An Exploration of Social Identity: The Structure of the BBC News-Sharing Community on Twitter

Julius Adebayo, Tiziana Musso, Kawandeep Virdee, Casey Friedman, and Yaneer Bar-Yam

New England Complex Systems Institute

238 Main Street Suite 319, Cambridge, Massachusetts 02142, US

(Dated: July 31, 2013)

Abstract

Online social media are host to networks that influence the flow of news and other information, potentially altering the social dynamics of collective social action while generating a large volume of data useful to researchers. Mapping these networks may make it possible to predict the course of social and political movements, technology adoption, and economic behavior. Here we map the network formed by Twitter users sharing British Broadcasting Corporation (BBC) articles. Members of the network primarily "follow" members sharing articles in the same language, while English-speaking users further differentiate themselves into clusters by interest, political orientation, and location. Unlike the previously studied New York Times news sharing network the largest scale structure of the BBC network does not include a densely-connected sub-group of globally interested and globally distributed users, which we attribute in part to the BBC's history as an alternative source of local news in regions outside the United Kingdom (UK).

Tracing the flow of information in society can offer insights into the structure of social groups–who associates with whom–and the patterns of response to various events [1–4]. Here we map the structure of a network composed of individuals who share British Broadcasting Corporation (BBC) news articles on the social networking website Twitter. We find that the BBC global network separates into linguistically distinct regional groups. The largest of these groups reads the BBC in English, with subgroups focused on sports, UK news, and business, among other topics. Other major language groups observed include Spanish, Russian, and Arabic. Significantly, there is little evidence of a cosmopolitan audience that follows news from around the globe—headline articles in each language category typically pertain to the corresponding region. These results differ from those we previously obtained for the New York Times news-sharing network, which includes a substantial cosmopolitan group in addition to subgroups focused on local and national news [5]. The organization of the BBC network along linguistic lines is consistent with provision of customized foreign language services and the historical prominence of the BBC in regions where political constraints have inhibited the reliability of local news sources.

Social media platforms including Facebook and Twitter are providing an explosion of information about interpersonal communication because these platforms provide a record of ongoing of communication and response to information [3, 5, 6]. Increasingly, these platforms serve as filters for, or determinants of, the information that individuals are exposed to. As they interact on these platforms, individuals preferentially associate with others from whom they want to receive information. These associations can be characterized as links in a network, and the users as nodes. The substructure of the network can be identified through the presence of densely interconnected components: users who are likely to reflect common interests and other shared characteristics. Mapping network structure can therefore elucidate the factors that drive social association, including shared interest, geography, age, class, language, religion, race, or profession [7–10]. Differences in the structure of distinct news-sharing networks reflect differences in news coverage and the audience these news sources attract. The ability to analyze social relationships and map the structure of social communities enable an understanding of collective response to news events [11, 12].

Here we identify social groups among users sharing BBC articles on Twitter. The BBC is a semi-autonomous British public media corporation, with television, radio, and online services. The BBC provides both United Kingdom (UK) and international news, as well

as other forms of programming. The BBC reaches 239 million people globally, and its news coverage reaches over 80% of the UK adult population [13]. The international reach of the BBC stems from the BBC World Service, founded in 1932 as a purely English-language "Empire Service." In 1938, the BBC began its first foreign language broadcasts, directed at the Arab World and Latin America, in order to counter propaganda by European rivals [14]. The run-up to war and World War II itself prompted the BBC to expand into many other languages, including those of Nazi-occupied countries. Throughout its history, the UK government, BBC staff, and historians have cast the World Service as both an objective news source and a public diplomacy tool to convey the British experience to far-flung peoples [15–17]. Historical accounts suggest that the BBC gained global reputation for reliability [14]. The BBC World Service has responded to changing conditions by initiating or expanding broadcasts in languages spoken in places of geopolitical importance or national crisis: introduction of a Latin American Service in 1938, German in 1939, Arabic after the 1956 Suez crisis, non-Russian languages of the former Soviet Union from 1991, and Albanian in 1993 [17, 18].

BBC articles are currently being shared hundreds of thousands of times per week on Twitter. Twitter is a micro-blogging platform through which individuals can publish and read posts of up to 140 characters called "tweets" that may contain links to webpages [19, 20]. Users subscribe to receive tweets from other users, indicating that they read their tweets on a regular basis. This is called *following*. The follower relationship among users identifies links that comprise a social network. We can characterize users by the subjects of the articles they tweet. Each BBC Online article belongs to a subject category, which is the first sub-directory in the URL of the article shared. For example, an article with the URL: http://www.bbc.co.uk/sports/0/football/... belongs to the *sports* category. For our analysis, we labeled users by their interest, inferred from the category of article they tweeted the most. For additional dimensions along which to characterize users, we performed text analysis of user provided fields: location and self-description. Identifying users by interest, location, and self-description enables a characterization of emergent highly connected groups. Methodological details are provided in the Appendix.

We collected tweets containing BBC online article links over a seven day period, March 4 - 10, 2011, resulting in 489,878 tweets posted by 153,172 users. We considered only users who shared at least three articles from a particular category resulting in 10,347 users. We

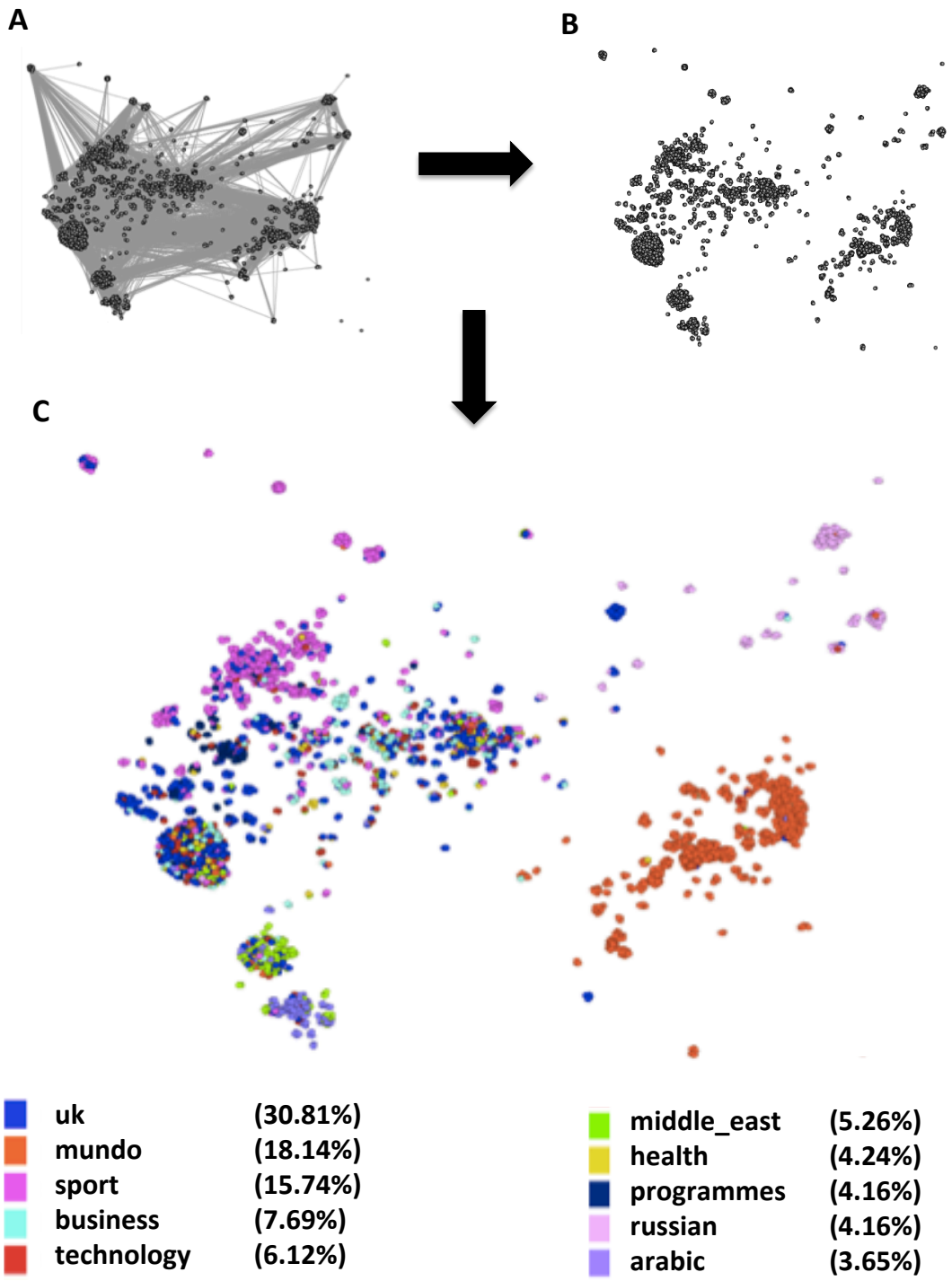

FIG. 1: (A) Network of Twitter users who share BBC online articles. Links are "follow" relationships between users. The network layout is obtained by pulling followers and followees close to each other, while pushing apart unconnected nodes. The long range links that are visible occur where relatively few links connect groups of nodes that are otherwise pushed far apart. (B) Same as A with links removed for clarity. (C) Same as B with user nodes colored according to the topic of articles they share the most.

generated a network of BBC article-sharers by representing each Twitter user in our sample as a node and each *follower-followee* relationship between users as an edge. The spatial layout of the network, shown in Figure 1, is determined by a force directed layout algorithm, which takes into account *only* the follow relationship between users. The layout algorithm pulls together users (nodes) that are connected together and pushes apart unconnected users, highlighting densely connected components of the overall network [21]. We performed a clustering analysis that quantitatively separates clusters into distinct subgroups [5, 22]. The clustering of the network further manifests, visually, the presence of distinct highly connected subcomponents.

Figure 1 shows the BBC news-sharer network with user nodes colored by interest. This labeling enables us to characterize the composition of the clusters. Topic colors are not randomly scattered; for part of the network, nodes of the same colors tend to be more closely associated spatially. We identified seven constituent cluster communities. Figure 2 shows the result of the characterization of these clusters. The seven clusters can be classified into four different groups primarily distinguished by language: English, Spanish, Russian, and Middle East. The latter includes Arabic- and English-speaking parts. We summarize the clusters found and their characterization as follows:

- The 'English group' consists of several clusters with users who primarily share English articles. About 70 percent of all users share English language articles. The English group further subdivides into four clusters, one focused on Sports and three focused on combinations of UK and business news.

- The 'Spanish group' consists of users reading articles from the Spanish-language BBC Mundo service. The BBC Mundo content in Spanish relates to Latin America and Spain. These articles address regional issues such as economics, political instability, and health care, with some global coverage as well. This cluster constitutes approximately 18 percent of the users studied. About 65 percent of users in this group are from Venezuela. Other countries represented include Chile, Mexico, and Spain.

- The 'Russian group' is primarily focused on BBC news articles in Russian. Articles in Russian generally cover the state of the Russian economy and government, as well as other international news. This cluster constitutes about 4 percent of the users studied. The group has a sparse substructure of multiple subgroups that are not well

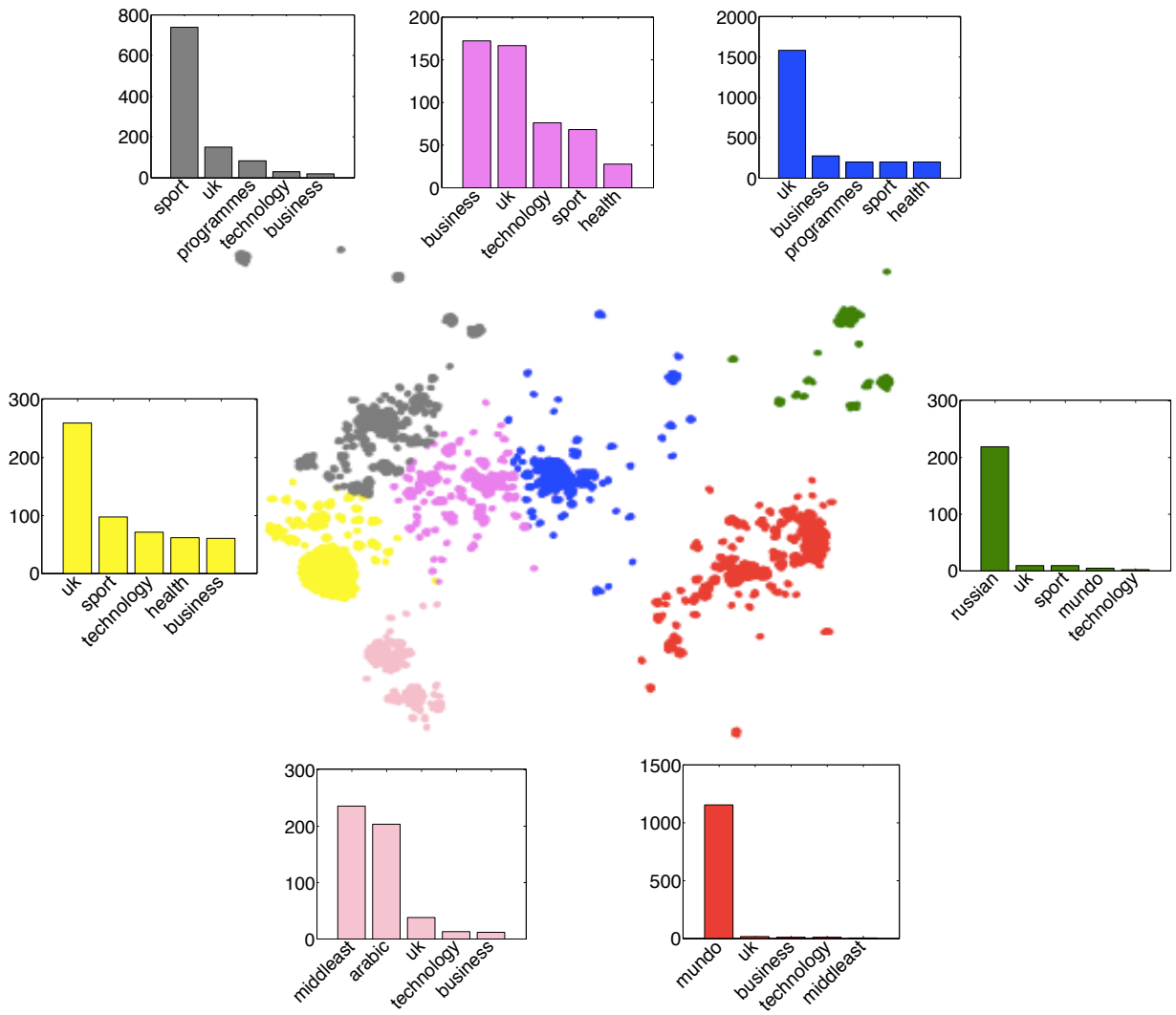

FIG. 2: Article subject frequency for article sharing in clusters of Figure 1. For each cluster the five most frequent user article sharing types are shown, with their frequency among users in the cluster. The correspondence between the graphs and the clusters is by color.

interconnected. Almost all users are from Russia, particularly Moscow. We also see users from Ukraine and Karachi, Pakistan.

- The 'Middle East group' consists of two clusters and is the only mixed language component. The first cluster is Arabic speaking while the other is English speaking. The latter is particularly tightly linked. The dominant subject category is the Middle East. Some users are in Arabic speaking countries such as Egypt and Kuwait, while

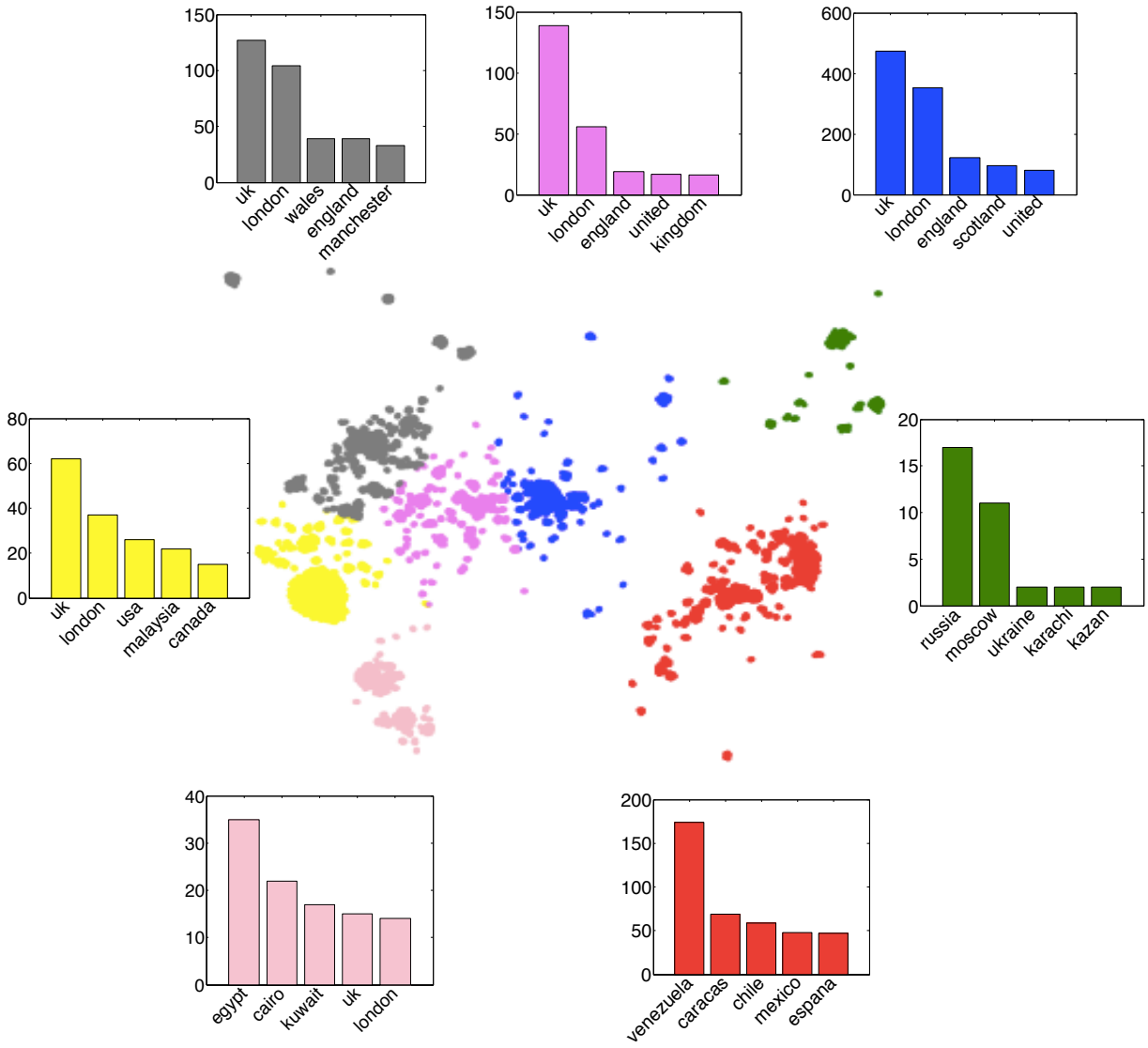

FIG. 3: Location frequencies for the clusters of the network in Figure 1. For each cluster, the five most frequent user locations are shown, with their frequency among users in the cluster. The correspondence between the graphs and the clusters is by color.

others are in the UK.

In order to clarify the substructure of the English speaking supercluster, we generated a network consisting only of users who tweeted mostly English-language articles. Figure 5 illustrates the structure of connectedness and the distribution of individual interests within this subnetwork. A clustering of the English subnetwork manifests four groups: one focused on Sports, two on National News, and the fourth on the Middle East. The fourth group is

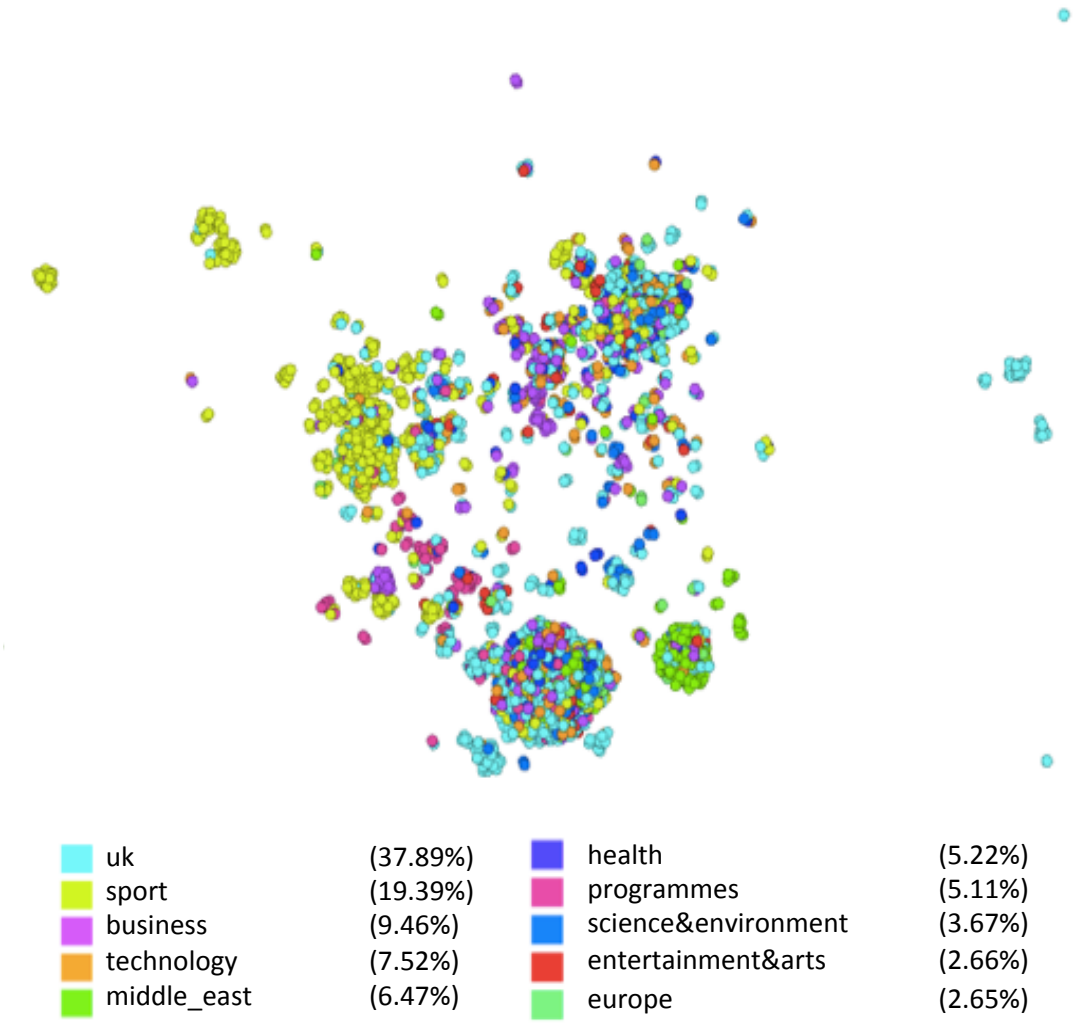

FIG. 4: Network of Twitter users who tweet BBC online articles in English, constructed similarly to Figure 1. User nodes are colored according to the topic of articles they share the most. Network edges have been omitted for clarity.

the English speaking part of the Middle East cluster previously identified. We summarize the clusters found in Figure 5 as follows:

- Cluster A is primarily focused on sports with other secondary interests such as UK news, business, programs, and technology. Users from this cluster are predominantly from the UK.

- Cluster B is primarily focused on UK news (news stories regarding England, Northern

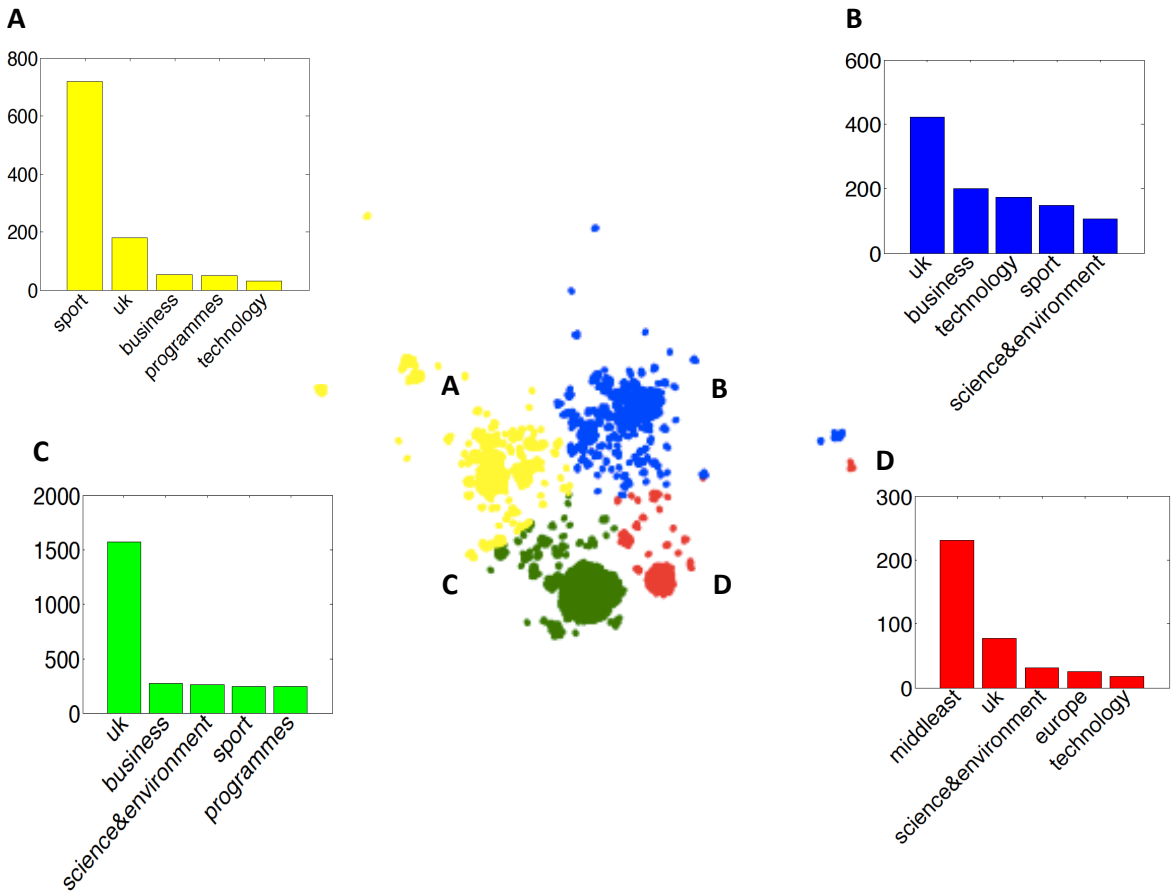

FIG. 5: Article subject category frequencies for article sharing in the clusters of the English group of Figure 4. For each cluster, the five most frequent categories are shown. The correspondence between the graphs and the clusters is by color.

Ireland, Scotland, and Wales) with secondary interests in business, technology, sports, and science & environment. A significant percentage of users in this cluster are located in the UK, with others indicating their location as United States, Canada, and Malaysia–countries with strong historical ties to the UK. However, many countries with large populations and strong British ties, such as India, are not represented.

- Cluster C is similar to Cluster B, though users exhibit a stronger topical focus on UK news. Other secondary interests include business, sports, programs, and science & environment. Users in this cluster are almost exclusively from the UK.

- Cluster D is focused on the Middle East. This is the English speaking portion of the Middle East group indicated in Figure 1. The BBC Middle East service publishes

international news stories about the Middle East region. A small percentage of users are also interested in UK news, science & environment, Europe, and technology.

Clusters B and C show a remarkable similarity of subject categories. To further distinguish between them, we performed a text analysis of the users' self-descriptions in each cluster. Figure 6 shows the relative frequency of the words obtained above a threshold frequency (0.007). The most commonly used words in Cluster B self-descriptions are "money," "business," "marketing," and "market," suggesting that members of Cluster B are highly business-oriented. The predominant words from Cluster C are "political", "student", "geek", "science", and "labour," indicating that members of this group include politically and technologically oriented individuals, students, and individuals who may be supportive of the UK Labour Party. We infer that these two clusters, B and C, may reflect polarization between two major sides of the UK ideological political spectrum associated with the dominant political parties, the Conservative party (Tory) and the Labour party respectively. This is similar to the liberal-conservative political divide observed for the US, in a mapping of the New York Times news sharing community on Twitter [5].

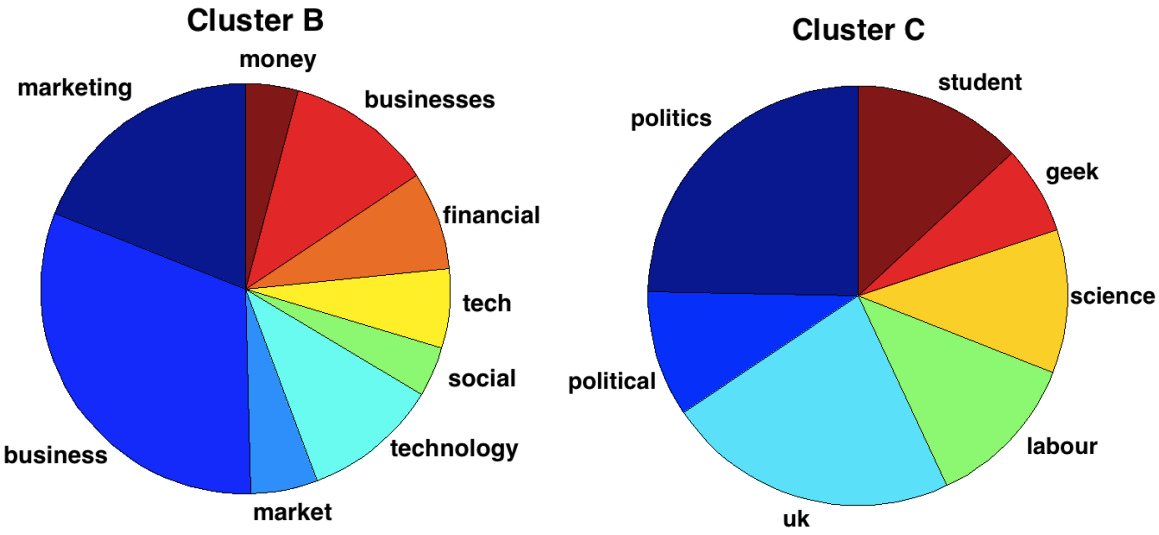

FIG. 6: Most frequent words contained in user self-description fields in sub-cluster B (left chart) and sub-cluster C (right chart).

Overall, the dominant structural determinants of the BBC article-sharing network are geography and language. The BBC serves, for the most part, isolated regional audiences

interested in regional news rather than a global audience interested in global affairs. Language, and therefore location, tend to be important barriers between groups. The cluster structure is likely to be rooted in the BBC's long history as an alternative local information source. One implication of the close link between language and geography is that BBC Online's Mundo and Russian coverage primarily serves foreign populations rather than immigrant populations resident in the UK. The sharp division between regional sub-networks highlights the BBC's need to provide content tailored to different audiences.

From the BBC news sharer network shown in Figure 1, we observe that a handful of countries–Chile, Egypt, Mexico, Pakistan, Russia, and Venezuela–account for most non-English article sharing. A potential explanation for pockets of BBC influence abroad is a historical absence of trustworthy or high-quality domestic news sources in these countries that led their inhabitants to seek regional news from an outside news service to which they have limited cultural ties. Indeed, there is a correspondence between these countries with historical periods of lower press freedom, high values (less free) on the world press freedom index, and the volume of article sharing in our sample. Chile, Egypt, Mexico, Pakistan, Russia, and Venezuela all have had notable limitations to press freedom in the past justifying a reliance on other news sources. There exist countries that have historically weak press freedoms but few tweets in the sample studied. In some of these cases governments censor news media and block social media. For example, even though the BBC has Chinese and Farsi language services, there is little sharing of articles in these languages in our sample—this can be readily understood because the governments of mainland China and Iran both block public Twitter access.

In summary, access to new social media platforms and the increased propagation of news has led to the emergence of new ways for individuals to organize, and created new channels for coordinating group collective action [5]. Knowledge of cluster and user traits can enable a better understanding of social dynamics and opportunities for news providing organizations. Insights gained from our work on BBC news sharing during a particular window of time reflects basic understanding about the BBC news organization and its audience globally. In recent years, financial pressures have led to closure of some of the BBC foreign language services [23, 24]. Information about the social networks formed may be helpful in developing corporate strategies.

Appendix

During March 4-10, 2011, we collected tweets that contain a URL from the "bbc.co.uk" (including redirects from "bbc.in" and other top level domains) domain, using the Twitter Application Programming Interface (API), resulting in 489,878 tweets posted by 153,172 unique users. Each user was labeled with a particular topic of interest corresponding to the specific genre of the articles tweeted. The topic genre is identified from the URL of the article shared. For example, a user sharing an article with URL: http://www.bbc.co.uk/sports/0/football/... would be in the *s*ports category. As of March 2011, BBC had more than 20 main categories including *sports*, *UK News*, *Mundo*, and *Russian*. We focused on the 10 most popular categories in our sample. Users were restricted to those who tweeted at least three BBC online URLs from the same category over the period in which the tweets were collected. We obtained the list of "follow" relationships among the set of users being studied.

Layout generation and clustering

The spatial layout of the network given in Figure 1 was determined solely by the topology of the network – based on a force-directed layout algorithm, which optimizes the layout so that nearby nodes push each other apart and edges pull the connected nodes closer [21, 25].

We applied the k-means unsupervised clustering algorithm on the two-dimensional coordinates of the nodes to divide users into 7 different communities (clusters) [22].

Analysis of geographical locations

The profile of each Twitter user contains an optional location attribute as an unstructured text field. We analyzed the text in this field to obtain the location of the users resulting in the name of the city, the name of the country or unidentified text [26].

Analysis of overrepresented words

The public profile of each Twitter user contains a short biography of the user as a self-provided text field. We analyzed this field to obtain information about the English speaking

groups. In general, these fields contain non-standard language or typos. Nevertheless, it is possible to capture the main characteristics of communities via aggregating text-based statistics over a group.

We employed a bag-of-words technique for analyzing the profile descriptions [27]. We lower-cased the texts, used the space character to split the text into words and did not carry out any lemmatization – the surface forms of the words such as "works", "working" were counted separately. We removed any word containing characters outside the English alphabet including é, è, and -. We also discarded stop words such as "and", "the", "of" from the text.

A characterization of a cluster can be obtained by the most over-represented words in the cluster. The bias of a word w, given two clusters c_1 and c_2 is $b(w, c_1, c_2) = p(w, c_1) - p(w, c_2)$ where $p(w, c)$, the prevalence of a term w in a cluster c, is the ratio of users in c who mention w in their bio fields.

---

[1] H. A. Simon, *The Sciences of the Artificial, 3rd ed.* (MIT Press, 1999).

[2] Y. Bar-Yam, *Dynamics of Complex Systems* (Westview Press, 1997).

[3] J. B. Thompson, *The Media and Modernity: A Social Theory of the Media* (Stanford University Press, Stanford, CA, USA, 1995).

[4] A. Briggs, P. Burke, *Social History of the Media: From Gutenberg to the Internet* (Polity Press, 2002).

[5] A. Herdağdelen, W. Zuo, A. Gard-Murray, Y. Bar-Yam, An Exploration of Social Identity: The Geography and Politics of News-Sharing Communities in Twitter, *arXiv:1202.4393* (2012).

[6] J. U. Henrikson, The growth of social media: An infographic, *Search Engine Journal* (2011).

[7] B. A. Nardi, D. J. Schiano, M. Gumbrecht, L. Swartz, Why we blog, *Commununications of the ACM*, **47**, 41–46 (2004).

[8] M. T. Poe, *A History of Communications: Media and Society from the Evolution of Speech to the Internet* (Cambrigde University Press, 2011).

[9] B. Winston, *A History of Communications: Media and Society from the Evolution of Speech to the Internet* (Routledge, New York, 2000).

[10] M. S. Granovetter, The strength of weak ties, *American Journal of Sociology*, **78**, 1360–1380 (1973).

[11] T. Heverin, L. Zach, *Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in the Seattle-Tacoma, Washington, Area* (ISCRAM, 2010).

[12] J. P. Bagrow, D. Wang, A.-L. Barabási, Collective Response of Human Populations to Large-Scale Emergencies, *PLoS ONE*, **6**, 17680 (2011).

[13] *The BBC Executive's Review and Assessment* (The British Broadcasting Corporation, London, United Kingdom, 2012).

[14] A. Briggs, *The History of Broadcasting in the United Kingdom: Volume II: The Golden Age of Wireless* (Oxford University Press, 1995).

[15] M. Thompson, The World Service can survive these cuts (2011). http://www.telegraph.co.uk/culture/tvandradio/bbc/8281797/The-World-Service-can-survive-these-cuts.html.

[16] T. Shaw, *Eden, Suez and the Mass Media: Propanganda and Persuasion During the Suez Crisis* (I. B. Tauris & Co. Ltd, 1996).

[17] Analysis: BBC's Voice in Europe (2005). http://news.bbc.co.uk/2/hi/europe/4375652.stm Accessed July 15, 2013.

[18] B. News, Años del servicio latinoamericano de la BBC (2000). http://www.bbc.co.uk/spanish/specials/1721_bbcmundoanivers/index.shtml Accessed July 15, 2013.

[19] Twitter, Numbers (2011). http://blog.twitter.com/2011/03/numbers.html Accessed October 22nd, 2011.

[20] J. Akshay, S. Xiaodan, F. Tim, T. Belle, Why We Twitter: Understanding Microblogging Usage and Communities, Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007 (2007).

[21] S. Martin, W. M. Brown, R. Klavans, K. W. Boyack, Openord: an open-source toolbox for large graph layout, Conference on Visualization and Data Analysis, **7868**, 786806 (2011).

[22] T. Hastie, R. Tibshirani, J. Friedman, J. Franklin, *The elements of statistical learning: data mining, inference and prediction* (Springer Verlag, 2011).

[23] BBC World Service in Government Funding Cut (June 2013). http://www.bbc.co.uk/news/entertainment-arts-22853598 Accessed July 15, 2013.

[24] BBC World Service to Lose 73 Jobs (October 2012). http://www.bbc.co.uk/news/entertainment-arts-19988316 Accessed July 15, 2013.

[25] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, International AAAI Conference on Weblogs and Social Media (2009).

[26] Geonames Web Services. http://www.geonames.org/ Accessed July 15, 2013.

[27] A. Kao, S. Poteet, *Natural language processing and text mining* (Springer Verlag, 2007).