**BANDARI — What**

<2 figures>

**What Do You Do with a Million Readers?**

Bandari, R., Tangherlini, T. R. and Roychowdhury, V.

In 2006, Gregory Crane posed the provocative question, ‘What do you do with a million books?’, a question that captured the increasing anxiety in the humanities that accompanied the rapid digitization of many library collections, and the distribution of more and more books in digital form (Crane, 2006). Far more books than could ever possibly be read by a single person were now machine actionable, requiring a response from the intellectual community. In the ensuing years, that response has emerged from the digital humanities community, and DH projects have focused on a broad range of approaches to the study of literature at scale, operationalizing the ‘distant reading’ that Franco Moretti had already proposed in his well-known article, ‘Conjectures on World Literature’ (Moretti, 2000; 2013). Yet what has been overlooked in many of these studies is that, alongside the explosion in the digitization of world literature, there has been an equally large explosion of readers commenting on books in online forums and other easily accessed electronic venues. The analysis of these responses allows for the consideration of reader response at scale. The goal of our work is to provide a preliminary answer to the question, ‘What can one do with a million readers?’ At the very least, we want to know what types of information one can extract automatically from thousands of reader posts about a particular work of literary fiction, and what this can tell us about how people (or classes of people) read.

The target data for our study are reader reviews of sixteen works of fiction, five of which we focus on in this presentation (*The Hobbit*; *Gone with the Wind*; *The Life of Pi*; *Frankenstein*; *Of Mice and Men*). The works were chosen from the list of the most frequently rated books on the Goodreads site (number of ratings > 500,000). From these books, we downloaded the maximum allowed 3,000 reviews. The sixteen works we ultimately chose were selected on the basis of the broad disparity in their narrative structures, number of characters, and character relationships. The initial goal was to develop a review-based summary of the novel in the form of a character graph (Figure 1), with dramatis personaeand pair-wise connections between them based on actions, events, or other relationships. These graphical representations can be seen as an abstraction of what the reviewers on Goodreads collectively imagine who the main characters and what the main events to be, and the relationships between these characters and events.

The reviews were harvested using a crawler specifically designed for this project. Readily available information on the reviewer was retained as metadata for use in addressing second-order questions related to classes of reviewers (e.g., gender, age, frequency of reviewing). To evaluate reader reviews at scale, we devised two metrics related to plot: completeness and accuracy. We used SparkNotes as a basis for ‘gold standard’ summaries, a choice motivated by SparkNotes’ high degree of completeness and accuracy for target features (entities and relationships), and the brevity of the summaries.

We devised several progressively more difficult challenges to test the type of information we could derive from the approximately 48,000 reader reviews we crawled from Goodreads. First, without any training data, could we successfully discover the main dramatis personae in each novel? Second, could we automatically discover the action-based relationships between characters as represented by the reviewers? Third, could we discover the ‘events’ in which dramatis personae played a role? And fourth, could we develop a visualization that captured these relationships in a clear and engaging fashion that also represented the varying degree or strength of relationships between these entities?

To achieve reproducible results, we devised a simple workflow that can be applied to other similar sites. This workflow consists of a preprocessing step, a statistical entity ranking step that surfaces the main dramatis personae in any given work, a pair-wise relationship discovery step, and a visualization step. As with most blogs and crowdsourced data sources, the Goodreads reviews are ‘noisy’. Consequently, preprocessing focused on reducing noise. We also ran language detection on the reviews to eliminate non-English reviews, and a stemmer to aggregate inflected words.

In our efforts to discover the dramatis personae for any target work, we experimented with three main approaches. LDA proved to be successful at separating the works; we intend to explore this approach more thoroughly in future work for classifying reviews that are not preclassified as they are on Goodreads. Traditional NER (named entity recognition) approaches proved to be less accurate, given the broad variance in orthography that typify these reviews. Ultimately, the most successful approach was based on a statistical ranking of tokens between the review corpus and the individual subcorpus of target work reviews.

The next challenge was to discover the relationships between dramatis personae as represented in the reviews. Here, we focused on the verbs, extracted using POS tagging, between two entities, after discovering all sentences with pairs of entities. In future work, we hope to refine these pair-wise relationships by collapsing verbs into a series of higher-level representations of the entity-relationship space. Even without this processing, the current approach discovers important relationships. For example, by looking at the reviews for *The Hobbit*, the relationship between ‘Tolkien’ and ‘Hobbit’ is dominated by the verb ‘write’ with a directional relationship in which ‘Tolkien’ inhabits the subject position and ‘Hobbit’ the object position. Similarly, ‘Bilbo’ has a relationship with ‘ring’ characterized by ‘find’, and a relationship with ‘adventure’ of ‘go on’, while the dragon ‘Smaug’ has a relationship with ‘treasure’ of ‘guard’. Interestingly, the reader reviews also generated a relationship between ‘Bilbo’ and ‘Smaug’ characterized by ‘kill’, an inaccurate depiction of the actual events (Bard the Bowman slays Smaug). Our automatic method is able to build a surprisingly large number of relationships between discovered entities without any training data, nor any preexisting lists of entities or relationships.

For evaluation, using only the SparkNotes plot summaries, experts created a list of nouns including (1) names, (2) locations, (3) objects, and (4) concepts that are explicitly mentioned in the summary; these were considered the true dramatis personae. Completeness was quantified by computing the proportion of this list also produced by the algorithm. We measured both Precision (proportion of main characters produced by the algorithm) and the False Detection Rate (proportion of produced characters not in the relevant set). Similarly, using only words explicitly found in the SparkNotes plot summary, experts derived relationships based on verbs that connected pairs of characters mentioned above. We again measured the Precision and the False Detection Rate (FDR) for these relationships. As an example, applying this to *The Hobbit* reviews produced a measure of Completeness (or precision) for dramatis personae of 0.6 and Accuracy (False Detection Rate) of 0.13.

The final challenge was to present these relationships in a visually engaging manner (see Figures 1 and 2). We have developed directional, multicolored graphs that represent the strength (or confidence) of a relationship by an edge of varying width. These graphs are easily compared with the ‘gold standard’ ground truth graphs, and they provide us with a visual representation of our Completeness and Accuracy measures. What is immediately clear is that reader reviews have significantly lower Completeness than a resource dedicated to providing comprehensive summaries, while the Accuracy of described relationships is good (Bilbo’s dragon-slaying feats notwithstanding). The comparison raises intriguing issues about memory—for example, why is it that certain events disappear from the user-driven graphs, while others become accentuated?

Other graphs are generated for classes of reviewers: e.g., female reviewers vs male reviewers of *The Hobbit*, which allow for a different type of comparison. Here the question is on which aspects of a story different types of reviewers tend to comment. Additional refinements could include metrics that reveal the number of reviews that mention particular entities or particular relationships. Currently, a missing component is a dynamic representation of reviewers’ concepts of plot (dynamics), which we are reserving for future work.

The approach we describe here is widely applicable to other crowdsourced response sites. Of particular interest are movie review sites such as Rotten Tomatoes that, much like Goodreads, allow viewers to present their own reviews of popular films. An intriguing aspect of many of these review sites is the propensity of reviewers to provide ‘plot summaries’ as opposed to critical engagements of more sophisticated thematic analysis. While this may drive many literary scholars toward 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 do post, can we use the scale of these sites to derive insight into how people (or groups of people) not only read but also remember? It is our contention that what people remember, and what people forget (or choose to leave out), can be very telling indicators of popular engagement with art.

**References**

**Crane, G.** (2006). What Do You Do with a Million Books? *D-Lib Magazine,* **12**(3): 1.

**Moretti, F.** (2000). Conjectures on World Literature. *New Left Review,* **1** (January–February): 54–68.

**Moretti, F.** (2013). Operationalizing: or, the Function of Measurement in Modern Literary Theory. *New Left Review,* **84** (November–December): 103–19.

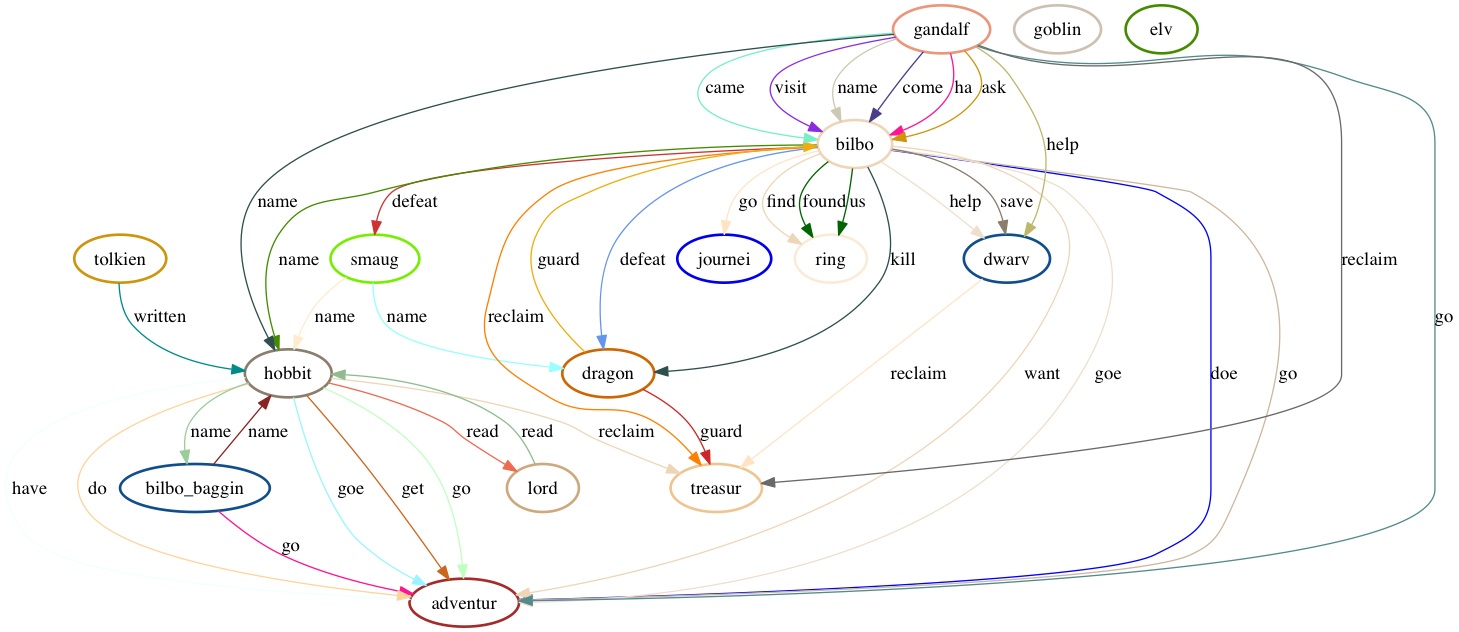


Figure 1. Character/relationship graph of *The Hobbit*.

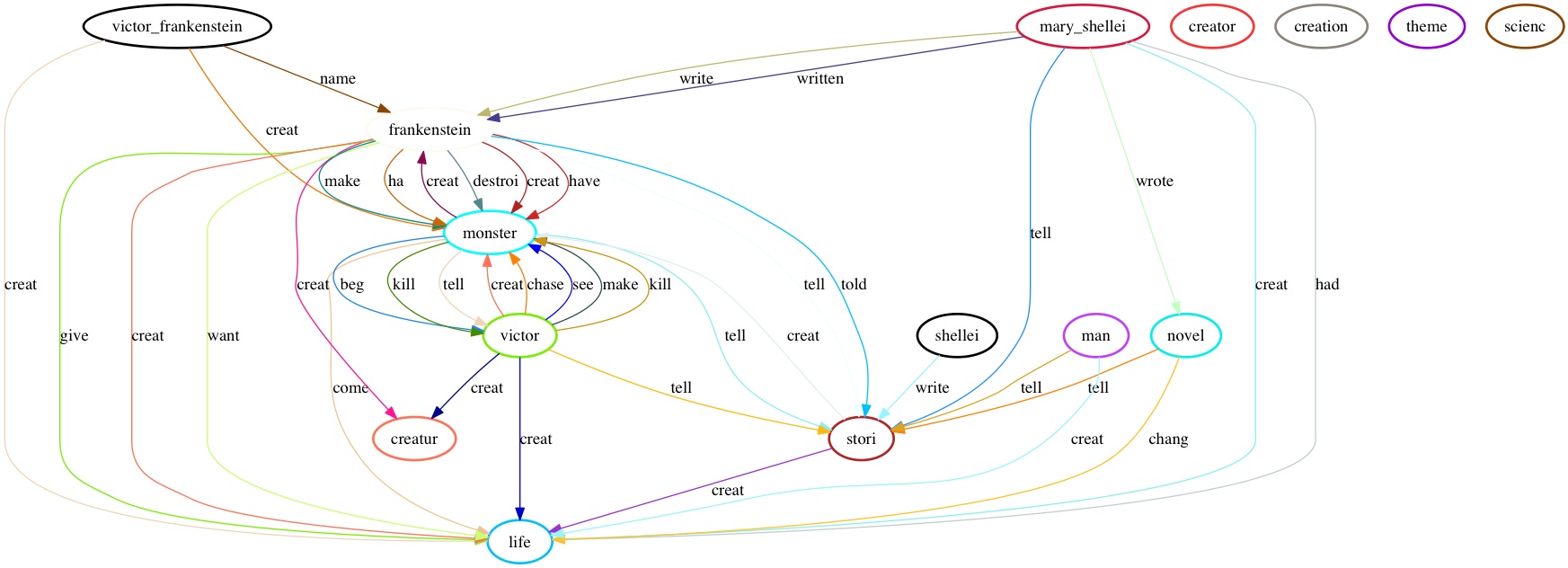


Figure 2. Character/relationship graph of *Frankenstein.*