# Traditional media seen from social media

**Jisun An**
The Computer Laboratory,
University of Cambridge, UK
jisun.an@cl.cam.ac.uk

**Daniele Quercia**
Yahoo! Labs,
Barcelona, Spain
daniele.quercia@gmail.com

**Meeyoung Cha**
Graduate School of Culture
Technology, KAIST
meeyoungcha@kaist.edu

**Krishna Gummadi**
Max Planck Institute for
Software Systems
gummadi@mpi-sws.org

**Jon Crowcroft**
The Computer Laboratory,
University of Cambridge, UK
jon.crowcroft@cl.cam.ac.uk

## ABSTRACT

With the advent of social media services, media outlets have started reaching audiences on social-networking sites. On Twitter, users actively follow a wide set of media sources, form interpersonal networks, and propagate interesting stories to their peers. These media subscription and interaction patterns, which had previously been hidden behind media corporations' databases, offer new opportunities to understand media supply and demand on a large scale. Through a map that connects 77 media outlets based on Twitter subscription patterns, we are able to answer a variety of questions: to what extent New York Times and the Wall Street Journal readers overlap? Are they competitors or potential collaborators? Are people who know each other more likely to subscribe to similar outlets?

### Author Keywords

Social media, Twitter, Visualization, Structural hole, Media landscape

## INTRODUCTION

For the past two decades, a large number of studies have been focusing on understanding the exposure or the readership of media sources in order to develop effective marketing strategies in the news business. The goal of these strategies has been to identify what news readers like to ultimately maximize the exposure of news products. For small news outlets, these ways of studying market segments and maximizing exposure are not sustainable, as they are costly and their results quickly become outdated.

This work proposes a new way of visualizing the relationship between media sources in a cost-effective manner by utilizing social media data. The visualization of *media landscape* can be repeated over multiple time snapshots, which then can provide an up-to-date view of the media industry. We focused on 77 major traditional media sources presenting on Twitter and their aggregate 14 million followers.


*WebSci'13*, May 1 – May 5, 2013, Paris, France.


We create a map by connecting pairs of media sources that are conceptually "close"–depending on what fraction of the subscribers of the two media sources overlap–and show that this map of media landscape can provide the following functions:

- Captures the competition structure among media sources by identifying the degree to which subscribers overlap;
- Identifies the influential media sources that would have not been identified by traditional marketing studies;
- Unveils hidden opportunities (structural holes) in the media landscape and suggests that these opportunities are currently exploited by very popular journalists.

## METHODOLOGY

To build a map of media landscape, we need data and a methodology–both are described next.

### The Twitter dataset

We used the Twitter dataset from previous work [1][1]. For the analysis, we chose seventy seven popular media sources in different categories by consulting Twitter's 'Find People' directory[2] and `http://wefollow.com`, a user powered directory service that lists popular Tweeters by topic. We only considered sources having at least 10,000 followers in order to make sure that each media source had a large audience.

For those 77 media sources, we extracted all follow links to them and their tweets. Also for each user following the media source, we extracted her follow links and tweets in order to study how users interact with media on Twitter. Users mentioning a media source were identified based on the inclusion of '@medianame' in their tweeted text.

The resulting dataset includes 77 media sources that posted 471,121 tweets and have a total of 14,236,029 subscribers. Media outlets were mentioned 594,573 times altogether by their subscribers. Some of the 14M media subscribers were interconnected among themselves. In total they produced 48M follow links. For convenience, Table 1 collates a summary of the data for the representative media sources in each category.

[1]It comprises the following three types of information: profiles of 54M users, 1.9B directed follow links among these users, and all 1.7B public tweets that were ever posted by those users.

[2]`http://twitter.com/#!/who_to_follow/interests`

| Topic | Account | Followers (Active) |
|---|---|---|
| News | nytimes | 1,755,740 (27.2%) |
| (38 sources) | TerryMoran | 895,157 (19.2%) |
| Politics | nprpolitics | 1,145,170 (28%) |
| (5) | jdickerson | 953,993 (19.3%) |
| Technology | BBCClick | 1,165,991 (22.3%) |
| (13) | mashable | 1,270,763 (31.8%) |
| Business | davos | 750,523 (26.4%) |
| (2) | alleyinsider | 861,715 (18.5%) |
| Sports | NBA | 1,172,755 (25.7%) |
| (7) | nfl | 981,309 (22.5%) |
| Music | MTV | 294,971 (75.9%) |
| (3) | iTunesTrailers | 814,011 (23.9%) |
| Fashion & | themoment | 1,094,496 (20.2%) |
| Gossip (4) | peoplemag | 1,289,415 (24.6%) |
| Leisure & | trazzler | 944,266 (17.9%) |
| Health (5) | goodhealth | 653,939 (20.8%) |

**Table 1. Summary the 77 media sources studied**

Out of the 77 media sources, 18 of them were individual reporters and journalists. A large fraction of them belonged to the NEWS or TECHNOLOGY category. For convenience, we identified four high-level categories–news, technology, sport, and entertainment–and mapped each media source into one of these four categories. Media sources in NEWS have 10M unique followers whereas 6.7M are in TECHNOLOGY, 3.5M in SPORT, and 7M in ENTERTAINMENT. The audience of these media sources show a restricted subscription pattern in that most people follow 2-3 different media types and only 20% follow more than 10 media sources.

### A closeness model of media relationship

To build a map of connections, we need to define when to connect two media outlets–that is, when to say two outlets are "close" to each other. As a measure of *closeness*, we calculate the fraction of common subscribers out of the union set of their subscribers, known as *Jaccard similarity coefficient*. Intuitively, the closer two media sources are, the more their audience overlap. Let $A$ represent the media of interest and $\{B_1, B_2, \cdots, B_n\}$ be the set of $n$ other media sources for which we would like to measure the distances from $A$. Then, the closeness value of $A$ from $B_i$ is defined as the probability that a random user who follows either $A$ or $B_i$ follows both of them at the same time :

$$\mathrm{c}(A, B_i) = \frac{|A \cap B_i|}{|A \cup B_i|} \qquad (1)$$

For every media source, we calculated *Eq.(1)*[3] to all other media sources and examined which pairs appear the closest. This results in 5,852 non-zero closeness values among 77 media sources.

### Visualization

To visualize a map of media sources on the basis of closeness, we proceeded as follows. Firstly, to discount those users who may no longer use Twitter, we selected subscribers who actively use Twitter by only considering users who have posted at least 10 tweets in last three months and have at least 10 followers and followees, respectively. This leaves us with 3.4M users and 12M follow links. These active users show a different subscription pattern in that individual users subscribe to multiple media sources across diverse topics (50% of users subscribe more than 4 topics).

The media sources were then positioned on a map by the Force-altras algorithm in Gephi[4] (Figure 1). This algorithm optimizes the placement of nodes (i.e., media sources) based on relationship strengths (i.e., the number of common subscribers).

Because there was at least one common subscriber for most pairs of media sources, the closeness values were also non-zero, including a full mesh-like network. To unclutter the map and show only the most relevant relationships, we retained the two strongest outbound relationships for each media source [2], and obtained 154 media relationship edges. While the closeness metric we used defines a symmetric relationship, by selecting the strongest links from each media source, the map can show a directional relationship.

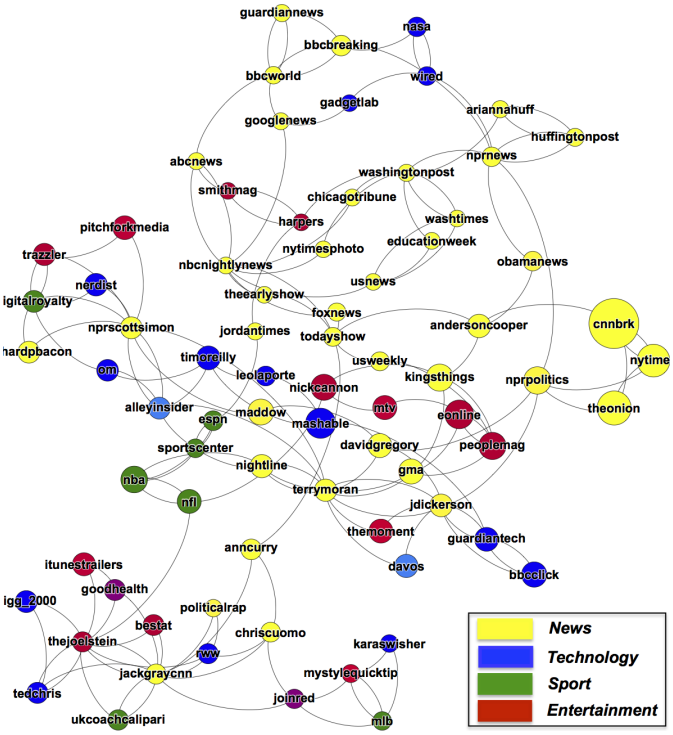


**Figure 1. Map of traditional media**

## RESULT

Here we first discuss the visual structure of the map, and then answer two questions: 1) who competes with whom; and 2) where are the hidden opportunities in the media industry.

### Map of traditional media

In Figure 1, nodes represent media sources and links reflect co-subscription relationships. Node size is proportional to the log of the number of subscribers, and color indicates genre value.

[3] We also tested other variations such as $\frac{|A \cap B|}{|A|} \cdot \frac{|A \cap B|}{|B|}$ which are symmetric similarity measures and $\frac{|A \cap B|}{min(|A|,|B|)}$, which are a symmetric similarity measures. In this work, we reported the result by *Eq.(1)* because the other metrics showed the similar results.

[4] `www.graphiviz.org`

**Media relationships** In the map, the proximity of nodes of the same color suggests that subscriptions are mainly motivated by topical interest. We observe a large cluster of tightly connected news media sources (yellow) in the upper part of the map, and smaller clusters in the center of the map: four SPORT media sources (espn, sportcenter, nba, and nfl) in green, four ENTERTAINMENT media sources (nick cannon, mtv, eonline, and peoplemag) in red, and five TECHNOLOGY media sources (timoreily, om, mashable, alleyinsider, and leolaporte) in blue. These smaller clusters have all strong intra-cluster connectivity but their connectivity structures differ– the ENTERTAINMENT and SPORT clusters unfold in a shape similar to a "long chain" (suggesting different views), while the TECHNOLOGY cluster is star-shaped (suggesting more homogeneous and potentially redundant views).

Besides being driven by topics, media subscribers were also influenced by geography of media sources. For instance, on the top center of the map, there is a clique of UK-based media sources (e.g., bbcworld and guardian news). Furthermore, washingtonpost and washtimes are present and are closely connected to each other because of geography (despite offering different political leanings in their reporting).

A third and final element that influences media subscriptions was the presence of very popular news reporters which are not media outlets in the traditional sense (e.g., terrymoran and gma or arianahuff and huffingtonpost). We find several cases where two media outlets become connected by their presence.

**Bridging clusters** Figure 1 depicts a picture of media relationships and offers several take-away messages. While the Twitter subscription network encompasses different relationships, it seems that users preferentially subscribe to media sources based on their topical interests. Yet, there are clusters that host more than one topic, especially media sources positioned in the center of the map (e.g., todayshow, mashable, nickcannon, kingsthings, and terrymoran). For example, the TECHNOLOGY cluster (left side) is indirectly connected to the ENTERTAINMENT cluster through certain key brokering nodes. These brokers include Larry King (kingsthings) who links the ENTERTAINMENT cluster to the NEWS cluster, and few "star" journalists (e.g., jdickerson, maddow, or jackcraycnn) who connect media outlets that revolve around a variety of topics (e.g., cnnbrk, nytimes, and the onion positioned in right center of the map). Next, we spell out the important function of these key brokering nodes.

### Who competes with whom

To determine the role of brokering nodes, we need to identify communities in our map and see which nodes serve as a broker *across* multiple communities. To this end, we identify groups of media outlets that are tightly connected together[5]– nodes in the same community have the same color.

Figure 2 shows seven communities, each of which hosts media outlets that compete with each other. Media outlets speak to similar audiences (belong to the same community), if they are based in the same or nearby geographic regions ($c1$ groups outlets in two specific regions); they cover similar topics ($c2$ is about NEWS, $c3$ is about ENTERTAINMENT and SPORT, and part of $c4$ is about TECHNOLOGY); or they are simply popular ($c5$ groups well-known media sources).

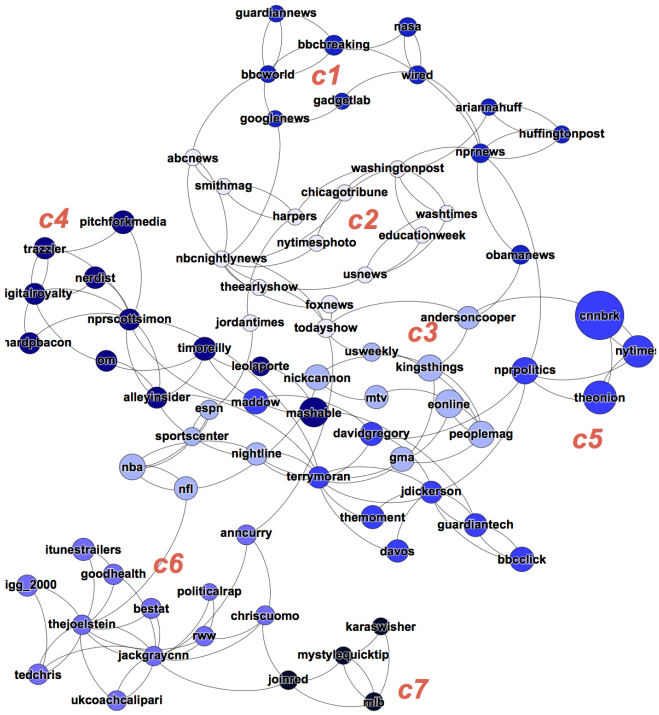


**Figure 2. Map of traditional media (color reflects the community).**

Unlike in the case for $c1$, $c2$, $c3$, and $c5$, the dynamics in $c4$, $c6$, and $c7$ cannot be explained by geography, topics, or popularity alone. Instead, what these communities have in common is that they have rich oppurtunities *in structural holes* [4]. In network parlance, structural holes are missing relations that inhibit information flow between people, and people who bridge structural holes (gaps between discrete groups of people) build considerable advantage in early exposure to diverse information; they get advantage in receiving diverse ideas with which they can create value. In our map, thejoelstein in $c6$ is a case in point, as the user fills a structural hole in that he bridges information flow between disconnected media sources (e.g., digg_2000 and bestat).

A broker can play a powerful role as advertiser, collaborator, or information provider since it reaches and talks to diverse audiences who are willing to receive heterogeneous information and opinions. Through a simple map based on a media subscription pattern, we unveil their presence and highlight their ability of brokering across diverse news audience.

### Where are the hidden opportunities

To identify brokers in a quantitative fashion, we compute Burt's constraint for each node in the graph. This measure reflects the extent to which a node occupies a *brokerage* position [4]. Burt's constraint is high (less brokerage opportunities) if one has mutually stronger related (i.e., redundant) contacts, whilst it is low if one has contacts that span different clusters.

[5]We use the "Louvain" community detection algorithm `http://perso.uclouvain.be/vincent.blondel/research/louvain.html`

| Rank | Degree centrality | Mentions | Burt's constraint |
|---|---|---|---|
| 1 | cnnbrk (987,529) | mashable (403,909) | jackgraycnn* (107,608) |
| 2 | theonion (519,715) | kingsthings* (313,006) | nbcnightlynews (9,417) |
| 3 | nytimes (478,197) | leolaporte* (98,054) | thejoelstein* (110,641) |
| 4 | mashable (403,909) | richardpbacon* (159,182) | nprscottsimon* (160,217) |
| 5 | eonline (361,086) | andersoncooper* (204,474) | terrymoran* (172,048) |
| 6 | nprpolitics (321,791) | nerdist (180,082) | washingtonpost (20,535) |
| 7 | peoplemag (317,354) | maddow (279,408)* | todayshow (70,293) |
| 8 | kingsthings* (313,006) | jackgraycnn* (107,608) | nprnews (87,652) |
| 9 | nba (301,037) | cnnbrk (987,529) | jdickerson* (183,787) |
| 10 | nickcannon (282,407) | timoreilly* (217,846) | nprpolitics (321,791) |

**Table 2. Top ranked media based on degree centrality, number of mentions, and Burt's constraint score. Journalists are marked with the * sign. The sheer number of active followers are denoted in parentheses.**

Table 2 compares the top 10 media outlets ranked: by degree centrality (total number of active subscribers); by number of mentions; and by lowest Burt's constraint score. Ranked by degree centrality, established media sources like CNN, Mashable, and NBA make up the top list. Ranked by number of mentions, popular journalists like Mashable, Larry King, and Anderson Coopoer make up the top list. Surprisingly, the Burt's constraint ranking offers a different picture: it shows only 10% overlap (only nprpolitics is in both lists) with the previous list and mainly consists of media journalists (like Jack Gray, Joel Stein, and Scott Simon) and of well-known news outlets (like nbcnightlynews and today show).

Those who take brokerage opportunities in the media landscape were mainly journalists, who do not necessarily have large audience. To understand how they differ from popular media outlets, we collected all tweets posted by the top 10 most popular nodes–highest degree centrality and by the top 10 nodes with lowest Burt's constraint. Figure 3 shows the word clouds that characterize the two groups.

Popular media outlets (Figure 3(a)) talk about politics (e.g., #superdelegates and gabfest), health (e.g., healthcare and Alzheimer's), and news media (e.g., nightlife and msnbc), and also post negative emotion words (e.g., unsolved, reduced, and pretend). On the other hand, journalists (Figure 3(b)) talk about diverse topics like basketball (e.g., kobe, orl, and nuggets), home (e.g., WilkinsonPlus), arts (e.g., Blab–comics anthology and ArtsBeat), entertainment (e.g., idol's, #dwts–Dancing With The Start, Bachelorette–TV show, Pattison–actor of Twilight, and Britney), politics (e.g., g.o.p.), business (e.g., economix and dealbook–financial news service), and humor (e.g., #fishtime). The diverse set of topics expressed in informal ways seems to be what distinguishes brokers from popular media outlets, yet it needs to be investigated further in the future.


(a) Degree centrality-based (b) Burt's constraint-based

**Figure 3. Tag clouds of tweets by popular media**

## DISCUSSION

Our result offers important implications, including:

**Analyzing big data through visualization.** It is challenging to make sense of large quantities of data. We have shown that one effective way of analyzing such data is to visualize it. This methodology is effective in the sense that simple data mapping has allowed us to find out who competes with whom and, more importantly, which are the hidden opportunities in the publishing industry and who is tapping into them.

**Tracking publishing industry.** We have shown how up-to-date maps of the publishing industry can be created free of charge from publicly available data. These maps suggest that market segmentation by demographics, which is the most commonly used way of running marketing campaigns, is not the only way of reaching new readers. Diverse audiences who cannot be segmented by demographics are generally reached by very popular and highly-reputable journalists, so it would be beneficial for marketing campaigns to also rely on those special individuals to increase exposure.

## CONCLUSION

This paper presented a study of the media industry based on a simple subscription graph derived from Twitter data. With limited cost, an up-to-date map can be obtained that can offer key insights for building new marketing strategies in the continuously changing news media industry. We are currently working on developing different ways of connecting news readers who share the same interests and yet do not subscribe to the same media outlets. By providing a platform that connects such users, we hope to facilitate the exchange of opinions among news readers.

**Acknowledgment** This research was partially supported by the EU SocialSensor FP7 project (contract no. 287975). J. An is supported in part by the Google European Doctoral Fellowship in Social Computing. M. Cha is supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2011-0012988).

## REFERENCES

1. J. An, M. Cha, K. Gummadi, and J. Crowcroft. 2011. Media landscape in twitter: A world of new conventions and political diversity. In *Proc. ICWSM*, 2011
2. J. Bollen, H. V. de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva. Clickstream data yields high-resolution maps of science. *PlosONE*, 2009
3. S. Aral and M. V. Alstyne. The Diversity-Bandwidth tradeoff. *American Journal of Sociology*, 2011.
4. Burt, R. S. Structural holes: The social structure of competition. *Harvard University Press*, 1992.