# A Semi-Automated Method of Network Text Analysis Applied to 150 Original Screenplays

**Starling David Hunter III**
Carnegie Mellon University
Tepper School of Business
starling@andrew.cmu.edu

## Abstract

In this paper I apply a novel method of network text analysis to a sample of 150 original screenplays. That sample is divided evenly between unproduced, original screenplays (n = 75) and those that were nominated for Best Original Screenplay by either the Academy of Motion Picture Arts & Sciences or by major film critics associations (n = 75). As predicted, I find that the text networks derived from unproduced screenplays are significantly less complex, i.e. they contain fewer concepts (nodes) and statements (links). Unexpectedly, I find that those same networks are more cohesive, i.e. they exhibit higher density and coreness.

## 1 Introduction

Diesner & Carley (2005, p. 83) employ the term *network text analysis* (NTA) 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, emphasis added). A second important assumption is that the position of concepts within a text network provides insight into the meaning or prominent themes of the text as a whole.

Broadly considered, creating networks from texts has two basic steps: (1) the assignment of words and phrases to conceptual categories and (2) the assignment of links to pairs of those categories. Approaches to NTA differ with regard to how these steps are performed, as well as to the level of automation or computer support, the linguistic unit of analysis (e.g. noun or verbs), and the degree and basis of concept generalization. In the social sciences, several studies in the last two decades have linked the structural properties of text networks to measures of individual, group/team, and organizational performance (Nadkarni & Narayaran, 2005). The quantitative empirical literature on this topic can be divided into two groups or streams—educational psychology (EP) and managerial and organizational cognition (MOC). The former typically links structural properties of text networks abstracted from documents like exams and case analyses to academic performance and learning outcomes. The latter abstracts text networks from reports generated by firm's managers, e.g. letters to shareholders and 10-K filings, and links those properties directly or indirectly to firm performance.

Across both streams, the structural properties of networks that have been examined fall into three broad categories—measures of *complexity* or size, measures of *cohesion* or connectedness, and measures of *centrality* or concentration. Another point of consensus concerns the underlying relationships from which the text networks are constructed. Most of the quantitative and empirical studies have relied upon logical relationships among concepts in documents for that purpose. These relationships include, but are not limited to, dependence, chronology, similarity, functionality, causality, and composition (Popping, 2003, pp. 94-5). The second and less commonly used type of relationship involves the co-occurrence of concepts within a user-defined window (e.g. Carley, 1997). Notably, grammatical and lexical relationships have received no attention in the empirical literature. However, Hunter (in press) recently described a "novel", semi-automated method of network text analysis whereby multimorphemic compounds (e.g. abbreviations, acronyms, blend words, clipped words, and compound words) in a text are linked via shared etymological roots. He applied that method to sample of seven recent winners of the Academy Award for Best Original Screenplay and found that the most centrally-positioned words in five of the seven networks corresponded very closely to the themes contained in the films' synopses



*Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 68–76,
Baltimore, Maryland USA, 27 June 2014. 

found on Wikipedia, IMDb and Rotten Tomatoes.

This study represents the first application of Hunter's method to a sample of screenplays of sufficient size to permit multivariate statistical analysis. The specific aim of the study is to examine the relationship between text networks' properties and performance outcomes. To that end, I herein develop and test two falsifiable hypotheses concerning that relationship on a sample of 150 contemporary screenplays—half winners and nominees of major awards and the other half unproduced screenplays obtained from two online screenplay portals. Consistent with the prior literature I find that the more favorably rated screenplays—i.e. the award winners and nominees—have significantly larger text networks than the unproduced ones. Unexpectedly, I find that text networks of these screenplays exhibit significantly lower cohesiveness, i.e. lower density and coreness.

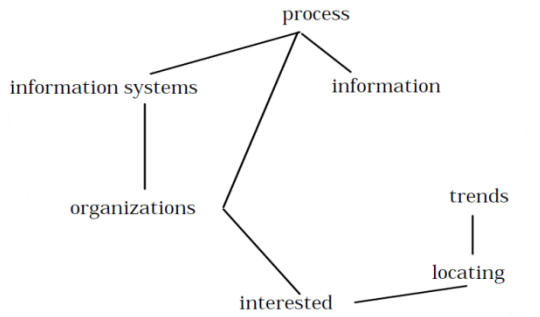
The remainder of this paper is organized as follows. In section 2, *Theory & Hypotheses*, I summarize the relevant social science literature on text network properties and performance and formulate two hypotheses concerning that relationship. In the third section, *Data & Methods*, I describe the data set and the method for constructing the text networks for each screenplay in the sample. In the fourth section, *Results & Discussion*, I report the level of statistical support found for each hypothesis and discuss the implication of the results for current and future research in this area.

## 2 Theory & Hypotheses

Figure 1, below, is adapted from Carley (1997) and it is typical of many network representations of texts. The network itself was constructed from the following two sentences: "Organizations use information systems to handle data. Information is processed by organizations who are interested in locating behavioral trends."

Several things about the network are noteworthy. First, observe that there are seven *concepts* depicted as nodes in the network, each of which appears only once. They are "organizations", "information systems", "process", "information", "interested", "locating", and "trends." Second, see that there are also seven *statements*, i.e. pairs of concepts: (1) "information systems" and "process" (2) "information systems" and "organizations" (3) "process" and 'information" (4) "process" and "organizations" (5) "interested" and "organizations" (6) "interested" and "locating" and (7) "locating" and "trends." Third, note that the *map* itself is comprised of the network formed by all seven *statements*. Typically, the analyst must read some or all of the *statements* in a *map* in order to extract the meaning of the text as a whole. In this regard, it is then notable that the seven *concepts* are implicated in varying numbers of *statements*. Specifically, the *concepts* labeled "organization" and "process" are found in three *statements* while all other *concepts* are found in either two or one.

Figure1: A Simple Text Network
(adapted from Carley, 1997)




In the social science literature, the most widely-investigated structural property of text networks are the number of concepts and the number of links between pairs of concepts. For example, Calori, Johnson & Sarnin (1994) studied the moderating effects of "environmental complexity", i.e. the scope of the organization as measured by the number of distinct businesses and geographic segments, on the relationship between the "cognitive complexity of the chief executive" and firm performance. One of their measures of cognitive complexity was the number of concepts abstracted from interviews with each CEO about their firm's environment. They hypothesized that cognitive maps of CEOs of more diverse firms had more "comprehensive", i.e. larger, cognitive maps than CEOs of more focused firms. This hypothesis was NOT supported. However, they also hypothesized that cognitive maps of CEOs in firms with greater international geographic scope would contain more concepts. This hypothesis was supported.

Nadkarni (2003, p. 336) employed the term "comprehensiveness" to refer to the "number of concepts in a mental model." In a study of students exposed to three different instructional methods, he hypothesized and found (1) significant differences in the comprehensiveness of the

mental models of students of student across methods and (2) greater comprehensiveness in said models among students with low-learning maturity who were exposed to a "hybrid" method of instruction, i.e. a mix of lecture-discussion and experiential learning.

Nadkarni & Narayaran (2005) examined the relationship of two measures of "complexity"—the number of concepts and the number of statements—on learning outcomes. Specifically, they reported a positive relationship between the number of concepts and links found in "text-based causal maps" abstracted from students' written case analyses and their course grades.

Carley (1997) compared the mental models 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 statements (links) compared to maps by members of non-successful groups. In light of the aforementioned studies, the first hypothesis (H1) is that *network complexity, measured as the number of concepts and/or links, is positively related to performance.*

As a class, measures of network *cohesion* indicate the degree to which the nodes in a network are connected to one another. Common measures of cohesion include, but are not limited to, density, fragmentation, connectedness, average path distance, and diameter (Borgatti, Everett, and Freeman, 2002). But while many such measures exist, very few empirical studies have directly examined the linkage between the cohesion in text networks and measures of performance. One such study is Nadkarni & Narayaran's (2005) aforementioned analysis of text-based causal maps abstracted from business case studies. They hypothesized and found network density—measured as the ratio of the number of links to the number of *possible* links—to be positively related to three measures of academic performance—test grades, case analysis grades, and class participation scores.

A second such study is Bodin's (2012) investigation of "university physics student's epistemic framing when solving and visualizing a physics problem using a particle-spring model system" (p. 1). In that study, concept networks were developed from two sets of interview transcripts where students described the task and (physics) problem they were about to solve, as well as their planned strategies for solving the problem. An analysis of networks drawn prior to and right after completion of the assignment revealed a 24% increase in the number of concepts, a 71% increase in the number of links, and 12% increase in network density. While all of these quantities were in the predicted direction, no statistical significance was indicated. Still, the existing empirical evidence suggests network density is positively related to performance. And because various network cohesion measures are closely related conceptually—and can be strongly correlated, as well (Borgatti, Everett, &Johnson, 2014)—then it is more appropriate to phrase the second hypothesis (H2) in more general terms, i.e. *that network cohesion is positively related to performance.*

## 3 Methods & Data

As indicated in the preceding section, the empirical literature has been focused on two kinds of texts—student assignments and firm reporting—and two kinds of performance—grades and financial performance. But there is nothing inherent in these network text analytic methods that limits investigation to the texts mentioned above. Nor has any of the research reviewed indicated otherwise. That said, a number of specific rationales motivated the selection of screenplays, in general, and original screenplays in particular. First, screenplays are highly structured texts, both logically and temporally, with the three-act structure in screenwriting being a prime example (Field, 1998). Second, there exists a large, widely-read, and broadly-disseminated body of knowledge concerning the theory and best practice of screenwriting (e.g. Snyder, 2005; McKee, 2010; Field, 2007). Third, screenplays are carefully evaluated by many interested parties on numerous dimensions, not the least of which are commercial success and artistic merit (Simonton, 2005; Pardoe & Simonton 2008). Finally, the performance of their authors is discrete and quite unambiguous: more than 15,000 screenplays are registered in the US each year with the Writer's Guild of America but fewer than 700 get "green-

lighted" and are subsequently produced (Eliashberg, Elberse, & Enders, 2006). Further, those screenplays that do get "green-lighted" either garner awards or critical acclaim or they do not (Simonton, 2004, 2005).

Somewhat surprisingly, textual analyses of screenplays are relatively rare when compared to analyses of other literary forms such as novels, plays, and poetry. The only studies of which I am aware that links textual variables of screenplays to performance are those by Elishaberg, Hui, & Zhang (2007, 2014) whose kernel-based approach to the study of 300 movies released between 1995 and 2010 significantly predicted Return on investment, i.e. box office revenues as a percentage of budget. The present study represents the first attempt to link textual measures of screenplays to a non-financial-related performance measure.

Screenplays contained in the sample were obtained from a variety of sources. The oldest and most prestigious awards in American cinema are the Academy Awards, aka the "Oscars" (Osborne, 1989) and several studies have been done explaining their artistic and commercial importance (e.g., Krauss, Nan, Simon, et al, 2008; Lee, 2009; Simonton, 2004). Academy Award nominated and winning screenplays are routinely studied by aspiring screenwriters (New York Film Academy, 2014) and widely available online either for free (Simply Scripts, 2014) or purchase (Script Fly, 2014). Winners and nominees of other awards are often available online, as are the screenplays of films which garner no particular artistic acclaim. There are, as well, numerous online forums, websites, and blogs devoted to their discussion and analysis. Moreover, the screenplays for award-nominated, award-winning, and critically-acclaimed films are usually made available by their producers or studios during the award season, but not all of them remain so. In this study, the "produced" or high-performing sample of screenplays are of two kinds. The first consists of nominees and winners of the Academy Award for Best Original Screenplay. Five screenplays are nominated each year making for a potential sample of 40 screenplays. However, two screenplays by Woody Allen—*Blue Jasmine* and *Midnight in Paris*—were not available. Another five nominees whose films were all or partially in foreign-languages were also excluded—*Pan's Labyrinth* (Spanish), *Amour* (French), *A Separation* (Farsi), *Babel* (Arabic, English, Spanish, and Japanese), and *Letters from Iwo Jima* (Japanese). Thus there were 32 remaining Academy Award nominated screenplays for films released in the years 2006-2013.

Another fifty-two (52) screenplays were nominated in the years 2006-13 for Best Original Screenplay by the 32 regional members of the American Film Critics Association, e.g. the New York, Washington D.C., and San Francisco Film Critics Circles. Several of these were not commercially or otherwise available. These include *Upstream Color*, *The Tree of Life*, *Frances Ha*, *World's End*, *Sound of My Voice*, *United 93*, and *Stranger than Fiction*. *Toy Story 3* was excluded because, while an original screenplay, it was part of a film franchise. The South African film *Black Book* was excluded, as well, because it was not in English. The remaining 43 screenplays were obtained. Thus there was a total of 75 screenplays contained in the produced and thus "high-performing" category.

Another 75 unproduced screenplays were randomly selected from two online screenplay databases—*Simply Scripts* and *Trigger Street Labs*. The former hosts pages within its site titled "Unproduced Scripts" where screenwriters are invited to upload their screenplays. Trigger Street Labs is a portal maintained by actor Kevin Spacey's Trigger Street Productions. It allows writers to post original short stories, short films, and screenplays. Thirty-eight (38) screenplays posted between January 1, 2006 and December 31st, 2013 and between 100 and 140 pages were randomly selected from both sites. One was then selected at random and eliminated, making the total number of unproduced screenplays seventy-five (75).

Diesner (2012) outlines four steps for the creation of a text network—(1) Selection (2) Abstraction (3) Relation and (4) Extraction. The first step involves identification of those words that will be subjected to subsequent analysis and the elimination of those that will not. Following Hunter (in press), this stage involved retention of all multi-morphemic compounds comprised of two or more free (unbound) morphemes. These included, but were not limited to, closed and hyphenated compound words, clipped words, blend words or portmanteau, and all acronyms, anacronyms, abbreviations, and initialisms.

Also included were selected instances of conversion, certain prefixes and suffixes, plus selected multi-word compounds and infixes, Examples are shown in Table 1 below. And though it may seem otherwise, this is no random grouping. Rather, they comprise a well-defined, inter-

related set that is extensively-studied in the field of morphology. Specifically, they all belong to the branch of morphology known as word-formation, the study of creation of new or "novel" words principally through changes in their form (Wisniewski, 2007).

Because no existing text mining software selects these groups words from a text, the process for identifying them was only semi-automated with the help of a software program called Automap 3.0.10 (Carley, 2001-2013).

Table 1: Examples of 12 Types of Novel Words in the Sample

| **Type** | **Examples** |
|---|---|
| ***Compounding*** *>Closed Compounds* | briefcase, cowboy, deadline, handcuffs, inmate |
| ***Compounding*** *>Copulative compounds* | attorney-client, actor/model |
| ***Compounding*** *> Open Compounds* | post office, fire alarm |
| ***Compounding*** *>Hyphenated Compounds* | open-minded, panic-stricken, tree-lined |
| ***Compounding*** *> Multi-word Compounds* | Over-the-top, jack-in-the-box, sister-in-law |
| ***Derivation*** *> Affixation> Prefix* | understand, overdrive, overhand, underhanded |
| ***Derivation*** *> Affixation> Suffix* | awesome, hardware, software, clockwise |
| ***Derivation*** *> Affixation> Infix* | *Un*bloody*believable*, *fan*blooming*tastic* |
| ***Derivation*** *> non-Affixation> Abbreviations, Acronyms* | DMV, MTV, FBI, VCR, Yuppie, radar, scuba, laser |
| ***Derivation*** *> non-Affixation> Blend Words* | medevac, motel, guesstimate, camcorder, helipad |
| ***Derivation*** *> non-Affixation> Clipped Words* | Internet, hi-fi, email, slo-mo, vid-com |
| ***Derivation*** *> non-Affixation> Conversion* | eyeball; photoshop |

The process was as follows. First the screenplay was converted to a text file and uploaded to Automap. After removing single letters, extra spaces, and spurious characters, two routines were run within Automap—*Identify Possible Acronyms* and *Concept List.* The former routine identified and extracted all words that were capitalized. Several of these turned out to acronyms or abbreviations. The latter routine used was *Concept List (Per Text)* which generated a list of all unique words for each text. Excluded from consideration were all proper nouns (Green Zone, Hollywood), place and organization names (South Pole, Scotland Yard, Burger King), product names (Land Rover), holidays (New Year's Eve, Christmas Eve), as well as any other word or phrase connoting a specific person, place, or thing through capitalization. Also eliminated were all instances of screenplay and film jargon, e.g. ECU (extreme close-up), off-screen, VO (voice-over) and POV (point of view), as well as multi-word exclamations and interjections such as good night, goodbye, OMG (oh my God), etc.

The second of the four steps of constructing a text network involves abstraction of the selected multi-morphemic compounds to higher-order concepts. In this study, each of the free (unbound) morphemes in each compound was assigned to its etymological root, typically the Indo-European, Latin, or Greek (Watkins, 2011).

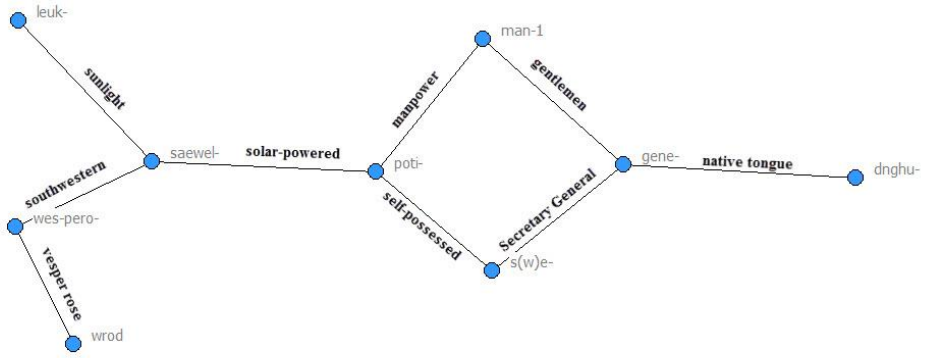
By definition, from every etymological root descends or originates at least one word, otherwise it is not a root. That relationship is genitive, i.e. a relational case typically expressing source, possession, or partition. It is hierarchical and directed—from the root (parent) to word (descendent). Thus, in the third step of network construction, two or more etymological roots were linked or related when words (free morphemes) descending from them co-occurred within the same word, as the following examples demonstrates.

Consider a text that contains the following nine words: the closed compound words *manpower*, *sunlight*, *southwestern*, and *gentlemen*; the open compounds *vesper rose*, and *native tongue*; the hyphenated compound *solar-powered self-possessed*; and the proper noun *Secretary General*. As shown in Table 2, below, these words are all multi-morphemic compounds, each element of which descends from two different etymological roots.

Table 2: Selected Indo-European Roots and their Derivatives (Watkins, 2011)

| **Roots (definition)** | **Selected Derivatives** |
|---|---|
| **wes-pero-** (evening) | West, Visigoth, vesper |
| **wrod-** (rose) | rose, julep, rhodium |
| **dnghu-** (tongue) | tongue, language, linguist |
| **leuk-** (light, brightness) | light, lux, illumination, lunar, luster, illustrate, lucid, |
| **man-1** (man) | man, mannequin, mensch |
| **poti** (powerful; lord) | possess, power, possible, potent, and pasha |
| **saewel** (the sun) | sun, south, solar, solstice |
| **gene-** (to give birth) | gender, general, gene, genius, engine, genuine, gentle, pregnant, nation, native. |
| **s(w)e-** (self) | self, suicide, secede, secret, secure, sever, sure, sober, sole, idiom, and idiot. |

Figure 2: A Text Network Based on Etymological Relationships among Selected Multi-morphemic Compounds Contained in Table 1




Recall that a *statement* in NTA is comprised of two concepts and the relationship that links them. In Figure 2, above, each of these words appears on the link between the two etymological roots—the concepts—that co-occur within the word. Put another way, the relationship is the co-occurrence of two different etymological roots in the same multi-morphemic compound or multi-word expression—co-occurrence in what is essentially a window of one word. For example, the etymological roots **gene-** (to give birth; beget) and **man-1** (man) are linked by their co-occurrence in the compound word *gentleman*. Taken together, that word and those two roots comprise a statement. And as shown below, it is possible to construct an entire map or network from these interconnected statements. Specifically, that network is comprised of eight concepts—namely, the Indo-European roots ***dnghu-, gene-, man1, s(w)e-, poti-, saewel-, leuk, wrod-*** and ***wes-pero***—and the nine multi-morphemic compounds—*native tongue*, *gentleman*, *Secretary General*, *self-possessed*, *manpower*, *solar power*, *sunlight*, *vesper rose*, and *southwestern*.

A similar approach was used to constructing text networks for each of the 150 screenplays in the sample. Specifically, after matching all of the above classes of multi-morphemic compounds to their corresponding etymological roots, all pairs of roots for each screenplay were converted into a symmetrical matrix which was then uploaded into version 6.487 of the UCINet software program (Borgatti, Everett, and Freeman, 2002). Text networks were then generated using version 2.118 of the NetDraw software program embedded in UCINet. Figure 3, below, is depicted main component of the text network for the screenplay of *Zero Dark Thirty*. which was nominated for Best Original Screenplay in 2012. The main component is the largest group of mutually-reachable nodes in a network. Note that the node labels are etymological roots, typically Indo-European (Watkins, 2011) In the case of words with non-Indo-European roots, the base form of the component of the multi-morphemic compound is used.

The fourth stage involves the extraction of meaning from the completed text network. But since the investigation of meaning is not a part of this analysis, it is excluded from further consideration. See Hunter (in press) for a detailed discussion and examples. Table 3, below, summarizes some basic statistics and network metrics for the 150 screenplays in the sample.

Table 3 Summary Statistics (n =150)

| **Variable** | **Mean** | **Range** |
|---|---|---|
| Words (000's) | 20.9 | 9.3 - 36.2 |
| Genre = Comedy Only | 0.17 | 0 - 1 |
| Concepts/Nodes | 176 | 84 - 320 |
| Statements/Pairs of Nodes | 173 | 72 - 337 |
| Density | 1.2% | 0.66 - 2.50% |
| Core-Periphery | 1.7% | 0.40 - 3.10% |
| Normalized Degree | 0.32 | 0.14 - 1.06 |
| Network Centralization | 1.7% | 0.83 - 4.43% |

## 4 Results & Discussion

Recall that the first hypothesis (H1) proposed that network complexity was positively related to performance. Following the prior literature, complexity was measured as both the number of concepts in a network, i.e. the number of unique etymological roots, and as the number of statements, i.e. the number of pairs of concepts.. Because these values were very highly correlated ( $\rho = 0.98$, $p < 0.0001$) their respective z-scores were averaged to obtain a single value for complexity. Table 4a, below, presents the results of a multinomial regression of screenplay genre, word count, and network complexity on screenplay type. The positive coefficients on complexity indicate that, as predicted, text networks of screenplays of winners and nominees of Academy Awards ($\beta = 0.574$, $p < 0.0001$) and critics' awards ($\beta = 0.359$, $p < 0.01$) have significantly greater complexity than text networks of unproduced screenplays. The Nagelkerke, Cox & Snell, and McFadden pseudo-R2 values were 30.6%, 26.7%, and 15.0%, respectively. Thus, H1 is strongly supported.

Table 4a: Multinomial Regression of Genre, Word Count, and Network Complexity on Type of Screenplay

| Category | Variable | Estimate |
|---|---|---|
| **Critic's Awards** | Genre = Comedy | -3.775 |
| | SQRT(Words/1000) | -0.114 |
| | Complexity | **0.359**** |
| **Academy Awards** | Genre = Comedy | -0.315* |
| | SQRT(Words/1000) | 0.011 |
| | Complexity | **0.574**** |

The second hypothesis (H2) predicted that network cohesion would be positively related to performance. In this study cohesion was measured by density and coreness (the degree to which the network is characterized by a tightly interconnected core and a much less tightly connected periphery).

Because these two values were highly correlated ($\rho = 0.55$, $p < 0.001$), the z-scores for each measure were averaged to obtain a single value for cohesion. Table 4b presents the results of a multinomial regression of screenplay genre, word count, and network cohesion on screenplay type. The negative and significant value of the coefficients indicates that cohesion is negatively associated with performance, the exact opposite of what was predicted.

Specifically, cohesion of text networks derived from screenplays in the Academy ($\beta = -0.870$, $p < 0.0001$) and critics award ($\beta = -0.413$, $p < 0.001$) categories is significantly lower than that for unproduced screenplays. That means they typically have both lower density and/or a less core-periphery type structure. The Nagelkerke, Cox & Snell, and McFadden pseudo-$R^2$ values were 36.8%, 32.1%, and 18.7%, respectively.

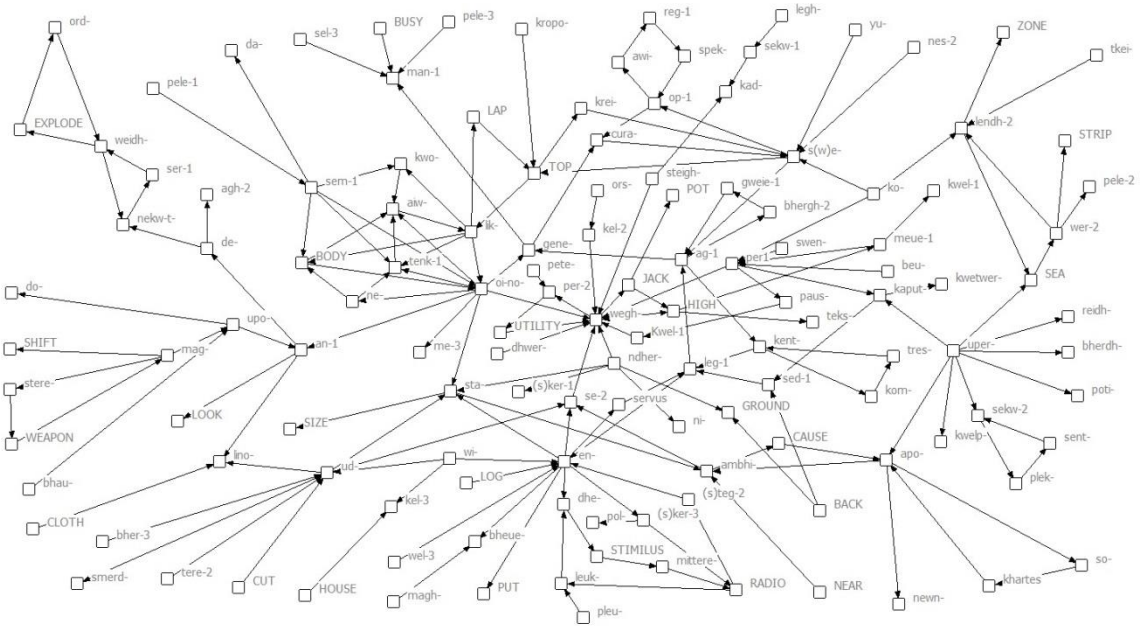
Taken together, the results suggest that text networks derived from original screenplays selected by the Academy and by film critics have very different structural properties than text networks derived from unproduced screenplays. In short, the former are larger, yet held together by proportionately fewer linkages.

Table 4b: Multinomial Regression of Genre, Word Count, and Network Cohesion on Type of Screenplay

| Category | Variable | β |
|---|---|---|
| **Critic's Awards** | Genre = Comedy | -3.809 |
| | SQRT(Words/1000) | -0.139 |
| | Cohesion | **-0.413***** |
| **Academy Awards** | Comedy | -0.395** |
| | SQRT(Words/1000) | -0.078 |
| | Cohesion | **-0.870****** |

The reason for this disparity may well have to do with the size of the networks under examination. In both the educational psychology and the managerial and organizational cognition literatures, the typical size of the networks is about 1/6 that of those examined here. Recall that as a network grows, the number of *possible* connections grows exponentially. As such, density becomes smaller at an exponential rate. In this study, the text networks derived from both sets of screenplays had concept to statement ratios of close to unity. Thus, given that the award winners and nominees had much larger text networks, it follows logically that those networks were also much less dense.

Figure 3: Main Component of the Text Network of *Zero Dark Thirty*




In closing, it is important to recognize three important limitations of the current study. Firstly, the sample size is relatively small—only half the size of that found in the only other study of screenplays that includes textual analysis, i.e. Eliashberg, Hui, & Zhang (2014). The sample is also not ideally constructed. Rather than comparing award winners to unproduced screenplays, a better sample would include entries from a screenplay competition and the study design would attempt to select finalists and winners from that sample. Winners of top competitions are very frequently optioned and eventually produced. Alternatively, another approach would be to increase sample size and to include screenplays which garnered no critical or artistic awards. Even though many such screenplays can be located, finding a representative sample of them is difficult because they are not uniformly available. Still, if that hurdle can be overcome, the sample could be used to study box-office revenues, potentially improving upon the current understanding of screenplays' contribution to a film's financial success.

A third limitation of this study concerns the labor-intensive nature of the coding. At present, the process outlined above is only semi-automated. Unless and until advanced computational methods of text analysis can be developed and applied, sample sizes will remain small and the coding prone to human error.

Finally, recall that the fourth stage of network text analysis involves the extraction of meaning from the network itself. Typically, this involves a study of the most central nodes (concepts), as well as the statements (links) associated with them. For example, in the case of Zero Dark Thirty the most influential node or concept, as measured by network constraint, is the Indo-European root ***wegh-*** which means "way" and from which descends such words as way and weight. One of the multi-morphemic compounds associated with that concept in the Figure 3, above, is "hijacker" which is a clipped word, the full form being "highwayjacker" (Online Etymological Dictionary, 2014). Other multi-morphemic compounds associated with highly influential nodes in the network include SEAL, CIA, WMD, QRF (Quick Reaction Force), and NSA. As noted by Hunter (in press), other award-winning and critically-acclaimed screenplays exhibit the same pattern, i.e. words associated with influential nodes are also closely related to key themes of the story. Anecdotally, the unproduced screenplays give less evidence of this tendency. Future research might examine whether this difference is systematic. To that end, all data involved in this study—the text of all 150 screenplays and all of the network coding—will be made freely available to interested and qualified researchers upon request.

## References


Bodin, M. (2012). Mapping university students' epistemic framing of computational physics using network analysis. *Physical Review Special Topics-Physics Education Research*, *8*(1), 1-14.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). Ucinet for Windows: Software for social network analysis.

Calori, R., Johnson, G., & Sarnin, P. (1994). CEOs' cognitive maps and the scope of the organization. *Strategic Management Journal*, *15*(6), 437-457.

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

Carley, K.M. (2001-13), Automap 3.0.10. Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Diesner, Jana, (2012). Uncovering and Managing the Impact of Methodological Choices for the Computational Construction of Socio-Technical Networks from Texts. *Dissertations.* Paper 194.

Eliashberg, Jehoshua, Anita Elberse, and Mark AAM Leenders (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. K., & Zhang, Z. J. (2014, forthcoming). Assessing Box Office Performance Using Movie Scripts: A Kernel-based Approach. *IEEE Transactions on Knowledge and Data Engineering*.

Field, S. (2007). *Screenplay: The foundations of screenwriting*. Random House LLC.

Hunter, S. (2014, forthcoming). A Novel Method of Network Text Analysis. Open Journal of Modern Linguistics.

Krauss, J., Nann, S., Simon, D., Gloor, P. A., & Fischbach, K. (2008). Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis. *Proceedings of the 16th European Conference on Information Systems*, 2026-2037.

Lee, F. L. (2009). Cultural discount of cinematic achievement: the academy awards and US movies' East Asian box office. *Journal of Cultural Economics*, *33*(4), 239-263.

McKee, R. (2010). Story: Substance, Structure, Style and the Principles of Screenwriting, Harper Collins, New York.

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.

Nadkarni, S., & Narayanan, V. K. (2007). Strategic schemas, strategic flexibility, and firm performance: the moderating role of industry clockspeed. *Strategic management journal*, *28*(3), 243-270.

*New York Film Academy: Bachelor of Fine Arts in Screenwriting*. (2014). Retrieved from http://www.nyfa.edu/bfa/screenwriting.php

Osborne, R., & Davis, B. (1989). *60 years of the Oscar: the official history of the Academy Awards*. New York: Abbeville Press.

Pardoe, I., & Simonton, D. K. (2008). Applying discrete choice models to predict Academy Award winners. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 375-394.

Popping, R. (2003). Knowledge graphs and network text analysis. *Social Science Information* 42(1):91-106.

*Script Fly*. (2014). Retrieved from http://scriptfly.com

*Simply Scripts*. (2014). Retrieved from http://simplyscripts.com

Simonton, D. K. (2004). Film awards as indicators of cinematic creativity and achievement: A quantitative comparison of the Oscars and six alternatives. *Creativity Research Journal*, *16*(2-3), 163-172.

Simonton, D. K. (2005). Film as art versus film as business: Differential correlates of screenplay characteristics. *Empirical Studies of the Arts*, *23*(2), 93-117.

Snyder, B. (2005). *Save the Cat*. Michael Wiese Productions.

Watkins, C. (2011), The American Heritage Dictionary of Indo-European Roots, 3rd Edition, Houghton Mifflin Harcourt, Boston MA.

Wisniewski, K. (2007). Word formation. http://www.tlumaczenia-angielski.info/linguistics/word-formation.htm