Characterization and Time-Series Prediction of Alternative News Propagation

Gupta, Shrey
shreyg19

Ray, Nathaniel (Scott)
nsray

Wang, Haiyin
haiyinw

December 11, 2017

1 Introduction

News exposure and opinion formation increasingly occurs through social media channels, and more specifically, alternative news has been enjoying increased popularity and readership. Alternative news sources tend to focus more on sensationalized and highly opinionated content, often with a lower regard for journalistic standards and unbiased analysis/reporting. While alternative news itself is not a new phenomenon, social networks provide a channel where information diffusion occurs rapidly, where news is difficult to tie to a source, and where the required level of journalistic quality is falling [9]. Moreover, the spread of alternative news has also been tied to the advancing of malicious agendas such as the spread of misleading of political information.

The past few years have seen a sharp increase in the prevalence of alternative news [1]. Coupled with the public's demonstrated inability to distinguish alternative news with its generally more credible mainstream counterpart, alternative news has become deeply entrenched in major social media networks. As these alternative news sources become confounded with accurate ones, false news begins to spread throughout many social networks.

Considering current online social networks, these networks contain high degrees of homophily for users, content on the platform is more readily consumed due to the relative trust users have in their network. However as the platform grows in relevance, the prevalence of alternative news grows alongside it. Users can quickly find themselves engaging with false information that is relative difficult to distinguish. A 2016 analysis determined that 62 percent of US adults obtained their news on social media and the most popular alternative news stories were more widely shared than the most popular mainstream news stories [4]. Motivated by these findings, we pursue the analysis of graph characteristics of the combined social networks and news sources graph to characterize and predict the spread of both mainstream and alternative news between social networks over time.

In our analysis we will begin by reviewing three related works that inspired the methods used in our work, then we will elaborate on the nature of the data and the graph models we employ, then we will dive into a detailed description of the implementation and results of our algorithms, and finally we will end with a brief discussion of possible future paths for our research.

2 Literature Review

2.1 The Link Prediction Problem for Social Networks

Motivated by a goal to understand the dynamic behavior of social networks, Kleinberg & Liben-Nowell investigate prediction models for link formation between nodes in the snapshot of a social network [7]. The experiments use scientific co-authorship networks over a span of time where link prediction models attempt to predict co-authorship endeavors in a future time period based off the initial graph structure of the data at the beginning time period. They implement a range of scoring algorithms to ascribe predictions of new edges created in the graph and also use graph transformations that reduce "noise" and create simplified structural representations to operate over. The results of the experiment indicate while there was no clear superior technique among the methods they attempted, the algorithms all out-

performed a random prediction and demonstrated the relevance of information within the network topology alone. Furthermore, the results indicated as the social network dataset grew to be more diverse, the predictors were more accurate compared to networks of similar social nodes where there was more random link formation. Overall the research suggests that link prediction can be partially attributed to the topology of social networks.

2.2 Everyone's an Influencer: Quantifying Influence on Twitter

The research analyzes the diffusion of information within the social network of Twitter to identify, characterize, and predict influencers within the network [2]. In the experiment, the design utilizes sharing of URLs through the Twitter network to form disjoint cascading trees of link sharing between nodes used to quantify an influence score per member of the network. The research then continues to implement a prediction model using a regression tree with user attributes around local network size such as number of followers to predict the influence score a user would have received from the URL sharing experiment. From the experiment, the authors found that past performance and a node's immediate network size (followers on Twitter) were the primary two factors predicting future influence. While this conclusion is intuitive, it also demonstrated that the number of users one followed on the network and the activity of a user in the network were surprisingly not important predictors of influence. Overall, the research details a successful approach to capturing the diffusion of information through Twitter and suggests the ability to predict influential nodes based on their social network attributes.

2.3 The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources

In their paper, Zannettou et al. attempt to offer the first network-focused analysis on the propagation of mainstream alternative news across various social networks [8]. They specifically analyze three key dimensions of influential social networks including Twitter, the /pol/ board on 4chan, and 6 news and current affairs oriented subreddits on Reddit. The first dimension involves generally characterizing the nature of alternative news content on each of the platforms by tracing posts and comments with URLs to popular mainstream and alternative news sites. Key results here included that various news sites (alternative and mainstream) source their traffic equally from the 3 platforms and that Twitter has the highest alternative to mainstream news ratio. The second dimension of analysis involved temporal dynamics: determining the rate at which alternative news travels and the order in which it propagates to each platform. Zannettou et al. found that while Twitter users were most aggressive in propagating alternative news on the platform (Twitter has the least average time elapsed between initial occurrence of an article URL and subsequent reposts), much of the media had been present on Reddit and 4chan first. Finally, Zannettou et al. performed a third dimension of analysis, influence estimation. Through this, they aimed to determine how much influence each platform had on the media shared on the other platforms by modeling influence over time via Hawkes Processes. The key insight from this influence estimation was that Twitter derives roughly 6% of its alternative news media from just `/pol/` on 4chan and `/The_Donald/` on Reddit combined. Overall, Zannettou et al. provide a cohesive model for understanding inter-platform alternative news media propagation, with 4chan and Reddit as early incubators and Twitter as an aggressive propagator/popularizer of such content.

2.4 Critique

From the reviewed literature, we formulated the research question: can we predict the diffusion of mainstream and alternative news through social networks using its structural properties? Motivated by the approach within Zannettou et al. which analyzes the spread of fake news between social networks, we extend the approach to incorporate predictions also based on the link prediction algorithms elaborated in Kleinberg and Liben-Nowell. The motivation being that the news source and social network graphs have distinct structural characteristics that are important in accurately predicting the spread of news articles into social networks. We use the influence characterization model discussed in Zannettou et al. to characterize each social network's influence in our data and then utilize link prediction models discussed in Liben-Nowell to predict new news articles being posted within each social network. Through this process we aim to more accurately simulate the spread of mainstream and alternative news between different social

networks. To accomplish this, we explore two primary algorithms touched on by Liben-Nowell on the bipartite graph of social networks composed of one group and news sources in the other group. Through these link prediction algorithms we aim to predict the propagation of alternative and mainstream news sources' articles between social networks.

3 Methods

We begin with an analysis of the Reddit dataset, then proceed to concretely define the prediction problem using a graph model and discuss the algorithms used to approach the problem.

3.1 Data

For our social networks we utilized data from Reddit where each subreddit was considered as a distinct social network. Reddit submission data for the years 2008 to 2017 (up to 2017-03) was retrieved from publicly available data collected by Reddit user Stuck_in_the_Matrix [3]. The submission data includes timestamps, so it can be treated as a time series for the purpose of link prediction. To identify alternative and mainstream news, we matched URLs found in Reddit submissions to the same list of labeled domains used by Zannettou et al. We found 924,367 URLs with relevant domains in total, including 37,147 alternative URLs and 887,220 mainstream URLs. We generated a bipartite network structure where nodes belong to the groups of either subreddits (social network) or news domains, and for a weighted edge between a subreddit and a news domain, the weight of the edge is the number of submissions of a URL from the domain to the subreddit. We chose to utilize this model as its changing states clearly parallel the spread of news by its publishing source across our social networks. Due to the immense amount of data we had access to we opted to focus on understanding the spread of news from the root domain sources which was a more constrained set of nodes compared to other attributes in our data which would have grown exponentially in size with our dataset.

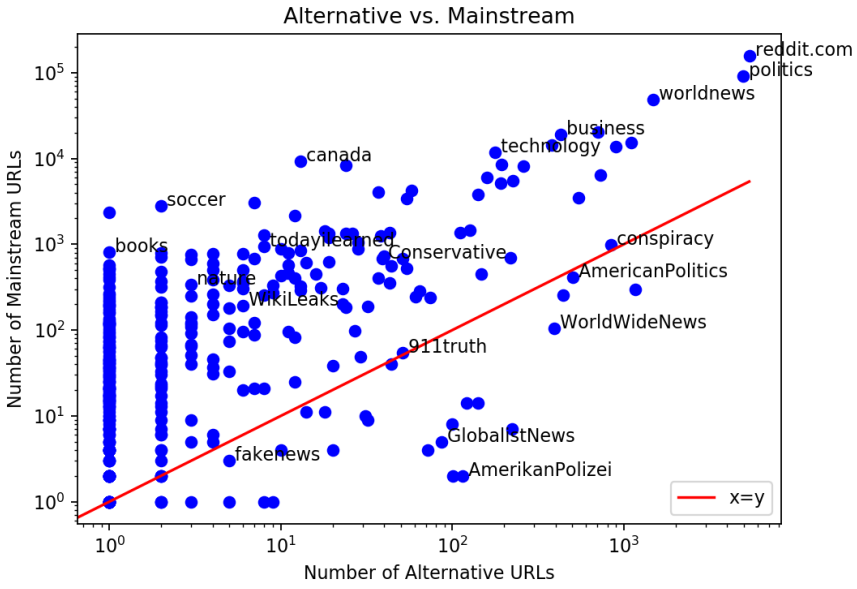

Figure 1: A loglog scale scatter plot of URL frequencies. Each data point represents a different subreddit, and points of interest are annotated with their subreddit names.

We can confirm from Figure 1 that the majority of Reddit communities get most of their news from mainstream news sources, with few exceptions. We can also identify, by measuring the vertical distance between a point and the identity line, some subreddits where alternative news is posted much more fre-

quently than mainstream news, such as the "GlobalistNews" subreddit, and we can also see some subreddits in which alternative URLs are posted very rarely compared to mainstream URLs, like the "canada" subreddit. Notably, in some conspiracy-themed communities like "911truth" and "conspiracy", mainstream sources are posted as heavily as alternative sources. The trend in the data is roughly linear, suggesting that the share of alternative news domains remains roughly constant as total URL activity increases. From this, it is clear that a key challenge of this problem will be designing link prediction models that accommodate a range of mainstream and alternative submission proportions across our data's subreddits.

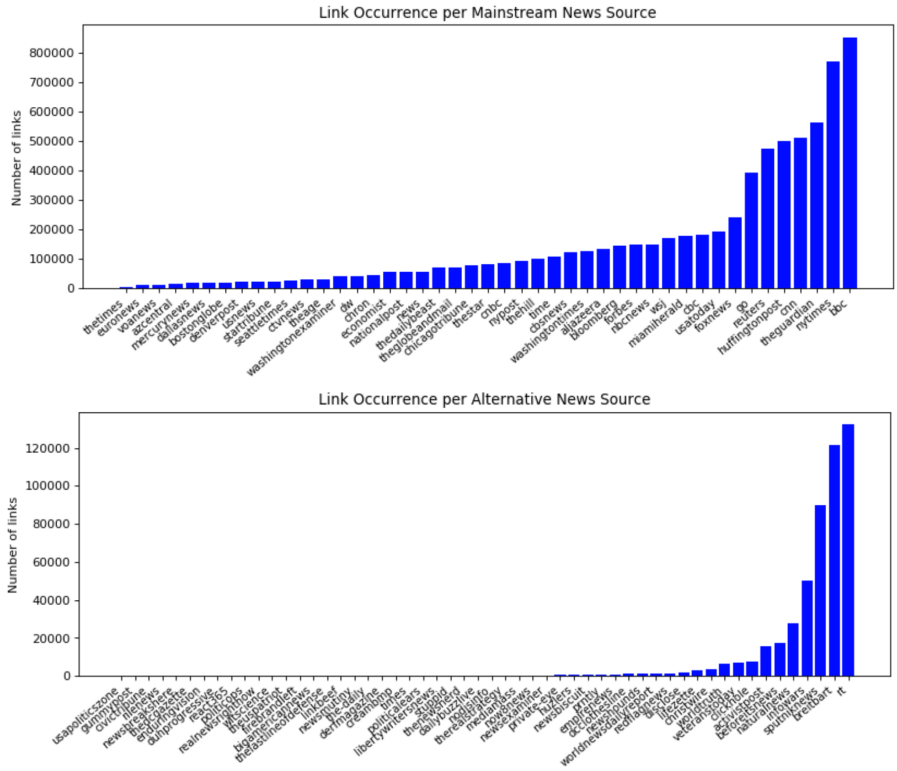

Figure 2: Link occurrence histograms for mainstream and alternative news. These histograms suggest that both mainstream and alternative news have large disparities in terms of amount of links shared per source, but this disparity appears to be much larger for alternative news.

Figure 2 shows a distribution of the amount of times a given news source appears in the overall data, from 2008 to 2017. We can see that both distributions obey a power law, and that the number of subreddits with a given number of alternative URLs is roughly proportional to the number of subreddits with that number of mainstream URLs. This trend is likely due to the fact that the share of alternative news domains remains fairly stable while the number of total URLs increases. From this analysis, it is unlikely that the share of alternative domains alone is enough to predict the future sharing of fake news within a community. Therefore we predict that selecting link prediction algorithms that draw more influence from the overall graph topology than the specific type of news source type they are connected to will be more successful.

Finally, we note that our dataset contains much more news link data for later years than previous ones, as shown in Figure 3. Considering how news sharing on social network platforms has recently skyrocketed, this trend is expected. However, this also implies that there may be some discrepancies when attempting to use prior time data to predict news flow for more recent years, as the underlying patterns and mechanisms may have themselves changed.

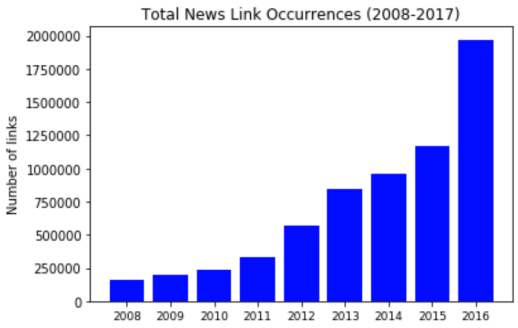

Figure 3: News Link Occurrences per Year.

3.2 Model and Problem Definition

First we explored our data to decide what graph structure and link prediction algorithms to implement. From our data classification step we were able to decide on the bipartite graph structure where nodes are either a news source domain or a subreddit social network and edges represent the number of articles from a news source on a subreddit indicated by the weight of the edge.

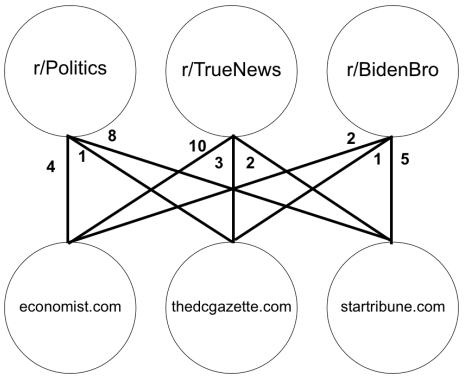

Figure 4: Subsection example of our graph model

With this graph state in mind and the goal of predicting the spreads of mainstream and alternative news throughout social networks, we chose to utilize the graph state consisting of Reddit data between 2008-2017 to create our initial time state bipartite graphs. For each year from 2009-17, we predict the new links generated during that year by running the link prediction algorithms described in the next section over a network generated from the data from all previous years. It is important to note that our model utilizes weighted edges but in the algorithms we implemented it is equivalent to consider our graph as a multigraph with a number of edges between two nodes equal to its weight. The output of the prediction can be interpreted as either a set of increments in the weights of edges of a weighted network or a set of new links in a multigraph.

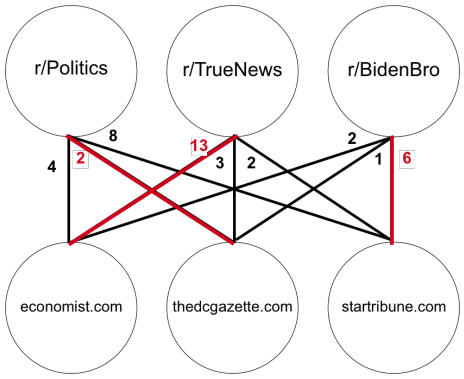

Figure 5: Red edges represent new news articles to predict

3.3 Algorithms

In the classical link prediction setting presented in [1], two networks are constructed: the first network is the train network and represents events in the time interval $[t_0, t_k]$; and the second network is the test network and represents events in the time interval $[t_{k+1}, t_n]$. A similarity score $score(x, y)$ is computed for each pair of nodes (x, y) in the train graph that are also present in the test graph - this set of nodes is called the "core set" - and to predict n new links, the n pairs with the highest scores are chosen. These similarity scores are often defined using graph topology and characteristics. Finally, the new links are added to the train network, and the result network is compared with the test network to judge performance.

However, this method of translating scores into predictions does not suit our prediction problem, because it predicts at most one new link will be added between any pair of nodes in the core set. In our data, it is common for weights to be much larger in the test network than in the train network. To predict n links, instead of taking the top n pairs, we let each pair of nodes increase in weight proportionally to its similarly score. That is, with respect to the core set C, if a pair of nodes $x, y \in C$ was originally linked by an edge of weight w, then its predicted weight is:

$$w + \frac{n * score(x, y)}{\sum_{i,j \in C} score(i, j)}$$

For the scoring metric, we observe that typical scoring

metrics used for link-prediction, such as those using Jaccard's coefficient or Adamic measures, will perform extremely poorly, given our unique weighted bipartite graph structure and the characteristics of our Reddit submission data set. This is because such algorithms are designed for unipartite graphs and often assume triadic closure and high amounts of clustering of nodes in the graph. However, in a bipartite graph, an edge can never form within a given disjoint set, rendering the intuitions behind these approaches invalid [5]. Hence, we opt to implement modified versions of two link prediction scoring algorithms discussed in the Liben-Nowell and Kleinberg literature, designed to work for weighted bipartite graphs. We elaborate on our implementation of each algorithm below.

3.3.1 Katz Prediction

First, we implement the Katz predictor described in Liben-Nowell and Kleinberg. The Katz predictor algorithm defines a score $score(x, y)$ that gives a relative estimate of how likely it is that a link will appear between nodes x and y which represent a news source and a subreddit in our graph. The smaller the score, the less likely the link is predicted to be formed.

The Katz predictor is defined as:

$$score(x, y) := \sum_{l=1}^{\infty} \beta^l * |paths^{<l>}(x, y)|$$

The score sums over all paths from x to y by length l. In the unweighted version, the term $paths^{<l>}(x, y) = 1$ if there is a path of length l and 0 otherwise. The term β^l, which includes a tuning parameter β, is used to ensure shorter paths contribute more to the sum.

The input networks have up to hundreds of thousands of weighted edges, and since the graph is nearly complete, it is requires a great deal of computational power to generate all the paths and compute the Katz predictor for a single pair of nodes. To reduce the computation time, we compute an approximation by summing over only the paths of length $< k \in 3, 5$:

$$score(x, y) := \sum_{l=1}^{k} \beta^l * |paths^{<l>}(x, y)|$$

Intuitively, edges with a higher weight should count more towards node similarity than edges with lower weights, so for the purposes of computing the Katz predictor, we reinterpret the graph as an unweighted multigraph, where an edge with weight w in the original graph corresponds to w individual unweighted edges in the multigraph.

3.3.2 Rooted PageRank

Our second algorithm is a modified version of the Rooted PageRank algorithm discussed in Liben-Nowell and Kleinberg [7]. The Rooted PageRank algorithm is inspired by the notion of "Hitting time", which is the expected number of steps needed to reach from a given node x to a given node y through a random walk. Loosely, this translates to a measure of proximity and/or influence between any 2 nodes in a graph. However, the hitting time algorithm by itself is prone to be unduly influenced by relatively distant parts of the graph. The Rooted PageRank algorithm improves on this by allowing the random walk to reset (teleport) back to the original node with probability $(1 - \alpha)$.

We define score(x, y) to be the stationary probability of reaching y in a random walk initiated from x; the higher the stationary probability, the higher chance the two nodes have high proximity or influence on each other. We can calculate these stationary probabilities (and hence score(x, y)) for all pairs of nodes in the graph as such [10]:

$$RPR = (1 - \alpha)(I - \alpha D^{-1} A)^{-1}$$

Where A refers to the adjacency matrix of the graph and D is a diagonal matrix where $D_{i,i} = \sum_j A_{i,j}$ (i.e. degree matrix).

The algorithm defined above has demonstrated good performance for unweighted, unipartite graphs, but in order to make the algorithm well suited for our graph, we employed a few modifications. Since the graph is weighted, we used a weighted version of A, where $A_{i,j}$ represents the weight of an edge between nodes i and j. Since the graph is bipartite, we extracted the submatrix of the resulting RPR matrix that corresponded to edges between news source nodes and subreddit nodes, and renormalized the submatrix. This was used to provide the scores for each node pair. Finally, after some initial experimentation with this algorithm, we found that obscure news articles would frequently be overpredicted. To help potentially curb this, we randomly set the lowest p percent of scores in RPR to be 0.

4 Results and Experiments

To test our models, we generated predictions for each year in our data set (from 2009 to 2017), using the prior years to generate the train network. The parameters $k = 3$ and $\beta = 0.5$ for the Katz predictor and $\alpha = 0.9$ and $percentile = 40$ for the PageRank predictor were found to yield the best results, so we report the results for these parameters only. The accuracy and precision for each of these predictors per year is presented in Figure 7, whereas Figures 8 present the results for just the mainstream and alternative links respectively. All metrics were calculated considering only the "Core" set of nodes mentioned earlier. This is since the models would not be able to predict links between news sources and subreddits not seen in the training data.

From Figure 7, we can see that the Katz Predictor significantly outperforms PageRank on all years with respect to accuracy and precision of the predicted links. For all years and both algorithms, the accuracy for alternative links is much higher than the accuracy for mainstream links, suggesting that every predictor may be overestimating the number of links formed between alternative news sites and subreddits. We can see this by the relatively low precision scores provided for alternative news link prediction versus mainstream news link prediction. However, since alternative news URLs account for only a tiny share of overall news URL submissions as seen in Figure 6, high accuracy on alternative links does not translate to high accuracy overall.

We believe that Katz may perform best because it does the best job exploiting the following simple but important features in the data: first, as the news URL activity on a subreddit increases, news domains are likely to have roughly the same share of the new activity that they had in the past; and second, the URL behavior on subreddits tends to be more similar if they get most of their news from the same domains. The first property is exploited by counting paths of length one, and the second property is exploited by counting paths of length three. This outperforms the PageRank approach as, despite the modifications made, the algorithm may still not be fully suited for use on a bipartite graph. This is because the algorithm will still influence stationary probabilities for impossible edges (i.e. edges between subreddit nodes). The PageRank approach may also overcompensate for highly weighted edges by adding an inordinate amount of weight for already highly weighted edges and not adding enough weight to lower weighted ones.

From analyzing the raw predictions data from the Katz model, we also see that this approach tends to add too much weight to subreddit nodes with a high degree (as sum of edge weights) (e.g. `/r/politics`) while giving less weight to subreddits on the lower end of the degree frequency distribution.

Year	Total	Mainstream	Alternative
2009	183591	177139	6452
2010	220710	212374	8836
2011	300575	289459	11116
2012	411896	388623	23273
2013	760958	715769	45189
2014	740395	692175	48220
2015	981321	874530	106791
2016	1145066	1022623	122443

Figure 6: Number of predicted links by year and category

Year	Metric	Katz	PageRank
2009	accuracy	0.646	0.378
	precision	0.636	0.378
2010	accuracy	0.615	0.379
	precision	0.602	0.379
2011	accuracy	0.601	0.393
	precision	0.485	0.393
2012	accuracy	0.568	0.337
	precision	0.547	0.337
2013	accuracy	0.577	0.334
	precision	0.540	0.334
2014	accuracy	0.547	0.336
	precision	0.517	0.336
2015	accuracy	0.488	0.275
	precision	0.463	0.275
2016	accuracy	0.221	0.153
	precision	0.219	0.153

Figure 7: Prediction statistics by year and algorithm

Year	Metric	Katz	PageRank
2009	accuracy	0.643	0.367
	precision	0.643	0.437
2010	accuracy	0.615	0.371
	precision	0.615	0.457
2011	accuracy	0.555	0.383
	precision	0.555	0.478
2012	accuracy	0.569	0.320
	precision	0.569	0.402
2013	accuracy	0.566	0.323
	precision	0.566	0.410
2014	accuracy	0.546	0.326
	precision	0.546	0.443
2015	accuracy	0.500	0.292
	precision	0.500	0.426
2016	accuracy	0.471	0.278
	precision	0.471	0.439
2017	accuracy	0.223	0.191
	precision	0.223	0.279

Year	Metric	Katz	PageRank
2009	accuracy	0.921	0.674
	precision	0.032	0.125
2010	accuracy	0.848	0.578
	precision	0.032	0.100
2011	accuracy	0.859	0.674
	precision	0.030	0.108
2012	accuracy	0.868	0.586
	precision	0.053	0.132
2013	accuracy	0.836	0.586
	precision	0.052	0.132
2014	accuracy	0.673	0.422
	precision	0.042	0.088
2015	accuracy	0.420	0.300
	precision	0.050	0.084
2016	accuracy	0.586	0.291
	precision	0.060	0.071
2017	accuracy	0.571	0.387
	precision	0.019	0.025

Figure 8: Mainstream (left) and Alternative (right) news prediction statistics by year and algorithm

5 Conclusion

We were able to achieve our goal of developing a model to predict the spread of both mainstream and alternative news through social networks with reasonable success. By formulating a unique bipartite graph structure to represent the relationship between social networks and news sources, we developed a graph state where link prediction algorithms could be utilized in the prediction of news article propagation. Referencing previous work in link prediction we implemented two algorithms nuanced to performing well on our graph structure and used them to predict the spread of news articles across Reddit subreddit communities throughout 2017. Our implementation demonstrated that the Katz link prediction algorithm was the most successful in predicting new news articles spreading throughout the Reddit subcommunities. Although the Katz predictor performed very well through 2015, it was not able to achieve good

predictions for 2016, suggesting news trends changed markedly in 2016 likely in parallel with world political events within the year. Overall, we were able to achieve results with parity to those seen in similar implementations in literature.

To build upon results found in this research, a number of directions may be considered. First, we had began initial implementations of modeling the graph state utilizing the Hawkes process discussed in Zannettou et al. which would best model the potential influence of social networks on each other in terms of spreading news articles but found the required amount of data processing required for a significant granularity of results was outside of the timescope of our project. However the initial graph characterization we saw from the Hawkes process modeling was promising and more accurate link prediction can be explored further utilizing this method. Secondly, an important future undertaking is the addition of other social network platforms such as Twitter. Due to constraints on data availability we limited our research to just Reddit communities, but the expansion to a larger number of social networks would reveal whether the methods and algorithms discussed here perform well on a wider range of social networks. Thirdly, we could explore ways of reducing the influence of high degree subreddits, and also compensating for the large disparity of data per year (as shown in Figure 3). Finally, another direction that could be expanded upon is the exploration of combinations of the link prediction algorithms with other common link prediction algorithms such as the unseen bigrams algorithm discussed by Liben-Nowell.

References

[1] Allcott, H., and Gentzkow, M. "Social Media and Fake News in the 2016 Election." Journal of Economic Perspectives, 31(2): 211-236, 2017.

[2] Bakshy, E. Hofman, J., Mason, W. and Watts, D. "Everyone's an Influencer: Quantifying Influence on Twitter." In Proceedings of WSDM, 2011.

[3] Dewarim. "Updated Reddit Comment Dataset as Torrents." Reddit (r/datasets), 2017.

[4] Gottfried, J., and Shearer, E. "News Use across Social Media Platforms 2016." Pew Research Center, 2016.

[5] Kunegis, J., De Luca, E., and Albayrak S. "The Link Prediction Problem in Bipartite Networks." In Proceedings of the 13th International Conference of Information Processing and Management of Uncertainty, 380-389, 2010.

[6] Lee, K. "What Analyzing 1 Million Tweets Taught Us." The Next Web, 2015.

[7] Liben-Nowell, D., and Kleinberg, J. "The Link-prediction Problem for Social Networks." Journal of the American Society for Information Science and Technology, 58(7): 1019–1031, 2007.

[8] Savvas, Z., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., Blackburn, J. "The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources." In Proceedings of the 17th ACM Internet Measurement Conference, 2017.

[9] Simons, M. "Journalism faces a crisis worldwide." The Guardian, 2017.

[10] Wang, P., Xu, B., Wu, Y., Zhou, X. "Link Prediction in Social Networks: the State-of-the-Art." Science China Information Sciences, 58: 38-76, 2015.