Consumers and Suppliers: Attention asymmetries. A Case Study of Aljazeera's News Coverage and Comments

Sofiane Abbar[1], Jisun An[1], Haewoon Kwak[1], Yacine Messaoui[2], and Javier Borge-Holthoefer[1]

[1]Qatar Computing Research Institute, HBKU , Doha, Qatar, {sabbar,jan,hkwak,jborge}@qf.org.qa

[2]Al-Jazeera Network , Doha, Qatar , yacine.messaoui@ajlazeera.net

ABSTRACT

In the last decades researchers have devoted a fair amount of effort to understand the news geography, i.e. a description of the general patterns for which countries are presented in which other countries' news. These efforts presume that published opinion is a good enough proxy to infer such perception; indeed, that is the case when no other evidence is available. In this work we uncover a new point of view, that of consumers, relying on a rich data set related to the comments posted by users on the largest news media organization in the Middle East region, namely Aljazeera Network. Each comment comes with a body content, a unique identifier of the user, and her IP address, which makes it possible to infer the countries from which comments are posted. All in all, our analysis encompasses over 20,000 articles and more than 2 million comments posted by 90,000 unique readers. Such a rich data set allows us confronting for the first time –to the best of our knowledge– the producer-consumer standpoints.

Keywords

foreign news, news geography, news supplier *vs.* news consumer

1. INTRODUCTION

Despite the communication deluge of social media, traditional news channels are still the main sources that most people rely on to access information [9]. Presumably, the choices by domestic news media regarding the selection of international news is shaping individuals' perception about foreign countries. Such assumption typically underlies news coverage studies, where it is taken for granted that news media, ubiquitous and regular, dominate opinion formation among citizens –above or together with other factors such as diplomacy, economic relations abroad, etc. [4]. The GDELT (Global Data on Events, Location, and Tone) project [10], a large-scale news coverage dataset that monitors news media in over 100 languages from the whole world, has strengthened such ideas, widening the breadth and depth of studies on news geography [7, 8].

And yet, whether such conclusions about *individuals* are true or not remains roughly tested. Researchers can gather data about news *production*, but it is typically much harder to actually find out information about news *consumption*: polls and surveys have served as an awareness proxy [13], but no data about readers' interest in this or that particular news is brought forward.

With these limitations in mind, a question is raised about a possible asymmetry between the interests of suppliers and consumers: how do citizens read news and perceive the world? Does such perception actually depend on how journalists produce news? It is known that strong regionalism has been typically found in news media production [6, 12, 8]. However, such regionalism does not guarantee that readers also pay more attention to events that occur in close countries. These questions are ever more relevant, as the trend towards the globalization of news media makes a systematic analysis even more interesting: readers of major news networks today are not confined to a single country any more; rather, they come from all around the world.

In order to address the previous inquiries, we exploit in this work a large collection of fine-grained data about readers' behavior: who reads what, and from where. With data from the Aljazeera Media Network at hand, we aim to reveal *attention asymmetries* at two levels: (i) consumer-consumer asymmetries, which inform us about mutual attention between pairs of countries. At this level, we expect that asymmetries will arise closely following the pattern of international news coverage: most of readers' attention is focused on those countries for which more news are produced, without evidence for reciprocity; and (ii) consumer-supplier asymmetries, which provide a valuable insight for news professionals on audience behavior and expectations: to what extent does the audience match the news supply? To answer these questions, we proceed to construct a "network of attention", linking readers' countries to reported-on countries. A systematic analysis of the resulting structure allows us to disentangle some interesting facts, regarding the "level of dissonance" between a news supplier and its consumers. Then, we measure to what level each country is over- or under-represented in the news, according to the level of interest it gathers from the audience.

Bringing forward a new data aspect, we add a new and fundamental dimension to the study of international news coverage, which opens the path for a more detailed understanding of audience interests and behavior.

2. DATASET

The data we use in this study is provided by Aljazeera Media Network which is the largest news organization in the MENA (Middle East and North Africa) region with more than 5 different TV channels including Aljazeera America, Aljazeera Arabic, Aljazeera Balkans, Aljazeera English serving millions of users from around the world. Each TV channel has its corresponding and dedicated news website, and our data set is actually collected from Aljazeera English website[1]. The data collection consists of 22K articles and 2.3M comments posted by 90K distinct users from 214 different countries. Comments cover articles posted between 21/11/2012 and 21/01/2015.

Comments. Comments come with a user email (unique user identifier), the IP address of the device used to post the comment, and the content of the comment itself. As we are interested in this study in the geo distribution of consumers, we use Python GEOIP[2] library to map each IP address into a country. We then use a heuristic to map users to their countries i.e., each user is mapped to the country from which she posted most of her comments.

Articles. Articles come with a title, a content, and a set of comments posted on them. We use Open Calais[3], an online named entity recognition and topic modeling web service to request for each article the set of named entities it mentions such as *Countries, Cities, Persons, Organizations* along with the generic topics the article refers to. Such topics include but are not limited to: *Politics, War Conflict, Social Issues, Sports, Disaster & Accident*, and *Environment*. In addition, we use Google Analytics[4] service to request the total number of visits (measured as the number of sessions) of each article in our data set.

In order to avoid the problem of abnormally long discussion threads –which could bias the geo-distribution analysis–, we determine that a user is accounted only once in a given comment thread, regardless of the number of posts she authors in it. This reduction step is important as it measures the number of individuals posting comments from a given country, rather than the number of comments posted from that country.

3. METHODS

In this section we outline only those methods with which the reader may be more unfamiliar, i.e. those related to network construction and analysis. Extensive related readings may be found in [2, 3].

3.1 Building the "network of attention"

The data from Aljazeera can be suitably represented as a network which encodes two complementary aspects: the producer and the consumer sides. Nodes in the network represent countries, and a directed link from node i to node j implies that audience from i has read some news in AJE about j. More importantly, each country (node) bears an associated quantity a that stands for the amount of attention they receive from AJE, understood as the proportion of articles AJE has released about it (producer aspect). On the other hand, each directed link from i to j has an associated weight (i, j, w), which proxies the fraction of attention the audience from i devotes to news about j (consumer aspect). As such, the total strength of out-going links is normalized, $s_i = \sum_j w_{ij} = 1$. Self-loops (i.e. reader attention on its own country) are not considered.

3.2 Network backbone extraction

Networks can be described from different levels of analysis. At the *micro* level, the focus lies on single nodes and their specific positions within the overall structure; this level can be described in terms of node degree, strength or clustering coefficient, among other metrics. At the *macro* level, the focus shifts to the aggregation of those metrics and the properties of their distribution. Between these two extremes, we have a third level of analysis, the mesoscale, which aims to account for the complexity of networks between the position of individual nodes and the relational properties of the collectives they form. It is at this level where reduction techniques like backbone extraction operate.

Network backbone extraction refers to the filtering techniques aimed at uncovering the relevant information; in general, such techniques aim at pruning the links of a network, keeping only those which are statistically relevant. Ideally, the reduced structure is computationally more tractable while it retains most of the interesting features of the original one. Here, we apply the backbone extraction proposed in [11] to the network of attention (see previous subsection). In their work, Serrano et al. exploit the empirical trend by which link weights are heavily fluctuating, i.e. only a few links carry the largest proportion of the node's total strength.

3.3 Overlapping community detection

A different approach to the analysis of the *meso* level of a network is that of community detection. Here, we exploit fuzzy or overlapping community detection, which means that as a result nodes (countries) will be grouped in relevant modules, *and* at the same time they can belong simultaneously to more than one group.

Among the many available techniques, we exploit a well established one: that of edge partitioning via modularity optimization [1]. The idea behind this approach is that community detection should try to classify edges. In doing so, since nodes may be attached to links that belong to different communities, we have a way of quantifying to level of implication of a node with a given module or community (i.e. a fraction of its edges devoted to that module). The algorithm to perform such analysis is publicly available in most popular programming languages[5].

4. INITIAL FINDINGS

In this section we present some general findings result of a high level analysis of our data set. In the following, producer's attention toward a given country is measured in terms of the number (or proportion) of articles devoted to that country. Consumers attention is measured as the number of unique users commenting on articles about the country. The decision to comments to measure consumers' attention rather than visits is motivated by two reasons: (i) The geographic distribution of visits retrieved from Google Analytics dashboard is at the aggregate level of the entire website; which makes it impossible to figure out the origin of visits per article. (ii) Commenting is considered as a more engaging action compared to visiting. Obviously, a user who posts a comment on an article is more engaged with the content of the article than a user who just browses it. Furthermore, unlike commenting, the browsing behavior presents more noise due to some exogenous factors such as the placement of the article in the website, mis-clicks, etc.

Using our data set, we compute for each country the following four different scores:

- Number of visits (sessions) from the country (*VisitsFrom*). This score is requested from the Google Analytics page of Aljazeera English website for the period for which we have collected the comments.

- Number of comments posted from the country (*Comments-From*). This score is directly inferred from our data set by mapping user IP addresses into countries. We recall that due to the normalization step we undertook, multiple comments

[1] http://www.aljazeera.com

[2] http://pythonhosted.org/python-geoip/

[3] http://opencalais.org

[4] http://www.google.com/analytics/

[5] http://barabasilab.neu.edu/projects/linkcommunities/

posted by the same user on the same articles are counted only once.

- Amount of articles in which the country is mentioned (*ArticlesAbout*). This score is computed using the countries identified by Open Calais. If there are n countries mentioned in an article, then the article equally contributes to all mentioned countries with a score of $1/n$, thus making sure that $ArticlesAbout$ scores of all articles sum to the total number of articles in our data set ($22K$).
- Amount of comments posted on articles about the country (*CommentsAbout*). This score is calculated in the same way as $ArticlesAbout$. If a comment is posted on an article that mentions n countries, then the comment contributes equally to the $CommentsAbout$ scores of these countries with $1/n$.

Table 1: Spearman's correlations

	VF	CF	CA	AA
Visits From	1.000	**0.967**	0.771	0.754
Comments From		1.000	0.735	0.722
Comments About			1.000	**0.980**
Articles About				1.000

Table 1 reports the Spearman's correlation scores between the four different country distributions. While all correlations are above 0.7, one can easily see that the highest correlations are achieved between CommentsFrom-VisitsFrom on one side and CommentsAbout-ArticlesAbout on the other, with scores of $\rho_1 = 0.967$ and $\rho_2 = 0.980$ respectively. ρ_1 hints at the fact that the number of comments coming from a country is a good proxy to the number of visits from that country, assuring us robust results when we exploit data about comments on specific articles; more importantly, ρ_2 suggests a *macro* trend, which indicates that at the global scale, supply and consumption are aligned: the number of comments posted on articles about a given country is proportional to the number of articles published about that country. As we shall see later, this general pattern is not so clear at the *micro* level.

5. CONSUMER-CONSUMER ASYMMETRIES

We tackle in this section the first research question which aims at accessing the consumer-consumer asymmetries. That is, we investigate whether readers from countries pay attention to a wide range of other countries or not. We also try to verify whether countries receive the same attention or not.

5.1 Who pays attention to whom

Following the method described in section 3.1, we built the AJE network of attention in which nodes are countries and edges (i, j, w) represent the proportion (w) of people from country i commenting on articles about country j. As expected, it was difficult to infer any valuable conclusion from a dense graph with almost each country connected to all other countries. Thus, we conducted a backbone extraction (section 3.2) to keep only statistically significant edges for each node (i.e., remove noisy edges). The resulting network –composed of 1,430 edges– is presented in Figure 1.

Many observations can be made here. (i) We clearly see that all countries do not get the same amount of attention. Actually, the attention distribution is very skewed with a handful number of countries concentrate most of the attention in the network. These countries –in the core of the network– are: United States, Israel, United Kingdom, Syria, Iraq, Russia and Egypt. This list of countries summarizes by itself to a large extent numerous major events covered by AJE in the last three years, such as the Arab spring, the Iraqi civilian war, and the Israeli-Palestinian conflict. Beyond mere visualization, Figure 3 shows that countries in the core have many of their out-links reciprocated (they receive attention back), that's why nodes with high degree have also high reciprocity; but those in the periphery are *not* reciprocated, i.e. they pay attention to other countries, but those countries do not pay it back. Reciprocity for each node has been measured as $r = k^{\leftrightarrow}/k_{out}$, where $k^{\leftrightarrow}$ is the number of reciprocal links, and k_{out} is the out-degree of the node. A correction factor has been applied to r afterwards (adapted from [5]). (ii) In terms of link analysis, the list of top 10 heaviest links (people from country i interested in country j) is found to be: $\{US \rightarrow Israel,\ Canada \rightarrow US,\ US \rightarrow Syria,\ US \rightarrow Iraq,\ US \rightarrow Egypt,\ UK \rightarrow US,\ US \rightarrow UK,\ Australia \rightarrow US,\ US \rightarrow Russia,\ US \rightarrow Palestine\}$. Notice that this list is sorted in the decreasing order of the weights of links. Surprisingly, we can see that most of the heaviest links are outgoing from the United States toward different countries in the inner circle of the network. This is an interesting observation for a news media organization that is based in the Middle East. While the backbone network shows that attention could be derived by other factors than the pure geographical proximity, such as geopolitics, economics, and human migration, a reduced version of the backbone in which we

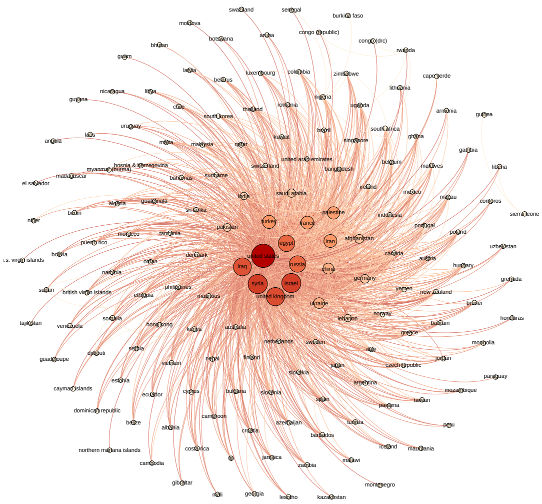


Figure 1: AJE consumer-consumer network backbone. With only the statistically relevant links left, the network shows a clear star shape, with a few core countries receiving most attention, and the rest of them laying at the periphery.

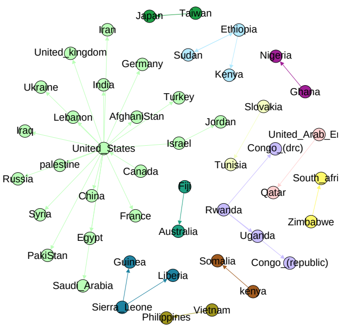


Figure 2: Connected components of countries that pay the most attention to other countries in the Backbone network.

keep for each country i, the country j that pays the most attention to it, reveals the existence of many cases where countries pay attention to their immediate neighboring countries. Figure 2 presents different connected components besides the one stared by the United States. We can see for instance that the country that pays the most attention to Qatar is its neighbor United Arab Emirates. The same observation stands for Zimbabwe and South Africa, Kenya and Somalia, and Japan and Taiwan.

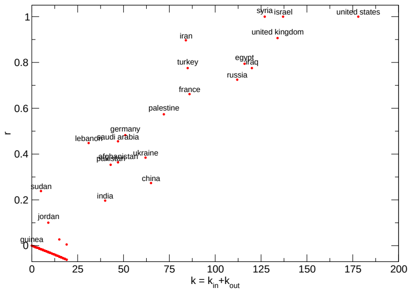


Figure 3: Scatter plot confronting reciprocity (r) to node total degree ($k = k_{in} + k_{out}$). Labels for those countries with $r < 0.1$ have been removed, to ease visualization.

5.2 Community structure

Overlapping community structure is determined in an unsupervised manner via the algorithm outlined in section 3.3. With the output of 14 modules at hand, it is possible to conduct different levels of analysis. Initial inspection tells us that we have an heterogeneous collection of communities, with sizes oscillating from over 180 countries (roughly informative, as almost the whole set of countries is included) to small clusters of size 4, which allow some interpretation. Clearly, the main driver for the formation of smaller modules is regionalism: such is the case of the grouping {Samoa, Tuvalu, Bermuda, Timor-Leste, United States} (Bermuda is, of course, far from the Pacific Islands, but under the influence of the same main actor, United States); or a module of size 15 in which we find countries grouped around the MENA region (with Egypt, Israel, Saudi Arabia or Qatar in them) along with United Kingdom and Nigeria, which hints at the fact that these countries are connected in terms of readers' interests and comments. (Note that countries can belong to more than one community, thus the sum of elements in each module is greater or equal than the total number of countries).

Relevant to news coverage and international relationships, our analysis gives room also for other insights. For example, Figure 4 shows those countries which participate in more communities (in decreasing order; white color indicates no participation). Noticeably, United Kingdom appears to have a significant role in all communities; India, in the second place, is present also in all but 1 of the modules (precisely the one dominated by countries in the MENA region). Interestingly, United States –the country most commented and written about– participates only in 2 communities: the one reported above (along with some Atlantic and Pacific Islands), and the most general one (with 182 countries in it). This result highlights the outstanding role of United States in the audience attention: its influence is so uniformly widespread that it fails to be significant *except for the most general module*.

6. CONSUMER-SUPPLIER ASYMMETRIES

In this section we focus on asymmetries between the news suppliers (AJE) and its consumers (readership). The main goal is to look more closely at the attention level of dissonance between these two standpoints.

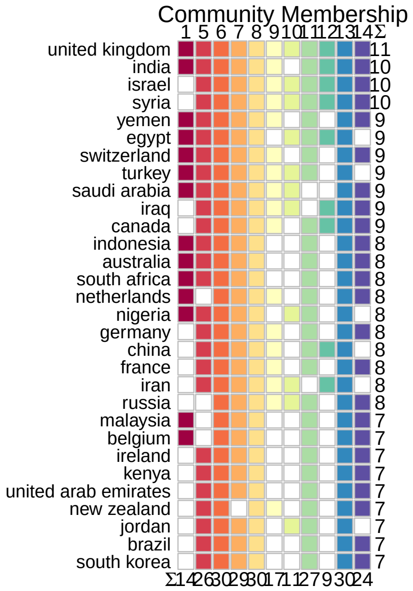
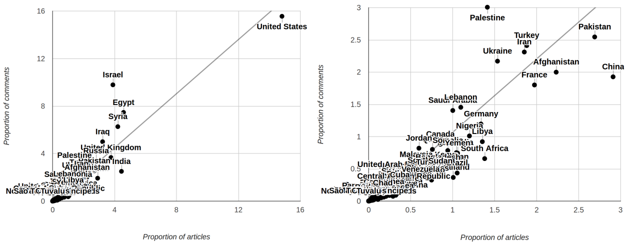


Figure 4: Overlapping community structure: participation of countries in different modules.

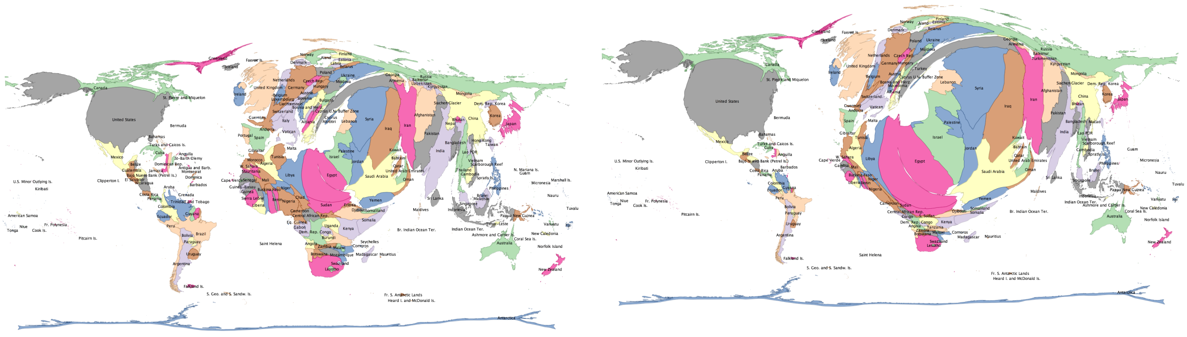


(a) (b)

Figure 6: Production-consumption scatter plots.

As a first attempt to access the differences between the attention level the supplier (AJE) and its consumers (readership) devote to each country, we produced two cartograms of the world map in which the territory of each country is reshaped to match a given score. Figure 5 puts side by side the supplier's (left) and the consumers' (right) cartograms. In the supplier cartogram, scores reflect the supplier's attention toward each country measured as the number of articles posted about that country. In the consumers cartogram, scores reflect consumers' attention in terms of comments posted on articles about each country. On visual inspection, one can easily see that both supplier and consumers attention focus goes to the Middle East region, which is over-represented in both maps. Yet, one can see that countries such as Egypt, Israel, and Syria are definitely catching more attention than what is expected.

In order to allow a more careful analysis of the level of dissonance between supplier's and consumers' attention, we generate for each country a pair of coordinates $(s_{attention}, c_{attention})$, we then plot all countries on a bi-dimensional space in which x-axis indicates the supplier's attention and y-axis indicates consumers' attention. Figure 6 represents a scatter plot confronting comments (consumers' standpoint) and articles (supplier' standpoint) for each of the considered countries. Both panels (top, bottom) are the same, but the lower one zooms in the figure to appreciate some details. A shallow inspection allows us to recognize several outliers, i.e. the amount of articles *and* comments are outstanding, if compared to



(a) (b)

Figure 5: AJE cartograms. (a): country size is proportional to the amount of articles from AJE website about it. (b): a country's size is proportional to the amount of comments that an article about it has elicited.

the rest. These are generally countries in the MENA region (Israel, Syria, etc.), except for the United States. If we focus on the comparison between consumption and production attention, we observe that the general trend is, for a given country, to have less comments compared to the number of articles about it. This is particularly surprising for countries like India and China, which receive a significant attention from news producers –and gather $\approx$ 30% of the global population and represent the 1st and 3rd economies of the world in terms of GDP. Additionally, many African countries seem to have less comments than expected.

7. CONCLUSIONS

In this contribution we focus on a new layer of information of great potential to journalists and data analysts. Digital traces from the audience of mass media networks can be exploited to attain unprecedented, article-level details about news consumption behavior. In particular, we have analyzed a rich dataset from the Aljazeera Media Network (English website) to provide evidence about asymmetries, regarding both inter-consumers and supplier-consumer attention. Our results suggest that AJE audience around the Globe is heavily biased towards most written-about countries, which fits well previous research regarding the role of media as a main actor driving population's opinion formation. However, we uncover also supplier-consumer asymmetries (countries for which audience interest is greater than news production about them, and viceversa) which open the path to a new level of actionable insight for media networks, in the quest to adjust or promote certain contents in their agenda.

Beyond these achievements, we are aware that this represents an initial study, provided that we bring forward a single source of data, in a single language, etc. While tools like GDELT guarantee a wide range of possibilities of studies *on the production side*, so far the audience aspect remains undisclosed. And yet, the methods presented here are valid and useful in general –beyond the particularities of AJE website.

We envisage promising directions of future research. Besides deepening our understanding in the news geography, analysis can be oriented towards more specific items, such as consumers' attention to given topics, supply/consumption adjustments in time (longitudinal study), etc. Along these lines, fine-grained audience analysis, above and beyond traditional survey data, will increasingly become a focus of research attention.

8. REFERENCES

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010.

[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.

[3] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[4] J. Galtung and M. H. Ruge. The structure of foreign news the presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

[5] D. Garlaschelli and M. I. Loffredo. Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):268701, 2004.

[6] G. Gerbner and G. Marvanyi. The many worlds of the world's press. *Journal of communication*, 27(1):52–66, 1977.

[7] H. Kwak and J. An. A first look at global news coverage of disasters by using the gdelt dataset. In *Social Informatics*, pages 300–308. Springer, 2014.

[8] H. Kwak and J. An. Understanding news geography and major determinants of global news coverage of disasters. *arXiv preprint arXiv:1410.3710*, 2014.

[9] S. Leckner, U. Facht, et al. *A sampler of international media and communication statistics 2010.* Nordic Information Center for Media and Communication Research, 2011.

[10] K. Leetaru and P. A. Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, page 4, 2013.

[11] M. Á. Serrano, M. Boguná, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *ational academy of sciences*, 106(16):6483–6488, 2009.

[12] A. Sreberny-Mohammadi. The "world of the news" study. *Journal of Communication*, 34(1):121–134, 1984.

[13] W. Wanta, G. Golan, and C. Lee. Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism & Mass Communication Quarterly*, 81(2):364–377, 2004.