Predicting Box Office from the Screenplay: An Empirical Model

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.

Keywords:

box office

screenwriting

screenplay

text analysis

network analysis

content analysis

[1] Corresponding author: starling@andrew.cmu.edu (email address can be published),

AUTHOR BIOGRAPHIES

Starling David Hunter III was born in California, USA and received his Ph.D. in Organization Theory from *Duke University's* Fuqua School of Business in 1999. He is currently an Associate Teaching Professor at *Carnegie Mellon University's* Tepper School of Business in Doha, Qatar. His research interests include organizational and text network analysis. Formerly he was a faculty member in the School of Business Administration at the *American University of Sharjah* in the United Arab Emirates and the Sloan School of Management at the *Massachusetts Institute of Technology* in Cambridge, Massachusetts.

Susan Smith is a documentary film-maker who was born in Oklahoma, USA. In 1994 she received her Master's in Visual Anthropology from the *University of Southern California's* (USC) School of Cinematic Arts. She is currently employed as a professor at the *American University of Sharjah* in the United Arab Emirates where she teaches courses in film and documentary studies.

Saba Singh was born in New Dehli, India. In 2014 she received her Bachelors of Science in Business Administration with a concentration in marketing from the Tepper School of Business at *Carnegie Mellon University* in Doha, Qatar. She is currently employed as a business coordinator at the 60Degrees Creative Advertising Agency in Doha, Qatar. In the fall of 2015 she will begin work on an MFA in Department of Interaction Design at the *School of Visual Arts* in New York City.

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 "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 seven independent predictors—adjusted production *costs*, critics *ratings*, the film's *genre*, whether the film was distributed by a *major* company or an independent company, the *date/season* of the film's release, and whether the film was nominated for or 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 variables, 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 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 plays, the number of awards it receives, and the size of its budget are all only known well after the film is

released. 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 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 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 of predictors of box office revenue. 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 & 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 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 hard 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 "hard" information 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 were *bag-of-words* properties of the individual words comprising the script, e.g. 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 of the story 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 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 defined in the study 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 the 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* (Bakker, 1987), and *morpho-etymological networks* (Hunter, 2014b; Hunter & Singh, 2015). Studies in educational psychology (EP) have reported a significant relationship between the structural properties of text networks and measures of academic performance (Carley, 1997; Nadkarni, 2003; Nadkarni & Narayanan, 2005). Hunter's (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 screenplay portal *Simplyscripts.com* written during the same time period as the award winners and nominees. He reported that the size of the "morpho-etymological" text networks of award winners and nominees were over 33% larger than those of the latter, 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 directly associated with a variety of measures of performance. 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 (Boxofficemojo.com) 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. Of these 406 films, 92 were for very low-budget, independently- or self-produced films released in only a few theaters and/or which earned less than $10K in their opening weekend. These were eliminated from further consideration.

We next searched several online databases to find screenplays for the remaining 314 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. The appendix contains a list of the titles of all films whose screenplays were analyzed in this study. The log-transformed value of the opening weekend box office for the 170 films whose screenplays were found had a higher mean (6.77 vs.

5.72, p< 0.0001, 1-tailed) and less than half the variance (0.87 vs. 1.83, p< 0.0001, 1-tailed) of the 144 films whose screenplays were not found.

3.1 Dependent variable

Following Nelson & Glotfelty (2012), who note Hollywood's increasing reliance on a "strong theatrical opening" (p.142), we selected opening weekend box office as our dependent variable. 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). The average opening weekend box office revenue for the sample was $15.3 million with standard deviation of $16.8 million. The median box office value was $10.9 million, meaning that half earned above that amount and half earned less. As with all prior research, the highly skewed nature of these returns necessitated that these values be log-transformed prior to use in the regression models detailed below.

3.2 Independent variable

As noted previously, our independent variable is the size of the text network for each of the 170 screenplays contained in the sample. Following Hunter (2014b, 2015), we opted for the construction of morpho-etymological networks which are constructed solely from a text's multi-morphemic compounds (MMCs). MMCs 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* (CIA, FBI, yuppie, radar, laser), *blend words* (camcorder, motel, guesstimate),

selected *clipped words* (internet, hi-fi, e-mail), and some affixed-compounds (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 CASOS Institute's *Automap* software to generate a word list for each screenplay (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. 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 20,000 unique MMC 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 an MMC 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 MMC *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. Down the first column of Table 1, below, is listed 20 of the MMCs identified in the screenplay of the *The Fighter* (2010), an *Academy Award* nominee for Best Original Screenplay in 2010. According to the *International Movie Database,* the film was based on the "the early years of boxer 'Irish' Micky Ward and his brother who helped train him before going pro in the mid-1980s." Columns 2-4 contain the etymological roots associated with each constituent morpheme. For example, regarding the MMC *middleweight*, Column 2 and 3 contains the IE roots ***medhyo-*** and ***wegh-*** respectively. The 20 MMCs were associated with 21 different etymological roots, 19 of which were Indo-European and defined by Watkins (2011) as follows: ***bha-2*** (to speak); ***bhoso-*** (naked); ***denk-*** (to bite); ***dheue-*** (to close, finish, come full circle); ***g[e]n-*** (to compress into a ball); ***kap-*** (to grasp); ***legwh-*** (light; having little weight); ***medhyo-*** (middle); ***ne-*** (not); ***oi-no-*** (one; unique); ***se-2*** (long; late); ***sent-*** (to head for; go); ***(s)kel-1*** (to cut); ***(s)ker-3*** (to turn; bend); ***tkei-*** (to settle, dwell. be home); ***ud-*** (up; out); ***wegh-*** (to go, transport in a vehicle); ***wel-3*** (to turn, roll); and ***wi-ro-*** (man).

Table 1: Twenty Multi-Morphemic Compounds Appearing in the Screenplay of *The Fighter* and Their Corresponding Etymological Roots

Table 2, below, is a symmetrical matrix—a socio-matrix—whose row and column labels are the names of the 21 aforementioned etymological roots. A "1" in a cell of the table indicates that the two corresponding roots co-occur in the same MMC, as with **medhyo-** and **wegh-** in the MMC *middleweight*. Because several morphemes appear in one or more MMCs—for example, the root ***wegh-*** is associated with seven of them while the root ***oi-no-*** is associated with six—several roots can be interconnected.

Table 2: Socio-matrix for the Twenty-One Etymological Roots

Figure 1, below, is a network map of the 21 etymological roots. They are connected by a linkages or ties which represent the MMCs associated with the associated roots. Notably, several of the MMCs represented in the network emphasize the boxing theme of the film, for example, *WBU* (World Boxing Union), *light-heavyweight, knock-down*, *knock-out*, *bare-knuckle, bare-chested, middleweight, once-tough*, *ringside*, *lightweight*, *welterweight*, and *outweighed*.

Insert Figure 1 Here

Once a socio-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 is referred to as the "main component" (Borgatti, 2006). In Figure 2, below, is displayed the entire morpho-etymological text network for *The Fighter*. The main component is encircled and appears on the right side of the graph. Scattered through the network are several smaller clusters of connected nodes—clusters of 2 to 10 nodes. 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.

Insert Figure 2 Here

3.3 Statistical Modeling

We employed an ordinary-least squares (OLS) regression analysis to model the effect of text 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), the opening weekend box office of the screenwriter's most recently completed film (RECORD), and whether the film was released in 2011 (Y2011). The sole independent variable was the natural 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 natural log of the opening weekend box office receipts for each film. Specifically, the OLS model specification was as follows: Log (Opening Weekend Box Office) = $\alpha + \beta_1$*LOGSIZE + β_2 * DRAMA + β_3* MPAA-R + β_4 + *ORIGINAL + β_5*RECORD + β_6*Y2011 + ε

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. Note that it 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—predict 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. Those films were *The Killer Inside Me* (2010),

Jack Goes Boating (2010), *Life During Wartime* (2009), and *City Island* (2009). There were no outliers 2.5 or more standard deviations above the mean. 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 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 the text network size's influence is stronger than that of any other variable in the model, i.e. genre, rating, originality, track record, and year of release. These results spell very strong support four our hypothesis, i.e. that text network size is positively associated with box office 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, it confirms 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 that 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 for the most part, significantly associated with performance. Taken together, both studies affirm a long-standing finding—that genre matters. Secondly, Eliashberg, Hui, & Zhang (2014) also reported that early exposition and strong nemesis were positive and significant predictors. We controlled for only one aspect of content—whether the film had a restricted ("R") rating from the MPAA. As shown above, that rating was negatively and significantly associated with performance. Taken together, both studies support another long-standing finding—that content matters for box office performance.

Fittingly, where our study adds the most to the current level of understanding of pre-production drivers of performance is in its conceptualization of textual variables. Recall that Eliashberg, Hui &Zhang (2014) found a negative and significant relationship between just one of their bag-of-words measures—the one named "LS2"—which captured aspects of the "style of languages in the dialogues" and whose higher values signified "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 predicted, it positively and very significantly predicted opening weekend box office performance. 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. The 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. it may not be representative of films and screenplays in the years before or after. Third, the films whose screenplays we did obtain were more successful at the box office than those that we did not find. So again, our sample may not be truly representative even of the years upon which we focused our efforts. Finally, there may have been many and substantial changes in the screenplays that occurred in 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 and practical implications associated with our approach and related findings that deserve mentioning. First, 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 similar network structural characteristics. Recall that in our data set, when screenplays were ranked by the size of the main component, 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, 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 $20 million or more 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.

What these results suggest is that this method may be best at raising red flags during the greenlighting phase, i.e. in flagging scripts that have a very low probability of earning large sums at the box office and/or that are of low quality. If so, then it means that the method may help to reduce left tail risk, i.e. the probability of downside outliers, without curtailing right-tail or upside potential. This is precisely what Litman (1983, p. 159) referred to in his seminal paper when he wondered whether the empirical approach might identify "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."

That said, it's worth recalling that one of the smallest text networks in our sample was for *Winter's Bone*—and adaptation of a novel by the same name. It was also one of the many in the bottom quartile that earned under $250K in the opening weekend. And yet, it was nominated for four Academy Awards—Best Motion Picture of the Year, Best Performance by and 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 onto the Hollywood A-list. What we should understand from this example is that because films made from screenplays with small text networks may have low initial revenue potential, 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 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.

Finally, aside from size itself, another class of network-of-words measures may also predict box office performance—measures like betweenness and degree centrality (Freeman, 1977)—both of which indicate the positions and roles of nodes in a network. Several recent studies have confirmed that more influentially-positioned concepts in text networks appear to be more thematically central or relevant. For example, Grbic, Hafferty, & Hafferty (2013) analyzed the semantic structure of mission statements of 132 US medical schools and found that the most centrally-positioned concepts in the resulting text networks included the terms *leader*, *biomedical*, *health*, *community*, and *research*. Nerghes, et al's (2014) semantic network analysis of European Central Bank (ECB) press releases before, during, and after the recent financial crisis revealed that MMCs like *MRO* (main refinancing operation), *LTRO* (long-term refinancing operations), and *interest rate* were among the influentially positioned. And in the area of film studies, Hunter & Singh (2015) analyzed the screenplay of *Fight Club* and found words and concepts evoking the themes of gender, social and individual identity, capitalism, and anarchism to be more centrally located in the network. In the present study's sample, we found anecdotal evidence suggesting that text networks of screenplays vary substantially—and maybe systematically—regarding the words and concepts that are most centrally located. For example, in the text network for the comic-book adaptation *Jonah Hex* (2010), two of the three most influentially-positioned nodes in the main component of its text network were the Indo-European roots **bhel-3**, from which descends the word *blood*, and **skei-** from which descends the word *shit.* Associated with these two very influentially-positioned nodes were fifteen MMCs either beginning with *blood* (blood-splattered, blood-crazed, blood-caked, bloodlust, blood-red, blood-

soaked, & blood-thirsty), or *shit* (shit-faced, shit-fire, shit-load, & shit-smelling), or referring animal feces itself (bird-shit, dog-shit, cow-shit, & horseshit). No other screenplay in the top quartile of the sample had so-much profanity, let alone had it so central in its text network. That *Jonah Hex* did might have had some bearing on its relative under-performance at the box office—only a $10.9 million opening weekend versus $19.8 million for the ten other films whose screenplay text networks were larger. Future research should examine whether the most influentially positioned words and concepts in the screenplay text network are systematically aligned with a story's core themes and the extent to which that alignment also predicts box office performance.

Appendix: Screenplays of Films Included in the Study Sample

1. 50/50
2. Abduction
3. Adjustment Bureau, The
4. After.Life
5. American, The
6. Arthur
7. A-Team, The
8. Atlas Shrugged: Part I
9. Back-up Plan, The
10. Bad Teacher
11. Battle: Los Angeles
12. Beastly
13. Beginners
14. Black Swan
15. Blue Valentine
16. Book of Eli, The
17. Bounty Hunter, The
18. Bridesmaids
19. Brooklyn's Finest
20. Burlesque
21. Cars 2
22. Cedar Rapids
23. Change-up, The
24. Charlie St. Cloud
25. City Island
26. Conan the Barbarian
27. Conspirator, The
28. Conviction
29. Cop Out
30. Cowboys & Aliens
31. Crazy, Stupid, Love.
32. Darkest Hour, The
33. Date Night
34. Dear John
35. Descendants, The
36. Despicable Me
37. Diary of a Wimpy Kid
38. Dilemma, The
39. Dinner for Schmucks
40. Drive
41. Drive Angry
42. Easy A
43. Eat Pray Love
44. Everything Must Go
45. Expendables, The
46. Extraordinary Measures
47. Extremely Loud & Incredibly Close
48. Faster
49. Fighter, The
50. Flipped
51. For Colored Girls
52. Fright Night
53. Frozen
54. Get Him to the Greek
55. Get Low
56. Going the Distance
57. Good Old Fashioned Orgy, A
58. Green Lantern
59. Greenberg
60. Hall Pass
61. Hereafter
62. Hesher
63. Horrible Bosses
64. Hot Tub Time Machine
65. How Do You Know
66. How to Train Your Dragon
67. Hugo
68. I Am Number Four
69. Ides of March
70. Immortals
71. It's Kind of a Funny Story
72. J. Edgar
73. Jack Goes Boating
74. Jonah Hex
75. Just Go With It
76. Just Wright
77. Kids Are All Right, The
78. Killer Inside Me, The
79. Killers
80. Knight & Day
81. Larry Crowne
82. Last Exorcism, The
83. Last Song, The
84. Leap Year
85. Legion
86. Letters to Juliet
87. Life as We Know It
88. Life During Wartime
89. Limitless
90. Lincoln Lawyer
91. Little Fockers
92. Losers, The
93. Lottery Ticket
94. Love and Other Drugs
95. MacGruber
96. Machete
97. Machine Gun Preacher
98. Margin Call
99. Marmaduke
100. Martha Marcy May Marlene
101. Mechanic, The
102. Megamind
103. Middle Men
104. Morning Glory

105. My Soul to Take
106. Next Three Days, The
107. Nightmare on Elm Street
108. No Strings Attached
109. Other Guys, The
110. Our Idiot Brother
111. Pariah
112. Piranha 3D
113. Pirates of the Caribbean: On Stranger Tides
114. Please Give
115. Predators
116. Priest
117. Prince of Persia
118. Prom
119. Rabbit Hole
120. Ramona and Beezus
121. Red
122. Remember Me
123. Rise of the Planet of the Apes
124. Roommate
125. Rum Diary, The
126. Runaways, The
127. Salt
128. Scream 4
129. Season of the Witch
130. Secretariat
131. Shark Night 3D
132. She's Out of My League
133. Shutter Island
134. Sitter, The
135. Skyline
136. Smurfs
137. Social Network, The
138. Solitary Man
139. Something Borrowed
140. Somewhere
141. Sorcerer's Apprentice, The
142. Stone
143. Straw Dogs
144. Super
145. Super 8
146. Switch, The
147. Take Shelter
148. Thor
149. Tourist, The
150. Tower Heist
151. Town, The
152. Toy Story 3
153. Tron Legacy
154. True Grit
155. Twelve
156. Twilight Saga: Eclipse
157. Unstoppable
158. Valentine's Day
159. Very Harold & Kumar 3D Christmas, A
160. Wall Street: Money Never Sleeps
161. War Horse
162. Warrior
163. Water for Elephants
164. We Bought a Zoo
165. When in Rome
166. Win Win
167. Winter's Bone
168. Wolfman, The
169. You Again
170. Young Adult

References

Bakker, R. R. (1987). *Knowledge Graphs: representation and structuring of scientific knowledge*.

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.

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.

Burt, R., Jannotta, J., & Mahoney, J. (1998). Personality correlates of structural holes. *Social Networks*, *20*(1), 63-87.

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 (2009), 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.

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.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35-41.

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.

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

Grbic, D., Hafferty, F. W., & Hafferty, P. K. (2013). Medical school mission statements as reflections of institutional identity and educational purpose: A network text analysis. *Academic Medicine*, *88*(6), 852-860.

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.

Life During Wartime (2009), 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.

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.

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.

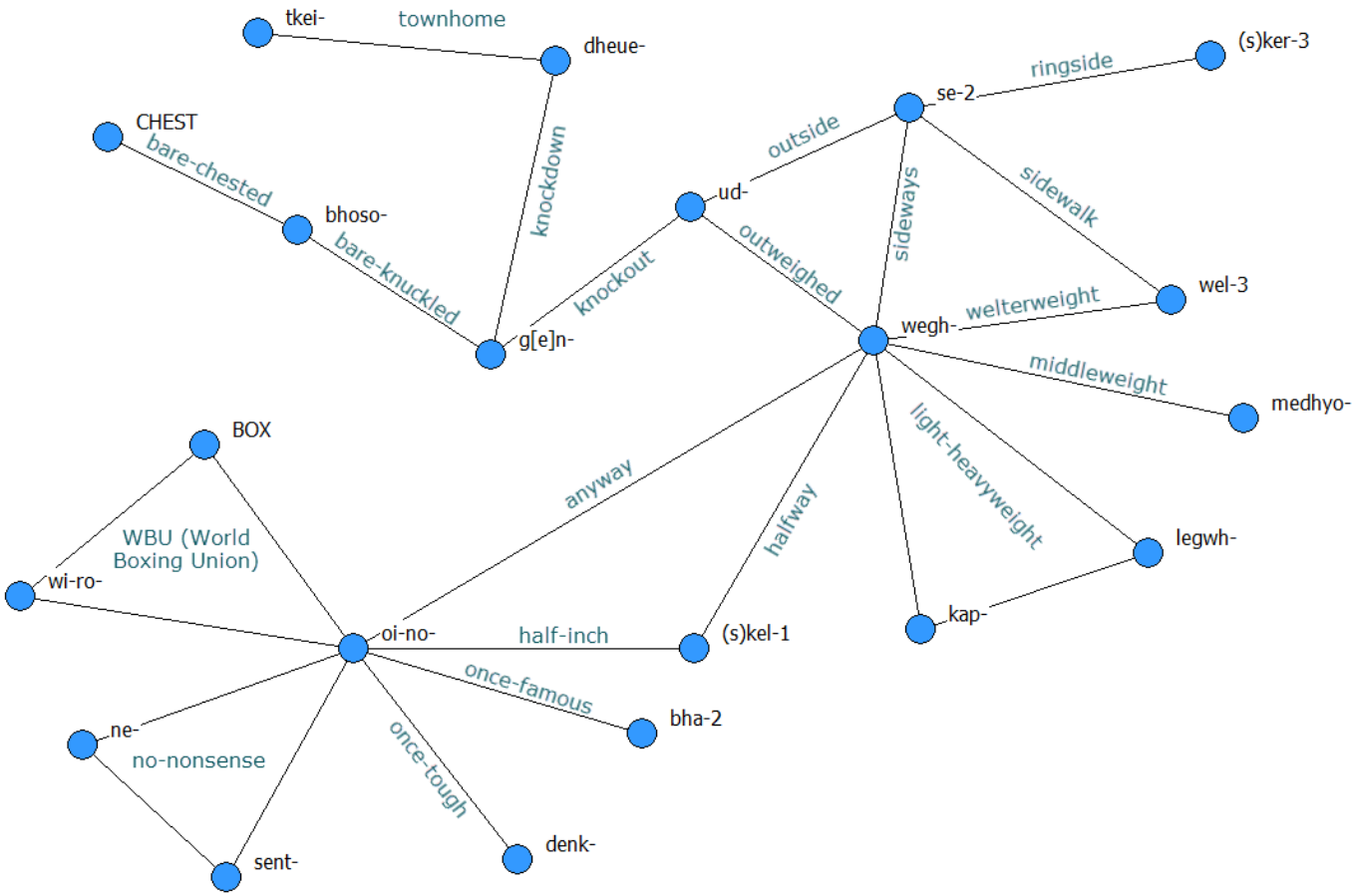

Figure 1 A Segment of the Morpho-Etymological Text Network for the Screenplay of *The Fighter*

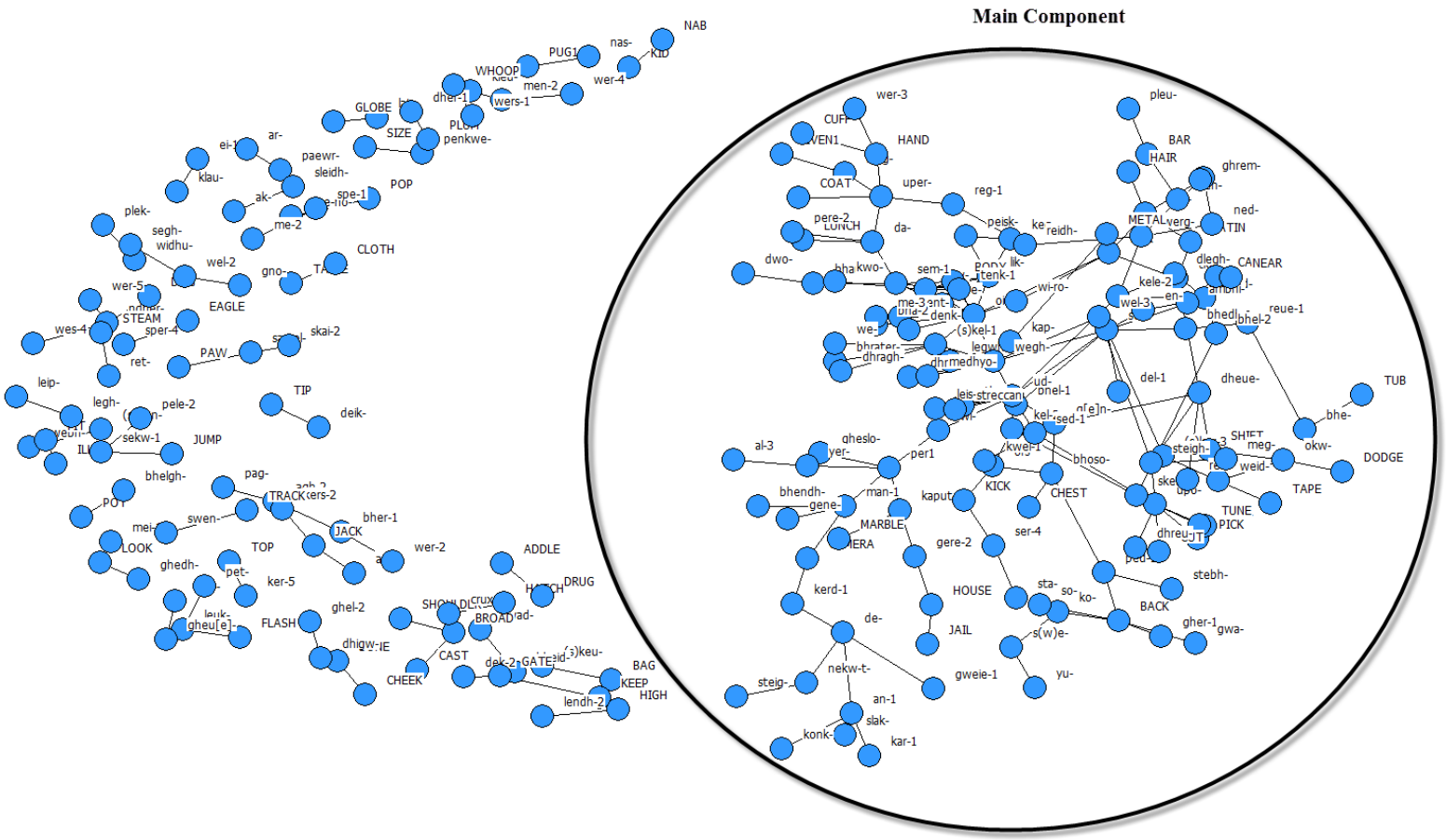

Figure 2 **Main Component and Sub-components of the Morpho-Etymological Text Network of *The Fighter***

Table 1: Twenty Multi-Morphemic Compounds Appearing in the Screenplay of *The Fighter* and Their Corresponding Etymological Roots

Multi-Morphemic Compound	Root 1	Root2	Root3
1. anyway	oi-no-	wegh-	
2. bare-chested	bhoso-	CHEST	
3. bare-knuckle	bhoso-	g[e]n-	
4. half-inch	(s)kel-1	oi-no-	
5. halfway	(s)kel-1	wegh-	
6. knock-down	g[e]n-	dheue-	
7. knockout	g[e]n-	ud-	
8. light-heavyweight	legwh-	kap-	wegh-
9. middle-weight	medhyo-	wegh-	
10. no-nonsense	ne-	oi-no-	sent-
11. once-famous	oi-no-	bha-2	
12. once-tough	oi-no-	denk-	
13. outside	ud-	se-2	
14. outweighed	ud-	wegh-	
15. ringside	(s)ker-3	se-2	
16. sidewalk	se-2	wel-3	
17. sideways	se-2	wegh-	
18. townhome	dheue-	tkei-	
19. WBU (World Boxing Union)	wi-ro-	BOX	oi-no-
20. welter-weight	wel-3	wegh-	

Table 2: Socio-matrix for the Twenty-One Etymological Roots

	wegh-	tkei-	BOX	oi-no-	ne-	wi-ro-	sent-	legwh-	kap-	(s)kel-1	(s)ker-3	se-2	bhoso-	CHEST	g[e]n-	dheue-	ud-	medhyo-	bha-2	denk-	wel-3
wegh-	0	0	0	1	0	0	0	1	1	1	0	1	0	0	0	0	1	1	0	0	1
tkei-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
BOX	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
oi-no-	1	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0
ne-	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wi-ro-	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
sent-	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
legwh-	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
kap-	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
(s)kel-1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(s)ker-3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
se-2	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1
bhoso-	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
CHEST	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
g[e]n-	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0
dheue-	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ud-	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
medhyo-	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bha-2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
denk-	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wel-3	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Table 3 Results of Multiple Regression Analysis

	1a	*1b*	*2a*	*2b*	*3a*	*3b*	*4a*	*4b*
LOGSIZE		**0.282****** (4.43)		**0.266****** (4.13)		**0.281****** (4.40)		**0.330****** (5.31)
DRAMA	**-0.282****** (-4.24)	**-0.237***** (-3.71)	**-0.300****** (-4.47)	**-0.260****** (-4.02)	**-0.281****** (-4.23)	**-0.236***** (-3.70)	**-0.326****** (-4.93)	**-0.281****** (-4.54)
MPAA-R	**-0.298****** (-4.45)	**-0.260****** (-4.06)	**-0.297****** (-4.40)	**-0.267****** (-4.12)	**-0.297****** (-4.43)	**-0.259****** (-4.05)	**-0.307****** (-4.63)	**-0.258****** (-4.16)
ORIGINAL	**-0.109*** (-1.64)	**-0.083#** (-1.31)	**-0.144*** (-2.13)	**-0.114*** (-1.76)	**-0.115*** (-1.73)	**-0.088#** (-1.40)	**-0.089#** (-1.36)	**-0.056** (-0.91)
RECORD	**0.261****** (3.92)	**0.214***** (3.35)	**0.234***** (3.49)	**0.195**** (3.01)	**0.260****** (3.90)	**0.213***** (3.33)	**0.242***** (3.67)	**0.183**** (2.95)
Y2011	**0.102#** (1.58)	**0.109*** (1.78)	**0.057** (0.86)	**0.069** (1.10)	**0.100#** (1.54)	**0.107*** (1.73)	**0.122*** (1.90)	**0.137*** (2.31)
R^2	34.2%	41.2%	34.7%	41.0%	34.2%	41.2%	36.4%	46.0%
Adj-R^2	32.2%	39.1%	32.6%	38.8%	32.2%	39.0%	34.5%	43.9%
Δ Adj-R^2	--	6.9%	---	6.2%	--	6.8%	--	9.4%
Mean-squared Error	0.554	0.497	0.466	0.424	0.545	0.490	0.499	0.427
% Change in MSE		-10.3%		-9.0%		-10.1%		-14.4%
F-statistic	17.03****	19.06****	16.99****	18.42****	17.02****	19.00****	18.46****	22.68****
N	170	170	166	166	170	170	167	167

Legend: LOGSIZE = the natural log of the number of nodes in the main component of the text network of the screenplay; DRAMA is a dummy variable coded 1 if *Box Office Mojo* classified it thus; MPAA-R is a dummy variable coded 1 if the Motion Picture Association of America (MPAA) gave the film a restricted (R) rating; ORIGINAL is a dummy variable coded 1 if the film is not a remake, adaptation, or sequel; RECORD is an eight-level Likert-scale variable based on the opening weekend box office of the screenwriter's most recent film; all p-values 1-tailed, # = $p < 0.10$; * = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$; **** = $p < 0.0001$