THE AXIOLOGY OF NECROLOGIES: USING NATURAL LANGUAGE PROCESSING TO EXAMINE VALUES IN OBITUARIES

by

JACOB G. LEVERNIER

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2016

DISSERTATION APPROVAL PAGE

Student: Jacob G. Levernier

Title: The Axiology of Necrologies: Using Natural Language Processing to Examine Values in Obituaries

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Gerard Saucier	Chair
Mark Alfano	Core Member
Caitlin Fausey	Core Member
Colin Koopman	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2016

© 2016 Jacob G. Levernier
This work, including text and images of this document but not including supplemental files (for example, not including software code and data), is licensed under a Creative Commons
Attribution 4.0 International License.

License terms for the code and other supplemental files written for this dissertation can be found at http://doi.org/10.7264/D3WC7S.

DISSERTATION ABSTRACT

Jacob G. Levernier

Doctor of Philosophy

Department of Psychology

December 2016

Title: The Axiology of Necrologies: Using Natural Language Processing to Examine Values in Obituaries

This dissertation is centrally concerned with exploring obituaries as repositories of values. Obituaries are a publicly-available natural language source that are variably written for members of communities that are wide (nation-level) and narrow (city-level, or at the level of specific groups therein). Because they are explicitly summative, limited in size, and written for consumption by a public audience, obituaries may be expected to express concisely the aspects of their subjects' lives that the authors (often family members living in the same communities) found most salient or worthy of featuring.

140,599 obituaries nested in 832 newspapers from across the USA were scraped with permission from *Legacy.com,* an obituaries publisher. Obituaries were coded for the age at death and gender (female/male) of the deceased using automated algorithms. For each publishing newspaper, county-level median income, educational achievement (operationalized as percent of the population with a Bachelor's degree or higher), and race and ethnicity were averaged across counties, weighting by population size.

A Neo4J graph database was constructed using WordNet and the University of South Florida Free Association Norms datasets. Each word in each obituary in

the corpus was lemmatized. The shortest path through the WordNet graph from each lemma to 30 Schwartz value prototype words published by Bardi, Calogero, and Mullen (2008) was then recorded. From these path lengths, a new measure, "word-by-hop," was calculated for each Schwartz value to reflect the relative lexical distance between each obituary and that Schwartz value.

Of the Schwartz values, Power, Conformity, and Security were most indicated in the corpus, while Universalism, Hedonism, and Stimulation were least indicated. A series of seven two-level regression models suggested that, across Schwartz values, newspaper community accounted for the greatest amount of word-by-hop variability in the corpus. The best-fitting model indicated a small, negative effect of female status across Schwartz values. Unexpectedly, Hedonism and Conformity, which had conceptually opposite prototype words, were highly correlated, possibly indicating that obituary authors "compensate" for describing the deceased in a hedonistic way by concurrently emphasizing restraint. Future research could usefully further expand word-by-hop and incorporate individual-level covariates that match the newspaper-level covariates used here.

CURRICULUM VITAE

NAME OF AUTHOR: Jacob G. Levernier

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
University of San Francisco, San Francisco, CA, USA
Blackfriars Hall, University of Oxford, Oxford, UK

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2016, University of Oregon
Master of Science, Psychology, 2012, University of Oregon
Bachelor of Arts, Psychology, 2010, University of San Francisco

AREAS OF SPECIAL INTEREST:

Values
Data mining to understand morals
Data ethics

PROFESSIONAL EXPERIENCE:

Instructor, Data Carpentry, 2016
Graduate Affiliate, Digital Scholarship Center, University of Oregon, 2014-2016
Workshop Instructor (Archival Book Digitization, Scientific Programming in R, Scientific Programming with Version Control), Digital Scholarship Center and Department of Psychology, University of Oregon, 2015-2016
Instructor, Software Carpentry, 2014-2016
Research Data Management Assistant, Science Library, University of Oregon, 2013-2014
Software Architect for unified participant recruitment tracking system, Department of Psychology, University of Oregon, 2012-2015
Research Lab Manager, University of Oregon, 2012-2013

Graduate Teaching Fellow (Graduate Statistics, Advanced Data Analysis, Research Methods, Child Development, Motivation and Emotion, Imagination, Mnemotechnics), Department of Psychology, University of Oregon, 2011-2016
Graduate Research Fellow, University of Oregon, 2012-2013
Staff Research Associate, NCIRE - The Veterans' Health Research Institute / San Francisco Veterans Affairs Medical Center, 2010-2011
Research Assistant, University of San Francisco, 2009-2010
Research Assistant, Circle of Security, Spokane, WA, 2008
Research Assistant, University of San Francisco / Numerical Reasoning Lab, University of California, Berkeley, 2007

GRANTS, AWARDS AND HONORS:

Gregores Graduate Research Award, University of Oregon, 2013
John Templeton Foundation grant, 2012-2014, with Marjorie Taylor and Candice Mottweiler
Graduate School Research Award, University of Oregon, 2012
Marthe E. Smith Memorial Science Fellowship, University of Oregon, 2012
Valedictorian in Arts and Social Sciences, University of San Francisco, 2010
Dean's Medalist in Arts, University of San Francisco, 2010
Fr. Hubert Flynn, SJ, Award for highest cumulative GPA in graduating class, University of San Francisco, 2010
Member, Alpha Sigma Nu Jesuit Honor Society, 2008
University Scholar, University of San Francisco, 2006-2010
Dean's List, University of San Francisco, 2006-2010

PUBLICATIONS:

Taylor, M., Mottweiler, C.M., Naylor, E.R., & Levernier, J.G. (2015). Imaginary worlds in middle childhood: A case study of two pairs of coordinated paracosms. Creativity Research Journal. doi:10.1080/10400419.2015.1030318

ACKNOWLEDGEMENTS

My graduate experience would have been neither lasting nor fruitful had it not been for the guidance of a wide network of individuals. I am particularly grateful to Gerard Saucier, who, as my primary advisor, has demonstrated not only patience as my interests developed over my course of study but also insisted that I retain and strengthen my desire to work interdisciplinarily. I am also lastingly grateful to Mark Alfano at Delft University of Technology in the Netherlands and Australian Catholic University, who provided conceptual and methodological advice throughout this project, offered advice beyond this project, and generated the initial creative spark that eventually led to this dissertation.

I similarly appreciate of the work of the two additional members of the dissertation committee, Caitlin Fausey and Colin Koopman, as well as the members of other committees who have advised me throughout my time as a graduate student, Sanjay Srivastava, Azim Shariff, Brian Westra, and Lou Moses. I am additionally grateful to Marjorie Taylor for her supervision during the first two years of my course of study, and for her help in naming and pursuing newly-developing interests. With this in mind, I also acknowledge the obviously conscious work of the entire faculty of the Department of Psychology not only to allow but also to *facilitate* collaborative work across their research labs.

I am grateful to Kim Evenson and *Legacy.com* for their vision and open-mindedness in allowing this project to go forward. *Legacy.com* graciously provided not only permission to use obituaries from its website, but also insight through several shared conversations about the technical and ethical considerations of a project such as this.

This project would have been infeasible without the teaching and support of John Russell and Karen Estlund, now at Pennsylvania State University, and the University of Oregon's Digital Scholarship Center and its Graduate Affiliates (both the program and the people). I am also grateful to Rose Hartman and Tyler Matta for their generosity in offering time and advice to help me design and complete my analyses.

The members of the Department of Psychology's 2011 cohort have shaped my experience incalculably, especially in my final year here. I carry forward a growing sense of discernment that their conversation has helped to engender.

I am deeply grateful to Lori Olsen, whose logistical support and care of students as Graduate Secretary in the Department of Psychology enabled this work.

This project has been a sobering and ultimately very special experience in witnessing the humanity of individuals at perhaps their most fragile but also optimistic points: making sense of the mortality of those they care about, and of their own mortality. I am grateful to all of the authors of the obituaries that I sampled in the creation of this dissertation for reaffirming in their writing and anecdotes the optimism of living on, celebrating shared values and virtuous actions, and cherishing shared memories and experiences, even following great loss. I hope that I have responsibly curated these texts, and that this dissertation not only provides at least some insight into the values of different communities, but also does so in a way that is sensitive and meaningful to both obituary authors and their subjects.

To Emmit Routson, whose steps I have tried to follow throughout my education, and who would have had many optimistic and thoughtful things to say as I completed this project if he had had time to do so;

To Paul and Susan Levernier, whose closeness, sagacity, and advice *I* value tremendously;

And to Natasha Kolosowsky, who walked with me through this process.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Summary Statement[1]

This dissertation is centrally concerned with exploring obituaries as repositories of values. It seeks to lay a foundation for detecting values, defined from the Schwartz values model (e.g., Schwartz, 2012), in obituaries, and to examine their relation to characteristics of the individuals described in those obituaries, and to their communities. Obituaries are a publicly-available natural language source that are variably written for members of communities that are wide (nation-level) and narrow (city-level, or at the level of specific groups therein). Because they are explicitly summative, limited in size, and written for consumption by a public audience, obituaries may be expected to express concisely the aspects of their subjects' lives that the authors (often family members living in the same communities) found most salient or worthy of featuring.

This study of values using a natural language approach with obituaries fundamentally takes a descriptive rather than normative stance. However, this dissertation is founded in the idea that descriptive research into values can provide insight into the normative makeup of the communities that espouse them and, potentially, can allow useful relationships to be seen about the communities and the individuals that they comprise (in this project, the obituaries' subjects). As, to some extent, rituals of praise, obituaries specifically can offer a route to understanding the values of the communities that write them.

[1]The code and other supplemental files written for this dissertation can be found at `http://doi.org/10.7264/D3WC7S`.

Background on the Schwartz Values model

The Schwartz values paradigm (e.g., Schwartz, 2012) is widely cited[2] in the values literature for its wide scope and a substantial body of evidence demonstrating its ability to describe value priorities across cultures and social groups therein. As Schwartz (2012) summarized, the model posits that 10 basic categories reliably encompass differences in values across cultures. The model defines values "as desirable, transsituational goals, varying in importance, that serve as guiding principles in people's lives" (Schwartz et al., 2001, p. 521). Thus, in the model, value-categories are primarily differentiated by the "[primary end-]goal[s] or motivation[s]" toward which they are oriented (Schwartz, 2012, p. 4; Schwartz et al., 2001)[3].

Graphical depictions of the model often arrange the value categories circularly, conceptually allowing pairs of given categories to be complementary or in opposition to one another, and to represent that they are more continuously than rigidly defined[4]. The approximate quadrants of this circle can also be named

[2]As of September, 2016, Google Scholar recorded that Schwartz (1992) has been cited 12,544 times.

[3]Schwartz' conceptualization of values is in line with that of Haybron and Tiberius (2012, p. 9), who defined values as "relatively robust pro-attitudes [i.e., attitudes in favor of certain actions], or clusters of pro-attitudes, that we take to generate reasons for action and furnish standards for evaluating how our lives are going."

Schwartz (1992, p. 4) further defined values as "(1) concepts or beliefs [that] (2) pertain to desirable end states or behaviors, (3) transcend specific situations, (4) guide selection or evaluation of behavior and events, and (5) are ordered by relative importance." This definition agrees with Haybron and Tiberius' (2012) at least in its second and fourth parts. To Haybron and Tiberius, values go beyond short-term preferences in generating impulses to action, to some extent becoming ends for action in themselves in a way that preferences do not (Haybron and Tiberius noted that values can be called "subjectively reason-grounding" in that they are used by individuals to justify actions even when no good or otherwise justifiable reasons for those actions exist).

[4]A pictorial representation of this model is provided by Schwartz (2012), whose report is freely available under a Creative-Commons-family license (albeit one that prohibits reuse of the image by itself in this document).

(All quotes below are from Schwartz, 2012, p. 5-9; the list below moves around the image of the circular model presented by Schwartz, 2012, in a counter-clockwise direction):

1. Openness to change
    1. *Self-Direction* ("Defining goal: independent thought and action–choosing, creating, exploring.")
    2. *Stimulation* ("Defining goal: excitement, novelty, and challenge in life.")
    3. *Hedonism* ("Defining goal: pleasure or sensuous gratification for oneself.")
2. Self-Enhancement
    1. *Hedonism* (Also above, in Openness to change)
    2. *Achievement* ("Defining goal: personal success through demonstrating competence according to social standards.")
    3. *Power* ("Defining goal: social status and prestige, control or dominance over people and resources.")
3. Conservation
    1. *Security* ("Defining goal: safety, harmony, and stability of society, of relationships, and of self.")
    2. *Conformity* ("Defining goal: restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.")

3. *Tradition* ("Defining goal: respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides.")

4. Self-Transcendence

    1. *Benevolence* ("Defining goal: preserving and enhancing the welfare of those with whom one is in frequent personal contact [the 'in-group'].)"
    2. *Universalism* ("Defining goal: understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.")

While, Schwartz (2012) noted, the model does posit (and has accumulated a large body of evidence for) 10 distinct values, the value-categories may more usefully be thought of as pragmatically-defined boundaries along a continuum. Thus, Schwartz argued, the value categories can be expected to blend into one another, and can reasonably be subdivided (or superdivided) depending on the goals of a particular project. Further, measures that assess values as defined in the model seek to allow interpretation of "priorities" from among the value categories, rather than of raw scores (Schwartz, 2012, p. 12; cf. Schwartz et al., 2001). With this in mind, this dissertation project examines *relative* relationships among texts in a corpus in the values that they express, and seeks to avoid taking an absolute approach (e.g., x text is linguistically closer / expresses value y to a greater extent than value z," vs. "x text is about y value").

Methods of Measuring the Schwartz Values

Self-Report Measures. The body of evidence in support of the Schwartz model has primarily utilized self-report measures, to which it is useful to devote attention in order to understand the history on which later attempts to analyze values in natural language sources (such as Bardi et al., 2008,

and the project reported below) are founded. The model has been assessed primarily through the use of two self-report questionnaires, the Schwartz Values Survey (SVS), which is appropriate for use with adults, and the Portrait Values Questionnaire (PVQ), which is appropriate for use with children (Schwartz, 2012). As summarized by Schwartz (2012, p. 10), the SVS contains an approximate total of 56 items in two parts that describe "potentially desirable end-states in noun form" and "ways of acting in adjective form," respectively. The SVS assesses motivations as "guiding principle[s] in MY life," and is constructed with an asymmetric response scale to account for respondents' tendencies to rate all values as at least somewhat important to them. In contrast, the PVQ assesses the same 10 values as the SVS, but in a more cognitively concrete (and thus child-friendly) way, by asking respondents to rate hypothetical others for how similar they are to the respondent. Further decreasing cognitive load, the PVQ, unlike the SVS, does not explicitly identify itself as a measure of values (Schwartz et al., 2001).

Schwartz et al. (2001) introduced and evaluated a preliminary version of the PVQ after finding that SVS samples from some areas of the world (specifically, "sub-Saharan Africa, India, Malaysia, and rural areas of less developed nations" [p. 519], as well as with "those with minimal schooling [and] the elderly" [p. 538]) had not fit the 10-value model well. Schwartz et al. sought to examine the extent to which this lack of fit had been a measurement issue caused by the SVS rather than an issue with the generalizability of the theory itself. Encouragingly, Schwartz et al. did find that fit with the values model *was* related to the measure used (the SVS vs. the PVQ), interpreting this as evidence for increased generalizability of the model (since, even if the model is not able to be called "truly universal" [p. 538],

the study showed that populations that previously had not been well-described by the model *could* be assessed using the newer measure).

Schwartz (2012) reported on a variety of evidence that the model replicates cross-culturally and across populations within cultures. Specifically, Schwartz (p. 12) noted, across 82 countries, "each of the ten basic values [has been]... distinguished in at least 90% of samples." This body of evidence is not based on sparse samples: as of their writing, Schwartz et al. (2001) stated that that the circular values model had been supported by over 200 samples (at that time, in 60 or more countries), including several representative adult community samples as well as samples with adolescents (Schwartz et al., 2001, e.g., reported on a sample of 13- to 14-year-old girls from Uganda when examining the properties of the PVQ).

Natural Language. More recently, in addition to self-report, the Schwartz values paradigm has also been explored using natural language corpora. Natural language, while messier to analyze than data from standardized questionnaires such as the SVS and PVQ, opens a variety of new research possibilities that would be more difficult using those traditional self-report approaches, including looking across wider levels of society in an ecologically valid way (Bardi et al., 2008). Natural Language Processing (NLP) allows not only the use of data gathered from a variety of sources, at both individual and societal levels (*ibid.*), but also allows the use of data that were not originally created *as research data.* This report thus now briefly shifts to Bardi et al.'s (2008) approach to bringing NLP concepts to the Schwartz values, and to expanding their approach with obituaries, a type of corpus that can be expected to be particularly laden with values.

Obituaries as a type of document are a subset of newspapers, which Bardi et al. (p. 486) analyzed[5], noting that "there is evidence in previous studies... that popular textual media in democratic societies largely represent the salient values, opinions, and concerns of people." Within not just newspapers but text sources generally, obituaries can be expected to be particularly charged with values (even if only expressed using positive wording). Bardi et al. applied their value lexicon to a collection of newspapers from the United States between 1900 and 2000, searching for instances of words from the lexicon across all newspaper pages. The approach described below for the current project expands on this approach: instead of solely searching for word instances in a given text, the approach below includes calculating the lexical "distance" of every word in the text from each of the words listed in Bardi et al.'s value lexicon, allowing for greater nuance in analyses and conclusions.

Keeping in mind the current project's use of the value lexicon, Bardi et al.'s (2008) findings are encouraging. Bardi et al. examined the convergent and discriminant validity of their lexicon by comparing the relationships between the 10 Schwartz values found in their newspaper corpus with patterns found in self-reports by contemporary participants who had completed the Schwartz Values Survey. Bardi et al. noted that they did not expect to find a perfect correspondence between inter-correlations among Schwartz values as expressed in newspapers vs. through the SVS; in line with this expectation, they did not find a perfect relationship, but *did* find "the same general pattern of correlations" (p. 487).

[5]Bardi et al. did not specifically mention exploring obituaries. However, as of this writing (albeit eight years after Bardi et al. published their report), the resource they used, *NewspaperArchive.com,* notes on its home page that its holdings include "online historical and genealogical newspaper articles, obituaries, [and] local and international old newspapers archives" (NewspaperArchive.com, 2016).

Further, and particularly interesting for the current project, Bardi et al. examined the *predictive* validity of the lexicon by comparing changes in lexicon-word usage over time with society-level behavioral indicators that could be expected to share the same motivations as specific Schwartz values (for example, for Power, the percentage of the population engaged in active military duty by year; for Benevolence, the percentage of the population who had been deported by year; and for Conformity, the number of births that occurred without the parents being married). With the exception of Achievement (relating to successful patent applications per capita) and Tradition (relating to churches per capita), which co-occured regularly with words from other Schwartz values, Bardi et al. did observe a "medium effect size" (p. 490) in the relationship between co-occurence of words within and across Schwartz values in the newspaper corpus. Strikingly, Bardi et al. demonstrated through a series of visualizations that usage of words from the value lexicon over time did track major historical events in the United States (e.g., Security words increasing during the 1940s and early 2000s).

Bardi et al. (p. 490) concluded that, "overall, the value lexicon represents words that can significantly discriminate between Schwartz's values in natural language use on the Internet." Given their findings, the extension of their value lexicon to allow for words that are *not* part of the lexicon but are nonetheless *related to* the lexicon, especially using a corpus that can be expected to contain heightened levels of value words, seems a useful path for new research. The current project thus expands Bardi et al.'s (p. 493-495) goal of "facilitating the measurement of values over time and in real-world settings" "without reliance on self-report questionnaire responses," in an ecologically-valid way (p. 495). Further, given Bardi et al.'s conclusion that their value lexicon was sensitive enough to be

used at both individual and group levels, this extension in the project described below, which includes analyses of both individual-level (age, gender) and group-level (income, education, and ethnicity and race) predictors, seems potentially fruitful.

Why Obituaries

As briefly mentioned above, obituaries are a particularly useful source of language for understanding the relative value priorities of communities.

A Context of Praising Examplars for Their Ideal Traits. When they are written to be more than simple death notices by containing biographical information, obituaries are part of a long tradition of education in both moral and non-moral values through the use of exemplars, especially surrounding death. In the tradition of Western philosophy, the admonishment "not to speak evil of the dead" has been repeated at least since ancient Greece (Diogenes Laertius, c. 350/1853, p. 33).

Pappas and Zelcer (2014, p. 100 ff.) noted that in Plato's dialogue *The Menexenus,* which primarily repeats and comments on an apocryphal oration by Pericles, moral education was explicitly referenced as a goal of the Athenian tradition of funerary rites. Although the dialogue expresses skepticism on the actual ability of stories about exemplary moral agents to produce substantive and positive moral change in an audience – Pappas and Zelcer (p. 100) paraphrased that "high praise for the dead will make the hearers who did not know those dead envious[, especially about] whatever exceeds an accomplishment they would be capable of," and that, "where envy is the most natural response, emulation is not likely to occur" – regardless of the actual pedagogical efficacy of funerary stories, the inclusion of valued traits in such stories is clear in this tradition.

Obituaries obviously share features with eulogies such as those in the *Menexenus,* and exaggerate some features of eulogies: they are intended for public consumption, but at a larger scale (whether for hundreds to thousands of people in the case of obituaries published in local newspapers, or for millions of people in the case of those published in national or international publications). In addition, they are more versatile, typically running shorter in length than a transcribed eulogy, and thus are presumably easier for audience members to digest; indeed, the publication format of obituaries (with many together on a page) encourages reading more than one in a sitting. The 153 obituaries published in December 2013 in *The Eugene Register-Guard,* the primary newspaper in a university town of approximately 160,000 residents (and wider circulation to the surrounding county), averaged only 230 words in length. Despite their brevity, they comprised descriptions of a wide variety of backgrounds and life experiences, with ages of the deceased ranging from 17 to 99 years. In aggregate, these smaller "snapshots" might form a much larger, richer picture than a single eulogy of equivalent cumulative length.

Within the stream of Western civilization, the tradition of moral education through stories has had additional intermediate steps to the present since Pericles' oration. Reading hagiographies famously converted Ignatius Loyola, the founder of the Jesuit order of Catholic priests (for a history, see, e.g., Decloux, 1991), and formed a central part of his book *Spiritual Exercises,* several principles of which historically prefaced many (including non-denominational) modern "spiritual retreats." Moral primers that include both fables and true stories about paragons have also been used to educate children. The 1910 *Ethics for Children* (Cabot, 1910), as well as the more recent *The Book of Virtues* (Bennett, 1996), e.g.,

include stories about personages such as Abraham Lincoln alongside accounts of fictional characters.

Life-summaries of the dead are, as Fowler (2005, p. 53) argued, "more than a series of recollections about random individuals." Rather, Fowler (p. 56) noted, obituaries, whether written by "friends, colleagues, or even journalists," form a foundation of "collective memory" that, as with Alfano's (2013) conception of Hacking's (1995) "looping kinds," "feed back and shape public [history]." Berger (1969, p. 43) called "the confrontation with death (be it through actually witnessing the death of others or anticipating one's own death in the imagination)... what is probably the most important [']marginal situation,[']" in which not only individuals but entire communities "often... experience... 'ecstasy' (in the literal sense of *ek-stasis* — standing, or stepping, *outside* reality as commonly defined)." As Berger implied, mourning or otherwise facing mortality provokes a liminal experience through which current practices and closely-held values are reconciled, and possibly challenged. In the 15th century, the *Ars Moriendi* was produced to educate Christian mourners how to address and behave around dying individuals, and to remind those experiencing death how to do so virtuously (specifically by avoiding vices; see Nicholson, Caxton, & de Worde, 1891, which is an English translation of the original Latin text; this work is also noted by Doughty, 2014).

For Berger (p. 44), a "'good death'" means one that "retain[s] to the end a meaningful relationship with the *nomos* [per p. 19, "a[n]... order... of common meaning" imposed on individuals by their society]... — subjectively meaningful to oneself and objectively meaningful in the minds of others." Doughty (2014, p. 214) drew a similar conclusion, stating that "every culture has death values. These

values are transmitted in the form of stories and myths, told to children starting before they are old enough to form memories. The beliefs children grow up with give them a framework to make sense of and take control of their lives." Berger (p. 43) further asserted that the death anxiety experienced by a society needs a ritualized, form-based method of reduction; mortality salience, Berger stated, must be met with "legitimations of the reality of the social world," presumably including the moral world. Obituaries, especially through the process of authoring them but also through that of reading them, may be seen from this perspective as offering a chance to mourners within a community to declare the meaningfulness and "goodness" of the lives of the deceased[6], and of the shared values of those left

[6]Any informal survey of obituaries in a local newspaper in the USA will likely reveal that obituaries are almost exclusively positive in disposition, apparently embracing the maxim to "speak no evil of the dead" (this point was anecdotally substantiated in a formal [but not yet published or fully-analyzed] survey of over 1,000 obituaries undertaken by this author and advisors). Negative obituaries are rare enough to sometimes garner media attention specifically for their novelty, as in the case of Marianne Theresa Johnson-Reddick, whose children wrote, in part, "we celebrate her passing from this earth and hope she lives in the after-life reliving each gesture of violence, cruelty, and shame that she delivered on her children. Her surviving children will now live the rest of their lives with the peace of knowing their nightmare finally has some form of closure." (Mikkelson, 2013). As of this writing, a World Wide Web search for Johnson-Reddick brings up numerous news articles reporting on her obituary and trying to make sense of it: three of the first page of Google search results are (1) a post from *Snopes.com,* a website that investigates "urban legends, folklore, myths, rumors, and misinformation" (*Snopes.com*, 2016), examining the evidently difficult-to-believe claim that "[a] family member runs caustic obituary about deceased parent" (Mikkelson, 2013); (2) a *New York Daily News* article titled, "Son who helped write vicious obit for Reno mom insists 'everything in there was completely true'" (The Associated Press, 2013); and (3) a UK *Daily Mail* article which notes in its first summary point that the child who authored the obituary is "completely unrepentant" (Bates, 2013). Given the perceived newsworthiness of a negative obituary such as this, it seems reasonable to assume that obituaries (at least in the USA, where the obituary was authored, but also possibly in the UK, where the *Daily Mail* is published) carry a taboo against negative sentiments, and thus have a function related to educating readers in shared values, typically through sharing positive traits (and thus virtues). Indeed, an interview (McAndrew, 2014) published with Johnson-Reddick's son, who authored the obituary, began, "It did what obituaries don't do. Instead of a celebration of a life found in newspapers every day, the obituary for Marianne Theresa Johnson-Reddick was scathing." Notably, however, despite having violated this taboo against writing an obituary with a negative tone, Johnson-Reddick's son explicitly stated that he hoped to reinforce a shared value (disgust with child abuse) in his readers: "'People may see this as something we did to shame our mother.... But this is to bring shame to the issue of child abuse. I want every single person to realize this could be your obituary.'"

living; following Berger's vocabulary, obituaries could be understood as a type of "theodicy," a society-level mechanism for reassuring individuals of meaning; in this view, the writing and reading of obituaries could be an approach to quiet "anomic" discomfort (i.e., individuals' discomfort based on personally having to face death or other triggers that cause the *nomos* to come into question). Summarizing the lives of decedents offers an opportunity to legitimize values by publicly invoking them in a semi-ritualized way.

More specifically within the cultural history of the USA, on which this paper focuses because of the prevalence of obituaries from it as a data source, exemplars have been publicly used and described as a means of values (re)affirmation for decades if not centuries; witnessing death often requires individuals and communities such discomfort that a reaffirmation is required. This need is expressed variably, to the point that, as Doughty (2014, p. 225) described from an interview with a supervising physician, medical students will sometimes choose not to communicate terminal diagnoses rather than "'face their own mortality... [by] fac[ing] a dying person;'" cf. Rosel (1978), who also discussed at length the difficulty faced by many contemporary Americans in acknowledging the existential discomfort brought on by thinking about death with limited ritualized means. Writing and reading obituaries are two processes that remain for addressing this psychological discomfort. Given their cultural history, it makes sense that, among all publicly-read media, obituaries would be among the best candidate bearers of values. Because obituaries are succinct and explicitly intended to summarize their subjects' lives, they may be expected to include only the features that their authors find most salient, to signal not only to their authors themselves as relatives

or biographers of the deceased but also to others in the community aspects of the character of the deceased that all in common might find important.

Obituaries are public and readily-available, not only in print newspapers but also more accessibly through newspaper websites and databases, allowing, with the ethical and legal caveats considered below, the ability to readily conduct text-mining analyses without having first to transcribe the texts. While obituaries are certainly not the only bearers of values, this ready availability makes them an attractive option for research. This attraction is particularly salient for ethnographic research to understand the differential values of geographically or otherwise-distinct communities across the USA (e.g., compared not only by physical location but also by gender). In addition, it provides a foundation for research of a psychometric mindset (e.g., using prospective obituaries written by living individuals to index their values "match" with different communities across the USA based on those communities' published obituaries, and to corroborate this index with survey data using existing measures of values such as the Schwartz Values Survey, Schwartz, 1992).

The Logistics of Using Obituaries as Values-Containing Data Sources

Obituaries and related formats of necrologies are not only a *potential* data source for community values; rather, they are currently in use, both for their existential weight and for their accessibility as public resources. Having argued above that obituaries may be expected to contain community values, this paper now turns to considering the logistical realities of extracting those values from this format of data. Specifically, instances in which obituaries have directly been used will be considered as examples from which to build new research.

Necrologies are in use Prospectively in Clinical Practice

In clinical psychological practice, Hayes, Strosahl, and Wilson (2011, p. 304 ff.) defined a prospective exercise called "What do you want your life to stand for?" for clinicians to use as part of an Acceptance and Commitment Therapy treatment approach. In this exercise, a therapist asks a client to imagine having died but being able to attend her own funeral. The client is encouraged to consider the "eulogies offered by [her spouse,]... children,... friends, [and] the people [she has]... worked with." This exercise is explicitly centered on considering what an ideal eulogy from each of those speakers would comprise, moving beyond simply thinking about the immediate effects of one's own death (e.g., one's spouse and children being distraught) and into the larger meanings that others in one's community might draw from one's life. Implying the gravity of the characteristics that eulogies (and, by extension, obituaries) typically include, Hayes et al. encouraged clinicians to point out to clients that clients' imagined eulogies typically comprise different, larger goals and values than individuals are often concerned with on a day-to-day basis — that is, these life-summaries are concerned more with values that speak to moral, spiritual, and character accomplishments, rather than things that a client might "berate [himself]... about and struggle with" on a more local scale, such as annual income or feelings of personal inadequacy.

Hayes et al. (2011) also proposed an alternative exercise, in which a client is encouraged to compose and then write an epitaph on an imagined or drawn tombstone, noting that it also "often... reveals wide discrepancies between the client's values and his or her current actions." Although Hayes et al. noted that this exercise can be overwhelming for some individuals for the amount of mortality salience it elicits, this writing of prospective epitaphs, eulogies, and, by

extension, obituaries, either about oneself or a loved one or friend, underscores the psychological gravity associated with summarizing an individual's life upon his or her *actual* death.

Obituaries are in Use Retrospectively in Social Research.

Obituaries have also been used non-clinically to examine social processes. In the last of a series of studies, Goodwin, Piazza, and Rozin (2014) used living participants' ratings of the individuals described in obituaries to examine the importance of moral character and social warmth on individuals' perceptions of others. Bridging definitions of virtue, values, and personality, Goodwin et al. (p. 148) defined "moral character" as "compris[ing] the moral[7] dimensions of a person's personality," implying its inclusion in modern personality research (as Saucier, 2009, also noted). To Goodwin et al., traits exist on a continuum of moral relevance, as well as on a continuum of general relevance in constructing global perceptions about others. Goodwin et al. collected 250 recent obituaries from *The New York Times,* which, as a major newspaper of record, typically runs longer-format, professionally-written, more in-depth biographical accounts than local newspapers. Obituaries published in *The New York Times* are also predominantly about particularly noteworthy individuals – CEOs of major companies, international military leaders (famous or infamous), well-known artists, etc. Goodwin et al. noted that obituaries offer detailed views into the lives of their subjects, and thus comprised a useful corpus for understanding person-perception (while controlling for coders' previous knowledge of the deceased).

Goodwin et al. (p. 162) employed (living) Research Assistants (RAs) and participants from Amazon's Mechanical Turk service to code the obituaries on

[7]Goodwin et al. did not define the word "moral."

several criteria. RAs first coded the obituaries on valenced Likert scales for the obituary subject's described "(1) abilities or lack of abilities, (2) moral character or immoral character,... [and] (3) social warmth or coldness." Twelve or more Mechanical Turk workers then independently rated each obituary on a wider Likert scale for overall impression. Goodwin et al. (p. 162) concluded, in part, that "moral character impressions are conveyed more prominently in [these] summary accounts of people's lives than are impressions of social warmth," implying that obituaries *did* carry significant information about moral character. Further, Goodwin et al. found that Mechanical Turk participants' overall impressions of obituary subjects were able to be predicted significantly using RA's ratings of "morality, warmth, and ability," indicating (in addition to a satisfactory ICC statistic among RA coders on these dimensions) that obituaries of this type carry information about these qualities that is consistently accessible to readers. While shorter-format obituaries from smaller newspapers almost certainly do not carry this information about individual subjects to the same extent as the long-format biographies of *The New York Times,* they may, in aggregate, similarly provide useful information about the community of authors and readers among whom the deceased may have lived.

Rodler, Kirchler, and Hölzl (2001, p. 829-831) used obituaries as an "unobtrusive method" for examining longitudinal change in stereotypes about male and female subjects who had, in life, occupied positions of organizational power (e.g., "director[s] or head[s] of a business firm, school or other public organization, chair[s] of a department, etc."). Replicating and updating previous work that sought to establish trends in the words used to describe male and female leaders, Rodler et al. gathered 992 obituaries authored by the corporations or

other organizations for which the deceased had served as leaders. The sample of obituaries allowed immediate conclusions to be drawn about the male-to-female leadership workforce distribution over the years included in the analysis: the corpus was heavily lopsided, with 757 obituaries about male leaders and 137 about female leaders, even after a second round of data collection to balance the sample as much as possible. Rodler et al. [p. 831] coded "all verbs, adjectives, and nouns" in each obituary for representation of 58 categories that had been established in previous work (including, e.g., "Caring," "Consensus-oriented," "Decision-maker," and "Skillful"). An iterative correspondence analysis on the frequencies at which obituaries of men vs. women invoked each of the categories revealed two primary dimensions: first, an authority dimension, which included "attributes like servant, committed, professional, and humane to intelligent, efficient, skillful, experienced, and expert" (p. 831); and second, a competence dimension, which was, "at the one pole, described by expert, experienced, professional, intelligent, and, at the other pole, by venerable." Rodler et al. concluded that while descriptions of male leaders had remained largely consistent over time (during the years 1992 and 1998, which were included in the sample, but also building on previous research that first used these analysis methods on obituaries during the 1970s and then 1980s), descriptions of female leaders moved from the 1970s to 1990s from focusing on "venerable" qualities to "professional[ism] and commit[ment]."

Rodler et al. (2001) may be seen to have been studying (even if indirectly) community values around gender. Additionally, Rodler et al.'s analyses included several categories that meet definitions of virtue, including "courageous," "ethical," "honest," "patriotic," "religious," "unselfish," and "venerable" (see their Appendix). From this work, as well as work like that of Goodwin et al. (2014),

obituaries may be seen by precedent to be an accessible, usable data source for examining social dynamics, including the expression of values generally.

Obituaries are Biased in a Useful Way. As data sources, obituaries likely reveal as much (and, regarding values, possibly more) about their authors and readers as about the individuals whom they describe. In contrast to existing measures of values and their antecedents such as the Moral Foundations Questionnaire (Graham et al., 2011), which includes items that address disvalues (i.e., negative values), but somewhat like the Schwartz Values Survey, which does not, obituaries may be expected to be almost entirely positively biased in the traits they ascribe to their subjects. Speaking conceptually on this type of sampling bias, Fowler (2005, p. 61) noted that, "admittedly, there are limits to the obituary as a form of witnessing, due to constraints on critical openness at the time of death [i.e., a taboo on mentioning negative aspects of the deceased] and the conflicting perspectives of different newspapers [(which might, e.g., each publish a death notice of noteworthy individuals with a unique political spin)]." However, Fowler suggested that obituaries presented together in newspapers and similar media create a tapestry of a larger process of grieving and reflecting than any single piece could show. In their values, two obituaries, individually read, might tell little about a community, reflecting too strongly the idiosyncratic grief or values of the authoring friend or family member (or, perhaps, the political or other ideological agenda of the authoring journalist) as he or she attempts to reconcile the loss of the deceased with some meaning in his or her own life (cf. Doughty (2014), who, as noted above, writes at length about the meaning-making that death engenders among the living, and the varied ways in which those left alive compensate for the loss of and grieve the dead). In aggregate, though, obituaries "[supply] factual

materials which can be read in terms of a wider relational perspective, [and thus]... contribute a vital resource for actively shaping and demystifying collective memory."

Fowler (2005, p. 61 ff.) did note that newspaper obituaries have become "less coded [and] more subtle" since the 19th century (i.e., less adherent to historical "rigid aristocratic formulae" that sought to shape public political and professional consciousness by offering "verdict[s]" and "last judgment[s]" about the deceased's professional and personal accomplishments). While obituaries of this "coded" form certainly still exist (especially, as noted above, in the commissioned biography-obituaries of major political figures in international newspapers such as *The New York Times*), given that obituaries of who might be called "everyday people" have become more available in major newspapers across the USA over the last century, Fowler implies in her argument that in modern times, obituaries are able to be read more at face value than in the past. However, obituaries in many newspapers continue to exhibit a heavy sampling bias toward what Fowler called "the dominants" of a community's social hierarchy. Fowler reported from a content analysis of approximately 100 British obituaries across major newspapers that the individuals summarized were disproportionately well-educated (i.e., college educated) and of a very high social class (i.e., "Oxbridge" educated).

Fowler also suggested that obituaries often disregard explicit mention of communalistic values in favor of individualistic ones, portraying the deceased in personal narratives that, even if not fully divorced from the social context that facilitated that personal journey, fail to accurately reflect the importance of that social context. Fowler interpreted this disregard as a remnant of the historical bias, noted above, toward aristocratic and otherwise socially "dominant" figures.

In line with this, Fowler (p. 63) also noted a decrease over time in the number of prominent obituaries of individuals who served communalistic social organizations, such as "military, clergy or elite civil servants," with a concomitant increase in focus on "artists, writers, musicians and actors, or academics and politicians." In future research, this discrepancy might be able to be controlled for or at least understood empirically, however, using other markers of presumably communalistic thinking on the part of obituary authors (e.g., percentage of family- and team-related words used to describe the deceased).

In summary, the use of obituaries as markers or carriers of community values may be expected to be beset by (at least partially) sampling bias in (a) the values themselves (with a preference for positive terms), (b) the type of values (with a preference for individualism over communalism), (c) the individuals described (with a preference for higher-educated and especially influential individuals), and (d) gender (at least in some contexts). These biases would be substantive sources of limitation in a research project seeking to catalog community values (Goodwin et al., 2014, also acknowledged the second form of bias in their work); however, once acknowledged (and effectively taken into account), they would not invalidate this research approach. One fourth type of bias that Fowler noted, but which would be beneficial to the use of obituaries for the study of values in a given place and time, is political motivation of authors, especially in highly-publicized obituaries. Noting that biographical sketches in obituaries can be written either according to or specifically to counter and raise awareness of changes in a prevailing societal view, the ability of obituaries to change over time in their foci makes them stronger candidates for longitudinal analyses of a given geographic or otherwise cultural location.

It should be noted that Fowler's content analysis used four major London newspapers, plus *The New York Times* and *Le Monde* (p. 70), all of which likely print obituaries that have been chosen on a highly selective and topical basis and have been written by professional, commissioned authors. With access to smaller publications (e.g., local newspapers for small- to medium-sized towns and cities across the USA), the causes of sampling bias in the larger papers could potentially be used as a feature rather than a bug of research, as they could allow the differential comparison of national-level values discourse with the discourse of smaller communities, possibly following research such as that of Fu, Plaut, Treadway, and Markus (2014), who analyzed and contrasted the value landscapes of quadrants of the USA. This approach could allow the comparison of local-level newspapers with major publishers of obituaries (such as *The New York Times*) that controls for geographic and ideological similarity between the publications.

Publicly-Available, but Possibly Ethically Problematic Considerations for the Deceased and Their Family Members and Friends when Using Obituaries in Research. Obituaries aim not only to broadcast to readers facts about the deceased but also to convey intimate, summative life portraits. With that context in mind, it is worth consideration that, especially for obituaries authored by friends and family with the intent to publish in a local newspaper for consumption by community members only, the work of "morality mining" (Christen, Alfano, Bangerter, & Lapsley, 2013) could be interpreted as intrusive, especially when researchers are neither members of authors' local communities nor manually reading and interpreting each obituary individually as a document of a once-living person. Aggregating public records that were likely originally intended by their authors to be read singly, while legally

unproblematic, could become emotionally distressing to authors or their relations if they felt that their words were being taken out of context or used for purposes for which they neither intended nor explicitly consented (such as generating profit or, possibly, academic prestige, where focus might fall more on the researching company or academic team than on the values of the authoring communities and legacies of their decedent members).

Valeski (2012, p. 217) argued that understanding ethical implications of the use of "new forms of communication" by third parties (including researchers) often centers on considering what social norms have developed around those new technologies' predecessors. These norms, Valeski noted, often include dynamics of control between authors and recipients (indeed, ownership and control of access are two of the four "elements of big data ethics" proposed by Davis & Patterson, 2012, p. 16). If obituaries are written not only to serve as official, public death notices but also to facilitate the grieving processes of both readers and authors by shaping reminiscence (i.e., if obituary authors are writing for themselves and their families as much as or more so than for the readers), a sense of control over the use of the text is likely important in that grieving process. It would be ethically problematic for researchers, even if well-intentioned, to exploit the publicness of the medium at the expense of authors' (and deceased individuals') interests.

This sentiment follows Tsvetovat and Kouznetsov (2011, p. 162), who argued that any type of data collection "on human beings *affects their lives,* in hard-to-perceive subtle ways." This argument can be extended and applied to the use of public datasets with previously-collected data (cf. O'Neil & Schutt, 2013, p. 354ff., who cite and re-print a "Hippocratic Oath of Modeling"). It also follows Scime and Murray (2013), who posited that social scientists, especially using large, aggregated

datasets, are responsible to the communities in which they are performing their research, including remaining sympathetic to changing norms and expectations of community members around data use. Tsvetovat and Kouznetsov's argument would also certainly apply in research that seeks to examine prospective obituaries from living participants (e.g., using the Acceptance and Commitment Therapy exercises proposed by Hayes et al., 2011).

A central theme around which ethical considerations for the use of public but emotionally-charged data such as obituary texts can center is a socially-based concept of privacy. In line with Scime and Murray (2013), Helbing (2015, p. 135) noted that "ethical problems... are related to cultural values and social norms;" following Solove (2007), this is especially the case with potential violations of privacy. Solove (p. 754 ff.) noted that "traditional" attempts to define privacy have done so, both narrowly and broadly, by searching for "the essence of privacy," a feature that extends across all contexts and types of concerns. Solove summarized and argued for an alternative conceptualization in which privacy is understood not merely in terms of individual rights but rather also with reference to social good. In this conceptual approach, privacy is about "power relationships between people and... institutions" (p. 757), and is often concerned with enabling "rules of behavior, decorum, and civility" that may relate to expressions of "people's autonomy and dignity" (p. 761-3).

With this conception in mind, a researcher working with this type of data may be seen to have a *greater* ethical responsibility than she otherwise would, as "misuse" of these data could be defined by the very *perception* of their misuse by members of the public, and could not only damage the experience of privacy (as control of use of the data, regardless of the data's public nature) of the

obituaries' authors and their relations, but could also diminish those and other individuals' *future* willingness to engage in this type of public ritual. Thus, for the potential gain that could come from responsible, thoughtful use of this type of data, there also exists a potential damage, born of perceived meddling in moments of particular fragility. This is especially the case with obituary data, as many of the recommendations provided by Helbing (2015, p. 137) "towards privacy-preserving data analysis" are infeasible: the "participants" in this type of research are the deceased individuals who are the subjects of the texts; they cannot consent to research, and are specifically excluded from the federal definition of human research subjects in the USA (Office for Human Research Protections, 2010, §46.102f). Their data cannot be substantively anonymized or randomized, as the full text even of an obituary scrubbed of names is often easily recoverable by searching the World Wide Web for a phrase or sentence from it. As with this dissertation project, it is possible to "coarse-grain" (Helbing, 2015, p. 141) this type of data by performing analyses only at aggregate levels and to release only datasets that are derived from (but do not actually include) the original texts; this, however, comes at the cost of full reproducibility by future researchers.

To some extent, ethical questions surrounding the use of this type of data may be empirical ones: many authors, if asked, may be happy to have the obituaries they authored be included in aggregative inquiry, especially if it is performed with the goal of understanding community values. Regardless, values research conducted with obituaries should take place with a mindset of sensitivity and respect to the obituaries' authors and their subjects, and should proceed with as much transparency as possible, perhaps also including explicit licenses on published datasets, clarifying expectations around data reuse. Ideally, these issues

should be explicitly considered in a Data Management Plan *before* beginning a project.

In summary, the potential for negative emotional reactions among authors reveals an ethical weight associated with this class of data, and also points to the value of intent in interacting with these records. In addition, this ethical weight indicates the richness that this type of corpus may hold for research that seeks to understand community values in order to promote the social good. Although this type of research could be undertaken to manipulate a ritual of public grieving into an abstract data-generating mechanism divorced from its original context, those wishing to understand the moral language of local communities responsibly, especially to perform value-relevant work within or translating across those author communities, may find valuable insight in this approach. To this end, it is worth reflecting on Schwartz' (2012) argument that "the critical focus of value transmission is to develop commitment to positive relations, identification with the group, and loyalty to its members." The potential importance of obituaries as values-transmitting artifacts, at least within the United States, may be expected to relate to the enormity of betrayal and upset that members of a community could experience if they perceived a violation of propriety surrounding their use.

Legal Considerations for Research that Uses Obituaries. The American Psychological Association's (APA) *Ethical Principles of Psychologists and Code of Conduct* guidelines (American Psychological Association, 2010, §8.14) advise that,

> After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through

> reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and *unless legal rights concerning proprietary data preclude their release*" [emphasis added].

Whether considered for allowing publication of a dataset or for using the dataset to begin with, the legal status of obituaries is unclear. Obituaries are to some extent public records (and thus especially useful as research data), since they are published in newspapers to serve as death notices (source: personal correspondence with *Legacy.com,* one of the USA's largest obituary publishers, which contracts with newspapers nationally), but their actual legal status is often ill-defined; e.g., many states in the USA do not *require* the publication of death notices, meaning that obituaries' publication is not necessarily protected under public notice laws in those states. Russell (2012)[8] noted that copyright of obituaries typically remains with their original authors (unless the authors have atypically signed over rights to the publishing newspaper), but that the Fair Use doctrine in the USA also protects their inclusion in research under certain circumstances. These points highlight that while obituary data begins "in the open" (i.e., published for public consumption), their availability for re-use remains uncertain, especially around issues of republication of datasets. Especially with obituaries, if data are to be made available for further use by other researchers, the admonitions from the APA's guidelines above may be seen to require the

[8]The analysis of Russell (2012) is not a source of official legal advice, but is a useful summary of legal issues specifically surrounding the use of obituaries.

researchers who initially gather the data to serve as ethical "gatekeepers" for re-use of the data[9].

The Goals of this Project

This dissertation project has two purposes. First, the project examines the Schwartz values paradigm in a naturalistic context (that of obituaries). To this end, the project seeks to describe which values are discussed most across contemporary newspaper-based "communities" across the United States, taking into account potential individual- and group-level predictors of values-differentiation: at the individual level, the age and gender of the deceased; and at the community level, median income, education level, and ethnic and racial demographic composition in the counties in which each obituary's publishing newspaper(s) are based.

Second, this project seeks to develop tools for a novel method of assessing the "fit" of obituary texts with the Schwartz values for future researchers to use. The creation and documentation of a graph database from Princeton University's (2010) WordNet thesaurus[10], following but expanding on the work of Nagi (2013), as well as readily-usable algorithms for employing it in the analysis of this obituary

[9]Indeed, when I negotiated a data use agreement with *Legacy.com,* one of the largest obituary warehousing websites in the USA, the agreement centered around avoiding breaches of public trust not only in my own research (specifically by conducting analyses unlikely to cause resentment or visceral unease on the part of obituary authors or other family members of the deceased) but also in future research that others might perform from data gathered and processed through this project (minimizing these through, e.g., data cleaning and partial anonymization techniques).

While, as discussed above, full anonymization of the obituaries would likely not be possible (as a simple internet search of any of the obituary's text would allow finding the original source), names (e.g., surnames) could be stripped out of published versions of derivative datasets, encouraging future researchers to perform analyses of the texts over potentially more problematic analyses such as the race of obituary subjects.

[10]During the course of writing this dissertation, the official WordNet project released a canonical Neo4J graph database, currently available at https://github.com/wordnet/wordnet. The inclusion of the University of South Florida Free Association Norms dataset (Nelson, McEvoy, & Schreiber, 1998) in this dissertation project's database, as well as the documentation provided as part of this project, will hopefully help to supplement this new, official Neo4J WordNet implementation.

corpus, can hopefully be utilized in future Natural Language Processing research in psychology surrounding "linguistic distance."

Although not its primary motivation, this project can also enable future research that could make clear areas in which the Schwartz values paradigm may *not* be comprehensive. By subtracting out of the corpus words that have low "word-by-hop" values (defined below) for all Schwartz categories, future research could use the graph database approach used in this project to examine new areas of consistent speech across or within communities in the USA that could be considered values without fitting well in the current Schwartz values paradigm (e.g., by topic modeling the words that remain after removing those that are highly connected to Bardi et al's [2008] value lexicon words). The project could also, using the same approach, potentially demonstrate wide comprehensiveness of the current Schwartz values model, contributing to its wide body of existing evidence.

Research Questions of this Project

The current project seeks to answer two primary research questions, each of which comprises additional sub-questions.

Research Question 1. First, the current project asks whether the Schwartz values can be detected in obituaries, using a thesaurus-based expansion of Bardi et al.'s (2008) value lexicon approach (described below). To this end, it additionally asks the following questions:

1. Are all of the Schwartz values indicated in the corpus?

Given the nature of values in the Schwartz model as differentially prioritized across communities, it is reasonable to expect that not all Schwartz values will be equally represented in different communities; further, it is also reasonable to expect that even in aggregate, not all Schwartz values will be equally represented in the

obituaries corpus overall. Given that obituaries are subject to cultural taboos around topic-appropriateness, as discussed above, I expect that Hedonism and Power specifically will be represented less than other values in aggregate, if at all[11].

2. In what relative proportions are the Schwartz values talked about in the corpus?

3. For each of the 10 Schwartz values, what is the mean word-by-hop numerical value across the obituary corpus?[12]

Research Question 2. Second, the current project asks the following question:

To what extent is discussion of particular Schwartz values related to characteristics of obituaries, at the levels of individuals and their communities?

The current project seeks to answer this question by expanding Bardi et al.'s (2008) approach with word-by-hop calculations as a dependent measure.

Individual level.

[11]I do not mean to imply that Hedonism and Power, especially as defined in the Schwartz values model, are "bad" values, but rather that they specifically may be indicated by a larger number of potential anecdotes that would be less likely to be printed in an obituary than anecdotes that reference the other values.

[12]In the originally-proposed version of this dissertation, Research Question 1 contained two additional components, both of which "fell away" as nonsensical as the word-by-hop calculation described below developed. First, sub-question 3 asked what the mean word-by-hop value across all Schwartz values was. This became less relevant when word-by-hop became a relative measure (i.e., one that is not meaningful unless compared with another word-by-hop value). Second, a fourth sub-question asked what percentage of words found in the corpus are linked with Schwartz values. This question became less relevant as it became clear that the WordNet graph database described below was much more robust than expected: *almost all* words ($> 90\%$) in the corpus had a connection to at least one of the value lexicon words described in the Methods section. Thus, to clarify, these additional questions have not been ignored in this complete draft; rather, they have fallen away as the attention that was originally directed to them has turned to exploring the properties of word-by-hop, in consultation with members of the Committee.

Approximate age of the deceased Schwartz et al. (2001, p. 533) stated that age is positively related to Tradition, Conformity and Security, and that it is negatively related with Self-direction, Stimulation, and Hedonism. In line with these past findings, Schwartz et al. found in an international pair of samples that age correlated significantly with "self-transcendence (Benevolence, Universalism) and negatively with self-enhancement (Power, Achievement)." It is reasonable to expect that age should show similar main effects in the obituary corpus.

Gender Schwartz et al. (2001) found small correlations (primarily less than or equal to .11) between several of the 10 Schwartz values and gender (defined as male vs. female)[13]. The largest relationships reported by Schwartz et al. ($r > .11$) included women over men in Benevolence and Tradition values, and men over women in Stimulation. However, these were localized to specific cultures from those that were sampled (Israel, South Africa, and Italy, respectively). Similarly, Schwartz and Rubel (2005) found increased importance of Tradition in women over men, and of Hedonism in men over women, but inconsistently in both cases over four studies. Schwartz and Rubel presented evidence, however, that this inconsistency may have been a result of a sample (student vs. community) x measure (SVS vs. PVQ) interaction. When controlling for sample type and measure, Schwartz and Rubel (p. 1019) found that men consistently rated Stimulation, Hedonism, Achievement, and Self-Direction higher than women, while

[13]Schwartz and Rubel (2005) called this variable "sex" rather than "gender." The current project uses "gender;" thus, I have re-worded Schwartz and Rubel's report in this summary, despite the differences between the two concepts, given the rationale for using "gender" as a binary variable in this project. Since obituaries are written about individuals as they presented themselves to others, but typically lack the level of detail that would be required for more nuanced classification than a simplified "man" vs. "woman" distinction, the current project refers to "gender" but still uses a "sex"-like binary classification.

women consistently rated Universalism and (to a lesser extent) Security higher than men, especially in older (non-student) samples. As in Schwartz et al.'s (2001) report, however, the effect sizes for these findings were small, and were not always consistent across countries (for example, Schwartz and Rubel, p. 1022, noted, for Self-Direction, "The sex difference (men higher) is smaller the more autonomous versus embedded... and the more individualist... the culture of the country," and is "smaller... the richer the country").

Thus, I expected gender to have a minimal main effect in the obituary corpus. However, for communities that value Tradition highly overall, given the ritual nature of obituaries as values-affirming documents intended to be read by members of the community, it is possible that this particular corpus would show higher gender effects than have been previously observed. Perhaps more usefully, age and gender were allowed to interact in the models explored below, as well as gender with community education level. These follow Schwartz and Rubel's (2005) call for future research to examine interactions between "[gender] and such demographic variables as age, education,... and social class."

Community level.

Median income and median education level Although education and income[14] could be aggregated into a single measure conceptually approaching "wealth," they were instead both included individually in the models below and allowed to interact.

[14]The Gini coefficient of income inequality could be useful in these analyses, but was not included. Instead, I chose to use median income level because obituaries are almost certain *not* to reflect the full range of individuals on the wealth spectrum. Assuming that obituaries across locations tend to be of those who are at a certain level of socio-economic status and above, it is reasonable to expect that variability that would otherwise be related to the Gini coefficient would not be available to a statistical model for explanation when using this corpus.

Schwartz et al. (2001, p. 534) found positive, significant relationships between education (from none through college ["beyond high school"]) and "self-direction and stimulation[,] and negative correlations with conformity and tradition values." I therefore expected similar effects in the corpus of obituaries.

Valuing Power, by its definition above, includes a motivation toward "control[ling]... resources" (Schwartz et al., 2001, p. 5), which conceptually relates to income. Thus, higher income was expected to be positively related to Power, as well as possibly to Achievement (which, as above, includes a motivation to "[demonstrate] competence according to social standards," Schwartz, 2012, p. 5).

Ethnic/Racial demographics. Additionally, a coarse indicator of the ethnic makeup of geographic communities will be included, in five covariates, respectively reflecting the percentage of individuals classified as "American Indian or Native Alaskan," "Black or African American," "Hispanic," "Native Hawaiian and other Pacific Islander," and "White" (each in isolation or in combination with other categories) by the United States Census Bureau in that area (at the county level)[15].

[15]The US Census dataset from which these data were gathered (United States Bureau of the Census Population Division, 2015) uses variable names that sometimes conflate race and ethnicity (e.g., "Black or African-American," "White," etc.). Thus, the admittedly vague phrase "racial and ethnic groups/categories" is used throughout this report to accurately reflect the data used.

CHAPTER II

METHODOLOGY

Data Collection and Initial Processing

Lexical Graph Database.

WordNet. Bardi et al. (2008) composed a "value lexicon" of three "lexical indicators" (i.e., prototype words) for each of the 10 Schwartz values categories, for a total of 36 words (1 category title + 3 prototype words for each of the 10 categories, with some titles overlapping with prototype words). This Schwartz values dictionary is reprinted in Table 1.

Table 1. The Schwartz value lexicon, reprinted from Bardi et al. (2008).

Value	Lexical indicators for each value
Power	power, strength, control
Achievement	achievement, ambition, success
Hedonism	luxury, pleasure, delight
Stimulation	excitement, novelty, thrill
Self-direction	independence, freedom, liberty
Universalism	unity, justice, equality
Benevolence	kindness, charity, mercy
Tradition	tradition, custom, respect
Conformity	restraint, regard, consideration
Security	security, safety, protection

In order to quantify the extent to which obituaries in the current project's corpus were concerned with each of the 10 Schwartz values, I paired Bardi et al.'s

value lexicon with a network-graph-based dictionary/thesaurus of the full English language. In doing so, I enabled analyses of "lexical distance," as explained below (see the section defining "word-by-hop"), between words in any given obituary and the words from Bardi et al.'s value lexicon.

Beginning in 2005, and with updates periodically through 2010 (when it released its current version as of this writing, 3.1), Princeton University (2010) published WordNet, a computer-readable dictionary/thesaurus of 147,478 words grouped in 117,791 definitions. WordNet version 3.1 was downloaded in SQLite[1] format from the WordNet SQL (WNSQL) project (Bou & Princeton University, 2014), which repackaged the WordNet database for ready use. The database tables were converted into separate CSV files, which were then imported into Neo4J v2.3.3 Community Edition, an open-source graph database platform.

A traditional "relational database" (e.g., WordNet in its original form), which models data in a collection of linked "tables" (each conceptually equivalent to a spreadsheet, with ID numbers from one table recorded in columns in linked subsequent tables allowing related data to be joined during queries). In contrast, a graph database models data directly as nodes and relationships between them. Graph databases can thus be used more straightforwardly than traditional relational databases to find paths between nodes of different types (for example, from a word x through a series of definitions and words to a final word, y). This graph approach is visualized in Figure 1, which shows Bardi et al.'s "Power" words in green, with the words' definitions in blue.

The import of the WordNet tables into Neo4J expanded on work by Nagi (2013), who published a schema for representing the primary tables of WordNet

[1]SQLite is an SQL database format that stores the entire database in a single file, and can thus be more straightforward than other SQL database formats to manage and inspect.

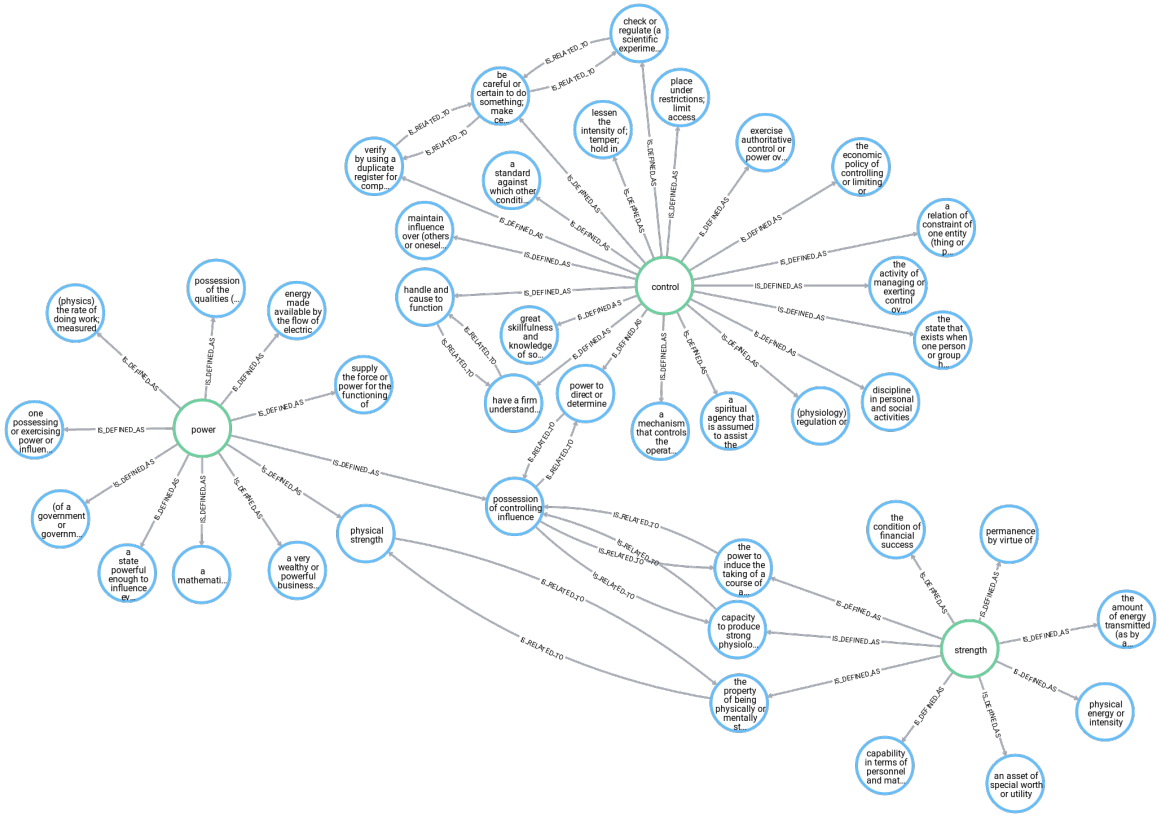

Figure 1. Bardi et al. (2008) value lexicon "Power" words ("power," "strength," and "control;" lighter/green) with their definitions (darker/blue). Words are connected to definitions, and some definitions are connected to other definitions.

in Neo4J. Nagi's schema specifically encompassed the Words, Senses, Synsets, Semlinks, and Linktypes tables of WordNet; the graph database constructed for this project additionally included WordNet's Morphs and Morphmaps tables. Lacking consistent documentation as of this writing, WordNet's tables are summarized here in Table 3. The graph database constructed for this project used three WordNet-derived relationship types ("Is a form of," "Is related to," and "Is defined as"), as well as one relationship type (not used in the analyses reported below) derived from the University of South Florida Free Association Norms dataset (Nelson et al., 1998, see below), "Is freely associated with." These relationships are summarized in Table 2. The full data schema is summarized visually in Figure 2.

Table 2. Node and relationship types in the WordNet graph database.

From Node	Relationship Name	To Node
Word	IS_FREELY_ASSOCIATED_WITH	Word
Morph	IS_A_FORM_OF	Word
Synset	IS_RELATED_TO	Synset
Word	IS_DEFINED_AS	Synset

Table 3. WordNet table names and descriptions.

Table Name	Table Description
Words	Lemmas (the basest forms of words – e.g., for "is," "be"; for "shouted," "shout")
Synsets	Part of speech, lexdomain reference, definition

Table Name	Table Description
Senses	Linking table for words and synsets
Postypes	Defines Part of Speech abbreviations
Semlinks	Links synset meanings with each other using a list of types of relationships
Linktypes	Definitions for the types of relationships between semlinks (e.g., "similar," "antonym,"[2] "hypernym," etc.)
Morphs	Gives different/alternative forms (e.g., noun, verb) of words in the Words table
Morphmaps	Records words and their different forms as written/recorded in the Morphs table

The "Postypes" table was not imported into the graph database, but was used when translating Part of Speech (POS) abbreviations between TreeTagger (described below) and WordNet.

University of South Florida Free Associations Network.

In addition to WordNet, the graph database utilized Nelson, McEvoy, and Schreiber's (1998) *University of South Florida word association, rhyme, and word fragment norms* dataset. As noted in the dataset's introductory webpage (http://w3.usf.edu/FreeAssociation/Intro.html), beginning in 1975, Nelson et al. (1998) asked approximately 6,000 participants each to list one word that the

[2]All linktypes were included when calculating hop numbers between obituary lemmas and value lexicon words, based on the understanding that (e.g., with antonyms), obituaries that include words that are the *opposite* of a given value *are* nonetheless invoking that value (even if with a negative disposition).

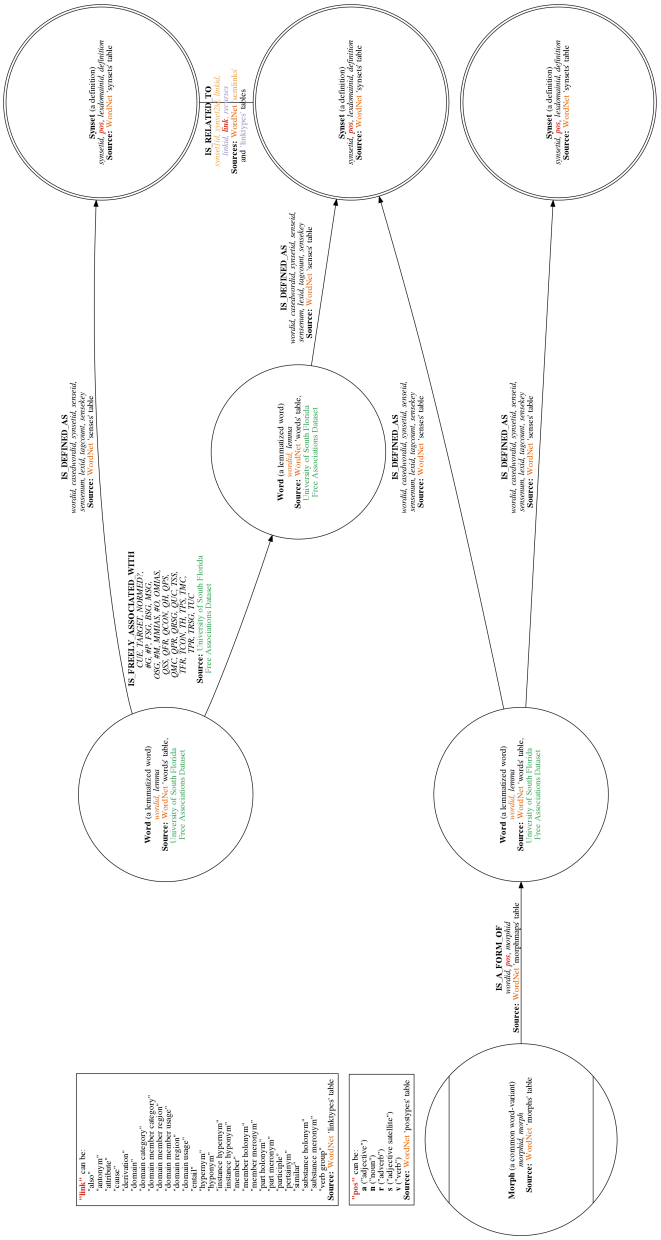

Figure 2. Visualization of Neo4J data model used in this project. Node and relationship attributes are printed in italics.

participant freely associated with each of a sample of 5,019 target words. These responses were then normed.

Because of its likely heavy dependence on the temporal and geographic context of its participants, Nelson et al.'s dataset was *not* used in the analyses described below; specifically, it was not used to generate "hop" numbers between lemmas from the obituary corpus and each of Bardi et al.'s value lexicon lemmas. However, because of the effort required to import the free association norms dataset into the graph database, as well as Dr. Nelson's generous permission to me (personal correspondence, May 26, 2016) to republish the dataset under a WordNet 3.1 license, this dataset *is* included in the supplemental material made available as part of this dissertation, and so is described here.

***Legacy.com* Newspaper Metadata.** *Legacy.com* is an international obituaries warehouser that contracts with 1,041 newspapers across 48 of the 50 United States, as well as Guam, to display newspapers' obituaries online[3]. Although, as discussed above, the copyright for a given obituary typically remains the property of the obituary's author, *Legacy.com* retains a transferable license to republish and use the content that is displayed on its site. In February 2015, I negotiated and signed a data use agreement with *Legacy.com's* Chief Marketing Officer, who graciously allowed me to scrape and analyze obituaries from *Legacy.com's* website for this and related projects. Thus, a series of downloading scripts was developed, first to scrape and process metadata about *Legacy.com's*

[3] *Legacy.com* notes on its "About Legacy.com, Inc." page (Legacy.com, 2016a; cf. *Legacy.com's* "Our International Newspaper Partners" page, Legacy.com, 2016c) that it partners with over 1,500 newspapers total, including newspapers in Canada, Australia, New Zealand, the United Kingdom, and Bermuda. *Legacy.com* claims in the "About Us" page of their Memorial Websites subdomain (Legacy.com, 2016b) that *Legacy.com* "partners with 76 of the 100 largest newspapers in the U.S. and features obituaries and Guest Books for more than 60 percent of people who die in the United States."

domestic affiliate newspapers, and then to locally download copies of obituaries displayed for each of those newspapers for further processing and analysis.

Newspaper Names and States Downloader. First, a downloader was written in Bash and used to scrape the *Legacy.com* map of affiliate newspapers[4] in order to record the website's internal state/location ID values. These ID values were then used to download each of a series of state- and territory-specific pages listing affiliate newspapers. The locally-downloaded files were subsequently scraped for each newspaper's name, unique *Legacy.com* internal ID shortcode, *Legacy.com* newspaper-specific website URL, and geographic state.

The geographic distribution of *Legacy.com* domestic newspaper affiliates can be seen in Figure 3.

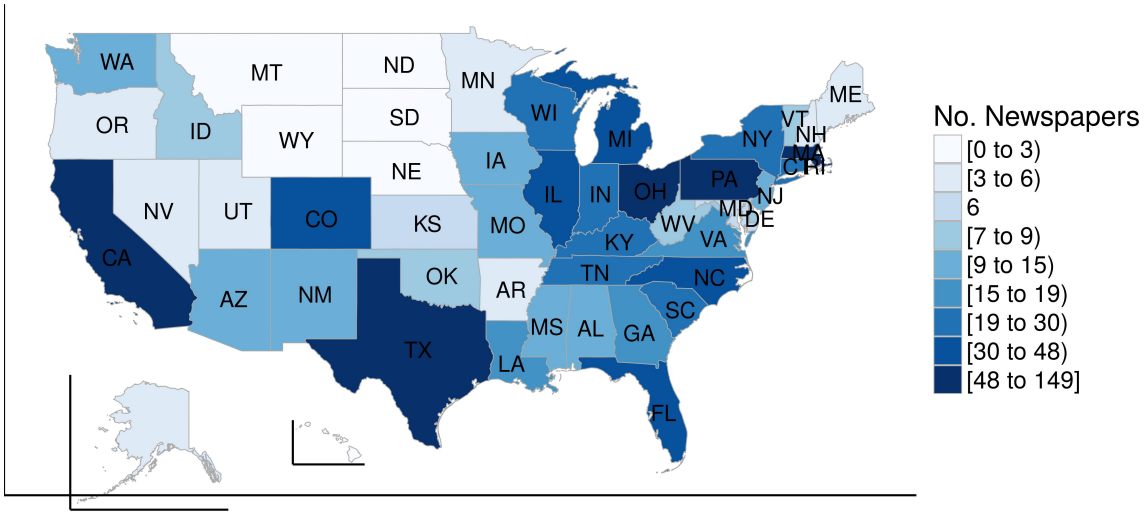

Figure 3. Newspaper sample size by state. WY and HI had 0 newspapers that contract with *Legacy.com*.

[4]As of this writing, the map is located at `http://www.legacy.com/ns/about/newspapers/`.

Finding Newspaper Distribution Points Using Google Maps. A script was then written in R[5] to query the Google Maps Application Programming Interface (API) in order to retrieve all addresses available for each newspaper's name and state abbreviation (e.g., in response to the search term "The Register Guard newspaper OR" for *The Register Guard* newspaper, which primarily operates in Lane County, which encompasses Eugene, Oregon). This approach was founded in the understanding that many newspapers have multiple counties of primary coverage (for example, *The Spokesman Review* newspaper has main offices in both Spokane county in Washington and Kootenai County in Idaho). Thus, this approach allowed incorporating all addresses returned by Google Maps for each newspaper; ideally, each of these addresses referred to a field or administrative office for a given newspaper.

For newspapers for which Google Maps returned more than six results (a threshold set simply by manually searching for several newspapers using the query format given above and noting at what approximate level search result sets became consistently filled with false positives), R's `agrep` function was used to selectively remove results. The `agrep` function allowed running a fuzzy (i.e., approximate) search among the list of returned business and place names for the "canonical" name taken from the *Legacy.com* pages that listed newspaper affiliates. `maximum.distance`, the argument that controlled the precision with which names were treated as matching the canonical name, was set to 0.3 after iterative trial-and-error tests with several high-search-result Google Maps queries (specifically, "The Arizona Republic," which yielded nine search results; the "Baltimore Sun," which yielded six search results; the "Times-News," which yielded 20 search results;

[5]The version numbers of all R packages used in this project that are not presented in-text can be found in the Appendix of this report.

and "The Birmingham News," which yielded 14 search results). This value was chosen with the intent to run searches that were conservative but not *overly* restrictive.

In these cases of high search results, a three-step process was used:

1. Always accept the first result (this was based on the assumption that Google Maps returned results ordered by relevance).

2. Search among the remaining returned business/place names for the canonical name of the newspaper, as defined using *Legacy.com's* affiliate newspaper list.

3. Filter for unique place names in the remaining Google Maps search set.

Each result's address, Google Maps place name, latitude, longitude, *Legacy.com* newspaper code, and date of download were recorded.

Of the original sample of 1,041 newspapers, 1.441% (15 newspapers) returned no search results, and thus were excluded from further analyses. Thus, 1,026 newspapers were included in the subsequent analyses. Excluded newspapers were not isolated to a single geographic region, as is shown in Table 4.

Table 4. State Abbreviations and Names of newspapers that did not return address results from the automated Google Maps query procedure and thus were excluded from further analyses.

State Abbreviation	Newspaper Name	Shortcode
CA	Pomerado News	pomeradonews
CA	Inside Bay Area	insidebayarea
IL	Chicago Sun-Times	chicagosuntimes
IL	Daily Southtown	daily-southtown
MA	The Country Gazette - Plainville	wickedlocal-plainville

State Abbreviation	Newspaper Name	Shortcode
MA	Falmouth Bulletin	wickedlocal-falmouth
MA	Waltham News Tribune	WalthamNewsTribune
MI	Muskegon Chronicle	muskegon
NJ	South Jersey Times	southjerseytimes
OH	The Marion Star	marionstar
OR	Hillsboro Argus/ Forest Grove Leader / Beaverton Leader	argus
PA	Susquehanna County Independent	independentweekender
SC	The Cheraw Chronicle	thecherawchronicle
VA	RichmondObitNews.com	richmond-VA
WI	Kenosha News	kenoshanews

Finding United States Census County Codes for Newspaper Locations Using the Federal Communications Commission's (FCC's) Data Conversions API. I then submitted the latitude and longitude of each unique recorded Google Maps search result to the Federal Communications Commission's (FCC's) Data Conversions API[6], a service provided by the FCC to translate location data such as latitude and longitude into United States Census Federal Information Processing Standard (FIPS) codes, unique numeric identifiers used by the United States Census Bureau for places down to the county level. FIPS codes are embedded in many US Census Bureau datasets and, further, are standardized geographically and temporally, unlike ZIP codes, the boundaries

[6]As of this writing, the API can be found at https://www.fcc.gov/general/census-block-conversions-api.

of which can be amended by the US Postal Service as postal requirements of a community change (put differently, FIPS codes seemed more appropriate than ZIP codes for use in this project given the relative stability of the former). The FCC's Data Conversions API provides a 15-digit FIPS code for a given latitude and longitude, which includes the following geographic identifiers:

- State (2 digits)
- County (3 digits)
- Census tract (6 digits)
- Block group (1 digit)
- Block (3 digits)

Thus, with an interest in the county level, the first five digits (State + County) of each location's FIPS code were recorded for further analyses.

Figure 3, a map displaying newspaper sample size by state, could be re-visualized with these county-level results as in Figure 4. Given the range of newspapers by county, Figure 4 is accompanied by Figure 5, which provides a histogram of newspaper frequency by county. Figure 3 is also accompanied by Figure 6, a visualization of US population by county in 2014, the latest year for which data was available at the time of writing (data is from the US Census Bureau Population Division's "Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2014" dataset, United States Bureau of the Census Population Division, 2015).

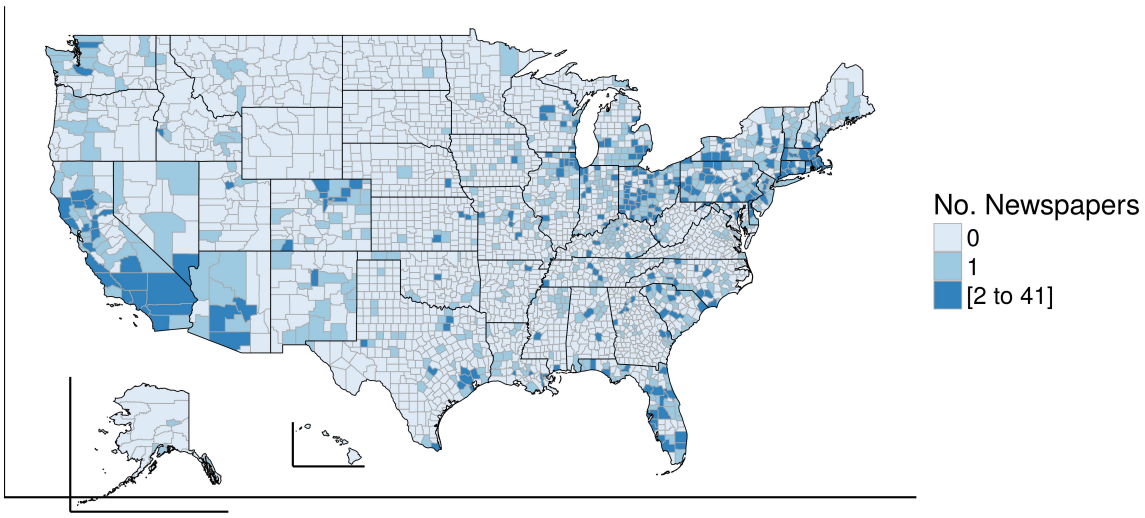

Figure 4. Newspaper sample size by county. WY and HI had 0 newspapers that contract with *Legacy.com.*

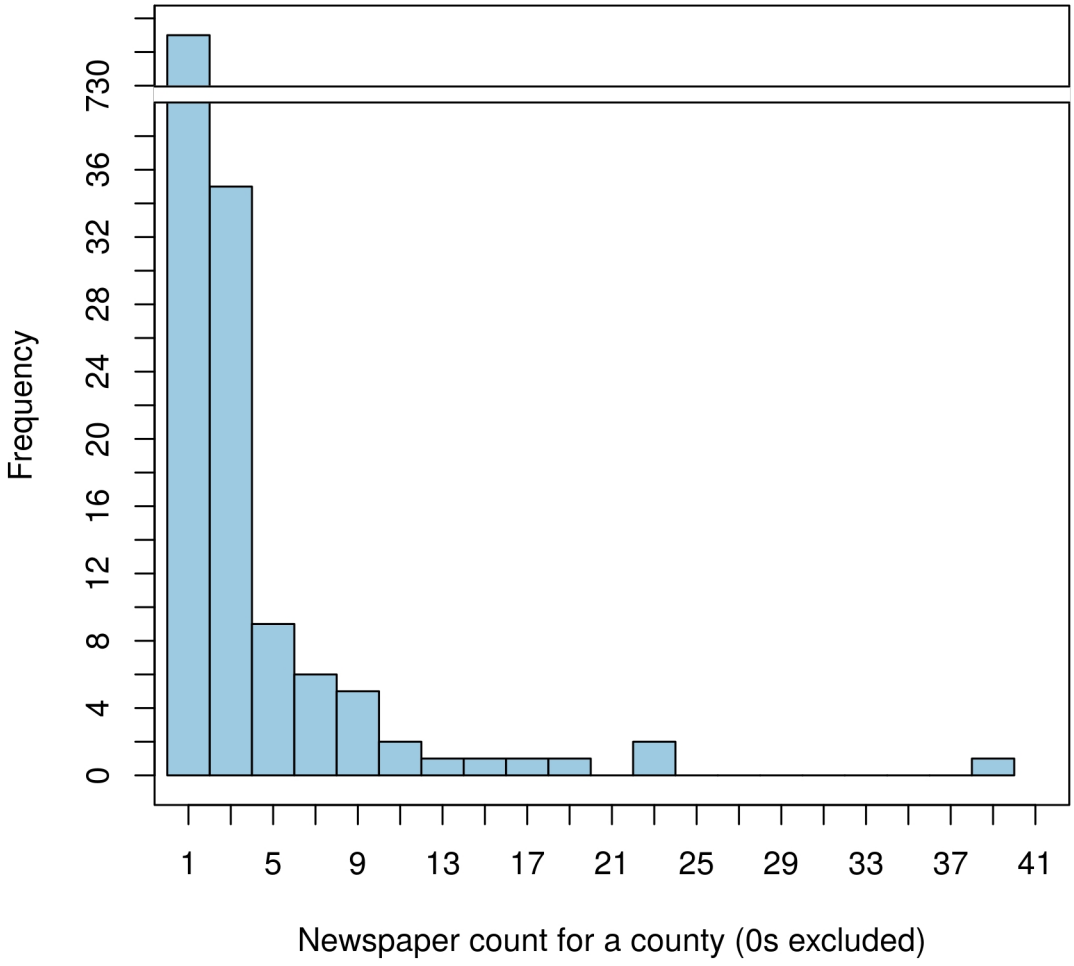

Figure 5. Histogram of newspaper sample size by county.

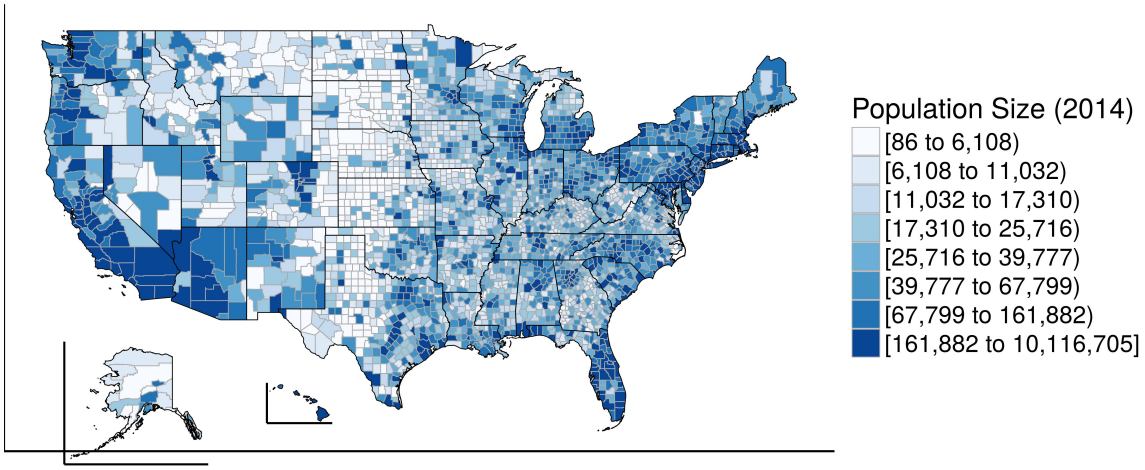

Figure 6. The United States' population by county in 2014, using the US Census Bureau Population Division's "Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2014" dataset (United States Bureau of the Census Population Division, 2015).

***Legacy.com* Obituaries Downloader.** A Bash script was written to use the *Legacy.com* newspaper-affiliate-specific URLs gathered as described above, appending to them query arguments to request the newspaper's "Last 7 Days" of results. Results from these requests were returned from *Legacy.com's* servers in pages of up to 10 results, and were wrapped in HTML <entryContainer> elements. Page numbers in the URLs being queried were incremented by 1 (e.g., "page=1", "page=2", etc.) until the most recently downloaded page contained no results (defined as a lack of <entryContainer> elements). The final (empty) downloaded page was then discarded, and the process re-started for the next newspaper. Downloads were rate-limited (all queries spaced two seconds apart) to minimize the intrusiveness of the process on *Legacy.com's* servers, and were launched during weekend nights. Downloads were run over seven consecutive weeks, on the following dates:

- Saturday, April 16, 2016
- Sunday, April 24, 2016
- Saturday, April 30, 2016
- Saturday, May 07, 2016
- Friday, May 13, 2016, to Saturday, May 14th, 2016
- Friday, May 20, 2016, to Saturday, May 21, 2016
- Friday, May 27, 2016, to Saturday, May 28, 2016

Combined, these downloads yielded 34,640 search result pages, totaling 9.1 GB of HTML data (i.e., the not-yet-scraped obituaries, as well as all of their HTML markup), including 313,607 obituaries[7].

***Legacy.com* Obituaries Data Scraper.** Using R, a scraper was developed to take the raw HTML content of each locally-downloaded search results pages and, for each result, extract the name of the deceased, the obituary text, the date of download, the URL of the individual obituary, and the name and state of the publishing newspaper. In addition, the scraper included algorithms for automatically coding the gender and age of the deceased (as necessary, calculating the latter by guessing date of birth from the information included in the obituary). Reliability statistics for these algorithms are included below.

Development of the Gender-Guessing Algorithm. The gender-coding algorithm utilized a two-pass approach to infer the gender of the deceased from the obituary text. Given the limited information available in obituaries and the fact that obituaries are very rarely written by the deceased and do not typically contain direct statements of gender identity, gender was coded using three categories: "female," "male," and "uncertain."

In the first pass, the gender algorithm counted the number of masculine ("he," "his," "him") and feminine ("she," "her") pronouns in the obituary text. If the obituary text contained more masculine than feminine pronouns, the gender of the deceased was coded as "male;" if more feminine than masculine pronouns were counted, the gender of the deceased was coded as "female."

In cases in which there were no gendered pronouns, or an equal number of masculine and feminine pronouns, the gender algorithm undertook a second

[7]It should be noted that, given date overlaps, this number includes some obituaries that were downloaded more than once. Pruning of obituaries is described in the Results section below.

pass, in which it attempted to use historical Census data to guess gender based on the forename of the deceased. The `gender` package for R was used for this purpose. While the package's `gender` function can tailor its guesses based on the approximate date of birth of the deceased, in order to keep the gender- and age-guessing algorithms separate from one another (such that imperfections in one would not unduly affect the output of the other), no date information was passed to the `gender` function, causing it to make a guess using Census data from across the timespan encompassed in its dataset.

In all cases, the gender-guessing algorithm recorded the source of its final output (pronouns vs. pronouns followed by forename), to facilitate debugging during the development process. Additionally, the gender algorithm was constructed to find and strip leading initials (e.g., "J. Henry Doe"), which were not usable by the `gender` Census-lookup function, from all names before sending them to the function.

Development of the Age-Guessing Algorithm. In order to guess age at death (or, as necessary, to calculate it by guessing year of birth), the age-guessing algorithm completed four separate searches through the text of each obituary and combined the information gathered from each to code age:

1. The lowest four-digit number (as a possible year of birth) in the obituary text
2. The first instance (if available) of the phrase "born… x," where x was a four-digit number.
3. A one- to three-digit number (as possible age at death) after a space and before a comma or period (for phrases such as "John Doe, 82, passed away…"). Candidate numbers needed to meet the following criteria:

1. Not be immediately preceeded by any long- ("January") or short-form ("Jan") month name or a one- to four-digit number followed by a space or forward-slash (e.g., "10/12," which likely signified a date rather than an age).
2. Be immediately preceeded by a space.
3. Be immediately followed by either one or more spaces or a comma.
4. Not be immediately followed by zero or more spaces followed by "AM," "am," "PM," "pm," or any combination of these abbreviations using uppercase and lowercase letters. This search also checked for versions of these abbreviations that included periods (e.g., "p.m."). This search avoided recognizing numbers in phrases such as "5 pm") as candidate ages.
5. Not be immediately followed by whitespace or a comma and then the year of the obituary's download (2016). This search avoided recognizing, e.g., the "9" in "April 9, 2016," which was assumed to be more likely a date of death or of scheduled funerary services. This search had the notable downside of systematically missing children who died as infants at less than one year. This decision was made after reading several hundred obituaries manually and seeing few infant obituaries, especially with a date of birth printed explicitly.

4. A two- to four-digit number that might be a year in the US-common date form of MM/DD/YY[YY] or MM-DD-YY[YY].

The age algorithm then removed any candidate year of birth numbers (not candidate age at death numbers) higher than the year of download (e.g., 2016) or

lower than the year of download minus 110, which was expected to be a liberal estimate of the highest age at death to be found in the corpus. Since no individuals were expected to be older than 110[8], potential age at death numbers higher than 110 were thus assumed to be misidentified, and were discarded.

The algorithm then considered all of the gathered candidate years for each obituary, and for each obituary, took the *lowest* candidate year as the likeliest year of birth. This choice was based on the assumption that, since obituaries typically craft a narrative around the life of the deceased, the earliest date was likely to be the correct one to use.

Following the rationale described above for systematically excluding guesses for infant children less than one year old at death, the age algorithm declined to make an age guess if the only guess to make was the year of download (e.g., 2016).

In cases in which the algorithm made a guess for year of birth but not for age, approximate age was calculated using the equation

$$year\ of\ publication - year\ of\ birth = approximate\ age$$

Validation of the Gender- and Age-Guessing Algorithms. The algorithms for categorizing gender and age did not need to be perfect; they did, however, need to be approximately as accurate as would be expected from a human coder working as part of a team (for example, a Research Assistant).

[8]Following a sample in Spain, Gómez-Redondo and García González (2010, p. 164-165) concluded that "there is no relationship between the population size of the regions or provinces, and the number of supercentenarians[, individuals who are 110 years or older]." With that in mind, it can be useful to consider historical rates of supercentenarianship. Kestenbaum and Ferguson (2010) found 325 individuals who had died at age 110 or greater between 1980 and 2003, a span of over two decades in a country of a minimum of 226.5 million (at the beginning of that span, in 1980, as reported by Forstall & United States Bureau of the Census Population Division, 1996).

Using a "development" corpus of 27,923 additional obituaries that were downloaded and scraped separately by approximately two weeks from the larger "production" corpus (on June 11th, 2016; this "development" corpus was *not* included in the final analyses reported below; rather, it was downloaded and scraped specifically for use in the code development required to enable those analyses), I randomly selected and manually coded 50 obituaries for gender and age or date of death (whichever was easiest given the information presented in each obituary) before looking at the guesses made by the automated algorithms. This allowed me to see initially whether the algorithms needed major adjustment. Seeing that inter-rater reliability estimates were high with this sample of 50 (following calculations described below), I took a new random sample (with replacement) of 100 obituaries. This new sample was coded separately (and, as before, without knowledge of the automated algorithms' output) by two coders: a member of the Dissertation Committee (Mark Alfano) and me.

Initial Data Processing Steps for Calculating Reliability Statistics. In order to increase the meaningfulness of the reliability statistics to be computed, the matrix of automatically- and manually-coded genders and ages was first processed using the following steps. These steps were performed separately between each of the two humans' codes and the automated algorithm's codes.

1. I replaced all blank values with 0s, in order that rows in which the computer or the human coder did not have enough information to make a guess would not be skipped in the calculations. This had the effect of penalizing cases in which the human coder made a guess and the automated algorithm did not,

and rewarding cases in which both the computer and the human coder agreed that there was no basis for a guess.[9]

2. For rows in which the automated algorithm and human coder's age guesses were within one year of each other, the two guesses were considered the same (by setting them equal to one another). This step was warranted because for both the human and computer coders, age guesses were sometimes calculated from a guess of year of birth. In those cases, age was calculated as "2016 [the current year] - year of birth". This equation yields a result one year off in cases in which the birthday of the deceased has not yet occurred in the current year.[10]

Gender Coding Reliability. Cohen's κ, computed between the automated algorithm and me was 0.886, and was 0.881 between the automated algorithm and Mark Alfano for gender codes. Percent agreement was 94% between the automated algorithm and each of the human coders.

Age Coding Reliability. Cohen's κ for age codes between the automated algorithm and Mark Alfano was 0.803, and was 0.865 between the automated algorithm and me. Respectively, percent agreement was 81% and 87% between the automated algorithm and the two coders. Respectively, the correlation of age guesses between the automated algorithm and the two coders was 0.760

[9]This processing step had the downside of making Pearson's correlation statistics less meaningful in this context (nevertheless, Pearson's correlations are still reported below, for completeness), in cases in which the computer made a guess and the human coder did not, or vice versa.

[10]For example, if an individual's date of birth were February 20th, 1987, and the person were 28 years old at the time of death, this equation would yield a date of birth one year off (2016 - 1987 = 29), until after February 20th (when the individual would have turned 29). I considered the potential one-year difference acceptable for the analyses described later in this report; for assessing reliability of the coding algorithm, however, it seemed necessary to allow a one-year tolerance when evaluating computer and human age guesses as matching.

and .830. While this statistic is reported for completeness, it is not as useful as the above statistics in this context, given that the data processing steps described above could cause disproportionate decreases in any case in which the automated algorithm guessed the age of the deceased in an obituary that a human coder decided did not contain sufficient information to support a guess, or vice versa.

Filtering of Obituaries by Word Length. Obituaries published in the USA follow two primary formats (defined as endpoints on a continuum). First, some obituaries comprise short-form death notices, which often contain the deceased's name and age or date of birth, but no or almost no biographical content. These short-form obituaries seem primarily motivated to serve as public notices and to provide logistical information about funerary rites for the deceased (e.g., the date and time of a wake or funeral, locations to which to send flowers, and the name of the managing funeral home). Short-form obituaries of this type might also contain limited familial information (using phrases such as "Survived by..."). Long-form obituaries, which might be several hundred words in length, by contrast, contain all of the information expected in a short-form obituary, but also include a richer biographical sketch of the deceased. In some newspapers, such as *The New York Times,* long-form obituaries can be commissioned and professionally-authored, reading much more like biographies possibly to the exclusion of more detailed logistical or familial information. Many obituaries, however, are written by friends or family members of the deceased (and occasionally by the deceased him- or herself in advance of death, perhaps during a long illness), either following templates provided by the publishing newspaper or an assisting funeral home (supplying unique biographical information but in a consistent format), or composed wholly by the author.

I expected that short-form obituaries, as death notices, would likely not contain information useful for understanding values that might differ from community to community, or from author to author; short-form death notices can be thought of as demonstrating shared values about notifying community members about deaths, sharing with community members in funerary rites, and enumerating ties to still-living relations, but these values are, by the very fact that obituaries are published across the USA, not likely to vary or be unique to particular communities nearly as much as richer biographical information. For this reason, I filtered obituaries from the corpus gathered over the seven-week recording period to contain as few short-form death notices as possible. Since short-form necrologies are defined in part by their length, obituaries were filtered based on word count. It should be noted that this filtering decision was quite granular: it is possible, for example, that an obituary could comprise a brief biographical sketch and eschew any logistical funeral information in such a way that it would not solely be a death notice but would still be excluded from the final corpus based on the word-count filter. I made the decision to use word count as a useful indicator simply from having read many obituaries manually and not remembering having seen any short-form obituaries that did not contain only logistical and basic familial information. While constructing a more advanced algorithm for pruning obituaries could be useful fodder for future research, the decision can be seen as valid for this project in its pragmatic, bottom-up approach (from manual reading), which enabled progression towards the project's larger questions and goals.

In order to set a word-count threshold below which to exclude obituaries from further analyses, I qualitatively examined a series of obituary word-count histograms of successively finer granulation (two of these are reproduced here as

Figures 7 and 8). An initial histogram of word-counts across the corpus showed a notable drop-off in obituaries just below the 100-word mark. Creating further histograms with smaller x-axis ranges and finer-grained bins suggested that a qualitative change might take place in obituaries (from the highly-templated and thus consistent-length short-form death notices to longer-form obituaries in which authors wrote about the deceased in more detail and with more idiosyncrasy) at lengths somewhere between 40 words and 120 words.

I thus took random samples (for each sample, $n = 30$) of obituaries within ranges of 10 words, starting with the range of 21 to 30 words, and manually read every obituary in the sample to look for non-templated biographical content. The intent of this approach was to stop sampling in this way and set a filter cutoff as soon as I began to see biographical information (even in just one or two obituaries in these small samples) in order to set the cutoff as conservatively as possible. This approach relied on the assumption that a random sample of 30 obituaries (from a larger corpus of several thousand for that 10-word range) would be sufficient to manually detect this qualitative change in the corpus from all or almost all short-form death notices to longer-form biographical obituaries. While this approach was admittedly imprecise, it was also highly functional. After examining samples of obituaries in the ranges of 21 to 30 words, 31 to 40 words, 41 to 50 words, 51 to 60 words, and greater than 100 words, I set the filter cutoff at 60 words.

Construction of a Matrix of Lemma - Part-of-Speech Distances from Bardi et al.'s Dictionary. Of the original 313,607 obituaries in the corpus gathered during the five-week collection period, 211,963 obituaries were above the

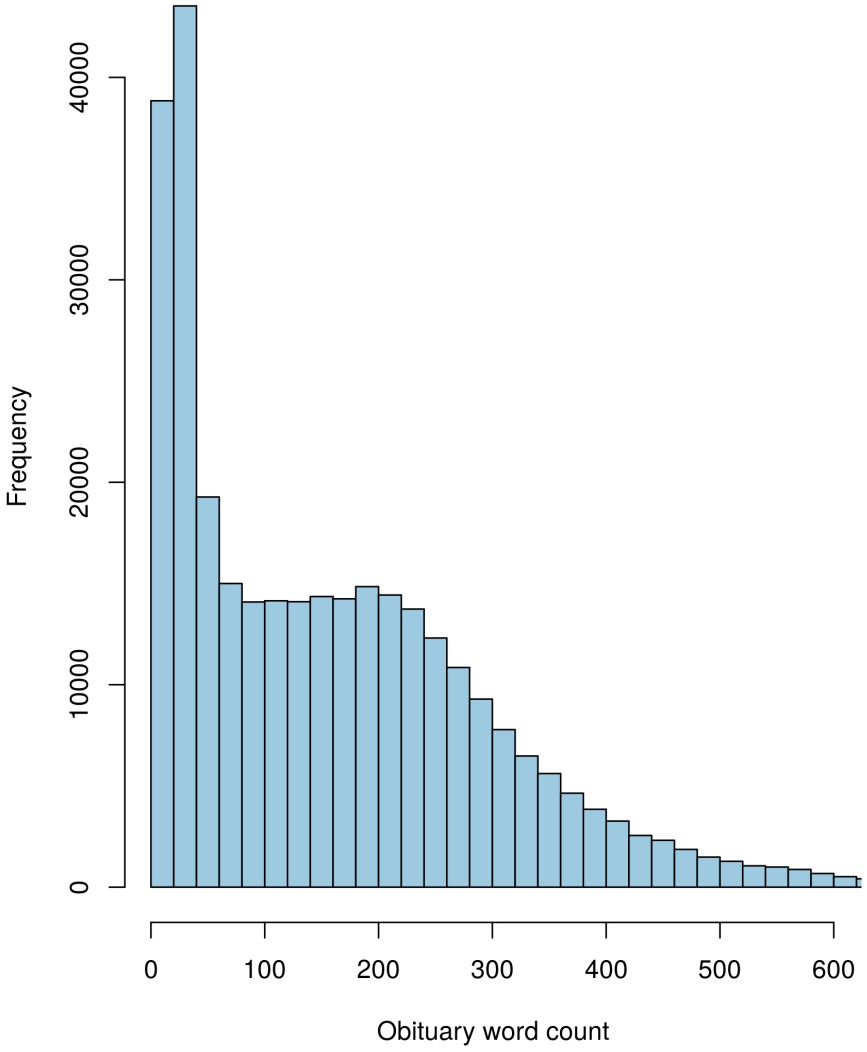

Figure 7. Histogram of obituary word counts, cut off at 600 words (the actual maximum word count in the corpus was 3,502 words, at the end of a long positively-skewed tail in the full histogram's distribution).

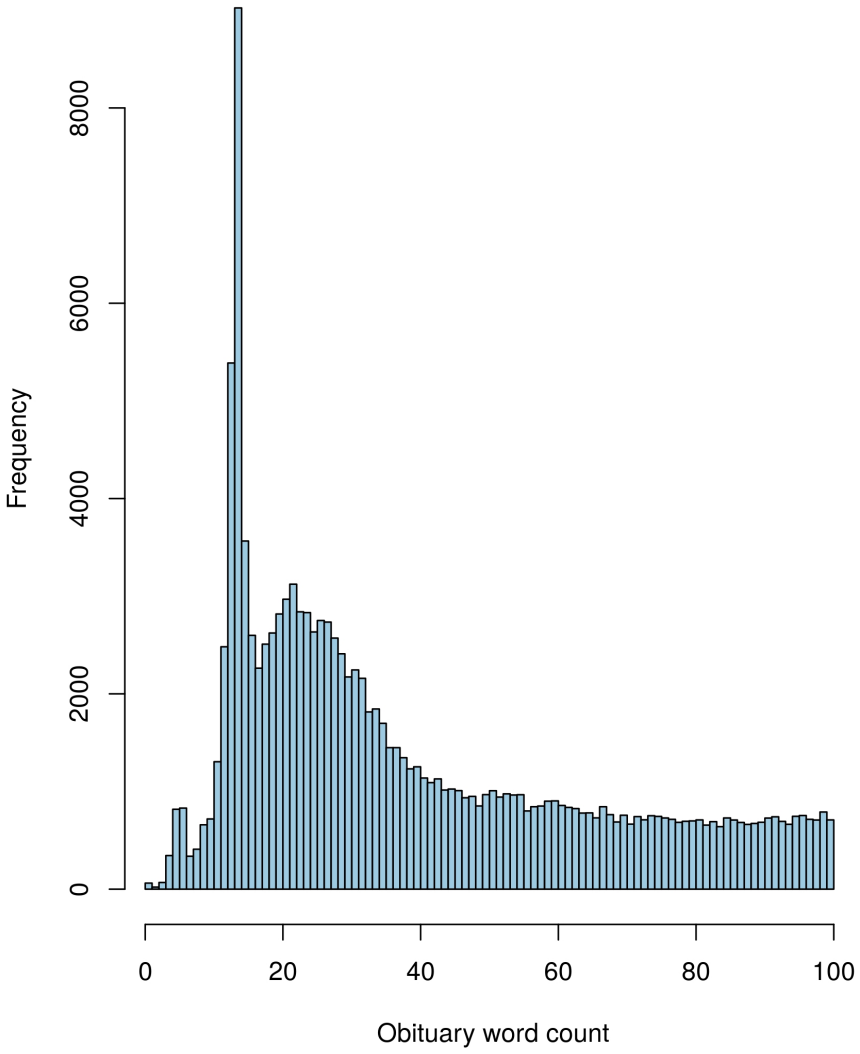

Figure 8. Histogram of obituary word counts, cut off at 100 words (the actual maximum word count in the corpus was 3,502 words, at the end of a long positively-skewed tail in the full histogram's distribution).

minimum word count cutoff, and were thus retained for further analyses[11]. For each of these remaining obituaries, the following steps were taken:

1. The text of the obituary was lower-cased (e.g., "John Doe was very brave" became "john doe was very brave") in order to normalize text across obituaries to correct for possible inconsistencies in capitalization, whether from typos or noun usage (e.g., "...was a member of the Methodist church" [a particular church building in an area] vs. "...was a member of the Methodist Church" [an overall religious organization]).

2. All words in the obituary were tagged for Part of Speech (POS) using TreeTagger (Schmid, 2013), an open-source, cross-platform language tagging and processing tool.

3. TreeTagger's POS dictionary (i.e., the list of Parts of Speech it had available for use) was substantially more detailed than WordNet's. Thus, in order to enable compatibility with the WordNet graph database (the usage of which is described below), TreeTagger POS tags were "translated" into WordNet POS tags, using a conversion table (Table 5).

Table 5. TreeTagger to WordNet Part Of Speech (POS) conversions.

TreeTagger POS	WordNet POS
verb	v
adjective	a, s
adverb	r

[11]The "development" dataset was similarly cleaned: of its initial 23,946 rows, 33.241% were removed for being under the minimum word length, leaving 15,986 obituaries.

TreeTagger POS	WordNet POS
noun	n

The TreeTagger "adjective" tags were given two matching WordNet tags, signifying WordNet's "adjective" and "adjective satellite" categories. In the WordNet query used in the steps below, these two tags were combined in an "OR" statement (e.g., "match any word in the WordNet database that is tagged as an adjective OR as an adjective satellite.").

4. TreeTagger was used to "lemmatize" all of the words in the obituary. Unlike coarser "stemming" methods (e.g., using the Porter stemming algorithm[12] – see, e.g., Porter, 1980), which normalizes words in a corpus by removing endings (changing "speaking" and "speaker" into "speak," and "happy" and "happiness" to "happ"), "lemmatizing" seeks to change a word into its most basic form, given its identified POS (e.g., for "saw" as a noun, "saw;" for "saw" as a verb, "see;" for "was" as a verb, "is"). Because it requires knowledge about POS, lemmatizing is more difficult than stemming, but also potentially more useful for its increased nuance (assuming the accuracy of the POS tags that it uses).

5. Words of the following Parts of Speech were removed from the list of lemmatized words, on the assumption that they would not be relevant for calculating distance from Bardi et al. words:

[12] As of this writing, the official website for the Porter stemming algorithm is at `http://tartarus.org/~martin/PorterStemmer/`. The citation provided here is for one of the algorithm's canonical publications.

1. "Cardinal numbers" (e.g., "76") were excluded, since the "lemma" value of each was replaced by TreeTagger with the string "@card@", and so was not expected to contribute to this project's analyses.
2. "Modal" lemmas (e.g., "will" in "will follow") were also excluded on the expectation that they would also not be relevant (especially as compared to the verbs they were paired with) for this project's analyses of distances from value lexicon words.

6. Stopwords were removed, following the R `tm` ("TextMining," v. 0.6.2) package's standard English stopword list: "i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "would", "should", "could", "ought", "i'm", "you're", "he's", "she's", "it's", "we're", "they're", "i've", "you've", "we've", "they've", "i'd", "you'd", "he'd", "she'd", "we'd", "they'd", "i'll", "you'll", "he'll", "she'll", "we'll", "they'll", "isn't", "aren't", "wasn't", "weren't", "hasn't", "haven't", "hadn't", "doesn't", "don't", "didn't", "won't", "wouldn't", "shan't", "shouldn't", "can't", "cannot", "couldn't", "mustn't", "let's", "that's", "who's", "what's", "here's", "there's", "when's", "where's", "why's", "how's", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under", "again", "further", "then",

"once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very"

7. Punctuation was removed using the following list: ".", ",", [single quote], [double quote], ";", ":", "!", "[", "]", "{", "}", "/", "_", "&", "$", "#", "!", "(", ")".

8. For each lemma-POS combination (e.g., "see" used as a verb), the WordNet graph database was queried to find the shortest distance through the graph between the lemma and each of the 36 words in Bardi et al.'s dictionary. Candidate paths were capped at 15 total hops (counting all node types). Within this 15-hop cap, separate counts were performed for the following (all from the same shortest path):

    1. Total number of hops
    2. Number of hops only counting Word nodes
    3. Number of hops only counting Morph nodes
    4. Number of hops only counting Synset nodes
    5. Number of Synset-to-Synset relationships
    6. Number of Word-to-Synset relationships
    7. Number of Morph-to-Word relationships
    8. Number of Free Association relationships (this was expected to be zero in every case, given the decision to exclude `IS_FREELY_ASSOCIATED_WITH` relationships from the current project's analyses, and was included

to both check that expectation, and to provide ready code for future projects that release this constraint)

Put differently, in the path `x:Word --> y:Synset --> z:Word --> a:Synset --> b:Word` (where "x:Word" means "a node, x, which is a Word"), the count of Word nodes between x and b would be 2, and the count of total hops would be 4. The cap of 15 total hops was based on an arbitrary but arguably common-sense decision that words more than 15 hops distant from each other would not be reasonably defined as "related" in a meaningful way in the context of this project[13]. The lemma's POS tag was used to filter the first Synset (i.e., definition) encountered in the shortest path (in plain language, this aspect of the query would have read, "Find the shortest path between 'saw' and 'benevolence,' where the first Synset node encountered is tagged as a 'verb.'"[14] Lemmas that were not successfully POS-tagged by TreeTagger were also queried, without the first-Synset filter (allowing any POS/definitions in the first hop). Since the words in Bardi et al.'s dictionary were not POS tagged by Bardi et al., their definitions in the graph were not filtered during these queries. The Bardi et al. words were not

[13]Although in this case doing so would be preferable, as of this writing, Neo4J is not capable of filtering to, e.g., 15 *Word* hops specifically, without setting a maximum constraint on total number of hops at an arbitrarily larger threshold. Setting such a threshold would be computationally expensive without necessarily succeeding in allowing the desired maximum number of Word hops, e.g., if a network path between a lemma and a value lexicon word contained enough Synset-to-Synset relationships.

[14]Unlike the constraint imposed on the first Synset encountered in the shortest path between a given obituary lemma and a target value lexicon word, I decided *not* to constrain the final Synset encountered before a value lexicon word to be a noun (i.e., the POS type of all of the value lexicon words: per Bardi et al., 2008, p. 485, "In order to hold part of speech constant across different values, we considered only nouns. For example,for the value Hedonism, the noun pleasure was used rather than the adjective pleasing."). While Bardi et al's value lexicon words are all nouns, given the exploratory aspect of this project, I decided not to impose this constraint in order that traversal through the network initially be allowed to be liberal rather than conservative (put differently, to allow more rather than fewer connections). However, POS of the final Synset *was* recorded when computing shortest paths, in order to explore whether the final Synset POS was notably variable without this constraint.

lemmatized during this graph query process, as they were understood to have been published as lemmas.

The graph database query was parallelized using an Amazon Web Services Virtual Machine running Ubuntu Linux 14.04 with 40 virtual CPUs and 160GB of memory. Query results were appended to a 12-column matrix:

1. From (the lemma in the corpus)
2. To (the target value lexicon word)
3. Part of Speech of first hop Synset (the POS of the lemma as it was used in the corpus)
4. POS of final Synset encountered (used to check whether not constraining the final Synset to be of a noun POS was likely to affect analysis outcomes)
5. Total number of hops in the shortest path
6. Shortest number of hops, only counting Word nodes
7. Shortest number of hops, only counting Morph nodes
8. Shortest number of hops, only counting Synset nodes
9. Number of Synset-to-Synset relationships in the shortest path
10. Number of Word-to-Synset relationships in the shortest path
11. Number of Morph-to-Word relationships in the shortest path
12. Number of free association relationships in the shortest path (this was used to check the constraint that free association relationships not be allowed in queries for this project, as explained above)

Lemma-POS combinations that did not connect to any Bardi et al. words were appended with a null value in all count columns, in order that they not be looked up again from subsequent obituaries.

Calculation of Schwartz Value Distances from Lemma - Bardi et al. Word Distances. The final matrix of terms showed that the filtered obituary corpus comprised 98,499 unique lemma-POS combinations, with 1,148,983 total rows (each for a unique lemma-POS-Bardi et al. word combination)[15]. The matrix of these combinations was grouped by the Schwartz values of the value lexicon words and aggregated using a new calculation, "word-by-hop," which is defined below.

[15]The "development" dataset contained 42,784 unique lemma-POS combinations, with 635,124 total rows (each for a unique lemma-POS-Bardi et al. word combination).

CHAPTER III

RESULTS

Further Cleaning and Pruning the Dataset

Beyond the initial data "cleaning" steps described above in the Methods section, I took several steps to better organize the dataset to make it more ready for use in analyses.

Pruning Within-Newspaper Duplicate Obituaries. As noted above, overlap in obituary download dates caused some obituaries to be downloaded twice. In addition, several obituaries were downloaded repeatedly from the same newspapers because they appeared over time in multiple forms in *Legacy.com's* search results. Several of these forms are non-exhaustively listed below:

- "Teaser" obituaries, which seemed to serve as placeholders for full, biographical texts, but presented only logistical death notice information (e.g., time and location of funerary services). These obituaries also sometimes contained a note that the full obituary would be published on a given date.
- Variants of the same text, but including a nickname of the deceased (e.g., "John Doe…" vs. "John 'Johnny' Doe…")
- Variants of the same text, but including spelling corrections (especially for place names).

To remove these duplicates, I grouped obituaries by newspaper and obituary text (i.e., condensing by exactly-matching obituary text strings)[1]. This

[1]I compared this method to grouping by newspaper, obituary text, and name of the deceased. The two methods produced a difference of 218 rows. I then compared the names of the deceased

removed approximately 16% of the original dataset's rows, leaving 177,272 rows. I then further condensed the dataset using the URL of each obituary (as each obituary listed in a *Legacy.com* search result page contains a link to a unique page comprising only that obituary). 1.66% of the remaining dataset contained a duplicate URL (e.g., from the "teaser" obituaries mentioned above). After confirming that no URLs were duplicated *across* newspapers (cross-newspaper duplicates are discussed below, as a separate area of consideration), for each URL in the dataset, I retained the obituary with the longest word count, assuming it to be the most updated version of the obituary text. This removed an additional 1,482 rows (0.836% of the remaining dataset) from the dataset (a portion of the 1.66% mentioned above), leaving 175,790 rows[2].

Removing Obituaries of Individuals of Unknown Age. Of the remaining dataset, 6.256% did not contain sufficient information for the automated age-coding algorithm to make a guess. Given the relatively low percentage of the dataset lacking an age guess, and the intention to use age as a predictor in the regression models below, I excluded these data from further analyses, rather than attempt (likely problematically) to impute age values. This left 164,792 rows in the dataset[3].

---

for each pair of rows that contained matching obituary text but non-matching names (in order to confirm that obituaries of two separate people had not both been written from similar templates, e.g.), and found that in all cases, the paired obituaries qualitatively obviously referred to the same individuals, differing only in punctuation, capitalization, or other minor differences in the printed name of the deceased.

[2]The "development" dataset was similarly cleaned: of its remaining rows, 0.181% were removed, leaving 15,957 rows (zero rows were removed from the development dataset for containing duplicate URLs).

[3]The "development" dataset was similarly cleaned: of its remaining rows, 5.634% were removed for being of uncertain age, leaving 15,058 rows.

Removing Obituaries of Individuals of Uncertain Gender. Of the remaining dataset, 0.439% (724 rows) were assigned a gender category of "uncertain" by the automated coding algorithm. Given the low percentage of data in this third category (vs. "female," which 48.867% of the remaining dataset was coded as; and "male," which 50.694% of the remaining dataset was coded as), I excluded those rows from further analyses. This left 164,068 rows in the dataset[4].

Assessing Across-Newspaper Duplicate Obituaries. A frequency table of obituary text strings (i.e., collapsing exactly-matching obituary texts and counting the frequency of each) revealed that 20.294% of the remaining dataset were duplicate obituaries *across newspapers* (within-newspaper duplicates having been removed following the description above; a check was also performed at this step to confirm that no within-newspaper duplicate obituary texts remained in the dataset).

These duplicate obituaries were not isolated to only a small number of newspapers: 393 papers in the sample printed at least one duplicate obituary. This is not conceptually problematic: it is understandable that an obituary be published in more than one newspaper, depending on the notoriety or width of the social circle of the deceased, and/or the number of newspapers operating simultaneously in the same geographic region. However, from an analysis perspective, this overlap required consideration. Steps taken to accommodate these duplicate obituaries are described further below.

[4]The "development" dataset was similarly cleaned: of its remaining rows, 0.578% were removed for being of uncertain gender, leaving 14,971 rows (before removing these unknown rows, the percentages of obituaries coded as female vs. male were 49.416% and 50.001%, respectively).

Analyses

Calculation of a Measure of Network Distance, "Word-by-Hop".

Because not all lemmas contained in the obituary corpus were related to all Schwartz values (or, more specifically, to all of the words from Bardi et al.'s [2008] value lexicon), it was necessary to calculate a numeric value for each word's relationship with each of the value lexicon words that was able to represent "no relationship." Number of hops through the WordNet graph, while intuitive to use when answering Research Question 1, was inapposite for use especially in the regression context of Research Question 2, since it was only able to represent a lack of relationship (i.e., no path within 15 hops between two lemmas) with missing values; "0" hops, rather than signifying no relationship, signified a *perfect* relationship, that the lemmas being compared were identical. Thus, in a regression context, lemmas that were unrelated to given Bardi words would need to be either a) excluded from analyses in either a listwise or pairwise fashion, or b) imputed with an arbitrary value (such as 100). Neither of these solutions seemed satisfactory; therefore, a new calculation, called "word-by-hop[5]," was used.

Word-by-hop values are not meaningful in isolation, but *can* be compared to one another to determine the relative fit between different pairs of lemmas. Larger word-by-hop values represent better "fit" with a Schwartz value, where fit is defined as "greater connection," rather than solely by "fewer hops."

Final Calculation Formula for Word-by-Hop. This report first presents the final calculation for "word-by-hop," in order that it be presented as early as possible. Following that, it explains the step-by-step development of the calculation.

[5] "Word-by-hop" is a shorthand for "Words divided by median hops."

The equation for word-by-hop is as follows:

$$\text{word-by-hop} = \frac{\frac{\text{Number of words that have a path to Value v}}{\text{Total number of words in the obituary}}}{1 + \text{Median number of hops of the connected words to Value v}}$$

Word-by-hop has boundaries at [0, 1][6].

Initial Calculation of Word-by-Hop. In its first stage of development, word-by-hop was initially calculated by counting the total number of lemmas (including repeats) in an obituary that have any path to a given word from Bardi et al.'s value lexicon (i.e., in this project, from 0 to 15 hops) and dividing that count by one plus the median number of hops in each lemma's shortest path to the given word. In cases in which there are no connections, word-by-hop was set to 0.

The addition of one to the median number of hops corrects for cases in which there is complete overlap between obituary and value lexicon words (i.e., a median hop number of 0) – rather than dividing by 0 or a fraction less than 1 (which would multiply the count of words unlike for non-zero median hop values), an ideal word-by-hop value is simply not penalized, by dividing by 1.

An example word-by-hop calculation between a hypothetical obituary that comprises only a single lemma and three Schwartz values is illustrated in Tables 6 and 7.

Table 6. Example data showing the number of hops through WordNet between "business" and words for three Schwartz values from Bardi et al's (2008) value lexicon

Lemma in Obituary	Value Lexicon Word	Number of Word Hops
business	novelty	6

[6]The notation is meant to denote a "closed" interval, i.e., one that includes its endpoints.

Lemma in Obituary	Value Lexicon Word	Number of Word Hops
business	excitement	5
business	thrill	4
business	charity	(No path)
business	kindness	(No path)
business	mercy	(No path)
business	tradition	(No path)
business	custom	7
business	respect	4

Table 7. Example implementations of the *initial* word-by-hop calculation, using data from Table 6.

Schwartz Value	word-by-hop calculation	word-by-hop value
Stimulation	3 words ÷ (1 + median[6, 5, 4]) hops	0.500
Benevolence	0 words	0
Tradition	2 words ÷ (1 + median[7, 4]) hops	0.308

Table 7 illustrates that the hypothetical obituary (which, in this example, contains only one word, "business"), is more connected to Stimulation (word-by-hop = 0.500) than to Tradition (word-by-hop = 0.308), and that it is not connected to benevolence (word-by-hop = 0), using this initial word-by-hop calculation.

This process can be aggregated over as many lemmas in the obituary that have a path to words from the value lexicon. Other, more informal, example calculations illustrate the properties of this initial version of word-by-hop:

- For a given Schwartz value, 1 connected word ÷ (1 + median 3 hops) = .250, whereas 5 connected words ÷ (1 + median 3 hops) = 1.250 (a higher value). Thus, the calculation considers obituaries that contain larger numbers of connected words "more connected" to a Schwartz value than those that contain smaller numbers of similarly-connected words.

- For a given Schwartz value, 7 connected words ÷ (1 + median 3 hops) = 1.75, whereas 4 connected words ÷ (1 + median 2 hops) = 1.333. Thus, the calculation considers obituaries that contain a larger number of connected words at a slightly larger median graph distance "more connected" to a Schwartz value than those that contain a smaller number of connected words at a slightly smaller median graph distance. This is unlike using median hop count only, as hop count would prefer the latter case over the former.

This word-by-hop approach is partially conceptually similar to Bardi et al.'s (2008), which included a preference for more vs. fewer indicator words (Bardi et al. only counted pages in their newspaper corpus that included all three indicator words for a given value).

The initial word-by-hop calculation defined above required several corrections for use in the current project, however, as described below.

Correcting Word-by-Hop Values for Discrepant Numbers of Value Lexicon Words. Bardi et al.'s (2008) value lexicon includes three indicators for each Schwartz value. For some Schwartz values, but not all, the *name*

of the Schwartz value could be included as an additional (fourth) indicator: for some Schwartz values, Bardi et al.'s lexicon uses the value's name as an indicator (e.g., for Power, "power," "strength," and "control"), while for others, the value's name is not included (e.g., for Hedonism, "luxury," "pleasure," and "delight"). Thus, in total, Bardi et al.'s value lexicon contains 36 words, comprising four Schwartz values of three indicators, and six Schwartz values of four indicators (when the value's name is included)[7].

Word-by-hop calculations could be affected by the inclusion of a fourth indicator for a given value by providing additional opportunities for connections to a value lexicon word. Thus, the analyses described below were run with only the three primary indicator words for each Schwartz value[8].

Correcting Word-by-Hop Values for Discrepant Numbers of Words across Obituaries. Calculating word-by-hop values using the count of words that have a connection to a given Schwartz value through Bardi et al.'s value lexicon potentially preferences longer obituaries over shorter obituaries.

This preference could be conceptually problematic, given that newspapers may charge fees to publish obituaries based on word count (potentially confounding wealth with word-by-hop calculations). This was remedied by replacing count of words that have a connection to a Schwartz value with that count divided by the total number of words in the obituary (i.e., using percentage of total words in

[7]Bardi et al. (2008, p. 485) explained this discrepancy in their report, stating, "A preference was given to use the value label itself (e.g., *power, security*) as one of the words to represent each value; however, this was not always possible because some of these value labels yielded prohibitively low word frequencies (e.g., *universalism, hedonism*)."

[8]Future follow-up work could repeat the analyses, including the fourth indicator words to examine the results' robustness. Given Bardi et al.'s (2008, p. 485) note that the value titles that were not used as indicators were not included because they were seen infrequently in their corpus, however, those value titles' addition can likely be expected not to change the results dramatically.

the obituary, and subsequently dividing that percentages by one plus the median number of hops of those words)[9].

This adjustment has the secondary effect of causing the word-by-hop calculation to be bounded at [0, 1], as can be seen in these examples:

- An obituary with "perfect" connection with a Schwartz value would be one that contained only words that exactly matched those from the value lexicon for that Schwartz value: for example, 10 words (including repeated words) that exactly correspond with value lexicon words for a given value ÷ 10 total words in the obituary ÷ (1 + 0 median hops) = 1.000

- 15 words (including repeated words) that have a connection to a given value ÷ 120 total words in the obituary ÷ (1 + 4 median hops) = 0.025, whereas 10 words (including repeated words) ÷ 120 total words in the obituary ÷ (1 + 4 median hops) = 0.0167, representing that the first case is "better connected" to the value than the second.

- 15 words (including repeated words) that have a connection to a given value ÷ 120 total words in the obituary ÷ (1 + 4 median hops) = 0.025, whereas 10 words (including repeated words) ÷ 120 total words in the obituary ÷ (1 + 3 median hops) = 0.021, representing that, as above, the calculation still considers obituaries that contain a larger number of connected words at a slightly larger median distance "more connected" to a Schwartz value than

[9]Alternatively, I could have replaced count with count divided by the *number of words in the obituary that have a connection to* any *Schwartz value* (i.e., using *percentage of values words* in the obituary, and subsequently dividing that by one plus the median number of hops of those words. However, this could build into the calculation the possibly hazardous assumption that that all obituaries are equally laden with Schwartz values overall.

those that contain a smaller number of connected words at a slightly smaller median distance.

Word-by-hop is conceptually related to several other measures of graph connectedness. Unlike in-degree and out-degree, which respectively measure the number of incoming and outgoing edges from a given node in a directed graph, word-by-hop ignores the direction of relationships in the graph, allowing bidirectional traversal between source and target lemmas. Word-by-hop does not consider the centrality of a given node, since it is primarily concerned with the existence and distance of relationships between a source node and a given set of target nodes, rather than whether a given node is a hub for other nodes; however, it does incorporate some elements that are similar to those used to calculate betweenness centrality, the equation for which is $\frac{\text{how many of the paths include node } n}{\text{number of shortest paths between node } a \text{ and node } b}$ (assuming that nodes a, b, and n are not the same) (see, e.g., O'Neil & Schutt, 2013, p. 258). Word-by-hop's equation could be re-written $\frac{\text{number of shortest paths between obituary word } a \text{ and the value lexicon words } b, c, \text{ and } d \text{ for a given value}}{\text{the median distance of those shortest paths}}$; like betweenness centrality, word-by-hop incorporates information about the number of paths between a set of nodes. Although word-by-hop is not the same as measures of cross-clique centrality, it does share the approach of assessing the extent to which a given node connects to "cliques" of nodes (here, defined using the value lexicon, which is divided following the Schwartz values paradigm; it is not necessarily the case, however, that all words in the value lexicon for a given value connect to one another).

Two potential limitations of word-by-hop are that it does not take into account the clustering coefficient of nodes (i.e., how prone to clustering an

obituary's words are) nor the baseline frequency of words as they are used in written English. These are both additions that would be useful to consider in future projects.

Statistically Modeling Word-by-Hop Values. Because it is a proportion, word-by-hop is theoretically bounded at [0, 1]. Although word-by-hop values of 1.0 are almost certain never to be seen in this type of corpus (as it would require that every word in an obituary exactly matched a word from the value lexicon for a particular Schwartz value), word-by-hop values of 0 *are* realistically possible; these values would indicate that a given obituary has no words that have any connection (within 15 hops through the lexical graph database) to any of the value lexicon words for a given Schwartz value. Although no word-by-hop values of 0 were observed in the dataset used for this project (see below, in response to Research Question 1), it is worth pausing to consider the analytic approaches that would be best suited to modeling this new measure, in case zeros *are* observed in future datasets to which it is applied.

The [0, 1] boundaries of the corrected word-by-hop calculation defined above could be problematic to use in a typical multiple regression (including multi-level multiple regression) context, for two reasons. First, multiple regression could poorly predict word-by-hop as a dependent measure if, as is reasonable to expect, much of the dataset pools at the bottom of its range. Put differently, the expectation that there will be a substantial number of values at and near 0 could cause difficulty for a regression model, which would otherwise mathematically assume that the model's Dependent Variable (DV) has a theoretical range from $-inf$ to inf and thus might present predicted values that are beyond the actual boundaries of the DV (Ferrari & Cribari-Neto, 2004). Second, beyond pooling near 0, word-by-hop values could

be considered more aptly described by a beta distribution than by the normal distribution assumed in a typical multiple regression context. Beta distributions can take a variety of both normal and non-normal shapes, and are bounded at (0, 1) (i.e., between 0 and 1, without including values of exactly 0 or 1).

For datasets that include word-by-hop as a dependent measure but do not include values of 0 or 1, several analytic possibilities exist. Using a Generalized Linear Model (GLM), word-by-hop values may be modeled directly as beta-distributed, using, for example, the `betareg`, `glmmADMB`, and `zoib` packages for R, each of which allows beta-distributed dependent variables as well as the inclusion of random effects (here, e.g., the newspaper that published a given obituary). This approach is described in the literature as a "beta regression," and was introduced by Ferrari and Cribari-Neto (2004). Alternatively, the data may be modeled using a wider variety of tools (such as R's `glmer` package) using a logit link function and a binomial distribution (University of California, Los Angeles Institute for Digital Research and Education, 2016). A test model had trouble converging using this approach, however, with the full obituaries dataset, possibly due to a large number of small (non-zero, but near zero) word-by-hop values[10].

As a third alternative, the data may be logit-transformed, using the formula $\log \frac{\text{word-by-hop}}{1-\text{word-by-hop}}$. The logit-transformation will re-scale values to be bounded at $(-inf, inf)$, allowing them to be used in a normal multi-level regression.

For future datasets that *do* include word-by-hop values of 0, the DV will not be able to be directly modeled by either a beta regression as described above or a normal regression using logit-transformed data, as in the former case,

[10]I include this speculative explanation because applying a test transformation to word-by-hop values such that all values were moved up by a constant until the maximum value was just below 1 caused the warning produced when modeling the raw word-by-hop values with this approach to cease.

the beta distribution cannot include zeros, and in the latter case, the normal distribution cannot include $-inf$ (which is the result of logit-transforming 0). Ospina and Ferrari (2010; cf. Ospina & Ferrari, 2012) proposed a "zero-inflated" beta regression for this case, which simultaneously models non-zero values of the DV as beta-distributed and DV values of 0 as binomial-distributed. The `zoib` package for R allows for zero-inflation (Liu & Kong, 2015); however, it is built on `JAGS` and `rjags`, which perform Markov-Chain Monte Carlo (MCMC) estimation, and thus performed very slowly in tests even for simple models when adding a random effect for newspaper with the development dataset of ˜15,000 obituary rows. As of this writing, R's `glmmADMB` package also allows for zero-inflation, but, per its documentation, only when modeling outcome variables using Poisson or binomial distributions.

For future datasets that do include word-by-hop values of 0, in contexts in which performing a zero-inflated beta regression is either not desirable or (as with this project) logistically untenable because of dataset size causing large model convergence times, Smithson and Verkuilen (2006, p. 55) proposed transforming the data using the formula $\frac{y\times(n-1)+\frac{1}{2}}{n}$, "where n is the sample size," in order to slightly inflate values of 0 and slightly deflate values of 1 (cf. the R `betareg` package vignettes documentation, Cribari-Neto & Zeileis, n.d., p. 3, which quotes Smithson and Verkuilen).

The current project featured 10 DVs, each comprising word-by-hop values for a different Schwartz value. Thus, each regression model that I sought to test in the current project needed to be run 10 times, once for each DV (e.g., in an iterative model-building exercise in which seven models were compared to one another, 70 total regressions would need to be run). Numerous tests indicated

that, even under conditions of parallelization, all beta regression options for R described above performed too slowly on the full dataset to be usable, particularly when adding a random intercept term for newspaper (even without random slope components for age of the deceased, gender of the deceased, and their interaction). Thus, for the regression analyses described below, word-by-hop values were logit-transformed and then passed into R's `lme4` package's `lmer` function, which is designed for more traditional multi-level modeling.

Logit-transforming the data in this way presents several potential issues. First, any deviation from modeling the data directly (i.e., as beta-distributed) is open to a rightful conceptual criticism that interpretation of the resulting models' output may be more difficult. In this case, however, it is useful to consider that word-by-hop is a new measure, which *could* be made to include the logit transformation in its definition. More strongly, word-by-hop values are not meant to be interpreted in isolation, but rather relative to one another (as higher values indicate greater "connectedness" to the target words). As higher word-by-hop values correspond to higher logit-transformed word-by-hop values, the potential issue of interpretation is minimized in this case. Second, logit-transforming data does not cause them to become normally-distributed. Thus, while word-by-hop values for several Schwartz values do appear to be normally distributed, several appear binomially-distributed at least in some newspapers (See Figures 23, 24, 25, and 26, which are further explored below); output from models that utilized those non-normal word-by-hop variables should be interpreted with this in mind. Future work could also usefully compare the results of using this analysis approach vs. a beta regression approach on a subset of word-by-hop data.

Multiple Membership Analysis Rationale. As noted above, 20.294% of the remaining dataset comprised obituaries that had been published across more than one newspaper. Because "communities" in this project were defined by newspaper, these data would ideally be modeled as "multiple members[11]" of the newspaper-communities (henceforward called simply "newspapers") that published them, following Leckie's (2013) terminology. Incorporating the multiple membership of obituaries in newspapers would include explicitly modeling obituaries (at level 1 of the model), collapsed by their text strings such that duplicate obituaries were all counted as a single row, as potentially having been published in more than one newspaper (at level 2 of the model), rather than being strictly nested each within one newspaper. Following the approach advocated by Leckie (2013), this process could also include weights to show each obituary's percentage of "membership" in each of the newspapers that published it (e.g., weighting could be done by the newspapers' counties' relative population sizes). In this case, a random-intercepts regression model with age and gender as obituary-level predictors and newspapers' counties' median income as a newspaper-level predictor would be of the following form:

[11]Leckie (2013) and the Center for Multilevel Modelling at the University of Bristol use the term "multiple membership" to describe this data structure. In other literatures, the same structure is alternatively called "mixed membership" and (in some cases) a "non-nested" design.

$$
\begin{aligned}
&\text{word-by-hop}_{\text{Schwartz value}_v} = \\
&\beta_0 \\
&+ \beta_1 age_o \\
&+ \beta_2 male_o \\
&+ \beta 3\overline{\text{newspaper median income}_o} \\
&+ \sum_{n \in newspapers(o)} w_{n,o}^{(2)} u_{0n}^{(2)} \\
&+ e_o
\end{aligned}
$$

In this equation, each β functions as in a typical regression, as a measure of the expected logit-transformed word-by-hop change for a one-unit increase in the associated predictor, holding all other predictors at a constant level (e.g., 0). "(2)" superscripts denote terms at level 2 of the model (the newspaper level). $\overline{\text{newspaper median income}_o}$ denotes the median income of the counties in which a newspaper is published. Following Leckie's approach and terminology, this term would be defined for each obituary as the weighted sum of the income values of each of the newspapers that published the obituary:

$$
\overline{\text{newspaper median income}_o} = \underbrace{\sum_{n \in newspapers(o)}}_{\text{Sum over obituary o's newspapers...}} \underbrace{w_{n,o}^{(2)} \text{newspaper median income}_o}_{\text{For each newspaper, weight the median income (e.g., by population)}}
$$

$u_{on}^{(2)}$ would denote a random effect at the newspaper level (i.e., the variance attributable to newspaper membership).

The term $\sum_{c \in counties(o)} w_{c,o}^{(3)} u_{0c}^{(3)}$ could also be added to the first equation above to explicitly model newspapers as being multiple members of counties (rather

than manually computing a weighted average by population size for each newspaper across the counties in which it operates, as was done for this project's analyses).

Unfortunately, as of this writing, software available for conducting analyses was not readily able to combine this type of multiple membership approach with the beta regression approach described above, or, in the case of R, to readily incorporate multiple membership data structures generally. Given that it was not logistically feasible (although it would be ideal) to incorporate multiple membership in either a weighted or non-weighted way, two primary options for analyses were considered. First, the multiple membership structure of the data could be ignored. In this approach, obituaries that were in fact duplicates across newspapers would be treated in each of their instances as unique. Previous research has demonstrated that ignoring multiple membership structure underestimates higher-level variance while overestimating level-1 variance (see Leckie, 2013), which would be especially problematic in this case, since the newspaper level was of primary interest. Alternatively choosing to consider only one instance of each obituary would be, according to Leckie, a "naive" approach (p. 3); nonetheless, this alternative approach seemed preferable to proceeding while ignoring the data's multiple membership structure. Systematically selecting a single obituary for each set of duplicates (e.g., by the newspaper serving the counties with largest populations) would be conceptually problematic, as it would build into all downstream models the assumption that obituaries are most intended for as wide an audience as possible (vs., e.g., being intended first for a smaller community, with which the deceased may have had particularly close ties, and secondly for a larger community). Thus, each set of across-newspaper duplicate obituaries was randomly downsampled to one instance of the obituary. This approach did

essentially discard data; however, it retained as much data as possible while not ignoring the data's multiple membership structure. Conceptually, this approach is equivalent in Leckie's schema to randomly weighting newspaper membership to 0, 0, and 1, vs. an (e.g., arbitrary) weighting scheme of $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$, for an obituary published in three newspapers.

As mentioned above, county-level predictors for the analyses below were manually aggregated to the newspaper level by creating averages for each newspaper weighted by county population size. The newspaper-in-county data structure could *itself* be modeled using a multiple-membership approach. Failing to do this likely resulted in shunting county-level variance elsewhere in the model (hopefully, but not necessarily, to the newspaper level). However, since counties were *not* the focus of this project (and were originally incorporated with the erroneous understanding that newspapers primarily would operate in only one county), this decision, which also simplified the regression models used below, seemed appropriate. With this in mind, future work focused more directly on these decisions' analytic implications could determine the extent to which choosing not to include a random county-level variance component substantially alters conclusions. Future work could also re-downsample the duplicate obituaries and re-run the regression models below in order to explore the analyses' robustness to this approach.

Covariate Descriptive Statistics. After randomly down-sampling obituaries published in multiple newspapers to single instances (as described above), the dataset contained 140,599 obituaries, nested in 832 newspapers. The remaining newspapers and their associated obituaries were eliminated for the analyses presented below because county-level information for them was unavailable

(i.e., for the remaining newspapers, an SQL `JOIN` of the newspaper data onto the county-level data using the Google Maps-derived county FIPS codes as shared identifiers returned no matches).

Age at death (as coded by the automated algorithm described above) ranged from 1 to 110 ($M = 73.5$, $SD = 21$). The percentages of obituaries coded as having been written about women vs. men in the final sample were approximately even, with 49.1% female and 50.9% male. 71.3% of gender codes (which were all assigned using the automated algorithm described above) were made solely using counts of gendered pronouns in the obituary text, while 27.9% were made using the `gender` package for R. 0.892% of codes were assigned using a combination of pronoun-counting and the gender package. Descriptive statistics for all newspaper-level predictors (mixed ethnic/racial demographic variables; county population size; median household income; and education, operationalized as percent of the population with a Bachelor's degree or higher) are presented in Table 8.

Table 8. Descriptive statistics for newspaper-level covariates.

Variable	min	max	mean	sd
Race/Ethnicity: "White"	20.210	99.08	79.634	13.687
Race/Ethnicity: "Black & African American"	0.569	79.14	16.157	13.251
Race/Ethnicity: "American Indian & Native Alaskan"	0.400	41.31	1.639	2.302
Race/Ethnicity: "Native Hawaiian & Pac. Islander"	0.020	3.14	0.273	0.334
Race/Ethnicity: "Hispanic"	0.496	88.69	12.639	12.852
Income (divided by 1,000 for this table)	26.351	108.48	54.245	12.115
Percentage with Bachelor's degree or higher	7.300	64.00	30.046	8.886

Figure 9 visualizes the densities of median household income (United States Bureau of the Census, 2015b) and education level (United States Bureau of the Census, 2015a). Education was operationalized as percentage of the population with a Bachelor's degree or higher after comparing its distribution across counties to the percentage of the population with a high school degree or higher, and finding the variability of the former to be greater than that of the latter, and thus of greater use in a regression context.

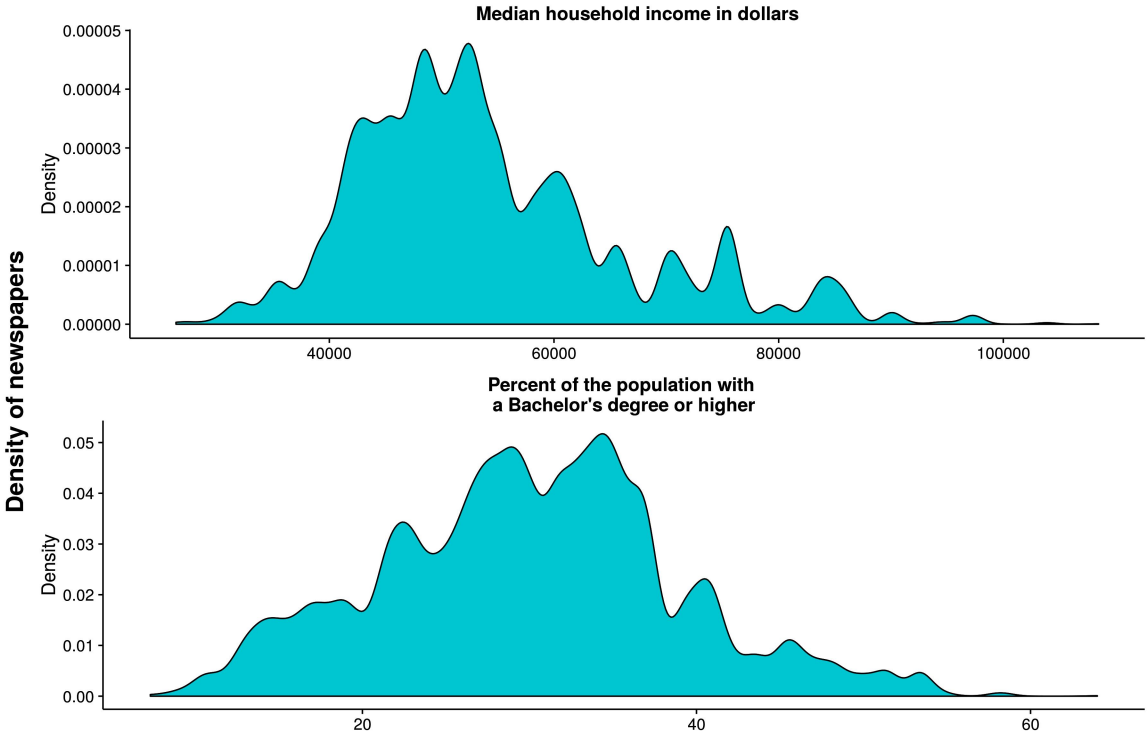

Figure 9. Density curves for median household income and education level (operationalized as percent of the population with a bachelor's degree or higher) from US census data averaged across counties for each newspaper (after weighting by county population size).

Figure 10 visualizes the densities of racial and ethnic groups across newspapers (United States Bureau of the Census Population Division, 2015).

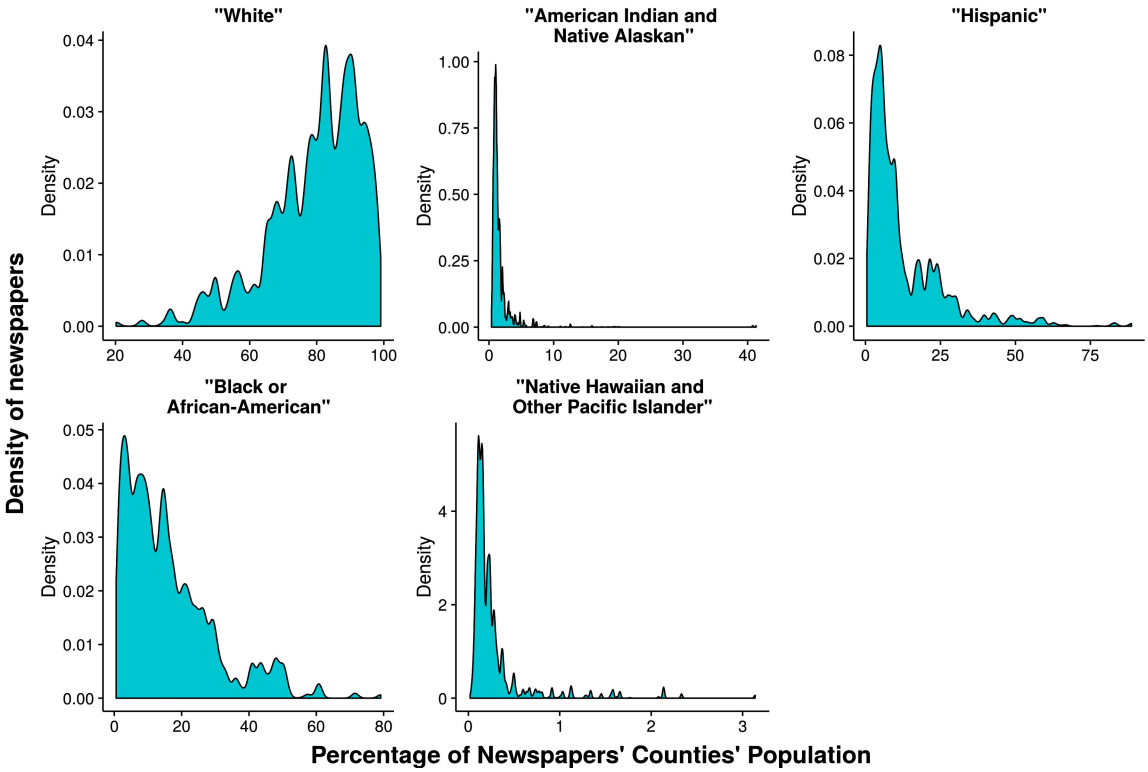

Figure 10. Density curves for each race or ethnicity variable from US census data averaged across counties for each newspaper (after weighting by county population size). Category titles are taken from the Census dataset, and are of that category "Alone or in Combination" with other race/ethnicity categories (i.e., an individual included in the "White" category could have self-described herself as both "White" and "Hispanic").

Exploring the Properties of the New Word-by-Hop Measure.

Because word-by-hop is a new measure, it is worthwhile to pause here to explore its properties before reporting statistics directly relevant to the project's two major research questions.

As noted in the Methods section above, for each shortest path between an obituary lemma-POS combination and a target value lexicon word, eight hop counts were recorded, each following a different definition of "shortest path" that could be relevant as part of this project:

1. Total number of hops
2. Number of hops only counting Word nodes
3. Number of hops only counting Morph nodes
4. Number of hops only counting Synset (i.e., Definition) nodes
5. Number of Synset-to-Synset relationships
6. Number of Word-to-Synset relationships
7. Number of Morph-to-Word relationships
8. Number of Free Association relationships

In order to facilitate understanding of word-by-hop computations, raw hop numbers for each of these definitions of "shortest path" are summarized in Table 9. The number of free association relationships was 0 across shortest paths, as expected; as noted above, because free association edges were excluded from the graph queries for this project, this descriptive statistic was included as a check that the query was working properly. Morph nodes were seen infrequently in the

shortest paths, as also expected, given that Morphs are uncommon variants of Words, and only connect to Words.

Table 9. Descriptive statistics for hop numbers in the shortest paths between obituary lemma-POS combinations and value lexicon words.

	min	max	mean	sd
Total Number of Hops	0	15	8.305	1.516
Shortest Number of Hops Only Counting Words	0	7	1.723	0.710
Shortest Number of Hops Only Counting Morphs	0	1	0.000	0.017
Shortest Number of Hops Only Counting Synsets	0	14	6.581	1.468
Number of Synset to Synset Relationships in Shortest Path	0	13	4.859	1.737
Number of Word to Synset Relationships in Shortest Path	0	14	3.445	1.420
Number of Morph to Word Relationships in Shortest Path	0	2	0.001	0.034
Number of Free Association Relationships in Shortest Path	0	0	0.000	0.000

The liberal decision not to constrain the final Synset node encountered in each shortest path when calculating hop numbers through WordNet can be evaluated with reference to Table 10, which indicates that although in the majority (˜95%) of cases, the final Synset encountered was of POS noun, some paths did finish on verb or adjective Synsets. This table suggests that allowing the final Synset POS to be unconstrained was likely useful given the exploratory nature of this work, but that future work that chooses to add a noun constraint to the final Synset POS encountered in each shortest path would likely not affect hop counts (and thus derived word-by-hop values) substantively.

Table 10. Percentage of final Synset (i.e., definition) node POS in shortest paths between obituary lemma-POS combinations and value lexicon word (excluding cases in which no path existed).

POS	Frequency as Percentage
a	0.105
n	94.501
v	5.393

Of the eight measurements above for shortest path length, the two that I considered most relevant for this project were number of hops only counting Word nodes, and number of hops counting only Synset-to-Synset relationships, as they seem especially to follow lay definitions of word closeness[12]. Thus, for each obituary, the word-by-hop value for each Schwartz value was calculated and recorded twice, first using Word node counts, and then using Synset-to-Synset relationship counts. These two word-by-hop calculations for each Schwartz value are compared below, both in response to Research Question 1 and because these comparisons created a foundation for choosing one computation over the other for use in the regressions reported below in response to Research Question 2. Table 11 presents the minimum, maximum, mean, and standard deviation for each computation of word-by-hop.

[12]Free association also seems to me to follow a lay definition of word closeness; for this project's analyses, however, free associations were not included in the graph, following the rationale above.

Table 11. Descriptive statistics for word-by-hop calculations, ordered by mean (descending) followed by standard deviation (ascending).

Schwartz Value	Calculated Using	min	max	mean	sd
power	word hop	0.005	0.320	0.207	0.029
conformity	word hop	0.005	0.320	0.202	0.033
security	word hop	0.005	0.320	0.201	0.034
self direction	word hop	0.005	0.320	0.190	0.039
benevolence	word hop	0.005	0.320	0.176	0.042
achievement	word hop	0.005	0.320	0.170	0.041
tradition	word hop	0.004	0.320	0.156	0.037
universalism	word hop	0.004	0.317	0.142	0.025
hedonism	word hop	0.004	0.269	0.139	0.018
stimulation	word hop	0.004	0.213	0.139	0.018
power	synset to synset relationship	0.002	0.148	0.086	0.013
universalism	synset to synset relationship	0.002	0.143	0.084	0.011
stimulation	synset to synset relationship	0.002	0.148	0.083	0.011
achievement	synset to synset relationship	0.002	0.130	0.081	0.012
security	synset to synset relationship	0.002	0.138	0.081	0.012
conformity	synset to synset relationship	0.002	0.136	0.081	0.012
tradition	synset to synset relationship	0.002	0.143	0.077	0.012
self direction	synset to synset relationship	0.002	0.123	0.076	0.013
hedonism	synset to synset relationship	0.002	0.122	0.075	0.012
benevolence	synset to synset relationship	0.002	0.123	0.071	0.010

Table 9 demonstrates that counts of Word nodes were smaller on average and with a smaller range than counts of Synset-to-Synset relationships. This had the effect of producing *larger* word-by-hop values with larger ranges, as can be seen in Table 11 and many of the figures below. Because the Word node count calculation provided higher levels of variance for to account for, word-by-hop values derived from Word node counts were chosen over those derived from Synset-to-Synset relationship counts for use in the regression models reported below.

The relationship between the two calculations of word-by-hop is visualized in Figures 11, 12, 13, and 14. These figures indicate a clear positive relationship between the two calculations, such that word-by-hop values calculated from Word node counts do tend to correspond to higher word-by-hop values calculated from Synset-to-Synset relationship counts. However, these figures also show a consistently multi-linear relationship between the two calculations, resulting in a low linear correlation value between the means of each calculation across Schwartz values (taken from Table 9), $r = 0.122$.

Answering Research Question 1: Which Values are Present in Relation to One Another in Obituaries. The distributions of male vs. female word-by-hop values plotted by age at death are visualized in Figures 15, 16, 17, and 18. These figures illustrate the differential variances of the two word-by-hop calculation approaches, and also, surprisingly, suggest that if an age effect exists in the data at a statistically noteworthy level, it is small enough perhaps not to pass the threshold for *conceptual* noteworthiness. Overall, the figures suggest that obituaries coded as written about women also tended to be coded as written about older individuals; it is worth noting, however, that the range of ages displayed in the figures (from 1, the minimum age allowed by the automated age-guessing

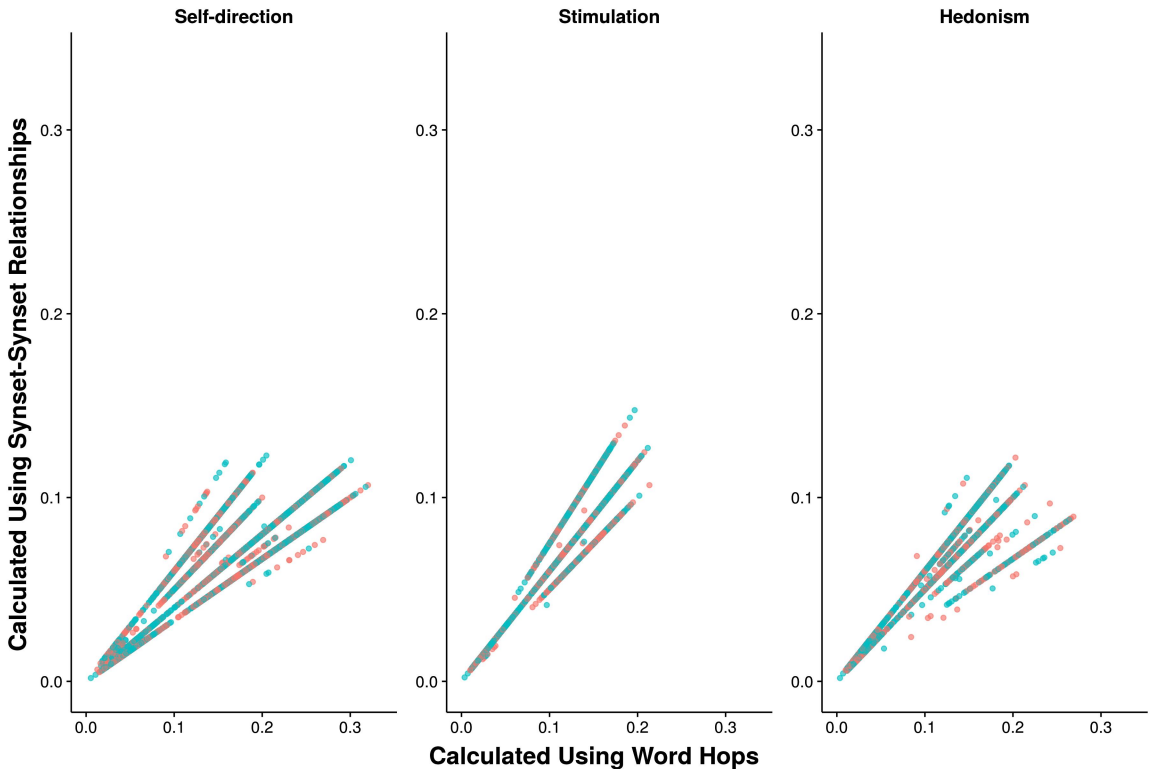

Figure 11. Schwartz "openness to change" values (Self-direction, Stimulation, and Hedonism) plotted by gender to compare word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word) vs. as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Pink/lighter points represent obituaries coded as written about women.

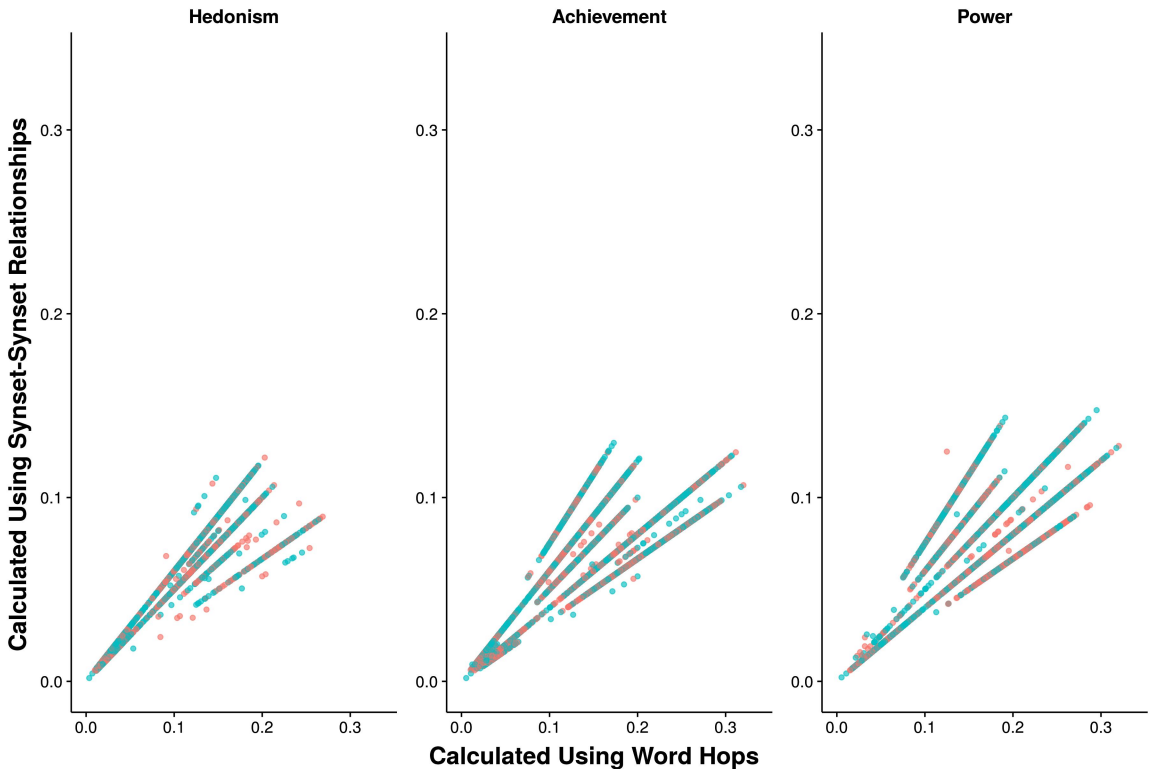

Figure 12. Schwartz "self-enhancement" values (Hedonism, Achievement, and Power) plotted by gender to compare word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word) vs. as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Pink/lighter points represent obituaries coded as written about women.

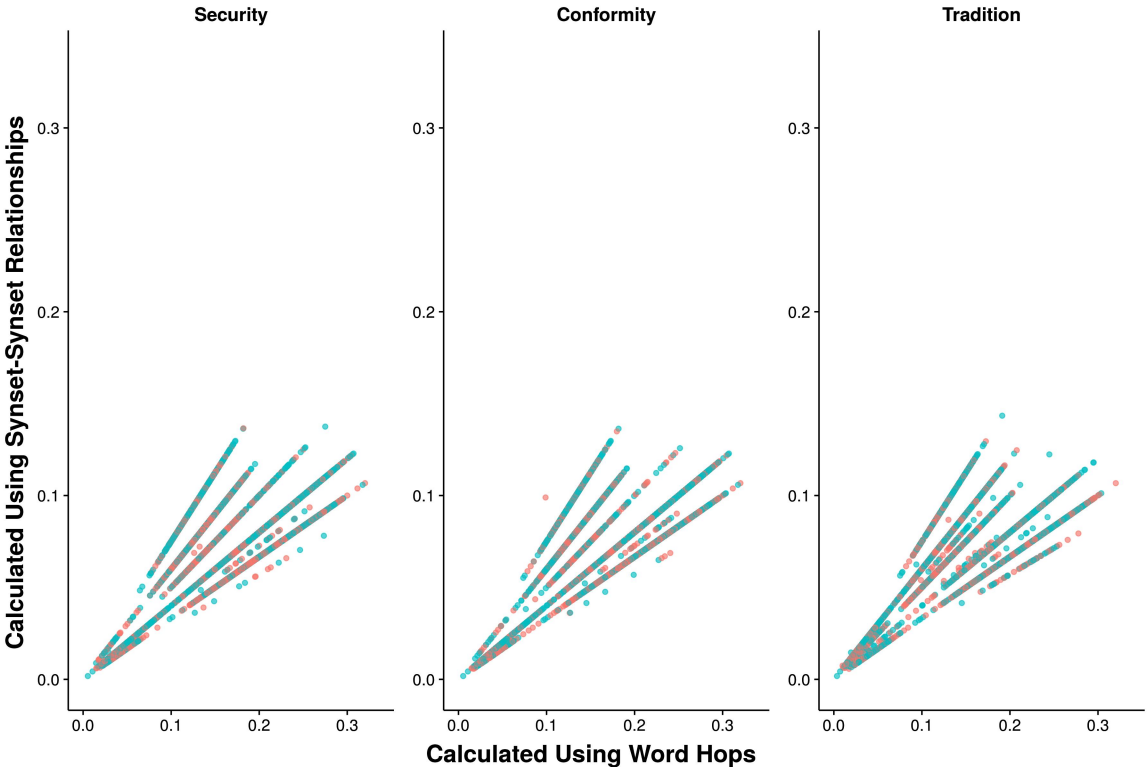

Figure 13. Schwartz "conservation" values (Security, Conformity, and Tradition) plotted by gender to compare word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word) vs. as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Pink/lighter points represent obituaries coded as written about women.

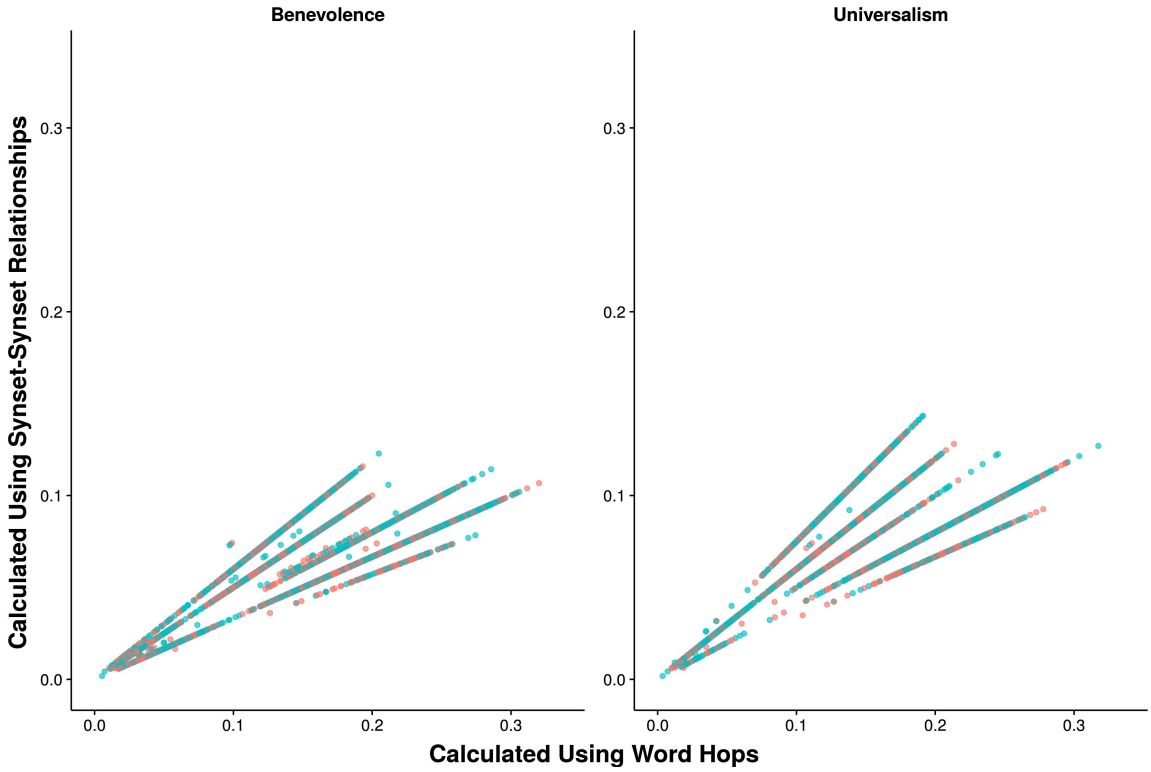

Figure 14. Schwartz "self-transcendence" values (Benevolence and Universalism) plotted by gender to compare word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word) vs. as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Pink/lighter points represent obituaries coded as written about women.

algorithm, to 110, the maximum age allowed) suggest that a sizable number of obituaries may have been misclassified (as it is unlikely that so many 110-year-olds actually died during the data collection period, which covered approximately one month). This is to be expected, given that the validation exercise for the age-coding algorithm reported above in the Methods section indicated that the algorithm was in high but not perfect agreement with the human coders, and because the full dataset is large (and thus provided ample opportunities for misclassification, even at a low overall level); however, it does suggest that age-related results should be interpreted with caution.

Consistent "banding" can be observed at the bottom of several of Figures 15, 16, 17, and 18. The obituaries that these bands comprise are low in Schwartz value connectedness but tend to be written for individuals who are of higher ages at death. It is possible that these indicate a particular template of obituary, which deflates its word-by-hop values while inflating its word count by naming large numbers of family members (cf. Figures 19, 20, 21, and 22, which are discussed below, and which also show banding at consistent ranges of word counts).

The distributions of male vs. female word-by-hop values plotted by obituary word count are visualized in Figures 19, 20, 21, and 22. The consistent conical distribution shapes in these figures may indicate several properties of word-by-hop. First, it is possible that the cones show that obituaries at lower word counts contain more "noise" in relation to clearly understanding their relationship to a given Schwartz value. This is reasonable, given that obituaries are generally expected to perform several roles simultaneously, including providing biographical information about the deceased, naming surviving family members, and providing logistical information about funerary services. Thus, obituaries with lower word

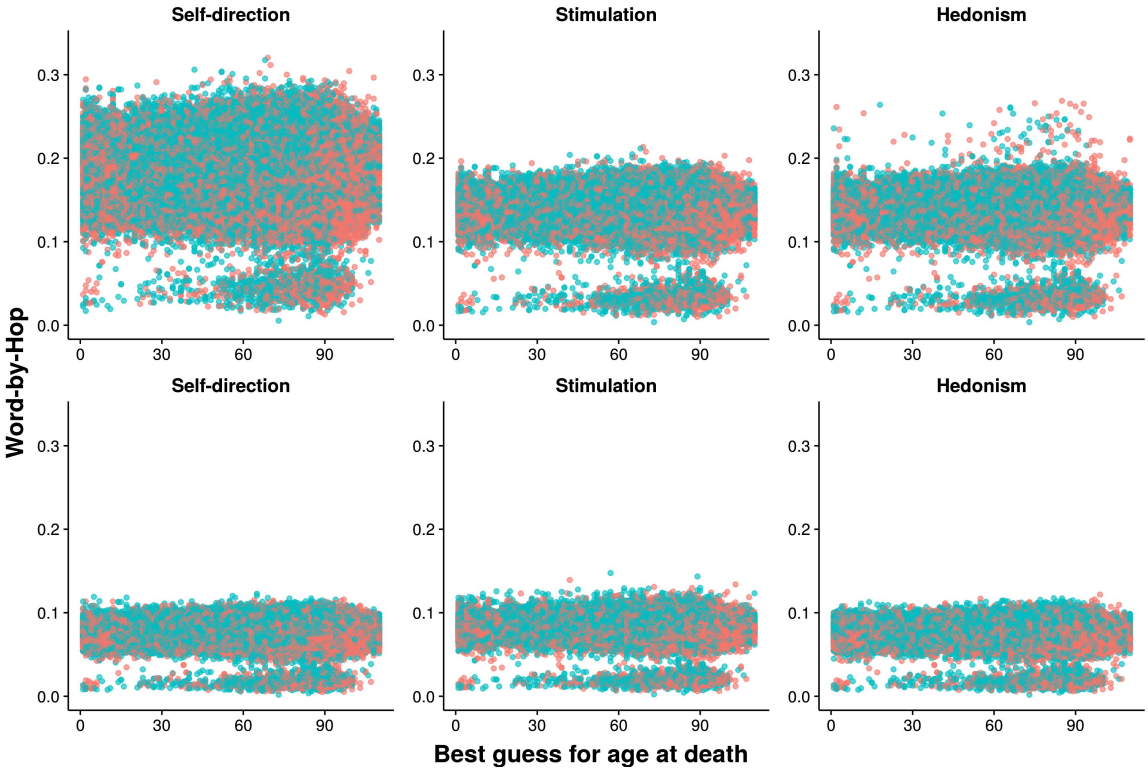

Figure 15. Schwartz "openness to change" values (Self-direction, Stimulation, and Hedonism) plotted by gender and age at death, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

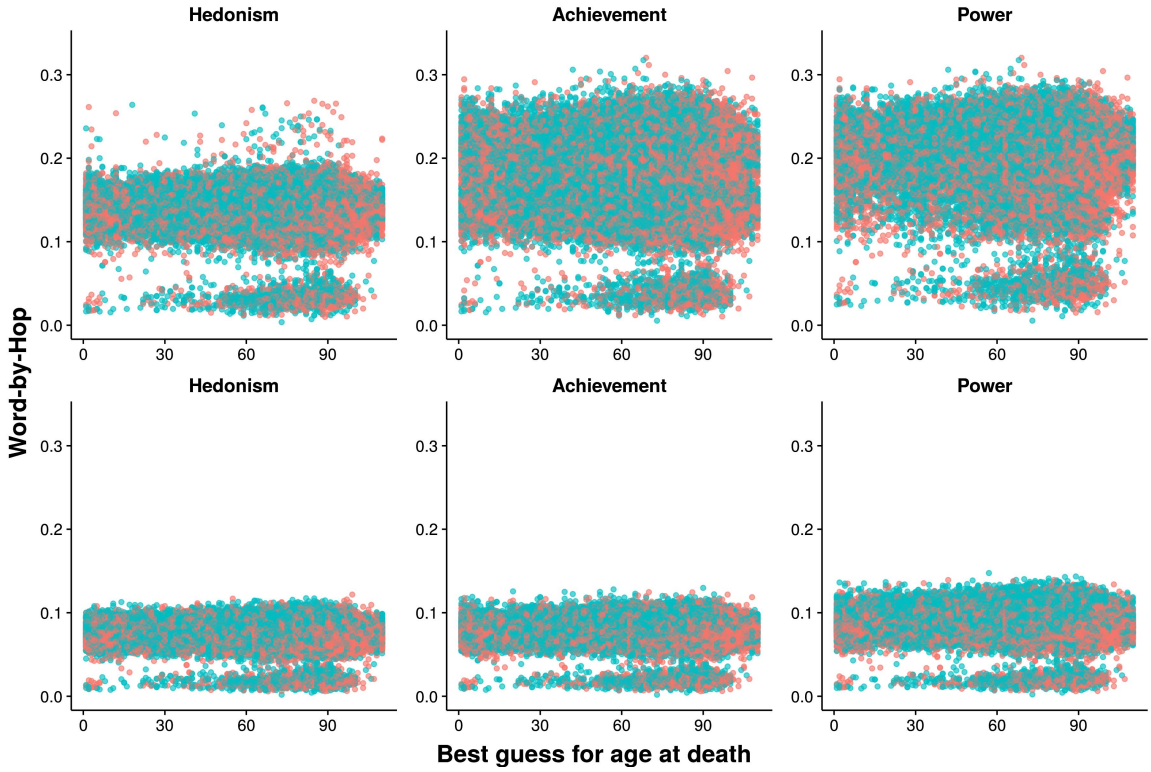

Figure 16. Schwartz "self-enhancement" values (Hedonism, Achievement, and Power) plotted by gender and age at death, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

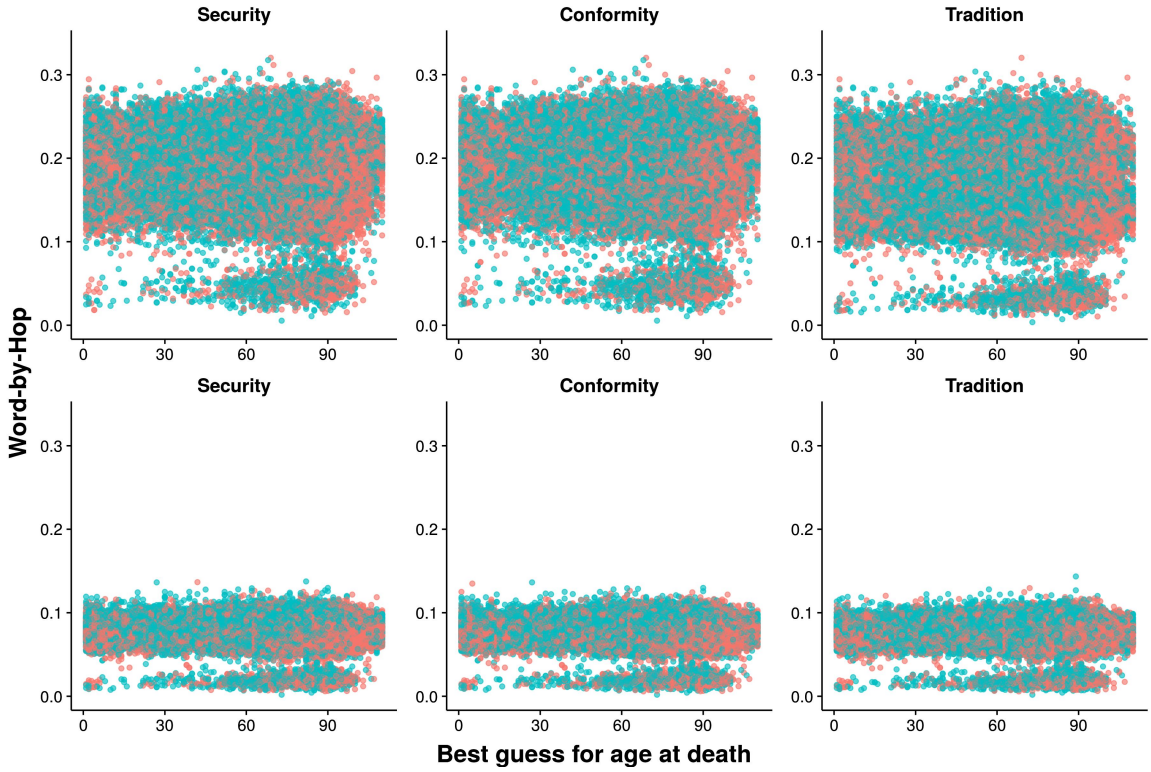

Figure 17. Schwartz "conservation" values (Security, Conformity, and Tradition) plotted by gender and age at death, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

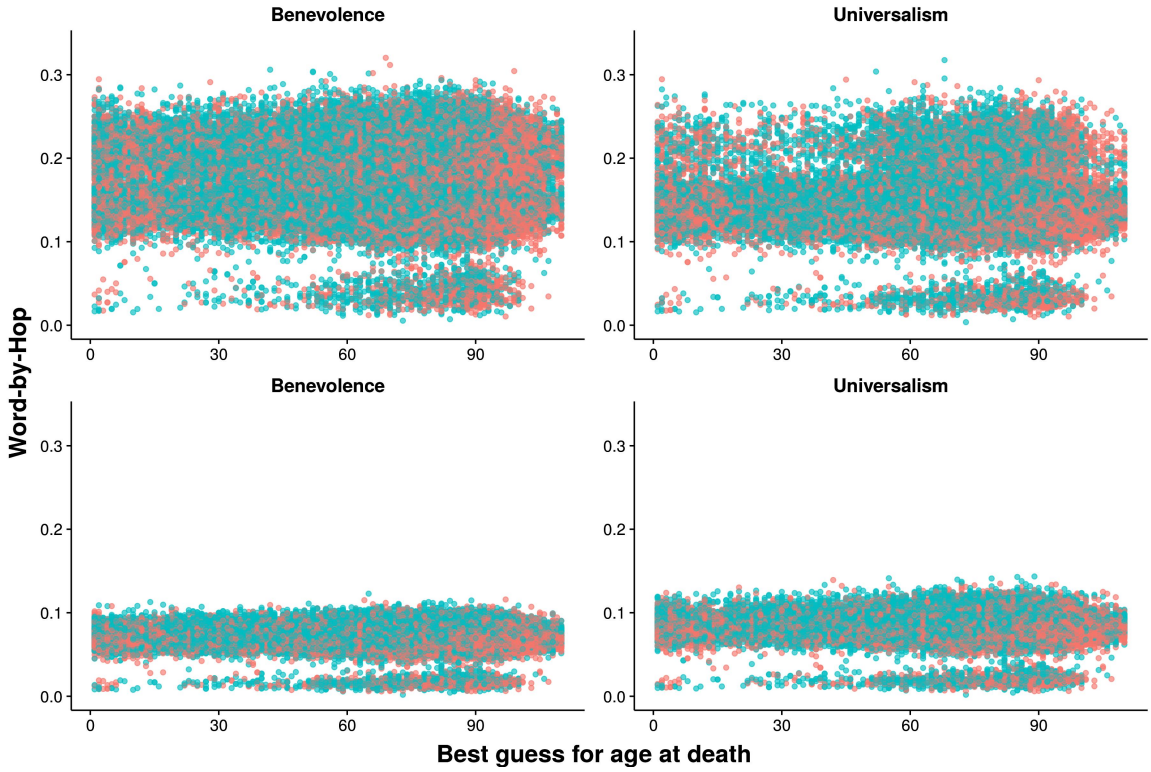

Figure 18. Schwartz "self-transcendence" values (Benevolence and Universalism) plotted by gender and age at death, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

counts may be more variable in their word-by-hop values because they include a higher percentage of words that may be only incidentally value-relevant while serving one of the more logistical functions of obituaries as death notices. Following this, it is possible that longer obituaries converge on a "true" level of connectedness with a Schwartz value, as expressed in their word-by-hop values. Less strongly, this may indicate that there exists a minimum word count threshold beyond which value words become more likely to appear, and that the 60-word threshold applied in this project for biographical content is below that threshold for values. If such an additional threshold exists, however, it is likely to be more complicated than based simply on word count (anecdotally, many obituaries of only a few hundred words can be seen readily to invoke values).

Second, it is possible that the cone-shaped distributions indicate a property of the word-by-hop formula itself, such that longer obituaries are less able to fit a larger number of word-by-hop values. This possibility is one that would be particularly useful to explore in follow-up research. The bimodal distributions of word-by-hop values in these figures for Self-direction, Achievement, and all of the Schwartz "conservation" values (Security, Conformity, and Tradition) and "self-transcendence" values (Benevolence and Universalism) are also noteworthy, and potentially point to differences across communities particularly in those Schwartz values (as explored below in the random-intercepts regression models, in which word-by-hop values are estimated for each newspaper separately). Looking from a different perspective, it is worth noting that only two Schwartz values (Stimulation, Hedonism, as well as, to a lesser extent, Power) *lack* a bimodal distribution when plotted against word count. This widespread bimodality could be an artifact of an interaction between the use of templates by obituary authors and the connections

between words in the WordNet graph. If a small number of highly-connected words tended to be used by obituary authors to describe the deceased (perhaps as part of a templated phrase), the inclusion of a small number of additional highly-connected words could presumably "bump up" the word-by-hop value for that obituary by a consistent level. If this were true, Hedonism and Stimulation, as values that occur less frequently, might have fewer templates available and thus show greater normality. This explanation is bolstered by the fact that bimodality is not seen in these figures where Synset-to-Synset-derived word-by-hop values are used as DVs; this difference in plots between the two calculations implies that distribution is linked to Word connectedness (as well as, possibly, an inherent property of the word-by-hop formula that manifests more when the median number of hops is lower, as when using Word node counts).

Figures 23, 24, 25, and 26 illustrate the densities of word-by-hop values by Schwartz value, showing distributions without reference to any covariates. These figures are plotted by newspaper (with darker areas indicating greater overlap among separate newspapers), and also show distributional variability (e.g., bimodality vs. normality) across communities not only in Self-direction, Achievement, Tradition, and Benevolence, but in *every* value when looking across newspapers.

The correlation matrix of word-by-hop values is presented in Figure 27. The minimum correlation was 0.462. Schwartz values in this matrix are listed following the Schwartz circumplex model in a counter-clockwise direction. A visual inspection of this matrix does indicate that, generally, Schwartz values as reconstructed using word-by-hop values do tend to correlate more highly with values that are closer to them in the circle (i.e., the beginning and end of each

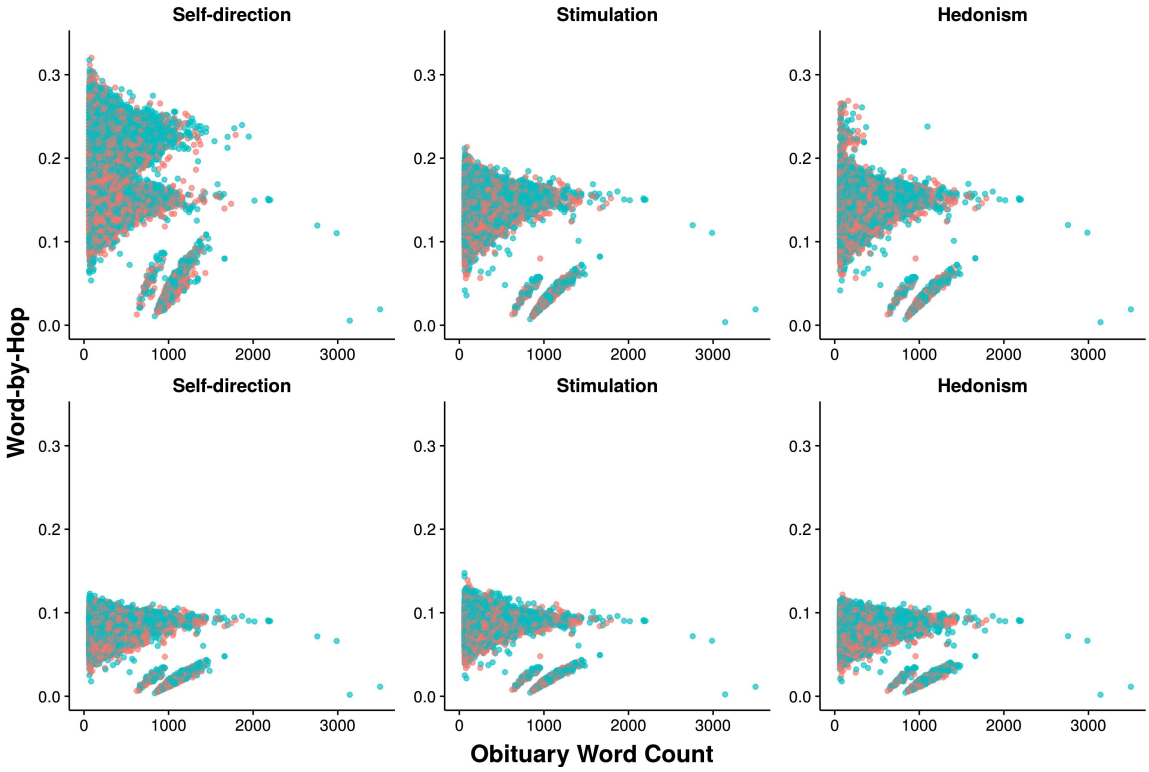

Figure 19. Schwartz "openness to change" values (Self-direction, Stimulation, and Hedonism) plotted by gender and obituary word count, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

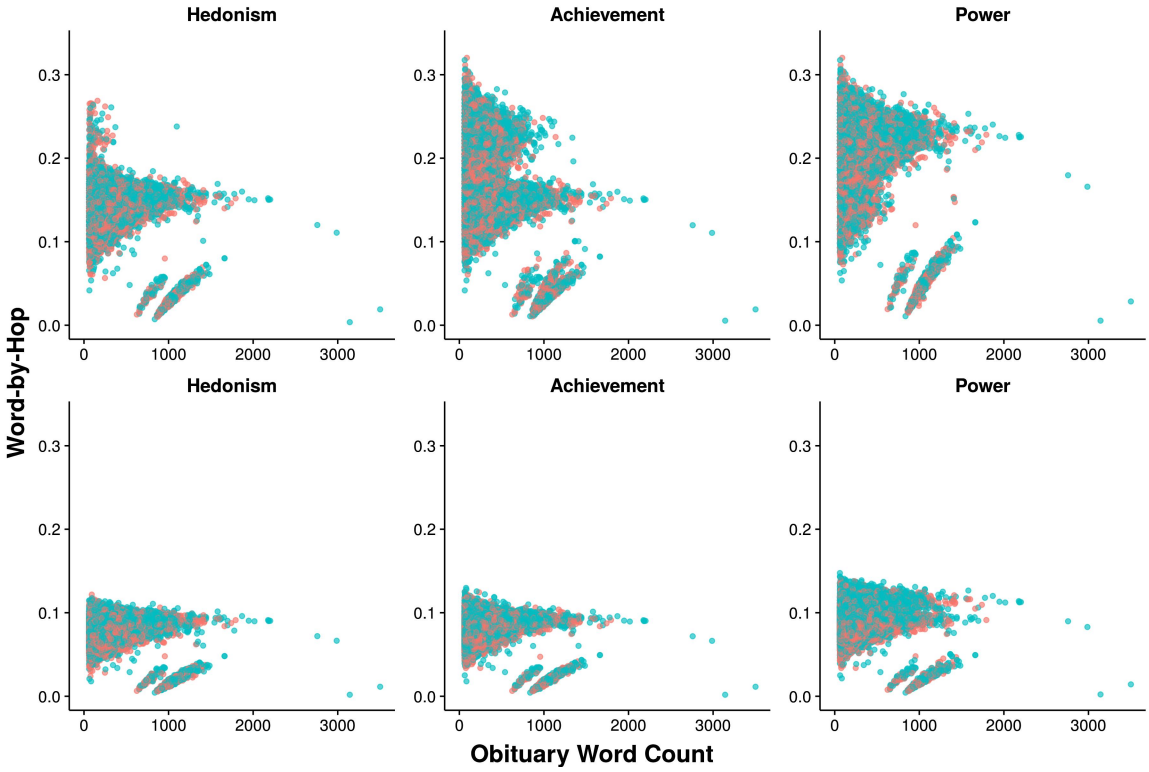

Figure 20. Schwartz "self-enhancement" values (Hedonism, Achievement, and Power) plotted by gender and obituary word count, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

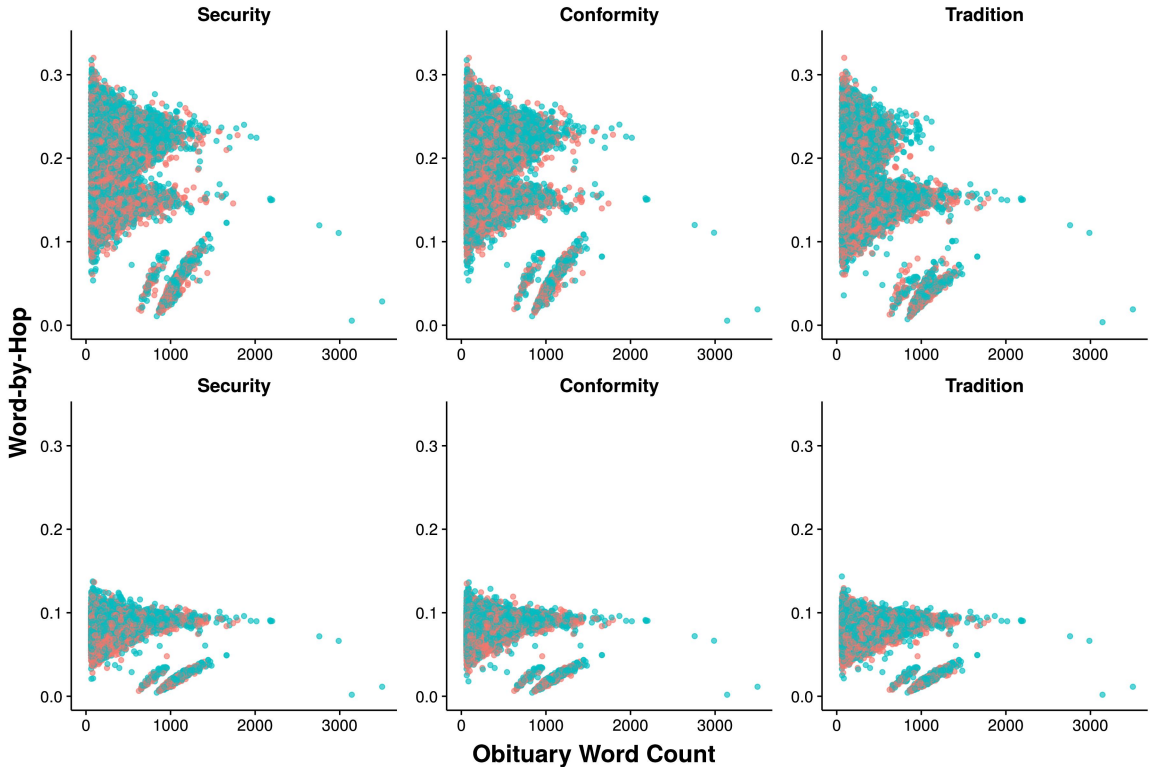

Figure 21. Schwartz "conservation" values (Security, Conformity, and Tradition) plotted by gender and obituary word count, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

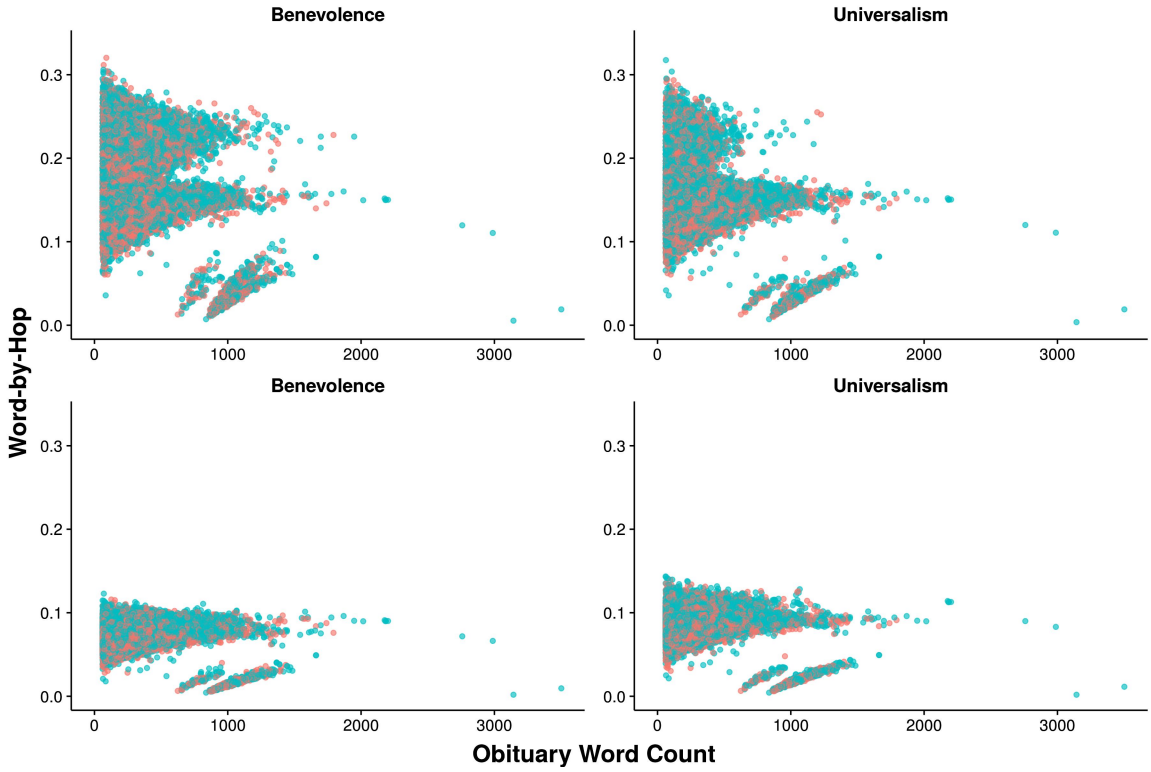

Figure 22. Schwartz "self-transcendence" values (Benevolence and Universalism) plotted by gender and obituary word count, as both coded using the automated algorithm described in the Methods section. Pink/lighter points represent obituaries coded as written about women. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word.

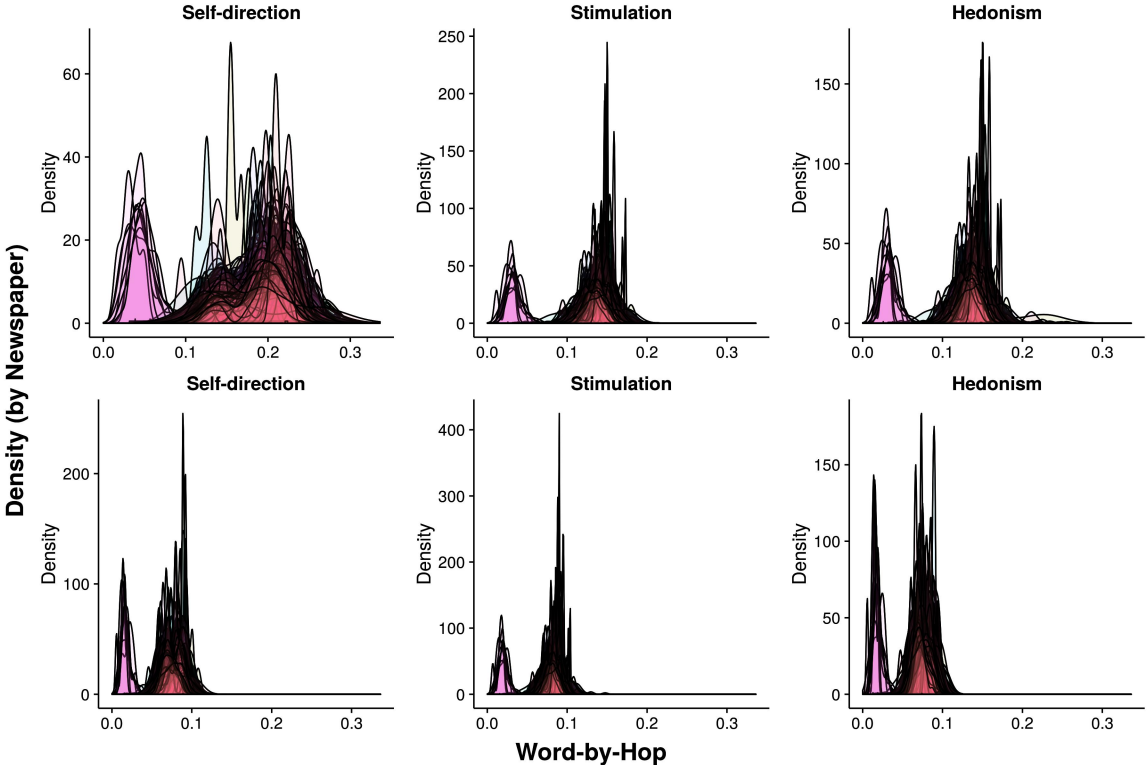

Figure 23. Densities of word-by-hop for Schwartz "openness to change" values (Self-direction, Stimulation, and Hedonism), plotted by newspaper to allow visual inspection of both overall distribution and variability across communities. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Darker areas indicate higher newspaper overlap.

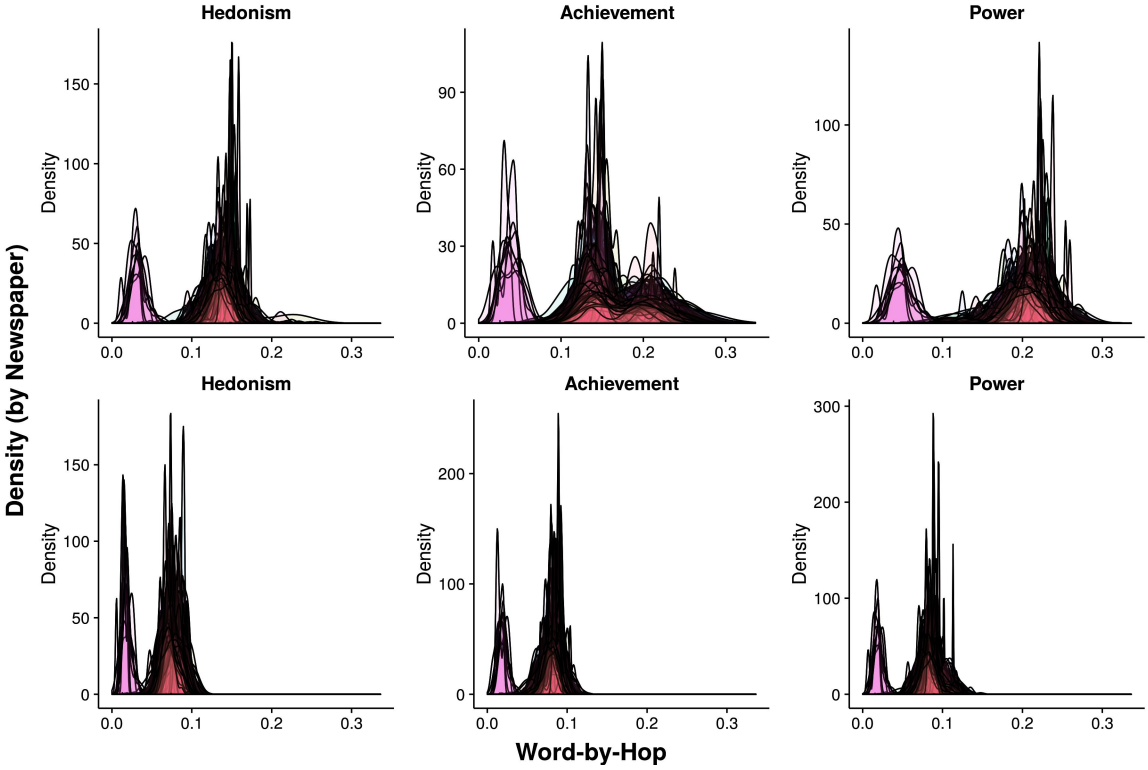

Figure 24. Densities of word-by-hop for Schwartz "self-enhancement" values (Hedonism, Achievement, and Power), plotted by newspaper to allow visual inspection of both overall distribution and variability across communities. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Darker areas indicate higher newspaper overlap.

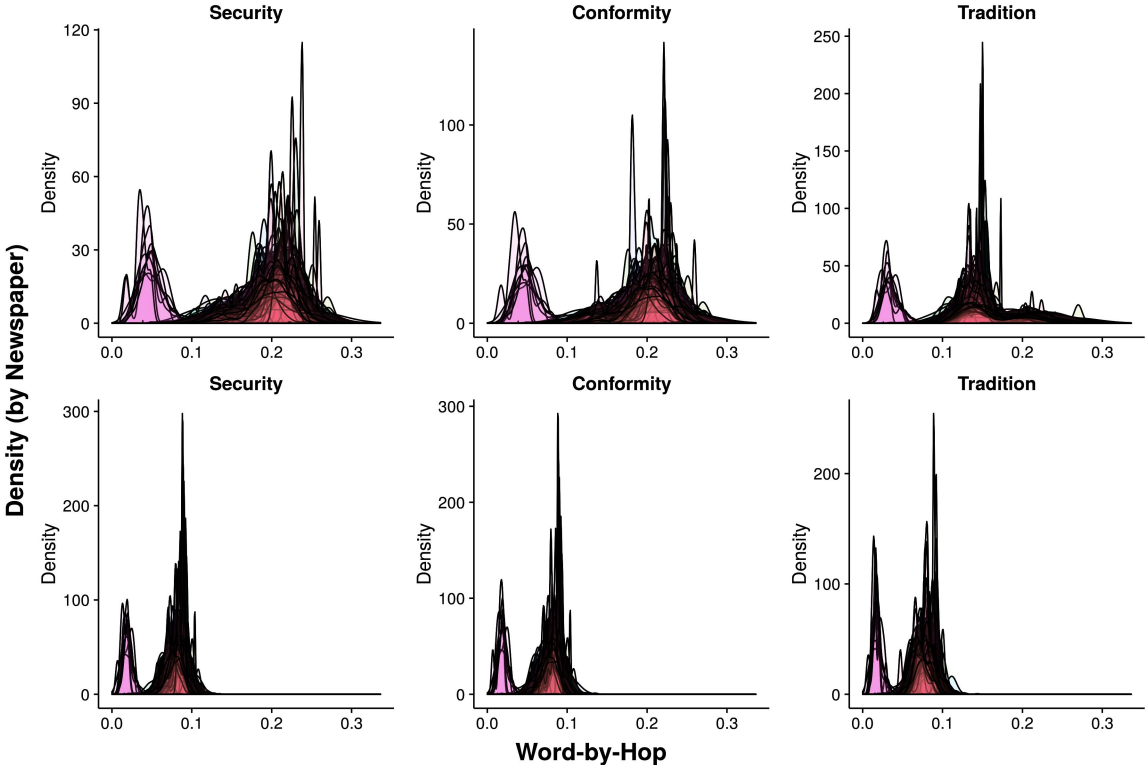

Figure 25. Densities of word-by-hop for Schwartz "conservation" values (Security, Conformity, and Tradition), plotted by newspaper to allow visual inspection of both overall distribution and variability across communities. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Darker areas indicate higher newspaper overlap.

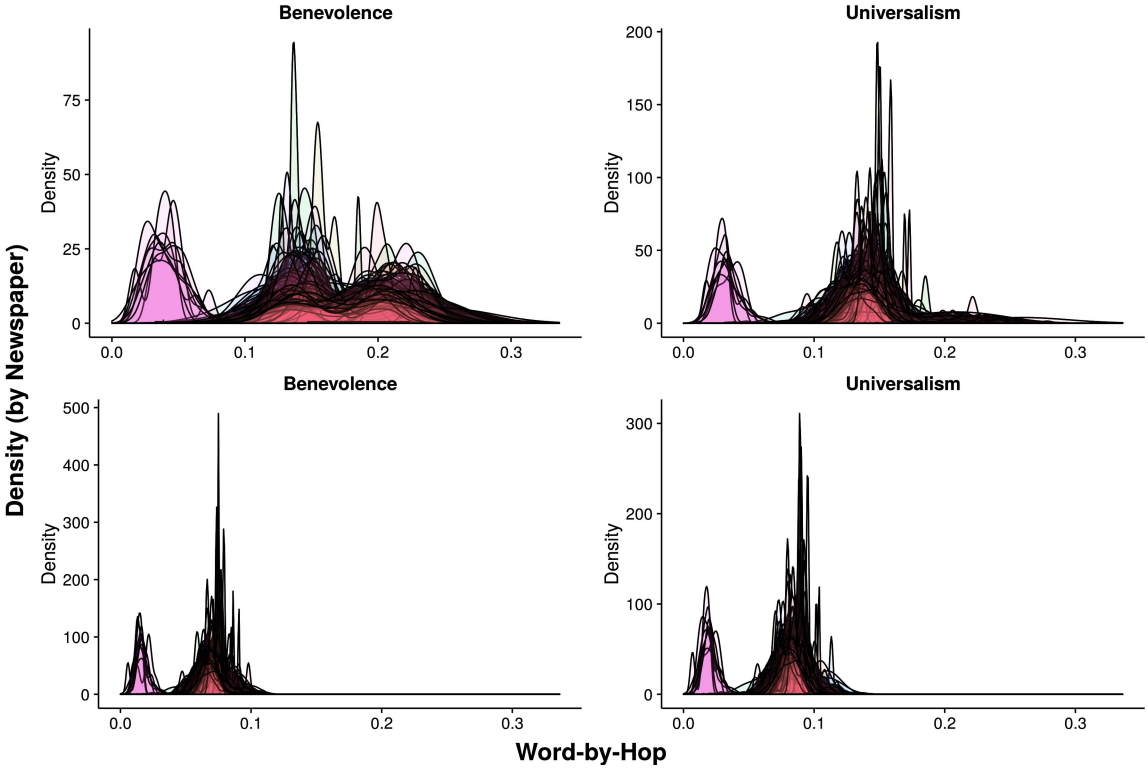

Figure 26. Densities of word-by-hop for Schwartz "self-trancendence" values (Benevolence and Universalism), plotted by newspaper to allow visual inspection of both overall distribution and variability across communities. The top row plots word-by-hop values as computed using the number of Word nodes between each obituary lemma / Part-of-Speech combination and a value lexicon word, less one (i.e., the number of Word node hops in the shortest path between a given source and target word). The bottom row plots word-by-hop values as computed using the number of Synset-to-Synset edges in the shortest path between a given source and target word. Darker areas indicate higher newspaper overlap.

row of the matrix) than with those that are further away. The matrix does not always follow this pattern, however: Benevolence and Universalism, which exist side-by-side in the circumplex model, had a low correlation with each other. It may be that the value lexicon words for Benevolence and Universalism (Table 1), while related, would be invoked in different contexts, especially based on the occupation of the deceased. "Kindness," "charity," and "mercy," the prototype words for Benevolence, seem able to be applied to a variety of individuals. The prototype words for Universalism, however, "unity," "justice," and "equality," might be invoked more for individuals of particular professions (e.g., lawyers, judges, and activists), diminishing the overall correlation between the two values.

Similarly, Hedonism was highly correlated with Conformity, its polar opposite in the circumplex model. Hedonism's relationship to Conformity does make sense with reference to Bardi et al.'s lexicon words for these two values. The lexicon words for Hedonism, "luxury," "pleasure," and "delight," *are* conceptually opposite of those for Conformity, "restraint," "regard," and "consideration." Given the expectation in this project that Hedonism is a taboo value to invoke in an obituary, this correlation may indicate that obituary authors "compensate" for describing the deceased in a hedonistic way by concurrently emphasizing restraint (and therefore Conformity).

Because it is possible that income and education might be associated with obituary word counts, which are used when calculating word-by-hop values, Figure 28 visualizes the relevant relationships between these variables. Visual inspection of this figure suggests that no noteworthy relationship exists (put differently, it does not appear to be the case that wealthier or more highly-educated communities systematically produced longer obituaries).

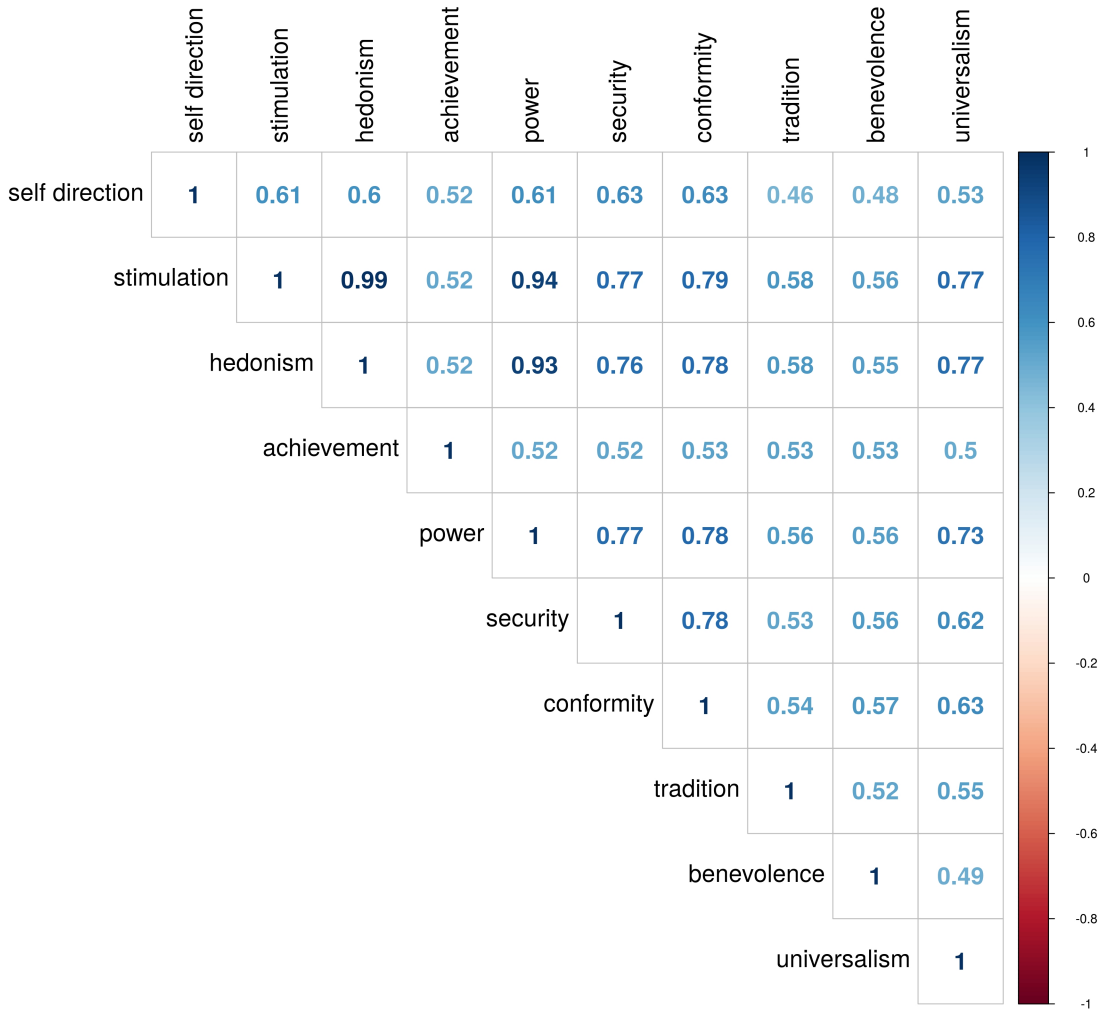

Figure 27. Correlation matrix of word-by-hop (calculated using number of Word nodes between a source and target word, rather than the number of Synset-to-Synset edges) for each pair of Schwartz values. Schwartz values in this matrix are listed following the Schwartz circumplex model in a counter-clockwise direction.

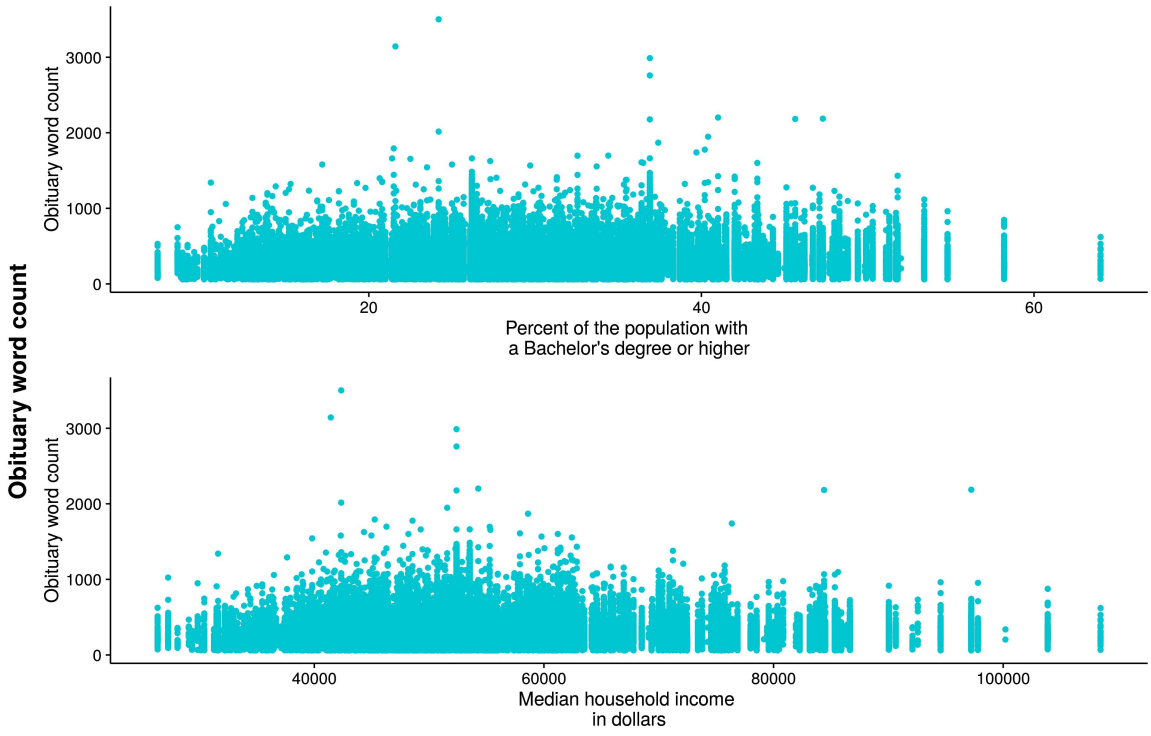

Figure 28. Scatterplot of median household income and education level (operationalized as percent of the population with a bachelor's degree or higher) from US census data averaged across counties for each newspaper (after weighting by county population size) vs. obituary word count.

Overall, these results demonstrate that all 10 Schwartz values are present in this obituary corpus, but that some Schwartz values *are* more lexically indicated than others[13]. Following Table 11, as expected, Hedonism was among the least-indicated Schwartz values in the corpus. However, counter to predictions, Power, rather than also being low in its mean word-by-hop value, was *the single most indicated Schwartz value in the corpus.* Schwartz "conservation" values (Conformity and Security) were also particularly highly-indicated in the corpus overall. With this in mind, it is useful to note that a substantial amount of variability exists across newspapers. This variability is addressed below in response to Research Question 2.

Answering Research Question 2: Does Match with the Schwartz Values Vary in Relation to Age, Gender, Race or Ethnicity, Income, and/or Education. Following the rationale discussed above, word-by-hop values were logit-transformed and entered into a series of two-level linear models, with obituaries as level 1 and newspapers as level 2. Also as discussed above, across-newspaper duplicate obituaries were randomly downsampled until one per duplicate set remained in the dataset. For each model, 10 regression analyses were performed, each using a word-by-hop DV aligned with a different Schwartz value. To be interpreted alongside Table 11, Table 12 presents the minimum, maximum, mean, and standard deviation for each Word-hop-derived logit-transformed word-by-hop DV.

[13]All results discussed from this point forward employed word-by-hop values calculated using the number of Word nodes between a source and target word, rather than the number of Synset-to-Synset edges.

Table 12. Descriptive statistics for Word-hop-derived, logit-transformed word-by-hop calculations, ordered by mean (descending) followed by standard deviation (ascending).

Schwartz Value	min	max	mean	sd
power	-5.21	-0.753	-1.36	0.230
conformity	-5.21	-0.753	-1.39	0.252
security	-5.21	-0.753	-1.40	0.259
self_direction	-5.21	-0.753	-1.48	0.296
benevolence	-5.21	-0.753	-1.58	0.326
achievement	-5.21	-0.753	-1.61	0.318
tradition	-5.62	-0.753	-1.72	0.303
universalism	-5.62	-0.765	-1.81	0.234
hedonism	-5.62	-1.001	-1.84	0.205
stimulation	-5.62	-1.304	-1.84	0.205

Five predictors were prepared for use in the regression models:

- Level 1 (Obituaries):
  - Age of the deceased (automatically coded using the algorithm described above)
  - Gender (female/male) of the deceased (automatically coded using the algorithm described above). This variable was dummy-coded (0 = male, 1 = female)

- Level 2 (Newspaper) (as described above, computed from data from the counties in which each newspaper operated, using a weighted average by county population size):
    - Median household income (United States Bureau of the Census, 2015b)
    - Education, operationalized as percentage of the population with a Bachelor's degree or higher (United States Bureau of the Census, 2015a)
    - Percentage of the population that identified (in isolation or in combination) as one of five racial and/or ethnic identities (each category was represented with its own predictor variable): "American Indian or Native Alaskan," "Black or African American," "Hispanic," "Native Hawaiian and other Pacific Islander," and "White" (United States Bureau of the Census Population Division, 2015)

This project's analyses were focused not only on the fixed effects of these predictors, but also centrally on the effects of adding random intercepts by newspaper, and random slopes by gender, age, and the interaction between gender and age. Thus, a series of seven nested or semi-nested regression models were run (for a total of 70 regression analyses, one for each word-by-hop Schwartz value DV). These models were written and run before the results of each were analyzed (put differently, these models were written in an *a priori* fashion), with the exception of one final model (listed as model 3 below, given that it conceptually builds on model 2), which was developed after viewing the results of the others. Given the number of models run, p-values are eschewed in the report below in favor of considering estimates and their standard errors directly, with a particular additional focus on comparing Akaike Information Criterion (AIC) values across models to aid

in determining fit. All models were run using the `lmer` function from R's `lme4` package, with the exception of the first, most basic model, which did not include any random effects and was thus run with R's related `lm` ("linear model") function. Because of the vast difference in scales among the predictor variables (e.g., income, which had a mean in the tens of thousands, vs. age at death, which had a mean in the tens), all continuous predictor variables were z-scored (i.e., made to have a mean of 0 and a Standard Deviation of 1).

The seven models are described below. Although the model descriptions do not mention it, all models contained an individual-level error term, as is standard in multiple regression. All formulae are in R's `lm`/`lmer` format, where a tilde "~" separates a DV from the rest of the formula, and a pipe symbol ("|") indicates a random factor grouped by the term on the right side of the pipe (in this notation, "(1 | group)" would indicate a random intercept, and "(predictor | group)" would indicate a random slope for the predictor).

1. A basic model, consisting of only a single fixed intercept (i.e., for each DV, modeling word-by-hop values using only the mean word-by-hop value). Labeled in tables and plots below as "Model with only intercept."

2. A model consisting of a *random* intercept term (i.e., for each DV, modeling word-by-hop values for each newspaper separately using the mean word-by-hop value of that newspaper). Labeled in tables and plots below as "Model with random intercept." The formula for this model was (for an example Schwartz value word-by-hop DV) "Achievement ~ (1 | newspaper_shortcode)".

3. Model #2 expanded to include both level-1 predictors (age and gender), as well as random slopes for age, gender, and the interaction of age and gender

(i.e., allowing the individual and interactive effects of age and gender to differ across newspapers in the same way that the intercept was allowed to vary across newspapers). Labeled in tables and plots below as "Full model with random interaction effects without level-2 predictors." The formula for this model was (for an example Schwartz value word-by-hop DV) "Achievement ~ Female + Age at death + (1 + Female * Age at death | newspaper_shortcode)".

4. Model #3 without the random effects for age, gender, and age-by-gender interaction (i.e., a model consisting of a random intercept by newspaper and fixed effects for age and gender, without their interaction; this modeled the effect of age and gender as being the same for every newspaper). Labeled in tables and plots below as "Full model except for random interaction and level-2 predictors." The formula for this model was (for an example Schwartz value word-by-hop DV) "Achievement ~ Female + Age at death + (1 | newspaper_shortcode)".

5. Model #2 with fixed effects for gender and age, and fixed effects for income, education, the interaction between income and education, and the interaction between gender and education. Labeled in tables and plots below as "Full model except for random interaction and race/ethnicity." The formula for this model was (for an example Schwartz value word-by-hop DV) "Achievement ~ Female + Age at death + Income + Education + Female * Education + Income * Education + (1 | newspaper_shortcode)".

6. Model #5 with all race/ethnicity predictors added as fixed effects. Adding these as a block allowed examining the overall effect of race/ethnicity.

Labeled in tables and plots below as "Full model except for random interactions." The formula for this model was (for an example Schwartz value word-by-hop DV) "Achievement ~ Female + Age at death + Income + Education + Female * Education + Income * Education + Race/Ethnicity:"White" + Race/Ethnicity: "Black or African American" + Race/Ethnicity: "American Indian or Native Alaskan" + Race/Ethnicity: "Native Hawaiian and Other Pac. Islander" + Race/Ethnicity: "Hispanic" + (1 | newspaper_shortcode)".

7. Model #6 combined with Model #3. Labeled in tables and plots below as "Full model." The formula for this model was (for an example Schwartz value word-by-hop DV) "Achievement ~ Female + Age at death + Income + Education + Female * Education + Income * Education + Race/Ethnicity:"White" + Race/Ethnicity: "Black or African American" + Race/Ethnicity: "American Indian or Native Alaskan" + Race/Ethnicity: "Native Hawaiian and Other Pac. Islander" + Race/Ethnicity: "Hispanic" + (1 + Female * Age at death | newspaper_shortcode)".

Given the large number of AIC values to present (as 70 regression analyses were run in total), rather than presenting these results in a table, Figure 29 graphically shows AIC values across models and Schwartz values. Figure 30 removes the worst-fitting model (Model #1, which included only a fixed intercept) from Figure 29 in order to show with more detail the differences among the remaining models for each Schwartz value.

Lower AIC values indicate better model fit. AIC incorporates both the number of parameters and the model's deviance; thus, given two models with similar deviance, AIC will prefer the more parsimonious model (i.e., the model with

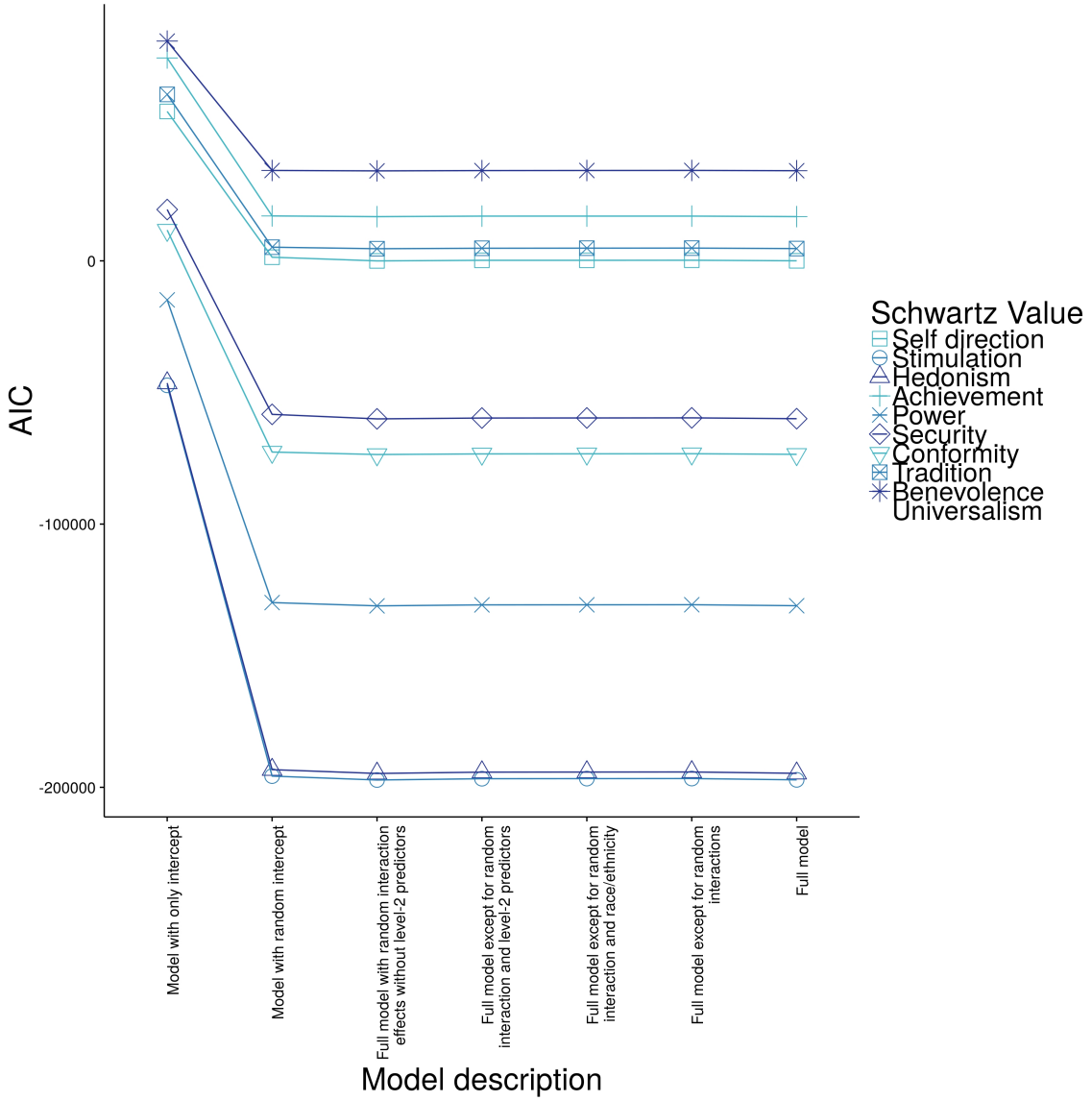

Figure 29. AIC values plotted across models and Schwartz values. The highest AIC (indicating worst fit) was found consistently across Schwartz values in the intercept-only model (Model #1). The lowest AIC was not in the full model (Model #7), but rather the full model lacking level-2 predictors (Model #3).

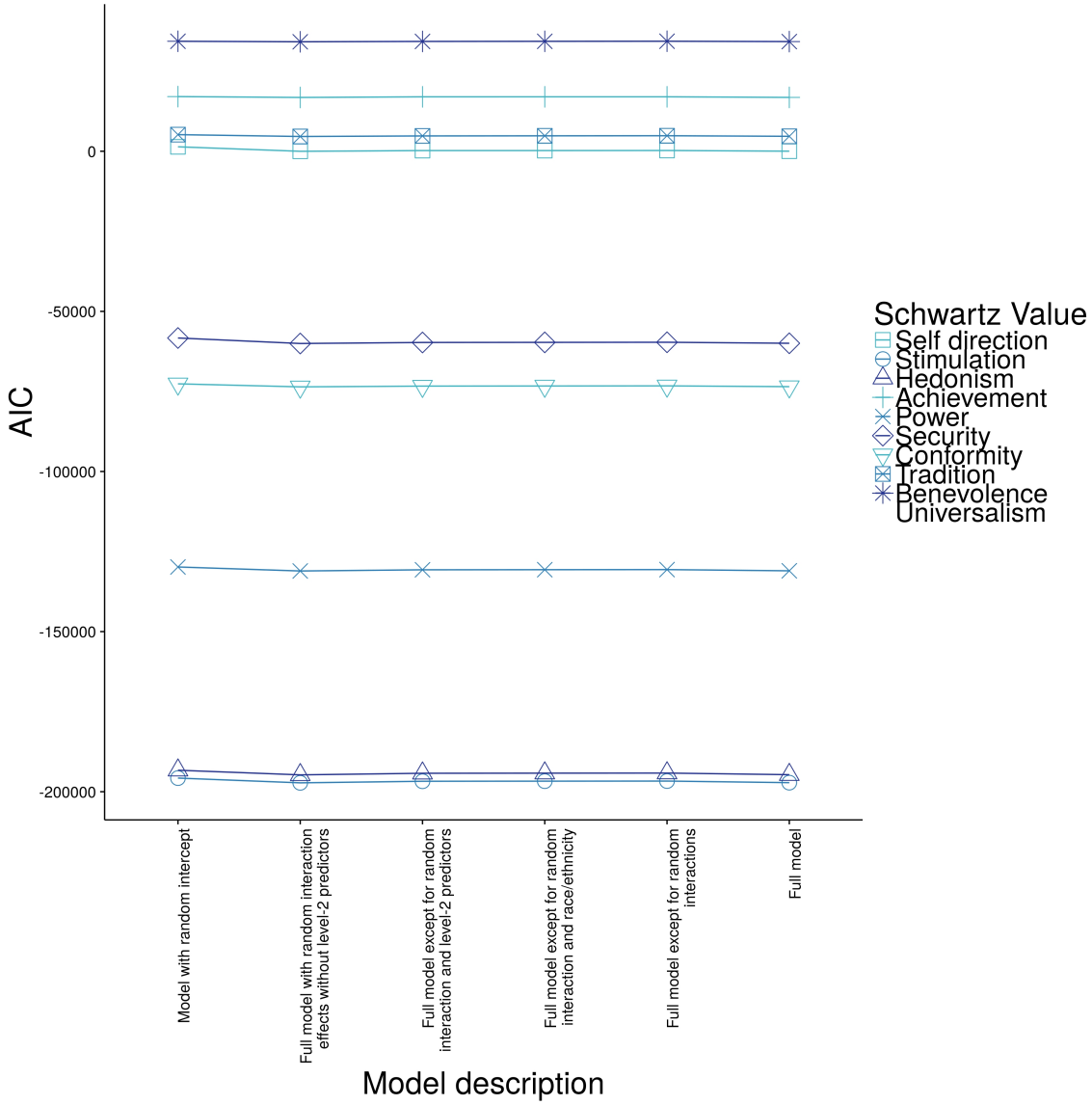

Figure 30. AIC values plotted across models and Schwartz values. The lowest AIC was not in the full model (Model #7), but rather the full model lacking level-2 predictors (Model #3).

fewer parameters). Looking at Figure 29, it is striking that, *by far,* the greatest amount of variability across Schwartz values, as measured using word-by-hop, was accounted for simply by including a random intercept by newspaper (moving from Model #1 to Model #2). As expected, this indicates, either or together, that, as expected, obituaries within a given newspaper *do* cluster around particular value levels, and that word-by-hop as a measure is sensitive to these differences. Figure 30 shows more clearly that, surprisingly, AIC values were consistently lowest across Schwartz values not for the full model (Model #7), but for Model #3, which allowed random slopes for age, gender, and their interaction, but excluded level-2 covariates. The full model consistently showed the next-lowest AIC values. While it is to be expected that adding parameters (even poorly predictive ones) to a model will increase its fit, AIC's penalization of larger numbers of parameters likely resulted in Model #3, which is more parsimonious, being given the smaller AIC values. Models #4 and #6 shows that it was not age and gender in themselves as fixed effects that resulted in better model fit, but specifically their being modeled as random effects.

Models #5 and #6, which lacked those random effects but differed from each other only in the inclusion of the race/ethnicity variables, indicate that the race/ethnicity covariates were not useful additions to this model. Taken together with Model #7, which had a higher AIC value than the model that lacked those covariates as well as income and education (Model #3), this could indicate that value levels in obituaries do not vary systematically by income, education, or race/ethnicity. However, this could also indicate that the measurement of these covariates was problematic in some way, perhaps either in that counties were not explicitly modeled and were instead aggregated by newspaper, or in that the

automated newspaper-location lookup algorithm included a large number of false positives and/or false negatives.

It is also worth remarking on the fact that AIC values were different across Schwartz values. This is likely the result of the word-by-hop DVs having different levels of variability for the models to account for. Word-by-hop values for Hedonism, for example, had higher uniformity (i.e., a lower Standard Deviation) than other values (as shown in Tables 11 and 12), perhaps because of norms against invoking Hedonism in obituaries (see the Discussion section below). For Benevolence, by contrast, the regression models had larger amounts of variability (indicated by a larger Standard Deviation) to account for: more so than Hedonism, it seems reasonable to expect that descriptions of the deceased in some obituaries would be much more indicative of Benevolence than in others.

As above, given the large number of analyses, Figures 31 and 33 graphically present the estimates and their standard errors for Models #3 and #7, respectively (i.e., the best-fitting model and the next-best fitting, full, model). Figures 32 and 34 feature the same respective information, but lacking the models' intercept estimates, allowing more detailed inspection.

Figure 34 indicates that in Model #3, of gender and age, gender was estimated to have had a lower but consistently non-zero effect across Schwartz values. This effect was most pronounced in Self-direction (with females having ˜.04 lower logit-transformed word-by-hop values than men, holding age constant), while the effect was least pronounced in Achievement. The estimates of age's effect encompassed zero for several values, and were not consistently positive or negative for the rest.

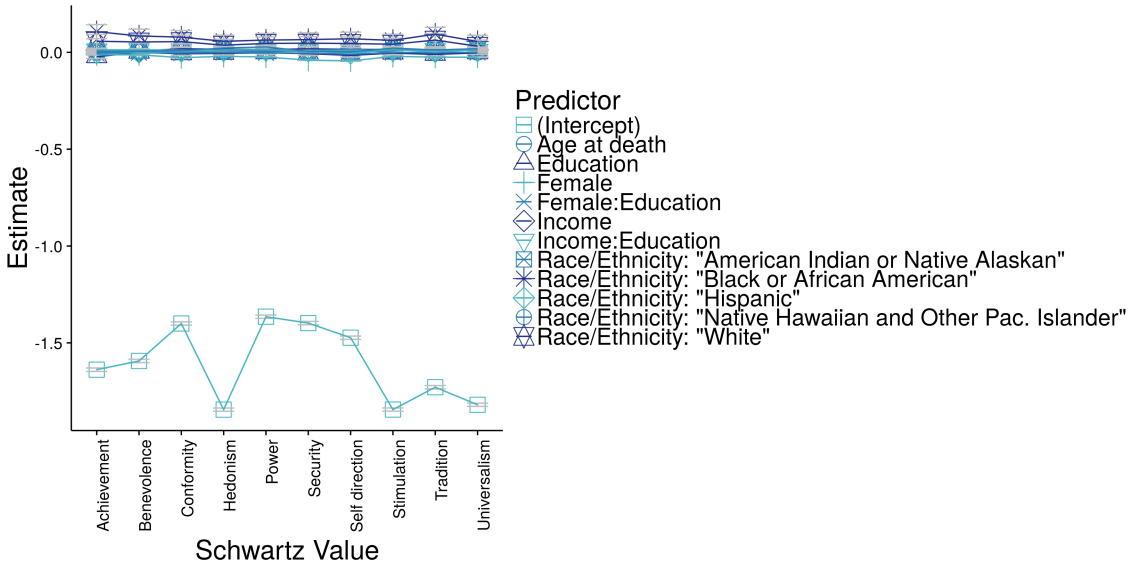

Figure 31. Plot showing coefficient estimates and standard errors for all Schwartz value word-by-hop Dependent Variables and all model predictors *for Model #7, the full model.* Note that all predictors were z-scored before being entered into the model, and that the word-by-hop DV was logit-transformed.

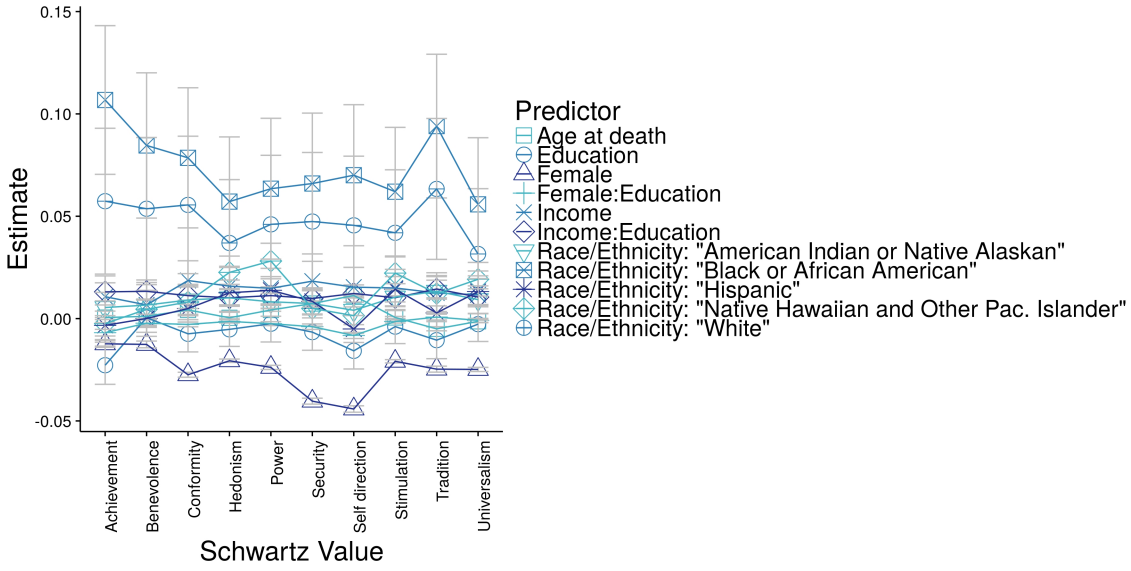

Figure 32. Plot showing coefficient estimates and standard errors for all Schwartz value word-by-hop Dependent Variables and all model predictors *for Model #7, the full model, excluding the intercept* (in order to see the other estimates in more detail). Note that all predictors were z-scored before being entered into the model, and that the word-by-hop DV was logit-transformed.

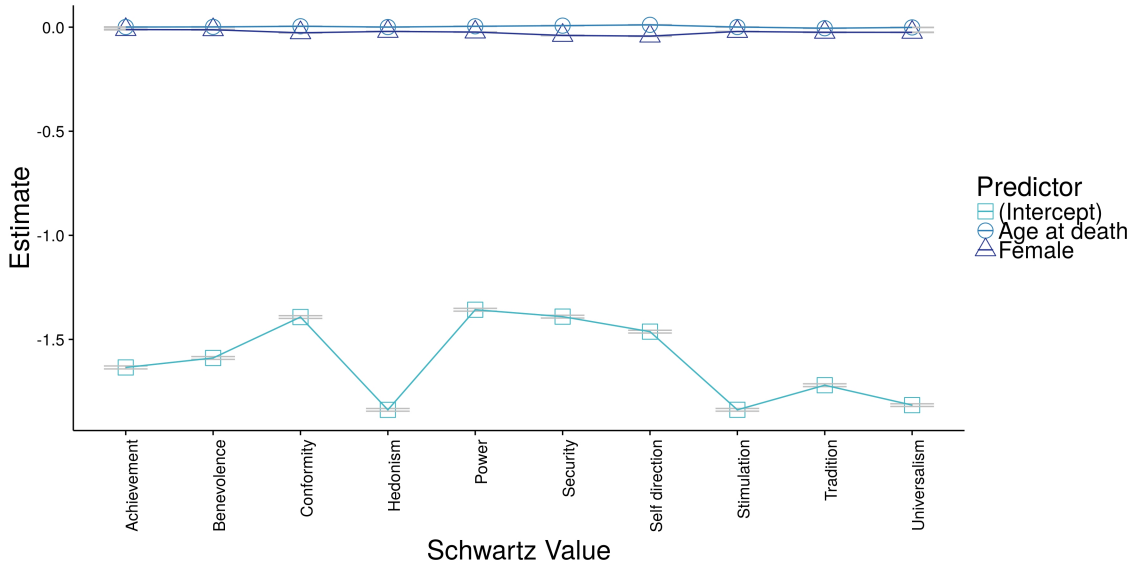

Figure 33. Plot showing coefficient estimates and standard errors for all Schwartz value word-by-hop Dependent Variables and all model predictors *for Model #3, the model with the lowest AIC.* Note that all predictors were z-scored before being entered into the model, and that the word-by-hop DV was logit-transformed.

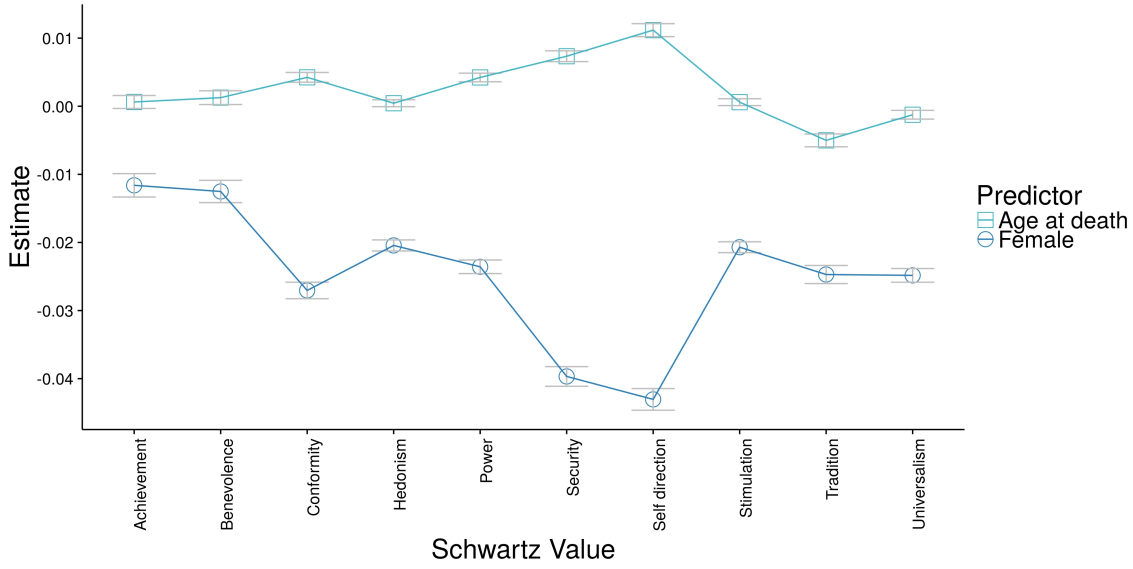

Figure 34. Plot showing coefficient estimates and standard errors for all Schwartz value word-by-hop Dependent Variables and all model predictors *for Model #3, the model with the lowest AIC, excluding the intercept* (in order to see the other estimates in more detail). Note that all predictors were z-scored before being entered into the model, and that the word-by-hop DV was logit-transformed.

Figure 32 indicates that, for the full model (Model #7), *some* race/ethnicity percentages did vary consistently with logit-transformed word-by-hop across Schwartz values. An increase of one SD in the county-level percentage of the population categorizing itself as "Black or African American" was associated with an estimated increase of over .10 logit-transformed word-by-hop values for Achievement, holding all other predictors constant. Percentage of the population that categorized itself as "White" consistently had the next-highest effect estimates across Schwartz values. Having said that, it is worth keeping in mind that race and ethnicity *as a whole* (i.e., as a set of predictors) did not substantially increase model fit.

CHAPTER IV

DISCUSSION

The current project explored the extent to which a sample of contemporary obituaries from across the USA used words related to values from Schwartz' (e.g., 2012) model. To do this, it employed word-by-hop, a new measure of lexical distance from a given set of target words and the Schwartz values they indicate. Obituaries were expected to be a particularly useful potential repository of values when viewed in aggregate, as, given their place in death-related ritual in the USA, they could be expected to be particularly charged with value-expressions (even if only of a positive valence), more so than general newspaper texts (such as those used by Bardi et al., 2008) or other public corpora.

The approach used in this project *did* reveal differences across Schwartz values as operationalized using word-by-hop. Power, the most-indicated Schwartz value, had a mean word-by-hop value across the corpus that was substantially (more than 2.25 of its Standard Deviations) above Hedonism and Stimulation, the least-indicated Schwartz values. Word-by-hop values were consistently small, as expected (as a value of 1 would indicate perfect overlap between lemmas from an obituary and target value lexicon words); despite this, it is encouraging and reasonable to interpret these findings as evidence that the word-by-hop measure is sensitive to relative variation, at least in aggregate, in the values (or at least the template-expressions of values) that a corpus indicates.

Stimulation and Hedonism were the least-indicated Schwartz values in the corpus. This value ordering makes sense with reference to Berger (1969, p. 43), who, as discussed above in this dissertation's introduction, argued that responses to death rely on "legitimations of the reality of the social world," presumably

including the moral world. Many systems of morality have to do with regulating Hedonism and Stimulation specifically, more so than the other Schwartz values. Further, compared with the other Schwartz values (including Power, Conformity, and Security, which were most indicated in the corpus), Hedonism and Stimulation are perhaps the least-social (even Self-direction, which, according to Schwartz, 2012, p. 5, is motivated by "independent thought and action–choosing, creating, exploring," is defined as belonging to a social context, even if eschewing it). These moral and less-social aspects of Hedonism and Stimulation may also interact with the *ritual* nature of obituary-writing in the USA, which places taboos on including statements that could be perceived by readers as immoral.

Power was the *most*-indicated value in the corpus. This result was unexpected, but does make sense following Berger, as above, as well as Fowler (2005), who noted that obituaries tend to emphasize individualism over communalism. Within Berger's framework of responding to death by emphasizing the social over the anti-social, Power is the Schwartz value that could be seen as the most individualistic while also social. While Achievement, e.g., is motivated by "personal success" (also emphasizing individualism; Schwartz, 2012, p. 5-9), Power is motivated by "social status and prestige, control or dominance over people and resources," emphasizing a distinct, more direct social concern.

One central question to address in follow-up work is what causes the bimodality seen when plotting word-by-hop values against obituary word-count (as in Figures 19, 20, 21, and 22). As noted above in the Results section, bimodal distributions were seen at least somewhat in all Schwartz values except for Hedonism and Stimulation, the two least-indicated values in the corpus. This could suggest that writing obituaries from templates is more common in longer

obituaries. If this is the case, it presents an interesting additional impetus for future research: following a template, whether explicit (e.g., "Fill in the names of the spouse, children, etc. of the deceased") or implicit (e.g., "Many people use the phrase 'Went to her eternal home,' so I will, as well"), is inherently a cultural act: it is based in imitation of others in one's community. Future work could thus not only explore obituaries from this corpus for markers of greater linguistic similarity (which could indicate that authors tended to follow an implicit or explicit template), but also the traits of authors of prospective obituaries, whose writing process could be observed in a more controlled setting. Incorporating the clustering coefficient of nodes (i.e., how prone to clustering an obituary's words are) into future analyses and even the word-by-hop calculation itself would be one way to assess templating tendencies.

In the best-fitting regression model, gender (male/female) had a consistently non-zero estimated effect on predicted logit-transformed word-by-hop values. This finding is in line with those of Rodler et al. (2001), who observed different changes over time in the attributes used in obituaries to describe male vs. female leaders. Given that the effect was small even in the most extreme case (for Self-direction), however, they also follow Schwartz et al. (2001), who observed that gender tended to be useful for explaining variance in Schwartz values data, but not to a large extent (and not in all cultures). Interestingly, disregarding the idiosyncrasy of individual newspapers, female status resulted in *lower* predictions across all Schwartz values. Acknowledging that these differences were small (the effect estimates are on the scale of *logit-transformed* word-by-hop values), this finding is in contrast with Schwartz and Rubel (2005), who found (albeit possibly as the result of a sample/measurement interaction) that women did tend

to place higher emphasis on some values (such as Universalism) than men. If this result in the current project is both accurate and large enough to warrant conceptual attention (vs. solely statistical attention), however, it is useful to note that obituaries as a data source are unlike typical self-report measures of Schwartz values: most relevantly, they are written by informants (family members, friends, or colleagues, commissioned biographers), and are possibly written not actually about the deceased individual herself, but rather about an *idealized (or at least curated) version* of the individual, reflecting perhaps as much in this context about the author as about the individual. Thus, this finding prompts the question, Regardless of whether women report prioritizing certain values at a higher rate than men at the individual level (as in Schwartz and Rubel's report), do authors (whether women or men) express these same value preferences *when publicly signalling them* about a gendered other (such as a decedent)? If not, obituaries could be expected to show higher gender differences for communities that have higher levels of Tradition, and could *possibly* indicate that some communities culturally see men more than women as containers of values when described by others.

The best-fitting statistical model (considering AIC) for predicting word-by-hop values was one that excluded all newspaper/county-level covariates, including race/ethnicity, income, and education. Given findings reported in previous literature, this result was unexpected. Specifically, it is counter to Schwartz et al. (2001, p. 534), who found positive, significant relationships between education and Self-direction and Stimulation: in the full model that included it, education had a predicted near-zero or *negative* effect for both of these values as measured through word-by-hop. It is also not in line with Schwartz et al.'s finding of negative correlations between education and Conformity and Tradition: while education's

predicted relationship in the full model was negative for both of these Schwartz values, the models that included education were not the best-fitting. It is worth considering, however, that education was operationalized in this project quite differently than in Schwartz et al.'s report. First, this project considered the percentage of the population with a Bachelor's degree or higher, rather than, as in Schwartz et al.'s report, a scale between no educational experience and experience "beyond high school" (p. 534). Second, in the current project, education was measured not at the individual level but at the county level. The inclusion of this covariate, as well as those for income and race/ethnicity, was based on the assumption that, *in aggregate,* community-level demographic indicators would provide a "signal" detectable even through the "noise" introduced by not modeling these characteristics at the individual level, which would have been infeasible or impossible to do with this dataset. The current project's findings cast doubt on this assumption and suggest that future research with a smaller number of obituaries that can be manually coded for at least some of these and related covariates (e.g., education, even if not income or self-identified race/ethnicity) could be fruitful for re-examining those covariates' statistical utility. As it was modeled in this project, even income, which, as discussed above in this dissertation's introduction, seemed *definitionally* related to Power, had among the lowest fixed effect estimates for Power in the full model.

Having acknowledged that the race/ethnicity covariates did not contribute to the model overall, it is worthwhile to tentatively interpret the full model (Model #7; Figure 32) as a starting point for future work. Even with a small effect, the model indicated most notably that for every increase of 1 SD in the community-level percentage of individuals who self-identified as "Black or African American"

and, to a lesser extent, of those who self-identified as "White," estimated logit-transformed word-by-hop values were expected to increase by a non-zero amount for all Schwartz values. This *may* indicate that areas with higher percentages of individuals who identify as part of those groups may have norms for less formalization or greater vividness in their language. However, they may also indicate an effect that may be true for other racial/ethnic groups as well but is masked by the data. Only English-language newspapers were considered in this project, possibly heavily misestimating the "true" word-by-hop values that would result from incorporating obituaries from Spanish-speaking media in areas with high populations that identify as "Hispanic." Somewhat similarly, Figure 10 indicates that most communities included in the sample contained less than 0.5% individuals who self-identified as "Native Hawaiian and Other Pacific Islander." It is an open question whether the effect estimate for this group would be similar if a newspaper from Hawaii had been included in the sample (as the caption for Figure 3 notes, *Legacy.com* does not currently contract with any newspapers in Hawaii).

This project expands Bardi et al.'s (2008) approach of understanding values through ecologically-valid text analysis in two primary aspects. First, the obituaries used in this project may be expected to comprise a more values-focused corpus than newspaper texts in general, potentially amplifying the ability to detect values in the text. Second, this project expands the approach of values-detection from counting instances of specific words to analyzing the lexical distance *for every word in a text* to those words, even if they do not appear explicitly in a given text. The development of word-by-hop will, I hope, be of use for future research not only in its concept but also in its implementation: the code that was written and documented as part of this project will be released alongside this dissertation

under an open license, ideally enabling future analyses of other sources of natural language, possibly with expanded or alternative target lexica. Additionally, the codebase includes automated algorithms for guessing gender and age; a WordNet specification (and documentation) for the Neo4J graph database platform; the University of South Florida Free Associations dataset released (to this author's knowledge) for the first time publicly under an explicit and open license, thanks to the generosity of Nelson, McEvoy, and Schreiber (2004); and the hop count dataset derived from the obituaries corpus[1].

The approach taken by this project included several notable limitations, several of which may serve as fodder for follow-up research. Most sweepingly, this project necessarily incorporated several conceptually "fuzzy" methodological assumptions and approaches, as is common in analyses of free-text data. As noted above, Figures 15, 16, 17, and 18 suggest that some obituaries were misclassified by the automated age-guessing algorithm (and thus likely also by the automated gender-coding algorithm). The use of some non-noun final Synset nodes when calculating word-by-hop values (as shown in Table 10) reflects the open, exploratory conceptual approach of this project, which had to release some definitional rigidity in order to utilize the free-text dataset. The lack of readily-available baseline English-language word frequency data (such as from the Corpus of Contemporary American English) prevented word-by-hop values from being normalized using lemmas' baseline frequencies; it is possible that future inclusion of this type of auxiliary data could substantively improve the word-by-hop calculation. These aspects require that this project be seen accurately as a first step toward expanding values research that uses this type of corpus. Moving forward from the

[1]With *Legacy.com's* permission, raw obituary data may also be released in the future.

foundation provided by this project, including not only its findings but also its associated code, algorithmic refinements would be a fruitful next area of focus.

This project's findings point toward several additional exciting areas of future research. Having demonstrated in a fully free-text corpus that word-by-hop values do show sensitivity to relative differences in value-expressions, future work could further explore both the properties of word-by-hop (seeking to explain, e.g., the cone-shaped distributions seen in Figures 19, 20, 21, and 22), and new ways to model it (e.g., comparing the logit-transformation approach used in the models for this project with beta- and/or multiple-membership regression approaches). Future analyses on this obituary corpus could explore additional individual-level characteristics, such as veteran status (e.g., by searching for phrases such as "Army," "Navy," etc.), family membership (by attempting to create a graph of types of familial relations from the descriptions of surviving family members), and the overall sentiment of words in each obituary (using, e.g., SentiWordNet, a fork of the WordNet database used in this project with an additional positive- vs. negative-sentiment layer; see Baccianella, Esuli, & Sebastiani, 2010). While obituaries may be expected to be generally positive in tone, i.e., not disparaging the deceased, they may vary in positive vs. negative word choice.

Given that newspaper/county-level predictors were not statistically useful (comparing models that included them with a model that excluded them), future work could usefully explore the effects of aggregating covariates either by *larger* geographic regions or, perhaps most likely to alter results, by explicitly modeling newspapers as multiple members of counties (and, relatedly, obituaries as multiple members of newspapers, and therefore counties; see Leckie, 2013). Alternatively, and given the gender-related question posed above, a future project could compare

the values invoked in free-text *prospective* obituaries (either about self or a close other) written by living participants with their responses to the Schwartz Values Survey (SVS).

As measured by word-by-hop, Hedonism was highly correlated with Conformity, its opposite in the Schwartz circumplex model, in the obituaries corpus. The reasons for this are unclear. However, they do suggest that the correlation matrix presented in Figure 27 may not reproduce the structure of Schwartz values that is expected based on previous analyses of individual-level responses to measures such as the SVS (Schwartz, 2012). This may be the result of obituary texts being able to express values that are "opposed" to one another *more consistently in aggregate* than value rankings made by living people (e.g., in response to the SVS). The structure of the correlation matrix from the obituaries corpus could, therefore, usefully be explored (e.g., through factor analysis) in future work in order to examine whether obituaries not only show differences in values but also reproduce the expected mappings of those values in relation to each other. Similarly, using the word-by-hop approach, all Schwartz-value-relevant words could be removed from the corpus (by removing all words that have a word-by-hop value above a given threshold for a Schwartz value). The occurrence of the remaining words in the corpus could then be factor-analyzed to look for clusters of words that fit Schwartz' (1992, p. 4) general definition of values, but do not fit any of the Schwartz values themselves.

This dissertation project had two goals at its outset: first, to describe which Schwartz values are discussed most across newspaper-communities in the USA, and second, to provide the initial development of new tools for future obituary-based "morality mining" work (Christen et al., 2013). Addressing the first goal, results

revealed that Power was the most-indicated value across the corpus, followed by Conformity, Security, Self-direction, and Benevolence. Stimulation and Hedonism were least-indicated, followed by Universalism, Tradition, and Achievement. Addressing the second goal, this dissertation is supplemented with algorithm definitions and code as well as datasets that I hope will facilitate responsible future research of this type. Results of this project indicated that the answer to Research Question 1, which asked whether the Schwartz values can be detected in obituaries, is "Yes," with the qualification that part of the "signal" being detected may be an artifact of authors following explicit or implicit templates (although template-following itself can indicate an adherence to community norms). In response to Research Question 2, whether Schwartz value expression is related to characteristics of obituaries at the individual and community levels, the answer is a more qualified "Yes." Results indicated that obituaries certainly vary by community in their distance from the Schwartz values; however, the effect sizes for individual and community-level covariates were small.

Obituaries are written in response to individual events, but can be viewed in aggregate as the output of a shared cultural ritual of externalizing grief and presenting it to one's wider community. This project explored an expanded methodology for traversing the constellation of obituaries' words' meanings, indicating that obituaries, whether read singly or together, can inform understanding of community value priorities.

APPENDIX

R BASE AND LIBRARY VERSION NUMBERS

Version numbers of R base and R packages used in this project. This table was generated automatically from package documentation within R; author names are therefore as the authors wished them to be printed.

Package Name	Version Number (NA if not installed through CRAN)	Package Citation
base	3.3.1	R Core Team (2016)
bbmle	1.0.18	Bolker and Team (2016)
betareg	3.1.0	Zeileis, Cribari-Neto, Gruen, and Kosmidis (2016)
boot	1.3.18	Canty and Ripley (2016)
choroplethr	3.5.2	Lamstein and Johnson (2016)
choroplethrMaps	1.0	Lamstein (2014)
choroplethrZip	(NA)	
coda	0.18.1	Plummer et al. (2015)
coefplot	1.2.4	Lander (2016)
coefplot2	(NA)	
corrplot	0.77	Wei and Simko (2016)
cowplot	0.6.3	Wilke (2016)
doParallel	1.0.10	Analytics and Weston (2015)

Package Name	Version Number (NA if not installed through CRAN)	Package Citation
dplyr	0.5.0	Wickham and Francois (2016)
fmsb	(NA)	
foreach	1.4.3	Revolution Analytics and Weston (n.d.)
foreign	0.8.66	R Core Team (2015)
gender	0.5.1	Mullen (2015)
ggmaps	(NA)	
ggmcmc	1.1	i Marn (2016)
ggplot2	2.1.0	Wickham and Chang (2016)
glmmadmb	(NA)	
glmmADMB	0.8.3.3	Skaug, Fournier, Nielsen, Magnusson, and Bolker (2016)
grid	3.3.1	**?**
gridExtra	2.2.1	Auguie (2016)
gtable	0.2.0	Wickham (2016a)
Hmisc	3.17.4	Harrell (2016)
irr	0.84	Gamer, Lemon, and ¡puspendra.pusp22@gmail.com¿ (2012)

Package Name	Version Number (NA if not installed through CRAN)	Package Citation
jsonlite	1.0	Ooms, Temple Lang, and Hilaiel (2016)
knitr	1.13	Xie (2016)
koRpus	0.6.5	m.eik michalke (2016)
lme4	1.1.12	Bates, Maechler, Bolker, and Walker (2016)
lme4a	(NA)	
magrittr	1.5	Bache and Wickham (2014)
methods	3.3.1	(Not provided by package author)
optparse	1.3.2	documentation et al. (2015)
parallel	3.3.1	(Not provided by package author)
plotrix	3.6.3	Lemon et al. (2016)
plyr	1.8.4	Wickham (2016b)
RColorBrewer	1.1.2	Neuwirth (2014)
readODS	1.6.2	Schutten and hong Chan (2016)
reshape	0.8.5	Wickham (2014b)
reshape2	1.4.1	Wickham (2014a)

Package Name	Version Number (NA if not installed through CRAN)	Package Citation
rjags	4.6	Plummer (2016)
rJava	0.9.8	Urbanek (2016)
RNeo4j	1.6.4	White (2016)
rstan	2.12.1	Guo, Gabry, and Goodrich (2016)
rvest	0.3.2	Wickham (2016c)
shiny	0.13.2	Chang, Cheng, Allaire, Xie, and McPherson (2016)
shinystan	2.2.1	Gabry (2016)
sqldf	0.4.10	Grothendieck (2014)
stringdist	0.9.4.1	van der Loo (2016)
stringr	1.0.0	Wickham (2015)
stringr_	(NA)	
testthat	(NA)	
visNetwork	1.0.1	Almende B.V. and Thieurmel (2016)
wordnet	0.1.11	Feinerer and Hornik (2016)
zoib	1.4.1	with contributions from Yunchuan Kong (2016)

REFERENCES CITED

Alfano, M. (2013). *Character as moral fiction.* Cambridge: Cambridge University Press.

Almende B.V., & Thieurmel, B. (2016). visnetwork: Network visualization using 'vis.js' library [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=visNetwork` (R package version 1.0.1)

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct.* `http://www.apa.org/ethics/code/index.aspx`.

Analytics, R., & Weston, S. (2015). doparallel: Foreach parallel adaptor for the 'parallel' package [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=doParallel` (R package version 1.0.10)

Auguie, B. (2016). gridextra: Miscellaneous functions for "grid" graphics [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=gridExtra` (R package version 2.2.1)

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 2200–2204).

Bache, S. M., & Wickham, H. (2014). magrittr: A forward-pipe operator for r [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=magrittr` (R package version 1.5)

Bardi, A., Calogero, R. M., & Mullen, B. (2008, May). A new archival approach to the study of values and value–behavior relations: Validation of the value lexicon. *The Journal of Applied Psychology*, *93*(3), 483–497. doi: 10.1037/0021-9010.93.3.483

Bates, D. (2013, September). *Son behind vicious obit to mother that went viral speaks.* `http://www.dailymail.co.uk/news/article-2419429/Patrick-Reddick-wrote-vicious-obituary-abusive-mother-unrepentant.html`.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2016). lme4: Linear mixed-effects models using 'eigen' and s4 [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=lme4` (R package version 1.1-12)

Bennett, W. J. (1996). *The book of virtues* (1st ed.). New York: Simon & Schuster.

Berger, P. L. (1969). *The sacred canopy: Elements of a sociological theory of religion*. Garden City, N.Y.: Doubleday.

Bolker, B., & Team, R. D. C. (2016). bbmle: Tools for general maximum likelihood estimation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=bbmle` (R package version 1.0.18)

Bou, B., & Princeton University. (2014). *WordNet SQL (WNSQL)*. `http://wnsql.sourceforge.net/`.

Cabot, E. L. (1910). *Ethics for children: A guide for teachers and parents*. Boston: Houghton Mifflin.

Canty, A., & Ripley, B. (2016). boot: Bootstrap functions (originally by angelo canty for s) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=boot` (R package version 1.3-18)

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2016). shiny: Web application framework for r [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=shiny` (R package version 0.13.2)

Christen, M., Alfano, M., Bangerter, E., & Lapsley, D. (2013, May). Ethical issues of 'morality mining:' Moral identity as a focus of data mining. In H. Rahman & I. Ramos (Eds.), *Ethical data mining applications for socio-economic development* (pp. 1–21). Idea Group Inc (IGI).

Cribari-Neto, F., & Zeileis, A. (n.d.). *Beta regression in R.*

Davis, K., & Patterson, D. (2012). *Ethics of big data* (1st ed.). Cambridge: O'Reilly Media.

Decloux, S. (1991). *The Ignatian Way* (C. M. Buckley, Trans.). Chicago: Loyola University Press.

Diogenes Laertius. (c. 350/1853). *The lives and opinions of eminent philosophers* (C. D. Yonge, Trans.). London: H. G. Bohn.

documentation, T. L. D. S., examples ported from Allen Day's getopt package. Some documentation from the optparse Python module by the Python Software Foundation. Contributions from Steve Lianoglou, Nikelski, J., Mller, K., Humburg, P., & FitzJohn., R. (2015). optparse: Command line option parser [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=optparse` (R package version 1.3.2)

Doughty, C. (2014). *Smoke gets in your eyes: And other lessons from the crematory* (1st ed.). New York: W. W. Norton & Company.

Feinerer, I., & Hornik, K. (2016). wordnet: Wordnet interface [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=wordnet` (R package version 0.1-11)

Ferrari, S., & Cribari-Neto, F. (2004, August). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, *31*(7), 799–815. doi: 10.1080/0266476042000214501

Forstall, R. L., & United States Bureau of the Census Population Division. (1996). *Population of States and Counties of the United States: 1790 - 1990*.

Fowler, B. (2005, January). Collective memory and forgetting components for a study of obituaries. *Theory, Culture & Society*, *22*(6), 53–72. doi: 10.1177/0263276405059414

Fu, A. S., Plaut, V. C., Treadway, J. R., & Markus, H. R. (2014). Places, products, and people "make each other up:" Culture cycles of self and well-being. In *Geographical psychology: Exploring the interaction of environment and behavior* (pp. 275–300). Washington, DC, US: American Psychological Association.

Gabry, J. (2016). shinystan: Interactive visual and numerical diagnostics and posterior analysis for bayesian models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=shinystan` (R package version 2.2.1)

Gamer, M., Lemon, J., & ¡puspendra.pusp22@gmail.com¿, I. F. P. S. (2012). irr: Various coefficients of interrater reliability and agreement [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=irr` (R package version 0.84)

Gómez-Redondo, R., & García González, J. M. (2010). Emergence and verification of supercentenarians in Spain. In H. Maier, J. Gampe, B. Jeune, J.-M. Robine, & J. W. Vaupel (Eds.), *Supercentenarians* (pp. 151–172). Springer.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014, January). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. doi: 10.1037/a0034726

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. doi: 10.1037/a0021847

Grothendieck, G. (2014). sqldf: Perform sql selects on r data frames [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=sqldf (R package version 0.4-10)

Guo, J., Gabry, J., & Goodrich, B. (2016). rstan: R interface to stan [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rstan (R package version 2.12.1)

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). Clarendon Press.

Harrell, F. E., Jr. (2016). Hmisc: Harrell miscellaneous [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=Hmisc (R package version 3.17-4)

Haybron, D. M., & Tiberius, V. (2012, February). *Normative foundations for well-being policy* (Papers on Economics and Evolution). Philipps University Marburg, Department of Geography.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2011). *Acceptance and commitment therapy, second edition: The process and practice of mindful change* (2nd ed.). New York: The Guilford Press.

Helbing, D. (2015). *Thinking Ahead - Essays on Big Data, Digital Revolution, and Participatory Market Society*. Cham: Springer International Publishing.

i Marn, X. F. (2016). ggmcmc: Tools for analyzing mcmc simulations from bayesian inference [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ggmcmc (R package version 1.1)

Kestenbaum, B., & Ferguson, B. R. (2010). Supercentenarians in the United States. In H. Maier, J. Gampe, B. Jeune, J.-M. Robine, & J. W. Vaupel (Eds.), *Supercentenarians* (pp. 219–230). Springer.

Lamstein, A. (2014). choroplethrmaps: Contains maps used by the choroplethr package [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=choroplethrMaps (R package version 1.0)

Lamstein, A., & Johnson, B. P. (2016). choroplethr: Simplify the creation of choropleth maps in r [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=choroplethr (R package version 3.5.2)

Lander, J. P. (2016). coefplot: Plots coefficients from fitted models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=coefplot (R package version 1.2.4)

Leckie, G. (2013). Module 13: Multiple Membership Multilevel Models. In (pp. 1–61). Bristol: Center for Multilevel Modelling.

Legacy.com. (2016a). *About Legacy.com - Life Stories* [Text]. http://www.legacy.com/about/about-us.

Legacy.com. (2016b, October). *About us: Our history and our dedication to obituaries and memorials.* http://memorialwebsites.legacy.com/aboutus.aspx.

Legacy.com. (2016c, October). *Our international newspaper partners.* http://www.legacy.com/ns/about/newspapers/international.aspx.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., ... Ogle, D. (2016). plotrix: Various plotting functions [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=plotrix (R package version 3.6-3)

Liu, F., & Kong, Y. (2015). Zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. *R Journal*, *7*(2), 34–51.

McAndrew, S. (2014, June). *Brutal obituary reveals lives of abuse, neglect.* http://www.rgj.com/story/news/2014/06/25/brutal-obituary-reveals-lives-of-abuse-neglect/11322269/.

m.eik michalke. (2016). korpus: An r package for text analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=koRpus (R package version 0.06-5)

Mikkelson, B. (2013, September). *Death penalty.* http://www.snopes.com/media/iftrue/obituary.asp.

Mullen, L. (2015). gender: Predict gender from names using historical data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=gender (R package version 0.5.1)

Nagi, K. (2013). A new representation of WordNet using graph databases. In *DBKDA 2013: The Fifth International Conference on Advances in Databases, Knowledge, and Data Applications* (pp. 1–8).

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* http://w3.usf.edu/FreeAssociation/.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004, August). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. doi: 10.3758/BF03195588

Neuwirth, E. (2014). Rcolorbrewer: Colorbrewer palettes [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=RColorBrewer (R package version 1.1-2)

NewspaperArchive.com. (2016). *NewspaperARCHIVE.com: Find Your Family in 4.6 Billion Historic Names.* http://newspaperarchive.com/.

Nicholson, E. W. B., Caxton, W., & de Worde, W. (1891). *Ars moriendi: That is to saye the craft for to deye for the helthe of mannes sowle.* London: Bernard Quaritch.

Office for Human Research Protections. (2010, January). *45 CFR 46* [Text]. http://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html.

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline* (1st ed.). O'Reilly Media.

Ooms, J., Temple Lang, D., & Hilaiel, L. (2016). jsonlite: A robust, high performance json parser and generator for r [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=jsonlite (R package version 1.0)

Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, *51*(1), 111–126.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623.

Pappas, N., & Zelcer, M. (2014). *Politics and philosophy in Plato's Menexenus* (1st ed.). Milton Park, Abingdon, Oxon; New York: Routledge.

Plummer, M. (2016). rjags: Bayesian graphical models using mcmc [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rjags (R package version 4-6)

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., ... Magnusson, A. (2015). coda: Output analysis and diagnostics for mcmc [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=coda (R package version 0.18-1)

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Princeton University. (2010). *About WordNet.* https://wordnet.princeton.edu.

R Core Team. (2015). foreign: Read data stored by minitab, s, sas, spss, stata, systat, weka, dbase, ... [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=foreign (R package version 0.8-66)

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Revolution Analytics, & Weston, S. (n.d.). foreach: Provides foreach looping construct for r [Computer software manual].

Rodler, C., Kirchler, E., & Hölzl, E. (2001, December). Gender stereotypes of leaders: An analysis of the contents of obituaries from 1974 to 1998. *Sex Roles*, *45*(11-12), 827–843. doi: 10.1023/A:1015644520770

Rosel, N. (1978). Toward a social theory of dying. *Omega: Journal of Death and Dying*, *9*(1), 49–55. doi: 10.2190/FGXD-GFGU-H3QM-GCR0

Russell, J. G. (2012, September). *Copyright and the obit.* http://www.legalgenealogist.com/blog/2012/09/12/copyright-and-the-obit/.

Saucier, G. (2009). What are the most important dimensions of personality? Evidence from studies of descriptors in diverse languages. *Social and Personality Psychology Compass*, *3*(4), 620–637. doi: 10.1111/j.1751-9004.2009.00188.x

Schmid, H. (2013). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing* (pp. 1–9).

Schutten, G.-J., & hong Chan, C. (2016). readods: Read ods files [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=readODS (R package version 1.6.2)

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*, *25*, 1–65.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1), 1–20.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001, January). Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross-Cultural Psychology*, *32*(5), 519–542. doi: 10.1177/0022022101032005001

Schwartz, S. H., & Rubel, T. (2005). Sex differences in value priorities: Cross-cultural and multimethod studies. *Journal of Personality and Social Psychology*, *89*(6), 1010–1028. doi: 10.1037/0022-3514.89.6.1010

Scime, A., & Murray, G. R. (2013, May). Social science data analysis: The ethical imperative. In H. Rahman & I. Ramos (Eds.), *Ethical data mining applications for socio-economic development* (pp. 131–147). Idea Group Inc (IGI).

Skaug, H., Fournier, D., Nielsen, A., Magnusson, A., & Bolker, B. (2016). glmmadmb: Generalized linear mixed models using 'ad model builder' [Computer software manual]. (http://glmmadmb.r-forge.r-project.org, http://admb-project.org)

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, *11*(1), 54.

Snopes.com. (2016). `http://www.snopes.com/`.

Solove, D. J. (2007). "I've Got Nothing to Hide" and Other Misunderstandings of Privacy. *San Diego Law Review*, *44*(GWU Law School Public Law Research Paper No. 289), 745.

The Associated Press. (2013, September). *Son who wrote vicious obit for Reno mom insists 'everything in there was completely true'.* `http://www.nydailynews.com/news/national/son-wrote-vicious-obit-reno-mom-insists-completely-true-article-1.1454890`.

Tsvetovat, M., & Kouznetsov, A. (2011). *Social Network Analysis for Startups: Finding connections on the social web* (1st ed.). Sebastopol, CA: O'Reilly Media.

United States Bureau of the Census. (2015a). *2010-2014 American Community Survey 5-Year Estimates (Generated by Jacob Levernier using American FactFinder)* (Tech. Rep.).

United States Bureau of the Census. (2015b). *Small Area Income and Poverty Estimates* (Tech. Rep.).

United States Bureau of the Census Population Division. (2015, June). *CC-EST2014-ALLDATA-[ST-FIPS]: Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2014* (Tech. Rep.).

University of California, Los Angeles Institute for Digital Research and Education. (2016). *Stata FAQ: How does one do regression when the dependent variable is a proportion?* `http://www.ats.ucla.edu/stat/stata/faq/proportion.htm`.

Urbanek, S. (2016). rjava: Low-level r to java interface [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rJava` (R package version 0.9-8)

Valeski, J. (2012, November). Social media: Erasable ink? In Q. E. McCallum (Ed.), *Bad data handbook: Cleaning up the data so you can get back to work* (1st ed., pp. 213–224). Sebastopol, CA: O'Reilly Media.

van der Loo, M. (2016). stringdist: Approximate string matching and string distance functions [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=stringdist` (R package version 0.9.4.1)

Wei, T., & Simko, V. (2016). corrplot: Visualization of a correlation matrix [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=corrplot` (R package version 0.77)

White, N. (2016). Rneo4j: Neo4j driver for r [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=RNeo4j` (R package version 1.6.4)

Wickham, H. (2014a). reshape2: Flexibly reshape data: A reboot of the reshape package. [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=reshape2` (R package version 1.4.1)

Wickham, H. (2014b). reshape: Flexibly reshape data. [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=reshape` (R package version 0.8.5)

Wickham, H. (2015). stringr: Simple, consistent wrappers for common string operations [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=stringr (R package version 1.0.0)

Wickham, H. (2016a). gtable: Arrange 'grobs' in tables [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=gtable (R package version 0.2.0)

Wickham, H. (2016b). plyr: Tools for splitting, applying and combining data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=plyr (R package version 1.8.4)

Wickham, H. (2016c). rvest: Easily harvest (scrape) web pages [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rvest (R package version 0.3.2)

Wickham, H., & Chang, W. (2016). ggplot2: An implementation of the grammar of graphics [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ggplot2 (R package version 2.1.0)

Wickham, H., & Francois, R. (2016). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 0.5.0)

Wilke, C. O. (2016). cowplot: Streamlined plot theme and plot annotations for 'ggplot2' [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=cowplot (R package version 0.6.3)

with contributions from Yunchuan Kong, F. L. (2016). zoib: Bayesian inference for beta regression and zero-or-one inflated beta regression [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=zoib (R package version 1.4.1)

Xie, Y. (2016). knitr: A general-purpose package for dynamic report generation in r [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=knitr (R package version 1.13)

Zeileis, A., Cribari-Neto, F., Gruen, B., & Kosmidis, I. (2016). betareg: Beta regression [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=betareg (R package version 3.1-0)