Predicting Box Office from the Screenplay: A Text Analytical Approach

Starling David Hunter III
Tepper School of Business, *Carnegie Mellon University,* Pittsburgh PA[1]

Susan Smith
Department of Mass Communication, *American University of Sharjah,* United Arab Emirates

Saba Singh
Department of Interaction Design, *School of Visual Arts,* New York, NY

Abstract: Empirical studies of the determinants of box office revenues have mostly focused on post-production factors, i.e. ones known *after* the film has been completed and/or released. Relatively few studies have considered pre-production factors, i.e. ones known *before* a decision has been made to greenlight a film project. The current study directly addresses this gap in the literature. Specifically, we develop and test a relatively parsimonious, pre-production model to predict the opening weekend box office of 170 US-produced, English-language, feature films released in the years 2010 and 2011. Chief among the pre-production factors that we consider are those derived from the textual and content analysis of the screenplays of these films. The most important of these is determined through the application of network text analysis—a method for rendering a text as a map or network of interconnected concepts. As predicted, we find that the size of the main component of a screenplay's text network strongly predicts the completed film's opening weekend box office.

[1] Corresponding author: starling@andrew.cmu.edu (email address can be published), phone: +1 786 352-8784 (please do not publish phone number)

1. Introduction

In 1983, two highly-influential works were published concerning the business of movie-making. One was a memoir entitled *Adventures in the Screen Trade.* It was authored by William Goldman, a two-time *Academy Award* winner for best screenplay. In said memoir Goldman succinctly summarized the conventional wisdom concerning Hollywood's (in)ability to predict box office success when he quipped that "nobody knows anything" (Goldman, 1983). According to Caves (2000, p. 371), what Goldman meant was that while "producers and executives know a great deal about what has succeeded commercially in the past and constantly seek to extrapolate that knowledge to new projects…their ability to predict at an early stage the commercial success of a new film project is almost nonexistent."

The second influential work was an academic journal article by Barry Litman entitled *Predicting Success of Theatrical Movies: An Empirical Study.* In the first paragraph of that paper Litman acknowledged the conventional wisdom concerning the "uncertainty and unpredictability associated with investments in the motion picture industry" (Litman, 1983, p. 159). To underscore that point he quoted Jack Valenti, then-president of the Motion Picture Association of America (MPAA), who had claimed five years previously that even "with all of its experience, with all the creative instincts of the wisest people in our business, *no one, absolutely no one,* can tell you what a movie is going to do in the marketplace…Not until the film opens in a darkened theatre and sparks fly up between the screen and the audience can you say this film is right" (cf. Valenti, 1978, p.7; italic emphasis in Litman, 1983, p. 159). But as empiricists are wont to do, Litman wondered whether there were "any signposts along the way, which while not guaranteeing success, nevertheless, might prevent one from taking the wrong fork in the road and thus, narrow the range of uncertainty" (Litman, 1983, p. 159). Accordingly, Litman went on to propose and test a predictive model of box office revenues, a statistical model that included the following predictors—adjusted production *costs*, critics' *ratings*, whether the film's *genre* was science-fiction, whether the film was distributed by a *major* or by an independent company, the *date/season* of the film's release, and whether the film was *nominated* for or if it *won* an Academy Award. The statistical measure of fit for the model was exceptionally high, a result that suggested that perhaps it was possible for someone to know something.

In the intervening 30+ years, subsequent empirical research has both confirmed the relevance of Litman's initial model and identified several other important predictors, as well. These include, but are not limited to, whether or not the film is a *sequel* (Sawhney & Eliashberg, 1996), the presence in the film of bankable *star* actors or directors (Neelamegham & Chintagunta, 1999), the film's MPAA *rating* (Wallace, Seigerman, & Holbrook 1993), the number of *screens* on which it appears (Ravid, 1999), and *competitive conditions*, e.g. the number of other films in theaters at the same time and with the same MPAA rating (Karniouchina, 2011). For some samples, combinations of these and other variables have explained in excess of 60% of the variation in box office revenues (e.g. Ravid, 1999; Elberse & Eliashberg, 2003). But no matter how impressive the statistical fit of these and other models, the core of Goldman's and Valenti's complaints are not fully resolved. In part this is because, like all goods and services, film production has a value chain (Eliashberg, Elberse, & Leenders, 2006) and the majority of the predictors listed above are only known in the later stages, i.e. post-production or post-release (Eliashberg, Hui, & Zhang, 2007, 2014). For example, critics cannot review films that haven't yet been produced and most of their reviews aren't penned until after the film has been released into theaters. Similarly, other predictors such as the number of screens on which a film played, the number of awards it received, and the size of its budget might not be known until after the film has been withdrawn from theaters. Knowledge of these predictors is of particularly high value to those in the industry involved in film promotion and distribution and marketing—the later stages of the value chain. But Goldman's and Valenti's concern was with the other end of the value chain, with executives' inability "to predict *at an early stage* the commercial success of a new film project" (Caves, 2000).

And so, despite their significant explanatory power, it can and has been argued that the aforementioned predictors of box office are of minimal use to the "knowledgeable", "wise", "experienced" and "creative" executives and producers whose influence is wrought mostly in these earlier stages. This all matters because there are decisions made in those earlier stages that precede box office success. Among the most important of these is the decision to "greenlight" a film. As Eliashberg, Hui, & Zhang (2007) note, "movie studios often have to choose among thousands of scripts to decide which ones to turn into movies. Despite the huge amount of money at stake, this process—known as green-lighting in the movie industry—is largely guesswork based on experts' experience and intuitions." To date, few studies have attempted to identify and model the influence of any pre-production factors on a film's subsequent box office revenues. One of the few that has is by Goetzman, Ravid, & Sverdlove (2013) who found that the price paid for a screenplay positively predicted box office revenues. Another such study is Eliashberg, Hui, & Zhang's (2014) textual analysis of 300 shooting scripts. They reported that several variables derived solely from an analysis of the scripts significantly predicted the ensuing film's box office revenues.

Like these two, the present study is concerned with predicting box office using early stage variables. And like the latter, in particular, it relies principally upon textual properties of the films' screenplays. What differentiates our study from the latter is the textual analysis strategy we employ. In place of the standard word-frequency approach, we apply network text analysis, a technique for rendering any text as a map or network of interconnected concepts (Carley & Diesner, 2005). As predicted, in a model comprised only of variables known or reasonably inferred during the pre-production phase, we find that the size of a screenplay's text network is a positive and statistically-significant predictor of the subsequent film's box office revenues—a finding that squarely rebuts Goldman's & Valenti's "nobody knows" principle (Caves, 2000; Walls, 2005).

The remainder of this paper is organized as follows. The next section contains the literature review and hypothesis. The third section describes the analytical methods and data that we have employed. The fourth section contains a discussion of the results while in the fifth and final section we discuss the implications of the same.

2. Literature Review

Over the last 30+ years, empirical researchers have identified several predictors of box office revenues. At least eleven predictors have appeared in a dozen or more research studies. In no particular order they are: the film's *genre* (Litman, 1983; Eliashberg, Hui, & Zhang, 2014); whether or not the film is a *sequel* (Litman & Kohl, 1989; Prag & Casavant, 1994; Terry, Butler, & D'Armond, 2005; Nelson & Glotfelty, 2012); the film's *star power*, i.e. whether or not top actors, actresses, and/or directors are associated with the film (Smith & Smith, 1986; Sochay, 1994; Basuroy, Chatterjee, & Ravid, 2003; Ghiassi, Lio, & Moon, 2015); the *date*, timing, or season of the film's release (Litman, 1983; Sawhney & Eliashberg, 1996; Zufryden, 2000; Sharda & Delen, 2006); the quantity and/or quality of *reviews* by film critics (Litman, 1983; Wallace, Seigerman, & Holbrook, 1993; Elberse & Eliashberg, 2003; Goetzman, Ravid, & Sverdlove, 2013); the film's MPAA or other content *rating* (Ravid, 1999; Walls, 2005;

Gopinath, Chintagunta, & Venkataraman, 2013); *awards* or *nominations* received by the film, its director, and/or the actors and actresses appearing therein (Litman & Kohl, 1989; Sochay, 1994; Nelson, Donihue, Waldman, & Wheaton, 2001); the number of *screens*, venues, or theaters in which the film plays (Wallace, Seigerman, & Holbrook, 1993; Neelamegham & Chintagunta, 1999; Zuckerman & Kim, 2003; McKenzie, 2013); the film's total *budget* and/or the budget for promotion and advertising (Litman & Kohl, 1989; Prag & Casavant, 1994; Stimpert, Laux, et al, 2008; Gopinath, Chintagunta, & Venkataraman, 2013); the *market power* of the film's distributor (Litman, 1983; Zuckerman & Kim, 2003); the *competitive conditions* faced by the film at its release and/or during its run in theaters (Litman & Kohl, 1989; Kulkarni, Kannan & Moe, 2012), and most recently the *"buzz"* surrounding the film on social media (Mestyan, Yasseri, & Kertesz, 2013; Kim, Hong & Kang, 2015)

As noted in the introduction, all of these predictors of box office are, for the most part, determined definitively in the latter stages of the value-chain, i.e. after the film is completed and/or released. Comparatively speaking, predictors associated with earlier stages, e.g. development and pre-production, are under-examined (Eliashberg, Hui, & Zhang, 2007). However, two recent studies have given long-overdue attention to them. The first of these is by Goetzman, Ravid, & Sverdlove (2013) who examined whether the prices paid for screenplays are "forward looking", that is to say, whether buyers (studios) "will pay more for screenplays that eventually lead to successful movies" (p. 277). As predicted, they found that price had a significant and positive effect on the completed film's revenues, suggesting thereby that "screenplay buyers make rational economic decisions" *and* that the "prices paid serve as a signal for the perceived quality of the subsequent project" (p. 297).

The second recent and relevant study is by Eliashberg, Hui, & Zhang (2014) who relied upon several textual, content, and genre properties of screenplays to predict box office performance. Unlike their previous work which involved textual analysis of movie spoilers (Eliashberg, Hui, & Zhang, 2007), this study relied on a sample of 300 shooting scripts of films released between 1995 and 2010. They focused on extracting objective information from screenplays because, they tell us, executives and producers are constantly faced with the decision about which of many potential film projects to fund, i.e. which scripts to turn into movies. This is referred to in the industry as the green-lighting decision. And at the point in time when this decision needs to be made, neither the future performance of the potential projects is known nor are any of the "post-production drivers of box-office performance" (p. 2639). Further complicating matters is the fact that while the new conventional wisdom holds that a "movie's story line is highly predictive of its ultimate financial performance" (ibid.), the best current methods of predicting that performance are idiosyncratic, intuitive, and highly dependent upon the comparison sample of scripts. Because objective properties of the screenplay itself are rarely taken into account in these decisions, the goal of their study was to identify a set of text-based measures useful for both comparing screenplays and predicting their performance. Those measures fell into four groups, each at a different level of analysis.

At the higher end was the story's *genre*, i.e. drama, action, comedy, etc. followed by story *content,* e.g. the presence of a surprise ending or the likability of the protagonist. At the lower end of the scale were placed *semantic* features of the text, e.g. the total number of scenes and the average length of dialogs. The fourth and final group of predictors consisted of *bag-of-words* properties of the individual words comprising the script, e.g. the styles and frequencies of individual words in the text. The latter two features were determined using fully-automated, natural language processing (NLP) methods while the first two were determined by human coders. In total, these four groups of parameters contained over three dozen different measures. The two most strongly predictive, in order of influence, were "early exposition" (communicating the general theme of the movie as early as possible) and the presence of a "strong nemesis" in the story. Notably, both of these were *content* features and were identified by human coders. The next two strongest predictors involved the story's *genre*, specifically, whether or not the film was a romance or a thriller. As with the content features, genre was again determined by human coders. The fifth strongest measure was one of the *bag-of-words* features which captured "styles of language" such as contractions, interjections, and the presence of profanity and vulgarity. Notably, none of the six *semantic* variables appeared among the top ten in terms of predictive power. These were the total number of scenes, the percentage of interior scenes, the total number of dialogs, the average length of the dialogs, and their concentration index. Thus, the best script-based predictors in their model were those determined by human coders—*content* and *genre*.

Also notable is fact that the variables that required the greatest amount of computational effort—the *bag-of-words* and the *semantic* factors—were the least predictive. That said, the relatively poor performance of the semantic and lexical measures is not dispositive of their predictive potential. Other measures and methods exist whose efficacy can be examined empirically. Our choice is network text analysis (NTA), a term employed by Diesner & Carley (2005, p. 83) to describe a wide variety of "computer supported solutions" that enable analysts to "extract networks of concepts" from texts and to discern the "meaning" represented or encoded therein. The key underlying assumption of such methods or solutions, they assert, is that the "language and knowledge" embodied in a text may be "modeled" as a network "of words and the relations between them " (ibid).

In short, creating networks from texts has two basic steps. The first involves the assignment of words and phrases to conceptual categories. The second concerns the assignment of linkages to pairs of those categories. Well over a dozen distinct approaches to NTA have been identified in the literature (Diesner, 2012). These include, but are not limited to, *text-based causal maps* (Nadkarni & Narayanan, 2005), *word network analysis* (Danowski, 2009), *map analysis* (Carley & Palmquist, 1992), *conceptual graphs* (Sowa, 1992), *semantic networks* (Nerghes, Lee, Groenewegen, Hellsten, 2014), *centering resonance analysis* (Corman, Kuhn, et al, 2002), *mental models* (Carley, 1997), *knowledge graphs* (Gomez, Moreno, et al 2000), and *morpho-etymological networks* (Hunter, 2014b; Hunter & Singh, 2015).

Several studies in the field of educational psychology have reported a significant relationship between the structural properties of text networks and measures of academic performance. For example, Nadkarni & Narayanan (2005) examined the relationship between two measures of the size of “text-based causal maps” and students’ learning outcomes. Specifically, they reported a positive and significant relationship between the number of concepts and the number links (pairs of concepts) in causal maps abstracted from students’ written case analyses and their subsequent course grades. Carley (1997) compared the cognitive maps of eight project teams, each with 4-6 members, enrolled in an information systems project course at a private university. Each team was required to “analyze a client’s need and then design and build an information system to meet that need within one semester.” Five of these teams were eventually deemed successful and three were not. At three points during the semester, each team was required to provide responses to two open-ended questions—“What is an information system?” and “What leads to information system success or failure?” Their answers were coded and used as data. On average, the cognitive maps of the members of successful groups had significantly more concepts and more pairs of concepts compared to maps by members of non-successful groups. Also, in a study of students exposed to three different instructional methods, Nadkarni (2003) reported that among students with low-learning maturity, those who were exposed to a mix of lecture-discussion and experiential learning had larger mental models than those exposed to just one of those learning modes.

Hunter (2014a) is the only study of which we are aware that examined the relationship between concept map size and performance in the motion picture industry. Specifically, he examined whether the size of text networks could distinguish between two groups of contemporary screenplays. One group consisted of the 75 winners of and nominees for the best original screenplay award given by the *Academy of Motion Picture Arts and Sciences* (aka the *Academy Awards*), as well as the *American Film Critics Associations*, between the years 2006-2012. The second group was comprised of 75 unproduced screenplays randomly-selected from the online portal *Simplyscripts.com* , scripts that were written during the same time period as the award winners and nominees. He reported that the text networks of award winners and nominees were over 33% larger than those of the amateurs, a difference that was highly statistically-significant.

In conclusion, while prior research on the relationship between text network properties and performance is limited, what research exists is unequivocal: text network size is positively and significantly associated with a variety individual-level performance outcomes. While none of the performance examined thus far is financial in nature, our expectation is that the same relationship holds, i.e. that a*ll else equal, the size of the text network of a screenplay will be positively associated with the completed film's box office performance.*

3. Methods and Data

We used the *Box Office Mojo* website to obtain a list of all films released in the US in the years 2010 and 2011. After eliminating all documentaries, foreign-produced and foreign-language films, re-releases/re-issues, films for which no box office revenues were reported, and films whose distribution or release was complicated by legal wrangling, a total of 200 films remained from 2010 and another 206 from the year 2011. We next searched several online databases to find screenplays for these 406 films. These included, but were not limited to, *Simply Scripts* (simplyscripts.com), *Write to Reel* (writetoreel.com), *JoBlo's Movie Screenplays* (www.joblo.com/movie-screenplays-scripts), the *Internet Movie Script Database* (imsdb.com), and *Scriptfly* (www.scriptfly.com). In all we found 170 screenplays in machine readable form.[2] The log-transformed value of the opening weekend box office for the 170 films whose screenplays were found had a significantly higher mean ($p< 0.0001$, 1-tailed) and less than half the variance ($p< 0.0001$, 1-tailed) of the 236 films whose screenplays were not found. These differences are principally due to the fact that screenplays associated with very low budget, independently- or self-produced, and small box-office films were not routinely made available online or for sale. Importantly, our analysis revealed no systematic differences in either the mean or the variance of the log-transformed value of opening weekend box-office among the top 30 and the top 40 films in these two groups of screenplays.

3.1 Dependent variable

Following several recent studies (Simonoff & Sparrowe, 2000; Dellarocas, Zhang, & Awad, 2007; Gemser, Van Oostrom, & Leenders, 2006; Hennig-Thurau, Houston, & Walsh, 2006;

[2] A list of the films whose screenplays were analyzed in this study is available from the corresponding author upon request.

Asur & Huberman, 2010; Hadida, 2010; Terry, King & Walker, 2010; Terry, King & Patterson, 2011; Kulkani, Kannan, & Moe, 2012; Mestyan Yasseri, & Kertesz, 2013; Gopinath, Chintagunta, and Venkataraman, 2013; McKenzie, 2013; Ghiassi, Lio, & Moon, 2015) , we selected opening weekend box office as our dependent variable. The primary justifications for doing so are that first- or opening-weekend box office has been shown to account "for 25% of the total domestic box office gross…(and is thus)… highly predictive for total gross" (Simonoff & Sparrowe, p.15; see also Gmerek, 2015). Secondly, first- or opening-weekend box office is the performance measure most proximal to the green-lighting stage. Additionally, "since competition for movie screens is fierce", opening-weekend revenues constitute the basis upon which "all major decisions pertaining to a film's ultimate financial destiny are made", most notable among them, the decision made by movie theater owners to keep a film running "more than the contractually obligated two weeks" (ibid, p. 19). Gross revenues are affected by such subsequent post-release decisions in a way that opening weekend revenues are not.

All revenue figures were obtained from either the *Box Office Mojo* website or the *International Movie Database* (imdb.com). For the 170 screenplays in our sample, opening weekend box office ranged from a low of $11,083 for *Killer Inside Me* (2010) to a high of $110.3 million for *Toy Story 3* (2010). When we recall Simonoff & Sparrowe's (2000) finding that opening weekend box office typically represents 25% of gross revenues, then our revenue range is comparable to that reported in the two most directly comparable studies, i.e. Eliashberg, Hui, & Zhang (2010, 2014), where the minimum gross revenue figures were $25.7K and $29.9K and the maximums were $424 and $757.5 million, respectively. Further, the average opening weekend box office revenue for our sample was $15.3 while the average gross in theirs was $44.3 million—about three time as much.

3.2 Independent variable

As noted previously, our hypothesis is that the size of a screenplay's text network will be positively and significantly associated with the box office performance of the subsequent film. Well over a dozen distinct families of methods for constructing networks from texts have been developed and applied in the last four decades (Diesner , 2012). They can be distinguished from one another on the basis of a variety of characteristics including the degree of automation involved, whether words are abstracted to higher order categories, and the nature of the relationship used to construct the network. In this study we opted for Hunter's (2014b) morpho-etymological approach, one which is semi-automated, which abstracts words into higher-order conceptual categories defined by common etymology, and which relates those categories based upon their co-occurrence within words known as "multi-morphemic compounds" (MMC). MMCs may include, but are not necessarily limited to, *closed compounds* (briefcase, cowboy, deadline), *copulative compounds* (attorney-client, actor/model), *hyphenated compounds* (open-minded, panic-stricken), *hyphenated multiword expressions* (jack-in-the-box, sister-in-law), *infixes* (un-bloody-believable, fan-blooming-tastic), *abbreviations* and *acronyms* (NASA, FBI, yuppie, radar, laser), and *blend words* (camcorder, motel, guesstimate), as well as selected *clipped words* (internet, hi-fi, sci-fi, e-mail), *open compounds* (post office, ice cream, full moon), and *pseudo-compound words* (understand, overcompensate).

Our first step in creation of the morpho-etymological text networks entailed identifying the MMCs in each screenplay. To accomplish this we used the *Generate Concept List* and the *Identify Possible Acronyms* commands in the CASOS Institute's *Automap* software (Carley & Diesner, 2005). This involved two steps, the first of which was eliminating from further consideration all words in the screenplay that were not MMCs. This was accomplished through the use of a “stop list”, i.e. a self-generated list of words that were previously determined to not be MMCs. Our stop list contained over 50,000 words which we developed for use on this and other research studies (Hunter, 2014b). It included such terms as *toast, apple, monotheism, wallet, pencil, boat, basket, pad, tire,* etc. The next step was to determine which of the remaining words were MMCs. We accomplished this by comparing the remaining words for each screenplay to Hunter’s (2014a) proprietary, Excel database which contains over 30,000 unique MMCs extracted from over 500 contemporary screenplays and teleplays. Approximately 75% of the MMCs in each screenplay were already contained in the database. All remaining words were then checked manually by all three authors with the intent of identifying those MMCs not currently contained in the database.

The next step involved decomposing every MMC in each screenplay into its constituent morphemes. For example, the closed compound *heavyweight* is comprised of two morphemes—*heavy* and *weight*. Next, each morpheme was assigned to a conceptual category defined by its most remote etymological root. Typically, the most remote root was Indo-European, as defined in the 3rd edition of the *American Heritage Dictionary of Indo-European Roots* (Watkins, 2010). That source assigns over 13,000 English words to over 1,300 Indo-European (IE) roots. Over 85% of the individual morphemes in our sample were assigned to IE roots. For example, the closed compound word *middleweight* has two constituent morphemes—middle and weight—which descend from the IE roots **medhyo-,** which means “middle” (Watkins, 2011, p. 53) and **wegh-**, which means “to go, transport in a vehicle” (ibid., p. 98), respectively. Where IE roots of constituent morphemes could not be identified, then etymological roots provided in the *American Heritage Dictionary of the English Language* were used. Most typically these were Latin, Greek, Germanic, or Old English.

After decomposing all MMCs into their constituent morphemes and assigning said morphemes to their etymological roots, the next step was to create a symmetrical matrix for each screenplay where the rows and column labels were the etymological roots associated with all MMCs in the screenplay. Once the matrix was created for each screenplay, the size of the resulting network was calculated using the UCINet software program (Borgatti, Everett, & Freeman, 2002). In social network analysis, the largest cluster of mutually-reachable nodes in a network is referred to as the "main component" (Borgatti, 2006). Our measure of the size of the text network is the number of nodes contained in the main component, *not* the total number of nodes in the network. Figure 1, below, depicts a portion of the main component of the text network constructed from the screenplay for the film *The Fighter* (2010).

→ Insert Figure 1 Here

3.3 Statistical Modeling

We employed an ordinary-least squares (OLS) regression analysis to model the effect of our measure of network size on box office revenues while controlling for whether the film was a drama (DRAMA), whether or not it was rated "R" (MPAA-R), whether the screenplay was original (ORIGINAL), a Likert-scaled variable based on the opening weekend box office of the screenwriter's most recently completed film (RECORD), and whether the film was released in 2011 (Y2011). The independent variable was the log of the number of unique etymological roots in the main component of the text network of the screenplay (LOGSIZE). The dependent variable was the log of the opening weekend box office receipts for each film. Specifically, the OLS model specification was as follows: Log (Opening Weekend Box Office) = α + β_1*LOGSIZE + β_2 * DRAMA + β_3* MPAA-R + β_4 + *ORIGINAL + β_5*RECORD + β_6*Y2011

+ ε. Descriptive statistics and correlations for all variables described above are contained in Table 1 and Table 2, below.

Insert Tables 1 & 2 Here

4. Results

Table 3, below, contains the results of four pairs of regression models used to predict opening weekend box office. They are labeled 1a & 1b, 2a & 2b, 3a & 3b, and 4a & 4b. The first model in each pair establishes the baseline prediction of box office and contains only the five control variables—DRAMA, MPAA-R, ORIGINAL, RECORD, and Y2011. The second model in each pair—the "b" model—adds the independent measure, LOGSIZE, to the baseline model. Each pair of models predicts the relationship between text network size and opening weekend box office under slightly different conditions. The first pair of models—1a & 1b—predicts box office for all 170 screenplays in the sample. The second pair—models 2a & 2b—excludes four outliers, namely the four films whose values on the dependent variable fell more than 2.5 standard deviations below the mean. There were no outliers 2.5 or more standard deviations *above* the mean. The four films on the lower end were *The Killer Inside Me* (2010)*, Jack Goes Boating* (2010)*, Life During Wartime* (2010)*,* and *City Island* (2010). Instead of excluding these four outliers, the third pair of models replaced those four values with a quantity equal to 2.5 standard deviations below the mean. In the fourth and final set of models, the minimum outliers in each quartile were excluded.

Insert Table 3 Here

All four of the "a" models have similar and highly significant goodness-of-fit statistics, i.e. adjusted-R^2 values. More specifically, in models 1a through 4a the adjusted-R^2 values are 32.2%, 32.6%, 32.2%, and 34.5%, respectively. In all of the "b" models the addition of the independent variable—the size of the main component of the text network—increases the model's adjusted-R^2. The increase in adjusted-R^2 ranges from a low of 6.2% (model 2a vs. 2b) to a high of 9.4% (model 4a vs. 4b). In every instance, the coefficient associated with text network size is positive ($0.266 < \beta < 0.330$) and highly, statistically-significant ($p < 0.0001$, 1-tailed; $4.13 < t < 5.31$). Moreover, the strength of size's influence is stronger than that of any other variable in the model, i.e. stronger than genre, rating, originality, track record, and year of release. These results indicate very strong support for our hypothesis, i.e. that text network size is positively associated with box office performance.

Finally, it's worth noting that when screenplays were ranked by the size of the main component alone, only three films in the bottom quartile earned $20 million or more in their opening weekend—*Immortals* (2011), *The Last Exorcism* (2010), and *Date Night* (2010)—while 18 earned less than $250K. In comparison, in the top quartile of network size, sixteen films garnered $20 million or more in the opening weekend while none earned less $250K. In the middle two quartiles the number of films earning over $20 million or less than $250K in the opening weekend were 15 and 4, respectively. That's not much different than the top quartile but very different from the bottom one. Further, we note that there were 13 films whose scores on the five key variables were in the less advantageous condition, i.e. network size was below the median, the story concept was original, the film was R-rated, the genre was drama, and the box office of the screenwriter's most recent prior film was below the median. Among these thirteen films, only one appeared in the top 25% in terms of opening weekend box office—*Immortals* with $31.2 million. None of them appeared in the next lowest quartile. Another two appeared in the next lowest quartile—*J. Edgar* ($10.9 million), *50/50* ($8.4 million).. The remaining ten all appeared in the bottom 25% in terms of opening weekend box office. Specifically those were *Margin Call* ($545K), *Blue Valentine* ($194K), *Win Win* ($146K), *Beginners* ($137K), *Martha Marcy May Marlene* ($134K), *Hesher* ($122K), *Solitary Man* ($95K), *Stone* ($76K), *Take Shelter* ($50K), and *Life During Wartime* ($31K). The last four in this group were found in the bottom decile—bottom 10%—of the sample with regards to box office. Thus, when it came to films whose screenplays had small main components *and* which were R-rated *and* whose concepts were original *and* which were dramas *and* whose writers' last project was sub-par—these films seriously underperformed as a group. Conversely, there were fifteen films whose screenplays whose networks had the opposite values on these five variables. They were just as disproportionately represented at the top as the others were at the bottom. Specifically, nine of them were found in the top 25% of opening weekend box office—*Toy Story 3, Pirates of the Caribbean: On Stranger Tides, Thor, Green Lantern, How to Train Your Dragon*, *Cowboys & Aliens, Dear John, Prince of Persia*, and *The A-Team*—with the first six placing in the top 10%. Of the remaining six films—*I am Number Four, Limitless, The Tourist, Arthur, The Losers*, and *Jonah Hex*—the first three were in the 3rd quartile and the last three were in the 2nd quartile. Not one of the fifteen were found in the bottom 25% of opening weekend performance.

5. Discussion & Conclusion

On the whole, our results are both comparable and complementary to prior research on the drivers of box office performance. Most importantly, they confirm the findings of Eliashberg, Hui, & Zhang (2014), the only other study to examine textual properties of screenplays and the subsequent financial performance of films made from them. Recall that their study found that genre and content variables were the strongest predictors of box office revenues while text-level and semantic variables were less so. Although their study and ours are not directly comparable, there are several points of similarity. First of all, even though we coded for content and genre differently than they, our results are essentially the same. They reported that the romance and thriller genres were positively associated with performance and we found that the drama genre was negatively and significantly associated with performance. Taken together, both studies affirm the long-standing finding that genre matters for box office.

Secondly, they also reported that two content variables—early exposition and strong nemesis—were positive and significant predictors. We controlled for only one aspect of content in our study—whether or not the film had a restricted ("R") rating from the MPAA. As shown above, that rating was negatively and significantly associated with box office performance. Taken together, both studies broadly support another long-standing finding, i.e. that content matters for box office performance.

Fittingly, our study adds the most to the current level of understanding of pre-production drivers of performance through its conceptualization of textual variables. Recall that Eliashberg, Hui &Zhang (2014) found a negative and significant relationship between performance and just one of their bag-of-words measures—the one named "LS2"—which captured aspects of the "style of language in the dialogues" and whose higher values signified the "more prevalent use of vulgarity" (p. 2642). Our study examined just one network-of-words measure—the size of the main component of the text network—and as expected, it positively and very significantly predicted opening weekend box office revenues. Taken together, the results of these two studies affirm that objective properties of the text of a screenplay matter.

All of the above having been said, there are a few caveats concerning this study and its data that should be explicitly noted. First, the data set is relatively small. As Eliashberg, Hui, & Zhang (2014) confirm, coding this kind of data is very time-consuming and labor-intensive. It can't, as of yet, be fully-automated. Our results would certainly be more reliable if the sample was larger and covered more years. Second, the two years of data that we did cover may have been abnormal, i.e. they may not be representative of films and their screenplays in the years before or after. Third, the films whose screenplays we did obtain were, on the whole, more successful at the box office than those whose screenplays we did not find. Moreover, low-budget, independently- and self-produced films were very under-represented. Finally, there may have been many and substantial changes in the screenplays that occurred during the production process. To the degree that we rely on shooting scripts rather than earlier drafts, the chances of this happening become more remote.

There are also a few important practical implications associated with our approach that deserve mentioning. Most importantly, this study establishes a new basis for identifying "comps", i.e. a comparable or benchmark set of screenplays that executives and producers can use during the green-lighting process. Instead of the current practice of relying only on recent films with similar genre and content characteristics, our approach suggests that the set be broadened to include films with similarly-sized main components where size, at its most basic level, indicates the number of distinct concepts that have been brought into relation with one another in the screenplay. Understanding why this quantity matters becomes more apparent when we first recognize that said concepts were brought into relation by one or more screenwriters; they reflect their word choices. Implicit in that recognition is another, one concerning a well-studied construct in personality psychology known as *cognitive complexity*. While definitions and measures of that construct abound, among those most relevant to our purposes are "the degree of differentiation in an individual's construct system" (Bieri, 1966, p. 16), "the number of independent dimensions of concepts that an individual brings to bear in describing a particular domain of phenomena" (Scott, 1962, p. 405), and "the number of independent constructs a person uses in perceiving and interpreting the environment" (Tinsley, Kass, Moreland, & Harren, 1983, p. 94). A large body of empirical research has linked higher levels of cognitive complexity to a wide variety of perceptual and intellectual abilities. For example, Bosgra (2009, p. 176) notes that individuals with higher cognitive complexity have better abilities "…to apply different points of view… to perceive contradictions…to deal with duality…" and that they have "a more complete view of their environment in the sense that they are able to distinguish the important factors that play a role and the relationships among those factors." Burleson (2007 , p. 122) treats cognitive complexity as a "communication skill" and states that those who possess more of it have "more acute social perception" and can "produce more effective messages in challenging circumstances, and appear to process others' messages more deeply." Other published research has associated high cognitive complexity with an enhanced ability to solve "unstructured problems" (Davidson, 1996, p. 219), an "increased tolerance for alternative viewpoints" (Ledgerwood, et al 2006, p. 460), the ability "to balance contradictions, ambiguities, and trade-offs" (Boyacigiller, et al, 2004, p. 83), and the possession of a "more sophisticated" understanding of "people and situations" (Allen, p. 231). If our text analytic approach does capture some aspects of this construct, then our size measure may also indicate the completeness and the internal workings of the world in which the writers' story unfolds, as well as their skill in communicating these things to others.

All that having been said, it's worth noting that one of the smallest text networks in our sample was the one for *Winter's Bone*—and adaptation of a novel by the same name. It was also one of the many in the bottom quartile of size that earned under $250K in the opening weekend. But, while the film almost perfectly conformed to the size-performance relationship predicted in this study, it went on to be highly profitable—ultimately earning almost $14 million worldwide on an estimated budget of $2 million. It was also highly critically-acclaimed, earning nominations for four Academy Awards—Best Motion Picture of the Year, Best Performance by an Actress in a Leading Role, Best Performance for an Actor in a Supporting Role, and Best Writing-Adapted Screenplay. Further, after the film's star, Jennifer Lawrence, won the Oscar for best actress in a leading role, she was lifted out of relative obscurity and right on to the Hollywood A-list.

What we should understand from this example is that it is precisely because films made from screenplays with small text networks have low initial revenue potential that their overall and promotional budgets must be set and managed accordingly, thereby increasing the likelihood that the films will be profitable. Acknowledging this fact could, in turn, have implications for several pre-production and post-production drivers of box office. For example, if a decision is made to produce a screenplay with a small text network —and thus one with low revenue-potential—then one option for managing the budget is to cast relatively unknown actors instead of bankable stars, or to cast bankable stars who will work on the project for substantially less than their standard fee. Similarly, the marketing budgets can be scaled and promotional campaigns focused accordingly. Another possibility is to consider re-writing the script in such a way that its text network is larger—not larger for its own sake, of course, but larger in a way that reflects more cognitive complexity, larger in a way that indicates a deeper, more detailed, and more nuanced understanding of the environment in which the story unfolds. Further research should be undertaken to determine whether there are other properties of a screenplay's text network aside from its size that both convey cognitive complexity and are linked to box office performance.

References

Allen, M. (2002). A synthesis and extension of constructivist comforting research. *Interpersonal communication research: Advances through meta-analysis*, 227-245.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, *67*(4), 103-117.

Bieri, J. (1966). Cognitive complexity and personality development. In *Experience Structure & Adaptability* (pp. 13-37). Springer Berlin Heidelberg.

Borgatti, S. (2006). Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory* 12(1): 21-34.

Borgatti, S., Everett, M. & Freeman, L. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Bosgra, S. (2007). Cognitive Complexity, industry dynamism, and risk-taking in entrepreneurial decision-making in *Entrepreneurial Strategic Decision-making: A Cognitive Perspective* (pp. 175-89). Elgar, Northampton, MA.

Boyacigiller, N., Beechler, S., Taylor, S., Levy, O., Lane, H. W., Maznevski, M. L., et al. (2004). The crucial yet elusive global mindset. *The Blackwell handbook of global management: A guide to managing complexity*, 81-93.

Burleson, B. R. (2007). Constructivism: A general theory of communication skill. *Explaining communication: Contemporary theories and exemplars*, 105-128.

Carley, K. M. (1997). Extracting team mental models through textual analysis. *Journal of Organizational Behavior*, *18*(1), 533-558.

Carley, K., & Diesner, J. (2005). "AutoMap: Software for network text analysis." *CASOS (Center for Computational Analysis of Social and Organizational Systems),* Carnegie Mellon University.

Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, *70*(3), 601-636.

Caves, R. E. (2000). *Creative industries: Contracts between art and commerce* (No. 20). Harvard University Press.

City Island (2010), Wr: Raymond De Felitta, Dir: Raymond De Felitta, USA, 104 mins.

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying Complex Discursive Systems. *Human communication research*, *28*(2), 157-206.

Danowski, J. A. (2009). Inferences from word networks in messages. *The content analysis reader*, 421-429.

Date Night (2010), Wr: Josh Klausner, Dir: Shawn Levy, USA, 88 mins.

Davidson, R. A. (1996). Cognitive complexity and performance in professional accounting examinations. *Accounting Education*, *5*(3), 219-231.

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. Carnegie-Mellon University, Pittsburgh PA. Institute of Software Research International.

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. *Causal mapping for information systems and technology research: Approaches, advances, and illustrations*, 81-108.

Elberse, A., & Eliashberg, J. (2003). Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. *Marketing Science*, *22*(3), 329-354.

Eliashberg, J., Elberse, A., & Leenders, M. A. (2006). The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science*, *25*(6), 638-661.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881-893.

Eliashberg, J., Hui, S., and Zhang, J. (2014). "Assessing Box Office Performance Using Movie Scripts: A Kernel-based Approach." *IEEE Transactions on Knowledge and Data Engineering*, *26(11):2639-2648.*

Fighter, The (2010), Wr: Scott Silver, Paul Tamasay, and Eric Johnson, Dir: David O. Russell, USA, 116 mins.

Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, *42(6): 3176-3193*

Goetzmann, W. N., Ravid, S. A., & Sverdlove, R. (2013). The pricing of soft and hard information: economic lessons from screenplay sales. *Journal of Cultural Economics*, *37*(2), 271-307.

Goldman, W. (1983),Adventures in the Screen Trade: A Personal View of Hollywood and Screenwriting. Warner Books: New York.

Gómez, A., Moreno, A., Pazos, J., & Sierra-Alonso, A. (2000). Knowledge maps: An essential technique for conceptualisation. *Data & Knowledge Engineering*, *33*(2), 169-190.

Gopinath, S., Chintagunta, P. K., & Venkataraman, S. (2013). Blogs, advertising, and local-market movie box office performance. *Management Science*, *59*(12), 2635-2654.

Hunter, S. & Singh, S. (2015). A Network Text Analysis of *Fight Club*. *Theory and Practice in Language Studies, 5(4)*, 737-49.

Hunter, S. (2014a). A Semi-Automated Method of Network Text Analysis Applied to 150 Original Screenplays. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 1(1):68-76.

Hunter, S. (2014b). A Novel Method of Network Text Analysis. *Open Journal of Modern Linguistics*, *4*(2), 350-66.

Immortals (2011), Wr: Charles Parlapanides and Vlas Parlapanides, Dir: Tarsem Singh, USA, 110mins.

Jack Goes Boating (2010) Wr: Robert Glaudini, Dir: Phillip Seymour Hoffman, USA, 89 mins.

Karniouchina, E. V. (2011). Impact of star and movie buzz on motion picture distribution and box office revenue. *International Journal of Research in Marketing*, *28*(1), 62-74.

Killer Inside Me (2010), Wr: John Curran, Dir: Michael Winterbottom, USA, 109 mins.

Kim, T., Hong, J., & Kang, P. (2015). Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting*, forthcoming.

Kulkarni, G., Kannan, P. K., & Moe, W. (2012). Using online search data to forecast new product sales. *Decision Support Systems*, *52*(3), 604-611.

Last Exorcism, The (2010), Wr: Huck Botko and Andrew Gurland, Dir: Daniel Stamm, USA, 87 mins.

Ledgerwood, A., Chaiken, S., Gruenfeld, D. H., & Judd, C. M. (2006). Changing minds: Persuasion in negotiation and conflict resolution.

Life During Wartime (2010), Wr: Todd Solondz, Dir: Todd Solondz, USA, 98 mins.

Litman, B. R. (1983). Predicting success of theatrical movies: An empirical study. *The Journal of Popular Culture*, *16*(4), 159-175.

Litman, B. R., & Kohl, L. S. (1989). Predicting financial success of motion pictures: The'80s experience. *Journal of Media Economics*, *2*(2), 35-50.

McKenzie, J. (2013). Predicting box office with and without markets: Do internet users know anything?. *Information Economics and Policy*, *25*(2), 70-80.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, *8*(8), e71226.

Nadkarni, S. (2003). Instructional methods and mental models of students: An empirical investigation. *Academy of Management Learning & Education*, *2*(4), 335-351.

Nadkarni, S., & Narayanan, V. K. (2005). Validity of the structural properties of text-based causal maps: An empirical assessment. *Organizational Research Methods*, *8*(1), 9-40.

Neelamegham, R., & Chintagunta, P. (1999). A Bayesian model to forecast new product performance in domestic and international markets. *Marketing Science*,*18*(2), 115-136.

Nelson, R. A., & Glotfelty, R. (2012). Movie stars and box office revenues: an empirical analysis. *Journal of Cultural Economics*, *36*(2), 141-166.

Nelson, R. A., Donihue, M. R., Waldman, D. M., & Wheaton, C. (2001). What's an Oscar worth?. *Economic Inquiry*, *39*(1), 1-6.

Nerghes, A., Lee, J. S., Groenewegen, P., & Hellsten, I. (2014). The shifting discourse of the European Central Bank: Exploring structural space in semantic networks. In *Tenth International Conference on Signal-Image Technology and Internet-Based Systems, 10*(1)*,* 447-455.

Prag, J., & Casavant, J. (1994). An empirical study of the determinants of revenues and marketing expenditures in the motion picture industry. *Journal of Cultural Economics*, *18*(3), 217-235.

Ravid, S. A. (1999). Information, Blockbusters, and Stars: A Study of the Film Industry*. *The Journal of Business*, *72*(4), 463-492.

Sawhney, M. S., & Eliashberg, J. (1996). A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, *15*(2), 113-131.

Scott, W. A. (1962). Cognitive complexity and cognitive flexibility. *Sociometry*, 405-414.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, *30*(2), 243-254.

Smith, S. P., & Smith, V. K. (1986). Successful movies: A preliminary empirical analysis. *Applied Economics*, *18*(5), 501-507.

Sochay, S. (1994). Predicting the performance of motion pictures. *Journal of Media Economics*, *7*(4), 1-20.

Sowa, J. F. (1992). Conceptual graphs as a universal knowledge representation. *Computers & Mathematics with Applications*, *23*(2), 75-93.

Stimpert, J. L., Laux, J. A., Marino, C., & Gleason, G. (2011). Factors influencing motion picture success: Empirical review and update. *Journal of Business & Economics Research (JBER)*, *6*(11).

Terry, N., Butler, M., & De'Armond, D. (2005). The determinants of domestic box office performance in the motion picture industry. *Southwestern Economic Review*, *32*(1), 137-148.

Terry, N., King, R., & Patterson, R. (2011). Vampires, Slashers, Or Zombies: Opening Weekend's Favorite Box Office Monster. *Journal of Business & Economics Research (JBER)*, *9*(2).

Terry, N., King, R., & Walker, J. J. (2010). The Determinants of Box Office Revenue for Horror Movies. *Journal of Global Business Management*, *6*(2), 10.

Tinsley, H. E., Kass, R. A., Moreland, J. R., & Harren, V. A. (1983). A longitudinal study of female college students' occupational decision making. *Vocational Guidance Quarterly*, *32*(2), 89-102.

Toy Story 3 (2010), Wr: Michael Arndt, Dir: Lee Unkrich, USA, 103 mins.

Valenti, J. (1978). *Motion Pictures and Their Impact on Society in the Year 2001*. Midwest Research Institute.

Wallace, W. T., Seigerman, A., & Holbrook, M. B. (1993). The role of actors and actresses in the success of films: How much is a movie star worth?. *Journal of Cultural Economics*, *17*(1), 1-27.

Walls, W. D. (2005). Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns. *Journal of Cultural Economics*, *29*(3), 177-190.

Watkins, C. (Ed.). (2000). *The American Heritage Dictionary of Indo-European Roots*. Houghton Mifflin Harcourt.

Zuckerman, E. W., Kim, T. Y., Ukanwa, K., & von Rittmann, J. (2003). Robust Identities or Nonentities? Typecasting in the Feature-Film Labor Market. *American Journal of Sociology*, *108*(5), 1018-1073.

Zufryden, F. (2000). New film website promotion and box-office performance. *Journal of Advertising Research*, *40*(1/2), 55-64.