**ALGEE-HEWITT­ — Measuring**

<1 table>

**Measuring Virality:** **Quantifying Formal and Paratextual Features Associated with ‘Viral’ Books**

Algee-Hewitt, M., Morgan, F., Fredner, E., Porter, J. D. and Walser, H.

In this poster, we outline a working methodology and present preliminary results from a new project in Stanford’s Literary Lab. The goal of this project is to measure a contemporary novel’s potential ‘virality’. This project is the result of a research partnership the Literary Lab made with a publishing company to analyze a corpus of more than 43,000 ebooks from across the 20th and 21st centuries.

**Why ‘Virality’?**

From a literary critical perspective, we have chosen to study the phenomenon of virality in order to approach broader questions of genre, reader reception, and the spread of popular fiction (Latour, 2005). We believe that a study of virality will also identify some formal characteristics associated with popular contemporary fiction, as well as a subset of formal characteristics that do *not* correlate with popularity. Answering these questions will provide some key information for making historically situated claims about the nature of pop fiction in the late 20th and early 21st centuries, especially as it relates to the rise of the Internet and the changing nature of novel reading.

As an object of quantitative analysis, virality poses an interesting challenge. Going into this project, we had no assurances that there exist formal features that correlate with viral texts. Virality may simply be a phenomenon that exists at the level of social networks or, even more broadly, ‘culture’. Moreover, it is undeniable that when a text goes viral it does so at least in part outside of the bounds set by its covers and spine. After all, a copy of a book is usually sold before it is read. The ambitious goal of our study is to identify formal markers at the level of the text that correlate with the larger social enactment of virality.

# Definition

As it is usually used in reference to digital media, ‘virality’ refers to the tendency of a piece of content to be ‘shared rapidly and widely among Internet users’.1 Unfortunately, this definition cannot be mapped directly onto novels. There are a number of technical and cultural factors that preclude the possibility of using the standard definition of virality for our project. First and foremost, viral Internet content is generally shared as a hyperlink. Readers access the link through any number of digital communication tools, including email, social media, and so on. There are also qualitative and quantitative differences between clicking on a link and choosing to read a book, such as time commitment (most viral Internet content is short; most novels require hours to read) and monetary investment (most viral content is free; most novels aren’t). Most importantly, this born-digital notion of virality suggests a direct person-to-person spread of information—clicking on a link someone tweeted or posted on your wall, for example. The closest analogue to this notion of virality would be lending a friend a book that you recommend. But the limits of this analogy are immediately apparent: A link can be viewed by as many people as the server can handle at once, while an ink-and-paper novel only serves one reader at a time.

Moreover, we argue that viral novels do not always have a one-to-one relationship between recommender and new readers, though that certainly can be the case. Like extremely viral Internet content, viral novels become visible at multiple levels of mainstream culture: social media, television, radio, references in conversation, reading in public, and so forth. While the mechanism is difficult to describe with precision, there’s a platitude that expresses the feeling associated with viral novels: ‘Everyone’s reading it’.

For the purposes of our project, we have defined viral texts as those works that achieve rapidly accelerating high sales without viral content antecedents. *Harry Potter and the Sorceror’s Stone* would be a prime example. However, by our definition, the second *Harry Potter* book in the series would *not* be considered viral, since its popularity stemmed in part from the first book’s virality. The same goes for sequels in other viral series, such as *The Hunger Games*. The restriction we have imposed on content antecedents also applies to character series, such as Dan Brown’s Robert Langdon books, including *Angels and Demons* (which was not initially viral) and *The Da Vinci Code* (a highly viral sequel to *Angels and Demons* that bolstered its sales long after *Angels and Demons* was released).

# Research Questions

Our overarching questions in this project include the following:

• What does it look like when a book ‘goes viral’? How does sales performance change relative to the book’s earlier sales? How does sales performance change relative to the book’s projected in-genre sales?

• Are there formal features shared across viral texts? Are these features constrained by genre?

• How does a single text operating within an established genre go viral? For example, why did *Twilight* significantly outperform other romances with vampires in 2005?

• Is there a set of lexical or syntactic similarities between viral texts that transcends genre or audience?

Building on methods formalized during the course of the Literary Lab’s previous projects—most recently in ‘Between Canon and Archive’ and previously in projects like ‘Quantitative Formalism’—this study will operationalize and quantify multiple aspects of virality suggested in the above research questions.

# Data

To assess sales performance, we use Nielsen BookScan, which provides in-depth point-of-sale data at major book retailers going back to 2001. For books sold from 1994 to 2001, we use the *New York Times*’ best seller list, which provides relativistic sales performance (i.e., the number-one book on the *NYT* best seller list outsold number two, but the *NYT* does not quantify that differential).

For BISAC codes and other means of categorizing novels in our corpus, we use data provided by the publishing company as well as Library of Congress data for novels not provided by the publisher.

Our corpus was constructed using novels from the publisher’s corpus, as well as ebook versions of viral novels not published by our partner organization.

# Corpus Characteristics

The texts under consideration in this study were originally published between 1994 and 2014. We decided to limit the study to this time period because it includes the transition from traditional brick-and-mortar bookselling to the rise of Amazon (which opened in 1994) as the dominant bookseller. This time period also encompasses the rise of the ebook, another market dominated by Amazon.

We have created several corpora for testing hypotheses related to specific genres, including viral and nonviral comparison versions of each genre corpus. Our nonviral corpora include works that showed average sales performance relative to other works within the same BISAC designation. The ‘one-off’ nature of many viral novels precludes the possibility of constructing a sufficiently large viral corpora to accurately represent certain subgenres.

# Research Methodology

To begin this project, we created a corpus of what we have termed ‘ultraviral’ novels. These works achieved extraordinary success—often but not always from first-time authors—and were sufficiently popular to have movie adaptations made. (In the case of some of the more recent novels, movie adaptations have been optioned and are still in progress.)

The table below shows the novels in our ultraviral corpus, each of which meets the definitional requirements we outlined above:

|  |  |  |
| --- | --- | --- |
| **Title** | **Author** | **Release Date (US)** |
| *The Horse Whisperer* | Nicholas Evans | 6/15/1995 |
| *Left Behind: A Novel of the Earth’s Last Days* | Tim LaHaye, Jerry Jenkins | 12/31/1995 |
| *Cold Mountain* | Charles Frazier | 6/15/1997 |
| *Harry Potter and the Philosopher’s Stone* | J. K. Rowling | 9/1/1998 |
| *Life of Pi* | Yann Martel | 9/1/2001 |
| *The Lovely Bones* | Alice Sebold | 6/15/2002 |
| *The Da Vinci Code* | Dan Brown | 4/1/2003 |
| *The Kite Runner* | Khaled Hosseini | 5/29/2003 |
| *The Time Traveler’s Wife* | Audrey Niffenegger | 6/15/2003 |
| *Eragon* | Christopher Paolini | 8/26/2003 |
| *Twilight* | Stephenie Meyer | 10/5/2005 |
| *Water for Elephants* | Sara Gruen | 5/26/2006 |
| *Diary of a Wimpy Kid* | Jeff Kinney | 4/1/2007 |
| *The Shack* | William Young | 5/1/2007 |
| *The Girl with the Dragon Tattoo* | Stieg Larsson | 6/15/2008 |
| *The Hunger Games* | Suzanne Collins | 9/14/2008 |
| *Divergent* | Veronica Roth | 4/25/2011 |
| *Fifty Shades of Grey* | E. L. James | 6/20/2011 |
| *Gone Girl* | Gillian Flynn | 6/1/2012 |
| *The Goldfinch* | Donna Tartt | 9/23/2013 |

Our research team read these novels to develop hypotheses about their shared features. Initial hypotheses identified worldbuilding, accessible vocabulary, a proliferation of questions, and the fusion of multiple distinct genres in a single work (e.g., a novel that is both a coming-of-age and fantasy story) as potential sites of virality at a formal level.

We are currently testing our ultraviral corpus against comparison corpora with average sales performance. Acknowledging that we cannot be sure we know what to look for across such a heterogeneous sample united by an external factor, we are running a bevy of tests on the ultraviral corpus to identify any extant similarities or patterns. These tests include assessments of age of acquisition for vocabulary; average lengths of sentences, paragraphs, chapters, and books; part-of-speech tagging; and type-token ratio analysis. Later phases of research will include topic modelling and an assessment of narration-to-dialogue ratio. Past research on popular fiction at Stanford has also suggested that virality correlates with a relatively high level of information density when compared with other texts in the same genre (Archer, 2014). We will test this information density hypothesis with our various corpora.

**Note**

1. ‘Virality’, Oxford Dictionaries, http://www.oxforddictionaries.com/us/definition/american\_english/virality.

**References**

**Archer, J.** (2014). Reading the Bestseller: An Analysis of 20,000 Novels*.* PhD dissertation, Stanford University*.*

**Latour, B.** (2005). *Reassembling the Social*. Oxford University Press, Oxford.