Connecting the Dots in the Blogsphere

Elena Agapie, Yifan Wu

Dec 2012

Abstract

Social media is arguably the most important new media of the century. Grassroots opinion leaders propagate a wealth of information over a dynamic network. The current paper analyses the dynamics of the network based on link structures for different types of media sources. We identified the most influencial set of sources using PageRank and HITS and gained more insights into the effectiveness of graph algorithms applied on a real data set of news and blogs articles. Focusing on the Mediacloud data from the Berkman Center, we analyze the characteristics of the complex network, as a result of the influence of different media sources, and present suggestions to improve the design of the data API for better analysis.

Keywords: social network, blogging, graph algorithms

1 Introduction

Analysing social media is interesting to data mining researchers for the theoretical challenge and social implications. Blogging represents one of the more detailed methods through which people can express opinions on various topics and is deeply rich. The Berkman Center data that we are working with contains a wide range of media sources and a clean data dump that's easy to work with, with the goal to find quantitative answers to questions about the type of media coverage in the news and blogs, differences between different blog platforms on certain topics, and information flow in general [8]. We thus

set off to "mine" two month's worth of blogs and news (June and July 2011) via their link structures for to identify "important" nodes, defined vageud as influential news sources and distributor, be they in mainstream media, political or technical blogs. We also conduct two more detailed case study with information around the Arab Springs and the Chile Volcano erruption arount that time.

The paper is structured according to the different aspects of data analysis we performed. We first look at the raw data: how connected the medias are to each other, what the most connected to medias are, and how different caragories differ from each other. We also explore the algorithms by finding the important media sources via PageRank and HITS; testing the robustness of the algorithm by perturbing different varibles set for the algorithms and evaluating the change, and the temporal analysis of how the importance of media change over the span of two months. Additionally, we try to explore the temporal aspect of the ranking algorithms, by analyzing how the ranking reveals changes in the media influence over the two months of data we are using. Lastly, we integrated the experiments and discuss the implications of our experience working with practical data.

2 Related Work

Our work is motivated by various research done to identify how information propagates through the online media spehere. We therefore focus on blogs and news sites and seek to understand how this layer of information contributes to the distribution of information through the networks by identifying the most influential news sources across different types of blogs and news sites, based on link analysis.
Research done by the MIT Media Lab is directly applicable to our work. Zuckerman et al identify the information propagation though news stories of the SOPA/PIPA debate and the Trayvon Martin case [7]. One level of information they use is link analysis to track when the stories became popular and what articles helped propagate the information to the public. The conversations we pursued with the Center for Civic Media and the Berkman Center revealed that most of the analysis they performed invoved only basic link counts and human hours to analyze the network data. Our goal is to provide algorithmic tools that support this type of research and make it easier to perform.

Other relavant reserach is focuse on the interaction of different types of media such as blogs, and cross citations [1]. Adamic looks at the interaction between liberal and conservative blogs before the 2004 elections. They analyze echo chamber effects of how communities of blogs cite each other. They show that liberal blogs are more highly connected and point mostly to liberal blogs. Conservative blogs are less likely to point to other blogs. They start with 1500 blogs and reduce them to the most popular 40 blogs by mainly eliminating the blogs with few citations. This serves as an inspiration to how we might want to filter through the data and focus on subsets of the blogs. Other work such as Nakajima [4] and Song [6] research how to detect popular topics and bloggers based on social media data.
PageRank [5] and Hubs and Authorities (HITS) [3] are widely used on a lot of graph/ranking related questions, thus a natural candidate as a baseline for our analysis. Song et al 2005 [6] has used PageRank as a way to detect popular topics in micro-blogging, and Nakajim et al 2005 [4] developed a "Agitator-Summarizer" model that is closely related to "Hubs-Authority". We see a continuum of reasoning and modeling approaches. More specifically, blogs form a subset of web pages and share many similar properties of being connected and readers follow similar patterns to those of the web surfer. However, blogs differ from webpages in significant ways, as described in [2], explicit URL links in blogs have a much less dense graph and as a result PageRank and HITS don't work as well.

Among the more recent works we found Modeling Blogger Influence in a Community by Agarwal et al [2] to be a very helpful foundation. It provides an overview of how to define influence for a blog and for a blogger, based on the following metrics: recognition - inlinkis, activity generation - number of comments, novelty - outlinks, inversely related, eloquence - length. Given that we don't have access to comments[1], we will focus our analysis only based on inlinks and outlinks.

[1] And this is much harder to scrape in general

3 Data Processing

3.1 Basic Process

We scrape the cleaned version of all the HTMLs from the media sources the MediaCloud contains and map the URL and "stories" (each HTML page is a sotry with a unique ID) back to the original media sources (please see appendix for listing). We then build the network and run different algorithms over them. For keyword specific HITS we employ MongoDB to load the HTML contents and filder for stories with those keywords. For the detailed process please refer to graph below, Figure 1, Processing Procedures.

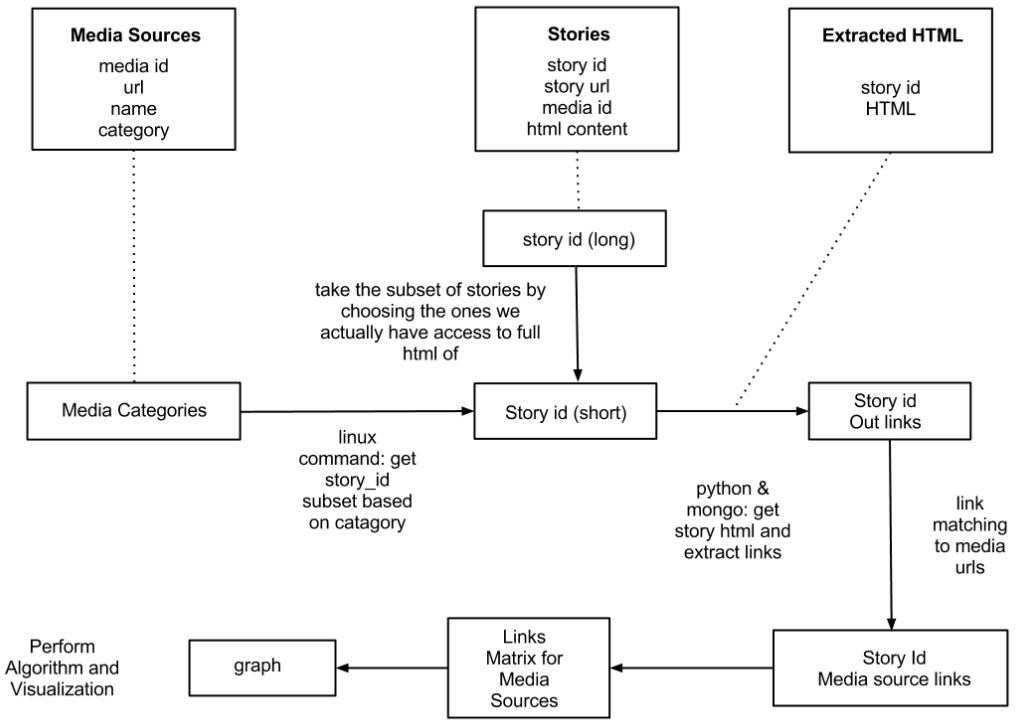
3.2 Limitation of Data

As an early disclaimer, we haven't yet doncucted a wider range of data to solidify the findings. The media soruces are not comprehensive, causing there to be unmatched links. To deal with this, we loosened the search criteria to achieve broader matches, and it worked out well. While the strict matches only had 2325 matches, the loose match had 498109, which is the majority of our matched edge. Some times the media sources share the same domain, like wordpress, and a match is really rare for a subdomain for a more specific one, so we allow for loose matching to make possible more candidate edges, at the cost of the granularity of original media source divide. This is a big compromise we chose to make, and we are looking into ways of alleviating this trade off. For a detailed description of how the loose matching works please refer to the appendix. The data is also not symmetric, for instance, there are links to mainstream russian media but we have no stories belonging to that media. This relates to a broader question of how best to organize information and break them down into consumable pieces for research and API access.

One important issue is that we don't have a clear understand nor evaluation of how "clean" the data is, and that makes a difference on the accuracy and approach of the algorithm we might think about. Note that this is very difference from previous resesarch perspectively where the researcher collaborate with the online platforms and have access to neat and organized data. By design, different categorizes could map to the same media, [2] which

[2] Which we realized later on; I guess we're helping Mediacloud find and fix bugs in the process.

Figure 1: Processing Procedures



	June	Percentatge	July	Percentatge
No. of raw edges	1,020,843	100%	1,135,708	100%
No. of matched edge	615,595	60.30%	699,374	61.58%
No. of self edge	164,049	16.07%	198,940	17.51%

Table 1: Edge Meta Data

	June	July
No. of stories	427,746	
No. of stories with out-links	186,112	128,149
Percentage Match	87.02%	59.92%

Table 2: Stories Meta Data

increases the representation of certain catagories. Since our importance analysis is mainly based on media sources as opposed to catagories, the integrety of the research finding is not too severly harmed.

4 Data Characteristics

4.1 Graph Characteristics

Out of the matched edges, a large portion is self-pointing. This is not surprising because sites often point to its own resoruces[3]. Given the loose match, around 60% of the edges we identified matched with our media sources, which demonstrates the comprensiveness of the media sources. This echos with the "long tail" nature of the internet. The average number of edges per node (conditioning on that it has edges to begin with) is **5.23** and the standard deviation is **9.67**. Details per the tables below.

[3]There are of course SEO incentives to do so as well

4.2 Media Categories

Media categories is hand picked by Berkman researchered and used for organization and comparison purposes. Examples of media categories are: political blogs - left wing/right wing/center, popular blogs, top technical blogs, mainstream media, random blogs, a variety of Russina blog categories. These categories are provided through the Mediacloud data. The categories have been annotated manually by humans and contain approximatelly 9000 of the media sources. The categories are provided in Table 9, which includes the number of stories associated.

5 Results

The graph algorithms focused on identifying the most influential news sources in the network graph: identifying the sites that are most linked based on link counts, using PageRank to identify the weights of the different news sources in the dataset, using HITS to identify which are the hubs and authorities in the dataset, a temporal analysis of the changes in ranking across the two months' data. The analysis maps the raw edges up a level to media sources to increase connectivity, rank their importance and identify inluentials in different catagories: are there news sources that are particlarly influential in each category? Do the different categories have different levels of influence?

5.1 Degrees of the Link Graph

Rankings of the Sights with the Highest Number of *In Edges* in July
We ranked the number of edges pointing to media sources and surprisingly, the simplest ranking indeed reflects the more cited media sources. Here the media sources are among social media sites and established news sites such as *New York Times* and *Wall Street Journal*. This confirms the general observation that social media such as *twitter, youtube, and flickr*serve as an important tool for the spread of information. However, we realize that although the absolute count may seem to be doing a good job at identifying the big players in the field, the results for the smaller players are very inacurate and prone to abuse.
Rankings of the Sites with the Highet Number of *Out Edges* in July

Table 3: In Degree Ranking

Count	Media Source	Category
326	twitter.com/#!/globalvoices/nigeria-elections-9	Colin's Nigeria Blogs
282	en.blog.wordpress.com	Top Tech Blogs
277	youtube.com/videos?s=mp	Popular Blogs
273	amazon.com/gp/goldbox	Popular Blogs
209	google.dirson.com	Popular Blogs
203	nytimes.com/pages/health/index.html?partner=rss	Popular Blogs
192	flickr.com/groups/abstract_art/pool	Popular Blogs
161	online.wsj.com	Popular Blogs
151	guardian.co.uk/environment/blog	Technorati Politics
131	washingtonpost.com/?nav=rss_email/components	Popular Blogs

The out-degree analysis reflects the count of media sources that link the most to other sources. We notice very popular news sources here, spanning from blogs, such as salon, to mainstream media, such as the Guardian, Forbes, and Hacker News (Y Combinator). The broad representation of media categories informs us that every category has powerful news aggregator and distributors. One issue however is that not only are news hubs represented but also potentially advertisement heavy sites that links to many junk sites just for profit.

Visualization of the Connectedness of the Network

The first pair is a log log graph of the connected components. The overall strucure is very similar across June and July, and both are pretty closely conncted, suggesting that although blogs tend to link less, their aggregation to the media level is engaged in active converstations.

As part of early project research, we were inspired by the political blogging behavior researched by Adamic etal in 2012 [1]. So we wanted to answer a similar question with our data. Here the Red represents the political right and the Blue the left. Overall it seems decently connected (as opposed to the drastic divide in the 2004 paper). It could be that we are looking on the media source level and they communicate more information across their platform than bloggers who generally share opinions, or it could be that the political

Count	Media Source	Category
478	examiner.com	Top 25 Mainstream Medi
210	forbes.com/	Top 25 Mainstream Medi
204	gawker.com	Popular Blogs
195	salon.com/rss/all_salon.rss	Popular Blogs
189	guardian.co.uk/	Popular Blogs
185	msnbc.msn.com/	Top 25 Mainstream Medi
179	time.com	Top 25 Mainstream Medi
178	metafilter.com/	Popular Blogs
176	huffingtonpost.com/raw_feed_index.rdf	Popular Blogs
174	news.ycombinator.com/	

Figure 2: Indegree LogLog Plot, June and July

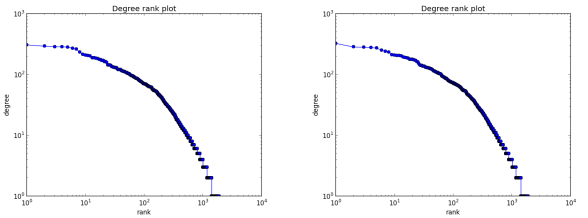
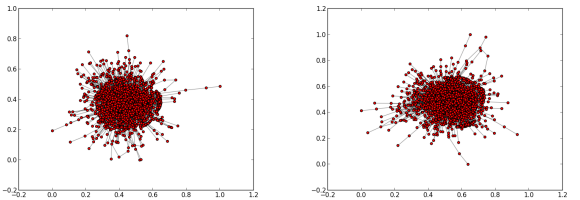
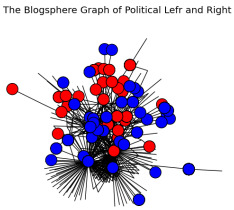
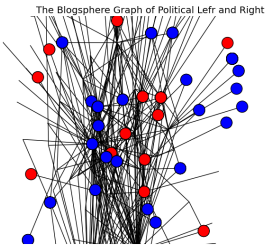


Figure 3: Indegree LogLog Plot, June and July



debates are now more integrated into both sides. To reach a more concrete hypothesis we need more data and analysis.





5.2 PageRank

We used PageRank to identify the relative importance of different nodes in the network relative to the entire structure of the network. The PageRank uses the connectivity matrix we build over the entire graph. We also experiment with different values for the probability of a random jump. We expect out user to have a low jumping likelyhood when the media sources in our dataset are rather specialized (political or technical), and since this comprises a large portion of our media sources, it is expected that users would have a particular preference and orientation of their opinion, but we experiment with a range of jumping probability regardless to test the hypothesis, per below.

After ranking all the media sources in our graph using PageRank we analyzed the results based on the category of the media sources, we have the following table: (note again that becuase of our loose matching it may not be the exact website)

The results are reflective of the nodes in the network that are *most linked to*, which in this dataset rank websites that represent social media or very commonly linked websites, because of reviews and product. Interestingly, this doesn't capture however the ranking of

Figure 4: Pagerank Results of Different Media Sources

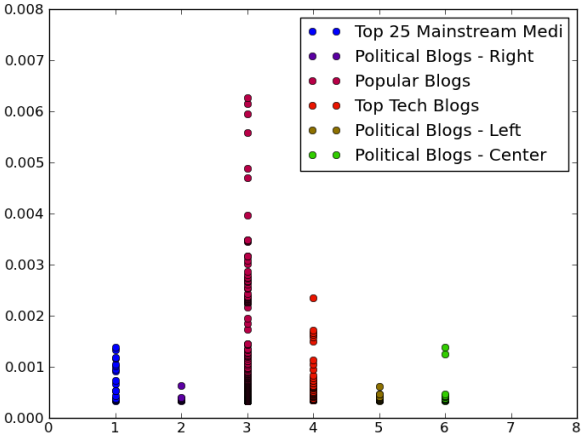


Table 4: Page Rank Top Ten List

0.08705	Popular Blogs	http://www.amazon.com/gp/goldbox
0.06035	Popular Blogs	http://itunes.apple.com/WebObjects/MZStore.woa/wa/
0.01990	Colin's Nigeria Blogs	https://twitter.com/#!/globalvoices/nigeria-elections-9
0.01695	Popular Blogs	http://www.youtube.com/videos?s=mp
0.01673	Top Tech Blogs	http://en.blog.wordpress.com
0.01631	Popular Blogs	http://www.flickr.com/groups/abstract_art/pool/
0.01205	Popular Blogs	http://www.apple.com/hotnews/
0.00627	Popular Blogs	http://google.dirson.com
0.00615	Popular Blogs	http://www.ted.com/talks/browse
0.00595	Popular Blogs	http://www.rottentomatoes.com/movies/

the most influential news sources in the political and tech space - which are the most represented ones in our dataset, demonstrating yet again the power of general purpose news sources.

We further looked at the ranking relative to categories. Figure 10 shows a summary of rankings across categories. We focused on the categories that have most stories. We notice that popular blogs, mainstream media, some tech blogs and political center blogs have the highest rankings amongs all the nodes in the graph. We also notice that political left and political right have overall smaller rankings than the other categories. It could be hypothesized that people pay more attention to the mainstream media and other popular blogs than they do to a more biased sample that focuses on very specific opinion orientations (such as the left and right wing blogs). This has relevant implications in gaining knowledge on how people behave and what they read and whether these rankings are reflective of user behavior - paying attention to more neutral news sources than to biased ones.
The table below shows a selection of the most influential media sources in different categories, going beyond the top 10 sources that might be biased by the nature of the sites they represent, with the corresponding ranking. These reflect the highest ranks across categories that you can notice in the in the overall plot of ranks too. The results below which are highest in their categories, are reflective of the highly ranked news sources, such as the New Yorker, the NY Yimes or NPR. The results also show the influential news sources in other categories that might not be as familiar to the user. We think these are succesful results, as we look at particular categories. The categories help to identify influential news sources that otherwise would be lost in the ranking values.
We have also analyzed the distribution of ranks across different categories. This can inform us on how the influence is distributed within each category and what is the density of influential blogs in each category. The different categories have a majority of sources with almost equally low ranks. The political blogs center have a slightly more dispersed range or rankings and the mainstream media has the broadest range of rankings - showing that mainstream media is has most sources that are more influential than the average. You can see that in the plots below representing rankings on Top 25 Mainstream Media, and Plitical Blogs Center and Right.

Figure 5: Top 25 Mainstream Media Page Rank

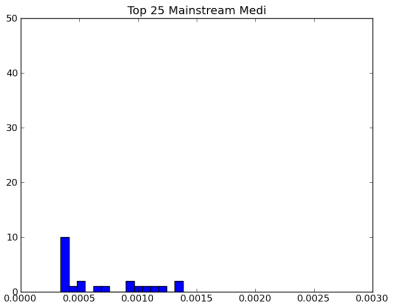


Figure 6: Top Political Left Blog Page Rank

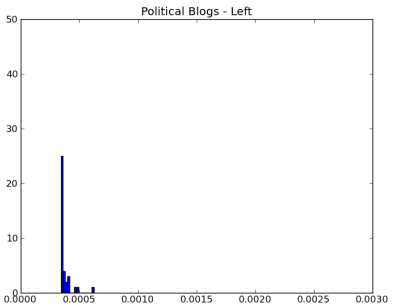


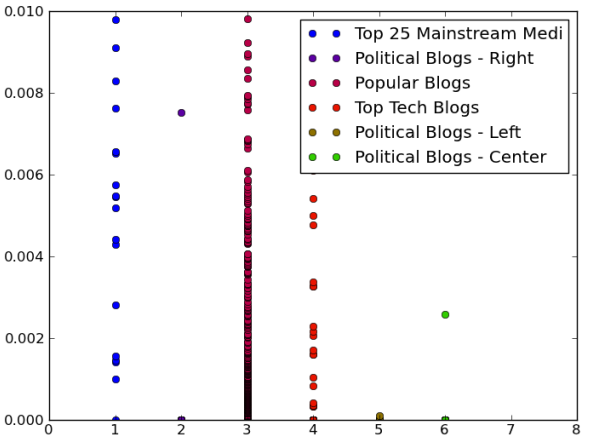
Table 5: Page Rank Result

0.00559	Popular Blogs	newyorker.com/rss/feeds/everything.xml
0.00488	Popular Blogs	nytimes.com/pages/health/index.html?partner=rss
0.00291	Technorati Politics Uncoded	guardian.co.uk/environment/blog
0.00226	US Random Blogs	which4u.co.uk
0.00138	Top 25 Mainstream Medi	nydailynews.com/
0.00134	Top 25 Mainstream Medi	dailymail.co.uk/home/index.html
0.00114	Top Tech Blogs	arstechnica.com/index.php
0.00105	Top Tech Blogs	latimesblogs.latimes.com/technology
0.00062	Political Blogs - Left	thinkprogress.org
0.00047	Political Blogs - Left	mediamatters.org
0.00063	Political Blogs - Right	pajamasmedia.com/victordavishanson
0.00040	Political Blogs - Right	newsbusters.org/blogs/jeff-poor
0.00138	Political Blogs - Center	npr.org/blogs/politicaljunkie
0.00124	Political Blogs - Center	www2.politicalbetting.com/

We inteded to analyze the temporal aspect of influence. The results we presented are obtained from analyzing the month of June. We ran the same analysis on the months of July but did not obtain very different result. In Figure 2, you can see that the overall ranks stay about the same between the two months.

In relation to the in-degree and out-degree analysis computed above we notice that PageRank is heavily influence by the similar factors as the out-degree analysis - which is the *advertisement and product oriented* sites that have a significant amount of out-links. We also tried to explore different parameters of the jumping probability: parameter of 70% and 97% probability of random jump were tested. We found that the top 1 ranking is very robust against the change. It seems that results are highly innfluenced by generic, popular sites without a specific focus but have either a lot of user generated contents (social networks), or e-commerce site with products, followed by big news sources.

Figure 7: Hubs Plot On all Media



5.3 HITS

We run the HITS algorithms to identify the hubs and authorities in the network. This give us extra information in addition to PageRank to identify the influential nodes. Although HITS performs better when it's ran on a subset of the network focused on one topic, we think it would provide us with valuable information on the hubs and authorities of the entire network as well.

Hubs Score

The results we obtain across the different categories emphasize the difference in ranking and spread of ranking of media sources. We notice that tech blogs, popular blogs and

Table 6: Top Ranked Hubs Scores of Media Souces

0.0140	Top 25 Mainstream Medi	examiner.com
0.0117	Top 25 Mainstream Medi	washingtonpost.com
0.0098	Popular Blogs	gawker.com
0.0079	Political Blogs - Left	dailykos.com
0.0075	Political Blogs - Right	hotair.com
0.0072	Top Tech Blogs	gigaom.com
0.0069	Political Blogs - Left	www.talkingpointsmemo.com
0.0068	Top Tech Blogs	techcrunch.com
0.0025	Political Blogs - Center	fivethirtyeight.com

mainstream media are the highest ranked hubs. The politica blogs left and right are verly low ranked as hubs, and so is most of the political center. This has implications at the level how different political spaces behave. It could be argued that the more biased or focused news spaces point very little to other media sources, as opposed to the more mainstream sources. Mainstream sources might also have strong biases, but they seem to be pointing to more sources overall. In the table below we present the top ranked media sources in different categories:

This type of ranking provides a perhaps hidden overview of the media sources that are providing most links to other media sources. We might expect that from the *Examiner*, a content aggregation site, but it might be surprising to see that for the Washington Post. Similarly, *Gigaom* is raned higher than *TechCrunch*. Futhermore, this table of the highest ranks across categories provides a refernece to the few exceptions of hubs in political left and right wing blogs: the *Dailykos* and the *Hotair*. This information is relevant to the reserchers who study media representation and makes it easier to identify which media sources are the most important hubs for the media sources that we are concerned with.

Authorities Score

Table 7: Top Authorities Score Table for All Media Sources

0.0106	Colin's Nigeria Blogs	twitter.com/#!/globalvoices/nigeria-elections-9
0.0103	Popular Blogs	youtube.com/videos?s=mp
0.0103	Top Tech Blogs	en.blog.wordpress.com
0.0080	Technorati Politics Uncoded	guardian.co.uk/environment/blog
0.0073	Top Tech Blogs	wired.com/epicenter/
0.0063	Top Tech Blogs	latimesblogs.latimes.com/technology
0.0061	Political Blogs - Center	npr.org/blogs/politicaljunkie
0.0056	Top 25 Mainstream Medi	msnbc.msn.com/
0.0056	Top Tech Blogs	techcrunch.com
0.0054	Top 25 Mainstream Medi	dailymail.co.uk/home/index.html
0.0053	Political Blogs - Center	www2.politicalbetting.com/
0.0034	Political Blogs - Left	thinkprogress.org
0.0020	Political Blogs - Left	dailykos.com
0.0015	Political Blogs - Right	hotair.com
0.0014	Political Blogs - Right	pajamasmedia.com/victordavishanson

On the level of the authorities in the graph we identified the following media sources:
The results capture some of the well known high authorities in the tech blogs space - *Wired* and *Techcrunch*, in mainstream media - *MSNBC* and *the Daily Mail*, and in the political center - *NPR* and in political left - *ThinkProgress* and right - *Hotair*. A lot of these are recognized as established authorities in their categories. We consider these results to be a success, capturing actual authorities in the news space. Again, we have to credit the availability of media categories - whithout them we would not be able to capture as high ranked all of the aboe news sources. Secondly, the top 10 media sources contain a lot of the similar news sources we identified in the in-degree analysis. These are related to the nodes that are most liked to. Authorities provides a better overview due to the fact that it accounts for high hubs that are pointing to the authorities.

We also analyzed the distribution of the authority levels over each category. The distribution of ranks is not as skewed as the hubs distribution. However, the categories have a similar pattern in terms of ranks. Tech blogs, popular blogs and mainstream media have the most representatives with high ranks, and the political left, right and cener have the overal lowest ranks.

We compared the results over the months of June and July and the rankings were extremely similar, so we are not including the plots for the month of July.

Compared to the in and out link analysis and with PageRank, we think that the hubs and authorities results are most representative to the actual authority levels of these media sources.

5.4 HITS on topics

To evaluate the HITS algorithm better we run it on topic subsets, as suggested in the original algorithm. We use the 2011 entry in Wikipedia to identify events that occured in June or July or 2011. We pick the topics of ”arab spring” and ”volcano chile” as representative for the a social and a natural event. We evaluate how these events are represented in the media and how the analysis per topic affects the ranking of the media sources covering the story.

Figure 8: Authority Plot for All Media

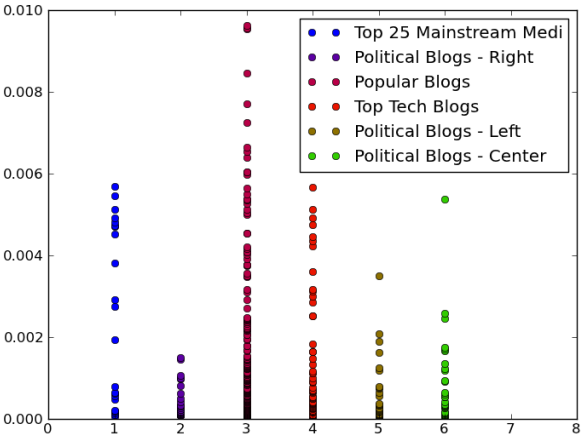


Table 8: Arab Spring HITS

rank	category	media source
0.1265	Popular Blogs	news.bbc.co.uk/go/rss/-/2/hi/default.stm
0.0899	Popular Blogs	www.huffingtonpost.com/raw_feed_index.rdf
0.0869	Top 25 Mainstream Medi	www.bbc.co.uk/?ok
0.0827	Popular Blogs	www.newscientist.com/
0.0786	Popular Blogs	news.bbc.co.uk/go/rss/-/1/hi/default.stm
0.0720	Top 25 Mainstream Medi	www.huffingtonpost.com/
0.0632	Popular Blogs	news.nationalgeographic.com
0.0494	Popular Blogs	news.bbc.co.uk/go/rss/-/1/hi/technology/default.stm
0.0494	Popular Blogs	news.bbc.co.uk/go/rss/-/2/hi/technology/default.stm
0.0419	Top 25 Mainstream Medi	washingtonpost.com

For the Arab Spring topic we selected all the articles in the dataset that contained the words *Arab* and *Spring*. We had a total of 1610 stories that generated 4683 edges covering 180 media sources. After running HITS on this graph we obtained similar results for both hubs and authorities, containing the following top 10 media sources:
The ranks differ slightly along the hubs and authorities but the ordering is about the same, suggesting a well connected graph where the hubs and the authorities are the same. This analysis shows that hits performs much better when it is focused on a smaller network that treats a specific topic.
We will also present the results we obtained while analyzing the topic of the volcano eruption in Chile. We selected all stories that contained the word *Chile* and *volcano*. The difference in this case is that the hubs and authorities are slightly different. the hubs contain mostly popular blogs and the authorities contain a combination of blogs and mainstream media. Let's look at the differences between the top sources for each.
This shows the big difference between hubs and authorities in this case. The hubs identify either generic sites, like *Google* or *Twitter*, or popular blogs such as *pcmag*, *Engadget*, or wheat can be identified as blog on a major news site such as *the Guardian* and *BBC*. On the other hand the authorities are prominent news sources such as *BBC technology feed* and

Table 9: Hubs Score for Volcano

rank	category	media source
0.0763	Technorati Politics Uncoded	guardian.co.uk/environment/blog
0.0647	Colin's Nigeria Blogs[4]	twitter.com/#!/globalvoices/nigeria-elections-9
0.0528	Popular Blogs	google.dirson.com
0.0360	Top Tech Blogs	mashable.com
0.0360	Popular Blogs	pcmag.com
0.0319	Top Tech Blogs	engadget.com
0.0314	Popular Blogs	articles.sitepoint.com
0.0314	Technorati Politics Uncoded	sacbee.com
0.0245	Popular Blogs	gawker.com/tag/valleywag
0.0245	Popular Blogs	news.bbc.co.uk/go/rss/-/2/hi/default.stm

Table 10: Authorities for Volcano

rank	category	media source
0.1265	Popular Blogs	news.bbc.co.uk/go/rss/-/2/hi/default.stm
0.0899	Popular Blogs	www.huffingtonpost.com/raw_feed_index.rdf
0.0869	Top 25 Mainstream Medi	www.bbc.co.uk/?ok
0.0827	Popular Blogs	www.newscientist.com/
0.0786	Popular Blogs	news.bbc.co.uk/go/rss/-/1/hi/default.stm
0.0720	Top 25 Mainstream Medi	www.huffingtonpost.com/
0.0632	Popular Blogs	news.nationalgeographic.com
0.0494	Popular Blogs	news.bbc.co.uk/go/rss/-/1/hi/technology/default.stm
0.0419	Top 25 Mainstream Medi	washingtonpost.com

Washington Post, or scientific source such as *New Scientist*, and *National Geographic*.

In summary, the two topic examples that we provided offer an overview on the different types of rankings HITS can generate with topics that might spread differently in the media.

5.5 Temporal Analysis

As described in each of the sections above, we compared results between June and July and didn't notice significant changes of the change of importance for media sources. One question for going into the future would be how to propagatee importance factor to new emerging articles.

6 Discussion

6.1 Discussion of Results

We started from the premise that we would identify the most influential nodes in the network based on link analysis. The algorithms we applied gave us an overall ranking of the different media sources in our network. We observe that having the categorization of the media sources significantly improves the relevance of the ranking. Using the categories we obtain several sources that are clearly influential (such as *NPR*, *Wired*, etc) which were not among the highest ranked in the full network. Overall, the algorithms pointed us to sources that we expected to be influential, but also to srouces that we were not as familiar with.

Secondly, using a combination of PageRank and HITS gives us a better sample of highly ranked sources than using each of them individually. PageRank will be skewed by the media sources that have intrinsically more links than the rest of the network because of their domain that has a lot of pages, hubs suffers of some of the same issues, and authorities gives some improvements in this sense. Looking at the results from all three gives a more comprehensive understanding of the space.

Third, the influence of a particular news source can manifest just at the level of a particular topic. Most news sources don't present stories on all topics. The topic examples we

provided illustrated some of this aspect of the problem. Thus finding an influential source should always be tied to a topic or range of topics.

6.2 Future Research

There are several research directions that we would like to approach in this space. Our analysis only involved the basic PageRank and HITS algorithms. First, edges don't always have the same weight. We might want to adjust the weight of an edge between two sources if those particular sources cite each other a lot. For example, edges could be weighted based on the frequency between the two nodes they connect. Secondly, we believe there could be a temporal value added to nodes, particularly in the blogs space. If a blog is highly cited during one period for some specific articles, how should that affect the future weights coming from a source that is currently highly ranked, but usually might not be? how can we incorporate the temporal aspect of news sources and their influence? over how long of a period of time does the influence of a blog oscillate and by how much?

Moreover, we would like to evaluate the results by comparing against existing network analysis. Towards this end we have been in touch with researchers from the Berkman Center and Media Lab who have already done this type of analysis on a couple of topics with network results validated by humans.

7 Reflections on the Research Process

7.1 External Problems

Getting the MediaCloud data took a while, and the data we were initially given was truncated and a lot of the mapping between stories and the media source was missing (aproximatelly 1 in 4). We didn't realize this until towards the end of our data processing when the results looked problematic. It is hard to have a clear expectation about what's correct for "big data" problems. The data was also not clearly described. There were repeated categories that we were not aware of, which made debugging confusing, since correct/incorrect signals were mixed. Another is that the relationships (e.g. one to one, on to manyetc.)

between different entities were not provided. We made the mistake of having extra dependencies on certain uniqueness and existence and ran into numerous bugs. Another layer of complexity which we completely did not expect is issues with computing resources at SEAS - we had our home directory permanantly deleted once, ran out of space the other, and several times the Python configuration stopped running and or RSA login failed. All these are solvable but involved communication and slowed the progress. That being said, we still received excellent help from Ian Stokes, academic computing, which made things easier.

Many of the links, despite having been scraped, still contain a lot of "extra" edges that are emails, links to ads, photos, etc. *Beautiful Soup*, a web scraping library that we are using, is not completely clean in terms of getting the most accurate link scraping so that introduces extra noise too that slows industrial level process because constant human intenvention.

7.2 Internal Problems

"Big" data mining is more ambiguous than we thought and it was hard to ask the right questions, especially given the constraints. Both of us were learning about the tools as we proceeded with the research, so there were a decent amount of ramp up time. It was also easy to be buried inside a mix of issues and not pin point exactly where the issues are, delaying the progress much beyond our initial expectations.

A meta level struggle with data mining in general is finding surprising yet accurate result. Most of our findings echos what we already know, but a few goes beyond to confirm unsure conjestures and even to suggest surprising findings. It is a skill, or even an art, to be able to ask and look at the right data.

8 Conclusion

Understanding how information flows in the network is important in the era of "big data", not only to help digest opinions and trends, but also to assist development of social theory. On the other hand, the amount of noise present on today's compact webpages is very large.

It requires a lot of preprocessing to be close to ready for PageRank and HITS. Even when employed for Pagerank and HITS it doesn't really provide enough granularity for smaller media sources. As a result, a specific evaluation on a page is even more inacurate.

Through preparing for this final project, we learned hands on tools to analyze data sets and to implement algorithms. We gained a better appreciation for researchers and engineers' phenomenal task at developing innovative theories and making complex analytics system work. We also developed some insights in certain aspect of the news source, and found new questions we wish to pursue in the future. Lastly, our use of MediaCloud's data and infrastructure also helped the Berkman Center reflect on how best to share their information and some existing issues.

A Visualization Details

The Pagerank Result of the Two Graphs analyzed, Side by Side

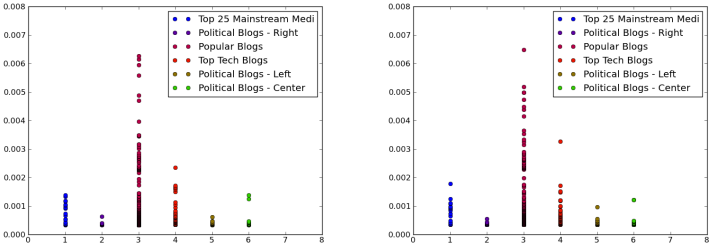


Figure 9: This is a comparison of the June to July change in ranks through the PageRank algorithm. There are some differences, but they are generally minimal

B Implementation Details

B.1 URL mapping

It is tricky to elevate entities to the same level to decrease sparsity but also keep them fairly paralell. Our first heuristic is to get rid of *html* and *www*, and try to match the rest, then if that failes we take just the part before the extention (e.g. .com) is matched[5].

Some of the links are shortened links like *bit.ly*, and a simple regex would not be sufficient to map to desired media sources, so we implemented a method to get the expanded URL to fix it. More heuristics could be applied to improve the analysis process. For instance filtering out social media plugin sites such as Google Plus and Facebook Like, which we saw numerous accurrences of.

[5] And more implemetation nuances for domaines with two parts like *.co.uk*

In general URL mapping is a huge part of preprocessing and is also a key limiting factor. We know of other research that just generates a new key if the mapping doesn't workout, but that would generate severely sparse graphs. We suggest that the Mediacloud provide an API tool that helps with the mapping of the data to unify the research experience and expectations.

B.2 Better Practices for Mediacloud Data Dump

There was overlap in the two files we received with data from the 2 months and it askewed our initial analysis. Again, dealing with big data could be tricky in terms of identifying issues and it is ideal to be provided with more informed and expected data.

B.3 False Dependencies

The media categories generally don't overlap but they could — we found overlaps in Political Blogs with Political Blogs - Left and Political Blogs - Right. As a result in our analysis we chose to use, as much as possible, categories that don't overlap in media sources.

B.4 Plotting

Visualization is a very important part of data presentation. We experimented with different types of histagrams, and other combination of axis, and chose to stick with the ones that are more capable of expressing the comparison between catagories, as we see that's where our biggest contribution lies. We tried various tools for network plotting, such as networkX and Gephi, but visualizing large graph data prooves to be very difficult.

C Meta Data on the Media Cloud Data

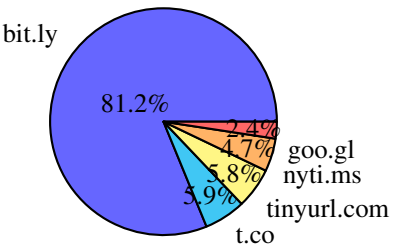
D Random Findings: Shortened URLs

While processing the data we found some interesting tangent info about the data: for the 1135708 links identified in the month of July, 2492 were *bit.ly* links, 178 were *t.co*, 177

Table 11: CategoryTable - all categories present in the dataset and the numver of corresponding stories

ID	Category	medias	stories
26	Popular Blogs	838	210148
879	US Random Blogs	996	1038
16744	Colin's Nigeria Blogs	62	0
16900	Top Tech Blogs	100	10885
16915	Iowa Mainstream Media and Popular Blogs	17	0
16932	Iowa Right Blogs	19	0
16959	Top 25 Mainstream Medi	24	262155
1878	Russian Top 25 Mainstream Media	26	0
1905	Russian Popular Blogs	1235	0
3144	Russian Random Blogs	1376	0
4628	Oil Spill Mainstream Media	2	0
4631	Russian TV	9	0
5719	Russian Government	7	0
5727	Political Blogs	393	10272
6112	Technorati Politics Uncoded	967	23605
7047	technorati top 100	99	10872
7125	Political Blogs	716	3970
7126	Political Blogs - Center	207	42
7127	Political Blogs - Right	304	921
7128	Political Blogs - Left	205	3007
7129	White House	1	0
7131	Technorati Politics Left Top 25	25	3007
7132	Technorati Politics Right Top 25	25	921
7133	Technorati Politics No Ideology Top 25	17	42

were *tinyurl.com*, 145 were *nyti.ms*, 76 were *goo.gl*, and others insignificant. So about 0.27% is shortened URLs total



References

[1] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 36–43, New York, NY, USA, 2005.

[2] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Modeling blogger influence in a community. *Social Netw. Analys. Mining*, 2(2):139–162, 2012.

[3] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. In Howard J. Karloff, editor, *SODA*, pages 668–677. ACM/SIAM, 1998.

[4] Shinsuke Nakajima, Junichi Tatemura, Yoichiro Hino, Yoshinori Hara, and Katsumi Tanaka. Discovering important bloggers based on analyzing blog threads. 2005.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[6] Shuangyong Song, Qiudan Li, and Xiaolong Zheng. Detecting popular topics in microblogging based on a user interest-based model. In *IJCNN*, pages 1–8. IEEE, 2012.

[7] E. Zuckerman. Mediacloud gephi.

[8] Ethan Zuckerman. Introducing mediacloud. 2009.