PREDICTING NIELSEN RATINGS FROM PILOT EPISODE SCRIPTS: A CONTENT ANALYTICAL APPROACH

STARLING DAVID HUNTER III

Name Starling David Hunter III
Academic centre Wayne State University
E-mail address starling@wayne.edu, starling@mit.edu

KEYWORDS
Pilot episode; script; screenplay; television; television series; television ratings; Nielsen ratings; statistical analysis; panel data; regression analysis; content analysis; network analysis; semantic network analysis.

ABSTRACT
Textual and content data were extracted from the pilot episode scripts of 183 new, dramatic television series appearing during ten consecutive seasons—2008/9 through 2017/8—on the four major American broadcast networks—ABC, CBS, Fox, and NBC. These data were used to predict the 18-49 demo ratings for the first five episodes of each series' first season. As expected, the *originality* of a series' premise, the *track record* of success of the its creator(s), and the *cognitive complexity* of its pilot episode script each explain a statistically significant proportion of the variance in the Nielsen ratings over the first five episodes.

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

1. INTRODUCTION

Every year in the middle of May, US television network executives converge on some of New York City's most prestigious venues for an event known in the trade as the "Upfronts" (Sobieck 2008). Among the most anticipated highlights of this event are said executives' announcements of their networks' programming schedules for the upcoming season, one which commences in earnest less than four months later.

Upfronts are also important because it is during this time that that these networks sell the majority of their advertisement inventory for the coming season (Sereday and Cui 2017). Given the large sums of money and commitments of resources that such transactions represent, all parties involved have strong economic incentives to accurately predict the performance of the forthcoming series. Over the last few decades, a wide variety of forecasting models have been developed and deployed—both by academic and industry-based researchers—to meet this pressing need, especially as it pertains to predictions of Nielsen ratings (Danaher, Dagger and Smith 2011).

In this pilot study, we emphasize the impact of three variables on one type of Nielsen rating—the 18-49 Demo—in a sample of 183 new dramatic series appearing on the four major broadcast networks (ABC, CBS, Fox, and NBC) during the ten most recently-completed seasons, i.e. the 2008-09 through the 2017-18 seasons. Those three variables are (a) the originality of the series' premise (b) the track record of success of the series' creator(s) and (c) the cognitive complexity of the pilot episode's script. Each of these variables has been shown in prior research to explain significant variation in several measures of performance of new, dramatic television series airing on major broadcast and cable networks (Hunter and Breen, 2017; Hunter, Smith, & Chinta, 2016). As expected, we find that each of these three factors explains a statistically significant proportion of variance in ratings *between* series.

2.1 Literature Review

Two groups of academic researchers have developed early-stage prediction models for film and television projects. Both groups rely on content analysis—using both human and machine coding—to extract a wide range of contextual and textual factors for subsequent investigation. Eliashberg, Hui and Zhang (2014) extracted several such factors from screenplays, many of which explained significant variation in box office and/or return on investment in a sample of 300 feature films. Among the more than two dozen factors they incorporated into their analysis were the average dialogue length, the film genre, the concentration of dialogue, the presence of a strong nemesis, and the familiarty of the setting.

Hunter, Smith and Singh (2016) extracted a different set of textual and content factors from a sample of over 170 feature-film screenplays, each of which explained significant variation in opening weekend box office. The key variables were the *originality* of the story's premise, the *track record* of the screenwriter, and the *cognitive complexity* of the screenplay itself.

Subsequent research investigated how these same three variables impact the study of the performance of new dramatic television series appearing on the four major broadcast networks (Hunter et al. 2016; Hunter, Smith and Chinta 2016; Hunter and Breen 2017). Performance outcomes explained by the three variables included the number of viewers (Hunter et al. 2016; Hunter, Smith and Chinta 2016) and the number of episodes in the first season as well as the likelihood of being renewed for a second season (Hunter and Breen 2017). To date there has been no examination of Nielsen-derived data as outcomes (performance measures). The aim of this analysis is to test whether the same variables that were previously used to predict several measures of TV series performance can also predict Nielsen ratings, the "gold standard" of series performance (Sereday and Cui 2017).

2.2 Hypotheses

Several recent studies by Hunter and colleagues (Hunter, Smith and Chinta 2016; Hunter et al. 2016; Hunter and Breen 2017) reported that the originality of a premise has a positive and statistically significant impact on the subsequent success of new, dramatic TV series. These measures included number of viewers per episode (Hunter, Smith, and Chinta 2016) and the number of episodes in the first season (Hunter and Breen 2017). In these studies, an original premise was one which credits no prior art, i.e. is not derived from pre-existing intellectual property such as other televisions series or franchises, novels, musicals, short stories, comic books, musicals, stage plays, etc. Nielsen ratings—the proprietary, and most-widely-used, metric for US television audience size—are widely accepted as the strongest leading indicator of the likelihood that a series will stay on the air (Sereday and Cui 2017; TV Series Finale 2019). Although none of these studies used Nielsen ratings as their measure of success, the relationship between ratings and other measures of success motivates the first hypothesis:

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

H1 All else equal, an *original premise* will be positively associated with the Nielsen ratings of a new dramatic series.

The same studies referenced immediately above also found a positive and significant impact from the track record of a series' creator(s) on its subsequent success. Specifically, they reported that new series from creators with at least one renewed series to their credit would have larger audiences than new series from creators with no renewed series to their credit. Notably, these findings paralleled research in the film economics literature which found that the prior performance of actors, directors, and screenwriters positively predicted box office (Goetzmann, Ravid and Sverdlove 2013). As such, our second hypothesis is that:

H2 All else equal, a strong creator track record will be positively associated with the Nielsen ratings of a new dramatic series.

Cognitive complexity is a construct common in the psychology literature. Two of its most common definitions are "the number of independent dimensions of concepts that an individual brings to bear in describing a particular domain of phenomena" (Scott 1962: 405) and "the number of independent constructs a person uses in perceiving and interpreting the environment" (Tinsley et al. 1983: 94). In the aforementioned studies by Hunter, et al, concept networks were constructed from selected words and concepts in the screenplay of the pilot episode script. The size of that network was their proxy for cognitive complexity. They also reported that measure to be very positively and significantly correlated with the success of new dramatic television series. Our third and final hypothesis is that

H3 All else equal, the cognitive complexity of the pilot episode screenplay will be positively associated with the Nielsen ratings of a new dramatic series.

3. METHODS & DATA

This analysis utilizes text- and content-based factors extracted from the pilot episodes of new dramatic series appearing on the four major broadcast networks over the 10 most recently-completed seasons, i.e. the 2008-9 through 2017-8 seasons. Excluded from consideration were series that had "*back-door"* pilots, e.g. *CSI Cyber* (2015); that premiered as *two-hour TV movies,* e.g. *Fringe* (2008); that were *anthology or event series*, e.g. *Law & Order: True Crime* (2017);that were exclusively *foreign-produced* series, e.g. *Crusoe* (2008); that premiered *without a pilot episode*, e.g. *Agent Carter* (2015), that moved to an *online platform* immediately after the network premiere, e.g. *The Good Fight* (2017), or that had a pilot episode broadcast in a *different season* than the second episode, e.g. *Glee* (2009).

Of the 193 series that remained, 183 (95%) had pilot episodes that could be incorporated into this analysis. The main sources for pilot episode scripts were the *TV Writing* blog (sites.google.com/site/tvwriting/) and the online screenplay broker *Scriptfly* (scriptfly.com). The ten pilot episode scripts that were not located and thus not included in the analysis are *The Ex-List* (2008), *Alcatraz* (2012), *Law & Order: Los Angeles* (2010), *Agents of Shield* (2013), *Once Upon a Time in Wonderland* (2013), *Super Girl* (2015), *Inhumans* (2017), *9-1-1* (2018), and *The Gifted* (2017). A complete list of these series included in the study is available upon request from the author.

3.1 Dependent Variable

The outcome measure for this study is the same-day "18-49 Demo" rating as provided by the Nielsen company. As its name suggests, the rating estimates the number of adults aged 18-49 watching an episode of a certain series and it is generally deemed to be more important than the total number of viewers (Storey 2009; Carter 2010). That's because the 18-49 demo is the one used to determine how much a network will charge for advertisements to be seen during commercial breaks (Santiago 2007).

Ratings data was located from various online sources including, but not limited to, *TV By the Numbers* (tvbythenumbers.zap2it.com), *TV Series Finale* (tvseriesfinale.com), and *The Futon Critic* (futoncritic.com) websites, only for the first five episodes of the first season of each series. Thus, our dataset is in the format of panel. That is to say, it contains ratings obtained for multiple, distinct time periods for the same series. Because seven series were cancelled before the fifth episode had aired—*Ironside* (2013), *Lucky Seven* (2013), *Lone Star* (2010), *My Generation* (2010), *The Playboy Club* (2011), and *Wicked City* (2015)—the panel is slightly unbalanced.

3.2 Predictors

The following factors were extracted from pilot episode scripts via content analysis or were inferred/calculated using information contained therein.

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

3.2.1 Originality of Concept

A categorical variable entitled "ORIGINALITY" was coded "0" if the series' Wikipedia page and/or *Internet Movie Database* (IMDb) page identified pre-existing source material upon which the series was based, e.g. a film, another TV series, a novel, a comic book, etc. and coded "1" otherwise.

3.2.2 Track Record

A categorical variable entitled "TRACK RECORD" was coded "1" only if one or more of the writers of the pilot episode script had previously written the pilot of one or more series that had been renewed for at least one full season, i.e. 20 or more episodes, and coded "0" otherwise.

3.2.3 Cognitive Complexity

In its continuous form, cognitive complexity was equal to the number of links in the main component of a semantic network constructed from the text of the series' pilot episode script. More specifically, the greater the number of links in the semantic network, the greater the cognitive complexity. Across the sample, the average number of links was 110.2 with a standard deviation of 46.6. The max was 293 links in the semantic network of the pilot episode script of CBS series *SEAL Team* (2017). Other pilot scripts with very large semantic networks included those for *Hawaii 50* (2010) with 272 links and *Scorpion* (2014) with 226 links. The smallest network—with only 8 links and thus the least cognitively complex—was that for the musical *Smash* (2012). Other pilot scripts with fewer than 30 links in their semantic networks included those for *Chicago Fire* (2012), *Harper's Island* (2008), *My Generation* (2010), and *Extant* (2013).

In the regression analyses below, the categorical variable entitled "Complexity" was coded "1" if the number of links was greater than 133, a number which marked the thirtieth percentile of the sample, and "0" otherwise. Detailed information on the processes by which screenplays were converted into text networks can be found in the appendix.

3.3 **Covariates**

Six co-variates not examined in prior research studies were also included in this analysis.

3.3.1 Female Show Creator

A categorical variable entitled "FEMALE" was coded "1" only if one or more of the writers of the pilot episode script were females and code "0" otherwise.

3.3.2 Length of First Act

Following Calvisi (2016), a measure of the length of the first act or "teaser" was created with a categorical variable that was coded "1" only if first act of the screenplay claimed 25% or more of the total pages of the pilot episode script and coded "0" otherwise.

3.3.3 Other Covariates

Five other non-content related variables were used in the statistical analysis that follows. These was one for each of the four major broadcast networks—ABC, CBS, Fox, and NBC. Specifically, a variable "ABC" was Coded "1" only if the series appeared on the *American Broadcasting Company* "ABC" and coded "0" otherwise. A variable named "CBS" was coded "1" only if the series appeared in the *Columbia Broadcasting System* (CBS) and coded "0" otherwise. Another named "FBC" was coded "1" only if the series appeared on the *Fox Broadcasting Company* (FBC) and coded "0" otherwise, and one named "NBC" was coded "1" only if the series appeared on the *National Broadcasting Company* (NBC) and coded "0" otherwise.

Finally, because Nielsen ratings are in a decade-plus long downtrend, I included a two-digit variable named "Season" to represent the year of the television season in which the series debuted, e.g. "16" for shows debuting at any point in the 2016-17 season or "09" for any series debuting at any time in the 2009-10 season, etc.

3.4 **Descriptive Statistics**

Table 1 contains descriptive statistics for the variables included in all subsequent statistical analyses. Bi-variate correlations among the three keys variables—originality, track record, and cognitive complexity—were all less than 4% and none were statistically significant. Specifically, the three bivariate correlations—*Originality-to-Track Record*, *Originality-to-Complexity*, and *Track Record-to-Complexity*—were -3.3% (p= 0.33), 3.8% (p=0.25) and 2.3% (p=0.48), respectively. One notable correlation was the significant and negative one

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

TABLE 1. DESCRIPTIVE STATISTICS

Variable	Mean	St. Dev	Min	Max
Log (18-49 demo)	0.140	0.241	-0.699	0.785
SeasonYear	12.85	2.74	8	17
Original	0.653	0.476	0	1
Track Record	0.209	0.407	0	1
Complex	0.283	0.451	0	1
Sum	1.146	0.780	0	3
Female Creator	0.241	0.428	0	1
Long First Act	0.383	0.486	0	1
ABC	0.312	0.464	0	1
CBS	0.230	0.421	0	1
NBC	0.277	0.447	0	1
FBC	0.181	0.386	0	1
Episode #	2.978	1.414	0	1

found between female creators and cognitive complexity ($p<0.001$, two-tailed). In other words, pilot episode scripts including one or more female writers had significantly lower cognitive complexity.

4. **RESULTS**

Two multiple regression models were specified in order to analyze the effect of the above-described variables on the 18-49 demo ratings over the first five episodes of the first season of the183 new dramatic series in our sample. For the panel data we specified a set of generalized least squares (GLS), random-effects regression models. To estimate the effects on individual episodes I specified a set of ordinary least squares (OLS) regression models. Table 2, below, contains results of the former while Table 3, also below, contains results for the latter.

4.1 **Results of Panel Data Regressions**

The first model includes only control variables—the season year, female creator, female lead, the length of the first act, and three network variables—ABC, CBS, and NBC. The model explains almost 35% of the variance *between* series and none of the variance *within* series. That means that these variables explain why some series have higher ratings than others but nothing about decline of ratings across episodes for a given series. Somewhat as expected, there is a strongly negative coefficient associated with the year, indicating thereby that 18-49 demo ratings are declining with each passing year. I also find that series from female creators have significantly lower ratings ($b = -0.071$, $p<0.05$, 1-tailed).

Models 2-4 add the three variables of interest—originality, track record, and cognitive complexity—and/or combinations thereof. In Model 2 we see that each of these variables are statistically significant at the $p<0.05$ level or better. From Model 3 we can see that the sum of these three factors is even more statistically significant, ($b = 0.066$, $p<0.0001$). In Model 4 we see that when all three variables were absent—not original, creator had no track record, and script was not cognitively complex—then ratings were again significantly lower ($b = -0.135$, $p<0.0001$).

In Model 5, the episode number and its number squared were added to the regression in Model 4. Both variables are very highly significant ($p<0.001$) and all other variables retain their relative level of significance as shown in Models 1-4. The important difference is that the two episode-related variables explain 54% of the variance *within* series and none of the variance *between* them. Put another way, the episode number is the best predictor of how much ratings of any series will decline over time but tells us nothing about why ratings differ between series.

4.2 **Results of Episode Regressions**

The results of the episode-specific regressions, as shown in Table 3, evidence a similar pattern. Specifically, the year of the series debut is the most statistically significant, and, as before, negatively so. Every control variable that was statistically significant in the panel regressions is also significant here. Notably, the model for the 3rd episode was the strongest both in terms of the percentage of variance explained (39.6% overall, 37.1% adjusted) while the model for the 4th episode is where the statistical significance of the key predictor was the greatest ($b = -0.144$, $p< 0.001$, 1-tailed).

4.3 **Summary**

The above results confirm that series with higher ratings were more likely (a) to be original rather than adapted (b) to be written by creators with strong track records of success and (c) to be more cognitively complex. In addition, combinations of these three factors were associated with even more significant differences in ratings.

The average 18-49 Demo ratings for the third episode of these series was 1.04 versus 1.62 for the rest of the sam-

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

TABLE 2. RANDOM-EFFECTS GENERALIZED LEAST SQUARES REGRESSION WHERE DEPENDENT VARIABLE IS THE LOG OF "18-49 DEMO" NIELSEN RATINGS

Variables	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Season Year	-0.040**** (-7.65)	-0.042**** (-7.83)	-0.041**** (-8.03)	-0.039**** (-7.54)	-0.038**** (-7.52)
Female Creator	-0.071* (-2.16)	-0.054* (-1.65)	-0.059* (-1.83)	-0.065* (-2.02)	-0.062* (-1.92)
Long First Act	-0.074** (-2.49)	-0.070** (-2.40)	-0.069** (-2.40)	-0.050* (-1.69)	-0.056* (-1.88)
Network = ABC	-0.085* (-2.07)	-0.091* (-2.28)	-0.091** (-2.29)	-0.101** (-2.51)	-0.102** (-2.55)
Network = CBS	-0.049 (-1.13)	-0.074* (-1.72)	-0.072* (-1.69)	-0.076* (-1.78)	-0.073* (-1.70)
Network = NBC	-0.086* (-2.04)	-0.095* (-2.32)	-0.095** (-2.33)	-0.094* (-2.29)	-0.093* (-2.27)
Original		0.052* (1.79)			
Track Record		0.071* (2.26)			
Complexity		0.078** (2.61)			
Sum			0.066**** (3.92)		
Sum =0				-0.135**** (-4.15)	-0.126**** (-4.16)
Episode					-0.110**** (-13.13)
Episode*Episode					0.011**** (7.99)
R-squared (within)	0.0%	0.35%	0.40%	1.2%	54.0%
R-squared (between)	34.6%	39.5%	39.3%	38.5%	37.4%
R-squared (overall)	30.1%	34.1%	33.9%	33.4%	40.3%
Wald Chi-squared	92.8****	114.5****	115.3****	115.3****	933.3****
N (observations)	893	893	893	893	893
N (series)	183	183	183	183	183

Note: regression coefficients are unstandardized

ple—a 56% difference. Also, the first seasons of these 37 series were 40% shorter. In addition, not one of them achieved a full first season (measured as 20 or more episodes) and only one went on to a full second season—*Lethal Weapon*. Not surprisingly, many more of these series than average had their cancellation announced before the last episode aired, were rescheduled to another day and/or time slot or were pulled from the altogether after just a few episodes.

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

TABLE 3: ORDINARY LEAST SQUARES REGRESSION WHERE DEPENDENT VARIABLE IS LOG OF "18-49 DEMO" NIELSEN RATINGS

Variables	**Episode 1**	**Episode 2**	**Episode 3**	**Episode 4**	**Episode 5**
Season Year	-0.036****	-0.041****	-0.039****	-0.037****	-0.041****
	(-7.18)	(-7.90)	(-7.40)	(-6.65)	(-6.59)
Female Creator	-0.067*	-0.064*	-0.065*	-0.075**	-0.058#
	(-2.11)	(-1.96)	(-1.97)	(-2.14)	(-1.51)
Long First Act	-0.061*	-0.035*	-0.057*	-0.057**	-0.064*
	(-2.08)	(-1.14)	(-1.85)	(-1.73)	(-1.76)
Network = ABC	-0.109**	-0.102**	-0.107**	-0.086*	-0.113*
	(-2.76)	(-2.50)	(-2.56)	(-1.92)	(-2.30)
Network = CBS	-0.106**	-0.071#	-0.085*	-0.058	-0.071#
	(-2.50)	(-1.62)	(-1.91)	(-1.22)	(-1.37)
Network = NBC	-0.094**	-0.078*	-0.107**	-0.108**	-0.104*
	(-2.32)	(-1.87)	(-2.53)	(-2.39)	(-2.09)
Sum = 0	-0.078*	-0.111**	-0.141****	-0.144****	-0.122**
	(-2.26)	(-3.15)	(-3.88)	(-3.72)	(-2.83)
Model R-squared	35.7%	37.8%	39.6%	36.6%	33.2%
Adjusted- R-squared	33.1%	35.3%	37.1%	33.9%	30.4%
Root MSE	17.8%	18.3%	18.3%	19.4%	21.3%
N (observations)	183	183	178	176	175

Note: regression coefficients are standardized

5. CONCLUSIONS

The models described in this paper do not explain ratings of all types of American television shows. Rather, they demonstrate only that variance in an important subset of them—new 1-hour dramatic series, which are a mainstay of American prime-time programming—can be predicted with some accuracy from factors contained in or derived from the pilot episode scripts. While that finding generally comports with prior research, to our knowledge, no other research has demonstrated a link between any script-based factors and Nielsen ratings.

Perhaps the most counter-intuitive of our findings concern the originality hypothesis—counter-intuitive because the results in the film industry are exactly the opposite. Specifically, our earlier work on film performance (Hunter, Smith and Singh 2016), as well as that of Basuroy and Chatterjee (2008), hypothesized and reported that sequels, re-makes, and adaptations were associated with significantly *higher* box office returns. However, the results do comport with our prior findings on televisions series which found originality positively associated with the number of viewers per episode (Hunter, Smith and Chinta 2016) and the number of episodes in the first season (Hunter and Breen 2017). That said, it is worth recalling the that results for originality were statistically the least powerful of the three hypothesized variables.

Two new results not reported elsewhere or in our previous work are the negative impacts of longer opening acts and the negative impact associated with female creators. The former result is consistent with predictions of script analysts writing in the professional screenwriting literature, e.g. Calvisi (2016) who argue for short, compelling, opening acts. The latter finding, however, has no precedent of which we are aware. There are many possible explanations for the finding concerning gender and ratings. One possibility concerns the apparent interaction of genre with both gender and ratings. Specifically, in this sample, female creators are significantly under-represented in the action genre ($r = -0.21$, $p < 0.0001$), one which is strongly and positively correlated with both Nielsen ratings ($r = 0.11$, $p < 0.01$) and with cognitive complexity ($r = 0.21$, $p < 0.0001$). In fact, of 37 action series in the sample, only three of them—a mere 8 percent—had one or more female creators—*American Odyssey* (2014,

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

NBC), *Chase* (2010, NBC), and *Off the Map* (2010, ABC). The sample contains not a single action series broadcast on either Fox or CBS with one or more female creators credited. In light of the stunning revelations of persistent and pervasive gender discrimination and sexual harassment at the highest levels of Hollywood recently brought to the fore by the Me-Too movement, readers could be forgiven for concluding that female under-representation in this specific area is not accidental.

As noted in the introduction, all players in the TV value chain have economic incentives to increase the accuracy of predictions about the ratings of new series. To the degree that models can predict which series are most likely to fall short ratings-wise, the less need there will be for make-goods and the more optimal media planning will be. Future research will endeavor to improve predictive accuracy by including data from scripts of each episode in the study sample.

ACKNOWLEDGEMENTS

An early version of this study was presented in April 2018 to Netflix's Content Data Sciences Group in Los Angeles. I thank William High, Tristan Cossio, Ashish Rastogi, and other members of the group for helpful suggestions and feedback on how to improve this work.

REFERENCES

Adalian, Josef (2012). "Six Female Showrunners Talk Ratings, Their Comedy Icons, and Internet Hate." *Vulture*, 25 May 2012. http://www.vulture.com/2012/05/female-showrunners-talk-ratings-internet-hate.html (last accessed 26-06-19).

Basuroy, Sumanand Subimal Chatterjee (2008). "Fast and frequent: Investigating box office revenues of motion picture sequels." *Journal of Business Research* 61(7): 798-803. http://dx.doi.org/10.1016/j.jbusres.2007.07.030.

Calvisi, Daniel (2016). *Story Maps: TV Drama—The Structure of the One-Hour Television Pilot*. Palos Verdes, CA: Act Four Screenplays.

Carter, Bill (2010). "An 'Idol' Ratings Loss, But Not in its Pocketbook." *The New York Times*, 6 April 2010. http://www.nytimes.com//2010//04//07//business//media//07adco.html (last accessed 26-06-19).

Danaher, Peter. J., Tracey S. Dagger and Michael Smith (2011). "Forecasting television ratings." *International Journal of Forecasting* 27(4): 1215-40. http://doi.org/10.1016/j.ijforecast.2010.08.002.

Eliashberg, Jehoshua, Sam K. Hui and Z. John Zhang (2014). "Assessing box office performance using movie scripts: A kernel-based approach." *IEEE Transactions on Knowledge and Data Engineering* 26(11): 2639-48. https://doi.org/10.1109/TKDE.2014.2306681.

Goetzmann, William. N., S. Abraham Ravid and Ronald Sverdlove (2013). "The pricing of soft and hard information: economic lessons from screenplay sales." *Journal of Cultural Economics* 37(2): 271-307. https://doi.org/10.1007/s10824-012-9183-5.

Hunter, Starling D. (2014). "A novel method of network text analysis." *Open Journal of Modern Linguistics* 4(2): 350-66. https://doi.org/10.4236/ojml.2014.42028.

Hunter, Starling and Yelitza Prada Breen (2017). "W(h)ither the Full Season: An Empirical Model for Predicting the Duration of New Television Series' First Season." *Advances in Journalism and Communication* 5(3): 38-97. https://doi.org/10.4236/ajc.2017.52005.

Hunter, Starling et al. (2016). "Moneyball for TV: A model for forecasting the audience of new dramatic television series." *Studies in Media and Communication* 4(2): 13-22.

Hunter, Starling., Susan Smith, S., & and Ravi Chinta, R. (2016). "Predicting new TV series ratings from their pilot episode scripts." *International Journal of English Linguistics,* 6(5):, 1-12. DOI: 10.5539/ijel.v6n5p1DOI: 10.5539/ijel.v6n5p1http://dx.doi.org/10.5539/ijel.v6n5p1.

Hunter, S., Susan Smith and Saba Singh (2016). "Predicting box office from the screenplay: A text analytical approach." *Journal of Screenwriting* 7(2): 135-154. https://doi.org/10.1386/josc.7.2.135_1.

Santiago, Rosario (2007). "For Advertising Purposes Grey's Anatomy May Well Be Colored Green." BuddyTV, 3 October 2007. http://www.buddytv.com/articles/greys-anatomy/for-advertising-purposes-greys-11849.aspx (now offline).

Scott, William A. (1962). "Cognitive complexity and cognitive flexibility." *Sociometry* 25(4): 405-14. http://doi.org/10.2307/2785779.

Sereday, Scott and Jingsong Cui (2017). "Using Machine Learning to Predict Future TV Ratings." https://www.nielsen.com/us/en/insights/article/2017/using-machine-learning-to-predict-future-tv-ratings/# (last accessed 26-06-19).

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

Sobieck, Ben (2008). “What are the Upfronts?” *Writer’s Digest*, 22 May 2008 http://www.writersdigest.com/editor-blogs/script-notes/industry-updates/what-are-the-upfronts (last accessed 26-06-19).

Storey, Michaal (2009). “Not in the 18-49 Age Group? TV Execs Write You Off.” *Arkansas Democrat Gazette*, 23 April 2009. http://www2.arkansasonline.com/news/2009/apr/23/tv-column-not-18-49-age-group-tv-execs-wr-20090423/ (last accessed 26-06-19).

Tinsley, Howard E. A. et al. (1983). “A Longitudinal Study of Female College Students’ Occupational Decision Making.” *The Career Development Quarterly*, 32(2): 89-102. https://doi.org/10.1002/j.2164-585x.1983.tb01563.x.

Watkins, Calvert (ed.) (2011). *The American Heritage Dictionary of Indo-European Roots*, 3rd ed. Boston: Houghton Mifflin Harcourt.

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

APPENDIX: DESCRIPTION OF THE PROCESS OF CONSTRUCTING A TEXT NETWORK FROM A SCREENPLAY

The purpose of this short appendix is to provide a detailed description of the steps by which screenplays in the sample were converted into text networks. For an understanding of the theoretical underpinnings for these steps, the reader is encouraged to consult Hunter (2014) which is entitled *A Novel Method of Network Text Analysis*.

The example screenplay is that for the political drama *Designated Survivor* (2016). Written by David Guggenheim, the series debuted on ABC and was cancelled after its second season but then picked up by Netflix for a third. The series follows the life of...

> Thomas Kirkman, an American politician named as the designated survivor for the State of the Union address, who suddenly ascends from the Secretary of Housing and Urban Development to the position of President after an explosion kills everyone ranked ahead of him in the line of succession[1].

Step 1: **Convert PDF to Text File**

The screenplay for *Designated Survivor* was downloaded as a machine-readable, PDF from the *TV Writing* website maintained by Lee Thomson[2]. The text of the body of the screenplay file was extracted and copied into Notepad and then saved as a text file.

Step 2: **Identifying Acronyms**

The text file was then uploaded into AutoMap, a network text analysis application developed by Prof. Kathleen Carley, Ph.D. of the *Center for Computational Analysis of Social and Organizational Systems* (CASOS) at Carnegie Mellon University[3]. The next step was to generate a list of potential acronyms. As a practical matter this means that Automap exported to an Excel file all words from the screenplay comprised of two or more letters, all of which were capitalized.

	A	B
1	**Potential Acronyms**	**Characters**
2	EEOC	4
3	SOTU	4
4	CIA	3
5	CSI	3
6	EMS	3
7	FBI	3
8	FHA	3
9	FHS	3
10	LEO	3
11	NSA	3
12	USS	3
13	DC	2
14	GW	2
15	NO	2
16	OR	2
17	PA	2
18	PM	2
19	UN	2
20	US	2
21	COMMUNICATIONS	14
22	TRANSPORTATION	14

Step 3: **Disambiguation of Acronyms**

The list of potential acronyms exported from Automap were then compared to a proprietary database of known acronyms in possession of the author. All acronyms that already appeared in the database were maintained. The remaining ones were evaluated by a human coder for their potential inclusion into the database.

1	**Disambiguated Acronyms**
2	EEOC (Equal Employment Opportunity Comission)
3	SOTU (State of the Union)
4	CIA (Central Intelligence Agency)
5	CSI (Crime Scene Investigation)
6	EMS (Emergency Medical Services)
7	FBI (Federal Bureau of Investigation)
8	FHA (Federal Housing Administration)
9	FHS (undetermined, possible typo)
10	LEO (Law Enforcement Officer)
11	NSA (National Security Agency)
12	USS (United States Ship)
13	DC (District of Columbia)
14	GW (George Washington)
15	NO (the word "no" caps)
16	OR (the word "or" in caps)
17	PA (Public Address)
18	PM (abbreviation for night time)
19	UN (United Nations)
20	US (United States)

1 https://en.wikipedia.org/wiki/Designated_Survivor_(TV_series)

2 http://www.zen134237.zen.co.uk/Designated_Survivor_1x01_-_Pilot.pdf

3 http://www.casos.cs.cmu.edu/projects/automap/

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

Step 3: **Application of Delete List**

The next step inside Automap was to employ the "text refinement" algorithm, specifically the application of a "delete list." That list consists of over 20,000 words which were previously determined not to be of interest in this analysis. After the delete list was applied, the remaining words were exported as an Excel file.

	A	B
1	**concept**	**frequency**
2	government-issued	1
3	forward-thinking	2
4	in-over-his-head	1
5	ambassadorship	1
6	Now-President	1
7	SPEECHWRITERS	2
8	TRANSPORATION	1
9	awe-inspiring	1
10	discriminated	1
11	informercials	1
12	over-stepping	1
13	speechwriting	1
14	three-by-five	1
15	INTERCUTTING	5
16	SPEECHWRITER	12
17	face-to-face	1
18	happenstance	1
19	nevertheless	1
20	notification	1
21	overwhelming	1
22	radiological	1
23	speechwriter	4

Step 4: **Transformation Coding**

The next step was the load the remaining words and acronyms into another proprietary database comprised of words and acronyms and the etymological roots that give rise to their underlying morphemes. Hunter (2014) coined the term "multi-morphemic compound" (MMC) to describe the compound, acronyms, and abbreviations which comprise the bulk of that database. For example, in this database there is an entry for the word *countdown*. In the database this word is transformed into "**pau-2**" which is the Indo-European root from which *count* descends and "**dheue-**" which is the root from which *down* descends (Watkins, 2011). There is also an entry for the acronym *CIA* which stands for *Central Intelligence Agency*. The three words comprising that acronym descend from three different Indo-European roots: *Central* descends from **kent-**; *Intelligence* descends from **leg-1**; and *Agency* descends from **ag-1**. And since both *motorcycle* and *CIA* appear in the screenplay of *Designated Survivor*, their transformation coding was done automatically. However, about 20% of the words and acronyms of interest in the screenplay did not appear the database. These had to be inspected individually and coded accordingly.

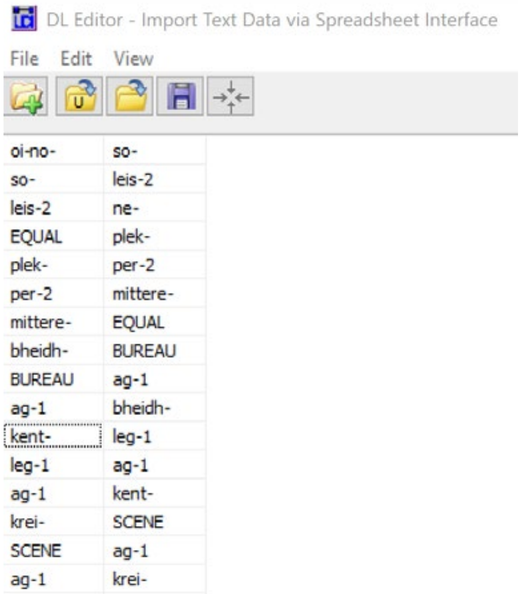
	A	B	C	D	E	F	G	H
1	Type	TITLE	CONCEPT	ROOT1	ROOT2	ROOT3	ROOT4	ROOT5
63574	DramaPilot	DesignatedSurvivor	bone-white	BONE	kweit-			
63575	DramaPilot	DesignatedSurvivor	breakfast	bhreg-	past-			
63576	DramaPilot	DesignatedSurvivor	briefcase	mregh-u-	kad-			
63577	DramaPilot	DesignatedSurvivor	broadcasted	BROAD	CAST			
63578	DramaPilot	DesignatedSurvivor	bystanders	ambhi-	sta-			
63579	DramaPilot	DesignatedSurvivor	CHAIRMAN	sed-1	man-1			
63580	DramaPilot	DesignatedSurvivor	Chairman	sed-1	man-1			
63581	DramaPilot	DesignatedSurvivor	CIA	kent-	leg-1	ag-1		
63582	DramaPilot	DesignatedSurvivor	cobblestone	COBBLE	stai-			
63583	DramaPilot	DesignatedSurvivor	comeback	gwa-	BACK			
63584	DramaPilot	DesignatedSurvivor	countdown	pau-2	dheue-			
63585	DramaPilot	DesignatedSurvivor	cross-eyed	crux-	okw-			
63586	DramaPilot	DesignatedSurvivor	cross-talk	crux-	del-2			
63587	DramaPilot	DesignatedSurvivor	CSI (Crime Scene Investigation)	krei-	SCENE	ag-1		
63588	DramaPilot	DesignatedSurvivor	DC	streig-	COLUMBIA			
63589	DramaPilot	DesignatedSurvivor	Defcon	gwhen-	reidh-	deik-		
63590	DramaPilot	DesignatedSurvivor	doorbell	dhwer-	bhel-4			
63591	DramaPilot	DesignatedSurvivor	doorway	dhwer-	wegh-			

DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

Step 5: **Conversion to Network Format**

Once all etymological roots of all relevant words and acronyms in the screenplay were identified, the next step was to put them in a format that allowed them to be represented as a text network. This step was performed by the "DL Editor" function within the social network analysis program known as UCINet. Specifically, it re-arranged the roots for each word as one or more pairs of roots, also known as an "edge list."

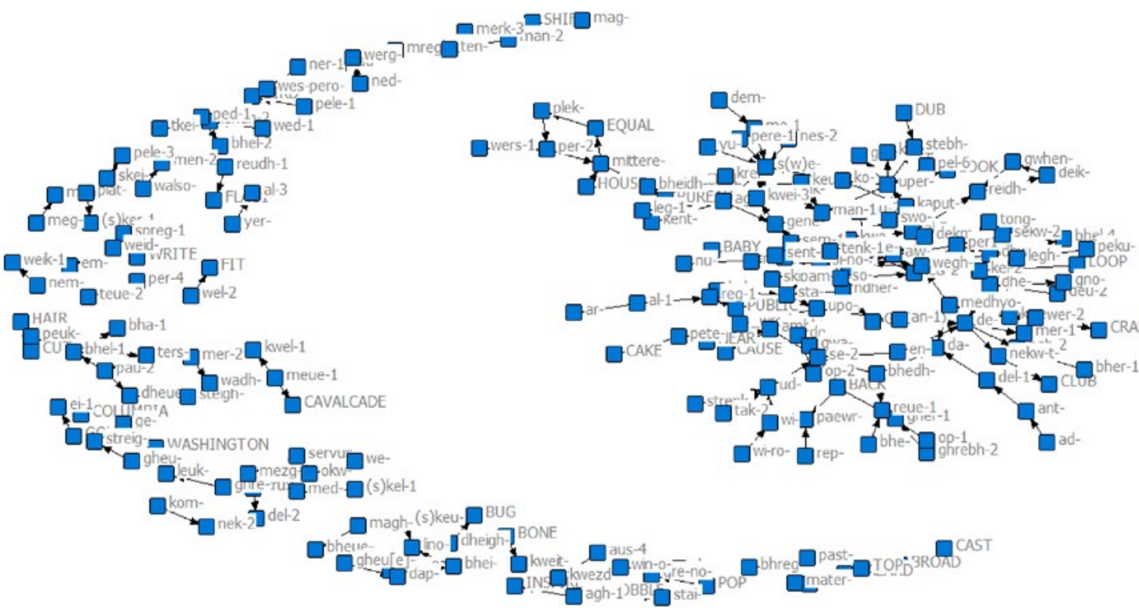
The resulting pairs of roots were exported to a file format that allowed them to be represented as a network graph.



oi-no-	so-
so-	leis-2
leis-2	ne-
EQUAL	plek-
plek-	per-2
per-2	mittere-
mittere-	EQUAL
bheidh-	BUREAU
BUREAU	ag-1
ag-1	bheidh-
kent-	leg-1
leg-1	ag-1
ag-1	kent-
krei-	SCENE
SCENE	ag-1
ag-1	krei-

Step 6: **Graphical Representation as a Network**

A program called NetDraw was used to first represent all pairs of concepts. This "whole" network, which is shown below, was comprised of 222 nodes broken down into 45 network components ranging in size from two (2) nodes to as many as 118 nodes. The largest cluster of nodes is clearly visible on the right-hand side of the diagram. In network parlance it is known as the "main component."



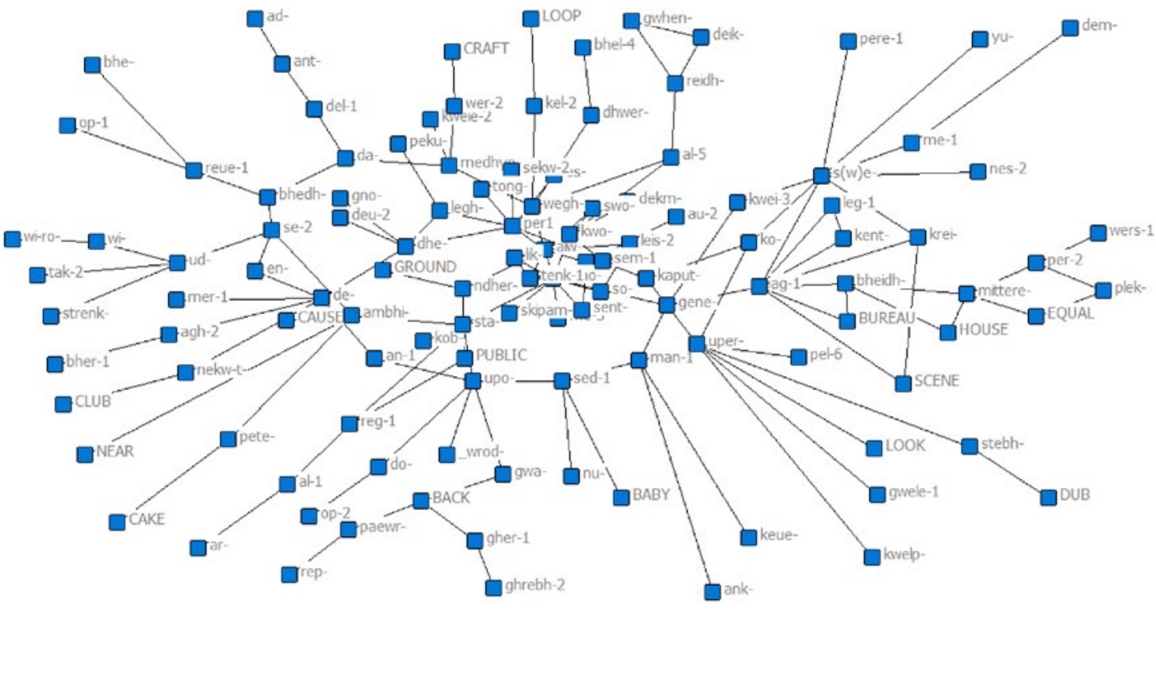
DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X

Step 7: **Measurement of the Main Component**

All measures used in the preceding statistical analysis were based on measures derived from the above network's main component which is the largest group of interconnected nodes. As noted above, the main component consisted of 118 nodes. These were interconnected by 160 ties, a number which put *Designated Survivor* in the top 10% of the sample.

Notably, the root **ag-1**, which means "to act," is clearly distinguishable in the center right of the diagram. Words and acronyms descending from that root include the *FBI* (Federal Bureau of *Investigation*), *CIA* (Central Intelligence *Agency*), *NSA* (National Security *Agency*), CSI (Crime Scene *Investigation*)—not an uncommon cluster of terms for contemporary political dramas in the USA. Also, connected to the cluster are the two other government agencies: the *Federal Housing Administration*—a government agency under the direction of Kirkman—and the *Equal Employment Opportunity Commission*—another government agency—as well as the words *Secret Service* and *Secretary of Housing and Urban Development*, the protagonist's position. Harder to discern in this graph is the center of the network where are found two important roots—**sta**- and **oi-no**- which give rise to the following relevant words and acronyms—SOTU (*State of the Union*), the President's annual address to Congress, *United States*, *United States Ship*, and *United Nations*.

One important point concerning the approach outlined above should be emphasized. Unlike all other methods of semantic network analysis of which the author is aware, the method use here places no emphasis or importance on frequency with which a given word appears in the screenplay. Rather, any word whose morphemes are derived from two or more etymological roots, i.e. a multi-morphemic compound, will appear somewhere in the semantic network, either in the main component or in one of the lesser ones.



DOI https://doi.org/10.6092/issn.2421-454X/9079
ISSN 2421-454X