shown in the grey box. The noun phrases have descriptive SVCop relationships, whose descriptive phrases are sampled in the **light red** and **green** blobs. The phrases in the **red blob** are classified as *threats* by our majority classifier and the phrases in the **green blob** are classified as *non-threats*. The highlighted and bold descriptive phrases are sample phrases for which the nearest neighbors are shown. The kNN classifier reasonably clusters phrases that are syntactically different but semantically similar using the BERT embedding. Darker nearest neighbors occur more frequently.

Table 5.4: **A qualitative overview of key relationships that refer to "Bill Gates" in social media and the news reports.** These relationships describe the role that the "Bill Gates" node plays in connecting the Corona virus to conspiracy theories.

Date	News Reports	4Chan Threads
04/04	[Bill {Gates}] → [{predicted}] → [the {outbreak} and the China biolabs]	[{5g}] → [{causes}] → [{coronavirus}], [regular {people}] → [{go}] → [{untested}]
04/07	[conspiracy theorist David {Icke}] → [{added}] →[that Bill {Gates}, who is helping fund vaccine research, should be jailed]	[{Gates}] → [{saying}] → [[...]we all [...] accept his discount mark of the beast]
04/09	[Bill {Gates}] →[{invented}] →[{5G} to depopulate the world]	[the satanic {cabal}] → [to {leverage}] → [crisis into a forced vaccination /I D {program}]

occurring in social media, thereby creating a feedback loop between the two. The interlocking computational methods described above facilitate the discovery of a series of important features of the (i) narrative frameworks that bolster conspiracy theories and their constituent rumors circulating on and across social media, and (ii) the interaction between social media and the news.

5.3.1 Conspiracy theories in social media

The main communities and their interconnections in the aggregated social media corpus reveal the centrality of several significant conspiracy theory narrative frameworks. In particular, groupings of large communities form expansive frameworks and may well represent the dominant conspiracy theory frameworks in the corpus. In other cases, coherent narrative frameworks can be discovered within a single community. These communities may have some connections or overlap with communities describing the contours of the pandemic, as well as to other small communities that provide support for aspects of the narrative framework.

We find four large community groupings which present easy-to-interpret conspiracy theory frameworks. The first of these groupings is comprised of nodes from communities 5, 56, and 82 (see Figure 5.9). The narrative framework suggests that the Corona virus is closely linked to the 5G cellular network, and Bill Gates's associations with both faulty research and wide-

scale vaccination programs. Eager to expand a global vaccination program to help prevent the explosion of the world's population, Gates has contributed to the design of the Corona virus, which can be characterized as a bio-weapon. Potentially activated by 5G signals (a technology that is also the result of faulty research), the virus is intended to eradicate various populations throughout the world.

Certain key sub-nodes play key roles in connecting these communities to create the conspiracy theory narrative. For example, the sub-node "facility, faulty, practice, research" interacts with "Bill Gates" and his supposed obsession with exploding populations and vaccination efforts, the "virus"' origin story, and the emerging "5g" technology, thereby offering one potential route traversed by conspiracy theorists. This traversal aligns three distinct communities as the conspiracy theorists create a unifying theory. None of these key nodes are innocuous, but rather have all been classified as threats (See Figure 5.12).

A second group is comprised of nodes from communities 1, 40, and 65. In this narrative framework, the limited information about the virus released by the Chinese Communist Party is coupled to the virus's origin either in Chinese wet markets selling pangolins, presumably for human consumption, or labs studying bats (or potentially both). The narrative framework is informed by bigoted discussions of Chinese food practices coupled to an ongoing critique of the truthfulness of Chinese researchers. Several intriguing elements of the narrative framework are the "fluoroquinolone" sub-node, an antibiotic which is also a favored medication in other narrative frameworks, and the inclusion of a Bill Gates sub-node. Both of these suggest clear points of potential attachment with other conspiracy theory frameworks, such as the 5G one described above, and another one focused on information cover-ups and the virus-as-hoax (See Figure 5.13).

A third group, comprised of communities 0, 23, 24, 121 and 150, presents an expansive narrative framework. Here, the virus is presented as an engineered bioweapon, either deliberately or accidentally released from a lab. Confirmation of the engineered nature of the virus can be provided by scientists (pulmonologists) or members of the military (researcher, soldier). The subnodes in the graph set up a clear dichotomy between western governments and the Chinese government, and the controlling Chinese Communist Party (CCP), all of which are

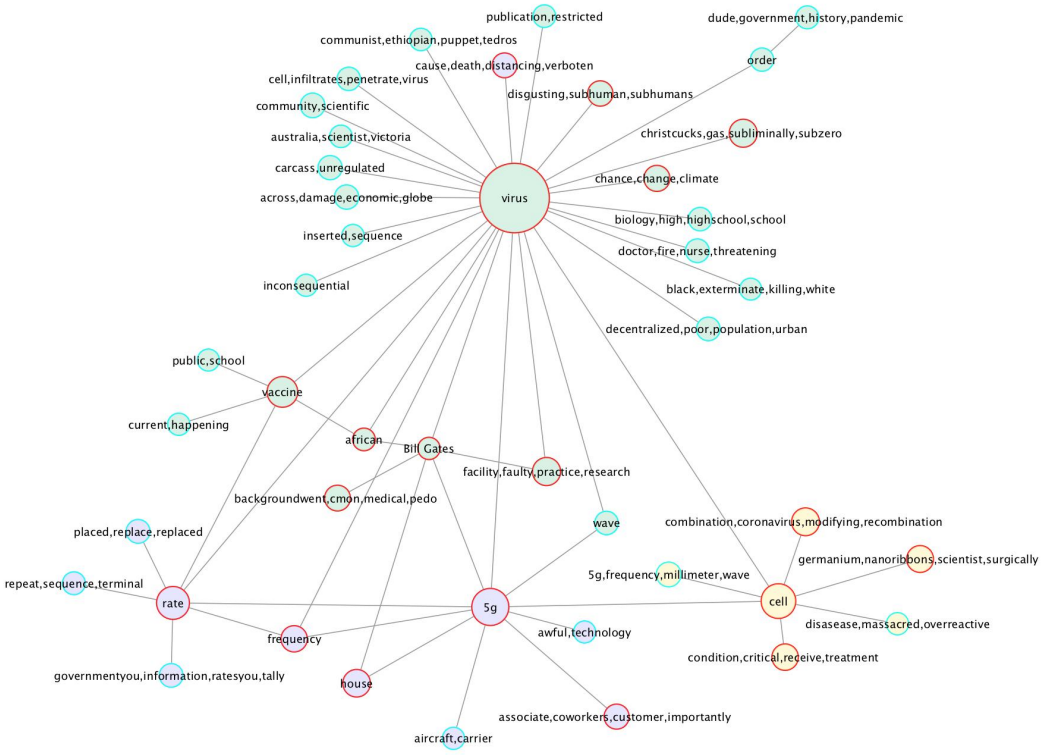

Figure 5.12: **A conspiracy theory narrative framework that links the virus to 5G, Bill Gates, and vaccination.** Nodes have been scaled by NER mentions; those with fewer than 250 mentions have been filtered for the sake of clarity. Nodes are colored by community, and outlined with red if they represent a threat.

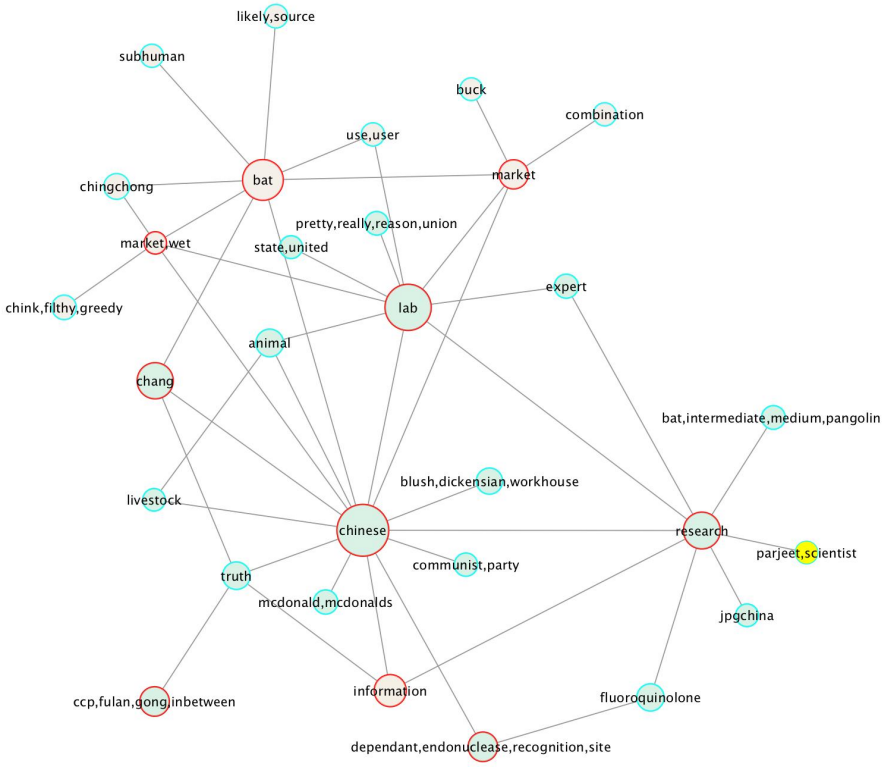

Figure 5.13: **Communities comprising the narrative framework suggesting that the virus is a result of Chinese wet markets and deliberate information cover-ups.** The narrative framework focuses heavily on markets, exotic animals such as pangolins, and the role of Chinese Communist Party in hiding information about the initial outbreak. Nodes are colored by community, and outlined with red if they represent a threat. The graph is filtered to show nodes with degree *geq* 2.

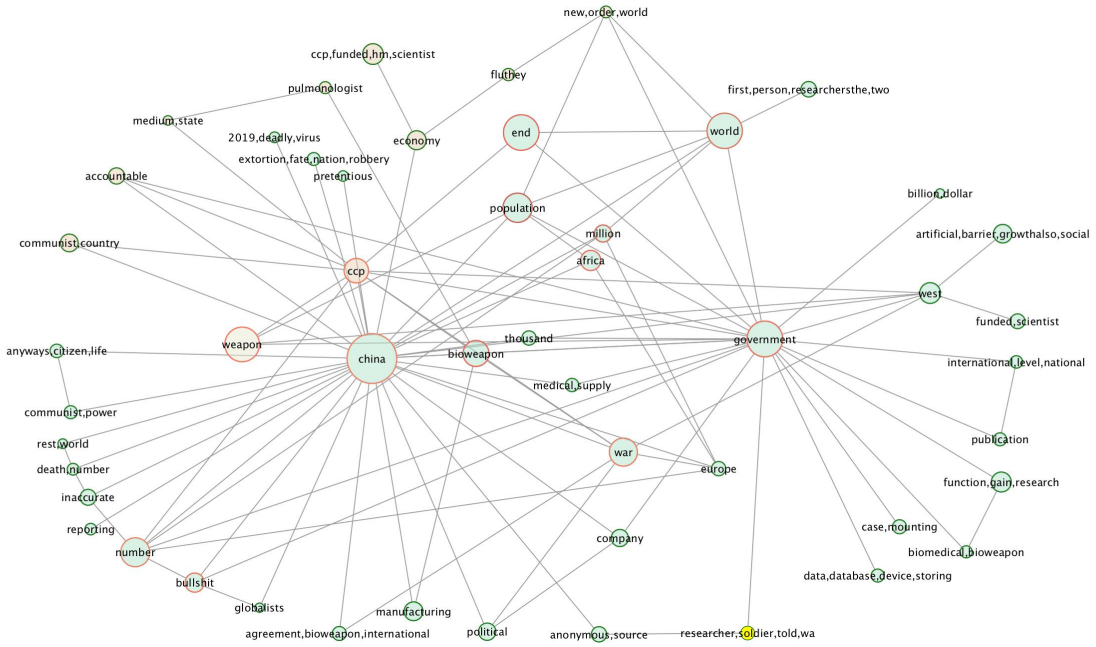

Figure 5.14: **Communities comprising the Covid-19-as-bioweapon narrative framework.** The narrative framework focuses heavily on laboratories and the potential role of the virus as a weapon. Nodes are colored by community, and outlined with red if they represent a threat. The graph is filtered to show nodes with degree *geq* 2.

classified as threats. It is worth noting that the CCP abbreviation is used by some social media contributors as a reference to the "Chinese Communist Plague", a racially derogatory term for the virus analogous to Trump's reference to the virus as the "Kung flu" [80]. Aspects of the framework also support discussions of the economic impact of the pandemic, as well as the role of "globalists" in promoting the danger of the virus through inaccurate reporting and inflated counts of victims across the world, including Europe (See Figure 5.14).

A fourth grouping comprised of communities 18, 21 and 75, constitutes a narrative framework proposing that the pandemic is a hoax on the same level as the global warming "hoax". This framework includes actants such as Trump, the American news commentator Sean Hannity, the right-wing podcaster Nick Fuentes and Republicans writ large who are fighting against globalists, Democrats, scientists such as Anthony Fauci and, in keeping with the long history of anti-Semitism in conspiracy theorizing, the "Jews", all of whom have conspired to perpetrate this hoax, which is wreaking havoc on the economy. The British conspiracy theorist,

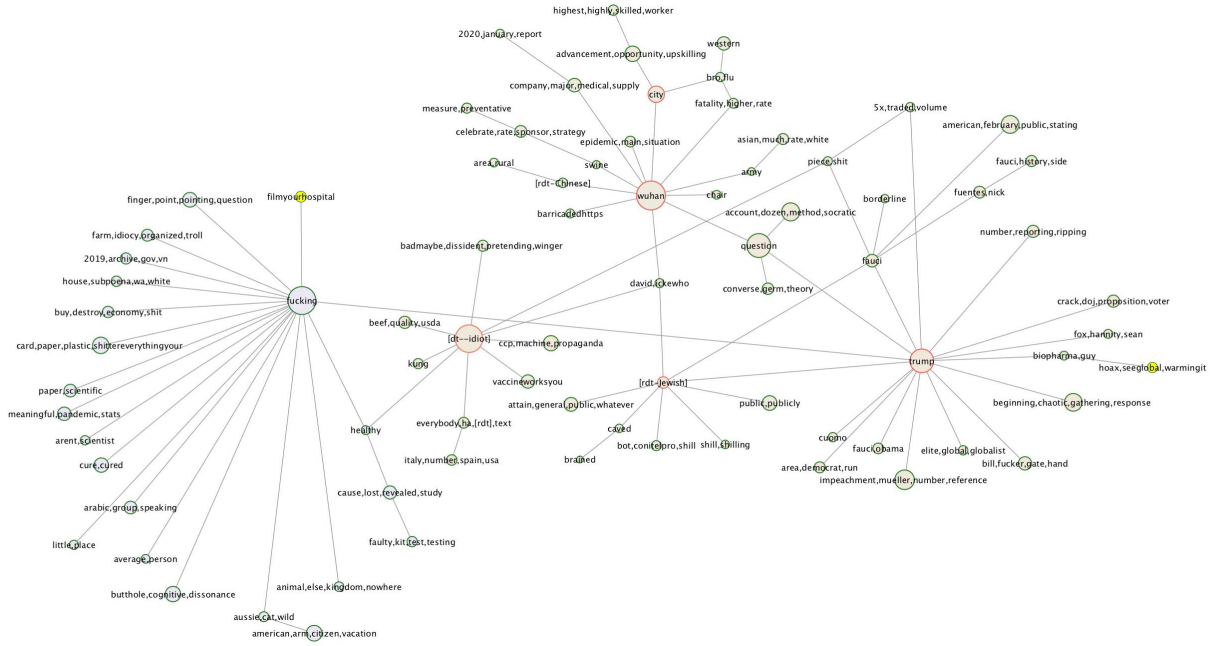

Figure 5.15: **The communities comprising the globalist hoax narrative framework:** Here, a globalist cabal has conspired to foist the hoax of the Corona virus on the world, with the virus presenting with mild flu-like symptoms. Trump and his allies are fighting against the Democrats and their surrogates to stave off the economic impact of the hoax. Nodes are colored by community, and outlined with red if they represent a threat. The graph is filtered to show nodes with degree *geq* 2. Two nodes, "filmyourhospital" and "hoax,see global warming" have been highlighted in yellow.

David Ickes appears with a direct link to a node representing the "Jewish globalists". Interestingly, albeit perhaps not surprisingly, Bill Gates appears once again in this framework, now more closely related to the Mueller inquiry and Democrats such as Obama. While the goal of the hoax is not made explicit, the framework bolster the erroneous suggestion that the virus presents with mild symptoms, and is no more dangerous than the flu (See Figure 5.15). The belief that the pandemic is a hoax inspires the "#filmyourhospital" movement as a means for publicizing the "discovery" that the virus poses no meaningful threat other than the economic threat of stay-at-home orders [81].

Several related narrative frameworks intersect with the main "hoax" framework in interesting ways. For example, a grouping of communities 6, 15 and 45, reveals a discussion of the

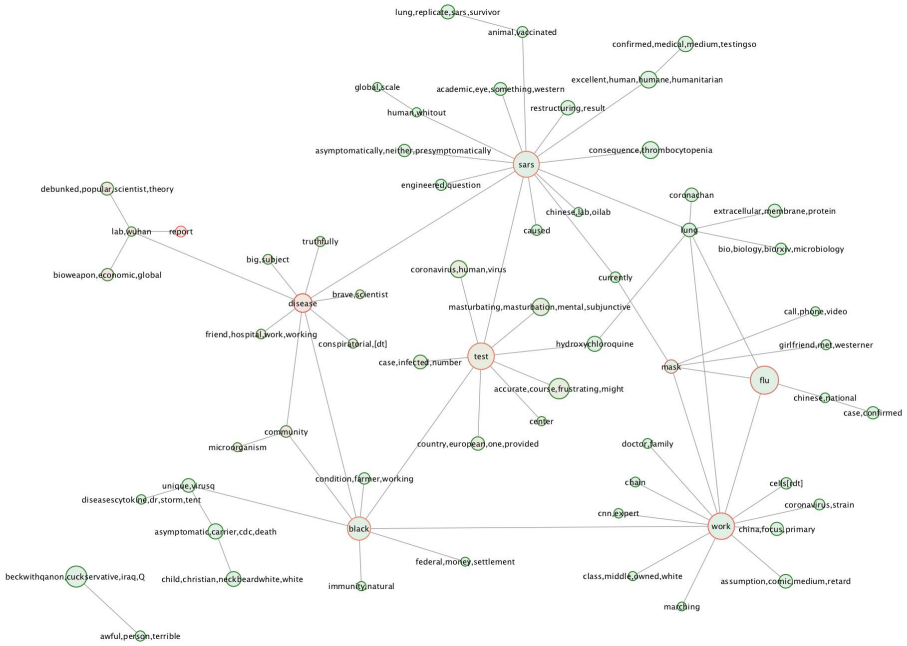

Figure 5.16: **A narrative framework that can be deployed by multiple groups.** The framework focuses on the relationship between the virus, SARS, the flu and the testing regimen. Also included are nodes representing research on the virus and questions of immunity. A small disconnected component on the filtered graph provides a critique of QAnon and Glenn Beck. Nodes are colored by community, and outlined with red if they represent a threat. The graph is filtered to show nodes with degree *geq* 2.

disease, its relation to SARS and the flu, the testing regimen, the accuracy of the tests and the efficacy of masks. It also includes an apparent critique of media figures who often endorse conspiracy theories (See Figure 5.16). While aspects of this narrative framework can be deployed as part of the more elaborated hoax framework—which seems to be the case particularly given the "threat" coding of "masks"—it can also be activated in the service of a counter-hoax narrative, given the inclusion of anti-conspiratorial content. In that regard, this particular grouping of communities captures the ongoing negotiation of the framework and the activation of parts of a framework as individuals come together (or move apart) to construct a totalizing narrative.

There are numerous other nucleations of narrative frameworks in the overall space that are worth noting. A particularly interesting community is 51 which has strong connections to

the well-known "Pizzagate" conspiracy theory, as well as connections to the much broader QAnon conspiracy theory [10]. The intrusion of QAnon, and the alignment of the pandemic with the broader narrative of a ring of pedophile human traffickers gained strong support in certain conversations associated with the pandemic-as-hoax frameworks. It also aligns well with the belief, noted above, that tents erected in Central Park were part of an operation to save children trafficked through underground tunnels, a key feature of the "Pizzagate" conspiracy theory [10].

These smaller frameworks suggest that there is a lively, ongoing negotiation of community beliefs about the pandemic. As the conversations progress, many of these smaller narratives are likely to become more closely connected with larger groupings, while others are likely to fade away. Community 42, for instance, describes the pandemic as a deliberate attack on the nation perpetrated by the Democrats; despite the impact on the global economy, the virus is no worse than a bad flu. Such a community could be easily subsumed in the broader virus-as-hoax narrative framework. Two additional examples of much smaller nucleations would be community 171 which consists of three sub-nodes: "cell phone", "ear" and "state surveillance", and community 123 with six semantically meaningful sub-nodes: "cancer", "cell phone", "cell tower", "microwave", "human cell", and "substance". One would expect, as discussions continue, that these communities would move closer to the 5G conspiracy theory narrative framework.

This tendency toward monological thinking already appears to be at work in the alignment of the 5G conspiracy theory with the biological weapons conspiracy theory, with both of those frameworks sharing close connections with the narrative framework describing the pandemic as a whole. Other alignments seem possible, with the 5G conspiracy and the hoax conspiracy potentially aligning through community 32, which in general focuses on Italy and quarantine measures across Europe. The inclusion of two peripheral sub-nodes, one labeled "5G, chemtrailz" and another one that labels the quarantines "ridiculous", not only provide an opportunity to challenge the meaningfulness of quarantine measures (thus providing a potential alignment with the hoax narrative), but also provide a connection between 5G and the longstanding "chemtrailz" conspiracy theory [82].

In earlier work on conspiracy theories, we discovered that conspiracy theorists, as part of their theorizing, tend to collaboratively negotiate a single explanatory narrative framework, often composed of a pastiche of smaller narratives, aligning otherwise unaligned domains of human interaction as they develop a totalizing narrative [24] [83] [10]. In many conspiracy theories, this coalescence of disparate stories into a single explanatory conspiracy theory relies on the conspiracy theorists' self-reported access to hidden, secret, or otherwise inaccessible information. They then use this information to generate "authoritative" links between disparate domains, engaging in what Goertzel has labeled "monological thinking" [83].

For the current pandemic, however, a single unifying corpus of special or secret knowledge does not yet exist—there are no "smoking guns" to which the conspiracy theorists can point, such as the Wikileaks emails on which Pizzagate conspiracy theorists relied [10]. Consequently the social media space is crowded by a series of conspiracy theories. In the various forums we considered, proponents of different narratives fight for attention, while also trying to align the disparate sets of actants and interactant relationships in a manner that allows for a single narrative framework to dominate and, by extension, to provide the "winning" theorists with the bragging rights of having uncovered "what is really going on." This type of jockeying for position is also reflected in the news.

5.3.2 Social Media and the News

Importantly, there is a lively interaction between the news media and the discussions about the pandemic on social media. Consequently, while the news media reports on the conspiracy theories that are evolving on social media, the social media groups point back toward reporting on the pandemic in the news media. The interaction is not, however, one of simple endorsement. Rather, the conversations on social media frequently contest, poke holes in, and otherwise challenge the narratives presented in the news media. In turn, the news media explores not only the content but also the veracity (or lack thereof) of the social media discussions. Unlike normal fact checking, however, the rejection in the news media of a particular social media position may be fuel for the conspiracy theorists, given their frequent

suspicion of the news media.

This interaction between social media and the news, modeled by the cross-correlation of relative coverage scores in Figures 5.3 and 5.4, indicates that the information flow between the two corpora is swift: the correlation-maximizing offset of days was 0 or nearly 0 for all considered actant groups. Since the data is smoothed over five days, this finding implies that the major actants appearing in narrative frameworks get aligned within days of appearing in either channel.

A qualitative example expanding upon this dynamic of knowledge synchronization between the news and social media is observed in Table 5.4 where "Bill Gates" was earlier highlighted as an important actant. News reports on April $4^{\texttt{th}}$ actively mentioned Gates's prediction of the Covid-19 outbreak. At the same time, 4Chan threads were embroiled in the discussion of "5g" causing the "Coronavirus". Perhaps the shock of such an accurate prediction—-and Bill Gates's continued investment in pandemic prevention and vaccine research—-helped motivate David Icke, an influential conspiracy theorist, to proclaim on April $7^{\texttt{th}}$ that "Bill Gates belongs in jail", echoing comments of a Florida pastor, Adam Fannin, who believes Gates is involved in a global effort to depopulate the world. In the ensuing days after Ickes's comments, 4Chan threads began denigrating "Gates", alleging him to be a part of a satanic cabal (thereby creating a direct link to "Pizzagate"), labeling him the anti-Christ, and accusing him of being an opportunist forcing the world into a crisis to further his *alleged* forced vaccination campaign. News reports, seemingly in response, summarized the conspiracy theories circulating on 4Chan communities with headlines such as, "The Dangerous Coronavirus Conspiracy Theories Targeting 5G Technology, Bill Gates, and a World of Fear" [84].

5.4 Conclusion

As the global Covid-19 pandemic continues to challenge societies across the globe, and as access to accurate information both about the virus itself and what lies in store for our communities continues to be limited, the generation of rumors and conspiracy theories will

continue unabated. Although news media have paid considerable attention to the well-known Q-Anon conspiracy theory (perhaps the most capacious of conspiracy theories of the Trump presidency), social media conversations have focused on four main conspiracy theories: (i) the virus as related to the 5G network, and Bill Gates's role in a global vaccination project aimed at limiting population growth; (ii) a cover-up perpetrated by the Chinese Communist Party after the virus leaped to human populations based largely on Chinese culinary practices; (iii) the release, either accidental or deliberate of the virus from, alternately, a Chinese laboratory or an unspecified military laboratory, and its role as a bio-weapon; and (iv) the perpetration of a hoax by a globalist cabal in which the virus is no more dangerous than a mild flu or the common cold. As the conversations evolve, these conspiracy theories appear to be connecting to one another, and may eventually form a single coherent conspiracy theory that encompasses all of these actants and their relationships. At the same time, smaller nucleations of emerging conspiracy theories can be seen in the overall social media narrative framework graph.

Because the news cycle appears to chase social media conversations, before feeding back into it, there is a pressing need for systems that can help monitor the emergence of conspiracy theories as well as rumors that might presage real-world action. We have already seen people damage 5G infrastructure, assault people of Asian heritage, deliberately violate public health directives, and ingest home remedies, all in reaction to the various rumors and conspiracy theories active in social media and the news. We have shown that a pipeline of interlocking computational methods, based on sound narrative theory, can provide a clear overview of the underlying generative frameworks for these narratives. Recognizing the structure of these narratives as they emerge on social media can assist not only in fact checking but also in averting potentially catastrophic actions. Deployed properly, these methods may also be able to help counteract various dangerously fictitious narratives from gaining a foothold in social media and the news. At the very least, our methods can help identify the emergence and connection of these complex, totalizing narratives that have, in the past, led to profoundly destructive actions.

5.5 Limitations

There are limitations with our approach, including those related to data collection, the estimation of the narrative frameworks, the labeling of threats, the validation of the extracted narrative graphs, and the use of the pipeline to support real time analytics.

Data derived from social media sources tends to be very noisy, with considerable amounts of spam, extraneous and off-topic conversations, as well as numerous links and images interspersed with meaningful textual data. Even with cleaning, a large number of text extractions are marred by spelling, grammatical and punctuation errors, and poor syntax. While these problems are largely addressed by our NLP modules, they produce less accurate entity and relationship extractions for the social media corpus than for the news corpus. Also, unlike news articles which tend to be well archived, social media posts, particularly on sites such as 4Chan, are unstable, with users frequently deleting or hiding posts. Consequently, re-crawling a site can lead to the creation of substantively different target data sets. To address this particular challenge, we provide all of our data as an OSF repository [79].

The lack of consistent time stamping across and within social media sites makes determining the dynamics of the narrative frameworks undergirding social media posts difficult. In contrast to the news data harvested from the GDELT project, the social media data is marked by a coarse, and potentially inaccurate, time frame due to inconsistent time stamps or no time stamps whatsoever. Comparing a crawl from one day to the next to determine change in the social media forums may help attenuate this problem. Given the potential for significant changes due to the deletion of earlier posts, or the move of entire conversations to different platforms, the effectiveness of this type of strategy is greatly reduced. Because of the limited availability of consistently time-stamped data, our current comparison between the social media conspiracy theory narrative frameworks, and those appearing in the news, is limited to a three week window.

There appears to be a fairly active interaction between the "Twittersphere" and other parts of the social media landscape, particularly Facebook. Many tweets, for instance, point to discussions on social media and, in particular, on Facebook. Yet, because of restrictions on

access to Facebook data for research purposes, we are unable to consider this phenomenon. Future work will incorporate tweets that link to rumors and other conspiracy theories in our target social media arena. As part of this integration, we also plan to include considerations of the trustworthiness of various Twitter nodes, and the amplification role that "bots" can play in the spread of these stories [85] [86].

As with a great deal of work on social media, there is no clear ground truth against which to evaluate or validate. This problem is particularly apparent in the context of folkloric genres such as rumor, legend and conspiracy theories, as there is no canonical version of any particular story. Indeed, since folklore is always a dynamically negotiated process, and predicated on the concept of variation, it is not clear what the ground truth of any of these narratives might be. To address this problem, we consider the narrative frameworks emerging from social media and compare them to those arising in the news media. The validation of our results confirms that our social media graphs are accurate when compared to those derived from news media.

Currently, our pipeline only works with English language materials. The modular nature of the pipeline, however, allows for the inclusion of language-specific NLP tools, for parsing of languages such as Italian or Korean, both areas hard hit by the pandemic, and likely to harbor their own rumors and conspiracy theories.

In addition, we believe that our semi-supervised approach to threat detection would require less human effort if we had more accurate semantic embeddings.

Finally, we must note that the social media threads, particularly those on 4Chan, are replete with derogatory terms and abhorrent language. While we have not deleted these terms from the corpus, we have, wherever possible, masked those terms in our tables and visualizations, with obvious swears replaced by asterisks, and derogatory terms replaced by "dt" for derogatory term, or "rdt" for racially derogatory term and a qualifier identifying the target group.

CHAPTER 6

Character and Relationship Extraction from Readers literary Book Reviews

6.1 Data

We use reader reviews of five works of fiction from the community forums on Goodreads: *Frankenstein* (1818); *Of Mice and Men* (1937); *The Hobbit* (1937); *Animal Farm* (1945); and *To Kill a Mockingbird* (1960) [87–90]. The works were chosen from the list of the most frequently rated books on the Goodreads site (number of ratings $> 500,000$). For highly rated novels, the number of reviews is also quite high, although significantly lower than the number of ratings. For example, *The Hobbit* has been rated over 2.5 million times, but has $44,831$ reviews (at the time of our data collection). For each of the novels, we downloaded the maximum allowed three thousand reviews given the Goodreads API limits on review requests.

The reviews were harvested using a crawler specifically designed for this project. Not all reviews were useful since numerous posts were either spam, posts on different topics, or written in languages other than English. Other reviews were either too short to include meaningful content, or so garbled as to be unintelligible. After filtering the reviews, we were left with a corpus of 8693 usable reviews: *Frankenstein* (2947), *The Hobbit* (2897), *Of Mice and Men* (2956), *Animal Farm* (2482) and *To Kill a Mockingbird* (2893). We discovered two types of phrases in the reviews: (i) Opinion phrases that reflected the readers' opinions about the book, the author, or the various characters and events. Relationships extracted from these phrases are the dominant ones when aggregated over all readers' posts, which is not surprising given that these posts are intended to be reviews. (ii) Plot phrases that

	# of posts	# of sentences
Frankenstein	2947	38432
The Hobbit	2897	37529
Of Mice and Men	2956	30205
Animal Farm	2482	27269
To Kill a Mockingbird	2893	33000

Table 6.1: Data description and size.

describe what happened to a subset of the actants, and how they interacted with each other. These phrases contain both the actants and their relationships, and are of primary interest to us.

Although our initial study corpus consisted of sixteen novels, we selected these five novels for detailed analysis on the basis of the broad disparity in their narrative structures, large variability in the number of characters, and a broad range of character relationships. For example, *The Hobbit* can be characterized as a multi-episodic, linear narrative that takes place across many different settings in an elaborate fantasy world, and includes a large cast of both human and non-human characters, instantiating an elaborate version of a standard hero's journey plot. *Of Mice and Men*, by way of contrast, is a short novella with a limited cast of characters that takes place in a highly localized, realistic setting, and represents a straightforward version of Vonnegut's "From bad to worse" plot. *Frankenstein*, although told partly in flashback, has a largely linear plot and a limited cast of characters, with a strong central figure and a relatively clear villain, although this is complicated by its use of nested narratives. Finally, *To Kill a Mockingbird* has an overlapping set of complex characters with multiple subplots.

For our ground truth narrative framework graphs, we relied on the online SparkNotes resource for each of the five chosen novels. SparkNotes is a corpus of freely available, professionally generated summaries of works of fiction, and provides us with a list of actants, as well as a chapter level plot summary. These fine-grained summaries allowed us to manually

create an actant-relationship narrative framework graph for each novel. These ground truth graphs were coded independently by two experts in literature, and a third expert was used to adjudicate any inter-annotator disagreements.

Reviewers who post to Goodreads have a variety of motivations for posting. The majority of reviewers use the site as part of a social network focused on reading, with the gender balance of active reviewers skewing slightly toward women [91]. There appear to be several categories of active reviewers on the Goodreads site, including students reviewing books as part of school assignments, members of book clubs, and people who aspire to become professional book reviewers. We make no discrimination as to classes of reviewers, but rather consider each review equally, as our goal is to understand the aggregate narrative model of a reviewed book. At the same time, we recognize that reviews of a book are often conditioned by the pre-existing reviews of that same book, including reviews such as those found in SparkNotes, Cliff Notes, and other similar resources. In certain cases, we recognize that these reviews may be influenced by the filmed adaptations of the target novels or professionally written summaries.

6.2 Results: Character Detection

We first examine the syntactic method of establishing actant-actant relationships for clustering. In Table 6.2, the Appos and SVCop relationships suggest not only limiting sentence-level associations, but also semantically invariant associations mentioned explicitly in the reviews. While this syntactic approach may work in many situations, book reviewers often *assume* a basic shared knowledge of the plot of a novel. This assumption dissuades reviewers from explicitly writing out the relationships between actants. In addition, book reviews are not very descriptive in general, focusing more on specific plot points or a character's trajectory. This tendency in book reviews further weakens direct Appos and SVCop actant-relationship extraction.

We applied our EMG algorithm to obtain the actants as documented in Table 6.5. Table 6.3 and Fig. 6.3 provide example statistics obtained during the execution of the EMG algorithm.

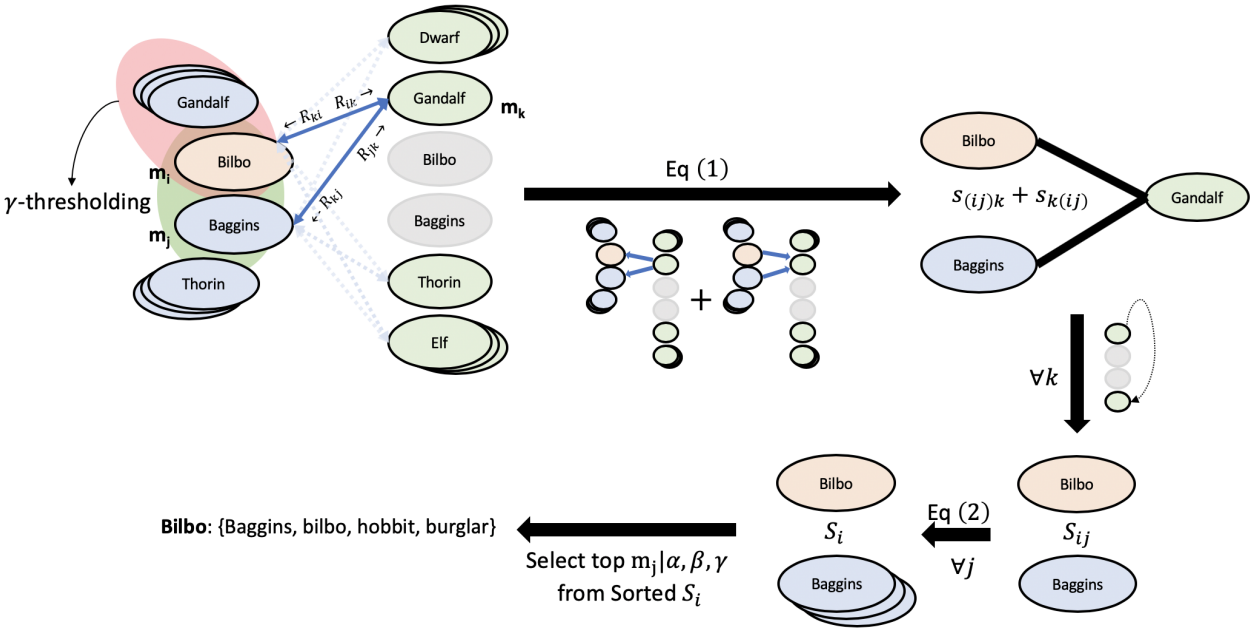

Figure 6.1: The pipeline of the EMG task shows the formation of the bipartite graph G with the computation of the Score Matrix S, along with hyperparameters α, β, γ

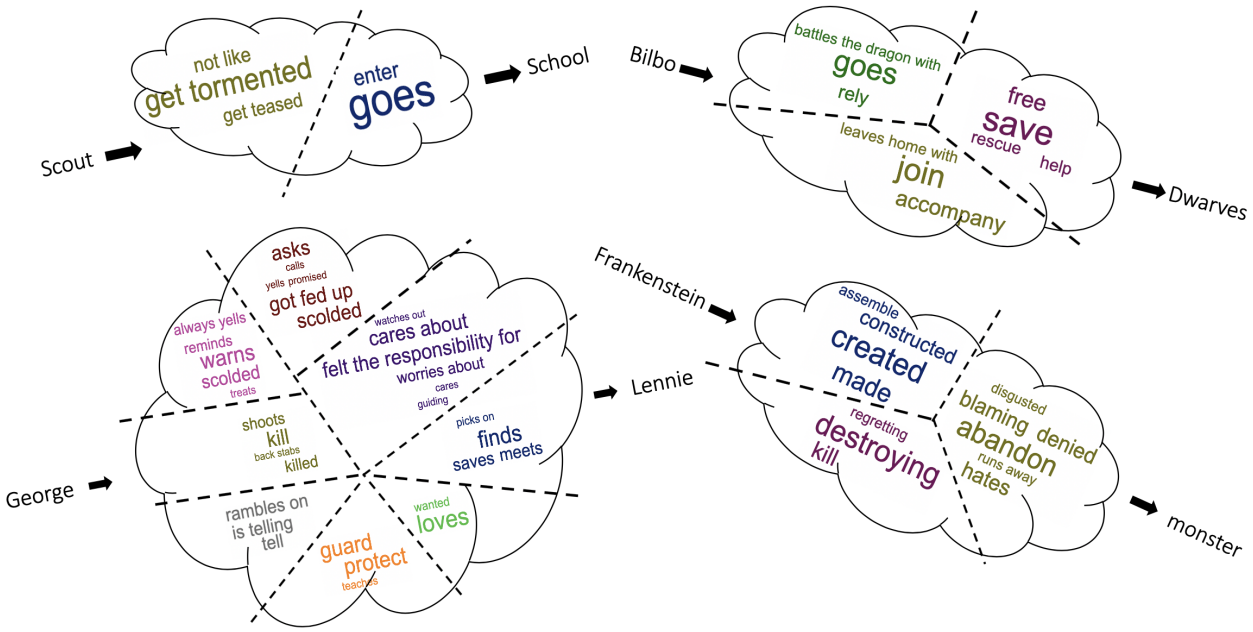

Figure 6.2: Directed and clustered relationships emergent after IARC between 2 actants per novel. In clockwise direction from top left: from Scout to School in *To Kill a Mockingbird*, from Bilbo to Dwarves in *The Hobbit*, from Frankenstein to Monster in *Frankenstein* and from George to Lennie in *Of Mice and Men*.

Entity	Descriptors
	The Hobbit
Bilbo	(a, the, simple, clean) hobbit, a burglar, baggins, hero, protagonist
Smaug	(a, the, horrible, vicious) dragon
Gandalf	(a, the, wise) wizard
	Frankenstein
Frankenstein	(a, the, fantasy) book, (the, a) creator, (a, the) doctor
Monster	(his, a, the) creation
	Of Mice and Men
George	a small (man,-, in height), Lennie's (caretaker, best friend, father figure, protector)
Lennie	(the, pitiful, unique, favorite) character, George's (foil, best friend)
	To Kill a Mockingbird
Jem	(big, the older, strong) brother
Atticus	(the, loving, ordinary, her) father
Scout	(a, hotheaded, young, an interesting) Tomboy

Table 6.2: Examples for Appos and SVcop candidate descriptors for entity mentions across the four novels.

Entity Mention (m_i)	**Ranked Similarity Scores for other Mentions (m_j) (S_{ij}'s, see Eq. 3.7)**
Bilbo	baggins,42.14 hobbit,14.47 burglar,3.80
Burglar	bilbo,3.80 dwarves,2.79
Wizard	gandalf,22.49 gandolf,7.00 grey,5.34 thorin,3.32
Hobbit	bilbo,14.47 baggins,6.06

Table 6.3: Given two entity mentions (m_i, m_j), the similarity score S_{ij} (see Eq. 3.7) measures the semantic "fungibility" of the mentions (i.e., whether both mentions are used interchangeably to refer to the same actant). The table shows several popular entity mentions (m_i's) and the similarity scores of other candidate mentions, m_j's, in *The Hobbit*. Clearly, the mentions [Bilbo, baggins, Hobbit, Burglar] form a clique representing the same actant, *Bilbo Baggins*. One can also see the emergence of another EMG [Wizard, Gandalf, Gandolf, Grey] for the actant *The wizard*.

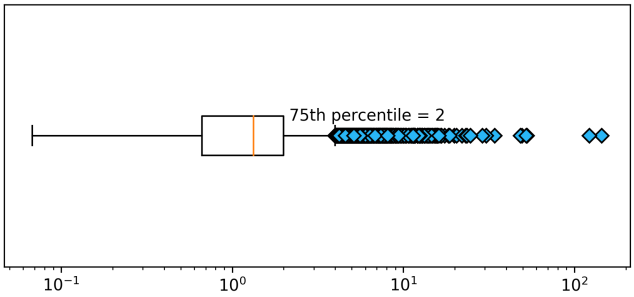

Figure 6.3: A Box plot of the similarity scores, S_{ij}'s (see Eq. 3.7), for all entity mention pairs (m_i, m_j) in *The Hobbit*. For any entity mention, m_i, its Entity Mention group (EMG) is first pruned to contain m_j's with scores, $S_{ij} \geq \alpha$, where α is the 75^{th} percentile of the score distribution. From the plot we find $\alpha = 2$. This EMG is further pruned by first sorting the list by their scores, and then ensuring that the ratio of any two successive scores is bounded below, i.e., $\frac{S_{i(j-1)}}{S_{ij}} \geq \beta$ (for $j \geq 2$). We found that $\beta = 2$ provided a good cutoff.

Each actant, once formed, aggregates relationships that the individual entity mentions imply. The clustering of relationships aggregated under the now-formed entity mention groups yield higher granularity and confidence in the IARC task, as semantic connections between entity mentions reinforce the relationships from one actant to another. This effect is observed across the four books as shown in Fig. 6.2. The relative size of words in the figure correlate to their frequency in the aggregated relationships between the entity mention groups.

The task of mapping relationship clusters to particular ground truth labels is shown for the "converse" and "warn" clusters from George to Lennie in *Of Mice and Men* (Figure 6.4). The rich clusters, in comparison to the ground truth labels from SparkNotes suggests recall as a good measure of performance for our pipeline. A summary of our results for all four books including recall is presented in Table 6.4.

In general, the relationships between actants reveal a high degree of consistency with the ground truth graph. The largest divergences consist of missed relationships rather than the identification of non-existent relationships, although these occur occasionally. This latter group of relationships is often the attribution of a relationship, such as the killing of Smaug (the dragon in *The Hobbit*), to an important character such as Bilbo Baggins. In other words, many readers *incorrectly believe* that Bilbo killed Smaug. Another small set of spurious relationships, including one that suggests that Jem killed Bob Ewell in *To Kill a Mockingbird*, are caused by reader confusion, "what-if" scenarios or, more commonly, incorrect pronoun resolution and aggregation. Apart from the relatively infrequent misattribution of relationships, the reduction in relationships aligns with the corresponding reduction in the number of actants connected to the central component of the story graph.

Figure 6.5 depicts the narrative framework graph for *The Hobbit* with blue nodes representing ground truth actants or meta-actants. We also show four examples of resolved actants or meta-actants (colored green) not found in the ground truth: **Tolkien**:[tolkein, author], **novel**:[book, fantasy, story, novel], **Fili**:[fili] and **Film**:[film, movie, scene]. Blue edges represent relationships in the ground truth found by using our methods (frequency threshold ≥ 5), while red edges represent undetected ground truth relationships. Green edges connecting to green nodes (frequency threshold ≥ 10) are edges that cannot be verified; we include

Figure 6.4: Evaluation phase: matching 2 clusters of relationships in *Of Mice and Men*, from George to Lennie, to ground truth labels, in accordance to Algorithm 2. β_c determines the set of edges.

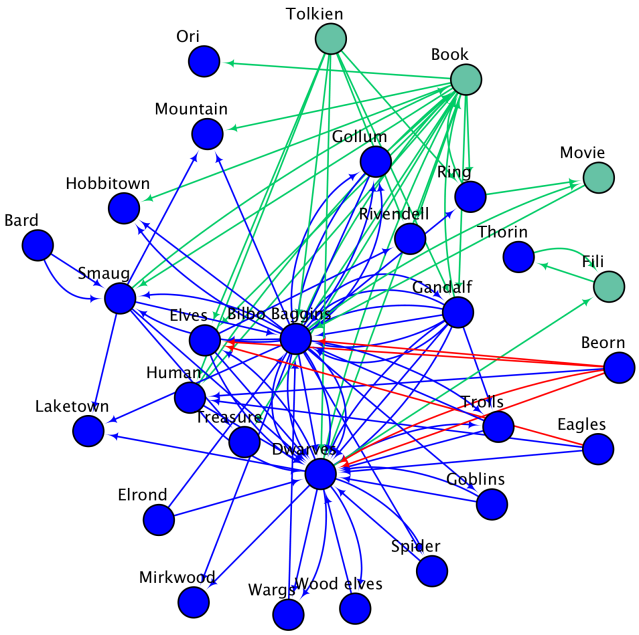

Figure 6.5: Narrative Framework graph of *The Hobbit*. Green nodes are extracted entities not part of the ground truth, red edges are ground truth edges which were not detected by the algorithm, blue edges are detected ground truth edges.

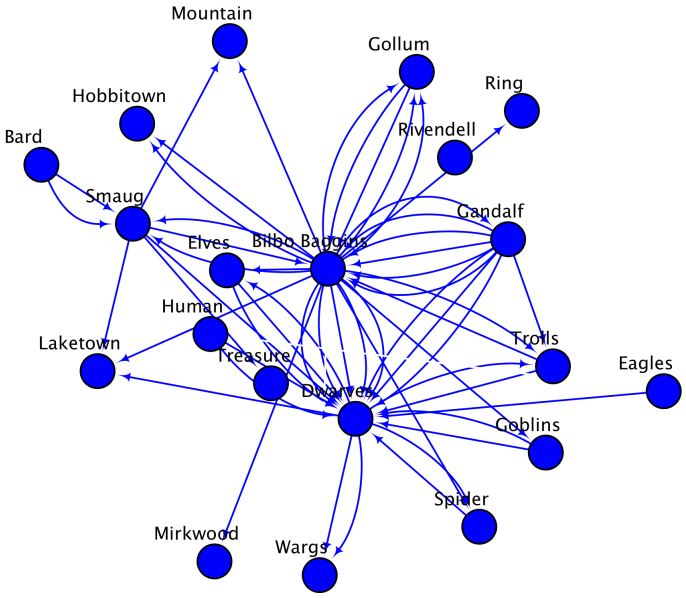

Figure 6.6: Narrative Framework graph of *The Hobbit after* thresholding on the frequency of relationship. Blue edges have at least 5 relationship instances.

them to indicate the richness of the extracted graph as opposed to the ground truth. Figure 6.6 shows a graph similar to Figure 6.5 after the deletion of low frequency edges (≤ 5), and represents the core structure of the narrative covered in the reviews conditioned on the SparkNotes ground truth.

There are shared structural properties (disregarding the specific relationships they encode) that can be used to automatically distinguish between actual characters in the novels and the various meta-actants. For example, the meta-actant **Tolkien** (the green node at the top center of Figure 6.5) has only outgoing edges, indicating that Tolkien appears only as the subject in any inferred relationship triplet. This lack of incoming edges is a significant feature of meta-actants: An important character in a novel usually has bi-directional relationships with other characters. An author of the novel, on the other hand, usually "acts" on the characters; hence the corresponding node is directionally isolated. The incoming edges for the meta-actant "Book" are all attributable to phrases such as " character XNZ is portrayed *in the* book/novel". A simple filtering of these preposition-induced relationships directionally isolates the meta-actant "Book." Further structural explorations of the derived networks,

	Of Mice and Men	**The Hobbit**	**Frankenstein**	**To Kill a Mockingbird**
Recall (%)	**88.33** (83.33)	**82.61** (59.42)	**69.04** (66.66)	**90.16** (68.85)
Edge detection rate (%)	**98.33** (96.66)	**92.75** (69.56)	**73.80** (73.80)	**93.44** (77.04)
Average Number of Relationships	**246.55** (209.15)	**139.34** (14.03)	**20.33** (13.38)	**72.09** (27.34)
Median Number of Relationships	**54** (48)	**43** (3)	**7** (7)	**36** (6)

Table 6.4: Performance on character relationship extraction with IARC after (in bold) and before (within parentheses) EMG. In the "before", scenario an actant group consisted of only the mention used in the ground truth. Thus for actant "Bilbo" only the mention "Bilbo" was used to compute its relationship. Post EMG, the mentions in the group **Bilbo**:[bilbo, baggins,burglar,hobbit] were aggregated to compute the actant Bilbo's relationships.

such as measures of centrality and importance of different characters, are part of our ongoing work.

6.3 Story Network Creation and Expansion

The resulting story network graphs for the five works are presented in Figures 6.11, 6.7, 6.10, 6.9 and 6.8, with a clear visual distinction between the diagetic nodes (i.e. the novel's characters) and the metadiscursive or extra-diegetic nodes (i.e. actants not in the novel *per se*).

These expanded graphs reveal interesting features not only about readers' perceptions of the stories, but also of how readers conceptualize authorship as well as other external features relevant to an understanding of the novel. For example, the authors of *Of Mice and Men* and *The Hobbit* are directly linked to main characters in those novels, whereas for the other novels, the authors are only connected to the main story graph through intermediary nodes. The close connection of both Tolkien and Steinbeck to the main story graphs possibly highlights the readers' perceptions of the author as equally important to any discussion of the novel for these two works. For *Frankenstein*, by way of contrast, the expanded graph captures the complex discussions of "authoriality" that pervade both the narrative and meta-narrative

space. For the other two novels in our corpus, the author appears to play a slightly more divorced role, at least in the reader discussions. While our data collection occurred prior to the release of Harper Lee's *Go Set a Watchman* [92], which may well have triggered greater awareness of Lee as an author, the reviews for *To Kill a Mockingbird* and *Animal Farm* may be capturing a reduced awareness of the authorships of Lee and Orwell, as opposed to the broadly recognized authorships of Tolkien and Steinbeck.

The extended networks include a proliferation of generic terms such as "people" and "story" possibly capturing readers' awareness of other readers – the generic "people" – and the narrativity of the work itself. Other terms such as "place" and "home" may reflect readers' efforts to tie the novel into their own understanding of the world and, possibly, considerations of locality and the domestic. The "ways" node appearing in most of the expanded story graphs accounts for the strategies that characters pursue in the target novel. For example, in *To Kill a Mockingbird*, "Scout" is connected to the "ways" node with relationships such as "looked" and "thought", reflecting readers' awareness of Scout's efforts to both comprehend and solve the fundamental challenges she encounters in the novel.

These metadiscursive nodes share an additional interesting feature: the edges connecting them to the character nodes in the main story graph are mostly in a single direction (either toward the story graph or away from it) as in, for example, *Of Mice and Men* (See fig. 6.7). By way of contrast, main story characters interact with each other generating a mixture of in- and out-edges. The *Animal Farm* graph presents an intriguing example of this directionality, with the "Orwell" node having only outwardly directed edges, and the remaining additional metadiscursive nodes only having inwardly directed edges. There are two exceptions in this graph to this general rule: the nodes "revolution" and "rebellion" share an outwardly directed edge connecting them to the diegetic node of the "farmhouse". Readers here collectively recognize not only the strategy of "revolution" that animates the novel, but also the focus of the uprising on the locus of institutional authority, here the "farmhouse".

The expanded novel graph for *Frankenstein* reveals a large number of metadiscursive nodes that are not directly connected to the main story component, creating a secondary network of high-degree metadiscursive and extra-diegetic nodes, thereby capturing a broad reader

conversation not centered on the novel itself. To highlight this aspect of the reader conversations, we extended the additional nodes by finding the inter-connections between a pair of candidate mentions. This secondary network reveals a lively conversation not only about the story plot(s) but also about meta-narrative considerations such as the composition of the novel, its epistolary frame narrative, Mary Shelley's authorship, and philosophical speculation about "God".

Given the implications of the expanded narrative graphs, we would be remiss to dismiss *Goodreads* reviews as amateurish plot-focused summaries, since they capture more sophisticated speculations of a broad readership, reflecting a latent diversity of opinion that may well be an echo of Fish's communities of interpretation. Indeed, the methods presented here capture the voices of emerging literary critics, and their engagement with the works of fictions and the other readers as they negotiate the boundaries of their interpretive communities. Consequently, we can see these graphs as capturing an emerging discussion of the complexity of a work of fiction, the relationship between authors and their works, the constitutive role that the acts of reading and reviewing play on the works in question, and the interpretive range of reader engagements that extends well beyond straight forward plot summary.

6.4 SENT2IMP: Character Impression

6.4.0.1 Single Character Impression

Our unsupervised method of character impression discovery provides insightful clusters about each character. A subset of these clusters for "Bilbo" in *The Hobbit* is described in Table 6.6. The first cluster provides a convincing argument that this character is *unpleasant*. The second one, in contrast, describes him as a hobbit of *impeccable personality*. These contradictory representations may capture a dichotomy of readers' impressions of Bilbo. The third and fourth clusters, however, are comparatively different, revealing disparate information about the character not related to sentiment at all, characterizing *Bilbo* as both a burglar and a hobbit, both of which are true. Such findings justify our assumption that SVcop relationships

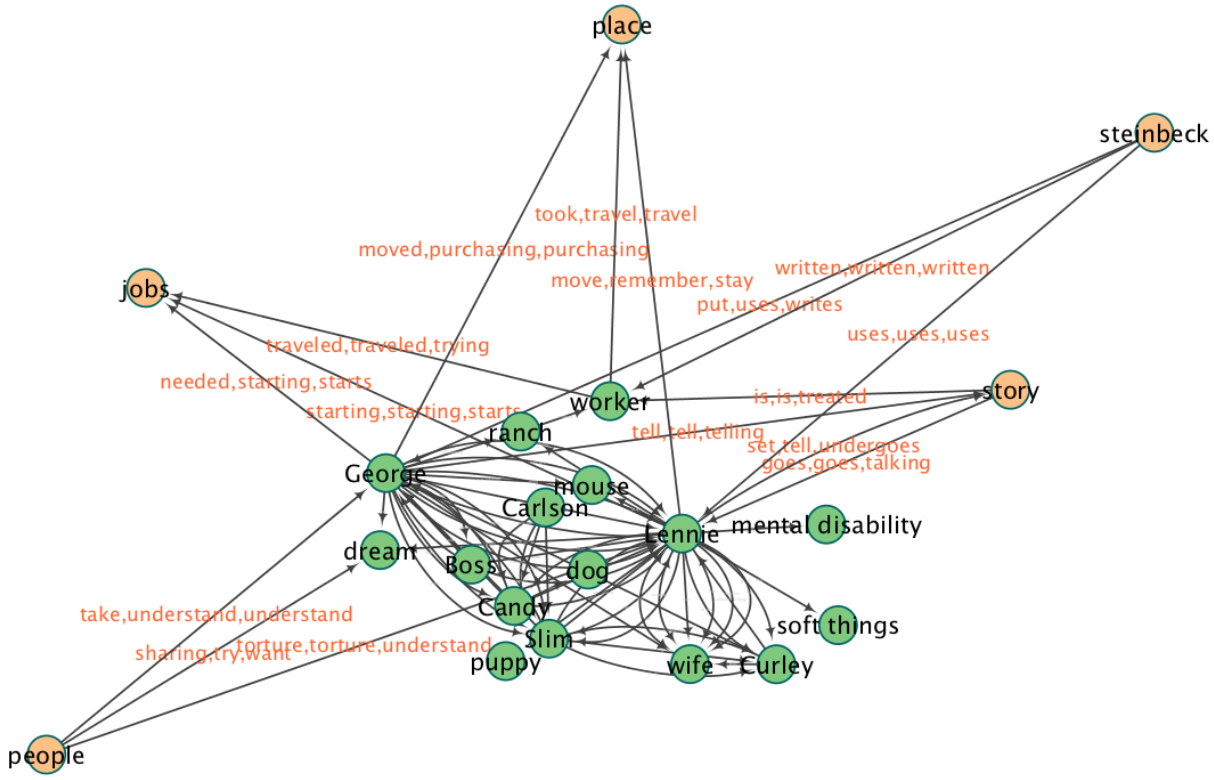

Figure 6.7: *Expanded Story Network Graph for "Of Mice and Men"*: Nodes that represent characters in the story are in green while the actants extending the original character story network are in orange. The node "steinbeck" has an in-degree of 0 suggesting readers' understanding of the author's impact on creating complex story actors, while the actants have no meaningful return engagement. Similarly, the "place" node cannot directly effect causal change in the story and as a result is very rarely found in the *subject* part of a relationship (the out-degree is 0).

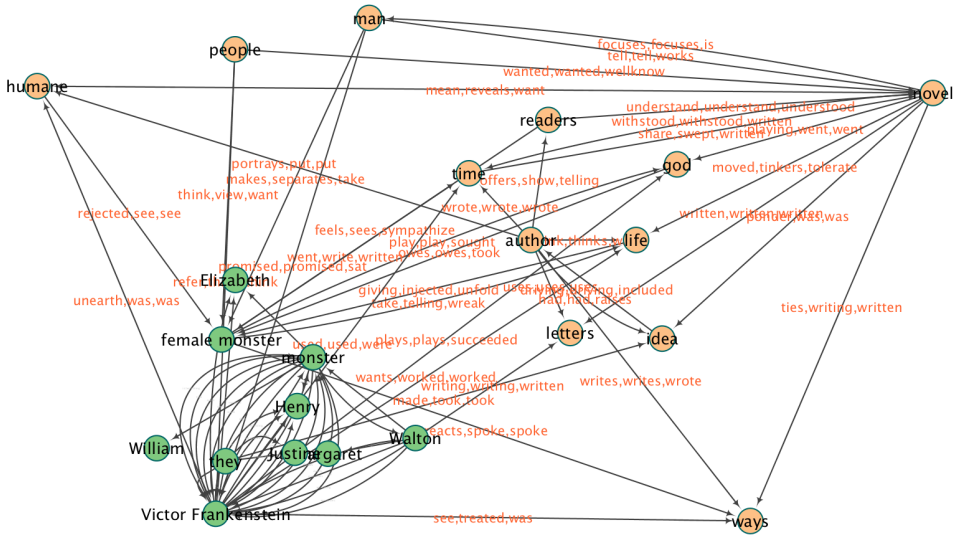

Figure 6.8: *Expanded Story Network Graph for "Frankenstein"*: Nodes that represent characters in the story are in green while the actants extending the original character story network are in orange. The subnetwork of "letters", "author" and "novel" indicate that readers recognize the epistolary nature of *Frankenstein*. The common node "people" (which is found in most of the graphs) represents the reviewers' perception of other reviewers.

are worth consideration, as they not only capture the readers' broad range of perspectives on a character but also because these phrases form rich clusters of semantically aligned meanings that are not captured by standard supervised sentiment detection methods.

In addition to extracting rich clusters of actant-conditioned impressions, the HDBSCAN algorithm (with the default and constant distance threshold) clusters all *noisy* impressions into a separate cluster labeled "-1". In this algorithm, we used a relatively high $eps = 2$ parameter to decrease the extreme sensitivity to noise. For example, for "Bilbo", phrases such as ['the uncle of Frodo', 'unbelievably lucky', 'nostalgic'] are classified in the noise cluster. More examples can be found in the last row of Table 6.6. Our results also show that there is a correlation between the perceived popularity of a character and the complexity of the impressions he or she elicits. These clusters of impressions can be informatively visualized with a dendrogram-heatmap that sorts similar clusters with respect to correlation scores (see Methodology for how this score is computed). In order to find a label for a cluster, we pick the most frequent word in the cluster's phrase list excluding stop words.

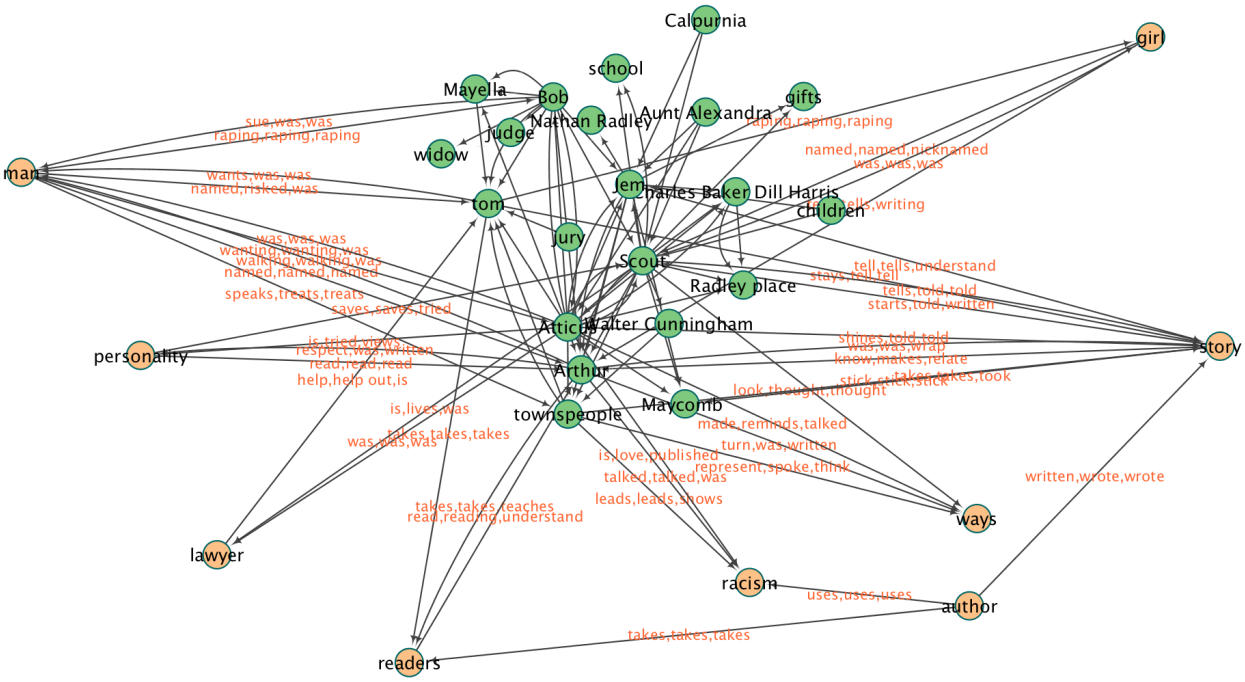

Figure 6.9: *Expanded Story Network Graph for "To Kill a Mockingbird"*: Nodes that represent characters in the story are in green while the actants extending the original character story network are in orange. Important and intangible actants such as "racism", "lawyer", "personality" compose the extended story network nodes in this graph. The "personality" node reflects the novel's dedication to character development, be it of "scout", "atticus" or even "arthur".

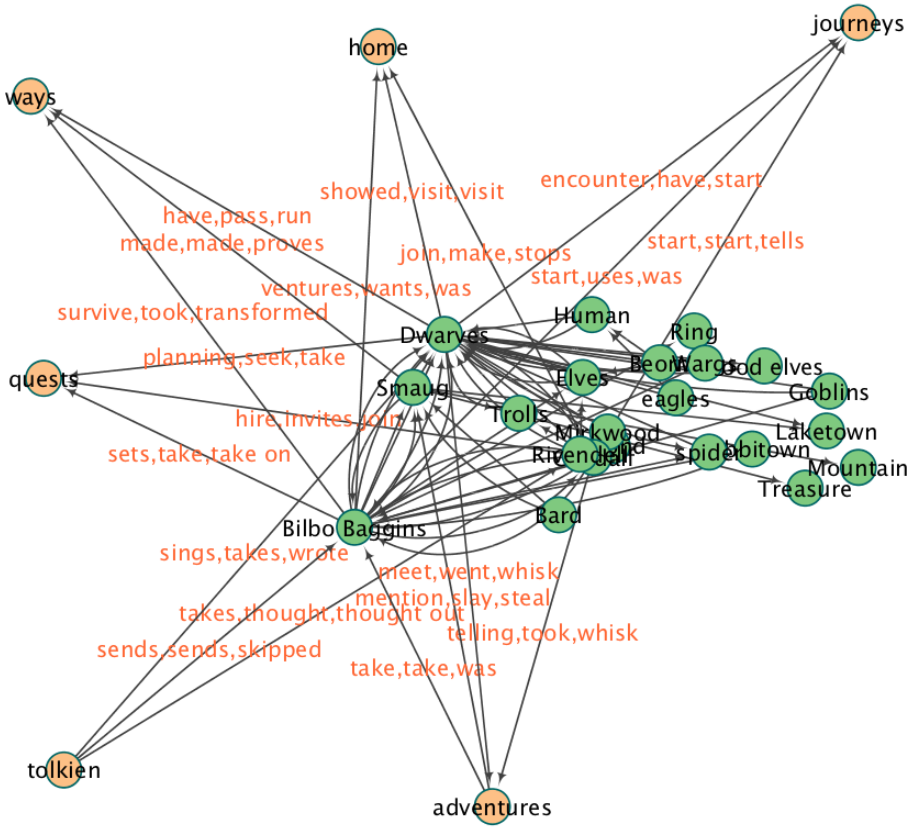

Figure 6.10: *Expanded Story Network Graph for "The Hobbit"*: Nodes that represent characters in the story are in green while the actants extending the original character story network are in orange. The readers' classification the novel's genre is immediately apparent in the nodes "adventures", "quest", "home" and "journeys". Inanimate actants such as "home", "journey", "quest" and "ways" typically have a very low out-degree (in this case 0) whereas "tolkien" has a very low in-degree. The node "ways" signals strategy: "Dwarves" *have* "ways" or "Bilbo Baggins" *took* "ways".

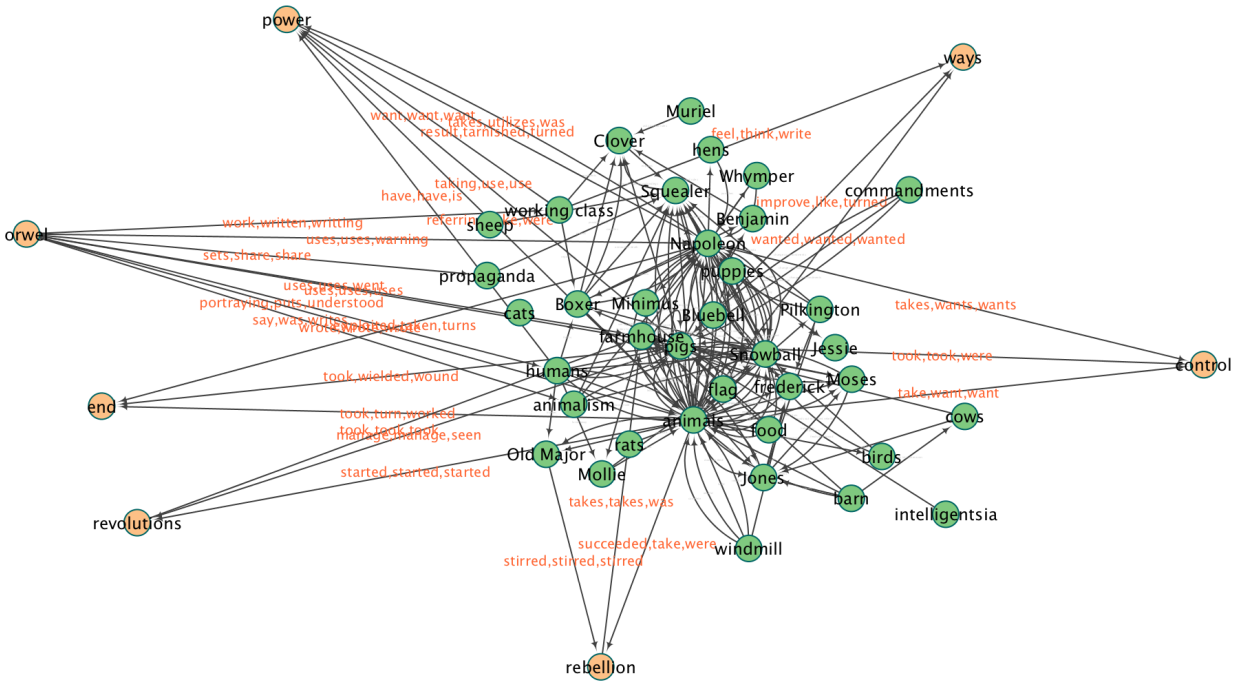

Figure 6.11: *Expanded Story Network Graph for "Animal Farm"*: Nodes that represent characters in the story are in green while the actants extending the original character story network are in orange. The nodes "rebellion" and "revolution", in conjunction with the nodes "power" and "control" highlight the sustained themes of power struggle, social dynamics and politics that lay at the ideological root of the novel. The author "orwell" once again has a high out-degree and the node "ways" once again signals strategy: "hens" *think* of "ways" and "pigs" *wanted* "ways".

A sample heatmap for “Victor Frankenstein” is shown in Figure 6.12. In this figure, there are three groups of impression clusters: 1) With labels such as, ”brilliant scientist”, ”scientist of story”, ”responsible” and ”instinctively good”; 2) A group that includes ”young student”, ”name of creator”, ”name of man”; and finally, 3) another group comprising, ”horrible person”, ”selfish brat”, ”real monster” and ”mad scientist”. On closer inspection, a cluster labeled ”Victor Frankenstein” stands out as a separate group with almost no correlation to other groups.

This representation also provides insight into the performance and limitations of BERT embeddings. Clusters that are similar should have a high similarity score (red) and clusters that are dissimilar should have a low similarity score (blue). For example, the clusters labeled “selfish,pitiful” and “horrible,person” have a high similarity score and the clusters labeled “scientist,mad” and “responsible” have a low similarity score. All the clusters are most similar to themselves so the major diagonal is deep red. There is a high similarity score between the clusters labeled “sympathetic,character” and “selfish,pitiful”, due to the similarity between representative phrase {“a sympathetic character”} in the first cluster and the phrases {“miserable”,“sad”,“pitiful”} in the second cluster: after all, Frankenstein was a “horrible person” for having created the monster but was also a person deserving of “sympathy” and “pity” for all the loss and grief that creation caused him.

Character	Descriptors
Bilbo	**The Hobbit**
Cluster 1	['not the interesting character', 'timid not', 'not enthusiastic', 'reluctant', 'not the type of hero', 'less cute', 'not as cool', 'unsure of situation', 'a small unadventurous creature', 'Perhaps just not the kind of character', 'not as important', 'less cute']
Cluster 2	['a true personality', 'an exemplary character', 'such a great character', 'resourceful', 'likable', 'still loveable', 'quite content', 'such a strong character', 'an amazing character', 'respectable', 'a great protagonist too', 'clever', 'such an amazing character', 'a peaceful', 'such an endearing character', 'a great choice', 'a fantastic lead character', 'quite engaging', 'cute', 'much charismatic character', 'such a fantastic Character', 'truly beautiful', 'enjoyable', 'just so charming', 'personable', 'able', 'the best character', 'quite skilled gets', 'awesome', 'smart']
Cluster 3	['of course the burglar', 'a thief', 'a thief go', 'to a burglar', 'to a thief', 'to a thief', 'the burglar', 'their designated burglar', 'could a burglar', 'of course the burglar', 'a Burglar', 'a Burglar']
Cluster 4	['a respectable hobbit', 'a respectable Hobbit', 'a sensible Hobbit', 'a clean well mannered hobbit', 'a respectable Hobbit', 'a sensible Hobbit', 'a proper hobbit']
Cluster 5	['small', 'small', 'little', 'small', 'little']
Cluster -1	['rich','the right man','a feisty character','the uncle of Frodo','unbelievably lucky','the perfect example of success','nostalgic','middle aged']

Table 6.6: **Example impression clusters for "Bilbo" in *The Hobbit***: Clusters 1 and 2 describe impressions of "Bilbo"'s character while clusters 3 and 4 describe his profession and community. Cluster marked -1 is noise. Labels for each cluster are aggregated based on the most frequent monograms per cluster.

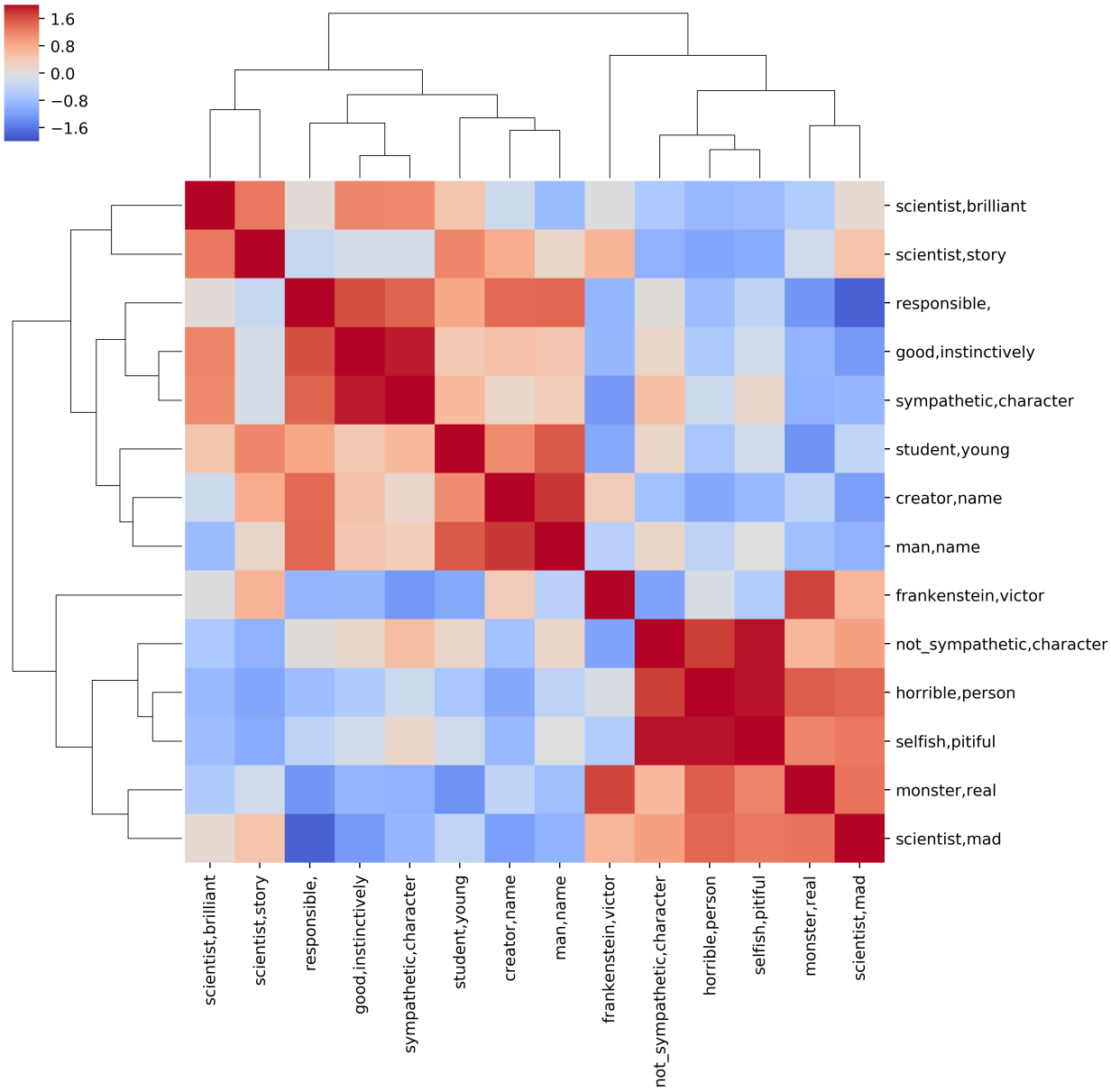

Figure 6.12: **The (symmetric) heatmap for the character "Victor Frankenstein"**: The similarity scores between clusters of impressions labelled by the row/column headers are computed by Algorithm 8. The sub-matrices that are deep red or blue imply a hierarchical structure to the mutual similarity or dissimilarity between groups of impression clusters. The diagonal entries are +2 as a cluster of impressions is most similar to itself.

6.4.0.2 Pairwise Character Impression Comparison

Generally, in a literary work, each character plays various roles across a wide range of events in the *syuzhet* or story line(s) of the novel. Characters most often also exhibit a diverse range of character traits. As such, each character is an individual, even if they share certain characteristics, or play similar roles, to other characters. For example, in *Animal Farm*, nearly all the characters are anthropomorphized animals, and live on a fictitious farm. Although there are multiple pigs in the story, each pig is distinctive from every other pig. In *To Kill a Mockingbird*, the characters are grounded in reality, sharing many recognizable characteristics (at least for American audiences) of small town America, and the central crisis of the novel and the myriad reactions of the characters creates an empathetic potential for the reader. Yet each reader brings to their experience of the novel a set of external experiences and conditions. These experiences allow each reader – and each reviewer – an opportunity to augment the construction of story lines and characters in the novel. To avoid falling prey to the "intentional fallacy" [93], where a critic tries to untangle the intentions of an author, the methods we devise here turn instead to an exploration of the constitutive nature of the reader reviews. Because each reviewer brings with them their own unique approach to reading, and given the wide range of characters and events in a novel, one might expect that these characters, especially mined from reviews, cannot be compared.

We find, however, that, while writing reviews, reviewers collate their character impressions into clusters of descriptors that are more semantically consistent *across characters* than the raw reviews would initially suggest. One possible reason for this finding could be a result of reviewers mapping their impressions into a shared consensus model of a character in an effort to write more convincing reviews and thereby receive more positive response from the broader community of reviewers. This could be based on reading other reviews of the same book, or reacting to comment threads on their own review or other reviews. Because of this semantic similarity in character descriptors, the impression clusters *enable inter-character comparison.*

The results of these inter-character comparisons may capture readers' broader understanding

of fictional characters, and the process by which communities of interpretation emerge. The alignments of character impressions across multiple fictional works may in turn reflect the consistency of approaches to reading, so that the text is constituted in a complex manner across many readings.

To illustrate these intriguing areas of character overlap across different works of fiction, we produce heatmaps for pair-wise comparison of distinct characters, as in Figure 6.13. Here we compare the impression clusters of "Victor Frakenstein" from *Frankenstein* to those of "Atticus Finch" from *To Kill a Mocking Bird*. The seemingly unlikely pair exhibit a surprising series of overlaps based on the readers' impressions of these characters. For example, the two have a high similarity score for clusters describing aspects of gender, responsibility and overall strength of character (as evidenced by the row/column labels in the figure).

One particularly interesting similarity is found in the clusters labeled "father,kids" for "Atticus" and "creator,name" for "Frankenstein". This similarity reflects a twofold process: first, the recognition of the readers of the similarity in these roles and second, the worldview encoded into BERT embeddings. BERT embeddings, as seen in single character heatmaps, carry additional artifacts into the realm of cross-character evaluation. For example, the cluster labeled "responsible" from "Frankenstein" and the cluster labeled "lawyer" from "Atticus" Finch have a highly negative similarity score. This suggests either that the readers have a negative bias against lawyers, or that pretrained BERT embeddings are biased, or both: regardless of the source of this bias, the combined model integrating reviewer comments and the cosine-distance measure when applied to BERT embeddings, seem to suggest that lawyers are *not* responsible.

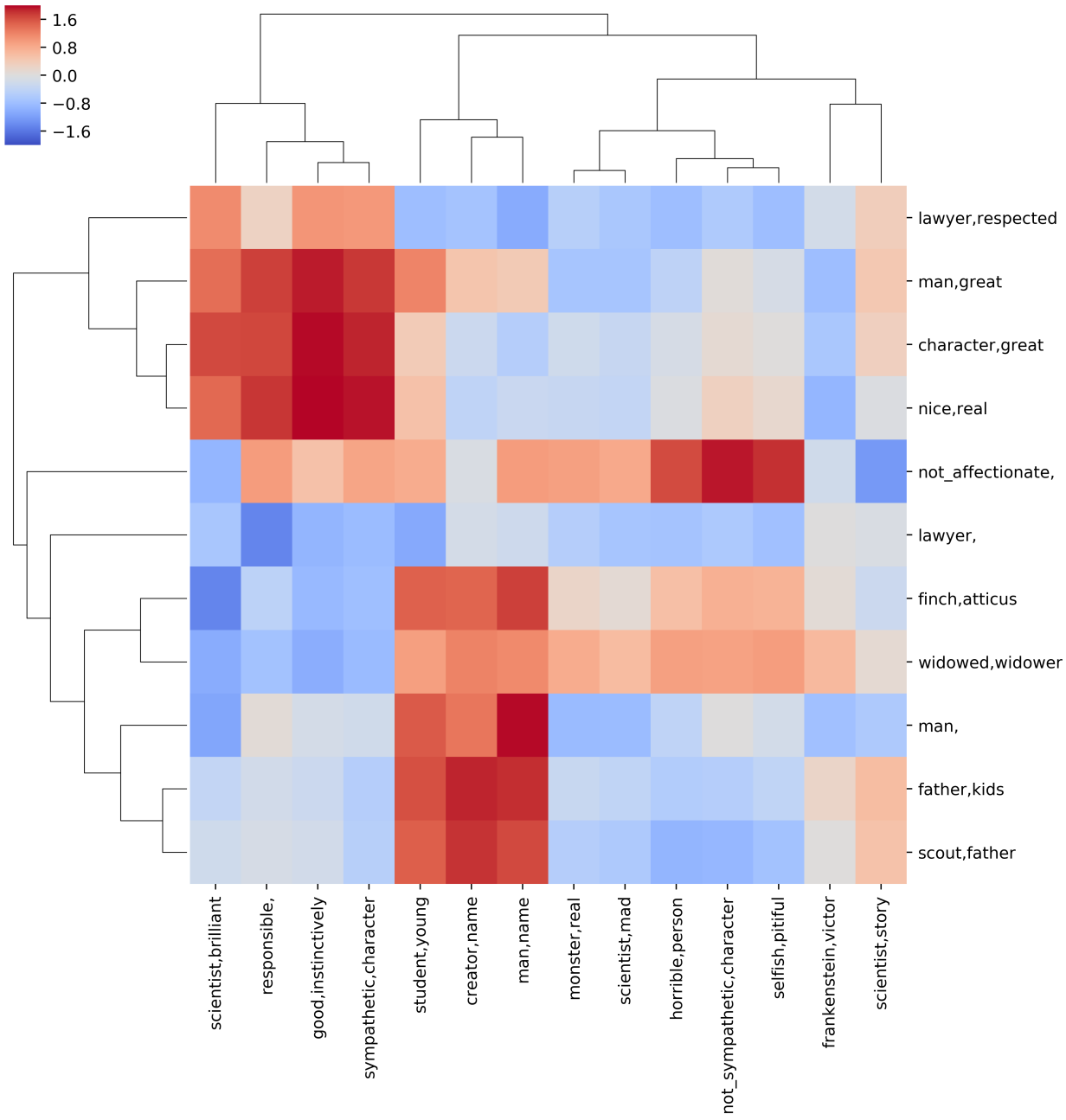

Figure 6.13: **The (asymmetric) heatmap comparing the character "Victor Frankenstein" from *Frankenstein* and "Atticus Finch" from *To Kill a Mockingbird***: The similarity scores between clusters of impressions labelled by the row/column headers are computed by Algorithm 8. The color coding of impression clusters suggests valuable information stored in these representations about pairwise character similarity across novels, capturing the readers' process of aligning impressions from one novel to impressions created while reading another novel.

Plotting the entropy of the single-character heatmaps can assist in the quantification of a character's perceived *complexity*. The resulting bar plot is presented in Figure 6.14. Not surprisingly, the relative number of impression clusters empirically correlates to the entropy: reviewers describe a wider range of impressions for complex characters than for less complex ones. However, this feature alone cannot explain all the trends observed in the plot. For "Jones", "Napolean" and "Boxer" in *Animal Farm*, each character is associated with a roughly equal fraction of impression clusters, yet "Napolean" emerges as a more complex character in the readers' conceptualizations of *Animal Farm* characters; this is not surprising, as "Napolean" is the most enduring villain in the plot. It is also noteworthy that the three actors are ascribed by readers similar roles in the plot and this similarity extends partially to their complexity measure. In *To Kill a Mockingbird*, "Atticus" is a central focus of the novel and it is his character that takes the spotlight as he defends "Tom". Indeed, "Boo" and "Scout" appear in the novel in many scenes to *support* "Atticus".

Of Mice and Men focuses on the dynamic between "George" and "Lennie", a pair of characters with notably different personalities, and the inherent complexity in their relationship. The resulting duality in character impressions, the limited number of additional characters in the novel, and a linear timeline results in a similar complexity profile for this pair of actants. The readers' impressions of characters that extend beyond the one-dimensional dismissal of inherently bad characters such as "Curley" may motivate them to focus more intently on these two, plumbing the depths of their personalities and trying to understand their decisions in the context of the cruel economic environment of Depression-era America.

The Hobbit rigorously follows the genre conventions of fantasy action-adventure, with a fairly clear delineation of "good guys" and "bad guys". As a result, "Bilbo", the main protagonist, attracts the most attention in discussion forums, which, in turn, contributes to a greater perceived complexity. Indeed, a feature of these complexity measures is that complexity increases with the amount of attention an actant attracts from the reviewers. This aspect is not a failing, however, and captures instead a part of the mental model that readers create as they read the novel and that they subsequently feel are important enough to share with other readers.

Last, in *Frankenstein*, while the "Monster" wreaks havoc, it is in fact "Frankenstein", the scientist, whose work raises ethical, moral and social concerns. Reader discussions about the character "Frankenstein" and his complex positioning in the novel ultimately foster debate about the purpose of science and frequently consider whether the scientist "Frankenstein" was perhaps the real monster. This effect is projected on both the relative number of impression clusters and the resulting complexity measure for the character in the reader reviews.

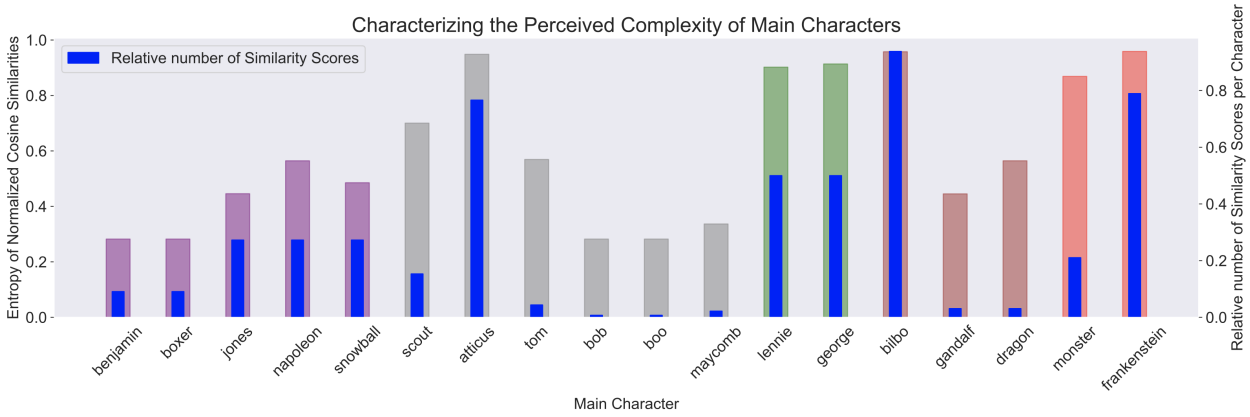

Figure 6.14: **A measure of perceived *complexity* per character across novels**: The color blue corresponds to the relative number of empirical samples per character-specific heatmap used to compute entropy (prior to smoothing). Each translucent color corresponds to a specific novel and plotted are the respective entropies of characters that have at least 4 impression clusters. We found $b = 50$, and $w = 3$ to be optimal hyperparameter choices to explore the differences in the complexity measure between characters.

The consensus impression and narrative models created by our framework enable one to turn the spotlight back on individual readers. While the individual reviews collectively reflect and encode the whole, the whole in turn constructs a rubric to better understand the parts, i.e. how individuals both align with and differ from the collective. For example, one might ask, what makes a review informative and useful? We may not be able to identify the exact features that constitute a good review but, according to our model, expanded interpretations of the overall story space, rich event sequences that emphasize main characters, and unique impressions of these characters constitute important proxy targets. If a review makes use of at least some of these features, it will consist of enough information about the novel to be

self-sustaining in the overall discussion space on the work without requiring access to that original work. Equally important is the presence of markers that indicate departures from the general consensus. Such departures not only set apart the individual review from the rest, but might be the seeds for the future emergence of new collective impressions in an evolving dialog over novels and characters.

Consider, for example, an impression cluster for "Atticus Finch" extracted by the SENT2IMP algorithm and labeled "Man,Great" (see Figure 6.13). This cluster of impressions collected from all the reviews of the novel consists of the phrases: {'a good father', 'the loving father', 'the best dad', 'a man of integrity' ...}. Similarly, there is another impression cluster labeled "Father,Kids," comprising the phrases { 'the father of protagonist', 'the father of Jem', ...}, and emphasizing his role as a father, while naming his children. Scout, the daughter and protagonist in the novel, has an impression cluster, labeled "Narrator,Smart" comprising the phrases: {'really smart', 'very thoughtful, 'a smart girl', ...}, bringing out a key attribute that has given the novel a lasting legacy. Tom Robinson, yet another pivotal character, has an impression cluster "Innocent,Man" with the phrases: {'a mere poor victim of circumstances', 'innocent', 'a good black man', ...}, and correctly portraying him as a victim of racial bias and violence.

In light of the preceding collective impression clusters, let us consider the following review for *To Kill a Mockingbird*:

> **Review 1**: "I think that To Kill A Mockingbird has such a prominent place in American culture because it is a naive, idealistic piece of writing in which naivete and idealism are ultimately rewarded. [...] Atticus is a good father, wise and patient; Tom Robinson is the innocent wronged; Boo is the kind eccentric; Jem is the little boy who grows up; Scout is the precocious, knowledgable child.",

The reviewer clearly aligns with and contributes to the majority views on the characters, Atticus Finch and Scout. Reviews such as this one contribute significantly more information to the review ecosystem than a more cryptic review such as the following about *The Hobbit*:

> **Review 2**: "Maybe one day soon I'll write a proper review of The Hobbit. In

the meantime, I want to say this: If you are a child, you need to read this for Gollum's riddles. [...]".

This review is not only brief, but also skips references to a majority of the story lines, event sequences or character impressions. It does, however, emphasize the role of Gollum, a hugely popular character in the movie adaptation of *The Hobbit*, and the reader's evaluation of the suitability for children of the character's riddles. Our model thus provides an evidential measure, and one can objectively conclude – as admitted by the reviewer – that the review plays only a peripheral role in the overall review space, yet retains the potential to seed further discussions.

6.5 Discussion

The results support the idea that readers, when summarizing a novel, tend to reduce the scope of the story and to focus on the most memorable aspects of the plot, here modeled as inter-actant relationships. In the reviews we studied, people converge on a set of main actants and relationships that map well to a core set of actants and relationships in the ground truth summaries, suggesting that people are relatively adept at summarizing even complex novels. As part of their summaries, however, people tend to simplify. This simplification may be related to cognitive limits on the number of real-world relationships that a person can keep in mind.

Since reviews tend to be short, when compared to the length of the work summarized, it is not surprising that people reduce both the number of actants, particularly in works with very large casts of characters such as *The Hobbit*, and the relationships between those actants. The inter-actant relationships are also simplified in the reader reviews. Readers can simplify complex plots, such as that in *To Kill a Mockingbird*, into relatively straight forward stories of conflict, strategies to address that conflict, and the result of the use of those strategies. The reduction of plot complexity may also be influenced by the abstraction of the novel in other media. For certain books, such as *The Hobbit*, recent films have been highly successful, and it is quite possible that movie watching has had some impact on reader reviews. The

same may apply to the other books in this study given, for example, the numerous references to the actor Gregory Peck in the reviews of *To Kill a Mockingbird*. Although we have not done so here, it may be interesting to compare reader reviews of filmatized novels to the summary story graphs for those films.

6.6 Conclusion

The approach we describe here is widely applicable to other crowd-sourced review sites such as Rotten Tomatoes and Metacritic (for films) and LibraryThing and Love Reading (for literature) that, much like Goodreads, allow viewers or readers to present their own reviews of fiction, be it literature or film. An intriguing aspect of many of these sites is the propensity of reviewers to provide "plot summaries" as opposed to critical engagements with more sophisticated thematic analysis. While this plot-based approach to reviewing works of fiction may drive literary scholars to the brink of insanity, it does allow us to consider questions regarding the popular engagement with literature and other forms of artistic production. Given the responses that people post, we can use the scale of these sites to derive insight into how people (or groups of people) not only read but also remember.

Aggregation techniques on entity mentions and their relationships provide an unparalleled level of resolution for constructing consensus knowledge representations be that for aggregating reviewers' character impressions. This method of extracting open-world, infinite vocabulary knowledge graphs from partial information samples (reviews) forms an explainable and powerful tool to *discover* narratives, potentially in an online fashion. This pipeline then has applications not only in plot synthesis from novel reception, but also in other similar settings that employ narrative theory such as social media where real-life event reception drives the generation of posts.

Reader reviews, such as those from *Goodreads*, are not often considered in the context of literary analysis. We believe, however, that they provide an intriguing window into the broad cultural memories of "what a book is about." Sophisticated analyses of theme, or the deep anchoring of a literary work in a detailed intellectual, social and historical context,

may at times elude the thousands of reviewers contributing individual reviews to these social reading sites. Yet, despite these failings, the reviews still capture the meaningful thoughts of thousands of readers, each with their own diverse motivations for reading and reviewing, and are thus reflective of these readers' literary engagement [12, 15, 20]. Although they are usually unknown to each other, the readers of a particular work of fiction implicitly create an imagined community that shares, at least for some time, an interest in that work [16]. Complex, dynamic characters are conceptualized as a series of impressions that, despite their simplicity in an individual review, capture in the aggregate some of the complexity of character that lies at the heart of fiction writing and literary analysis.

Importantly, our approach allows us to preserve an awareness of the individual reader who carries with them their own compact representation of a complex work of fiction while also contributing to a collective, and often more complicated, overview of that work. Because our methods capture both how an individual reads and reviews, and how the broader community of readers of the same work read and review, it is possible to glimpse the relationship between a reader and the communities of interpretation that they are writing with, against and across. The numerous pathways through the narrative framework, in that sense, capture the multiple ways that people understand, remember, and recount their own individual engagement with the work of fiction.

Although a frequent refrain of teachers of literature is that amateur or otherwise "uninformed" engagements with literature are nothing more than "plot summary", our exploration of *Goodreads* countermands this criticism. The reviews we considered, for example, encode far more information than simple plot summary. Inevitably, reviewers include their impressions of one or two characters as well as some small number of events meaningful to them in their understanding of the novel. Readers, of course, draw their impressions of characters in any work of fiction not only from that work itself, but also from all of their experiences of other characters and events, both real and fictional. Consequently, by considering these reviews in the aggregate, one can derive insight into readers' attempts to draw comparisons across novels, both on the basis of genre and story structure, and also on the level of character. As we show, readers' impressions of a character from one novel resonate with similar

impressions of a character from another novel – even if those novels are as unalike as *To Kill a Mockingbird* and *Frankenstein* – thereby establishing a network of inference and allusion that resonates throughout the collective reservoir of reading. What we discover in these reader reviews, when taken collectively, echoes – in a data-driven manner – some of the fundamental literary critical ideas of the relationship between readers and texts.

In short, our methods allow one to explore the individual and collective reimaginings of a novel – the constitutive aspects of reader response that have been at the foundation of several strands of literary criticism from the early phenomenological reader response theories of Iser and others [94], through the explorations of communities of interpretation advocated by Fish [17], to concepts of intertextuality rooted in the work of Julia Kristeva [95] and its resonance in the work of Roland Barthes [96] among others. So, while individual reviews might not tell the whole story, and may on the individual level fail to capture the complexity of characters, the collective impressions of thousands of readers provide important insight into how people read, remember, retell and review. In so doing, these methods allow us to do many things, including reassemble a portrait of a tortured scientist and his monster.

6.7 Limitations

Data can be noisy, particularly when social media posts, which are informal by nature, are the primary source. This informality creates noise in the relationship extraction phase. A missing punctuation mark, for example, can significantly change the dependency tree structure and lead to erroneous extractions of both the arguments and the relationship phrases. Other parts of our pipeline are equally sensitive to noise, including pronoun resolution and BERT embeddings. While pronoun resolution is needed to improve coverage (that is, to capture relationships amongst entity mention references when they are expressed in terms of pronouns), the process adds additional noise by occasionally resolving pronouns to the wrong entity mentions. Error from pronoun resolution is more noticeable in relation to rare words. For example, in the sentence, "The example their single father Atticus sets for *them* is one all parents wish we could parallel.", *them* is mapped to the single character *Dill*. *Dill*

is among the characters mentioned least frequently in reviews of *To Kill a Mockingbird*. In such a scenario, the extracted relationships have a low fidelity because of the sparse sample space. In addition, while the BERT embeddings that we use for this work provide useful vectors in cosine-measured k-means clustering, the approach also suffers from sensitivity to noise. Using SparkNotes as a ground truth also raises some issues, as the summaries in these reader guides are less detailed than the novels that they summarize. Consequently, comparing our extractions to the limited relationships described in SparkNotes means that some of our discovered relationships, which may be in the novel but not in the SparkNotes summary, are improperly evaluated (i.e. the relationship exists in both the target novel and our extractions but is missing in SparkNotes). For example, while our extractions reveal that George cares for or loves Lennie in *Of Mice and Men*, this relationship is missing from the SparkNotes summary. Similarly, certain actants or relationships that exist in the ground truth summaries may simply be absent from the reader review corpus, as is the case for certain Frankenstein actants such as M. Krempe. Our methods are not able to discover actants or relationships that do not appear in reader reviews–this elision of characters and relationships, however, may be indicative of interesting aspects of reader review practice.

Book Name	Entity Mention Groups
Of Mice and Men	**Lennie**:[Lennie, lenny], **George**:[george, milton], **Curley's Wife**:[curley's wife, tart, wife], **Aunt Clara**:[aunt clara, aunt, clara], **men**:[workers, men], **ranch**:[ranch, farm], **soft things**:[soft things, soft, things], **mental disability**:[mental disability, mental, disability]
The Hobbit	**Bilbo**:[bilbo, baggins, burglar, hobbit], **Rivendell**:[rivendell, middleearth], **Gandalf**:[gandalf, wizard, gandolf, grey], **dwarf**: [dwarf, dwarves], **Thorin**: [thorin, company], **trolls**:[trolls, orcs], **elf**:[elf, elves], **Hobbitown**:[hobbitown, shire, hobbiton], **man**: [human, man, lakemen], **dragon**:[dragon, smaug]
Frankenstein	**monster**:[monster, creature, adam], **Frankenstein**:[frankenstein, victor, doctor, creator], **Mary Shelley**: [mary, shelley, author, mary shelley], **Elizabeth**:[elizabeth, wife], **Walton**:[walton, robert], **Henry**:[henry, clerval], **Justine**:[justine, moritz], **Caroline**:[caroline, beaufort]
To Kill a Mockingbird	**Scout**:[scout, sister], **Atticus**:[atticus, dad, father, finch], **Jem**:[jem, brother], **Harper Lee**: [lee, harper lee, author, harper], **Tom**: [tom, robinson, negro, mockingbird, africanamerican], **Bob**:[bob, ewell], **Boo**: [boo, arthur, arthur radley, boo radley], **Mayella**: [mayella, daughter], **aunt**: [aunt, alexandra], **Maycomb**: [maycomb, alabama, town], **Heck**:[heck, tate], **Cunningham**:[cunningham, walter]

Table 6.5: Final actants after EMG per book. Each actant group is labeled with the most frequent mention in the group. Empirically, these automatically computed labels match the ground truth entities as derived from SparkNotes.

CHAPTER 7

Concluding Remarks and Future Work

Throughout this dissertation, we aim to understand the collective narrative that spread across thousands of social media posts. We create a mirror-like tool to display the information told on social media. In 2 we introduce our generative narrative model. We represent narrative framework as a graph that consists of characters and their relationships. We explain the intuition behind these elements and how taken together they form the narrative graph. We introduce a graph-based algorithm, Entity Mention Grouping (EMG) that helps us discover characters based on their role in narrative networks. EMG runs after we extract the narrative graph. It looks into the nodes' roles in the narrative graph and discovers the characters called with different mentions. Given the expert-generated novel plots, we devise an unsupervised method to compare them with our extracted graphs. The increase in the accuracy after applying EMG proves its success in determining the characters. In chapter 3, we propose computational methods to extract the narrative graphs that include sentence-level relationship extraction, embedding-based supernode/subnodes derivation, graph formation, and a statistical method to mine multiple stories. We propose a threat detection method to find critical nodes in our graphs given the narrative graph. A threat node is one of the critical elements in the narrative model of a conspiracy theory. This classification problem opens the door to perform supervised machine learning based on our narrative graphs. We provide the detailed implementation of the EMG algorithm and evaluation methods further in 3.

We demonstrate our computational results in three different chapters. In chapter 4, we study two datasets. The first dataset is about the Pizzagate conspiracy theory, and the second is the Bridgegate conspiracy. We derive the narrative networks based on the models

earlier described in 2. We demonstrate how a conspiracy theory consists of multiple loosely connected layers.

In chapter 5 we perform the narrative extraction on a dataset on Covid-19 conspiracy theories, and we disentangle a narrative graph made from multiple conspiracy theories. We present our computational results comparing news-generated communities and our discovered stories. We report the homogeneity and completeness comparing two sets of communities. In chapter 6 we study the literary book reviews posted on social media. We develop EMG and interactant relationship clustering discussed earlier in 2. We show improvements in the edge detection accuracies after applying EMG on four different novels. This algorithm relies on the semantic role of each mention in the extracted narrative graph. We compare the narrative graphs extracted from social media with the original plot. While individual reviews might not tell the whole story and may, on the individual level, fail to capture the complexity of characters, the collective impressions of thousands of readers provide essential insight into how people read, remember, retell and review. These methods allow us to do many things, including reassembling a portrait of a tortured scientist and his monster. In future work, we will expand our approach in several directions itemized below:

- The clustering method used on noun phrases embeddings significantly impacts the network's quality and size in nodes and edges. We have explored Kmeans clustering with the elbow point, HDBscan, DBscan, and their different variations. Designing an optimization problem on the actant discovery using word embeddings would be worthwhile that can achieve the optimized set of characters and their relationships.

- Time series analysis: Currently, we generate our daily networks on the most recent social media posts. Every day, we collect posts on multiple platforms, and after removing the Personal Identifying Information(PII), we extract the narrative graphs. Due to their structural nature, these networks provide us with notorious information on daily happenings. Among the recent highlights, we observe many discussions on vaccine hesitancy, and our system has been able to collect valuable social media posts on this topic. One area of promising future work is time series analysis. We believe

that the dynamic of the discussions is valuable for social science studies. Our narrative extraction pipeline provides a fundamental step to observe such dynamics.

- Toward supervised narrative extraction: The threat detection classifier proved the achievable high accuracy on one of the story elements. Recall that a story has a set of characters, each with different roles such as an insider or outsider, a threatening action or character, a strategy to combat the threat, and finally, an outcome. Recent access to labeling mechanical crowd workers makes designing and getting labeled training data sets easier. We may use the language models and fine-tuning pre-trained networks to obtain the story elements and expand our misinformation detection to post-level. Post-level analysis helps us avoid noise in the network community detection and lets us have topic filtering prior to building the narrative graphs.

- In this work, we studied five classical novels. These novels were selected based on their unique story structure; however, studying other genres, writers, or popular books would be interesting. Other online platforms that might attract different age groups will present exciting results.

To conclude, the unsupervised methods developed in this dissertation take the first steps to understand the narrative and the context behind conspiracy theories from thousands of social media posts. It becomes more evident that conspiracy theories are loosely connected domains. To develop a more refined understanding of narratives, we study online book readers' reviews and track the book's original plots in the collective narratives extracted from these posts.

REFERENCES

[1] Homogeneity, completeness and v-measure. `https://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness`.

[2] Carl Wilhelm Von Sydow. On the spread of tradition. *CW von Sydow, Selected papers in folklore*, pages 11–43, 1948.

[3] Juha Pentikainen. Oral repertoire and world view. an anthropological study of marina takalo's life history. *FF Communications Turku*, 93(219):1–366, 1978.

[4] Siikala Anna-Leena. Interpreting oral narrative. *Folklore Fellows' Communications*, 245, 1990.

[5] Timothy R Tangherlini. Toward a generative model of legend: Pizzas, bridges, vaccines, and witches. *Humanities*, 7(1):1, 2018.

[6] Algirdas Julien Greimas. Éléments pour une théorie de l'interprétation du récit mythique. *Communications*, 8(1):28–59, 1966.

[7] William Labov and Joshua Waletzky. Narrative analysis: Oral versions of personal experience. 1997.

[8] George Boole. *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities*. Dover, 1854.

[9] Timothy R Tangherlini, Vwani Roychowdhury, Beth Glenn, Catherine M Crespi, Roja Bandari, Akshay Wadia, Misagh Falahi, Ehsan Ebrahimzadeh, and Roshan Bastani. "mommy blogs" and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR public health and surveillance*, 2(2):e166, 2016.

[10] Timothy R. Tangherlini, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, pizzagate and storytelling on the web. *PLOS One*, 15(6):1–39, 06 2020.

[11] John Laudun. Talk about the past in a midwestern town:"it was there at that time.". *Midwestern Folklore*, 27(2):41–54, 2001.

[12] Simone Rebora, Peter Boot, Federico Pianzola, Brigitte Gasser, J Berenike Herrmann, Maria Kraxenberger, Moniek Kuijpers, Gerhard Lauer, Piroska Lendvai, Thomas C Messerli, et al. Digital humanities and digital social reading. 2019.

[13] Federico Pianzola, Simone Rebora, and Gerhard Lauer. Wattpad as a resource for literary studies. quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS One*, 15(1):e0226708, 2020.

[14] Maria Antoniak, Melanie Walsh, and David Mimno. Tags, borders, and catalogs: Social re-working of genre on librarything. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), April 2021.

[15] Ed Finn. *The Social Lives of Books: Literary Networks in Contemporary American Fiction*. Stanford University, 2011.

[16] Benedict Anderson. *Imagined communities: Reflections on the origin and spread of nationalism*. Verso books, 2006.

[17] Stanley Eugene Fish. *Is there a text in this class?: The authority of interpretive communities*. Harvard University Press, 1980.

[18] David Peplow, Joan Swann, Paola Trimarco, and Sara Whiteley. *The discourse of reading groups: Integrating cognitive and sociocultural perspectives*. Routledge, 2015.

[19] Simone Rebora and Federico Pianzola. A new research programme for reading research: analysing comments in the margins on wattpad. *DigitCult-Scientific Journal on Digital Cultures*, 3(2):19–36, 2018.

[20] Joan Swann and Daniel Allington. Reading groups and the language of literary texts: A case study in social reading. *Language and Literature*, 18(3):247–264, 2009.

[21] Shadi Shahsavari, Ehsan Ebrahimzadeh, Behnam Shahbazi, Misagh Falahi, Pavan Holur, Roja Bandari, Timothy R. Tangherlini, and Vwani Roychowdhury. An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In *12th ACM Conference on Web Science*, WebSci '20, page 277–286, New York, NY, USA, 2020. Association for Computing Machinery.

[22] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[23] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[24] Mattia Samory and Tanushree Mitra. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[25] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[26] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[27] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592, 2014.

[28] Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Samuel Bowman, Timothy Dozat, and Christopher D Manning. More constructions, more genres: Extending stanford dependencies. In *Proceedings of the second international conference on dependency linguistics (depling 2013)*, pages 187–196, 2013.

[29] Sebastian Schuster and Christopher D Manning. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, 2016.

[30] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[31] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[32] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 523–534. Association for Computational Linguistics, 2012.

[33] Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366, 2013.

[34] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[37] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[38] Sergei Vassilvitskii and David Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

[39] Jerome R Bellegarda, John W Butzberger, Yen-Lu Chow, Noah B Coccaro, and Devang Naik. A novel word clustering algorithm based on latent semantic analysis. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 172–175. IEEE, 1996.

[40] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[41] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[42] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[43] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[44] Thomas Aynaud. Community detection for networkx's documentation, 2018.

[45] Reza Zafarani and Huan Liu. Evaluation without ground truth in social media research. *Communications of the ACM*, 58(6):54–60, 2015.

[46] Silvio Waisbord. Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies*, 19(13):1866–1878, 2018.

[47] Lars Wissler, Mohammed Almashraee, Dagmar Monett Díaz, and Adrian Paschke. The gold standard in corpus annotation. *IEEE GSC*, 21, 2014.

[48] Gregor Aisch, Jon Huang, and Cecilia Kang. Dissecting the# pizzagate conspiracy theories. *The New York Times*, 10:2, 2016.

[49] Bill Marsh and Zernike K Chris Christie. the lane closings: a spectator's guide. *The New York Times*, 2015.

[50] Evelyn Gius, Nils Reiter, and Marcus Willand. Foreword to the special issue "a shared task for the digital humanities: annotating narrative levels". *Journal of Cultural Analytics*, 4(3):11202, 2019.

[51] Ron Artstein. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer, 2017.

[52] Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523, 2010.

[53] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[54] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[55] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[56] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[57] Claude Sammut and Geoffrey I. Webb, editors. *TF–IDF*, pages 986–987. Springer US, Boston, MA, 2010.

[58] D. N. Joanes and C. A. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):183–189, 1998.

[59] S. Kokoska and D. Zwillinger. Crc standard probability and statistics tables and formulae, student edition. 1999.

[60] Simon DeDeo, Robert X. D. Hawkins, Sara Klingenstein, and Tim Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013.

[61] Abby Ohlheiser. Fearing yet another witch hunt, reddit bans 'pizzagate.'. *The Washington Post*, 2016.

[62] Colin Klein, Peter Clutton, and Adam G Dunn. Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum. *PloS one*, 14(11):e0225098, 2019.

[63] Panagiotis Metaxas and Samantha T Finn. The infamous# pizzagate conspiracy theory: Insight from a twittertrails investigation. 2017.

[64] Misagh Falahi. *A cognition-driven approach to modeling document generation and learning underlying contexts from documents*. University of California, Los Angeles, 2017.

[65] Mattia Samory and Tanushree Mitra. 'the government spies using our webcams': The language of conspiracy theories in online discussions. *Proc. ACM Hum. Comput. Interact.*, 2:152:1–152:24, 2018.

[66] Roja Bandari, Zicong Zhou, Hai Qian, Timothy R. Tangherlini, and Vwani P. Roychowdhury. A resistant strain: Revealing the online grassroots rise of the antivaccination movement. *Computer*, 50:60–67, 2017.

[67] Mikolaj Morzy. On mining and social role discovery in internet forums. *2009 International Workshop on Social Informatics*, pages 74–79, 2009.

[68] Jim Maddock, Kate Starbird, Haneen J. Al-Hassani, Daniel E. Sandoval, Mania Orand, and Robert M. Mason. Characterizing online rumoring behavior using multidimensional signatures. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015.

[69] Colin Klein, Peter Clutton, and Vince Polito. Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in Psychology*, 9, 2018.

[70] Peter Shawn Bearman and Katherine Stovel. Becoming a nazi: A model for narrative networks. *Poetics*, 27:69–90, 2000.

[71] Wendy G Lehnert. Narrative text summarization. In *AAAI*, pages 337–339, 1980.

[72] Ethan Zuckerman. Qanon and the emergence of the unreal. *Issue 6: Unreal*, 2019.

[73] David Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David M. Rothschild, Michael Schudson, Steven A. Sloman, Cass Robert Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan Zittrain. The science of fake news. *Science*, 359:1094 – 1096, 2018.

[74] Véronique Campion-Vincent and Jean-Bruno Renard. Conspiracy theories today. *Diogenes*, 249(1):1–264, 2015.

[75] Walter Anderson. *Kaiser und Abt: Die Geschichte eines Schwanks*, volume 42. Suomalainen Tiedeakatemia, Academia Scientiarum Fennica, 1923.

[76] Kate Starbird. Information wars: A window into the alternative media ecosystem. *Medium.(Mar. 2017)*, 2017.

[77] Russell Muirhead and Nancy Rosenblum. The new conspiracists. *Dissent*, 65(1):51–60, 2018.

[78] Peter Pomerantsev and Michael Weiss. The menace of unreality: How the kremlin weaponizes information, culture and money. 2014.

[79] Shadi Shahsavari, Pavan Holur, Timothy R Tangherlini, and Vwani Roychowdhury. "covid-19_conspiracy-theories.".

[80] David Nakamura. With'kung flu,'trump sparks backlash over racist language-and a rallying cry for supporters. *Washington Post. June*, 24, 2020.

[81] Elahe Izadi and Sarah Ellison. Americans want to see what's happening in hospitals now, but it's hard for journalists to get inside. *The Washington Post*, 2020.

[82] J Eric Oliver and Thomas J Wood. Conspiracy theories and the paranoid style (s) of mass opinion. *American journal of political science*, 58(4):952–966, 2014.

[83] Ted Goertzel. Belief in conspiracy theories. *Political psychology*, pages 731–742, 1994.

[84] Amy Davidson Sorkin. The dangerous coronavirus conspiracy theories targeting 5g technology, bill gates, and a world of fear. *The New Yorker*, 2020.

[85] Emilio Ferrara. # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*, 2020.

[86] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016.

[87] Mary Shelley. Frankenstein. london: Lackington, hughes, harding, mavor, and jones, 1818. ed. stuart curran. *Romantic Circles Electronic Editions*, 16, 2015.

[88] John Steinbeck. Of mice and men. new york: Covici & friede, 1937.

[89] John Ronald Reuel Tolkien. *The hobbit*. Houghton Mifflin Harcourt, 2012.

[90] Harper Lee. To kill a mockingbird. philadelphia & new york, 1960.

[91] Mike Thelwall and Kayvan Kousha. Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, 68(4):972–983, 2017.

[92] Harper Lee. *Go Set a Watchman*. Harper Collins, 2015.

[93] William K. Wimsatt and Monroe C. Beardsley. *The Verbal Icon*. University of Kentuck Press, 1954.

[94] Wolfgang Iser. *The act of reading: A theory of aesthetic response*. JHU Press, 1979.

[95] Julia Kristeva. *Desire in language: A semiotic approach to literature and art*. Columbia University Press, 1980.

[96] Roland Barthes. The death of the author. *Contributions in Philosophy*, 83:3–8, 2001.