Submitted on: 01.06.2017

2017 IFLA International News Media Conference
27-28 April, 2017
Reykjavík, Iceland

Conspiracy Stories: Building Archives to Facilitate Narrative Analyses of Online Fake News

Peter Broadwell
Digital Library Program, UCLA Library, Los Angeles, California, USA
E-mail address: broadwell@library.ucla.edu

Abstract:

Misinformation masquerading as news is not a new phenomenon, but social media and other digital methods of dissemination arguably have rendered so-called "fake news" more pernicious in recent years. This project seeks to apply narrative analytical models first used to scrutinize online anti-vaccination theories to various species of misleading news stories. These techniques move beyond basic text mining by incorporating internal network models that are able to identify "actants" (people, organizations) within and across stories and to characterize interactions between them. Such models can assist ongoing efforts to develop more sophisticated taxonomies of fake news (propaganda, slanted news, pseudo-satire, etc.) by contributing novel network/narrative-oriented features to such discussions. Perhaps more importantly, understanding the narrative features of harmful misinformation can aid efforts to mitigate its influence.

This project requires the accumulation of a large corpus of "fake" news from online sources, as well as collaborations with institutions like the Internet Archive to assemble comparative corpora of plausibly "legitimate" news from both web-based and broadcast television sources. Recent discoveries concerning the role of social media and online advertising in the propagation of fake news are certainly relevant to this work, but the focus of this inquiry remains on making the actual contents of "legitimate" and "fake" news stories available for comparison via narrative analysis. This task involves a considerable amount of automated curation, which natural language processing tools such as the Internet Archive's experimental WANE (Web Archive Named Entity) extraction service are increasingly helping to facilitate. This paper will conclude with an overview of the materials gathered so far and the salient entities, actant groups, and narrative features extracted from initial case studies of the "Bridgegate" and "Pizzagate" incidents.

Keywords: fake news, web archiving, entity recognition, network analysis

Introduction

Propaganda, misinformation, yellow journalism, fake news: slanted, exaggerated, or simply false reports that seek to advance a particular agenda by assuming the superficial trappings of more objective news reporting have a long lineage in the history of news media. Yet recent events suggest that the growth of web-based news and social media have rapidly lowered the threshold of effort and resources necessary to produce and disseminate such fake news widely, while at the same time greatly amplifying its potential influence on large audiences. The same trends have simultaneously expanded both the opportunities and challenges associated with the large-scale collection and archiving of online news and related media, as the rising potency of online propaganda and misinformation masquerading as legitimate news also magnifies the importance of born-digital news archiving. In particular, the intertwined nature of these trends supports the assertion that instead of being conceived of as sedentary repositories of the "first draft of history" that latter-day historians may eventually sift through at their leisure, archives of web news – both "fake" and otherwise – should be designed to support and if possible accelerate the efforts of present-day researchers to investigate and ultimately gain a better understanding of new and possibly pernicious modes of online social discourse. Debates continue about the actual degree of influence fake news exerted upon political events such as the 2016 U.S. presidential election, and the meaningfulness of the term "fake news" continues to be diluted through overuse [1]. Yet given the trends cited above, the need to understand and hopefully mitigate the effects of online misinformation has arguably never been more urgent.

This project developed from a previous study into a different but related kind of online misinformation, namely the spread of anti-vaccination narratives via the postings and replies on parenting social media sites (colloquially referred to as "mommy blogs") [2]. The authors of the earlier study indexed millions of posts on parenting discussion sites over a multi-year period, then used statistical machine learning techniques to generate an aggregate narrative network model of these discussions, in which "actants" (entities such as people, places, and organizations) are connected by transitive relations inferred from the processed text of the posts (e.g., X "believes in" Y). This analysis identified a few dominant narratives that urge parents to seek vaccine exemptions for their children, findings that may eventually prove helpful to public health officials seeking to discourage these exemptions. In particular, it is likely that effective counter-narratives may prove more persuasive than the mere citation of facts when the target audience is emotionally committed to a certain point of view.

What follows here is a discussion of the initial steps taken to assemble a corpus of web-based "fake news" in order to facilitate a massive narrative framework analysis of online misinformation masquerading as news, using a modified version of the software previously applied to the study of anti-vaccination narratives. Accompanying the data-gathering discussion is a commentary on how current web-archiving approaches and frameworks might be enhanced to help achieve such research-oriented objectives. This paper additionally presents some initial results of small pilot studies conducted to test the narrative analytical techniques that ultimately will be scaled up to the level of millions of online postings. Because these subsequent studies are likely to compare the narrative "shapes" of news stories along a continuum from hoaxes to verifiable reporting, the pilot studies focus on archives of web materials based around two conspiracies: one that turned out to be real, namely, the so-called "Bridgegate" scandal of politically motivated lane closures on the George Washington Bridge, and one that was false: the so-called "Pizzagate" hoax.

Related work

Analysis of fake news in its various guises has been a growing area of research over the past several years. The "Hoaxy" project, from Indiana University's Observatory on Social Media, is part of a long-running effort to analyse the spread of information – and especially politically motivated misinformation – on social media [3]. The 2016 U.S. presidential election significantly increased the amount of attention the phenomenon has received from academics, archivists and journalists, with recent investigations reinforcing the key role social media plays in the dissemination of slanted or entirely fabricated news to already polarized online groups who are more likely to be receptive to its message [4]. Sensationalist headlines also tend to spur devotees of these communities to follow links from social media to the host of rapidly assembled websites presenting fabricated news stories within layouts that often mimic the gravitas of legitimate sites. Automated online advertising systems such as Google's AdSense, which generate a small amount of revenue each time such a "clickbait" page is visited, are the primary economic motivator of the systematic generation of fake news sites and articles, which proves especially lucrative when targeting the web's vast English-language readership [5]. In addition, it is plausible that alleged fake news "factories," such as the infamous case of the city of Veles, Macedonia, contributed to an "echo chamber" effect by sometimes using rumors and conspiracy theories emerging from polarized groups on Twitter, Facebook, Reddit, and other social media platforms as the inspiration for the fake news articles that they subsequently peddled to these groups [6].

Access to the comprehensive data holdings of large social media firms like Facebook and Twitter is quite difficult to obtain for non-associates of the companies. The semi-closed nature of these commercial platforms therefore makes it difficult and sometimes impossible for researchers and archivists to assemble collections that directly facilitate studies of the role these social media behemoths play in the promulgation of fake news. And although journalists and other researchers have in recent months developed a host of social and technological approaches to mitigate the corrosive effect of fake news on public discourse, including education campaigns, linking to countervailing facts on sites such as snopes.com, and various collaborative tagging, voting, and filtering systems, it ultimately falls to the social media firms themselves to apply such approaches directly to their platforms if they are to be truly effective [6].

It is certainly possible to incorporate small-scale collections or samples of social media information into particular case studies, however, as is the case with analysis of the Pizzagate "archive" discussed below. Given the previously discussed difficulties of accessing and working with social media data at scale, though, the contents of websites that contain large quantities of fake news articles seem to offer a more promising and readily available source of materials for a large-scale narrative analysis of fake news. Yet such work is not without its own challenges, as shown in the case studies below.

Data gathering: large-scale web crawling

Efforts to gather a large collection of fake news from the web began in December of 2016 via daily crawls of a series of "seed" URLs using the Internet Archive's Archive-It subscription service (built upon their Heritirix web crawler). These seeds were derived from a list of suspected fake news web sites that Professor Melissa Zimdars from the Department of Communications and Media at Merrimack College released shortly after the 2016 U.S. presidential election. As Prof. Zimdars and others have discussed, compiling such a list is fraught with complications: many of even the most suspicious sites may contain a mixture of news stories that range from truly fraudulent to merely slanted, while others can claim (correctly or not) to be primarily satirical and thus undeserving of accusations that they hawk "fake" news [8]. To minimize this categorical ambiguity, the crawl focused on 41 sites judged most likely to contain the highest proportion of fabricated stories.

Even assuming that the choice of seeds is ideal, however, building a reliable web archive that enables the study of fake news stories across time – an important objective of narrative analyses – requires nearly constant monitoring, as some sites periodically disappear and reappear at new URLs, and even well-established sites may change their internal formats in a way that makes archiving them more complicated (intentionally or not). Indeed, the inadequacy of web archiving technologies to keep up with the rapid pace of technological and content-based "churn" online has prompted some recent critiques of the entire practice of news web archiving [9]. A partial solution to some of these objections would involve increased coordination and collaboration between memory institutions, much of it likely automated and based on registries of the sites being archived [10]. The potential benefits of this approach – which include reduced duplication of effort, among others – were underscored when a simple search of public Archive-It collections, conducted soon after the fake news crawl began, revealed that staff members at the Internet Archive were building a more or less identical collection based on nearly the same set of seed URLs [11]. Given that the Internet Archive literally owns the technological underpinnings of the Archive-It service, delegating this task to them or to some other comparatively research- and expertise-rich organization seems likely to result in a larger and more complete data set. As a final point, it is important to remember that simply amassing a plausibly representative web archive does not by any means indicate that the data set should be considered ready for analysis, as the case of the "Bridgegate" archive below illustrates.

Case studies: preliminary analytical techniques and results

Assuming that large-scale web archiving eventually will produce a viable collection of fake news for large-scale narrative analysis, the next logical step of this research was to conduct small-scale pilot studies of news (or "newsy") materials from the web in order to determine how well the existing narrative analysis pipeline – which was built to process blog posts and their associated user comments – would deal with "fake news" articles and possibly their associated comments and social media content, as well as whether further modifications to the pipeline would be necessary. As discussed above, these pilot studies focused on small-scale archives related to the strange-but-true "Bridgegate" conspiracy scandal, and the bizarre-and-false "Pizzagate" conspiracy theory. In addition to providing fodder for basic functional testing, the case studies also offered opportunities to test various research-related questions and hypotheses: Could the narrative analysis software extract meaningful "actants" and relations from the new data sources? How would these results compare to other, less automated but sometimes quite tech-savvy analyses and visualizations of these events in the media? Would this work discover meaningful, categorical differences between the "shapes" of real and false conspiracy stories? These inquiries are still proceeding, but the experiences and discoveries thus far provide some useful insights into the relationship between online archives and studies of fake news.

Pizzagate

The "Pizzagate" data set consists primarily of the texts of approximately 20,000 Reddit posts from the /r/pizzagate conspiracy board. The posters used the stolen emails of Clinton campaign chairman John Podesta as the basis for a hoax which, according to Abby Ohlheiser of the Washington *Post*, "combines the Clinton diaspora with accusations of a secret pedophile ring" centered upon a pizzeria in Washington, DC [12]. This archive is unusual in that it was "self-archived" by devotees of the Pizzagate conspiracy after the discussion board was banned from Reddit for violating the site's policies regarding online bullying; the data set was subsequently removed from the open-source software and data site Github for similar reasons. It finally found a new home on Voat, a fringe Reddit-like site [13]. Even after the Pizzagate conspiracy's primary assertions were discredited (but not before inciting one adherent to fire a rifle in the pizzeria in question), the theorists established a "Pizzagate Wiki," which claims to document "facts and sources about crimes, potential abuses of power and disinformation conducted by governments around the world." [14]

sentence	**arg1**	**rel**	**arg2**
Fearing yet another witch hunt, Reddit bans Pizzagate - The Washington Post	{Reddit}	{bans}	{Pizzagate}
Citing its policy against posting the personal information of others, Reddit has banned the "Pizzagate" conspiracy board from the site.	{Reddit}	has {banned}	the Pizzagate conspiracy {board}
Assassination Of Finnish Reporters Investigating Pizzagate	{Assassination} Reporters	{Investigating}	{Pizzagate}
Strange: Police found various other items including a child's jumpsuit, pictured, in the basement	{Police}	{found}	various other including a child's {jumpsuit}{items} pictured
Washington gunman was investigating Pizza Gate fake news	Washington {gunman}	was {investigating}	Pizza Gate fake {news}
TURKEY AND OTHER COUNTRIES ARE TALKING ABOUT #PIZZAGATE ON TWITTER!	{TURKEY}	ARE {TALKING}	ABOUT {PIZZAGATE}
Short and to the point: For anyone wonerding whether Trump knows about Hillary and PizzaGate, here: Donald Trump drops Haiti joke on Hillary at Al Smith charity dinner	Donald {Trump}	{drops}	Haiti {joke} on Hillary
I'm sure there are tons of Jews who aren't pedos, and tons of pedos who aren't Jews.	{Jews}	{aren}	t {pedos}
The jews are attacking and defending pedophilia.	The {jews}	are {attacking}	{pedophilia}

Table 1: Pairs of "actants" and the relations between them discovered automatically in the /r/pizzagate Reddit posts by narrative network analysis tools.

As a downloadable archive, the Pizzagate data set is nearly 900 megabytes in size, most of which is images. The actual Reddit posts, as well as a small amount of related Web content, required some text normalization and cleaning in order to be used with the analysis software, but given that the content was to some degree already "curated," the amount of work required was relatively limited. And because the format of Reddit posts is fairly similar to the "mommy blog" comments analyzed in the previous vaccination study, the data preparation and analysis code required little modification. These efforts have yet to generate large-scale narrative networks from these posts, but Table 1 illustrates some sample actant-to-actant relations derived automatically from the texts (including hate speech).

It has also been possible to generate analogues of the relational graphs that the narrative analysis software eventually will produce by processing the Pizzagate archive text contents with the DBpedia Spotlight named-entity resolution tool and visualizing the results. The Spotlight tool resolves detected entity names in texts to discrete entities in a massive semantic database derived from Wikipedia. For example, the tool is able to use statistical calculations to resolve references to both "Hillary" and "Hillary Clinton" to the authoritative URI http://dbpedia.org/page/Hillary_Clinton, thereby eliminating ambiguous and redundant references from the data set [15]. By limiting the data set of resolved entities to people, places, and organizations and counting the number of times two different entities "co-occur" (i.e., are mentioned in the same post), it is possible to generate a network whose structure provides a rough narrative "snapshot" of the Pizzagate conspiracy as represented in the Reddit posts. Visualizations generated via network analysis software packages [16] share notable similarities with explanatory "infographics" of the Pizzagate conspiracy such as one published in the New York *Times* to explain the links – some undisputed, others farfetched and often libellous – that constitute the "web" of the conspiracy [17].

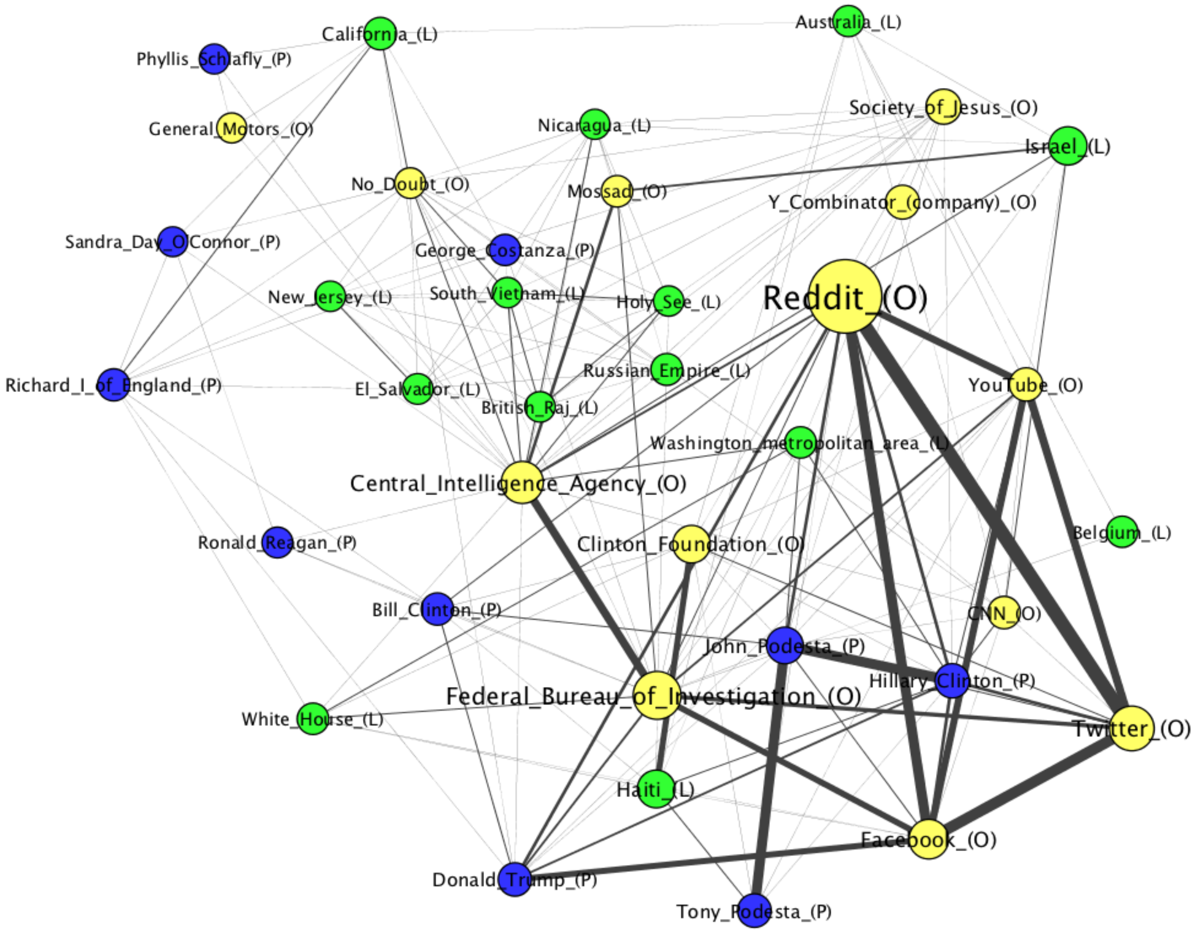

Figure 1: Network visualization of the most prominent people (P), locations (L) and organizations (O) in the Pizzagate data set as identified by the DBpedia Spotlight named entity resolution tool. The number of connections between nodes is determined by how frequently they co-occur in a single post. Nodes and edges are scaled in proportion to their degree (number of connections).

Figure 1 presents such a visualization of the Pizzagate conspiracy network generated from the actual Reddit posts at the center of the hoax, based on the 2,092 unique entities identified by

DBpedia Spotlight and their 3,395 co-occurrences. Entities with few connections (which constitute the majority of the names in the data set) have been removed from the graph in order to improve the legibility of its most noteworthy features. Some of the network's attributes, including the high "degree" of the Reddit and Twitter nodes, are likely artifacts of the data source, but others, such as the prominence of the FBI and CIA (something that is absent from the New York *Times* visualization) highlight what could be salient attributes of many conspiracy theories, e.g., a preoccupation with government intelligence agencies. The computer-aided quantification of such tendencies may eventually lead to an automated process to rank news articles on a scale from "suspected hoax" to "likely legitimate" based on the prevalence of these features.

A prime limitation of these preliminary visualizations is that using raw entity co-occurrence as the only means of linking entities neglects many other, potentially revelatory classes of relationships, which might be inferred if more conceptual entity types and ontological relations (some of which are available in DBpedia's semantic web materials) were included. In addition, social media posts are especially challenging for the DBpedia Spotlight tool to process. The limited amount of context and use of colloquial language make the software especially prone to false positives (inaccurately interpreting certain irrelevant terms, usually nouns, as subjects in DBpedia) as well as false negatives (failing to recognize certain terms as significant; these can be either nonstandard references to DBpedia entities or references to entities that are not present in DBpedia). The full narrative analysis tools from the vaccination study, however, are specifically designed to address these concerns. For example, they are likely to recognize that the pejorative label "Shillary" actually refers to Hillary Clinton, and also are able to "learn" a wide range of conceptual relations. As a consequence, we defer these enhancements to further iterations of the project.

Bridgegate

Online news coverage of the "Bridgegate" lane-closure scandal – an outlandish conspiracy story that nevertheless was revealed to be true, largely through the efforts of investigative journalists – provides a useful contrast to the case of the Pizzagate hoax, and involves data sources that more closely resemble the large-scale web archives to be used for the full-scale fake news narrative analysis study in the future. The actual data sources for this pilot study were two semi-curated online archives of news coverage of the conspiracy and resulting scandal from mainstream online media sources: specifically, all articles on the *Huffington Post* new aggregator site tagged with the "bridgegate" keyword, and the archive of award-winning coverage contributed by reporters and editors at the *Record* newspaper in northern New Jersey [18] [19].

The texts of the news coverage were first obtained for analysis by running one-time crawls targeted at the root pages of the archives with the Archive-It web-crawling subscription service (see the references section for the specific URLs). In addition to the HTML of the crawled websites, made available in the customary aggregated WARC (web archive) format, specialists at the Internet Archive also provided WANE (web archive named entity) files from the nascent "Archive-It Research Services" feature, which contain person, place, and organization names automatically identified from the archived texts via NLP (natural language processing) techniques.

In retrospect, given the highly constrained nature of the target data set, it would have been much easier for us to crawl the web pages directly via a custom script. The "by-catch" of irrelevant articles and other web detritus returned by Archive-It, even when using fairly restrictive crawl "scope" settings, was enormous. In all, Archive-It saved 124,355 unique URLs linked from the two "seed" pages, of which 47,398 URLs were identified through various heuristics to be advertisements and other irrelevant content, and 73,544 of which were articles. Yet after much further filtering, it was determined that the actual target set of designated Bridgegate-related stories was a grand total of 415: 163 from the *Huffington Post* and 252 from the *Record*. Nevertheless, it is probable that analyzing the large-scale web archives of fake news currently being collected also will require this type of extensive content categorization and filtering, so developing such expertise likely was not wasted effort.

Another finding was that the contents of the Archive-It WANE files are somewhat excessively inclusive – at least, they required further post-processing to be useful for this study's purposes, even after being filtered to include entities extracted from only the 415 target pages. The WANE files contain each crawled page's detected person, place, or organization names, which have not been disambiguated or resolved to authoritative identifiers (though it seems that the Archive-It Research Services may provide this ability in the future). For this initial inquiry, it proved more efficient to disregard the WANE files and instead run the DBpedia Spotlight service to perform both named-entity recognition and authority resolution to entities from DBpedia on the text extracted from the page HTML, similar to the processing of the Pizzagate texts as described above. Illustrating the degree to which named-entity resolution reduces the total number of detected entities, the number of unique strings identified as likely person, place, and organization names for the 415 target pages as reported in the Archive-It WANE files was 3,375 names, while for the same texts, DBpedia Spotlight highlighted 954 unique identifiers for people, places, and organizations.

Figure 2 presents a network visualization of the entity co-occurrences in the Bridgegate data set, using similar techniques to those that produced Figure 1. The highly political nature of this scandal, and particularly its relevance to both regional and presidential politics, is immediately apparent from the people and organizations that figure prominently in the graph. The overall more "conventional" nature of this conspiracy narrative is evident in the lack of farfetched entity connections – other than those that result from false positives in the named entity resolution process – compared to the Pizzagate network. One important methodological difference in this plot relative to Figure 1 is that nodes and edges have been sized in proportion to their "betweenness" level, which, rather than merely counting connections, instead takes into account the extent to which a node or edge connects otherwise fragmented portions of the network. Visualizing betweenness serves both representational and interpretive purposes. It mitigates against the disproportionate amount of attention viewers often pay to the layout of the network, which is desirable because the multiple interrelations that network layouts attempt to represent via node proximity typically lead to placements that reveal little if any real information about the actual degree of association between most nodes. Using node and edge sizing to highlight the most prominent "bridge" nodes instead draws attention to this much more meaningful visual cue. In the case of the Bridgegate graph, the emphasis on betweenness reveals the degree to which New Jersey Governor Chris Christie was central to the conspiracy narrative as expressed in press coverage, despite the fact that he was never directly linked to the politically motivated bridge lane closures, which were determined to have been carried out by his lieutenants. Christie and his associates all have high "degree" ratings in the network due to their prominence in the story, but the betweenness-based visualizations shows that only Christie (in addition to the primary locations involved in the story) truly ties the entire network together.

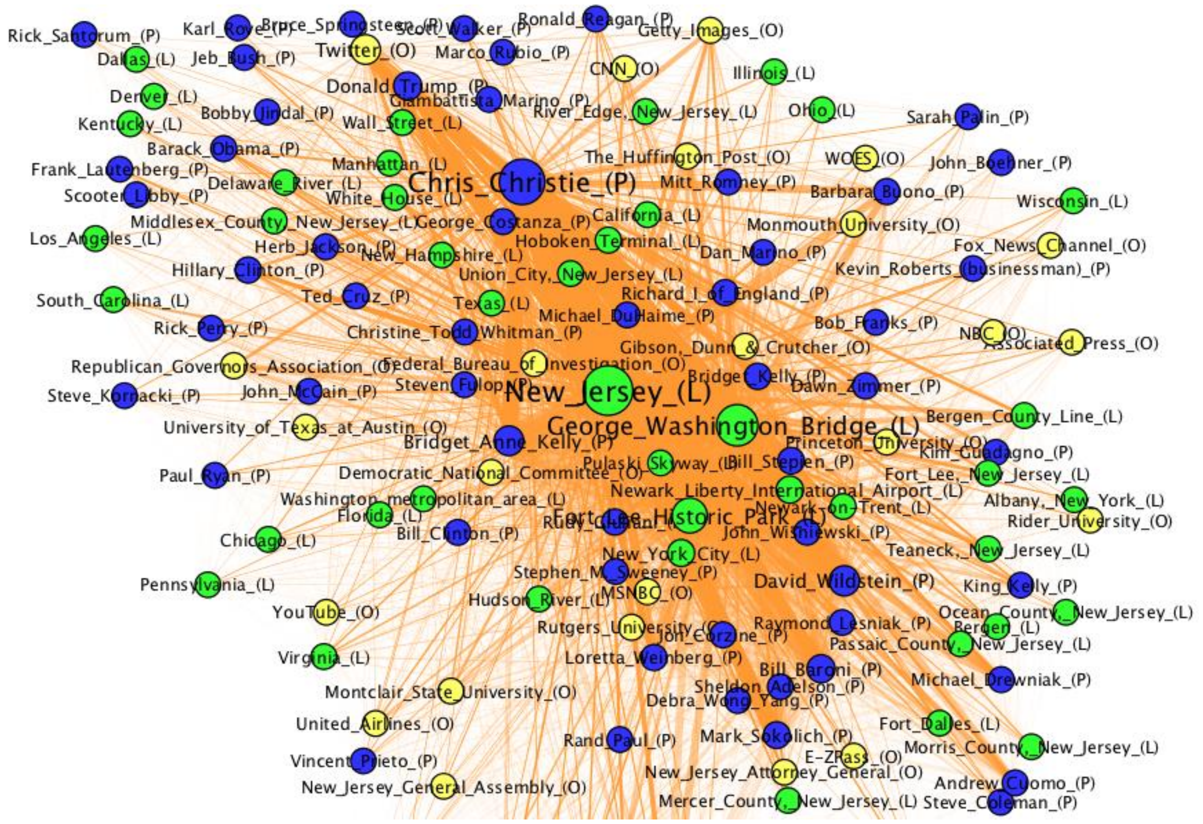

Figure 2: Network visualization of the most prominent people (P), locations (L) and organizations (O) detected in the Bridgegate data set by the DBpedia Spotlight named entity resolution tool. The number of connections between nodes is determined by how often they co-occur in a single news article. Nodes and edges are scaled in proportion to their "betweenness" – the degree to which they "bridge" otherwise disconnected regions of the network.

Final comments

This paper has presented archival efforts and pilot studies oriented towards the goal of conducting large-scale network-based narrative analyses of stories posted on "fake news" websites. Some primary considerations identified related to news media archiving include the challenges inherent in identifying target "fake news" web sites for bulk article collection, as well as the difficulty of maintaining web crawls of such sites over extended periods of time. Even assuming the web crawls are successful, it is important for researchers to realize that they will be working with (at best) a relatively large "bagged" region of the web that may contain an extremely high proportion of extraneous materials. Therefore, it is to be expected that considerable effort must be devoted to the identification and extraction of relevant content from web archives. Some suggestions for addressing these challenges include the development of an infrastructure for coordinating web crawls between institutions and the enhancement of research-oriented tools to make web archives more usable. The latter should include tools to facilitate filtering, versioning and deduplication of archival content, named entity detection, disambiguation of named entities, and incorporation of semantic linkages and external authorities into this augmented data. By providing these enhancements, archivists and librarians can play a more prominent role in the advancement of research to understand and counter the threat that fake news poses to journalistic reliability and public discourse.

Acknowledgments

I am especially grateful to UCLA Professors Vwani Roychowdhury and Timothy Tangherlini, and to the members of Prof. Roychowdhury's research group, particularly the Ph.D. students Ehsan Ebrahimzadeh, Behnam Shahbazi, and Misagh Falahi. The preliminary narrative analysis examples presented in the results section would not have been possible without their expertise, which included not only running the analysis but also helping to prepare the archived data sets for processing.

Thanks as well to Mark Graham and Karl Blumenthal at the Internet Archive for their helpful feedback and input regarding archiving and analysis of news web sites via the Archive-It service.

References

1 Jacob L. Nelson. 2017. "Is 'fake news' a fake problem?" *Columbia Journalism Review*, January 31, 2017 [http://www.cjr.org/analysis/fake-news-facebook-audience-drudge-breitbart-study.php].

2 TR Tangherlini, V Roychowdhury, B Glenn, CM Crespi, R Bandari, A Wadia, M Falahi, E Ebrahimzadeh, R Bastani. 2016. "'Mommy Blogs' and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites." *JMIR Public Health Surveillance* 2:2 (e166, 2016) [DOI: 10.2196/publichealth.6586].

3 Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, Filippo Menczer. 2016. "Hoaxy: A Platform for Tracking Online Misinformation." Submitted to Third Workshop on Social News On the Web, 4 Mar 2016 [arXiv:1603.01511].

4 Craig Silverman. 2016. "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook." *Buzzfeed News*, November 16, 2016 [https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.svwd1xGYj#.erYM3L7od].

5 Craig Silverman and Jeremy Singer-Vine. 2016. "The True Story Behind The Biggest Fake News Hit Of The Election." *Buzzfeed News*, December 16, 2016 [https://www.buzzfeed.com/craigsilverman/the-strangest-fake-news-empire?utm_term=.gh427bkoZ#.nk60wDR5P].

6 Emma Jane Kirby. 2016. "The city getting rich from fake news." *BBC News*, December 5, 2016 [http://www.bbc.com/news/magazine-38168281].

7 Craig Silverman. "Facebook Wants To Teach You How To Spot Fake News On Facebook." 2017. *Buzzfeed News*, April 6, 2017 [https://www.buzzfeed.com/craigsilverman/facebook-wants-to-teach-you-how-to-spot-fake-news-on?utm_term=.iuGwkr70R#.mjE1Py2DR].

8 Steve Annear. 2016. "A Mass. teacher made a list of 'false' or 'clickbaity' websites. You'll never guess what happened next." 2016. *The Boston Globe*, November 16, 2016 [https://www.bostonglobe.com/metro/2016/11/16/mass-teacher-made-list-false-clickbaity-websites-you-never-guess-what-happened-next/5XCP5T4I4wJz2D4sBd2deI/story.html].

9 Kalev Leetaru. 2017. "Why Are Libraries Failing At Web Archiving And Are We Losing Our Digital History?" *Forbes*, March 27, 2017 [https://www.forbes.com/sites/kalevleetaru/2017/03/27/why-are-libraries-failing-at-web-archiving-and-are-we-losing-our-digital-history/#709dfb26ecd4].

10 Stephen Abrams, Andrea Goethals, Martin Klein, Rosalie Lack. 2016. “Cobweb: A Collaborative Collection Development Platform for Web Archiving.” *Research Ideas and Outcomes* 2:e8760 (April 12, 2016) [DOI: 10.3897/rio.2.e8760].

11 https://archive-it.org/collections/8142.

12 Abby Ohlheiser. 2016. “Fearing yet another witch hunt, Reddit bans ‘Pizzagate.’” *The Washington Post*, November 24, 2016 [https://www.washingtonpost.com/news/the-intersect/wp/2016/11/23/fearing-yet-another-witch-hunt-reddit-bans-pizzagate/?utm_term=.61e1c7c53619].

13 “GitHub censored pizzagate repo!!!” [https://voat.co/v/pizzagate/1451672].

14 “The #Pizzagate Wiki.” [http://pizzagate.wiki/Main_Page].

15 Joachim Daiber, Max Jakob, Chris Hokamp, Pablo N. Mendes. 2013. “Improving Efficiency and Accuracy in Multilingual Entity Extraction.” *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*. Graz, Austria, September 4-6, 2013 [http://jodaiber.de/doc/entity.pdf].

16 P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, T Ideker. 2003. “Cytoscape: a software environment for integrated models of biomolecular interaction networks.” *Genome Research* 13:11, pp. 2498-2504 (November 2013).

17 Gregor Aisch, Jon Huang, Cecilia Kang. 2016. “Dissecting the #PizzaGate Conspiracy Theories.” *New York Times*, December 10, 2016 [https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html].

18 http://www.huffingtonpost.com/news/bridgegate/.

19 “The Record’s Bridgegate coverage – from start to finish.” [http://archive.northjersey.com/news/chris-christie-and-the-george-washington-bridge-scandal-on-northjersey-com-1.737481].