A Systematic Approach to Analyzing Syndication of Viral News Stories

17-631 Information Security, Privacy & Policy

Gaurav Deshpande
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
`gdeshpan@andrew.cmu.edu`

Sam E. Teplov
Institute for Software Research
Carnegie Mellon University
Pittsburgh, PA 15213
`steplov@andrew.cmu.edu`

Alon Peer
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
`apeer@andrew.cmu.edu`

Dr. Nicolas Christin
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
`nicolasc@andrew.cmu.edu`

December 7, 2019

paper/img/university.png

Abstract

In this work, we present a novel, semiautonomous methodology for analyzing syndicated news articles. We first detail our manual process of news syndication analysis from which we observe a number of unsettling patterns. From there, we go on to categorize and define two different forms of syndication: top-down and bottom-up. We then transform our manual process into a semiautomated one which combines web scraping and machine learning techniques to create and then analyze a data set of viral news stories. From the analysis, we are able to make some inferences about the nature of syndicated news articles, as well as trace back a viral news story to its original publisher and article. By presenting this work, we hope to minimize the effects that blind news syndication has on the spread of misinformation.

***K*eywords** Fake news · News propagation · News syndication

Contents

1 Introduction

News are everywhere- television, newspapers, social media. We consume it whether we like it or not. But the question is, how do we know which news are true, and which are false?

Fake news has been a real issue since the spark of social media, especially in the recent years, with its spike coming during the 2016 U.S. presidential election. In fact, a study from Ohio State University shows that fake news had a significant impact on the results of the elections. The researchers found a correlation between the amount of Obama voters who believed fake news about Clinton and the amount of them who chose to vote for Trump. At that time, the media was swarmed with fake news, and one could hardly blame people who couldn't distinguish them from real news.

But what exactly is fake news? According to Wikipedia, "fake news is a form of news consisting of deliberate disinformation or hoaxes spread via traditional news media (print and broadcast) or online social media". In addition to this definition, fake news may also include clickbait stories, propaganda, satire, biased news, or even sometimes just bad journalism.

These days, anyone can publish content on the internet, whether it is on a website, a blog, or a social media profile, and potentially reach large audiences. The fact that people nowadays rely on the internet to consume news doesn't make this problem any easier to solve. Fake news is very profitable for the publisher, as viral stories generate large amounts of traffic, thus increasing advertisement revenue. For that reason, fake news are widely common around the internet. Another motive to publish fake news is to promote a political agenda. Fake political news stories are meant to persuade consumers to accept a biased of false beliefs (like the fake news regarding Hillary Clinton during the 2016 U.S. presidential election). This is the main reason detecting fake news is such a challenge. These false stories were created to sound true and interesting.

A different variation of fake news is syndication. A website can take a real news story, and change it a bit, so it will fit that website's beliefs. A news story can be syndicated in various ways- changing the title, adding interpretations that aren't necessarily true, and much more. A popular form of syndication is news stories that are posted by local newspapers and websites, and as they get viral, these stories get picked up by bigger and more popular news websites. On the way from the local news source to the national (or international) news website, the story can get syndicated. Sometimes, it's not necessarily on purpose, but it happens due to the lack of due diligence by the mainstream media.

One can try and detect fake news by checking the source of the story, read beyond the title (which is intended to attract users and thus contains more interesting and controversial topics), try to trace the article back to the original source, or even make sure the article is not a satire by any chance. Even though it's possible, doing so is not an easy task, and that is why there has been a lot of work put into automating this process. A step towards this, is detecting and analyzing the syndication of viral news stories. That is, how stories change through time and the difference between articles regrading the same stories but posted in different websites.

The thing is, that identifying news propagation path is tedious, too. To identify the propagation path of a given article, one would have to dive deep into the results of Google search, manually identify the original source of the story, and look through many news websites and see the differences between the ways each website covers the same story. That's why we decided to try and automate this process using machine learning techniques and data analysis.

2 Background

The field of misinformation and fake news has become widely researched by academics ever since the Russian misinformation campaign against the 2016 United States presidential election came to light[1]. The most straightforward approach to detecting fake news is applying various machine learning techniques, and this is what many researchers have focused on thus far.

2.1 Fake News Detection

Shu et al. goes through a comprehensive overview of all of the various ways of attacking the problem of fake news detection and classification from a data mining perspective [2]. From the perspective of detecting fake news via machine learning techniques, Shu et al. outlines that there are generally four different types of features that can be used to detect fake news: news content features, linguistic-based features, visual-based features, and social context features. These features are then applied to one of two different types of models: news content models, or social context models . Just as they sound, news content models focus on fact checking claims in the news article or looking at stylistic features that may indicate that the article is fake. Social context models, on the other hand, focus on how users interact with the article, the virality of the article, and the propagation path of the news story. These two different methods are rarely used independently, but rather compliment each other in many hybrid approaches, such as ensemble models [2].

2.2 Open Issues in Fake News Detection

Shu et al. closes by giving an overview of many of the various open areas of research in this field, as well as proposing some future research ideas. Shu et al. outlines four future research directions: data-oriented, feature-oriented, model-oriented, and application-oriented. Within the area of data-oriented research, Shu et al. brings forward a couple of issues that need to be addressed. For one, there is no large, encompassing fake news data set which can be used as a benchmark to facilitate future research [2]. In fact, this was such a large problem that in 2018, Shu et al. published another work detailing their methodology for building a large scale, open source, data set consisting of fake news, which is now available on Github [3]. Another idea that Shu et al. proposes for future research within the data-oriented direction is early fake news detection [2]. Many of the current state-of-the-art fake news detection methods rely on data that needs to be accumulated over time and information that is not readily available when news is just starting to go viral [4]. Shu et al. suggests creating a system that provides fake news alerts during the dissemination process so that misinformation can be flagged before it is viewed by a large audience. Another suggestion that is made is to look at social media posts made within some time delay of the original posts as a means of verifying the veracity of the post [2].

3 Related Work

There has been work done on early fake news detection by analyzing propagation paths and time series, as well as analyzing the dynamics of news syndication. Most of this work, however, has been limited in scope in terms of the domains that it covers, as well as the type of media being analyzed. For example, past research has focused on a either a few social media platforms, a specific media outlet, or a single country [4],[5], [6]. The research that has spanned across multiple different domains of news (social media, various news outlets, etc.) has been limited in scope when considering the type of media that is being analyzed [7]. There has been little work, if any, done specifically on the syndication of viral news and trying to automatically analyze and track back a viral news story to its original publisher.

3.1 Domain-Specific Approaches

Liu et al. propose a method of early fake news detection by modeling the propagation paths as a multivariate time series. They build a classifier that uses recurrent and convolutional networks to detect fake news. They were able to achieve an accuracy of 85% on Twitter and an accuracy of 92% on Sina Weibo when it came to correctly classifying fake news. Liu et al. claims that they were able to detect the misinformation five minutes after it began to spread [4]. The downside to this approach is that it relies on user interaction with the article to generate the data needed to feed into their classifier. This means that some users will be exposed to the fake news before it can be flagged as misinformation just due to the nature of the data set that they are using. Also, this approach has only been tested on social media platforms and micro blogs, meaning that this methodology cannot be applied on a global scale that spans hundreds of different news outlets and social media sites.

Jin et al., in their work, tackle the problem of misinformation on the popular website Micro blog. They propose a three-layer hierarchical model that establishes a credibility score for each post and then propagates this score throughout the network. They then formulate this propagation process as a graph optimization problem and find a globally optimal solution. Jin et al. were able to boost accuracy by 6% over the baseline model [5]. Again, just like Liu et al.'s approach, Jin et al. focus solely on one website: Micro blog. Their model is reliant on data to be taken specifically from Micro blog. Although their technique could in theory be applied to other platforms, their approach is mostly domain-specific and does not analyze misinformation on a global scale.

Wang et al. explore and analyze the patterns of news propagation and syndication in Chinese news media. They draw comparisons between the way news diffuses through time and space and how an epidemic spreads. Something interesting that they found was that 80% of news outlets were responsible for re-printing news articles directly from the source [6]. Some of our findings also substantiate this claim. Like much of the other research on this topic, Wang et al. focused on only Chinese media outlets and they also did not attempt to trace back an article back to its original source without knowing what that source was in the first place. Although Wang et al.'s research brings to light some interesting patterns in the ways that news article propagate and are syndicated, the work is confined mostly to Chinese news media, which is known to be heavily regulated by the Chinese Communist Party [8].

3.2 Media-Specific Approaches

Zannettou et al. take a different approach to analyzing propagation; instead of looking at social media posts or news stories, they focus on politically motivated memes. Unlike any of the previous research talked about, Zannettou et al. pool data from a number of various sources: Twitter, Reddit, 4chan, and Gab. Using memes from these sites, they are able to trace the propagation path of each meme and even draw conclusions about the influence that each meme outlet has. Zannettou et al. use Hawkes process to model how memes from various sources interact with each other and quantify the influence that each meme has on others [7]. Although this research encompasses many different media outlets and social media platforms, it fails to analyze any other form of media besides memes. Some of the processes outlined in this work could be applied to news articles, but much of it is very specific to memes and images and cannot be generalized to news articles due to the difference in data between an image and a set of text. Nevertheless, Zannettou et al. showed that memes are often syndicated by larger media outlets which causes fringe web communities to have a much broader reach than they ever could have before [7].

4 Phase I - Manual Approach

Our research and methodology can be split up into a manual process (phase I), and a semiautonomous process (phase II). In the manual approach, we created a data set of popular news articles and then attempted to trace them back to the original publisher. Once we identified the propagation path, we attempted to analyze how various article features changed over time.

4.1 Dataset Creation

We started by collecting a diverse range of articles from a number of different news sources. As seen in Appendix A, we selected five articles from each news source and recorded each articles' associated comment count. These articles were picked out between October 5 2019 and October 8 2019. We looked for articles that were published in the past two weeks. We picked media outlets that tended to be biased either towards the political left or right as defined by the chart in Appendix B. The reason we did not pick any mainstream, unbiased news outlets (New York Times, Washington Post, Economist, etc.) is because these news sources typically do their due diligence in ensuring that the news that they are publishing is true, and the public typically trusts these sources [9]. It is also worth noting that we did not consider any articles that were about, or mentioned, President Donald Trump, as these could possibly skew results. Aside from that, we did not consider the actual content of the article when picking articles for the initial dataset.

4.2 Case Study: Chinese Gold

One of the first articles that stuck out to us from the initial dataset was the article from the Daily Mail titled *Thirteen and a half tonnes of gold worth up to £520million is found in a corrupt Chinese official's home and £30BILLION in suspected bribe money in his bank account*.

4.2.1 Propagation Analysis

The reason this article jumped out at us is because the number of comments for this article was much higher than any article in the dataset. Once we identified this specific story, we decided to try to trace it back to its original source. The process of trying to trace an article back to its origin is an extremely tedious and time consuming task. This is because there is no systematic approach already out there that can help us gather the information that we needed. The data used to build the propagation path for the Chinese Gold Story can be found in Appendix C

As it turned out, the original content for the news story was actually a twitter post. What was fascinating was that none of the articles actually attributed that Twitter was the first platform where the content for the story appeared; many of the news outlets did link to the Twitter post, but not once did any of the news sources explicitly say "this content was originally posted on Twitter by a user." It is also important to notice the time delays between when the story was first posted on Twitter, and when it was picked up by mainstream media.

4.2.2 Article Feature Analysis

Besides just looking at how the article traveled through space and time as it was syndicated, we were interested in looking at how the articles' properties changed over time.

Figure 1: First Article About Chinese Gold That We Found

4.3 Findings

We found that

4.3.1 Different Types of Syndication

Top-down and bottom-up

5 Phase II - Semiautonomous Approach

In the semiautomated approach, we wanted to see if our findings from phase I would still hold true against a larger data set. We automated the process of tracing the propagation path and conducting analysis on the resulting articles. The resulting methodology from phase II is the main contribution of this work.

The approach can be broadly categorized into two stages.

1. *Dataset Generation*: As a part of this data generation we identify the viral news articles, scrape for articles that are similar to it and gather metadata on these set articles. The raw data is passed through machine learning models to group them as relevant articles and discard non relevant articles.
2. *Analysis*: Numerous data visualization methods were applied based on the findings of phase I. These included time series analysis and histograms on certain features.

5.1 Dataset Creation

5.1.1 Identification of Viral News Article

Like in phase I, we did not use articles published by unbiased media sources and steered clear of stories related to Donald Trump. An article with more than *ten thousand reactions*[1] on facebook was quantified as a viral news article.

[1] reaction is the sum of likes, comments and shares

`paper/img/chinese-gold-timeline.png`

Figure 2: Timeline of Chinese Gold Story

The articles selected in the end were a heterogenous mixture of local and international events. These new articles were used to manually build*url.csv* dataset. This dataset consists of a single feature, namely the url pointing to the article.

5.1.2 Scraping similar news articles

5.1.3 Building Dataset

Add code snippet here

Figure 3: Article Length Over Time

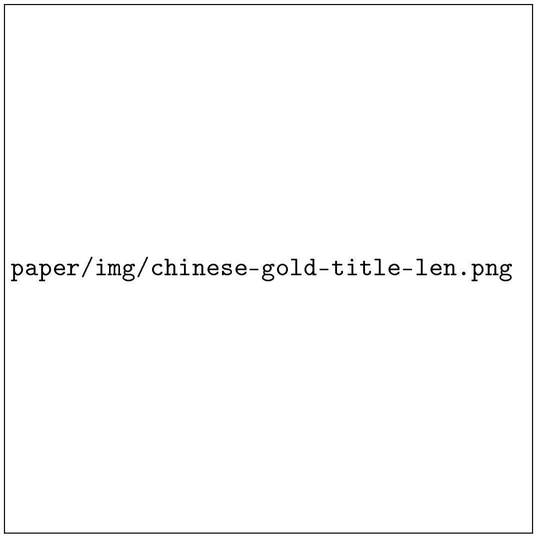

Figure 4: Title Length Over Time

5.2 Data Clustering

5.2.1 Similarity Score

Stemming

TF-IDF

Cosine Distance

5.2.2 Hierarchical Clustering

5.3 Data Analysis

5.4 Case Study : The Arkansas Story

5.4.1 Overview

5.4.2 Findings

6 Future Work

7 Discussion and Conclusion

References

[1] Constanze Stelzenmüller. The impact of russian interference on germany's 2017 elections. *Testimony before the US Senate Select Committee on Intelligence June*, 28, 2017.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[3] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[4] Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining*, pages 230–239. IEEE, 2014.

[6] Youzhong Wang, Daniel Zeng, Xiaolong Zheng, and Feiyue Wang. Propagation of online news: dynamic patterns. In *2009 IEEE International Conference on Intelligence and Security Informatics*, pages 257–259. IEEE, 2009.

[7] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202. ACM, 2018.

[8] Qiuqing Tai. China's media censorship: A dynamic and diversified regime. *Journal of East Asian Studies*, 14(2):185–210, 2014.

[9] Michael Kearney. Trusting news project report 2017. *Reynolds Journalism Institute*, 2017.

A Initial Dataset Used for Phase I Analysis

Media Outlet	Article	Comments
Patribotics	Dear Mr. Putin, Let's Play Chess	148
Patribotics	Wikileaks is Connected to Russia – Despite Their Claims	42
Patribotics	Planespotting: Michael Cohen's Amazing Journey	49
Patribotics	Putin's Hacker, Wikileaks Host Pyotr Chayanov, Hacked America's Vote System And the DNC	11
Patribotics	Wikileaks Hands "Keys" to Putin's Russian Hacker – Readers, Leakers Tracked	19
InfoWars	De-Dollarization: Europe Joins The Party	6
InfoWars	Unemployment Falls to Lowest Level Since 1969	13
InfoWars	China Unveils 'doomsday Bomb' While U.s. Military Concentrates on "diversity"	83
InfoWars	'no Precedent in Human Experience': Study Finds Nuclear War Between India and Pakistan Could Leave 125 Million Dead	6
InfoWars	China Reveals New Photos of Strange Substance From Dark Side of Moon	26
Daily Caller	Iranian Foreign Minister Uses Instagram To Resign	22
Daily Caller	Israel Holding Early Elections As Bribery Allegations Engulf Netanyahu	12
Daily Caller	Turkey's President Arrests More Than 100 People For Connections To Failed 2016 Coup	4
Daily Caller	Saudi Crown Prince Fires Entertainment Chief Because Of Tightly Clad Female Circus Performers	4
Daily Caller	Hong Kong Police Unload Live Rounds On Protesters, Shoot 18-Year-Old: Report	23
Daily KOS	How to Support the Hong Kong Protesters	2
Daily KOS	Who knew? Ukraine-gate is actually a Rick Perry crime spree!	76
Daily KOS	NYT: Second Ukraine-related whistleblower may soon come forward	96
Daily KOS	Have we learned nothing about wars in the Middle East?	119
Daily KOS	Open thread for night owls: 'Itching for a War' with Iran	93
Daily Mail	Russia is helping China build a new missile attack warning system in 'response' to US plans to deploy missiles in Asia	44
Daily Mail	British sausage makers claim nation's bangers are under threat from pork shortage in China that has seen prices rise by 45 per cent	184
Daily Mail	Thousands of pro-democracy activists rally in Hong Kong ahead of four days of protests to overshadow anniversary celebrations in Beijing	0
Daily Mail	Thirteen and a half tonnes of gold worth up to £520million is found in a corrupt Chinese official's home and £30BILLION in suspected bribe money in his bank account	451
Daily Mail	'Most-wanted' Chinese fugitive, 63, hides in a cliff-side cave for 17 YEARS after escaping from prison	6
The Washington Times	Man pulls gun in road rage incident over Elizabeth Warren sticker, police say	2
The Washington Times	Man pulls gun in road rage incident over Elizabeth Warren sticker, police say	5
The Washington Times	Man pulls gun in road rage incident over Elizabeth Warren sticker, police say	124
The Washington Times	FBI runs Russian-language Facebook ads asking for help neutralizing 'hostile foreign intelligence'	1
The Washington Times	FBI runs Russian-language Facebook ads asking for help neutralizing 'hostile foreign intelligence'	0
Mother Jones	It's No Coincidence That the Top Presidential Candidates Are All So Old	6
Mother Jones	Columnist at the Center of Ukraine Scandal Joins Fox News	5
Mother Jones	Microsoft Says Iranian Hackers Are Targeting a 2020 Presidential Campaign	9

Media Outlet	Article	Comments
Mother Jones	Researchers Assembled over 100 Voting Machines. Hackers Broke Into Every Single One.	7
Mother Jones	The Biden Campaign Is Demanding That TV Execs Stop Booking Guiliani	69
Reason	Supreme Court Will Finally Hear Arguments Over Federal LGBT Discrimination Protections	100
Reason	China Banned South Park After the Show Made Fun of Chinese Censorship	105
Reason	The NBA Cares More About Making Money in Mainland China Than Supporting Freedom in Hong Kong	94
Reason	The U.K. Must Ban Pointy Knives, Says Church of England	76
Reason	The New York Times Says ’Free Speech Is Killing Us.’ But Violent Crime Is Lower Than Ever.	120

B Media Bias Diagram

paper/img/media-bias.jpg

Media Bias Diagram

C Chinese Gold Story Dataset

Media Outlet	Article Title	Date of Publication
Twitter	N/A (Twitter post by user @h1300062810)	. 09-24-2019 10:09
PowerApple	Secretary of Haikou copied 13.5 tons of cash, booked 268 billion in gold (translated)	09-26-2019 08:08
CrimeRussia	Chinese official hides 13 tons of gold in basement	09-26-2019 09:50
MenaFN	13.5 Tons Of Gold Found In Chinese Ex Mayors Basement	09-26-2019 18:20
Novinite	13 Tonnes of Gold Found in the Basement of Former Chinese Mayor	09-27-2019 13:49
RT	13.5 TONS of gold found piled in Chinese ex-governor's home	10-01-2019 13:37
Daily Star	13.5 tons of gold and $37billion cash found during police raid on mayor in China	10-01-2019 18:21
Daily Mail	Thirteen and a half tonnes of gold worth up to £520million is found in a corrupt Chinese official's home and £30BILLION in suspected bribe money in his bank account	10-02-2019 04:54
Mirror	Corrupt Chinese official found with £520million worth of gold bullion in home	10-03-2019 02:01