Appendix B

APPLICATION AND RESOURCES USED IN THE RESEARCH

GEPHI (http://gephi.org/)

Gephi is a tool for data analysts and scientists keen to explore and understand graphs. Like Photoshop™ but for graph data, the user interacts with the representation, manipulate the structures, shapes and colors to reveal hidden patterns. Gephi's goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. Gephi is used for Exploratory Data Analysis, a paradigm appeared in the Visual Analytics field of research. Gephi's feature include dynamic graph analysis. Users can visualize how a network evolve over time by manipulating the embedded timeline (see Figure 1).

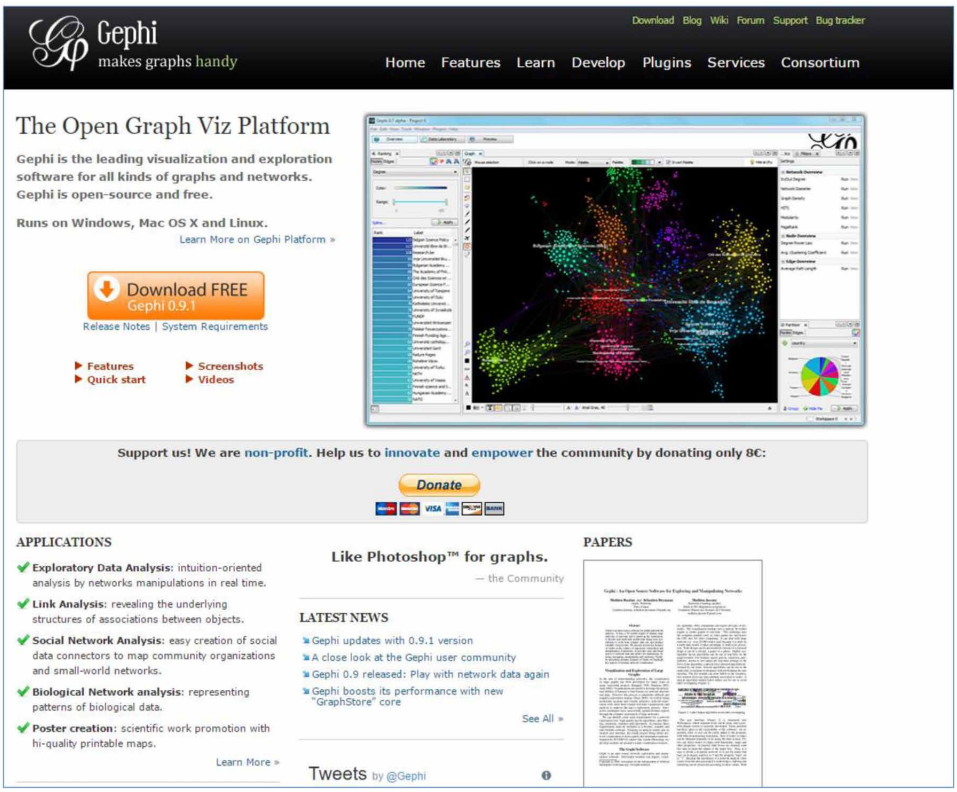
Sanford Topic Modeling Toolbox (https://nlp.stanford.edu)

The Stanford Topic Modeling Toolbox (TMT) brings topic modeling tools to social scientists and others who wish to perform analysis on datasets that have a substantial textual component. The toolbox features that ability to:

- Import and manipulate text from cells in Excel and other spreadsheets;
- Train topic models (LDA, Labeled LDA, and PLDA new) to create summaries of the text;
- Select parameters (such as the number of topics) via a data-driven process;

Appendix B

Figure 1. Snapshot of Gephi's website



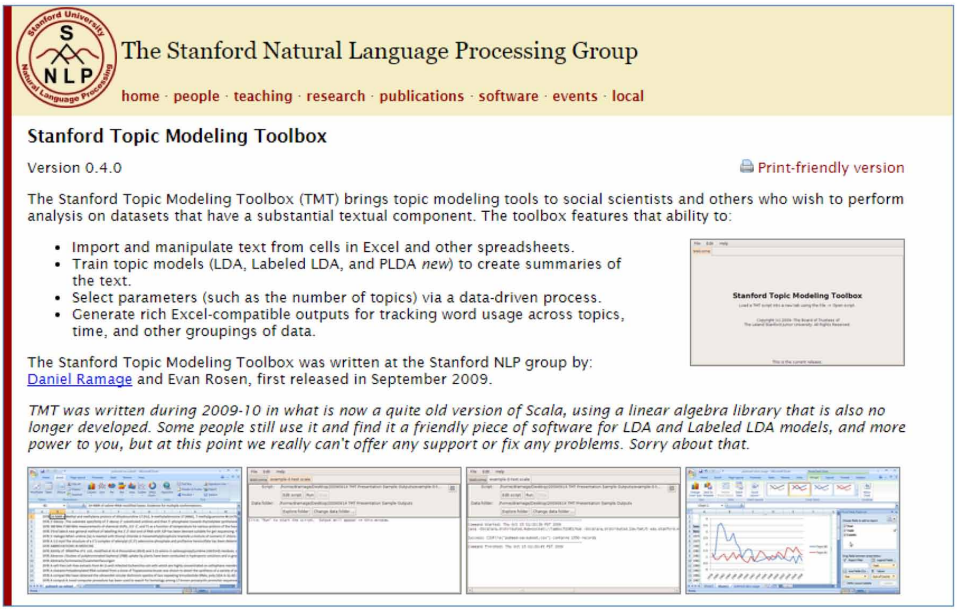
- Generate rich Excel-compatible outputs for tracking word usage across topics, time, and other groupings of data.

The Stanford Topic Modeling Toolbox (see Figure 2) was written at the Stanford NLP group by Daniel Ramage and Evan Rosen, first released in September 2009.

NodeXL (http://www.smrfoundation.org)

The Social Media Research Foundation is the home of NodeXL – Network Overview Discovery and Exploration for Excel (2010, 2013 and 2016) – extending the familiar spreadsheet so user can collect, analyze and visualize complex social networks from Twitter, Facebook, Youtube and Flickr. NodeXL

Figure 2. Snapshot of Stanford Topic Modeling Toolbox (TMT)'s website



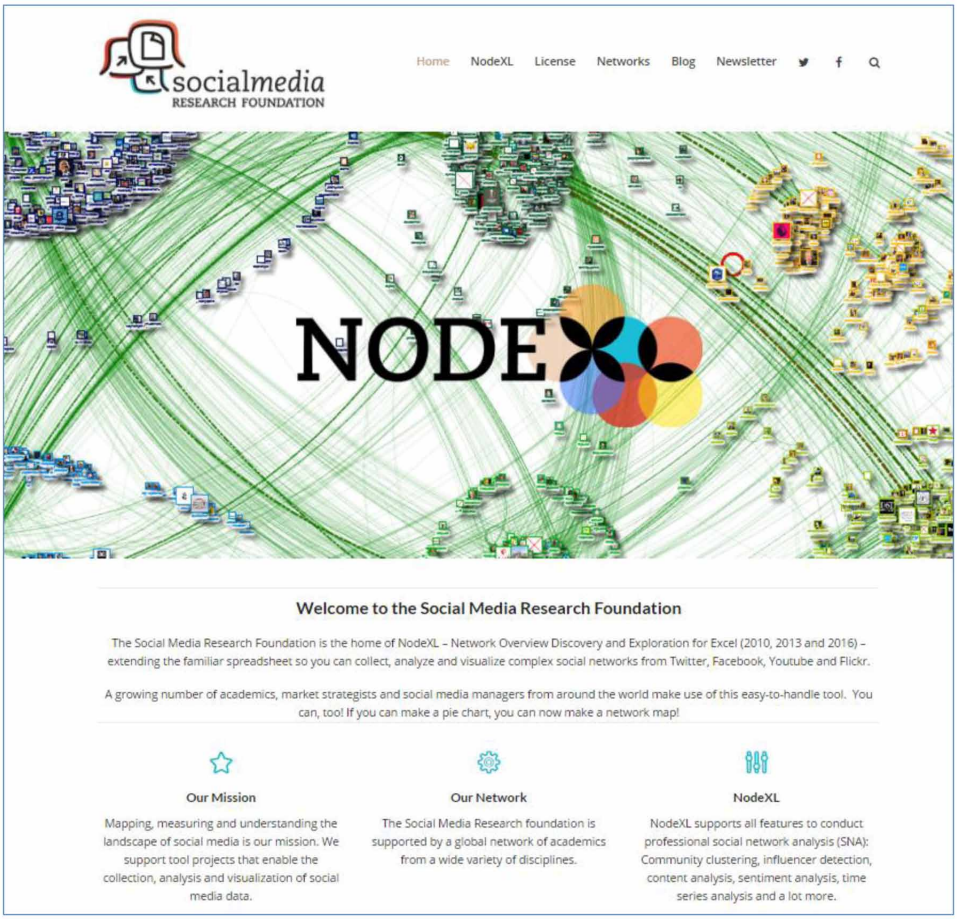
supports all features to conduct professional social network analysis (SNA): Community clustering, influencer detection, content analysis, sentiment analysis, time series analysis and a lot more (see Figure 3).

OYESTER Entity Resolution (https://sourceforge.net/projects/oysterer/)

The OYSTER Open Source Project is sponsored by the Center for Advanced Research in Entity Resolution and Information Quality (ERIQ) at the University of Arkansas at Little Rock. It is intended to provide an entity resolution system that includes functionality for entity identity information management (EIIM). Originally developed as a teaching tool, it now has enough capability to support record linking, EIIM, and master data management (MDM) processes in small to medium-sized organizations. OYSTER is designed to be easily configurable through the use of several, run-time XML scripts that define such things as the format and locations of reference sources to be processed, access to previously defined identity structures, identity rules and associated matching algorithms, as well as many parameters that adjust system

Appendix B

Figure 3. Snapshot of NodeXL's website

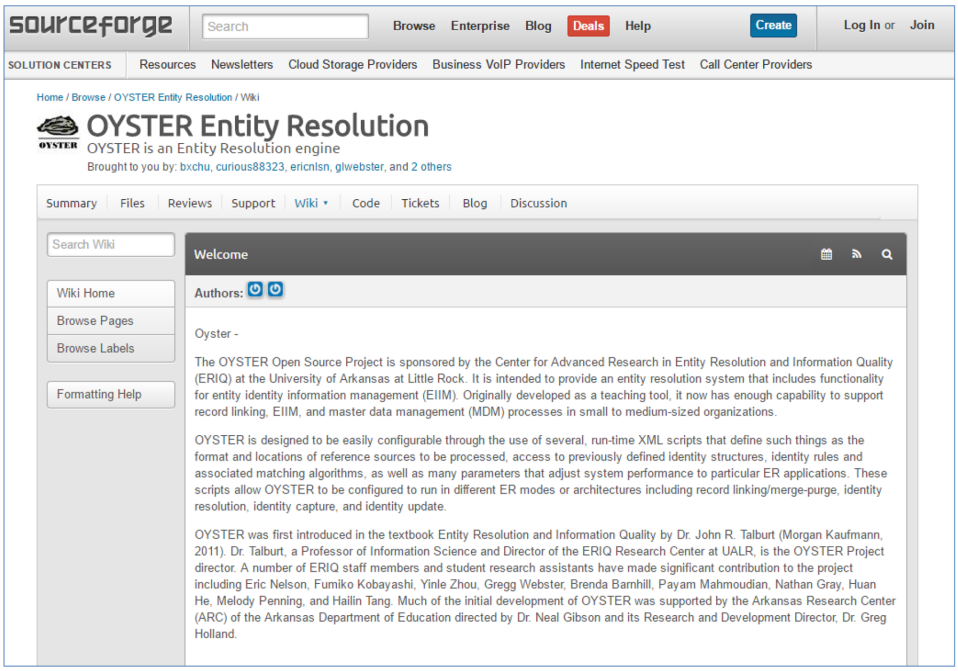


performance to particular ER applications. These scripts allow OYSTER to be configured to run in different ER modes or architectures including record linking/merge-purge, identity resolution, identity capture, and identity update (see Figure 4).

LIWC (Linguistic Inquiry and Word Count) (http://liwc.wpengine.com/)

LIWC is a powerful research and learning tool based on solid science. LIWC allows users to look under the hood of works of literature. The way that the Linguistic Inquiry and Word Count (LIWC) program works is fairly simple.

Figure 4. Snapshot of Oyster's website



Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. Because LIWC was developed by researchers with interests in social, clinical, health, and cognitive psychology, the language categories were created to capture people's social and psychological states. The psychometric validation of LIWC categories is significant because it allows LIWC users to draw justified inferences *from* word frequencies *to* psychological states of the authors. The LIWC program includes the main text analysis module along with a group of built-in dictionaries. The text analysis module was created in the Java programming language and runs identically on PC and Mac computers. LIWC reads written or transcribed verbal texts which have been stored in a digital, computer-readable form (such as text files). The text analysis module then compares each word in the text against a user-defined dictionary. As described below, the dictionary identifies which words are associated with which psychologically-relevant categories. After the processing module has read and accounted for all words in a given text, it calculates the percentage of total words that match each of the dictionary categories. For example, if LIWC analyzed a single speech that was 2,000 words and compared them

to the built-in LIWC dictionary, it might find that there were 150 pronouns and 84 positive emotion words used. It would convert these numbers to percentages, 7.5% pronouns and 4.2% positive emotion words.

Whereas the text analysis module identifies and categorizes words, the heart of the program is a group of dictionaries that tell the text analysis module which words to identify and classify. LIWC comes with three internal dictionary systems. The LIWC master dictionary is composed of almost 6,400 words, word stems, and selected emoticons. For each dictionary word, there is a corresponding dictionary entry that defines one or more word categories. For example, the word cried is part of five word categories: Sadness, Negative Emotion, Overall Affect, Verb, and Past Focus. Hence, if the word cried was found in the target text, each of these five subdictionary scale scores would be incremented. As in this example, many of the LIWC categories are arranged hierarchically. All sadness words, by definition, will be categorized as negative emotion and overall affect words. All anger words, by definition, will be categorized as negative emotion and overall emotion words.

Each of the 82 preset LIWC categories used in this research is composed of a list of dictionary words that define that scale. Table 1 provides a comprehensive list of these LIWC categories with sample scale words.

Table 1. LIWC categories with sample words

LIWC Dimensions and Sample Words	
Dimension	**Examples**
I. Standard Linguistic Dimensions	
Pronouns	I, them, itself
Articles	a, an, the
Past tense	walked, were, had
Present tense	Is, does, hear
Future tense	will, gonna
Prepositions	with, above
Negations	no, never, not
Numbers	one, thirty, million
Swear words	*****
II. Psychological Processes	
Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin

continued on following page

Table 1. Continued

LIWC Dimensions and Sample Words	
Dimension	**Examples**
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad
Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without
Perceptual Processes	see, touch, listen
Seeing	view, saw, look
Hearing	heard, listen, sound
Feeling	touch, hold, felt
Biological Processes	eat, blood, pain
Body	ache, heart, cough
Sexuality	horny, love, incest
Relativity	area, bend, exit, stop
Motion	walk, move, go
Space	Down, in, thin
Time	hour, day, oclock
III. Personal Concerns	
Work	work, class, boss
Achievement	try, goal, win
Leisure	house, TV, music
Home	house, kitchen, lawn
Money	audit, cash, owe
Religion	altar, church, mosque
Death	bury, coffin, kill
IV. Spoken Categories	
Assent	agree, OK, yes
Nonfluencies	uh, rr*
Fillers	blah, you know, I mean

Appendix B

Table 2a. LIWC 2015 output variable information

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word count	WC	-	-	-	-
Summary Language Variables					
Analytical thinking	Analytic	-	-	-	-
Clout	Clout	-	-	-	-
Authentic	Authentic	-	-	-	-
Emotional tone	Tone	-	-	-	-
Words/sentence	WPS	-	-	-	-
Words > 6 letters	Sixltr	-	-	-	-
Dictionary words	Dic	-	-	-	-
Linguistic Dimensions					
Total function words	funct	it, to, no, very	491	.05	.24
Total pronouns	pronoun	I, them, itself	153	.25	.67
Personal pronouns	ppron	I, them, her	93	.20	.61
1st pers singular	i	I, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
Impersonal pronouns	ipron	it, it's, those	59	.28	.71
Articles	article	a, an, the	3	.05	.23
Prepositions	prep	to, with, above	74	.04	.18
Auxiliary verbs	auxverb	am, will, have	141	.16	.54
Common Adverbs	adverb	very, really	140	.43	.82
Conjunctions	conj	and, but, whereas	43	.14	.50
Negations	negate	no, not, never	62	.29	.71
Other Grammar					
Common verbs	verb	eat, come, carry	1000	.05	.23
Common adjectives	adj	free, happy, long	764	.04	.19
Comparisons	compare	greater, best, after	317	.08	.35
Interrogatives	interrog	how, when, what	48	.18	.57
Numbers	number	second, thousand	36	.45	.83
Quantifiers	quant	few, many, much	77	.23	.64
Psychological Processes					
Affective processes	affect	happy, cried	1393	.18	.57
Positive emotion	posemo	love, nice, sweet	620	.23	.64
Negative emotion	negemo	hurt, ugly, nasty	744	.17	.55
Anxiety	anx	worried, fearful	116	.31	.73
Anger	anger	hate, kill, annoyed	230	.16	.53
Sadness	sad	crying, grief, sad	136	.28	.70
Social processes	social	mate, talk, they	756	.51	.86
Family	family	daughter, dad, aunt	118	.55	.88

"Words in category" refers to the number of different dictionary words and stems that make up the variable category. All alphas were computed on a sample of ~181,000 text files from several of our language corpora (see Table 3). Uncorrected internal consistency alphas are based on Cronbach estimates; corrected alphas are based on Spearman Brown. See the Reliability and Validity section below. Note that the LIWC2015 dictionary generally arranges categories hierarchically. There are some exceptions to the hierarchy rules. For example, Social processes include a large group of words that denote social processes, including all nonfirstpersonsingular personal pronouns as well as verbs that suggest human interaction (talking, sharing) many of these

Table 2b. LIWC 2015 output variable information continued

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Friends	friend	buddy, neighbor	95	.20	.60
Female references	female	girl, her, mom	124	.53	.87
Male references	male	boy, his, dad	116	.52	.87
Cognitive processes	cogproc	cause, know, ought	797	.65	.92
Insight	insight	think, know	259	.47	.84
Causation	cause	because, effect	135	.26	.67
Discrepancy	discrep	should, would	83	.34	.76
Tentative	tentat	maybe, perhaps	178	.44	.83
Certainty	certain	always, never	113	.31	.73
Differentiation	differ	hasn't, but, else	81	.38	.78
Perceptual processes	percept	look, heard, feeling	436	.17	.55
See	see	view, saw, seen	126	.46	.84
Hear	hear	listen, hearing	93	.27	.69
Feel	feel	feels, touch	128	.24	.65
Biological processes	bio	eat, blood, pain	748	.29	.71
Body	body	cheek, hands, spit	215	.52	.87
Health	health	clinic, flu, pill	294	.09	.37
Sexual	sexual	horny, love, incest	131	.37	.78
Ingestion	ingest	dish, eat, pizza	184	.67	.92
Drives	drives		1103	.39	.80
Affiliation	affiliation	ally, friend, social	248	.40	.80
Achievement	achieve	win, success, better	213	.41	.81
Power	power	superior, bully	518	.35	.76
Reward	reward	take, prize, benefit	120	.27	.69
Risk	risk	danger, doubt	103	.26	.68
Time orientations	TimeOrient				
Past focus	focuspast	ago, did, talked	341	.23	.64
Present focus	focuspresent	today, is, now	424	.24	.66
Future focus	focusfuture	may, will, soon	97	.26	.68
Relativity	relativ	area, bend, exit	974	.50	.86
Motion	motion	arrive, car, go	325	.36	.77
Space	space	down, in, thin	360	.45	.83
Time	time	end, until, season	310	.39	.79
Personal concerns					
Work	work	job, majors, xerox	444	.69	.93
Leisure	leisure	cook, chat, movie	296	.50	.86
Home	home	kitchen, landlord	100	.46	.83
Money	money	audit, cash, owe	226	.60	.90
Religion	relig	altar, church	174	.64	.91
Death	death	bury, coffin, kill	74	.39	.79
Informal language	informal		380	.46	.84
Swear words	swear	fuck, damn, shit	131	.45	.83
Netspeak	netspeak	btw, lol, thx	209	.42	.82
Assent	assent	agree, OK, yes	36	.10	.39
Nonfluencies	nonflu	er, hm, umm	19	.27	.69
Fillers	filler	Imean, youknow	14	.06	.27

Table 3. Summary information for LIWC2015 statistics

	Blogs	Expressive Writing	Novels	Natural Speech	NY Time	Twitter
Total files	37295	6179	875	3232	34929	35269
Total authors	37295	2510	441	2174	Unknown	35269
Total words	119449058	2526709	57467183	2566446	26007632	23172994

words do not belong to any of the Social processes subcategories. Another example is Relativity, which includes a large number of words that cannot be found in any of its subcategories.